The Review of Financial Studies

Implied Stochastic Volatility Models

Yacine Aït-Sahalia
Department of Economics, Princeton University, and NBER

Chenxu Li
Guanghua School of Management, Peking University

Chen Xu Li
School of Business, Renmin University of China

This paper proposes "implied stochastic volatility models" designed to fit option-implied volatility data and implements a new estimation method for such models. The method is based on explicitly linking observed shape characteristics of the implied volatility surface to the coefficient functions that define the stochastic volatility model. The method can be applied to estimate a fully flexible nonparametric model, or to estimate by the generalized method of moments any arbitrary parametric stochastic volatility model, affine or not. Empirical evidence based on S&P 500 index options data show that the method is stable and performs well out of sample. (*JEL* G12, C51, C52)

Received April 22, 2019; editorial decision January 20, 2020 by Editor Ralph Koijen. Authors have furnished an Online Appendix, which is available on the Oxford University Press Web site next to the link to the final published paper online.

We are grateful to the Editor and two anonymous referees for very helpful comments and suggestions. We also benefited from the comments of participants at the 2017 Stanford-Tsinghua-PKU Conference in Quantitative Finance, the 2017 Fifth Asian Quantitative Finance Conference, the 2017 BCF-QUT-SJTU-SMU Conference on Financial Econometrics, the Second PKU-NUS Annual International Conference on Quantitative Finance and Economics, the 2017 Asian Meeting of the Econometric Society, the Third Annual Volatility Institute Conference at NYU Shanghai, the 2018 Review of Economic Studies 30th Anniversary Conference, the 2018 FERM Conference, the 2019 Econometric Society Asian Meetings, and seminars at the University of Chicago, London School of Economics and Political Science, University College London, University of Tokyo, Central European University, Center for Economic Research and Graduate Education, University of Zurich, and Scuola Normale in Pisa. The research of Chenxu Li was supported by the Guanghua School of Management, the Center for Statistical Science, and the Key Laboratory of Mathematical Economics and Quantitative Finance (Ministry of Education) at Peking University, as well as the National Natural Science Foundation of China (Grant 71671003). Chen Xu Li is grateful for a graduate scholarship and funding support from the Graduate School of Peking University as well as the support from the Bendheim Center for Finance at Princeton University and the School of Business at Renmin University of China. All authors contributed equally. Supplementary data can be found on *The Review of Financial Studies* website. Send correspondence to Chen Xu Li at the School of Business, Renmin University of China, Beijing, 100872, China; telephone: 86-10-82500564. E-mail: lichenxu@rmbs.ruc.edu.cn.

The Review of Financial Studies 34 (2021) 394–450

doi:10.1093/rfs/hhaa041 Advance Access publication March 30, 2020

In an ideal econometric situation, one specifies a model directly for the variables to be observed and estimates it. This is not possible in option pricing, since specifying a panel data model directly for a set of option prices of different maturities and strikes would most likely entail arbitrage opportunities. No-arbitrage pricing arguments instead start with an assumed dynamic model for the underlying asset price. That model is typically of the stochastic volatility type, as in, for example, Hull and White (1987), Heston (1993), Bates (1996), Duffie, Pan, and Singleton (2000), and Pan (2002). Difficulties arise because the relationship between the observed option prices and the state variables of the model is not fully explicit. Option prices can only be computed numerically or approximated, even under affine stochastic volatility models, for which option prices admit analytical Fourier transforms (see, e.g., Duffie, Pan, and Singleton 2000 and the references therein).

As the variety of affine or non-affine specifications suggests, there is no accepted consensus in the literature on what constitutes a correct model specification. There is, however, agreement that a stochastic volatility model should seek to produce option prices, or equivalently implied volatilities, sharing the main features of the observed data. A prevalent approach relies then on fitting a prespecified model to data by minimizing root mean squared pricing errors. Alternatively, models can be calibrated to fit a set (a continuum is often required) of options or other derivative prices exactly. Prominent examples of the latter approach are the local volatility model of Dupire (1994) and the results of Andersen and Andreasen (2000), Carr et al. (2004), Carr and Cousot (2011), and Carr and Cousot (2012) including local Lévy jumps.

We provide in this paper a new method that uses directly and explicitly the information contained in implied volatility data to construct an underlying stochastic volatility model, rather than a local volatility one. We call the resulting models "implied stochastic volatility models," since they are stochastic volatility models whose coefficient functions have been constructed to reproduce the salient empirical characteristics of the market-implied volatility surface. In the context of the Black-Scholes model, a single parameter, namely implied volatility, is all that is needed to characterize the model that is consistent with the market data. In the stochastic volatility context, the equivalent notion requires that we identify the implied coefficient functions driving the volatility dynamics to characterize the model that is consistent with the market data; this is what this paper does.

At each point in time, implied volatility data take the form of a surface representing the implied volatility of the option as a function of its moneyness and time-to-maturity. We use a small number of observable and practically useful "shape characteristics" of the implied volatility surface, including but not limited to the slope of the implied volatility smile/smirk that has attracted most of the attention, to fully characterize the underlying stochastic volatility

model. We will do so both nonparametrically, in order to retain maximum flexibility to fit the data, as well as parametrically.[1]

The implied volatility shape characteristics we employ for this purpose are along both the log-moneyness dimension and the term-structure dimension, and typically take the form of level, slope, and curvature. Although a finite number of shape characteristics cannot fully represent a given surface, these specific shape characteristics have been shown to be useful descriptions of the implied volatility data when trading portfolios of options; see, for example, Carr and Wu (2007), Bakshi, Carr, and Wu (2008), and Durrleman and El Karoui (2008). Various strategies have been designed to expose and/or hedge the risks reflected by these shape characteristics, such as straddle, risk reversal, and butterfly spread, exposing the risks of at-the-money level, slope, and convexity, respectively, along the log-moneyness dimension, as well as the calendar spread exposing the risk of the slope of at-the-money implied volatility curve along the term-structure dimension.

In the nonparametric case, we show how the coefficient functions of the stochastic volatility model can be recovered (or inverted) from the previously mentioned shape characteristics by standard local polynomial regression. In the parametric case, the observed shape characteristics can be set up as a set of explicit generalized method of moments (GMM) conditions to estimate the parameters of the stochastic volatility model. Our construction of implied stochastic volatility models is simple to implement, even for the non-affine models commonly regarded as analytically intractable, thanks to the closed-form formulae we provide for the shape characteristics of arbitrarily general stochastic volatility models. Indeed, the main method in the literature to estimate an option pricing model from option prices consists of minimizing the root mean squared pricing errors between the model and the market prices. This requires the numerical computation of the model's prices each time the minimization algorithm adjusts the model's parameter values; such minimization on top of the numerical re-pricing at the slightly altered parameter values can be numerically very challenging, especially outside the affine class of models. By contrast, the estimation method in this paper requires no numerical computation of option prices, or any symbolic calculation, since we already provide the relevant closed-form formulae for any model's coefficients to be plugged into. And in the parametric case, our estimation method reduces to a standard GMM procedure for which we provide explicit moment conditions.

[1] This paper is not the first one to estimate a nonparametric stochastic volatility model, or more broadly, a continuous-time diffusion or jump-diffusion model: see Jiang and Knight (1997), Bandi and Phillips (2003), and Renò (2008) for estimating scalar diffusions, as well as Comte, Genon-Catalot, and Rozenholc (2010), Kanaya and Kristensen (2016), and Bandi and Renò (2018) for estimating stochastic volatility models. These methods focus, however, on estimating nonparametrically the dynamics of a model under the physical probability measure based on time-series observations. By contrast, the method we propose nonparametrically estimates the risk-neutral dynamics for derivative pricing, using cross-sectional implied volatility data as inputs.

Applying the method to S&P 500 index options, we take as inputs the level, slope, and convexity of the implied volatility surface in the log-moneyness direction, and the slope in the term-structure direction. We find empirically that the resulting nonparametric implied stochastic volatility model has the following features: a strong leverage effect between the innovations in returns and volatility, mean reversion in volatility, monotonicity, and state dependency in volatility of volatility. We find that the model also matches surprisingly well the convexity in the term-structure dimension and the mixed slope characterizing the sensitivity to the change of smile/smirk slope with respect to a change of time-to-maturity, even though those are not employed as inputs. As a result, the nonparametric implied stochastic volatility model is capable of fitting all six observable (and useful in practice) shape characteristics up to the second order. We also find that the estimated nonparametric characteristics of the model are stable over time, reducing the need for constant refitting of the model, and that the model performs well out of sample.

From a technical perspective, part of our procedure for constructing implied stochastic volatility models is made possible by virtue of a closed-form bivariate expansion of implied volatilities in time to maturity and log-moneyness. Various types of expansions for implied volatilities or option prices, obtained using different methods, are available in the literature. They include: small volatility-of-volatility expansions near a nonstochastic volatility, also known as small ε or small noise expansions (Kunitomo and Takahashi 2001 and Takahashi and Yamada 2012); expansion based on slow-varying volatility (Sircar and Papanicolaou 1999 and Lee 2001); expansion based on fast-varying and slow-varying analysis (Fouque, Lorig, and Sircar 2016); short maturity expansions (see Medvedev and Scaillet 2007 for an expansion with respect to the square of time-to-maturity with expansion term sorted in terms of moneyness scaled by volatility, Durrleman 2010 with a correction due to Pagliarani and Pascucci 2017, and Lorig, Pagliarani, and Pascucci 2017); expansion using partial differential equation methods (Berestycki, Busca, and Florent 2004); singular perturbation expansion, (Hagan and Woodward 1999); expansion around an auxiliary model (Kristensen and Mele 2011); expansion using transition density expansion (Gatheral et al. 2012 and Xiu 2014); expansion of the characteristic function (Jacquier and Lorig 2015). Some of these methods apply generally, while others apply only to specific models, such as the Heston model, as in Forde, Jacquier, and Lee (2012) (short maturity), Forde and Jacquier (2011) (long maturity), or exponential Lévy models as in Andersen and Lipton (2013). The asymptotic behavior of implied volatilities as time-to-maturity approaches zero is important: for the continuous case, see Ledoit, Santa-Clara, and Yan (2002) and Berestycki, Busca, and Florent (2002), and with jumps, see Carr and Wu (2003) and Durrleman (2008). Finally, a number of asymptotic results concerning long-dated, short-dated, far out of the money strike, and jointly varying strike-expiration regimes are available; see Lee (2004), Gao and Lee (2014), and Tehranchi (2009).

The expansion we employ is different from existing ones; it takes the form of a bivariate series in time-to-maturity and log-moneyness, applies to general stochastic volatility models, and produces closed-form expressions for arbitrary stochastic volatility models with or without jumps. Given the extensive literature on expansions, however, the novelty in this paper is not its expansion (although it is new). The existing literature on implied volatility expansions has primarily been concerned with the mathematical derivation of specific expansions and their properties, rather than with using such expansions for the purposes of estimating the model that underlies the expansion. The main contribution of the paper is a new method that uses restrictions derived from the expansion to construct and estimate an implied stochastic volatility model that is consistent with observed characteristics of the implied volatility data, either parametrically or nonparametrically. A main feature of the method is that it is fully closed-form, yet applicable to general stochastic volatility models, including non-affine ones, without the need to optimize numerically computed option prices.

1. Stochastic Volatility Models and Implied Volatility Surfaces

Consider a generic continuous bivariate stochastic volatility (SV) model.[2] Under an assumed risk-neutral measure, the price of the underlying asset S_t and its volatility v_t jointly follow a diffusion process

$$\frac{dS_t}{S_t} = (r-d)dt + v_t dW_{1t}, \tag{1}$$

$$dv_t = \mu(v_t)dt + \gamma(v_t)dW_{1t} + \eta(v_t)dW_{2t}. \tag{2}$$

We will add jumps in returns to the model in Section 6. Here, r and d are the risk-free rate and the dividend yield of the underlying asset, both assumed constant for simplicity, and observable; W_{1t} and W_{2t} are two independent standard Brownian motions; μ, γ, and η are scalar functions. The generic specification in Equations (1)–(2) nests all existing continuous bivariate SV models. For models conventionally expressed in terms of instantaneous variance rather than volatility (e.g., the model of Heston 1993), it is straightforward to obtain the equivalent form of Equations (1)–(2) by Itô's lemma. Our objective is to fully identify the model, that is, v_t at each discrete instant at which data sampling occurs, and the unknown functions $\mu(\cdot)$, $\gamma(\cdot)$, and $\eta(\cdot)$. This is a natural extension to stochastic volatility models of the question answered in Dupire (1994) for local volatility models, which relied on a method that cannot be used in the stochastic volatility context.[3]

---

[2] We focus on the bivariate case for illustrating the method, which in principle applies to SV models with any finite number of factors.

[3] Local volatility models are of the form $dS_t/S_t = (r-d)dt + \sigma(S_t)dW_t$. The approach of Dupire (1994), based on inverting the pricing equation for the function $\sigma(\cdot)$, cannot be extended from the local to the stochastic volatility

We are also interested in the leverage effect coefficient function as the correlation function between asset returns and innovations in spot volatility, defined as in Aït-Sahalia, Fan, and Li (2013) by

$$\rho(v_t)=\frac{\gamma(v_t)}{\sqrt{\gamma(v_t)^2+\eta(v_t)^2}}. \tag{3}$$

This coefficient function is identified once the other components of the model are. In general, $\rho(v_t)$ is empirically found to be negative, and is in general stochastic since the dependence in v_t need not cancel out between the numerator and denominator in Equation (3): see, for example, the models of Jones (2003) and Chernov et al. (2003), among others. For $\rho(v_t)$ to be independent of v_t, that is, $\rho(v)\equiv\rho$ for some constant ρ, it must be that $\eta(v)=\gamma(v)\sqrt{1-\rho^2}/\rho$, that is, the two functions $\eta(v)$ and $\gamma(v)$ are uniformly proportional to each other. This is the case in the model of Heston (1993), for instance.

The arbitrage-free price of a European-style put option with maturity T, that is, time-to-maturity $\tau=T-t$, and exercise strike K is (in terms of log-moneyness $k=\log(K/S_t)$),

$$P(\tau,k,S_t,v_t)=e^{-r\tau}\mathbb{E}_t[\max(S_te^k-S_T,0)],$$

where $\mathbb{E}_t$ denotes the risk-neutral conditional expectation given the information up to time t. In practice, the market price of an option is typically quoted through its Black-Scholes implied volatility (IV) Σ, that is, the value of the volatility parameter, which, when plugged into the Black-Scholes formula $P_{\mathrm{BS}}(\tau,k,S_t,\sigma)$, leads to a theoretical value equal to the observed market price of the option:[4]

$$P_{\mathrm{BS}}(\tau,k,S_t,\Sigma)=P(\tau,k,S_t,v_t).$$

Viewed simply as mapping actual option prices into a different unit, using implied volatilities does not require that the assumptions of the Black-Scholes model be satisfied, and has a few advantages: implied volatilities are independent of the scale of the underlying asset value or strike price, deviations from a flat IV surface denote deviations from the Black-Scholes model (or equivalently deviations from the Gaussianity of log-returns), and such deviations can be monotonically interpreted (the higher the IV above the flat level, the more expensive the option, and similarly below), so differences in IV allow for relative value comparisons between options.

1.1 From stochastic volatility to implied volatility

The IV depends on S_t only through k, that is, $\Sigma=\Sigma(\tau,k,v_t)$. This is because the option price can be written in a form proportional to the time-t price S_t,

$$P(\tau,k,S_t,v_t)=S_t\bar{P}(\tau,k,v_t)$$

---

situation: when employed in a stochastic volatility setting, it can only characterize $\mathbb{E}[v_T|S_T,S_0]$ rather than the full dynamics in Equation (2).

[4] Model-implied volatilities calculated from put and call options are identical by put-call parity.

with

$$\bar{P}(\tau,k,v_t)=e^{-r\tau}\mathbb{E}_t\left[\max\left(e^k-\frac{S_T}{S_t},0\right)\right]. \tag{4}$$

For a given log-moneyness k, the function $\bar{P}(\tau,k,v_t)$ is independent of the initial underlying asset price S_t since the dynamics in Equation (1) of the underlying asset price imply that the ratio S_T/S_t is independent of S_t, so is the expectation function in Equation (4) for defining $\bar{P}(\tau,k,v_t)$. Writing $P_{\mathrm{BS}}(\tau,k,S_t,\sigma)=S_t\bar{P}_{\mathrm{BS}}(\tau,k,\sigma)$, the IV Σ is determined by

$$\bar{P}_{\mathrm{BS}}(\tau,k,\Sigma)=\bar{P}(\tau,k,v_t), \tag{5}$$

and therefore

$$\Sigma(\tau,k,v_t)=\bar{P}_{\mathrm{BS}}^{-1}(\tau,k,\bar{P}(\tau,k,v_t)). \tag{6}$$

The mapping $(\tau,k)\longmapsto\Sigma(\tau,k,v_t)$ at a given time t is the (model) IV surface at that time. We will consider several shape characteristics of the IV surface, such as its slope and convexity along the log-moneyness and the term-structure dimensions, defined by the partial derivative $\partial^{i+j}\Sigma/\partial\tau^i\partial k^j$ for integers $i,j\geq 0$. In particular, we will focus on the at-the-money ($k=0$) and short maturity ($\tau\to 0$) shape characteristics

$$\Sigma_{i,j}(v_t)=\lim_{\tau\to 0}\frac{\partial^{i+j}}{\partial\tau^i\partial k^j}\Sigma(\tau,0,v_t). \tag{7}$$

To illustrate, we show in Figure 1 the S&P 500 index IV surface on a given day along with the two slopes $\Sigma_{0,1}(v_t)$ (log-moneyness slope, or IV smile/smirk) and $\Sigma_{1,0}(v_t)$ (term-structure slope) as red and blue dashed lines, respectively. Under standard regularity conditions on the coefficients functions of the SV model, the limit in Equation (7) exists for any arbitrary integers i and j under the continuous SV model in Equations (1)–(2); this follows from results in Durrleman (2010) and Pagliarani and Pascucci (2017).

We treat the shape characteristics of the IV surface as observable from IV data—we will describe how below—and then use them to recover the SV model in Equations (1)–(2) that is compatible with them. We show that such a recovery can be achieved either in a parametric way, or more flexibly, in a nonparametric way without a priori parametrization: this is the main contribution of this paper. We call the resulting model an implied stochastic volatility (ISV) model.

To achieve this, we need to express the shape characteristics $\Sigma_{i,j}(\cdot)$ in terms of v_t and the model coefficient functions $\mu(\cdot)$, $\gamma(\cdot)$, and $\eta(\cdot)$ in Equation (2). We first note that $\Sigma_{0,0}$, representing the level of the IV surface, must be given by the instantaneous volatility

$$\Sigma_{0,0}(v_t)=v_t, \tag{8}$$

which is a well-known fact (see, e.g., Ledoit, Santa-Clara, and Yan 2002 and Durrleman 2008). Beyond this leading term, we focus on the at-the-money level $\Sigma_{0,0}$, slope $\Sigma_{0,1}$, and convexity $\Sigma_{0,2}$ along the log-moneyness dimension, as

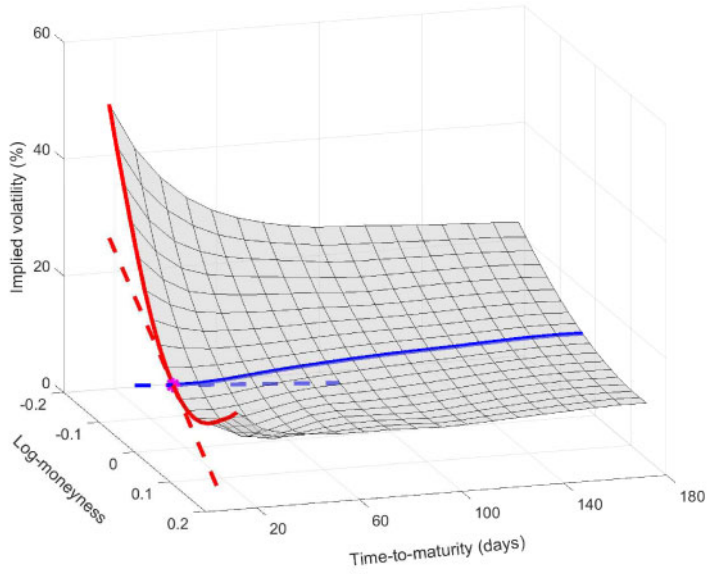

Figure 1
IV surface and shape characteristics of S&P 500 index options on January 3, 2017
This plot represents the IV surface $(\tau,k)\longmapsto\Sigma(\tau,k,v_t)$ on January 3, 2017, for S&P 500 index options. The two slopes $\Sigma_{0,1}(v_t)$ (log-moneyness slope, or implied volatility smile) and $\Sigma_{1,0}(v_t)$ (term-structure slope) are approximated and represented as red and blue dashed lines, respectively, with each partial derivative $\partial^{i+j}\Sigma(\tau,0,v_t)/\partial\tau^i\partial k^j$ evaluated at $\tau=1$ month.

well as the slope $\Sigma_{1,0}$ along the term-structure dimension, all for short time-to-maturity. These four basic shape characteristics describe the skeleton of the IV surface. Mathematical details are contained in the Appendix. We show in Appendix A that

$$\Sigma_{0,1}(v_t)=\frac{\gamma(v_t)}{2v_t},\ \Sigma_{0,2}(v_t)=\frac{1}{6v_t^3}[2v_t\gamma(v_t)\gamma'(v_t)+2\eta(v_t)^2-3\gamma(v_t)^2], \tag{9}$$

$$\Sigma_{1,0}(v_t)=\frac{1}{24v_t}[2\gamma(v_t)(6(d-r)-2v_t\gamma'(v_t)+3v_t^2)+12v_t\mu(v_t)$$
$$+3\gamma(v_t)^2+2\eta(v_t)^2]. \tag{10}$$

Since $\Sigma_{i,j}$ are the coefficients (up to constants) of the bivariate Taylor expansion of IV $\Sigma(\tau,k,v_t)$:

$$\Sigma^{(J,\mathbf{L}(J))}(\tau,k,v_t)=\sum_{j=0}^{J}\sum_{i=0}^{L_j}\sigma^{(i,j)}(v_t)\tau^i k^j,\ \text{with}\ \sigma^{(i,j)}(v_t)=\frac{\Sigma_{i,j}(v_t)}{i!j!}, \tag{11}$$

up to integer expansion orders J and $\mathbf{L}(J)=(L_0,L_1,\cdots,L_J)$ with $L_j\geq 0$, Equation (11) provides a tool for inferring $\Sigma_{i,j}$ as coefficients in a (nonlinear)

regression of IV on log-moneyness and time-to-maturity. So we can treat the shape characteristics $\Sigma_{0,0}(\cdot)$, $\Sigma_{0,1}(\cdot)$, $\Sigma_{0,2}(\cdot)$, and $\Sigma_{1,0}(\cdot)$ (and higher-order ones if necessary) as observable from options data.

1.2 From implied volatility to stochastic volatility

The main idea in this paper is to now invert the explicit relations in Equations (8)–(10) back into the unknown coefficients functions of the SV model, $\mu(\cdot)$, $\gamma(\cdot)$, and $\eta(\cdot)$. In other words, we can view these relations as providing a system of equations to be solved for $\gamma(\cdot)$, $\eta(\cdot)$, and $\mu(\cdot)$, given the IV surface characteristics. This leads to a useful estimation method because it turns out that the system can be inverted in fully closed form, so no further approximation, numerical solution of a differential equation, or other numerical inversion is required.

