Journal of Econometrics 184 (2015) 242–261

Contents lists available at ScienceDirect

Journal of Econometrics

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

Semi-nonparametric estimation of the call-option price surface under strike and time-to-expiry no-arbitrage constraints☆

Matthias R. Fengler[a,*], Lin-Yee Hin[b]

[a] University of St. Gallen, Department of Economics, Bodanstrasse 6, CH-9000 St. Gallen, Switzerland
[b] Department of Mathematics and Statistics, Curtin University, Kent Street, Bentley WA 6102, Australia

ARTICLE INFO

Article history:
Received 26 April 2013
Received in revised form
1 September 2014
Accepted 1 September 2014
Available online 16 September 2014

JEL classification:
C14
C58
G13

Keywords:
B-splines
No-arbitrage constraints
Option pricing function
Semi-nonparametric estimation
Shape-constrained regression
State-price density

ABSTRACT

We suggest a semi-nonparametric estimator for the call-option price surface. The estimator is a bivariate tensor-product B-spline. To enforce no-arbitrage constraints across strikes and expiry dates, we establish sufficient no-arbitrage conditions on the control net of the B-spline surface. The conditions are linear and therefore allow for an implementation of the estimator by means of standard quadratic programming techniques. The consistency of the estimator is proved. By means of simulations, we explore the statistical efficiency benefits that are associated with estimating option price surfaces and state-price densities under the full set of no-arbitrage constraints. We estimate a call-option price surface, families of first-order strike derivatives, and state-price densities for S&P 500 option data.

1. Introduction

Option prices carry information on the risk factors that drive the underlying asset price process. This information can be exploited to price other, more complex contingent claims consistently with the market, to study policy events, and to learn about the risk perception and risk attitude of the representative agent in the market. Numerous strategies have therefore been suggested for estimating an option pricing function from randomly observed option price data in order to extract the relevant information.

Unlike many other financial data, option price data have the particular feature that a number of expiry dates across different exercise prices are traded concurrently. This feature allows the simultaneous study of cross-sections of option price data over various time horizons. Jackwerth and Rubinstein (1996) and Aït-Sahalia and Lo (1998), for example, compare the state-price density implied from option prices across different expiry dates; in a similar vein, Aït-Sahalia and Lo (2000), Jackwerth (2000) and Bliss and Panigirtzoglou (2004) study the empirical pricing kernel and implied risk aversion. An assumption that is implicit in many of these studies is that option prices observed contemporaneously over multiple time horizons are realizations from a smooth surface defined across exercise prices and expiry dates.

In this paper, we suggest an estimator for the pricing function of a European-style call-option that extends across all available expiry dates. In other words, we explicitly take account of two dimensions and estimate a call-option price surface. The estimator is a bivariate tensor product B-spline. It therefore belongs to the flexible class of series estimators also called semi-nonparametric estimators; see Gallant (1987). As financial theory requires, the estimator obeys shape constraints both across the strike and the expiry dimension to ensure that the resulting option price surface

☆ We thank Francesco Audrino, Wolfgang Härdle, Jens Jackwerth and the seminar participants at the Erasmus University Rotterdam, the Konstanz workshop "The Pricing Kernel Puzzle", the Frankfurt School of Finance and Management, the Humboldt-Universität zu Berlin, the CFE 2011 conference, the MathFinance Conference, the 7th World Congress of the Bachelier Finance Society, the 5th Annual SoFiE Conference and the EEA-ESEM 2012 for their comments. The paper has particularly benefited from valuable suggestions made by Patrick Gagliardini, Enno Mammen and three anonymous referees. Any remaining errors are ours.

* Corresponding author. Tel.: +41 71224 2457.
E-mail addresses: matthias.fengler@unisg.ch (M.R. Fengler), linyeehin@gmail.com (L.-Y. Hin).

http://dx.doi.org/10.1016/j.jeconom.2014.09.003
0304-4076/

is free of arbitrage. Since option data may be sparsely distributed, which could lead to an ill-posed estimation problem, we also study a regularized version of the estimator that is well-behaved.

Estimating an option price surface under shape constraints has rarely been achieved in the literature so far. Two-dimensional estimators, such as those suggested in Aït-Sahalia and Lo (1998), Cont and da Fonseca (2002), and Fengler et al. (2007), do not accommodate no-arbitrage constraints. It is therefore a common practice to estimate univariate option pricing functions for each expiry date independently, usually in combination with some interpolation of the data across the expiry dates.

The literature on flexible modeling techniques designed to estimate a univariate option price function and/or its second-order strike derivative (the state-price density) under shape constraints is vast. The most important rivals of our estimator, however, are the fully nonparametric approaches, such as kernel smoothers, and semi-nonparametric (SNP) regression techniques that are based on series estimators, such as polynomial regression splines or other series expansions. Within the nonparametric stream, the pioneering work of Aït-Sahalia and Duarte (2003) suggests a local linear smoother to estimate the option pricing function and its strike derivatives. Alternative kernel regression estimators are proposed by Birke and Pilz (2009) and Fan and Mancini (2009). As regards SNP techniques, a number of polynomial spline methods have been suggested to date: B-splines in Wang et al. (2004), Laurini (2011), and Corlay (2013); smoothing splines in Yatchew and Härdle (2006), Monteiro et al. (2008), and Fengler (2009); linear splines in Härdle and Hlávka (2009). Other SNP-type estimators are based on the Edgeworth expansion as in Jarrow and Rudd (1982), on Hermite polynomials as in Madan and Milne (1994) and Jondeau and Rockinger (2001), or on approximation methods, such as the positive convolution approximation as in Bondarenko (2003) and the nonparametric density mixtures as in Yuan (2009). Finally, there are flexible estimation approaches that are nonparametric in nature, but do not properly fit into either strand, such as the neural network used in Hutchinson et al. (1994), the maximum entropy method suggested in Stutzer (1996) and the regularized calibration approach devised by Jackwerth and Rubinstein (1996).[1]

First approaches to include no-arbitrage constraints for surface estimation are Benko et al. (2007) and Glaser and Heider (2012). Both studies build on local polynomials, but they do not analyze the asymptotic properties of the constrained estimators, nor do they make an attempt to assess the efficiency benefits that are associated with implementing no-arbitrage constraints in the time-to-expiry dimension. Here, we exploit a projection framework for constrained smoothing devised by Mammen et al. (2001) to prove the consistency of the estimators and to provide an upper bound for their rates of convergence. For both estimators, this upper bound is the optimal rate for regression estimation owed to Stone (1982). In addition, we show by means of simulations that substantial efficiency gains are to be expected for the estimation of the option price surface and its derivatives, if one implements calendar conditions as well as strike constraints.

As noted above, the use of splines to estimate the option pricing function is not new in itself. Moreover, splines have a long tradition in the statistical literature on smoothing under shape constraints like positivity, monotonicity, and convexity; see Delecroix and Thomas-Agnan (2000) for a survey. Two main avenues for constraints implementation in spline spaces can be distinguished. The first exploits the fact that in the considered spaces the shape constraints can be represented by a finite number of linear inequality constraints to achieve the desired shape constraint globally; see, inter alia, Hildreth (1958), Brunk (1970), Dierckx (1980), Ramsay (1988), He and Shi (1998), and Meyer (2008). Alternatively, one seeks only approximately to satisfy the constraints on a finite subset of the domain of the function; see, e.g., Villalobos and Wahba (1987) and Mammen and Thomas-Agnan (1999). Both strands exploit specific properties of the spline spaces under consideration and impose the conditions directly on the unknown regression function.[2] For example, to impose convexity in a cubic spline space, one can utilize the linearity of second-order derivatives, which in turn leads to conditions on the coefficients of the spline.

Our approach differs. We do not impose the no-arbitrage shape constraints directly on the unknown regression function, but derive sufficient conditions for no-arbitrage on the *control net* of the tensor product (TP) B-spline. The notion of the control net, which is a set of points with certain averages of the knot sequences as abscissae and the B-spline coefficients as ordinates, originates in the literature on computer-aided geometric design; see Prautzsch et al. (2002). In some senses, the control net spans the shape of the TP spline surface. As an important property, it is independent of the degree of the B-spline, thereby allowing B-splines of arbitrary degree to be used for estimation (within numerical limitations). This is in contrast to the aforementioned literature that seldom generalizes beyond the polynomial degree for which it is developed. Nevertheless, our no-arbitrage conditions on the control net are linear. Constrained estimation can therefore be carried out by standard quadratic programming techniques. As we will discuss in detail in Section 4.3, apart from mechanically forcing a polynomial call-option price surface to be free of arbitrage, our conditions also have an economic interpretation. This is because they embed as a special case the findings of Carr and Madan (2005) and Davis and Hobson (2007), which apply to the linear call-option price surface.

A natural question is why we should use polynomial regression splines rather than kernel methods like Aït-Sahalia and Duarte's (2003) local polynomial method, and why specifically B-splines. Concerning the first point, it is difficult to argue that one approach is uniformly better than the other. The local polynomial method fits a polynomial in a local neighborhood around a given point and the estimate is given by a sequence of such fits. Consequently, the asymptotic behavior of the kernel estimator at a point is very well understood; see Fan (1992, 1993). For regression splines, one fits a piecewise polynomial by minimizing a global loss criterion. This makes it hard to obtain precise asymptotic bias expressions and could be seen as a disadvantage; see Zhou et al. (1998) and Huang (2003). On the other hand, regression splines and their derivatives are exhaustively characterized by the coefficients in a basis expansion. Since the number of basis functions is smaller than the sample size, one obtains a complete yet parsimonious summary of the underlying data. This feature may be of practical importance. As we also confirm in our simulations and empirical applications, in most practical situations, however, the two methods are likely to deliver similar estimates. This observation has a theoretical justification since a spline estimator can be shown to be asymptotically equivalent to a certain kernel smoother; see Silverman (1984) and Huang and Studden (1993).

With regard to the second question, B-splines are attractive among the polynomial splines, since they are compactly supported functions that form a basis for a polynomial spline space with a given degree, smoothness, and domain partition. Moreover, by means of the de Boor recursion formula, a stable algorithm for evaluating splines in B-spline form is available; see Appendix A. Hence, B-splines behave in a numerically favorable way and are easy to

[1] This short overview is far from complete. Most importantly, we omit the fully parametric models; see Jackwerth (2004) for more references.

[2] All of the studies cited above, which use polynomial splines to estimate a shape-constrained option pricing function or state-price density, can be assigned to either strand.

handle. The main contribution of our work is that with the TP B-spline we are able to describe the problem of estimating an entire arbitrage-free call-option price surface within a single quadratic program with linear inequality constraints.

The study is structured as follows. The next section gives an overview on no-arbitrage constraints on the call-option price surface. In Section 3, we present the TP B-spline estimator and study consistency. Least squares implementation under sufficient no-arbitrage constraints, optimal knot placement and the selection of the parameter that controls the amount of regularization are discussed in Section 4. In Section 5, we investigate the efficiency gains to be expected for the estimation of option price surfaces and state-price densities under the full set of no-arbitrage constraints. Section 6 demonstrates the estimator, and Section 7 concludes. The Appendix offers basic facts about B-splines, details the consistency proofs, and presents supplementary results on monotonicity-constrained TP B-spline surfaces.

2. The call-option price surface

2.1. Overview and notation

We are given a probability space $(\Omega, \mathcal{F}, \mathrm{P})$ with a filtration $(\mathcal{F}_t)$, to which a stock price process S is adapted. By the fundamental theorem of asset pricing, no-arbitrage is equivalent to the existence of a risk-neutral probability measure Q equivalent to P under which the discounted stock price process is a martingale (Delbean and Schachermeyer, 1994). One can therefore write the price of a call-option with European exercise style as a discounted expectation of its terminal payoff, i.e.,

$$\begin{aligned} C(S_t, K, \tau, r, \delta) &= e^{-r\tau}\, \mathrm{E}_t[(S_T - K)^+] \\ &= e^{-r\tau} \int_0^\infty (S_T - K)^+ \, d\,\mathrm{Q}_\tau(S_T), \end{aligned} \tag{1}$$

where S_t is the time-t underlying asset price, K the exercise price, T the expiry date, $\tau = T - t$ time-to-maturity, r is the risk-free interest rate and δ the dividend yield of the asset. We assume r and δ to be constant for ease of exposition, but we emphasize that all arguments of Sections 2.2 and 2.3 hold true when both are deterministic functions of time. We denote the call payoff by $(u)^+ = \max(u, 0)$. The expectation operator E_t is with respect to Q_τ, i.e., the restriction of Q on $\mathcal{F}_t$. Q_τ is the time-t risk-neutral pricing probability measure for time T and may depend, apart from the variables S_t, τ, r and δ, on further $\mathcal{F}_t$-adapted state variables, such as latent volatility factors, which we suppress here.

The call-option price $C(S_t, K, \tau, r, \delta)$ is a function in five variables. Focusing attention to the two variables K and τ, one can consider the option price as a surface extending across strikes and time-to-maturities. For C to qualify as a call-option price surface, it must obey shape constraints which are dictated by no-arbitrage theory.

In Section 2.2, we summarize the necessary and sufficient no-arbitrage constraints on a European style call-option price surface for future reference.[3] The focus is on a setting with zero interest rates and zero dividends. In Section 2.3, allowing for the more general setting with non-zero interest rates and dividend yields, we introduce a transformation of call-option prices under a dimension reduction and a homogeneity assumption. The transformation allows us to express the initial five-dimensional estimation problem as a two-dimensional one. It also facilitates the implementation of no-arbitrage constraints in the more general setting.

2.2. No-arbitrage conditions for the call-option price surface

Across the strike dimension, the necessary no-arbitrage constraints for the call-option price function have been known for a long time; see, e.g., Aït-Sahalia and Duarte (2003) for a thorough account. Reiner (2000) and Gatheral (2004) are the first to obtain necessary constraints for no calendar arbitrage for European-style options; see Fengler (2009). A more recent literature establishes that these constraints are also sufficient for no-arbitrage. Carr and Madan (2005) and Davis and Hobson (2007) investigate under which conditions a set of discretely observed call-option price quotes is free of arbitrage. Roper (2010) studies the case of a continuous call-option price surface. In these three works, it is shown that the conditions imply the existence of a non-negative Markov martingale such that any element of the surface can be written as an expected value of a call on this martingale. Hence, according to the fundamental theorem of asset pricing, trading of these options does not admit arbitrage.

Since we are interested in estimating a continuous call-option price surface, we summarize the results for no-arbitrage conditions by drawing on Roper's Theorem 2.1. Similarly, we assume that $r = \delta = 0$. We denote by $C(K, \tau)$ a call price surface suppressing all other arguments. We assume that $C(K, \tau)$ is observed at time t and associated with a known stock price S_t.

Proposition 2.1. *Let $S_t > 0$. Define the function*

$$C : [0, \infty) \times [0, \infty) \to \mathbb{R}$$

such that $C(K, \tau)$

(C1) *is convex in K for all $\tau \geq 0$,*
(C2) *has bounds*

$$(S_t - K)^+ \leq C(K, \tau) \leq S_t,$$

for all $K \geq 0$, $\tau \geq 0$,
(C3) *has expiry value $C(K, 0) = (S_t - K)^+$ for all K,*
(C4) *satisfies $\lim_{K \to \infty} C(K, \tau) = 0$ for all τ,*
(C5) *and is non-decreasing in τ for all $K \geq 0$.*

Then

(*i*) *there exists a family of measures Q_τ indexed by τ that admit the existence of a non-negative Markov martingale M such that*

$$C(K, \tau) = \mathrm{E}[(M_T - K)^+ | M_t = S_t]$$

for all $K, \tau \geq 0$;
(*ii*) *under the assumption that M is a non-negative martingale, properties (C1)–(C5) are necessary properties of C to be a conditional expectation of a call option on M.*

Proof. Theorem 2.1 in Roper (2010). ■

The assumptions (C1)–(C5) are weak; for instance, the homogeneity assumption introduced in Section 2.3 is not a consequence of these assumptions. Furthermore, the measures Q_τ may not admit continuous densities. Under the assumption, however, that C is twice continuously differentiable[4] with respect to K for all $\tau > 0$, it is evident that (C1), (C2), and (C4) imply

$$-1 \leq \partial C / \partial K \leq 0 \tag{2}$$

$$\text{and} \quad \partial^2 C / \partial K^2 \geq 0. \tag{3}$$

This yields the familiar representations of the no-arbitrage conditions; see Aït-Sahalia and Duarte (2003). Moreover, $\partial^2 C / \partial K^2$ can be associated with the transition density (state price density) of M; see Section 5.2.

[3] Using the put–call parity it is easy to obtain the corresponding put constraints.

[4] By the properties of convex functions, C is already differentiable for all $K \in (0, \infty)$ for $\tau > 0$.

2.3. A transformation under homogeneity and dimension reduction

The assumption of zero interest rates and zero dividends is not realistic and therefore difficult to maintain. We now introduce a transformation under a homogeneity and a dimension reduction assumption that allows us to accommodate non-zero interest rates and dividend yields. Let $F_t^T = e^{(r-\delta)(T-t)}S_t$ denote the forward price for delivery of the underlying asset at time T.

(C6) For some function f, it holds that

$$C(S_t, K, \tau, r, \delta) = e^{-r\tau} f(F_t^T, K, \tau).$$

(C7) The call price function is homogeneous of degree one in the strike and the forward, i.e.,

$$f(\alpha F_t^T, \alpha K, \tau) = \alpha f(F_t^T, K, \tau)$$

for every positive scalar α.

Because of (C6), the option price is not a function of (S_t, r, δ) separately, but decomposes multiplicatively into the discount factor $e^{-r\tau}$ and a function f that depends on these variables only through F_t^T; this follows Aït-Sahalia and Lo (1998). A homogeneity property in the strike and the contemporaneous underlying asset price is advocated in Merton (1973) as a natural property of an option pricing function. In (C7), we strengthen this property to hold for the forward, but with deterministic interest rates and dividend yields this is not restrictive. Following Renault (1997, Proposition 2), a necessary and sufficient condition for homogeneity is that under the pricing measure future returns and the contemporaneous asset price (or the forward) are independent conditionally on the current information set. This holds for many option pricing models, such as standard stochastic volatility and jump–diffusion models.[5]

With these assumptions, we are ready to define a transformed call-option price surface as

$$z(x, \tau) \equiv \frac{e^{r\tau} C(S_t, K, \tau, r, \delta)}{F_t^T} \tag{4}$$

$$= \frac{f(F_t^T, K, \tau)}{F_t^T} = f(1, x, \tau)\ , \tag{5}$$

where we denote the strike measured in forward moneyness terms by $x = K/F_t^T$. The second line follows from (C6) and (C7) by letting $\alpha = 1/F_t^T$. It clarifies that in contrast to the original option price surface, z only involves the two variables forward moneyness x and time-to-maturity τ.[6]

It is essential to realize that z has a succinct interpretation as a claim on a certain relative price process of the underlying asset defined by $M_T = S_T/F_t^T$. To see this apply the transformation to the right-hand side of (1). This yields

$$\begin{aligned} z(x, \tau) &= \int_0^\infty \left(\frac{S_T}{F_t^T} - x\right)^+ d\,\mathrm{Q}_\tau\,(S_T) \\ &= \int_0^\infty (M_T - x)^+\ d\,\mathrm{Q}_\tau\left(M_T F_t^T\right) \\ &= \mathrm{E}_t[(M_T - x)^+]\ , \end{aligned} \tag{6}$$

where the second line follows from the variable substitution $S_T = M_T F_t^T$. In consequence, $z(x, \tau)$ is the price of a call-option on M_T struck at strike x. The process M itself can be interpreted as the "forward moneyness process relative to time t". It has initial price $M_t = 1$ and is a martingale under the risk-neutral measure, since for all times t', such that $t \le t' \le T$, it holds that

$$\mathrm{E}_{t'}[M_T] = \mathrm{E}_{t'}[S_T/F_t^T] = \frac{e^{(r-\delta)(T-t')}S_{t'}}{e^{(r-\delta)(T-t)}S_t} = \frac{S_{t'}}{F_t^{t'}} = M_{t'}, \tag{7}$$

where the second step follows from the fact that the discounted stock price process with continuously reinvested dividends is a martingale under the risk-neutral measure. By the law of iterated expectations, we have furthermore that $\mathrm{E}[M_T] = 1$.

