Heliyon 8 (2022) e11388

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Analysis of parametric and non-parametric option pricing models ☆

Qiang Luo, Zhaoli Jia *, Hongbo Li, Yongxin Wu

School of Mathematics, Hefei University of Technology, Hefei 230009, PR China

ARTICLE INFO

Keywords:
Heston model
Bi-Heston model
Option
XGBoost

ABSTRACT

In this paper, a closed-form analytical solution of option price under the Bi-Heston model is derived. Through empirical analysis, the advantages and disadvantages of the parametric pricing model are compared and analysed with those of the non-parametric model. The analysis shows that: (1) the parametric pricing model significantly outperforms the machine learning model in terms of in-sample pricing effects, while the Bi-Heston model slightly outperforms the Heston model. (2) In terms of out-of-sample pricing, the machine learning model is inferior to the parametric model for call options, while the Bi-Heston model is significantly better than the other two models for put options, and the other two models are similar. (3) In the robustness analysis of the three pricing models, the machine learning model shows strong instability, while the Bi-Heston model shows a more stable side.

1. Introduction

In recent years, China's financial markets have developed at a rapid pace, and a large number of financial products have emerged as derivatives. Financial derivatives, which are financial contracts with functions such as risk management, price discovery and hedging, have become an integral part of the financial market as they can efficiently enhance market allocation efficiency. Options, as a financial derivative product and one of the most dominant products traded in the financial derivatives market, are valued by investors for their flexibility in risk management. While options can bring great convenience to investors, they also have some potential drawbacks if they are not priced correctly. Therefore, the correct pricing of options is critical to the risk control of investors. In the literature, there are two types of pricing models, parametric option pricing models and non-parametric option pricing models.

In all the parametric models, the most influential one was proposed by Black and Scholes [1] in 1973, from which an option pricing formula in the form of a closed solution can be derived. This formula was considered to be epoch-making and was well known among researchers and practitioners due to an easiness of use. Soon after, researchers discovered that it did not correspond well to some of the realities of the market due to its strict assumptions. And then, Heston [2] improved the Black-Scholes model by making a new assumption that volatility is not constant, but follows a CIR [3] process. This solves the volatility estimation problem. The Heston model is arguably one of the most widely used volatility models, but later scholars have found that the Heston model does not fit the implied volatility smile phenomenon very well. To address this problem, Christoffersen et al. [4] improved the Heston model by proposing the Bi-Heston stochastic volatility model which specifies a two-factor structure for volatility. Its additional parameters allow it to estimate option prices more accurately as well as fit implied volatility. Very recently, more scholars joined in the study of the model. Rouah [5] showed the implied volatility of the Bi-Heston model and found that it was closer to the actual market implied volatility than the Heston model. Fallah [6] first proved the existence and uniqueness of solutions of stochastic differential equations under the Bi-Heston model. Mehrdoust [7] proposed an extension of the Bi-Heston model to include jumps in the financial modelling of stock prices and found that it was an efficient tool for option pricing with the Fast Fourier Transform method. Their findings all suggest that the Bi-Heston model is highly relevant. Therefore, in this paper, we aim to explore the pricing performance of the Bi-Heston model, and the Heston model is used to compare with.

However, there is one more issue that needs to be addressed when using these parametric models in practice. As the name implies, to use the parametric model to price options, the model parameters need to be estimated first. Models usually have many parameters to be estimated, and the

☆ This paper was supported by the National Natural Science Foundation of China under Grant No. 72071068; Innovation and entrepreneurship project for College Students under Grant No. 202110359075.

* Corresponding author.
E-mail addresses: luoqiang0930@163.com (Q. Luo), jiazhaoli@hfut.edu.cn (Z. Jia), lihongbo214274@163.com (H. Li), Wuyongxin0709@163.com (Y. Wu).

https://doi.org/10.1016/j.heliyon.2022.e11388
Received 1 June 2022; Received in revised form 3 September 2022; Accepted 31 October 2022

2405-8440/© 2022 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

objective function is complex. To solve this problem, there are also many researchers who have given their solutions. Cui [8] changed the structure of the characteristic functions in the Heston model to give an analytic gradient for solving the problem of parameter estimation and proposed a new fast and efficient multi-parameter correction algorithm for the Heston model based on the analytic gradient. In recent studies, Cacace [9] developed a new polynomial filtering method to estimate parameters in the Heston model that relied on a linear filter using a polynomial state-space formulation of a discrete version of the continuous-time model. In this paper, we reduce this problem to an optimization problem and give a solution based on numerical calculations.

Even if the problem of parameter estimation is solved, traditional parametric models still have some drawbacks that they are strictly limited in some statistical or economic assumptions and do not facilitate the judgement of the merit of the model fit. Machine learning, as an alternative to non-parametric models, is also intensively developed recently. Ma [10] proposed a price forecasting method based on cost-sensitive deep forest, which maintains the high accuracy of deep forest and can significantly reduce the forecast cost error while maintaining high forecast accuracy. Hitam [11] proposed a support vector machine based on particle swarm algorithm optimization to predict the futures price of cryptocurrencies. It is found that the hybrid model can effectively predict the price of cryptocurrencies, and the prediction performance is better than that of a single support vector machine. Yang et al. [12] combined hybrid genetic algorithms with support vector machines and found that support vector machines constructed based on genetic algorithms and radial basis functions are more suitable for real values than other methods. Kumar [13] proposed four hybrid prediction models that are combinations of four different feature selection techniques with proximal support vector machine classifier. Due to the variety of machine learning algorithms, researchers have proposed a large number of related pricing models, and among these, the XGBoost model stands out. Le et al. [14] used an improved XGBoost algorithm to build a non-parametric machine learning model for analyzing corporate financial statements and demonstrated that the optimized XGBoost algorithm has a great advantage in processing time and has the best prediction results. More recently, Ivacu [15] analyzed the prediction performance of the XGBoost algorithm for option pricing and affirmed its feasibility. Chang [16] constructed a credit risk assessment model using an extreme gradient boosting machine classifier and empirically tested that the XGBoost classifier had better prediction results than logistic regression, clustered data processing techniques, and three support vector machine classifiers.

In this paper, we use traditional parametric models and non-parametric machine learning models to price 50ETF options respectively and compare the prediction accuracy of all models. Of these models, the Heston and Bi-Heston model represent the traditional parametric models, and the XGBoost model represents the non-parametric model. Our article is organized as follows: In Section 2, we give a brief description of the Heston model, and a detailed derivation of the Bi-Heston model is provided. In Section 3, the research methodology and data are presented. Firstly, the selected data sources are presented, followed by the machine learning model using the XGBoost algorithm. Besides, the methods for parameter calibration and the metrics for model evaluation are given. In Section 4, the detailed process of the empirical analysis is given. The prediction results of the traditional parametric model and the machine learning model are compared to provide a more comprehensive comparison of the predictive validity of them. Furthermore, the robustness tests are conducted for both. Some summaries are provided in Section 5.

