Universität St.Gallen

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

Matthias R. Fengler, Lin-Yee Hin

May 2013 Discussion Paper no. 2011-36

School of Economics and Political Science, Department of Economics

University of St. Gallen

Editor:	Martina Flockerzi University of St. Gallen School of Economics and Political Science Department of Economics Varnbüelstrasse 19 CH-9000 St. Gallen Phone +41 71 224 23 25 Fax +41 71 224 31 35 Email seps@unisg.ch
Publisher:	School of Economics and Political Science Department of Economics University of St. Gallen Varnbüelstrasse 19 CH-9000 St. Gallen Phone +41 71 224 23 25
Electronic Publication:	Fax +41 71 224 31 35 http://www.seps.unisg.ch

Semi-nonparametric estimation of the call price surface under strike and time-to-expiry no-arbitrage constraints[1]

Matthias R. Fengler, Lin-Yee Hin

Author's address:

Matthias R. Fengler
School of Economics and Political Science
University of St. Gallen
Bodanstrasse 6
CH-9000 St. Gallen
Phone +41 (0)71 224-2457
Fax +41 (0)71 224-2894
Email matthias.fengler@unisg.ch

Lin-Yee Hin
Department of Mathematics and Statistics
Curtin University of Technology
Kent Street
Bentley WA 6102
Australia
Email linyeehin@gmail.com

[1] We thank Francesco Audrino, Wolfgang Härdle, Anselm Ivanovas, Jens Jackwerth, Simon Knaus, Enno Mammen, Pirmin Meier and all participants of the Konstanz workshop on the "Pricing Kernel Puzzle" for valuable suggestions and comments. All remaining errors are our own.

Abstract

We suggest a semi-nonparametric estimator for the entire call price surface based on a tensor-product B-spline. To enforce no-arbitrage constraints in strike and calendar dimensions we establish sufficient no-arbitrage conditions on the control net of the tensor product (TP) B-spline. Since these conditions are independent of the degrees of the underlying polynomials, the estimator can be parametrized with TP B-splines of arbitrary order. We derive consistency and explore the statistical efficiency benefits from surface estimation. As an application, we estimate families of state price densities and a local volatility surface for S&P500 option data.

Keywords

Option pricing function, no-arbitrage constraints, state price density, implied volatility, local volatility, semi-nonparametric estimation, B-splines.

JEL Classification

C14, C58, G13

1 Introduction

Option prices bear information on the risk factors driving 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. Accordingly, for extraction of the information content inherent in option prices, numerous strategies have been suggested to estimate an option pricing function from an ensemble of observed option price data.

Unlike other financial instrument data such as stocks, futures, and bonds, 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, while Jackwerth (2000) and Bliss and Panigirtzoglou (2004) study the empirical pricing kernel and the implied risk aversion in a similar way. An assumption that is implicit in many of these studies is that option prices observed contemporaneously over multiple time horizons are thought of as realizations from a smooth surface defined across exercise prices and expiry dates.

In this paper, we propose an estimator for a European-style option pricing function that explicitly accounts for the surface structure of option data, i.e., we not only estimate a single option pricing function, but an entire option pricing surface across all available strikes and expiry dates. Since option data often are sparsely distributed, which might lead to ill-posed estimation problems, we additionally study a regularized version of the estimator that behaves in a numerically benign way. Both versions of the estimator are based on a tensor product (TP) B-spline of arbitrary degree and therefore belong to the class of semi-nonparametric estimators. As required by financial theory, we formalize sufficient no-arbitrage constraints across both the strike and the expiry dimension to warrant that the resulting estimate is free of arbitrage.

Accounting for the surface structure when estimating an option pricing function under shape constraints has rarely been achieved in the literature so far. In fact, it is common practice to estimate the option pricing function for each expiry date independently, i.e. "slice by slice", usually in combination with some interpolation of the data across the expiry dates. The simplest approach would be to specify a suitably flexible risk-neutral density from some parametric family and to minimize a cost functional of model and market prices of a fixed expiry date.[1] But most non- and semiparametric contributions for estimating the option pricing function also act on each expiry date separately.[2] On the other hand, common two-dimensional models, such as Aït-Sahalia and Lo (2000), Cont and da Fonseca (2002), Fengler et al. (2007), Giacomini et al. (2009), and Audrino and Colangelo (2010), do not accommodate financial no-arbitrage constraints. Pioneering approaches to include no-arbitrage constraints for surface estimation are Benko et al. (2007) and Glaser and Heider (2010) who employ locally polynomial estimators. Neither study, however, analyzes the asymptotic properties of the constrained estimators, nor do they attempt to assess the statistical efficiency benefits associated with implementing no-arbitrage constraints in the time-to-expiry dimension. Exploiting a projection framework for constrained smoothing by Mammen et al. (2001), we prove consistency of the estimator and provide an upper bound for the rate of convergence, which is Stone's optimal rate for unconstrained estimates. Through simulation results we show that aside from strike no-arbitrage constraints, substantial efficiency gains can ensue from additionally implementing time-to-expiry no-arbitrage constraints.

Naturally, usage of polynomial regression splines, such as B-splines, to estimate the option pricing function or the state price density is not new in itself; see Wang et al. (2004), Yatchew and Härdle (2006), Fengler (2009) and Monteiro et al. (2008), Härdle and Hlávka (2009), Laurini (2011). Yet our implementation strategy differs crucially from previous work. Rather than imposing the no-arbitrage constraints directly on the unknown regression function as does the extant literature, we derive sufficient conditions for no-arbitrage on the *control net* of the TP B-spline. The notion

[1] See Abadir and Rockinger (2003) as a prime example of such an approach, but the literature on alternative parametric specifications is vast. We refer to Jackwerth (2004) for a comprehensive account.

[2] See Jarrow and Rudd (1982), Madan and Milne (1994), Jackwerth and Rubinstein (1996), Jondeau and Rockinger (2001), Aït-Sahalia and Duarte (2003), Bondarenko (2003), Wang et al. (2004), Yatchew and Härdle (2006), Monteiro et al. (2008), Birke and Pilz (2009), Fan and Mancini (2009), Fengler (2009), Härdle and Hlávka (2009), Yuan (2009), and Grith et al. (2013); see Garcia et al. (2010, Section 5) for an excellent account of this literature.

of the control net, which is the set of points having certain averages of the knot sequences as abscissae and the B-spline coefficients as ordinates, originates from the Computer Aided Geometric Design literature, where spline curves rather than spline functions are studied. In some sense 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 for B-splines of arbitrary degree to be used for estimation (up to numerical limitations). This is in stark contrast to the aforementioned literature that is confined to employing at most cubic splines, given the complexity of imposing no-arbitrage constraints, such as convexity. Nevertheless, the no-arbitrage conditions on the control net are linear, which is in contrast to many other methods where constraints can assume quite involved forms. Constrained estimation can therefore be carried out by standard quadratic programming techniques. Remarkably, as it turns out, the sufficiency conditions also nicely embed as a special case recent theoretical results from Carr and Madan (2005) and Davis and Hobson (2007), which were obtained for linear option price surfaces. They thus have a natural interpretation in the context of the mathematical finance literature.

The study is structured as follows. The next section gives an overview on no-arbitrage constraints on the call price surface. In Section 3, we present the TP B-spline estimator and study the consistency. Least squares implementation under sufficient no-arbitrage constraints and optimal knot placement are discussed in Section 4. Simulations exploring the efficiency of surface no-arbitrage constraints and the choice of the regularization parameter are given in Section 5. Section 6 demonstrates the estimator for various applications, and Section 7 concludes. An appendix offers basic facts about B-splines, details the consistency proofs, and presents further supplementary results on monotonicity constrained smoothing.

2 Properties of an arbitrage-free call price surface

In this section, we summarize the no-arbitrage constraints on a European style call price surface for future reference. We focus on call options only, since from the put-call parity it is easy to obtain the corresponding put constraints. Before deriving the constraints, we transform observed call prices under a homogeneity assumption. More precisely, we express them as compounded call prices relative to the forward and estimate them as a function in terms of forward moneyness and time-to-expiry. The purpose of this transformation is twofold. First, from a statistical perspective, we resolve a potential nonstationarity issue, when the underlying asset price process is assumed to be nonstationary (Gouriéroux et al., 1995; Renault and Touzi, 1996). Second, the transformation will substantially alleviate the constraints implementation for the estimator. Consider, e.g., the case of a constant interest rate r. The no-arbitrage slope of the zero-strike call is then $-e^{-ry}$ and hence is a function of time-to-maturity y. A surface estimator must account for this. As it turns out, no-arbitrage shape constraints assume particularly convenient forms under this transformation even in the presence of deterministic interest rates and dividend yields.

2.1 Transformation under homogeneity

By the fundamental theorem of asset pricing, no-arbitrage implies the existence of a risk-neutral probability measure under which the discounted stock price process is a martingale (Dalang et al., 1990). This measure may depend, aside from the underlying asset price, on further state variables, such as latent volatility factors. At time 0, the price of a European-style call option with strike price X to be exercised at some future date y can be written as

$$\begin{aligned} C_0(S_0, X, y, r_y, \delta_y) &= e^{-r_y y}\mathrm{E}[(S_y - X)^+ | \mathcal{F}_0] \\ &= e^{-r_y y} \int_0^\infty (S_y - X)^+ \, \mathrm{Q}_y(dS_y) \,, \end{aligned} \tag{1}$$

where r_y is the risk-free interest rate and δ_y the dividend yield of the asset (both from 0 to y) . Both rates are assumed to be at most functions of time. By $u^+ = \max(u, 0)$, we denote the call payoff, and $\mathrm{E}[\cdot|\mathcal{F}_0]$ is the expectation conditional on the time-0 information set $\mathcal{F}_0$ taken under the risk-neutral measure Q.

For estimation of the option pricing function $C_0(S_0, X, y, r_y, \delta_y)$ simultaneously across strikes and several maturities, we transform observed call price data under a mild dimension reduction assumption and a homogeneity assumption. First, following Aït-Sahalia and Lo (2000), we assume that C_0 is not a function of (S_0, r_y, δ_y) separately, but depends on these variables through the forward price $F_0^y = e^{(r_y-\delta_y)y} S_0$ times a discount factor. More precisely, we assume $C_0(S_0, X, y, r_y, \delta_y) = e^{-r_y y} C_0(F_0^y, X, y)$. Second, we presume that C_0 is homogeneous of degree one in strikes and the forward.[3] Defining the strike measured in forward moneyness terms by $x = X/F_0^y$, homogeneity allows us to write

$$C_0(S_0, X, y, r_y, \delta_y) = e^{-r_y y} F_0^y C_0(1, x, y) \tag{2}$$

which implies the notation

$$\begin{aligned} z(x, y) \equiv C_0(1, x, y) &= \frac{e^{r_y y}\, C_0(S_0, X, y, r_y, \delta_y)}{F_0^y} \\ &= \mathrm{E}[(M_y - x)^+ | \mathcal{F}_t] = \int_x^\infty (M_y - x)\, \mathrm{Q}_y(dM_y)\ . \end{aligned} \tag{3}$$

Here, $M_y = S_y/F_0^y$ denotes the moneyness process relative to time 0. Note that M_y is a martingale under Q. In (3), $z(x, y)$ is the time 0 compounded call price surface relative to the forward and measured in forward moneyness x and time-to-maturity y.[4] As is obvious, under these transformations the shape constraints on the option pricing surface take the same form as in a situation with zero interest rate and zero dividend yield.

2.2 No-arbitrage conditions for the call price surface

Across strikes, necessary no-arbitrage shape constraints of the call price function are very well-known; see, e.g., Aït-Sahalia and Duarte (2003) for a thorough account. Necessary calendar arbitrage constraints for European options, similar to those known for American-style options, were first obtained by Reiner (2000) and Gatheral (2004). A rather recent literature established sufficient shape constraints on a call price surface. In Carr and Madan (2005) and Davis and Hobson

[3]Implications of this homogeneity property, first discussed in Merton (1973), are found in more detail in Garcia and Renault (1998a;b), Bates (2005), Alexander and Nogueira (2007), and Garcia et al. (2010, Section 2.1). As outlined in Renault (1997, Proposition 2) a necessary and sufficient condition for this homogeneity property to hold is that the pricing measure does not depend on F_0^y, i.e., under the pricing measure, future returns and the current asset price are independent conditionally on the current information set. This assumption does not exclude leverage effects. For a given informational setting, it "must be understood as a non-causality relationship in the Granger sense between the current [forward] price and future returns [...] and not as independence property" (Renault, 1997, pp. 244).

