Rev Deriv Res (2007) 10:227–267
DOI 10.1007/s11147-008-9021-2

Discount curve construction with tension splines

Leif Andersen

Published online: 7 June 2008

Abstract Polynomial splines are popular in the estimation of discount bond term structures, but suffer from well-documented problems with spurious inflection points, excessive convexity, and lack of locality in the effects of input price perturbations. In this paper, we address these issues through the use of shape-preserving splines from the class of generalized tension splines. Our primary focus is on the classical hyperbolic tension spline which we derive non-parametrically from a penalized least squares criterion, but extensions to generalized tension splines—such as rational splines and exponential splines—are also covered. Our methodology allows both for best-fitting of noisy bonds and for the construction of an exact interpolatory term structure to a set of liquid instruments. We work with a local tension B-spline basis and support both fully non-parametric and user-imposed knot location strategies.

Keywords Discount curves · Hyperbolic tension splines · Bond pricing · Swap pricing · Perturbation locality · Optimality

1 Introduction

A key concept in pricing and risk-management of fixed income securities is the risk-free term structure of interest rates, represented, for instance, as a curve of zero-coupon bond prices (the so-called *discount curve*) for a continuum of maturities in some interval. As only a finite set of fixed income securities trade, very few of which are zero-coupon bonds, in practice a computational procedure is required to interpolate between adjacent maturities of observable securities, and to extract zero-coupon bond prices from more complicated securities such as coupon bonds, swaps, and Eurodollar

L. Andersen (✉)
Banc of America Securities, 1 Bryant Park, New York, NY 10036, USA
e-mail: leif.andersen@bofasecurities.com

futures. Imperfections in some markets will introduce considerable noise in observable security prices, in which case the computational procedure must additionally be capable of smoothing and regularizing data.

One approach to the construction of a term structure of interest rates involves the specification of a smooth functional form with parameters to be determined by nonlinear least-squares regression. Examples include Nelson and Siegel (1987); Diament (1993), and Svensson (1994). While a functional form can offer significant insight in theoretical work, the resulting fit to observed security prices is typically too loose for mark-to-market purposes and may also, as in the case of the polynomial model in Chambers et al. (1984), result in highly unstable term structure estimates. As a consequence, financial institutions involved in actual trading of fixed income securities rarely, if ever, rely on functional forms.

The most common method of estimating the term structure of interest rates is based on polynomial splines. Application of linear splines on term yields[1] is particularly popular and, in the pure interpolation case, gives rise to an efficient iterative approach known as *bootstrapping*, covered in many introductory textbooks (see e.g. Hull 2000). An often-seen variation uses piecewise flat forward rates, corresponding to a linear spline in the logarithm of zero-coupon bond prices; see e.g. the survey article by Hagan and West (2004) for a discussion. Both methods are stable and fast, but ultimately produce forward rate term structures that are discontinuous. This curve shape is at odds with economic reality and gives rise to technical problems in many dynamic models of the term structure of interest rates.

To produce continuous term structures of forward rates, a number of approaches based on cubic spline interpolation have been proposed in the literature, starting with the cubic regression splines proposed in McCulloch (1975). While McCulloch (1975) applied cubic splines directly to the zero-coupon bond price curve, Sheah (1984) and many others have noted that this leads to instabilities in yields and forward rates. Consequently, cubic splines are most often applied either on term yields, the logarithm of zero-coupon bond prices, or a similar transformation; see Tanggaard (1997); Hagan and West (2004), and McCulloch and Kochin (2000) for a discussion. In the actual application of cubic splines, a variety of algorithms and spline choices have been proposed in the literature, ranging from the C^1 Hermite splines (such as the Catmull-Rom spline, see Catmull and Rom 1974) to C^2 splines with boundary conditions on first, second, or third derivatives. Of particular relevance to our work here is the non-parametric approach of Tanggaard (1997) which optimizes a least-squares pricing norm penalized with a smoothness term to arrive at a natural cubic spline with knot placement determined by the data. Hagan and West (2004) surveys a number of other cubic spline algorithms, and also review approaches based on quartic splines (see e.g. Adams (2001) or Adams and van Deventer (1994)). Not surprisingly, quartic splines appear to offer no obvious advantages over cubic splines.

Despite considerable popularity in financial institutions and in software packages, term structure estimation with cubic splines inherits a number of well-known

[1] The concepts of term yields, forward rates, and similar quantities used to characterize the term structure of interest rates are defined in Sect. 2.

problems of cubic splines in general. For instance, cubic splines cannot be guaranteed to preserve any convexity or monotonicity properties that may characterize the original data, and occasionally introduce excess convexity and spurious inflection points, giving rise to curves with a "wiggly" appearance. Apart from lack of realism, an overly wiggly curve is likely to cause violation of economic constraints inherent in the data, such as positivity of forwards and yields. C^2 cubic splines[2] also have an inherent lack of locality, in the sense that a local perturbation of curve input data will cause "ringing" and modify sections of the discount curve far away from the perturbed data point. As common practice for risk management and hedging of a fixed income portfolio relies heavily on such perturbation of curve inputs, lack of locality in the discount curve construction methodology can lead to misleading hedge information where, say, the hedge of a short-term security is reported to contain a significant position in a long-dated security.

Systematic efforts to overcome the drawbacks of cubic spline interpolation in term structure estimation are of considerable interest to the banking industry, but have so far been fairly limited. Hagan and West (2004) discuss applications of the approach in Hyman (1983) to preserve convexity of the input data. These authors also propose an algorithm based on a quadratic spline for discrete forward rates; when combined with a series of modifications, this approach can be guaranteed to produce a strictly positive, but non-differentiable, forward curve.

The approach in this article is similar to that of Tanggaard (1997), but we modify the smoothness penalty term to include a component that measures curve length. The curve that optimizes this modified norm is proven to be a *hyperbolic tension spline*. Introduced by Schweikert (1966), the hyperbolic tension spline can be considered the result of adding a pulling force (tension) to each end point of a cubic spline; as the force is increased, excess convexity and extraneous inflection points are gradually reduced until the curve eventually approaches a linear spline. Basic algorithms for construction and parameter selection of purely interpolating hyperbolic tension splines can be found in Cline (1974); Renka (1987), and Rentrop (1980), among others. For the purpose of term structure estimation (where discount curve function values are initially unknown) algorithms for interpolating splines are generally less natural than those based on a formal expansion of tension splines in local basis functions. Such work was undertaken in Koch and Lyche (1989, 1993), and was later generalized in Kvasov (2000). We rely on a local base representation throughout.

While hyperbolic tension splines are appealing due to their characterization as the optimal solution to a fairly natural regularization norm, a number of other families of tension splines have been proposed in the literature. Common for all tension spline is the existence of one or more control parameters that allow for a gradual (and uniformly convergent) transition from a cubic spline to a linear spline. It is this property of tension splines that ultimately allows us to smoothly manipulate locality and shape preservation, and thereby to overcome the problems of cubic spline interpolation. Although our primary focus is on hyperbolic tension splines, we present elements of GB-spline theory (Kvasov 2000) and describe one simple approach to generalize our

[2] C^1 cubic Hermite splines are local, with one point on the discount curve being linked only to a few neighboring ones.

work to generalized tension splines. A further extension of our basic approach allows for externally specified spline knots, allowing for additional control of the resulting curve.

Application of hyperbolic tension splines to term structure estimation was first proposed in a note by Barzanti and Corradi (1998), where a classically formulated hyperbolic tension spline was used in a least-squares fit to a given set of simple zero-coupon bonds, with knot-placement based on a rule proposed by Sheah (1984). Our approach is different in many respects, starting with our basic formulation of the fitting problem, where we properly work with coupon-bearing instruments (zero-coupon bonds are rarely traded and then only in short maturities) and operate in a space of curves more general, and more suitable, than that of the discount function itself. Our basic estimator is not of the least-squares type, but is the result of non-parametric procedure allowing us to systematically balance price accuracy against curve regularity properties. In this approach, knot placement is entirely data-driven, although we, as discussed, also supply useful extensions of our algorithm that facilitates external knot placement. In the numerical implementation, our use of a local basis improves computational efficiency and allows us to easily generalize the algorithm to the broader class of generalized tension splines with a GB-spline basis. In their brief numerical results, Barzanti and Corradi (1998) are primarily interested in the ability of tension splines to repair artifacts associated with the choice of building a least-squares spline directly on the discount function; in contrast, our formulation of the problem essentially circumvents this issue and we can focus our numerical tests on matters of more importance to practitioners, such as the construction of interpolating Libor yield curves with desired locality and smoothness properties.

The rest of the paper is organized as follows. In Sect. 2, we briefly outline the basic problem of discount curve estimation, and provide a discussion of issues involved in the formulation of a suitable mathematical model. Section 3 provides some elementary results for cubic splines and hyperbolic tension splines in an interpolatory setting. In particular, we introduce the local B-spline basis for hyperbolic tension splines and discuss methods to incorporate boundary conditions. In Sect. 4, we state the problem of term structure estimation as the solution to a particular minimization criterion, the solution of which is a hyperbolic tension spline. An efficient numerical procedure is presented, along with certain results for properly establishing the trade-off between curve regularity and precision of market data fit. Section 5 lists several extensions to the basic algorithm in Sect. 4, and provides support for the use of user-specified spline knots, non-uniform curve tension, and generalized (GB) tension splines. Selected numerical results are shown in Sects. 6 and 7 contains our conclusion. Three appendices contain proofs, a discussion of forward curve overlays to produce "turn" effects, and additional numerical results.

2 Financial markets basics

2.1 The discount curve construction problem

Consider a riskless claim (a *zero-coupon bond*) paying \$1 at time t; let the time 0 value of this claim be denoted $P(t)$. Interpreted as a function of t, the continuous mapping $P : \mathbb{R}_+ \to [0, 1]$ is known as the *discount curve*. In the absence of arbitrage, $P(t)$ is never-increasing in t, starting at $P(0) = 1$. The discount curve is of fundamental importance in the valuation and risk-management of financial securities, yet is essentially never directly observable in the market. Instead, the discount curve must be inferred out from securities the observable prices of which are functions of the discount curve.

In practice, the securities selected for the purpose of discount curve construction will nearly always be *linear* in a finite set of points on the discount curve. A prototypical example of such a security is an ordinary coupon bond, paying non-random cash amounts $c_1, c_2, \ldots, c_M$ at times $0 < t_1 < t_2 < \cdots < t_M$. Let the time 0 price of this bond be V; by elementary arguments, we must have

$$V = \sum_{j=1}^{M} c_j P(t_j). \tag{1}$$

Extending our setup to N such cash-paying securities—the *benchmark set*—we write

$$V_i = \sum_{j=1}^{M} c_{ij} P(t_j), \quad i = 1, \ldots, N, \tag{2}$$

where c_{ij} denotes the cash amount paid at time t_j on the ith security. Note that the time-line $\{t_j\}_{j=1}^{M}$ in practice would be obtained by merging together the cash-flow schedules of each of the N securities; as a consequence, many of the c_{ij} may be zero.

The valuation expression (2) can conveniently be expressed in matrix form. For this, define the M-dimensional discount bond vector

$$\mathbf{P} = (P(t_1), \ldots, P(t_M))^\top,$$

and let $\mathbf{V} = (V_1, \ldots, V_N)^\top$ be the vector of observable security prices. Also let $\mathbf{c} = \{c_{ij}\}$ be an $N \times M$ dimensional matrix containing all the cash-flows produced by the chosen set of securities; as mentioned, $\mathbf{c}$ would typically be quite sparse. The relation (2) can be written as

$$\mathbf{V} = \mathbf{cP}, \tag{3}$$

an equation that serves as the fundamental starting point for the estimation of $\mathbf{P}$.