Since the transformation is time-t measurable, there is no loss in information about the pricing measure Q_τ. Hence estimating C and z is equivalent. Yet the transformed price surface makes the implementation of no-arbitrage constraints simpler. Indeed, Eq. (6) can be read as a call-option valuation relationship with zero interest rates, zero dividends, and an initial spot price equal to one. We can therefore apply the theory provided in Section 2.2 to model z.

To appreciate this fact fully, consider the case of the monotonicity constraints for two $\tau_1 < \tau_2$. By (C5), we must have $z(x, \tau_1) \le z(x, \tau_2)$. Inserting (4), we obtain

$$\frac{e^{r\tau_1} C(S_t, K_1, \tau_1, r, \delta)}{F_t^{T_1}} \le \frac{e^{r\tau_2} C(S_t, K_2, \tau_2, r, \delta)}{F_t^{T_2}} \tag{8}$$

$$\Leftrightarrow \quad C(S_t, K_1, \tau_1, r, \delta) \le e^{\delta(T_2-T_1)} C(S_t, K_2, \tau_2, r, \delta) \tag{9}$$

where the forward-moneyness relation implies that the two absolute strikes are given by $K_2 = K_1 F_t^{T_2}/F_t^{T^1} = K_1 e^{(r-\delta)(T_2-T_1)}$. Eq. (9) derives the monotonicity constraint for European call-options in the presence of non-zero interest rates and dividend yields; see, e.g., Eq. (7) in Fengler (2009). Note that the previous axial monotonicity constraint is now a monotonicity constraint along a nonlinear path of strikes interrelated by forward moneyness. It is hard to implement such a constraint on an estimator.

The original call-option price function can be obtained by applying the reverse transformation

$$C(S_t, K, \tau, r, \delta) = e^{-r\tau} F_t^T z(x, \tau). \tag{10}$$

Furthermore, if the option price surface is sufficiently smooth, by the chain rule of differentiation, we obtain the relations

$$\frac{\partial C}{\partial K} = e^{-r\tau}\frac{\partial z}{\partial x} \tag{11}$$

$$\text{and} \quad \frac{\partial^2 C}{\partial K^2} = e^{-r\tau}\frac{1}{F_t^T}\frac{\partial^2 z}{\partial x^2}\ . \tag{12}$$

The last expression can be exploited for state-price density estimation; see Sections 5.2 and 6.2.2.

3. The tensor-product B-spline call-option price surface

3.1. Notation and definition of the estimators

We now consider our model for the call-option price surface. Suppose we observe a sample of call-option prices $\{(X_i, Y_i), Z_i\}_{i=1}^n$, where Z_i is a call-option price, which is submitted to the transformation in (4), X_i is its relative strike (expressed in terms of forward moneyness) and Y_i is its time-to-maturity. The model is given by

$$Z_i = z_0(X_i, Y_i) + \varepsilon_i, \tag{13}$$

where $z_0 : [\underline{x}, \overline{x}] \times [\underline{y}, \overline{y}] \to (0, 1]$ is the unknown call-option price surface that is assumed to be free of arbitrage opportunities. The collection of relative strikes and time-to-maturities $\{(X_i, Y_i)\}_{i=1}^n \in [\underline{x}, \overline{x}] \times [\underline{y}, \overline{y}]$ (with lower and upper bounds $\underline{x} \ll \overline{x}$ and $\underline{y} \ll \overline{y}$) are considered as random design points, ε_i is an error term with mean zero and finite variance, and $\{(X_i, Y_i), Z_i\}_{i=1}^n$ are assumed to be i.i.d. as $\big((X, Y), Z\big)$.

[5] This assumption does not exclude leverage effects. Further implications and discussions of the homogeneity property are found in Garcia and Renault (1998a,b), Bates (2005), Alexander and Nogueira (2007), and Garcia et al. (2010, Section 2).

[6] This transformation also appears in Davis and Hobson (2007), but without reference to the homogeneity property.

The purpose is to approximate the functional parameter of interest $z_0(x, y) = \mathrm{E}[Z|x, y]$ by a two-dimensional polynomial spline that is an element in the space

$$\mathcal{S}(p_1, p_2; \boldsymbol{\xi}, \boldsymbol{\upsilon}) = \mathcal{S}(p_1; \boldsymbol{\xi}) \otimes \mathcal{S}(p_2; \boldsymbol{\upsilon}).$$

This space is constructed by the tensor product of two univariate polynomial spline spaces $\mathcal{S}(p_1; \boldsymbol{\xi})$ and $\mathcal{S}(p_2; \boldsymbol{\upsilon})$, where each is defined over the forward moneyness and the time-to-maturity dimension, respectively. We define

$$\mathcal{S}(p_1; \boldsymbol{\xi}) = \{s \in \mathcal{C}^{p_1-1} : s \text{ is a polynomial of degree } p_1 \text{ on each (nonempty) subinterval in } \boldsymbol{\xi}\},$$

where $\mathcal{C}^p$ is the set of p-times continuously differentiable functions and $\boldsymbol{\xi}$ a knot sequence with (p_1+1)-fold coalescent boundary knots of the form

$$\underline{x} = \xi_0 = \cdots = \xi_{p_1} < \xi_{p_1+1} < \cdots < \xi_{q_1} < \xi_{q_1+1} = \cdots = \xi_{q_1+p_1+1} = \bar{x}, \tag{14}$$

defined for some integer $q_1 > p_1$. For two integers $q_2 > p_2 > 0$, $\mathcal{S}(p_2; \boldsymbol{\upsilon})$ is defined likewise, where $\boldsymbol{\upsilon}$ denotes the second knot sequence given by

$$\underline{y} = \upsilon_0 = \cdots = \upsilon_{p_2} < \upsilon_{p_2+1} < \cdots < \upsilon_{q_2} < \upsilon_{q_2+1} = \cdots = \upsilon_{q_2+p_2+1} = \bar{y}. \tag{15}$$

It is well known that $\mathcal{S}(p_1; \boldsymbol{\xi})$ and $\mathcal{S}(p_2; \boldsymbol{\upsilon})$ are spanned by a set of non-negative basis functions $B_{j_1,p_1}(x), j_1 = 0, \ldots, q_1$, and $B_{j_2,p_2}(y), j_2 = 0, \ldots, q_2$, respectively, the so-called B-splines of degree p_1 and p_2 that are defined over $\boldsymbol{\xi}$ and $\boldsymbol{\upsilon}$; see de Boor (2001) and, for some basic facts on B-splines, Appendix A. In adopting $(p_k + 1)$-fold coincident boundary knots, $k = 1, 2$, the support of the basis functions is contained in $[\underline{x}, \bar{x}]$ and $[\underline{y}, \bar{y}]$. It then follows that $\mathcal{S}(p_1, p_2; \boldsymbol{\xi}, \boldsymbol{\upsilon})$ has basis functions $B_{j_1,p_1}(x)B_{j_2,p_2}(y)$, $j_1 = 0, \ldots, q_1, j_2 = 0, \ldots, q_2$, that form a partition of unity and are non-negative on $[\underline{x}, \bar{x}] \times [\underline{y}, \bar{y}]$. Any $s \in \mathcal{S}(p_1, p_2; \boldsymbol{\xi}, \boldsymbol{\upsilon})$ can be written as

$$s(x, y) = \sum_{j_1=0}^{q_1} \sum_{j_2=0}^{q_2} \theta_{j_1,j_2} B_{j_1,p_1}(x) B_{j_2,p_2}(y) \tag{16}$$

with coefficients $\theta_{j_1,j_2} \in \mathbb{R}$. The dimension of $\mathcal{S}(p_1, p_2; \boldsymbol{\xi}, \boldsymbol{\upsilon})$ is given by $J_n = (q_1 + 1)(q_2 + 1)$. With the notation J_n, we stress that as n grows we regard finer and finer approximations to the call-option price surface.

We now introduce several estimators for z_0. An unconstrained estimate of the call price surface $\hat{z}$ is defined by means of the least squares criterion

$$\hat{z} = \underset{s \in \mathcal{S}(p_1,p_2;\boldsymbol{\xi},\boldsymbol{\upsilon})}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (Z_i - s(X_i, Y_i))^2 . \tag{17}$$

This estimate is likely to violate no-arbitrage restrictions. Following the theory presented in Section 2, an arbitrage-free estimate would be given by

$$\hat{z}^C = \underset{s \in \mathcal{S}^C(p_1,p_2;\boldsymbol{\xi},\boldsymbol{\upsilon})}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (Z_i - s(X_i, Y_i))^2 , \tag{18}$$

where $\mathcal{S}^C(p_1, p_2; \boldsymbol{\xi}, \boldsymbol{\upsilon}) = \{s : s \in \mathcal{S}(p_1, p_2; \boldsymbol{\xi}, \boldsymbol{\upsilon}),\ 0 \le s \le 1,\ -1 \le \partial s/\partial x \le 0,\ \partial^2 s/\partial x^2 \ge 0,\ \partial s/\partial y \ge 0\}$. Note that in the definition of $\mathcal{S}^C$ we replace the strict inequality in the lower bound on s by a weak inequality to avoid an open interval.[7]

As we discuss below, because of the special structure of option data, it may be that (17) and (18) will lead to ill-posed optimization problems. We therefore additionally introduce the regularized estimators

$$\hat{z}_\lambda = \underset{s \in \mathcal{S}(p_1,p_2;\boldsymbol{\xi},\boldsymbol{\upsilon})}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (Z_i - s(X_i, Y_i))^2 + \lambda_n |\boldsymbol{\theta}|^2 \tag{19}$$

$$\text{and} \quad \hat{z}^C_\lambda = \underset{s \in \mathcal{S}^C(p_1,p_2;\boldsymbol{\xi},\boldsymbol{\upsilon})}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (Z_i - s(X_i, Y_i))^2 + \lambda_n |\boldsymbol{\theta}|^2, \tag{20}$$

which are well-defined for a sequence of non-negative numbers λ_n. Here, $|\cdot|$ denotes the Euclidean vector norm, and $\boldsymbol{\theta}$ is the vector comprising all B-spline coefficients θ_{j_1,j_2}, for $j_1 = 0, \ldots, q_1, j_2 = 0, \ldots, q_2$. In (19) and (20) the function $|\boldsymbol{\theta}|$ acts as a penalty term that improves the conditioning of the estimation problem, and the sequence λ_n controls the weight that is given to this term. Regularization by penalization follows the idea of ridge regression and has a long tradition in the literature on splines and ill-posed problems; see Chapter 8 in Wahba (1990) and references cited therein. The regularization considered here differs from the penalized B-splines suggested in Eilers and Marx (1996a), which build on penalties constructed from differences in the coefficients of the B-splines.

3.2. Large sample properties of the shape-restricted estimator

Large sample properties of unconstrained series estimators based on B-splines are well understood and have been studied by Agarwal and Studden (1980), Gallant (1982), Andrews (1991), Stone (1994), Newey (1997), Zhou et al. (1998) and Huang (2003); see Chen (2007) for an exhaustive overview. Monotone regression estimators are explored in Brunk (1970), Wright (1981) and Mammen (1991a). Hanson and Pledger (1976) and Mammen (1991b) establish consistency and rates of convergences for certain concave/convex regression functions, and Groeneboom et al. (2001a,b) derive the limiting distribution of a convex function at a fixed point. As a general result of this literature, under convexity, a constrained estimator *may* achieve faster rates of convergence than an unconstrained estimator. This is proved in Mammen and Thomas-Agnan (1999) for smoothing splines; see also Meyer (2008). Moreover, under monotonicity as well as under convexity the limit distributions are involved and non-standard.

Given the multiplicity of constraints in the present estimation problem (positivity, monotonicity and convexity), a complete analysis of the large sample properties of the estimator is beyond the scope of the paper. We therefore only consider the consistency of our B-spline estimator.

3.2.1. Assumptions

We adopt the following assumptions:

(A1) $\{(X_i, Y_i), Z_i\}_{i=1}^n$ is i.i.d. as $\big((X, Y), Z\big)$, has compact support and a density that is absolutely continuous and bounded away from zero and infinity.

(A2) $\sigma^2(x, y) = \mathrm{Var}[Z|x, y]$ is bounded and bounded away from zero.

(A3) $z_0(x, y) = \mathrm{E}[Z|x, y]$ is r-smooth[8] for some $r \ge 3$.

(A4) As $n \to \infty, J_n \to \infty$ and $J_n^2/n \to 0$.

(A5) The knot sequences $\boldsymbol{\xi}$ and $\boldsymbol{\upsilon}$ have a bounded mesh ratio.[9]

[7] Strictly speaking, $s = 0$ leads to an arbitrage. Nevertheless, this is immaterial to the following discussions, since we can always set the lower bound to some small $\epsilon > 0$ thanks to the compactness of the support.

[8] Let k denote an integer and set $r = k + \beta$, $0 < \beta \le 1$. A function f on $[\underline{x}, \bar{x}] \times [\underline{y}, \bar{y}]$ is said to be r-smooth if f is k times continuously differentiable and $D^{\alpha_1,\alpha_2} f = \partial^{\alpha_1+\alpha_2} f / \partial^{\alpha_1} x \partial^{\alpha_2} y$ satisfies the Hölder condition of order β,

$$|D^{\alpha_1,\alpha_2} f(x_2, y_2) - D^{\alpha_1,\alpha_2} f(x_1, y_1)| \le \gamma \left((x_2 - x_1)^2 + (y_2 - y_1)^2\right)^{\beta/2},$$

where γ is some finite constant for any $\alpha_1 + \alpha_2 = r - \beta$ with $k = \alpha_1 + \alpha_2$.

[9] The knot sequence $\boldsymbol{\xi}$ is said to have bounded mesh ratio if

$$\frac{\max_{p \le j \le q+1} (\xi_{j+1} - \xi_j)}{\min_{p \le j \le q+1} (\xi_{j+1} - \xi_j)} \le \gamma$$

for some constant γ (since boundary knots are coinciding, they are excluded here).

Assumption (A1) is standard for series estimators and required for identification; see Chen (2007, p. 5602). The compactness of the support could be given up, but then the asymptotic discussion would be dominated by the tail behavior of the design. We therefore keep it for simplicity.[10] The continuity assumption regarding the design density may be critical for estimates which are obtained from the data of a single trading day. This is because only a discrete set of expiry days is typically listed on a futures exchange. It should be noted, however, that for most underlying assets there exist liquid broker markets, on which other expiries are quoted over-the-counter. If the goal is to estimate a functional relationship that is assumed to be stable across time, as in Aït-Sahalia and Lo (1998), this issue is not relevant, since estimation can be based on a long time-series of cross-sections of option price data. (A2) allows for heteroscedasticity, e.g., induced by a bid–ask spread that varies in x and y. As described in Section 3.1, however, we study convergence under a simple L_2-norm only and therefore will not look for efficient estimates that take care of heteroscedasticity. Assumption (A3) is a typical smoothness assumption made for series estimators; $r \geq 3$ ensures the existence of a family of continuous state price densities. (A4) is owed to Newey (1997), but can be weakened to $J_n \log n/n \to 0$; see Huang (2003). (A5) is a regularity condition which excludes pathological knot sequences; see Condition A.2 in Huang (2003) for further discussion.

3.2.2. Consistency

To study consistency and the rates of $\hat{z}$ and $\hat{z}_\lambda$, we interpret the estimation problem within the general projection framework for constrained smoothing suggested in Mammen et al. (2001). In this framework, Mammen et al. (2001) show that constrained smoothing is equivalent to two iterative projections in suitably defined vector spaces. First, data Z_i are projected on a vector space of unconstrained candidate regression functions which yields $\hat{z}$; in a second step, $\hat{z}^C$ is obtained by projecting $\hat{z}$ on a subspace with constrained candidate regression functions. The framework, which is described in some detail in Appendix B, is applicable whenever the unconstrained estimate can be written as a weighted average of the Z_i, i.e., as $\hat{z}(x,y) = \sum_{i=1}^n w_i(x,y)Z_i$ where $w_i \geq 0$ are some weights. For the TP B-splines this holds true, since the estimated B-spline coefficients $\hat{\theta}_{j_1,j_2}$ are determined through least squares, and therefore are linear functions of the Z_i.

The following proposition, which relies on $\mathcal{S}^C(p_1,p_2;\boldsymbol{\xi},\boldsymbol{\upsilon})$ being a convex set, is a consequence of interpreting constrained smoothing as two consecutive projections.

