Expert Systems with Applications 41 (2014) 5227–5237

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Parametric models and non-parametric machine learning models for predicting option prices: Empirical comparison study over KOSPI 200 Index options

Hyejin Park [a], Namhyoung Kim [b,*], Jaewook Lee [c]

[a] Department of Industrial and Management Engineering, Pohang University of Science and Technology (POSTECH), San 31 Hyoja, Pohang 790-784, South Korea
[b] Department of Applied Statistics, Gachon University, 1342 Seongnamdaero, Sujeong-gu, Seongnam-si, Gyeonggi-do 461-701, South Korea
[c] Department of Industrial Engineering, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 151-744, South Korea

ARTICLE INFO

Keywords:
Option pricing
Gaussian processes
Support vector machines
Artificial neural network
Black–Scholes model
Heston model
Merton model

ABSTRACT

We investigated the performance of parametric and non-parametric methods concerning the in-sample pricing and out-of-sample prediction performances of index options. Comparisons were performed on the KOSPI 200 Index options from January 2001 to December 2010. To verify the statistical differences between the compared methods, we tested the following null hypothesis: two series of forecasting errors have the same mean-squared value. The experimental study reveals that non-parametric methods significantly outperform parametric methods on both in-sample pricing and out-of-sample pricing. The outperforming non-parametric method is statistically different from the other models, and significantly different from the parametric models. The Gaussian process model delivers the most outstanding performance in forecasting, and also provides the predictive distribution of option prices.

1. Introduction

Despite the recent global financial crisis, options are still one of the most important financial instruments in risk management. The recent surge in option trading volume demonstrates that options are crucial means to hedge against risk. To hedge risks using options, an investor must be aware of the fair price of the option when buy or sell. Option pricing initiated by the Black–Scholes model suffers from several unrealistic assumptions that conflict with the characteristics of option data traded in the real market. For example, the key assumption that the return of stock prices follows a geometric Brownian motion with constant drift and volatility is wrong. The implied volatility from the real market exhibits a “volatility-smile” pattern with different values depending on the time-to-maturity and strike prices. Thus, other advanced alternative parametric models were developed. These models are generally divided into two classes, namely, parametric and non-parametric.

Researchers have developed parametric models that can explain the volatility smile in the market data under no-arbitrage conditions. One widely used alternative parametric model is the stochastic volatility model, which assumes that volatility follows a random diffusion process (Heston, 1993). Another widely used model is the jump-diffusion model (Merton, 1976), wherein the movement of underlying assets follows a stochastic process with jumps to Brownian motion. Since the 1990s, other advanced models have been studied actively including more generalized version of Lévy models (Carr, Gaman, Madan, & Yor, 2003; Madan, Carr, & Chang, 1998).

Along with the development of IT technology, non-parametric models based on machine learning techniques have been developed to determine the option prices of real market data. Neural network (NN) models have been used to initiate these attempts, and have been extensively discussed in option pricing. Researchers have investigated a variety of non-parametric methodologies for option pricing, since Hutchinson, Lo, and Poggio (1994) demonstrated that NN models obtain a positive result compared with the performance of out-of-sample pricing and delta-hedging (Amilon, 2003; Gencay & Qi, 2001; Gradojevic, Gencay, & Kukolj, 2009; Han & Lee, 2008; Lajbcygier & Connor, 1997; Liang, Zhang, Xiao, & Chen, 2009; Park & Lee, 2012; Yang & Lee, 2011). Most researchers argued that prediction accuracy is improved with constant historical volatility compared with that of the Black–Scholes model.

This paper investigated the efficiencies of non-parametric machine learning techniques on financial option pricing compared with parametric methods. This study is not limited to traditional

* Corresponding author.
E-mail address: nhkim@gachon.ac.kr (N. Kim).

http://dx.doi.org/10.1016/j.eswa.2014.01.032
0957-4174/

comparison effects on forecasting, and verifies the power of non-parametric methods for prediction (including recently developed state-of-the-art machine learning techniques such as support vector machines and Gaussian processes). To verify the performance of each model, we conducted an extensive empirical study by applying state-of-arts parametric and non-parametric methods to heavily traded index options. The comparison was executed using three measures, namely, measurement of option pricing error with various methods, measurement of option price prediction, and statistical tests of each forecasting series.

The motivation and contribution of this study are as follows. It is the first comprehensive study to compare the performance of both state-of-arts parametric no-arbitrage models and non-parametric machine learning models including Gaussian process regression model, the most popular Bayesian kernel model, for predicting options prices especially on KOSPI 200 Index options from January 2001 to December 2010. So far, many comparison studies have been made to compare the performance of only no-arbitrage parametric models or only non-parametric machine learning models or some simple parametric and non-parametric methods. In addition, the empirical results made over KOSPI 200 Index options reveal that non-parametric methods show statistically better performance compared with parametric ones on both in-sample pricing and out-of-sample pricing. Especially the Gaussian process regression method delivers the overall outstanding performance in prediction accuracy as well as in its capability to provide the predictive distribution of option prices.

