Estimating State-Price Densities with Nonparametric Regression

Kim Huynh,[1] Pierre Kervella and Jun Zheng

1 Introduction

Derivative markets offer a rich source of information to extract the market's expectations of the future price of an asset. Using option prices, one may derive the whole risk-neutral probability distribution of the underlying asset price at the maturity date of the options. Once this distribution also called State-Price Density (SPD) is estimated, it may serve for pricing new, complex or illiquid derivative securities.

There exist numerous methods to recover the SPD empirically. They can be separated in two classes:

- methods using option prices as identifying conditions
- methods using the second derivative of the call pricing function with respect to K

The first class includes methods which consist in estimating the parameters of a mixture of log-normal densities to match the observed option prices, Melick and Thomas (1997). Another popular approach in this class is the implied binomial trees method, see Rubinstein (1994), Derman and Kani (1994) and Härdle and Zheng (2002). Another technique is based on learning networks suggested by Hutchinson, Lo and Poggio (1994), a nonparametric approach using artificial neural networks, radial basis functions, and projection pursuits.

The second class of methods is based on the result of Breeden and Litzenberger (1978). This methodology is based on European options with identical time to maturity, it may therefore be applied to fewer cases than some of the techniques in the first class. Moreover, it also assumes a continuum of strike prices on $\mathbb{R}^+$ which can not be found on any stock exchange. Indeed, the strike prices are always discretely spaced on a finite range around the actual underlying price. Hence, to handle this problem an interpolation of the call pricing function inside the range and extrapolation outside may be performed. In the following, a semiparametric technique using nonparametric regression of the implied volatility surface will be introduced to provide this interpolation task. A new approach using constrained least squares has been suggested by Yatchew and Härdle (2002) but will not be explored here.

The concept of Arrow-Debreu securities is the building block for the analysis of economic equilibrium under uncertainty. Rubinstein (1976) and Lucas (1978) used this concept as a basis to construct dynamic general equilibrium models in order to determine the price of assets in an economy. The central idea of this methodology is that the price of a financial security is equal to the expected net present value of its future payoffs under the risk-neutral probability density function (PDF). The net present value is calculated using the risk-free interest rate, while the expectation is taken with respect to the weighted-marginal-rate-of-substitution PDF of the payoffs. The latter term is known as the state-price

[1] Queen's University

density (SPD), risk-neutral PDF, or equivalent martingale measure. The price of a security at time t (P_t) with a single liquidation date T and payoff $Z(S_T)$ is then:

$$P_t = e^{-r_{t,\tau}\tau}\mathrm{E}_t^*[Z(S_T)] = e^{-r_{t,\tau}\tau}\int_{-\infty}^{\infty} Z(S_T)f_t^*(S_T)dS_T \tag{1}$$

where E_t^* is the conditional expectation given the information set in t under the equivalent martingale probability, S_T is the state variable, $r_{t,\tau}$ is the risk-free rate at time t with time to maturity τ, and $f_t^*(S_T)$ is the SPD at time t for date T payoffs.

Rubinstein (1985) shows that if one has two of the three following pieces of information:

- representative agent's preferences
- asset price dynamics or its data-generating process
- SPD

then one can recover the third. Since the agent's preferences and the true data-generating process are unknown, a no-arbitrage approach is used to recover the SPD.

2 Extracting the SPD using Call-Options

Breeden and Litzenberger (1978) show that one can replicate Arrow-Debreu prices using the concept of *butterfly spread* on European call options. This spread entails selling two call options at exercise price K, buying one call option at $K^- = K - \Delta K$ and another at $K^+ = K + \Delta K$, where ΔK is the stepsize between the adjacent call strikes. These four options constitute a butterfly spread centered on K. If the terminal underlying asset value S_T is equal to K then the payoff $Z(\cdot)$ of $\frac{1}{\Delta K}$ of such butterfly spreads is defined as:

$$Z(S_T, K; \Delta K) = P(S_{T-\tau}, \tau, K; \Delta K)|_{\tau=0} = \left.\frac{u_1 - u_2}{\Delta K}\right|_{S_T=K,\tau=0} = 1 \tag{2}$$

where

$$u_1 = C(S_{T-\tau}, \tau, K + \Delta K) - C(S_{T-\tau}, \tau, K),$$
$$u_2 = C(S_{T-\tau}, \tau, K) - C(S_{T-\tau}, \tau, K - \Delta K).$$

$C(S, \tau, K)$ denotes the price of a European call with an actual underlying price S, a time to maturity τ and a strike price K. Here, $P(S_{T-\tau}, \tau, K; \Delta K)$ is the corresponding price of this security ($\frac{1}{\Delta K} * butterfly\ spread(K; \Delta K)$) at time $T - \tau$.

As ΔK tends to zero, this security becomes an Arrow-Debreu security paying 1 if $S_T = K$ and zero in other states. As it is assumed that S_T has a continuous distribution function on $\mathbb{R}^+$, the probability of any given level of S_T is zero and thus, in this case, the price of an Arrow-Debreu security is zero. However, dividing one more time by ΔK, one obtains the price of ($\frac{1}{(\Delta K)^2} * butterfly$

spread$(K;\Delta K)$) and as ΔK tends to 0 this price tends to $f^*(S_T)e^{-r_{t,\tau}}$ for $S_T = K$. Indeed,

$$\lim_{\Delta K \to 0} \left(\frac{P(S_t, \tau, K; \Delta K)}{\Delta K} \right) \Bigg|_{K=S_T} = f^*(S_T)e^{-r_{t,\tau}}. \tag{3}$$

This can be proved by setting the payoff Z_1 of this new security

$$Z_1(S_T) = \left(\frac{1}{(\Delta K)^2} (\Delta K - |S_T - K|) \mathbf{1}(S_T \in [K - \Delta K, K + \Delta K]) \right)$$

in (1) and letting ΔK tend to 0. Indeed, one should remark that:

$$\forall (\Delta K) : \int_{K-\Delta K}^{K+\Delta K} (\Delta K - |S_T - K|) dS_T = (\Delta K)^2.$$

If one can construct these financial instruments on a continuum of states (strike prices) then at infinitely small ΔK a complete state pricing function can be defined.

Moreover, as ΔK tends to zero, this price will tend to the second derivative of the call pricing function with respect to the strike price evaluated at K:

$$\begin{aligned} \lim_{\Delta K \to 0} \left(\frac{P(S_t, \tau, K; \Delta K)}{\Delta K} \right) &= \lim_{\Delta K \to 0} \frac{u_1 - u_2}{(\Delta K)^2} \\ &= \frac{\partial^2 C_t(\cdot)}{\partial K^2}. \end{aligned} \tag{4}$$

Equating (3) and (4) across all states yields:

$$\frac{\partial^2 C_t(\cdot)}{\partial K^2} \Bigg|_{K=S_T} = e^{-r_{t,\tau}\tau} f_t^*(S_T)$$

where $r_{t,\tau}$ denotes the risk-free interest rate at time t with time to maturity τ and $f_t^*(\cdot)$ denotes the risk-neutral PDF or the SPD in t. Therefore, the SPD is defined as:

$$f_t^*(S_T) = e^{r_{t,\tau}\tau} \frac{\partial^2 C_t(\cdot)}{\partial K^2} \Bigg|_{K=S_T}. \tag{5}$$

This method constitutes a no-arbitrage approach to recover the SPD. No assumption on the underlying asset dynamics are required. Preferences are not restricted since the no-arbitrage method only assumes risk-neutrality with respect to the underlying asset. The only requirements for this method are that markets are perfect (i.e. no sales restrictions, transactions costs or taxes and that agents are able to borrow at the risk-free interest rate) and that $C(\cdot)$ is twice differentiable. The same result can be obtained by differentiating (1) twice with respect to K after setting for Z the call payoff function $Z(S_T) = (S_T - K)^+$.