2. Derivation of the bi-Heston model

It is well-known that the Heston model overcomes the Black-Scholes model's shortcoming of assuming constant volatility well and is one of the most popular models in the option valuation literature. It is given by

$$\begin{cases} dS_t = rS_t dt + \sqrt{v_t} S_t dZ_{1,t} \\ dv_t = \kappa(\theta - v_t)dt + \sigma\sqrt{v_t} dZ_{2,t} \\ Cov(Z_{1,t}, Z_{2,t}) = \rho, \end{cases}$$

where S_t denotes the underlying asset price, which is modelled by the square root process with a stochastic instantaneous variance v_t, $\{Z_{i,t}\}$ $(i = 1, 2)$ denote the standard Brownian motion, r is the risk-free rate, κ is the mean reversion speed, θ is the mean reversion of variance, σ is the volatility of variance, and ρ is the correlation coefficient. However, some scholars have pointed out that the Heston model sometimes fails to fit the implied volatility smile very well. To overcome this shortcoming, the Bi-Heston model is generalized from the Heston model, which is a straightforward way to incorporate a stochastic correlation by using multiple stochastic volatility factors. Assume that risky asset prices satisfy the Bi-Heston model as follows

$$\begin{cases} dS_t = rS_t dt + \sqrt{v_{1t}} dZ_{1,t} + \sqrt{v_{2t}} dZ_{2,t} \\ dv_{1t} = \kappa_1(\theta_1 - v_{1t})dt + \sigma_1\sqrt{v_{1t}} dZ_{3,t} \\ dv_{2t} = \kappa_2(\theta_2 - v_{2t})dt + \sigma_2\sqrt{v_{2t}} dZ_{4,t}, \end{cases}$$

where S_t is modelled by two square root processes with two stochastic instantaneous variances v_{1t} and v_{2t}, $\{Z_{i,t}\}$ $(i = 1, 2, 3, 4)$ denote the standard Brownian motion, v_{1t}, v_{2t} are extended from v_t. In order to ensure that the square root processes are always positive, we require that $2\kappa_1\theta_1 \geq (\sigma_1)^2, 2\kappa_2\theta_2 \geq (\sigma_2)^2$ [2]. We assume that the covariances between these Brownian motions are equal to zero except $Cov(Z_{1,t}, Z_{3,t}) = \rho_1 t$, $Cov(Z_{2,t}, Z_{4,t}) = \rho_2 t$.

Consider undefined assets denoted by $V(S_t, v_{1t}, v_{2t}, t)$. The maturity payout function for European options is $(S_T - K)^+$. By the Faman-Kac theorem, we have

$$\begin{cases} \dfrac{\partial V}{\partial t} + rS\dfrac{\partial V}{\partial S} + \kappa_1(\theta_1 - v_{1t})\dfrac{\partial V}{\partial v_{1t}} + \kappa_2(\theta_2 - v_{2t})\dfrac{\partial V}{\partial v_{2t}} + \dfrac{1}{2}(v_{1t} + v_{2t})S^2\dfrac{\partial^2 V}{\partial S^2} + S\sigma_1 v_{1t}\rho_1\dfrac{\partial^2 V}{\partial S \partial v_{1t}} \\ \qquad + S\sigma_2 v_{2t}\rho_2\dfrac{\partial^2 V}{\partial S \partial v_{2t}} + \dfrac{1}{2}{\sigma_1}^2 v_{1t}\dfrac{\partial^2 V}{\partial {v_{1t}}^2} + \dfrac{1}{2}{\sigma_2}^2 v_{2t}\dfrac{\partial^2 V}{\partial {v_{2t}}^2} - rV = 0 \\ V(S, v_{1t}, v_{2t}, T) = (S - K)^+. \end{cases}$$

Decompose $V(S_t, v_{1t}, v_{2t}, t)$ into two parts

$$V(S, v_{1t}, v_{2t}, t) = SP_1(S, v_{1t}, v_{2t}, t) - Ke^{-r(T-t)}P_2(S, v_{1t}, v_{2t}, t), \tag{1}$$

where $P_j(S, v_{1t}, v_{2t}, t)$ $(j = 1, 2)$ satisfies the following equation at $t = T$

$$P_1(S, v_{1t}, v_{2t}, T) = P_2(S, v_{1t}, v_{2t}, T) = H(S - K),$$

where $H(x)$ is Heaviside function, $H(x) = I_{\{x>0\}}$.

Let $x = \ln S$, then $P_j(x, v_{1t}, v_{2t}, t)$, $(j = 1, 2)$ satisfies the following differential equation

$$\begin{cases} \dfrac{\partial P_j}{\partial t} + [r + \alpha_j(v_{1t} + v_{2t})\dfrac{\partial V}{\partial x}] + (\kappa_1\theta_1 - \beta_j v_{1t})\dfrac{\partial V}{\partial v_{1t}} + (\kappa_2\theta_2 - r_j v_{2t})\dfrac{\partial V}{\partial v_{2t}} + \dfrac{1}{2}(v_{1t} + v_{2t})\dfrac{\partial^2 V}{\partial x^2} \\ \qquad + \sigma_1 v_{1t}\rho_1\dfrac{\partial^2 V}{\partial v_{1t}\partial x} + \sigma_2 v_{2t}\rho_2\dfrac{\partial^2 V}{\partial v_{2t}\partial x} + \dfrac{1}{2}{\sigma_1}^2 v_{1t}\dfrac{\partial^2 V}{\partial {v_{1t}}^2} + \dfrac{1}{2}{\sigma_2}^2 v_{2t}\dfrac{\partial^2 V}{\partial {v_{2t}}^2} = 0 \\ P_j(x, v_{1t}, v_{2t}, T) = H(x - \ln K), \end{cases}$$

where

$$\begin{cases} \alpha_1 = \dfrac{1}{2}, \alpha_2 = -\dfrac{1}{2}, \beta_1 = \kappa_1 + \lambda_1 - \rho_1\sigma_1, \beta_2 = \kappa_1 + \lambda_1, \\ \gamma_1 = \kappa_2 + \lambda_2 - \rho_2\sigma_2, \gamma_2 = \kappa_2 + \lambda_2. \end{cases}$$