As typically $M \neq N$, it is unlikely that (3) allows for a unique solution for $\mathbf{P}$: if $M > N$ (which is most often the case), an infinite set of solutions will typically

exist and additional structure must be imposed on the problem to chose a single **P**; if $M < N$, (3) will generally not have a solution, and we will need to look for a solution for **P** that minimizes some measure of price error. In any case, we are normally not interested solely in finding the discrete set of discount factors **P**, but also to construct a smooth continuous function for all t in some interval on the real line, to be used for the pricing and risk-management of securities outside of the benchmark set (see Sects. 2.3 and 2.4 below). This will involve imposing interpolation and, if we wish to move outside the interval $[t_1, t_M]$, extrapolation rules, in addition to enforcing the basic valuation constraint of (3).

2.2 Yield and forward formulation

The discount function $P(t)$ is fundamentally an exponentially decaying function, due to compounding interest. As such, it is common and appropriate to perform the curve fitting exercise on a logarithmic transformation of $P(t)$. One possible curve to consider is the *continuously compounded (term) yield* $y(t)$ defined by

$$e^{-y(t)t} = P(t) \Rightarrow y(t) = -t^{-1} \ln P(t). \tag{4}$$

The mapping $t \mapsto y(t)$ is known as the *yield curve*; it is related to the discount curve by the simple transformation (4). Of related interest is also the *instantaneous forward curve* $f(t)$, given by

$$P(t) = e^{-\int_0^t f(u)du} \Rightarrow f(t) = -d \ln P(t)/dt. \tag{5}$$

An alternative transformation considered in the literature (see e.g. McCulloch and Kochin 2000) is defined as

$$z(t) = y(t)t = -\ln P(t) \tag{6}$$

(such that $dz(t)/dt = f(t)$). Alternatively, we could consider

$$w(t) = -\frac{(\varepsilon + t)}{t} \ln P(t) = y(t)(\varepsilon + t) = \varepsilon y(t) + z(t), \tag{7}$$

where $\varepsilon > 0$ is some constant weight.

There is little econometric evidence to suggest that any particular transformation of the discount curve is fundamentally better than any other. From a curve construction purpose, however, some forms may be more convenient to work with than others, and may lead to more natural boundary conditions. For instance, $z(t)$ and $w(t)$ are often chosen due the fact that imposing linearity for large t (consistent with the specification of a standard *natural spline* boundary condition) will imply that both $y(t)$ and $f(t)$ will approach a flat asymptote for large t. Such asymptotic behavior is theoretically attractive (see e.g. Cox et al. 1985), and also avoids the uncontrolled extrapolatory growth in yields and forward rates associated with, say, the popular method of using

a natural cubic spline directly on the yield curve $y(t)$ (see Hagan and West 2004). On the other hand, $y(t)$ is often of more direct interest to market participants, so to the extent that extrapolatory behavior is either of secondary importance or can be controlled (for instance through expression of explicit boundary conditions), in some situations it may be most appropriate to work with $y(t)$. We also note that $y(t)$ is often flatter as a function of time than are transformations such as (7) and (6), which may in some circumstances simplify curve construction. In any case, going forward we keep the discussion general enough to allow for a user-specified choice of discount curve transformation.

2.3 Pricing of non-benchmark fixed income securities

Assume now that we have managed to construct a continuous discount curve or, more likely, some invertible transformation of $P(t)$. What may we then want to use such a curve for? Fundamentally, the resulting curve can be applied to price any fixed future cashflow; indeed, the time 0 value of a known cash-flow c paid at time T is just $cP(T)$, where $P(T)$ can be read off the constructed discount curve. This, in turn, allows us to price any fixed income security that pays streams of fixed cashflows, e.g. coupon-bearing bonds. More generally, however, a *random* cash-flow $X(T)$ paid at time T can be priced from the expression

$$V(0) = P(T)\mathrm{E}^T\left(X(T)\right) \tag{8}$$

where $\mathrm{E}^T\left(X(T)\right)$ is the expectation of $X(T)$ in the so-called T*-forward martingale measure* (see Jamishidian 1991). While it is outside the scope of this paper to deal in detail with arbitrage pricing relationships such as (8), we still want to make the point that for many interest rate derivatives, the expectation $\mathrm{E}^T\left(X(T)\right)$ is itself a function of the time 0 discount curve. For instance, if $X(T)$ represents a Libor rate published at time $T - \Delta$ (with Δ typically being 3 months), then

$$\mathrm{E}^T\left(X(T)\right) = \Delta^{-1}\left(\frac{P(T-\Delta)}{P(T)} - 1\right) \approx f(T), \tag{9}$$

where the approximation holds for small Δ. Important derivative contracts such as fixed-floating swaps, caps, and swaptions all involve payments of floating Libor rates or transformations thereof. The expression (9) alerts us to an important fact: securities valuation outside the benchmark set may very well require use of the discount function $P(\cdot)$ in ways other than a simple scale on fixed cash-flows. This, in turn, will impose additional requirements on the curve, e.g. that forward rates $f(T)$ computed from the model be smooth and well-behaved functions of T (as would certainly benefit (9)).

Along similar, but even more fundamental, lines is the observation that many modern models for fixed income derivatives valuation apply their stochastic dynamics to the instantaneous forward curve; see e.g. Heath et al. (1992). In these frameworks the curve $t \mapsto f(t)$ constitutes the *initial condition* for a dynamic diffusion model that moves all forward rates through calendar time in random, arbitrage-free fashion. If

$f(t)$ is not well-behaved as a function of t, it is easy to demonstrate that the dynamics of the points on the forward curve themselves become non-smooth, with drift-terms that may take very large absolute values. For instance, in the popular Gaussian model (see Jamishidian (1991) or Hull and White (1990)), the relevant (risk-neutral) dynamics of the short rate[3] $r(t)$ are

$$\begin{aligned} dr(t) &= \kappa\left(\theta(t) - r(t)\right)dt + \sigma dW(t), \\ \theta(t) &= \kappa^{-1}\frac{df(t)}{dt} - f(t) + \frac{1}{2}\kappa^{-3}\sigma^2\left(1 - e^{-2\kappa t}\right), \end{aligned} \tag{10}$$

where $\kappa, \sigma > 0$ are model parameters, and $W(t)$ is a scalar Brownian motion. Notice that $\theta(t)$ depends on the t-derivative of the forward curve; hence, if the forward curve f is not smooth in t, the drift term in this stochastic differential equation may become unwieldy. To the extent that we believe that a very rapid change in interest rates may be possible at a given date (e.g. around a rate policy meeting of the central bank) large gradients in f may be economically justifiable, but otherwise it is generally more reasonable to assume that the forward curve be smooth and well-behaved as a function of t.

2.4 Risk computations and hedging

Beyond pricing of non-benchmark securities, an important application of discount curve construction technology is the generation of *sensitivity reports*, to aid in risk management of securities positions. To describe common practices, consider a portfolio of securities with value V_0, where V_0 is a functional of the discount curve $P(t)$. The securities in the portfolio would typically not be in the benchmark set and could, say, contain a variety of interest rate options. As the discount curve is a function of the benchmark set values $\mathbf{V} = (V_1, \ldots, V_N)^\top$, we may write

$$V_0 = V_0\left(V_1, \ldots, V_N; \mathbf{p}\right),$$

where the vector $\mathbf{p}$ contains model parameters (e.g. option volatilities) and where the function $V_0(\cdot)$ is determined both from the valuation model of the security in question, and from the curve construction algorithm employed. For finite sized moves in $\mathbf{V}$ and $\mathbf{p}$, we have to first order

$$\Delta V_0 \approx \sum_{i=1}^{N} \frac{\partial V_0}{\partial V_i}\Delta V_i + \sum_i \frac{\partial V_0}{\partial p_i}\Delta p_i. \tag{11}$$

For the purpose of risk-managing first-order risk exposure to moves in the discount curve, (11) suggests that the collection of derivatives $\partial V_0/\partial V_i$, $i = 1, \ldots, N$ form a natural metric for portfolio risk. In particular, if all these derivatives are zero, our portfolio would, to first order, be immunized against any possible move in the discount

[3] That is, the interest rate at time t for deposits maturing at time $t + \epsilon$, with $\epsilon \downarrow 0$.

curve. On the other hand, if some or all of the derivatives were non-zero, we could manage our risk by setting up a hedge portfolio of benchmark securities, with notional $-\partial V_0/\partial V_i$ on the ith security. We emphasize that the resulting hedge would typically *not* be model-consistent: most interest models assume that yield curve risk originating from only a few stochastic yield curve factors that tend to move the curve smoothly, in a predominantly parallel fashion. Theoretically, a bucket-by-bucket immunization against all terms ΔV_i may then be considered an "overkill"—we typically hedge against far too many risk factors (N)—but is nevertheless standard industry practice and has proven to be very robust. Notice that bucket-hedging along these lines would, for instance, correctly reject the notion that we could perfectly hedge a 20-year bond with a 1-year bond, something that a one-factor interest rate model such as (10) would happily accept.

The simplest approach to computation of the derivative $\partial V_0/\partial V_i$ involves a manual bump[4] to V_i, followed by a reconstruction of the yield curve, and a subsequent repricing of the portfolio V_0. For it to work properly, it is important that the curve construction algorithm is fast and produces clean, local perturbations of the yield curve when benchmark prices are shifted. For instance, perturbing a short-dated bond price should not cause noticeable movements in the long-term part of the discount curve, lest we reach the erroneous conclusion (again) that we can hedge a 20-year bond with a 1-month bond. The requirement of perturbation locality tends to conflict with requirements of smoothness (Sect. 2.3), a trade-off we shall examine in some detail later in this article.

2.5 Choice of benchmark securities

We round off our description of the discount curve construction problem with a few words on the selection of the securities in the benchmark set. Mathematically, all that is required of securities in the benchmark is that their prices are observable and can be written as linear combinations of zero coupon bond prices. As we discussed, the latter requirement is satisfied by coupon bonds, but many other non-callable securities qualify as well; for instance, it is not difficult to verify that the generic valuation expression (1) also applies to such securities as forward rate agreements (FRAs) and fixed-floating interest rate swaps. (For an introduction to these and other basic fixed-income securities, see e.g. Hull 2000). However, rather than arbitrarily mix and match different kinds of securities, in practice one would always aim to pick benchmark securities that originate from the same market as the financial instruments that one ultimately would like to apply the discount curve to. For instance, to construct a curve for pricing Treasury bond securities, it is natural to choose a set of Treasury coupon bonds and T-Bills as the benchmark set, with maturities spanning the period for which we wish to construct a discount curve. Similarly, if we are interested in constructing a discount curve applicable for bonds issued by a particular firm, we would naturally use bonds and loans used by the firm in question. For capital markets purposes, the

[4] In practice, rather than bumping the price V_i outright, one may instead bump the quoted yield of the ith benchmark security (typically by 1 basis points). See also footnote 7.

most important yield curve is no doubt the *Libor curve*, constructed out of market quotes for Libor (London Interbank Offered Rate) deposits, fixed-floating interest rate swaps, and Eurodollar futures.

