Modeling Implied Volatility

Rongjiao Ji
Instituto Superior Técnico, Lisboa, Portugal

November 2017

Abstract

With respect to the valuation issue of a derivative's contracts in finance, the volatility of the price of the underlying asset is often unknown. Volatility is a measure of randomness, allowing us to assess how uncertain the price movement is in the future.

In this work we first derive the implied volatility for each contract, using the Black-Scholes formula. Since it is not possible to determine the implied volatility analytically, one needs to resort to numerical methods. Here we propose to use the bisection method to compute the estimate value of the implied volatility. The determination of implied volatility is discussed afterwards, using the future price and the asset price as input value in the Black-Scholes formula. In addition, two calculation methods are presented, in order to increase the accuracy of the estimates obtained.

Several models are presented for adjustment of the obtained values, namely models based on linear quantile regression and random forests. Using these models, we may forecast the implied volatility, and then use these values to predict the price of an option contract in the future.

Keywords: implied volatility, Black-Scholes formula, quantile regression, random forests

1. Introduction

Black and Scholes [2] and Merton [11] proposed in 1973 the Black-Scholes model to fit the market prices of options. The Black-Scholes formula is applied to transform the market prices into an expression in terms of implied volatility [9]. Implied volatility is the square root of the quadratic variation of the asset's log price process, defined as a parameter in the option pricing model Black-Scholes formula that yields the price of a particular option. Given observed option prices in the market, the value of implied volatility can be deduced by matching the corresponding Black-Scholes price with the market price [4]. The implied volatility from the real market shows a pattern named volatility smile, influenced by time to maturity and strike price of the option.

Different from the theoretical parametric models, such as stochastic volatility model [5] and jump-diffusion model [12], non-parametric models based on machine learning techniques, which has less restrictions in usual cases, have been developed for pricing options [14]. Additionally, quantile method [6], beyond the conditional mean, gives more information about the quantile distribution of response variable as a function of the explanatory variables. The combination of non-parametric models and quantile methods, in this way, has less restrictions on the required assumptions and can bring a new perspective for estimating the implied volatility [16].

Our work firstly implements several financial criterion to a dataset composed by options, future contracts and discount values, and secondly builds the quantile linear and random forest for the computed implied volatility, in order to generate a clear cognition of the option trading process and to predict the Black-Scholes implied volatility.

1.1. Dataset Description

The dataset used in this work is real option trades provided by BNP Paribas bank, including information over three types of derivatives (option contracts, discount values and future contracts). These three derivatives are all based on the same underlying asset, the EURO STOXX50 index, which is an Europe's leading Blue-Chip index for the Eurozone.

Specifically, as the main object, option contracts include trades in 424 different dates, ranged from January 02, 2014 to October 29, 2015. Almost all the maturities of options concentrate on Friday, and the months with more maturities are March, June, September and December. There are only ten distinct maturities of future contracts and these maturities spread over the third Friday of every quarter from 2014 to the mid year of 2016. These kind of maturities are denoted as 'major maturities', and Triple Witching Phenomenon occurs at the major maturities. At this time, the contracts for stock index futures, stock index options and stock options expire on the same day. Trading activity is more active and volatility is larger because contracts that are allowed to expire may necessitate the purchase or sale of the underlying security.

In Table 1 we present the descriptions of relevant terms. In the original data set, some non-relevant information is discarded (such as the lot size, bid and ask size). Moreover we also include four variables which are marked by star and are relevant for the rest of the study, namely: time to maturity, moneyness, constructed price and interest rate.

Table 1: The descriptions of relevant terms which are provided from three types of derivatives (option, future and discount). The number of observations for each type is shown in brackets. The notation star(*) indicates that the marked variables are generated afterwards.

Term	Description	Example Set
Date t	The beginning date of a derivative contract.	2014-01-03
Maturity T	The expiration date of a derivative contract.	2014-03-21
Option (312339)		
Type	Call options C or put options P.	Both
Strike K	The price paid for the asset if the option is exercised.	Shown in Figure 1
Bid price	The highest price a buyer is willing to pay.	
Ask price	The lowest price a seller is willing to accept.	
*Moneyness M	Deduced by $M = \frac{KD}{S} = \frac{K}{F}$ and used to label options.	
*Time to maturity τ	The number of days from date to maturity.	77
*Constructed price $\hat{S}$	Constructed for the true market price S by $\hat{S} = FD$.	3062.73
Future Contracts (1153)		
Bid Price	The highest price a buyer is willing to pay	3061
Ask Price	The lowest price a seller is willing to accept	3062
Discount Value (7255)		
Discount Value D	Discount from future back to current time.	0.9996
Interest Rate r	Deduced by $D = e^{-r(T-t)}$.	5.5×10^{-6}

Here we display specifically the trades (denoted as 'Example Set') with trading date in January 03, 2014 and maturity in March 21, 2014, as an example to explain the micro structure of the dataset and the following operations. The information in detail is shown in the right side of Table 1.

Combination of options

Due to the Put-Call Parity, volatility is the same for a call option and a put option with the same combination of date, maturity and strike. After combining the call with corresponding put options, we have now 107571 pairs of call and put options.

Following our 'Example Set', in Figure 1, we can see that, under the condition of fixing other variables, for call options, when strike value goes up, the price of call option declines. Oppositely, the price of put option increases. However, the range of call option prices is larger than the range of put option prices. One of the possible reasons might be that traders believe the market price would go up.