To solve the above system of differential equations, we first consider its characteristic functions, namely $f_j(x, v_{1t}, v_{2t}, t; \varphi)$, $(j = 1, 2)$. When $t = T$, the termination condition is

$$f_j(x, v_{1t}, v_{2t}, T; \varphi) = e^{i\varphi x}, \ (j = 1, 2).$$

Similar to the derivation of the Heston model, we assume that the characteristic function has the form of the following solution

$$f_j(x, v_{1t}, v_{2t}, t; \varphi) = e^{C(T-t;\varphi) + D(T-t;\varphi)v_{1t} + E(T-t;\varphi)v_{2t} + i\varphi x}. \tag{2}$$

Taking the partial derivatives separately into equation (2) yields the following three ODEs

$$\begin{cases} \dfrac{\mathrm{d}D}{\mathrm{d}t}v_{1t} + i\varphi\alpha_j v_{1t} - \dfrac{1}{2}v_{1t}\varphi^2 + (i\varphi\sigma_1\rho_1 - \beta_j)v_{1t}D + \dfrac{1}{2}\sigma_1 v_{1t}D^2 = 0 \\ \dfrac{\mathrm{d}E}{\mathrm{d}t}v_{2t} + i\varphi\alpha_j v_{2t} - \dfrac{1}{2}v_{2t}\varphi^2 + (i\varphi\sigma_2\rho_2 - \gamma_j)v_{2t}E + \dfrac{1}{2}\sigma_2 v_{2t}E^2 = 0 \\ \dfrac{\mathrm{d}C}{\mathrm{d}t} + \gamma i\varphi + \kappa_1\theta_1 D + \kappa_2\theta_2 E = 0 \end{cases}$$

By solving the above ODE, we can easily obtain a display solution of the characteristic function, as follows

$$\begin{cases} C(\tau, \varphi) = r\varphi i\tau + \dfrac{\kappa_1\theta_1}{{\sigma_1}^2}\left\{(\beta_j - \rho_1\sigma_1\varphi i + d_1)\tau - 2\ln[\dfrac{1 - g_1 e^{d_1\tau}}{1 - g_1}]\right\} \\ \qquad\qquad + \dfrac{\kappa_2\theta_2}{{\sigma_2}^2}\left\{(\gamma_j - \rho_2\sigma_2\varphi i + d_2)\tau - 2\ln[\dfrac{1 - g_2 e^{d_2\tau}}{1 - g_2}]\right\} \\ D(\tau, \varphi) = \dfrac{\beta_j - \rho_1\sigma_1\varphi i + d_1}{{\sigma_1}^2}[\dfrac{1 - e^{d_1\tau}}{1 - g_1 e^{d_1\tau}}] \\ E(\tau, \varphi) = \dfrac{\gamma_j - \rho_2\sigma_2\varphi i + d_2}{{\sigma_2}^2}[\dfrac{1 - e^{d_2\tau}}{1 - g_2 e^{d_2\tau}}], \end{cases}$$

where

$$\begin{cases} g_1 = \dfrac{\beta_j - \rho_1\sigma_1\varphi i + d_1}{\beta_j - \rho_1\sigma_1\varphi i - d_1}, \quad d_1 = \sqrt{(\rho_1\sigma_1\varphi i - \beta_j)^2 - {\sigma_1}^2(2\alpha_j\varphi i - \varphi^2)}, \\ g_2 = \dfrac{\gamma_j - \rho_2\sigma_2\varphi i + d_2}{\gamma_j - \rho_2\sigma_2\varphi i - d_2}, \quad d_2 = \sqrt{(\rho_2\sigma_2\varphi i - \gamma_j)^2 - {\sigma_2}^2(2\alpha_j\varphi i - \varphi)^2}. \end{cases}$$

By the Fourier inversion, $P_j(S, v_{1t}, v_{2t}, t)$ $(j = 1, 2)$ can be calculated in the following form.

$$P_j(S, v_{1t}, v_{2t}, t) = \frac{1}{2} + \frac{1}{\pi}\int_0^\infty \mathrm{Re}(\frac{e^{-i\varphi\ln K} f_j(x, v_{1t}, v_{2t}, \varphi)}{i\varphi})d\varphi. \tag{3}$$

The above expressions can be obtained numerically by programming. As a result, substitute the equation (3) into (1), we obtain a closed-form solution of the Bi-Heston model.

3. Data and price by XGBoost

3.1. Data

In this paper, we use the 50ETF stock Index options data as the research data to calibrate the parameters of the parametric model and train the machine learning model. A total of 123,999 options were selected from 4 January 2016 to 31 December 2020, with the option code, trade date, option closing price, option volume, option trade amount, option type, option name, strike price, expiry date, remaining expiry time, 50ETF price, risk-free return and implied volatility recorded as the original data set (Tables 9 and 10).

3.2. The XGBoost model

The option pricing model is established based on XGBoost, where the feature variables $(x_i^{(m)}, m = 1, ..., 7)$ are: option trading volume, execution price, remaining maturity time, 50ETF price, 50ETF daily return, risk-free return and implied volatility, and the target variable (p) is the closing price of the option. Let the training data set with n data sets and 7 feature indicators be, $D = \{(x_1, p_1), ..., (x_n, p_n)\}$, where $x_i = \{x_i^{(1)}, ..., x_i^{(7)}\}$. Because its essence is boosting iterative algorithm, which includes two key parts: addition model and forward distribution algorithm. According to the addition model, the predicted value of option price can be expressed as $\hat{p}_i = \sum_{t=1}^{k} f_t(x_i)$, f_t is the t-th tree model, and $\hat{p}_i$ denotes the predicted price of the i-th option sample x_i. At the same time, we need to minimize the objective function to learn the function in the supervision model. During the training of option pricing model, the objective function is:

$$Obj = \sum_{i=1}^{n} l(p_i, \hat{p}_i) + \sum_{t=1}^{k} \Omega(f_t).$$

p_i is the market price of the option. The objective function Obj consists of a traditional loss function l and a regularization term Ω that suppresses the complexity of the model and prevents the model from over fitting. The loss function reflects the difference between the predicted value and the real value. As the penalty term of tree model complexity, Ω should be regularized by the number of leaf nodes and the weight of leaf nodes. So we define the complexity function of the tree model