First, observe that $v_t = \Sigma_{0,0}(v_t)$ and Equations (9)–(10) imply

$$\gamma(v_t) = 2\Sigma_{0,0}(v_t)\Sigma_{0,1}(v_t), \tag{12}$$

and

$$\eta(v_t) = \left[3\Sigma_{0,0}(v_t)^3\Sigma_{0,2}(v_t) - \Sigma_{0,0}(v_t)\gamma(v_t)\gamma'(v_t) + \frac{3}{2}\gamma(v_t)^2\right]^{\frac{1}{2}}, \tag{13}$$

$$\mu(v_t) = 2\Sigma_{1,0}(v_t) + \frac{\gamma(v_t)}{6}(2\gamma'(v_t) - 3\Sigma_{0,0}(v_t)) - \frac{\eta(v_t)^2}{6\Sigma_{0,0}(v_t)} - \frac{\gamma(v_t)}{\Sigma_{0,0}(v_t)}\left(d - r + \frac{1}{4}\gamma(v_t)\right). \tag{14}$$

Second, plug in Equation (12) into Equation (13), and then plug in both expressions into Equation (14) to obtain:

Theorem 1. The coefficient functions $\gamma(\cdot)$, $\eta(\cdot)$, and $\mu(\cdot)$ of the SV model in Equations (1)–(2) can be recovered in closed form as functions of the level $\Sigma_{0,0}(\cdot)$, log-moneyness slope $\Sigma_{0,1}(\cdot)$ and convexity $\Sigma_{0,2}(\cdot)$, as well as term-structure slope $\Sigma_{1,0}(\cdot)$ as follows:

$$\gamma(v_t) = 2\Sigma_{0,0}(v_t)\Sigma_{0,1}(v_t), \tag{15}$$

and

$$\eta(v_t) = \left[3\Sigma_{0,0}(v_t)^3\Sigma_{0,2}(v_t) + 2\Sigma_{0,0}(v_t)^2\Sigma_{0,1}(v_t)^2 - 4\Sigma_{0,0}(v_t)^2\Sigma_{0,1}(v_t)\Sigma'_{0,1}(v_t)\right]^{\frac{1}{2}}, \tag{16}$$

$$\mu(v_t) = \Sigma_{0,0}(v_t)^2\left[\Sigma_{0,1}(v_t)(2\Sigma'_{0,1}(v_t) - 1) - \frac{1}{2}\Sigma_{0,2}(v_t)\right] - 2(d - r)\Sigma_{0,1}(v_t) + 2\Sigma_{1,0}(v_t). \tag{17}$$

where $\Sigma'_{0,1}(v_t)$ represents the first-order derivative of $\Sigma'_{0,1}(v_t)$ with respect to v_t.

This new result characterizes the closed-form relation between the SV model coefficient functions and the IV shape characteristics. It has the following interesting implications. First, Equation (15) shows that for a given IV level $\Sigma_{0,0}(v_t)$, the log-moneyness slope $\Sigma_{0,1}(v_t)$ plays an important role in determining the volatility function $\gamma(v_t)$ attached to the common Brownian shocks W_{1t} of the asset price S_t and its volatility v_t. For a given level of $\Sigma_{0,0}(v_t)$, a steeper log-moneyness slope $\Sigma_{0,1}(v_t)$ results in a higher absolute value of the volatility function $\gamma(v_t)$. Second, from Equation (16), a steeper slope $\Sigma_{0,1}(v_t)$ has an effect on the volatility function $\eta(v_t)$ attached to the idiosyncratic Brownian shock W_{2t} in the volatility dynamics, which can be of either sign. Besides the level $\Sigma_{0,0}(v_t)$ and slope $\Sigma_{0,1}(v_t)$, the log-moneyness convexity $\Sigma_{0,2}(v_t)$ also matters for the volatility function $\eta(v_t)$. The total spot volatility of volatility is $\sqrt{\gamma(v_t)^2+\eta(v_t)^2}$, so for a given level of $\Sigma_{0,0}(v_t)$ and $\Sigma_{0,1}(v_t)$, a greater convexity $\Sigma_{0,2}(v_t)$ results in a larger volatility of volatility. Third, from Equations (3) and (15), we see that the sign of the leverage effect coefficient $\rho(v_t)$ is determined by the sign of the slope $\Sigma_{0,1}(v_t)$: as is typically the case in the data, a downward-sloping IV smile/smirk, $\Sigma_{0,1}(v_t)<0$, translates directly into $\rho(v_t)<0$. Further, $\rho(v_t)$ is monotonically decreasing (in absolute value) in $\eta(v_t)$, so it follows from Equations (3) and (16) that, for a given level of $\Sigma_{0,0}(v_t)$ and $\Sigma_{0,1}(v_t)$, a greater convexity $\Sigma_{0,2}(v_t)$ leads to a larger volatility of volatility, and consequently, a weaker leverage effect $\rho(v_t)$. Finally, Equation (17) shows that for a given level of $\Sigma_{0,0}(v_t)$, $\Sigma_{0,1}(v_t)$, and $\Sigma_{0,2}(v_t)$, an increase of the term-structure slope $\Sigma_{1,0}(v_t)$ on the IV surface results in an increase in the drift $\mu(v_t)$, that is, a faster expected change of the instantaneous volatility v_t.

2. Constructing a Nonparametric Implied Stochastic Volatility Model

We now use Theorem 1 to construct a nonparametric ISV model. By construction, the estimated model will generate option prices that match the observed features of the IV surface.

We start by estimating the shape characteristics $\Sigma_{i,j}$ from the observed IV surfaces. Assume that a total of n IV surfaces are observed with equidistant time interval Δ, without loss of generality. On day l, we observe n_l implied volatilities $\Sigma^{\text{data}}(\tau_l^{(m)},k_l^{(m)})$ along with time-to-maturity $\tau_l^{(m)}$ and log-moneyness $k_l^{(m)}$ for $m=1,2,\ldots,n_l$.

We can view the bivariate expansion in Equation (11) as a standard polynomial regression of IV on time-to-maturity τ and log-moneyness k, where the expansion coefficient $\sigma^{(i,j)}(v_t)$ corresponds to the regression coefficient of $\tau^i k^j$. On day l, we regress

$$\Sigma^{\text{data}}(\tau_l^{(m)},k_l^{(m)})=\sum_{j=0}^{J}\sum_{i=0}^{L_j}\beta_l^{(i,j)}(\tau_l^{(m)})^i(k_l^{(m)})^j+\epsilon_l^{(m)}, \tag{18}$$

for $m=1,2,\ldots,n_l$, where $\epsilon_l^{(m)}$ represent observation errors, assumed independent and identically distributed (i.i.d.) as well as exogenous with zero means.[5] From the estimator $\hat{\beta}_l^{(i,j)}$, $i,j\geq 0$, the shape characteristics $\Sigma_{i,j}(v_{l\Delta})$ on the lth day are estimated as

$$[\Sigma_{i,j}]_l^{\text{data}}=i!\,j!\,\hat{\beta}_l^{(i,j)},\ \text{for } i,j\geq 0; \tag{19}$$

in particular, $v_{l\Delta}=[\Sigma_{0,0}]_l^{\text{data}}=\hat{\beta}_l^{(0,0)}$. Note that while the objects of interest $\Sigma_{i,j}$ are derivatives of the IV surface Σ evaluated at $(\tau,k)=(0,0)$, the regression in Equation (18) includes observations with (τ,k) away from $(0,0)$ in order to estimate these partial derivatives.

Theorem 1 can now be employed to construct the ISV model. First, to estimate $\gamma(\cdot)$ nonparametrically, we rely on Equation (15). Let

$$[\gamma]_l^{\text{data}}=2[\Sigma_{0,0}]_l^{\text{data}}[\Sigma_{0,1}]_l^{\text{data}}, \tag{20}$$

and consider the nonparametric regression

$$[\gamma]_l^{\text{data}}=\gamma(v_{l\Delta})+\epsilon_l, \tag{21}$$

where $v_{l\Delta}$ is the explanatory variable, and ϵ_l represents an exogenous observation error. The function $\gamma(\cdot)$ can be estimated based on Equation (21) using any number of nonparametric regression methods; in what follows, we employ locally linear kernel regression (see, e.g., Fan and Gijbels 1996).

To estimate the coefficient functions $\eta(\cdot)$ and $\mu(\cdot)$, we implement the closed-form relations in Equations (13)–(14).[6] Note that these equations require estimating both the function γ and its derivative γ'. One advantage of locally linear kernel regression is that it provides in one pass an estimator not only of the regression function but also of its derivative. For two arbitrary points v and w, suppose that $\gamma(w)$ can be approximated by its first-order Taylor expansion around $w=v$, that is, $\gamma(w)\approx\gamma(v)+\gamma'(v)(w-v)$. Then, for any arbitrary value v of the independent variable, $[\gamma]_l^{\text{data}}$ is regarded as being approximately generated from the local linear regression as follows:

$$[\gamma]_l^{\text{data}}\approx\alpha_0+\alpha_1(v_{l\Delta}-v)+\epsilon_l,$$

where the localization argument makes the intercept α_0 and slope α_1 coincide with γ and its first-order derivative γ' evaluated at v, respectively, that is,

$$\hat{\gamma}(v)=\hat{\alpha}_0 \text{ and } \hat{\gamma}'(v)=\hat{\alpha}_1.^{7}$$

---

[5] This is a generalization of the linear regression in Dumas, Fleming, and Whaley (1998) of implied volatilities on τ and $K=S_t e^k$. The expansion in Equation (11) provides a justification for the design of this regression equation.

[6] It is mathematically equivalent to implement the closed-form formulae in Equations (16)–(17) in Theorem 1.

[7] Note that $\hat{\gamma}'(v)$ is an estimator of $\gamma'(v)$ but is not the derivative of $\hat{\gamma}(v)$.

The estimators $\hat{\alpha}_0$ and $\hat{\alpha}_1$ are obtained from the following weighted least squares minimization problem

$$(\hat{\alpha}_0, \hat{\alpha}_1) = \underset{\alpha_0, \alpha_1}{\operatorname{argmin}} \sum_{l=1}^{n} ([\gamma]_l^{\text{data}} - \alpha_0 - \alpha_1(v_{l\Delta} - v))^2 \mathcal{K}\left(\frac{v_{l\Delta} - v}{h}\right), \tag{22}$$

where $\mathcal{K}$ denotes a kernel function and h the bandwidth. In practice, we use the Epanechnikov kernel

$$\mathcal{K}(z) = \frac{3}{4}(1 - z^2)1_{\{|z|<1\}},$$

and a bandwidth h selected either by the standard rule of thumb or by standard cross-validation, which minimizes the sum of leave-one-out squared errors. The sum of leave-one-out squared errors, for example, in the case of the volatility function γ, is given by $\sum_{l=1}^{n}([\gamma]_l^{\text{data}} - \hat{\alpha}_{0,-l})^2$, where $\hat{\alpha}_{0,-l}$ is the local linear estimator $\hat{\alpha}_0$, at $v = v_{l\Delta}$, obtained from the weighted least squares problem in Equation (22) but without using the lth observation $(v_{l\Delta}, [\gamma]_l^{\text{data}})$. For a choice of kernel function $\mathcal{K}$ with bandwidth h, the solution of the weighted least squares problem in Equation (22) is explicitly given by

$$\hat{\alpha}_0 = \left(\sum_{i,j=1}^{n} s_{ij}(v)(v_{i\Delta} - v)\right)^{-1} \left(\sum_{i,j=1}^{n} s_{ij}(v)(v_{i\Delta} - v)[\gamma]_j^{\text{data}}\right), \tag{23}$$

and

$$\hat{\alpha}_1 = -\left(\sum_{i,j=1}^{n} s_{ij}(v)(v_{i\Delta} - v)\right)^{-1} \left(\sum_{i,j=1}^{n} s_{ij}(v)[\gamma]_j^{\text{data}}\right), \tag{24}$$

where

$$s_{ij}(v) = \mathcal{K}\left(\frac{v_{i\Delta} - v}{h}\right)\mathcal{K}\left(\frac{v_{j\Delta} - v}{h}\right)(v_{i\Delta} - v_{j\Delta}).$$

Next, in light of Equation (13), it is natural to calculate the data of $\eta(v_{l\Delta})$ as

$$[\eta]_l^{\text{data}} = \left[3([\Sigma_{0,0}]_l^{\text{data}})^3[\Sigma_{0,2}]_l^{\text{data}} - [\Sigma_{0,0}]_l^{\text{data}}\hat{\gamma}(v_{l\Delta})\hat{\gamma}'(v_{l\Delta}) + \frac{3\hat{\gamma}(v_{l\Delta})^2}{2}\right]^{\frac{1}{2}},$$

given $[\Sigma_{0,0}]_l^{\text{data}}$ and $[\Sigma_{0,2}]_l^{\text{data}}$, as well as the estimators of γ and γ' obtained previously. In practice, on the right-hand side of the above equation, the quantity inside the bracket $[\cdot]^{1/2}$ may take a negative value, owing to sampling noise in the data $[\Sigma_{0,0}]_l^{\text{data}}$ and $[\Sigma_{0,2}]_l^{\text{data}}$. To avoid this problem, we work instead with data $[\eta^2]_l^{\text{data}}$ calculated as

$$[\eta^2]_l^{\text{data}} = 3([\Sigma_{0,0}]_l^{\text{data}})^3[\Sigma_{0,2}]_l^{\text{data}} - [\Sigma_{0,0}]_l^{\text{data}}\hat{\gamma}(v_{l\Delta})\hat{\gamma}'(v_{l\Delta}) + \frac{3\hat{\gamma}(v_{l\Delta})^2}{2}. \tag{25}$$

We then estimate the coefficient function $\eta^2(\cdot)$ at each value v by a kernel regression that localizes the data $[\eta^2]_l^{\text{data}}$ at each point $v = v_{l\Delta}$, as we did in Equation (22) for $\gamma(\cdot)$. In our experience, the estimator $\hat{\eta}^2(\cdot)$ is always nonnegative thanks to the kernel smoothing (even though a small number of the plugged-in data points $[\eta^2]_l^{\text{data}}$ may be negative). We then define $\hat{\eta}(\cdot) \equiv \left[\hat{\eta}^2(\cdot)\right]^{1/2}$.

Finally, in light of Equation (14), we define

$$[\mu]_l^{\text{data}} = 2[\Sigma_{1,0}]_l^{\text{data}} + \frac{\hat{\gamma}(v_{l\Delta})}{6}(2\hat{\gamma}'(v_{l\Delta}) - 3[\Sigma_{0,0}]_l^{\text{data}}) - \frac{\hat{\eta}(v_{l\Delta})^2}{6[\Sigma_{0,0}]_l^{\text{data}}} - \frac{\hat{\gamma}(v_{l\Delta})}{[\Sigma_{0,0}]_l^{\text{data}}}\left(d - r + \frac{1}{4}\hat{\gamma}(v_{l\Delta})\right), \tag{26}$$

given the data $[\Sigma_{1,0}]_l^{\text{data}}$, estimators of γ, γ', and η^2 obtained previously, and estimate the coefficient function $\mu(\cdot)$ at each value v using the data computed from Equation (26) and the same kernel localization procedure in Equation (22) as employed for $\hat{\gamma}(\cdot)$ and $\hat{\eta}^2(\cdot)$.

To summarize, the construction of a nonparametric ISV model consists of two steps: first, we estimate the shape characteristics from the IV surface data via standard regression; second, we recover the model coefficient functions nonparametrically by combining the closed-form relations in Equations (12)–(14).

3. Estimating a Parametric Implied Stochastic Volatility Model

Although the nonparametric ISV model illustrated in Section 2 offers a maximal degree of flexibility for a bivariate SV model to generate option prices matching the observed features of the IV surface, one may still be interested in a parametric version of an ISV model. We show how the same explicit formulae relating the shape characteristics to the (now parametric) coefficient functions can now be cast as a system of moment conditions in GMM. Importantly, Theorem 1 applies generically so the formulae it contains do not need to be re-derived: simply plug in the assumed specification of the parametric model.

The SV model in Equations (1)–(2) is now a parametric one, so that $\mu(\cdot) = \mu(\cdot;\theta)$, $\gamma(\cdot) = \gamma(\cdot;\theta)$, and $\eta(\cdot) = \eta(\cdot;\theta)$, where θ denotes the vector of unknown parameters to be estimated in a compact space $\Theta \subset \mathbb{R}^K$, and θ_0 denotes their true values. We further assume that the parametric functions are known, and twice continuously differentiable in θ, and also assume that the data are stationary and strong mixing with rate greater than two.

To estimate θ, we propose to form moment conditions as follows:

$$g^{(i,j)}(v_{l\Delta};\theta) = [\Sigma_{i,j}]_l^{\text{data}} - [\Sigma_{i,j}(v_{l\Delta};\theta)]^{\text{model}}, \tag{27}$$

where $[\Sigma_{i,j}]_l^{\text{data}}$ once again denotes the data of shape characteristics $\Sigma_{i,j}$, and $[\Sigma_{i,j}(v_{l\Delta};\theta)]^{\text{model}}$ denotes the closed-form formulae of $\Sigma_{i,j}$ provided in

Equations (9)–(10), and additional higher orders if necessary. We gather the different moment conditions $g^{(i,j)}$ into a vector

$$g(v_{l\Delta};\theta)=(g^{(i,j)}(v_{l\Delta};\theta))_{(i,j)\in I}$$

for some integer index set I consisting of nonnegative integer pairs (i,j) such that $i+j\geq 1$: typically, $I=\{(1,0),(0,1),(0,2)\}$. The choice of moment conditions is flexible, and those of higher-order shape characteristics may need to be included, depending on the shape characteristics one decides to fit, and the number of parameters to be estimated. We assume that

$$\mathbb{E}[g(v_{l\Delta};\theta_0)]=0$$

and $\mathbb{E}[g(v_{l\Delta};\theta)]\neq 0$ for $\theta\neq\theta_0$ holds. We also assume that θ_0 is in the interior of Θ. As the moments are given by coefficients of an expansion (up to constants), a bias term of small order is left, an effect similar to that in Aït-Sahalia and Mykland (2003). We treat this term as negligible on the basis of fitting each IV surface near its at-the-money and short maturity point, and verify in simulations in Section 4 that this is indeed the case in practice.

As in the construction above of the nonparametric ISV model, the data for the shape characteristics $\Sigma_{i,j}$ are obtained by the polynomial regression in Equation (18) of IV on time-to-maturity and log-moneyness. The rest of the procedure is standard GMM (see Hansen 1982): to estimate the parameters θ, we construct the sample analog of $\mathbb{E}[g(v_{l\Delta};\theta)]$ as

$$g_n(\theta)\equiv\frac{1}{n}\sum_{l=1}^{n}g(v_{l\Delta};\theta).$$

The estimator $\hat{\theta}$ is defined as the solution of the quadratic minimization problem

$$\hat{\theta}=\operatorname*{argmin}_{\theta} g_n(\theta)^{\intercal}W_n g_n(\theta), \tag{28}$$

where W_n is a positive definite weight matrix. If the number of moment conditions is equal to that of parameters to estimate, that is, the model is exactly identified, the estimator $\hat{\theta}$ is the solution of the (system of) equations

$$g_n(\hat{\theta})=0,$$

and the choice of W_n does not matter. Otherwise—that is, if the number of moment conditions is greater than that of parameters to estimate—the model is overidentified and the optimal choice of the weight matrix W_n follows from a standard two-step estimation. For both of the exactly identified and overidentified cases, the asymptotic behavior of $\hat{\theta}$ is given by

$$\hat{\theta}\xrightarrow{P}\theta_0 \text{ and } \sqrt{n}(\hat{\theta}-\theta_0)\xrightarrow{d}\mathcal{N}(0,V(\theta_0)), \text{ as } n\to\infty, \tag{29}$$

where the asymptotic variance matrix $V(\theta_0)$ can be consistently estimated by an estimator $\hat{V}(\hat{\theta})$, which is given along with further technical details in Appendix C.

We provide in Section 4.1 an example showing how to estimate a parametric ISV model, and the results of Monte Carlo simulations where the model is either exactly identified or overidentified. We find that for each parameter, the bias of the estimator is less than the corresponding finite-sample standard deviation and that the estimator $\sqrt{\hat{V}(\hat{\theta})/n}$ of the asymptotic standard deviations provides a reliable way of approximating standard errors for the parameters.

We have illustrated our method for constructing/estimating nonparametric and parametric ISV models in a bivariate setting. At least in principle, constructing/estimating models with more factors, such as, for example, the two-factor parametric SV models proposed in Duffie, Pan, and Singleton (2000) and Christoffersen, Heston, and Jacobs (2009) as well as their nonparametric generalizations, is possible. It requires closed-form formulae for high-order shape characteristics in order to estimate the coefficient functions (resp. parameters) in the nonparametric (resp. parametric) model. And the data of the corresponding high-order shape characteristics ought to be computed by incorporating additional terms in the bivariate regression in Equation (18). Nevertheless, the benefits from constructing trivariate models or even beyond appear to be limited, in terms of fitting the practically useful shape characteristics. This is because, as we will show in Section 5, a bivariate nonparametric model is already flexible enough to reproduce all the shape characteristics up to second order, that is, all $\Sigma_{i,j}$ with $i+j \leq 2$. In the case of parametric models, however, there are likely gains to be achieved from including (one or more) additional state variable(s).

4. Monte Carlo Simulation Results

4.1 A parametric implied stochastic volatility model

Our parametric method provides a new approach to estimate a model with a given parametrization. We illustrate the method with the SV model of Heston (1993), because its numerical tractability for generating option prices makes it easy to compute accurate prices that serve as the input data for the simulations. Recall that such tractability is not needed in real applications of our method: we take the observed prices as data and do not require any model-based computation of option prices.

Under the assumed risk-neutral measure, the underlying asset price S_t and its spot variance $V_t = v_t^2$ follow

$$\frac{dS_t}{S_t} = (r-d)dt + \sqrt{V_t}dW_{1t}, \tag{30}$$

$$dV_t = \kappa(\alpha - V_t)dt + \xi\sqrt{V_t}[\rho dW_{1t} + \sqrt{1-\rho^2}dW_{2t}], \tag{31}$$

where W_{1t} and W_{2t} are independent standard Brownian motions. Here, the parameter vector is $\theta = (\kappa, \alpha, \xi, \rho)$, and we assume that Feller's condition holds: $2\kappa\alpha > \xi^2$. The leverage effect parameter is $\rho \in [-1, 1]$. To estimate the

four parameters in $\theta=(\kappa,\alpha,\xi,\rho)$, we successively employ the four moment conditions in $g=(g^{(1,0)},g^{(0,1)},g^{(0,2)},g^{(1,1)})^\intercal$ to exactly identify the parameters or employ the five moment conditions in $g=(g^{(1,0)},g^{(0,1)},g^{(0,2)},g^{(1,1)},g^{(2,0)})^\intercal$ to overidentify the parameters. We impose $\alpha>0$, $\kappa>0$, $\xi>0$, and Feller's condition as constraints during the GMM minimization in Equation (28).