Information of index price

EURO STOXX50 index is the asset under study, i.e. the objective financial asset that our three derivatives are associated with. The information of historical index price can be easily obtained [1]. Note that the newly obtained index prices were captured at 18:00 CET in each trading day while

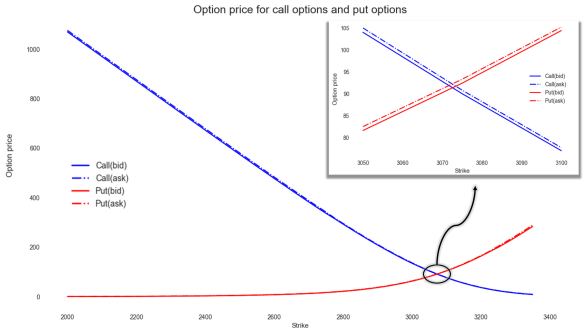


Figure 1: The tendencies of option prices of the call and corresponding put options with strike price in Example Set. The enlargement of the crosspoint of lines is shown in the top right.

the options contracts were gathered at 17:15 CET. Following the 'Example Set', the index price in Jan 03, 2014 is 3074.43, while the constructed price is 3062.73. The constructed price, by the way, is a non-standard terminology used in this work. Because it is deduced from the discounted future price, in theory it is supposed to be the best representation of the current asset price.

The constructed price should worth the same as current true price of underlying asset. Thus, we have two different situations depending on whether

a) Those options whose maturities belong to the major maturities of future contracts. Thus the index price obtained can variate a bit from the constructed price within 45 minutes at very active trading days, in particular at the major maturities due to Triple Witching Phenomenon.

b) Those options whose maturities do not belong to the major maturities of future contracts, or their maturities do belong to the major maturities but the influence of Triple Witching Phenomenon is limited. In this situation, the price of the underlying asset is very similar to the index price.

Thus the index price can be used as a supplement for those options which have no information about the relevant future price. It can give a rough direction for the investors although its accuracy is not assured.

1.2. Outline

The paper is structured as follows. Section 2 introduces most important criteria, the Put-Call Parity and Black-Scholes Formula. Section 3 interprets quantile linear regression and the main idea of decision trees and random forest. Section 4 demonstrates how to calculate the implied volatility. We generate the estimate prices as a function of implied volatility deduced by two forms of Black-Scholes formula and two methods of calculation, and then derive the implied volatility by the Bisection Method, given the market price. Section 5 mainly focuses on using quantile methods together

[1] From the website 'Yahoo Finance' `https://finance.yahoo.com/quote/\%5ESTOXX50E/history?period1=1388620800\&period2=1446076800\&interval=1d\&filter=history\&frequency=1d`

with linear regression and random forests method to model the subset which contains some significant features. The work concludes with Section 6, where we also point possible future work. We skip the introduction of financial products (more details in [2] and [17]) and some statistical tools (details in [1], [13], [10] and [6]) due to the limitation of pages.

2. Financial Theoretical Overview

We define that for a fixed date t, an European call option (whose call option price is $C(T,K)$ and the corresponding put option price is $P(T,K)$) with asset price S_t, maturity T and strike price K is defined as a contingent claim with payoff $max\{S_T - K, 0\}$ at maturity. The related future price for the underlying asset is $F(t,T)$, discount value is $D(t,T)$ and interest rate is r. **Put-Call Parity** describes the relationship for different portfolios, shown as:

$$C(T,K) - P(T,K) = S_t - Ke^{-r(T-t)},$$
or: $C(T,K) - P(T,K) = [F(t,T) - K]D(t,T).$

The **Black-Scholes Formula** [2] proposes the estimate value of the call option price $C(T,K)$ and put option price $P(T,K)$ as

$$C(T,K) = S_t\Phi(d_1) - Ke^{-r(T-t)}\Phi(d_2), \quad (1)$$
$$P(T,K) = -S_t\Phi(-d_1) + Ke^{-r(T-t)}\Phi(-d_2), \quad (2)$$
$$d_1 = \frac{ln(\frac{S_t}{K}) + (r + 0.5\sigma^2)(T-t)}{\sigma\sqrt{T-t}}, d_2 = d_1 - \sigma\sqrt{T-t},$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution, and σ is the volatility of returns of the underlying asset.

Another common form for Black-Scholes formula based on future price and discount value for call option price $C(T,K)$ and put option price $P(T,K)$ are proposed as:

$$C(T,K) = D(t,T)(F(t,T)\Phi(d_1) - K\Phi(d_2)), \quad (3)$$
$$P(T,K) = D(t,T)[\Phi(-d_2)K - \Phi(-d_1)F(t,T)], \quad (4)$$
$$d_1 = \frac{ln(\frac{F(t,T)}{K}) + 0.5\sigma^2(T-t)}{\sigma\sqrt{T-t}}, d_2 = d_1 - \sigma\sqrt{T-t}.$$

3. Statistical Theoretical Overview

In this section we do a brief overview of quantile regression and random forests.

3.1. Quantile Regression Method

Quantile regression is a generalization of linear regression where it models a quantile of interest as a linear function of the explanatory variables [6]. It is more robust then ordinary least squares, and it has shown that it can lead to good results when there are complex relations between variables.

Definition of quantile

Suppose that there is a dataset with n observations, $\{y_i, \mathbf{x}_i\}$ for $i = 1, \ldots, n$, where y_i is response variable and the explanatory variables $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$, and $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ is the coefficients vector.