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} {\omega_j}^2, I_j = \{i | q(x_i) = j\},$$

where γT denotes the penalty term for the complexity of the number of leaf nodes T, ω_j is the weight of the leaf node and also represents the predicted value on each leaf node, $\frac{1}{2}\lambda \sum_{j=1}^{T} {\omega_j}^2$ punishes the weight of leaf nodes in the form of norm. In addition, XGBoost, as the antecedent distribution algorithm, the learning process is to learn the first tree first, then learn the second tree based on the first, and so on. In the t-th iteration, the price prediction for the i-th option sample x_i is

$$\hat{p}_i^t = \hat{p}_i^{t-1} + f_t(x_i),$$

where $f_t(x_i)$ is the tree model we are going to find in this round. By optimizing the above objective function, we can get $f_t(x_i)$. Specifically, carry out the second-order Taylor expansion of the function l and remove the constant term $l(p_i, \hat{p}_i^{t-1})$ of the second iteration to obtain the simplified result:

$$Obj = \sum_{i=1}^{n} (g_i f_t(x_i) + \frac{1}{2} h_i {f_t}^2(x_i)) + \sum_{t=1}^{k} \Omega(f_t), \tag{4}$$

where g_i, h_i are the first and second derivatives of the function l to the option price prediction value $\hat{p}_i^{t-1}$ obtained by the passed $t-1$ iteration. Next, define the sample set of leaf nodes

$$I_j = \{i | q(x_i) = j\},$$

$q(x)$ represents the specific leaf node into which the sample falls, $\omega_{q(x)}$ represents the predicted value of each sample, i.e., we can convert $f_t(x)$ to $\omega_{q(x)}$. Rewriting the set of samples to the set of leaf nodes. In the equation (4), let $G_j = \sum_{i \in I_j} g_i$, $H_j = \sum_{i \in I_j} h_i$, the new objective function can be obtained

$$Obj = \sum_{j=1}^{T} [G_j \omega_j + \frac{1}{2}(H_j + \lambda){\omega_j}^2] + \gamma T. \tag{5}$$

The predicted value ω_j is unknown. The problem of requiring the most value of the above function, i.e., making the value of the first-order derivative of the objective function (5), finding the value ${\omega_j}^* = -\frac{G_j}{H_j + \lambda}$ corresponding to the leaf node j, and obtaining the minimum value of the objective function as

$$Obj = -\frac{1}{2} \sum_{j=1}^{T} \frac{{G_j}^2}{H_j + \lambda} + \gamma T.$$

The final sample prediction value we want to get is the sum of the weights of the leaf nodes Ω that the sample finally falls into in each tree model. When implementing XGBoost algorithm, we should consider the optimal parameter combination of the algorithm to ensure the optimality of the model. The super parameters of XGBoost algorithm are mainly divided into three categories: task parameters, conventional parameters and lifter parameters. The following is the range of a set of super parameters of XGBoost selected by grid search method combined with cross validation to find the best combination of super parameters (Table 1):

Table 1. Algorithm Parameters of XGBoost Model.

Parameters of XGBoost	Implication of parameter	Range of parameter
n-estimators	Number of decision trees	300~1000
max-depth	Maximum depth of tree	5~10
min-child weight	Sum of leaf node weights	5~10
learning rate	Learning Rate	0.1~0.5
subsample	Proportion of subsamples to training sample set	0.6~0.8
Colsample bytree	Feature random sampling ratio	0.3~0.5

In the empirical analysis, we control the scope of the above hyperparameters within the above range to obtain multiple sets of prediction results and then compare the models.

3.3. Parameter calibration and model estimation

There is no doubt that, after observing the option pricing formulae described in the previous section, when we want to use the Heston and Bi-Heston models for option pricing, we should determine the values of the parameters in the models firstly. The objective of parameter determination is to minimise the error between the true price and the theoretical price, and the method of judgement used in this paper is non-linear least squares, so the parameter calibration problem is transformed into an optimisation problem as shown in the following equation.

$$\arg\min_{\Theta} \ MSE = \frac{1}{n}\sum_{i=1}^{n}(p_i - \hat{p}_i(\Theta))^2,$$

where n denotes the number of samples used for parameter calibration, and p_i denotes the market price of option i and $\hat{p}_i$ denotes the price predicted by the Heston or Bi-Heston model of option i. The Heston model requires the determination of the values of 5 parameters $\Theta = (v, \theta, \kappa, \sigma, \rho)$, while the Bi-Heston requires 10 parameters $\Theta = (v_1, \theta_1, \kappa_1, \sigma_1, \rho_1, v_2, \theta_2, \kappa_2, \sigma_2, \rho_2)$. At the same time, the parameter calibration problem is transformed into the problem of finding the minimum value of a high-dimensional function. We use a particle swarm algorithm based on parallel computing to calibrate the parameters of the Heston and Bi-Heston models. To avoid the influence of the parameter calibration part on the empirical study of the models later on, we selected the same objective function and the same sample for the parameter calibration of the two models mentioned above.

Firstly, we use the root mean squared error (RMSE) as a way of assessing the in-sample pricing effectiveness of the model. A distinct advantage of the RMSE is that it squares the bias term, which allows it to give greater weight to error terms with larger weights, thus penalising models with larger error terms. We expect the better model to show a smaller RMSE, which in turn reflects the good fit of the model.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - \hat{p}_i)^2}.$$

Secondly, we use the implied volatility root mean square error (IVRMSE) as a way of assessing the out-of-sample pricing effectiveness of the model. IVRMSE is a frequently used method when evaluating the pricing effectiveness of options and is represented by the following equation

$$IVRMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\sigma_i - \hat{\sigma}_i)^2},$$

where σ_i denotes the implied volatility of the underlying asset as inverse solved from the B-S model using the price of option i on that day; and $\hat{\sigma}_i$ denotes the implied volatility of the underlying asset as inverse solved from the B-S model by using the price of option i obtained from the pricing model. As implied volatility is an important factor in option analysis, a smaller IVRMSE indicates a smaller deviation between the market price of the option and the theoretical price, and the model is better.

4. Empirical analysis

4.1. Experimental steps

In the first step, since the dataset has some invalid fields and some data are missing, we need to pre-process the dataset to facilitate the subsequent study; in the second step, the valid dataset is divided into experimental samples, part of which is used for in-sample parameter calibration and model training, and the other part is used for out-of-sample model prediction; in the third step, the parameters of the Heston and Bi-Heston models are calibrated and the XGBoost model is trained for the first part of the sample; in the fourth step, the parameters and the model obtained in the third step are used for out-of-sample pricing for the second part of the sample (Fig. 1).

4.2. Analysis of in-sample pricing effects

Based on the processed dataset, we perform a hybrid parameter calibration using options data from 30 September 2020 to 17 December 2020, i.e., call and put options are placed in the same dataset. Finally, the calibrated parameters and the trained model are used to perform in-sample pricing on this dataset. As the PSO algorithm is an intelligent algorithm and XGBoost is a machine learning model, the results obtained are stochastic in nature. Therefore, for the in-sample pricing of the three models, we will perform 10 operations each and evaluate the in-sample pricing effectiveness of the three models using the evaluation metric RMSE described in the previous section.