Itô's lemma applied to $v_t=\sqrt{V_t}$ yields

$$\mu(v)=\frac{\kappa(\alpha-v^2)}{2v}-\frac{\xi^2}{8v},\ \gamma(v)=\frac{\xi\rho}{2},\ \eta(v)=\frac{\xi\sqrt{1-\rho^2}}{2}. \tag{32}$$

Then, applying the results of Section 1.1 and the general method for deriving higher orders in Appendix A, we can calculate the shape characteristics $\Sigma_{0,0}(v)$, $\Sigma_{0,1}(v)$, $\Sigma_{0,2}(v)$, $\Sigma_{1,0}(v)$, $\Sigma_{1,1}(v)$, and $\Sigma_{2,0}(v)$:

$$\Sigma_{0,0}(v)=v,\ \Sigma_{0,1}(v)=\frac{\rho\xi}{4v},\ \Sigma_{0,2}(v)=-\frac{1}{24v^3}\left(5\rho^2-2\right)\xi^2, \tag{33}$$

and

$$\begin{aligned}
\Sigma_{1,0}(v)=&\frac{1}{96v}\left(\xi\left(24\rho(d-r)+\xi\left(\rho^2-4\right)\right)+v^2(12\xi\rho-24\kappa)+24\kappa\alpha\right),\\
\Sigma_{1,1}(v)=&-\frac{\xi}{384v^3}\left(16\left(2-5\rho^2\right)(r-d)\xi+\rho\left(40\kappa\alpha+3\left(3\rho^2-4\right)\xi^2\right.\right.\\
&\left.\left.+v^2(4\rho\xi-8\kappa)\right)\right),\\
\Sigma_{2,0}(v)=&\frac{1}{15360v^3}\left[\xi^2\left(-640(r^2+d^2)\left(5\rho^2-2\right)+80d\left(3\rho\left(4-3\rho^2\right)\xi\right.\right.\right.\\
&+16\left(5\rho^2-2\right)r\left.\right)+\left(59\rho^4-88\rho^2-16\right)\xi^2+240\rho\left(3\rho^2-4\right)r\xi\left.\right)\\
&+320v^4\left(5\kappa^2-5\kappa\rho\xi+\left(2\rho^2-1\right)\xi^2\right)-80\kappa\alpha\xi(40d\rho-40\rho r\\
&+\left(5\rho^2-8\right)\xi\left.\right)-40v^2(2\kappa-\rho\xi)\left(\xi\left(-8d\rho+3\rho^2\xi+8\rho r\right.\right.\\
&\left.\left.-4\xi\right)+8\kappa\alpha\right)-960\kappa^2\alpha^2\left.\right].
\end{aligned}$$

We start with a set of simulations designed to approximate the setting in which we will employ the method in real data. For each simulation trial, we generate a time series of the state variables (S_t,V_t) containing $n=1{,}000$ consecutive values at the daily frequency, that is, with time increment $\Delta=1/252$, by subsampling higher-frequency data simulated using the Euler scheme. The parameter values are $r=0.03$, $d=0$, $\kappa=3$, $\alpha=0.04$, $\xi=0.2$, and $\rho=-0.7$. Each day, we calculate option prices with time-to-maturity τ equal to 5, 10, 15, 20, 25, and 30 days, and for each time-to-maturity τ, we include 20 log-moneyness values k within $\pm v_t\sqrt{\tau}$, where τ is annualized and v_t is the spot volatility. The rationale for choosing such a region of (τ,k) for simulations is discussed in the next paragraph. Due to the affine nature of the model of Heston (1993),

these option prices can be calculated by Fourier transform inversion, and the corresponding IV values can be consequently computed by numerical root-finding. To mimic a realistic market scenario, we add observation errors to these implied volatilities, sampled from a normal distribution with mean zero and constant standard deviation equal to 15 basis points (bps) and further assumed to be uncorrelated across time-to-maturity and log-moneyness, as well as over time. Then, for each IV surface, we follow the regression procedure described at the beginning of Section 2 to extract the estimated coefficients $\hat{\beta}_l^{(i,j)}$ of the bivariate regression in Equation (18). As a result, for each simulation trial, we obtain $n=1,000$ vectors of shape characteristics data (up to constants according to Equation (19)) for the subsequent GMM estimation, that is, $(\hat{\beta}_l^{(i,j)})_{(i,j)\in I}$ for $l=1,2,\ldots,n$.

In practice, one needs to choose the orders J, L_0, L_1, …, L_J in the bivariate polynomial regression in Equation (18) and the region in (τ,k) of the IV surface data to compute the regression. On the one hand, we need at a minimum to include enough orders in the regression to estimate the coefficients of interest for the estimation method; recall that we need the shape characteristics $\Sigma_{0,0}$, $\Sigma_{0,1}$, $\Sigma_{0,2}$, $\Sigma_{1,0}$, and $\Sigma_{1,1}$ for estimating an exactly identified ISV model, and we need to include an additional shape characteristic $\Sigma_{2,0}$ for estimating an overidentified one. But we can consistently estimate all these lower-order coefficients from a higher-order regression, discarding the estimates of the higher-order coefficients. On the other hand, the orders cannot be chosen as too high, and the region in (τ,k) cannot be chosen as too narrow, to avoid overfitting the regression. Specifically, we set the order to be $(J,\mathbf{L}(J))=(2,(2,2,1))$, so:

$$\begin{aligned}&\Sigma^{\text{data}}(\tau_l^{(m)},k_l^{(m)})\\&=\beta_l^{(0,0)}+\beta_l^{(1,0)}\tau_l^{(m)}+\beta_l^{(2,0)}(\tau_l^{(m)})^2+\beta_l^{(0,1)}k_l^{(m)}+\beta_l^{(1,1)}\tau_l^{(m)}k_l^{(m)}\\&\quad+\beta_l^{(2,1)}(\tau_l^{(m)})^2k_l^{(m)}+\beta_l^{(0,2)}(k_l^{(m)})^2+\beta_l^{(1,2)}\tau_l^{(m)}(k_l^{(m)})^2+\epsilon_l^{(m)},\end{aligned}\tag{34}$$

for $m=1,2,\ldots,n_l$. The estimated coefficients from this regression estimate the IV surface characteristics that we need (recall Equation (19)). Figure 2 plots the contours of approximation errors involved in the expansion $\Sigma^{(2,(2,2,1))}(\tau,k,v_t)$, as a function of (τ,k). The figure shows that the approximation errors in the region where we employ the expansion is less than 3 bps (in IV units), which is dominated by the typical level of noise, that is, 15 bps. This suggests that the estimated coefficients $\hat{\beta}_l^{(i,j)}$ are approximately unbiased.

We then implement the method proposed in Section 3 to estimate the model parametrically. We consider two cases. The first one is exactly identified using $g=(g^{(1,0)},g^{(0,1)},g^{(0,2)},g^{(1,1)})^\intercal$, while the second adds one more moment condition for $g^{(2,0)}$, to overidentify the parameters. Table 1 summarizes the results, based on $M=500$ simulation replications. We find that, for each parameter, the absolute bias is relatively small and is less than the corresponding finite-sample standard deviation. In the exactly identified (resp. overidentified)

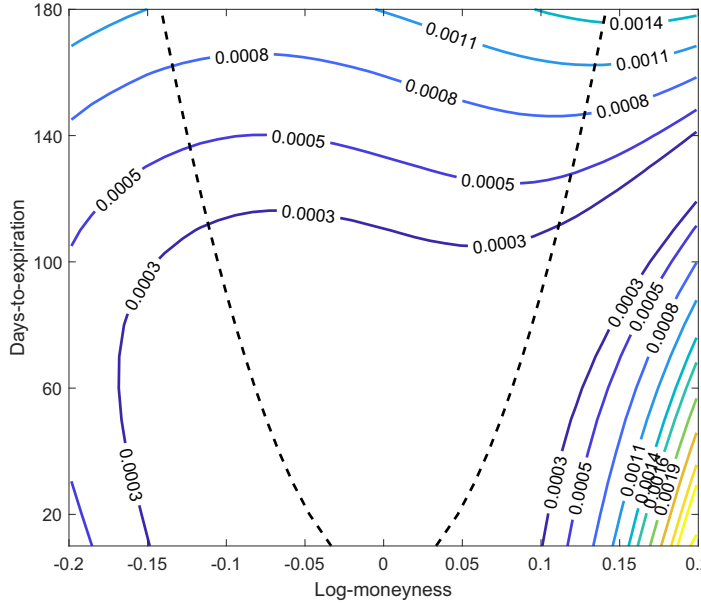

Figure 2
Contour plot of the approximation errors for the expansion
This plot reports the approximation error of the IV expansion $\Sigma^{(2,(2,2,1))}(\tau, k, v_t)$ as a function of log-moneyness k and time-to-maturity τ under the Heston model in Equations (30)–(31) with parameters $r = 0.03$, $d = 0$, $\kappa = 3$, $\alpha = 0.04$, $\xi = 0.2$, $\rho = -0.7$, and $v_t = 0.2$. Units on the plot are the same as those of IV, so, for example, 0.2 represents an annualized IV of 20% per year. The two dashed curves are the parabolas $k = v_t\sqrt{\tau}$ and $k = -v_t\sqrt{\tau}$, respectively, delineating the region in (τ, k) where we employ the expansion. The approximation error in the region where we employ the expansion is less than 3 bps in IV units, compared with a typical level of noise in the data of about 15 bps.

case, we compare for each parameter the finite-sample standard deviation exhibited in the fourth (resp. sixth) column of Table 1 with the consistent estimator of its asymptotic counterpart. Figure 3 compares the finite-sample standard deviation for each parameter with the distribution of sample-based asymptotic counterpart in the exactly identified case. Consider the upper left panel of Figure 3 as an example. The histogram characterizes the distribution of sample-based asymptotic standard deviation $\sqrt{\hat{V}_{11}(\hat{\theta})/n}$ for parameter κ, where $\hat{V}_{11}$ represents the $(1,1)$th entry of the estimated variance matrix $\hat{V}$. The red star marks the corresponding finite-sample standard deviation shown in the fourth cell from the first row of Table 1. As shown in Figure 3, for each parameter, the finite-sample standard deviation falls within the range of its sample-based asymptotic counterpart, suggesting that the sample-based approximation $\sqrt{\hat{V}(\hat{\theta})/n}$ of the asymptotic standard deviations is a reasonable estimator of the standard errors. The corresponding results for the overidentified case are in Figure S.1 of the Online Appendix.

Table 1
Parametric ISV model: Monte Carlo simulations

		Exact identification		Overidentification	
Parameter	True	Bias	Std. dev.	Bias	Std. dev.
κ	3.00	-0.031	0.554	-0.259	0.488
α	0.04	3.21×10^{-4}	0.0022	0.0012	0.0029
ξ	0.20	0.0021	0.0109	3.53×10^{-4}	0.0106
ρ	-0.70	0.0058	0.0374	0.0017	0.0374

The model is given in Equations (30)–(31). In the fourth and sixth columns, the standard deviation of each parameter is calculated by the finite-sample standard deviation of estimators based on $M = 500$ simulation trials.

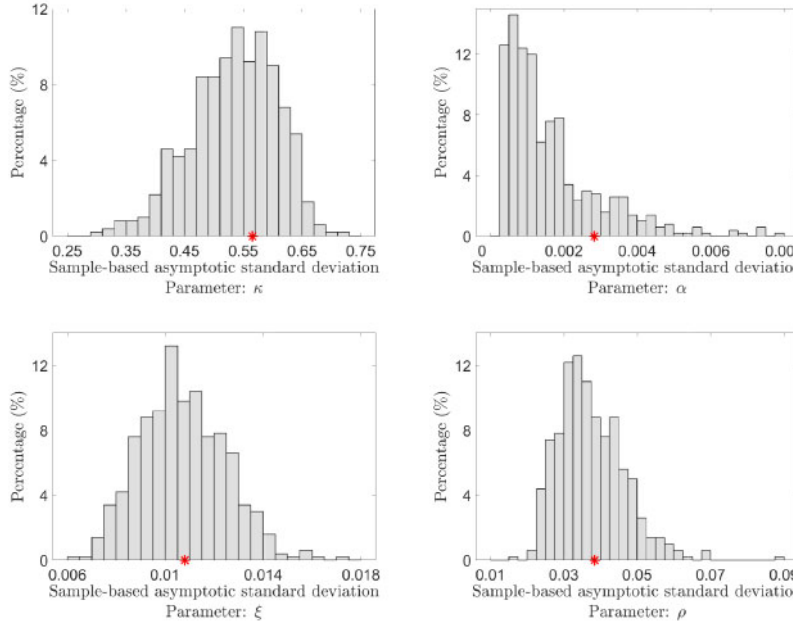

Figure 3
Parametric ISV model: Distribution of the sample standard deviations of the estimators in the exactly identified GMM case
The model is given in Equations (30)–(31). In the third and fifth columns of each panel, the standard error of each parameter is obtained from the estimator in Equations (29) and (C1). For instance, the standard error of the parameter κ is given by $\sqrt{\hat{V}_{11}(\hat{\theta})/n}$, where $\hat{V}_{11}$ represents the $(1,1)$th entry of the matrix $\hat{V}$.

The simulations we just described are designed to approximate the setting in which we will employ the method in real data, and they revealed that the method works well in that setting. We now examine how the method performs as the data availability and quality deteriorate, namely as the data becomes more illiquid or sparse, for example, with shorter time series, fewer strikes and time-to-maturities, and/or with more noise in the observations. The purpose is to gain an understanding of the limitations of the method. Specifically, we

consider the exactly identified case and make matters worse in the following four dimensions: decreasing the number of days available in the sample, that is, the sample size n, decreasing the number of different time-to-maturities available on each day, decreasing the number of strikes available for each time-to-maturity, and increasing the level of noise for IV observations. The benchmark setting is similar to the case of Table 1 consisting of $n=1,000$ daily IV surfaces for each simulation trial; on each surface, we observe six time-to-maturities: 10, 20, 30, 40, 50, and 60 days; for each time-to-maturity, we observe 20 equidistant strikes ranging from $-v_t\sqrt{\tau}$ to $v_t\sqrt{\tau}$. Observation errors are i.i.d. sampled from a normal distribution with mean zero and standard deviation equal to 15 bps.

We depart from this benchmark setting by successively halving n from 1,000 (the benchmark) to 500, 250, 125, and 63; reducing the number of time-to-maturities on each IV surface from 6 (the benchmark) to 5, 4, and 3; reducing the number of strikes for each maturity from 20 (the benchmark) to 16, 12, 8, and 4; and finally increasing the level of noise from 15 bps to 30, 45, 60, and 75 bps. As we vary each relevant dimension, we hold the settings of the other three dimensions fixed at the corresponding benchmark level, and repeat $M=500$ simulations. The true parameter values are the same as in the benchmark case.

The results are in Figure 4. For each parameter, we measure the estimation performance using the relative root mean squared error (RRMSE): separately for each parameter ϑ in $\theta=(\kappa,\alpha,\xi,\rho)$, the RRMSE for that parameter is defined by $\mathrm{RMSE}(\hat{\vartheta})/|\vartheta_0|$, where ϑ_0 (resp. $\hat{\vartheta}$) represents the true value (resp. estimator) of ϑ. Figure 4 shows in graphical form how the performance deteriorates as the data setting gets more challenging. As expected, the performance behaves monotonically, starting from the benchmark scenario. The worst result is consistently obtained for the speed of volatility mean reversion κ, with an RRMSE greater than 0.2. Given the (realistic) low assumed value of κ_0, this is indeed the hardest parameter to estimate, particularly on the basis of short time series. Overall, for a desired level of parameter accuracy, the figure can be used to determine up to what limit the method can be employed.

Results (not shown to save space) are fairly similar in the overidentified case, with the following differences. The RRMSE profile is flatter than in the corresponding panels of Figure 4, so the performance in the overidentified case is more robust than in the exactly identified case as the quantity and/or quality of data deteriorate. For κ, however, although more robust to data deterioration, overidentification leads to larger RRMSE overall than exact identification, as noted in Table 1. It is hard to tell whether this is a generic result, due to the specific additional moment condition or the model/parameter values, but it is not an unheard-of situation in GMM estimation.

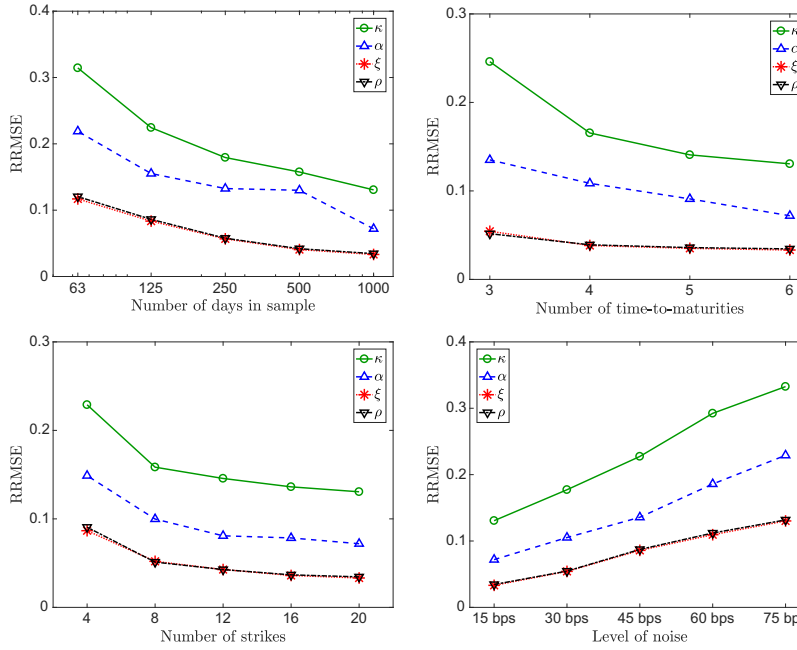

Figure 4
Stress test of the method
This figure examines in Monte Carlo simulations the performance of the parametric ISV model estimation when the estimation conditions get progressively worse, as the number of days in sample decreases (upper left panel), the number of time-to-maturities available decreases (upper right), the number of option strikes decreases (lower left), and the amount of noise in the data increases (lower right). The performance is measured for each of the four parameters in $\theta=(\kappa,\alpha,\xi,\rho)$ separately, as that parameter's root mean squared error divided by the true value of the parameter. The benchmark setting is similar to that of Table 1, where the number of days in the sample is 1,000, the number of time-to-maturities each day is 6, the number of strikes for each time-to-maturity is 20, and the level of noise is 15 bps. The benchmark setting is at the extremity of each plot: in the upper left, upper right, and lower left panels, it corresponds to the points on the right of the plot, whereas in the lower right panel, it corresponds to the points on the left of the plot.

4.2 Nonparametric implied stochastic volatility model

Next, we apply the nonparametric method of Section 2 to the simulated data that was generated under the Heston model. Through the following Monte Carlo experiments, we validate the nonparametric construction, showing that for each coefficient function, its nonparametric estimator is close to the corresponding true function. Moreover, with an eye toward the empirical implementation in Section 5, we validate a bootstrap procedure for calculating the small sample standard deviation of the nonparametric estimator of each coefficient function.

In Figure 5, the upper left, upper right, middle left, and middle right panels, respectively, plot the nonparametric estimators of the functions μ, $-\gamma$, η^2, and η for the model given in Equations (1)–(2). Consider the upper left panel for the function μ. We perform local polynomial regression at equidistantly distributed values of v in the interval $[0.1,0.3]$. For each $v\in[0.1,0.3]$, we mark the true value of $\mu(v)$ by a blue dot, according to Equation (32), and plot the mean

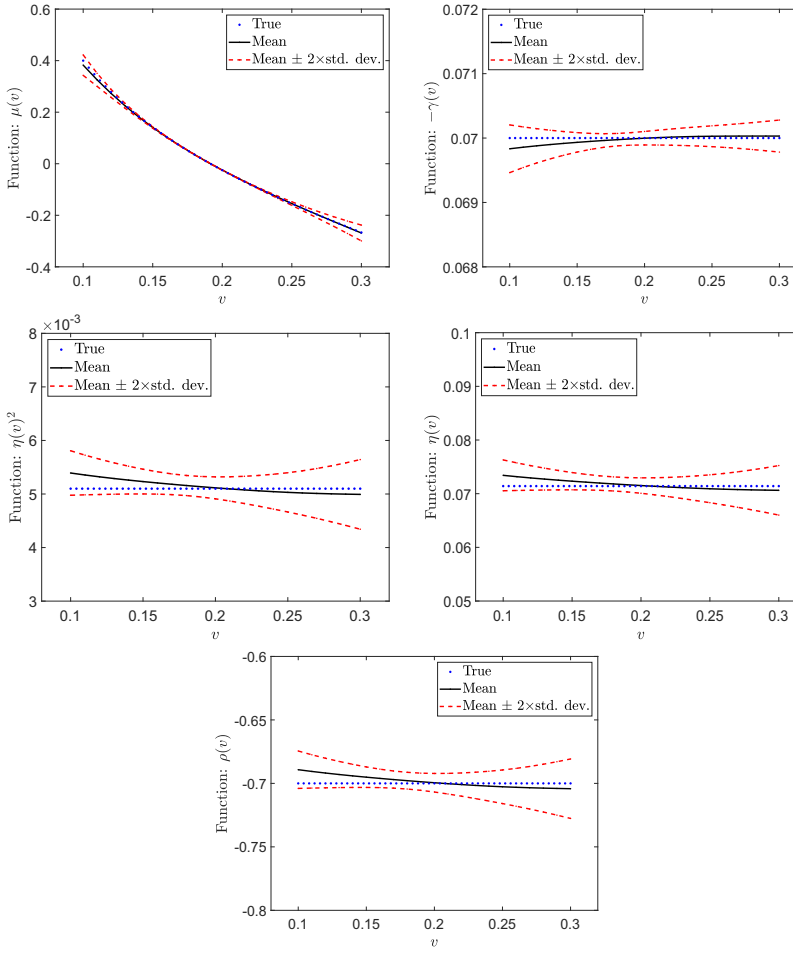

Figure 5
Nonparametric ISV model: Monte Carlo simulations
In each panel, the true function is determined by Equation (32). The black solid curve represents the mean of nonparametric estimators from the 500 simulation trials. Each point on the upper (resp. lower) red dashed curve is plotted by vertically upward (resp. downward) shifting the corresponding one on the black mean curve by a distance equal to twice the corresponding finite-sample standard deviation.

of estimators of $\mu(v)$ on a black solid curve. We then compute a pointwise confidence interval: each point on the upper (resp. lower) dashed curve is obtained by shifting vertically upward (resp. downward) the corresponding one on the mean curve by a distance equal to twice the corresponding finite-sample standard deviation. As seen from the figure, the shape of the estimated nonparametric function matches that of the true one. The results reveal that, at each level of v, the nonparametric estimator is sufficiently close to the true value relative to the pointwise confidence interval.

We then combine the estimators $\hat{\gamma}(\cdot)$ and $\hat{\eta}^2(\cdot)$ to estimate the leverage effect $\rho(v_t)$ under the nonparametric ISV model in Equations (1)–(2) by

$$\hat{\rho}(v_t) = \frac{\hat{\gamma}(v_t)}{\sqrt{\hat{\gamma}(v_t)^2 + \hat{\eta}(v_t)^2}}. \tag{35}$$

As in the other four panels of Figure 5, we exhibit the estimation results for $\rho(v)$ in the lowest panel. We find that the shape of the estimated function $\rho(v)$ is approximately constant at the level of parameter ρ, as it ought to be under the model of Heston (1993).

We propose in what follows a bootstrap estimator of the (pointwise) standard errors. It is based on multiple bootstrap replications out of a simulation trial, in order to mimic a realistic estimation scenario in real data. In each bootstrap replication, we reproduce an IV surface each day. The surface contains the same number of implied volatilities as that of the original surface, and the implied volatilities on the reproduced surface are sampled as i.i.d. replications of the volatilities on the original surface. Based on the bootstrap "data," we apply the same estimation method proposed in Section 2 to obtain the bootstrap estimators of functions $\mu(\cdot)$, $-\gamma(\cdot)$, $\eta^2(\cdot)$, $\eta(\cdot)$, and $\rho(\cdot)$. The bootstrap standard error of each function is accordingly calculated as the standard deviations of its multiple bootstrap estimators.

To validate this method, which we will employ below in real data, we randomly select one simulation trial and calculate the bootstrap standard error of each coefficient function out of 500 corresponding bootstrap estimators. Figure 6 summarizes the estimation result of this trial. In each panel of Figure 6, the blue dotted (resp. black solid) curve represents the true function (resp. nonparametric estimator). Each point on the upper (resp. lower) dashed curve is a pointwise band and is plotted by shifting vertically upward (resp. downward) the corresponding one on the black solid curve by a distance equal to twice the corresponding bootstrap standard error. Figure 6 reveals that the nonparametric estimators are all close to the corresponding true functions. More importantly, the bootstrap method seems to provide a reliable way for calculating standard deviations, as seen from comparing each panel in Figure 6 with the corresponding one in Figure 5. Compare the upper right panels of Figures 6 and 5 as an example: for any v, the lengths of intervals bounded by the two dashed curves in these two panels are close to each other. In light of the simulations evidence, we will employ bootstrap standard errors in the coming empirical analysis.