Proposition 3.1. *Suppose* $\mathcal{S}(p_1,p_2;\boldsymbol{\xi},\boldsymbol{\upsilon}) = \mathcal{S}(p_1;\boldsymbol{\xi}) \otimes \mathcal{S}(p_2;\boldsymbol{\upsilon})$ *such that* $p_1, p_2 \geq r-1$. *Under Assumptions* (A1)–(A5) *we have*

$$\left(\frac{1}{n}\sum_{i=1}^n (\hat{z}^C(X_i,Y_i) - z_0(X_i,Y_i))^2\right)^{1/2}$$
$$\leq \left(\frac{1}{n}\sum_{i=1}^n (\hat{z}(X_i,Y_i) - z_0(X_i,Y_i))^2\right)^{1/2} + O_P(J_n^{-r}) \tag{21}$$
$$= O_P\left(\left(\frac{J_n}{n} + J_n^{-r}\right)^{1/2}\right) + O_P(J_n^{-r}). \tag{22}$$

For the regularized estimator, let $\lambda_n \to 0$ *such that* $\lambda_n|\hat{\boldsymbol{\theta}}_z - \boldsymbol{\theta}_{z_0^*}|^2 = O_P\left(J_n/n + J_n^{-r}\right)$, *where* $\boldsymbol{\theta}_{z_0^*}$ *is the B-spline weight vector of an approximation* z_0^* *in* $\mathcal{S}(p_1,p_2;\boldsymbol{\xi},\boldsymbol{\upsilon})$ *to the true regression function* z_0 *such that* $\left(\frac{1}{n}\sum_{i=1}^n (z_0^*(X_i,Y_i) - z_0(X_i,Y_i))^2\right)^{1/2} = O_P(J_n^{-r})$; *see Appendix B for further details on* z_0^*. *We then obtain*

$$\left(\frac{1}{n}\sum_{i=1}^n (\hat{z}_\lambda^C(X_i,Y_i) - z_0(X_i,Y_i))^2\right)^{1/2}$$
$$\leq \left(\frac{1}{n}\sum_{i=1}^n (\hat{z}_\lambda(X_i,Y_i) - z_0(X_i,Y_i))^2\right)^{1/2}$$
$$+ O_P\left((J_n/n + J_n^{-r})^{1/2}\right) \tag{23}$$
$$= O_P\left(\left(\frac{J_n}{n} + J_n^{-r}\right)^{1/2}\right). \tag{24}$$

Proof. See Appendix B. ■

The intuition of this result is evident. In projecting the unconstrained estimate on a (sub)space obeying the required shape constraints, we ought to be at least as good as (if not better than) in the first projection step, when the constraints are ignored. Hence, the constrained estimator must attain at least the convergence rates of the unconstrained one. Provided that λ_n converges sufficiently fast, similar considerations apply to $\hat{z}_\lambda^C$. Given Proposition 3.1, one cannot choose an optimal rate of λ_n. Yet, as we will demonstrate in Section 4.5, for the estimation of option price surfaces the selection of the regularization parameter has little influence on the estimate since strong shape constraints like convexity provide a lot of smoothing. For practical use of the estimator, we therefore suggest choosing a small but otherwise arbitrary value.

In Proposition 3.1, we make use of the fact that for the unconstrained estimates, it holds that $\left(\frac{1}{n}\sum_{i=1}^n (\hat{z}(X_i,Y_i) - z_0(X_i,Y_i))^2\right)^{1/2} = O_P\left(J_n/n + J_n^{-r}\right)$. In this expression, $O_P(J_n/n)$ is induced by the variance and $O_P(J_n^{-r})$ by the squared bias; see Newey (1997). Since the remaining terms in (22) and (24) converge at the same rate or faster, they are asymptotically negligible. Optimally balancing the variance and the squared bias term, i.e., choosing J_n such that $J_n/n \asymp J_n^{-r}$, one obtains the rate $n^{-r/(r+1)}$. This is the optimal rate of convergence uniformly over the class of r-smooth functions; see Stone (1982). In consequence, our constrained estimators – whether regularized or not – attain at least these optimal rates.

4. Least squares fitting under linear no-arbitrage constraints

4.1. Quadratic programming

Given a sample $\{(X_i,Y_i),Z_i)\}_{i=1}^n$, we denote the $(q_1+1)(q_2+1)\times 1$ vector of approximating functions by $\mathbf{b}_{(p_1,p_2)}(x,y) = \left(\mathbf{b}_{p_1}(x) \otimes \mathbf{b}_{p_2}(y)\right)$, where $\mathbf{b}_{p_1}(x) = (B_{0,p_1}(x),\ldots,B_{q_1,p_1}(x))^\top$ and $\mathbf{b}_{p_2}(y) = (B_{0,p_2}(y),\ldots,B_{q_2,p_2}(y))^\top$, and by $\mathbf{B}$ the $n\times(q_1+1)(q_2+1)$ matrix whose ith row is given by $\mathbf{b}_{(p_1,p_2)}(X_i,Y_i)^\top$, i.e.,

$$\mathbf{B} = \left(\mathbf{b}_{(p_1,p_2)}(X_1,Y_1),\ldots,\mathbf{b}_{(p_1,p_2)}(X_n,Y_n)\right)^\top. \tag{25}$$

Furthermore, let $\mathbf{z}$ be the $n\times 1$ vector of observed relative call-option prices and $\boldsymbol{\theta}$ the $(q_1+1)(q_2+1)\times 1$ vector of unknown coefficients. Then the finding of the minimizers of (18) and (20) can be cast as the quadratic program

$$\widehat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}} \frac{1}{2}\boldsymbol{\theta}^\top\mathbf{D}\boldsymbol{\theta} - \boldsymbol{\theta}^\top\mathbf{d},$$
$$\text{subject to} \quad \hat{z}_\lambda^C = \mathbf{b}^\top\widehat{\boldsymbol{\theta}} \in \mathcal{S}^C(p_1,p_2;\boldsymbol{\xi},\boldsymbol{\upsilon}) \tag{26}$$

where $\mathbf{D} = \mathbf{B}^\top\mathbf{B} + \lambda\mathbf{I}$, $\mathbf{d} = \mathbf{B}^\top\mathbf{z}$, and $\mathbf{I}$ is the unit matrix. Because the sample size is fixed, we drop the explicit dependence of λ

[10] Infinitely supported B-splines exist; see Corlay (2013) on estimating option price functions and state-price densities.

on n for the remaining part of this work. We set $\lambda > 0$ for the regularized estimator, and $\lambda = 0$ otherwise. In the following, we will show that it is possible to formalize the condition $\hat{z}^C_\lambda \in \mathcal{S}^C(p_1, p_2; \boldsymbol{\xi}, \boldsymbol{\upsilon})$ in terms of a set of linear inequality constraints such that (26) can be solved by means of quadratic programming such as the Goldfarb and Idnani (1983) algorithm.

The quadratic program (26) is well-posed if $\mathbf{D}$ is strictly positive-definite. Practically, this condition may fail because of the quoting conventions on options markets. In contrast to univariate B-splines, there are unfortunately no conditions of the Schoenberg–Whitney type for multivariate B-spline approximations that would easily allow one to check for rank-deficiency of the collocation matrix $\mathbf{B}$; see Dierckx (1993, Chap. 9.1.2) for a detailed account of this problem. In this case, we recommend the regularized estimator. In Section 4.5 we will discuss the selection of λ in more detail.

4.2. Sufficient conditions for no-arbitrage on the control net

We now discuss how to guarantee that $\hat{z}^C$ and $\hat{z}^C_\lambda$ are elements of $\mathcal{S}^C(p_1, p_2; \boldsymbol{\xi}, \boldsymbol{\upsilon})$. To this end, we develop sufficient no-arbitrage conditions on the control net associated with the knot sequences $\boldsymbol{\xi}$ and $\boldsymbol{\upsilon}$. The control net determines the shape of the spline surface. Since the conditions imposed on the control net do not depend on the degree of the B-spline basis, B-splines of arbitrary polynomial degree can be handled. The constraints are formulated as linear inequalities on the B-spline coefficients θ_{j_1,j_2}. We emphasize that the key to constraint implementation is the selection of coalescent boundary knots in the definitions of the knot sequences (14) and (15).

4.2.1. Control net

For two strictly increasing sequences of abscissae $\xi^*_0, \ldots, \xi^*_{q_1}$ and $\upsilon^*_0, \ldots, \upsilon^*_{q_2}$, the control net of the surface is defined as the unique function $c : [\xi^*_0, \xi^*_{q_1}] \times [\upsilon^*_0, \upsilon^*_{q_2}] \to \mathbb{R}$ which satisfies the interpolation conditions $c(\xi^*_{j_1}, \upsilon^*_{j_2}) = \theta_{j_1,j_2}, j_1 = 0, \ldots, q_1$ and $j_2 = 0, \ldots, q_2$, and which is bilinear in the rectangle

$$R_{j_1,j_2} = [\xi^*_{j_1}, \xi^*_{j_1+1}] \times [\upsilon^*_{j_2}, \upsilon^*_{j_2+1}]. \tag{27}$$

Define $\alpha_{j_1,1}(x) = (x - \xi^*_{j_1})/(\xi^*_{j_1+1} - \xi^*_{j_1})$ and $\alpha_{j_1,0}(x) = 1 - \alpha_{j_1,1}(x)$ and $\beta_{j_2,1}(y) = (y - \upsilon^*_{j_2})/(\upsilon^*_{j_2+1} - \upsilon^*_{j_2})$ and $\beta_{j_2,0}(y) = 1 - \beta_{j_2,1}(y)$. The control net over the rectangle R_{j_1,j_2} is then given by

$$c_{j_1,j_2}(x, y) = \sum_{k=0}^{1} \sum_{\ell=0}^{1} \alpha_{j_1,k}(x)\, \beta_{j_2,\ell}(y)\, \theta_{j_1+k,j_2+\ell}. \tag{28}$$

For TP B-splines, the sequence of abscissae spanning the control net does not coincide with the knot sequences $\boldsymbol{\xi}$ and $\boldsymbol{\upsilon}$, but with certain averages of these known as Greville sites. The Greville sites $\xi^*_{j_1}, j_1 = 0, \ldots, q_1$, and $\upsilon^*_{j_2}, j_2 = 0, \ldots, q_2$, associated with the knot sequences $\boldsymbol{\xi}$ and $\boldsymbol{\upsilon}$ are defined by

$$\xi^*_{j_1} = \frac{\xi_{j_1+1} + \cdots + \xi_{j_1+p_1}}{p_1} \quad \text{and} \quad \upsilon^*_{j_2} = \frac{\upsilon_{j_2+1} + \cdots + \upsilon_{j_2+p_2}}{p_2}. \tag{29}$$

The set of points $(\xi^*_{j_1}, \upsilon^*_{j_2}, \theta_{j_1,j_2}), j_1 = 0, \ldots, q_1, j_2 = 0, \ldots, q_2$ are known as control points of the surface. This interpretation of this set of points derives from the linear precision property of the B-spline which allows one to interpret the spline function in fact as a spline curve; see Appendix A for further details.

Fig. 1 illustrates the univariate case. A cubic B-spline is shown along with a knot sequence ξ_j, the Greville sites ξ^*_j and the univariate control net, a control polygon. As in (14) and (15), we choose $(p + 1)$-coalescent, i.e., four-fold boundary knots for the knot sequence. It is immediately obvious that the spline (blue) mimics the shape of the control polygon (red). Moreover, the following pivotal properties of the univariate B-spline can be seen (Prautzsch et al., 2002, Chapter 5.7): The spline evaluated at the boundaries of the support coincides with the end-points of the control polygon, and the tangents of the spline evaluated at the boundaries of the support coincide with the legs of the control polygon. These two properties hold if coincident boundary knots are chosen for the knot sequence. Finally, since any point of the pth degree spline $(x, s(x))$ where $x \in [\xi_j, \xi_{j+1})$ is an affine and convex combination of the $p + 1$ control points $(\xi^*_j, \theta_j), \ldots, (\xi^*_{j+p}, \theta_{j+p})$, the spline segment $s[\xi_j, \xi_{j+1})$ lies in the convex hull of these points; see Appendix A. For ease of visual perception, the convex hull of the control polygons belonging to the first and the last segment of the spline is shaded.

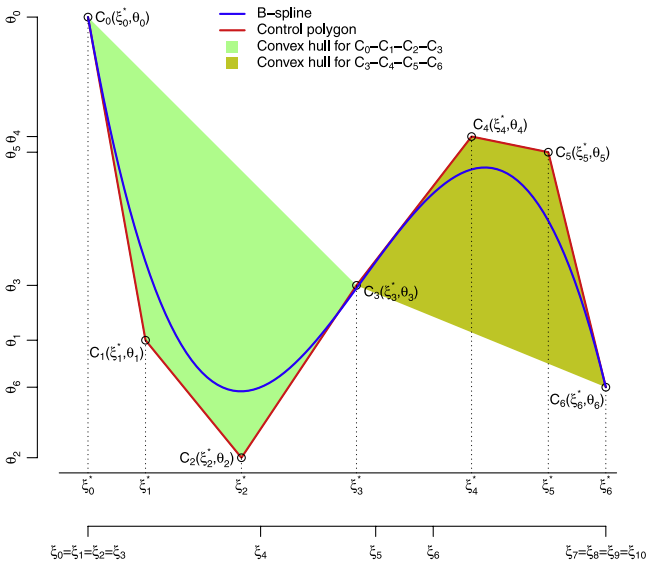


Fig. 1. Illustration of the univariate B-spline (blue line), its knot sequence $\xi_0 = \xi_1 = \xi_2 = \xi_3 < \xi_4, \ldots, \xi_6 < \xi_7 = \xi_8 = \xi_9 = \xi_{10}$ (second horizontal axis), the associated control polygon (broken red line) and Greville sites $\xi^*_0, \ldots, \xi^*_6$. The plot demonstrates that the spline segment $s[\xi_3, \xi_4)$ lies in the convex hull (dark-green shaded area) of its control polygon $c_0 - c_1 - c_2 - c_3$ given by the set of points $\{(\xi^*_0, \theta_0), \ldots, (\xi^*_3, \theta_3)\}$, where θ_i is the weight of the corresponding B-spline basis function. Likewise, the segment $s[\xi_6, \xi_7)$ lies in the convex hull (yellowish shaded area) of the control polygon $c_3 - c_4 - c_5 - c_6$ spanned by the points $\{(\xi^*_3, \theta_3), \ldots, (\xi^*_6, \theta_6)\}$. This property also applies to the segments $s[\xi_4, \xi_5)$ and $s[\xi_5, \xi_6)$ that lie in the convex hull of their respective control polygons (not highlighted). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

For a univariate spline, these properties make the implementation of strike conditions by means of the control polygon immediately obvious. In the following, we show that the intuition gained from Fig. 1 carries over to the tensor product case. We also handle the calendar dimension.

4.2.2. Price bounds

According to Proposition 2.1, no-arbitrage relative forward prices must be non-negative and less than one. We have

Proposition 4.1. *Let* $z(x, y) = \sum_{j_1=0}^{q_1} \sum_{j_2=0}^{q_2} \theta_{j_1,j_2}\, B_{j_1,p_1}(x)\, B_{j_2,p_2}(y)$ *be a TP B-spline defined on the knot sequences* (14) *and* (15)*. If*

$$\theta_{j_1,j_2} \geq 0 \quad \text{and} \quad \theta_{j_1,j_2} \leq 1\,, \tag{30}$$

for $j_1 = 0, \ldots, q_1$ *and* $j_2 = 0, \ldots, q_2$, *then* $0 \leq z(x, y) \leq 1$.

Proof. This is a trivial consequence of the positivity property of B-splines and the fact that TP B-splines form a partition of unity, i.e., $\sum_{j_1=0}^{q_1} \sum_{j_2=0}^{q_2} B_{j_1,p_1}(x)\, B_{j_2,p_2}(y) = 1$ for $(x, y) \in [\underline{x}, \bar{x}] \times [\underline{y}, \bar{y}]$. ■

4.2.3. Bounds on first-order derivatives and convexity

The recursion formula of B-splines owed to de Boor (1972) and Cox (1972) implies that the univariate B-spline over each knot interval is an affine and convex combination of its control points. The univariate B-spline over each knot interval therefore lies in the convex hull of these points (see Fig. 1 and the further explanations and references in Appendix A). A convex set of control points therefore implies a globally convex function. As we show in Proposition 4.2, this also holds true for the TP B-spline.

As explained in Section 2.2, it follows from (C1), (C2), and (C4) and additional smoothness assumptions that first-order derivatives of the call-option price surface must be between minus one and zero; see Eq. (2). To achieve this constraint, we exploit the fact that the slope of a univariate B-spline coincides at the boundary with the slope of the outer legs of the control polygon if the boundary knots of the knot sequences $\boldsymbol{\xi}$ and $\boldsymbol{\upsilon}$ are coincident. In Proposition 4.2, it is shown that this property extends to the TP B-spline. Note that with convexity in the forward moneyness dimension in place, it is sufficient only to constrain the slope at the boundaries of the support, i.e., in $\underline{x}$ and $\bar{x}$ for any y.

Proposition 4.2. *Let* $(\xi^*_{j_1}, \upsilon^*_{j_2}, \theta_{j_1,j_2})$, $j_1 = 0, \ldots, q_1$, $j_2 = 0, \ldots, q_2$, *be the control points associated with the spline* $z(x,y) = \sum_{j_1=0}^{q_1} \sum_{j_2=0}^{q_2} \theta_{j_1,j_2} B_{j_1,p_1}(x) B_{j_2,p_2}(y)$ *defined on the knot sequences* $\boldsymbol{\xi}$ *and* $\boldsymbol{\upsilon}$. *The condition*

$$-1 \leq \frac{\theta_{j_1+1,j_2} - \theta_{j_1,j_2}}{\xi^*_{j_1+1} - \xi^*_{j_1}} \leq \frac{\theta_{j_1+2,j_2} - \theta_{j_1+1,j_2}}{\xi^*_{j_1+2} - \xi^*_{j_1+1}} \leq 0, \tag{31}$$

is a sufficient condition such that $z(x,y)$ *is convex in* x *and it holds that* $-1 \leq \partial z(x,y)/\partial x \leq 0$ *for any* y.

Proof. For any $\check{y} \in [\underline{y}, \bar{y}]$, consider the iso-parameter curve across forward moneyness given by

$$z(x,\check{y}) = \sum_{j_1=0}^{q_1} B_{j_1,p_1}(x) \sum_{j_2=0}^{q_2} \theta_{j_1,j_2} B_{j_2,p_2}(\check{y}) \tag{32}$$

$$= \sum_{j_1=0}^{q_1} B_{j_1,p_1}(x) Q_{j_1}(\check{y}) \tag{33}$$

where $Q_{j_1}(\check{y}) = \sum_{j_2=0}^{q_2} \theta_{j_1,j_2} B_{j_2,p_2}(\check{y})$. This is a univariate spline curve with control points $(\xi^*_{j_1}, Q_{j_1}(\check{y}))$, $j_1 = 0, \ldots, q_1$. Given the above observations preceding the proposition, it is sufficient to show that (31) implies

$$-1 \leq \frac{Q_{j_1+1}(\check{y}) - Q_{j_1}(\check{y})}{\xi^*_{j_1+1} - \xi^*_{j_1}} \leq \frac{Q_{j_1+2}(\check{y}) - Q_{j_1+1}(\check{y})}{\xi^*_{j_1+2} - \xi^*_{j_1+1}} \leq 0. \tag{34}$$

To see this, take any of the interior inequalities from (31) and multiply by $B_{j_2,p_2}(\check{y})$. Summing over all $j_2 = 0, \ldots, q_2$ and noting that $B_{j_2,p_2}(\check{y}) \geq 0$ yields

$$\frac{\sum_{j_2=0}^{q_2} B_{j_2,p_2}(\check{y}) \left(\theta_{j_1+1,j_2} - \theta_{j_1,j_2}\right)}{\xi^*_{j_1+1} - \xi^*_{j_1}} \leq \frac{\sum_{j_2=0}^{q_2} B_{j_2,p_2}(\check{y}) \left(\theta_{j_1+2,j_2} - \theta_{j_1+1,j_2}\right)}{\xi^*_{j_1+2} - \xi^*_{j_1+1}}, \tag{35}$$

which produces the inner set of inequalities. The outer inequality is obtained from the same argument and by noting that B-splines form a partition of unity. The proposition follows from the properties of the univariate B-spline and the associated control polygon. ■

4.2.4. Monotonicity

To derive sufficient conditions for monotonicity along forward moneyness, we need two results obtained by Floater and Peña (1998). Because of their importance, the proofs are given in Appendix C. As preparation, recall that a bivariate function f is called increasing in direction $d = (d_1, d_2) \in \mathbb{R}^2$, if

$$f(x + \gamma d_1, y + \gamma d_2) \geq f(x,y) \quad \text{for } \gamma > 0.$$

The following lemma characterizes an increasing control net by a set of four linear inequalities.