As can be seen from Table 2, the Bi-Heston model outperforms the Heston model in terms of parameter calibration, although both models use the same calibration technique. In terms of in-sample pricing, the Bi-Heston model and the Heston model are significantly better than the XGBoost model, and the Heston model is more consistent with the Bi-Heston model in terms of pricing, but the Bi-Heston model is slightly better than the Heston model. It is only a coincidence that the in-sample pricing effect is not good enough, but the out-of-sample pricing effect is better, so we

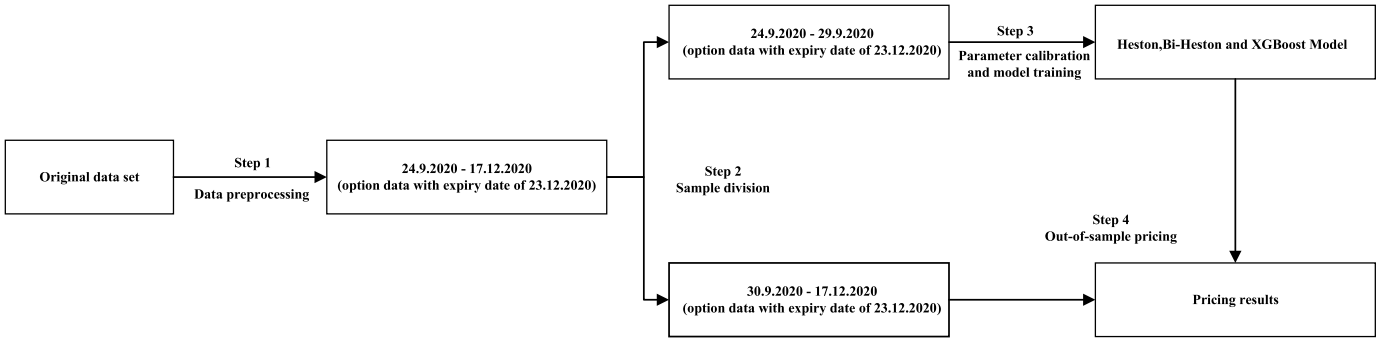

Fig. 1. Experimental Flow Chart.

Table 2. Comparison of Parametric Calibration Results.

Number of experiment	Heston model	Bi-Heston model	XGBoost model
1	0.0207	0.0191	0.0375
2	0.0200	0.0190	0.0387
3	0.0199	0.0191	0.0454
4	0.0199	0.0193	0.0366
5	0.0200	0.0190	0.0375
6	0.0199	0.0190	0.0399
7	0.0209	0.0190	0.0366
8	0.0199	0.0191	0.0375
9	0.0199	0.0190	0.0451
10	0.0199	0.0190	0.0375

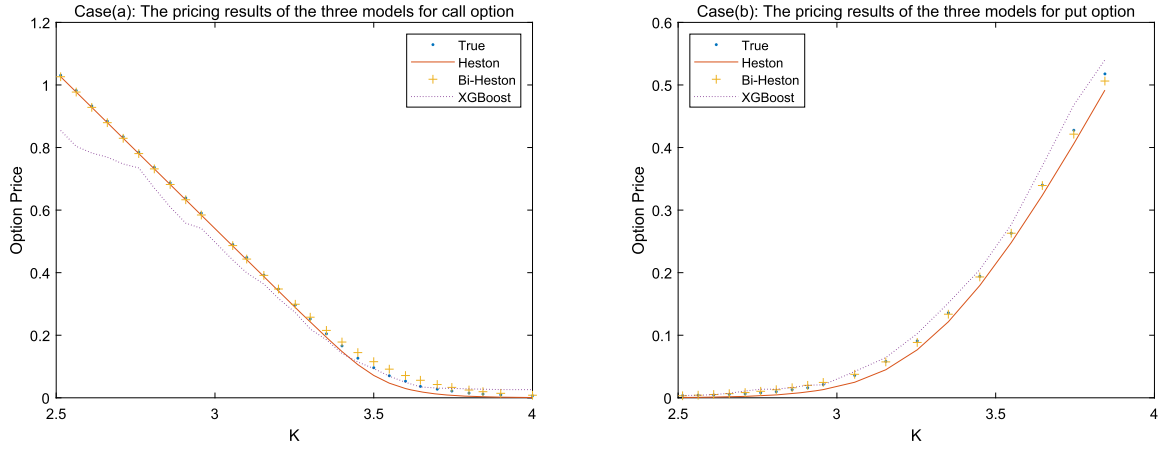

Fig. 2. Out of sample pricing effect comparison of the three models.

select the group with the best in-sample pricing effect as the parameters for out-of-sample pricing, and the specific parameter values are as follows (Table 3):

Table 3. Model parameters.

Heston	ν	θ	κ	σ	ρ
	0.0086	0.7644	0.8579	1.1452	0.4069
Bi-Heston	ν_1	θ_1	κ_1	σ_1	ρ_1
	0.0471	0.0010	0.8579	0.0025	-0.9997
	ν_2	θ_2	κ_2	σ_2	ρ_2
	0.5038	1.0000	0.0422	0.0010	0.2901

4.3. Analysis of out-sample pricing effects

Based on the model parameters obtained as described previously, we can easily calculate the out-of-sample option price based on the option pricing formula. Fig. 2 shows a comparison of the actual results of the three models for call and put options respectively.

It is clear from the graphs that the pricing results of the Heston and Bi-Heston models match the actual option prices, while the XGBoost model is less effective than the first two models. As the strike price of the call option increases, the theoretical price of the option decreases for all models; as the strike price of the put option increases, the theoretical price of the option increases for all models. This phenomenon is consistent with objective reality. The Bi-Heston model is slightly more effective than the Heston model, as can be seen from the images, and specifically, the actual pricing results are shown in Table 4, 5.

Table 4. Call options for Selected Days and Maturities.

K	True	Heston	Bi-Heston	XGBoost
2.514	1.031	1.026	1.026	0.855
2.612	0.933	0.928	0.928	0.783
2.711	0.835	0.830	0.830	0.747
2.809	0.737	0.732	0.732	0.671
2.908	0.639	0.633	0.633	0.558
3.100	0.449	0.444	0.444	0.399
3.200	0.348	0.341	0.348	0.317
3.300	0.252	0.242	0.258	0.220
3.400	0.166	0.148	0.178	0.142
3.500	0.096	0.071	0.115	0.094
3.600	0.053	0.030	0.071	0.050
3.700	0.028	0.012	0.042	0.031
3.800	0.015	0.005	0.025	0.029
3.900	0.010	0.002	0.014	0.026
4.000	0.007	0.001	0.008	0.026

Table 5. Put options for Selected Days and Maturities.