2.1 Black-Scholes SPD

The Black-Scholes call option pricing formula is due to Black and Scholes (1973) and Merton (1973). In this model there are no assumptions regarding preferences, rather it relies on no-arbitrage conditions and assumes that the evolution

of the underlying asset price S_t follows a geometric Brownian motion defined through

$$\frac{dS_t}{S_t} = \mu dt + \sigma dW_t. \tag{6}$$

Here μ denotes the drift and σ the volatility assumed to be constant.

The analytical formula for the price in t of a call option with a terminal date $T = t + \tau$, a strike price K, an underlying price S_t, a risk-free rate $r_{t,\tau}$, a continuous dividend yield $\delta_{t,\tau}$, and a volatility σ, is:

$$\begin{aligned} C_{BS}(S_t, K, \tau, r_{t,\tau}, \delta_{t,\tau}; \sigma) &= e^{-r_{t,\tau}} \int_0^\infty \max(S_T - K, 0) f^*_{BS,t}(S_T) dS_T \\ &= S_t e^{-\delta_{t,\tau}\tau} \Phi(d_1) - K e^{-r_{t,\tau}\tau} \Phi(d_2) \end{aligned}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function and

$$\begin{aligned} d_1 &= \frac{\log(S_t/K) + (r_{t,\tau} - \delta_{t,\tau} + \frac{1}{2}\sigma^2)\tau}{\sigma\sqrt{\tau}}, \\ d_2 &= d_1 - \sigma\sqrt{\tau}. \end{aligned}$$

As a consequence of the assumptions on the underlying asset price process the Black-Scholes SPD is a log-normal density with mean $(r_{t,\tau} - \delta_{t,\tau} - \frac{1}{2}\sigma^2)\tau$ and variance $\sigma^2\tau$ for $\log(S_T/S_t)$:

$$\begin{aligned} f^*_{BS,t}(S_T) &= e^{r_{t,\tau}\tau} \left. \frac{\partial^2 C_t}{\partial K^2} \right|_{K=S_T} \\ &= \frac{1}{S_T\sqrt{2\pi\sigma^2\tau}} \exp\left[- \frac{[\log(S_T/S_t) - (r_{t,\tau} - \delta_{t,\tau} - \frac{1}{2}\sigma^2)\tau]^2}{2\sigma^2\tau} \right]. \end{aligned}$$

The risk measures Delta (Δ) and Gamma (Γ) are defined as:

$$\Delta_{BS} \stackrel{\text{def}}{=} \frac{\partial C_{BS}}{\partial S_t} = \Phi(d_1)$$

$$\Gamma_{BS} \stackrel{\text{def}}{=} \frac{\partial^2 C_{BS}}{\partial S_t^2} = \frac{\Phi(d_1)}{S_t\sigma\sqrt{\tau}}$$

The Black-Scholes SPD can be calculated in XploRe using the following quantlet:

```
bsspd = spdbs(K,s,r,div,sigma,tau)
      estimates the Black-Scholes SPD
```

The arguments are the strike prices (`K`), underlying price (`s`), risk-free interest rate (`r`), dividend yields (`div`), implied volatility of the option (`sigma`), and the time to maturity (`tau`). The output consist of the Black-Scholes SPD (`bsspd.fbs`), Δ (`bsspd.delta`), and the Γ (`bsspd.gamma`) of the call options. Please note that `spdbs` can be applied to put options by using the Put-Call parity.

However, it is widely known that the Black-Scholes call option formula is not valid empirically. For more details, please refer to Fengler et al. (2002). Since

the Black-Scholes model contains empirical irregularities, its SPD will not be consistent with the data. Consequently, some other techniques for estimating the SPD without any assumptions on the underlying diffusion process have been developed in the last years.

3 Semiparametric estimation of the SPD

3.1 Estimating the call pricing function

The use of nonparametric regression to recover the SPD was first investigated by Aït-Sahalia and Lo (1998). They propose to use the Nadaraya-Watson estimator to estimate the historical call prices $C_t(\cdot)$ as a function of the following state variables $(S_t, K, \tau, r_{t,\tau}, \delta_{t,\tau})^\top$. Kernel regressions are advocated because there is no need to specify a functional form and the only required assumption is that the function is smooth and differentiable, Härdle (1990). When the regressor dimension is 5, the estimator is inaccurate in practice. Hence, there is a need to reduce the dimension or equivalently the number of regressors. One method is to appeal to no-arbitrage arguments and collapse S_t, $r_{t,\tau}$ and $\delta_{t,\tau}$ into the forward price $F_t = S_t e^{(r_{t,\tau}-\delta_{t,\tau})\tau}$ in order to express the call pricing function as:

$$C(S_t, K, \tau, r_{t,\tau}, \delta_{t,\tau}) = C(F_{t,\tau}, K, \tau, r_{t,\tau}). \tag{7}$$

An alternative specification assumes that the call option function is homogeneous of degree one in S_t and K (as in the Black-Scholes formula) so that:

$$C(S_t, K, \tau, r_{t,\tau}, \delta_{t,\tau}) = KC(S_t/K, \tau, r_{t,\tau}, \delta_{t,\tau}). \tag{8}$$

Combining the assumptions of (7) and (8) the call pricing function can be further reduced to a function of three variables $(\frac{K}{F_{t,\tau}}, \tau, r_{t,\tau})$.

Another approach is to use a semiparametric specification based on the Black-Scholes implied volatility. Here, the implied volatility σ is modelled as a nonparametric function, $\sigma(F_{t,\tau}, K, \tau)$:

$$C(S_t, K, \tau, r_{t,\tau}, \delta_{t,\tau}) = C_{BS}(F_{t,\tau}, K, \tau, r_{t,\tau}; \sigma(F_{t,\tau}, K, \tau)). \tag{9}$$

Empirically the implied volatility function mostly depends on two parameters: the time to maturity τ and the moneyness $M = K/F_{t,\tau}$. Almost equivalently, one can set $M = \tilde{S}_t/K$ where $\tilde{S}_t = S_t - D$ and D is the present value of the dividends to be paid before the expiration. Actually, in the case of a dividend yield δ_t, we have $D = S_t(1 - e^{-\delta_t})$. If the dividends are discrete, then $D = \sum_{t_i \leq t+\tau} D_{t_i} e^{-r_{t,\tau_i}}$ where t_i is the dividend payment date of the i^{th} dividend and τ_i is its maturity.

