



Research article

Stochastic volatility modeling of high-frequency CSI 300 index and dynamic jump prediction driven by machine learning

Xianfei Hui¹, Baiqing Sun¹, Indranil SenGupta², Yan Zhou^{1,*} and Hui Jiang³

¹ School of Management, Harbin Institute of Technology, Harbin 150001, China

² Department of Mathematics, North Dakota State University, Fargo ND 58108-6050, USA

³ College of Management and Economics, Tianjin University, Tianjin 300072, China

* **Correspondence:** Email: zyhittxzz@126.com.

Abstract: This paper models stochastic process of price time series of *CSI 300* index in Chinese financial market, analyzes volatility characteristics of intraday high-frequency price data. In the new generalized Barndorff-Nielsen and Shephard model, the lag caused by asynchrony of market information and market microstructure noises are considered, and the problem of lack of long-term dependence is solved. To speed up the valuation process, several machine learning and deep learning algorithms are used to estimate parameter and evaluate forecast results. Tracking historical jumps of different magnitudes offers promising avenues for simulating dynamic price processes and predicting future jumps. Numerical results show that the deterministic component of stochastic volatility processes would always be captured over short and longer-term windows. Research finding could be suitable for influence investors and regulators interested in predicting market dynamics based on high-frequency realized volatility.

Keywords: stochastic volatility modeling; jump; Lévy process; high-frequency data; machine learning and deep learning

1. Introduction

As we all know, financial fluctuations may come not only from the financial system itself, but also from other aspects of social and economic life. For example, COVID-19, has caused frequent and violent fluctuations in global financial markets [1, 2]. In the post-COVID-19 era, affected by internal and external factors in the market, the price of financial assets has been unstable during the first half of 2021. Facing a world with more dynamic economic situation, enterprises and research circles are realising the importance of the challenges and opportunities presented by financial fluctuations. The volatility of financial assets, which is the intensity of changes in the rate of return of financial

assets over a period of time, is unobservable [3]. The measurement of volatility, which describes the potential deviation from the expected value, is the core issue in the study of financial volatility. The accurate prediction of financial volatility is the key factor for successful financial asset pricing [4], economic forecasting [5], risk management [6], portfolio optimization [7], and quantitative investment [8]. Volatility Analysis of financial time series is a practical method to study the law of volatility and estimate volatility.

An effective way to fit dynamic asset changes is stochastic volatility modeling in the research of financial quantification. There are a lot of derivative pricing models that could be used to model stock prices. He et al. (2021) [9] proposed a new stochastic volatility model to provide a better fit to real data and showed numerically the validity by comparing the results with the Monte Carlo simulation results. He et al. (2021) [10] used the FMLS (finite moment log-stable) model with the stochastic volatility to analyse the effect of both jumps and stochastic volatility. The numerical experiments that it is effective and converges very rapidly. Most interestingly, new progress has been made in the field of volatility estimation under high frequency environment in recent years. A large amount of literature focuses on the three directions of high-frequency volatility estimation, namely, model establishment [11], model evaluation [12], and model application [13]. Asset price process [14] (continuous, finite jump or Lévy jump) and data characteristics [15] (whether there is microstructure noise and whether regular sampling is implemented) are the two main contents in the field of volatility estimation under high frequency environment.

Jump, excessive fluctuation of asset price in a certain period of time, is one of the key issues in asset price dynamics research. Theoretically, when there is no jump in asset price, the realized fluctuation is an unbiased and consistent estimation of potential fluctuation. However, the jump phenomenon of price volatility in the capital market is widespread. The jump leads to consistent overestimation of continuity fluctuations, causes realized volatility and realized range volatility to no longer be an unbiased and consistent estimate of potential volatility. In response to this jump phenomenon of asset price fluctuations, estimation of realized bipower variation, which was first proposed by Barndorff-Nielsen and Shephard, was used to decompose realized fluctuations into continuous fluctuations and jump fluctuations [16].

The Barndorff-Nielsen and Shephard model (BN-S) model [17], which is used to describe the random behavior of price process in the research field of non-parametric methods for high-frequency time series, is a popular stochastic volatility model with a Lévy process as driving factor of financial asset price. From academic points of view, the classic BN-S model has many attractive properties. But its theoretical framework is not completely satisfied in many application scenarios. Problems such as lack of long-term dependence may lead to the failure of the model in use. Recently, a variety of improvement schemes to the basic model are proposed, generalized BN-S models are constructed, and multiple dimensional applications, such as jump capture [18], pricing [19], and risk management [20], are implemented in the process of random fluctuations in asset prices. Roberts et al. (2020) [21] computed infinitesimal generators of generalized BN-S model driven by Lévy processes, and captured large or small jumps from each sensor receives data. SenGupta (2014) [22] derived the price expression for the put asian options in financial market when the asset is driven by the generalized BN-S model with stochastic volatility. Arai et al. (2017) [23] built generalized BN-S model under many additional conditions to obtain locally risk-minimizing strategies for call and put options.

Artificial intelligence in big data environment provides new tools for financial research and

enriches the previous research on volatility estimation [24]. Over the past few years, data processing classifiers based on machine learning and deep learning have always shown excellent performance in the field of financial prediction [25]. Machine learning [26] could be used to deal with the nonlinear relationship and interaction between variables, which would effectively solve the collinearity problem between variables. The nonlinearity and randomness in financial time series could be captured by machine learning, so machine learning algorithms show bigger prediction ability than linear models. Monfared et al. (2014) [27] used machine learning algorithms to predict the volatility of the National Association of Securities Dealers Automated Quotations index in the United States. The results show that machine learning could improve the prediction ability. Compared with machine learning, deep learning [28] has stronger optimization capabilities and more advantages when dealing with big data sets. Recurrent neural network (RNN) is widely used because of its strong learning capability, stability and simulation ability for volatility prediction. Liu (2019) [29] deployed RNN for predicting the volatility of the S&P 500 index, and found that RNN has stronger prediction ability than Generalized AutoRegressive Conditional Heteroskedasticity (GARCH). Fischer et al. (2018) [30] used Long Short Term Memory (LSTM) to predict out-of-sample directional movements for the S&P 500, and found that the performance of the stock portfolio constructed by this method is better than that of other linear models. The contribution of this paper is that the proposed approach through the use of machine learning algorithms enhances the forecasting ability of high-frequency *CSI 300* index volatility, which is essential for dynamic jump prediction in hedging, and purposes of arbitrage, and that the new model solves the problem of the classical BN-S model and enhances the forecasting efficiency of high-frequency *CSI 300* index volatility by fusing machine learning and deep learning algorithms with a generalized BN-S model.

In the research of asset volatility under random uncertain environment, the classic BN-S model including a single OU process is often constructed in the previous literature. However, the model will fail in application due to the lack of long-term dependence. Some existing studies solve this problem by superimposing the OU process, but the actual economic significance of different stochastic processes is less considered. The generalized BN-S model is used to study the volatility of daily sampled commodity prices by many authors in the US and European markets. There are fewer relevant studies on the Asia-Pacific market, and fewer relevant studies using the minute sampling frequency.

As one of the fastest-growing markets in the world, the Asia-Pacific securities market has attracted more and more attention [31]. The *CSI 300* index covers most of the domestic market value of China (the largest economy in the Asia-Pacific region), and reflects the market's mainstream investment returns and changes in the trader structure. The use of samples with high sampling frequency in the day can retain more market information and discover more detailed fluctuation characteristics caused by the impact of various information on the market. This research focuses on the price dynamics of the *CSI 300* index with high sampling frequency, and uses the generalized BN-S model to quantitatively analyze the volatility process of financial time series to capture the deterministic component of the random process of price fluctuations. The impact of overnight information [32] on the market is avoided in data preprocessing. Samples with high sampling frequency in the day are used to retain market information to a greater extent and discover more volatility characteristics caused by abnormal information shocks on the market.

Our approach to exploring stochastic process of high-frequency asset price dynamics has several advantages. First, the certainty element (θ) in the new model helps us freely fit stock index prices

and dynamic volatility in a correlated but different way. Because the superposition of Lévy process is considered, it can solve the problem that the classical BN-S model does not have enough dependence for a long time. Many previous studies believe that future volatility is unobservable and completely random. It is noteworthy that, in [33], a generalized BN-S model with the superposition of Lévy process extracted deterministic component in daily random volatility of crude oil out of low frequency price time series. Is there a deterministic component in random volatility that can improve the accuracy of volatility estimation? This paper studies the extension and application of the BN-S model in high-frequency financial time series field. The high-frequency data incorporate more market information than the daily price. Model description shows that the generalized new BN-S model is more superior in long-term volatility fitting and forecasting. Numerical results show that the deterministic component of stochastic volatility processes would always be captured over short and longer-term windows.