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

(2007), for the case of a discrete ensemble of call options, and in Roper (2010), for the case of a continuous call price surface, sufficient shape constraints are derived. These constraints are shown to imply the existence of a (non-negative) Markov martingale M_y with $M_0 = 1$ on some stochastic basis such that $z(x, y) = \mathrm{E}[(M_y - x)^+|\mathcal{F}_0]$ for every $(x, y) \in [0, \infty) \times [0, \infty)$. Hence, by the fundamental theorem of asset pricing, the call price surface does not admit arbitrage. Due to the assumptions and transformations outlined in Section 2.1, we can summarize the necessary and sufficient no-arbitrage constraints by citing Theorem 2.1 of Roper (2010), who assumes absence of interest rates and dividend yields.[5]

Proposition 2.1 *Define the function*

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

such that $z(x, y)$

(C1) is convex in x *for all* $y \geq 0$*,*

(C2) is bounded according to

$$(1 - x)^+ \leq z(x, y) \leq 1 \ , \tag{4}$$

for all $x \geq 0, y \geq 0$*,*

(C3) obeys $z(x, 0) = (1 - x)^+$ *for all* x*,*

(C4) has $\lim_{x\to\infty} z(x, y) = 0$ *for all* y*,*

(C5) and is monotonic in y *for all* $x \geq 0$*.*

Then

(i) there exists a non-negative Markov martingale M_y *such that*

$$z(x, y) = \mathrm{E}[(M_y - x)^+|\mathcal{F}_0]$$

for all $x, y \geq 0$*, and*

[5] We present the theorem in a slightly less general version than in Roper (2010).

(*ii*) *under no-arbitrage, $z(x, y)$ fulfills properties (C1)-(C5) for all $x, y \geq 0$.*

Proof (*i*) We only sketch Roper's argument here. For each fixed y, (C1)-(C4) give rise to a complementary distribution function, which defines a measure Q_y such that

$$z(x, y) = \int_0^\infty (v - x)^+ \mathrm{Q}_y(dv) . \tag{5}$$

The key now is that (C5) implies that the measures Q_y indexed by y form a family of measures which is in convex stochastic order. It then follows from Kellerer (1972, Theorem 3) that there exists a Markov martingale M_y with marginals given by Q_y.

(*ii*) (C1)-(C4) are obtained from elementary arbitrage considerations. (C5) follows from noting that for $0 \leq y_1 \leq y_2 < \infty$

$$z(x, y_2) = \mathrm{E}[(M_{y_2} - x)^+ | \mathcal{F}_0] = \mathrm{E}[\mathrm{E}[(M_{y_2} - x)^+ | \mathcal{F}_{y_1}] | \mathcal{F}_0] \geq \mathrm{E}[(M_{y_1} - x)^+ | \mathcal{F}_0] = z(x, y_1)$$

by Jensen's inequality and the martingale property. ∎

3 The tensor-product B-spline call price surface

3.1 Model

We consider the model

$$Z_i = z_0(X_i, Y_i) + \varepsilon_i , \tag{6}$$

where $z_0 : [\underline{x}, \overline{x}] \times [\underline{y}, \overline{y}] \to (0, 1]$ is a smooth unknown call price surface free of arbitrage opportunities, $\{(X_i, Y_i)\}_{i=1}^n \in [\underline{x}, \overline{x}] \times [\underline{y}, \overline{y}]$ (with lower and upper bounds $\underline{x} \ll \overline{x}$ and $\underline{y} \ll \overline{y}$) are random design points, ε_i is an error term, and $\{(X_i, Y_i), Z_i\}_{i=1}^n$ is *i.i.d.* as $((X, Y), Z)$, with $z_0(x, y) = \mathrm{E}[Z|x, y]$ and $\sigma^2 = \mathrm{Var}[Z|x, y]$ being bounded.

The purpose is to approximate z_0 by a two-dimensional polynomial spline that is an element in the space

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

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 moneyness and time-to-maturity dimension, respectively. More precisely, we define

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

where 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

$$\underline{x} = \xi_0 = \cdots = \xi_{p_1} < \xi_{p_1+1} < \cdots < \xi_{q_1} < \xi_{q_1+1} = \cdots = \xi_{q_1+p_1+1} = \overline{x} , \tag{7}$$

defined for some integer $q_1 > p_1$. For two integers $q_2 > p_2 > 0$ and a knot sequence $\boldsymbol{\upsilon}$:

$$\underline{y} = \upsilon_0 = \cdots = \upsilon_{p_2} < \upsilon_{p_2+1} < \cdots < \upsilon_{q_2} < \upsilon_{q_2+1} = \cdots = \upsilon_{q_2+p_2+1} = \overline{y} \tag{8}$$

$\mathcal{S}(p_2;\boldsymbol{\upsilon})$ is defined likewise. It is well known (de Boor, 2001) that $\mathcal{S}(p_1;\boldsymbol{\xi})$ and $\mathcal{S}(p_2;\boldsymbol{\upsilon})$ are spanned by a set of nonnegative basis functions $B_{j_1,p_1}(x)$, $j_1 = 0,\dots,q_1$, and $B_{j_2,p_2}(y)$, $j_2 = 0,\dots,q_2$, respectively, the so called B-splines of degree p_1 and p_2 that are defined over $\boldsymbol{\xi}$ and $\boldsymbol{\upsilon}$; see Appendix A for some facts on B-splines. In adopting (p_k+1)-fold coincident boundary knots, $k = 1,2$, the support of the basis functions is contained in $[\underline{x},\overline{x}]$ and $[\underline{y},\overline{y}]$. It then follows that $\mathcal{S}(p_1,p_2;\boldsymbol{\xi},\boldsymbol{\upsilon})$ has basis functions $B_{j_1,p_1}(x)B_{j_2,p_2}(y)$, $j_1 = 0,\dots,q_1, j_2 = 0,\dots,q_2$, that form a partition of unity and are nonnegative on $[\underline{x},\overline{x}] \times [\underline{y},\overline{y}]$. Any $s \in \mathcal{S}(p_1,p_2;\boldsymbol{\xi},\boldsymbol{\upsilon})$ can be written as $s(x,y) = \sum_{j_1=0}^{q_1}\sum_{j_2=0}^{q_2}\theta_{j_1,j_2}\, B_{j_1,p_1}(x)\, B_{j_2,p_2}(y)$ 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 $J_n = (q_1+1)(q_2+1)$. By the notation J_n we stress that as n grows we regard finer and finer approximations to the call 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

$$\hat{z} = \underset{s\in\mathcal{S}(p_1,p_2;\boldsymbol{\xi},\boldsymbol{\upsilon})}{\arg\min}\ \frac{1}{n}\sum_{i=1}^{n}(Z_i - s(X_i,Y_i))^2 \ . \tag{9}$$

This estimate is likely to violate no-arbitrage restrictions. An appropriate estimate (i.e. arbitrage-free) would be given by

$$\hat{z}^C = \underset{s \in \mathcal{S}^C(p_1,p_2;\boldsymbol{\xi},\boldsymbol{\upsilon})}{\arg\min} \; \frac{1}{n}\sum_{i=1}^{n}(Z_i - s(X_i, Y_i))^2 \;, \tag{10}$$

where $\mathcal{S}^C(p_1, p_2; \boldsymbol{\xi}, \boldsymbol{\upsilon}) = \{s : s \in \mathcal{S}(p_1, p_2; \boldsymbol{\xi}, \boldsymbol{\upsilon}),\ 0 \le s \le 1,\ -1 \le s_x \le 0,\ s_{xx} \ge 0,\ s_y \ge 0\}$, where s_x denotes the partial derivative of s with respect to x. Note that in the definition of $\mathcal{S}^C$ we replaced all strict inequalities by weak inequalities to avoid open intervals.[6]

As we will discuss below, due to the special structure of option data, it may happen that (9) and (10) lead to ill-posed optimization problems. We therefore additionally introduce the penalized estimators

$$\hat{z}_\lambda = \underset{s \in \mathcal{S}(p_1,p_2;\boldsymbol{\xi},\boldsymbol{\upsilon})}{\arg\min} \; \frac{1}{n}\sum_{i=1}^{n}(Z_i - s(X_i, Y_i))^2 \;+\; \lambda_n|\boldsymbol{\theta}|^2 \tag{11}$$

and

$$\hat{z}_\lambda^C = \underset{s \in \mathcal{S}^C(p_1,p_2;\boldsymbol{\xi},\boldsymbol{\upsilon})}{\arg\min} \; \frac{1}{n}\sum_{i=1}^{n}(Z_i - s(X_i, Y_i))^2 \;+\; \lambda_n|\boldsymbol{\theta}|^2 \;, \tag{12}$$

which are well-defined for a sequence of nonnegative numbers λ_n. Here, $|\cdot|$ denotes the Euclidian vector norm, and $\boldsymbol{\theta}$ is the vector comprising all B-spline coefficients θ_{j_1,j_2}, for $j_1 = 0, \ldots, q_1, j_2 = 0, \ldots, q_2$. The penalization follows the ideas of ridge regression.

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 were studied by, among others, Agarwal and Studden (1980), Gallant (1982), Andrews (1991), Stone (1994), Newey (1997), Zhou et al. (1998) and Huang (2003). An exhaustive treatment is found in Chen (2007). Shape-restricted regression estimators, such as monotone regression estimators, were explored in Brunck (1970), Wright (1981) and Mammen (1991a). Consistency and rates of convergences for certain concave/convex regression functions were established in Hanson and Pledger (1976) and Mammen (1991b), while the limiting distribution of a convex function at

[6] Strictly speaking, $s = 0$ would lead to arbitrage. The 0, however, can always be replaced by some small $\epsilon > 0$, since the support is bounded.

a fixed point has recently been derived by Groeneboom et al. (2001a;b). As a bottom line from this literature, under convexity, a constrained estimate *may* achieve faster rates of convergence than the corresponding unconstrained estimate. This is proved in Mammen and Thomas-Agnan (1999) for smoothing splines. 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 only consider consistency.

3.2.1 Assumptions

We adopt the following standard assumptions:

(A1) z_0 is r-smooth[7] for some $r \geq 3$.

(A2) The design density of (x, y) has compact support, is absolutely continuous and is bounded away from zero and infinity.

(A3) $\{(X_i, Y_i), Z_i\}_{i=1}^n$ is *i.i.d.* as $((X, Y), Z)$, and $z_0(x, y) = \mathrm{E}[Z|x, y]$ and $\sigma^2 = \mathrm{Var}[Z|x, y]$ are bounded.

(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.[8]

Assumption (A1) is a typical smoothness assumptions made for series estimators, and $r \geq 3$ en-

[7]Denote by k an integer and set $r = k + \beta$, $0 < \beta \leq 1$. A function f on $[\underline{x}, \overline{x}] \times [\underline{y}, \overline{y}]$ 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,

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

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

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

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

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

sures the existence of a family of continuous state price densities of z_0. (A2) is standard and required for identification. The continuity assumption on the design density may be seen critically for conditional estimates, since on a single trading day only a discrete set of expiries is listed on a futures exchange. It should be noted, however, that aside from listed options there typically exist very liquid broker markets, on which other expiries are actively quoted over-the-counter. For unconditional estimates that comprise several months of data, this issue is obviously not relevant. The homoscedasticity assumption (A3) might by questioned for financial data, but is made for ease of exposition only. Heteroscedastic error terms can be accommodated by endowing the least squares criteria and norms introduced below and in the appendix with nonnegative weights $w(X_i, Y_i)$ that are bounded away from zero and infinity. Asymptotically, however, the model will be locally homoscedastic by smoothness of z_0. (A4) is due 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 Conditions 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*

$$\left(\frac{1}{n}\sum_{i=1}^{n}(\hat{z}^C(X_i, Y_i) - z_0(X_i, Y_i))^2\right)^{1/2} \leq \left(\frac{1}{n}\sum_{i=1}^{n}(\hat{z}(X_i, Y_i) - z_0(X_i, Y_i))^2\right)^{1/2} + O_P(J_n^{-r}) \tag{13}$$

$$= O_P\left(\left(\frac{J_n}{n} + J_n^{-r}\right)^{1/2}\right) + O_P(J_n^{-r})\ . \tag{14}$$

For the penalized estimate, 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 get

$$\left(\frac{1}{n}\sum_{i=1}^{n}(\hat{z}_\lambda^C(X_i, Y_i) - z_0(X_i, Y_i))^2\right)^{1/2} \leq \left(\frac{1}{n}\sum_{i=1}^{n}(\hat{z}_\lambda(X_i, Y_i) - z_0(X_i, Y_i))^2\right)^{1/2} + O_P\left((J_n/n + J_n^{-r})^{1/2}\right) \tag{15}$$

$$= O_P\left(\left(\frac{J_n}{n} + J_n^{-r}\right)^{1/2}\right)\ . \tag{16}$$

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, one ought to be at least as good as (if not better than) in the first projection step, when constraints are neglected. Hence, $\hat{z}^C$ and $\hat{z}_\lambda^C$ must attain at least the convergence rates of the unconstrained estimates.

In Proposition 3.1, in the expressions $O_P\left(J_n/n + J_n^{-r}\right)$, J_n/n corresponds to the variance and J_n^{-r} to the squared bias of the unconstrained estimates. Optimally balancing these terms, i.e., choosing $J_n/n \asymp J_n^{-r}$, and observing that the remainder terms in (14) and (16) are of smaller order, one receives the optimal rate of convergence for a bivariate unconstrained function $n^{-r/(r+1)}$; see Stone (1982). In consequence, both $\hat{z}^C$ and $\hat{z}_\lambda^C$ converge at Stone's optimal rate as an upper bound.