Therefore, the dimension of the implied volatility function can be reduced to $\sigma(K/F_{t,\tau}, \tau)$. In this case the call option function is:

$$C(S_t, K, \tau, r_{t,\tau}, \delta_{t,\tau}) = C_{BS}(F_{t,\tau}, K, \tau, r_{t,\tau}; \sigma(K/F_{t,\tau}, \tau)). \tag{10}$$

Once a smooth estimate of $\hat{\sigma}(\cdot)$ is obtained, estimates of $\hat{C}_t(\cdot)$, $\hat{\Delta}_t = \frac{\partial \hat{C}_t(\cdot)}{\partial S_t}$, $\hat{\Gamma}_t = \frac{\partial^2 \hat{C}_t(\cdot)}{\partial S_t^2}$, and $\hat{f}_t^* = e^{r_{t,\tau}\tau}\left[\frac{\partial^2 \hat{C}_t(\cdot)}{\partial K^2}\right]$ can be calculated.

3.2 Further dimension reduction

The previous section proposed a semiparametric estimator of the call pricing function and the necessary steps to recover the SPD. In this section the dimension is reduced further using the suggestion of Rookley (1997). Rookley uses intraday data for one maturity and estimates an implied volatility surface where the dimension are the intraday time and the moneyness of the options.

Here, a slightly different method is used which relies on all settlement prices of options of one trading day for different maturities to estimate the implied volatility surface $\sigma(K/F_{t,\tau}, \tau)$. In the second step, these estimates are used for a given time to maturity which may not necessarily correspond to the maturity of a series of options. This method allows one to compare the SPD at different dates because of the fixed maturity provided by the first step. This is interesting if one wants to study the dynamics and the stability of these densities.

Fixing the maturity also allows us to eliminate τ from the specification of the implied volatility function. In the following part, for convenience, the definition of the moneyness is $M = \tilde{S}_t/K$ and we denote by σ the implied volatility. The notation $\frac{\partial f(x_1,\ldots,x_n)}{\partial x_i}$ denotes the partial derivative of f with respect to x_i and $\frac{\mathrm{d}f(x)}{\mathrm{d}x}$ the total derivative of f with respect to x.

Moreover, we use the following rescaled call option function:

$$\begin{aligned} c_{it} &= \frac{C_{it}}{\tilde{S}_t}, \\ M_{it} &= \frac{\tilde{S}_t}{K_i}. \end{aligned}$$

where C_{it} is the price of the i^{th} option at time t and K_i is its strike price.

The rescaled call option function can be expressed as:

$$\begin{aligned} c_{it} &= c(M_{it}; \sigma(M_{it})) = \Phi(d_1) - \frac{e^{-r\tau}\Phi(d_2)}{M_{it}}, \\ d_1 &= \frac{\log(M_{it}) + \left\{r_t + \frac{1}{2}\sigma(M_{it})^2\right\}\tau}{\sigma(M_{it})\sqrt{\tau}}, \\ d_2 &= d_1 - \sigma(M_{it})\sqrt{\tau}. \end{aligned}$$

The standard risk measures are then the following partial derivatives (for notational convenience subscripts are dropped):

$$\Delta = \frac{\partial C}{\partial S} = \frac{\partial C}{\partial \tilde{S}} = c(M, \sigma(M)) + \tilde{S}\frac{\partial c}{\partial \tilde{S}},$$

$$\Gamma = \frac{\partial \Delta}{\partial S} = \frac{\partial^2 C}{\partial S^2} = \frac{\partial^2 C}{\partial \tilde{S}^2} = 2\frac{\partial c}{\partial \tilde{S}} + \tilde{S}\frac{\partial^2 c}{\partial \tilde{S}^2}.$$

where

$$\begin{aligned}\frac{\partial c}{\partial \tilde{S}} &= \frac{\mathrm{d}c}{\mathrm{d}M}\frac{\partial M}{\partial \tilde{S}} = \frac{\mathrm{d}c}{\mathrm{d}M}\frac{1}{K},\\ \frac{\partial^2 c}{\partial \tilde{S}^2} &= \frac{\mathrm{d}^2 c}{\mathrm{d}M^2}\left(\frac{1}{K}\right)^2.\end{aligned}$$

The SPD is then the second derivative of the call option function with respect to the strike price:

$$f^*(\cdot) = e^{r\tau}\frac{\partial^2 C}{\partial K^2} = e^{r\tau}\tilde{S}\frac{\partial^2 c}{\partial K^2}. \tag{13}$$

The conversion is needed because $c(\cdot)$ is being estimated not $C(\cdot)$. The analytical expression of (13) depends on:

$$\frac{\partial^2 c}{\partial K^2} = \frac{\mathrm{d}^2 c}{\mathrm{d}M^2}\left(\frac{M}{K}\right)^2 + 2\frac{\mathrm{d}c}{\mathrm{d}M}\frac{M}{K^2}$$

The functional form of $\frac{\mathrm{d}c}{\mathrm{d}M}$ is:

$$\frac{\mathrm{d}c}{\mathrm{d}M} = \Phi'(d_1)\frac{\mathrm{d}d_1}{\mathrm{d}M} - e^{-r\tau}\frac{\Phi'(d_2)}{M}\frac{\mathrm{d}d_2}{\mathrm{d}M} + e^{-r\tau}\frac{\Phi(d_2)}{M^2}, \tag{14}$$

while $\frac{\mathrm{d}^2 c}{\mathrm{d}M^2}$ is:

$$\begin{aligned}\frac{\mathrm{d}^2 c}{\mathrm{d}M^2} &= \Phi'(d_1)\left[\frac{\mathrm{d}^2 d_1}{\mathrm{d}M^2} - d_1\left(\frac{\mathrm{d}d_1}{\mathrm{d}M}\right)^2\right]\\ &- \frac{e^{-r\tau}\Phi'(d_2)}{M}\left[\frac{\mathrm{d}^2 d_2}{\mathrm{d}M^2} - \frac{2}{M}\frac{\mathrm{d}d_2}{\mathrm{d}M} - d_2\left(\frac{\mathrm{d}d_2}{\mathrm{d}M}\right)^2\right]\\ &- \frac{2e^{-r\tau}\Phi(d_2)}{M^3}\end{aligned} \tag{15}$$

The quantities in (14) and (15) are a function of the following first derivatives:

$$\begin{aligned}\frac{\mathrm{d}d_1}{\mathrm{d}M} &= \frac{\partial d_1}{\partial M} + \frac{\partial d_1}{\partial \sigma}\frac{\partial \sigma}{\partial M},\\ \frac{\mathrm{d}d_2}{\mathrm{d}M} &= \frac{\partial d_2}{\partial M} + \frac{\partial d_2}{\partial \sigma}\frac{\partial \sigma}{\partial M},\\ \frac{\partial d_1}{\partial M} &= \frac{\partial d_2}{\partial M} = \frac{1}{M\sigma\sqrt{\tau}},\\ \frac{\partial d_1}{\partial \sigma} &= -\frac{\log(M) + r\tau}{\sigma^2\sqrt{\tau}} + \frac{\sqrt{\tau}}{2},\\ \frac{\partial d_2}{\partial \sigma} &= -\frac{\log(M) + r\tau}{\sigma^2\sqrt{\tau}} - \frac{\sqrt{\tau}}{2}.\end{aligned}$$