This paper is organized as follows. Section 2 provides a brief background of parametric and non-parametric derivative pricing models. Section 3 describes the design of the experiment for comparing the models. Section 4 presents the simulation results using real market data of the KOSPI 200 Index option and the statistical performance of the methods. Section 5 concludes.

2. Preliminaries

This section presents the evolution of non-parametric methodologies in option pricing. Thereafter, a brief review of each methodology prior to the design of the experiments is presented.

2.1. Non-parametric methods for option pricing

Since the 1990s, vast amounts of data have been accumulated in data warehouses and analyzed by high-performance computers. Rather than rely on mathematical models, attempts have been made to build financial pricing models of supervised learning using the characteristics of the data.

Hutchinson et al. (1994) showed that NN models exhibit positive performance in out-of-sample pricing and delta-hedging compared with ordinary least squares, Radial basis function networks, and Black–Scholes formulae. This researcher divided the daily data of the S&P 500 future option from 1987 to 1991 into 10 subperiods and trained the NN in the formal subperiod to test the next subperiod. These NN models consisted of four hidden nodes using the sigmoid activation function. Lajbcygier and Connor (1997) improved the prediction accuracy using a hybrid-artificial NN by understanding the residuals between option prices and modified Black–Scholes model prices; daily Australian SPI option data was used from 1992 to 1994 to verify the performance of the model. Gencay and Qi (2001) suggested that NN models should used Bayesian regulation, early stopping, and bagging to increase generalization. The Bayesian regulation exhibited significantly higher accuracy in out-of-sample pricing and hedging for the daily S & P 500 index option data. Amilon (2003) predicted the bid call price and ask call price by modeling a NN with 10 hidden nodes using the daily call options of the Swedish stock index and compared the results with those of the Black–Scholes model. Gradojevic et al. (2009) improved the prediction performance using a modular NN, containing three to nine modules regarding moneyness and time-to-maturity. Liang et al. (2009) reduced the pricing errors of conventional option pricing methods such as the binomial tree, finite difference method, and Monte–Carlo method by using NN and support vector regression (SVR). The estimated prices of each model were used as input factors to monitor the option market data of 122 firms in Hong Kong from 3d to 23d to forecast the next day. However, a data-snooping problem occurred wherein the model only learned using a short-time window instead of a number of firms.

Many kinds of NN option-pricing models estimate only a point forecast of option prices. Han and Lee (2008), Yang and Lee (2011), and Park and Lee (2012) grafted Gaussian process (GP) models as popular Bayesian kernel machines to forecast the distribution of option prices. The GP models overcame the typical shortcoming of NN models and improved the option pricing performance. Han and Lee (2008) proved that GP models achieve more accurate performance than advanced NN models for forecasting equity-linked warrant data, which have larger values than traditional, theoretical option values. Yang and Lee (2011) forecasted the implied volatility of the next day using one-day KOSPI 200 call prices of ELW and then used the Black–Scholes equation and predicted volatility to compute the option prices. The predicted distribution of prices did not preserve the original distribution of price ranges especially for deep in-the-money (ITM) options. Park and Lee (2012) suggested using the positive GP to predict the distribution of call option prices with non-negative values and verify the performance of out-of-sample pricing using KOSPI 200 Index option prices from 2008 to 2010. Table 1 shows the summary of the literatures for non-parametric methods with their data prescriptions.

2.1.1. Artificial neural network

NN model refers to the structure of the network connecting with a simple function of neurons like large-scale human nerve tissue. Artificial neural network (ANN) is composed of the following layers; the input layer consisting of nodes corresponding to each input variable, the output layer corresponding to the target variable, and the hidden layer treated as a non-linear function of the linear combination of the values that are passed from the input layer and other hidden layers. Each function $\phi(x)$ of node is a non-linear weighted sum of other functions $h_i(x)$ given by

$$\phi(x) = K\left(\sum_i w_i h_i(x)\right), \quad (1)$$

where K represent the activation function such as hyperbolic tangent and logistic function.

ANN model is optimized by the well-known back-propagation algorithm depending on the initial values. It minimize the sum of the squared error between the output value of network and the real target value. As the network training method, (Dan Foresee & Hagan, 1997) proposed the Bayesian regulation back-propagation method. This algorithm updates weight and bias values according to Levenberg–Marquardt optimization and generalizes a combination of weight for minimizing the squared error of the network to determine the correct combination. However, it is also difficult to interpret to identify the relationship between input variables and output variables.

2.1.2. Support vector regression

Support vector machine (SVM) is one of the supervised learning method to find a hyperplane of separation for given data using kernel trick. This technique is successfully applied to non-linear classification and regression problems as well as clustering (Jung,