Curve-Fitting Method for Implied Volatility

Article *in* The Journal of Derivatives · October 2018

DOI: 10.3905/jod.2018.26.2.019

Tianxiang Liu
Industrial and Commercial Bank of China

Curve-Fitting Method for Implied Volatility

Desheng Wu and Tianxiang Liu

JOD 2018, 26 (2) 19-37
doi: https://doi.org/10.3905/jod.2018.26.2.019
http://jod.iijournals.com/content/26/2/19

Institutional Investor Journals
1120 Avenue of the Americas, 6th floor,
New York, NY 10036, Phone: +1 212-224-3589

Curve-Fitting Method for Implied Volatility

DESHENG WU AND TIANXIANG LIU

DESHENG WU
is a distinguished professor in the School of Economics and Management at the University of Chinese Academy of Sciences in Beijing, China, and a professor at Stockholm University in Stockholm, Sweden.
dash@risklab.ca

TIANXIANG LIU
is a manager at Asset Management Department, Industrial and Commercial Bank of China in Beijing, China, and a researcher in the School of Economics and Management at the University of Chinese Academy of Sciences in Beijing, China.
tianxiang.liu@icbc.com.cn

Curve-fitting methods construct a curve, or a surface, that has the best fit to a series of data points and disregards unnecessary fluctuations empirically. There are three types of curve-fitting methods: interpolation, in which the results are required to be an exact fit to the data; smoothing, in which a function is built with a good smoothness to approximate the data; and regression, which focuses more on statistical inferences.

Curve-fitting methods are widely implemented in the financial industry. In the literature, curve fitting for interest rate curves (term structure) is heavily discussed. Lin (2002) fitted the term structure of the Taiwanese government bond yields with B-splines. Ron (2000) provided a detailed curve-fitting method for the swap term structure of marking-to-market fixed-income products. Hagan and West (2006) summarized a large range of curve-fitting methods used in interest rate term structure fitting and discussed the connection between the traditional interest rate bootstrap methods and curve-fitting methods.

In this article, we examine the curve-fitting methods that are deployed in a rarely discussed area: fitting the implied volatility surface (IVS) in the options market. Curve-fitting methods are implemented to provide an intuitive expression of market conditions to options traders. Furthermore, curve-fitting methods can interpolate or extrapolate implied volatilities of new options using the existing options data.

IVS is a concept derived from most widely accepted option pricing model, the Black–Scholes–Merton (BSM) model (see also Black and Scholes 1973 and Merton 1973). IVS is the implied volatility plotted against the moneyness and time to maturity. Moneyness is a description of a derivative relating its strike price to the price of its underlying asset; it describes the intrinsic value of an option in its current state. Compared with later option-pricing models, the BSM model is easier to understand, calculate, and explain to others. Moreover, financial information providers, such as Bloomberg and Wind, only provide the BSM model in their software. Therefore, most market practitioners still prefer the BSM model over the more advanced models. However, the BSM model has its own disadvantages. The assumption of the BSM model is that volatility, as the only input, is a constant that does not vary with the strike price and maturity, but the implied volatility calculated by the actual price of options is a function of exercise price and duration. The function with respect to strike is called an implied volatility skew, and the function with respect to strike and time to maturity is usually called an IVS. Curve-fitting methods are tightly integrated into the management of options positions and provide a data visualization for practitioners.

Curve-fitting methods for IVS can be classified into two groups: theoretical and statistical. The theoretical methods attempt to explain typical features of the underlying asset, which may include how the underlying asset evolves through time, whereas the statistical methods are independent of any model or theory of the underlying asset. They only attempt to find a close representation of the implied volatility at any point in time, given some observed options data. Statistical methods include linear regression, quadratic regression, nearest neighbor, bilinear interpolation, bicubic interpolation, locally weighted scatterplot smoothing (LOWESS), LOESS, the thin plate interpolation, biharmonic interpolation, the Nadaraya–Watson kernel regression, and the artificial neural network method. These statistical methods can be divided into three categories: regression, interpolation, and machine learning (ML) algorithm.

First, we examine the traditional parametric regression models, including linear regression and quadratic regression, as well as the nonparametric regression models, including the LOWESS, the LOESS, and the Nadaraya–Watson kernel regression. Linear regression is the simplest method used as a benchmark curve-fitting method. Quadratic regression is a reasonable candidate; Gatheral (2011) attempted to empirically model the implied volatility along the strike axis and the volatility skew by quadratic specifications. Nonparametric regression models fit simple models, such as moving average and parametric regression, to localized subsets of the data. In fact, nonparametric regressions do not specify a global function of any form; they only fit segments of the data. Fengler, Härdle, and Villa (2003), Fengler and Wang (2003), and Fengler (2006) employed the Nadaraya–Watson estimator for fitting the IVS. Higher-order local polynomial smoothing of the IVS was implemented by Rookley (1997).

We then consider interpolation methods, which is a process for estimating values between known data points. Bilinear interpolation and bicubic interpolation are used when values at random position on a regular two-dimensional grid need to be determined. The thin plate interpolation and biharmonic interpolation are spline-based techniques for data interpolation and smoothing. Wallmeier and Hafner (2001) fitted quadratic splines to one-dimensional implied volatility skews.

There is otherwise a lack of literature adopting interpolation methods to fit IVSs.

We also evaluate ML algorithms as fitting methods. These algorithms can solve nonlinear regression, classification, and pattern recognition. Famous examples are artificial neural networks, support vector machines, and Bayesian networks, among others. In this study, we pick the artificial neural network approach as a representative of ML algorithms to fit the IVSs.

As opposed to statistical methods, theoretical methods directly model IVSs by describing the dynamics of underlying asset process (see also Zhu and Avellaneda 1998 and Schönbucher 1999). Unlike traditional curve-fitting methods, implied volatility cannot be specified freely in theoretical methods because the specification of implied volatility is incorporated into the BSM model. The no-arbitrage condition should be satisfied by the specification of implied volatility and dictates the shape of surfaces generated by the model (see also Daglish, Hull, and Suo 2007 and Carr and Wu 2016).

We choose three criteria—the goodness of fit, the smoothness, and the economic meaning—to evaluate a curve-fitting method because these criteria concern option practitioners most. These three criteria are

(1) How well does it fit a set of observations? In other words, is the degree of error in the created curve sufficiently small? A good fitting method should adequately capture the underlying structure of the data and sufficiently fit the data. We use mean squared error (MSE), R2, and the Akaike information criterion (AIC) as measures of the goodness of fit. Please note that some interpolation methods always pass through the data points, and thus the goodness of fit of the total sample is meaningless in such case. We use cross validation here to determine how good the fit is.

(2) How smooth is the generated curve? The smoother the curve, the less noisy it appears to be. We will want to have continuity and the smoothness of the implied volatilities if we want to calculate the implied distribution from the fitted curve. When estimating implied distribution from IVS by iron butterfly spreads (see also Breeden and Litzenberger 1978), a smooth volatility surface is required to generate a stable distribution result.

(3) Does the fitting method reveal the economic meaning behind the curve? Some methods provide a mechanism to extract a few economically meaningful states from implied volatilities. For example, the Carr–Wu model extracts instantaneous volatility, volatility of volatility, correlation, time decay, and so on from the current implied volatilities, and these economic states give practitioners the economic implications of the current market.

We will discuss these criteria in the methodology section and empirical result section.

In the empirical analysis of the 2003–2016 S&P 500 samples, three interpolation methods (thin plate interpolation, biharmonic interpolation, and cubic interpolation) provide the best goodness of fit and relatively good smoothness. Quadratic regression, the Nadaraya–Watson kernel regression, and the Carr–Wu model generate the smoothest surface, but only the Carr–Wu model can explain the economic states behind the surfaces.

The situation in emerging markets may be quite different from that in developed markets for the following reasons: First, the options exchanges in emerging markets have very narrow strike price bands (the difference between the upper and lower bound of strike prices) and an insufficient intensity in strike price intervals; and second, the observed data points in emerging markets are not on a regular grid, and the range of data points is quite limited, which can cause problems. We also discuss the performance of curve-fitting methods in emerging options markets.

Because of the limited moneyness range of observed data in emerging markets, extrapolation along the strike axis is required. Lee (2004) proved that the extreme strike tails of implied volatility are bounded by $O(\sqrt{|\ln k|})$ where k is the moneyness. The Lee's condition should be satisfied when extrapolating, but no methods except the Carr–Wu model meet the condition. We propose a transformation method to improve quadratic regression, the thin plate interpolation, and biharmonic interpolation to satisfy the Lee's condition.

In regard to China's 50ETF options market, quadratic regression has the best goodness of fit satisfying the Lee's condition. In addition, the Carr–Wu model is a very good alternative considering that it natively satisfies the Lee's condition and has good economic implications.

The rest of this article is organized as follows: First, we review the existing literature on curve-fitting methods in the financial area. We then introduce 12 curve-fitting methods of IVS in the methodology section. We evaluate them from three aspects—the goodness of fit, the smoothness, and the economic meaning—by feeding them the US S&P 500 options data. To implement the curve-fitting methods in emerging markets, we propose a transformation method to improve curve-fitting methods to meet the Lee's condition. We compare the improved methods using the China 50ETF options market data.

LITERATURE REVIEW

In the existing literature, the curve fitting for the interest rate curve (term structure) is widely discussed. As explained by Hagan and West (2006) term structure estimation methods can be classified into two groups: statistical and theoretical. Statistical methods are independent of any model or theory of term structure, whereas theoretical methods attempt to explain typical features of the term structure, which may include how the term structure evolves through time. Statistical methods include simple interpolation methods, cubic splines, quartic splines, forward monotone convex spline, and minimalist quadratic interpolator.

The theoretical curve-fitting methods for interest rate curves typically posit an explicit structure for interest rates known as the short rate of interest. Examples of theoretical methods in interest rates are as follows (see also Lin 2002). In all the following r_t is the short rate and W_t is a Wiener process.