Second, the characteristics of information transmission and high-frequency data in the real financial market are fully considered in this paper. The new model realize the estimation of delay parameter (b) in the case of the jump in volatility caused by sluggish market response is not synchronized with the jump of asset price. Unlike low-frequency random volatility models, high-frequency random volatility models need to consider the impact of market microstructure noises based on data characteristics. Overnight information is another key factor affecting the accuracy of prediction results in high-frequency data processing. The impacts of market microstructure noises and overnight information are also considered in data analysis. The numerical results on *CSI 300* index show the effectiveness of the generalized BN-S model when it is used for high-frequency data with noise.

Third, the new model supports the cooperation of machine learning and stochastic volatility models, applies big data techniques to feature learning and parameter estimation on empirical datasets of stock index prices, and captures the deterministic components of the intraday price volatility of *CSI 300* stock index. It is easy to estimate the dynamic deterministic parameter with the help of machine learning algorithms and deep learning algorithms. It shows the application of data science in obtaining “deterministic components” from processes that are generally considered to be completely random. 11 kinds of machine learning and deep learning algorithms (Logistic regression, Support vector machine, K-nearest neighbors, K-means, Naive bayes, Gradient boost, Decision tree, Random forest, Neural network, LSTM and LSTM network with batch) are used to process data to estimate parameters. We believe that more algorithms could further prove the validity of the model and ensure the accuracy of the results. In general, the results offer promising avenues for simulating dynamic price processes and predicting future jumps. Research finding could be suitable for influence investors and regulators interested in predicting market dynamics based on high-frequency realized volatility.

The paper is organized as follows. In Section 2, the generalized BN-S model is introduced. In Section 3, the high-frequency *CSI 300* stock index price data is selected as the research sample, the high-frequency financial time series are preprocessed, the descriptive statistical characteristics of the data set are obtained, and the distribution of price fluctuations is analysed. Based on the research results obtained in Section 3, a deterministic component out of high-frequency price stochastic processes is derived by using machine learning and deep learning algorithms to realize parameter analysis and estimation in Section 4. In Section 5, a brief conclusion is provided.

2. Barndorff-Nielsen and Shephard model

Financial time series of different assets share many common features (heavy tailed distributions of log-returns, aggregational gaussianity, quasi long-range dependence). Many of these facts are successfully captured by stochastic models with Lévy processes. Lévy processes can be used to characterize the dynamic changes of the time series of financial asset prices with jump processes. BN-S model, which is a widely used stochastic model with Lévy processes, is used to describe the stochastic behavior of random time series in the research field of nonparametric methods of high-frequency time series. A brief introduction to this model is given as follows.

Consider a frictionless financial market in which a risk-free asset with a constant rate of return r and a stock are traded on a fixed horizon date T . The classical BN-S model assumes that the price process of a stock (or, a commodity) $S = (S_t)_{t \geq 0}$, which is defined in a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{0 \leq t \leq T}, \mathbb{P})$, is given by

$$S_t = S_0 \exp(X_t), \quad (2.1)$$

the log-return X_t is given by

$$dX_t = (\mu + \beta\sigma_t^2)dt + \sigma_t dW_t + \rho dZ_{\lambda t}, \quad (2.2)$$

where σ_t is the volatility at time t , the parameters $\mu, \beta, \rho \in \mathbb{R}$, and $\rho \leq 0$. The variance process is given by

$$d\sigma_t^2 = -\lambda\sigma_t^2 dt + dZ_{\lambda t}, \quad \sigma_0^2 > 0, \quad (2.3)$$

where $\lambda > 0$.

With respect to the probability measure \mathbb{P} , the process $W = (W_t)$ is a standard Brownian motion. Observe that the Ornstein-Uhlenbeck process in this model is driven by an incremental Lévy process, which is a random process of positive mean recovery. The process $Z = (Z_{\lambda t})$ is the subordinator (also known as *background driving Lévy process* or “BDLP”). The processes W and Z are independent of each other. Also, (\mathcal{F}_t) is a conventional augmentation of the filtering produced by (W, Z) .

Solving (2.3) we obtain

$$\sigma_t^2 = e^{-\lambda t} + \int_0^t e^{-\lambda(t-s)} dZ_{\lambda s}. \quad (2.4)$$

Clearly, the process $\sigma^2 = (\sigma_t^2)$ is strictly positive. The classical BN-S model has excellent performance in describing the dynamic characteristic response mode of stable asset prices in a short time. It is commonly used to capture some stylized features of time series observed in financial markets, such as semiheavy tails, aggregational Gaussianity, quasi long range dependency and self-similarity.

However, the results and theoretical framework of the classical BN-S model are not completely satisfactory in empirical situations. There are several problems in the classical model, which may make the model difficult to use in practice. For example, empirical results show that the jump in volatility is positively correlated with stock or commodity prices. But the jump phenomenon in volatility does not usually occur at the same time with the change in price because of the lag in market response. The topic of delayed response in the financial market has been studied in papers such as [34]. On the other hand, the study in [35] handles this problem with a delayed price formula, where the price volatility obeys the form $\sigma(S_t - b)$, for some delay parameter $b > 0$. However, the parameter b is also stochastic, and this makes the resulting model unnecessarily involved.

Furthermore, the classical BN-S model does not have long-term dependence property. Consequently, due to the high sequence correlation between hidden variables and parameters, for the analysis of the empirical data based on this model the convergence rate is slow. The classical BN-S model contains a single BDLP, which makes the logarithmic return, volatility and variance in the model completely dependent on each other. When the model is used over a long period of time, this absolute correlation may lead to inaccurate results. As a result, the model encounters serious failures in volatility estimation.

These problems are overcome in a new generalized model. It is clear that for the long-term implementation of the classical BN-S model, a single Lévy subordination is obviously ineffective. The research results [36] show that the superposition of Ornstein-Uhlenbeck (OU) type processes can achieve long-range dependence. The superposition of Lévy subordinations successfully fits the asynchronous changes from price and volatility in an interrelated but independent way. Referencing the previous research results [33], the structure of a generalized BN-S model will be introduced as follows.

The key point of our research is to capture the deterministic components out of high-frequency price stochastic processes. As proposed in [33], suppose Z_t and Z_t^* , with same (finite) variance, are two independent Lévy subordinators. There exists a Lévy subordinate $\bar{Z}_{\lambda t}$ independent of W , such that

$$d\bar{Z}_{\lambda t} = \rho' dZ_{\lambda t} + \sqrt{1 - \rho'^2} dZ_{\lambda t}^*, \quad 0 \leq \rho' \leq 1. \quad (2.5)$$

For $0 \leq \rho' \leq 1$, Z and \bar{Z} are positively correlated Lévy subordinators. Assume that the dynamics of S_t are given by (2.1) and (2.2), where σ_t is given by

$$d\sigma_t^2 = -\lambda\sigma_t^2 dt + d\bar{Z}_{\lambda t}, \quad \sigma_0^2 > 0. \quad (2.6)$$

In (2.6), the OU process $\bar{Z} = (\bar{Z}_{\lambda t})$ is related to the corresponding Z in (2.3) and is also independent of W .

In the following study, delay parameter b and the long range dependence property of model are considered. As shown in [33], the price $S = (S_t)_{t \geq 0}$ on some risk-neutral filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{0 \leq t \leq T}, \mathbb{P})$ is modeled by (2.1). And the convex combination of two independent subordinators Z and $Z^{(b)}$ would be implemented to express the dynamics of X_t in (2.2) by

$$dX_t = (\mu + \beta\sigma_t^2)dt + \sigma_t dW_t + \rho((1 - \theta)dZ_{\lambda t} + \theta dZ_{\lambda t}^{(b)}), \quad (2.7)$$

where $0 \leq \theta \leq 1$, θ is a deterministic parameter. At time t , $\lambda > 0$ is the proportional parameter. $Z_{\lambda t}$ and $Z_{\lambda t}^{(b)}$ are independent Lévy processes. Compared to $Z_{\lambda t}$, the process $Z_{\lambda t}^{(b)}$ corresponds to the greater Lévy intensity. For instance, if the Lévy densities of Z and $Z^{(b)}$ are given by $v_1 a e^{-ax}$ and $v_2 a e^{-ax}$, respectively (for $a > 0$, $v_1 > 0$, $v_2 > 0$, and $x > 0$), then $v_2 > v_1$. Also, in (2.8) the processes W , Z and $Z^{(b)}$ are independent, and (\mathcal{F}_t) is the usual augmentation of the filtration generated by $(W; Z; Z^{(b)})$.