For the remainder of this chapter, we define:

$$\begin{aligned}V &= \sigma(M),\\ V' &= \frac{\partial \sigma(M)}{\partial M},\\ V'' &= \frac{\partial^2 \sigma(M)}{\partial M^2}.\end{aligned} \tag{16}$$

The quantities in (14) and (15) also depend on the following second derivative functions:

$$\begin{aligned}\frac{\mathrm{d}^2 d_1}{\mathrm{d}M^2} &= -\frac{1}{M\sigma\sqrt{\tau}}\left[\frac{1}{M}+\frac{V'}{\sigma}\right]+V''\left(\frac{\sqrt{\tau}}{2}-\frac{\log(M)+r\tau}{\sigma^2\sqrt{\tau}}\right)\\ &+ V'\left[2V'\frac{\log(M)+r\tau}{\sigma^3\sqrt{\tau}}-\frac{1}{M\sigma^2\sqrt{\tau}}\right], \qquad (17)\end{aligned}$$

$$\begin{aligned}\frac{\mathrm{d}^2 d_2}{\mathrm{d}M^2} &= -\frac{1}{M\sigma\sqrt{\tau}}\left[\frac{1}{M}+\frac{V'}{\sigma}\right]-V''\left(\frac{\sqrt{\tau}}{2}+\frac{\log(M)+r\tau}{\sigma^2\sqrt{\tau}}\right)\\ &+ V'\left[2V'\frac{\log(M)+r\tau}{\sigma^3\sqrt{\tau}}-\frac{1}{M\sigma^2\sqrt{\tau}}\right]. \qquad (18)\end{aligned}$$

Local polynomial estimation is used to estimate the implied volatility smile and its first two derivatives in (16). A brief explanation will be described now.

3.3 Local Polynomial Estimation

Consider the following data generating process for the implied volatilities:

$$\sigma = g(M,\tau)+\sigma^*(M,\tau)\varepsilon,$$

where $\mathrm{E}(\varepsilon)=0$, $\mathrm{Var}(\varepsilon)=1$. M,τ and ε are independent and $\sigma^*(m_0,\tau_0)$ is the conditional variance of σ given $M=m_0,\tau=\tau_0$. Assuming that all third derivatives of g exist, one may perform a Taylor expansion for the function g in a neighborhood of (m_0,τ_0):

$$\begin{aligned}g(m,\tau)\approx g(m_0,\tau_0) &+ \left.\frac{\partial g}{\partial M}\right|_{m_0,\tau_0}(m-m_0)+\frac{1}{2}\left.\frac{\partial^2 g}{\partial M^2}\right|_{m_0,\tau_0}(m-m_0)^2\\ &+ \left.\frac{\partial g}{\partial \tau}\right|_{m_0,\tau_0}(\tau-\tau_0)+\frac{1}{2}\left.\frac{\partial^2 g}{\partial \tau^2}\right|_{m_0,\tau_0}(\tau-\tau_0)^2\\ &+ \frac{1}{2}\left.\frac{\partial^2 g}{\partial M\partial \tau}\right|_{m_0,\tau_0}(m-m_0)(\tau-\tau_0). \qquad (19)\end{aligned}$$

This expansion suggests an approximation by local polynomial fitting, Fan and Gijbels (1996). Hence, to estimate the implied volatility at the target point (m_0,τ_0) from observations σ_j $(j=1,\ldots,n)$, we minimize the following expression:

$$\begin{aligned}&\sum_{j=1}^{n}\Big\{\sigma_j-\Big[\beta_0+\beta_1(M_j-m_0)+\beta_2(M_j-m_0)^2+\beta_3(\tau_j-\tau_0)\\ &+\beta_4(\tau_j-\tau_0)^2+\beta_5(M_j-m_0)(\tau_j-\tau_0)\Big]\Big\}^2 K_{h_M,h_\tau}(M_j-m_0,\tau_j-\tau_0) \qquad (20)\end{aligned}$$

where n is the number of observations (options), h_M and h_τ are the bandwidth controlling the neighborhood in each directions and K_{h_M,h_τ} is the resulting kernel function weighting all observation points. This kernel function may be a product of two univariate kernel functions.

For convenience use the following matrix definitions:

$$X = \begin{pmatrix} 1 & M_1 - m_0 & (M_1 - m_0)^2 & \tau_1 - \tau_0 & (\tau_1 - \tau_0)^2 & (M_1 - m_0)(\tau_1 - \tau_0) \\ 1 & M_2 - m_0 & (M_2 - m_0)^2 & \tau_2 - \tau_0 & (\tau_2 - \tau_0)^2 & (M_2 - m_0)(\tau_2 - \tau_0) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & M_n - m_0 & (M_n - m_0)^2 & \tau_n - \tau_0 & (\tau_n - \tau_0)^2 & (M_n - m_0)(\tau_n - \tau_0) \end{pmatrix},$$

$$\sigma = \begin{pmatrix} \sigma_1 \\ \vdots \\ \sigma_n \end{pmatrix}, \quad W = diag\{K_{h_M,h_\tau}(M_j - m_0, \tau_j - \tau_0)\} \quad \text{and} \quad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_5 \end{pmatrix}.$$

Hence, the weighted least squares problem (20) can be written as

$$\min_{\beta} (\sigma - X\beta)^\top W (\sigma - X\beta). \tag{21}$$

and the solution is given by

$$\hat{\beta} = \left(X^\top W X\right)^{-1} X^\top W \sigma. \tag{22}$$

A nice feature of the local polynomial method is that it provides the estimated implied volatility and its first two derivatives in one step. Indeed, one has from (19) and (20):

$$\begin{aligned} \widehat{\frac{\partial g}{\partial M}}\bigg|_{m_0,\tau_0} &= \hat{\beta}_1, \\ \widehat{\frac{\partial^2 g}{\partial M^2}}\bigg|_{m_0,\tau_0} &= 2\hat{\beta}_2. \end{aligned}$$

One of the concerns regarding this estimation method is the dependence on the bandwidth which governs how much weight the kernel function should place on an observed point for the estimation at a target point. Moreover, as the call options are not always symmetrically and equally distributed around the ATM point, the choice of the bandwidth is a key issue, especially for estimation at the border of the implied volatility surface. The bandwidth can be chosen global or locally dependent on (M, τ). There are methods providing "optimal" bandwidths which rely on plug-in rules or on data-based selectors.