Quantile regression focuses on the conditional quantiles of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$. Assume the distribution function for a real-valued response variable $\mathbf{Y}$ is $F_{\mathbf{Y}}(\mathbf{y}) = P(\mathbf{Y} \leq \mathbf{y})$, then the τ-th quantile of $\mathbf{Y}$ is defined as the minimum value of $\mathbf{y}$ which satisfies $F(\mathbf{y}) \geq \tau$, i.e.,

$Q_{\mathbf{Y}}(\tau) = F_{\mathbf{Y}}^{-1}(\tau) = inf\{\mathbf{y} : F(\mathbf{y}) \geq \tau\}, 0 < \tau < 1.$

Definition of quantile loss function

Given a quantile $\tau \in (0,1)$, the L_1-norm quantile regression is used to minimize the loss function [7]:

$$L(y_i, \hat{y}_i)_\tau = \sum_{i=1}^{n} \rho_\tau(y_i - \hat{y}_i), \quad (5)$$

where $\hat{y}_i$ is the estimate values and $\rho_\tau(y_i - \hat{y}_i)$ (named as check function in [6]) is defined as:

$$\rho_\tau(y_i - \hat{y}_i) = \begin{cases} \tau(y_i - \hat{y}_i), & y_i - \hat{y}_i > 0 \\ -(1-\tau)(y_i - \hat{y}_i), & \text{otherwise} \end{cases}. \quad (6)$$

The quantile regression is going to be discussed in more detail in Section 3.1.1, and in Section 3.1.2 estimation methods based on decision trees and random forest are briefly reviewed.

3.1.1 Linear regression

Linear quantile regression fits a conditional quantile of the response variable by a linear function $\mathbf{x}^\top\beta$. In Table 2 we compare the general quantile linear regression with the most common alternatives: ordinary least squares regression and least absolute deviation regression [15].

Table 2: Comparison between ordinary least squares, least absolute deviation and quantile regression methods.

Ordinary Least Squares (OLS)
Conditional mean function $E(\mathbf{Y}\|\mathbf{X} = \mathbf{x}) = \mathbf{x}^\top\beta$.
Loss function: $\sum_{i=1}^{n}(y_i - \mathbf{x_i}^\top\beta)^2$,
minimize the sum of square of residuals.
Estimates $\hat{\beta} = \arg\min_{\beta\in\mathbb{R}^{p+1}} \sum_{i=1}^{n}(y_i - \mathbf{x_i}^\top\beta)^2$.
Least Absolute Deviation (LAD)
Conditional median function $Q_{\tau=0.5}(\mathbf{Y}\|\mathbf{X} = \mathbf{x}) = \mathbf{x}^\top\beta(\tau = 0.5)$.
Loss function: $\sum_{i=1}^{n}\|y_i - \mathbf{x_i}^\top\beta\|$,
minimize the sum of absolute errors.
Estimates $\hat{\beta} = \arg\min_{\hat{y}_i\in\mathbb{R}^{p+1}} \sum_{i=1}^{n}\|y_i - \mathbf{x_i}^\top\beta\|$,.
τ-Quantile Regression
Conditional quantile function $Q_\tau(\mathbf{Y}\|\mathbf{X} = \mathbf{x}) = \mathbf{x}^\top\beta(\tau)$.
Loss function: $\sum_{i=1}^{n}\rho_\tau(y_i - \mathbf{x_i}^\top\beta)$,
minimize the sum of weighted absolute errors.
Estimates $\hat{\beta} = \hat{\beta}(\tau) = \arg\min_{\beta\in\mathbb{R}^{p+1}} \sum_{i=1}^{n}\rho_\tau(y_i - \mathbf{x_i}^\top\beta)$.

3.1.2 Tree-based regressors

If the relationship between response and explanatory variables is highly non-linear or complicated,

the tree-based models may give better results and explanations than linear regression.

Decision Tree

Decision tree is a non-parametric supervised learning method in the form of a tree structure. It splits a dataset from the entire space (denoted as R) into several non-overlapping regions (denoted as $R_j, j = 1, \ldots, J$) from the top of the tree to each leaf in the bottom by a series of binary if-then rules. These rules identify distinct regions where the observations inside share the most homogeneous responses to explanatory variables. Due to the fact that it is difficult to take every possible partition of the dataset into J regions into account, we need to apply a top-down, greedy recursive binary splitting approach to split successively further down to the leaf and to choose the best split at each step of tree-building process. In each internal node, the dataset is split and explanatory variables are judged to minimize the prediction error. The leaves, named terminal nodes, represent the final division of regions. At a leaf node, the mean of the response values assigned to that node is the predicted value returned by the decision tree.

This work uses CART, namely Classification and Regression Tree algorithm (please check [3] for more details). In order to perform this approach, for a explanatory variable X_j, there is a cutpoint s such that the variable can be split to two regions, either the region where X_j is smaller than s, i.e. $\{X_j < s\}$ or the region where X_j is greater or equal to s, i.e. $\{X_j \geq s\}$. The best cutpoint for each variable $X_j, j = 1, \ldots, p$ is chosen such that the tree has the lowest sum of square of residuals. We define two regions $R_1(j, s) = \{X_j : X_j < s\}$ and $R_2(j, s) = \{X_j : X_j \geq s\}$,, and we want to find the best pair of (j, s) to minimize

$$\sum_{i:x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2,$$

where $\hat{y}_{R_1}$ ($\hat{y}_{R_2}$) represents the mean response over all observations in $R_1(j, s)$ ($R_2(j, s)$).

Finally, it is known that decision trees can cause the over-fitting problem if the tree is too complex and full of details, leading to bad performance on prediction of new observations. It is necessary to set constraints on tree size or to prune the grown tree. (Note that in this work, we mainly use decision trees following the random forest method mentioned in next subsection. Therefore the pruning is not necessary and its introduction is omitted here.)

3.1.3 Random Forest

Random forest is an ensemble learning method generated by generating many decision trees on selected training samples. Random forests employs randomness each time it selects a different bootstrap sample of the train set for building each tree, and each node is split using the best variable among a subset of variables which is randomly chosen. This strategy turns out to perform well. It is robust and prevents over-fitting [3].