Lemma 4.3. *The control net* c *increases in direction* $d = (d_1, d_2) \in \mathbb{R}^2$ *if and only if for* $j_1 = 0, \ldots, q_1 - 1$ *and* $j_2 = 0, \ldots, q_2 - 1$,

$$d_1 \Delta_1 \theta_{j_1,j_2+\ell} + d_2 \Delta_2 \theta_{j_1+k,j_2} \geq 0, \quad k, \ell \in \{0, 1\} \tag{36}$$

where $\Delta_1 \theta_{j_1,j_2} = (\theta_{j_1+1,j_2} - \theta_{j_1,j_2})/(\xi^*_{j_1+1} - \xi^*_{j_1})$ *and* $\Delta_2 \theta_{j_1,j_2} = (\theta_{j_1,j_2+1} - \theta_{j_1,j_2})/(\upsilon^*_{j_2+1} - \upsilon^*_{j_2})$.

Proof. Lemma 2.1 in Floater and Peña (1998); see Appendix C. ■

The subsequent result states that the TP B-spline basis preserves monotonicity if the control net is increasing.

Theorem 4.4. *Let the abscissae of the control net* c *be the Greville sites (29). If* c *is increasing in any direction* d *in* $\mathbb{R}^2$, *so is the TP B-spline* z *on the rectangle* $[\xi_{p_1}, \xi_{q_1+1}] \times [\upsilon_{p_2}, \upsilon_{q_2+1}]$.

Proof. Theorem 4.1 in Floater and Peña (1998); see Appendix C. ■

Lemma 4.3 and Theorem 4.4 allow the construction of an increasing control net which is monotonic in a direction d. The transformations in Section 2, however, cast monotonicity along a path on the surface into an axial monotonicity requirement, i.e., $d = (0, 1)$. In this case, the four inequalities in Lemma 4.3 collapse and we have

Corollary 4.5. *Let* $(\xi^*_{j_1}, \upsilon^*_{j_2}, \theta_{j_1,j_2})$, $j_1 = 0, \ldots, q_1$, $j_2 = 0, \ldots, q_2$, *be the control points associated with the spline* $z(x,y) = \sum_{j_1=0}^{q_1} \sum_{j_2=0}^{q_2} \theta_{j_1,j_2} B_{j_1,p_1}(x) B_{j_2,p_2}(y)$ *defined on the knot sequences* $\boldsymbol{\xi}$ *and* $\boldsymbol{\upsilon}$. *If*

$$\theta_{j_1,j_2+1} - \theta_{j_1,j_2} \geq 0 \tag{37}$$

for $j_1 = 0, \ldots, q_1$ *and* $j_2 = 0, \ldots, q_2 - 1$, *then* $z(x,y)$ *monotonically increases in* y.

4.3. Economic characterization of no-arbitrage conditions

The sufficiency conditions for no-arbitrage have a remarkable relation to previous theoretical results. Carr and Madan (2005) and Davis and Hobson (2007), hereafter jointly referred to as CMDH, investigate under which conditions a discrete ensemble of call prices is free of arbitrage. The solution has a striking resemblance to our conditions.

For greater precision, assume zero interest rates and zero dividend yields. We are given a regular grid of call-option prices Z_{j_1,j_2} with strikes X_{j_1}, $j_1 = 1, \ldots, q_1$, and time-to-maturities Y_{j_2}, $j_2 = 1, \ldots, q_2$, that are observed without error. The strikes and time-to-maturities are non-negative, strictly increasing and ordered in increasing magnitude. For each j_2, it is assumed that $Z_{1,j_2} = S_0$, where $X_1 = 0$ and S_0 is the current asset price; furthermore, assume that $Y_1 = 0$ with $Z_{j_1,1} = (S_0 - X_{j_1})^+$, for each j_1. Then, for each j_2, define the quantities

$$\Pi_{j_1,j_2} = \frac{Z_{j_1,j_2} - Z_{j_1-1,j_2}}{X_{j_1} - X_{j_1-1}}, \quad j_1 > 1, \ \Pi_{1,j_2} = -1. \tag{38}$$

For each $j_1 > 1$, Π_{j_1,j_2} is the price of a bear call spread that is long $1/(X_{j_1} - X_{j_1-1})$ calls struck at strike X_{j_1} and short the same

amount of calls struck at the lower strike X_{j_1-1}.[11] Such option positions are traded, for instance, to profit from range bound or falling asset prices; see Cohen (2005).

The CMDH sufficiency conditions for no-arbitrage can be summarized as follows:

$$S_0 \geq Z_{j_1,j_2} \quad \text{and the } Z_{j_1,j_2} \text{ tend to zero for increasing } j_1, \tag{39}$$

$$-1 \leq \Pi_{j_1,j_2} \leq \Pi_{j_1+1,j_2} \leq 0, \tag{40}$$

$$Z_{j_1,j_2+1} - Z_{j_1,j_2} \geq 0, \tag{41}$$

for all j_1, j_2. The intuition of (40) is that the bear call spreads Π_{j_1,j_2} should be non-positively priced (i.e., they yield a positive cash-flow at inception of the trade) and their prices should be nondecreasing in strikes. Moreover, the calendar spreads as defined in (41), where one buys a longer dated call and sells a shorter dated call with the same strike, should be priced non-negatively. It is apparent that (39)–(41) are a discretized version of the conditions of Proposition 2.1.

When $S_0 = 1$, the similarity to the conditions articulated in Propositions 4.1 and 4.2 and Corollary 4.5 is self-evident. The difference, however, is that the CMDH conditions are defined on the call-option price surface itself, whereas ours are stated on the control net. Despite this, both sets of conditions become identical if one uses the sequence of strikes and time-to-maturities as knot sequences and chooses a linear degree spline.[12] Indeed, for $p_1 = p_2 = 1$, it follows from (29) that the sets of Greville sites are identical to the respective knot sequences. Moreover, it follows from the recursion formula of the B-spline in (49)–(51) in Appendix A that all linear B-spline basis functions evaluated *at the knots* are zero, except for one that is equal to one. Thus, the TP B-spline evaluated at some arbitrary knot pair $(\xi_{j_1}, \upsilon_{j_2})$ is just $z(\xi_{j_1}, \upsilon_{j_2}) = \theta_{j_1,j_2}$. Choosing $\theta_{j_1,j_2} = Z_{j_1,j_2}$, for all j_1, j_2, therefore yields a bilinear TP B-spline surface that linearly interpolates the data set and that is identical to its own control net by definition.[13] Hence, condition (31) in Proposition 4.2 collapses to (40) and condition (37) in Corollary 4.5 collapses to (41). In consequence, the CMDH sufficiency conditions are embedded as the linear degree case into our set of conditions that allow for surfaces of arbitrary polynomial degree.

These considerations shed new light on Proposition 4.2 and Corollary 4.5. Both statements can be thought of as providing no-arbitrage conditions for the "prices" of bear spreads and calendar spreads that are set up in the control net of the TP B-spline. In this way, these two conditions can be given a neat financial intuition.

4.4. Knot placement

Up to this point, the knot sequence has been treated as an exogenous parameter, since asymptotically the precise placement of knots is irrelevant. In finite samples, however, knot placement is an important parametrization device. We address knot placement here.[14]

Typically, one chooses knots by minimizing some criterion function over a set of candidate knots. Traditional knot placement schemes proceed in a stepwise manner, starting with a forward knot selection followed by a backward knot deletion; see, e.g., Friedman and Silverman (1989) and He and Shi (1996). Zhou and Shen (2001) refine this idea by adding an adaptive knot search followed by local knot optimizations. Yet, because of the highly complex and nonlinear nature of the knot selection problem, these placement strategies may yield suboptimal knot locations, in particular for functions with kinks or other irregular features. As shown in Miyata and Shen (2003), this can be improved by the use of stochastic optimization techniques, which, however, come at the cost of a substantial computational burden.

For the estimation of the call-option price surface, we are guided by the following considerations. First, in order to avoid the complexity of a two-dimensional knot selection problem, we disregard the expiry dimension. The algorithms could be extended to the two-dimensional case, but we expect little gain from this. In our experience, the precise placement of time-to-maturity knots is not essential for capturing the features of the term structure. Second, since all constraints are implemented through the control net, one would like to place the Greville sites optimally rather than the knots. Unfortunately, there is no way to do so, because (29) maps $(q_k - p_k)$ interior knots into $(q_k - 1)$ interior Greville sites, $k = 1, 2$. As discussed in Section 4.3, only for $p_1 = p_2 = 1$ is this mapping one-to-one. To account for this, we start our knot search with a linear TP B-spline. This yields a good working knot sequence (and hence control net) for the linear call-option price surface, but not necessarily one for higher degree B-splines. Therefore, a knot relocation and deletion process follows, where splines at the desired degrees $p_1, p_2 > 1$ are employed. In this final process, the optimal knot sequence is determined.

We now give a more formal description of the algorithms which build with only minor modifications on Zhou and Shen (2001). As discussed, there is no guarantee that the procedure can yield a globally optimal knot sequence, but it converges quickly and gives a satisfactory performance for our purposes.

Algorithm 1: Knot search

For the knot search, we set $p_1 = p_2 = 1$. Among the no-arbitrage conditions, the convexity constraint is not imposed. This is because convexity is so strong a condition that the knot search breaks down. One starts with no interior knots, i.e., the initial working knot sequence is $\boldsymbol{\xi} = \{\xi_0, \xi_1, \xi_2, \xi_3\}$ where $\underline{x} = \xi_0 = \xi_1 < \xi_2 = \xi_3 = \bar{x}$. The search is only over the forward moneyness dimension. Denote by $\boldsymbol{\xi}^u$ the knot sequence which stores the updated knots. In the algorithm, a new knot is added if a sufficiently big reduction $\Delta_\Xi > 0$ in a decision criterion Ξ is achieved and if the new knot observes a sufficient distance from the respective interval boundaries $\Delta_\xi > 0$.

Step 1: The working knot sequence is assumed to be $\boldsymbol{\xi} = \{\xi_0, \ldots, \xi_{q_1+2}\}$. Set $\boldsymbol{\xi}^u = \boldsymbol{\xi}$. For each existing subinterval $[\xi_i, \xi_{i+1}]$, $i = 1, \ldots, q_1$, find an additional knot $\xi^o \in [\xi_i, \xi_{i+1}]$ that minimizes the selection criterion Ξ.
Insert ξ^o into $\boldsymbol{\xi}^u$, if
$\Xi(\boldsymbol{\xi}) - \Xi(\xi^o \cup \boldsymbol{\xi}) > \Delta_\Xi$
and if
$\xi^o - \xi_i > \Delta_\xi \wedge \xi_{i+1} - \xi^o > \Delta_\xi$.

[11] For avoidance of confusion, note that in Carr and Madan (2005) the corresponding quantities are defined as bull call spreads. This explains the differences between the two expositions. We have changed the definition to conform with our notation.

[12] More precisely, it is mandatory to augment the sequences by two additional knots to obtain two-fold coalescent boundary knots; see the definitions of the knot sequences in Section 3.1.

[13] See Gope and Fries (2011) for a related interpolation idea.

[14] Meyer (2008) introduces the convex cubic C-spline and investigates knot placement. She finds that only very few knots are required and that exact placement is of minor importance because of the flexibility of the spline. Although cubic C-splines are related to B-splines, these findings do not translate to our setting, since our constraints implementation differs from hers. To achieve convexity of the spline function, Meyer (2008) imposes rather mild conditions on the coefficients, such as non-negativity. Our constraints, however, are more restrictive, because they require coefficients to form a convex set on the Greville sites. Moreover, being defined on Greville sites, the conditions are strongly tied to the knot sequence. This comes at the cost of some flexibility, which in turn requires more and judiciously placed knots. For derivatives estimation as discussed in Section 5.2, good knot placement also appears to be vital.

Step 2: Put $\boldsymbol{\xi} = \boldsymbol{\xi}^u$. Repeat Step 1 for a user-specified number of runs, or stop if no further knot can be added.

Algorithm 2: Knot relocation and deletion

For relocation and deletion, a TP B-spline with the desired degrees p_1 and p_2 is used. The full set of no-arbitrage constraints is imposed to ensure that knot relocation and deletion are achieved optimally in the presence of the constraints. The algorithm leads to the deletion of a knot if deleting is associated with a value of Ξ lower than the one associated with relocating or with keeping it. If relocation improves Ξ, relocate. If neither deletion nor relocation results in a reduction of Ξ, no adjustment is made.

Step 1: Denote by $\boldsymbol{\xi}^u$ the knot sequence storing the updated knots. The working knot sequence is assumed to be $\boldsymbol{\xi} = \{\xi_0, \ldots, \xi_{q_1+p_1+1}\}$, where $q_1 > p_1 + 1$. Set $\boldsymbol{\xi}^u = \boldsymbol{\xi}$. For a subinterval $[\xi_i, \xi_{i+2}]$, $i = p_1, \ldots, (q_1 - 1)$, construct a knot sequence $\boldsymbol{\xi}_i$ such that $\xi_{i+1} \notin \boldsymbol{\xi}_i$ and $\left(\xi_{i+1} \cup \boldsymbol{\xi}_i\right) = \boldsymbol{\xi}$. Find an additional knot $\xi^o_{i+1} \in [\xi_i, \xi_{i+2}]$ that minimizes the selection criterion Ξ subject to

$$\xi^0_{i+1} - \xi_i > \Delta_\xi \quad \wedge \quad \xi_{i+2} - \xi^0_{i+1} > \Delta_\xi.$$

(a) If
$\Xi(\boldsymbol{\xi}_i) < \Xi\left(\xi^o_{i+1} \cup \boldsymbol{\xi}_i\right) \wedge \Xi(\boldsymbol{\xi}_i) < \Xi(\boldsymbol{\xi})$,
set $\boldsymbol{\xi}^u = \boldsymbol{\xi}_i$;

(b) if
$\Xi\left(\xi^o_{i+1} \cup \boldsymbol{\xi}_i\right) < \Xi(\boldsymbol{\xi}_i) \wedge \Xi(\xi^o_{i+1} \cup \boldsymbol{\xi}_i) < \Xi(\boldsymbol{\xi})$,
set $\boldsymbol{\xi}^u = (\xi^o_{i+1} \cup \boldsymbol{\xi}_i)$;

(c) otherwise set $\boldsymbol{\xi}^u = \boldsymbol{\xi}$.

Step 2: Perform Step 1 for each subinterval $[\xi_i, \xi_{i+2}]$, $i = p_1, \ldots, (q_1 - 1)$, sequentially in ascending magnitude of i.

Determining the number and position of knots is a model selection problem. Information criteria of the Akaike-type to approximate the Kullback–Leibler divergence are therefore a natural candidate for this task; see Shi (1995), He and Shi (1998), He and Ng (1999). For our work, we use the modified Akaike Information Criterion of Hurvich et al. (1998). It is defined as

$$\Xi_{\text{mAIC}} = \log(\text{ASR}) + 1 + \frac{2(\text{tr}(\mathbf{S}) + 1)}{n - \text{tr}(\mathbf{S}) - 2}, \tag{42}$$

where $\text{ASR} = \sum_{i=1}^n (z_i - \hat{z}_i)^2/n$ and $\mathbf{S} = \mathbf{B}(\mathbf{B}^\top\mathbf{B} + \lambda\mathbf{I})^{-1}\mathbf{B}^\top$ is the linear smoother matrix. Note that this is the smoother matrix of the unconstrained estimation problem; see Hastie and Tibshirani (1990). The degrees of freedom of the constrained estimation problem are at least as large as in the unconstrained case or smaller, depending on the number of active constraints. Since the number of active constraints is random, using the unconstrained smoother matrix is a practical solution. Alternative selection criteria used in the literature on knot selection are the Schwarz criterion and the generalized cross-validation criterion of Craven and Wahba (1979). We do not find any particular differences between these different criteria.

As is apparent, the bounds of the domain of estimation, $\underline{x}$ and $\bar{x}$, are not part of the optimization process. One should therefore not place the bounds too close to the domain of interest. One possibility, but an extreme one, is to set $\underline{x} = 0$. By assumption (C2), this forces the spline to pass through the set of points $(x, y, z) = (0, y, 1)$ for all y. In our experience, this selection leads to more stable derivatives estimates as it gives more room for optimal interior knot placements and it makes better use of the knowledge about the surface to be estimated. Placing the upper bound $\bar{x}$ far out to the right has a similar, yet weaker, effect since we implement (C4) only by means of the negative slope and the positivity of call-option prices. Following Eilers and Marx (1996b, p. 116), one can perceive this placement of boundary knots as a form of "extrapolation" beyond the given data range.

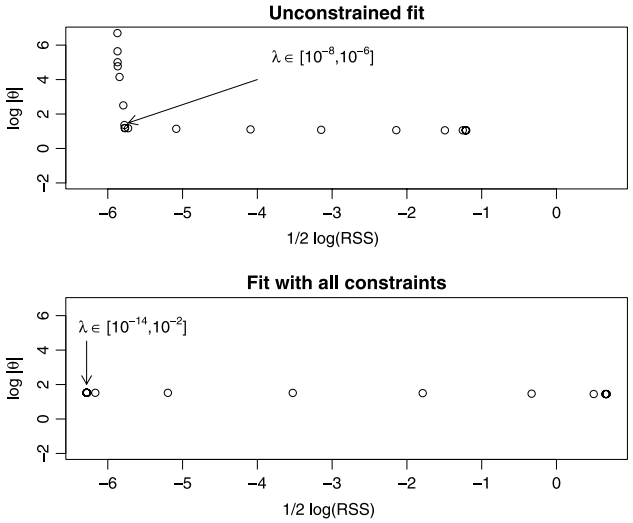


Fig. 2. The L-curve. Log-penalty plotted against the log-residual sum of squares (RSS) for different values of the regularization parameter $\lambda \in [10^{-14}, 10^5]$. The top panel shows the case of an unconstrained fit, the bottom panel the fit using all no-arbitrage constraints.