In the case of the volatility surface, it is vital to determine one bandwidth for the maturity and one for the moneyness directions. An algorithm called Empirical-Bias Bandwidth Selector (EBBS) for finding local bandwidths is suggested by Ruppert (1997) and Ruppert, Wand, Holst and Hössler (1997). The basic idea of this method is to minimize the estimate of the local mean square error at each target point, without relying on asymptotic result. The variance and the bias term are in this algorithm estimated empirically.

Using the local polynomial estimations, the empirical SPD can be calculated with the following quantlet:

```
lpspd = spdbl(m,sigma,sigma1,sigma2,s,r,tau)
      estimates the semi-parametric SPD.
```

The arguments for this quantlet are the moneyness (`m`), V (`sigma`), V' (`sigma1`), V'' (`sigma2`), underlying price (`s`) corrected for future dividends, risk-free interest rate (`r`), and the time to maturity (`tau`). The output consist of the local polynomial SPD (`lpspd.fstar`), Δ (`lpspd.delta`), and the Γ (`lpspd.gamma`) of the call-options.

4 An Example: Application to DAX data

This section describes how to estimate the Black-Scholes and local polynomial SPD using options data on the German DAX index.

4.1 Data

The dataset was taken from the financial database MD*BASE located at CASE (Center for Applied Statistics and Economics) at Humboldt-Universität zu Berlin. Since MD*BASE is a proprietary database, only a limited dataset is provided for demonstration purposes.

This database is filled with options and futures data provided by Eurex. Daily series of 1, 3, 6 and 12 months DM-LIBOR rates taken from the *Thomson Financial Datastream* serve as riskless interest rates. The DAX 30 futures and options settlement data of January 1997 (21 trading days) were used in this study. Daily settlement prices for each option contract are extracted along with contract type, maturity and strike. For the futures, the daily settlement prices, maturities and volumes are the relevant information. To compute the interest rates corresponding to the option maturities a linear interpolation between the available rates was used.

The DAX is a performance index which means that dividends are reinvested. However, assuming no dividend yields when inverting the Black-Scholes formula results in different volatilities for pairs of puts and calls contrary to the no-arbitrage assumption contained in the Put-Call parity. This remark can be explained by the fact that until January 2002 domestic investors have an advantage as they may receive a portion or all of the dividend taxes back depending on their tax status. Dividend tax means here the corporate income tax for distributed gains from the gross dividend.

Since the dividends are rebated to domestic investors the DAX should fall by an amount contained between 0 and these dividend taxes. Indeed, the value of this fall depends on the level of these taxes which may be equal to zero and on the weights of domestic and foreign investors trading the DAX. These dividend taxes have the same effects as ordinary dividends and should therefore be used for computing the implied volatilities and the future price implicit in the Black Scholes formula.

Hafner and Wallmeier (2001) suggest a method in order to get around this problem which consists in computing dividends implied by the Put-Call parity. Indeed, combining the futures pricing formula

$$F_{t,\tau_F} = S_t e^{r_{t,\tau_F}\tau_F} - D_{t,\tau_f}$$

and the Put-Call parity

$$C_t - P_t = S_t - D_{t,\tau_O} - Ke^{-r_{t,\tau_o}\tau_o}$$

we obtain:

$$C_t - P_t = F_{t,\tau_F} e^{-r_{t,\tau_F}} + D_{t,\tau_F,\tau_O} - K e^{-r_{t,\tau_O}\tau_O} \tag{23}$$

where τ_O is the maturity of the options, τ_F is the maturity of the nearest forward whose volume is positive and $D_{t,\tau_F,\tau_O} = D_{t,\tau_F} - D_{t,\tau_O}$ is the difference between the present values of the dividends.

Using (23), implied dividends were computed for each pair of put and call with the same strike. Theoretically, for a given time to maturity there must be only one value for these implied dividends. For each maturity the average of these implied dividends was used to compute the corrected price. Using this method implied volatilities are more reliable as the systematic "gap" between put and call volatilities disappears. The only uncertainty at this stage is due to the interpolated rates for the maturity τ_O.

The dataset consists of one file `XFGData9701` with 11 columns.

1	Day
2	Month
3	Year
4	Type of option (1 for calls, 0 for puts)
5	Time to maturity (in calendar days)
6	Strike prices
7	Option prices
8	Corrected spot price (implied dividends taken into account)
9	Risk-free interest rate
10	Implied volatility
11	Non-corrected spot price

The data can be read into XploRe by loading the quantlib `finance` and then issuing the following command:

```
data=read("XFGData9701.dat")
```

Next extract all call options on January 3, 1997 with the `paf` command:

```
data=paf(data,(data[,1]==3)&&(data[,4]==1))
```

4.2 SPD, delta and gamma

This section provides an example using XploRe to calculate the semiparametric SPD using DAX index options data. It is assumed that the quantlib `finance` has been loaded.

`XFGSPDonematurity.xpl` plots the SPD of the series of options closest to maturity. This first example only uses smoothing method in one dimension.

`XFGSPDoneday.xpl` calculates and plots the local polynomial SPD for January 10, 1997 for different times to maturity ($\tau = 0.125, 0.25, 0.375$). After loading the data, the implied volatility is estimated using the `volsurf` quantlet, while the first and second derivatives are estimated using `lpderxest` quantlet.

In this example the grid size is 0.01. The bandwidth is chosen arbitrarily at 0.15 and 0.125 for the moneyness and maturity directions respectively. The criteria used is a visual inspection of the first and second derivatives to ensure that they are continuous and smooth. Next the quantlet `spdbl` is used to calculate the SPD which is finally displayed in Figure 1.

This figure shows the expected effect of time to maturity on the SPD, which is a loss of kurtosis. The x-axis represents the terminal prices S_T. The local polynomial SPD displays a negative skew compared to a theoretical Black-Scholes SPD. The major reason for the difference is the measure of implied volatility. Using the local polynomial estimators one captures the effect of the "volatility smile" and its effects on the higher moments such as skewness and kurtosis. This result is similar to what Aït-Sahalia and Lo (1998) and Rookley (1997) found in their study.

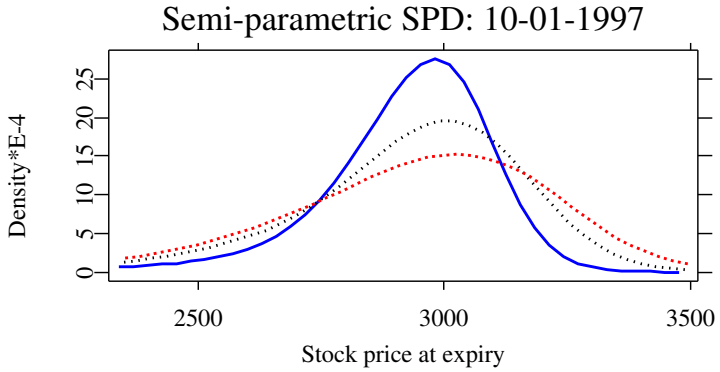

Figure 1: Local Polynomial SPD for $\tau = 0.125$ (blue,filled), $\tau = 0.25$ (black,dashed) and $\tau = 0.375$ (red,dotted).