The algorithm for generating a random forest regression model is as follow [8]:

1. Before building each tree, firstly draw several samples from training set by bootstrap resampling method (with replacement) from the original data. Here we denote the number of samples as n_{tree}, indicating that n_{tree} trees will be built.

2. For each selected sample, grow a regression decision tree without pruning. At each node of a decision tree, a random subset of m_{try} variables is chosen from the entire p variables, as the target to be split at this node, rather than choosing the best split among all variables. Only one of these m_{try} variables can be used to generate the best split rule and corresponding subregions at this node. Thus for the selected training sets in step 1, n_{tree} trees are fully grown and combined into a random forest.

3. Predict new observations by aggregating the predictions of this n_{tree} trees. For regression, the prediction value of random forests for a new data point is the averaged response of all the trees.

Quantile Random Forest

The key difference between quantile regression forest and random forest is that: in each tree, random forest keeps only the mean response values of the observations that fall into each leaf node, and neglects all other information. In contrast, quantile regression forests keeps recording every quantile response values in the leaf node, not just their means, and assesses the conditional distribution based on this broader and more comprehensive information.

In prediction, each observation in test set will go through every tree and get a set of prediction values of the same size as the number of trees. Inside of this set, different quantiles of the prediction values can be reached, and we can even set a prediction interval with lower (higher) quantile prediction value as lower (upper) bound.

4. Computation and Analysis of Implied Volatility

The main goal of this section is to demonstrate how to calculate the implied volatility from the dataset introduced in Section 1.1 and analyze the results.

4.1. Calculation Processes

Thanks to Put-Call Parity, we can relate call options prices with put options prices for the same underlying asset, strike, date and maturity. Based on the combination of call and put options we did in Section 1.1, as a matter of choice, we have decided to focus on call options prices and thus the implied volatility mentioned afterwards refers to call options.

If we define a call option price generated by

Black-Scholes formula with the only unknown parameter implied volatility σ as $F_{BS}(\sigma)$, it should be equal to the true market value C_{market} of the call option contract with the same asset, date, maturity and strike price: $F_{BS}(\sigma) = C_{market}$.

We need to invert the Black-Scholes formula, because it is a non-linear function which does not have a closed form solution for implied volatility. Therefore we need to resort to numerical approximations and we consider the bisection method. Then we need to estimate:

$$\hat{\sigma} = \arg\min_{\sigma}[F_{BS}(\sigma) - C_{market}]. \quad (7)$$

What is more, in fact, the market price for an option contract C_{market} is not provided in the dataset. We have information concerning bid and ask prices, hereby denoted by C_{bid} and C_{ask} respectively, as the range of C_{market} (also known as bid-ask spread $[C_{bid}, C_{ask}]$). The true price of the option, C_{market}, may not exactly fall in this range, although most likely it will match inside. Therefore we propose the following two alternative methods and show the ideas briefly in Figure 2:

Method 1: Assume that $C_{market} = \frac{1}{2}(C_{bid} + C_{ask})$, and compute the implied volatility from this new value.

Method 2: First, compute the implied volatility σ_{bid} (σ_{ask}) using C_{bid} (C_{ask}) respectively as the input variable each time, and compute the resulting implied volatility as the average of σ_{bid} and σ_{ask}.

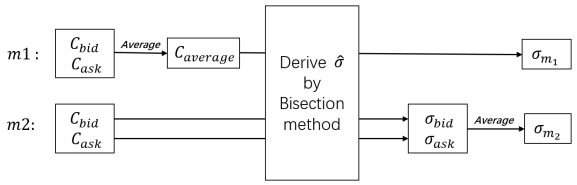


Figure 2: The framework of computation Method 1 (m1) and Method 2 (m2).

Next, two forms of Black-Scholes formula (**by using asset price** in Equation 1 and **by using future price** in Equation 3) are utilized to generate the implied volatility based on two combinations of the data sets. The specific formula is chosen depending on whether the price of asset or the future is involved.

In Figure 3 we show the framework of how we organize the related data sets, and as a result, there are in total four sets of implied volatility, noted as $IV_F^{m1}, IV_F^{m2}, IV_S^{m1}$ and IV_S^{m2}, where the subscript F (S) means that we use future price (asset price), and the superscript $m1$ ($m2$) means that we have used method 1 (method 2).

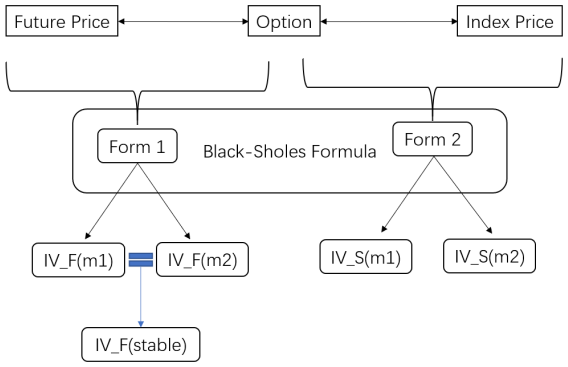


Figure 3: The framework of computation of implied volatility.

4.2. Analysis of the computed implied volatility

We start our analysis by showing the implied volatilities computed from the contracts in the subset mentioned in Section 1.1, which is named 'Example Set'. Afterwards, we focus on the most reliable result, IV_F, and present some plots and descriptive analysis.

4.2.1 Analysis based on the example

The 'Example Set' contains 42 contracts of options from Jan 3, 2014 to March 1, 2014, and the computed implied volatility is shown in Figure 4.