4.5. Impact of the regularization parameter

A classical way to determine the value of the regularization parameter is generalized cross-validation; see Wahba (1977), Craven and Wahba (1979), Golub et al. (1979). Standard cross-validation, however, only applies to unconstrained estimation problems, since the degrees of freedom in constrained estimation depend on the number of active constraints. Villalobos and Wahba (1987) and Wahba (1990, Chap. 9.4) suggest cross-validation schemes that correct for this fact, yet their approaches add an additional computational burden and do not immediately translate to our estimator, which differs from theirs in terms of constraints implementation. We therefore study the impact of regularization in constrained estimation by means of the graphical approach offered by the L-curve analysis. The L-curve is a plot of the regularization norm $|\boldsymbol{\theta}|$ against the square root of the residual sum of squares for a range of admissible regularization parameters. Plotted on a log–log-scale, an L-shaped curve is typically obtained; see Hansen and O'Leary (1993) for the theory of the L-curve analysis.

The estimations are based on the option data set described in Section 6.1 which uses a TP B-spline with degrees $p_1 = 5$ and $p_2 = 3$ in the forward moneyness and the time-to-maturity direction, respectively. The knot placements obtained from our algorithms can be found in Section 6.2. To obtain the L-curves, we estimate the surfaces for a wide range of possible values of λ. We contrast unconstrained fits with estimates which take advantage of the full set of no-arbitrage constraints, both in strike and in calendar dimension.

The top panel of Fig. 2 shows a typical L-curve for the unconstrained fits. For $\lambda \leq 10^{-6}$, which corresponds to the left uppermost part of the curve up to the kink, noise dominates the solution. For bigger values of λ, the penalty starts to dominate the solution, which results in deteriorating fits as expressed in the increasing residual sum of squares. This is in stark contrast to the shape-constrained solution seen in the lower panel of Fig. 2. No L-curve is obtained. For all $\lambda \in [10^{-14}, 10^{-2}]$ almost the same combinations of the penalty term and the residual sum of squares are recovered. This is because the monotonicity and convexity constraints that are imposed on the call-option price surface act as strong smoothing devices and enforce very similar solutions for a wide range of the regularization parameter. Interestingly, comparing the behavior of a variety of spline estimators, Meyer (2008, her Fig. 2a) comes to

the same conclusion for shape-restricted smoothing splines; see also the remarks in Dole (1999, p. 446). We therefore conclude that the shape-restricted option price surface is robust to a precise selection of λ. For our subsequent estimations, we will set $\lambda = 10^{-6}$, which corresponds to a small but numerically relevant value well in the middle of the powers of ten that produce similar solutions according to the L-curve in Fig. 2.

5. Monte Carlo simulations

We first consider the estimation of call-option price surfaces and gauge the efficiency benefits that are associated with enforcing no-arbitrage calendar constraints. We also investigate the usefulness of the proposed method for estimating state-price densities. For both studies, we use the regularized estimator on simulated data samples that mimic the situation of a single-day cross-sectional option data set. This is because one can expect efficiency gains to be more visible for sparse data sets.

5.1. Efficiency study for estimating the option price surface

Our set-up is as in Birke and Pilz (2009), but based on the jump–diffusion model suggested by Bates (1996) in order to obtain an empirically relevant price process for the underlying asset. Option price data are generated for the time-to-maturities of 1, 2, 3, 6, 12, and 24 months. Interest rates and dividends are assumed to be zero, and the underlying asset price is $S_0 = 1$. Strike spacing is 2% on the interval [0.6, 1.4], which leads to 41 observations per expiry. This spacing and the selection of expiry dates correspond to the usual reality for options on large indices, such as the S&P 500. Prices are additively perturbed with i.i.d. mean zero normally distributed errors with variance $\sigma_\varepsilon^2 = 0.01^2$. This implies a signal to noise ratio of about 12.

For the analysis, we use the regularized estimator with $\lambda = 10^{-6}$; compare the discussions in Section 4.5. The call-option price surface is estimated for a number of different smoothness assumptions and under three scenarios: (a) a single expiry date (0.5 years) option pricing function while disregarding the surface (univariate approach); (b) a surface without imposing calendar no-arbitrage constraints; (c) a surface employing all constraints. The number of repetitions is 5000 and knots are kept fixed at $\boldsymbol{\upsilon} = \{0.08, 0.15, 0.20, 0.30, 0.60, 1.20, 2.02\}$ for time-to-maturity and $\boldsymbol{\xi} = \{0.00, 0.49, 0.64, 0.75, 0.83, 0.97, 1.04, 1.20, 1.23, 1.31, 1.50, 2.00\}$ for forward moneyness.[15] In Table 1, we report the pointwise mean squared error (MSE), the variance and the squared bias from the exercise. All figures are expressed in relative terms to the MSE of the estimation approach (c). MSE numbers larger than one can therefore be interpreted as efficiency gains vis-à-vis the univariate estimator and the bivariate estimator without calendar constraints.

Skimming the numbers in Table 1, we see that the gains in MSE from imposing the full set of constraints ranges between 6% and 95%. The smallest gains of about 10% and less are obtained for a moneyness of about one. For moneyness of 0.8 and 1.2, we find gains soar to 20% and more. The largest efficiency gains are obtained at the boundary of the estimation domain, where the reduction in bias is also largest. These observations appear to be natural, but the size of the efficiency gains is striking.

Comparing the univariate approach in the left column block with the simple surface estimator that does not exploit time-to-maturity constraints in the right column block, we find that differences are small. Thus, using a surface estimator without calendar constraints hardly improves efficiency. The univariate approach performs similarly. When one allows for a more flexible parametrization in the strike dimension, i.e., a higher degree p_1, but keeps p_2 fixed, one sees little impact on MSE. This can be expected since strike constraints are present in all three approaches. On the other hand, varying p_2 with fixed p_1 shows mixed results. Typically, the MSE profile improves from the linear spline ($p_2 = 1$) to the quadratic one ($p_2 = 2$), but starts to deteriorate for the cubic one ($p_2 = 3$). Thus, calendar constraints tend to be beneficial when the estimator is either too restrictive or too flexible.[16] In summary, substantial efficiency gains from fully implementing calendar constraints can be expected, even when parsimonious parametrizations and moderate ranges of moneyness only are considered.

5.2. State-price density estimation and kernel estimates

If the pricing measure Q_τ is sufficiently smooth it admits a density q_τ, also called state-price density. Then (1) can be written as

$$C(S_t, K, \tau, r, \delta) = e^{-r\tau} \int_0^\infty (S_T - K)^+ \, \mathrm{q}_\tau(S_T) \, dS_T. \tag{43}$$

Differentiating twice with respect to K, one obtains

$$\mathrm{q}_\tau(S_T) = e^{r\tau} \left. \frac{\partial^2 C}{\partial K^2} \right|_{K=S_T}. \tag{44}$$

Thus, from an estimate of the call-option price function one can obtain estimates of q_τ.

From the B-spline representation of the estimator, the family of second-order derivatives in the forward moneyness dimension is given by

$$\frac{\partial^2 \hat{z}(x,y)}{\partial x^2} = \sum_{j_1=0}^{q_1} \sum_{j_2=0}^{q_2} \widehat{\theta}_{j_1,j_2} \, B^{(2)}_{j_1,p_1}(x) \, B_{j_2,p_2}(y), \quad (x,y) \in [\underline{x}, \bar{x}] \times [\underline{y}, \bar{y}] \tag{45}$$

where $B^{(k)}_{j,p}$ denotes the kth-order derivative of the p-degree B-spline basis function; see Appendix A. Since the TP B-spline estimate has a bounded support, it might be necessary to normalize the estimate with

$$K(y) = \sum_{j_1=0}^{q_1} \sum_{j_2=0}^{q_2} \widehat{\theta}_{j_1,j_2} \left\{ B^{(1)}_{j_1,p_1}(\bar{x}) - B^{(1)}_{j_1,p_1}(\underline{x}) \right\} B_{j_2,p_2}(y) \tag{46}$$

to ensure that the densities integrate to one across the strike domain. By construction of the estimator, $0 < K(y) \leq 1$. Using (12), one obtains the state-price densities on the original scale.

We compare our state-price density estimator with the kernel smoother by Aït-Sahalia and Duarte (ASD). The ASD estimator consists of two steps. First, one runs a constrained least squares regression that transforms the data such that they obey no-arbitrage constraints. In a second step, a locally linear estimator is applied to the transformed data and differentiated twice; see Aït-Sahalia and Duarte (2003) for the details.[17] For the kernel regression step, we use a Gaussian kernel.

The exercise is based on the simulated data set described in Section 5.1. The purpose is to estimate the state-price density of

[15] The coalescent boundary knots are not listed. The forward moneyness knots are obtained for a single, randomly selected surface, on which Algorithm 1 was applied for five runs followed by Algorithm 2.

[16] Note that in the univariate approach, any changes in the numbers along the p_2-dimension are entirely because of variation in the denominator (the full constraints case). The numerator necessarily stays constant.

[17] We thank Kay Pilz for generously sharing his codes with us.

Table 1
Efficiency study based on the jump–diffusion model: $dS_t/S_t = (r-d)dt + \sqrt{V_t}dW_t^1 + JdN_t$, $dV_t = \kappa(\theta - V_t)dt + \eta\sqrt{V_t}dW_t^2$, $dW_t^1 dW_t^2 = \rho dt$; N_t is a compound Poisson process with intensity λ_J and log-normal distribution of jump sizes $\log(1+J) \sim \mathcal{N}(\log(1+\mu_J) - \sigma_J^2/2, \sigma_J^2)$. Parameters are $S_0 = 1$, $r = 0$, $d = 0$, $V_0 = 0.1$, $\kappa = 2.03$, $\theta = 0.02$, $\eta = 0.38$, $\rho = -0.57$, $\lambda_J = 0.59$, $\mu_J = -0.05$, and $\sigma_J = 0.07$ as reported in Bakshi et al. (1997, Table III, column 'SVJ', p. 2018). Simulation results pertain to the option pricing function with time-to-maturity of 0.5 years. p_1 is B-spline degree for forward moneyness; p_2 for time-to-maturity. All numbers are expressed relative to the MSE of the bivariate fit with all constraints in place and computed from 5000 simulation runs.

Forward moneyness	p_2	Univariate fit					Bivariate fit without calendar constraint				
		0.6	0.8	1	1.2	1.4	0.6	0.8	1	1.2	1.4
		$p_1 = 3$									
rel. MSE	1	1.688	1.327	1.167	1.337	1.812	1.647	1.214	1.096	1.203	1.566
rel. Var.	1	1.532	1.327	1.166	1.319	1.412	1.488	1.214	1.095	1.191	1.196
rel. Bias^2 ($\times 10^{-2}$)	1	15.516	0.005	0.097	1.820	39.934	15.898	0.003	0.098	1.252	37.075
		$p_1 = 3$									
rel. MSE	2	1.603	1.315	1.119	1.385	1.664	1.587	1.217	1.067	1.271	1.481
rel. Var.	2	1.455	1.315	1.118	1.366	1.297	1.428	1.217	1.066	1.253	1.149
rel. Bias^2 ($\times 10^{-2}$)	2	14.734	0.005	0.093	1.885	36.684	15.918	0.001	0.093	1.800	33.187
		$p_1 = 3$									
rel. MSE	3	1.667	1.410	1.167	1.427	1.796	1.624	1.268	1.103	1.281	1.426
rel. Var.	3	1.513	1.410	1.166	1.407	1.400	1.458	1.268	1.102	1.268	1.110
rel. Bias^2 ($\times 10^{-2}$)	3	15.324	0.005	0.097	1.941	39.588	16.643	0.011	0.102	1.347	31.604
		$p_1 = 4$									
rel. MSE	1	1.566	1.336	1.181	1.321	1.853	1.542	1.237	1.111	1.191	1.601
rel. Var.	1	1.437	1.336	1.176	1.312	1.448	1.407	1.237	1.107	1.186	1.224
rel. Bias^2 ($\times 10^{-2}$)	1	12.903	0.017	0.493	0.898	40.505	13.548	0.031	0.458	0.524	37.691
		$p_1 = 4$									
rel. MSE	2	1.507	1.367	1.140	1.381	1.713	1.487	1.276	1.072	1.271	1.524
rel. Var.	2	1.383	1.367	1.135	1.371	1.339	1.353	1.276	1.067	1.262	1.180
rel. Bias^2 ($\times 10^{-2}$)	2	12.418	0.017	0.476	0.939	37.452	13.415	0.051	0.490	0.883	34.306
		$p_1 = 4$									
rel. MSE	3	1.564	1.475	1.185	1.448	1.840	1.529	1.343	1.107	1.301	1.475
rel. Var.	3	1.436	1.475	1.180	1.438	1.437	1.385	1.342	1.103	1.294	1.145
rel. Bias^2 ($\times 10^{-2}$)	3	12.892	0.018	0.494	0.984	40.214	14.403	0.097	0.425	0.719	32.935
		$p_1 = 5$									
rel. MSE	1	1.619	1.462	1.206	1.339	1.943	1.565	1.338	1.126	1.212	1.641
rel. Var.	1	1.496	1.462	1.202	1.329	1.442	1.439	1.338	1.122	1.206	1.188
rel. Bias^2 ($\times 10^{-2}$)	1	12.290	0.043	0.367	1.071	50.082	12.604	0.007	0.353	0.659	45.263
		$p_1 = 5$									
rel. MSE	2	1.533	1.427	1.144	1.320	1.788	1.502	1.328	1.094	1.228	1.605
rel. Var.	2	1.417	1.426	1.141	1.310	1.327	1.376	1.328	1.091	1.217	1.172
rel. Bias^2 ($\times 10^{-2}$)	2	11.636	0.042	0.348	1.055	46.105	12.572	0.009	0.361	1.154	43.266
		$p_1 = 5$									
rel. MSE	3	1.625	1.538	1.193	1.388	1.938	1.569	1.386	1.129	1.257	1.543
rel. Var.	3	1.502	1.537	1.190	1.377	1.438	1.434	1.386	1.124	1.248	1.134
rel. Bias^2 ($\times 10^{-2}$)	3	12.334	0.045	0.363	1.109	49.961	13.591	0.001	0.455	0.905	40.895

0.5 years to expiry by both methods. For each of the 5000 samples, a B-spline estimator with $p_1 = 5$ and $p_2 = 1$ is applied to the data for all expiry dates with the full set of no-arbitrage constraints. For comparison purposes, we also apply the univariate B-spline estimator with $p_1 = 5$ under strike no-arbitrage constraints. In consequence, the resulting state-price densities of the B-spline estimators are cubic polynomials. They are contrasted with the ASD estimator. As does the univariate B-spline, the ASD estimator only uses data of a single expiry date. For all estimators, we ensure that the densities integrate to one and have the same mean.[18] With this set-up, we aim to gain insights into the potential benefits for derivatives estimation if one implements no-arbitrage constraints on a surface rather than on a single expiry date only.

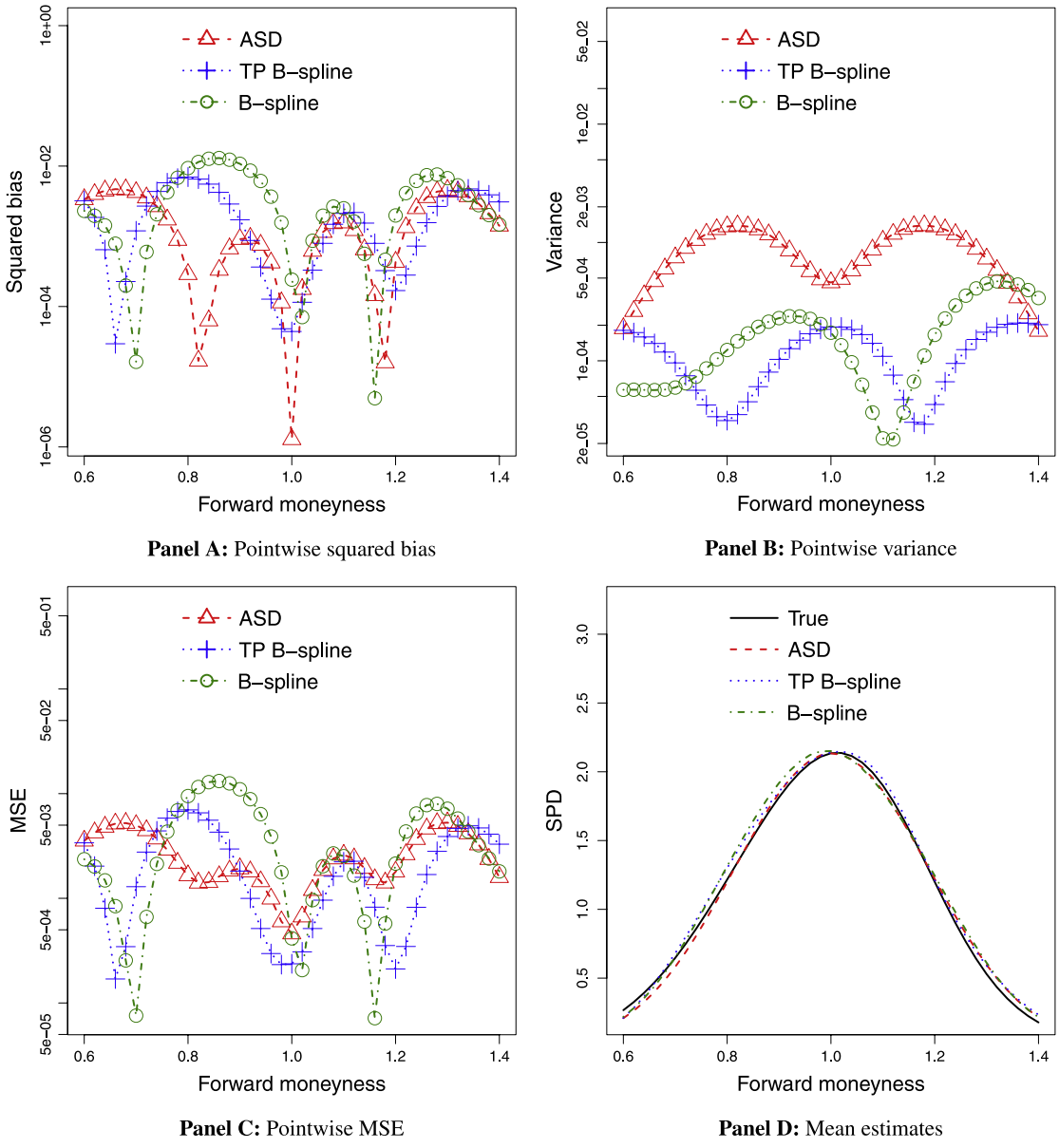
To make the different estimates comparable, we adopt the approach of Aït-Sahalia and Duarte (2003, Section 4) and select a bandwidth that minimizes the unweighted finite sample mean integrated squared error (MISE)

$$\Xi_{MISE} = \int \left\{ \left\{ \mathrm{E}\left[\frac{\partial^2 \hat{z}(x,y)}{\partial x^2} - \frac{\partial^2 z_0(x,y)}{\partial x^2} \middle| x, y \right] \right\}^2 + \mathrm{Var}\left[\frac{\partial^2 \hat{z}(x,y)}{\partial x^2} \middle| x, y \right] \right\} dx \Big|_{y=0.5}. \tag{47}$$

This is easily achieved for the 5000 samples for the ASD estimator. The optimal bandwidth we find is $h^* = 0.22$.