XFGSPDoneday.xpl

Figure 2 and Figure 3 show Delta and Gamma for the full range of strikes and for three different maturities. This method allows the user to get in one step both greeks in one estimation for all strikes and maturities.

A natural question that may arise is how do the SPDs evolve over time. In this section an illustrative example is used to show the dynamics of the SPD over the month of January 1997. XFGSPDonemonth.xpl estimates and plots the SPD for each trading day in January 1997. The x-axis is the moneyness, y-axis is the trading day, and the z-axis is the SPD. Figure 4 shows the local polynomial SPD for the three first weeks of January, 1997.

4.3 Bootstrap confidence bands

Rookley's method serves to estimate the SPD, where V, V' and V'' from (16) are computed via local polynomials. The method is now applied to estimate a SPD whose maturity is equal to the maturity of a series of options. In this case, the nonparametric regression is a univariate one.

With a polynomial of order $p = 2$ and a bandwidth $h = \left(n^{-1/9}\right)$, it can be shown that

$$E|\hat{f}_n^* - f^*|^2 = \mathcal{O}\left(n^{-4/9}\right),$$

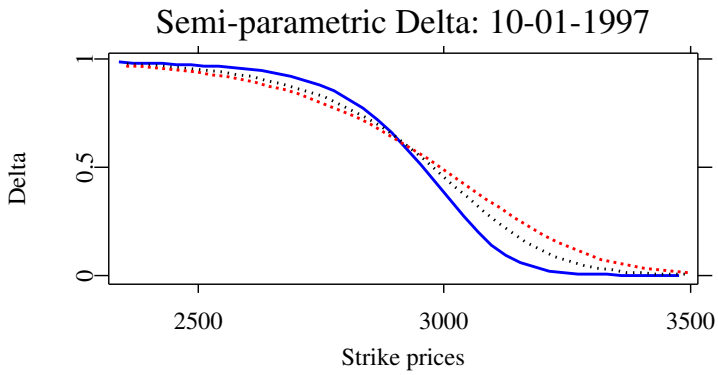

Figure 2: Local Polynomial Delta for $\tau = 0.125$ (blue,filled), $\tau = 0.25$ (black,dashed) and $\tau = 0.375$ (red,dotted).

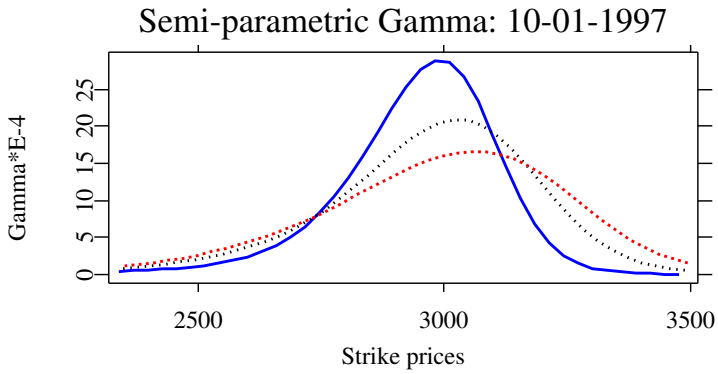

Figure 3: Local Polynomial Gamma for $\tau = 0.125$ (blue,filled), $\tau = 0.25$ (black,dashed) and $\tau = 0.375$ (red,dotted).

XFGSPDoneday.xpl

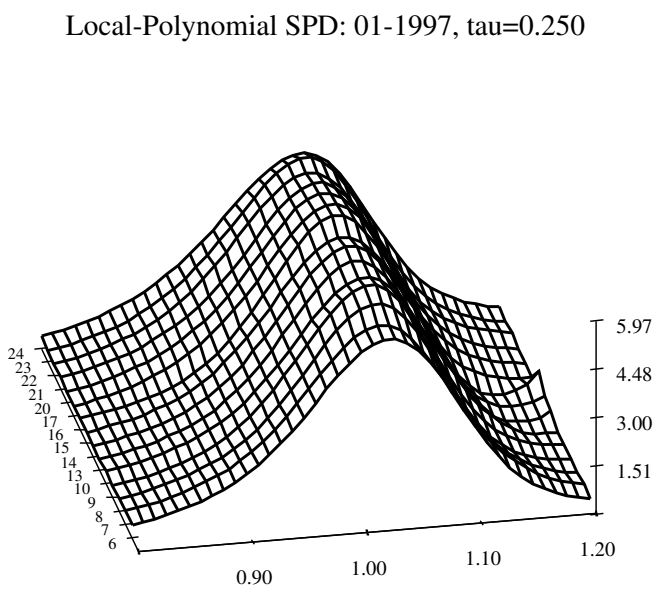

Figure 4: Three weeks State-Price Densities on a moneyness scale.

XFGSPDonemonth.xpl

because

$$\begin{aligned} E|\hat{V}_n - V|^2 &= \mathcal{O}\left(n^{-8/9}\right), \\ E|\hat{V}_n^{'} - V^{'}|^2 &= \mathcal{O}\left(n^{-4/9}\right), \\ E|\hat{V}_n^{''} - V^{''}|^2 &= \mathcal{O}\left(n^{-4/9}\right). \end{aligned}$$

This result can be obtained using some theorems related to local polynomial estimation, for example in Fan and Gijbels (1996), if some boundary conditions are satisfied.

An asymptotic approximation of $\hat{f}_n^*$ is complicated by the fact that $\hat{f}_n^*$ is a non linear function of V, V' and V''. Analytical confidence intervals can be obtained using delta methods proposed by Aït-Sahalia (1996). However, an alternative method is to use the bootstrap to construct confidence bands. The idea for estimating the bootstrap bands is to approximate the distribution of

$$\sup_k |\hat{f}^*(k) - f^*(k)|.$$

The following procedure illustrates how to construct bootstrap confidence bands for local polynomial SPD estimation.

1. Collect daily option prices from MD*BASE, only choose those options with the same expiration date, for example, those with time to maturity 49 days on Jan 3, 1997.

2. Use the local polynomial estimation method to obtain the empirical SPD. Notice that when τ is fixed the forward price F is also fixed. So that the implied volatility function $\sigma(K/F)$ can be considered as a fixed design situation, where K is the strike price.