4 Least squares fitting under linear no arbitrage constraints

4.1 Quadratic programming

Given data (X_i, Y_i, Z_i), $i = 1, \ldots, n$, 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.,

$$\mathbf{B} = \left(\mathbf{b}_{(p_1,p_2)}(X_1, Y_1), \ldots, \mathbf{b}_{(p_1,p_2)}(X_n, Y_n)\right)^\top . \tag{17}$$

Furthermore, let $\mathbf{z}$ be the $n \times 1$ vector of observed relative call prices and $\boldsymbol{\theta}$ the $(q_1 + 1)(q_2 + 1) \times 1$ vector of unknown coefficients. Then (10) and (12) can be cast into the quadratic program

$$\begin{aligned} & \widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \ \frac{1}{2}\boldsymbol{\theta}^\top \mathbf{D}\boldsymbol{\theta} - \boldsymbol{\theta}^\top \mathbf{d} \ , \\ \text{subject to} \quad & \hat{z}^C_\lambda = \mathbf{b}^\top \widehat{\boldsymbol{\theta}} \in \mathcal{S}^C(p_1, p_2; \boldsymbol{\xi}, \boldsymbol{\upsilon}) \end{aligned} \tag{18}$$

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. We set $\lambda > 0$ for the penalized 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 (18) can be solved by means of quadratic programming, such as the Goldfarb and Idnani (1983) algorithm.

The quadratic program (18) is well-posed if $\mathbf{D}$ is strictly positive definite. Practically, this condition may fail due to the quoting conventions on options markets. In contrast to univariate B-splines, unfortunately conditions of the Schoenberg-Whitney type for multivariate B-spline approximations do not exist 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 penalized estimator. In Section 5.2 we will discuss the choice of λ in more detail.

4.2 Sufficient conditions for no-arbitrage on the control net

We now discuss how to warrant 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-splines coefficients θ_{j_1,j_2}. We emphasize that the key to constraints implementation is the choice of coalescent boundary knots in the definitions of the knot sequences (7) and (8).

4.2.1 Control net

For two strictly increasing sequences of abscissae $\xi_0^*, \dots, \xi_{q_1}^*$ and $\upsilon_0^*, \dots, \upsilon_{q_2}^*$, 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 $c(\xi_{j_1}^*, \upsilon_{j_2}^*) = \theta_{j_1,j_2}$, $j_1 = 0, \dots, q_1$ and $j_2 = 0, \dots, q_2$, and which is bilinear on the rectangle

$$R_{j_1,j_2} = [\xi_{j_1}^*, \xi_{j_1+1}^*] \times [\upsilon_{j_2}^*, \upsilon_{j_2+1}^*] \ . \tag{19}$$

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

$$c_{j_1,j_2}(x,y) = \sum_{k=0}^{1} \sum_{\ell=0}^{1} \alpha_{j_1,k}(x) \beta_{j_2,\ell}(y) \, \theta_{j_1+k,j_2+\ell} \ . \tag{20}$$

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 $\xi_{j_1}^*$, $j_1 = 0, \dots, q_1$, and $\upsilon_{j_2}^*$, $j_2 = 0, \dots, q_2$, associated with the knot sequences $\boldsymbol{\xi}$ and $\boldsymbol{\upsilon}$ are defined by

$$\xi_{j_1}^* = \frac{\xi_{j_1+1} + \cdots + \xi_{j_1+p_1}}{p_1} \qquad \text{and} \qquad \upsilon_{j_2}^* = \frac{\upsilon_{j_2+1} + \cdots + \upsilon_{j_2+p_2}}{p_2} \ . \tag{21}$$

The set of points $(\xi_{j_1}^*, \upsilon_{j_2}^*, \theta_{j_1,j_2})$, $j_1 = 0, \dots, q_1$, $j_2 = 0, \dots, q_2$ are known as control points of the surface. This interpretation of this set of points is due to the linear precision property of the B-spline which allows one to interpret the spline function in fact as a spline curve; see the appendix for further details.

Figure 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 (7) and (8), 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). For ease of visual perception, the convex hull of the control polygons belonging to the first and the last segment of the spline are shaded.

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 Figure 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* (7) *and* (8). *If*

$$\theta_{j_1,j_2} \geq 0 \qquad \textit{and} \qquad \theta_{j_1,j_2} \leq 1 \ , \tag{22}$$

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 $(x, y) \in [\underline{x}, \overline{x}] \times [\underline{y}, \overline{y}]$. ∎

4.2.3 Bounds on first-order derivatives and convexity

The recurrence relation of B-splines due 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 Figure 1 and the further explanations and references in the appendix). A convex set of control points therefore implies a globally convex function. As we show in Proposition 4.2 below, this also holds true for the TP B-spline.

It follows from the assumptions in Proposition 2.1 that first-order derivatives of the call price surface must be between minus one and zero. 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 to only constrain the slope at the boundaries of the support, i.e. in $\underline{x}$ and $\overline{x}$ 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*

$$-1 \leq \frac{\theta_{j_1+1,j_2} - \theta_{j_1,j_2}}{\xi^*_{j_1+1} - \xi^*_{j_1}} \leq \frac{\theta_{j_1+2,j_2} - \theta_{j_1+1,j_2}}{\xi^*_{j_1+2} - \xi^*_{j_1+1}} \leq 0 \, , \tag{23}$$

is a sufficient condition such that $z(x,y)$ *is convex* x *and it holds that*

$$-1 \leq \frac{\partial z(x,y)}{\partial x} \leq 0 \tag{24}$$

for any y*.*

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

$$z(x, \breve{y}) = \sum_{j_1=0}^{q_1} B_{j_1,p_1}(x) \sum_{j_2=0}^{q_2} \theta_{j_1,j_2} B_{j_2,p_2}(\breve{y}) \tag{25}$$

$$= \sum_{j_1=0}^{q_1} B_{j_1,p_1}(x) Q_{j_1}(\breve{y}) \tag{26}$$

where $Q_{j_1}(\breve{y}) = \sum_{j_2=0}^{q_2} \theta_{j_1,j_2} B_{j_2,p_2}(\breve{y})$. This is a univariate spline curve with control points $(\xi^*_{j_1}, Q_{j_1}(\breve{y}))$, $j_1 = 0, \dots, q_1$. From the remarks preceding the proposition it is sufficient to show that (23) implies

$$-1 \leq \frac{Q_{j_1+1}(\breve{y}) - Q_{j_1}(\breve{y})}{\xi^*_{j_1+1} - \xi^*_{j_1}} \leq \frac{Q_{j_1+2}(\breve{y}) - Q_{j_1+1}(\breve{y})}{\xi^*_{j_1+2} - \xi^*_{j_1+1}} \leq 0 \ . \tag{27}$$

To see this, take any of the interior inequalities from (23). Multiplying with $B_{j_2,p_2}(\breve{y})$ and summing over all $j_2 = 0, \dots, q_2$ yields by $B_{j_2,p_2}(\breve{y}) \geq 0$ that

$$\frac{\sum_{j_2=0}^{q_2} B_{j_2,p_2}(\breve{y}) \left(\theta_{j_1+1,j_2} - \theta_{j_1,j_2}\right)}{\xi^*_{j_1+1} - \xi^*_{j_1}} \leq \frac{\sum_{j_2=0}^{q_2} B_{j_2,p_2}(\breve{y}) \left(\theta_{j_1+2,j_2} - \theta_{j_1+1,j_2}\right)}{\xi^*_{j_1+2} - \xi^*_{j_1+1}} \ , \tag{28}$$

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 of Floater and Peña (1998). Due to their importance, the proofs are given in Appendix C.

First, recall that a bivariate function f is called increasing in direction $d = (d_1, d_2) \in \mathbb{R}^2$, if

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

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

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

$$d_1 \Delta_1 \theta_{j_1,j_2+\ell} + d_2 \Delta_2 \theta_{j_1+k,j_2} \geq 0 \ , \qquad k, \ell \in \{0, 1\} \tag{29}$$

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 next results states that 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* (21). *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 together with Theorem 4.4 allows 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* $(\xi^*_{j_1}, \upsilon^*_{j_2}, \theta_{j_1,j_2})$, $j_1 = 0, \dots, q_1$, $j_2 = 0, \dots, q_2$ *be the control points associated with the call price 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*

$$\theta_{j_1,j_2+1} - \theta_{j_1,j_2} \geq 0 \tag{30}$$

for $j_1 = 0, \dots, q_1$ *and* $j_2 = 0, \dots, q_2 - 1$, *the call price surface* $z(x, y)$ *is monotonically increasing in y.*

4.3 Further characterization of no-arbitrage conditions

The sufficient conditions for no-arbitrage have a remarkable relation to previous theoretical results. In Carr and Madan (2005) and Davis and Hobson (2007) it is investigated under which conditions a discrete ensemble of call prices (observed without error) is free of arbitrage.[9] The solution is given

[9] See Gope and Fries (2011) for a related approach.

by a bilinear support function defined on call prices and strikes which obeys the very same convexity and monotonicity conditions as stated above in Propositions 4.1 and 4.2 and Corollary 4.5. Given this striking resemblance, it is natural to ask how these results are related.

As a main difference, the Carr-Madan-Davis-Hobson conditions are directly defined on the (linearly interpolated) call price surface itself, whereas ours are stated on a meta-object, the control net, which enforces no-arbitrage on the resulting call price surface. Both notions, however, become identical when choosing a linear TP B-spline. In fact, the degree-one TP B spline surface is the only case at finitely many knots, for which the control net and the spline surface coincide. The Carr-Madan-Davis-Hobson sufficiency conditions are therefore embedded as a special case into our general set of conditions allowing for surfaces of arbitrary polynomial degree.

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. Practically, however, knot placement is an important parametrization device. We address knot placement in this section.

For estimation of the call price surface, knot selection is challenging from two perspectives. First, it is a two-dimensional optimization problem. This fact introduces a higher degree of complexity than usually considered in the literature. Second, rather than optimizing knots, one ideally would like to 'optimize' the placement of Greville sites, since all constraints are implemented through the control net. Unfortunately, there is no direct way to do this, because (21) maps $(q_k - p_k)$ interior knots into $(q_k - 1)$ interior Greville sites, $k = 1, 2$. As explained in Section 4.3, only for degree $p_1 = p_2 = 1$, is this mapping invertible, since both sequences are identical in this case.

In the algorithms described below we address these two issues as follows. Taking into consideration that there is comparably very little variation in the term structure, we ignore the expiry dimension in the knot selection problem. The algorithms could be extended for a two-dimensional case, but we expect little gains through further optimization. To our experience, typically three to four interior knots, depending on the length of the expiry axis and the degree of the B-spline in this direction, are sufficient to capture the salient features of the term structure of the call price surface. The solution of the second issue proceeds from the aforementioned observation that Greville sites and the knot sequence coincide for the linear B-spline. We therefore start our knot search with a linear TP B-spline. This yields an optimal knot sequence (and hence control polygon) for the linear call price surface, but a suboptimal one for higher order B-splines. In the subsequent knot relocation and deletion process we then employ TP B-splines of the desired degrees $p_1, p_2 > 1$. In this final process, the optimal knot sequence is determined. We now give a more formal description of these algorithms. The algorithms build with only minor modifications on schemes with adaptive knot search and subsequent knot optimizations, such as in Zhou and Shen (2001).

Algorithm 1: Knot search

For the knot search a linear TP B-spline is used, i.e. $p_1 = p_2 = 1$. No-arbitrage constraints are not imposed. The algorithm starts with no interior knots, i.e., initially the working knot sequence is $\boldsymbol{\xi} = \{\xi_0, \xi_1, \xi_2, \xi_3\}$ where $\underline{x} = \xi_0 = \xi_1 < \xi_2 = \xi_3 = \overline{x}$, and searches only over the forward moneyness dimension. Denote by $\boldsymbol{\xi}^u$ the knot sequence book-keeping 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 is sufficiently far away from interval boundary, $\Delta_\xi > 0$.

Step 1: The working knot sequence is supposed to be $\boldsymbol{\xi} = \{\xi_0, \dots, \xi_{q_1+2}\}$. Set $\boldsymbol{\xi}^u = \boldsymbol{\xi}$. For each existing subinterval $[\xi_i, \xi_{i+1}]$, $i = 1, \dots, q_1$, find an additional knot $\xi^o \in [\xi_i, \xi_{i+1}]$ that minimizes a selection criterion Ξ.

Insert ξ^o into $\boldsymbol{\xi}^u$, if

$$\Xi(\boldsymbol{\xi}) - \Xi(\xi^o \cup \boldsymbol{\xi}) > \Delta_\Xi$$

and if

$$\xi^o - \xi_i > \Delta_\xi \ \wedge \ \xi_{i+1} - \xi^o > \Delta_\xi \ .$$

Step 2: Put $\boldsymbol{\xi} = \boldsymbol{\xi}^u$. If one or more knots are added in Step 1, return to Step 1; otherwise stop.