5. Empirical Results

We now turn to real data. We employ S&P 500 index options data covering the period from January 3, 2007, to December 29, 2017, obtained from OptionMetrics. Panel A of Table 2 reports the basic descriptive statistics of the raw sample of 1,411,612 observations, including both call and put options. This

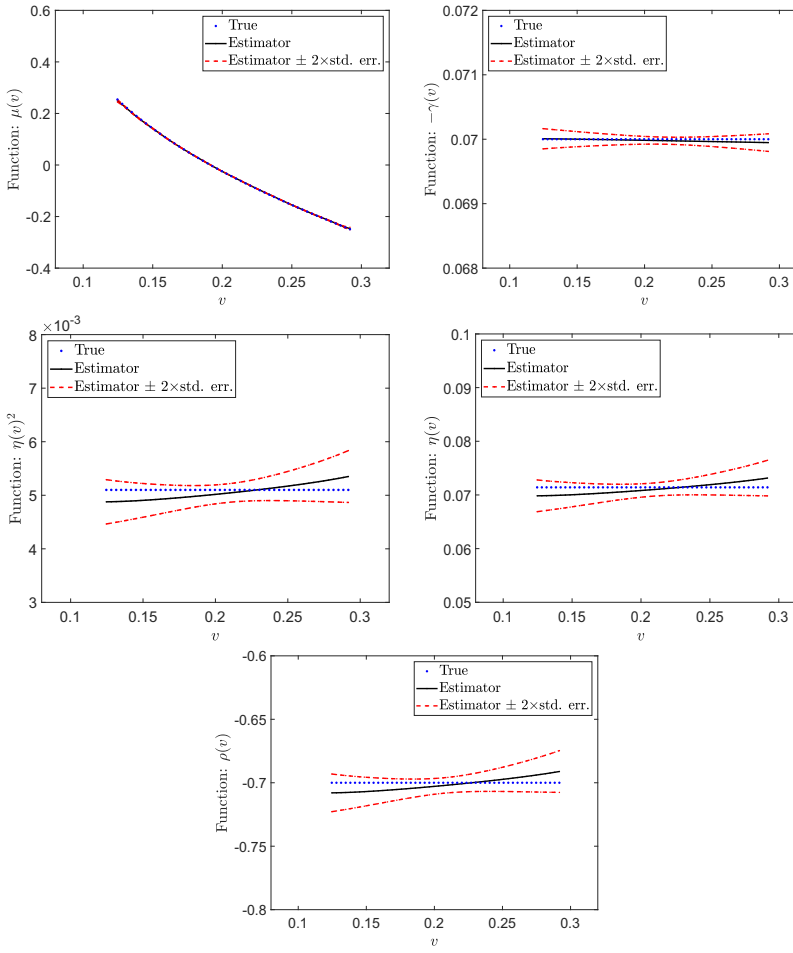

Figure 6
Nonparametric ISV model: Single sample bootstrap standard errors
This plot shows the results of employing the bootstrap method described in Section 4.2 to compute standard errors on a single sample. In each panel, the true function is determined or calculated according to Equation (32). The black solid curve represents the one-trial nonparametric estimator. Each point on the upper (resp. lower) red dashed curve is plotted from vertically upward (resp. downward) shifting the corresponding one on the black curve by a distance equal to twice of the corresponding standard error.

panel divides the data into three (calendar) days-to-expiration categories and six log-moneyness categories. For each category, we report the total number, mean, and standard deviation of implied volatilities therein. A comparison of the number of implied volatilities in each crossed category suggests that the most actively traded options are around the at-the-money region, especially for time-to-maturity between 10 and 60 days. This observation is also supported by the evidence from daily average trading volumes: Figure 7 plots a heat map

Table 2
Descriptive statistics for the S&P 500 index IV data, 2007–2017

	Number			Mean			Standard deviation		
Days-to-expiration	<10	$[10,60]$	>60	<10	$[10,60]$	>60	<10	$[10,60]$	>60
Panel A: Sample of raw options data									
Log-moneyness k									
$k<-5\%$	960	111,933	184,763	50.01	24.12	22.84	15.64	10.10	6.91
$-5\%\le k\le -2.5\%$	7,943	146,237	56,514	29.45	16.92	17.93	12.90	5.85	5.38
$-2.5\%\le k<0$	50,458	224,988	85,795	17.13	13.74	16.75	9.04	5.65	5.54
$0\le k<2.5\%$	37,785	204,737	76,109	15.39	12.01	15.61	9.14	5.85	5.88
$2.5\%\le k<5\%$	2,433	67,391	52,048	31.56	14.66	14.66	13.99	7.18	6.21
$k\ge 5\%$	345	20,789	80,384	53.11	25.98	17.86	17.98	12.21	7.23
Total	99,924	776,075	535,613	18.24	15.79	18.78	11.23	8.13	7.11
Panel B: Sample of filtered options data									
Log-moneyness k									
$k<-5\%$	62	19,101	7,414	61.58	32.72	31.13	13.24	12.92	10.12
$-5\%\le k\le -2.5\%$	437	60,487	5,603	35.70	20.25	24.06	12.14	7.16	7.41
$-2.5\%\le k<0$	1,749	205,395	9,031	22.73	14.00	22.59	9.63	5.81	7.34
$0\le k<2.5\%$	1,753	199,743	8,441	21.18	12.11	21.55	10.50	5.88	7.45
$2.5\%\le k<5\%$	467	60,421	5,966	29.93	15.23	20.60	12.10	7.29	7.85
$k\ge 5\%$	82	18,882	10,127	54.32	26.16	23.37	15.74	12.25	8.73
Total	4,550	564,029	46,582	25.22	15.17	23.85	12.99	8.21	8.89

The sample consists of daily implied volatilities of European options written on the S&P 500 index covering the period January 3, 2007–December 29, 2017. The columns "Mean" and "Standard deviation" are reported as percentages. The log-moneyness is $k=\log(K/S_t)$, where K is the exercise strike of the option and S_t is the spot price of the S&P 500 index. The data filters employed to go from panel A to B are described in the text.

of the trading volume of the options as a function of their time-to-maturity τ and log-moneyness k. The figure shows that the most actively traded options are in the short-maturity near at-the-money region, which is the data we will select when applying our method.

We select options from the raw sample using the following data filters. We select options with time-to-maturity between 10 and 60 calendar days, thereby excluding both extremely short-maturity ones, which are subject to significant trading effects and biases, and long-maturity ones for which the IV expansion becomes less accurate. As suggested by the evidence in the upper right panel of Figure 4, we include at least four time-to-maturities between 10 and 60 days. This requirement is satisfied by the vast majority of the data sample after 2012, which is when short-maturity options, in particular weekly options, become more actively traded. Before 2012, if there exists at least one time-to-maturity, but fewer than the four we seek between 10 and 60 days, we select the four time-to-maturities from an extended range of 8 to 160 days; if four do not exist in that range, we exclude that day from the sample. In the re-selection procedure, if the number of available time-to-maturity in the range of $[8,160]$ days exceeds four, we select the four closest to the range of $[10,60]$ days. Over the 11-year sample, only 26 days (24 in 2007 and 2 in 2008) are discarded as a result. For each time-to-maturity, we select option with log-moneyness within $\pm v_t\sqrt{\tau}$ where τ is the annualized time-to-maturity and v_t is the instantaneous volatility proxied by the observed IV with the shortest time-to-maturity longer than 10 days and log-moneyness k closest to 0 on that day. Panel B of Table 2

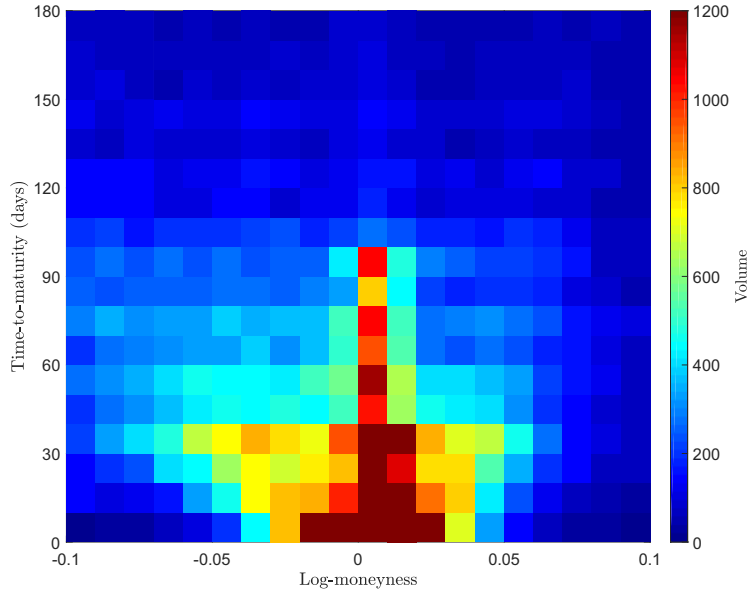

Figure 7
Liquidity of options, 2007–2017
This heat map shows the daily average trading volume of S&P 500 index options over 2007–2017 as a function of each option's time-to-maturity τ and log-moneyness k. For clarity, volumes greater than $1,200$ are truncated at $1,200$.

reports the basic descriptive statistics of the resulting sample after we apply these data filters.

Since our approach is based on a regression of the IV data, one approach is to include IV inferred from both calls and puts and rely on the regression to directly smooth out any IV discrepancies between the two; another is to employ strict filters to exclude from the regression options that violate arbitrage relations such as put-call parity, but this requires making a potentially arbitrary choice concerning which is the right data point. A further possibility is to start from the more liquid out-of-the-money option and replace the price of the corresponding in-the-money one by its value implied by put-call parity. The approach we employ is a mix of the two: use the filters described earlier to exclude options that are likely to be illiquid and hence noisy, but otherwise use both calls and puts in the IV surface regression and let the regression do the smoothing out of the errors.

To choose the order of the polynomial regression in Equation (18), a reasonable compromise is to set $(J, \mathbf{L}(J)) = (2, (2, 2, 0))$, that is,

$$\Sigma^{\text{data}}(\tau_l^{(m)}, k_l^{(m)})$$

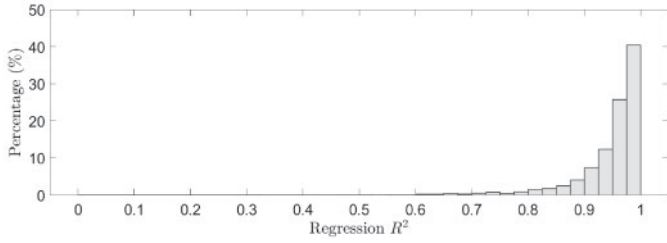

Figure 8
Histogram of R^2 for the polynomial IV surface regression, daily 2007–2017
Each day in the sample we run the regression in Equation (36) to estimate that day's IV surface shape characteristics. The time series of these shape characteristics serve as the input to estimate the ISV model, either parametrically or nonparametrically. This plot shows the empirical distribution of the regression's R^2 over the sample period 2007–2017. Each day's IV surface regression estimates seven parameters with an average of 225 options.

$$= \beta_l^{(0,0)} + \beta_l^{(1,0)}\tau_l^{(m)} + \beta_l^{(2,0)}(\tau_l^{(m)})^2 + \beta_l^{(0,1)}k_l^{(m)} + \beta_l^{(1,1)}\tau_l^{(m)}k_l^{(m)} + \beta_l^{(2,1)}(\tau_l^{(m)})^2 k_l^{(m)} + \beta_l^{(0,2)}(k_l^{(m)})^2 + \epsilon_l^{(m)}. \quad (36)$$

Comparing with the bivariate regression in Equation (34) employed in the Monte Carlo experiments, we remove a high-order regression coefficient $\beta_l^{(1,2)}$ to reduce the standard errors of the estimators of other low-order coefficients without loss of accuracy.

Figure 8 plots a histogram of the R^2 values achieved each day by the parametric regressions in Equation (36) across the full sample of IV surfaces. To keep these R^2 values in perspective, each day, we estimate seven parameters with an average of 225 options. We find that for over 95% of the sample the R^2 are greater than 0.8, and essentially none are lower than 0.7, suggesting that Equation (36) is quite successful at fitting the data. Practitioners often use polynomial regression to fit the short-maturity near at-the-money region of the IV surfaces in their own internal models,[8] so it is not surprising that the market data we collect end up reflecting this feature. As an example, Figure 9 plots the IV data and the corresponding fitted surface produced by the regression in Equation (36) on a randomly selected day in our sample (January 3, 2017).

5.1 Parametric implied stochastic volatility model

We now implement the method described in Section 3 to estimate a parametric ISV model of the Heston (1993) type for illustration, and compare our estimation results with those in the literature to understand, in particular, any differences. Table 3 reports the GMM results for both the exactly identified and overidentified cases. First, the estimators of ρ are around -0.6 in both cases, consistent with the alternative estimators in the literature and also with what can

[8] See, for example, Gatheral (2006).

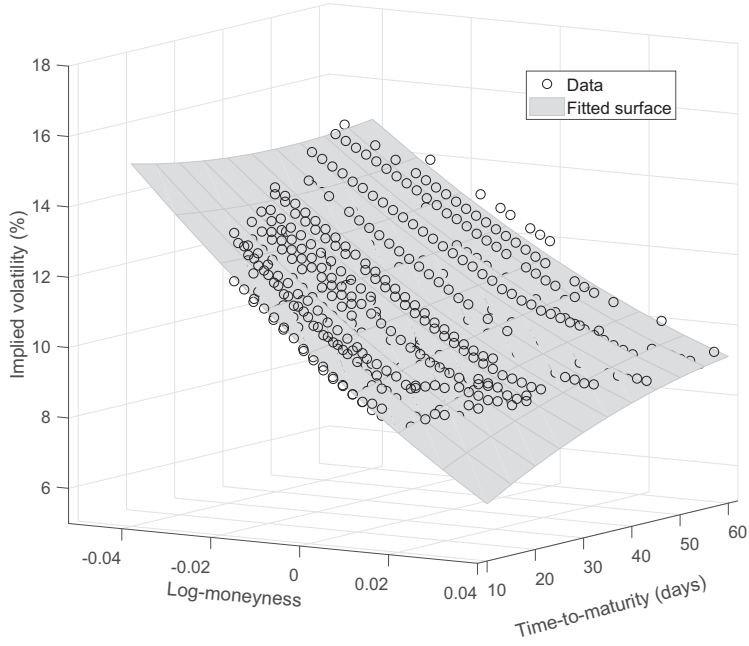

Figure 9
IV data on January 3, 2017, and the corresponding parametric fitted surface
The parametric fitted surface is estimated by the bivariate regression in Equation (36). The regression R^2 is 0.94.

be heuristically inferred directly from the level $[\Sigma_{0,0}]^{\text{data}}$ and log-moneyness convexity $[\Sigma_{0,2}]^{\text{data}}$, depicted by the corresponding histograms in Figure 10. The mean and standard deviation of the multiplicative data $([\Sigma_{0,0}]^{\text{data}})^3[\Sigma_{0,2}]^{\text{data}}$ are 0.0083 and 0.0284, respectively. Thus, there is no evidence for the mean of $([\Sigma_{0,0}]^{\text{data}})^3[\Sigma_{0,2}]^{\text{data}}$ to be significantly different from zero. On the other hand, it follows from the closed-form formulae for $\Sigma_{0,0}(v)$ and $\Sigma_{0,2}(v)$ given in Equation (33) that

$$\Sigma_{0,0}(v)^3\Sigma_{0,2}(v) = -\frac{1}{24}\left(5\rho^2 - 2\right)\xi^2.$$

Heuristically, moment matching by equating the estimated zero mean requires $-\left(5\rho^2-2\right)\xi^2/48=0$. This would approximately estimate ρ as -0.63, independently of the values of v and ξ.

Second, the estimator of ξ is around 1 (resp. 0.9) in the exactly identified (resp. overidentified) case. We find that, for both cases, the estimators of ξ are somewhat greater than those in the literature, which are usually less than 0.55 (see, e.g., Pan 2002; Aït-Sahalia and Kimmel 2007; and Christoffersen, Jacobs, and Mimouni 2010 among others for estimations by combining the data of options and time-series dynamics of the underlying asset price under the physical measure). As pointed out in, for example, Bates (2000),

Table 3
Parametric ISV model: Estimates

	Exact identification		Overidentification	
Parameter	Estimator	Standard error	Estimator	Standard error
κ	10.5	0.64	10.0	0.65
α	0.039	0.0019	0.037	0.0018
ξ	1.10	0.029	0.86	0.017
ρ	−0.615	0.0013	−0.603	0.0019

The model is given in Equations (30)–(31). In the third and fifth columns, the standard error of each parameter is obtained from the estimator in Equations (29) and (C1). For instance, the standard error of the parameter κ is given by $\sqrt{\hat{V}_{11}(\hat{\theta})/n}$, where $\hat{V}_{11}$ represents the $(1,1)$th entry of the matrix $\hat{V}$.

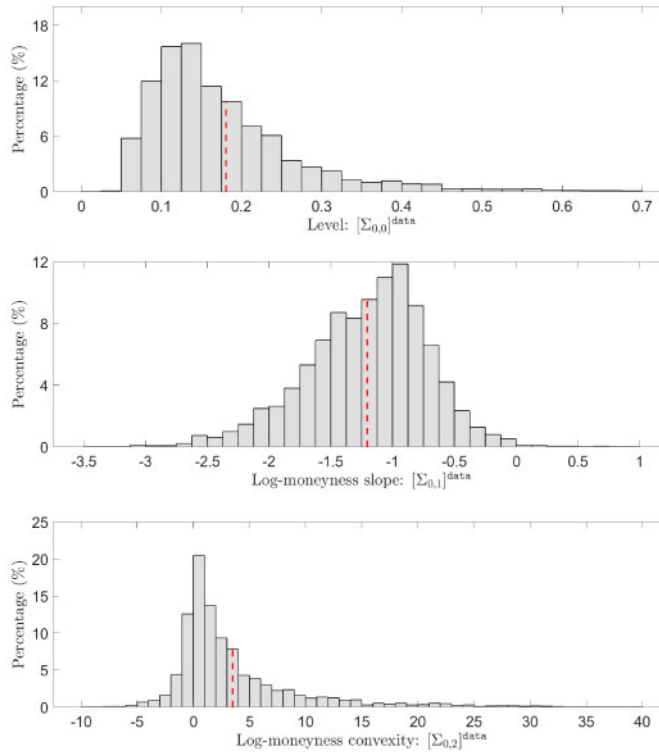

Figure 10
Distribution of the level $[\Sigma_{0,0}]^{\text{data}}$, log-moneyness slope $[\Sigma_{0,1}]^{\text{data}}$, and log-moneyness convexity $[\Sigma_{0,2}]^{\text{data}}$
$[\Sigma_{0,0}]^{\text{data}}$, $[\Sigma_{0,1}]^{\text{data}}$, and $[\Sigma_{0,2}]^{\text{data}}$ are the data of shape characteristics $\Sigma_{0,0}$, $\Sigma_{0,1}$, and $\Sigma_{0,2}$, respectively. They are computed from the bivariate regression in Equation (36) across the whole sample. In each panel, we plot a red dashed vertical bar to represent the mean of the corresponding histogram.

Eraker, Johannes, and Polson (2003), and Broadie, Chernov, and Johannes (2007), based on the time-series dynamics of underlying asset price, the volatility of volatility parameter ξ of the Heston model is usually estimated to be a small value and thus prevents the model to generate a sufficiently steep IV slope as observed from the real surface. However, for the purpose of constructing/estimating a model solely under the risk-neutral measure, our approach forces the ISV model to fit this slope by construction but without considering any constraints from the time-series dynamics of the underlying asset price under the physical measure. Recall that the closed-form formula for the log-moneyness slope $\Sigma_{0,1}$, given in Equation (33), is $\Sigma_{0,1}(v)=\rho\xi/(4v)$. Thus, for fitting the usually steep slope, the corresponding moment condition requires (given ρ) ξ to be larger than other methods, and this is what our GMM estimation procedure produces. Furthermore, based on the data $[\Sigma_{0,0}]^{\text{data}}$ and $[\Sigma_{0,1}]^{\text{data}}$ shown in Figure 10, the mean of the multiplicative data $[\Sigma_{0,0}]^{\text{data}}[\Sigma_{0,1}]^{\text{data}}$ is -0.21. On the other hand, it follows from Equation (33) that

$$\Sigma_{0,0}(v)\Sigma_{0,1}(v)=\frac{\rho\xi}{4}. \tag{37}$$

Similar to the aforementioned determination of ρ via the heuristic moment matching, we plug the estimated mean -0.21 of the multiplicative data and the estimator -0.619 of ρ as shown in Table 3 into Equation (37) to solve the parameter ξ as 1.37, which basically agrees with our GMM estimator.

Third, in both the exactly identified and overidentified cases, the estimators for κ are no less than 10, and in particular, larger than those estimated in the literature. This is necessary given the large values of the volatility of variance ξ, to keep the volatility process v_t mean-reverting sufficiently fast and consequently diminish the likelihood of having extreme volatilities. Fourth, again in both cases, the estimators of the long-term variance value α are around 0.04, corresponding to 20% volatility, which is consistent with the relatively low average values recorded by the S&P 500 index volatility during the overall sample period.

5.2 Nonparametric implied stochastic volatility model

Using the same data, and the same shape characteristics level $\Sigma_{0,0}$, log-moneyness slope $\Sigma_{0,1}$ and convexity $\Sigma_{0,2}$, as well as term-structure slope $\Sigma_{1,0}$ estimated from Equation (36), we now follow the method proposed in Section 2 to construct a nonparametric ISV model. In addition to analyzing the properties of the estimated coefficient functions, we demonstrate that our model is capable of fitting not only the four shape characteristics employed in the construction but also the remaining two second-order shape characteristics that do not serve as inputs: the mixed slope $\Sigma_{1,1}$ and term-structure convexity $\Sigma_{2,0}$.

The estimation results are summarized in Figure 11. The upper left, upper right, and middle left panels show the estimators $\hat{\mu}(\cdot)$, $-\hat{\gamma}(\cdot)$, and $\hat{\eta}^2(\cdot)$, respectively. Consider the upper left panel. The dots represent realizations of

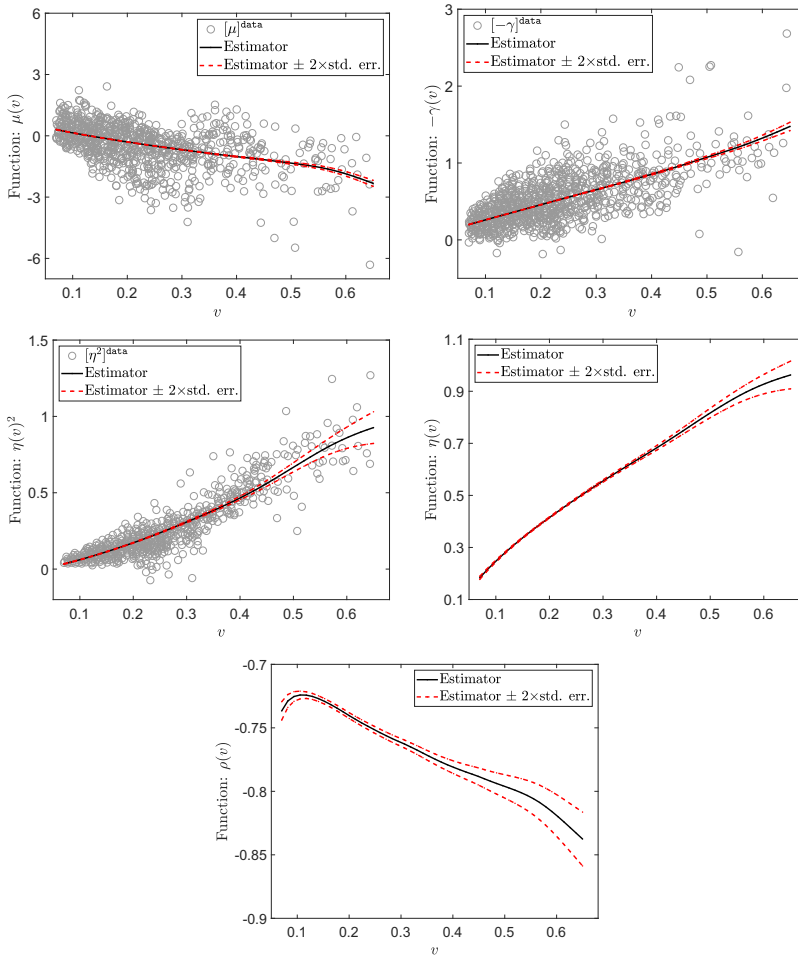

Figure 11
Nonparametric ISV model based on the full sample 2007–2017
In the upper left and middle left panels, the data $[\mu]^{\text{data}}$ and $[\eta^2]^{\text{data}}$ are calculated according to Equations (26) and (25), respectively. In the upper right panel, the data $[-\gamma]^{\text{data}}$ are simply the opposite numbers of the data $[\gamma]^{\text{data}}$, which are calculated according to Equation (20). In all these three panels, the nonparametric estimators are obtained by local linear regressions according to the method proposed in Section 2. In the middle right panel, the nonparametric estimator of η follows by taking the square root of the estimator of η^2. In the lowest panel, the nonparametric estimator of ρ follows from Equation (35). In all the panels, the standard errors of estimators are calculated by the bootstrap strategy described in Section 4.2.