3. Obtain the confidence band using the wild bootstrap method. The wild bootstrap method entails:

   - Suppose that the regression model for the implied volatility function $\sigma(K/F)$ is:
   $$Y_i = \sigma\left(\frac{K_i}{F}\right) + \varepsilon_i, \quad i = 1, \cdots, n.$$
   - Choose a bandwidth g which is larger than the optimal h in order to have oversmoothing. Estimate the implied volatility function $\sigma(K/F)$ nonparametrically and then calculate the residual errors:
   $$\tilde{\varepsilon}_i = Y_i - \hat{\sigma}_h\left(\frac{K_i}{F}\right).$$
   - Replicate B times the series of the $\{\tilde{\varepsilon}_i\}$ with wild bootstrap obtaining $\{\varepsilon_i^{*,j}\}$ for $j = 1, \cdots, B$, Härdle (1990), and build B new bootstrapped samples:
   $$Y_i^{*,j} = \hat{\sigma}_g\left(\frac{K_i}{F}\right) + \varepsilon_i^{*,j}.$$

- Estimate the SPD $f^{*,j}$ using bootstrap samples, Rookley's method and the bandwidth h, and build the statistics

$$T_f^* = \sup_z |f^{*,j}(z) - \hat{f}^*(z)|.$$

- Form the $(1-\alpha)$ bands $[\hat{f}^*(z) - t_{f^*,1-\alpha},\ \hat{f}^*(z) + t_{f^*,1-\alpha}]$, where $t_{f^*,1-\alpha}$ denotes the empirical $(1-\alpha)$-quantile of T_f^*.

Two SPDs (Jan 3 and Jan 31, 1997) whose times to maturity are 49 days were estimated and are plotted in Figure (5). The bootstrap confidence band corresponding to the first SPD (Jan 3) is also visible on the chart. In Figure (6), the SPDs are displayed on a moneyness metric. It seems that the differences between the SPDs can be eliminated by switching to the moneyness metric. Indeed, as can be extracted from Figure 6, both SPDs lie within the 95 percent confidence bands. The number of bootstrap samples is set to $B = 100$. The local polynomial estimation was done on standardized data, h is then set to 0.75 for both plots and g is equal to 1.1 times h. Notice that greater values of g are tried and the conclusion is that the confidence bands are stable to an increase of g.

4.4 Comparison to Implied Binomial Trees

In Härdle and Zheng (2002), the Implied Binomial Trees (IBT) are discussed. This method is a close approach to estimate the SPD. It also recovers the SPD nonparametrically from market option prices and uses the Black Scholes formula to establish the relationship between the option prices and implied volatilities as in Rookely's method. In Härdle and Zheng (2002), the Black Scholes formula is only used for Barle and Cakici IBT procedure, but the CRR binomial tree method used by Derman and Kani (1994) has no large difference with it in nature. However, IBT and nonparametric regression methods have some differences caused by different modelling strategies.

The IBT method might be less data-intensive than the nonparametric regression method. By construction, it only requires one cross section of prices. In the earlier application with DAX data, option prices are used with different times to maturity for one day to estimate the implied volatility surface first in order to construct the tree using the relation formula between option prices and risk-neutral probabilities. The precision of the SPD estimation using IBT is heavily affected by the quality of the implied volatility surface and the choice of the levels of the implied tree. Furthermore, from the IBT method only risk-neutral probabilities are obtained. They can be considered as a discrete estimation of the SPD. However, the IBT method is not only useful for estimating SPD, but also for giving a discrete approximation of the underlying process.

The greatest difference between IBTs and nonparametric regression is the requirement of smoothness. The precision of Rookley's SPD estimation is highly dependent on the selected bandwidth. Even if very limited option prices are given, a part of the SPD estimation still can be obtained using nonparametric regression, while the IBT construction has to be given up if no further structure is invoked on the volatility surface. Rookley's method has on first sight no obvious difference with Aït-Sahalia's method theoretically, Aït-Sahalia and Lo (1998). But investigating the convergence rate of the SPD estimation using

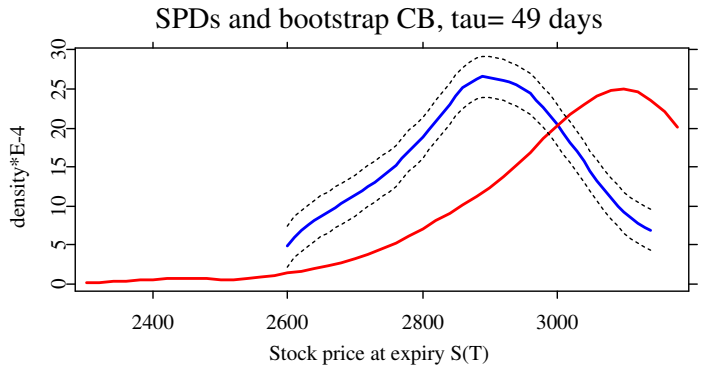

Figure 5: SPD estimation and bootstrap confidence band.

XFGSPDcb.xpl

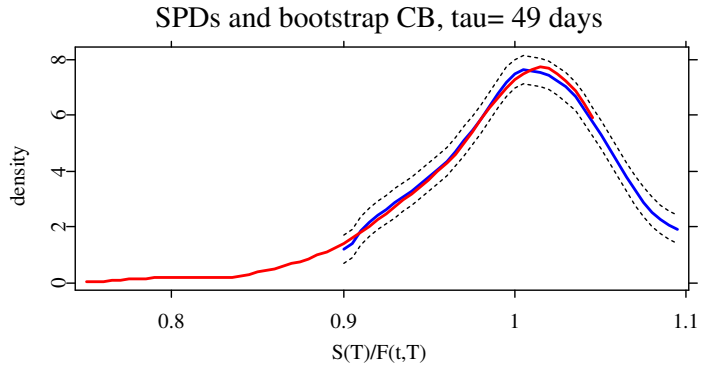

Figure 6: SPD estimation and bootstrap confidence band (moneyness metric).

XFGSPDcb2.xpl

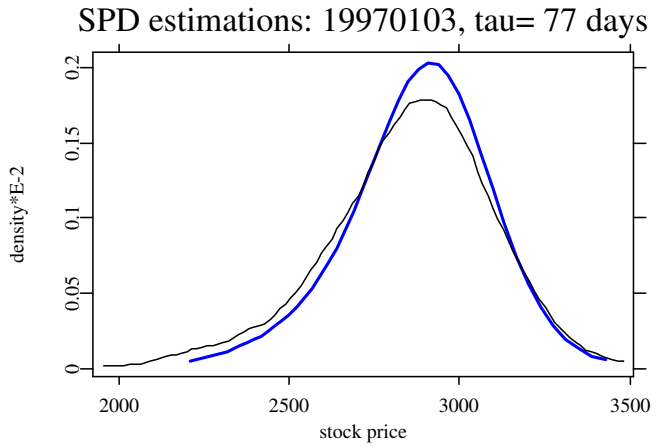

Figure 7: Comparison of different SPD estimations, by Rookley's method (blue) and IBT (black, thin).

XFGSPDcom.xpl

Aït-Sahalia's method allows one to conduct statistical inference such as test of the stability of the SPD and tests of risk neutrality.

The quantlet XFGSPDcom.xpl shows a comparison of the SPD estimates by IBT and Rookley's methods. The differences between these two SPD estimates may be due to the selection of the bandwidths in Rookley's method, the choice of steps in the construction of the IBT and the use of DAX implied dividends in Rookley's method. Figure 7 shows the implied binomial trees and the local polynomial SPDs for January, 3 1997.