In fact, the real financial market often deviates from the efficient market. The short-term changes of asset prices are usually affected by various market factors (e.g., transaction cost, asymmetric information of traders, etc.), which makes the observed prices deviate from the real prices. This deviation is called market microstructure noise. The observed asset price process, generally bought in noise, could be divided into two components

$$\bar{X}_t = X_t + \varepsilon_t, \quad (2.8)$$

where \bar{X}_t is the observed log price, X_t is the factual true value of log price, and ε_t is market microstructure noise. The asset log price processes and market microstructure noise are independent of each other. ε_t is independent identically distributed, $E\varepsilon_t = 0$ and $Var(\varepsilon_t) = E\varepsilon_t^2$. Market microstructure noise has a significant impact on estimation of high-frequency covariance matrix. Market microstructure noise has a significant impact on estimation of high-frequency covariance matrix. For high sampling frequency, the estimated value of realized covariance is in fact not the covariance of assets price, but the covariance of market microstructure noise. The asset price process S_t in (2.1) is given by

$$S_t = S_0 \exp(\bar{X}_t - \varepsilon_t). \quad (2.9)$$

The convex combination of two independent subordinators $Z_{\lambda t}$ and $Z_{\lambda t}^{(b)}$ would be implemented to expressed the dynamics of log price in (2.7) is given by

$$d(\bar{X}_t - \varepsilon) = (\mu + \beta\sigma_t^2)dt + \sigma_t dW_t + \rho((1 - \theta)dZ_{\lambda t} + \theta dZ_{\lambda t}^{(b)}). \quad (2.10)$$

The variance process in (2.3) in this case is given by

$$d\sigma_t^2 = -\lambda\sigma_t^2 dt + (1 - \theta')dZ_{\lambda t} + \theta' dZ_{\lambda t}^{(b)}, \sigma_0^2 > 0, \quad (2.11)$$

where $\theta' \in [0, 1]$ is deterministic parameter. For simplicity, assume $\theta = \theta'$ for the rest of this paper. The sum of $(1 - \theta')dZ_{\lambda t}$ and $\theta' dZ_{\lambda t}^{(b)}$, is a Lévy process, which is positively correlated with $Z_{\lambda t}$ and $Z_{\lambda t}^{(b)}$.

After a simple calculation, the solution of (2.11) can be explicitly written as

$$\sigma_t^2 = e^{-\lambda t} \sigma_0^2 + \int_0^t e^{-\lambda(t-s)} ((1 - \theta')dZ_{\lambda s} + \theta' dZ_{\lambda s}^{(b)}). \quad (2.12)$$

This enforces positivity of σ_t^2 . Thus, the process σ_t^2 is strictly positive and it is bounded from below by the deterministic function $e^{-\lambda t} \sigma_0^2$. The instantaneous variance of log returns is given by

$$(\sigma_t^2 + \rho^2(1 - \theta)^2 \lambda Var[Z_1] + \rho^2 \theta^2 \lambda Var[Z_1^{(b)}])dt.$$

The short-range-dependence problem of the classical BN-S model can be improved in the new model. The dynamics given by the new model incorporates a long-range dependence. Assume that J_Z is a jump measure related to the subordinate Z of the Lévy process, $J_Z^{(b)}$ corresponds to the subordinate $Z^{(b)}$ of the Lévy process, and $J(s) = \int_0^s \int_{\mathbb{R}_+} J_Z(\lambda d\tau, dy)$, $J(s)^{(b)} = \int_0^s \int_{\mathbb{R}_+} J_Z^{(b)}(\lambda d\tau, dy)$. Considering the logarithmic regression of the classical BN-S model and the new model, the covariances of X_t and X_s are given by

$$Corr(X_t, X_s) = \frac{\int_0^s \sigma_\tau^2 d\tau + \rho^2 J(s)}{\sqrt{(\int_0^t \sigma_\tau^2 d\tau + t\rho^2 \lambda Var(Z_1))(\int_0^s \sigma_\tau^2 d\tau + s\rho^2 \lambda Var(Z_1))}}, \quad t > s \quad (2.13)$$

and

$$Corr(X_t, X_s) = \frac{\int_0^s \sigma_\tau^2 d\tau + \rho^2(1 - \theta)^2 J(s) + \rho^2 \theta^2 J(s)^{(b)}}{\sqrt{\alpha(t)\alpha(s)}}, \quad t > s \quad (2.14)$$

respectively, where $\alpha(v) = \int_0^v \sigma_\tau^2 d\tau + v\rho^2 \lambda((1-\theta)^2 \text{Var}(Z_1) + \theta^2 \text{Var}(Z_1^{(b)}))$. When s takes a fixed value, for the classical BN-S model, $\text{Corr}(X_t, X_s)$ rapidly becomes smaller with the increase of t . Such attenuation may cause the failure of the classical model in applications with a long time span. It can be seen that the BN-S model, when used to fit the random fluctuation process of risky assets, may get inaccurate fluctuation simulation results, affected by the change of the time parameter t .

On the other hand, variance of the log-returns X_t and X_s (as shown in (2.14)) are

$$\int_0^t \sigma_\tau^2 d\tau + v\rho^2 \lambda((1-\theta)^2 \text{Var}(Z_1) + \theta^2 \text{Var}(Z_1^{(b)}))$$

and

$$\int_0^s \sigma_\tau^2 d\tau + v\rho^2 \lambda((1-\theta)^2 \text{Var}(Z_1) + \theta^2 \text{Var}(Z_1^{(b)}))$$

respectively.

Affected by the value of the parameter θ , $\text{Corr}(X_t, X_s)$ will never become “too small”. Because the value of t must have an upper limit when s takes a fixed value. This is the main difference between the results of (2.13) and (2.14). It can be clearly seen from the results that the generalized new BN-S model incorporates a long-range dependence and provides more accurate characteristics for the dynamic volatility analysis in the asset price process. The new model can accurately capture the essential characteristics of the random fluctuation process of financial time series.

In addition, compared to the classical BN-S model, the parameter θ in the new model can help us freely fit asset prices and volatility in a correlated but different way. For dynamic prices, the jump is not completely random, and there is a deterministic element (θ) that can be implemented to be effectively applied to the new BN-S model in a longer time. The large fluctuations can be captured in the future from historical experience data ($\theta = 1$), and the initial Lévy subordinate function $Z_{\lambda t}$ could be converted into a stronger Lévy subordinate function $Z_{\lambda t}^{(b)}$ to correspond to the large fluctuations. If there is no big jump apprehended for the upcoming time, the Lévy subordinate function $Z_{\lambda t}^{(b)}$ could be converted into Lévy subordinate function $Z_{\lambda t}$ based on historical data ($\theta = 0$) by using machine learning and deep learning algorithms.

Obviously, an important challenge in the application of the new model is to obtain an estimate of the value of a deterministic component of the empirical data. In this paper, the new model is used to analyze the price dynamics of high frequency *CSI 300* stock index. Several machine learning algorithms and deep learning algorithms are implemented to forecast parameter θ .

3. Data

3.1. Sources of data

The *CSI 300* index is always considered to have strong market representation. It covers most of China’s domestic circulating market value and reflects the overall trend of China’s Shanghai and Shenzhen markets. In particular, its constituent stocks include many mainstream investment stocks with market representation, liquidity and trading activity. So it is often used to study the returns of mainstream investments and changes in financial price fluctuations in the market.

The main purpose of this paper is to explore the volatility characteristics of intra-day high-frequency price data, and then to study the quantitative indicators in the process of random fluctuations in financial

time series. The generality and extensiveness of the application of the new model in the previous section are considered, and the *CSI 300* index price is considered as the empirical data of analysis. The corresponding intra-day high-frequency data is selected as the research sample. It is conducive to maximally retain market information to select research samples with a higher sampling frequency. The intraday closing price data of the *CSI 300* index on consecutive trading days from January 1, 2021 to June 30, 2021 is considered as a sample. The sampling frequency of this sample is 1 minute. The data set contains a total of 28,320 observations in 118 consecutive trading days (Data source: Wind*).

The fluctuation curve of the historical data over time is shown in Figure 1.

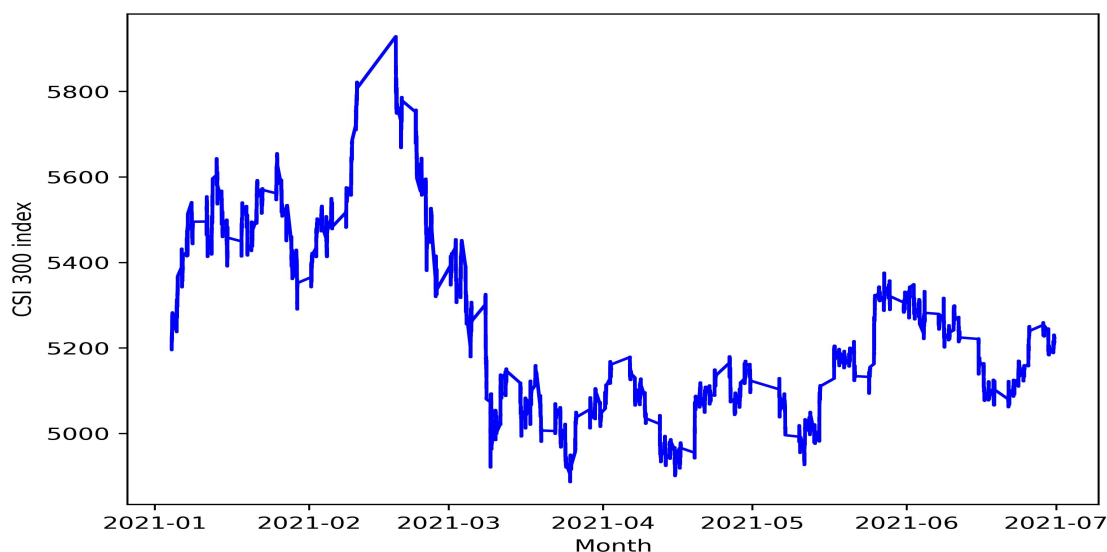


Figure 1. Curve of closing price per minute.