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 algorithm still searches only over the forward moneyness dimension, but the full set of no-arbitrage constraints is imposed to ensure that knot relocation and deletion is achieved optimally in the presence of the constraints. The algorithm leads to the deletion of a knot if deleting ξ_{i+1} is associated with a value of Ξ lower than the one associated with relocating or with keeping ξ_{i+1}. If relocation of ξ_{i+1} improves Ξ overall relocate. If neither deletion nor relocation of ξ_{i+1} results in a reduction of Ξ, no adjustment is made.

Step 1: Denote by $\boldsymbol{\xi}^u$ the knot sequence book-keeping 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 $(\xi_{i+1} \cup \boldsymbol{\xi}_i) = \boldsymbol{\xi}$. Find an additional knot $\xi^o_{i+1} \in [\xi_i, \xi_{i+2}]$ that minimizes a 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

$$\Xi(\boldsymbol{\xi}_i) < \Xi\left(\xi^o_{i+1} \cup \boldsymbol{\xi}_i\right) \quad \wedge \quad \Xi(\boldsymbol{\xi}_i) < \Xi(\boldsymbol{\xi}) \ ,$$

set $\boldsymbol{\xi}^u = \boldsymbol{\xi}_i$;

(b) if

$$\Xi\left(\xi^o_{i+1} \cup \boldsymbol{\xi}_i\right) < \Xi(\boldsymbol{\xi}_i) \quad \wedge \quad \Xi(\xi^o_{i+1} \cup \boldsymbol{\xi}_i) < \Xi(\boldsymbol{\xi}) \ ,$$

set $\boldsymbol{\xi}^u = (\xi^o_{i+1} \cup \boldsymbol{\xi}_i)$;

(c) else 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.

As a selection criterion we use the modified Akaike Information Criterion of Hurvich et al. (1998) which is designed for model choice based on regression splines. It is given by

$$\Xi_{\text{mAIC}} = \log(\text{RSS}) + 1 + \frac{2(\text{tr}(\mathbf{S}) + 1)}{n - \text{tr}(\mathbf{S}) - 2} \ , \tag{31}$$

where RSS $= \sum_{i=1}^{n}(z_i - \hat{z}_i)^2/n$ 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 and at the same time conservative procedure. Suitable alternative criteria could be the generalized cross-validation statistic or the Schwarz criterion.

5 Simulations

In this section, we gauge the efficiency benefits from estimating a call price surface while imposing no-arbitrage constraints and study the practical impact of the regularization parameter.

5.1 Efficiency study

The setting of this Monte Carlo study is as in Birke and Pilz (2009). We generate option price data from the jump-diffusion model introduced by Bates (1996) for a time-to-maturity of 1, 2, 3, 6, 12, and 24 months. Interest rates and dividends are assumed to be zero, 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 corresponds to the usual reality for options written on large indices, such as the S&P 500. Prices are additively perturbed with *i.i.d.* mean zero normally distributed error with variance $\sigma_\varepsilon^2 = 0.01^2$. This implies a signal to noise ratio of about 12. Given perturbed prices, we estimate the call price surface using the spline estimator in three ways: (a) estimate a single-date (6 months) option pricing function while disregarding the surface (univariate, 'slice-by-slice' approach); (b) estimate the surface without imposing calendar no-arbitrage constraints; (c) estimate the surface using all constraints. We use the penalized estimator with $\lambda = 10^{-6}$, three time-to-maturity knots and eleven knots in the strike dimension (counting interior knots only). We do this 5000 times for a number of different smoothness assumptions on the splines. 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 can therefore be interpreted as efficiency gains.

insert Table 1 about here

Skimming the numbers in Table 1, we see that the gains in MSE from imposing the full set of constraints mostly range between 10% and 80% percent. The smallest gains of about 1% are only obtained for moneyness in the neighborhood of one using the most parsimonious time-to-maturity specification, the linear spline ($p_2 = 1$). But even in this specification, moving away into the neighborhood of 0.8 or 1.2 moneyness gains are already as large as 12%. The largest efficiency gains are obtained at the boundary of the estimation domain, particularly due to a reduction in bias. These observations appear to be natural, but the size of the efficiency gains is striking.

Comparing (a) the slice-by-slice approach and (b) the simple surface estimator, we find that differences are tiny. Thus, using a surface estimator without calendar constraints hardly improves efficiency. Slice-by-slice performs equally well. Also, allowing for a more flexible parametrization in the strike dimension (p_1) scarcely has any impact on MSE. This can be expected since strike constraints are present in all three approaches. For higher degree splines along the time-to-maturity dimension (p_2), however, calendar constraint become increasingly beneficial.[10]

In summary, we find substantial efficiency gains from fully implementing calendar constraints, even when considering parsimonious parametrizations and moderate ranges of moneyness only.

5.2 Impact of the regularization parameter

As discussed in Section 4.1, due to the quoting conventions of futures exchanges, only a discrete set of expiries is listed. Other expiries might only be observed on the over-the-counter market.

[10]Note that in the slice-by-slice approach, any changes in the numbers along the p_2-dimension are entirely due to variation in the denominator (the full constraints case). The numerator necessarily stays constant.

As a consequence, the sample might be thinly populated for time-to-maturities between exchange-traded expiries. This could lead to an ill-posed QP, which is addressed by choosing the penalized estimator.

In order to understand the impact of penalization in the constrained estimation problem, we do an L-curve analysis as suggested in Hansen and O'Leary (1993). 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 more details on this fact.

In Figure 2, we display the resulting curve for a wide range of possible values of λ. We compare an unconstrained estimate with a fit taking advantage of the full set of no-arbitrage constraints, both in strike and calendar dimension. The estimations are carried out on the option data set from 1 December, 2010; see Section 6.1 for the description. In the top panel, for the unconstrained fit, a typical L-curve is obtained. For $\lambda \leq 10^{-12}$, 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 a deteriorating fit as is expressed in the increasing residual sum of squares.

This is in stark contrast to the shape-constrained solution. No L-curve is obtained. For all $\lambda \in [10^{-20}, 10^{-1}]$ almost the same combinations of the penalty term and the residual sum of squares are recovered. This is because monotonicity and convexity constraints that are imposed on the call price surface act as strong smoothing devices. They therefore enforce very similar solutions for a wide range of the regularization parameter; see Dole (1999) for similar observations. In conclusion, a precise choice of λ is practically immaterial.

6 Empirical applications

6.1 Data

For the demonstration of the estimators we use end-of-day quotes of S&P500 option data obtained from the OptionMetrics database. Along with the option data, S&P500 closing prices and zero rates are provided. The data include all observations from 2009-2010 and 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. (2011) for details. After filtering we compute mid quotes 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 to be the one with the minimal (absolute) difference between put and call price.

The unconditional estimates make use of the entire data set; the conditional estimates use the data of a single date, namely from Wednesday, December 1, 2010. In the first case 168,419 observations are available, for the latter single-date case 397. For estimation, puts are converted into calls by put-call parity.

6.2 Conditional estimates

To demonstrate the penalized estimator, we choose data of a single date. We discuss parametrization and knot sequences and display the estimate of the call price surface. As a most useful application of a conditional fit, we also derive the local volatility surface.

6.2.1 Parametrizations and knot sequences

Due to data availability of calls, we limit the analysis to the rectangle $[\underline{x}, \overline{x}] \times [\underline{y}, \overline{y}] = [0.60, 1.30] \times [0.04, 1.00]$. We employ a B-spline of degree five across forward moneyness and of a cubic degree for the time to expiry direction, i.e. $p_1 = 5, p_2 = 3$. The regularization parameter is set to $\lambda = 10^{-7}$, and the knot search and relocation algorithms are parametrized with $\Delta_{\text{Crit.}} = 10^{-5}$ and $\Delta_{\xi} = 10^{-4}$.

The knot sequence $\boldsymbol{\upsilon}$ in time to expiry is fixed at $\boldsymbol{\upsilon} = \{0.047, 0.150, 0.500, 0.834\}$ without enumerating coinciding boundary knots. After Algorithm 1 is run, the selected raw knot sequence in forward moneyenss direction is given by $\boldsymbol{\xi} = \{0.603, 0.868, 0.936, 1.007, 1.072, 1.133, 1.304\}$ with $\Xi_{\text{mAIC}} = -12.902$. Subsequent knot relocation and deletion in Algorithm 2 leads to an adjusted knot sequence $\boldsymbol{\xi} = \{0.603, 0.872, 0.936, 1.005, 1.072, 1.133, 1.304\}$ with no knots being deleted. The selection criterion is only very moderately improved: $\Xi_{\text{mAIC}} = -12.914$. We interpret the tiny adjustments occurring in the relocation process as an indication that the initial set of knots obtained for the linear TP B-splines are already a very good approximation to the optimal set of knots of the higher degree spline.

The control net that is the outcome of these optimizations is depicted in Figure 3. As explained, the control net mimics the shape of the underlying surface. At the same time it represents the set of sufficient no-arbitrage conditions on the call price surface.

6.2.2 The conditional call price surface

Figure 4 displays the estimated call price surface. Input data are shown as black dots. It is seen that it smoothly adapts to the curvature delineated by the input data. The correct implementation of the no-arbitrage constraints can be best inferred from the derivatives. For instance, in Figure 5 we present the first-order derivative with respect to forward moneyness. Obviously, the no-arbitrage constraint requiring that $-1 \leq \partial z/\partial x \leq 0$ is exactly matched for very short dated options only. This implies that second-order strike derivatives will only be interpretable as a density after suitable norming; see Section 6.3.1 for the details.

6.2.3 Local volatility surface

The sufficiency argument for no-arbitrage in a call price surface follows from an existence statement of a non-negative Markov martingale which has the same marginals as those implied from the call price functions over the various expiry dates. Dupire (1994) and Derman and Kani (1998) were the first to show how to construct such a Markov martingale from a call price surface.[11] Restricting the moneyness process M_t to follow a continuous Markov martingale, its dynamics under the risk-neutral measure are given by

$$dM_t = \sigma(M_t, t) M_t \, dW_t \,, \tag{32}$$

where W_t denotes a Wiener process and $\sigma(\cdot)$ a deterministic local volatility function of unknown functional form. Using the Kolmogorov forward equation, Dupire (1994) showed that local variance can be expressed by

$$\sigma_0^2(x, y) = 2 \frac{\partial z}{\partial y} \left(x^2 \frac{\partial^2 z}{\partial x^2} \right)^{-1} \,. \tag{33}$$

By the zero subindex we stress that this function is entirely determined by the time zero snapshot of the call price surface. For all $t > 0$, setting $\sigma(M_t, t) = \sigma_0(x, y)|_{x=M_t, y=t}$ the moneyness process (32) has the marginals as prescribed by the various time-to-maturity slices of the call price surface.

Nowadays, local volatility models are widely used in the practice of quantitative finance for pricing and hedging mildly path-dependent derivatives. This is because these models neatly embed in a portfolio context as they provide an almost perfect fit to observed option prices; at the same time option pricing is fast and efficient. Numerous methods have therefore been suggested for estimating the local volatility function; see, e.g., Jackson et al. (1998), Coleman et al. (1999), or McIntyre (2001). Given the present series representation of the call price surface, a convenient direct way of recovering local volatility is to replace the first-order time-to-expiry and second-order strike derivatives in (33) by the derivatives estimates we obtain from our single conditional estimate of the price surface. The resulting local volatility estimate could then be placed into finite difference schemes to price other options, such as barrier options or American options consistently with given European call prices; see, among others, Andersen and Brotherton-Ratcliffe (1997) and Dempster and Richards (2000).

In Figure 6 we plot the local volatility surface. It displays similar smiley patterns for the short horizons that is known for implied volatility surfaces, and it becomes generally flatter for longer

[11] See Madan and Yor (2002) for an exposition of this theme.

time horizons. It should be noted, however, that the precise local volatility shapes defy any reasonable economic interpretation, a failing for which this model is heavily criticized; see Dumas et al. (1998) and Hagan et al. (2002).

6.3 Unconditional estimates

Alternatively, one might be interested in studying an unconditional call price surface and its derivatives. We take this as an example to estimate the family of unconditional state price density that applies the unpenalized estimator. We take all available data between 2009-2010.

6.3.1 State price densities

Since the TP B-spline estimate has a bounded support, the state price density estimate does not possess tails beyond the domain of estimation. A natural estimator of the family of state price densities obtained from non-discounted relative call prices is therefore given by

$$\hat{\phi}(x,y) = \frac{1}{\widehat{K}(y)} \frac{\partial^2 \hat{z}(x,y)}{\partial x^2} = \frac{1}{\widehat{K}(y)} \sum_{j_1=0}^{q_1} \sum_{j_2=0}^{q_2} \widehat{\theta}_{j_1,j_2} \, B^{(2)}_{j_1,p_1}(x) \, B_{j_2,p_2}(y) \,, \quad (x,y) \in [\underline{x}, \overline{x}] \times [\underline{y}, \overline{y}] \tag{34}$$

where $\phi(x,y)$ is the state price density and $B^{(k)}_{j,p}$ denotes the kth order derivative of the p-degree B-spline basis function (see appendix for details on derivatives of the B-spline basis function). The normalizing function

$$\widehat{K}(y) = \sum_{j_1=0}^{q_1} \sum_{j_2=0}^{q_2} \widehat{\theta}_{j_1,j_2} \left\{ B^{(1)}_{j_1,p_1}(\overline{x}) - B^{(1)}_{j_1,p_1}(\underline{x}) \right\} B_{j_2,p_2}(y) \tag{35}$$

ensures that each density integrates to one across the strike domain. By construction of the estimator, $0 < K(y) \leq 1$. Since we model all prices relative to the forward, normalizing with $K(y)$ is equivalent to requiring that all forwards be correctly priced.