Both densities seems to be quiet different. Indeed, the IBTs SPD shows a fatter left tail than the Rookley's one and the Rookley's SPD shows a larger kurtosis. To test which of both densities is more reliable, a cross-validation procedure is performed. The idea of this test is to compare the theoretical prices based on (1) with those observed on the market. However, as the whole tails are not available for the Rookley's SPD, the test is done on butterfly spreads defined in Section 2 since their prices should not be influenced by the tails of the SPDs. For cross-validation, we remove the three calls used to calculate the observed butterfly prices from the sample before estimating the SPD. Moreover, since the largest difference between both SPDs is observed at the ATM point (see Figure 7), the test is applied only to the two butterfly spreads whose centers surround the ATM point. The width $2\Delta K$ of the butterfly spread is set to 200. This procedure is done for the 21 days of January 1997. Figure 8 displays the results in term of relative pricing error E:

$$E = \frac{P_{observed} - P_{SPD}}{P_{observed}}$$

where $P_{observed}$ is the observed price of the butterfly spread and P_{SPD} is the

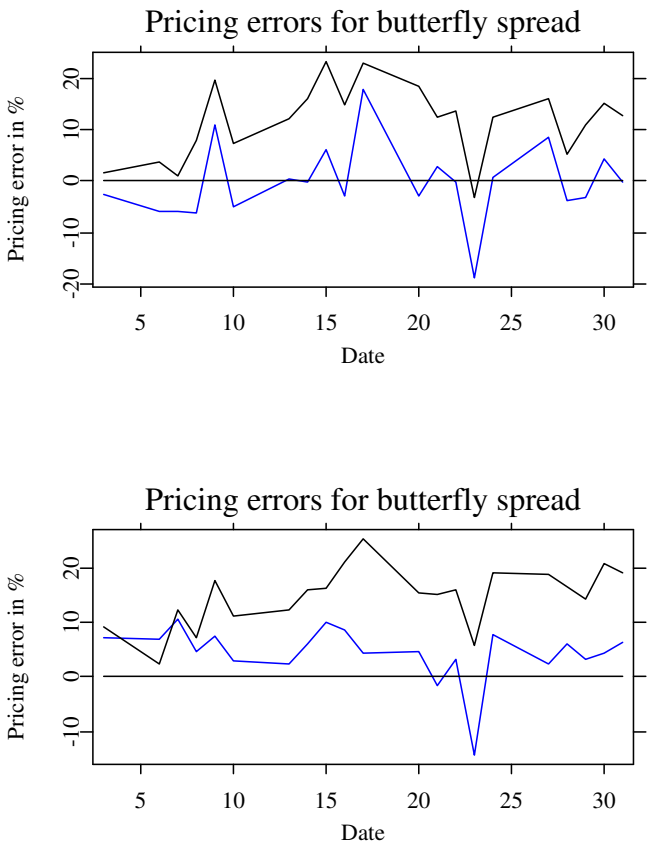

Figure 8: The upper graph display the relative pricing errors for the butterfly spread centered on the nearest strike on the left side of the ATM point. The second graph corresponds to the butterfly spread centered on the nearest strike on the right side of the ATM point. The black lines represent the IBT's pricing errors and the blue the Rookley's errors.

price computed using the SPD estimate and (1). It seems that both SPDs have a too small kurtosis since the observed prices of butterfly spreads are larger than those of both SPDs in most of the cases. However, Rookley's SPD is in mean nearer to the observed price than the IBT's one.

Acknowledgement

The authors acknowledge support by the Deutsche Forschungsgemeinschaft via Sonderforschungsbereich 373 "Quantifikation and Simulation ökonomischer Prozesse" at Humboldt-Universität zu Berlin.

References

Aït-Sahalia, Y. (1996). The Delta method for Nonparametric Kernel Functionals, mimeo.

Aït-Sahalia, Y. and Lo, A. W. (1998). Nonparametric estimation of state-price densities implicit in financial asset prices, *Journal of Finance* **53**: 499–547.

Arrow, K. (1964). The role of securities in the optimal allocation of risk bearing, *Review of Economic Studies* **31**: 91–96.

Bahra, B. (1997). Implied risk-neutral probability density functions from option prices: theory and application. Bank of England Working Paper 66.

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

Breeden, D. and Litzenberger, R. H. (1978). Prices of state-contingent claims implicit in option prices, *Journal of Business* **51**: 621–651.

Debreu, G. (1959). *Theory of Value*, John Wiley and Sons, New York.

Derman, E. and Kani, I. (1994). Riding on the smile, *Risk* **7**: 32–39.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Apllications*, Chapman and Hall, New York. Vol. 66 of Monographs on Statistics and Applied Probability.

Fengler, M., Härdle, W. and Schmidt, P. (2002). The Analysis of Implied Volatilities, *in* W. Härdle, T. Kleinow and G. Stahl (eds), *Applied Quantitative Finance*, Springer Finance, Springer Verlag.

Hafner, R. and Wallmeier, M. (2001). The dynamics of DAX implied volatilities, *Quarterly International Journal of Finance* **1**: 1–27.

Härdle, W. (1990). *Applied Nonparametric Regression*, Cambridge University Press, New York.

Härdle, W., Hlávka, Z. and Klinke, S. (2000). *XploRe Application Guide*, Springer-Verlag, Berlin.

Härdle, W. and Zheng, J. (2002). How Precise Are Price Distributions Predicted by Implied Binomial Trees?, *in* W. Härdle, T. Kleinow and G. Stahl (eds), *Applied Quantitative Finance*, Springer Finance, Springer Verlag.

Hutchinson, J., Lo, A. and Poggio, A. (1994). A nonparametric approach to the pricing and hedging of derivative securities via learning networks, *Journal of Finance* **49**: 851–889.

Lucas, R. E. (1978). Asset prices in an exchange economy, *Econometrica* **46**(1429-1446).

Melick, W. and Thomas, C. (1997). Recovering an Asset's Implied PDF from Option Prices: An application to Crude Oil During the Gulf Crisis, *Journal of Financial and Quantitative Analysis* **32**: 91–115.

Merton, R. B. (1973). Rational theory of option pricing, *Bell Journal of Economics and Management Science* **4**: 141–183.

Rookley, C. (1997). Fully exploiting the information content of intra-day option quotes: Applications in option pricing and risk management. mimeo.

Rubinstein, M. (1976). The valuation of uncertain income streams and the pricing of options, *Bell Journal of Economics* **7**(407-425).

Rubinstein, M. (1985). Nonparametric tests of alternative option pricing models using all reported trades and quotes on the 30 most active cboe option classes from august 23, 1976 to august 31, 1978, *Journal of Finance* **40**: 455–480.

Rubinstein, M. (1994). Implied binomial trees, *Journal of Finance* **49**: 771–818.

Ruppert, D. (1997). Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation, *Journal of the American Statistical Association* **92**: 1049–1062.

Ruppert, D., Wand, M. P., Holst, U. and Hössler, O. (1997). Local polynomial variance-function estimation, *Technometrics* **39**: 262–273.

Yatchew, A. and Härdle, W. (2002). Dynamic nonparametric state price density estimation using constrained least squares and the bootstrap. *Journal of Econometrics, forthcoming.*