$[\mu]^{\text{data}}$, which we recall are calculated according to Equation (26), while the nonparametric estimator of the function μ is shown as the solid curve. The confidence intervals on the curve are pointwise and represent two standard errors. The standard errors are calculated by the bootstrap procedure based on 500 replications, as validated in Section 4.2. Given the nonparametric

estimator of η^2 obtained based on the real sample (or bootstrap sample), we calculate the corresponding estimator of η by taking its square root. The results are in the middle right panel. Our suggestion to a reader interested in implementing the nonparametric estimator on a less liquid option contract would be to produce the corresponding figures on their specific data: overfitting if any should be visually clear in such a figure, and bandwidths adjusted accordingly.

We find that $\hat{\mu}(\cdot)$ is positive (resp. negative) when its argument is relatively small (resp. large), consistent with mean reversion in v_t. The upper right panel indicates that the coefficient function $\hat{\gamma}(\cdot)$ is always negative (the upper right panel shows $-\hat{\gamma}(\cdot)$) and approximately linear. As shown in the middle right panel, $\hat{\eta}(\cdot)$ is always positive and slightly concave when its argument is less than 0.3, as opposed to being approximately linear as γ is. The nonparametric ISV model can be used to formally assess the validity of a given parametric specification by testing the functional restrictions it imposes. For example, the linear mean-reverting parametric restriction $\mu(v_t;\theta)=\kappa(\alpha-v_t)$ that is employed in the vast majority of SV models can be tested using the statistic

$$\min_{\theta}\mathbb{E}\left[d\left(\mu(v_t;\theta);\hat{\mu}(v_t)\right)\right],$$

where d is a distance (such as the L^2 norm), and similarly for $\gamma(\cdot;\theta)$ and $\eta(\cdot;\theta)$: see Aït-Sahalia (1996) for a related use of a nonparametric estimator to test a parametric specification. Alternatively, the specification of a parametric model can be assessed by estimating the model using the parametric method of Section 3 and then employing GMM specification tests: see Newey (1985) for such tests.

Given the estimators of γ and η^2 based on the real sample (or bootstrap sample), we calculate the corresponding estimator of the leverage effect function ρ, that is, $\hat{\rho}(v_t)=\hat{\gamma}(v_t)/\sqrt{\hat{\gamma}(v_t)^2+\hat{\eta}(v_t)^2}$. The results are shown in the lowest panel of Figure 11. The estimated leverage effect estimator $\hat{\rho}(\cdot)$ is consistently negative, non-constant, and monotonically more negative when volatility increases (above 0.1). The negativeness of $\rho(\cdot)$ is a direct consequence of that of $\gamma(\cdot)$, given Equation (3). Furthermore, the results suggest that there is a continued increase (in absolute value) of $\rho(v_t)$ toward -1 as volatility increases.

This non-constant shape of $\hat{\rho}(\cdot)$ versus v_t implies that the leverage effect $\rho(\cdot)$ is indeed stochastic, unlike the assumption in the Heston model and in fact most existing parametric SV models. One consequence of this result is that the usual practice in the literature of specifying a SV model in the form

$$\frac{dS_t}{S_t}=(r-d)dt+v_t dW_{1t}, \tag{38}$$

$$dv_t=\mu(v_t)dt+\gamma(v_t)dW_{3t}, \tag{39}$$

with $\mathbb{E}[dW_{1t}dW_{3t}]=\rho dt$, ρ constant, should be replaced by the more appropriate formulation in Equations (1)–(2). The standard formulation in

Equations (38)–(39) in effect imposes a restriction on the model (constraining $\eta(\cdot)$ to be proportional to $\gamma(\cdot)$) that is not justified empirically.

5.3 Robustness checks

To check the robustness of our empirical results, we begin by verifying the goodness-of-fit of the shape characteristics $\Sigma_{i,j}$ involved in Equation (36). In each panel of Figure 12, we plot the inputted data for the shape characteristics $\Sigma_{i,j}$ as well as its fitted values $\hat{\Sigma}_{i,j}$. Here, the inputted data are inferred from the bivariate regression in Equation (36), while the fitted values $\hat{\Sigma}_{i,j}$ are obtained by plugging in $\hat{\mu}$, $\hat{\gamma}$, and $\hat{\eta}$, as well as $\hat{\gamma}'$ (which we recall is estimated at the same time as $\hat{\gamma}$ by locally linear kernel regression) in the corresponding formula for $\Sigma_{i,j}$ given in Equations (9)–(10). The fitted shape characteristics $\hat{\Sigma}_{0,1}$, $\hat{\Sigma}_{0,2}$, and $\hat{\Sigma}_{1,0}$ match the data well, which is expected since they are inputs in the construction. Surprisingly, however, we find that the fitted shape characteristics $\hat{\Sigma}_{1,1}$ and $\hat{\Sigma}_{2,0}$ also match the data well, as shown in the middle right and lowest panels of Figure 12, even though the mixed slope $\Sigma_{1,1}$ and term-structure convexity $\Sigma_{2,0}$ of the IV surface are not employed in the nonparametric construction of the implied model, and the calculation of their fitted values requires higher-order derivatives of the coefficient functions. This indicates that the nonparametric ISV model is flexible enough to reproduce all the second-order shape characteristics of IV surface, or conversely that all the shape characteristics of the IV surface up to the second order are consistent with the nonparametric ISV model. These six shape characteristics fitted by the model are more than enough to characterize an IV surface in the short-maturity and near at-the-money region, and represent the key characteristics of the IV surface that are relevant in pricing applications, such as straddle, risk reversal, butterfly spread, and calendar spread.

Next, we examine the impact of using calls versus puts versus both types of options together on the nonparametric estimation, as well as the stability for both of the parametric and nonparametric estimates by splitting the full sample into two: 2007–2011 (covering the financial crisis) and 2012–2017 (in the aftermath of the crisis). Results are in the Online Appendix. The use of put or call options does not alter the qualitative conclusions we have drawn regarding the shape of the coefficient functions of the nonparametric ISV model, and the estimated scale of model parameters (in the parametric case) as well as the estimated shape of the coefficient functions (in the nonparametric case) are remarkably consistent across the two subsamples, despite a vastly different financial environment.

5.4 Comparison with alternative methods and out-of-sample performance

The main existing method in the literature to estimate a parametric option pricing model using options data consists in minimizing, over the parameter

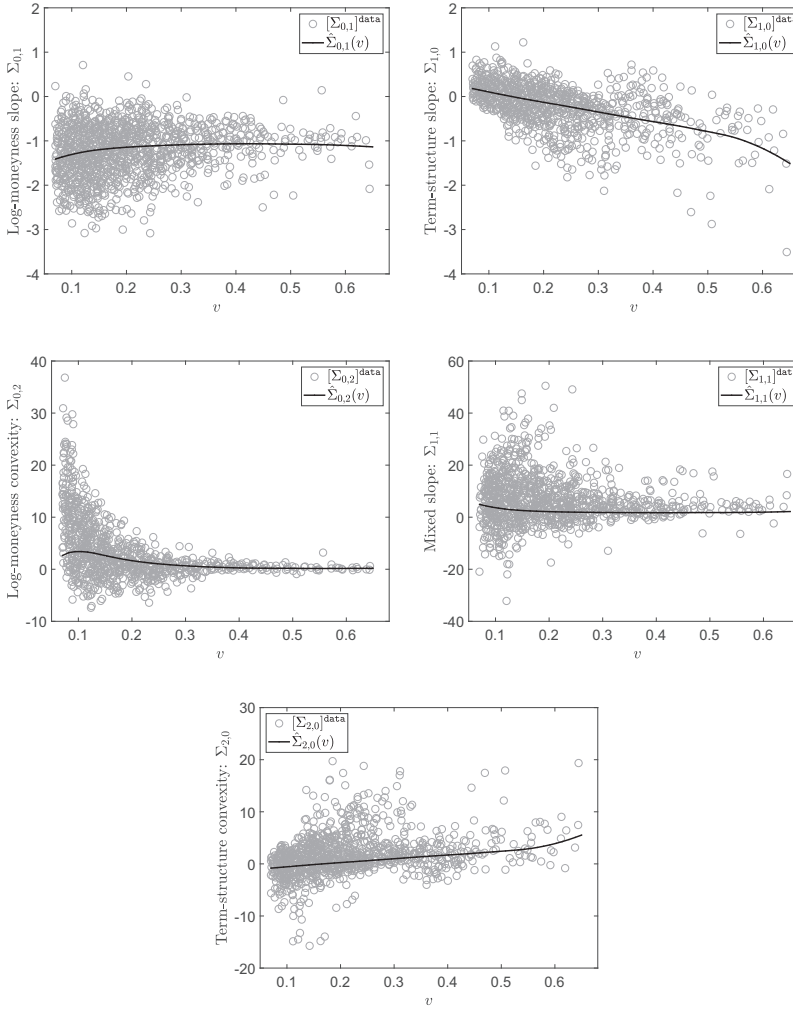

Figure 12
Fitting performance of the shape characteristics
In each panel, the data $[\Sigma_{i,j}]^{\text{data}}$ are obtained from the bivariate regression in Equation (36), while the fitted shape characteristics $\hat{\Sigma}_{i,j}$ are obtained from replacing the functions μ, γ, and η, as well as their derivatives by their nonparametric estimators in the formulae for $\Sigma_{i,j}$

values of the models, the root mean squared differences between the observed prices and model prices, or between the observed implied volatilities and those produced by the model, often including some weighting of the pricing errors (see, e.g., Huang and Wu 2004; Broadie, Chernov, and Johannes 2007; and Christoffersen, Heston, and Jacobs 2009). Either way, this approach requires the numerical computation of the model's prices, to be repeated

each time the minimization algorithm takes a new step. In special cases such as affine models, this computation can be facilitated by the fast Fourier transform (see, e.g., Heston 1993 and Carr and Madan 1998), but outside these cases, the computation-cum-minimization algorithm can be quite difficult in practice since it requires the continuous re-computation of the option prices as parameter values change by simulations, trees, or numerical solution of a partial differential equation.

By contrast, the estimation method in this paper requires no numerical computation of option prices, or in fact any symbolic calculation. For any given parametric model, the relevant formulae of shape characteristics in Equations (9)–(10) and additional high orders if necessary are already provided for that model's $\mu(\cdot;\theta)$, $\gamma(\cdot;\theta)$, and $\eta(\cdot;\theta)$ to be plugged into, and the estimation simply requires regressions of the IV observations, followed by a standard GMM procedure. Figure S.5 in the Online Appendix provides a quantitative look at the trade-off between accuracy for the various parameters and computational cost, as a comparison between our method and the conventional one. The results show that our estimation method is substantially more efficient.

On the other hand, our method is based on a local regression in (τ,k) near $(0,0)$, which is where most of the options' liquidity resides, but does not attempt to fit every available option. In principle, the conventional method can be made arbitrarily accurate as long as one spends enough computational time, for example, by densifying grid points for the involved Fourier numerical integration or by including more paths in simulations. Although we demonstrate that the approximation error of our method is much smaller than the bid-ask spread or violations of put-call parity, as shown in Figure 2, our method nevertheless remains based on an approximation and its ability to fit long-dated or far out-of-the-money options is going to be limited. It is mathematically possible to extend the method to compute local expansions in (τ,k) near different values (τ_0,k_0), in case these are empirically relevant for the application at hand, and apply the same estimation procedure with IV data that is selected in a neighborhood of that point. This said, as a practical matter, there are no natural alternative locations to $(\tau_0,k_0)=(0,0)$. First, most of the trading liquidity is in that neighborhood (near the money and slightly out of the money, and for maturities below six months, as shown in Figure 7), so this is where the data is most accurate, with the provision that extremely short maturities should be excluded for data reliability reasons. Second, the volatility state variable v_t is directly revealed in the limit at $(0,0)$, as the intercept in the IV surface regression, so the resulting expressions and implementation are much simpler there. Nevertheless, in the parametric case, expansions at different points could be incorporated as additional moment conditions, and fitted to different options, each in the neighborhood of that specific point, to provide further overidentification: there is no need necessarily to choose a single expansion point if more than one are available.

Table 4
In- and out-of-sample fit in terms of IVRMSE

Year		In-sample IVRMSE		Out-of-sample IVRMSE		Ratio: out/in	
In-sample	Out-of-sample	NP. Est.	P. Est.	NP. Est.	P. Est.	NP. Est.	P. Est.
2007	2008	0.0201	0.0818	0.0500	0.0528	2.4842	0.6455
2008	2009	0.0293	0.0568	0.0288	0.0401	0.9842	0.7062
2009	2010	0.0237	0.0342	0.0182	0.0604	0.7660	1.7688
2010	2011	0.0144	0.0341	0.0227	0.0455	1.5722	1.3342
2011	2012	0.0216	0.0749	0.0194	0.0695	0.8959	0.9278
2012	2013	0.0251	0.0339	0.0201	0.0392	0.8000	1.1568
2013	2014	0.0143	0.0369	0.0159	0.0458	1.1101	1.2405
2014	2015	0.0118	0.0316	0.0304	0.0261	2.5873	0.8242
2015	2016	0.0177	0.0458	0.0197	0.0607	1.1120	1.3244
2016	2017	0.0207	0.0234	0.0157	0.0549	0.7586	2.3464
Average		0.0199	0.0453	0.0241	0.0495	1.3071	1.2275

The third, fifth, and seventh (resp. fourth, sixth, and eighth) columns show the results for nonparametric estimation (resp. parametric estimation for the exactly identified case). In each of the first 10 rows, the "Ratio: out/in" is calculated as the ratio between the out-of-sample and in-sample IVRMSE's (defined in Equation (40)) from the same row. In the last row, each cell represents the average of the quantities for the 10 subsamples listed above it.

Finally, we examine the performance of our method out of sample. We follow the literature on estimating parametric option pricing models by computing in- and out-of-sample pricing errors for both nonparametric and parametric ISV models. We do this even though the ISV models are not estimated by minimizing such a criterion, so it is inherently at a disadvantage in such an exercise. To weigh the pricing errors at different time-to-maturities and log-moneyness, we employ the implied volatility root mean squared error (IVRMSE) criterion proposed in Christoffersen, Heston, and Jacobs (2009):

$$\text{IVRMSE} = \left(\frac{1}{\sum_{l=1}^{n} n_l} \sum_{l=1}^{n} \sum_{m=1}^{n_l} \left(\Sigma^{\text{data}}(\tau_l^{(m)}, k_l^{(m)}) - \Sigma^{\text{fitted}}(\tau_l^{(m)}, k_l^{(m)}, v_{l\Delta}) \right)^2 \right)^{\frac{1}{2}}, \tag{40}$$

where n (resp. n_l) denotes the total number of IV surfaces (resp. implied volatilities on the lth surface) as in Section 2. Here, Σ^{data} is the IV data, and Σ^{fitted} is the corresponding fitted value calculated from plugging in the fitted shape characteristics $\hat{\Sigma}_{i,j}$ into the bivariate expansion $\Sigma^{(2,(2,1,0))}$, which consists of all the terms up to the second order.

Table 4 reports the in- and out-of-sample IVRMSE over one-year moving windows, in each of which we treat the sample in the current (resp. next) year as the in-sample (resp. out-of-sample) data: we fit models to one full calendar year of data, and then generate out-of-sample forecasts for the next calendar year. Both the in- and out-of-sample IVRMSE can be compared with those shown in the standard literature, for example, Table 3 of Christoffersen, Heston, and Jacobs (2009), with the caveat that we use in our method a subset of the options (those employed in our estimation procedure) rather than the full universe of options that could otherwise be included.

We report results for both the nonparametric and parametric (Heston) ISV models. We find that the nonparametric ISV model produces not only lower in-sample IVRMSE, but also lower out-of-sample IVRMSE. This highlights the limitations of the Heston model in terms of fitting the varying shapes of the IV surface, and the fact that the nonparametric ISV model is not overfitting in-sample at the expense of out-of-sample performance. In the last two columns of Table 4, we calculate the ratio between out-of- and in-sample IVRMSE in each moving window experiment. The average ratio for the nonparametric ISV model is 1.31. These results show that our method performs well out of sample even when judged by a criterion that it is not designed to optimize. Table S.2 in the Online Appendix reports the in- and out-of-sample IVRMSE for different classes of time-to-maturities and log-moneyness of options, showing that the nonparametric model outperforms the parametric one in almost all cases.

6. Extension: Adding Jumps to the Model

A continuous nonparametric ISV model has been empirically demonstrated, in Section 5, to be robust in terms of fitting all the observable and practically useful shape characteristics up to the second order. We now generalize our approach to include jumps in returns to the model in Equations (1)–(2):

$$\frac{dS_t}{S_{t-}} = (r - d - \lambda(v_t)\bar{\mu})dt + v_t dW_{1t} + (\exp(J_t) - 1)dN_t, \tag{41}$$

$$dv_t = \mu(v_t)dt + \gamma(v_t)dW_{1t} + \eta(v_t)dW_{2t}. \tag{42}$$

N_t is a doubly stochastic Poisson process (or Cox process) with stochastic intensity $\lambda(v_t)$. J_t represents the size of log-price jump, which is assumed to be independent of the asset price S_t. When a jump occurs at time t, the log-price $\log S_t$ changes according to $\log S_t - \log S_{t-} = J_t$, that is, $S_t - S_{t-} = (\exp(J_t) - 1)S_{t-}$, where S_{t-} denotes the pre-jump price of the asset. The constant $\bar{\mu}$ is

$$\bar{\mu} = \mathbb{E}[\exp(J_t)] - 1,$$

where $\mathbb{E}$ denotes risk-neutral expectation. Based on this choice of $\bar{\mu}$, the drift term $-\lambda(v_t)\bar{\mu}dt$ compensates the jump component $(\exp(J_t) - 1)dN_t$ in the sense that the process $\int_0^t (\exp(J_s) - 1)dN_s - \int_0^t \lambda(v_s)\bar{\mu}ds$ becomes a martingale under the risk-neutral measure.

A typical example (introduced in Merton 1976) is one where the jump size J_t is normally distributed with mean μ_J and variance σ_J^2, in which case

$$\bar{\mu} = \exp\left(\mu_J + \frac{\sigma_J^2}{2}\right) - 1. \tag{43}$$

For future reference, we also define

$$\mu_+ = \frac{\mu_J}{\sigma_J} + \sigma_J \text{ and } \mu_- = \frac{\mu_J}{\sigma_J}, \tag{44}$$

and let $\mathcal{N}$ denote the standard normal cumulative distribution function.

Adding jumps to the volatility dynamics, or infinite activity jumps to either returns or volatility dynamics, has the potential to improve the fit and realism of the model even further but would substantially alter the approach we employ to derive the closed-form formulae for constructing ISV models. So for now we consider only the case of jumps in returns and leave these further extensions to future work.

6.1 The effect of jumps on the implied volatility surface

Following the same analysis as in Section 1, it is straightforward to see that the IV Σ remains as in the continuous model a trivariate function of τ, k, and v_t in the form given by Equation (6). However, under the model with jumps in Equations (41)–(42), the definition in Equation (7) for the IV shape characteristics $\Sigma_{i,j}$ is no longer valid, since the limit does not exist for $i \geq 1$ or $j \geq 2$. Fortunately, the bivariate expansion in Equation (11) for continuous models can be generalized to the case of jumps by incorporating the square root of time-to-maturity $\sqrt{\tau}$, as well as possibly its negative powers

$$\Sigma^{(J,\mathbf{L}(J))}(\tau,k,v_t)=\sum_{j=0}^{J}\sum_{i=\min(0,1-j)}^{L_j}\varphi^{(i,j)}(v_t)\tau^{\frac{i}{2}}k^j, \tag{45}$$

where $J \geq 0$ and $\mathbf{L}(J)=(L_0,L_1,\cdots,L_J)$ with $L_j \geq \min(0,1-j)$ are integers. The expansion in Equation (45) includes negative powers of $\sqrt{\tau}$ if the lowest power $\min(0,1-J)$ of $\sqrt{\tau}$ in the double summation is less than or equal to -1, that is, if $J \geq 2$. With the presence of jumps in return, the away-from-the-money IV will possibly explode to infinity as the time-to-maturity shrinks to zero: this limiting behavior was noted by Carr and Wu (2003), who used this divergence to construct a test for the presence of jumps in the data, and by Andersen and Lipton (2013).

Based on the generalized bivariate expansion in Equation (45), the expansion terms $\varphi^{(i,j)}$ correspond to at-the-money IV shape characteristics or combinations thereof, as the time-to-maturity shrinks to zero, up to time scalings. So, it is natural to view them as "generalized shape characteristics." In the construction of ISV models, the expansion terms $\varphi^{(i,j)}$ will play the same roles as $\Sigma_{i,j}$ in the continuous case in Equation (11), and can similarly be estimated from the IV surface data by a polynomial regression based on Equation (45).

We provide in Appendix B explicit formulae for the expansion terms $\varphi^{(i,j)}$ and use them to express the IV shape characteristics in terms of the model's coefficient functions.

6.2 Constructing an implied stochastic volatility model with jumps

We now use these formulae to construct a nonparametric ISV model with jumps. In the parametric case, the formulae provided in Appendix B for $\varphi^{(i,j)}$ can be directly used to form moment conditions as in the continuous case in Section

3. In the nonparametric case, these formulae imply a system of equations for recovering the model coefficient functions from the IV surface's shape characteristics that must be solved. Theorem B.2 in Appendix B.1 shows that this system of equations can be inverted in closed form, so that all the model coefficient functions can be recovered explicitly in terms of the $\varphi^{(i,j)}$ shape characteristics.

Using these results, here is how the jump size parameters μ_J and σ_J are determined from the IV surface:[9]

$$\frac{\bar{\mu}+2-2(\bar{\mu}+1)\mathcal{N}(\mu_+)-2\mathcal{N}(\mu_-)}{-\bar{\mu}+2(\bar{\mu}+1)\mathcal{N}(\mu_+)-2\mathcal{N}(\mu_-)}$$

$$=\frac{1}{\lim\limits_{\tau\to 0}\Sigma(\tau,0,v_t)^2}\left[\frac{\lim\limits_{\tau\to 0}2\sqrt{\tau}\frac{\partial^2\Sigma}{\partial\tau\partial k}(\tau,0,v_t)}{\lim\limits_{\tau\to 0}\frac{\sqrt{\tau}}{2}\frac{\partial^2\Sigma}{\partial k^2}(\tau,0,v_t)}-2(d-r)\right] \tag{46}$$

and

$$\frac{\sqrt{2}\bar{\mu}}{3\sqrt{\pi}[-\bar{\mu}+2(\bar{\mu}+1)\mathcal{N}(\mu_+)-2\mathcal{N}(\mu_-)]}$$

$$=\frac{\lim\limits_{\tau\to 0}\Sigma(\tau,0,v_t)\lim\limits_{\tau\to 0}\frac{\tau}{6}\frac{\partial^3\Sigma}{\partial k^3}(\tau,0,v_t)}{\lim\limits_{\tau\to 0}\frac{\sqrt{\tau}}{2}\frac{\partial^2\Sigma}{\partial k^2}(\tau,0,v_t)}, \tag{47}$$

where we recall that $\bar{\mu}$, μ_+, and μ_- are deterministic functions of μ_J and σ_J defined in Equations (43) and (44). According to these equations, one needs various at-the-money IV shape characteristics in both log-moneyness and time-to-maturity dimensions to pin down μ_J and σ_J, without requiring any prior identification of any of the coefficient functions $\lambda(\cdot)$, $\mu(\cdot)$, $\gamma(\cdot)$, or $\eta(\cdot)$. In particular, we see that the third-order derivative $\partial^3\Sigma/\partial k^3$ plays a crucial role. This is one of the main differences compared with the continuous case developed in Theorem 1. This theoretical finding is somewhat unfortunate from an empirical perspective, as it implies that the jump size parameters μ_J and σ_J depend on a higher-order structure of the IV surface that will be difficult to estimate precisely in the absence of large amounts of high-quality options data.