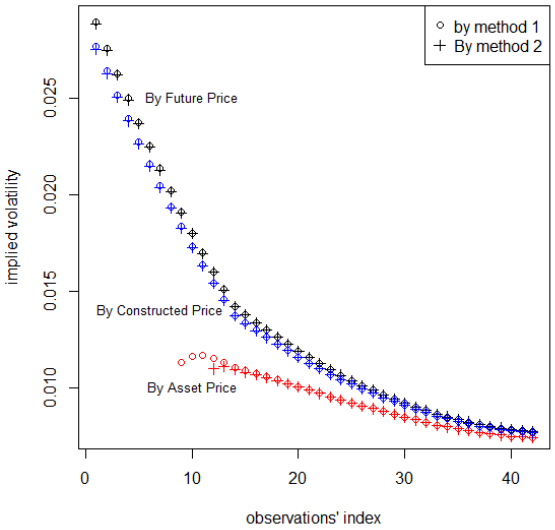


Figure 4: Different types of implied volatility based on the Example set. Implied volatility computed by future price, asset price and constructed price are shown in different colors. The results computed by method 1 (denoted as m1 in the legend) are denoted by the notation 'o', while the ones by method 2 (denoted as m2) are denoted by the notation '+'.

Note that although the constructed price is de-

duced from future prices, here the constructed price is used as input in the same form of Black-Scholes formula as index price.

We can see that the computation methods do not cause much difference. What is more, the implied volatilities based on future price and constructed price are similar as expected. However, the computed implied volatility based on index price is not very accurate especially when the implied volatility is larger in the major maturities, due to the fact that index price can variate quite a bit from deduced constructed price at the last 45 minutes. In a nutshell, the implied volatility based on future price is the most accurate and reliable result. Only when the future price is not available, which means that the maturity of an option is not one of the major maturities, the implied volatility based on the index price can be considered as a supplement.

4.2.2 Analysis on the entire IV_F

Now it is the time to have a general look of the entire implied volatility based on future price, IV_F'. Pair plot shown in Figure 5 illustrates the correlation of seven variables as follows: time to maturity, strike, constructed price, $IV(\text{m1})$, $IV(ask)$, $IV(bid)$ and $IV(\text{m2})$. The strike seems to follow a normal distribution, while the constructed price has two obvious normal distributions combined. All types of implied volatility obtained in this dataset follow similar distributions (almost normal distribution with right heavy tail) and they are highly correlated.

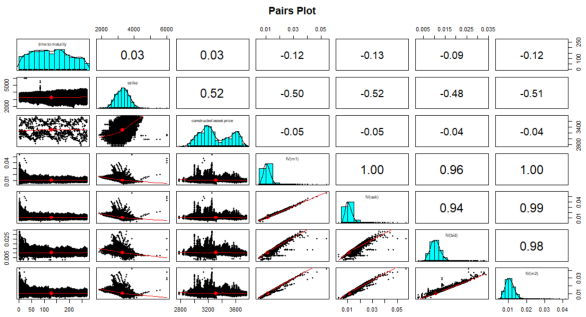


Figure 5: Pair plot of relevant variables. Here we plot the seven variables as follows: time to maturity, strike, constructed price, $IV(\text{m1})$, $IV(ask)$, $IV(bid)$ and $IV(\text{m2})$.

Stable and Unstable dataset in IV_F

Based on the phenomenon shown in Figure 6, we define two datasets as 'Stable' set (marked black) and 'Unstable' set (marked red) according to whether or not the observations are insensitive to the computation methods, i.e. whether the computed implied volatility remains relatively the same when the computation method changes. Here we define the meaning of 'relatively the same' as the difference between $IV_F(m1)$ and $IV_F(m2)$ smaller than 0.001. Here the cut-off point 0.001 is chosen manually by visualization.

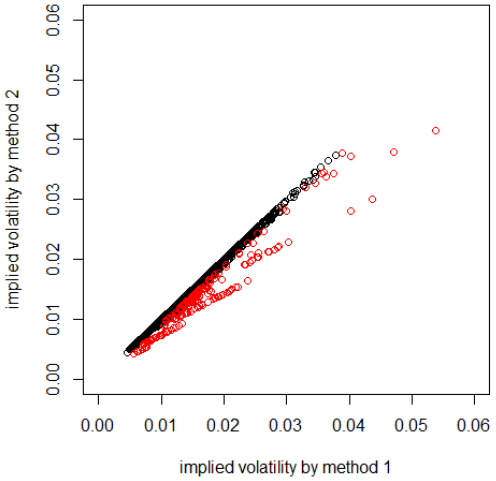


Figure 6: Comparison of implied volatility derived by both methods. Most of points marked black fall around the diagonal line. Only a small set of option contracts (only 0.65% of the whole dataset) is marked red.

After analysis, we find that the observations in the unstable set all belong to deep 'in-the-money' options or deep 'out-of-the-money' options, which are hard to be exercised successfully in reality. We decide not to focus on the unstable set anymore, and continue our study base on the stable set. The dataset mentioned afterwards refers to the stable set in the computed implied volatility with future price as input, denoted by IV_F^{stable}.

5. Modeling the Implied Volatility

Now we move forward, trying to fit the dataset concerned by regression models.

5.1. The process of choosing a subset

Due to the fact that one of the biggest challenges of this work is that the observations are highly mixed and overlapped, clear patterns of implied volatility are hard to attain.

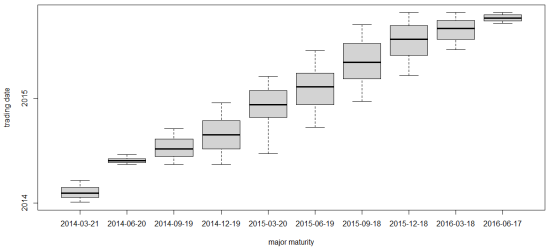


Figure 7: The range of trading dates for every major maturities.