The Merton (1973) model assumes the interest rate to be

$$r_t = r_0 + at + \sigma W_t \tag{1}$$

where a is the growth rate of the interest rate.

The Vasicek (1977) model describes the interest rate as

$$dr_t = a(b - r_t)dt + \sigma dW_t \tag{2}$$

where b is the long-term equilibrium interest rate and a is the mean-reverting rate.

The Rendleman and Bartter (1980) model explains the interest rate as

$$dr_t = \theta r_t dt + \sigma r_t dW_t \tag{3}$$

where θ is the growth rate of the interest rate.

The Cox, Jonathan, and Ross (1985) model assumes the interest rate to be

$$dr_t = a(b - r_t)dt + \sqrt{r_t}\sigma dW_t \tag{4}$$

where b is the long-term equilibrium interest rate and a is the mean-reverting rate.

The Ho and Lee (1986) model writes the interest rate as

$$dr_t = \theta_t dt + \sigma dW_t \tag{5}$$

where θ_t is the stochastic drift rate.

The Hull and White (1990) model supposes the interest rate to be

$$dr_t = (\theta_t - \alpha_t r_t)dt + \sigma_t dW_t \tag{6}$$

where θ_t is the stochastic drift rate, α_t is the stochastic mean-reverting rate, and σ_t is the stochastic volatility.

In the options trading area, some researchers have attempted to fit the IVS, but they did not achieve satisfactory results. Fengler, Härdle, and Villa (2003); Fengler and Wang (2003); and Fengler (2006) employed the Nadaraya–Watson estimator for fitting IVSs, and Rookley (1997) adopted higher-order local polynomial smoothing of the IVSs. Wallmeier and Hafner (2001) fitted quadratic splines to one-dimensional implied volatility skews. There is a lack of literature regarding interpolation methods to fit IVSs. Fengler (2009) proposed an arbitrage-free two-stage cubic spline method to avoid negative transition probabilities.

For theoretical methods, Zhu and Avellaneda (1998) and Schönbucher (1999) first attempted to describe the volatility surfaces theoretically. Daglish, Hull, and Suo (2007) and Carr and Wu (2016) managed to build their models on implied volatility dynamics and the no-arbitrage condition. Unlike stochastic volatility models, the implied volatility cannot be specified freely because the no-arbitrage condition must be satisfied by the specification of implied volatility.

METHODOLOGY

In this article, we discuss 12 curve-fitting methods, including linear regression, quadratic regression, nearest neighbor, bilinear interpolation, bicubic interpolation, LOWESS, LOESS, thin plate interpolation, biharmonic interpolation, Nadaraya–Watson kernel regression, artificial neural networks, and the Carr–Wu model. We attempt to include as many methods as possible. These 12 methods are either widely accepted in the engineering and financial engineering field (the first 11 methods) or were proposed recently in academia (the Carr–Wu model). For example, quadratic regression, Nadaraya–Watson kernel regression, thin plate interpolation, and biharmonic interpolation are implemented in fitting IVS in the aforementioned literature. Bilinear interpolation, bicubic interpolation, the LOWESS, and the LOESS are not mentioned in the literature but are widely implemented in financial software to plot IVS and other financial data on the screen. Linear regression and nearest neighbor are widely used as benchmark methods, so we mention these two methods. We also include artificial neural networks as a representative of ML algorithms.

These models fall into four categories: regression, interpolation, ML algorithm, and theoretical models, as described in Exhibit 1.

We first discuss the parametric regressions, including linear regression with three parameters

$$I_{k,\tau} = \beta_0 + \beta_1 k + \beta_2 \tau + \epsilon \tag{7}$$

and quadratic regression with six parameters

$$I_{k,\tau} = \beta_0 + \beta_{11}k^2 + \beta_{12}k\tau + \beta_{22}\tau^2 + \beta_1 k + \beta_2 \tau + \epsilon \tag{8}$$

where $I_{k,t}$ is fitted implied volatility with moneyness k and time to maturity τ. Panels A and B of Exhibit 2 illustrate these two methods.

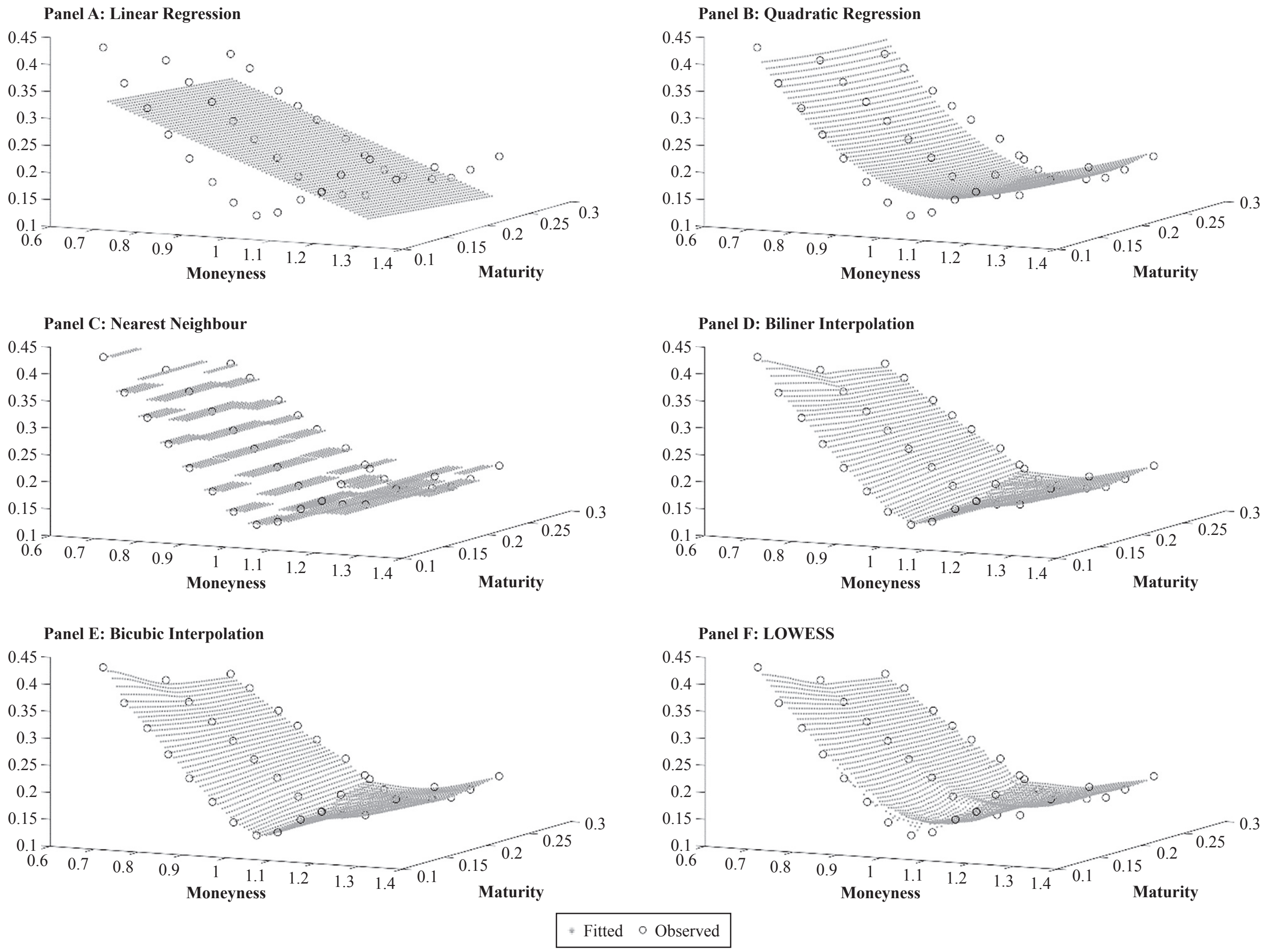
LOESS and LOWESS are two strongly related locally weighted linear regression methods used to smooth data. (see also Cleveland 1979). LOWESS is not a true acronym; it may be understood as standing for *locally weighted regression*. LOESS is a later generalization of LOWESS.

LOESS and LOWESS are local versions of least squares regression methods. The fitting process is

Exhibit 1
Curve-Fitting Method Summary

	Method	Category	Parametric	Parameter Number
1	Linear regression	Regression	Yes	3
2	Quadratic regression	Regression	Yes	6
3	Nearest neighbor	Interpolation	No	Sample size × 3
4	Bilinear interpolation	Interpolation	No	Sample size × 3
5	Bicubic interpolation	Interpolation	No	Sample size × 3
6	LOWESS	Regression (smoothing)	No	Sample size × 3
7	LOESS	Regression (smoothing)	No	Sample size × 3
8	Thin plate interpolation	Interpolation	No	Sample size × 2 + 6
9	Biharmonic interpolation	Interpolation	No	Sample size × 3
10	Nadaraya–Watson	Regression (smoothing)	No	Sample size × 3
11	Neural network	Machine learning algorithm	Yes	Hidden layer size × 5 + 9
12	Carr and Wu	Theoretical model	Yes	6

Exhibit 2
Illustration of Curve-Fitting Methods (based on US S&P 500 options data of January 1, 2013)