To find a knot sequence that minimizes MISE for the TP B-spline estimator, one could employ our Algorithms 1 and 2 with Ξ_{MISE} as decision criterion. Unfortunately, it is computationally infeasible for us to do this over all samples. We therefore skip Algorithm 1 and directly feed Algorithm 2 with knot sequences each consisting of three, four, five, …, etc. equidistant knots; moreover, we run Algorithm 2 only on a subset of 50 randomly selected samples. The forward moneyness knot sequence with the smallest MISE obtained from this procedure is given by $\boldsymbol{\xi} = \{0.00, 0.42, 0.60, 1.11, 1.20, 1.65, 2.00\}$. It necessarily contains fewer knots than the sequence in the price efficiency study, since estimators of derivatives converge at lower rates; see Zhou and Wolfe (2000). As discussed in all previous cases, the knot sequence in time-to-maturity direction is kept constant and is as reported in Section 5.1.

[18] As explained in Aït-Sahalia and Duarte (2003, p. 25), this amounts to norming and translating the state-price densities. With regard to the TP B-spline estimator, we find that these adjustments are almost never necessary if one places the forward moneyness boundary knots as we do, i.e., $\underline{x} = 0$ and $\bar{x} = 2.0$. This is because, up to a small error, this range corresponds to the numerically relevant domain for not too long-dated option price functions. Then these properties hold by construction.



Panel A: Pointwise squared bias

Panel B: Pointwise variance

Panel C: Pointwise MSE

Panel D: Mean estimates

Fig. 3. Comparison of the pointwise squared bias, variance and MSE profiles for the ASD estimator, the univariate B-spline and the TP B-spline state-price density estimators. Sample size is 5000. Panel A: Pointwise squared bias. Panel B: Pointwise variance. Panel C: Pointwise MSE. Panel D: True state-price density and mean estimates.

Sorting the three estimators based on the minimal MISE gives the following ranking: 0.0019 (TP B-spline), 0.0021 (ASD), and 0.0023 (univariate B-spline). To obtain a more detailed picture of the local behavior of the estimators, we additionally inspect the pointwise squared bias, variance, and MSE for all estimators based on the 5000 runs; see Fig. 3. According to Panel A, the ASD estimator most often has the smallest squared bias, followed by the TP B-spline estimator. Both B-spline estimators exhibit a large bias in the moneyness range from 0.75 to 0.9. The reason could be a suboptimal knot placement because of the sequential nature of the knot relocation process; see the discussion in Section 4.4. Panel B shows that the ASD estimator has the highest variance. On average it is larger by at least one order of magnitude. Note that the ordinates of Panels A–C are in log-scale, which implies drastic differences in relative terms between the estimators. In Panel C, the pointwise MSE is shown. Except for the moneyness range from 0.75 to 0.9 and at the very right boundary, the TP B-spline is best, followed by the ASD estimator. The univariate B-spline both achieves the lowest and the largest MSE among all estimators. Thus, the global criterion MISE puts the TP B-spline first, mainly because of its lower variance. Inspection of the pointwise MSE also shows, however, that the TP B-spline does not dominate uniformly. In contrast to the B-spline estimators, the kernel method achieves a more balanced MSE profile with less excessive variations.

It is wise not to over-interpret these finite-sample results. Both methods are different in nature, in particular in terms of their underlying smoothness assumptions but also in terms of the data that are exploited (all option data of a surface versus data of a single expiry date). Other, more sophisticated simulation environments could yield more substantive results. Yet the tendency uncovered by the price simulations in Section 5.1 is confirmed. We can expect improved estimates of derivatives of the option price surface, if we exploit the full informational content implicit in an entire option data surface, provided we make strict use of all no-arbitrage constraints both across the strike and across the expiry dimension.

Table 2
Summary statistics of S&P 500 call-option (mid) prices from December 1, 2010; closing price of S&P 500 index is $S_0 = 1206.07$ and the annualized risk-free rate is expressed in basis points;.
Source: OptionMetrics.

Time to maturity	Observations	Average	Option price percentile					Implied	Risk-free rate
(days)		strike spacing	5th	10th	50th	90th	95th	forward	(bps/year)
17	76	5.67	0.20	0.44	76.28	242.96	266.61	1205.30	27.43
30	31	15.00	1.51	3.65	48.84	230.47	267.61	1204.85	29.55
52	94	8.06	0.24	1.38	85.50	328.01	351.16	1204.00	33.01
80	48	17.55	0.44	1.78	83.13	362.35	418.71	1201.95	36.98
108	33	32.81	1.62	5.38	167.32	545.10	604.84	1200.35	40.37
199	26	47.00	9.19	14.35	161.52	756.90	806.57	1195.79	48.09
290	12	100.00	7.63	13.35	131.60	671.48	735.07	1191.57	52.82

6. Empirical applications

We continue to focus on the regularized estimator based on a single day of observed option data. An application of the unregularized estimator is not considered because of space considerations.

6.1. Data

We use end-of-day quotes of S&P 500 option data obtained from the OptionMetrics database from Wednesday, December 1, 2010. Along with the option data, S&P 500 closing prices and zero rates are provided. The data are subjected to the standard filtering rules applied in the literature to eliminate the influence of illiquid or otherwise error-prone observations; see, e.g., Constantinides et al. (2013) for details. After filtering we compute mid-prices from the bid–ask quotes. Since the options are European-style, we then calculate, for each available expiry date, the implied forwards by means of the put–call-parity from the put–call pair that is closest to at-the-money. This put–call pair is defined as the one with the minimal (absolute) difference between put and call price. Table 2 provides a summary of the call-option price data.

6.2. Surface estimate and first-order derivatives

6.2.1. Parametrizations and knot sequences

As in the previous sections, we use a B-spline of degree five across forward moneyness and a cubic degree for the time-to-expiry direction for our illustrations, i.e., $p_1 = 5, p_2 = 3$. In line with the discussions in Section 4.5, the regularization parameter is set to $\lambda = 10^{-6}$, and the knot search and relocation algorithms are parametrized with $\Delta_{\text{Crit.}} = 10^{-5}$ and $\Delta_{\xi} = 10^{-4}$.

As domain of estimation we choose the rectangle $[\underline{x}, \overline{x}] \times [\underline{y}, \overline{y}] = [0.00, 2.00] \times [0.04, 0.83]$. As discussed in Section 4.4, the reason for setting the forward moneyness range well outside the range of available data is our experience that this yields more stable estimates of second-order derivatives to be considered in Section 6.2.2. The knot sequence in time to expiry is fixed at $\boldsymbol{\upsilon} = \{0.047, 0.15, 0.50, 0.83\}$ without enumerating coinciding boundary knots. After seven runs of Algorithm 1 are completed, a raw knot sequence with 33 interior forward moneyness knots is obtained with $\Xi_{\text{mAIC}} = -15.68$. The subsequent knot relocation and deletion in Algorithm 2 lead to the adjusted knot sequence $\boldsymbol{\xi} = \{0, 0.30, 0.57, 0.69, 0.85, 0.92, 0.98, 1.03, 1.06, 1.07, 1.10, 1.11, 1.20, 1.21, 1.49, 2.00\}$ again without listing the coinciding boundary knots. Because of this dramatic reduction of knots, the selection criterion improves to $\Xi_{\text{mAIC}} = -16.72$.

The control net that is the outcome of the estimation is depicted in Fig. 4. According to our theory as described in Section 4, it represents the set of sufficient no-arbitrage conditions on a polynomial call-option price surface in B-spline form. As is clear from our discussions in Sections 2 and 4, the control net mimics the shape of the underlying surface.

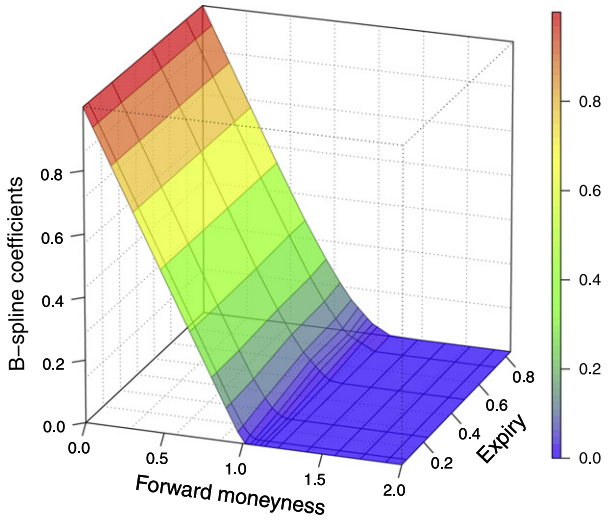


Fig. 4. Control net of the TP B-spline estimates of Figs. 5 and 6 with degrees $p_1 = 5$ in moneyness and $p_2 = 3$ in time-to-maturity direction. The control net represents the set of sufficient no-arbitrage conditions on the spline.

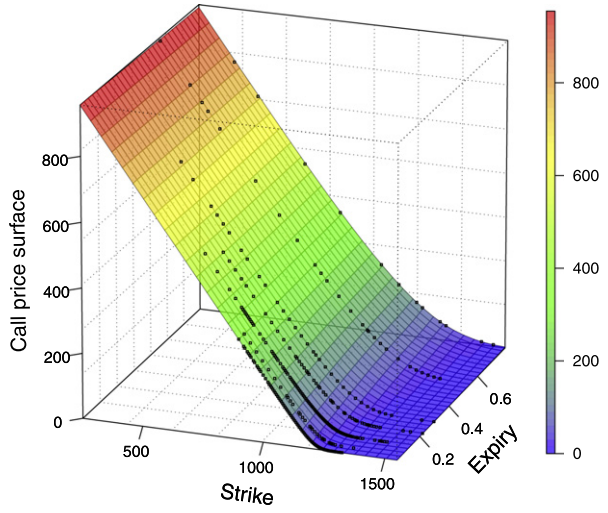


Fig. 5. Arbitrage-free TP B-spline estimate of the call-option price surface after application of the reverse transformation formula (10), based on the control net in Fig. 4. Degrees are $p_1 = 5$ in moneyness and $p_2 = 3$ in time-to-maturity direction. Surface is estimated from S&P 500 call-option price data as of December 1, 2010, shown as black dots.

Fig. 5 displays the estimated arbitrage-free call-option price surface on its original scale, i.e., after application of the reverse transformation formula given in (10). The input data are shown as black dots. It is seen that the surface smoothly adapts to the curvature delineated by the data. The correct implementation of

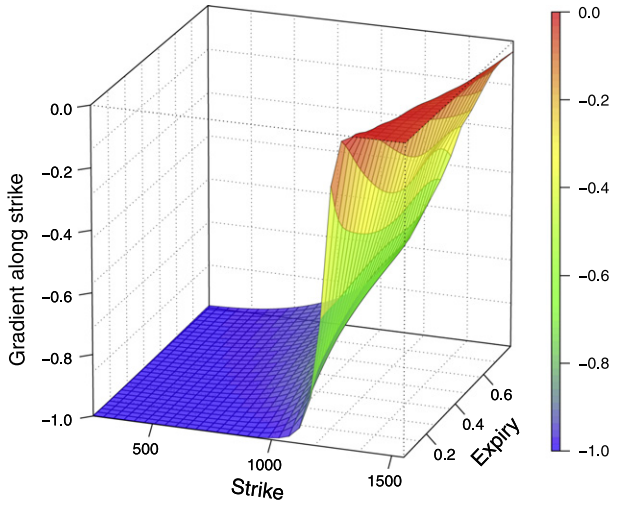


Fig. 6. Arbitrage-free TP B-spline estimate of the first-order strike derivative of the call-option price surface in Fig. 5. After differentiation, degrees are $p_1 = 4$ in moneyness and $p_2 = 3$ in time-to-maturity direction.

the no-arbitrage constraints can be best inferred from the derivatives. In Fig. 6, we present the first-order derivative with respect to forward moneyness. This plot is obtained by differentiating the TP B-spline. Then the reverse transformation relationship of the first-order derivative in (11) is applied. It is evident that the first-order derivatives monotonically increase along strikes as required by no-arbitrage theory.

6.2.2. State-price density estimation

For state-price density estimation, the value of the regularization parameter, the degrees of the B-spline estimator and the knot sequence in time-to-maturity direction remain unchanged. For the forward moneyness knots, we run Algorithm 1 four times, which after knot deletion and relocation yields the sequence $\boldsymbol{\xi} = \{0.00, 0.17, 0.73, 0.89, 0.91, 0.95, 1.04, 1.06, 1.10, 1.19, 1.42, 2.00\}$. We differentiate the estimate of the call-option price surface twice and substitute (12) into (44) to obtain the family of state-price densities on the original strike scale.

Aside from the TP B-spline estimates, we additionally display the ASD state-price densities. For bandwidth selection, we proceed as follows. Based on the bandwidth that is found to minimize MISE in the simulation study, we use a proportionally scaled bandwidth for each expiry date such that the ratio of the average strike spacing and the bandwidth is equal to the corresponding ratio in the simulation study. As the strike spacing increases over the different expiry dates, the bandwidths become larger. This may seem like an ad hoc approach for bandwidth selection, but it is founded on the insights in Section 5.2 and appears to deliver reasonable estimates. It can be given a rough asymptotic justification, as the average strike spacing tends to scale inversely with the number of observations per expiry date; see Table 2.

We display the state-price density estimates for six expiry dates in Fig. 7. As in the simulations, we make sure that the estimates integrate to one and the forwards are matched. For the first two expiry dates in Panels A–B, the estimates are close. In Panels C–E, differences become more manifest, as local bulges start to emerge in the estimates. The biggest differences appear in Panel F, where

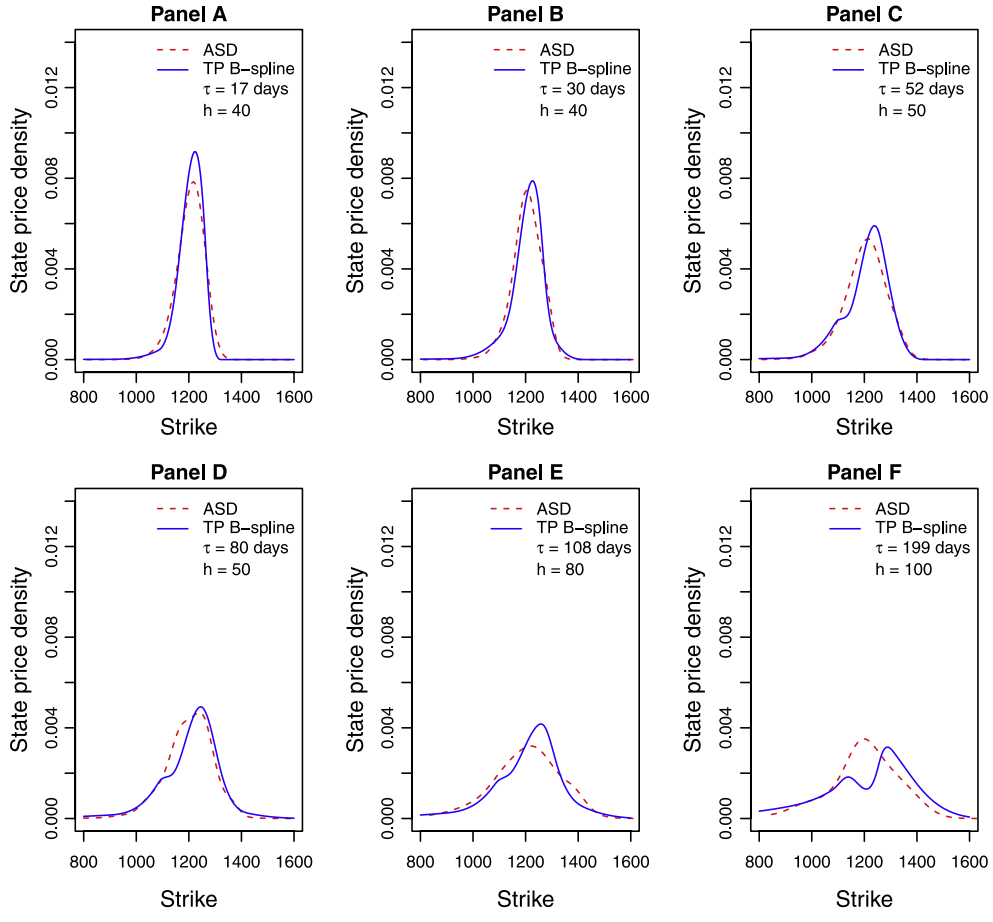


Fig. 7. State-price densities obtained as second-order strike derivatives of an estimate of the S&P 500 call-option price surface as of December 1, 2010 (solid blue line). After differentiation, the polynomial degree is $p_1 = 3$ in moneyness direction. The ASD estimator for expiry data is also given (dashed red line); h indicates the bandwidths used for estimation.

the TP-B state-price density displays a bimodal shape. Bimodal shapes are reported in the literature, typically in the context of major economic events; see, e.g., Melick and Thomas (1997) and Taylor et al. (2012). It is also possible, however, that the bimodality is an artifact from the sparse data distribution for these expiry dates. Whereas for the kernel estimator one can easily take account of sparse data by enlarging the bandwidth, the knot sequences of the TP B-spline estimator are static. This property may then give rise to oscillations. A way to address this issue could be locally adaptive knot meshes. We suggest this for future research.

The practical drawback of nonparametric or semi-nonparametric state-price density estimation is that one may receive no or implausible tail estimates because of the lack of option data for very small or very large moneyness. With the wide range of moneyness we use for the TP B-spline estimator we could have evaluated the estimates far outside the given data range. In our experience, the estimator along with our knot selection algorithm tends to deliver very reasonable tail estimates. This property can already be inferred from Panels A and B in Fig. 7, since for these two expiries there are no option price observations available for strikes larger than 1325. It seems hard, however, to characterize the tail behavior of such estimates. One might therefore prefer to model the tails by directly borrowing from known parametric densities. Such techniques are suggested, for instance, in Brunner and Hafner (2003) and Figlewski (2008); another alternative is offered in Corlay (2013), where the tails of the state-price density are inherited by calibration to a prior density. In this way, one can retain full control of the tail behavior of the state-price density while fully profiting from a flexible estimator such as ours by not taking an a priori stance on its main body.

7. Concluding remarks

This paper suggests a new approach to shape-constrained estimation of the call-option price surface. We build on tensor product (TP) B-spline methods and regularize the estimator in the presence of sparse data. To ensure that the estimated call-option price surface is arbitrage-free, we derive sufficient no-arbitrage conditions on the control net of the TP B-spline. The constraints are linear and independent of the degree of the underlying spline space. This allows for a least squares implementation using standard quadratic programming techniques. We show the consistency of the estimator and provide an upper bound of convergence, which is Stone's optimal rate. By means of simulations, we study the efficiency gains that ensue from estimating option price surfaces and state-price densities under the full set of surface no-arbitrage constraints. As an empirical demonstration, we apply the estimator to S&P 500 option price data.