Figure 7 shows, respectively, the range of the trading dates for those contracts which expire at ten distinct major maturities appearing in the dataset. The trading dates seem to concentrate on roughly three months before each maturity. Note that the information of options was gathered from January 2, 2014 to October 29, 2015. It is worth to mention that the records in the original dataset lacks the information from March 21, 2014 to May 16, 2015, due to unclear reasons.

According to our knowledge for financial activities, the movement of the market and actions of traders tend to have periodic variations. It is considered to be a good choice for studying the contracts with **maturity at September 18, 2015** as our current target, as it firstly contains relatively complete information, holding 17.7% of the total contracts. Secondly it is the last maturity before the end of the records, supposedly containing the most comprehensive information throughout the period of records.

5.2. Regression Modeling

As mentioned before, the implied volatility is influenced by time to maturity and strike price of the option. Thus our goal here is to generate alternative parametric models for implied volatility based on time to maturity and strike price.

We first separate the dataset into train set (75% of the entire observations) and test set (25% of the entire observations). Here we come up with both linear regression models and random Forest model to fit the chosen dataset by train set and compare the results at the end by test set.

Response variable (implied volatility) for train and test are positive-skewed, so we adopt Box-Cox method to calibrate the distribution of the implied volatility into normalization. The transformation parameter $\lambda = -1.59$ generated by train set is applied on response variable for both train and test sets to keep the transformed implied volatility staying in the same scale.

5.2.1 Linear Regression

We start now to fit linear models. Table 3 provides an overview in terms of the response variable together with original explanatory variables and ones generated afterwards.

The coefficient of determination R^2, as an effective performance statistic, is used to measure the goodness of the fitted models.

In fact, there is a more robust way to estimate the performance of models by prediction intervals for linear methods, because there is no requirement for the assumption of normally distributed residuals. In particular, using prediction interval can explain better and visualize straightforward the result. In this case, we set two quantiles 0.05 and 0.95 to create a 90% prediction interval, then calculate the response values in 5% quantile and 95% quantile and use them as the boundaries values. The observations in test set can be checked later on how correct it is for their true values to be contained in the prediction interval.

Table 3: An overview in terms of the response variable together with explanatory variables. Both notations and descriptions are displayed representatively.

	Variables	Descriptions
Response	$\mathbf{Y}$	Transformed implied volatility
Covariate	$\mathbf{X}_1$	Strike
	$\mathbf{X}_2$	Time to maturity
	$\mathbf{X}_3$	time to maturity $\times$ strike
	$\mathbf{X}_4$	1/ time to maturity
	$\mathbf{X}_5$	strike^2

The general expressions, coefficients and evaluations of three linear models are shown in Table 4. Specifically, the models are generated as the following steps:

As the initial model, $Model$ 1 is generated based on a simple addition of two original explanatory variables, time to maturity and strike. Taken into consideration of the relationship between strike and time to volatility, which seems to follow a non-linear curve. It indicates that the interaction of these two explanatory variables might be vital, so we combine the interaction term into the initial model and create $Model$ 2. Two more terms are added in $Model$ 3, which are the reciprocal value of time to maturity and the square of strike. Implied volatility seems to be inversely proportional to the time to maturity. What is more, the phenomenon of 'volatility smile' is visible, showing the square of strike and implied volatility are quite related.

Table 4: Summary of explanatory variables on three regression functions. The evaluation of adjusted R^2 for train and multiple R^2 for train and test are shown at the bottom of each case.

	Estimate	Std. Error	t value	Pr(>|t|)
Model 1: $Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \epsilon_i$,				
β_0	1394.7363	16.2883	85.63	0.0000
β_1	-0.5904	0.0045	-131.42	0.0000
β_2	-1.3634	0.0236	-57.68	0.0000
Multiple R^2: 0.8186, Adjusted R^2: 0.8185				
Model 2: $Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \epsilon_i$,				
β_0	2395.9840	34.4390	69.57	0.0000
β_1	-0.8872	0.0101	-87.97	0.0000
β_2	-7.5530	0.1940	-38.94	0.0000
β_3	0.0019	0.0001	32.10	0.0000
Multiple R^2: 0.8541, Adjusted R^2: 0.8540				
Model 3: $Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \beta_4 X_{4,i} + \beta_5 X_{5,i} + \epsilon_i$,				
β_0	1424.2046	107.3273	13.27	0.0000
β_1	-0.3830	0.0587	-6.53	0.0000
β_2	-6.4615	0.1970	-32.79	0.0000
β_3	0.0016	0.0001	27.58	0.0000
β_4	447.5932	23.7503	18.85	0.0000
β_5	-0.0001	0.0000	-8.14	0.0000
Multiple R^2: 0.8670, Adjusted R^2: 0.8668				

Table 4 illustrates that for the train set, the ad-

justed R^2 values for three models are gradually increase from 0.8185, 0.8540 to 0.8668, stating that the last model can explain 86.68% of the variance of the response variable. Compared with the multiple R^2 values of train set, test set has even better multiple R^2 values in these three models, and the highest one is 0.8695 from Model 3. Note that here we consider the multiple R^2 instead of adjusted R^2 because we want to compare it later on with the random forest, so we need to keep the evaluation in the same standard.

All explanatory variables appeared in three models are significant illustrated by very small p-values. It indicates that underestimation problem may exist, since for such a shifting response variable, we only use the information of time to maturity and strike. Beside the models mentioned above, we also tried some more complicated combinations of time to maturity and strike, like high-order polynomial regression. However, the resultant R^2 values do not have distinct improvement and the meanings of the models are harder to be explained. It is a trade-off on the complexity versus conciseness. Since we are going to apply both robust regression and quantile regression which can bring more perspectives of the relation between explanatory variables and response variable, we think it is a better choice to use a relative simple but efficient model ($Model$ 3), regarded as the best one till now.