Panel A: Linear Regression

Panel B: Quadratic Regression

Panel C: Nearest Neighbour

Panel D: Biliner Interpolation

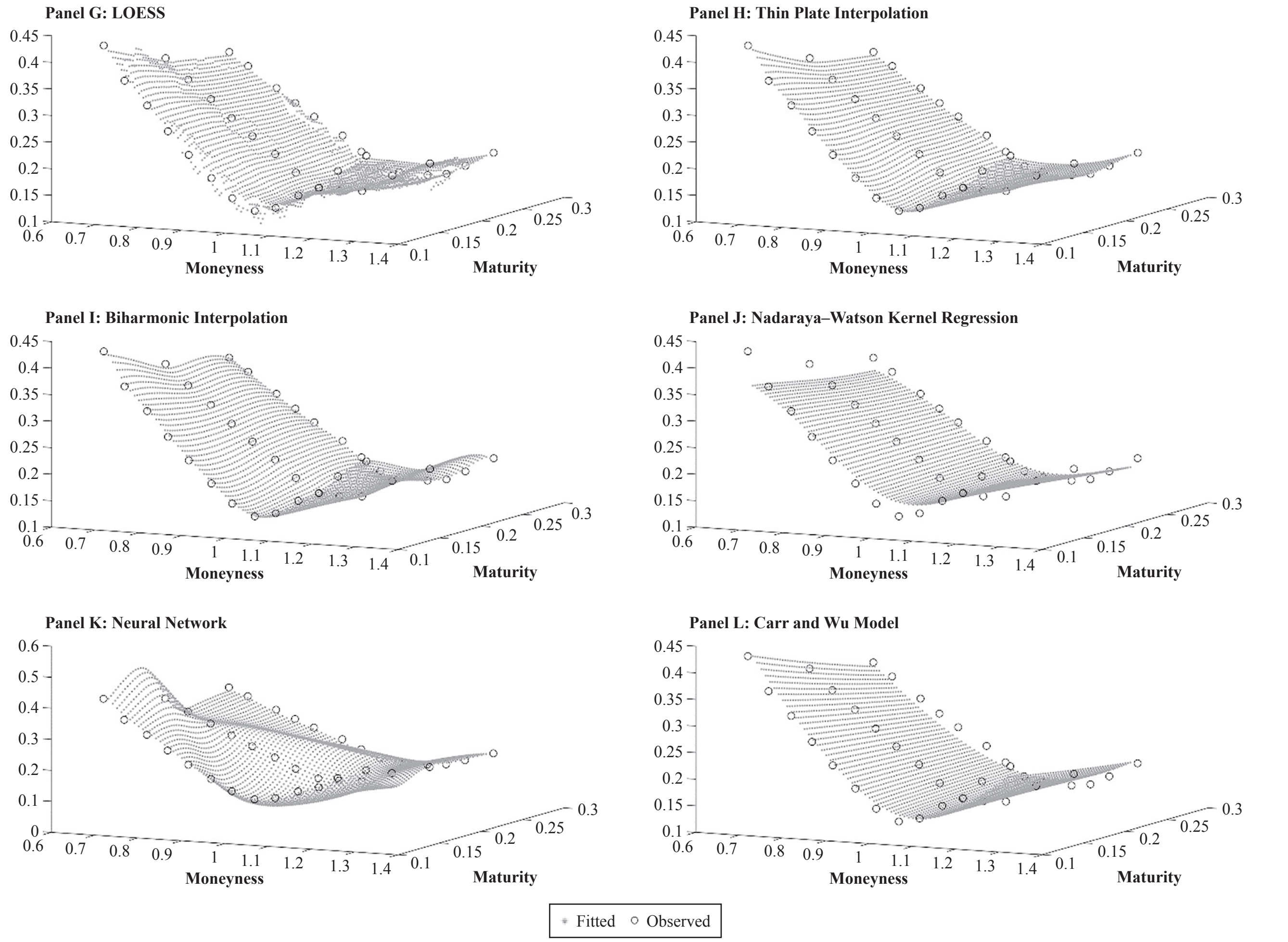
Panel E: Bicubic Interpolation

Panel F: LOWESS

(continued)

EXHIBIT 2 *(continued)*

Illustration of Curve-Fitting Methods (based on US S&P 500 options data of January 1, 2013)


Panel G: LOESS

Panel H: Thin Plate Interpolation

Panel I: Biharmonic Interpolation

Panel J: Nadaraya–Watson Kernel Regression

Panel K: Neural Network

Panel L: Carr and Wu Model

Notes: The 12 curve-fitting methods are shown. The circles are the original data points, and the gray points are fitted points.

considered local because each fitted value is determined by neighboring data points defined within a local area. In this article, we define the local area as the nearest 10% of the data points. The fitting process is weighted by a tricubic weight function, which is defined for the data points contained within the local area:

$$w_i = \left(1 - \left(\frac{\sqrt{(k-k_i)^2 + (\tau-\tau_i)^2}}{\max_i \sqrt{(k-k_i)^2 + (\tau-\tau_i)^2}}\right)^3\right)^3 \qquad (9)$$

where k and τ are the strike and moneyness associated with the response value to be fitted, respectively. k_i and τ_i are the strike and moneyness associated with i-th nearest neighbors of the fitted point with the local area, respectively.

These two methods are differentiated by the model used in the regression: LOWESS uses a linear polynomial model, whereas LOESS uses a quadratic polynomial model. The fitted values are given by the weighted least squares regression at a certain strike and maturity. Panels F and G of Exhibit 2 illustrate these two methods.

The Nadaraya–Watson local regression is another local regression method (see also Nadaraya 1964; Watson 1964; Härdle 1990). Instead of weighted least squares methods, which are used in LOWESS and LOESS, moving average is used in the Nadaraya–Watson local regression. The weight function in Nadaraya–Watson local regression is defined for the data points contained within the local area:

$$w_{i,Nadaraya-Watson} = \frac{K(k-k_i,\tau-\tau_i)}{\sum_j K(k-k_j,\tau-\tau_j)} \tag{10}$$

The $K(k, t)$ is a kernel function. In this article, we use a Gaussian kernel:

$$K(k,\tau) = e^{-\frac{1}{2}(k^2+\tau^2)} \tag{11}$$

Therefore, the Nadaraya–Watson estimator is the weighted average of the nearest neighbors:

$$\widehat{IV}_{k,\tau,\text{Nadaraya-Watson}} = \sum_i \mathrm{w}_{i,\text{Nadaraya-Watson}} \times IV_i \tag{12}$$

The local area in Nadaraya–Watson local regression was selected optimally by Bowman and Azzalini (1997). An illustration of Nadaraya–Watson local regression is displayed in Exhibit 2, Panel J.

In contrast to regressions, interpolation is a technique for adding new data points within a range of a set of known data points. Compared with the smoothing and regression methods, interpolation methods focus more on the goodness of fit of the original data. In most interpolation methods, the fitted curve must pass through the original data, so the R^2 value must be one. To compare different curve-fitting methods, we adopt cross validation to calculate the goodness of fit measures.

The simplest interpolation method, the nearest neighbor method, is to locate the nearest data value and assign the same value, which results in a discontinuous surface.

Bilinear interpolation is an extension of one-dimensional linear interpolation for interpolating functions of two variables in a two-dimensional regular grid. The interpolated value is based on bilinear interpolation of the values at the two nearest points in each respective dimension. Bilinear interpolation generates a continuous function without a continuous derivative. Bicubic interpolation is an extension of cubic interpolation for interpolating data points on a two-dimensional regular grid. The interpolated values are based on bicubic interpolation of the values at the four nearest points in each respective dimension. The function generated by bicubic interpolation is twice continuously differentiable. An illustration of these three methods is shown in Exhibit 2, Panels C–E.

On the other hand, spline-based interpolation generates a smoother fitted surface but requires more memory and computation. A polyharmonic spline (see also Harder and Desmarais 1972; Fasshauer 2007) is a linear combination of polyharmonic radial basis functions ϕ plus a polynomial term:

$$f(x) = \sum_{i=1}^{N} a_i \phi\left(\sqrt{(k-k_i)^2+(\tau-\tau_i)^2}\right) + b_0 + b_1 k + b_2 \tau \tag{13}$$

The weights a_i and b_i are determined such that the function must pass the N observed data points and fulfill the following conditions:

$$\sum_{i=1}^{N} a_i = 0,\ \sum_{i=1}^{N} a_i k_i = 0,\ \sum_{i=1}^{N} a_i \tau_i = 0 \tag{14}$$

In this article, we discuss biharmonic interpolation and thin plate interpolation. The biharmonic radial basis function is

$$\phi(r) = r^2 \ln(r) \tag{15}$$

Thin plate interpolation (see also Duchon 1977; Bookstein 1989) is a modified biharmonic interpolation with thin plate spline

$$f(k,\tau) = \sum_{i=1}^{N} a_i \phi\left(\sqrt{(k-k_i)^2+(\tau-\tau_i)^2}\right) \tag{16}$$

The weights a_i are calculated by minimizing the residual sum of squares

$$\sum_{i=1}^{N} (IV_i - f(k_i,\tau_i))^2 \tag{17}$$

where IV_i is the implied volatility of the i-th data point. These two spline-based methods produce smooth surfaces, which are infinitely differentiable. They have closed-form solutions for parameter estimation, but the calculation is memory- and time-consuming. An illustration of these two methods is shown in Exhibit 2, Panels H and I.

We also consider the application of ML in curve fitting. ML is an interdisciplinary subject involving probability theory, statistics, optimization, convex analysis, algorithm complexity theory, and so on. ML research addresses how computers simulate or implement human learning behavior to acquire new knowledge or skills and reorganize the existing knowledge structure to improve their performance. In this article, we consider one particular method, the artificial neural network, which is an ML system inspired by biological brains (see also Werbos 1974; Schmidhuber 2015). The 2-10-1 neural network is used to fit the IVS, and 10% of the data is reserved for validation. One of the problems that occurs during neural network training with a small size sample is overfitting. The neural network generates a surface with too much nonlinearity, but IVS usually does not require such high nonlinearity. An illustration of the neural network is shown in Exhibit 2, Panel K.

In theoretical models, we assume that the underlying asset follows the stochastic process

$$\frac{dS_t}{S_t} = \sqrt{v_t}\,dW_t \qquad (18)$$

where ν_t is the instantaneous variance and W_t is a Wiener process. In stochastic models, the instantaneous volatility is specified (see also Hull and White 1987; Scott 1987; Heston 1993). To model implied volatility, Carr and Wu (2016) specified the implied volatility rather than instantaneous volatility. This specification corresponds more strongly than stochastic volatility models with how practitioners manage their options because we model what practitioners quote directly. From Carr and Wu (2016)

$$dI_t(k,\tau) = Ie^{-\eta_t\tau}(m_t)dt + Iw_t e^{-\eta_t\tau}dZ_t \qquad (19)$$

where $\mathrm{I}_t(k,\tau)$ is the implied volatility at time t with respect to moneyness k and maturity τ; Z_t is a Wiener process; and variables m_t, j_t, w_t and η_t are stochastic processes that do not depend on k, τ, and $\mathrm{I}_t(k,\tau)$. μ_t is the drift of implied volatility, w_t is the volatility of volatility (volvol) process, $\rho_{i,t}$ is the stochastic process taking values in an interval from −1 to 1, and m_t describes the average drift of volatility.