We present the estimated family of unconditional state price densities in Figure 7. Without any specific optimization, we chose as knot sequences $\boldsymbol{\upsilon} = \{0.04, 0.10, 0.15, 0.20, 0.25, 0.30, 0.40, 1.04\}$ and $\boldsymbol{\xi} = \{0.20, 0.45, 0.71, 0.85, 1.02, 1.16, 1.30, 1.53, 1.82\}$, and $\lambda = 0$. Since degrees are $p_1 =$

5, $p_2 = 3$ for the estimate, the second-order derivative is still a cubic polynomial. This gives rise to the very smooth appearance of the densities. The densities exhibit a slight asymmetry and thicker tails on the down-side, as is usually found in the literature.

7 Concluding remarks

In this work, we suggest a TP B-spline estimator for the call price surface. To ensure the estimate is arbitrage-free, we formalize sufficient no-arbitrage constraints on the control net of the TP B-spline. Since the control net is independent of the degrees of the underlying spline, we are not limited to using cubic splines as is the concurrent literature. We show consistency of the estimator, provide an upper bound of convergence, which is Stone's optimal rate, and discuss least squares implementation. In simulations, we find substantial efficiency gains from implementing surface constraints. We also study the choice of the regularization parameter. As a demonstration, we apply our procedure to S&P500 call price data.

Further refinements could be implemented. For instance, a natural extension would be to combine both puts and call data in the estimator. In this way, the put-call parity, as a linear constraint as well, could be directly imposed in the quadratic program. This might be particularly interesting when the purpose is to back out a Black-Scholes implied volatility surface from the estimate of the option price surface. We leave this for further research.

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 $\xi_j \le \xi_{j+1}$ 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 $\xi_{j-1} < \xi_j = \ldots = \xi_{j+r-1} < \xi_{j+r}$. On 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

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

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 (de Boor, 1972; Cox, 1972)

$$B_{j,p}(x) = \alpha_{j,p-1}(x) B_{j,p-1}(x) + \{1 - \alpha_{j+1,p-1}(x)\} B_{j+1,p-1}(x) , \tag{36}$$

where

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

and

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

In the presence of multiple knots, the B-splines are defined according to the same recurrence 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

$$B_{j,p}^{(1)}(x) = p\left\{\frac{B_{j,p-1}(x)}{\xi_{j+p}-\xi_j} - \frac{B_{j+1,p-1}(x)}{\xi_{j+p+1}-\xi_{j+1}}\right\} . \tag{37}$$

From this it can be shown that

$$s'(x) = p\sum_{j=1}^{q}\frac{\theta_j-\theta_{j-1}}{\xi_{j+p}-\xi_j}B_{j,p-1}(x) . \tag{38}$$

Recursive application of the identity (37) and relationship (38) gives $B_{j,p}^{(k)}(x)$, the k-th order derivative of $B_{j,p}(x)$ with respect to x, where $k \le (p-1)$, and higher order derivatives of $s(x)$.

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

$$x \to (x, s(x)) .$$

As an important fact, this curve can be interpreted as a spline curve due to the linear precision property of the B-spline stating that

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

where the sequence of points $\xi_j^* = (\xi_{j+1} + \cdots + \xi_{j+p})/p$ is known as Greville sites. Due to 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 Figure 1.

B Asymptotic analysis

B.1 General projection framework

As described in the main text, (9) and (10) have the interpretation of a projection 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

$$\mathcal{V} = \left\{ \underset{\rightarrow}{z} = \begin{pmatrix} z_1(x,y) \\ \vdots \\ z_n(x,y) \end{pmatrix} : \; z_i \in \mathcal{S}(p_1, p_2; \boldsymbol{\xi}, \boldsymbol{\upsilon}), \; i = 1, \ldots, n \right\} , \tag{39}$$

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 $\mathcal{V}^z_C \subset \mathcal{V}^z$ 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 element of $\mathcal{V}$. We make z_0 element of $\mathcal{V}$ by replacing it by 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 standard spline theory; see (13.69) in connection with Theorem 12.8 in Schumaker (2007). As for the candidate regression functions, we set $z_{0i}^*(x,y) = z_0^*(x,y)$, $i = 1, \ldots, n$, such that 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 $\mathcal{V}^z$ are demeaned, i.e., are 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 the corresponding weight vector $\theta_{j_1,j_2} = 0$ for all j_1, j_2 whenever the element of $\mathcal{V}$ is a constant. This construction will be vital for the analysis of penalized estimator.

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

$$\langle \underset{\rightarrow}{f}, \underset{\rightarrow}{g} \rangle_n = \frac{1}{n} \sum_{i=1}^n f_i(X_i, Y_i)\, g_i(X_i, Y_i) \tag{40}$$

and its induced squared norm by

$$\| \underset{\rightarrow}{f} \|_n^2 = \frac{1}{n} \sum_{i=1}^n f_i^2(X_i, Y_i) \; . \tag{41}$$

With this notation, (9) and (10) can be written as

$$\underset{\to}{\hat{z}} = \arg\min_{\underset{\to}{z} \in \mathcal{V}^z} \|\underset{\to}{Z} - \underset{\to}{z}\|_n^2 \,, \tag{42}$$

and

$$\underset{\to}{\hat{z}}^C = \arg\min_{\underset{\to}{z} \in \mathcal{V}_C^z} \|\underset{\to}{Z} - \underset{\to}{z}\|_n^2 \,, \tag{43}$$

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

$$\|\underset{\to}{Z} - \underset{\to}{z}\|_n^2 = \|\underset{\to}{Z} - \underset{\to}{\hat{z}}\|_n^2 + \|\underset{\to}{\hat{z}} - \underset{\to}{z}\|_n^2 \,.$$

This holds since $\underset{\to}{Z} - \underset{\to}{\hat{z}}$ must be orthogonal to $\underset{\to}{\hat{z}} - \underset{\to}{z}$ with respect to the inner product defined in (40), since $\underset{\to}{\hat{z}}$ is the projection of the data $\underset{\to}{Z}$ on $\mathcal{V}^z$. As a consequence of this observation, it follows (Mammen et al., 2001, Prop. 1) that the minimization in (43) can be simplified to

$$\underset{\to}{\hat{z}}^C = \arg\min_{\underset{\to}{z} \in \mathcal{V}_C^z} \|\underset{\to}{\hat{z}} - \underset{\to}{z}\|_n^2 \,. \tag{44}$$

Thus (43) 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 (41) in $\mathcal{V}$. We have

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

and from that

$$\|\underset{\to}{\hat{z}} - \underset{\to}{z_0^*}\|_n^2 - \|\underset{\to}{\hat{z}^C} - \underset{\to}{z_0^*}\|_n^2 = \|\underset{\to}{\hat{z}} - \underset{\to}{\hat{z}^C}\|_n^2 \geq 0 \,.$$

This follows from the fact that $\langle \underset{\to}{\hat{z}} - \underset{\to}{\hat{z}^C}, \underset{\to}{\hat{z}^C} - \underset{\to}{z_0^*}\rangle_n \geq 0$, since $\underset{\to}{\hat{z}^C}$ is the projection of $\underset{\to}{\hat{z}}$ 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:

$$\left(\frac{1}{n}\sum_{i=1}^{n}(\hat{z}^C(X_i,Y_i)-z_0^*(X_i,Y_i))^2\right)^{1/2} \leq \left(\frac{1}{n}\sum_{i=1}^{n}(\hat{z}(X_i,Y_i)-z_0^*(X_i,Y_i))^2\right)^{1/2} . \tag{45}$$

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

$$\left(\frac{1}{n}\sum_{i=1}^{n}(\hat{z}^C(X_i,Y_i)-z_0(X_i,Y_i))^2\right)^{1/2} \leq \left(\frac{1}{n}\sum_{i=1}^{n}(\hat{z}(X_i,Y_i)-z_0(X_i,Y_i))^2\right)^{1/2} + O_P(J_n^{-r}) . \tag{46}$$

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) due to Newey (1997). This proves the first assertion.

Analysis of the penalized estimate. For the penalized estimator we alter the inner product (40) to

$$\langle \underset{\rightarrow}{f}, \underset{\rightarrow}{g} \rangle_n = \frac{1}{n}\sum_{i=1}^{n} f_i(X_i,Y_i)\, g_i(X_i,Y_i) + \lambda_n \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\theta}_{f_i}\cdot\boldsymbol{\theta}_{g_i} \tag{47}$$

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

$$\|\underset{\rightarrow}{f}\|_n^2 = \frac{1}{n}\sum_{i=1}^{n} f_i^2(X_i,Y_i) \;+\; \lambda_n \frac{1}{n}\sum_{i=1}^{n} |\boldsymbol{\theta}_{f_i}|^2 . \tag{48}$$

With these definitions and recalling that constant elements have a zero B-spline weight vector by construction, the estimation problems (11) and (12) can be expressed as minimizations of the squared norms (48) over the spaces $\mathcal{V}^z$ and $\mathcal{V}^z_C$. The same line of arguments as for the standard estimate then leads to the conclusion that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(\hat{z}^C_\lambda(X_i,Y_i)-z_0^*(X_i,Y_i))^2 + \lambda_n|\hat{\boldsymbol{\theta}}_{z^C}-\boldsymbol{\theta}_{z_0^*}|^2 \leq \\ \frac{1}{n}\sum_{i=1}^{n}(\hat{z}_\lambda(X_i,Y_i)-z_0^*(X_i,Y_i))^2 + \lambda_n|\hat{\boldsymbol{\theta}}_{z}-\boldsymbol{\theta}_{z_0^*}|^2 .\end{aligned} \tag{49}$$

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

$$\left(\frac{1}{n}\sum_{i=1}^{n}(\hat{z}_\lambda^C(X_i,Y_i) - z_0(X_i,Y_i))^2\right)^{1/2} \leq \left(\frac{1}{n}\sum_{i=1}^{n}(\hat{z}_\lambda(X_i,Y_i) - z_0(X_i,Y_i))^2\right)^{1/2} + O_P\left((J_n/n + J_n^{-r})^{1/2}\right) , \tag{50}$$

whereupon one concludes as above.

C Monotonicity constraints on control net

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

Proof of Lemma 4.3: By differentiating (20) with respect to x, we get

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

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 is

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

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

$$\frac{\partial z}{\partial x} = \sum_{j_2=0}^{q_2}\theta_{j_1,j_2}B_{j_2,p_2}(y)\sum_{j_1=0}^{q_1}\frac{\partial}{\partial x}B_{j_1,p_1}(x) \tag{51}$$

$$= \sum_{j_2=0}^{q_2}B_{j_2,p_2}(y)\sum_{j_1=1}^{q_1}(\theta_{j_1,j_2} - \theta_{j_1-1,j_2})\frac{p_1}{\xi_{i+p_1} - \xi_i}B_{j_1,p_1-1}(x) \tag{52}$$

$$= \sum_{j_2=0}^{q_2}B_{j_2,p_2}(y)\sum_{j_1=1}^{q_1}\left(\frac{\theta_{j_1,j_2} - \theta_{j_1-1,j_2}}{\xi_{j_1}^* - \xi_{j_1-1}^*}\right)B_{j_1,p_1-1}(x) \tag{53}$$

$$= \sum_{j_2=0}^{q_2}B_{j_2,p_2}(y)\sum_{j_1=1}^{q_1}\Delta_1\theta_{j_1-1,j_2}B_{j_1,p_1-1}(x) , \tag{54}$$

where the second equality follows from (38) and the third from noting that

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

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

Applying now the recurrence formula (36) to $B_{j_2,p_2}(y)$ and inserting into (54) yields

$$\frac{\partial z}{\partial x} = \sum_{j_1=1}^{q_1}\sum_{j_2=1}^{q_2} B_{j_1,p_1-1}(x)B_{j_2,p_2-1}(y)\sum_{\ell=0}^{1}\beta_{j_2,\ell}(y)\Delta_1\theta_{j_1-1,j_2-\ell} \ . \tag{55}$$

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 gets

$$\frac{\partial z}{\partial y} = \sum_{j_1=1}^{q_1}\sum_{j_2=1}^{q_2} B_{j_1,p_1-1}(x)B_{j_2,p_2-1}(y)\sum_{k=0}^{1}\alpha_{j_1,k}(x)\Delta_2\theta_{j_1-k,j_2-1} \ , \tag{56}$$

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

$$D_d\, z = \sum_{j_1=1}^{q_1}\sum_{j_2=1}^{q_2}\sum_{k=0}^{1}\sum_{\ell=0}^{1} B_{j_1,p_1-1}(x)B_{j_2,p_2-1}(y)\alpha_{j_1,k}(x)\beta_{j_2,\ell}(y)\left(d_1\Delta_1\theta_{j_1-1,j_2-\ell} + d_2\Delta_2\theta_{j_1-k,j_2-1}\right) \tag{57}$$

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

References

Abadir, K. and Rockinger, M. (2003). Density functionals, with an option-pricing application, *Econometric Theory* **19**: 778–811.