Robust linear regression

The robust L_1-norm regression is one specific case of quantile regression when the quantile required is equal to 0.5. Compared with the classical linear model $Model$ 3, the coefficients and their standard deviation of the L_1 norm robust linear model have slight changes, but not significantly. The R^2 for train set is 0.8640 and for test set it 0.8669, lower by 0.2% compared with classical linear regression, showing that robust linear regression does not perform better than classical linear regression in this case.

Quantile linear regression

We extent our study from the median to a set of quantiles $\{0.05, 0.10, 0.25, 0.75, 0.90, 0.95\}$ of the response variable.

Figure 8 displays the change of coefficient and its confidence interval for each explanatory variable on different quantile response values. The red lines are the ordinary least squares estimate and its confidence interval. We find that when the quantile of implied volatility is larger, the absolute value of explanatory variable 'time to maturity' is smaller (i.e. it has negatively less influence on response variables) and the absolute value of explanatory variable 'time to maturity $\times$ strike' is also smaller, showing less influence in a positively way. Variable 'strike^2' has more significant impact on lower quantiles of implied volatility and then keep the same influence as estimated through ordinary least squares method.

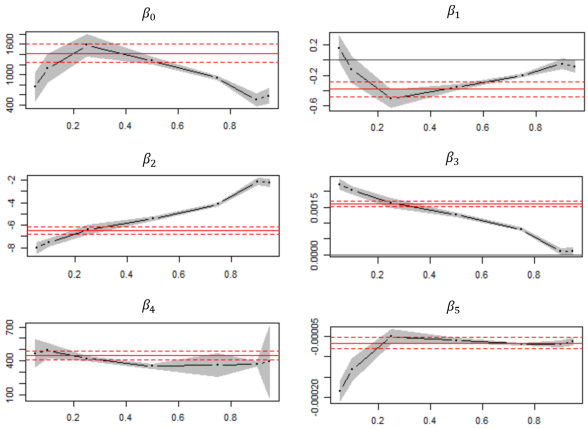


Figure 8: Plots for coefficients in quantile linear regression. Each black dot is the corresponding variable's coefficient for the quantile τ chosen in set $\{0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95\}$. The red lines are the ordinary least squares estimate and its confidence interval.

In order to display intuitively the effect and goodness of fit of quantile regressions, Figure 9 shows the fitted quantile values of the response variable respectively in lower quantiles, median and higher quantiles, for different strikes. We can see that the fitted values of lower and higher quantile regression can basically cover the boundaries of response variable and median regression catches the major feature of response variable as well.

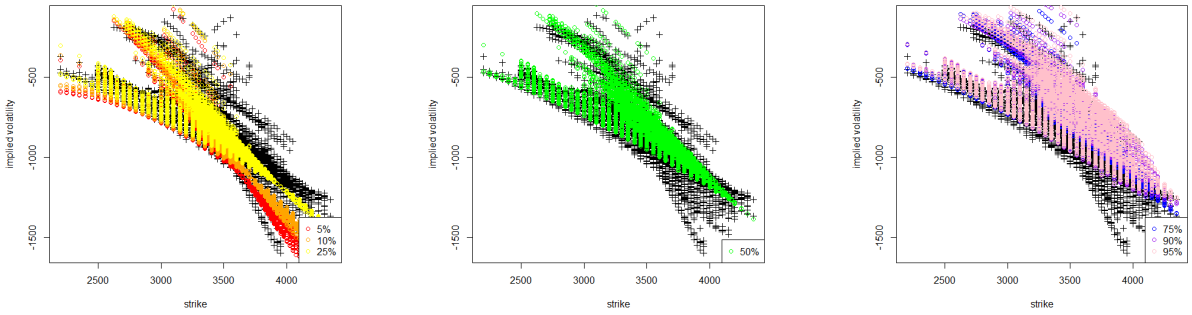
5.2.2 Quantile Random Forest

As for random forests, one of the main advantages is that: it can avoid the over-fitting problem as long as the number of trees is large enough. Therefore we build 1000 trees without the limitation on the depth, based on mean square error evaluation, and bootstrap resampling.

Table 5: Summary of two regression models. Correct rate means the probability of observations in test falling in 90% prediction interval

	Model Structure	Correct rate
Quantile Linear Regression	Model 3	83.27%
Quantile Random Forests	1000 trees	82.84%

A brief summary of two regression models is shown in Table 5. The 90% prediction interval is generated respectively for each model. There are 83.27% observations in test set covered correctly by prediction interval in linear regression. For random forests, 82.84% observations in test set are predicted correctly within 90% prediction interval. This correct rate is lower than the one got from linear re-



(a) Lower quantiles. (b) Median. (c) Higher quantiles.

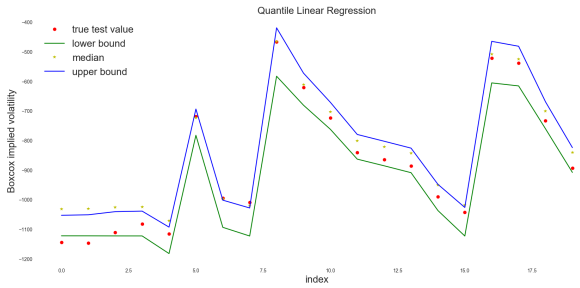
Figure 9: Quantile Regression for transformed implied volatility with strike. The quantile fitted values of response variable respectively in lower quantiles, median and higher quantiles are shown.

gression. However, the R^2 calculated based on the median prediction for this random forest is surprisingly 0.9907, stating that the model explains approximately 99.07% of the variance of the response variable. But note that since different observations contribute different amount of information, we do not have any way now to adjust R^2 from one observation occupying one degree of freedom.