K	True	Heston	Bi-Heston	XGBoost
2.563	0.004	0.001	0.004	0.004
2.661	0.006	0.002	0.007	0.007
2.711	0.006	0.002	0.008	0.011
2.809	0.010	0.005	0.013	0.014
2.908	0.016	0.009	0.020	0.020
3.056	0.036	0.025	0.038	0.042
3.154	0.058	0.045	0.057	0.064
3.253	0.091	0.077	0.088	0.103
3.351	0.136	0.122	0.134	0.152
3.450	0.194	0.179	0.193	0.205
3.549	0.263	0.248	0.263	0.277
3.647	0.340	0.324	0.339	0.371
3.746	0.428	0.406	0.421	0.468
3.844	0.518	0.492	0.506	0.540

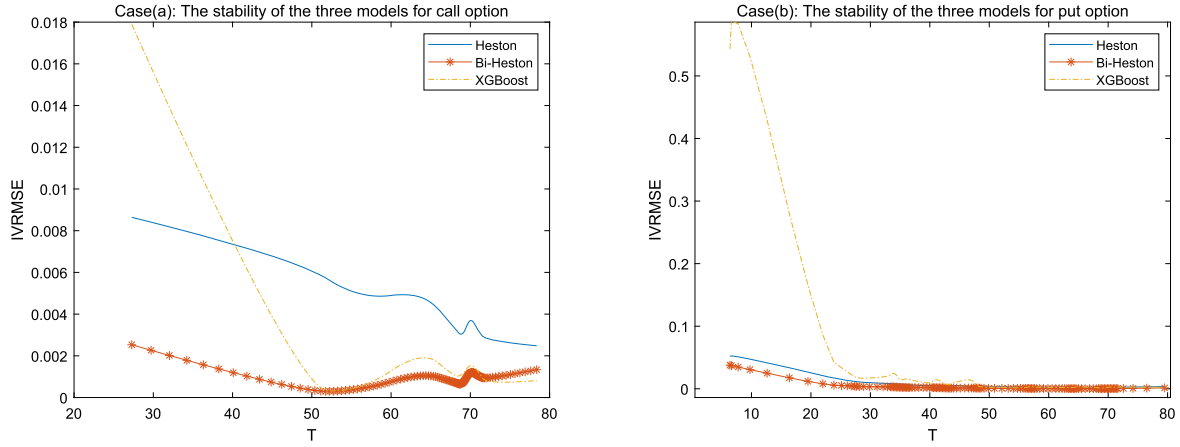

Fig. 3. Comparison of the Stability of three models.

In response to the above actual results, we can analyse that as the strike price (K) increases, the errors of all three models also increase, but the Bi-Heston model is more stable than the remaining two models. We can also see from the specific values that the Heston model does not differ significantly from the Bi-Heston model at low strike prices for call options, but as the strike price increases, the Bi-Heston model significantly outperforms the Heston model; for put options, the Bi-Heston model significantly outperforms the other two models.

On the other hand, we also examine the effect of out-of-sample pricing of options using IVRMSE. The out-of-sample pricing spans the period from 30 September 2020 to 17 December 2020, and the expiry date of the options is 23 December 2020. We can see that the data set has a relatively long time horizon and we can consider studies that address the stability of the option pricing model. The actual results are shown in Fig. 3.

As can be seen from the graph, the Bi-Heston model is more stable than the other two models for both call and put options. All three models show an increase in IVRMSE as the expiry date approaches, and this is particularly evident in the XGBoost model. Otherwise, the IVRMSE reflects the deviation between the market price of the option and the theoretical price, and the smaller the IVRMSE, the better the option is priced. From the above graph, we can see that the IVRMSE of the Bi-Heston model is smaller than that of the Heston and XGBoost models for both call and put options when the expiry dates are the same, which means that the pricing effect of the Bi-Heston model is better than that of the other two models.