Agarwal, G. G. and Studden, W. J. (1980). Asymptotic integrated mean square error using least squares and bias minimizing splines, *Annals of Statistics* **8**(6): 1307–1325.

Aït-Sahalia, Y. and Duarte, J. (2003). Nonparametric option pricing under shape restrictions, *Journal of Econometrics* **116**: 9–47.

Aït-Sahalia, Y. and Lo, A. (1998). Nonparametric estimation of state-price densities implicit in financial asset prices, *Journal of Finance* **53**: 499–548.

Aït-Sahalia, Y. and Lo, A. (2000). Nonparametric risk management and implied risk-aversion, *Journal of Econometrics* **94**: 9–51.

Alexander, C. and Nogueira, L. M. (2007). Model-free hedge ratios and scale-invariant models, *Journal of Banking and Finance* **31**(6): 1839–1861.

Andersen, L. B. G. and Brotherton-Ratcliffe, R. (1997). The equity option volatility smile: An implicit finite-difference approach, *Journal of Computational Finance* **1**(2): 5–37.

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

Audrino, F. and Colangelo, D. (2010). Semi-parametric forecasts of the implied volatility surface using regression trees, *Statistics & Computing* **20**(4): 421–434.

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

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

Bates, D. S. (2005). Hedging the smirk, *Finance Research Letters* **2**(4): 195–200.

Benko, M., Fengler, M. R., Härdle, W. and Kopa, M. (2007). On extracting information implied in options, *Computational Statistics* **22**(4): 543–553.

Birke, M. and Pilz, K. F. (2009). Nonparametric option pricing with no-arbitrage constraints, *Journal of Financial Econometrics* **7**(2): 53–76.

Bliss, R. and Panigirtzoglou, N. (2004). Option-implied risk aversion estimates, *Journal of Finance* **59**(1): 407–446.

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

Brunck, H. D. (1970). Estimation of isotonic regression, *in* M. L. Puri (ed.), *Nonparametric Techniques in Statistical Inference*, Cambridge University Press, pp. 177–195.

Carr, P. and Madan, D. B. (2005). A note on sufficient conditions for no arbitrage, *Finance Research Letters* **2**: 125–130.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models, *in* J. J. Heckman and E. E. Leamer (eds), *Handbook of Econometrics*, Vol. 6B, Elsevier, chapter 76, pp. 5549–5632.

Coleman, T. F., Li, Y. and Verma, A. (1999). Reconstructing the unknown local volatility function, *Journal of Computational Finance* **2**(3): 77–102.

Constantinides, G. M., Jackwerth, J. C. and Savov, A. (2011). The puzzle of index option returns, *Technical report*, Konstanz University.

Cont, R. and da Fonseca, J. (2002). The dynamics of implied volatility surfaces, *Quantitative Finance* **2**(1): 45–60.

Cox, M. G. (1972). The numerical evaluation of B-splines, *Journal of the Institute of Mathematics and its Applications* **10**: 134–149.

Dalang, R. C., Morton, A. and Willinger, W. (1990). Equivalent martingale measures and no-arbitrage in stochastic securities market models, *Stochastics* **29**: 185–201.

Davis, M. and Hobson, D. G. (2007). The range of traded options, *Mathematical Finance* **17**(1): 1–14.

de Boor, C. (1972). On calculating with B-splines, *Journal of Approximation Theory* **6**: 50–62.

de Boor, C. (2001). *A practical guide to splines*, revised edn, Springer-Verlag, New York.

Dempster, M. A. H. and Richards, D. G. (2000). Pricing American options fitting the smile, *Mathematical Finance* **10**(2): 157–177.

Derman, E. and Kani, I. (1998). Stochastic implied trees: Arbitrage pricing with stochastic term and strike structure of volatility, *International Journal of Theoretical and Applied Finance* **1**(1): 61–110.

Dierckx, P. (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, *Journal of Business and Economic Statistics* **17**(4): 444–455.

Dumas, B., Fleming, J. and Whaley, R. E. (1998). Implied volatility functions: Empirical tests, *Journal of Finance* **53**(6): 2059–2106.

Dupire, B. (1994). Pricing with a smile, *RISK* **7**(1): 18–20.

Fan, J. and Mancini, L. (2009). Option pricing with model-guided nonparametric methods, *Working paper*.

Fengler, M. R. (2009). Arbitrage-free smoothing of the implied volatility surface, *Quantitative Finance* **9**(4): 417–428.

Fengler, M. R., Härdle, W. and Mammen, E. (2007). A semiparametric factor model for implied volatility surface dynamics, *Journal of Financial Econometrics* **5**(2): 189–218.

Floater, M. S. and Peña, J. (1998). Tensor-product monotonicity preservation, *Advances in Computational Mathematics* **9**: 353–362.

Gallant, A. R. (1982). Unbiased determination of production technologies, *Journal of Econometrics* **20**(2): 285–323.

Garcia, R., Ghysels, E. and Renault, E. (2010). The econometrics of option pricing, *in* Y. Aït-Sahalia and L. Hansen (eds), *Handbook of Financial Econometrics*, Vol. 1, North Holland, Amsterdam, pp. 479–552.

Garcia, R. and Renault, E. (1998a). A note on hedging in ARCH and stochastic volatility option pricing models, *Mathematical Finance* **8**(2): 153–161.

Garcia, R. and Renault, E. (1998b). Risk aversion, intertemporal substitution and option pricing, *Working Paper 98s-02*, CIRANO.

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, España.

Giacomini, E., Härdle, W. and Krätschmer (2009). Dynamic semiparametric factor models in risk neutral density estimation, *AStA Advances in Statistical Analysis* **93**(4): 387–402.

Glaser, J. and Heider, p. (2010). Arbitrage-free approximation of call price surfaces and input data risk, *Quantitative Finance* . Forthcoming.

Goldfarb, D. and Idnani, A. (1983). A numerically stable dual method for solving strictly convex quadratic programs, *Mathematical Programming* **27**(1): 1–33.

Gope, P. and Fries, C. P. (2011). Volatility surface interpolation on probability space using normed call prices, *Technical report*. Available at SSRN: http://ssrn.com/abstract=1964634.

Gouriéroux, C., Monfort, A. and Tenreiro, C. (1995). Kernel M-estimators and functional residual plots, *Document de travail 9546*, CREST, Paris.

Grith, M., Härdle, W. K. and Park, J. (2013). Shape invariant modelling of pricing kernels and risk aversion, *Journal of Financial Econometrics* **11**(2): 370–399.

Groeneboom, P., Jongbloed, G. and Wellner, J. A. (2001a). A canonical process for estimation of convex functions: the "invelope" of integrated Brownian motion + t^4, *Annals of Statistics* **29**(6): 1620–1652.

Groeneboom, P., Jongbloed, G. and Wellner, J. A. (2001b). Estimation of a convex function: characterizations and asymptotic theory, *Annals of Statistics* **29**(6): 1653–1698.

Hagan, P., Kumar, D., Lesniewski, A. and Woodward, D. (2002). Managing smile risk, *Wilmott Magazine* **1**: 84–108.

Hansen, P. C. and O'Leary, D. P. (1993). The use of the L-curve in the regularization of discrete ill-posed problems, *SIAM Journal on Scientific Computing* **14**(6): 1487–1503.

Hanson, D. L. and Pledger, G. (1976). Consistency in concave regression, *Annals of Statistics* **4**: 1038–1050.

Härdle, W. and Hlávka, Z. (2009). Dynamics of state price densities, *Journal of Econometrics* **150**(1): 1–15.

Hastie, T. and Tibshirani, R. (1990). *Generalized additive models*, Chapman and Hall, London.

Huang, J. Z. (2003). Local asymptotics for polynomial spline regression, *Annals of Statistics* **31**(5): 1600–1635.

Hurvich, C., Simonoff, J. and Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion, *Journal of the Royal Statistical Society B* **60**(2): 271–293.

Jackson, N., Süli, E. and Howison, S. (1998). Computation of deterministic volatility surfaces, *Journal of Computational Finance* **2**(2): 5–32.

Jackwerth, J. C. (2000). Recovering risk-aversion from option prices and realized returns, *Review of Financial Studies* **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. and Rubinstein, M. (1996). Recovering probability distributions from contemporaneous security prices, *Journal of Finance* **51**: 1611–1631.

Jarrow, R. and Rudd, A. (1982). Approximate valuation for arbitrary stochastic processes, *Journal of Financial Economics* **10**(3): 347–369.

Jondeau, E. and Rockinger, M. (2001). Gram-Charlier densities, *Journal of Economic Dynamics & Control* **25**: 1457–1483.

Kellerer, H. G. (1972). Markov-Komposition und eine Anwendung auf Martingale, *Mathematische Annalen* **198**: 99–122.

Laurini, M. P. (2011). Imposing no-arbitrage conditions in implied volatilities using constrained smoothing splines, *Applied Stochastic Models in Business and Industry* . Forthcoming.

Li, Q. and Racine, S. (2007). *Nonparametric Econometrics*, Princeton University Press, Princeton, New Jersey.

Madan, D. B. and Milne, F. (1994). Contingent claims valued and hedged by pricing and investing in a basis, *Mathematical Finance* **4**(3): 223–245.

Madan, D. B. and Yor, M. (2002). Making Markov martingales meet marginals: with explicit constructions, *Bernoulli* **8**(4): 509–536.

Mammen, E. (1991a). Estimating a smooth monotone regression function, *The Annals of Statistics* **19**(2): 724–740.

Mammen, E. (1991b). Nonparametric regression under qualitative smoothness assumptions, *Annals of Statistics* **19**(2): 741–759.

Mammen, E., Marron, J. S., Turlach, B. A. and Wand, M. P. (2001). A general framework for constrained smoothing, *Statistical Science* **16**(3): 232–248.

Mammen, E. and Thomas-Agnan, C. (1999). Smoothing splines and shape restrictions, *Scandinavian Journal of Statistics* **26**: 239–252.

McIntyre, M. L. (2001). Performance of Dupire's implied diffusion approach under sparse and incomplete data, *Journal of Computational Finance* **4**(4): 33–84.

Merton, R. C. (1973). Theory of rational option pricing, *Bell Journal of Economics and Management Science* **4**(Spring): 141–183.

Monteiro, A. M., Tütüncü, R. H. and Vicente, L. N. (2008). Recovering risk neutral probability density functions from options prices using cubic splines and ensuring nonnegativity, *European Journal of Operational Research* **187**: 525–542.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators, *Journal of Econometrics* **79**(1): 147–168.

Prautzsch, H., Boehm, W. and Paluszny, M. (2002). *Bézier and B-spline techniques*, Springer-Verlag, Berlin, Heidelberg.

Reiner, E. (2000). Calendar spreads, characteristic functions, and variance interpolation. Mimeo.

Renault, E. (1997). Econometric models of option pricing errors, *in* D. M. Kreps and K. F. Wallis (eds), *Advances in Economics and Econometrics, Seventh World Congress*, Econometric Society Monographs, Cambridge University Press, pp. 223–278.

Renault, E. and Touzi, N. (1996). Option hedging and implied volatilities in a stochastic volatility model, *Mathematical Finance* **6**(3): 279–302.

Roper, M. (2010). Arbitrage free implied volatility surfaces, *Working paper*, School of Mathematics and Statistics, The University of Sydney, NSW 2006, Australia.

Schumaker, L. (2007). *Spline Functions: Basic Theory*, 3rd edn, Cambridge University Press, Cambridge, UK.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression, *Annals of Statistics* **10**(4): 1040–1053.

Stone, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation, *Annals of Statistics* **22**(1): 118–184.

Wang, Y., Yin., H. and Qi, L. (2004). No-arbitrage interpolation of the option price function and its reformulation, *Journal of Optimization Theory and Applications* **120**(3): 627–649.

Wright, F. T. (1981). The asymptotic behavior of monotone regression estimates, *Annals of Statistics* **9**: 443–448.

Yatchew, A. and Härdle, W. (2006). Dynamic state price density estimation using constrained least squares and the bootstrap, *Journal of Econometrics* **133**(2): 579–599.

Yuan, M. (2009). State price density estimation via nonparametric mixtures, *The Annals of Applied Statistics* **3**(3): 963–984.

Zhou, S. and Shen, X. (2001). Spatially adaptive regression splines and accurate knot selection schemes, *Journal of the American Statistical Association* **96**(453): 247–259.

Zhou, S., Shen, X. and Wolfe, D. A. (1998). Local asymptotics for regression splines and confidence regions, *Annals of Statistics* **26**(5): 1760–1782.