Using stochastic calculus and a no-arbitrage condition, the implied volatility can be expressed as

$$\frac{\tau^2 w^2 e^{-2\eta\tau}}{4} I^4 + (1 - \rho\tau\sqrt{\nu}we^{-\eta\tau} - 2m\tau e^{-\eta\tau})I^2 + (-\nu - (lnk)^2 w^2 e^{-2\eta\tau} - 2lnk\rho\sqrt{\nu}we^{-\eta\tau}) = 0 \qquad (20)$$

as done by Carr and Wu (2016).

The Carr–Wu model has successfully established the link between stochastic processes and IVSs. The shape of IVS is only affected by the current state of parameters ($\nu_t, m_t, w_t, \rho_t, \eta_t$) but does not depend on the stochastic dynamics of the parameters. We use ordinary least squares to estimate parameters. An illustration of the Carr–Wu model is shown in Exhibit 2, Panel L.

We do not adopt the state-space method mentioned by Carr and Wu (2016). They treated parameters as hidden states and implied volatility as measurements with the error. The result of the state-space model is quite unsatisfactory: It fails to explain the variation in IVS because update speed cannot catch up with parameters. The state-space model is not included in this article because of its unsatisfactory performance.

The number of parameters in parametric methods can be easily obtained. The neural network can be treated as a parametric model because it is a combination of matrix calculation and sigmoid functions. The number of parameters in the neural network depends on the size of the hidden layer. The nonparametric methods rely on the entire data sample; thus, the number of parameters for most nonparametric methods equals the sample size times the dimension of the sample (three in our case). In thin plate interpolation, the number of parameters can be compressed to the sample size times 2 plus 6. The number of parameters is highly related to the overfitting problem, and we will discuss it in the next section.

EMPIRICAL RESULTS

We use the Chicago Board Options Exchange's end-of-day listed S&P 500 options market data from 2003 to 2016 to test our method. The end-of-day mid-quotes are used to construct the implied volatilities. Data points without valid bid price in a day are removed from the sample.

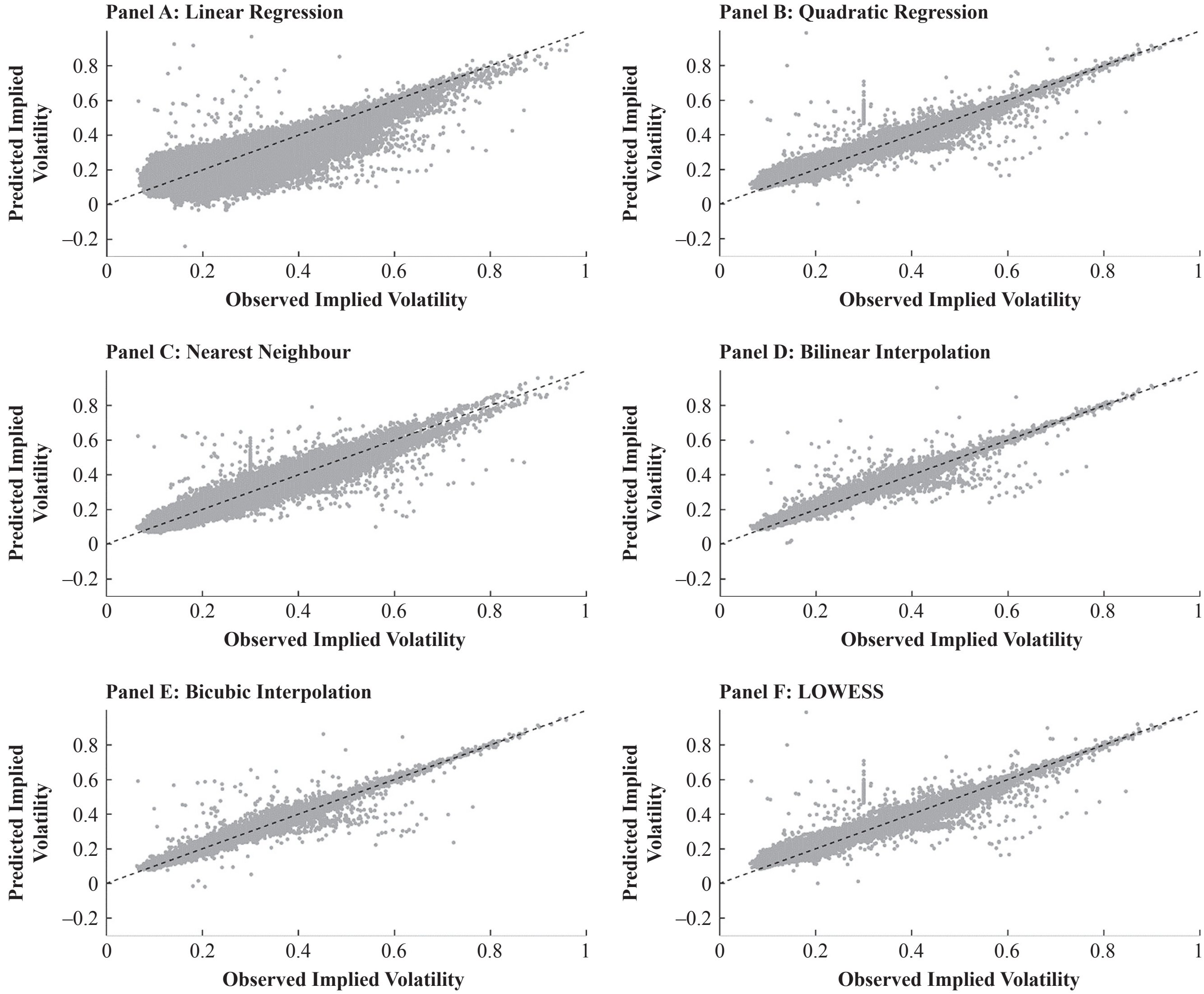
We construct implied volatility on a grid of 13 fixed relative strikes ranging from 70 to 130 and three fixed times to maturity, from one month to three months. During the construction, we use only out-of-the-money volatilities because out-of-the-money options are more actively traded and more sensitive to implied volatility. We use the average of put and call implied volatilities to present the at-the-money implied volatility.

Because many curve-fitting methods, especially interpolation methods, require the fitted surfaces to pass through the observed data points, the goodness of fit of the whole sample is meaningless. For comparison, we use cross validation. There are 13 times 3 data points per day. One data point is selected for testing, and other data points are used for training. We repeat this process until all data points are tested. We obtain MSE, R^2, and AIC values from the cross-validation testing results.

As suggested by Exhibit 3, the nearest neighbor method and quadratic regression also do not show a good fit because of their overly simple structures.

Exhibit 3

Fitting Plots of Predicted Implied Volatility against Observed Implied Volatility in All Data of the US S&P 500 Options from 2003 to 2016



Panel A: Linear Regression

Panel B: Quadratic Regression

Panel C: Nearest Neighbour

Panel D: Bilinear Interpolation

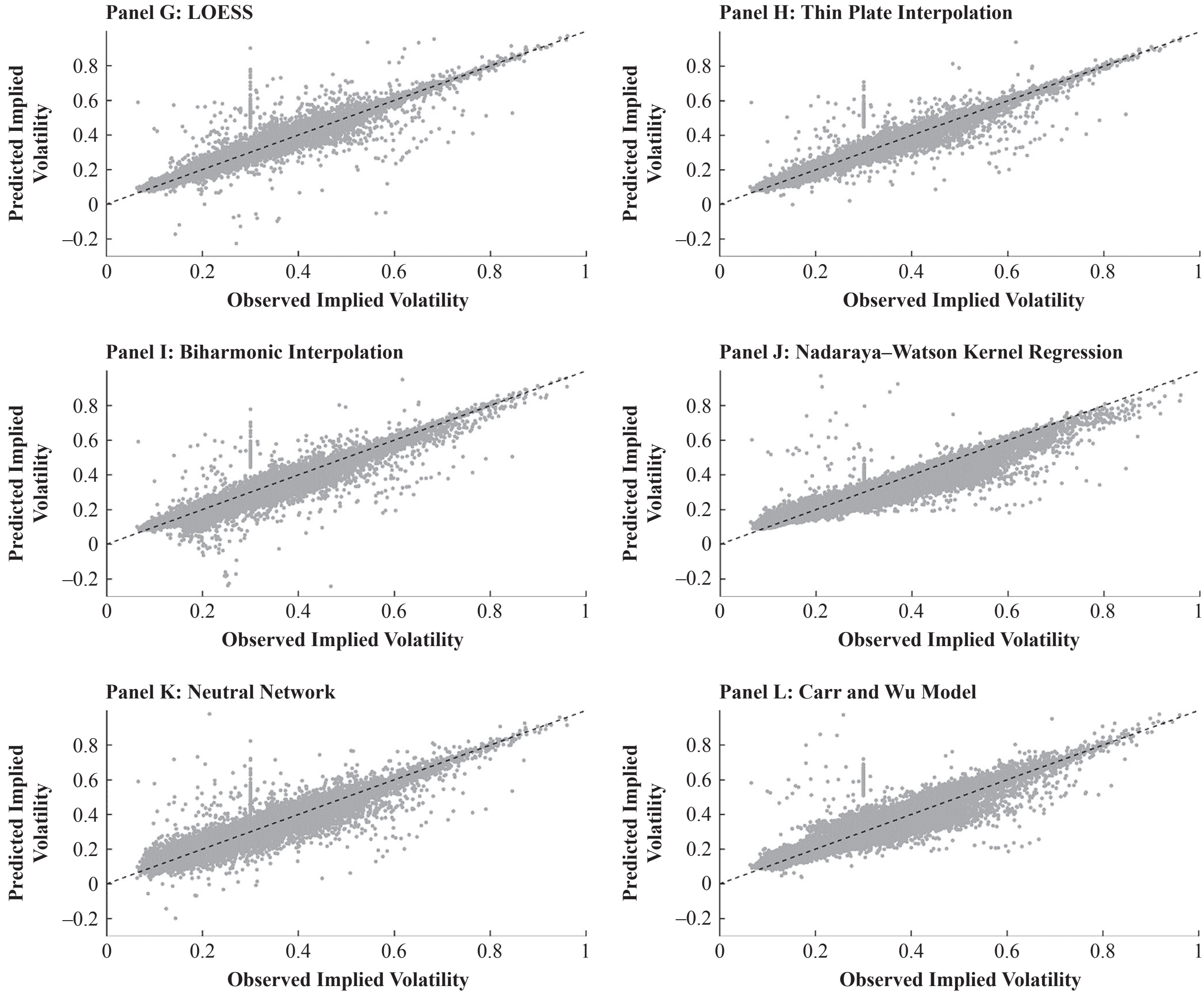
Panel E: Bicubic Interpolation

Panel F: LOWESS

(continued)