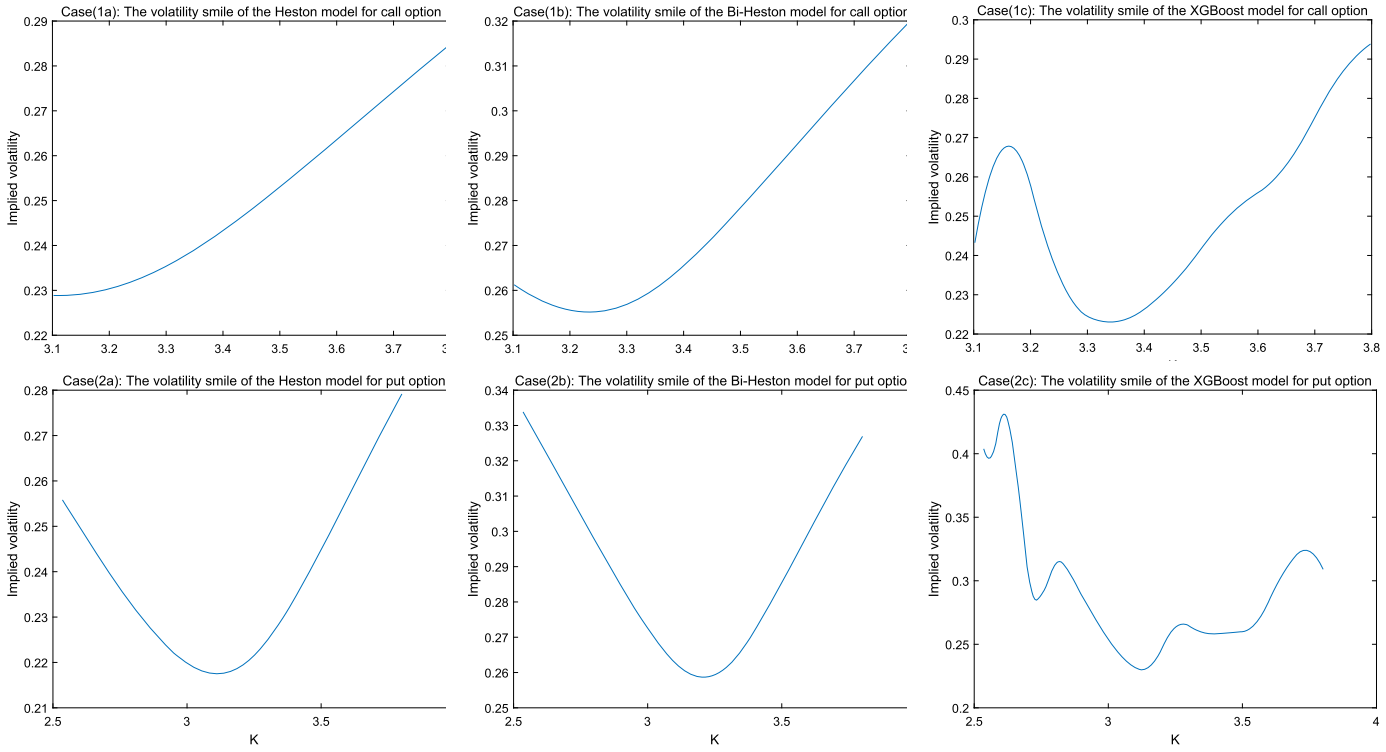

Fig. 4. Volatility Smile for Selected Days.

Table 6. RMSE for in-sample pricing results.

Dataset Name	Heston model	Bi-Heston model	XGBoost model
50ETF	0.0199	0.0190	0.0366
300ETF(510300)	0.0260	0.0250	0.0521
300ETF(159919)	0.0166	0.0162	0.0327

Table 7. IVRMSE for pricing results of out-sample call option.

Dataset Name	Heston model	Bi-Heston model	XGBoost model
50ETF	0.0268	0.0206	0.0894
300ETF(510300)	0.3479	0.3166	0.0800
300ETF(159919)	0.0524	0.0569	0.0830

For call options, the XGBoost model is significantly better than the Heston model when the expiry date is far away, while the XGBoost model is less stable than the Heston model as the expiry date approaches, and the pricing effect is inferior to that of the Heston model. For put options, both the Heston and Bi-Heston models are more stable and better priced than the XGBoost model.

According to the assumptions of the traditional B-S model, the implied volatility in the model should be a constant that is independent of the strike price. However, many scholars have pointed out that this assumption is not valid in the real market, and the most direct evidence of that is the volatility smile phenomenon. The volatility smile gets its name from the shape of this curve: when the option price deviates from the strike price, the implied volatility of the option rises, thus showing a low middle and high sides, in the form of a smiling mouth. We then also explored whether the three pricing models mentioned above could exhibit the phenomenon of a volatility smile, as shown in Fig. 4.

The first row shows the relationship between the implied volatility of the call option and the strike price, while the second row shows the relationship between the implied volatility of the put option and the strike price. We can see from the above graph that both the Heston model and the Bi-Heston model reflect the phenomenon of volatility smiles well, while the XGBoost model does not reflect the phenomenon well.

4.4. Analysis of the impact of dataset on the models

In order to explore the impact of the selected data on the models, we designed empirical analyses for different option datasets. In the process of the experiment, we only changed the types of datasets, and did not change the specific parameters of the three models and the experimental process, so as to exclude the adverse effects of other possible factors. In addition to the 50ETF, we also selected two other stock index options, namely, 300ETF(510300) and 300ETF(159919). The number of datasets and variables are basically consistent with the 50ETF described above. The experimental process is still divided into the mixed pricing of call and put options in the sample, and the separate pricing of call and put options out of the sample. The pricing effect is analyzed according to the evaluation metrics RMSE and IVRMSE. Detailed experimental data are shown in Table 6, 7, 8.

From the data in Table 6, 7, 8, it can be concluded that the pricing effect of the model has nothing to do with the type of dataset, and the relevant results are basically consistent with those above.

Table 8. IVRMSE for pricing results of out-sample put option.

Dataset Name	Heston model	Bi-Heston model	XGBoost model
50ETF	0.0072	0.0146	0.0868
300ETF(510300)	0.0232	0.0207	0.2002
300ETF(159919)	0.0073	0.0068	0.1271

Table 9. The data of 50ETF Call options.

Panel A: Number of Call options[1]					
Moneyness[2]	DTM[3] <30	30≤DTM<90	90≤DTM<180	DTM≥180	ALL
DOTM(S/K<0.94)	4171	5826	4694	1966	16657
OTM(0.94≤S/K<0.97)	1693	2578	1992	1224	7487
ATM(0.97≤S/K<1.03)	3510	5298	4169	2709	15686
ITM(1.03≤S/K<1.06)	1198	2046	1789	1089	6122
DITM(S/K≥1.06)	2491	4169	3428	1629	11717
ALL	13063	19917	16072	8617	57669
Panel B: Average Call Price					
Moneyness	DTM<30	30≤DTM<90	90≤DTM<180	DTM≥180	ALL
DOTM(S/K<0.94)	0.003	0.016	0.044	0.083	0.029
OTM(0.94≤S/K<0.97)	0.014	0.039	0.080	0.117	0.057
ATM(0.97≤S/K<1.03)	0.055	0.086	0.131	0.166	0.105
ITM(1.03≤S/K<1.06)	0.146	0.166	0.205	0.238	0.186
DITM(S/K≥1.06)	0.400	0.392	0.369	0.363	0.383
ALL	0.108	0.132	0.158	0.187	0.142
Panel C: Average Implied Volatility from Call Options[4]					
Moneyness	DTM<30	30≤DTM<90	90≤DTM<180	DTM≥180	ALL
DOTM(S/K<0.94)	0.268	0.211	0.182	0.166	0.212
OTM(0.94≤S/K<0.97)	0.183	0.163	0.148	0.142	0.160
ATM(0.97≤S/K<1.03)	0.188	0.158	0.139	0.129	0.155
ITM(1.03≤S/K<1.06)	0.264	0.171	0.138	0.126	0.171
DITM(S/K≥1.06)	0.550	0.270	0.161	0.143	0.279
ALL	0.288	0.199	0.158	0.142	0.199

[1] We use the closing price of options for each business day from 4 January 2016 to 31 December 2020.
[2] The moneyness and maturity filters used by Bakshi [17] are applied here as well.
[3] DTM represents the time to maturity of the option.
[4] The implied volatilities are extracted using the Black-Scholes formula.

Table 10. The data of 50ETF Put options.