List of tables and figures

	p_2	Univariate fit ('slice-by-slice')					Bivariate fit without calendar constraints				
Forward moneyness		0.6	0.8	1	1.2	1.4	0.6	0.8	1	1.2	1.4
						$p_1 = 3$					
rel. MSE	1	1.522	1.119	1.013	1.140	1.495	1.522	1.119	1.013	1.140	1.495
rel. Var.	1	1.331	1.117	1.013	1.139	1.323	1.331	1.117	1.013	1.139	1.323
rel. Bias^2($\times 10^{-2}$)	1	19.056	0.263	0.005	0.116	17.224	19.056	0.263	0.005	0.116	17.224
						$p_1 = 3$					
rel. MSE	2	1.655	1.523	1.201	1.464	1.602	1.677	1.418	1.123	1.364	1.339
rel. Var.	2	1.447	1.520	1.201	1.463	1.417	1.438	1.412	1.123	1.364	1.201
rel. Bias^2($\times 10^{-2}$)	2	20.721	0.358	0.006	0.150	18.451	23.863	0.617	0.011	0.004	13.789
						$p_1 = 3$					
rel. MSE	3	1.802	1.467	1.268	1.441	1.810	1.705	1.282	1.147	1.285	1.490
rel. Var.	3	1.576	1.464	1.268	1.440	1.602	1.487	1.278	1.147	1.283	1.302
rel. Bias^2($\times 10^{-2}$)	3	22.567	0.345	0.007	0.147	20.856	21.795	0.334	0.039	0.107	18.792
						$p_1 = 4$					
rel. MSE	1	1.505	1.117	1.012	1.150	1.511	1.505	1.117	1.012	1.150	1.511
rel. Var.	1	1.319	1.117	1.012	1.149	1.341	1.319	1.117	1.012	1.149	1.341
rel. Bias^2($\times 10^{-2}$)	1	18.633	0.027	0.013	0.129	16.994	18.633	0.027	0.013	0.129	16.994
						$p_1 = 4$					
rel. MSE	2	1.639	1.523	1.191	1.500	1.636	1.672	1.414	1.129	1.397	1.359
rel. Var.	2	1.436	1.523	1.191	1.498	1.452	1.434	1.412	1.129	1.397	1.224
rel. Bias^2($\times 10^{-2}$)	2	20.288	0.037	0.016	0.169	18.395	23.850	0.191	0.034	0.007	13.499
						$p_1 = 4$					
rel. MSE	3	1.774	1.471	1.244	1.450	1.829	1.690	1.292	1.136	1.282	1.490
rel. Var.	3	1.554	1.471	1.244	1.448	1.623	1.473	1.291	1.135	1.280	1.307
rel. Bias^2($\times 10^{-2}$)	3	21.957	0.035	0.016	0.163	20.567	21.723	0.084	0.056	0.127	18.343
						$p_1 = 5$					
rel. MSE	1	1.495	1.117	1.015	1.142	1.492	1.495	1.117	1.015	1.142	1.492
rel. Var.	1	1.310	1.117	1.015	1.139	1.328	1.310	1.117	1.015	1.139	1.328
rel. Bias^2($\times 10^{-2}$)	1	18.541	0.053	0.019	0.268	16.417	18.541	0.053	0.019	0.268	16.417
						$p_1 = 5$					
rel. MSE	2	1.608	1.471	1.195	1.478	1.605	1.656	1.369	1.126	1.367	1.343
rel. Var.	2	1.409	1.470	1.195	1.474	1.429	1.417	1.369	1.125	1.366	1.211
rel. Bias^2($\times 10^{-2}$)	2	19.944	0.070	0.022	0.347	17.666	23.824	0.001	0.028	0.064	13.140
						$p_1 = 5$					
rel. MSE	3	1.743	1.449	1.255	1.449	1.805	1.673	1.278	1.144	1.275	1.491
rel. Var.	3	1.527	1.449	1.254	1.445	1.606	1.456	1.278	1.144	1.273	1.308
rel. Bias^2($\times 10^{-2}$)	3	21.618	0.069	0.023	0.340	19.859	21.718	0.013	0.068	0.236	18.228

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 6 months. 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 5,000 simulation runs.

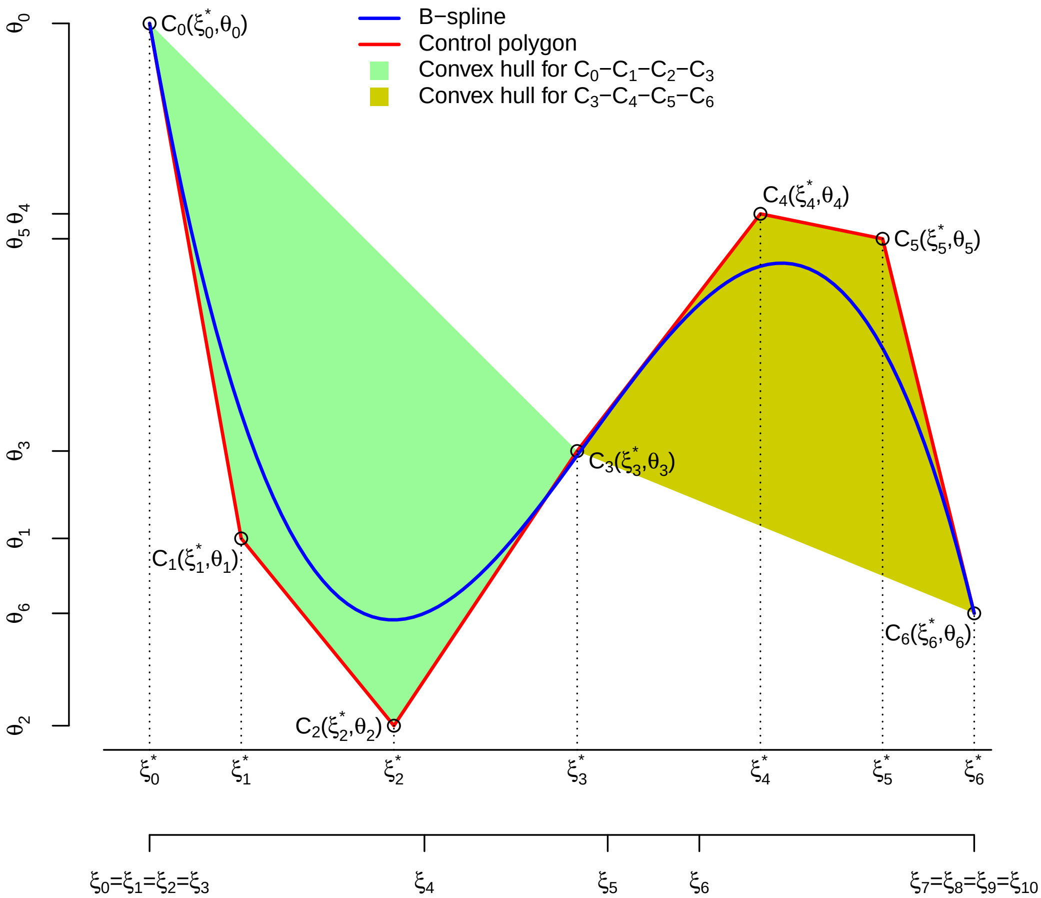

Figure 1: Illustration of the univariate B-spline (blue line), its knot sequence $\xi_0 = \xi_1 = \xi_2 = \xi_3 < \xi_4, \ldots, \xi_6 < \xi_7 = \xi_8 = \xi_9 = \xi_{10}$ (second horizontal axis), the associated control polygon (broken red line) and Greville sites $\xi_0^*, \ldots, \xi_6^*$. As examples 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 high-lighted).

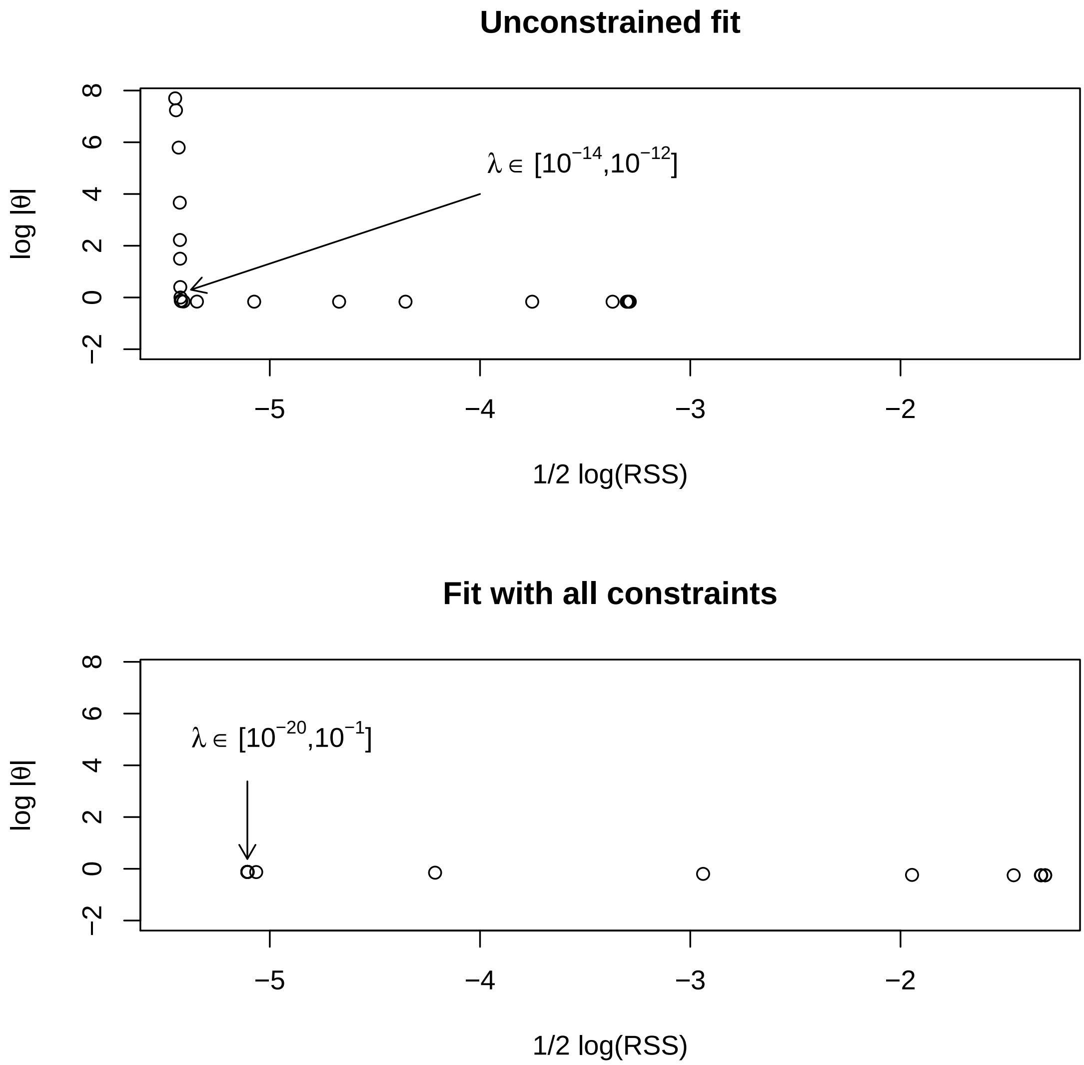

Figure 2: L-curve of log-penalty against log residual sum of squares (RSS). Penalty weight factor λ ranges in $[10^{-20}, 10^{5}]$. Top panel shows the unconstrained fit. Bottom panel the fit imposing all no-arbitrage constraints.