Beyond selection of a market, to generate a clean benchmark set, selection criteria may be applied to avoid illiquid, noisy, or otherwise non-representative securities affecting the discount curve. These criteria are typically market-specific, and may be associated with tax issues or seasoning; for instance, in Treasury markets, newly issued bonds ("on-the-run") are often substantially more liquid than older Treasury bonds, and a systematic price bias will often exist between new and old bonds. In the selection of benchmark securities, one would normally select bonds with equal liquidity characteristics. Rather than outright pruning of securities, it is also possible to mark certain securities in the benchmark set as less important than others, through the choice of weights in a fitting norm; we return to this in Sect. 4. On the flip-side, if the universe of observable security prices is sparse, one could contemplate adding fictitious securities to the benchmark set, constructed by interpolation rules applied to known securities. This practice is apparently common, but should be approached with some care, we think, as it may lead to odd curve behavior and suboptimal hedges. Also, interpolation rules are more naturally applied to fundamental quantities such as forward rates and term yields.

Finally, as we explained in Sect. 2.4 above, it is often the case that the securities in the benchmark set will ultimately be used in a hedging exercise where positions in the benchmark set is used to immunize a portfolio of non-benchmark securities against interest rate risk. In this case, we should obviously use only benchmark securities that we deem (a) liquid enough to allow for hedge trading; and (b) having near-zero *basis risk* to the securities in portfolio we want to hedge. The latter point essentially (re-) states the basic idea that we want the securities in the benchmark set to be "similar" to the non-benchmark securities we are interested in, lest we want to expose ourselves to the risk of the two markets diverging. For instance, hedging a swap portfolio with Treasury bonds would expose the hedger to moves in the spread (the so-called *swap spread basis*) between the Libor and Treasury discount curves.

3 Tension spline basics

Having now outlined basic issues in the discount curve construction problem, we temporarily move away from the financial context in order to provide some required background material from the tension spline literature. To start our discussion, consider a standard C^2 cubic spline $g(t)$ interpolating a set of data points (t_j, g_j), $j = 1, \ldots, M$. Here, the t_j are said to be *knots*. By necessity, a cubic spline interpolant is piecewise linear in its second derivative, i.e.

$$g''(t) = \frac{t_{j+1} - t}{h_j} g''_j + \frac{t - t_j}{h_j} g''_{j+1}, \quad t \in [t_j, t_{j+1}], \tag{12}$$

where $h_j \equiv t_{j+1} - t_j$ and where $g''_i \equiv g''(t_i)$, with hyphens denoting differentiation with respect to time. As is well-known, explicit equations for the interpolating cubic

spline can classically be recovered in $O(M)$ operations by integration of (12) and subsequently requiring the curve to pass through given data points as well as having continuous first derivatives across knots (see e.g. Press et al. (1992) for details and computer code). To uniquely specify the cubic spline, derivative boundary conditions must be expressed at $t = t_1$ and $t = t_M$. A classical boundary condition is $g_1'' = g_M'' = 0$, leading to the *natural cubic spline*.

While cubic splines have a number of useful features, they have, loosely speaking, a built-in aversion to make tight turns (which will cause large values of g''). As discussed earlier, this in turn will often produce extraneous inflection points and non-local behavior, in the sense that perturbation of a single g_j will affect the appearance of the curve for t-values far from t_j. Also, monotonicity and convexity properties of the original data-set will typically not be preserved. An attractive remedy to these shortcomings of the cubic spline is to permit some tension in the spline, that is, to apply a tensile force to the end-points of the spline. Formally, this can be accomplished (see Schweikert 1966) by replacing Eq. 12 with

$$g''(t) - \sigma^2 g(t) = \frac{t_{j+1} - t}{h_j}\left(g_j'' - \sigma^2 g_j\right) + \frac{t - t_j}{h_j}\left(g_{j+1}'' - \sigma^2 g_{j+1}\right), \quad t \in [t_j, t_{j+1}], \tag{13}$$

where $\sigma > 0$ is a measure of the tension applied to the cubic spline.[5] Notice that we have replaced the assumption of a piecewise linear second derivative with the assumption that the quantity $g''(t) - \sigma^2 g(t)$ is linear on each sub-interval $[t_j, t_{j+1}]$.

3.1 Basic properties

Before turning to an explicit representation of the hyperbolic tension spline, let us consider a few important characteristics of this class of splines. First, we notice that when the tension parameter $\sigma^2 = 0$, (13) and (12) are identical, i.e. the tension spline degenerates into a regular cubic spline. On the other hand, when $\sigma^2 \gg 1$, (13) asymptotically reduces to linear interpolation, as

$$\lim_{\sigma^2 \to \infty} g(t) = \frac{t_{j+1} - t}{h_j} g_j + \frac{t - t_j}{h_j} g_{j+1}, \quad t \in [t_j, t_{j+1}]. \tag{14}$$

For positive, finite value of σ, (13) evidently defines a curve that is a hybrid between a cubic spline and a linear spline.

The convergence of the tension spline towards a piecewise linear curve as $\sigma^2 \to \infty$ can be shown to be *uniform*, i.e. (14) holds uniformly in $[t_j, t_{j+1}]$ for $j = 1, \ldots, M-1$. Similarly

[5] Extension to non-uniform tension parameter is straightforward and involves replacing σ with σ_j in (13), with σ_j then being a measure of the tension applied locally to the curve in the interval $[t_j, t_{j+1}]$. We return to non-linear tension in Sect. 5.

$$\lim_{\sigma^2 \to \infty} g'(t) = \frac{g_{j+1} - g_j}{h_j} \quad \text{and} \quad \lim_{\sigma^2 \to \infty} g''(t) = 0$$

uniformly in any closed subinterval of $[t_j, t_{j+1}]$. See Pruess (1976) for details and a proof. The uniform convergence is important as it guarantees us that we can preserve the monotonicity and convexity properties of the underlying discrete data set, simply by choosing a sufficiently high value of the tension factor. Due to this property, hyperbolic tension splines are said to be *shape-preserving*. Generalizing, suppose we introduce constraints on function values, first derivatives, or second derivatives. As long as these constraints are satisfied by linear interpolation, there will exist some value of the tension parameter σ^2 (possibly $\sigma^2 = 0$) which will make the tension spline satisfy the constraints. This observation is key to algorithms for automatic selection of σ^2 from externally specified function constraints. See, for instance, Lynch (1982) and Renka (1987) for details and efficient algorithms for automatic tension selection.

3.2 B-spline basis

A classical derivation of the equations for hyperbolic tension splines parameterized by function values in knots is given in Cline (1974) and closely mimics the construction of cubic splines. As it turns out, on all intervals $[t_j, t_{j+1}]$ hyperbolic tension splines can be written as linear combinations of the functions 1, t, $e^{-\sigma t}$, $e^{\sigma t}$. We omit the detailed results of Cline (1974) here, as they are best-suited for simple interpolation problems where function values in all knots are (i) explicitly given and (ii) must be matched perfectly. For yield curve construction purposes, neither is necessarily true: we rarely, if ever, know discount function values in all knots and we may often be content with an imperfect fit to observed prices. In this situation, it is often useful to avoid parameterizing the spline directly through known function values, but instead rely on a basis representation of the form

$$g(t) = \sum_{k=0}^{M+1} b_k x_k(t) \tag{15}$$

where the $x_k(t)$, $k = 0, \ldots, M+1$ is a set of $M+2$ basis functions, and the b_k are constant weights. We need to use $M+2$ (and not M) basis functions to ensure that specified boundary conditions at t_1 and t_M can be satisfied; in other words, the dimension of the space of hyperbolic tension splines with M knots is $M+2$.

In the choice of basis-functions, it is particularly convenient if the basis is local, in the sense that the individual $x_k(t)$ functions are zero for most values of t. This is the case for the *exponential B-spline* basis suggested by Koch and Lyche (1989, 1993). Briefly, to construct this basis, we first extend our knot set with six new points t_{-2}, t_{-1}, $t_0, t_{M+1}, t_{M+2}, t_{M+3}$ satisfying $t_{-2} < t_{-1} < t_0 < t_1$ and $t_M < t_{M+1} < t_{M+2} < t_{M+3}$ but otherwise arbitrary. We then define the *exponential hat functions*

$$B_{j,2}(t) = \begin{cases} \Psi''_j(t), & t_j \le t < t_{j+1}, \\ \Phi''_{j+1}(t), & t_{j+1} \le t \le t_{j+2}, \\ 0, & \text{otherwise}, \end{cases}$$

where

$$\Psi_j(t) = \frac{\sinh\left(\sigma\left(t - t_j\right)\right) - \sigma(t - t_j)}{\sigma^2 \sinh\left(\sigma h_j\right)},$$
$$\Phi_j(t) = \frac{\sinh\left(\sigma\left(t_{j+1} - t\right)\right) - \sigma(t_{j+1} - t)}{\sigma^2 \sinh\left(\sigma h_j\right)}. \tag{16}$$

Notice that the exponential hat function $B_{j,2}(t)$ is non-zero only for $t \in [t_j, t_{j+2}]$, where the j's can take values $-1, 0, \ldots, M$. For $k = 3, 4$, we recursively define *quadratic* and[6] *cubic tension B-splines* as

$$B_{j,k}(t) = \Lambda_{j,k-1}(t) - \Lambda_{j+1,k-1}(t),$$

where

$$\Lambda_{j,k}(t) = \begin{cases} 0, & t < t_j, \\ c_{j,k}^{-1} \int_{t_j}^{t} B_{i,k}(y)\, dy, & t_j \le t \le t_{j+k}, \\ 1, & \text{otherwise}, \end{cases}$$

and

$$c_{j,k} = \int_{t_j}^{t_{j+k}} B_{j,k}(y)\, dy.$$

The cubic tension B-splines are particularly important to us here, as they can be shown to form a basis for the hyperbolic tension spline introduced earlier. Specifically, in (15) we can set $x_k(t) = B_{k-2,4}(t)$ such that

$$g(t) = \sum_{k=0}^{M+1} b_k B_{k-2,4}(t). \tag{17}$$

We list an explicit expression for the functions $B_{j,4}$ in the next section; we emphasize that $B_{j,4}(t) = 0$ if $t \notin [t_j, t_{j+4}]$, and $B_{j,4}(t) > 0$ if $t \in [t_j, t_{j+4}]$. In other words, the cubic tension B-spline basis is local, and (note the upper and lower limits on the sum)

$$g(t) = \sum_{k=j-1}^{j+2} b_k B_{k-2,4}(t), \quad \text{if } t \in [t_j, t_{j+1}], \tag{18}$$

[6] In the limit $\sigma \downarrow 0$, the tension B-splines become identical to the classical (polynomial) B-splines discussed in, e.g., de Boor (1978).

as for any value of t only four B-splines will be non-zero.