Panel A: Number of Put options[1]					
Moneyness[2]	DTM[3] <30	30≤DTM<90	90≤DTM<180	DTM≥180	ALL
DOTM(S/K<0.94)	2735	4658	4694	1966	14053
OTM(0.94≤S/K<0.97)	1188	2363	1993	1224	6768
ATM(0.97≤S/K<1.03)	3384	5387	4281	2824	15876
ITM(1.03≤S/K<1.06)	1709	2482	2078	1321	7590
DITM(S/K≥1.06)	5096	7600	6715	2633	22044
ALL	14112	22490	19761	9968	66331
Panel B: Average Put Price					
Moneyness	DTM<30	30≤DTM<90	90≤DTM<180	DTM≥180	ALL
DOTM(S/K<0.94)	0.382	0.391	0.409	0.384	0.394
OTM(0.94≤S/K<0.97)	0.138	0.157	0.196	0.229	0.178
ATM(0.97≤S/K<1.03)	0.045	0.075	0.121	0.155	0.095
ITM(1.03≤S/K<1.06)	0.011	0.032	0.070	0.100	0.050
DITM(S/K≥1.06)	0.003	0.012	0.033	0.067	0.027
ALL	0.099	0.123	0.162	0.179	0.138
Panel C: Average Implied Volatility from Put Options[4]					
Moneyness	DTM<30	30≤DTM<90	90≤DTM<180	DTM≥180	ALL
DOTM(S/K<0.94)	0.432	0.303	0.267	0.231	0.306
OTM(0.94≤S/K<0.97)	0.224	0.192	0.203	0.212	0.205
ATM(0.97≤S/K<1.03)	0.164	0.170	0.188	0.198	0.178
ITM(1.03≤S/K<1.06)	0.173	0.168	0.180	0.189	0.176
DITM(S/K≥1.06)	0.264	0.201	0.192	0.199	0.215
ALL	0.259	0.212	0.209	0.205	0.220

[1] We use the closing price of options for each business day from 4 January 2016 to 31 December 2020.
[2] The moneyness and maturity filters used by Bakshi [17] are applied here as well.
[3] DTM represents the time to maturity of the option.
[4] The implied volatilities are extracted using the Black-Scholes formula.

5. Conclusions

In this paper, the Bi-Heston pricing model is derived and a closed-form analytical solution is given. A relevant empirical analysis is also done to compare and analyse the advantages and disadvantages of the parametric pricing model with the machine learning pricing model. Due to the

high latitude and complex nature of the estimated reference function of the parametric model, we use a particle swarm algorithm based on parallel computing for parameter calibration, and the actual results show that the parameter calibration is excellent. In the final empirical analysis, we obtain the following conclusions: (1) In terms of in-sample pricing effects, the parametric pricing model is significantly better than the machine learning model, while the Bi-Heston model is slightly better than the Heston model. (2) In terms of out-of-sample pricing, the machine learning model is inferior to the parametric model for call options, while the Bi-Heston model is significantly better than the other two models for put options, and the other two models are similar. (3) In the robustness analysis of the three pricing models, the machine learning model shows strong instability, while the Bi-Heston model shows a more stable side. (4) When plotting the implied volatility images of the three pricing models, the machine learning model fails to capture the phenomenon of volatility smiles.

Declarations

Author contribution statement

Zhaoli Jia, Ph.D.: Conceived and designed the experiments; Wrote the paper.

Qiang Luo: Conceived and designed the experiments; Performed the experiments; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Hongbo Li: Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data.

Yongxin Wu: Analyzed and interpreted the data.

Funding statement

This work was supported by National Natural Science Foundation of China [72071068]; Innovation and entrepreneurship project for College Students [202110359075].

Data availability statement

The data that has been used is confidential.

Declaration of interests statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

References

[1] F. Black, M. Scholes, The pricing of options and corporate liabilities, J. Polit. Econ. 81 (3) (1973) 637–654.
[2] S.L. Heston, A closed-form solution for options with stochastic volatility with applications to bond and currency options, Rev. Financ. Stud. 6 (2) (1993) 327–343.
[3] J.C. Cox, J.E. Ingersoll, S.A. Ross, A theory of the term structure of interest rates, Econometrica 53 (1985) 385–408.
[4] P. Christoffersen, S. Heston, K. Jacobs, Review of the shape and term structure of the index option smirk: why multifactor stochastic volatility models work so well, Manag. Sci. 55 (12) (2009) 1914–1932.
[5] F.D. Rouah, The Heston Model and Its Extensions in Matlab and C, John Wiley & Sons, 2013.
[6] S. Fallah, F. Mehrdoust, On the existence and uniqueness of the solution to the double Heston model equation and valuing lookback option, J. Comput. Appl. Math. 350 (2019) 412–422.
[7] F. Mehrdoust, N. Saber, A.R. Najafi, Modeling asset price under two-factor Heston model with jumps, Int. J. Appl. Comput. Math. 3 (4) (2017) 3783–3794.
[8] Y. Cui, S. del Bano Rollin, G. Germano, Full and fast calibration of the Heston stochastic volatility model, Eur. J. Oper. Res. 263 (2) (2017) 625–638.
[9] F. Cacace, A. Germani, M. Papi, On parameter estimation of Heston's stochastic volatility model: a polynomial filtering method, Decis. Econ. Finance 42 (2) (2019) 503–525.
[10] C. Ma, Z. Liu, Z. Cao, W. Song, J. Zhang, W. Zeng, Cost-sensitive deep forest for price prediction, Pattern Recognit. 107 (2020) 107499.
[11] N.A. Hitam, A.R. Ismail, F. Saeed, An optimized support vector machine (svm) based on particle swarm optimization (pso) for cryptocurrency forecasting, Proc. Comput. Sci. 163 (2019) 427–433.
[12] R. Yang, L. Yu, Y. Zhao, H. Yu, G. Xu, Y. Wu, Z. Liu, Big data analytics for financial market volatility forecast based on support vector machine, Int. J. Inf. Manag. 50 (2020) 452–462.
[13] D. Kumar, S.S. Meghwani, M. Thakur, Proximal support vector machine based hybrid prediction models for trend forecasting in financial markets, J. Comput. Sci. 17 (2016) 1–13.
[14] T. Le, B. Vo, H. Fujita, N.T. Nguyen, S.W. Baik, A fast and accurate approach for bankruptcy forecasting using squared logistics loss with gpu-based extreme gradient boosting, Inf. Sci. 494 (2019) 294–310.
[15] C.F. Ivaşcu, Option pricing using machine learning, Expert Syst. Appl. 163 (2021) 113799.
[16] Y.C. Chang, K.H. Chang, G.J. Wu, Application of extreme gradient boosting trees in the construction of credit risk assessment models for financial institutions, Appl. Soft Comput. 73 (2018) 914–920.
[17] G. Bakshi, C. Cao, Z. Chen, Empirical performance of alternative option pricing models, J. Finance 52 (5) (1997) 2003–2049.