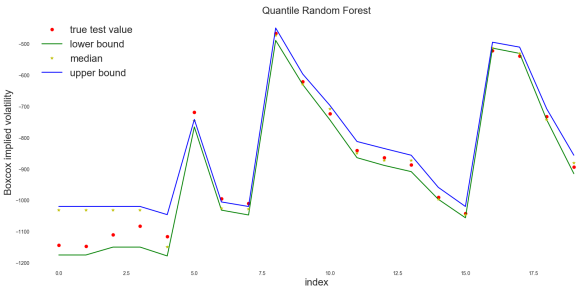
5.2.3 Comparison on 20 test observations

Because in test set we have 1411 observations, in order for better visualization, we choose the first 20 observations to show the prediction of these observations.

Figure 10(a) shows that for each observation, the quantile linear regression generates the boundaries of prediction interval with relatively equal width. For quantile random forests in Figure 10(b), the width of prediction interval is larger when the difference between true value and median prediction is larger, in which case the observation receives quite different prediction values from different trees (badly-fitted case). The prediction interval is narrower when the trees give similar prediction values which is a well-fitted case. The widths of prediction intervals for most observations in quantile random forests are generally tighter than those in quantile linear regression. This is probably the reason why quantile random forest has such a large R^2 but less correct rate regarding the prediction interval. Thus even though the points failing to fall into the prediction interval are five, four points among them are very close to the boundaries. While in quantile linear regressions there are four points beyond the prediction interval, only two are not close to the boundaries. This indicates that quantile random forest is more feasible to give accurate prediction and prediction interval for well-fitted observations, and for the badly-fitted points it sacrifices its accuracy (by increasing the prediction range) but increases the correct rate (i.e., to cover the true value under the prediction interval).



(a) Linear regression.



(b) Random Forests.

Figure 10: Two methods with median prediction and quantile boundaries.

6. Conclusions

The motivation for this work was to develop the understanding in the field of options trading, and to propose a way for computation and estimation of the important parameter, implied volatility.

6.1. Achievements

We firstly combined the options based on Put-Call Parity and prepared the environment for computation of implied volatility. Secondly, we derived the implied volatility using bisection method, through two forms of Black-Scholes formula and two calculation methods. Afterwards, we analyzed the com-

putation stability and compared the differences of computed implied volatilities. Thirdly we selected a subset and used it to estimate and predict the implied volatility by linear and tree-based regressions. Not only was the median regression considered, but quantile method were applied to establish the prediction interval and more general perspective of understanding the dataset. The features of both regression models were analyzed. Based on our analysis, the models we generated could explain most of the observations and give an acceptable prediction.

6.2. Directions for Future Work

As for fitting models, we have already been benefit from using the subset with significant features instead of studying the entire dataset. In the future the subsets can be chosen more specifically.

We also explored a bit on Gradient Boosting Regression Trees, but due to the tight time and computation limitation of the computer, it is not easy to calibrate the parameters through cross-validation. By setting the specific quantile for the Loss function, a Gradient Boosting Regression Tree used for explaining the distribution of response variable at that quantile is generated, thus its conditional quantile interval can be predicted as well. We believe that it could give better results in the further study.

In Section 4 we explained the computation of implied volatility with asset price involved IV_S. Due to the limitation of time, we have not found very straightforward way to compare the details of the two datasets. The two datasets are supposed to supplement with each other, which needs to have further attentions to dig out this latent relationship.

References

[1] D. Bachrathy and G. Stépán. Bisection method in higher dimensions and the efficiency number. *Periodica Polytechnica. Engineering. Mechanical Engineering*, 56(2):81, 2012.

[2] F. Black and M. Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3):637–654, 1973.

[3] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[4] T. E. Day and C. M. Lewis. Stock market volatility and the information content of stock index options. *Journal of Econometrics*, 52(1-2):267–287, 1992.

[5] S. L. Heston. A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The Review of Financial Studies*, 6(2):327–343, 1993.

[6] R. Koenker and G. Bassett Jr. Regression quantiles. *Econometrica: Journal of the Econometric Society*, pages 33–50, 1978.

[7] Y. Li and J. Zhu. L 1-norm quantile regression. *Journal of Computational and Graphical Statistics*, 17(1):163–185, 2008.

[8] A. Liaw, M. Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[9] J. D. MacBeth and L. J. Merville. An empirical examination of the black-scholes call option pricing model. *The Journal of Finance*, 34(5):1173–1186, 1979.

[10] P. C. Mahalanobis. On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India, 1936*, pages 49–55, 1936.

[11] R. C. Merton. Theory of rational option pricing. *The Bell Journal of Economics and Management Science*, pages 141–183, 1973.

[12] R. C. Merton. Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics*, 3(1-2):125–144, 1976.

[13] J. W. Osborne. Improving your data transformations: Applying the box-cox transformation. *Practical Assessment, Research & Evaluation*, 15(12):1–9, 2010.

[14] H. Park, N. Kim, and J. Lee. Parametric models and non-parametric machine learning models for predicting option prices: Empirical comparison study over kospi 200 index options. *Expert Systems with Applications*, 41(11):5227–5237, 2014.

[15] D. Pollard. Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, 7(2):186–199, 1991.

[16] D. Pradeepkumar and V. Ravi. Forecasting financial time series volatility using particle swarm optimization trained quantile regression neural network. *Applied Soft Computing*, 58:35–52, 2017.

[17] J. Voit. *The statistical mechanics of financial markets*. Springer Science & Business Media, 2013.