It is necessary to discuss the data distribution characteristics of intra-day price changes and yield fluctuations, which help us to explore the basic laws of *CSI 300* stock index time series fluctuations. In order to study the change trend and distribution characteristics of *CSI 300* stock index over time in different time intervals, an intuitive way is chosen to visually analyze the data structure. Figure 2 shows the moving average curve of the *CSI 300* index under different time spans.

Normally, the trading hours of the *CSI 300* index are each working day in 9 : 30 – 11 : 30 and 13 : 00 – 15 : 00, Beijing time (the effective trading time per day is a total of 4 hours). So four time spans of 1 minute, 30 minutes, 120 minutes (half a day) and 240 minutes (1 day) are chosen to observe the data set. In Figure 2, blue represents the price change curve per minute, and red represents daily price fluctuations. It can be clearly seen that the general trends of the two curves are similar, but there are fewer repetitions. The blue curve fluctuates more sharply than the red one, which shows that the high-frequency data during the day contains more market information than the closing price. The yellow line (representing the price change every 30 minutes) and the blue line overlap more severely than the red line. The green curve, which represents price changes every 2 hours, is more stable than the yellow line. These results are also considered to confirm the above view, that is, the data set at a higher sampling frequency is more effective for us to find the realized volatility estimator. Compared

*In the field of financial data, Wind has built a complete and accurate large-scale financial engineering and financial data center on financial and securities data in China. Uniform Resource Locator of wind is: <https://www.wind.com>.

with previous studies, the data set used in this paper has more advantages.

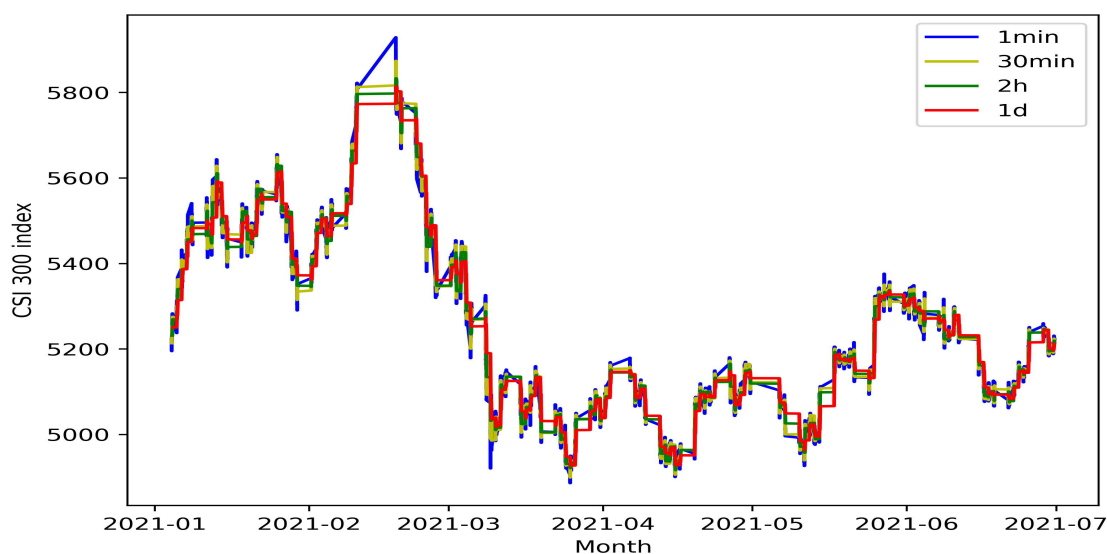


Figure 2. Moving average for *CSI 300* index.

3.2. Data preprocessing

A lot of misleading information exists in the unprocessed empirical data for various reasons. Unprocessed data is used directly, which may lead to undesirable results such as a decrease in the prediction accuracy of the time series. Therefore, the observed samples should be filtered before doing data analysis.

Compared with the fluctuations in daily stock index yields and trading volume, the impact of overnight information on the market should not be ignored. Most of these price changes on overnight information are concentrated within ten minutes of the opening. In other words, price fluctuations within ten minutes of the opening could not represent changes in stock index fluctuations throughout the trading day. In order to avoid shocking intra-day fluctuations and causing abnormal data (such as high kurtosis, increasing outliers, etc.), the data within 10 minutes of the opening of the daily observation sample should be excluded. After the overnight information was digested by the market, the empirical data would reflect the daily operation of the *CSI 300* index price more accurately. In addition, we remove the outliers and zero-value data from the observed data to keep the observed sample data tidy.

After preprocessing the empirical data according to the above filter conditions, the usable sample data are filtered out (28,081 observations in total). The rejection rate of sample data is 0.84%. It shows that the observed samples have both liquidity and validity, and the intra-day high-frequency price information is effectively stored in the empirical data.

3.3. Descriptive analysis

For the convenience of research, the observed samples are divided and numbered in chronological order. For example, Sample 1, Sample 2, ..., Sample 6, represent the samples from January through

June of 2021. Variance of realized volatility and realized covariance generally decrease with increased sampling frequency. Market microstructure noise has a significant impact on estimation of high-frequency covariance. Therefore, in the realized covariance estimation of financial high-frequency data, it is necessary to balance the variance and bias to select the appropriate sampling frequency. As proposed in [37], the sampling frequency of 5 minutes is selected in this paper. The statistical descriptions of the samples of high-frequency *CSI 300* index intraday prices are given in Table 1.

Table 1. Statistical description of *CSI 300* index high-frequency prices.

	Overall sample	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
Count	28081	940	705	1081	987	846	987
Mean	5245.46	5471.63	5563.75	5116.65	5067.71	5151.67	5202.71
Median	5187.84	5485.11	5523.59	5074.30	5087.45	5161.70	5221.97
Minimum	4886.40	5209.49	5319.68	4891.91	4902.15	4929.86	5065.91
Maximum	5908.34	5653.55	5894.03	5454.05	5177.60	5364.37	5343.28
Skewness	0.60	-0.68	0.40	0.94	-0.73	0.11	-0.21
Kurtosis	-0.50	0.23	-0.86	-0.17	-0.66	-1.19	-1.16

The fluctuation characteristics of *CSI 300* prices could be seen from the statistical results in Table 1. The highest price was 5908.34 in February, and the lowest price of 4886.4 appeared in March. Sample 2 has the least amount of data, but its mean and median are higher than those of other samples. It shows that the price in February is more advantageous compared with other months. The distribution of price data is not completely symmetrical. The skewness of the overall sample of Sample 1, Samples 4 and 6, are all less than zero. Their distributions have negative deviations, and the tail on the left is longer than the right. Because there are a few variables with small values, the left tail of the curve is dragged longer. In contrast to them, there are heavy-tailed distributions on the right side of Samples 2, 3 and 5 (the skewness of these three samples are all greater than 0). This phenomenon is most obvious in May, followed by June. The kurtosis of the observed samples are less than 3, which shows that the observed samples do not have leptokurtic characteristics. We believe this is related to sampling frequency. In the case of sampling frequency per minute, the kurtosis of the sample is less than that of normal distribution. The generalized BN-S model mentioned in Section 2 is suitable to discuss the above data characteristics, because Lévy processes in the model could be used to characterize the dynamic changes of the time series of financial asset prices with jump processes.

Figure 3 provides the difference in the distribution of intra-day high-frequency price samples of *CSI 300* in different time periods through a box plot. Compared with other samples, the prices in January and February are more advantageous, and the price fluctuation in February is also the largest.

A histogram of price distribution explains the dispersion and distribution of *CSI 300* in half a year in Figure 4. Obviously, the *CSI 300* index is the densest in the range of 5000–5200. Together with Figure 3, it could be seen that the prices from March to June are mostly within this range. It shows that the price fluctuations from January to February are likely to be more volatile, and the fluctuates in the smooth from March to June. The larger jumps we are concerned about are most likely to occur in Samples 1 and 2.