Exhibit 3 *(continued)*

Fitting Plots of Predicted Implied Volatility against Observed Implied Volatility in All Data of the US S&P 500 Options from 2003 to 2016



Panel G: LOESS

Panel H: Thin Plate Interpolation

Panel I: Biharmonic Interpolation

Panel J: Nadaraya–Watson Kernel Regression

Panel K: Neutral Network

Panel L: Carr and Wu Model

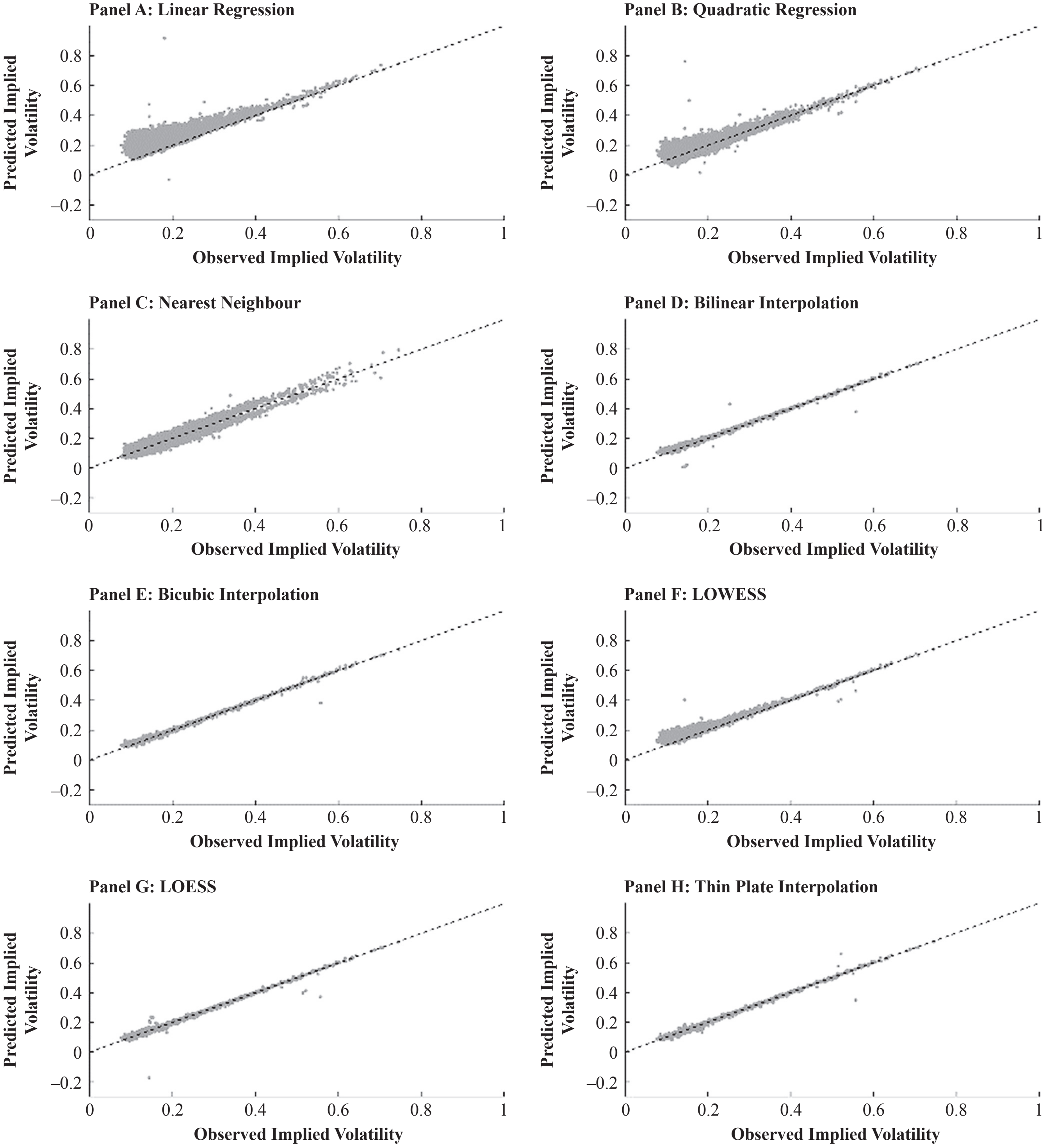
Notes: The fitting plots of 12 methods are shown in this exhibit. The diagonal lines represent the perfect fitting of the data. The more dispersed the data points, the worse the fitting. Linear regression unexpectedly appears to be the worst fit.

The interpolation methods, especially bilinear interpolation, bicubic interpolation, and thin plate interpolation, achieve the best fits of data. The plots in two nonparametric regressions, LOWESS and LOESS, also are relatively concentrated on the diagonal lines. Biharmonic interpolation, Nadaraya–Watson kernel, neural network, and the Carr–Wu model yield medium performance. Considering very few parameters are used in the Carr–Wu model, its performance is acceptable.

Before 2003 the Chicago Board Options Exchange used the at-the-money volatility as the volatility index. Because of the importance of at-the-money volatilities, we check their goodness of fit in Exhibit 4. Generally, all curve-fitting methods provide better fits to at-the-money volatilities than whole volatilities. Empirically,

Exhibit 4

Fitting Plots of Predicted Implied Volatility against Observed Implied Volatility in At-the-Money Data of the US S&P 500 Options from 2003 to 2016



Panel A: Linear Regression

Panel B: Quadratic Regression

Panel C: Nearest Neighbour

Panel D: Bilinear Interpolation

Panel E: Bicubic Interpolation

Panel F: LOWESS

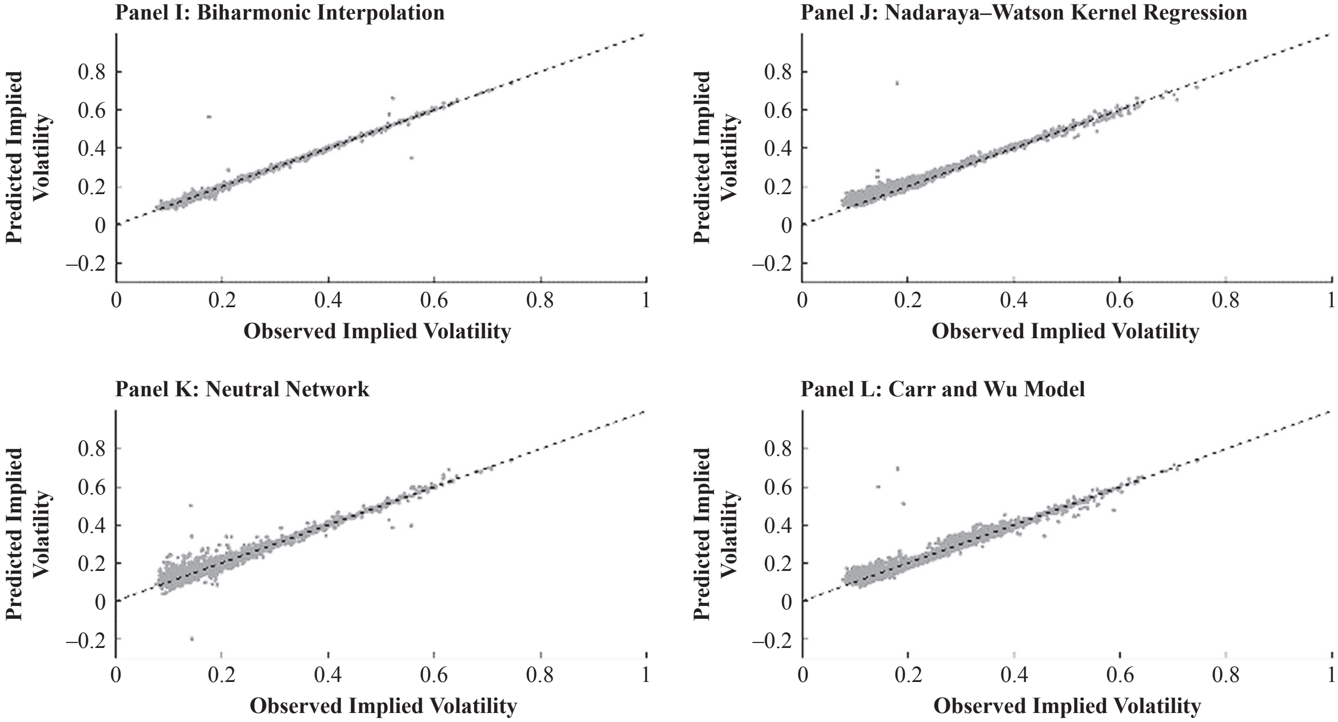
Panel G: LOESS

Panel H: Thin Plate Interpolation

(continued)

EXHIBIT 4 *(continued)*

Fitting Plots of Predicted Implied Volatility against Observed Implied Volatility in At-the-Money Data of the US S&P 500 Options from 2003 to 2016



Notes: The goodness of fit of at-the-money volatilities is plotted. The at-the-money options, which are options with 100 moneyness, are widely traded, and their volatilities are often considered as representative of volatility.