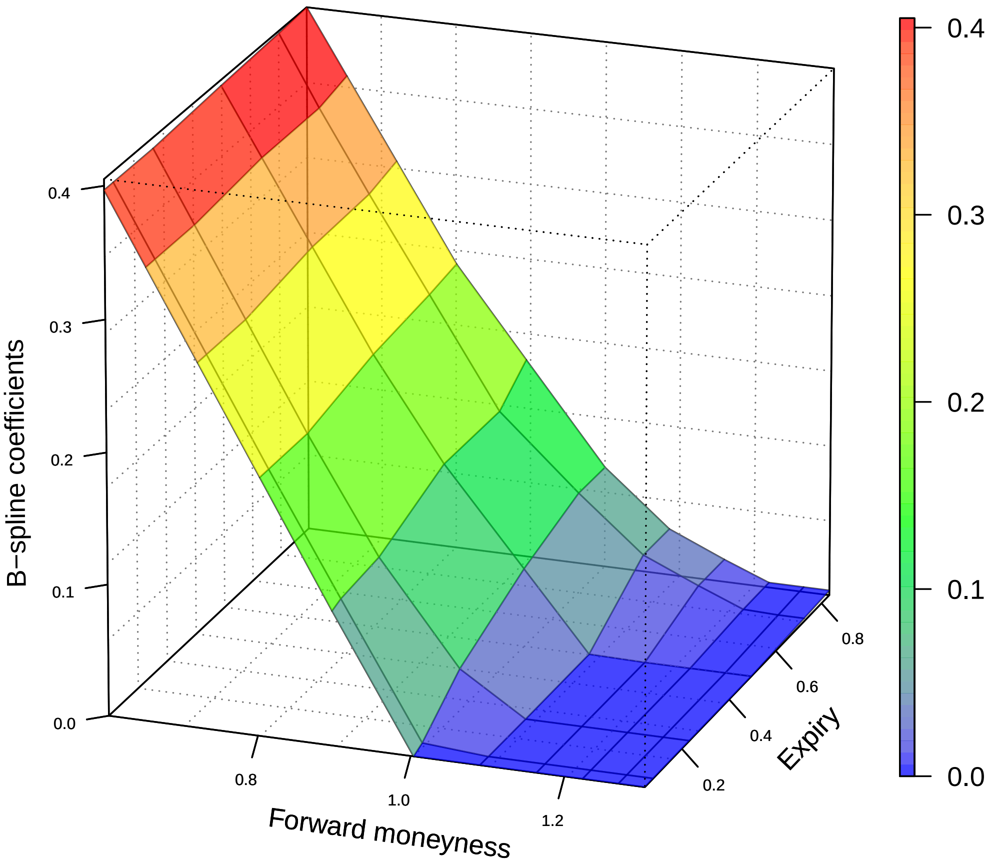

Figure 3: Control net of the estimates of Figure 4 and 5. The control represents the sufficient no-arbitrage conditions on the TP B-spline.

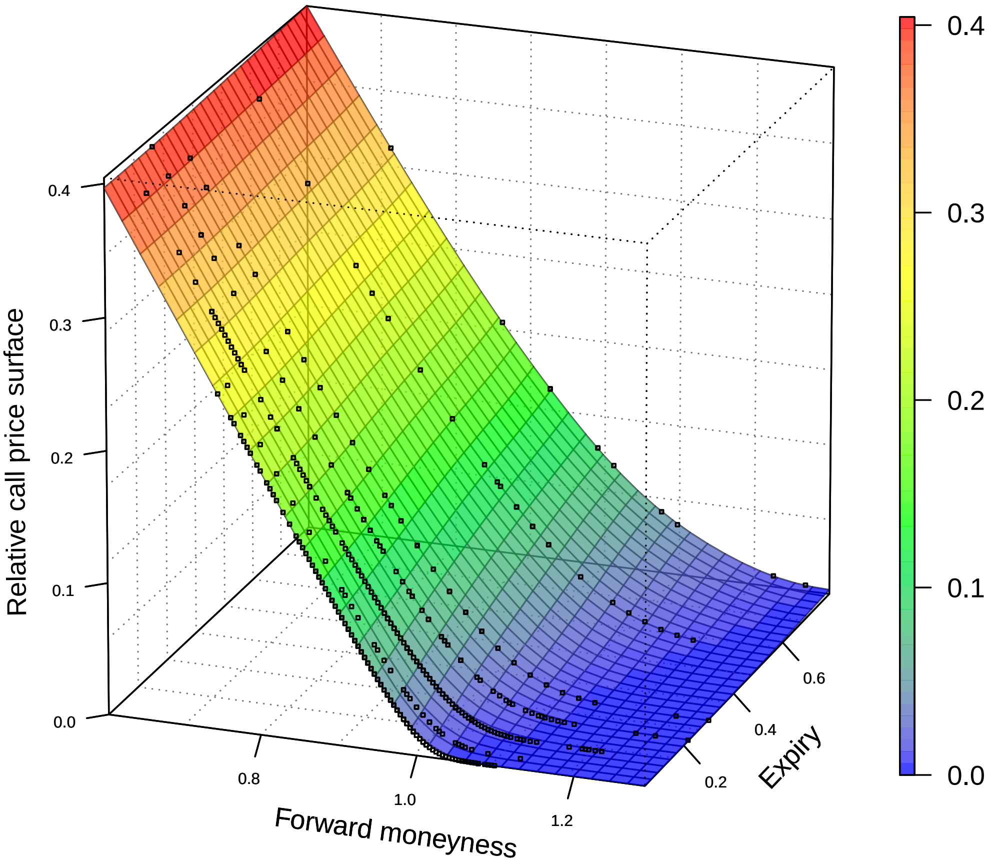

Figure 4: Arbitrage-free estimate of the transformed call price surface. TP B-spline is a five degree polynomial in moneyness and cubic polynomial in time to expiry direction, i.e. $p_1 = 5, p_2 = 3$. Surface is estimated from S&P 500 option call price data as of December 1, 2010, shown as black dots.

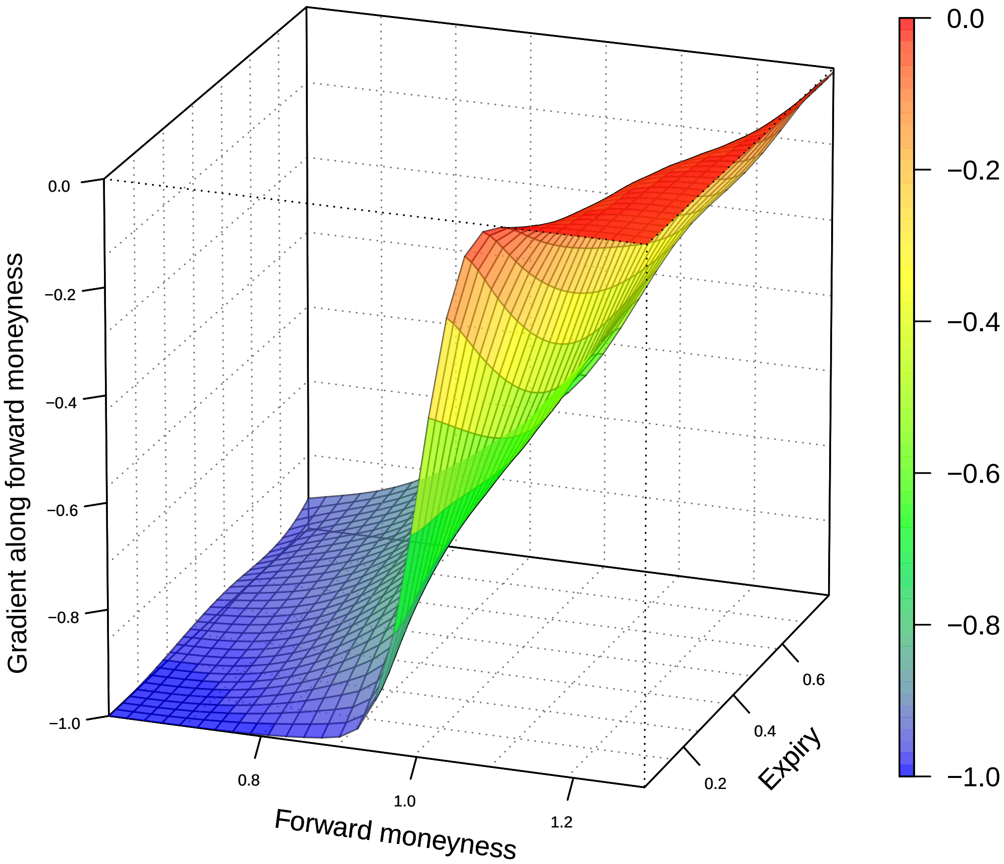

Figure 5: First-order strike derivative of the estimate in Figure 4.

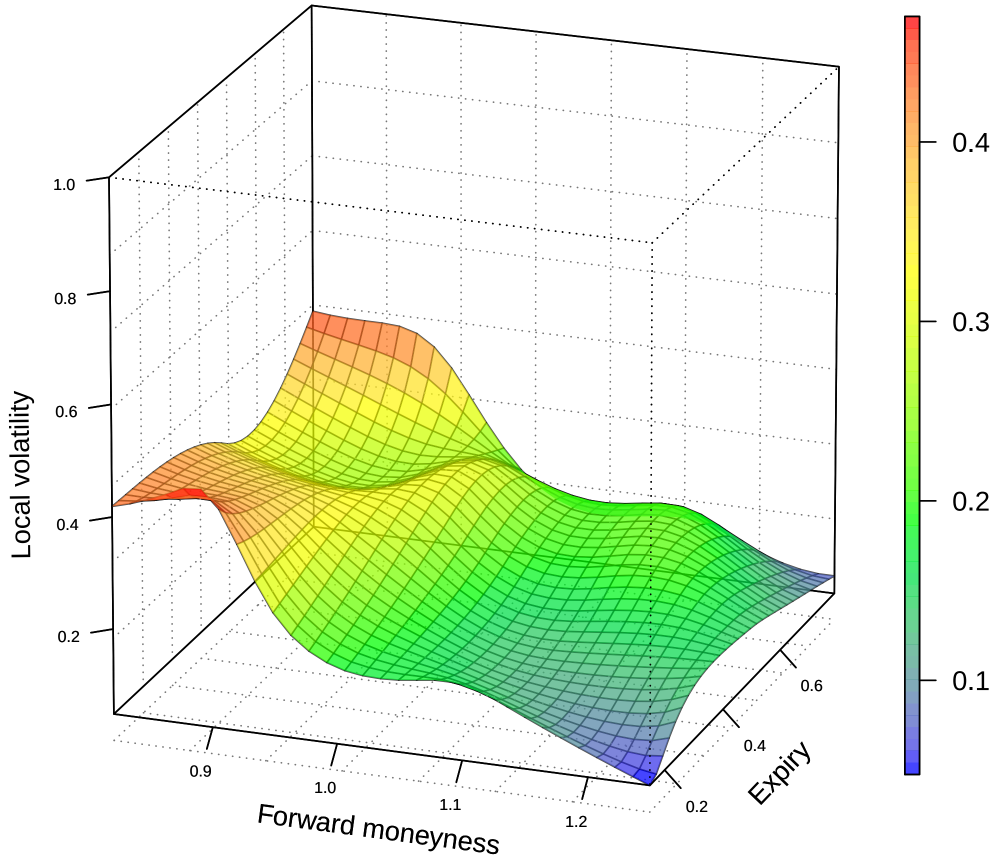

Figure 6: Local volatility surface obtained according to (33) from the estimated call price surface and its derivatives.

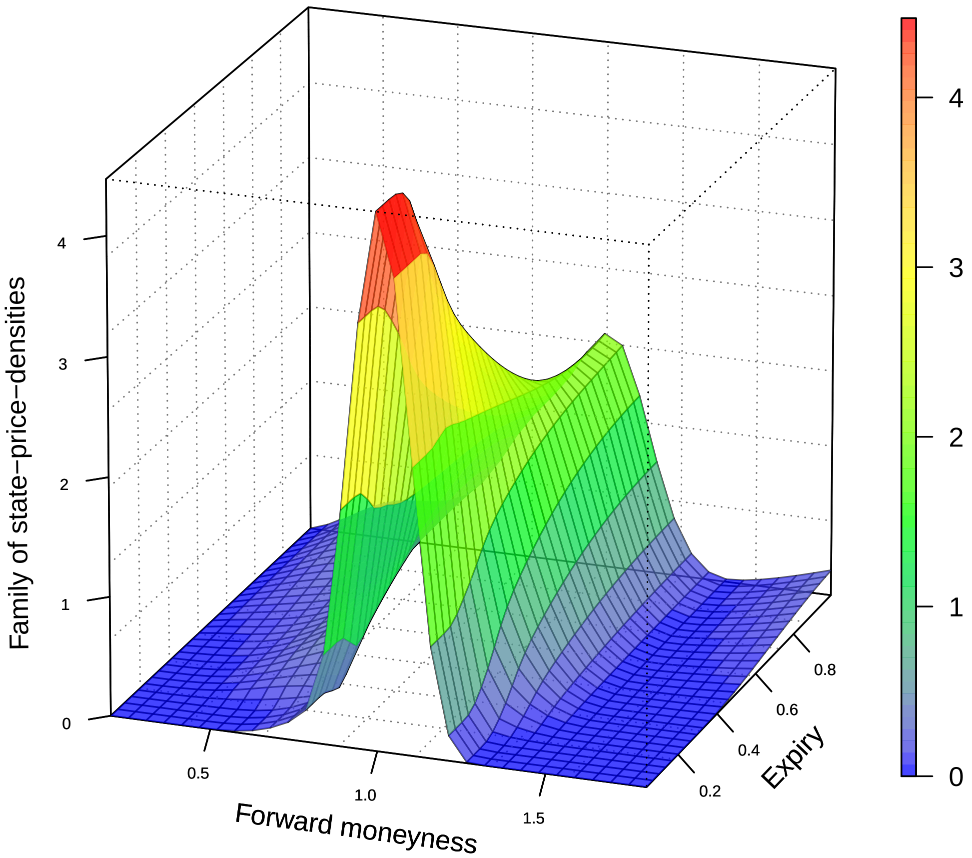

Figure 7: Family of unconditional state price densities obtained as second-order strike derivative of the estimate in Figure 4 and normed by (35). S&P 500 option data from 2009-2010.