3.3 Explicit basis representation

The recursion for B-splines $B_{j,4}$ above can be written explicitly (see Koch and Lyche 1989, 1993). To state the result, we need some notation (which we borrow from Kvasov 2000):

$$z_j = \Psi_{j-1}(t_j) - \Phi_j(t_j), \quad z'_j = \Psi'_{j-1}(t_j) - \Phi'_j(t_j), \quad y_j = t_j - z_j/z'_j,$$

$$b_j^{(1)} = \frac{b_{j+2} - b_{j+1}}{y_{j+2} - y_{j+1}}, \quad b_j^{(2)} = \frac{b_j^{(1)} - b_{j-1}^{(1)}}{z'_{j+1}}.$$

Then, for $t \in [t_j, t_{j+1}]$,

$$\begin{aligned} g(t) &= \sum_{k=j-1}^{j+2} b_k B_{k-2,4}(t) = b_j + b_{j-1}^{(1)}(t-y_j) + b_{j-1}^{(2)}\Phi_j(t) + b_j^{(2)}\Psi_j(t) \\ &= b_{j-1}\frac{\Phi_j(t)/z'_j}{y_j - y_{j-1}} + b_j\left(1 - \frac{t-y_j + \Phi_j(t)/z'_j}{y_{j+1}-y_j} - \frac{\Phi_j(t)/z'_j}{y_j - y_{j-1}} + \frac{\Psi_j(t)/z'_{j+1}}{y_{j+1}-y_j}\right) \\ &\quad + b_{j+1}\left(\frac{t-y_j + \Phi_j(t)/z'_j}{y_{j+1}-y_j} - \frac{\Psi_j(t)/z'_{j+1}}{y_{j+2}-y_{j+1}} - \frac{\Psi_j(t)/z'_{j+1}}{y_{j+1}-y_j}\right) \\ &\quad + b_{j+2}\frac{\Psi_j(t)/z'_{j+1}}{y_{j+2}-y_{j+1}}. \end{aligned} \tag{19}$$

We have deliberately kept definitions written in terms of $\Phi_j(t_j)$ and $\Psi_{j-1}(t_j)$, as this will allow us to later generalize results to non-hyperbolic tension splines.

Let us briefly note that to evaluate (19) for arbitrary values of σ, we need a robust way to compute the hyperbolic functions sinh and cosh for large and small arguments. A number of standard techniques exist for this, see e.g. Renka (1987) and Rentrop (1980). For small σ, a suitably truncated Taylor-expansion around zero is sufficient.

3.4 Boundary conditions

From (19), it follows that, for $j = 1, \ldots, M$,

$$g_j = b_j + \frac{b_{j-1}^{(1)}\Psi_{j-1}(t_j) - b_{j-2}^{(1)}\Phi_j(t_j)}{z'_j}, \tag{20}$$

$$g'_j = \frac{b^{(1)}_{j-1}\Psi'_{j-1}(t_j) - b^{(1)}_{j-2}\Phi'_j(t_j)}{z'_j}, \tag{21}$$

$$g''_j = b^{(2)}_{j-1}, \tag{22}$$

where we have used the fact that $\Psi_j(t_j) = \Psi'_j(t_j) = 0$ (see (16)). We can use this result to enforce boundary conditions at t_1 and t_M, by setting the free basis weights b_0 and b_{M+1} to specific combinations of neighboring b_j. For instance, suppose we are interested in a natural spline boundary condition $g''(t_1) = 0$. This requires

$$b^{(2)}_0 = \frac{\frac{b_2-b_1}{y_2-y_1} - \frac{b_1-b_0}{y_1-y_0}}{z'_1} = 0 \Rightarrow b_0 = b_1 - \frac{(b_2-b_1)(y_1-y_0)}{y_2-y_1}. \tag{23}$$

Similarly, if we want $g''(t_M) = 0$, we must set

$$b_{M+1} = b_M + \frac{(b_M - b_{M-1})(y_{M+1}-y_M)}{y_M - y_{M-1}}. \tag{24}$$

3.5 An explicit integral

For later use, we are interested in the evaluation of the integral

$$\int_{t_j}^{t_{j+1}} \left(g''(t)^2 + \sigma^2 g'(t)^2\right) dt$$

where g is a hyperbolic tension spline. We write this as

$$\int_{t_j}^{t_{j+1}} \left(g''(t)\cdot g''(t) + \sigma^2 g'(t)\cdot g'(t)\right) dt$$

and integrate by parts:

$$\begin{aligned}\int_{t_j}^{t_{j+1}} \left(g''(t)^2 + \sigma^2 g'(t)^2\right) dt &= \left[g''(t)g'(t)\right]_{t_j}^{t_{j+1}} \\ &\quad - \int_{t_j}^{t_{j+1}} \left(g^{(3)}(t) - \sigma^2 g'(t)\right) g'(t)\, dt \\ &= g''_{j+1}g'_{j+1} - g''_j g'_j - d_j\left(g_{j+1} - g_j\right)\end{aligned} \tag{25}$$

where

$$d_j \equiv \frac{g''_{j+1} - \sigma^2 g_{j+1}}{h_j} - \frac{g''_j - \sigma^2 g_j}{h_j}.$$

We note that we have used that, by definition, hyperbolic tension splines have $g^{(3)}(t) - \sigma^2 g'(t)$ piecewise constant and equal to d_j on each interval $[t_j, t_{j+1}]$ (see Eq. 13). We can express the d_j's directly as functions of the B-spline coefficients b_{j-1}, b_j, b_{j+1} through (20)–(22). A few rearrangements show that the correct result is

$$\begin{aligned} g''_j - \sigma^2 g_j = b_{j-1}&\left(\frac{[1-\sigma^2\Phi_j(t_j)]/z'_j}{y_j - y_{j-1}}\right) + b_{j+1}\left(\frac{[1-\sigma^2\Psi_{j-1}(t_j)]/z'_j}{y_{j+1} - y_j}\right) \\ &+ b_j\left(1 - \frac{[1-\sigma^2\Phi_j(t_j)]/z'_j}{y_j - y_{j-1}} - \frac{[1-\sigma^2\Psi_{j-1}(t_j)]/z'_j}{y_{j+1} - y_j}\right) \end{aligned}$$

such that

$$d_j = \alpha^j_{j-1} b_{j-1} + \alpha^j_j b_j + \alpha^j_{j+1} b_{j+1} + \alpha^j_{j+2} b_{j+2}$$

for easily computed constants α^j_k, $k = j-1, j, j+1, j+2$.

4 Non-parametric discount curve construction algorithm

Our basic discount curve construction algorithm is non-parametric and based on minimization of a penalized least-squares term, with the penalty term aiming to provide the user with control over pricing precision, curve shape, and perturbation locality. To state our algorithm, let φ be our representation of the yield curve (for instance, φ could be w, as defined in Sect. 2.2), such that

$$P(t) = P\left(t, \varphi(t)\right).$$

Let

$$\varphi = \left(\varphi(t_1), \ldots, \varphi(t_M)\right)^\top,$$

with φ being the curve we have used as a representation of the discount curve. Our starting point for the construction of φ is (3), which we state in the form

$$\mathbf{V} = \mathbf{cP}(\varphi) \tag{26}$$

to highlight the dependence of discount bonds on φ. For instance, if we set $\varphi(t) = w(t)$ (with $w(t)$ defined in Sect. 2.2), then

$$\begin{aligned} \mathbf{P}(\varphi) &= \left(P\left(t_1, \varphi(t_1)\right), \ldots, P\left(t_N, \varphi(t_N)\right)\right)^\top \\ &= \left(e^{-\varphi(t_1)t_1/(\varepsilon+t_1)}, \ldots, e^{-\varphi(t_N)t_N/(\varepsilon+t_N)}\right)^\top. \end{aligned}$$

4.1 Norm formulation

For reasons discussed earlier, we may not be able to (or even want to) solve (26) exactly for φ. Instead, we consider minimization of the norm

$$\mathscr{I}(\varphi) = \frac{1}{N}\left(\mathbf{V} - \mathbf{cP}(\varphi)\right)^{\top} \mathbf{W}^2 \left(\mathbf{V} - \mathbf{cP}(\varphi)\right) + \lambda\left(\int_{t_1}^{t_M} \left[\varphi''(t)^2 + \sigma^2\varphi'(t)^2\right] dt\right) \tag{27}$$

where $\mathbf{W}$ is a diagonal $N \times N$ matrix with elements W_i, and λ and σ are positive constants. The norm consists of three separate terms:

- A least-squares term

$$\frac{1}{N}\left(\mathbf{V} - \mathbf{cP}(\varphi)\right)^{\top} \mathbf{W}^2 \left(\mathbf{V} - \mathbf{cP}(\varphi)\right) = \frac{1}{N}\sum_{i=1}^{N} W_i^2 \left(V_i - \sum_{j=1}^{M} c_{ij} P\left(t_j, \varphi(t_j)\right)\right)^2$$

  where W_i is the ith diagonal element of $\mathbf{W}$. This term is an outright precision-of-fit norm and measures the degree to which the constructed discount curve can replicate input security prices. The weights W_i can be used to assign different importance to the various securities in the benchmark set, and/or to translate the precision of the fit from raw dollar amounts into a more intuitive quantities, such as security-specific quoted yields.[7]
- A weighted smoothness term $\lambda \int_{t_1}^{t_M} \varphi''(t)^2\, dt$, penalizing high second-order gradients of φ to avoid kinks and discontinuities.
- A weighted curve-length term $\lambda\sigma^2 \int_{t_1}^{t_M} \varphi'(t)^2\, dt$, penalizing oscillations, lack of perturbation locality, and excess convexity/concavity.

Sections 2.3 and 2.4 provide additional economic justification for the terms $\lambda \int \varphi''(t)^2\, dt$ and $\lambda\sigma^2 \int \varphi'(t)^2\, dt$, respectively. We note that for the case $\sigma^2 = 0$ the norm $\mathscr{I}(\varphi)$ coincides with the one chosen in Tanggaard (1997) and the norm-minimizing curve will be a cubic spline. By additionally adding a curve-length term, we aim to control oscillations and other undesired behavior of the cubic smoothing spline.

4.2 Norm minimization

As our estimate of the term structure of interest rates, we use the curve $\hat{\varphi}$ which minimizes $\mathscr{I}(\varphi)$ over the space $\mathscr{A} = C^2[t_1, t_M]$ of all twice differentiable functions $[t_1, t_M] \to \mathbb{R}$. That is,

[7] Most fixed-income securities are quoted through some type of yield, e.g. $V_i = g_i(r_i)$ where r_i is the quoted yield and g_i is a function that encapsulates the quoting convention. The quantity $D_i = -dg_i/dr_i$ is known as the *duration* of V_i. Setting $W_i = 1/D_i$ in the least-squares norm will turn price deviations into yield deviations.

$$\hat{\varphi} = \arg\min_{\varphi \in \mathscr{A}} \mathscr{I}(\varphi). \tag{28}$$

Examination of the circumstances under which the Gateaux variation of $\mathscr{I}(\varphi)$ equals zero (the necessary condition for a minimum) reveals that $\hat{\varphi}$ must be a natural[8] hyperbolic tension spline with tension factor σ and knots at all t_j, $j = 1, \ldots, M$. See Appendix A for details.