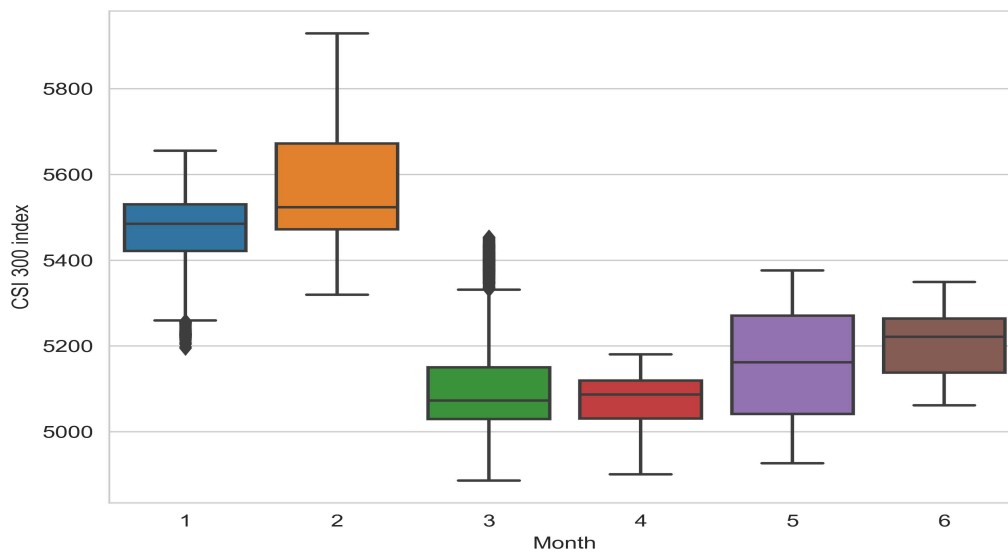


Figure 3. Daily boxplot for *CSI 300* index.

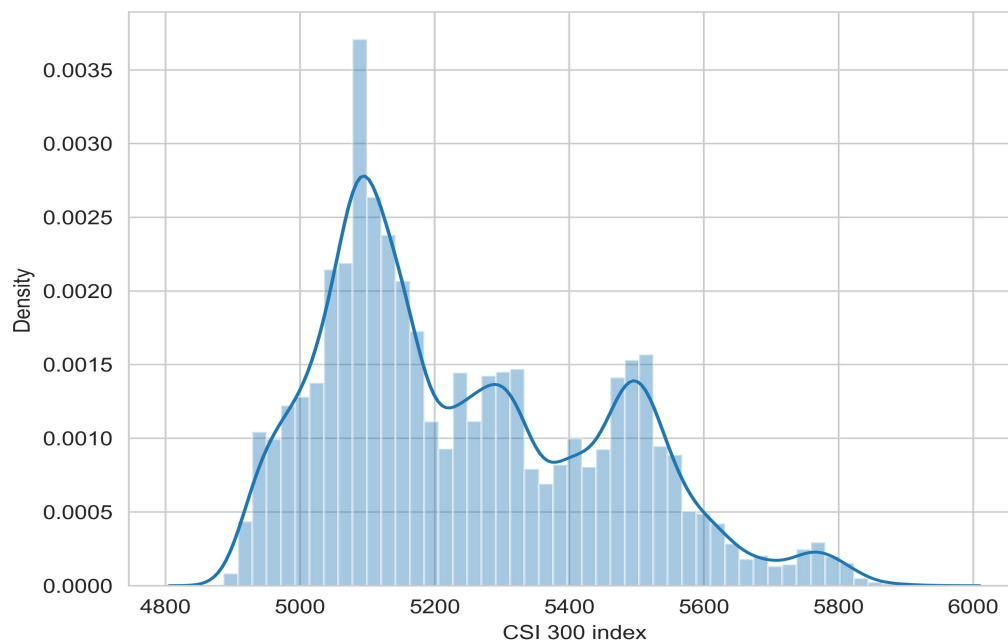


Figure 4. Distribution plot for *CSI 300* index.

Generally, the high-frequency price has a smaller range of changes than the low-frequency price in the same period of time. For example, the price change per minute is often smaller than the change every two minutes in the same upward trend. In order to observe small changes in high-frequency data, the value of the percentage of price change is more suitable to be used as observational data than price data in the analysis of the volatility distribution of high-frequency data.

The histogram of the *CSI 300* price change percentage is shown in Figure 5. It's seen that *CSI 300* price change statistics per minute, which do not follow the normal distribution, are mainly concentrated

in 0 (both positive and negative values exist). The graph is skewed to the right, which indicates that there are more rising empirical data than falling ones in the overall sample.

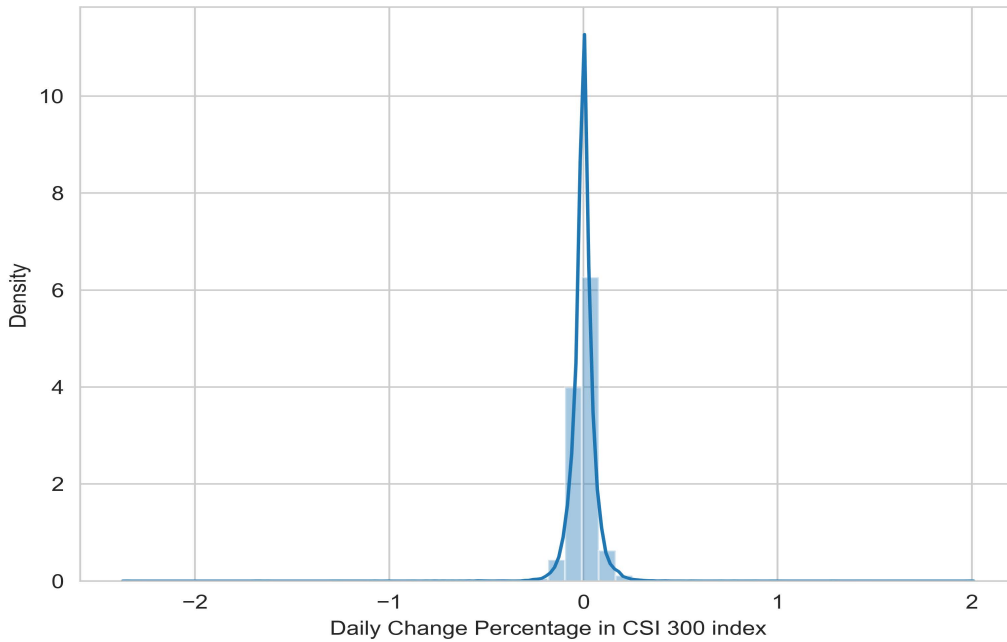


Figure 5. Histogram for daily change percentage in *CSI 300* index.

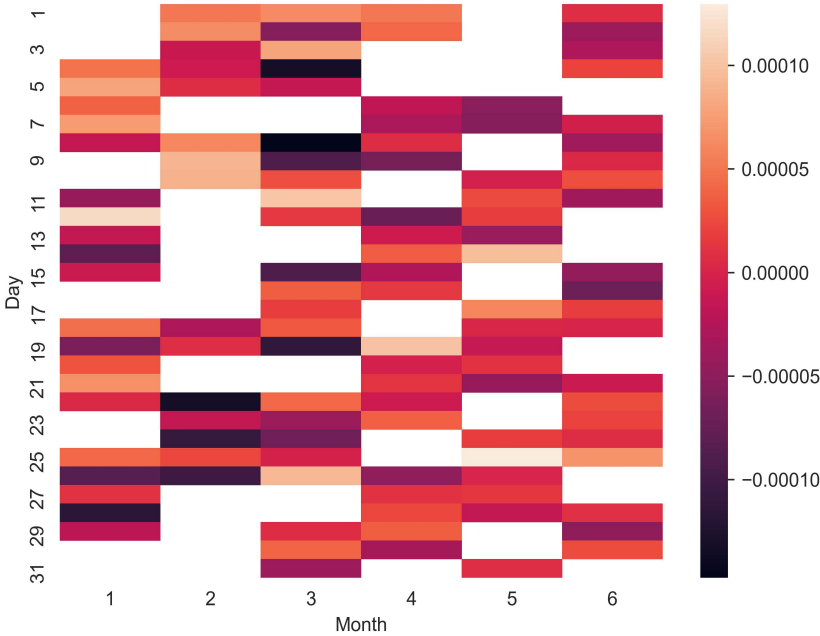


Figure 6. Heatmap for the realized volatility of high-frequency *CSI 300* index.

In order to explore the characteristics of the volatility change of the *CSI 300*, the realized volatility is described separately from the perspective of value distribution and change trend. Figure 6 is a

heatmap of the realized volatility with the sample month as the horizontal axis and the date as the vertical axis. Through the black and white areas in the figure, the realized volatility date with a volatility change of more than 1% could be identified. Obviously, the realized volatility of *CSI 300* experienced more frequent fluctuations in each month of the first half of the year. February contains the largest number of days with large fluctuations.