Being a monotonicity constraint only, the calendar no-arbitrage condition could seem to be a weak constraint. Yet, according to our simulations, if calendar no-arbitrage is enforced jointly with the strike conditions, it becomes a powerful ingredient that helps estimate the functional form of an option pricing function more precisely. Hence our results may be of significance for many applications beyond the ones studied here. Examples include estimating the empirical pricing kernel as in Jackwerth (2000), the local volatility surface as in Dupire (1994), the risk-neutral variance or other risk neutral moments from option data as in Bakshi et al. (2003). More research is warranted to examine the efficacy of surface estimation in these cases.

Appendix A. B-splines: definition and properties

This Appendix gives a brief review of the univariate B-splines and their properties. We draw on Chapter IX in de Boor (2001) and Chapter 5 in Prautzsch et al. (2002).

A function $s(x)$ is said to be a spline of degree p (order $p+1$) with knots $\xi_0, \ldots, \xi_{q+p+1}$ with $\xi_j \leq \xi_{j+1}$ and $\xi_j < \xi_{j+p+1}$ for all possible j, if $s(x)$ is $p-r$ times differentiable at any r-fold knot, and $s(x)$ is a polynomial of degree equal to or less than p over each knot interval $[\xi_j, \xi_{j+1}]$. A knot ξ_j is called r-fold if $\xi_{j-1} < \xi_j = \cdots = \xi_{j+r-1} < \xi_{j+r}$. For this sequence of knots denoted by $\boldsymbol{\xi}$, $q+1$ linearly independent B-splines $B_{j,p}(x)$ of degree p with minimal support and certain continuity properties can be defined. The B-spline representation of $s(x)$ is an affine combination of these basis functions $B_{j,p}(x)$ such that

$$s(x) = \sum_{j=0}^{q} \theta_j B_{j,p}(x). \tag{48}$$

Definition of B-splines. Denote by $\boldsymbol{\xi}$ a strictly increasing sequence of knots. We define B-splines with degree $p > 0$ by the recursion formula of de Boor (1972) and Cox (1972):

$$B_{j,p}(x) = \alpha_{j,p-1}(x) B_{j,p-1}(x) + \{1 - \alpha_{j+1,p-1}(x)\} B_{j+1,p-1}(x), \tag{49}$$

where

$$\alpha_{j,p-1}(x) = \frac{x - \xi_j}{\xi_{j+p} - \xi_j}, \tag{50}$$

and

$$B_{j,0}(x) = \begin{cases} 1, & x \in [\xi_j, \xi_{j+1}), \\ 0, & \text{otherwise.} \end{cases} \tag{51}$$

In the presence of multiple knots, the B-splines are defined according to the same recursion formula with the additional convention that $B_{j,r-1}(x) = B_{j,r-1}(x)/(\xi_{j+r} - \xi_j) = 0$ if $\xi_j = \xi_{j+r}$.

The definition makes the following properties of B-splines evident. $B_{j,p}(x)$ is a piecewise polynomial of degree p; it is positive on (ξ_j, ξ_{j+p+1}), zero outside $[\xi_j, \xi_{j+p+1}]$, and right continuous.

Derivatives of B-splines. The first-order derivative of the basis function $B_{j,p}(x)$ with respect to x is

$$B^{(1)}_{j,p}(x) = p \left\{ \frac{B_{j,p-1}(x)}{\xi_{j+p} - \xi_j} - \frac{B_{j+1,p-1}(x)}{\xi_{j+p+1} - \xi_{j+1}} \right\}. \tag{52}$$

From this it can be shown that

$$\frac{\partial s(x)}{\partial x} = p \sum_{j=1}^{q} \frac{\theta_j - \theta_{j-1}}{\xi_{j+p} - \xi_j} B_{j,p-1}(x). \tag{53}$$

Recursive application of the identity (52) and relationship (53) gives $B^{(k)}_{j,p}(x)$, the kth-order derivative of $B_{j,p}(x)$ with respect to x, where $k \leq (p-1)$, and higher order derivatives of $s(x)$.

Control polygon. The graph of a spline function defines a planar curve

$$x \to (x, s(x)). \tag{54}$$

Importantly, this curve can be interpreted as a spline curve because of the linear precision property of the B-spline stating that

$$x = \sum_{j=0}^{q} \xi_j^* B_{j,p}(x), \tag{55}$$

where the sequence of points $\xi_j^* = (\xi_{j+1} + \cdots + \xi_{j+p})/p$, for $j = 1, \ldots, q$, is known as Greville sites. Because of this property, the collection of points $(\xi_j^*, \theta_j) \in \mathbb{R}^2$ is known as control points of the spline function $s(x) = \sum_j \theta_j B_{j,p}(x)$. They form the vertices of the control polygon. Since B-splines form a partition of unity, any point $(x, s(x)) = \left(\sum_j \xi_j^* B_{j,p}(x), \sum_j \theta_j B_{j,p}(x)\right)$ is a convex combination of its control points (barycentric combination). Therefore, the spline segments defined by the associated control polygon lie in its convex hull; see Fig. 1.

Appendix B. Asymptotic analysis

B.1. A general projection framework

As discussed in the main text, (17) and (18) can be interpreted as projections in a suitably defined normed vector space. Following Mammen et al. (2001), we introduce the vector space $\mathcal{V}$ which contains n-tuples of the form

$$\mathcal{V} = \left\{ \underset{\rightarrow}{z} = \begin{pmatrix} z_1(x,y) \\ \vdots \\ z_n(x,y) \end{pmatrix} : z_i \in \mathcal{S}(p_1, p_2; \boldsymbol{\xi}, \boldsymbol{\upsilon}), \quad i = 1, \ldots, n \right\}, \tag{56}$$

in which several subspaces are identified. First, the observed data vector $Z = (Z_1, \ldots, Z_n)^\top$ is seen as an element $\underset{\rightarrow}{Z}$ of $\mathcal{V}$ which is an n-tuple of constant functions, i.e., $z_i(x,y) = Z_i, i = 1, \ldots, n$. This subspace of constant functions is denoted by $\mathcal{V}^Z$. A candidate regression function z is interpreted as an element $\underset{\rightarrow}{z}$ of $\mathcal{V}$, which has identical entries everywhere, i.e., $z_i(x,y) = z(x,y), i = 1, \ldots, n$. The subspace of such n-tuples is called $\mathcal{V}^z$. Finally, there is the subspace of constrained candidate regression functions $\mathcal{V}_C^z \subset \mathcal{V}^z$ whose elements have entries $z_i(x,y) = z(x,y), i = 1, \ldots, n$ with $z \in \mathcal{S}^C(p_1, p_2; \boldsymbol{\xi}, \boldsymbol{\upsilon})$.

Since the true regression function z_0 is only assumed to be r-smooth, we cannot directly construct a vector $\underset{\rightarrow}{z_0}$ that is an element of $\mathcal{V}$. We make z_0 an element of $\mathcal{V}$ by replacing it with z_0^*, by which we denote an *approximation* of $z_0(X_i, Y_i)$ in $\mathcal{S}^C(p_1, p_2; \boldsymbol{\xi}, \boldsymbol{\upsilon})$ such that $\left(\frac{1}{n}\sum_{i=1}^n (z_0^*(X_i, Y_i) - z_0(X_i, Y_i)^2\right)^{1/2} = O_P(J_n^{-r})$. Such an approximation exists by virtue of standard spline theory; see (13.69) in connection with Theorem 12.8 in Schumaker (2007). With regard to the candidate regression functions, we set $z_{0i}^*(x,y) = z_0^*(x,y), i = 1, \ldots, n$, such that $\underset{\rightarrow}{z_0^*} \in \mathcal{V}$. Furthermore, w.l.o.g., we assume that the B-spline representation of all elements of $\mathcal{V}^z$ is demeaned, i.e., is represented by $z(x,y) = \bar{\theta} + \sum_{j_1=0}^{q_1}\sum_{j_2=0}^{q_2} \theta_{j_1,j_2} B_{j_1,p_1}(x) B_{j_2,p_2}(y)$, where $\bar{\theta}$ is constant. This implies that we have $\theta_{j_1,j_2} = 0$ for all j_1, j_2 whenever the element of $\mathcal{V}$ is a constant. This construction will be needed for the analysis of the regularized estimator.

We define an inner product on $\mathcal{V}_S$ by means of

$$\langle \underset{\rightarrow}{f}, \underset{\rightarrow}{g} \rangle_n = \frac{1}{n}\sum_{i=1}^n f_i(X_i, Y_i)\, g_i(X_i, Y_i) \tag{57}$$

and its induced squared norm by

$$\| \underset{\rightarrow}{f} \|_n^2 = \frac{1}{n}\sum_{i=1}^n f_i^2(X_i, Y_i). \tag{58}$$

With this notation, (17) and (18) can be written as

$$\hat{z} = \underset{\underset{\rightarrow}{z} \in \mathcal{V}^z}{\operatorname{argmin}} \| \underset{\rightarrow}{Z} - \underset{\rightarrow}{z} \|_n^2, \tag{59}$$

and

$$\hat{z}^C = \underset{\underset{\rightarrow}{z} \in \mathcal{V}_C^z}{\operatorname{argmin}} \| \underset{\rightarrow}{Z} - \underset{\rightarrow}{z} \|_n^2. \tag{60}$$

By noting that for $\underset{\rightarrow}{z} \in \mathcal{V}^z$, we have

$$\| \underset{\rightarrow}{Z} - \underset{\rightarrow}{z} \|_n^2 = \| \underset{\rightarrow}{Z} - \underset{\rightarrow}{\hat{z}} \|_n^2 + \| \underset{\rightarrow}{\hat{z}} - \underset{\rightarrow}{z} \|_n^2. \tag{61}$$

This holds since $\underset{\rightarrow}{Z} - \underset{\rightarrow}{\hat{z}}$ must be orthogonal to $\underset{\rightarrow}{\hat{z}} - \underset{\rightarrow}{z}$ with respect to the inner product defined in (57), since $\underset{\rightarrow}{\hat{z}}$ is the projection of the data $\underset{\rightarrow}{Z}$ on $\mathcal{V}^z$. As a consequence of this observation, it follows (Mammen et al., 2001, Prop. 1) that the minimization in (60) can be simplified to

$$\hat{z}^C = \underset{\underset{\rightarrow}{z} \in \mathcal{V}_C^z}{\operatorname{argmin}} \| \underset{\rightarrow}{\hat{z}} - \underset{\rightarrow}{z} \|_n^2. \tag{62}$$

Thus (60) is *equivalent* to projecting an unconstrained estimate on $\mathcal{V}_C^z$, the subspace with constrained candidate regression functions.

B.2. Proof of Proposition 3.1

Standard estimate. To arrive at the desired conclusion, we compare the empirical MSE of the constrained versus the unconstrained estimator under the norm (58) in $\mathcal{V}$. We have

$$\| \underset{\rightarrow}{\hat{z}} - \underset{\rightarrow}{z_0^*} \|_n^2 = \| \underset{\rightarrow}{\hat{z}} - \underset{\rightarrow}{\hat{z}^C} \|_n^2 + \| \underset{\rightarrow}{\hat{z}^C} - \underset{\rightarrow}{z_0^*} \|_n^2 + 2\langle \underset{\rightarrow}{\hat{z}} - \underset{\rightarrow}{\hat{z}^C}, \underset{\rightarrow}{\hat{z}^C} - \underset{\rightarrow}{z_0^*} \rangle_n \tag{63}$$

and from that

$$\| \underset{\rightarrow}{\hat{z}} - \underset{\rightarrow}{z_0^*} \|_n^2 - \| \underset{\rightarrow}{\hat{z}^C} - \underset{\rightarrow}{z_0^*} \|_n^2 = \| \underset{\rightarrow}{\hat{z}} - \underset{\rightarrow}{\hat{z}^C} \|_n^2 \geq 0. \tag{64}$$

This follows from the fact that $\langle \underset{\rightarrow}{\hat{z}} - \underset{\rightarrow}{\hat{z}^C}, \underset{\rightarrow}{\hat{z}^C} - \underset{\rightarrow}{z_0^*} \rangle_n \geq 0$, since $\underset{\rightarrow}{\hat{z}^C}$ is the projection of $\underset{\rightarrow}{\hat{z}}$ on $\mathcal{V}_C^z$ and $\mathcal{V}_C^z$ is a convex set.

In consequence, under the empirical norm, the constrained estimate $\hat{z}^C$ is at least as close to z_0^* as the unconstrained estimate $\hat{z}$, if not closer:

$$\left(\frac{1}{n}\sum_{i=1}^n (\hat{z}^C(X_i, Y_i) - z_0^*(X_i, Y_i))^2\right)^{1/2} \leq \left(\frac{1}{n}\sum_{i=1}^n (\hat{z}(X_i, Y_i) - z_0^*(X_i, Y_i))^2\right)^{1/2}. \tag{65}$$

Applying twice the triangular inequality and making use of the approximation by which the true regression function z_0 was replaced, one obtains

$$\left(\frac{1}{n}\sum_{i=1}^n (\hat{z}^C(X_i, Y_i) - z_0(X_i, Y_i))^2\right)^{1/2} \leq \left(\frac{1}{n}\sum_{i=1}^n (\hat{z}(X_i, Y_i) - z_0(X_i, Y_i))^2\right)^{1/2} + O_P(J_n^{-r}). \tag{66}$$

Finally, under Assumptions (A1)–(A5), it is a textbook result that $\frac{1}{n}\sum_{i=1}^n (\hat{z}(X_i, Y_i) - z_0(X_i, Y_i))^2 = O_P(J_n/n + J_n^{-r})$; see Theorem 15.1($ii$) in Li and Racine (2007) owed to Newey (1997). This proves the first assertion.

Analysis of the regularized estimate. For the regularized estimator, we alter the inner product (57) to

$$\langle \underset{\rightarrow}{f}, \underset{\rightarrow}{g} \rangle_n = \frac{1}{n}\sum_{i=1}^n f_i(X_i, Y_i)\, g_i(X_i, Y_i) + \lambda_n \frac{1}{n}\sum_{i=1}^n \boldsymbol{\theta}_{f_i} \cdot \boldsymbol{\theta}_{g_i} \tag{67}$$

where the regularization parameter $\lambda_n \to 0$ is a deterministic sequence and where $\boldsymbol{\theta}_{f_i} \cdot \boldsymbol{\theta}_{g_i}$ denotes the Euclidean inner product applied to the B-spline weight vectors of the functions f_i and g_i. Accordingly, the squared norm becomes

$$\| \underset{\rightarrow}{f} \|_n^2 = \frac{1}{n}\sum_{i=1}^n f_i^2(X_i, Y_i) + \lambda_n \frac{1}{n}\sum_{i=1}^n |\boldsymbol{\theta}_{f_i}|^2. \tag{68}$$

With these definitions and recalling that constant elements have a zero B-spline weight vector by construction, we can express

the estimation problems (19) and (20) as minimizations of the squared norms (68) over the spaces $\mathcal{V}^z$ and $\mathcal{V}^z_C$. The same line of argument as for the standard estimate then leads to the conclusion that

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{z}^C_\lambda(X_i,Y_i)-z^*_0(X_i,Y_i))^2+\lambda_n|\hat{\boldsymbol{\theta}}_{z^C}-\boldsymbol{\theta}_{z^*_0}|^2$$
$$\leq \frac{1}{n}\sum_{i=1}^{n}(\hat{z}_\lambda(X_i,Y_i)-z^*_0(X_i,Y_i))^2+\lambda_n|\hat{\boldsymbol{\theta}}_z-\boldsymbol{\theta}_{z^*_0}|^2. \tag{69}$$

Choosing $\lambda_n \to 0$ such that $\lambda_n|\hat{\boldsymbol{\theta}}_z-\boldsymbol{\theta}_{z^*_0}|^2 = O_P(J_n/n+J_n^{-r})$ and proceeding as before, yields

$$\left(\frac{1}{n}\sum_{i=1}^{n}(\hat{z}^C_\lambda(X_i,Y_i)-z_0(X_i,Y_i))^2\right)^{1/2}$$
$$\leq \left(\frac{1}{n}\sum_{i=1}^{n}(\hat{z}_\lambda(X_i,Y_i)-z_0(X_i,Y_i))^2\right)^{1/2}$$
$$+O_P\left((J_n/n+J_n^{-r})^{1/2}\right), \tag{70}$$

whereupon one concludes as above.

Appendix C. Proofs on monotonicity-constrained TP B-splines

Here we reproduce the proofs of Lemma 4.3 and Theorem 4.4 owed to Floater and Peña (1998, Lemma 2.1 and Theorem 4.1).

Proof of Lemma 4.3: By differentiating (28) with respect to x, we obtain

$$\frac{\partial c_{j_1,j_2}}{\partial x} = \sum_{\ell=0}^{1}\beta_{j_2,\ell}(y)\Delta_1\theta_{j_1,j_2+\ell}$$
$$= \sum_{k=0}^{1}\sum_{\ell=0}^{1}\alpha_{j_1,k}(x)\,\beta_{j_2,\ell}(y)\Delta_1\theta_{j_1,j_2+\ell} \tag{71}$$

by noting that $\alpha_{j_1,1}(x)+\alpha_{j_1,0}(x)=1$. An analogous argument is made for $\partial c_{j_1,j_2}/\partial y$. Then the directional derivative with direction $d=(d_1,d_2)\in\mathbb{R}^2$ is

$$D_d\,c_{j_1,j_2} = d_1\frac{\partial c_{j_1,j_2}}{\partial x}+d_2\frac{\partial c_{j_1,j_2}}{\partial y}$$
$$= \sum_{k=0}^{1}\sum_{\ell=0}^{1}\alpha_{j_1,k}(x)\,\beta_{j_2,\ell}(y)\left(d_1\Delta_1\theta_{j_1,j_2+\ell}+d_2\Delta_2\theta_{j_1+k,j_2}\right), \tag{72}$$

from which the assertion follows. ■

Proof of Theorem 4.4: Differentiate $z(x,y)=\sum_{j_1=0}^{q_1}\sum_{j_2=0}^{q_2}\theta_{j_1,j_2}B_{j_1,p_1}(x)\,B_{j_2,p_2}(y)$ with respect to x. This yields

$$\frac{\partial z}{\partial x} = \sum_{j_2=0}^{q_2}\theta_{j_1,j_2}B_{j_2,p_2}(y)\sum_{j_1=0}^{q_1}\frac{\partial}{\partial x}B_{j_1,p_1}(x) \tag{73}$$
$$= \sum_{j_2=0}^{q_2}B_{j_2,p_2}(y)\sum_{j_1=1}^{q_1}(\theta_{j_1,j_2}-\theta_{j_1-1,j_2})\frac{p_1}{\xi_{i+p_1}-\xi_i}B_{j_1,p_1-1}(x) \tag{74}$$
$$= \sum_{j_2=0}^{q_2}B_{j_2,p_2}(y)\sum_{j_1=1}^{q_1}\left(\frac{\theta_{j_1,j_2}-\theta_{j_1-1,j_2}}{\xi^*_{j_1}-\xi^*_{j_1-1}}\right)B_{j_1,p_1-1}(x) \tag{75}$$
$$= \sum_{j_2=0}^{q_2}B_{j_2,p_2}(y)\sum_{j_1=1}^{q_1}\Delta_1\theta_{j_1-1,j_2}B_{j_1,p_1-1}(x), \tag{76}$$

where the second equality follows from (53) and the third from noting that

$$\xi^*_{j_1}-\xi^*_{j_1-1} = \frac{\xi_{j_1+1}+\cdots+\xi_{j_1+p_1}}{p_1}-\frac{\xi_{j_1}+\cdots+\xi_{j_1+p_1-1}}{p_1}$$
$$= \frac{\xi_{j_1+p_1}-\xi_{j_1}}{p_1}. \tag{77}$$

The last equation uses the notation for $\Delta_1\theta_{j_1,j_2}$ from Lemma 4.3.