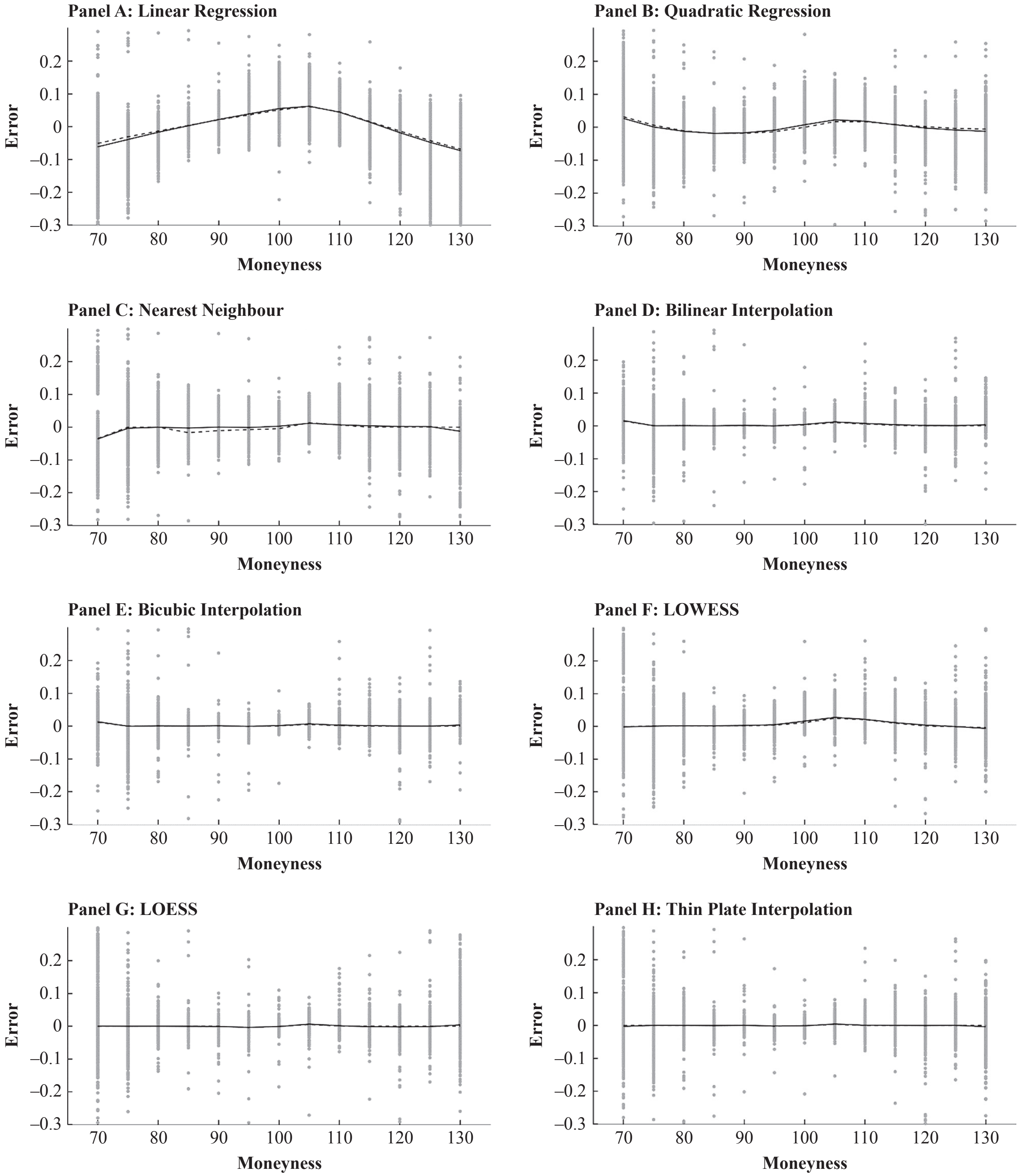
out-of-the-money volatilities tend to be higher than at-the-money volatilities. As a result, at a certain maturity, the volatility curve is usually a smile-like curve, and the volatility surface is usually a U-shape that rolls to 100 moneyness. For most methods, the fitted surface does not have as much curvature as the observed data, so the fitted at-the-money volatilities are always higher than the true values. The extreme case is seen in linear regression. Because there is no curvature in linear regression at all, the at-the-money volatilities are highly overestimated. The interpolation methods and nonparametric regressions show the highest goodness of fit of at-the-money volatilities. These local methods use the near-the-money data to fit the at-the-money data and, of course, have better performance than global methods in fitting at-the-money data.

As is suggested in Exhibit 5, all 12 methods perform better in at-the-money volatilities than in out-of-the-money volatilities. Some methods, including linear regression and the Nadaraya–Watson local regression, show a reverse U shape of error, which means that at-the-money volatilities are overestimated and out-of-the-money volatilities are underestimated. Quadratic regression and the Carr–Wu model show an S shape of error. These methods overestimate volatilities with very low and medium-high moneyness and underestimate volatilities with very high and medium-low moneyness. Please note that quadratic regression has a quadratic structure, and the Carr–Wu model has a biquadratic structure, and thus they share the same error pattern. Other methods generally show a stable error pattern, which means that the fit performance does not vary much across different moneyness.

Exhibit 6 shows the goodness of fit of all volatilities and at-the-money volatilities. With respect to mean square error in all data, three interpolation methods (thin plate interpolation, biharmonic interpolation, and cubic interpolation) get the best results, 0.000191, 0.000200,

EXHIBIT 5

Error Plots across Different Moneyness in All Data of US S&P 500 Options from 2003 to 2016



Panel A: Linear Regression

Panel B: Quadratic Regression

Panel C: Nearest Neighbour

Panel D: Bilinear Interpolation

Panel E: Bicubic Interpolation

Panel F: LOWESS

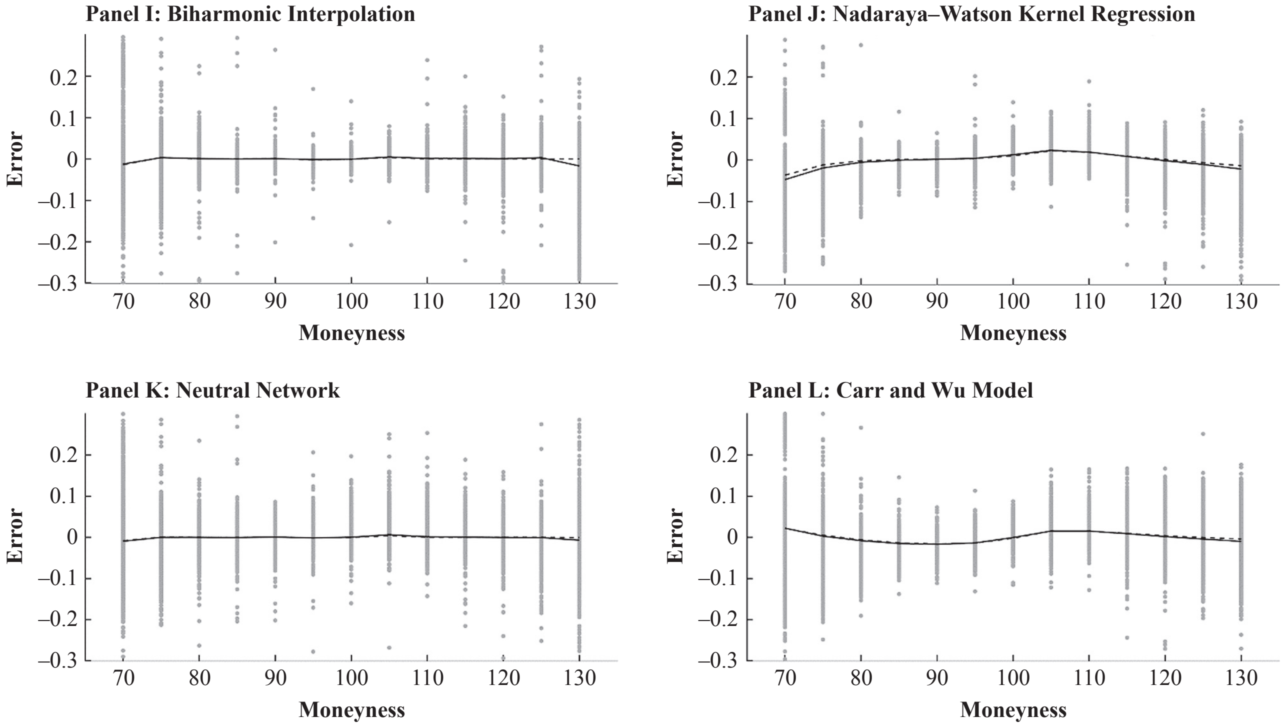
Panel G: LOESS

Panel H: Thin Plate Interpolation

(continued)

EXHIBIT 5 *(continued)*

Error Plots across Different Moneyness in All Data of US S&P 500 Options from 2003 to 2016



Panel I: Biharmonic Interpolation

Panel J: Nadaraya–Watson Kernel Regression

Panel K: Neutral Network

Panel L: Carr and Wu Model

Notes: Errors are plotted against moneyness under different fitting methods. The solid lines show the mean of errors, and the dashed lines show the median.

and 0.000214, respectively. The neural network and local regression methods (LOWESS and LOESS) get the second best results, 0.000460, 0.000481, and 0.000483. The R^2 result is consistent with the MSE. When we take the overfitting problem into consideration, the AIC includes a penalty that is an increasing function of the number of estimated parameters. The same three interpolation methods remain in the first tier with respect to AIC. The neural network and local regression methods (LOWESS and LOESS) remain in the second tier. For quadratic structure–based methods (quadratic regression and the Carr–Wu model), the goodness of fit is relatively bad, even considering the penalty of overfitting.