Figure 7 shows the trend of the realized volatility within half a year. As can be seen from the figure, the autocorrelation exists in the realized volatility data of the high-frequency *CSI 300* stock index (Volatility Clustering). The widest range of realized volatility changes occurred in February and March.

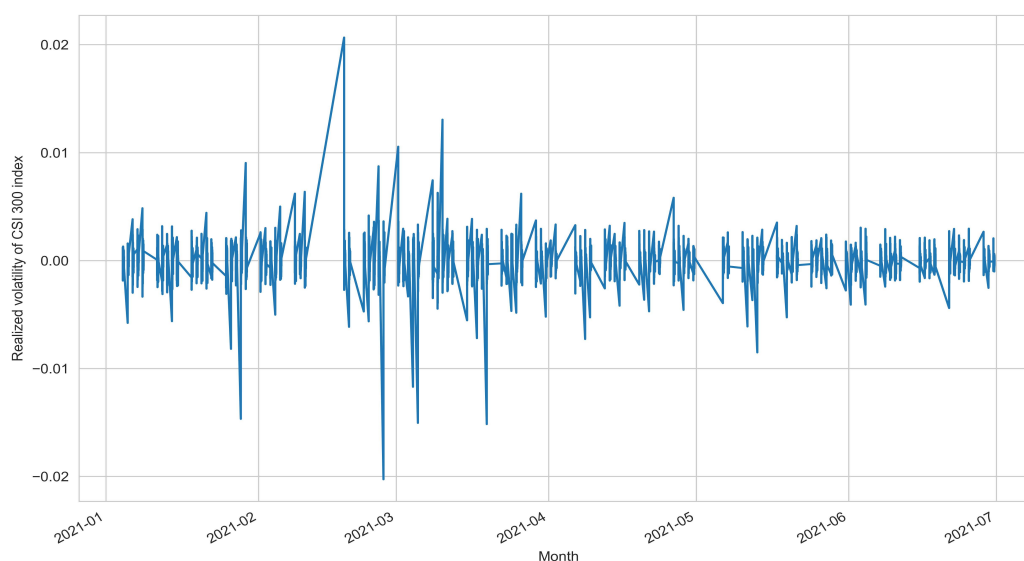


Figure 7. Line plot for the realized volatility of high-frequency *CSI 300* index.

Volatility jumps with different amplitudes and frequencies exist in each sample of the *CSI 300*. In the following section, the information of data characteristics shown in the above charts is used for learning and parameter estimation of empirical data.

4. Parameter analysis and estimation

By using the data analysis results in Section 3, the value of θ in the generalized new BN-S model in Section 2 is found, and the deterministic component in the random process of high-frequency price data fluctuations is captured in this section. In order to achieve the above goals, the classification problem based on the historical data set was created by implementing the following steps.

Step 1. Index the available historical price data and price change percentage data per minute of *CSI 300* in chronological order.

Step 2. According to the data fluctuation characteristics obtained in Section 3, create new data structures from historical data sets. Take the percentage of change in the closing price for 10 consecutive minutes as a subset of the rows, stacking layer by layer. Divide the empirical data according to the above rules to form a new *CSI 300* index price data matrix.

Step 3. Consider the volatility of the closing price per minute in the historical price data of the *CSI*

300 stock index, and determine the value of K to define the big jumps (large increases) in the high-frequency closing price fluctuation. Each time closing price is K lower than the price of the previous minute needs to be identified (for example, if $K = 0.1\%$, the date and time, when the closing price of *CSI 300* index is 0.1% lower than the previous minute's one, should be marked).

Step 4. Create a target column θ in the new data matrix and assign values. If there are at least two big jumps in the next 10 minutes, the parameter θ in the target column of the row is 1. Otherwise, the θ corresponding to the row is 0.

Step 5. Use several different machine learning algorithms and deep learning algorithms to learn from empirical data sets and estimate the value of θ . Substitute the obtained value of θ into (2.8) in Section 2, which means that the deterministic component of the *CSI 300* fluctuation random process is captured.

The variables involved in the above steps could be adjusted according to the characteristics of the data set. The adjustment rules could refer to the following reminders.

1) In step 2, if a multi-dimensional data structure is created by adjusting the division of data, the effectiveness of the result will be improved. In general, the more elements contained in a row subset, the more information is carried in the new matrix, and the accuracy of the result may be improved. At the same time, it also increases the workload of calculation and reduces the predictable time span.

2) In step 3, adjusting the value of K is believed to be an effective way to improve the results. Different values of K are suitable for different retrieval targets. Generally, the shorter the time, the smaller the change range of the observed data. A higher sampling frequency is often suitable for using a smaller value of K , which can identify more big jumps the same period.

3) In step 4, the value of θ is related to the number of the big jumps identified in a period.

In the same data set, the more subset elements selected in step 2, the more big jumps recognized in each row of the matrix, and the greater the possibility of $\theta = 1$. Setting the threshold for identifying the number of big jumps to be small will lead to a high probability of $\theta = 1$, and a low probability of $\theta = 0$.

The above steps could be used to calculate θ with reasonable accuracy to prove that these steps are feasible.

Various machine learning and deep learning algorithms are used for the new matrix formed in step 2 on Python. The input are the subset elements in each row of the matrix, and the output is the value of θ (0 or 1) in the target column of the new matrix mentioned in step 3. The algorithms we used are (A) Logistic regression, (B) Support vector machine, (C) K-nearest neighbors, (D) K-means, (E) Naive bayes, (F) Gradient boost, (G) Decision tree, (H) Random forest, (I) Neural network, (J) Long and short-term memory neural network (LSTM) and (K) LSTM network with batch normalizer. Specifically, logistic regression realize the estimation of θ through maximum likelihood estimation. Support vector machine solves the maximum-margin hyperplane of the sample to binary classify the θ value. K-nearest neighbors predict the classification of θ by identifying whether there are jumps during the adjacent time intervals. The evaluation of θ could be realized by calculating the distance of the object and the cluster center in K-means algorithm. Naive bayes estimates the classification of θ based on bayesian theorem and attribute conditional independence assumption. Gradient boost realize the estimation of θ by training the new weak classifier on the basis of negative gradient information of the loss function and accumulating the weak classifiers. After the decision tree is constructed, a reasonable θ is found through pruning. Random forest contains many decision trees. Two hidden and

output layers are built in the neural network in algorithm (I). If the output probability of the softmax activation function corresponding to $\theta = 1$ is greater than 0.3, the parameter θ is 1. In algorithm (J), Long short-term memory (LSTM) neural network realize forward calculation and back propagation through forward method and backward method. In algorithm (K), Batch normalizer has a positive effect on the training speed of LSTM.

Referring to the data characteristics analyzed in Section 3, four sets of training set and test set dates are selected from the empirical data set to run the classifiers, and find the corresponding index in step 1. The date selection is shown in Table 2.

Table 2. Time and index of the classifiers.

	Training Time (Index)	Testing Time (Index)
T1	01/01/2021 9:40:00 (0) to 02/10/2021 15:00:00 (1315)	02/18/2021 9:40:00 (1316) to 02/26/2021 15:00:00 (1644)
T2	01/01/2021 9:40:00 (0) to 02/26/2021 15:00:00 (1644)	03/01/2021 9:40:00 (1645) to 03/12/2021 15:00:00 (2114)
T3	04/01/2021 9:40:00 (2726) to 05/31/2021 15:00:00 (4558)	06/01/2021 9:40:00 (4559) to 06/30/2021 15:00:00 (5545)
T4	06/25/2021 9:40:00 (5358) to 06/29/2021 13:59:00 (5485)	06/29/2021 14:00:00 (5486) to 06/30/2021 15:00:00 (5545)

The fluctuation patterns of monthly data, weekly data, daily data, and intraday data are all considered in selecting test samples. The analysis results in Section 3 show that the average price of Sample 2 is the highest, and the volatility is the most intense in February. Therefore, the daily data of February (T1) where the maximum volatility is located could be selected as the estimated sample. After experiencing huge ups and downs, CSI 300 continued to fall in March, so the price changes in the first two weeks of Sample 3 deserve attention and related weekly data (T2) could be estimated. Monthly data forecast within a more stable range (T3) is also worthy of attention. It is also meaningful to estimate the parameters on intra-day historical data in the last five days in the data set (T4). The daily historical data in the last five days in the data set is selected to estimate the value of θ .