Going forward, we let $\hat{\varphi}(t)$ be a hyperbolic tension spline with natural boundary conditions,[9] such that

$$\hat{\varphi}(t) = \sum_{k=0}^{M+1} \hat{b}_k x_k(t), \quad t \in [t_1, t_M],$$

where $x_k(t) = B_{k-2,4}(t)$ are the B-spline functions defined earlier and the $\hat{b}_k$'s are constant weights, with $\hat{b}_0$ and $\hat{b}_{M+1}$ satisfying the linear constraints (23)–(24). We assume, as discussed earlier, that additional points $t_{-2}, t_{-1}, t_0, t_{M+1}, t_{M+2}, t_{M+3}$ have been added to our coupon time line, such that $x_0(t)$ and $x_{M+1}(t)$ are well-defined. Let $\hat{\mathbf{b}} = (\hat{b}_1, \ldots, \hat{b}_M)^\top$, and $\mathbf{x}(t) = (x_1(t), \ldots, x_M(t))^\top$, and express the constraints on $\hat{b}_0$ and $\hat{b}_{M+1}$ as

$$\hat{b}_0 = \mathbf{a}_0^\top \hat{\mathbf{b}}, \quad \hat{b}_{M+1} = \mathbf{a}_{M+1}^\top \hat{\mathbf{b}},$$

where $\mathbf{a}_0$ and $\mathbf{a}_{M+1}$ are M-dimensional vectors (with only two non-zero elements). It then follows that

$$\hat{\varphi}(t) = \left(\mathbf{x}(t)^\top + x_0(t)\mathbf{a}_0^\top + x_{M+1}(t)\mathbf{a}_{M+1}^\top\right)\hat{\mathbf{b}} \equiv \mathbf{y}(t)^\top \hat{\mathbf{b}},$$

where $\mathbf{y}$ is an M-dimensional natural spline basis vector. Also, with $\hat{\varphi} = (\hat{\varphi}(t_1), \ldots, \hat{\varphi}(t_M))^\top$ let

$$\hat{\varphi} = \mathbf{Y}\hat{\mathbf{b}}, \tag{29}$$

where $\mathbf{Y} = \{Y_{jk}\}$ is an $M \times M$ dimensional matrix with elements $Y_{jk} = y_k(t_j)$. Note that $\mathbf{Y}$ is tri-diagonal due to the locality of our B-spline basis. From the results of Appendix A, we have

[8] It follows from the result in Appendix A that if we prefer to explicitly specify boundary conditions of the type $\varphi'(t_1) = a$, $\varphi'(t_1) = b$, the tension spline still minimizes (28), but now on the smaller space $\mathscr{A} = \{\varphi \in C^2[t_1, t_M] : \varphi'(t_1) = a, \varphi'(t_M) = b\}$.

[9] Extensions to other boundary conditions are straightforward, using the results of Sect. 3.4. In particular, we note that prescribing explicit values for φ or its first or second derivative at t_1 and t_M will always result in constraints on $\hat{b}_0$ and $\hat{b}_{M+1}$ of the form $\hat{b}_0 = \mathbf{a}_0^\top \hat{\mathbf{b}} + c_0$ and $\hat{b}_{M+1} = \mathbf{a}_{M+1}^\top \hat{\mathbf{b}} + c_{M+1}$, where c_0 and c_{M+1} are constants, and $\mathbf{a}_0$ and $\mathbf{a}_{M+1}$ are M-dimensional vectors with mostly zero elements.

$$\hat{\mathbf{b}} = \arg\min_{\mathbf{b}} \frac{1}{N} \left(\mathbf{V} - \mathbf{cP}(\mathbf{Yb})\right)^{\top} \mathbf{W}^2 \left(\mathbf{V} - \mathbf{cP}(\mathbf{Yb})\right) + \lambda \left(\sum_{j=1}^{M} \varphi(t_j)(d_j - d_{j-1}) \right).$$

Here, $d_0 = d_M = 0$ and, for $1 \le j \le M-1$,

$$d_j = \sum_{k=j-1}^{j+2} \alpha_k^j b_k$$

for constants α_k^j defined in Sect. 3.5. Again using the boundary conditions $b_0 = \mathbf{a}_0^{\top}\mathbf{b}$ and $b_{M+1} = \mathbf{a}_{M+1}^{\top}\mathbf{b}$, it follows that

$$\sum_{j=1}^{M} \varphi(t_j)(d_j - d_{j-1}) = \mathbf{b}^{\top}\mathbf{Y}^{\top}\mathbf{Ab}$$

where $\mathbf{A}$ is a banded (5 diagonal bands) $M \times M$ matrix. Thereby

$$\hat{\mathbf{b}} = \arg\min_{\mathbf{b}} \frac{1}{N} \left(\mathbf{V} - \mathbf{cP}(\mathbf{Yb})\right)^{\top} \mathbf{W}^2 \left(\mathbf{V} - \mathbf{cP}(\mathbf{Yb})\right) + \lambda \left(\mathbf{b}^{\top}\mathbf{Y}^{\top}\mathbf{Ab}\right). \tag{30}$$

A necessary condition for the optimum can be found by differentiating with respect to $\mathbf{b}$ and set the resulting expression equal to zero. By standard matrix calculus, this yields[10]

$$\frac{2}{N}\mathbf{Y}^{\top}\mathbf{B}(\mathbf{Y}\hat{\mathbf{b}})\mathbf{c}^{\top}\mathbf{W}^2 \left(\mathbf{V} - \mathbf{cP}(\mathbf{Y}\hat{\mathbf{b}})\right) - \lambda \left(\mathbf{Y}^{\top}\mathbf{A} + \mathbf{A}^{\top}\mathbf{Y}\right)\hat{\mathbf{b}} = \mathbf{0}, \tag{31}$$

where $\mathbf{B}(\mathbf{Yb}) = \mathbf{B}(\varphi)$ is an $M \times M$ diagonal matrix of derivatives, $B_{jj} = \partial P(t_j, \varphi(t_j))/\partial\varphi(t_j)$. For instance, for the case $\varphi(t) = w(t)$, we get

$$B_{jj} = -e^{-\varphi(t_j)t_j/(\varepsilon+t_j)} \frac{t_j}{\varepsilon + t_j}, \quad j = 1, \ldots, M.$$

4.3 Numerical solution

Solution of (31) for $\hat{\mathbf{b}}$ must be done numerically, for instance by the usage of (quasi-) Newton methods or similar. To briefly give a specific algorithm, consider for instance a basic Gauss–Newton iterative scheme, in which the pth iteration estimate $\hat{\mathbf{b}}^{(p)}$ is updated from the first-order Taylor approximation

$$\mathbf{P}\left(Y\hat{\mathbf{b}}^{(p+1)}\right) \approx \mathbf{P}\left(\mathbf{Y}\hat{\mathbf{b}}^{(p)}\right) + \mathbf{B}\left(\mathbf{Y}\hat{\mathbf{b}}^{(p)}\right)\mathbf{Y}\left(\hat{\mathbf{b}}^{(p+1)} - \hat{\mathbf{b}}^{(p)}\right). \tag{32}$$

[10] Appendix A writes these matrix equations out in detail.

Inserting (32) and the first-order approximation $\mathbf{B}\left(\mathbf{Y}\hat{\mathbf{b}}^{(p+1)}\right) \approx \mathbf{B}\left(\mathbf{Y}\hat{\mathbf{b}}^{(p)}\right)$ into (31), we get the simple updating scheme

$$\frac{2}{N}\mathbf{Y}^{\top}\mathbf{B}\left(\mathbf{Y}\hat{\mathbf{b}}^{(p)}\right)\mathbf{c}^{\top}\mathbf{W}^{2}\left[\mathbf{V}-\mathbf{cP}\left(\mathbf{Y}\hat{\mathbf{b}}^{(p)}\right)-\mathbf{cB}\left(\mathbf{Y}\hat{\mathbf{b}}^{(p)}\right)\mathbf{Y}\left(\hat{\mathbf{b}}^{(p+1)}-\hat{\mathbf{b}}^{(p)}\right)\right] \\ -\lambda\left(\left(\mathbf{Y}^{\top}\mathbf{A}+\mathbf{A}^{\top}\mathbf{Y}\right)\hat{\mathbf{b}}^{(p+1)}\right)=\mathbf{0}$$

which can be rearranged to

$$\left[\mathbf{G}\left(\mathbf{Y}\hat{\mathbf{b}}^{(p)}\right)^{\top}\mathbf{G}\left(\mathbf{Y}\hat{\mathbf{b}}^{(\mathbf{p})}\right)+\frac{N}{2}\lambda\mathbf{H}\right]\hat{\mathbf{b}}^{(p+1)} \\ =\mathbf{G}\left(\mathbf{Y}\hat{\mathbf{b}}^{(p)}\right)^{\top}\left[\mathbf{WV}-\mathbf{WcP}\left(\mathbf{Y}\hat{\mathbf{b}}^{(p)}\right)+\mathbf{G}\left(\mathbf{Y}\hat{\mathbf{b}}^{(p)}\right)\hat{\mathbf{b}}^{(p)}\right]$$

where $\mathbf{H}\equiv\mathbf{Y}^{\top}\mathbf{A}+\mathbf{A}^{\top}\mathbf{Y}$ and $\mathbf{G}\left(\mathbf{Y}\hat{\mathbf{b}}^{(p)}\right)\equiv\mathbf{WcB}\left(\mathbf{Y}\hat{\mathbf{b}}^{(p)}\right)\mathbf{Y}$. For our applications we find that this scheme converges rapidly; see footnote 18 for some representative computation times. Tanggaard (1997) also reports good results in application of the Gauss–Newton algorithm to the simpler cubic spline problem.

In numerical implementation of the iteration above, we obviously should take care to exploit the efficiencies arising from the fact that many of the matrices ($\mathbf{Y}$, $\mathbf{H}$, $\mathbf{W}$, $\mathbf{B}$, $\mathbf{H}$, ..) have a sparse band-structure. To ensure quick convergence, we should also attempt to supply a good guess for $\hat{\mathbf{b}}^{(0)}$. When constructing a curve for the first time, it is most convenient to let the user input a guess $\mathbf{P}_g$ for the discount bond array, which can be inverted into a guess φ_g for the values of the tension spline in the knots. We then face a standard interpolation problem governed by the simple tri-diagonal system (29), allowing us to write

$$\mathbf{Y}\hat{\mathbf{b}}^{(0)}=\varphi_g \tag{33}$$

which can be solved by LU decomposition for $\hat{\mathbf{b}}^{(0)}$ in $O(M)$ operations. For small perturbations on an existing curve—such as those that happen in a perturbation analysis or naturally through the passage of time—the basis vector for the unperturbed curve serves as a natural guess for $\hat{\mathbf{b}}^{(0)}$.

4.4 Choice of λ

So far, we have assumed that the parameter λ has been specified exogenously by the user. In practice, however, a good magnitude of λ may sometimes be hard to ascertain by inspection, and a procedure to estimate λ directly from the data is often useful. Such data-driven estimation of λ is a complex problem that remains a subject of active research in the literature; we consequently limit ourselves to a rather cursory outline of a few techniques below. Additional methods (including information criteria, such as AIC) can be found in standard references for non-parametric regression, e.g. Eubank (1988). As discussed in (Eubank 1988), Chapter 5, we should notice that automatic

selection of λ—no matter how sophisticated the method—does not relieve the user from visually inspecting results and fine-tuning settings as needed.

4.4.1 User-specified RMS norm

In many settings, users may be able to specify their tolerance for the least-squares component of the norm (27), for instance by observation of bid-offer spreads on the securities in question. In this case, it is natural to replace the optimization problem (30) with the constrained optimization problem

$$\hat{\mathbf{b}} = \arg\min_{\mathbf{b}} \mathbf{b}^\top \mathbf{Y}^\top \mathbf{A}\mathbf{b}, \tag{34}$$

$$\frac{1}{N} \left(\mathbf{V} - \mathbf{cP}(\mathbf{Yb})\right)^\top \mathbf{W}^2 \left(\mathbf{V} - \mathbf{cP}(\mathbf{Yb})\right) = \gamma^2, \tag{35}$$

where γ is an exogenously specified constant. Notice in particular that if a perfect fit to the benchmark set is required (which is standard in construction of a Libor curve, as discussed in Hagan and West (2004)), we set $\gamma = 0$.