In at-the-money volatilities for all methods except linear regression, quadratic regression, nearest neighbor, and LOWESS have better R^2 than all data. Three interpolation methods (thin plate interpolation, biharmonic interpolation, and cubic interpolation) and LOESS have the best R^2. With respect to AIC, the four methods keep leading.

We next consider the smoothness of these methods. The smoothness brings not only the good-looking appearance of the fitted surface but also robustness when interpolating new volatility using the fitted surface. Theoretically, all the parametric regressions and theoretical models have an analytical expression of the surface; therefore, we can prove that they are infinitely continuously differentiable. The neural network can be regarded as a linear combination of sigmoid functions, so it is also infinitely continuously differentiable. The nearest neighbor method is not continuous, and bilinear interpolation is continuous but not differentiable. Bicubic interpolation and spline-based interpolations are twice continuously differentiable. There is lack of discussion of smoothness of nonparametric regressions in the existing literature.

We propose a numerical measure of smoothness defined as the finite difference of the following expression

$$\iint\left[\left(\frac{\partial^2 IV}{\partial k^2}\right)^2+2\left(\frac{\partial^2 IV}{\partial k\,\partial\tau}\right)^2+\left(\frac{\partial^2 IV}{\partial\tau^2}\right)^2\right]dk\,d\tau \qquad (21)$$

EXHIBIT 6

The Goodness of Fit in All Data and At-the-Money Data of US S&P 500 Options from 2003 to 2016

		All Data			At-the-Money		
		MSE	R^2	AIC	MSE	R^2	AIC
1	Linear regression	0.002880	0.726	764,261.5	0.004259	0.299	41,948.58
2	Quadratic regression	0.000997	0.905	627,301.3	0.001028	0.831	26,431.5
3	Nearest neighbor	0.001206	0.885	652,023.6	0.001120	0.816	27,577.73
4	Bilinear interpolation	0.000486	0.954	534,679.8	0.000065	0.989	–3,546.76
5	Bicubic interpolation	0.000214	0.980	428,757.3	0.000027	0.996	–13,221.7
6	LOWESS	0.000481	0.954	533,272.8	0.000490	0.919	18,550.25
7	LOESS	0.000483	0.954	534,009.0	0.000041	0.993	–8,568.94
8	Thin plate interpolation	0.000191	0.982	413,712.2	0.000022	0.996	–15,557.6
9	Biharmonic interpolation	0.000200	0.981	420,074.4	0.000036	0.994	–9,922.25
10	Nadaraya–Watson	0.000777	0.926	595,364.8	0.000351	0.942	14,915.78
11	Neural network	0.000460	0.956	527,409.3	0.000170	0.972	6,909.902
12	Carr and Wu	0.000898	0.914	613,705.8	0.000253	0.958	11,116.03

EXHIBIT 7

Smoothness of Curve-Fitting Methods in All Data of US S&P 500 Options from 2003 to 2016

		Continuous	Continuously Differentiable	Twice Continuously Differentiable	Smoothness
1	Linear regression	Yes	Yes	Yes	0
2	Quadratic regression	Yes	Yes	Yes	0.864615
3	Nearest neighbor	No	No	No	34,474.53
4	Bilinear interpolation	Yes	No	No	23.42391
5	Bicubic interpolation	Yes	Yes	Yes	8.850382
6	LOWESS	Not necessarily	Not necessarily	Not necessarily	5,295.323
7	LOESS	Not necessarily	Not necessarily	Not necessarily	27,383.39
8	Thin plate interpolation	Yes	Yes	Yes	12.69575
9	Biharmonic interpolation	Yes	Yes	Yes	29.40249
10	Nadaraya–Watson	Not necessarily	Not necessarily	Not necessarily	0.867198
11	Neural network	Yes	Yes	Yes	417.2152
12	Carr and Wu	Yes	Yes	Yes	1.181538

This expression is an estimator of smoothness in many spline-based smoothing methods. The second derivatives can be estimated by the finite difference method:

$$\frac{\partial^2 IV}{\partial k^2} = \frac{IV(k+\Delta k) + IV(k-\Delta k) - 2IV(k)}{(\Delta k)^2} \qquad (22)$$

$$\frac{\partial^2 IV}{\partial \tau^2} = \frac{IV(\tau+\Delta\tau) + IV(\tau-\Delta\tau) - 2IV(\tau)}{(\Delta\tau)^2} \qquad (23)$$

$$\begin{aligned}\frac{\partial^2 IV}{\partial k\,\partial\tau} = \frac{1}{4\Delta k\Delta\tau}[&IV(k+\Delta k,\tau+\Delta\tau) - IV(k-\Delta k,\tau+\Delta\tau)\\ &- IV(k+\Delta k,\tau-\Delta\tau) + IV(k-\Delta k,\tau-\Delta\tau)]\end{aligned} \qquad (24)$$

As suggested in Exhibit 7, linear regression has perfect smoothness because it generates a plane. Quadratic regression, Nadaraya–Watson kernel regression, and the Carr–Wu model generate the first-tier results. Four interpolation methods, including bilinear interpolation, bicubic interpolation, the thin plate interpolation, and biharmonic interpolation, generate the second-tier results. Of course, the discontinuous nearest neighbor method generates the worst result.

The Carr–Wu model provides not only the fit of IVS but also the economic explanation of the surface. The Carr–Wu model provides a mechanism to reduce the dimension of the surface to a few economically meaningful parameters. Given a volatility surface, one can easily understand the economic states behind the surface, such as instantaneous volatility, volatility of volatility, the drift of volatility, and correlation. Furthermore, these economic states proved to be pricing factors of future stock returns (see also Carr and Wu 2016). From this perspective, although the theoretical method has low goodness of fit, it has unique advantages in that it can reveal the economic meaning behind the surfaces.

CURVE-FITTING IN EMERGING MARKET

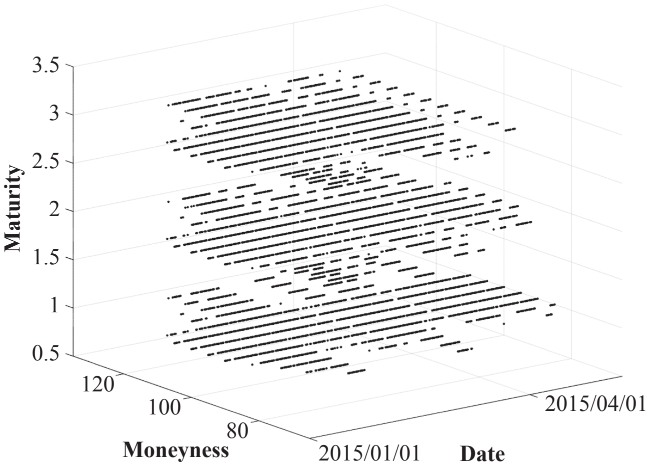
As pointed out by Wu and Liu (2018), the options exchanges in emerging markets have very narrow strike price bands (the difference between the upper and lower bound of strike prices) and insufficient intensity in strike price intervals. Furthermore, the exchanges often place more emphasis on season–month contracts (March, June, September, and December) than on non-season–month contracts. Thus, season–month contracts have broader strike bands. For these reasons, observed data points in emerging markets are not on a regular grid. The range of observed data points is limited, irregular, and time varying, which may have a great impact on the performance of curve fitting.

We take the China options market as an example here. China 50ETF options are physically settled on the SSE 50 Index ETF in China. Our China 50ETF options data consisted of monthly contracts from the beginning of the market on February 9, 2015 to April 22, 2016. In Exhibit 8, we see that the observed data barely cover the moneyness range from 70 to 130. In each day, observed data form an irregular grid, and the range of observed data changes over time. The limited irregular grid causes two problems.

First, some interpolation methods, including bilinear interpolation and bicubic interpolation, are required to input data on a two-dimensional regular grid. Amidror (2002) provided triangulation-based techniques to extend bilinear interpolation and bicubic interpolation to accommodate irregular grid data. Bilinear interpolation is extended to linear tetrahedral interpolation, and bicubic interpolation is extended to two-dimensional Clough–Tocher interpolation. Note that these two methods cannot be used for extrapolation.

EXHIBIT 8

Irregular Data Grids in China 50ETF Options Market from February 9, 2015 to April 22, 2016



Note: Each point represents the existence of an option contract with that maturity and that moneyness on that date.

Another problem is that extrapolation along the strike axis is required because of the limited moneyness range of observed data in emerging markets. Lee (2004) proved that the large strike tail of implied volatility is bounded by $O(\sqrt{\ln k})$, and the small strike tail of implied volatility is bounded by $O(\sqrt{\ln 1/k})$. He also did not recommend that either tail grow more slowly than the bound.

The Carr–Wu model satisfies exactly the bounds because the function $I^2 = f(\ln k)$ has two oblique asymptotes when ln k tends to +∞ or −∞. However, the statistical methods usually do not satisfy the bounds. In quadratic regression, the large strike tail grows by $O(k^2)$, and the small strike tail grows by $O(1)$. In thin plate interpolation and biharmonic interpolation, the large strike tail grows by $O(k^2 \ln k)$, and the small strike tail grows by $O(1)$. These three methods grow too fast as strike goes to positive infinity and grow too slowly as strike goes to zero. Here we propose a moneyness transform to improve these statistical methods to satisfy the Lee's condition.