In step 5, it is worth noting that the estimation results of θ by using 6 algorithms are not necessarily the same. The prediction results of different machine learning and deep learning algorithms often have different accuracy. In order to avoid possible misjudgments and make the results more accurate, the prediction results of various algorithms are evaluated. “Support” refers to the number of responsive samples that appeared during the calculation process. “Precision” is used to express the accuracy rate in all the prediction results, where θ takes 1 or 0. It is defined as the ratio of the number of accurate predictions $\theta = 1$ ($\theta = 0$) to the number of all prediction results $\theta = 1$ ($\theta = 0$). “Recall” shows the efficiency that $\theta = 1$ ($\theta = 0$) is accurately predicted. It represents the ratio of the quantity accurately predicted $\theta = 1$ ($\theta = 0$) to the true quantity $\theta = 1$ ($\theta = 0$). The accuracy of parameter prediction results could be represented by the values of “precision” and “recall”. The harmonic average of “precision” and “recall” is considered suitable to show the predictive effect of different algorithms directly. Its value is indicated by “F1-score”.

Several machine learning algorithms and deep learning algorithms are used for the empirical data

set in the above time periods, and the results of the classification report of the accuracy evaluation of θ are recorded in Tables 3–6.

It can be seen that the number of “support” in the report results, not affected by different algorithms, is only related to the time window T. It shows the number of jumps could be accurately identified by the eleven algorithms we used for the empirical price data of *CSI 300 Index*. Comparing the number of “supports” ($T4 < T1 < T2 < T3$), we find that as the time window T grows, the more jumps would be identified. It’s just like we thought.

The classification results in Table 3 illustrate that when threshold of the asset return K for identifying jumps is 0.1, there’s a strong possibility that θ is equal to 1, while $\theta = 0$ is still a possibility, not a probability. And the results in Tables 4–6 have reached the same conclusion.

Table 3. Accuracy report about θ estimation in $T1$.

$T1$	precision	recall	f1-score	support	precision	recall	f1-score	support
	$\theta = 0$	$\theta = 0$	$\theta = 0$	$\theta = 0$	$\theta = 1$	$\theta = 1$	$\theta = 1$	$\theta = 1$
(A)	0.00	0.00	0.00	37	0.89	1.00	0.94	293
(B)	0.25	0.46	0.33	37	0.92	0.83	0.87	293
(C)	0.21	0.22	0.21	37	0.90	0.90	0.90	293
(D)	0.16	0.11	0.13	37	0.89	0.11	0.20	293
(E)	0.03	0.03	0.03	37	0.88	0.87	0.88	293
(F)	0.17	0.03	0.05	37	0.89	0.98	0.93	293
(G)	0.15	0.41	0.22	37	0.90	0.71	0.80	293
(H)	0.14	0.05	0.08	37	0.89	0.96	0.92	293
(I)	0.17	0.05	0.08	37	0.89	0.97	0.93	293
(J)	0.00	0.00	0.00	37	0.89	0.99	0.93	293
(K)	0.14	0.03	0.05	37	0.89	0.98	0.93	293

Table 4. Accuracy report about θ estimation in $T2$.

$T2$	precision	recall	f1-score	support	precision	recall	f1-score	support
	$\theta = 0$	$\theta = 0$	$\theta = 0$	$\theta = 0$	$\theta = 1$	$\theta = 1$	$\theta = 1$	$\theta = 1$
(A)	0.00	0.00	0.00	72	0.85	1.00	0.92	399
(B)	0.08	0.10	0.09	72	0.83	0.79	0.81	399
(C)	0.21	0.29	0.25	72	0.86	0.80	0.83	399
(D)	0.19	0.11	0.14	72	0.92	0.12	0.22	399
(E)	0.36	0.68	0.47	72	0.93	0.78	0.85	399
(F)	0.36	0.11	0.17	72	0.86	0.96	0.91	399
(G)	0.17	0.33	0.22	72	0.85	0.70	0.77	399
(H)	0.23	0.08	0.12	72	0.85	0.95	0.90	399
(I)	0.18	0.08	0.11	72	0.85	0.93	0.89	399
(J)	0.26	0.17	0.20	72	0.86	0.91	0.88	399
(K)	0.36	0.25	0.30	72	0.87	0.92	0.90	399

Table 5. Accuracy report about θ estimation in T3.

T3	precision $\theta = 0$	recall $\theta = 0$	f1-score $\theta = 0$	support $\theta = 0$	precision $\theta = 1$	recall $\theta = 1$	f1-score $\theta = 1$	support $\theta = 1$
(A)	0.42	0.34	0.38	454	0.51	0.59	0.54	525
(B)	0.45	0.51	0.48	454	0.52	0.46	0.49	525
(C)	0.48	0.58	0.53	454	0.56	0.45	0.50	525
(D)	0.43	0.09	0.15	454	0.48	0.14	0.21	525
(E)	0.54	0.78	0.64	454	0.69	0.43	0.53	525
(F)	0.49	0.53	0.51	454	0.57	0.53	0.55	525
(G)	0.48	0.49	0.49	454	0.55	0.54	0.55	525
(H)	0.49	0.57	0.53	454	0.57	0.50	0.53	525
(I)	0.46	0.31	0.37	454	0.53	0.68	0.60	525
(J)	0.47	0.34	0.40	454	0.54	0.67	0.60	525
(K)	0.45	0.24	0.31	454	0.53	0.74	0.62	525

Table 6. Accuracy report about θ estimation in T4.

T4	precision $\theta = 0$	recall $\theta = 0$	f1-score $\theta = 0$	support $\theta = 0$	precision $\theta = 1$	recall $\theta = 1$	f1-score $\theta = 1$	support $\theta = 1$
(A)	0.78	0.90	0.83	39	0.43	0.23	0.30	13
(B)	0.77	0.92	0.84	39	0.40	0.15	0.22	13
(C)	0.76	0.90	0.82	39	0.33	0.15	0.21	13
(D)	0.80	0.10	0.18	39	0.09	0.08	0.08	13
(E)	0.83	0.90	0.86	39	0.60	0.46	0.52	13
(F)	0.70	0.67	0.68	39	0.13	0.15	0.14	13
(G)	0.74	0.59	0.66	39	0.24	0.38	0.29	13
(H)	0.79	0.85	0.81	39	0.40	0.31	0.35	13
(I)	0.78	0.72	0.75	39	0.31	0.38	0.34	13
(J)	0.81	0.64	0.71	39	0.33	0.54	0.41	13
(K)	0.84	0.69	0.76	39	0.40	0.62	0.48	13

The difference, however, is that the probability of $\theta = 1$ is different when using different algorithms in different time windows. Comparing the results in Tables 3 and 5, the possibility of $\theta = 1$ in the parameter estimation results of daily data is greater than that of monthly data. It shows that the price fluctuation in T1 is wider than that in T3, which is also entirely consistent with our analysis results in Section 3. The similar conclusions can be obtained in the comparison of Tables 4 and 6.

Machine learning algorithms can more effectively extract information from large amounts of structured or unstructured data, and analyze data with less understanding of data structure or the relationship between input and output (including nonlinear relationship) to realize the quantification and prediction of high-frequency transaction data. Machine learning and deep learning algorithms automate the decision-making process to overcome the limitations of human decision-making.

Machine learning and deep learning algorithms automate the decision-making process to overcome the limitations of human decision-making. The efficiency of the analysis results shows the advantages of using machine learning and deep learning algorithms to deal with nonlinear and collinear relations, the accuracy of stock price fitting and the ability to predict future jumps.

In the end, the results of the classification reports could be used to determine the value of θ . After the dynamic value of θ , as deterministic component in stochastic process of *CSI 300* index price fluctuations, is substituted into formula (2.10), the dynamic process of the price fluctuation of *CSI 300* index would be described flexibly and effectively by the generalized BN-S model, formulas (2.9)–(2.11).

5. Conclusions

The fluctuation of price is a regular phenomenon, which has been empirically seen to widely exist in financial markets. This paper introduces a new generalized BN-S model to describe stochastic fluctuations in high-frequency asset price dynamics. The new model considers the lag caused by the asynchrony of market information, and adds new parameters to the classic BN-S model, which effectively expands the application range.

The high-frequency *CSI 300* index in Chinese financial market was selected as the research sample. The empirical data was preprocessed (the influence of overnight information on the market was removed), and a series of statistical analyses were performed to estimate its volatility characteristics. With the help of machine learning and deep learning algorithms, we analyzed dynamic prices in different time spans (monthly data, weekly data, daily data and intraday data), and estimated the deterministic component in the stochastic fluctuation process of high-frequency price data to show good operability of the new model.

Our work provides a new perspective for the analysis of price fluctuations in the financial sector, and is of positive significance for improving the accuracy of dynamic fluctuation estimation. In ongoing work, we are working on extensions of new models to accommodate to more complicated financial market scenarios and investor needs. Future research will include, but not be limited to, the study of predicting the existence and magnitude of dynamic jumps in high-frequency fluctuations, confirming quantified trading timing and avoiding the risk of abnormal fluctuations.

Acknowledgments

This work is jointly supported by the National Natural Science Foundation of China under Nos. 11662001 and 11771105, the Science Foundation of Guangxi Province under Nos. 2017GXNSFFA198012 and 2018GXNSFAA138177.