The Lagrangian for the above problem becomes

$$\hat{\mathbf{b}} = \arg\min_{\mathbf{b}} \mathbf{b}^\top \mathbf{Y}^\top \mathbf{A}\mathbf{b} + \rho \left[\frac{1}{N} \left(\mathbf{V} - \mathbf{cP}(\mathbf{Yb})\right)^\top \mathbf{W}^2 \left(\mathbf{V} - \mathbf{cP}(\mathbf{Yb})\right) - \gamma^2 \right] \tag{36}$$

where the Lagrange multiplier ρ must be determined such that the constraint (35) is satisfied at the optimum of (36). Apart from a constant scale, (36) is identical to (30), so we solve the constrained optimization problem (34)–(35) through the following iteration over λ:

1. Given a guess for λ, find the optimum value of $\hat{\mathbf{b}}$, from solution of (30);
2. Compute $\mathscr{I}_{ls} = \frac{1}{N} \left(\mathbf{V} - \mathbf{cP}(\mathbf{Y}\hat{\mathbf{b}})\right)^\top \mathbf{W}^2 \left(\mathbf{V} - \mathbf{cP}(\mathbf{Y}\hat{\mathbf{b}})\right)$;
3. If $\mathscr{I}_{ls} = \gamma^2$, stop; otherwise update λ and go to step 1.

In general, the optimum precision norm $\mathscr{I}_{ls} = \mathscr{I}_{ls}(\lambda)$ will be a declining function in λ and, provided that a root exists[11] to $\mathscr{I}_{ls}(\lambda) = \gamma^2$, the updating in Step 3 can be done by any standard root search algorithm.

4.4.2 Morozov discrepancy principle

Suppose that we have reason to believe that benchmark prices (or, rather, weighted prices $W_i V_i$) are noisy, with the noise having standard deviation s. Then, according to the *Morozov Discrepancy Principle* Morozov (1966), in the algorithm in Sect. 4.4.1, we choose $\gamma = s$. To estimate s, we may rely on empirical market observations or,

[11] For ill-posed benchmark sets, there may be instances where $\mathscr{I}_{ls}(0) > \gamma^2$ (see Appendix C for an example). If the desired precision is unattainable, we can either increase γ^2 or perhaps prune the benchmark security set.

for a fully data-driven approach, on an in-sample estimator for s, such as the Rice estimator in Rice (1984).

4.4.3 L-curve

The Morozov Discrepancy Principle originally was formulated to handle ill-posed systems of linear equations. A popular technique with similar origins is that of the *L-curve* (Hansen 1992). In a nutshell, the L-curve is simply a log-log parametric plot of the least-squares term of the norm (27) against the regularity penalty term. Specifically, set

$$\begin{aligned} \mathscr{I}_{ls} &= \frac{1}{N}\left(\mathbf{V} - \mathbf{cP}(\hat{\varphi})\right)^{\top} \mathbf{W}^2 \left(\mathbf{V} - \mathbf{cP}(\hat{\varphi})\right), \\ \mathscr{I}_{reg} &= \int_{t_1}^{t_M} \left[\hat{\varphi}''(t)^2 + \sigma^2 \hat{\varphi}'(t)^2\right] dt, \end{aligned} \tag{37}$$

where we emphasize that we use the optimal value $\hat{\varphi}$ in evaluation of the two terms. As $\hat{\varphi}$ is function of λ, we may write $\mathscr{I}_{ls} = \mathscr{I}_{ls}(\lambda)$ and $\mathscr{I}_{reg} = \mathscr{I}_{reg}(\lambda)$; letting λ vary allows us to plot $\log_{10} \mathscr{I}_{reg}$ against $\log_{10} \mathscr{I}_{ls}$.

For a noisy, ill-posed problem—which can arise, say, if two benchmark securities have (near-)identical cash-flows but different prices—the curve $\hat{\varphi}$ becomes increasingly irregular as λ is lowered, causing a rapid increase in $\mathscr{I}_{reg}(\lambda)$; as a consequence, the plot of $\log_{10} \mathscr{I}_{reg}$ against $\log_{10} \mathscr{I}_{ls}$ will take a characteristic L-shape. According to Hansen (1992), a good value for λ is the one at the "corner" of the L-curve, i.e. at the point of maximum curvature.

For a well-posed problem, $\mathscr{I}_{reg}(\lambda)$ approaches a constant plateau as $\lambda \downarrow 0$, and the L-curve changes shape accordingly. While the L-curve method is traditionally applied to ill-posed problems only, the L-curve plot is, in fact, generally useful to visualize the trade-off between precision-of-fit and curve regularity; see Appendix C for a brief numerical example.

4.4.4 Cross-validation

In basic *leave-one-out cross-validation (CV)*, security i is removed from the benchmark set and the discount curve estimation procedure is applied to the reduced benchmark set. The resulting curve is then used to *predict* the value of the ith benchmark security, resulting in an estimate V_i^{cv} and a (weighted) prediction error of

$$e_i^{cv} = (V_i - V_i^{cv}) W_i.$$

Repeating this prediction exercise of all N securities in the benchmark set allows us to compute a total CV prediction error criterion

$$CV = CV(\lambda) = \sqrt{N^{-1}\sum_{i=1}^{N}\left(e_i^{cv}\right)^2}, \tag{38}$$

where we have emphasized that CV is a function of λ. According to the CV approach, the optimal value of λ is the one that minimizes the prediction error (38). The determination of λ can be done numerically, using a one-dimensional optimizer. As each iteration of the optimization procedure will involve the construction of N separate discount curves, the cross-validation approach will be computationally intensive.

4.4.5 GCV

Assume for a moment that our curve construction algorithm were to write

$$\mathbf{cP}(\hat{\varphi}) = \mathbf{Q}(\lambda)\mathbf{V} + \mathbf{Q}_0, \tag{39}$$

where $\mathbf{Q}_0$ is an N-dimensional vector and $\mathbf{Q}$ is an $N \times N$ *smoother matrix* that depends on λ. In this case, the CV criterion becomes approximately equal to[12]

$$GCV(\lambda) = \frac{\mathscr{I}_{ls}(\lambda)N^2}{\mathrm{tr}\,(\mathbf{I} - \mathbf{Q}(\lambda))}, \tag{40}$$

where tr is the usual trace operator and $\mathscr{I}_{ls}(\lambda)$ is the mean-square-error defined in (37). Determining λ to minimize the quantity GCV is known as *generalized cross-validation*; see Craven and Wahba (1979) for details.

While $GCV(\lambda)$ is much faster to compute than $CV(\lambda)$, the justification for its use hinges upon (39) being true. In reality, however, we have the more complicated relation (from 31)

$$\frac{2}{N}\mathbf{Y}^\top\mathbf{B}(\hat{\varphi})\mathbf{c}^\top\mathbf{W}^2\left(\mathbf{V} - \mathbf{cP}(\hat{\varphi})\right) = \lambda\mathbf{H}\hat{\mathbf{b}}, \quad \mathbf{H} = \mathbf{Y}^\top\mathbf{A} + \mathbf{A}^\top\mathbf{Y}, \tag{41}$$

which is generally not of the correct form. To overcome this, we can follow the linearization approach in Tanggaard (1997) and assume that $\mathbf{P}(\hat{\varphi})$ is approximately linear in $\hat{\varphi}$

$$\mathbf{P} \approx \mathbf{q}_0 + \mathbf{B}\hat{\varphi} = \mathbf{q}_0 + \mathbf{BY}\hat{\mathbf{b}},$$

which assumes that the matrix of derivatives $\mathbf{B}$ is constant. We then get, after a bit of rearrangement of (41),

$$\mathbf{cP}(\hat{\varphi}) = \mathbf{cBY}\left[\left(\frac{2}{N}\mathbf{Y}^\top\mathbf{Bc}^\top\mathbf{W}^2\mathbf{cBY} + \lambda\mathbf{H}\right)^{-1}\frac{2}{N}\mathbf{Y}^\top\mathbf{Bc}^\top\mathbf{W}^2\right]\mathbf{V} + \mathbf{Q}_0$$

[12] More precisely, GCV is a weighted version of CV.

where $\mathbf{Q}_0$ is a constant offset vector, the exact value of which is irrelevant. In the GCV criterion (40) we may then, as an approximation, use

$$\mathbf{Q}(\lambda) \approx \mathbf{cB}(\hat{\varphi})\mathbf{Y}\left[\left(\frac{2}{N}\mathbf{Y}^\top\mathbf{B}(\hat{\varphi})\mathbf{c}^\top\mathbf{W}^2\mathbf{cB}(\hat{\varphi})\mathbf{Y}+\lambda\mathbf{H}\right)^{-1}\frac{2}{N}\mathbf{Y}^\top\mathbf{B}(\hat{\varphi})\mathbf{c}^\top\mathbf{W}^2\right].$$

This expression is straightforward to compute after the optimum $\hat{\varphi}$ has been determined as described in Sect. 4.3.

4.5 Choice of σ

As we pointed out earlier, in Sect. 3.1, there is some literature on auto-selection of the tension parameter, starting from, say, given constraints on convexity and/or extrema of the tension spline. As functions such as $z(t)$, $y(t)$, $w(t)$ (see Sect. 2.2) are subject to no strong[13] requirements on their convexity properties, it is, however, not an easy problem to set up practical constraints that imply unique values for σ. While auto-selection of tension may occasionally be of use for benchmark sets that are very close to violating arbitrage constraints, most often the selection of σ must be based on an analysis of the economics of the discount curve problem in some detail. Elements of this analysis will include gauging the perceived "stiffness" of various parts of the discount curve and taking a stance of the importance of perturbation locality versus smoothness of the yield/forward curve. We illustrate the process through a detailed numerical example in Sect. 6. For now, we notice that it may, in fact, be quite useful to use different values of the tension parameter at different sections of the discount curve, as a means to express views about *local* behavior of the discount curve. Fortunately, extending out methodology to non-constant tension in the discount curve is straightforward; see Sect. 5.1 below. Finally, let us note (again) that the tension parameter has a strong impact on curve shape and perturbation behavior, so the fact that determination of the tension parameter may require careful analysis should, obviously, not tempt one to opt for a naive approach where one selects either $\sigma = 0$ or $\sigma = \infty$.

5 Extensions

Before turning to numerical results, we briefly cover a number of extensions to the basic algorithm above.

5.1 Non-uniform tension

So far, we have assumed that the tension parameter σ is constant. In practice, however, it is often useful to relax this assumption. For instance, as pointed out in Cline

[13] A basic no-arbitrage constraint of any discount curve is that the forward curve $f(t)$ is positive for all t. This condition, however, typically fails to put any constraint on σ (i.e. no values of $\sigma \in [0, \infty)$ will actually violate arbitrage).

(1974), tension splines with constant tension are affected by scaling, with the tension spline equations changing in non-linear fashion when knot intervals are scaled up or down. To remove this effect, we can use a local tension parameter σ_j for each interval $[t_j, t_{j+1}]$ determined to satisfy $\sigma_j h_j = q$, where q is a global "effective" tension parameter. Further, if we wish to apply any of the techniques to automatically select tension parameters (see e.g. Renka 1987), we normally need the flexibility of controlling curve shape locally. Finally, we may also simply be interested in expressing a trading view on different parts of the yield curve, through user-manipulation of local curve shape by means of changes in curve tension.