The stochastic intensity function $\lambda(v_t)$ is characterized by:

$$\lambda(v_t)=\frac{2\sqrt{2}\lim\limits_{\tau\to 0}\Sigma(\tau,0,v_t)^2\cdot\lim\limits_{\tau\to 0}\frac{\sqrt{\tau}}{2}\frac{\partial^2\Sigma}{\partial k^2}(\tau,0,v_t)}{\sqrt{\pi}[-\bar{\mu}+2(\bar{\mu}+1)\mathcal{N}(\mu_+)-2\mathcal{N}(\mu_-)]}. \tag{48}$$

So the short-maturity at-the-money IV convexity $\partial^2\Sigma(\tau,0,v_t)/\partial k^2$ is involved in determining the stochastic intensity function $\lambda(v_t)$ but not any third-order characteristics, except of course that those were already needed to

[9] This is obtained by combining the algebraic equation system in Equations (B34)–(B35) with the geometric interpretations of the involved expansion terms $\varphi^{(0,0)}$, $\varphi^{(-2,3)}$, $\varphi^{(1,1)}$, and $\varphi^{(-1,2)}$ provided in Equations (B24)–(B27).

identify $\bar{\mu}$, μ_+, and μ_-, which enter Equation (48). In the continuous case, the at-the-money convexity is finite as the time-to-maturity shrinks to zero, since the expansion in Equation (11) implies $\lim_{\tau\to 0}\partial^2\Sigma(\tau,0,v_t)/\partial k^2 = 2\sigma^{(0,2)}(v_t)$. By contrast, under the discontinuous model, the convexity $\partial^2\Sigma(\tau,0,v_t)/\partial k^2$ explodes to infinity as the time-to-maturity shrinks to zero, since Equation (45) directly implies that $\partial^2\Sigma(\tau,0,v_t)/\partial k^2 \sim 2\varphi^{(-1,2)}(v_t)/\sqrt{\tau}$ as $\tau\to 0$ with $\varphi^{(-1,2)}(v_t)$ finite. The formula in Equation (48) remains valid in the limiting case where the intensity $\lambda(v_t)$ tends to zero, that is, the jumps degenerate. This is because the convexity behaves in that case according to $\lim_{\tau\to 0}\sqrt{\tau}\partial^2\Sigma(\tau,0,v_t)/\partial k^2 = 0$, which obviously results in the right-hand side of Equation (48) tending to zero.

The volatility functions $\gamma(v_t)$ and $\eta(v_t)$ and the drift function $\mu(v_t)$ are all affected by the presence of jumps. Compared with the continuous case, the third-order mixed partial derivative $\partial^3\Sigma/\partial k^2\partial\tau$ (resp. term-structure slope $\partial\Sigma/\partial\tau$ and term-structure convexity $\partial^2\Sigma/\partial\tau^2$) participate in determining the volatility function $\eta(v_t)$ (resp. the drift function $\mu(v_t)$). By contrast in the continuous case, the term-structure slope $\partial\Sigma/\partial\tau$ is the only IV shape characteristic along the term-structure dimension that matters.

In Appendix B.2, we specialize the formulae to the special case of the jump-diffusion model of Merton (1976) and show how the above estimation procedure applies to this model to identify all the parameters based on observable shape characteristics.

6.3 Empirical challenges when jumps are present in the model

So, we have shown that it is possible in theory to imply a SV model with jumps from the shape characteristics of the IV surface. However, given the current liquidity of options markets and resulting availability of options data, one would encounter significant practical challenges when implementing the previously described strategy. As we did in the continuous case, it is natural to interpret the expansion in Equation (45) as the following regression

$$\Sigma^{\text{data}}(\tau_l^{(m)},k_l^{(m)}) = \sum_{j=0}^{J}\sum_{i=\min(0,1-j)}^{L_j}\beta_l^{(i,j)}(\tau_l^{(m)})^{\frac{i}{2}}k^j + \epsilon_l^{(m)}, \tag{49}$$

for $m=1,2,\ldots,n_l$, and the estimator of the coefficient $\beta_l^{(i,j)}$ serves as the data of the generalized shape characteristic $[\varphi^{(i,j)}]_l^{\text{data}}$.

Similar to the case for the regression in Equation (18), the choice of the orders J, L_0, L_1, …, L_J and the regions of IV surfaces data employed in the regression in Equation (49) should strike a balance between, on the one hand, the accuracy of the expansion $\Sigma^{(J,\mathbf{L}(J))}$ and, on the other hand, overfitting the regression to the IV data. Most importantly, the presence of jumps necessitates the estimation of third-order characteristics of the IV surface, which in our experience is effectively impossible to do accurately given the limitations of the data currently available. A substantially denser set of observations on option

prices or implied volatilities would be necessary to accurately estimate third-order derivatives without the error introduced by the strike and maturity surface interpolation implicit in Equation (49). Furthermore, the divergence of the IV surface due to the presence of negative powers of τ also requires very short maturity options to be accurately observed (as in Carr and Wu 2003's test for the presence of jumps in options data); such data can be affected by trading patterns specific to options with, for example, time-to-maturity τ less than one week, and log-moneyness k within $\pm 0.1 v\sqrt{\tau}$, where v represents the instantaneous volatility. This makes inferring the desired data $[\varphi^{(i,j)}]_l^{\text{data}}$ from the IV surface and the subsequent procedures for constructing ISV models substantially more difficult since we do not only need to just identify the divergence as in Carr and Wu (2003) but also estimate higher-order coefficients.

7. Conclusions and Future Directions

We proposed to construct ISV models that are consistent with observed shape characteristics of the IV market data, either nonparametrically or parametrically. To the best of our knowledge, this estimation method is new, and this paper is the first to estimate, from derivatives prices, a nonparametric SV model with fully flexible continuous volatility dynamics. The SV coefficient functions are estimated by exploiting the restrictions provided by the IV shape characteristics. Our method hinges on closed-form formulae linking the SV coefficients to the IV shape characteristics, which are themselves extracted from the IV data as a simple polynomial regression of IV on time-to-maturity and log-moneyness.

When estimating a parametric ISV model, whether the model is analytically tractable or not (i.e., affine or not) from an option pricing perspective does not constrain our method. The GMM conditions we provide are fully explicit and require no other computations (such as option prices) that are inherently difficult in other methods. Indeed, the main alternative method in the literature to estimate an SV model from option prices consists in minimizing the root mean squared pricing errors between the model's and the market prices. This requires the numerical computation of the model's prices each time the minimization algorithm adjusts the model's parameter values; such minimization on top of the numerical computation of prices evaluated at the current parameter values can be very challenging, especially outside of the affine class of models. Our GMM method also minimizes squared deviations between market data (in the form of the shape characteristics) and the corresponding model's output. The difference is that, unlike option prices, the model's output in the form of shape characteristics is given by explicit closed-form formulae. Our method effectively replaces a GMM procedure where the moment condition (market minus model's option prices) needs to be numerically computed along the way with a GMM procedure where the moment conditions are given in closed form (market minus model's shape characteristics).

New insights from the empirical results relate to the shape of the estimated functions as they depend nonparametrically on v_t and the information this provides regarding the relative fit of different parametric models. Using the level, slope, and convexity of the IV surface in the log-moneyness direction, and the slope in the term-structure direction, as inputs, we find empirically that the resulting nonparametric ISV model exhibits a strong leverage effect between the innovations in returns and volatility, weak mean reversion in volatility, strong monotonicity, and state dependency in volatility of volatility. Being able to conveniently estimate a vast range of models using this new method opens the door to the exploration of alternative, possibly better-fitting SV specifications than what current methods allow. We showed that the model also matches surprisingly well the convexity in the term-structure dimension and the mixed slope, even though those are not employed as inputs to the estimation procedure. As a result, the nonparametric ISV model is capable of fitting all six observable (and useful in practice) shape characteristics up to the second order. We also found that the estimated nonparametric coefficient functions of the model are stable over time, preserving the main features of the volatility dynamics.

While the paper focuses on data-based estimation, another interesting application of the method is a trading one that is not data-dependent: in an illiquid market, or in directional trading, a trader might express views as to the shape of the IV surface and seek an SV model consistent with those views, in order to price/hedge options or other contracts. Our method delivers the necessary SV model for that purpose.

At least in principle, ISV models in higher dimensions can be constructed using the same principle, although a bivariate nonparametric ISV model, as constructed here, is flexible enough to fit the shape characteristics of the IV surface up to the second order, that is, the level, slope, and convexity along both the moneyness and term-structure dimensions. When jumps are added to the model, we showed that the same ideas continue to work in principle and that a full characterization of the SV model can still be obtained in closed form, at least for models with jumps only in the returns dynamics. However, higher-order shape characteristics become necessary, for example, the third-order shape characteristic in the log-moneyness dimension, for inferring the jump components. The estimation of these higher-order derivatives requires substantially denser options observations in both time and moneyness than is currently available, even though options with shorter maturities, such as weekly, have recently become more liquid. So while we show that it is theoretically possible, in practice we have limited ability to identify the jump components due to the current lack of liquidity, which could be circumvented in the future as more and better data becomes available. Adding jumps to the volatility dynamics, or infinite activity jumps to either returns or volatility dynamics, would substantially alter the approach we employ to derive the IV expansion as a tool, and require in practice more accurate and delicate shape characteristics for fully recovering the SV model's components. Finally, the method in the

paper (under the risk-neutral measure) could in principle be combined with additional moment conditions calibrated to the underlying returns data (under the physical probability measure), and estimate the volatility risk premium as a result.

Appendix A. Closed-Form Shape Characteristics of Implied Volatility for Continuous Stochastic Volatility Models

In this appendix, we sketch how to obtain the shape characteristics $\Sigma_{i,j}$ in Equation (7) in closed form for the continuous SV model in Equations (1)–(2). Owing to the proportional relation between the shape characteristic $\Sigma_{i,j}$ and the expansion term $\sigma^{(i,j)}$ provided in Equation (11), it suffices to calculate the expansion terms $\sigma^{(i,j)}$ explicitly in what follows.

To simplify notations, we write $S_t = s$ and $v_t = v$ at time t. The main idea hinges on expanding both sides of the identity in Equation (5) with respect to the square root of time-to-maturity $\epsilon = \sqrt{\tau}$ and log-moneyness k and then matching expansion terms of the same orders. Thus, we propose the following $(J, \mathbf{L}(J))$th-order expansion of $\bar{P}(\tau, k, v_t)$ introduced in Equation (4) and appearing on the right-hand side of Equation (5):

$$\bar{P}^{(J,\mathbf{L}(J))}(\epsilon^2, k, v) = \sum_{j=0}^{J} \sum_{i=1-j}^{L_j} p^{(i,j)}(v) \epsilon^i k^j, \text{ with } \epsilon = \sqrt{\tau}, \tag{A1}$$

for any orders $J \geq 0$ and $L_j \geq 1 - j$, $j = 0, 1, \ldots, J$. The coefficients $p^{(i,j)}$ can be calculated explicitly by following Li (2014), in which the option price $P(\epsilon^2, k, s, v)$ admits a pseudo univariate expansion with respect to ϵ with closed-form expansion terms depending on both ϵ and k. The bivariate expansion in Equation (A1) follows from taking $s = 1$ in this univariate expansion and further expanding the coefficients with respect to k and ϵ.

Now, based on the bivariate expansion in Equation (A1) of $\bar{P}(\tau, k, v_t)$, which appears on the right-hand side of Equation (5), in what follows, we accordingly expand the composite function $\bar{P}_{\mathrm{BS}}(\tau, k, \Sigma(\tau, k, v))$ on the left-hand side. By matching the expansion term on both sides, we establish a set of iterations and solve the expansion terms $\sigma^{(i,j)}$ recursively.

We start from the following expansion of at-the-money IV $\Sigma(\epsilon^2, 0, v)$ with respect to ϵ:

$$\Sigma^{(L_0)}(\epsilon^2, 0, v) = \sum_{i=0}^{L_0} \sigma^{(i,0)}(v) \epsilon^{2i}, \tag{A2}$$

which is obtained by setting $k = 0$ in the bivariate expansion in Equation (11). According to Durrleman (2008), $\Sigma(\epsilon^2, 0, v)$ converges to the instantaneous SV of the asset price v_t, as the time-to-maturity $\tau = \epsilon^2$ approaches zero. Thus, $\sigma^{(0,0)}(v) = v$. By taking $\sigma^{(0,0)}(v)$ as the initial input, all other expansion terms can be solved recursively.

To compute the expansion terms $\sigma^{(i,0)}$, we apply the at-the-money condition $k = 0$ on both sides of Equation (5) to obtain

$$\bar{P}(\epsilon^2, 0, v) = \bar{P}_{\mathrm{BS}}(\epsilon^2, 0, \Sigma(\epsilon^2, 0, v)). \tag{A3}$$

Expanding both sides of Equation (A3) with respect to ϵ and matching the coefficients, we can obtain a system of equations. The closed-form formulae of expansion terms $\sigma^{(i,0)}$ follow by solving the equations recursively. Indeed, for the left-hand side of Equation (A3), the expansion of $\bar{P}(\epsilon^2, 0, v)$ with respect to ϵ can be obtained from Equation (A1) by setting $k = 0$, that is,

$$\bar{P}^{(L_0)}(\epsilon^2, 0, v) = \sum_{l=0}^{L_0} p^{(l,0)}(v) \epsilon^l. \tag{A4}$$

For the right-hand side of Equation (A3), the expansion of $\bar{P}_{\mathrm{BS}}(\epsilon^2, 0, \Sigma(\epsilon^2, 0, v))$ with respect to ϵ follows by combining the expansion of the function $\bar{P}_{\mathrm{BS}}(\epsilon^2, 0, \sigma)$, which is obtained by expanding

the explicit formula of $\bar{P}_{\mathrm{BS}}(\epsilon^2,0,\sigma)$, and the expansion of at-the-money IV $\Sigma(\epsilon^2,0,v)$, which is proposed in Equation (A2) with the undetermined expansion terms $\sigma^{(i,0)}$. Then, the closed-form expansion of $\bar{P}_{\mathrm{BS}}(\epsilon^2,0,\Sigma(\epsilon^2,0,v))$ has the form

$$\bar{P}_{\mathrm{BS}}^{(J)}(\epsilon^2,0,\Sigma(\epsilon^2,0,v))=\sum_{l=1}^{J}\tilde{p}^{(l,0)}(v)\epsilon^l, \tag{A5}$$

for any integer $J\geq 1$. In particular, for any odd integer $l\geq 3$, the expansion term $\tilde{p}^{(l,0)}$ by computation consists of IV expansion terms $\sigma^{(i,0)}$ for all $i\leq(l-1)/2$. Matching the coefficients of the expansions in Equations (A5) and (A4) yields the system of equations

$$p^{(l,0)}(v)=\tilde{p}^{(l,0)}(v),\ \text{for any odd integer } l\geq 1. \tag{A6}$$

For any integer $i\geq 1$, the closed-form formula of the expansion term $\sigma^{(i,0)}(v)$ follows from solving Equation (A6) with $l=2i+1$.

Finally, to compute the expansion terms $\sigma^{(i,j)}$ for $j\geq 1$, we resort to the identity

$$\frac{\partial^j}{\partial k^j}\bar{P}(\epsilon^2,0,v)=f_j(\epsilon,v), \tag{A7}$$

which is obtained from differentiating the identity in Equation (5) j times with respect to k and then applying the at-the-money condition $k=0$. Here, the function f_j is defined by

$$f_j(\epsilon,v)=\sum_{0\leq m_1\leq m_2\leq j}\binom{j}{m_2}\frac{\partial^{j-m_2+m_1}\bar{P}_{\mathrm{BS}}}{\partial k^{j-m_2}\partial\sigma^{m_1}}(\epsilon^2,0,\Sigma(\epsilon^2,0,v))G^{(m_1,m_2)}(\epsilon,v), \tag{A8}$$

where the nonnegative integers m_1 and m_2 satisfy that $m_1=0$ if and only if $m_2=0$. The function $G^{(m_1,m_2)}$ is defined by $G^{(0,0)}(\epsilon,v)=1$ and

$$G^{(m_1,m_2)}(\epsilon,v)=\sum_{\mathbf{l}\in\mathcal{S}_{m_1,m_2}}\frac{m_2!}{R(\mathbf{l})}\prod_{\ell=1}^{m_1}\frac{1}{i_\ell!}\frac{\partial^{i_\ell}\Sigma}{\partial k^{i_\ell}}(\epsilon^2,0,v), \tag{A9}$$

for $1\leq m_1\leq m_2$. Here, the integer index set $\mathcal{S}_{m_1,m_2}$ is given by

$$\mathcal{S}_{m_1,m_2}=\{(i_1,i_2,\cdots,i_{m_1}):1\leq i_1\leq i_2\leq\cdots\leq i_{m_1},\ \textstyle\sum_{\ell=1}^{m_1}i_\ell=m_2\}, \tag{A10}$$

and the function $R(\mathbf{l})$ is a constant defined by the product of factorials of the repeating times of distinct nonzero entries appearing more than once in index $\mathbf{l}$. For example, in index $\mathbf{l}=(1,1,2,2,2)$, distinct entries 1 and 2 appear twice and thrice, respectively. Then, the constant $R(\mathbf{l})$ is calculated as $2!\times 3!=24$. Similar to the previous case of $j=0$, by expanding both sides of Equation (A7) with respect to ϵ and matching the coefficients, we can obtain a system of equations for solving the expansion terms $\sigma^{(i,j)}(v)$ recursively.

Indeed, the expansion of $\partial^j\bar{P}(\epsilon^2,0,v)/\partial k^j$ on the left-hand side of Equation (A7) is

$$\frac{\partial^j\bar{P}^{(L_j)}}{\partial k^j}(\epsilon^2,0,v)=\sum_{i=1-j}^{L_j}j!\,p^{(i,j)}(v)\epsilon^i, \tag{A11}$$

which is obtained from differentiating the expansion in Equation (A1) j times with respect to k and then setting $k=0$. According to the definition in Equation (A8), the expansion of the function f_j on the right-hand side of Equation (A7) hinges on the expansions of two types of ingredients

$$\frac{\partial^{j-m_2+m_1}\bar{P}_{\mathrm{BS}}}{\partial k^{j-m_2}\partial\sigma^{m_1}}(\epsilon^2,0,\Sigma(\epsilon^2,0,v))\text{ and }G^{(m_1,m_2)}(\epsilon,v). \tag{A12}$$

As to the first ingredient, its expansion can be obtained by combining the expansions of the Black-Scholes sensitivities $\partial^{j-m_2+m_1}\bar{P}_{\mathrm{BS}}(\epsilon^2,0,\sigma)/\partial k^{j-m_2}\partial\sigma^{m_1}$, which is obtained based on the

explicit formula of $\bar{P}_{\mathrm{BS}}$, and the expansion of at-the-money IV $\Sigma(\epsilon^2,0,v)$, which is explicitly computed from the preceding iteration for $j=0$. By combining these two types of expansions, we obtain the Jth-order expansion of the first ingredient in Equation (A12) as

$$\frac{\partial^{j-m_2+m_1}\bar{P}_{\mathrm{BS}}^{(J)}}{\partial k^{j-m_2}\partial\sigma^{m_1}}(\epsilon^2,0,\Sigma(\epsilon^2,0,v))=\sum_{l=1-j+m_2}^{J}H_{l,m_1}^{(j-m_2)}\epsilon^l, \tag{A13}$$

for any integer order $J\geq 1-j+m_2$, where the expansion terms $H_{l,m_1}^{(j-m_2)}$ consist of various Black-Scholes sensitivities and at-the-money IV expansion terms $\sigma^{(i,0)}$.

To obtain the expansion of the second ingredient $G^{(m_1,m_2)}(\epsilon,v)$ in Equation (A12), according to its definition in Equation (A9), it suffices to combine the expansions of various at-the-money IV shape characteristics $\partial^{i_\ell}\Sigma(\epsilon^2,0,v)/\partial k^{i_\ell}$, while the expansion of $\partial^{i_\ell}\Sigma(\epsilon^2,0,v)/\partial k^{i_\ell}$ follows

$$\frac{\partial^{i_\ell}\Sigma^{(L_{i_\ell})}}{\partial k^{i_\ell}}(\epsilon^2,0,v)=\sum_{l=0}^{L_{i_\ell}}i_\ell!\sigma^{(l,i_\ell)}(v)\epsilon^{2l},$$

by differentiating Equation (11) i_ℓ times with respect to k and then setting $k=0$. Then, the function $G^{(m_1,m_2)}$ admits a Jth-order expansion in the form

$$G^{(m_1,m_2)}(\epsilon,v)=\sum_{l=0}^{J}G_l^{(m_1,m_2)}\epsilon^{2l}, \tag{A14}$$

for any integer order $J\geq 0$. Here, the expansion term $G_l^{(m_1,m_2)}$ is defined according to

$$G_l^{(m_1,m_2)}=\sum_{\mathbf{l}\in\mathcal{S}_{m_1,m_2},\,\mathbf{v}\in\mathcal{T}_{\mathbf{l},l}}\frac{m_2!}{R(\mathbf{l})}\prod_{\ell=1}^{m_1}\sigma^{(v_\ell,i_\ell)}(v),$$

for any integers $m_2\geq m_1\geq 1$ and $l\geq 0$, with the integer index set $\mathcal{S}_{m_1,m_2}$ given in Equation (A10) and the function $R(\mathbf{l})$ provided right after Equation (A10). Moreover, for any index $\mathbf{l}\in\mathcal{S}_{m_1,m_2}$, the integer index set $\mathcal{T}_{\mathbf{l},l}$ is defined by

$$\mathcal{T}_{\mathbf{l},l}=\left\{\mathbf{v}=\left(v_1,v_2,\cdots,v_{m_1}\right):v_1+\cdots+v_{m_1}=l \text{ and } v_\ell\geq 0, \text{ for } \ell=1,2,\ldots,m_1\right\}.$$

Based on the expansions in Equations (A13) and (A14), it follows from the definition in Equation (A8) that the function $f_j(\epsilon,v)$ admits the following Jth-order expansion

$$f_j^{(J)}(\epsilon,v)=\sum_{l=1-j}^{J}\tilde{p}^{(l,j)}(v)\epsilon^l, \tag{A15}$$

for any integer $J\geq 1-j$, where the expansion term $\tilde{p}^{(l,j)}$ satisfies

$$\tilde{p}^{(l,j)}(v)=\sum_{0\leq m_1\leq m_2\leq j}\binom{j}{m_2}\sum_{l_1+2l_2=l,\ l_1\geq 1-j+m_2,\ l_2\geq 0}H_{l_1,m_1}^{(j-m_2)}G_{l_2}^{(m_1,m_2)},$$

for any integer $l\geq 1-j$. In particular, for any odd integer $l\geq 1$, the expansion term $\tilde{p}^{(l,j)}(v)$ consists of IV expansion terms $\sigma^{(i,m)}$ for all $0\leq m\leq j$ and $0\leq i\leq(l-1)/2+\lfloor j-m\rfloor/2$, where the notation $\lfloor a\rfloor$ represents the integer part of any arbitrary real number a. By matching the coefficients of expansions in Equations (A15) and (A11), we obtain the following system of equations:

$$j!p^{(l,j)}(v)=\tilde{p}^{(l,j)}(v), \text{ for any integer } l\geq 1-j. \tag{A16}$$

For any integer $i\geq 1$, the closed-form formula of the expansion term $\sigma^{(i,j)}(v)$ follows from solving Equation (A16) with $l=2i+1$.