Conflict of interest

The authors declare there is no conflicts of interest.

References

1. C. T. Albuлесcu, COVID-19 and the United States financial markets' volatility, *Finance Res. Lett.*, **38** (2021), 101699. <https://doi.org/10.1016/j.frl.2020.101699>
2. S. Corbet, Y. G. Hou, Y. Hu, L. Oxley, D. Xu, Pandemic-related financial market volatility spillovers: Evidence from the Chinese COVID-19 epicentre, *Int. Rev. Econ. Finance*, **71** (2021), 55–81. <https://doi.org/10.1016/j.iref.2020.06.022>
3. J. Danielsson, M. Valenzuela, I. Zer, Learning from history: Volatility and financial crises, *Rev. Financ. Stud.*, **31** (2018), 2774–2805. <https://doi.org/10.1093/rfs/hhy049>
4. X. He, W. Chen, A closed-form pricing formula for european options under a new stochastic volatility model with a stochastic long-term mean, *Math. Financ. Econ.*, **15** (2021), 381–396. <https://doi.org/10.1007/s11579-020-00281-y>
5. M. Chauvet, Z. Senyuz, E. Yoldas, What does financial volatility tell us about macroeconomic fluctuations, *J. Econ. Dyn. Control*, **52** (2015), 340–360. <https://doi.org/10.1016/j.jedc.2015.01.002>
6. C. Arellano, Y. Bai, P. J. Kehoe, Financial frictions and fluctuations in volatility, *J. Polit. Econ.*, **127** (2019), 2049–2103. <https://doi.org/10.1086/701792>
7. M. Lin, I. SenGupta, Analysis of optimal portfolio on finite and small-time horizons for a stochastic volatility market model, *SIAM J. Financ. Math.*, **12** (2021), 1596–1624. <https://doi.org/10.1137/21M1412281>
8. F. Caselli, M. Koren, M. Lisicky, S. Tenreyro, Diversification through trade, *Q. J. Econ.*, **135** (2020), 449–502. <https://doi.org/10.1093/qje/qjz028>
9. X. He, S. Lin, A new nonlinear stochastic volatility model with regime switching stochastic mean reversion and its applications to option pricing, *Expert Syst. Appl.*, **212** (2023), 118742. <https://doi.org/10.1016/j.eswa.2022.118742>
10. X. He, S. Lin, A fractional black-scholes model with stochastic volatility and european option pricing, *Expert Syst. Appl.*, **178** (2021), 114983. <https://doi.org/10.1016/j.eswa.2021.114983>
11. X. He, C. Wnting, Pricing foreign exchange options under a hybrid heston-coxingersoll-ross model with regime switching, *IMA J. Manage. Math.*, **33** (2022), 255–272. <https://doi.org/10.1093/imaman/dpab013>
12. H. K. Baker, S. Kumar, K. Goyal, A. Sharma, International review of financial analysis: A retrospective evaluation between 1992 and 2020, *Int. Rev. Financ. Anal.*, **78** (2021), 101946. <http://doi.org/10.1016/j.irfa.2021.101946>
13. J. Baruník, T. Křehlík, Combining high frequency data with non-linear models for forecasting energy market volatility, *Expert Syst. Appl.*, **55** (2016), 222–242. <https://doi.org/10.1016/j.eswa.2016.02.008>
14. A. Dutta, E. Bouri, D. Roubaud, Modelling the volatility of crude oil returns: Jumps and volatility forecasts, *Int. J. Finance Econ.*, **26** (2021), 889–897. <https://doi.org/10.1002/ijfe.1826>
15. J. Jacod, Y. Li, X. Zheng, Statistical properties of microstructure noise, *Econometrica*, **85** (2017), 1133–1174. <https://doi.org/10.3982/ECTA13085>

16. O. E. Barndorff-Nielsen, N. Shephard, Power and bipower variation with stochastic volatility and jumps, *J. Financ. Econom.*, **2** (2004), 1–37. <https://doi.org/10.1093/jjfinec/nbh001>
17. O. E. Barndorff-Nielsen, Superposition of ornstein–uhlenbeck type processes, *Theory Probab. Appl.*, **45** (2001), 175–194. <https://doi.org/10.1137/S0040585X97978166>
18. M. Roberts, I. Sengupta, Sequential hypothesis testing in machine learning, and crude oil price jump size detection, *Appl. Math. Finance*, **27** (2020), 374–395. <https://doi.org/10.1080/1350486X.2020.1859943>
19. I. Sengupta, Generalized BN-S stochastic volatility model for option pricing, *Int. J. Theor. Appl. Finance*, **19** (2016), 1650014. <https://doi.org/10.1142/S021902491650014X>
20. I. SenGupta, W. Wilson, W. Nganje, Barndorff-nielsen and shephard model: oil hedging with variance swap and option, *Math. Financ. Econ.*, **13** (2019), 209–226. <https://doi.org/10.1007/s11579-018-0225-4>
21. M. Roberts, I. SenGupta, Infinitesimal generators for two-dimensional Lévy process-driven hypothesis testing, *Ann. Finance*, **16** (2020), 121–139. <https://doi.org/10.1007/s10436-019-00355-y>
22. I. SenGupta, Pricing asian options in financial markets using mellin transforms, *Electron. J. Differ. Equations*, **234** (2014), 1–9.
23. T. Arai, Y. Imai, R. Suzuki, Local risk-minimization for barndorff-nielsen and shephard models, *Finance Stochastics*, **21** (2017), 551–592. <https://doi.org/10.1007/s00780-017-0324-8>
24. S. Mullainathan, J. Spiess, Machine learning: an applied econometric approach, *J. Econ. Perspect.*, **31** (2017), 87–106. <https://doi.org/10.1257/jep.31.2.87>
25. Y. Qian, K. Zhang, J. Li, X. Wang, Adaptive neural network surrogate model for solving the implied volatility of time-dependent american option via bayesian inference, *Electron. Res. Arch.*, **30** (2022), 2335–2355. <https://doi.org/10.3934/era.2022119>
26. B. M. Henrique, V. A. Sobreiro, H. Kimura, Literature review: Machine learning techniques applied to financial market prediction, *Expert Syst. Appl.*, **124** (2019), 226–251. <https://doi.org/10.1016/j.eswa.2019.01.012>
27. S. A. Monfared, D. Enke, Volatility forecasting using a hybrid gjr-garch neural network model, *Procedia Comput. Sci.*, **36** (2014), 246–253. <https://doi.org/10.1016/j.procs.2014.09.087>
28. J. Zheng, X. Fu, G. Zhang, Research on exchange rate forecasting based on deep belief network, *Neural Comput. Appl.*, **31** (2019), 573–582. <https://doi.org/10.1007/s00521-017-3039-z>
29. Y. Liu, Novel volatility forecasting using deep learning–long short term memory recurrent neural networks, *Expert Syst. Appl.*, **132** (2019), 99–109. <https://doi.org/10.1016/j.eswa.2019.04.038>
30. T. Fischer, C. Krauss, Deep learning with long short-term memory networks for financial market predictions, *Eur. J. Oper. Res.*, **270** (2018), 654–669. <https://doi.org/10.1016/j.ejor.2017.11.054>
31. H. Zhou, P. S. Kalev, Algorithmic and high frequency trading in asia-pacific, now and the future, *Pac. Basin Finance J.*, **53** (2019), 186–207. <https://doi.org/10.1016/j.pacfin.2018.10.006>
32. N. Todorova, M. Souček, Overnight information flow and realized volatility forecasting, *Finance Res. Lett.*, **11** (2014), 420–428. <https://doi.org/10.1016/j.frl.2014.07.001>

33. I. SenGupta, W. Nganje, E. Hanson, Refinements of barndorff-nielsen and shephard model: an analysis of crude oil price with machine learning, *Ann. Data Sci.*, **8** (2021), 39–55. <https://doi.org/10.1007/s40745-020-00256-2>
34. K. Grobys, When the blockchain does not block: on hackings and uncertainty in the cryptocurrency market, *Quant. Finance*, **54** (2020), 1267–1279. <https://doi.org/10.1080/14697688.2020.1849779>
35. D. Xiao, J. Wang, Dynamic complexity and causality of crude oil and major stock markets, *Energy*, **193** (2020), 116791. <https://doi.org/10.1016/j.energy.2019.116791>
36. S. Habtemicael, M. Ghebremichael, I. SenGupta, Volatility and variance swap using superposition of the barndorff-nielsen and shephard type Lévy processes, *Sankhya B*, **81** (2019), 75–92. <https://doi.org/10.1007/s13571-017-0145-y>
37. T. G. Andersen, T. Bollerslev, F. X. Diebold, P. Labys, Modeling and forecasting realized volatility, *Econometrica*, **71** (2003), 579–625. <https://doi.org/10.1111/1468-0262.00418>



AIMS Press

©2023 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>)