For quadratic regression, we transform the moneyness into the following form:

$$x = \sqrt[4]{|\ln k|} \tag{27}$$

EXHIBIT 9

Performance of Different Methods in China 50ETF Options Market from February 9, 2015 to April 22, 2016

	Method	Irregular Grid	Extrapolation	Lee's Condition	MSE	R^2	AIC
1	Linear Regression	Yes	Yes	No	0.000709	0.921	–488.0
2	Quadratic Regression	Yes	Yes	Yes (transformed)	0.000286	0.968	–1,106.4
3	Nearest Neighbor	Yes	Yes	No	0.003819	0.575	892.6
4	Bilinear Interpolation	Yes (triangulated)	No	No	0.000388	0.957	–680.4
5	Bicubic Interpolation	Yes (triangulated)	No	No	0.000468	0.948	–551.0
6	LOWESS	Yes	Yes	No	0.000386	0.957	–683.7
7	LOESS	Yes	Yes	No	0.000428	0.952	–612.5
8	Thin Plate Interpolation	Yes	Yes	Yes (transformed)	0.000395	0.956	–732.5
9	Biharmonic Interpolation	Yes	Yes	Yes (transformed)	0.000560	0.938	–427.8
10	Nadaraya–Watson	Yes	Yes	No	0.000299	0.967	–859.8
11	Neural Network	Yes	Yes	No	0.001584	0.824	177.0
12	Carr and Wu	Yes	Yes	Yes	0.000488	0.946	–738.4

Then we regress implied volatilities against the transformed moneyness x. We can easily prove that either tail of transformed regression grows by $O(\sqrt{|\ln k|})$ and satisfies the Lee's condition.

For thin plate interpolation and biharmonic interpolation, the transform is required to satisfy

$$x^2 \ln x = \sqrt{|\ln k|} \tag{28}$$

Therefore, the transform is

$$x = \frac{\sqrt{2}\sqrt[4]{|\ln k|}}{\sqrt{W(2\sqrt{|\ln k|})}} \tag{29}$$

The transformed thin plate interpolation and biharmonic interpolation satisfy the Lee's condition. Other curve-fitting methods cannot meet the Lee's condition.

Exhibit 9 describes the performance of curve-fitting methods in the China 50ETF options market. Quadratic regression and the Nadaraya–Watson kernel regression have the highest R^2. Considering overfitting problems, these two methods still have the best AIC. The Carr–Wu model and thin plate interpolation are on the second tier. Therefore, transformed quadratic regression has the best goodness of fit with the Lee's condition, but the Carr–Wu model is a very good alternative considering that the theoretical model natively satisfies the Lee's condition and has economic implications.

CONCLUSION

We have reviewed 12 statistical and theoretical curving-fitting methods in options markets. IVSs are widely used in options markets, and the curve-fitting methodology of IVSs were discussed. We compared the methods based on three aspects: goodness of fit, smoothness, and economic meaning.

Based on empirical results, one needs to choose the appropriate method according to specific requirements. Three interpolation methods (thin plate interpolation, biharmonic interpolation, and cubic interpolation) are shown to provide the best goodness of fit and relatively good smoothness. In addition, quadratic regression, the Nadaraya–Watson kernel regression, and the Carr–Wu model generate the smoothest surface. However, only the Carr–Wu model can explain the economic states behind the surfaces.

We also discussed the curve-fitting approaches for irregular grid data in emerging options markets. We propose a transformation method to improve quadratic regression, thin plate interpolation, and biharmonic interpolation to satisfy the Lee's condition. When handling the China 50ETF options market data, quadratic regression achieves the best goodness of fit under the Lee's condition. The Carr–Wu model provides a very good alternative considering that the Carr–Wu model natively satisfies the Lee's condition and has economic implications.

ACKNOWLEDGMENTS

This work is supported by the Ministry of Science and Technology of the People's Republic of China under grant 2016YFC0503606, by the National Natural Science Foundation of China (grants 71825007, 71471055, 91546102), by the Chinese Academy of Sciences Frontier Scientific Research Key Project under grant QYZDB-SSW-SYS021, and by the Marianne and Marcus Wallenberg Foundation under grant MMW 2015.0007.

REFERENCES

Amidror, I. 2002. "Scattered Data Interpolation Methods for Electronic Imaging Systems: A Survey." *Journal of Electronic Imaging* 11 (2): 157–177.

Black, F., and M. Scholes. 1973. "The Pricing of Options and Corporate Liabilities." *The Journal of Political Economy* 81 (3): 637-54.

Bookstein, F. L. 1989. "Principal Warps: Thin-Plate Splines and the Decomposition of Deformations." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11 (6): 567–585.

Bowman, A. W., and A. Azzalini. *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*. New York: OUP Oxford, 1997.

Breeden, D. T., and R. H. Litzenberger. 1978. "Prices of State-Contingent Claims Implicit in Option Prices." *The Journal of Business* 51 (4): 621–651.

Carr, P., and L. Wu. 2016. "Analyzing Volatility Risk and Risk Premium in Option Contracts: A New Theory." *Journal of Financial Economics* 120 (1): 1–20.

Cleveland, W. S. 1979. "Robust Locally Weighted Regression and Smoothing Scatterplots." *Journal of the American Statistical Association* 74 (368): 829–836.

Cox, J. C., I. E. Jonathan Jr, and S. A. Ross. 1985. "A Theory of the Term Structure of Interest Rates." *Econometrica* 53 (2): 385–407.

Daglish, T., J. Hull, and W. Suo. 2007. "Volatility Surfaces: Theory, Rules of Thumb, and Empirical Evidence." *Quantitative Finance* 7 (5): 507–524.

Duchon, J. "Splines Minimizing Rotation-Invariant Semi-Norms in Sobolev Spaces." In *Constructive Theory of Functions of Several Variables*, edited by W. Schempp and K. Zeller, pp. 85–100. Berlin: Springer, 1977.

Fasshauer, G. E. *Meshfree Approximation Methods with Matlab (with CD-ROM)*. Singapore: World Scientific Publishing Company, 2007.

Fengler, M. R. *Semiparametric Modeling of Implied Volatility*. 1st ed. Berlin: Springer Science & Business Media, 2006.

——. 2009. "Arbitrage-Free Smoothing of the Implied Volatility Surface." *Quantitative Finance* 9 (4): 417–428.

Fengler, M. R., W. K. Härdle, and C. Villa. 2003. "The Dynamics of Implied Volatilities: A Common Principal Components Approach." *Review of Derivatives Research* 6 (3): 179–202.

Fengler, M. R., and Q. Wang. "Fitting the Smile Revisited: A Least Squares Kernel Estimator for the Implied Volatility Surface." Working paper, Social Science Electronic Publishing, 2003.

Gatheral, J. *The Volatility Surface: A Practitioner's Guide*. 1st ed. Hoboken, NJ: John Wiley & Sons, 2011.

Hagan, P. S., and G. West. 2006. "Interpolation Methods for Curve Construction." *Applied Mathematical Finance* 13 (2): 89–129.

Harder, R. L., and R. N. Desmarais. 1972. "Interpolation Using Surface Splines." *Journal of Aircraft* 9 (2): 189–191.

Härdle, W. *Applied Nonparametric Regression*. London: Cambridge University Press, 1990.

Heston, S. L. 1993. "A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options." *Review of Financial Studies* 6 (2): 327–343.

Ho, T. S. Y., and S. B. Lee. 1986. "Term Structure Movements and Pricing Interest Rate Contingent Claims." *The Journal of Finance* 41 (5): 1011–1029.

Hull, J., and A. White. 1987. "The Pricing of Options on Assets with Stochastic Volatilities." *The Journal of Finance* 42 (2): 281–300.

——. 1990. “Pricing Interest-Rate-Derivative Securities.” *Review of Financial Studies* 3 (4): 573–592.

Lee, R. W. 2004. “The Moment Formula for Implied Volatility at Extreme Strikes.” *Mathematical Finance* 14 (3): 469–480.

Lin, B. H. 2002. “Fitting Term Structure of Interest Rates Using B-Splines: The Case of Taiwanese Government Bonds.” *Applied Financial Economics* 12 (1): 57–75.

Merton, R. C. 1973. “Theory of Rational Option Pricing.” *Bell Journal of Economics & Management Science* 4 (1): 141–183.

Nadaraya, E. A. 1964. “On Estimating Regression.” *Theory of Probability and Its Applications* 9 (1): 157–159.

Rendleman, R. J., and B. J. Bartter. 1980. “The Pricing of Options on Debt Securities.” *Journal of Financial & Quantitative Analysis* 15 (1): 11–24.

Ron, U. *A Practical Guide to Swap Curve Construction*. Ottawa: Bank of Canada, 2000.

Rookley, C. 1997. “Fully Exploiting the Information Content of Intra Day Option Quotes: Applications in Option Pricing and Risk Management.” Working paper.

Schmidhuber, J. 2015. “Deep Learning in Neural Networks: An Overview.” *Neural Networks* 61: 85–117.

Schönbucher, P. J. 1999. “A Market Model for Stochastic Implied Volatility.” *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* 357 (1758): 2071–2092.

Scott, L. O. 1987. “Option Pricing When the Variance Changes Randomly: Theory, Estimation, and an Application.” *Journal of Financial and Quantitative Analysis* 22 (4): 419–438.

Vasicek, O. 1977. “An Equilibrium Characterization of the Term Structure.” *Journal of Financial Economics* 5 (4): 627.

Wallmeier, M., and R. Hafner. 2001. “The Dynamics of Dax Implied Volatilities.” *International Quarterly Journal of Finance* 1 (1): 1–27.

Watson, G. S. 1964. “Smooth Regression Analysis.” *The Indian Journal of Statistics* 26 (4): 359–372.

Werbos, P. 1974. “Beyond Regression: New Fools for Prediction and Analysis in the Behavioral Sciences.” Ph.D. thesis, Harvard University.

Wu, D., and T. Liu. 2018. “New Approach to Estimating VIX Truncation Errors Using Corridor Variance Swaps.” *The Journal of Derivatives* 25 (4): 54–70.

Zhu, Y., and M. Avellaneda. 1998. “A Risk-Neutral Stochastic Volatility Model.” *International Journal of Theoretical and Applied Finance* 1 (2): 289–310.