Appendix B. Closed-Form Generalized Shape Characteristics of Implied Volatility for Stochastic Volatility Models with Jumps

Similar to the derivation for the continuous case, the generalized shape characteristics $\varphi^{(i,j)}$, that is, the expansion terms introduced in Equation (45), can be solved by iterations. These iterations can be obtained by expanding both sides of the identity in Equation (5) with respect to the square root of time-to-maturity $\epsilon=\sqrt{\tau}$ and log-moneyness k and then matching expansion terms of the same orders. Solving these matched equations leads to the desired iterations. Thus, by omitting the similar arguments, it suffices to the following indispensable ingredient for completing the derivation: Under the general SV model with jumps in Equations (41)–(42), we propose the following closed-form bivariate expansion of $\bar{P}(\tau,k,v_t)$ introduced in Equation (4) and appearing on the right-hand side of Equation (5):

$$\bar{P}^{(J,\mathbf{L}(J))}(\epsilon^2,k,v)=\sum_{j=0}^{J}\sum_{i=1-j}^{L_j}\bar{p}^{(i,j)}(v)\epsilon^i k^j,\ \text{with}\ \epsilon=\sqrt{\tau}, \tag{B1}$$

for any orders $J\geq 0$ and $L_j\geq 1-j,\ j=0,1,2,\ldots,J$. This expansion generalizes that for the continuous model in Equations (1)–(2) provided in Equation (A1) and can be developed from the following three steps. Without loss of generality, by the time-homogeneity property of the model in Equations (41)–(42), the time span from t to T can be translated to that from 0 to $\tau=T-t$ for simplicity. We assume $S_0=s$ and $v_0=v$.

Step 1 – Representing $\bar{P}(\tau,k,v)$ under an auxiliary measure: We will rewrite the expectation representation of $\bar{P}(\tau,k,v)$ in Equation (4) under an auxiliary probability measure, under which the expectation becomes easier to handle. We denote by $\mathbb{Q}$ the assumed risk-neutral measure and denote by $\mathcal{F}_t$ the filtration generated by the process $(S_t,v_t)^\top$. The new probability measure $\tilde{\mathbb{Q}}$ is induced by a Radon-Nikodým derivative Λ_t according to

$$\left.\frac{d\mathbb{Q}}{d\tilde{\mathbb{Q}}}\right|_{\mathcal{F}_t}=\Lambda_t,\ \text{with}\ \Lambda_t\ \text{defined as}\ \Lambda_t=\left(\prod_{i=1}^{N_t}\lambda(v_{\tau_i})\right)\exp\left\{t-\int_0^t\lambda(v_s)ds\right\}, \tag{B2}$$

where τ_i denotes the arrival time of the ith jump, that is, $\tau_i=\inf\{t\geq 0: N_t=i\}$; in particular, $\Lambda_0=1$. According to Theorem T3 in Chapter VI of Brémaud (1981), N_t is a Poisson process with constant jump intensity 1 under the measure $\tilde{\mathbb{Q}}$. Changing the measure from $\mathbb{Q}$ to $\tilde{\mathbb{Q}}$ yields the following equivalent expectation representation of $\bar{P}(\tau,k,v)$:

$$\bar{P}(\tau,k,v)=e^{-r\tau}\tilde{\mathbb{E}}\left[\Lambda_\tau\max\left(e^k-\frac{S_\tau}{s},0\right)\right],$$

where $\tilde{\mathbb{E}}$ represents the expectation under the measure $\tilde{\mathbb{Q}}$. Then, by conditioning on the number of jumps between 0 and τ, we reformulate $\bar{P}(\tau,k,v_t)$ as the summation form

$$\bar{P}(\tau,k,v)=\sum_{\ell=0}^{\infty}e^{-r\tau}\tilde{\mathbb{Q}}(N_\tau=\ell)\bar{P}_\ell(\tau,k,v), \tag{B3}$$

with

$$\bar{P}_\ell(\tau,k,v)=\tilde{\mathbb{E}}^{(\ell)}\left[\Lambda_\tau\max\left(e^k-\frac{S_\tau}{s},0\right)\right], \tag{B4}$$

where the multiplier $\tilde{\mathbb{Q}}(N_\tau=\ell)$, as the probability of $N_\tau=\ell$ under the measure $\tilde{\mathbb{Q}}$ can be explicitly calculated as $\tau^\ell e^{-\tau}/\ell!$, and the notation $\tilde{\mathbb{E}}^{(\ell)}[\cdot]$ serves as the abbreviation of the conditional expectation $\tilde{\mathbb{E}}[\cdot|N_\tau=\ell]$.

According to the relation in Equation (B3), to expand $\bar{P}$, it suffices to multiply the expansions of $e^{-r\tau}$ and $\tilde{\mathbb{Q}}(N_\tau=\ell)=\tau^\ell e^{-\tau}/\ell!$ with respect to τ, which are trivial, and the expansion of conditional expectation $\bar{P}_\ell$ with respect to $\epsilon=\sqrt{\tau}$ and k for any $\ell\geq 0$. To expand $\bar{P}_\ell$ for $\ell=0$, in the beginning

of Step 2, we propose a decomposition of Λ_τ and S_τ. Then, based on this decomposition, we apply the method proposed in Li (2014) and develop a pseudo expansion of $\bar{P}_0$ with respect to ϵ with coefficients depending on both ϵ and k. The desired bivariate expansion of $\bar{P}_0$ follows from further expanding those coefficients with respect to k and ϵ. To expand $\bar{P}_\ell$ for $\ell \geq 1$, based on the decomposition of Λ_τ and S_τ introduced in Step 2, we apply the operator-based expansion in Aït-Sahalia (2002) to obtain the desired result in Step 3.

Step 2 – Expanding the conditional expectation $\bar{P}_\ell$ in Equation (B4) for $\ell = 0$: It follows from the dynamics in Equation (41) that the underlying asset price S_u admits the following decomposition form

$$S_u = s S_u^c S_u^J, \tag{B5}$$

where S_u^c and S_u^J are the continuous and jump components of S_u/s given by

$$S_u^c = \exp\left\{\int_0^u \left(r - d - \lambda(v_t)\bar{\mu} - \frac{1}{2}v_t^2\right)dt + \int_0^u v_t dW_{1t}\right\} \tag{B6}$$

and

$$S_u^J = \exp\left\{\sum_{i=1}^{N_u} J_{\tau_i}\right\},$$

respectively. Likewise, the Radon-Nikodým derivative Λ_u by the definition in Equation (B2) is decomposed as

$$\Lambda_u = \Lambda_u^c \Lambda_u^J, \tag{B7}$$

with the continuous component Λ_u^c and jump component Λ_u^J given by

$$\Lambda_u^c = \exp\left\{u - \int_0^u \lambda(v_t)dt\right\} \text{ and } \Lambda_u^J = \prod_{i=1}^{N_u} \lambda(v_{\tau_i}), \tag{B8}$$

respectively. Apparently, the continuous components S_u^c and Λ_u^c satisfy

$$\frac{dS_u^c}{S_u^c} = (r - d - \lambda(v_u)\bar{\mu})du + v_u dW_u,\ S_0^c = 1, \tag{B9}$$

and

$$d\Lambda_u^c = (1 - \lambda(v_u))\Lambda_u^c du,\ \Lambda_0^c = \Lambda_0 = 1, \tag{B10}$$

respectively, with the volatility v_u governed by Equation (42).

In the case of $\ell = 0$, the jump components in the decompositions in Equations (B5) and (B7) are disabled, so that the conditional expectation $\bar{P}_0$ in Equation (B4) simplifies to

$$\bar{P}_0(\tau, k, v) = e^{-r\tau}\tilde{\mathbb{E}}\left[\Lambda_\tau^c \max(e^k - S_\tau^c, 0)\right],$$

since $S_\tau = sS_\tau^c$ and $\Lambda_\tau = \Lambda_\tau^c$. By regarding $\Lambda_\tau^c \max\left(e^k - S_\tau^c, 0\right)$ as the payoff function of a derivative security with the underlying asset $(S_\tau^c, \Lambda_\tau^c)$ evolving according to the dynamics in Equations (B9), (B10), and (42), we apply the method proposed in Li (2014) and arrive at the following Jth-order univariate expansion of $\bar{P}_0(\tau, k, v)$:

$$\bar{P}_0^{(J)}(\epsilon^2, k, v) = e^{-r\epsilon^2}\epsilon v \sum_{l=0}^{J} \Phi_0^{(l)}\left(\frac{e^k - 1}{v\epsilon}\right)\epsilon^l,$$

where the coefficients $\Phi_0^{(l)}$ can be calculated in closed form. The desired bivariate expansion of $\bar{P}_0$ follows by further expanding the coefficients $\Phi_0^{(l)}$ with respect to k and ϵ.

Step 3 – Expanding the conditional expectation $\bar{P}_\ell$ in Equation (B4) for $\ell \geq 1$: Plugging in the decompositions in Equations (B5) and (B7) into Equation (B4) yields

$$\bar{P}_\ell(\tau,k,v)=\tilde{\mathbb{E}}^{(\ell)}\left[\Lambda_\tau^c \Lambda_\tau^J \max(e^k - S_\tau^c S_\tau^J, 0)\right].$$

Conditioning on Λ_τ^c, Λ_τ^J, and S_τ^c, we reformulate the above expectation as

$$\bar{P}_\ell(\tau,k,v)=\tilde{\mathbb{E}}^{(\ell)}[\Lambda_\tau^c \Lambda_\tau^J \tilde{\mathbb{E}}^{(\ell)}[\max(e^k - S_\tau^c S_\tau^J, 0)|\Lambda_\tau^c, \Lambda_\tau^J, S_\tau^c]]. \tag{B11}$$

We note that the component S_τ^J inside the inner expectation is independent with all the conditioning arguments S_τ^c, Λ_τ^c and Λ_τ^J defined in Equations (B6), (B8), and (B8), respectively, simply because jump sizes J_{τ_i} are assumed to be independent with the asset price S_u, the volatility v_u, and the Poisson process N_u for any $u \in [0,\tau]$ under the measure $\tilde{\mathbb{Q}}$. Consequently, the inner expectation in Equation (B11) can be expressed as $\phi_\ell(k, S_\tau^c)$ for some function ϕ_ℓ determined by the following integral form:

$$\phi_\ell(k,S_\tau^c)=\tilde{\mathbb{E}}^{(\ell)}[\max(e^k - S_u^c S_u^J, 0)|\Lambda_\tau^c, \Lambda_\tau^J, S_\tau^c] \tag{B12}$$

$$= \int_{\mathcal{J}^\ell} \max(e^k - S_\tau^c e^{u_1+u_2+\cdots+u_\ell}, 0)$$

$$\times f(u_1)f(u_2)\cdots f(u_\ell)du_1 du_2 \cdots du_\ell, \tag{B13}$$

where $\mathcal{J}$ and f represent the state space and the probability density function of the jump size J_t, respectively. The integral in Equation (B13) can be explicitly calculated if, for example, the jump size J_t follows a normal distribution with mean μ_J and variance σ_J^2 as commonly employed since the breakthrough invention of the jump-diffusion model by Merton (1976). Under this case, the closed-form formula of the integral in Equation (B13) is given by

$$\phi_\ell(k,S_\tau^c)=e^k \mathcal{N}\left(\frac{k-\log S_\tau^c - \ell\mu_J}{\sqrt{\ell}\sigma_J}\right) - S_\tau^c \exp\left(\ell\mu_J + \frac{\ell\sigma_J^2}{2}\right)$$

$$\times \mathcal{N}\left(\frac{k-\log S_\tau^c - \ell\mu_J}{\sqrt{\ell}\sigma_J} - \sqrt{\ell}\sigma_J\right).$$

It follows from Equations (B11) and (B12) that

$$\bar{P}_\ell(\tau,k,v)=\tilde{\mathbb{E}}^{(\ell)}[\Lambda_\tau^c \Lambda_\tau^J \phi_\ell(k,S_\tau^c)].$$

By conditioning on the components S_τ^c and Λ_τ^c, as well as the whole path of the volatility v_u for all $u \in [0,\tau]$, denoted by V for simplicity, the law of iterated expectation implies

$$\bar{P}_\ell(\tau,k,v)=\tilde{\mathbb{E}}[\Lambda_\tau^c \phi_\ell(k,S_\tau^c)\tilde{\mathbb{E}}^{(\ell)}[\Lambda_\tau^J | S_\tau^c, \Lambda_\tau^c, V]]. \tag{B14}$$

Plugging in the explicit expression of the jump component Λ_τ^J given in Equation (B8), we write the inner expectation as

$$\tilde{\mathbb{E}}^{(\ell)}[\Lambda_\tau^J | S_\tau^c, \Lambda_\tau^c, V] \equiv \tilde{\mathbb{E}}\left[\prod_{i=1}^{\ell} \lambda(v_{\tau_i}) \middle| S_\tau^c, \Lambda_\tau^c, V, N_\tau = \ell\right]. \tag{B15}$$

Given the conditions in Equation (B15), the randomness of $\prod_{i=1}^{\ell} \lambda(v_{\tau_i})$ solely hinges on those of the jump arrival times τ_i. Since N_t follows a Poisson process with constant intensity 1 independent with S_τ^c, Λ_τ^c, and V under the measure $\tilde{\mathbb{Q}}$, the conditional joint distribution of $(\tau_1, \tau_2, \cdots, \tau_\ell)$ given S_τ^c, Λ_τ^c, V, and $N_\tau = \ell$ is equivalent to that of $(\tau_1, \tau_2, \cdots, \tau_\ell)$ given $N_\tau = \ell$, which distributes as the

order statistics of ℓ independent observations sampled from the uniform distribution on $[0,\tau]$ (see, e.g., Theorem 2.3 in Chapter 4.2 of Karlin and Taylor 1975). Then, direct computation leads to

$$\tilde{\mathbb{E}}^{(\ell)}[\Lambda_\tau^J|S_\tau^c,\Lambda_\tau^c,V]=\left(\int_0^\tau \frac{1}{\tau}\lambda(v_u)du\right)^\ell=\left(1-\frac{1}{\tau}\log\Lambda_\tau^c\right)^\ell, \tag{B16}$$

where the second equality follows from the representation of Λ_τ^c in Equation (B8). Hence, by plugging Equation (B16) into Equation (B14), we simplify $\bar{P}_\ell(\tau,k,v)$ in Equation (B11) to

$$\bar{P}_\ell(\tau,k,v)=\tilde{\mathbb{E}}\left[\Lambda_\tau^c\phi_\ell(k,S_\tau^c)\left(1-\frac{1}{\tau}\log\Lambda_\tau^c\right)^\ell\right].$$

Finally, based on the dynamics of S_u^c, v_u, and Λ_u^c given in Equations (B9), (42), and (B10), respectively, an application of the operator-based expansion in Aït-Sahalia (2002) to the conditional expectation $\bar{P}_\ell(\tau,k,v)$ yields the Taylor expansion with respect $\tau=\epsilon^2$ in the form:

$$\bar{P}_\ell^{(J)}(\tau,k,v)=\sum_{l=0}^{J}\Phi_\ell^{(l)}(k,v)\tau^l,$$

for any integer order $J\geq 0$, where the expansion terms $\Phi_\ell^{(l)}$ are in closed form. The desired bivariate expansion of $\bar{P}_\ell$ follows from further expanding the coefficients $\Phi_\ell^{(l)}(k,v)$ with respect to k.

Then, given the price expansion in Equation (B1), we follow the same approach as in Appendix A to establish a set of iterations and solve the IV expansion terms $\varphi^{(i,j)}$ recursively. The specific expressions for the shape characteristics $\varphi^{(i,j)}$ depend on the specification of the jump term. With normally distributed jumps, the $(3,(2,1,0,-2))$th-order expansion of Equation (45) is given by

$$\begin{aligned}&\Sigma^{(3,(2,1,0,-2))}(\tau,k,v_t)\\&=\varphi^{(0,0)}(v_t)+\varphi^{(1,0)}(v_t)\tau^{\frac{1}{2}}+\varphi^{(2,0)}(v_t)\tau+\varphi^{(0,1)}(v_t)k+\varphi^{(1,1)}(v_t)\tau^{\frac{1}{2}}k\\&\quad+\varphi^{(-1,2)}(v_t)\tau^{-\frac{1}{2}}k^2+\varphi^{(0,2)}(v_t)k^2+\varphi^{(-2,3)}(v_t)\tau^{-1}k^3,\end{aligned} \tag{B17}$$

where $\varphi^{(0,0)}(v_t)=v_t$, and

$$\varphi^{(-1,2)}(v_t)=\frac{\lambda(v_t)\sqrt{\pi}}{2\sqrt{2}v_t^2}(-\bar{\mu}+2(\bar{\mu}+1)\mathcal{N}(\mu_+)-2\mathcal{N}(\mu_-)), \tag{B18}$$

$$\varphi^{(-2,3)}(v_t)=\frac{\lambda(v_t)\bar{\mu}}{3v_t^3}, \tag{B19}$$

$$\varphi^{(0,1)}(v_t)=\frac{1}{2v_t}[2\lambda(v_t)\bar{\mu}+\gamma(v_t)],\ \varphi^{(1,0)}(v_t)=2v_t^2\varphi^{(-1,2)}(v_t), \tag{B20}$$

$$\varphi^{(1,1)}(v_t)=\frac{\sqrt{\pi}\lambda(v_t)}{2\sqrt{2}v_t^2}[2(r-d)\bar{\mu}+2(\bar{\mu}+1)\mathcal{N}(\mu_+)(2(d-r)-v_t^2)+\bar{\mu}v_t^2+2v_t^2$$
$$-2\mathcal{N}(\mu_-)(2(d-r)+v_t^2)], \tag{B21}$$

$$\varphi^{(0,2)}(v_t)=\frac{1}{12v_t^3}[-3\gamma(v_t)^2+2v_t\gamma(v_t)\gamma'(v_t)-3\pi\lambda(v_t)^2(\bar{\mu}-2(\bar{\mu}+1)\mathcal{N}(\mu_+)$$
$$+2\mathcal{N}(\mu_-))^2+2\eta(v_t)^2+6\lambda(v_t)(\bar{\mu}(2(d-r)-\gamma(v_t))-2(\bar{\mu}+2)v_t^2)], \tag{B22}$$

$$\varphi^{(2,0)}(v_t)=\frac{1}{24v_t}[6v_t^2\gamma(v_t)+2\eta(v_t)^2+12\lambda(v_t)(\bar{\mu}(2(d-r)+\gamma(v_t))$$
$$-(\bar{\mu}+2)v_t^2)+3\gamma(v_t)^2+12(d-r)\gamma(v_t)+2v_t(6\mu(v_t)-2\gamma(v_t)$$
$$\times\left(3\bar{\mu}\lambda'(v_t)+\gamma'(v_t)\right))+12\bar{\mu}^2\lambda(v_t)^2]. \tag{B23}$$

Note that if we set the jump intensity function $\lambda(v)$ to zero, the expansion in Equation (45) reduces to the expansion in Equation (11) under the continuous SV model: under the model in Equations (1)–(2), the expansion term $\varphi^{(i,j)}(v)$ is identically zero for any negative or odd integer i and the expansion term $\varphi^{(i,j)}(v)$ coincides with $\sigma^{(i/2,j)}(v)$ for any nonnegative even integer i.

The last part of this appendix shows the calculations to link the coefficients $\varphi^{(i,j)}$ to the IV surface shape characteristics in Section 6.2. The expansion terms $\varphi^{(0,0)}(v_t)$, $\varphi^{(-2,3)}(v_t)$, $\varphi^{(0,1)}(v_t)$, $\varphi^{(1,1)}(v_t)$, and $\varphi^{(-1,2)}(v_t)$ satisfy

$$\varphi^{(0,0)}(v_t)=\lim_{\tau\to 0}\Sigma(\tau,0,v_t), \tag{B24}$$

$$\varphi^{(0,1)}(v_t)=\lim_{\tau\to 0}\frac{\partial\Sigma}{\partial k}(\tau,0,v_t), \tag{B25}$$

$$\varphi^{(1,1)}(v_t)=\lim_{\tau\to 0}2\sqrt{\tau}\frac{\partial^2\Sigma}{\partial\tau\partial k}(\tau,0,v_t), \tag{B26}$$

$$\varphi^{(-1,2)}(v_t)=\lim_{\tau\to 0}\frac{\sqrt{\tau}}{2}\frac{\partial^2\Sigma}{\partial k^2}(\tau,0,v_t), \tag{B27}$$

$$\varphi^{(-2,3)}(v_t)=\lim_{\tau\to 0}\frac{\tau}{6}\frac{\partial^3\Sigma}{\partial k^3}(\tau,0,v_t), \tag{B28}$$

while the expansion terms $\varphi^{(0,2)}(v_t)$ and $\varphi^{(2,0)}(v_t)$ satisfy

$$\varphi^{(0,2)}(v_t)=\lim_{\tau\to 0}\left(\frac{1}{2}\frac{\partial^2\Sigma}{\partial k^2}(\tau,0,v_t)+\tau\frac{\partial^3\Sigma}{\partial k^2\partial\tau}(\tau,0,v_t)\right), \tag{B29}$$

and

$$\varphi^{(2,0)}(v_t)=\lim_{\tau\to 0}\left(\frac{\partial\Sigma}{\partial\tau}(\tau,0,v_t)+2\tau\frac{\partial^2\Sigma}{\partial\tau^2}(\tau,0,v_t)\right). \tag{B30}$$

This is shown as follows. Setting $k=0$ in the bivariate expansion in Equation (45) implies that

$$\Sigma^{(L_0)}(\tau,0,v_t)=\sum_{i=0}^{L_0}\varphi^{(i,0)}(v_t)\tau^{\frac{i}{2}}. \tag{B31}$$

Differentiating both sides of Equation (45) with respect to k once, twice, or thrice, and then taking k to be zero, we obtain

$$\frac{\partial}{\partial k}\Sigma^{(L_1)}(\tau,0,v_t)=\sum_{i=0}^{L_1}\varphi^{(i,1)}(v_t)\tau^{\frac{i}{2}}, \tag{B32}$$

$$\frac{\partial^2}{\partial k^2}\Sigma^{(L_2)}(\tau,0,v_t)=\sum_{i=-1}^{L_2}2\varphi^{(i,2)}(v_t)\tau^{\frac{i}{2}}, \tag{B33}$$

and

$$\frac{\partial^3}{\partial k^3}\Sigma^{(L_3)}(\tau,0,v_t)=\sum_{i=-2}^{L_3}6\varphi^{(i,3)}(v_t)\tau^{\frac{i}{2}}.$$

Equation (B31) (resp. Equation (B32)) implies Equation (B24) (resp. Equation (B25)) as τ approaches zero, that is, $\varphi^{(0,0)}(v_t)=\lim_{\tau\to 0}\Sigma(\tau,0,v_t)$ (resp. that is, $\varphi^{(0,1)}(v_t)=\lim_{\tau\to 0}\partial\Sigma(\tau,0,v_t)/\partial k$). The rest of the formulae listed in Equations (B24)–(B30) hinge on finding the univariate Taylor expansions with respect to $\sqrt{\tau}$ of the time-scaled shape characteristics or their combinations appearing on the right-hand sides of these equations. Consider Equation (B29). It follows from Equation (B33) that

$$\frac{1}{2}\frac{\partial^2}{\partial k^2}\Sigma^{(L_2)}(\tau,0,v_t)=\sum_{i=-1}^{L_2}\varphi^{(i,2)}(v_t)\tau^{\frac{i}{2}},$$

and

$$\tau\frac{\partial^3}{\partial k^2\partial\tau}\Sigma^{(L_2)}(\tau,0,v_t)=\sum_{i=-1}^{L_2}i\varphi^{(i,2)}(v_t)\tau^{\frac{i}{2}}.$$

Adding the above two equations yields

$$\frac{1}{2}\frac{\partial^2}{\partial k^2}\Sigma^{(L_2)}(\tau,0,v_t)+\tau\frac{\partial^3}{\partial k^2\partial\tau}\Sigma^{(L_2)}(\tau,0,v_t)=\sum_{i=0}^{L_2}(i+1)\varphi^{(i,2)}(v_t)\tau^{\frac{i}{2}},$$

which is a Taylor expansion with leading term $\varphi^{(0,2)}(v_t)$, and Equation (B29) follows by letting τ approach zero.