Applying the recursion formula (49)–(51) to $B_{j_2,p_2}(y)$ and inserting it into (76) yields

$$\frac{\partial z}{\partial x} = \sum_{j_1=1}^{q_1}\sum_{j_2=1}^{q_2}B_{j_1,p_1-1}(x)B_{j_2,p_2-1}(y)\sum_{\ell=0}^{1}\beta_{j_2,\ell}(y)\Delta_1\theta_{j_1-1,j_2-\ell} \tag{78}$$

where $\beta_{j_2,0}(y)=(\upsilon_{j_2+p_2}-y)/(\upsilon_{j_2+p_2}-\upsilon_{j_2})$ and $\beta_{j_2,1}(y)=1-\beta_{j_2,0}(y)$. With analogous manipulations, one obtains

$$\frac{\partial z}{\partial y} = \sum_{j_1=1}^{q_1}\sum_{j_2=1}^{q_2}B_{j_1,p_1-1}(x)B_{j_2,p_2-1}(y)\sum_{k=0}^{1}\alpha_{j_1,k}(x)\Delta_2\theta_{j_1-k,j_2-1}, \tag{79}$$

where $\alpha_{j_1,0}(x)=(\xi_{j_1+p_1}-x)/(\xi_{j_1+p_1}-\xi_{j_1})$ and $\alpha_{j_1,1}(x)=1-\alpha_{j_1,0}(x)$. It follows for the directional derivative that

$$D_d\,z = \sum_{j_1=1}^{q_1}\sum_{j_2=1}^{q_2}\sum_{k=0}^{1}\sum_{\ell=0}^{1}B_{j_1,p_1-1}(x)B_{j_2,p_2-1}(y)\alpha_{j_1,k}(x)$$
$$\times\,\beta_{j_2,\ell}(y)\left(d_1\Delta_1\theta_{j_1-1,j_2-\ell}+d_2\Delta_2\theta_{j_1-k,j_2-1}\right) \tag{80}$$

with $d=(d_1,d_2)\in\mathbb{R}^2$. Combined with Lemma 4.3, this yields $D_d\,z\geq 0$ on the rectangle $[\xi_{p_1},\xi_{q_1+1}]\times[\upsilon_{p_2},\upsilon_{q_2+1}]$. ■

References

Agarwal, G.G., Studden, W.J., 1980. Asymptotic integrated mean square error using least squares and bias minimizing splines. Ann. Statist. 8 (6), 1307–1325.

Aït-Sahalia, Y., Duarte, J., 2003. Nonparametric option pricing under shape restrictions. J. Econometrics 116, 9–47.

Aït-Sahalia, Y., Lo, A., 1998. Nonparametric estimation of state-price densities implicit in financial asset prices. J. Finance 53, 499–548.

Aït-Sahalia, Y., Lo, A., 2000. Nonparametric risk management and implied risk aversion. J. Econometrics 94, 9–51.

Alexander, C., Nogueira, L.M., 2007. Model-free hedge ratios and scale-invariant models. J. Bank. Finance 31 (6), 1839–1861.

Andrews, D.W.K., 1991. Asymptotic normality of series estimators for nonparametric and semiparametric regression models. Econometrica 59 (2), 307–345.

Bakshi, G., Cao, C., Chen, Z., 1997. Empirical performance of alternative option pricing models. J. Finance 52 (5), 2003–2049.

Bakshi, G., Kapadia, N., Madan, D., 2003. Stock return characteristics, skew laws, and the differential pricing of individual equity options. Rev. Financ. Stud. 16 (1), 101–143.

Bates, D.S., 1996. Jumps and stochastic volatility: Exchange rate processes implicit in deutsche mark options. Rev. Financ. Stud. 9, 69–107.

Bates, D.S., 2005. Hedging the smirk. Finance Res. Lett. 2 (4), 195–200.

Benko, M., Fengler, M.R., Härdle, W., Kopa, M., 2007. On extracting information implied in options. Comput. Statist. 22 (4), 543–553.

Birke, M., Pilz, K.F., 2009. Nonparametric option pricing with no-arbitrage constraints. J. Financ. Econom. 7 (2), 53–76.

Bliss, R., Panigirtzoglou, N., 2004. Option-implied risk aversion estimates. J. Finance 59 (1), 407–446.

Bondarenko, O., 2003. Estimation of risk-neutral densities using positive convolution approximation. J. Econometrics 116 (1–2), 85–112.

Brunk, H.D., 1970. Estimation of isotonic regression. In: Puri, M.L. (Ed.), Nonparametric Techniques in Statistical Inference. Cambridge University Press, pp. 177–195.

Brunner, B., Hafner, R., 2003. Arbitrage-free estimation of the risk-neutral density from the implied volatility smile. J. Comput. Finance 7 (1), 75–106.

Carr, P., Madan, D.B., 2005. A note on sufficient conditions for no arbitrage. Finance Res. Lett. 2, 125–130.

Chen, X., 2007. Large sample sieve estimation of semi-nonparametric models. In: Heckman, J.J., Leamer, E.E. (Eds.), Handbook of Econometrics, vol. 6B. Elsevier, pp. 5549–5632. (Chapter 76).

Cohen, G., 2005. The Bible of Options Strategies: The Definitive Guide for Practical Trading Strategies. FT Press, New Jersey.

Constantinides, G.M., Jackwerth, J.C., Savov, A., 2013. The puzzle of index option returns. Rev. Asset Pricing Stud. 3, 229–257.

Cont, R., da Fonseca, J., 2002. The dynamics of implied volatility surfaces. Quant. Finance 2 (1), 45–60.
Corlay, S., 2013. B-spline Techniques for Volatility modeling. Technical Report, arXiv.org, http://arxiv.org/abs/1306.0995.
Cox, M.G., 1972. The numerical evaluation of B-splines. J. Inst. Math. Appl. 10, 134–149.
Craven, P., Wahba, G., 1979. Smoothing noisy data with spline functions. Numer. Math. 31, 377–390.
Davis, M., Hobson, D.G., 2007. The range of traded options. Math. Finance 17 (1), 1–14.
de Boor, C., 1972. On calculating with B-splines. J. Approx. Theory 6, 50–62.
de Boor, C., 2001. A practical guide to splines, revised ed. Springer-Verlag, New York.
Delbean, F., Schachermeyer, W., 1994. A general version of the fundamental theorem of asset pricing. Math. Ann. 300, 463–520.
Delecroix, M., Thomas-Agnan, C., 2000. Spline and kernel regression under shape restrictions. In: Schimek, M. G. (Ed.), Smoothing and Regression. John Wiley & Sons.
Dierckx, P.H., 1980. An algorithm for for cubic spline fitting with convexity constraints. Computing 24, 349–371.
Dierckx, P.H., 1993. Curve and Surface Fitting With Splines. Clarendon Press, Oxford, U.K.
Dole, D., 1999. CoSmo: A constrained scatterplot smoother for estimating convex, monotonic transformations. J. Bus. Econom. Statist. 17 (4), 444–455.
Dupire, B., 1994. Pricing with a smile. Risk 7 (1), 18–20.
Eilers, P., Marx, B.D., 1996a. Flexible smoothing with B-splines and penalties. Statist. Sci. 11 (2), 89–102.
Eilers, P., Marx, B.D., 1996b. Rejoinder to comments on "Flexible smoothing with B-splines and penalties". Statist. Sci. 11 (2), 115–121.
Fan, J., 1992. Design adaptive nonparametric regression. J. Amer. Statist. Assoc. 87 (420), 998–1004.
Fan, J., 1993. Local linear regression smoothers and their minimax efficiences. Ann. Statist. 21 (1), 196–216.
Fan, J., Mancini, L., 2009. Option pricing with model-guided nonparametric methods. J. Amer. Statist. Assoc. 104 (488), 1351–1372.
Fengler, M.R., 2009. Arbitrage-free smoothing of the implied volatility surface. Quant. Finance 9 (4), 417–428.
Fengler, M.R., Härdle, W., Mammen, E., 2007. A semiparametric factor model for implied volatility surface dynamics. J. Financ. Econom. 5 (2), 189–218.
Figlewski, S., 2008. Estimating the implied risk neutral density for the U.S. market portfolio. In: Bollerslev, T., Russell, J.R., Watson, M. (Eds.), Volatility and Time Series Econometrics: Essays in Honor of Robert F. Engle. Oxford University Press, Oxford, UK.
Floater, M.S., Peña, J., 1998. Tensor-product monotonicity preservation. Adv. Comput. Math. 9, 353–362.
Friedman, J.H., Silverman, B.W., 1989. Flexible parsimonious smoothing and additive modeling. Technometrics 31, 3–21.
Gallant, A.R., 1982. Unbiased determination of production technologies. J. Econometrics 20 (2), 285–323.
Gallant, A.R., 1987. Identification and consistency in semi-nonparametric regression. In: Bewley, T.F. (Ed.), Advances in Econometrics, Fifth World Congress, vol. 1. Cambridge University Press, Cambridge.
Garcia, R., Ghysels, E., Renault, E., 2010. The econometrics of option pricing. In: Aït-Sahalia, Y., Hansen, L. (Eds.), Handbook of Financial Econometrics, vol. 1. North Holland, Amsterdam, pp. 479–552.
Garcia, R., Renault, E., 1998a. A note on hedging in ARCH and stochastic volatility option pricing models. Math. Finance 8 (2), 153–161.
Garcia, R., Renault, E., 1998b. Risk aversion, intertemporal substitution and option pricing, Working Paper 98s-02, CIRANO, Montréal.
Gatheral, J., 2004. A parsimonious arbitrage-free implied volatility parameterization with application to the valuation of volatility derivatives, Presentation at the ICBI Global Derivatives and Risk Management, Madrid, Spain.
Glaser, J., Heider, P., 2012. Arbitrage-free approximation of call price surfaces and input data risk. Quant. Finance 12 (1), 61–73.
Goldfarb, D., Idnani, A., 1983. A numerically stable dual method for solving strictly convex quadratic programs. Math. Program. 27 (1), 1–33.
Golub, G.H., Heath, M.T., Wahba, G., 1979. Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics 21 (2), 215–223.
Gope, P., Fries, C.P., 2011. Volatility Surface Interpolation on Probability Space Using Normed Call Prices. Technical Report, Available at SSRN: http://ssrn.com/abstract=1964634.
Groeneboom, P., Jongbloed, G., Wellner, J.A., 2001a. A canonical process for estimation of convex functions: the "invelope" of integrated Brownian motion $+ t^4$. Ann. Statist. 29 (6), 1620–1652.
Groeneboom, P., Jongbloed, G., Wellner, J.A., 2001b. Estimation of a convex function: characterizations and asymptotic theory. Ann. Statist. 29 (6), 1653–1698.
Hansen, P.C., O'Leary, D.P., 1993. The use of the L-curve in the regularization of discrete ill-posed problems. SIAM J. Sci. Comput. 14 (6), 1487–1503.
Hanson, D.L., Pledger, G., 1976. Consistency in concave regression. Ann. Statist. 4, 1038–1050.
Härdle, W., Hlávka, Z., 2009. Dynamics of state price densities. J. Econometrics 150 (1), 1–15.
Hastie, T., Tibshirani, R., 1990. Generalized Additive Models. Chapman and Hall, London.
He, X., Ng, P., 1999. COBS: Qualitatively constrained smoothing via linear programming. Comput. Statist. 14 (3), 315–337.
He, X., Shi, P., 1996. Bivariate tensor-product B-splines in a partly linear model. J. Multivariate Anal. 58 (2), 162–181.
He, X., Shi, P., 1998. Monotone B-spline smoothing. J. Amer. Statist. Assoc. 93 (442), 643–650.
Hildreth, C., 1958. Point estimates of ordinates of concave functions. J. Amer. Statist. Assoc. 49, 598–619.
Huang, J.Z., 2003. Local asymptotics for polynomial spline regression. Ann. Statist. 31 (5), 1600–1635.
Huang, S.-Y., Studden, W.J., 1993. An equivalent kernel method for least squares spline regression. Statist. Decisions (Suppl. 3), 179–201.
Hurvich, C., Simonoff, J., Tsai, C.-L., 1998. Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. J. R. Stat. Soc. Ser. B Stat. Methodol. 60 (2), 271–293.
Hutchinson, J.M., Lo, A.W., Poggio, T., 1994. A nonparametric approach to hedging and pricing derivative securities via learning networks. J. Finance 49, 851–889.
Jackwerth, J.C., 2000. Recovering risk-aversion from option prices and realized returns. Rev. Financ. Stud. 13 (2), 433–451.
Jackwerth, J.C., 2004. Option-implied risk neutral distributions and risk aversion, Research Foundation of AIMR, Charlotteville, USA.
Jackwerth, J.C., Rubinstein, M., 1996. Recovering probability distributions from contemporaneous security prices. J. Finance 51, 1611–1631.
Jarrow, R., Rudd, A., 1982. Approximate valuation for arbitrary stochastic processes. Journal of Financial Economics 10 (3), 347–369.
Jondeau, E., Rockinger, M., 2001. Gram-Charlier densities. J. Econom. Dynam. Control 25, 1457–1483.
Laurini, M.P., 2011. Imposing no-arbitrage conditions in implied volatilities using constrained smoothing splines. Appl. Stoch. Models Bus. Ind. 27 (6), 649–659.
Li, Q., Racine, S., 2007. Nonparametric Econometrics. Princeton University Press, Princeton, New Jersey.
Madan, D.B., Milne, F., 1994. Contingent claims valued and hedged by pricing and investing in a basis. Math. Finance 4 (3), 223–245.
Mammen, E., 1991a. Estimating a smooth monotone regression function. Ann. Statist. 19 (2), 724–740.
Mammen, E., 1991b. Nonparametric regression under qualitative smoothness assumptions. Ann. Statist. 19 (2), 741–759.
Mammen, E., Marron, J.S., Turlach, B.A., Wand, M.P., 2001. A general framework for constrained smoothing. Statist. Sci. 16 (3), 232–248.
Mammen, E., Thomas-Agnan, C., 1999. Smoothing splines and shape restrictions. Scand. J. Stat. 26, 239–252.
Melick, W.R., Thomas, C.P., 1997. Recovering an asset's implied PDF from option prices: An application to crude oil during the Gulf crisis. J. Financ. Quant. Anal. 32 (1), 91–115.
Merton, R.C., 1973. Theory of rational option pricing. Bell J. Econ. Manage. Sci. 4 (Spring), 141–183.
Meyer, M.C., 2008. Inference using shape-restricted regression splines. Ann. Appl. Stat. 2 (3), 1013–1033.
Miyata, S., Shen, X., 2003. Adaptive free-knot splines. J. Comput. Graph. Statist. 12 (1), 197–213.
Monteiro, A.M., Tütüncü, R.H., Vicente, L.N., 2008. Recovering risk neutral probability density functions from options prices using cubic splines and ensuring non-negativity. European J. Oper. Res. 187, 525–542.
Newey, W.K., 1997. Convergence rates and asymptotic normality for series estimators. J. Econometrics 79 (1), 147–168.
Prautzsch, H., Boehm, W., Paluszny, M., 2002. Bézier and B-spline Techniques. Springer-Verlag, Berlin, Heidelberg.
Ramsay, J.O., 1988. Monotone regression splines in action. Statist. Sci. 3 (4), 425–461.
Reiner, E., 2000. Calendar spreads, characteristic functions, and variance interpolation. Mimeo.
Renault, E., 1997. Econometric models of option pricing errors. In: Kreps, D.M., Wallis, K.F. (Eds.), Advances in Economics and Econometrics, Seventh World Congress, Econometric Society Monographs. Cambridge University Press, pp. 223–278.
Roper, M., 2010. Arbitrage free implied volatility surfaces, Working paper, School of Mathematics and Statistics, University of Sydney, NSW, Australia.
Schumaker, L., 2007. Spline Functions: Basic Theory, third ed.. Cambridge University Press, Cambridge, UK.
Shi, P., 1995. Automatic selection of parameters in spline regression via Kullback–Leibler information. Syst. Sci. Math. Sci. 8 (4), 346–354.
Silverman, B.W., 1984. Spline smoothing: the equivalent variable kernel method. Ann. Statist. 12, 898–916.
Stone, C.J., 1982. Optimal global rates of convergence for nonparametric regression. Ann. Statist. 10 (4), 1040–1053.
Stone, C.J., 1994. The use of polynomial splines and their tensor products in multivariate function estimation. Ann. Statist. 22 (1), 118–184.
Stutzer, M., 1996. A simple nonparametric approach to derivative security valuation. J. Finance 51 (5), 1633–1652.
Taylor, S., Tzeng, C.F., Widdicks, M., 2012. Bankruptcy probabilities inferred from option prices. Technical report, Lancaster University Management School, Lancaster, UK.
Villalobos, M., Wahba, G., 1987. Inequality-constrained multivariate smoothing splines with application to the estimation of posterior probabilities. J. Amer. Statist. Assoc. 82 (397), 239–248.
Wahba, G., 1977. Practical approximate solutions to linear operator equations when the data are noisy. SIAM J. Numer. Anal. 14, 651–667.
Wahba, G., 1990. Spline Models for Observational Data. SIAM, Philadelphia.
Wang, Y., Yin, H., Qi, L., 2004. No-arbitrage interpolation of the option price function and its reformulation. J. Optim. Theory Appl. 120 (3), 627–649.

Wright, F.T., 1981. The asymptotic behavior of monotone regression estimates. Ann. Statist. 9, 443–448.
Yatchew, A., Härdle, W., 2006. Dynamic state price density estimation using constrained least squares and the bootstrap. J. Econometrics 133 (2), 579–599.
Yuan, M., 2009. State price density estimation via nonparametric mixtures. Ann. Appl. Stat. 3 (3), 963–984.
Zhou, S., Shen, X., 2001. Spatially adaptive regression splines and accurate knot selection schemes. J. Amer. Statist. Assoc. 96 (453), 247–259.
Zhou, S., Shen, X., Wolfe, D.A., 1998. Local asymptotics for regression splines and confidence regions. Ann. Statist. 26 (5), 1760–1782.
Zhou, S., Wolfe, D.A., 2000. On derivative estimation in spline regression. Statist. Sinica 10, 93–108.