Consider thus the situation, where each curve section $[t_j, t_{j+1}]$ is equipped with a knot-specific tension parameter σ_j, $j = 1, \ldots, M-1$. For $t \in [t_j, t_{j+1}]$ the curve $\varphi(t)$ is then characterized by (compare to 13)

$$\varphi''(t) - \sigma_j^2 \varphi(t) = \frac{t_{j+1} - t}{h_j} \left(\varphi''(t_j) - \sigma_j^2 \varphi(t_j) \right) + \frac{t - t_j}{h_j} \left(\varphi''(t_{j+1}) - \sigma_j^2 \varphi(t_{j+1}) \right),$$

and required to be twice differentiable. Allowing σ to depend on the knot index is, fortunately, straightforward in our B-spline setting: we simply need to modify the functions Φ and Ψ in (16) to

$$\Psi_j(t) = \frac{\sinh\left(\sigma_j \left(t - t_j\right)\right) - \sigma_j (t - t_j)}{\sigma_j^2 \sinh\left(\sigma_j h_j\right)},$$
$$\Phi_j(t) = \frac{\sinh\left(\sigma_j \left(t_{j+1} - t\right)\right) - \sigma_j (t_{j+1} - t)}{\sigma_j^2 \sinh\left(\sigma_j h_j\right)}.$$

With this minor modification, all results in Sect. 3 hold as written.

As for the yield curve estimation procedure in Sect. 4, all results again hold as written, provided that we change the optimization norm to reflect the non-constancy of the tension-spline parameter. Specifically, we alter $\mathscr{I}(\varphi)$ in (27) to

$$\mathscr{I}(\varphi) = \frac{1}{N} \left(\mathbf{V} - \mathbf{cP}(\varphi) \right)^\top \mathbf{W}^2 \left(\mathbf{V} - \mathbf{cP}(\varphi) \right) + \lambda \left(\sum_{j=1}^{M-1} \int_{t_j}^{t_{j+1}} \left[\varphi''(t)^2 + \sigma_j^2 \varphi'(t)^2 \right] dt \right). \tag{42}$$

5.2 User-specified knots

Specification of the tension-spline as the solution to an optimization problem dictates the placement of knots in all coupon payment dates, as discussed earlier. In many situations, knots will thereby be spaced quite closely together, making the curve appear quite smooth, even for large values of the tension parameter. In practice, it may be of interest to make the curve coarser, for instance to further improve locality under perturbation (see Sect. 6). This can be accomplished by allowing the user to specify

directly the positions of knots. To formalize this idea, consider setting $\varphi(t)$ to a tension spline with a set of knots $\{\tau_j\}_{j=1}^{Q}$, for simplicity assumed to satisfy $\tau_1 = t_1$ and $\tau_Q = t_M$. Given these knots, and the imposed constraint that $\varphi(t)$ be a tension spline on $\{\tau_j\}_{j=1}^{Q}$, we now still proceed to minimize the norm (27); the resulting minimum can obviously never improve the minimum obtained with knots in $\{t_j\}_{j=1}^{M}$. Following steps similar to those in Sect. 4.2, we write for the optimal tension spline

$$\hat{\varphi}(t) = \sum_{k=0}^{Q+1} \hat{b}_k x_k(t), \quad t \in [t_1, t_M],$$

where the $\hat{b}_k$'s are weights on B-spline basis functions $x_k(t)$, $k = 0, \ldots, Q+1$ (and where we have extended the knot grid $\{\tau_j\}_{j=1}^{Q}$ with six additional points, as before). We set $\hat{\mathbf{b}} = (\hat{b}_1, \ldots, \hat{b}_Q)^\top$ and

$$\hat{\varphi} = \left(\hat{\varphi}(t_1), \ldots, \hat{\varphi}(t_M)\right)^\top = \mathbf{Y}\hat{\mathbf{b}}$$

where $\mathbf{Y}$ is now an $M \times Q$ matrix with elements $Y_{jk} = x_k(t_j)$, with an adjustment for boundary conditions at $j = 1$ and $j = M$ (see Sect. 4.2). While $\mathbf{Y}$ will be sparse due to the local nature of our basis, $\mathbf{Y}$ is generally no longer square or tri-diagonal. The Q-dimensional vector $\hat{\mathbf{b}}$ can be found by solving

$$\hat{\mathbf{b}} = \arg\min_{\mathbf{b}} \frac{1}{N} \left(\mathbf{V} - \mathbf{cP}(\mathbf{Yb})\right)^\top \mathbf{W}^2 \left(\mathbf{V} - \mathbf{cP}(\mathbf{Yb})\right) + \lambda \left(\mathbf{b}^\top \mathbf{Z}^\top \mathbf{A}\mathbf{b}\right) \tag{43}$$

where $\mathbf{Z}$ and $\mathbf{A}$ are banded $Q \times Q$ matrices with $\left(\hat{\varphi}(\tau_1), \ldots, \hat{\varphi}(\tau_Q)\right)^\top = \mathbf{Z}\hat{\mathbf{b}}$ and

$$\mathbf{b}^\top \mathbf{Z}^\top \mathbf{A}\mathbf{b} = \sum_{j=1}^{Q} \varphi(\tau_j)(d_j - d_{j-1}), \quad d_j = \int_{\tau_j}^{\tau_{j+1}} \left[\varphi''(t)^2 + \sigma_j^2 \varphi'(t)^2\right] dt.$$

The solution of (43) problem can be done, as before, with the Gauss–Newton algorithm applied to the system

$$\frac{2}{N} \mathbf{Y}^\top \mathbf{B}(\mathbf{Y}\hat{\mathbf{b}})\hat{c}^\top \mathbf{W}^2 \left(\mathbf{V} - \mathbf{cP}(\mathbf{Y}\hat{\mathbf{b}})\right) - \lambda \left(\mathbf{Z}^\top \mathbf{A} + \mathbf{A}^\top \mathbf{Z}\right) \hat{\mathbf{b}} = \mathbf{0}.$$

λ can be found as in Sect. 4.4.1 or could be set to zero for a pure least-squares minimization solution. The latter would obviously require the chosen knots to be set sparsely enough to allow for a unique least-squares solution.

5.3 Generalized tension B-splines

So far, our attention has focused primarily on hyperbolic tension splines which conveniently solve the fairly natural minimization problem (27). Hyperbolic tension splines

are, however, just a particular member of a larger class of splines, all of which share the property that tension parameters allows us to move smoothly from cubic splines (when tension is zero) to linear splines (when tension is infinite). As discussed in Kvasov (2000), all splines in this class can be characterized by a particular choice of the generating functions Φ and Ψ used in the B-spline representation in Sect. 3.3. Below we list a few possible choices of such generalized tension B-splines (also known as GB-splines).

Rational spline (linear denominator):

$$\Psi_j(t) = \frac{(t-t_j)^3}{h_j\left(1+\sigma_j(t_{j+1}-t)\right)\left(6+6\sigma_j h_j+2\sigma_j h_j^2\right)},$$
$$\Phi_j(t) = \frac{(t_{j+1}-t)^3}{h_j\left(1+\sigma_j(t-t_j)\right)\left(6+6\sigma_j h_j+2\sigma_j h_j^2\right)}.$$

Rational spline (quadratic denominator):

$$\Psi_j(t) = \frac{(t-t_j)^3}{h_j\left(1+\sigma_j(t-t_j)(t_{j+1}-t)/h_j\right)\left(6+6\sigma_j h_j+2\sigma_j h_j^2\right)},$$
$$\Phi_j(t) = \frac{(t_{j+1}-t)^3}{h_j\left(1+\sigma_j(t-t_j)(t_{j+1}-t)/h_j\right)\left(6+8\sigma_j h_j+2\sigma_j h_j^2\right)}.$$

Exponential spline:

$$\Psi_j(t) = \frac{(t-t_j)^3\exp\left(-\sigma_j(t_{j+1}-t)\right)}{h_j\left(6+6\sigma_j h_j+\sigma_j h_j^2\right)},$$
$$\Phi_j(t) = \frac{(t_{j+1}-t)^3\exp\left(-\sigma_j(t-t_j)\right)}{h_j\left(6+6\sigma_j h_j+\sigma_j h_j^2\right)}.$$

We note that the rational splines involve no computationally expensive transcendental functions, making them a popular alternative to hyperbolic tension splines. In any case, substituting any of these splines for the hyperbolic tension spline in our algorithms above is straightforward. Obviously, however, if we elect to directly solve systems (30) (or (43) for $\lambda > 0$), it must be understood that the matrix terms representing the non-RMS part of the optimization norm no longer can be interpreted as exactly representing integrals of weighted sums of $\varphi''(t)^2$ and $\varphi'(t)^2$. We could obviously consider computing explicitly the representation of the norm $\mathscr{J}(\varphi)$ in (27) for each choice of tension spline type, but the practical importance of this may be relatively modest.

6 Numerical examples

Broadly speaking, yield curve construction tends to take place in one of two settings: either we (i) attempt to construct a curve from a disorderly and noisy set of securities with maturity and cash-flow dates not aligned in any particular order or pattern; or (ii) we have an orderly set of liquid benchmark securities, arranged in strictly increasing order of maturity and making payments on a (nearly) homogenous time line. Case (i) may arise when we, say, attempt to construct a corporate bond curve from whatever still-alive debt securities a specific firm may have issued over time, many of which will have different liquidity and tax characteristics. A similar situation often arises in government-issued debt, unless the benchmark-set is pruned to only include liquid (on-the-run) securities. Case (ii) typically arises in swap markets, where a very liquid set of securities (deposits, futures, and interest rate swaps) with standardized and non-overlapping maturities are quoted actively. While a relatively loose-fitting curve would be most appropriate for the first case, the second case normally requires a very tight fit to all benchmark securities, to reflect the high liquidity of the market prices.

While our curve construction algorithm is designed to deal with both case (i) and (ii) above, in the numerical experiments in his paper we will stay mostly in the spirit of case (ii) and we shall work with a small dataset designed to mimic the market input used to construct a Libor discount curve. As mentioned earlier, and discussed further in Golub and Tilman (2000), the Libor curve is considered the most important benchmark curve in fixed income pricing, so it is natural for us to focus on this setting. Moreover, the Libor curve is conveniently based on a relatively small set of securities, so it is feasible for us to give an example that can be reproduced (and tested) easily without the need to communicate large and noisy data-sets. We also note that hedging and pricing of most fixed income derivatives is virtually always done with a Libor curve, making an analysis of locality to benchmark security perturbation (which we shall undertake shortly) particularly important for the Libor curve construction. In any case, virtually all of the characteristics of our curve construction module can be demonstrated in a Libor curve setting. While careful numerical tests of case (i) must be relegated to a dedicated paper,[14] Appendix C contains a few representative results, primarily to illuminate the methods for auto-selection of λ listed in Sect. 4.4.

6.1 Benchmark data

While in practice a Libor curve is constructed out of short-term deposits, Eurodollar futures, and interest rate swaps, for clarity of exposition we simplify somewhat and only use swaps. In particular, all our benchmark securities are here assumed to par-valued unit-notional fixed-for-floating swaps that pay coupons on a semi-annual schedule. For our purposes, the swaps can be represented as coupon bonds with a time 0 value of $V = 1$. For a swap with annualized coupon θ and maturity $T = 0.5 \cdot k$, the

[14] For the zero tension case, Tanggaard (1997) shows many illuminating test results for case (i), employing multiple methods for the choice of λ.