Appendix B.1 From implied volatility to stochastic volatility and jumps

The following result generalizes Theorem 1 to the case where jumps are present. The closed-form formulae in Equations (B18)–(B23) and $\varphi^{(0,0)}(v_t)=v_t$ form a system of equations to be used for solving the coefficient functions $\mu(\cdot)$, $\gamma(\cdot)$, $\eta(\cdot)$, and $\lambda(\cdot)$, as well as the jump size parameters μ_J and σ_J, given the generalized shape characteristics $\varphi^{(0,0)}$, $\varphi^{(-1,2)}$, $\varphi^{(-2,3)}$, $\varphi^{(0,1)}$, $\varphi^{(1,1)}$, $\varphi^{(0,2)}$, and $\varphi^{(2,0)}$ as inputs. As in the continuous case, this system can be inverted in closed form, with all the model coefficient functions recovered explicitly in terms of the $\varphi^{(i,j)}$ shape characteristics:

Theorem B.2. The jump size parameters μ_J and σ_J of the model in Equations (41)–(42) can be recovered by the following coupled algebraic equations

$$\frac{\bar{\mu}+2-2(\bar{\mu}+1)\mathcal{N}(\mu_+)-2\mathcal{N}(\mu_-)}{-\bar{\mu}+2(\bar{\mu}+1)\mathcal{N}(\mu_+)-2\mathcal{N}(\mu_-)}=\frac{1}{\varphi^{(0,0)}(v_t)^2}\left[\frac{\varphi^{(1,1)}(v_t)}{\varphi^{(-1,2)}(v_t)}+2(r-d)\right] \tag{B34}$$

and

$$\frac{2\sqrt{2}\bar{\mu}}{3\sqrt{\pi}[-\bar{\mu}+2(\bar{\mu}+1)\mathcal{N}(\mu_+)-2\mathcal{N}(\mu_-)]}=\frac{\varphi^{(0,0)}(v_t)\varphi^{(-2,3)}(v_t)}{\varphi^{(-1,2)}(v_t)}. \tag{B35}$$

The coefficient functions $\lambda(\cdot)$, $\gamma(\cdot)$, $\eta(\cdot)$, and $\mu(\cdot)$ can be recovered in closed form as

$$\lambda(v_t)=\frac{2\sqrt{2}\varphi^{(0,0)}(v_t)^2\varphi^{(-1,2)}(v_t)}{\sqrt{\pi}[-\bar{\mu}+2(\bar{\mu}+1)\mathcal{N}(\mu_+)-2\mathcal{N}(\mu_-)]}, \tag{B36}$$

$$\gamma(v_t)=2\varphi^{(0,0)}(v_t)\varphi^{(0,1)}(v_t)-2\lambda(v_t)\bar{\mu}, \tag{B37}$$

and

$$\begin{aligned}\eta(v_t)=\Bigg[&6\varphi^{(0,0)}(v_t)^3\varphi^{(0,2)}(v_t)-\varphi^{(0,0)}(v_t)\gamma(v_t)\gamma'(v_t)+\frac{3}{2}\pi\lambda(v_t)^2(\bar{\mu}\\&+2\mathcal{N}(\mu_-)-2(\bar{\mu}+1)\mathcal{N}(\mu_+))^2+\frac{3}{2}\gamma(v_t)^2+3\lambda(v_t)(2\bar{\mu}(r-d)\\&+(\bar{\mu}+2)\varphi^{(0,0)}(v_t)^2+\bar{\mu}\gamma(v_t))\Bigg]^{\frac{1}{2}},\end{aligned} \tag{B38}$$

as well as

$$\begin{aligned}\mu(v_t)=\frac{1}{12\varphi^{(0,0)}(v_t)}&[24\varphi^{(0,0)}(v_t)\varphi^{(2,0)}(v_t)-6\varphi^{(0,0)}(v_t)^2\gamma(v_t)-2\eta(v_t)^2\\&-12\lambda(v_t)(\bar{\mu}(2(d-r)+\gamma(v_t))-(\bar{\mu}+2)\varphi^{(0,0)}(v_t)^2)-12(d-r)\gamma(v_t)\\&-3\gamma(v_t)^2-12\bar{\mu}^2\lambda(v_t)^2+4\varphi^{(0,0)}(v_t)\gamma(v_t)\left(3\bar{\mu}\lambda'(v_t)+\gamma'(v_t)\right)].\end{aligned} \tag{B39}$$

Equations (B34)–(B39) constitute a complete mapping from the expansion terms $\varphi^{(i,j)}(v_t)$, that is, the generalized shape characteristics of the IV surface, to the specification of the SV model in Equations (41)–(42).

To construct a parametric ISV model, based on the closed-form formulae for the generalized shape characteristics $\varphi^{(i,j)}$, we can then use the moment conditions

$$\mathbb{E}[g^{(i,j)}(v_{l\Delta};\theta_0)]=0, \text{ with } g^{(i,j)}(v_{l\Delta};\theta)=[\varphi^{(i,j)}]_l^{\text{data}}-[\varphi^{(i,j)}(v_{l\Delta};\theta)]^{\text{model}},$$

where $[\varphi^{(i,j)}]_l^{\text{data}}$ denotes the data of $\varphi^{(i,j)}(v_{l\Delta})$. Then apply the same GMM estimation approach proposed in Section 3 to estimate the parameters θ.

To construct a nonparametric model, we can in principle estimate the jump size parameters μ_J and σ_J before estimating the coefficient functions $\lambda(\cdot)$, $\mu(\cdot)$, $\gamma(\cdot)$, and $\eta(\cdot)$ as discussed. Indeed, the estimators of μ_J and σ_J can be obtained by the two (exactly identified) conditions as the sample analogs of algebraic equations (B34) and (B35)

$$\begin{aligned}&\frac{\bar{\mu}+2-2(\bar{\mu}+1)\mathcal{N}(\mu_+)-2\mathcal{N}(\mu_-)}{-\bar{\mu}+2(\bar{\mu}+1)\mathcal{N}(\mu_+)-2\mathcal{N}(\mu_-)}\\&=\frac{1}{n}\sum_{l=1}^{n}\frac{1}{\left([\varphi^{(0,0)}]_l^{\text{data}}\right)^2}\left(\frac{[\varphi^{(1,1)}]_l^{\text{data}}}{[\varphi^{(-1,2)}]_l^{\text{data}}}+2(r-d)\right),\end{aligned}$$

and

$$\frac{2\sqrt{2}\bar{\mu}}{3\sqrt{\pi}[-\bar{\mu}+2(\bar{\mu}+1)\mathcal{N}(\mu_+)-2\mathcal{N}(\mu_-)]}=\frac{1}{n}\sum_{l=1}^{n}\frac{[\varphi^{(0,0)}]_l^{\text{data}}[\varphi^{(-2,3)}]_l^{\text{data}}}{[\varphi^{(-1,2)}]_l^{\text{data}}}.$$

Then, regarding the estimators of jump size parameters as inputs, Equations (B36)–(B39) allow us to estimate coefficient functions $\lambda(\cdot)$, $\gamma(\cdot)$, $\eta(\cdot)$, and $\mu(\cdot)$ one after another iteratively, by following a similar approach proposed in Section 2 for constructing a nonparametric ISV model without jumps.

Appendix B.2 Example: The Merton jump-diffusion model

We now illustrate our expansion formulae in Equation (B17) in a special case, the jump-diffusion model of Merton (1976):

$$\frac{dS_t}{S_{t-}} = (r - d - \lambda\bar{\mu})dt + v_0 dW_t + (\exp(J_t) - 1)dN_t, \tag{B40}$$

where λ represents a constant jump intensity and v_0 a constant volatility. We obtain the expansion in Equation (B17) under this model simply by letting the SV components be zero and let the jump intensity function be the constant λ, that is,

$$v_t = v_0 \text{ and } \lambda(v_t) = \lambda. \tag{B41}$$

The expansion terms $\varphi^{(0,1)}(v_t)$, $\varphi^{(-1,2)}(v_t)$, and $\varphi^{(1,1)}(v_t)$ reduce to

$$\varphi^{(0,1)}(v_0) = \frac{\lambda\bar{\mu}}{v_0},\ \varphi^{(-1,2)}(v_0) = \frac{\lambda\sqrt{\pi}}{2\sqrt{2}v_0^2}(-\bar{\mu} + 2(\bar{\mu}+1)\mathcal{N}(\mu_+) - 2\mathcal{N}(\mu_-)),$$

and

$$\varphi^{(1,1)}(v_0) = \frac{\sqrt{\pi}\lambda}{2\sqrt{2}v_0^2}[2(r-d)\bar{\mu} + 2(\bar{\mu}+1)\mathcal{N}(\mu_+)(2(d-r) - v_0^2) + \bar{\mu}v_0^2 + 2v_0^2$$
$$-2\mathcal{N}(\mu_-)(2(d-r) + v_0^2)],$$

from the general formulae provided in Equations (B20), (B18), and (B21), respectively.

Equations (46), (47), and (48) (equivalently, Equations (B34)–(B36) in Theorem B.2) apply to the jump-diffusion model of Merton (1976), by plugging in the specification assumptions in Equation (B41). One is able to identify all the model components, that is, the constant volatility v_0, intensity λ, as well as jump size parameters μ_J and σ_J. Combining the following equations

$$\frac{\lambda\bar{\mu}}{v_0} = \lim_{\tau\to 0}\frac{\partial\Sigma}{\partial k}(\tau, 0, v_t),$$

$$\frac{\lambda\sqrt{\pi}}{2\sqrt{2}v_0^2}(-\bar{\mu} + 2(\bar{\mu}+1)\mathcal{N}(\mu_+) - 2\mathcal{N}(\mu_-)) = \lim_{\tau\to 0}\frac{\sqrt{\tau}}{2}\frac{\partial^2\Sigma}{\partial k^2}(\tau, 0, v_t),$$

$$\frac{\sqrt{\pi}\lambda}{2\sqrt{2}v_0^2}[2(r-d)\bar{\mu} + (\bar{\mu}+2)v_0^2 + 2(\bar{\mu}+1)\mathcal{N}(\mu_+)(2(d-r) - v_0^2) - 2\mathcal{N}(\mu_-)(2(d-r) + v_0^2)]$$

$$= \lim_{\tau\to 0} 2\sqrt{\tau}\frac{\partial^2\Sigma}{\partial\tau\partial k}(\tau, 0, v_t),$$

with Equation (B24), that is, $v_0 = \lim_{\tau\to 0}\Sigma(\tau, 0, v_0)$, we can identify the parameters of the Merton model v_0, λ, μ_J, and σ_J, given observations on the following four observable short-maturity IV shape characteristics – at-the-money level Σ, log-moneyness slope $\partial\Sigma/\partial k$ and convexity $\partial^2\Sigma/\partial k^2$, as well as the mixed slope $\partial^2\Sigma/\partial\tau\partial k$, all evaluated at $(\tau, 0, v_0)$. If employing Equation (47) instead, the much less easily observable third-order shape characteristic $\partial^3\Sigma/\partial k^3$ would become necessary.

Appendix C. Parametric ISV Model: GMM Asymptotics

The asymptotic distribution of the GMM estimator follows standard theory. The asymptotic variance matrix $V(\theta)$ in Equation (29) satisfies

$$V(\theta) = \left[G(\theta)^\intercal \Omega^{-1}(\theta)G(\theta)\right]^{-1},$$

with

$$G(\theta)=\mathbb{E}\left[\frac{\partial g(v_{l\Delta};\theta)}{\partial\theta}\right],\ \Omega(\theta)=\Omega_0(\theta)+\sum_{j=1}^{n-1}\left(\Omega_j(\theta)+\Omega_j(\theta)^{\intercal}\right),$$

and

$$\Omega_j(\theta)=\mathbb{E}[g(v_{l\Delta};\theta)g(v_{(l+j)\Delta};\theta)^{\intercal}],\ \text{for } j=0,1,2,\ldots,n-1.$$

A consistent estimator of the matrix $V(\theta_0)$ is given by $\hat{V}(\hat{\theta})$, where

$$\hat{V}(\theta)=\left[\hat{G}(\theta)^{\intercal}\hat{\Omega}^{-1}(\theta)\hat{G}(\theta)\right]^{-1},\ \text{with } \hat{G}(\theta)=\frac{1}{n}\sum_{l=1}^{n}\frac{\partial g(v_{l\Delta},\theta)}{\partial\theta}. \tag{C1}$$

The matrix $\hat{\Omega}(\theta)$ is the Newey-West estimator with ℓ lags:

$$\hat{\Omega}(\theta)=\hat{\Omega}_0(\theta)+\sum_{j=1}^{\ell}\left(\frac{\ell+1-j}{\ell+1}\right)\left(\hat{\Omega}_j(\theta)+\hat{\Omega}_j(\theta)^{\intercal}\right), \tag{C2}$$

where

$$\hat{\Omega}_0(\theta)=\frac{1}{n}\sum_{l=1}^{n}g(v_{l\Delta};\theta)g(v_{l\Delta};\theta)^{\intercal}\ \text{and}\ \hat{\Omega}_j(\theta)=\frac{1}{n}\sum_{l=j+1}^{n}g(v_{l\Delta};\theta)g(v_{(l-j)\Delta};\theta)^{\intercal},$$

for $j=1,2,\ldots,\ell$. In principle, the number of lags ℓ should grow with n at the rate $\ell=\mathcal{O}(n^{1/3})$. In the overidentified case, the optimal choice of W_n ought to be a consistent estimator of $\hat{\Omega}^{-1}(\theta_0)$. For this, the estimator $\hat{\theta}$ is obtained by the following two steps: First, set the initial weight matrix W_n in Equation (28) as the identity matrix and arrive at a consistent estimator $\tilde{\theta}$. Second, compute $\hat{\Omega}(\tilde{\theta})$ according to Equation (C2), so that its inverse $\hat{\Omega}^{-1}(\tilde{\theta})$ is a consistent estimator of $\Omega^{-1}(\theta_0)$. Then set the weight matrix W_n in Equation (28) as $\hat{\Omega}^{-1}(\tilde{\theta})$ and update the estimator to $\hat{\theta}$.

References

Aït-Sahalia, Y. 1996. Testing continuous-time models of the spot interest rate. *Review of Financial Studies* 9:385–426.

———. 2002. Maximum-likelihood estimation of discretely-sampled diffusions: A closed-form approximation approach. *Econometrica* 70:223–62.

Aït-Sahalia, Y., J. Fan, and Y. Li. 2013. The leverage effect puzzle: Disentangling sources of bias at high frequency. *Journal of Financial Economics* 109:224–49.

Aït-Sahalia, Y., and R. Kimmel. 2007. Maximum likelihood estimation of stochastic volatility models. *Journal of Financial Economics* 83:413–52.

Aït-Sahalia, Y., and P. A. Mykland. 2003. The effects of random and discrete sampling when estimating continuous-time diffusions. *Econometrica* 71:483–549.

Andersen, L., and J. Andreasen. 2000. Jump-diffusion processes: Volatility smile fitting and numerical methods for option pricing. *Review of Derivatives Research* 4:231–62.

Andersen, L., and A. Lipton. 2013. Asymptotics for exponential Lévy processes and their volatility smile: Survey and new results. *International Journal of Theoretical and Applied Finance* 16:1–98.

Bakshi, G., P. Carr, and L. Wu. 2008. Stochastic risk premiums, stochastic skewness in currency options, and stochastic discount factors in international economies. *Journal of Financial Economics* 87:132–56.

Bandi, F. M., and P. C. B. Phillips. 2003. Fully nonparametric estimation of scalar diffusion models. *Econometrica* 71:241–83.

Bandi, F. M., and R. Renò. 2018. Nonparametric stochastic volatility. *Econometric Theory* 34:1207–55.

Bates, D. S. 1996. Jumps and stochastic volatility: Exchange rate processes implicit in Deutsche Mark options. *Review of Financial Studies* 9:69–107.

———. 2000. Post-'87 crash fears in the S&P 500 futures option market. *Journal of Econometrics* 94:181–238.

Berestycki, H., J. Busca, and I. Florent. 2002. Asymptotics and calibration of local volatility models. *Quantitative Finance* 2:61–9.

———. 2004. Computing the implied volatility in stochastic volatility models. *Communications on Pure and Applied Mathematics* 57:1352–73.

Brémaud, P. 1981. *Point processes and queues: Martingale dynamics*. New York: Springer-Verlag.

Broadie, M., M. Chernov, and M. Johannes. 2007. Model specification and risk premia: Evidence from futures options. *Journal of Finance* 62:1453–90.

Carr, P., and L. Cousot. 2011. A PDE approach to jump-diffusions. *Quantitative Finance* 11:33–52.

———. 2012. Explicit constructions of martingales calibrated to given implied volatility smiles. *SIAM Journal on Financial Mathematics* 3:182–214.

Carr, P., H. Geman, D. B. Madan, and M. Yor. 2004. From local volatility to local Lévy models. *Quantitative Finance* 4:581–8.

Carr, P., and D. B. Madan. 1998. Option valuation using the fast Fourier transform. *Journal of Computational Finance* 2:61–73.

Carr, P., and L. Wu. 2003. What type of process underlies options? A simple robust test. *Journal of Finance* 58:2581–610.

———. 2007. Stochastic skew for currency options. *Journal of Financial Economics* 86:213–47.

Chernov, M., A. R. Gallant, E. Ghysels, and G. T. Tauchen. 2003. Alternative models for stock price dynamics. *Journal of Econometrics* 116:225–57.

Christoffersen, P., S. Heston, and K. Jacobs. 2009. The shape and term structure of the index option smirk: Why multifactor stochastic volatility models work so well. *Management Science* 55:1914–32.

Christoffersen, P., K. Jacobs, and K. Mimouni. 2010. Volatility dynamics for the S&P500: Evidence from realized volatility, daily returns, and option prices. *Review of Financial Studies* 23:3141–89.

Comte, F., V. Genon-Catalot, and Y. Rozenholc. 2010. Nonparametric estimation for a stochastic volatility model. *Finance and Stochastics* 14:49–80.

Duffie, D., J. Pan, and K. J. Singleton. 2000. Transform analysis and asset pricing for affine jump-diffusions. *Econometrica* 68:1343–76.

Dumas, B., J. Fleming, and R. E. Whaley. 1998. Implied volatility functions: Empirical tests. *Journal of Finance* 53:2059–106.

Dupire, B. 1994. Pricing with a smile. *RISK* 7:18–20.

Durrleman, V. 2008. Convergence of at-the-money implied volatilities to the spot volatility. *Journal of Applied Probability* 45:542–50.

———. 2010. From implied to spot volatilities. *Finance and Stochastics* 14:157–77.

Durrleman, V., and N. El Karoui. 2008. Coupling smiles. *Quantitative Finance* 8:573–90.

Eraker, B., M. S. Johannes, and N. Polson. 2003. The impact of jumps in equity index volatility and returns. *Journal of Finance* 58:1269–300.

Fan, J., and I. Gijbels. 1996. *Local polynomial modelling and its applications*. London: Chapman & Hall.

Forde, M., and A. Jacquier. 2011. The large-maturity smile for the Heston model. *Finance and Stochastics* 17:755–80.

Forde, M., A. Jacquier, and R. Lee. 2012. The small-time smile and term structure of implied volatility under the Heston model. *SIAM Journal on Financial Mathematics* 3:690–708.

Fouque, J.-P., M. Lorig, and R. Sircar. 2016. Second order multiscale stochastic volatility asymptotics: Stochastic terminal layer analysis and calibration. *Finance and Stochastics* 63:1648–65.

Gao, K., and R. Lee. 2014. Asymptotics of implied volatility to arbitrary order. *Finance and Stochastics* 18:349–92.

Gatheral, J. 2006. *The volatility surface: A pactitioner's guide*. Hoboken, NJ: John Wiley and Sons.

Gatheral, J., E. P. Hsu, P. Laurence, C. Ouyang, and T.-H. Wang. 2012. Asymptotics of implied volatility in local volatility models. *Mathematical Finance* 22:591–620.

Hagan, P. S., and D. E. Woodward. 1999. Equivalent Black volatilities. *Applied Mathematical Finance* 6:147–57.

Hansen, L. P. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* 50:1029–54.

Heston, S. 1993. A closed-form solution for options with stochastic volatility with applications to bonds and currency options. *Review of Financial Studies* 6:327–43.

Huang, J., and L. Wu. 2004. Specification analysis of option pricing models based on time-changed Lévy processes. *Journal of Finance* 59:1405–40.

Hull, J., and A. White. 1987. The pricing of options on assets with stochastic volatilities. *Journal of Finance* 42:281–300.

Jacquier, A., and M. Lorig. 2015. From characteristic functions to implied volatility expansions. *Advances in Applied Probability* 47:837–57.

Jiang, G. J., and J. Knight. 1997. A nonparametric approach to the estimation of diffusion processes – with an application to a short-term interest rate model. *Econometric Theory* 13:615–45.

Jones, C. S. 2003. The dynamics of stochastic volatility: Evidence from underlying and options markets. *Journal of Econometrics* 116:181–224.

Kanaya, S., and D. Kristensen. 2016. Estimation of stochastic volatility models by nonparametric filtering. *Econometric Theory* 32:861–916.

Karlin, S., and H. M. Taylor. 1975. *A first course in stochastic processes*. Second ed. Cambridge: Academic Press.

Kristensen, D., and A. Mele. 2011. Adding and subtracting Black-Scholes: A new approach to approximating derivative prices in continuous-time models. *Journal of Financial Economics* 102:390–415.

Kunitomo, N., and A. Takahashi. 2001. The asymptotic expansion approach to the valuation of interest rate contingent claims. *Mathematical Finance* 11:117–51.

Ledoit, O., P. Santa-Clara, and S. Yan. 2002. Relative pricing of options with stochastic volatility. Working Paper, University of California at Los Angeles.

Lee, R. 2001. Implied and local volatilities under stochastic volatility. *International Journal of Theoretical and Applied Finance* 4:45–89.

———. 2004. The moment formula for implied volatility at extreme strikes. *Mathematical Finance* 14:469–80.

Li, C. 2014. Closed-form expansion, conditional expectation, and option valuation. *Mathematics of Operations Research* 39:487–516.

Lorig, M., S. Pagliarani, and A. Pascucci. 2017. Explicit implied volatilities for multifactor local-stochastic volatility models. *Mathematical Finance* 27:927–60.

Medvedev, A., and O. Scaillet. 2007. Approximation and calibration of short-term implied volatilities under jump-diffusion stochastic volatility. *Review of Financial Studies* 20:427–59.

Merton, R. C. 1976. Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics* 3:125–44.

Newey, W. K. 1985. Generalized method of moments specification testing. *Journal of Econometrics* 29:229–56.

Pagliarani, S., and A. Pascucci. 2017. The exact Taylor formula of the implied volatility. *Finance and Stochastics* 21:661–718.

Pan, J. 2002. The jump-risk premia implicit in options: Evidence from an integrated time-series study. *Journal of Financial Economics* 63:3–50.

Renò, R. 2008. Nonparametric estimation of the diffusion coefficient of stochastic volatility models. *Econometric Theory* 24:1174–206.

Sircar, K. R., and G. C. Papanicolaou. 1999. Stochastic volatility, smile & asymptotics. *Applied Mathematical Finance* 6:107–45.

Takahashi, A., and T. Yamada. 2012. An asymptotic expansion with push-down of Malliavin weights. *SIAM Journal on Financial Mathematics* 3:95–136.

Tehranchi, M. R. 2009. Asymptotics of implied volatility far from maturity. *Journal of Applied Probability* 46:629–50.

Xiu, D. 2014. Hermite polynomial based expansion of European option prices. *Journal of Econometrics* 179:158–77.