Table 1 Benchmark data

Maturity	0.5	1	1.5	2	2.5	3	4	5	7	10	12	15	20	30
Par Coupon (%)	2.75	3.10	3.30	3.43	3.53	3.30	3.78	3.95	4.25	4.50	4.65	4.78	4.88	4.85
Duration	0.49	0.98	1.45	1.92	2.37	2.83	3.68	4.50	6.00	7.98	9.12	10.62	12.68	15.72

cash flow $c(t_i)$ at time $t_i = 0.5 \cdot i, i > 0$, is written as

$$c(t_i) = \begin{cases} 0.5 \cdot \theta, & 0 < i < k \\ 1 + 0.5 \cdot \theta, & i = k \\ 0, & i > k \end{cases} \tag{44}$$

Notice that the cash flow at $t_i = T$ (where $i = k$) includes redemption of the notional (here \$1).[15]

We use $N = 14$ swaps in our test, with maturities spanning six months to 30 years; our coupon payment time line is thereby $t_j = j \cdot 0.5$, $j = 1, \ldots, M$, where $M = 60$. Our time-line for spline construction purpose thus spans the interval $[t_1, t_M] = [0.5, 30]$. Table 1 lists maturities and par coupons (θ) for our 14 benchmark securities. We have also included in the table the duration of each benchmark security; see footnote 7 and any textbook on finance on how to compute bond durations.

While somewhat idealized, our data-set emulates a number of stylistic properties of real-life data: (a) short- and medium-term securities trade in maturities spaced closer than is the case for long-dated securities; (b) there is a sudden drop in par coupons in the first part of the curve (in our example, from $t = 2.5$ to $t = 3$);[16] (c) the term structure of par coupons is for the most part upward-sloping; and (d) the term structure of par coupons is for the most part concave.

6.2 Basic results

For spline construction purposes, let us start by using a non-parametric approach with knots in all 60 payment dates t_j, $j = 1, \ldots, 60$. We assume, for now, that our spline $\varphi(t)$ is hyperbolic and represents the term yield $y(t)$ (see Sect. 2.2). We use the method in Sect. 4.4.1 to determine the weight λ; our precision norm is weighted by the inverse of the individual bond durations as in footnote 7, and constrained to produce a (weighted) RMS value of 0.1 basis points (= 1/100,000). Figures 1 and 2 below show the resulting term structures for the term yield $y(t)$ and the instantaneous forward rate

[15] For the reader familiar with swap mechanics, we notice that the upfront value $V = 1$ represents the value of a floating leg with back-end exchange of notional, whereas the coupon stream in (44) represents the fixed leg of a swap.

[16] While a thorough explanation of this effect is beyond the scope of this paper, we note that this discontinuity reflects the fact that the short end of the swap curve in practice would be constructed from (convexity-adjusted) Eurodollar futures, whereas the medium- and long-term parts of the curve would be constructed from swaps. The two markets normally trade with a certain basis relative to each other, causing a jump in the yield curve as we move from one type of instrument to the next.

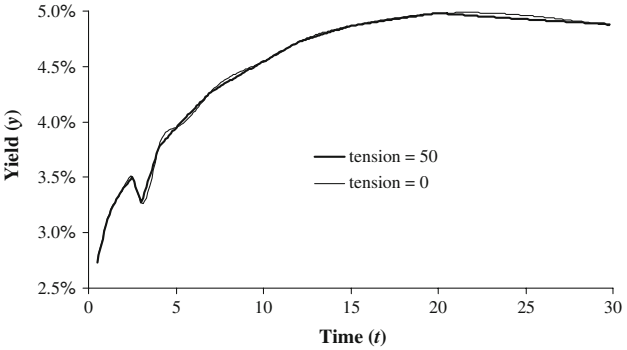

Fig. 1 Yield curve. *Notes*: The figure shows the yield curve $y(t)$ as computed from a fully non-parametric approach where weighted RMS error is constrained to 0.1 basis points. The discount map is $\varphi(t) = y(t)$. The tension parameter σ is uniform and equal to the values indicated in the figure

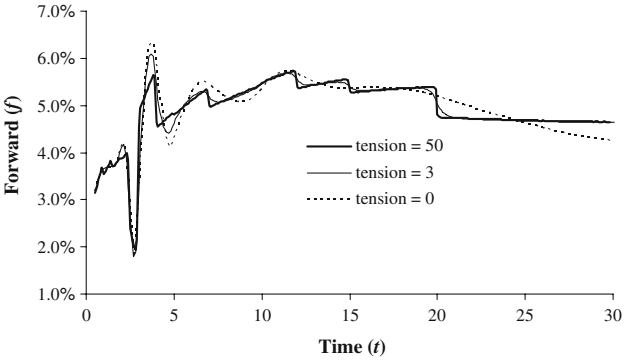

Fig. 2 Forward curve. *Notes*: The figure shows the instantaneous forward curve $f(t)$ consistent with the yield curve in Fig. 1

curve $f(t)$, for various choices of a constant tension parameter σ. We do not graph the discount function $P(t)$, which always take the basic shape of an exponentially decaying function with no particularly interesting features.

As we would expect, increasing the tension parameter moves us from a smooth yield curve to a piecewise linear one. Similarly, the forward curve (which involves differentiating the yield curve, see Sect. 2) moves from a smooth—but rather wiggly—forward curve toward a "saw-tooth" curve with discontinuities at each swap maturity. Around $t = 3$, the yield curve displays a sharp drop as expected, generating high whip-sawing gradients in the forward curve; the higher the tension parameter, the faster the effect of the yield curve drop disappears from the forward curve. For sufficiently high values of the tension parameter, the yield curve is concave everywhere, except around $t = 3$. For small values of the tension parameter, on the other hand, the yield curve has multiple inflection points along the entire curve, and displays what may be excessive concavity for maturities beyond 20 years.

Which value of the tension parameter gives the "best" curve is obviously situation-dependent, but the intermediate value of $\sigma = 3$ does a decent job at producing

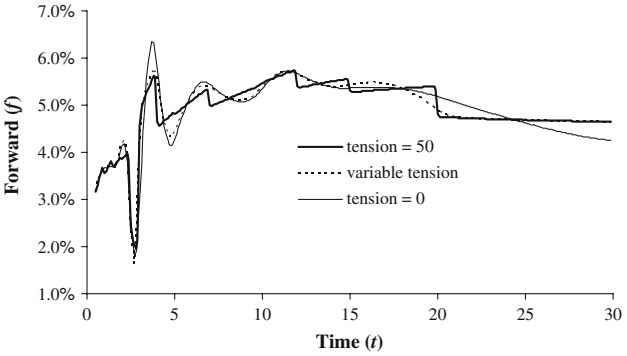

Fig. 3 Forward curve. *Notes*: The figure shows the instantaneous forward curve $f(t)$, generated from the same algorithm as in Fig. 2. The "variable tension" graph is produced by using tension parameters of $\sigma = 10$ for $3 \leq t < 4$, $\sigma = 2$ for $t \geq 20$, and $\sigma = 0$ everywhere else

a relatively smooth forward curve with limited yield concavity for $t > 20$. Still, it would probably be preferable to dampen the forward rate overshoot at $t = 3$ further. To do so, we could consider using non-uniform tension parameters in the spline. In particular, should our primary concern be to dampen the forward rate overshoot at $t = 3$ and the yield concavity above $t = 20$, it may be reasonable to use a high value of the tension parameter in these regions, and a low one elsewhere. Figure 3 illustrates the forward curve shape that can be produced this way; notice how the curve behaves like a simply bootstrapped curve in the region around $t = 3$ and $t > 20$, yet smoothly pastes onto cubic spline behavior in other regions.

The technique of using high tension values in certain parts of the curve may also be useful around dates where the forward curve is expected to jump (sometimes known as a "turn"), e.g. at dates coinciding with central bank policy meetings. If one has a concrete view of the magnitude of the forward curve jump, then one can also use a forward curve *overlay* to enforce the desired shape. The idea is straightforward and is included in Appendix B for completeness. Use of forward curve overlays has no impact on perturbation locality.

6.3 User-specified knots

So far we have relied on a non-parametric approach that dictates spline knots at all coupon payments dates. For the construction of Libor curves where the benchmark securities normally have orderly spaced maturities, it is obvious to experiment with the effect of placing knots only at maturities of the benchmark securities.[17] While this will necessarily result in curves that are sub-optimal in the sense defined in Sect. 4.2,

[17] Needless to say, location of knots for noisier and less orderly datasets is more challenging, and relying on a fully non-parametric approach is often the most straightforward approach. See Tanggaard (1997) for an empirical comparison of non-parametric cubic splines and regression splines with "rule of thumb" knot placement, in the context of Danish government bonds.

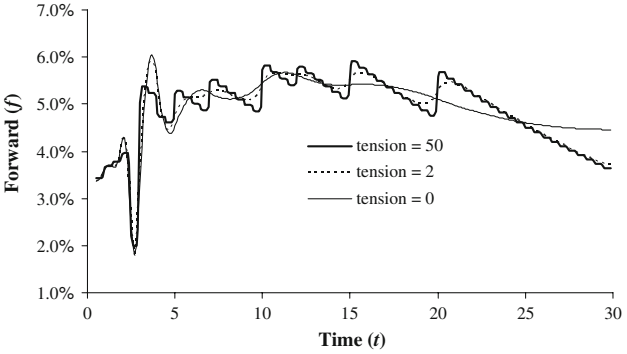

Fig. 4 Forward curve. *Notes*: The figure shows the instantaneous forward curve $f(t)$ as computed from a fully non-parametric approach where weighted RMS error is constrained to 0.1 basis points. The discount map is $\varphi(t) = ty(t)$. The tension parameter σ is uniform and equal to the values indicated in the figure

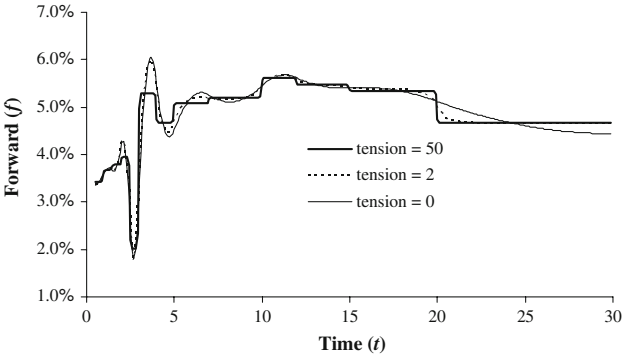

Fig. 5 Forward curve. *Notes*: The figure shows the instantaneous forward curve $f(t)$ as computed with knots set to benchmark security maturities, and where weighted RMS error is constrained to 0.1 basis points. The discount map is $\varphi(t) = ty(t)$. The tension parameter σ is uniform and equal to the values indicated in the figure

this suboptimality may be outweighed by computational benefits[18] (the dimension of the basis vector is reduced from 60 to 14) and possibly more intuitive manipulation of the curve.

As it turns out, when $\varphi(t) = y(t)$ placing knots as described above results in yield and forward curves that are essentially indistinguishable from those computed in Sect. 6.2. This, however, is not the case for arbitrary specifications of φ. To demonstrate, we now set $\varphi(t) = ty(t)$ and repeat the curve construction examples of Fig. 2, with and without user-specified knots. The results are in Figs. 4 and 5.

[18] For a user-specified λ, the fully non-parametric curve construction algorithm here takes less than 2/100 second on a single-processor 1.6 GHz Pentium PC. This time includes input-output transfer and construction of a first basis vector guess through (33), starting from a guess that the yield curve is flat. For the case of user-located knots, the computation time is roughly cut in half. If we iterate for λ (as in Sect. 4.4.1), the computation time depends on how good our initial guess for λ is, but in our example the total computation time (including the fixed data transfer overhead) typically roughly doubles.

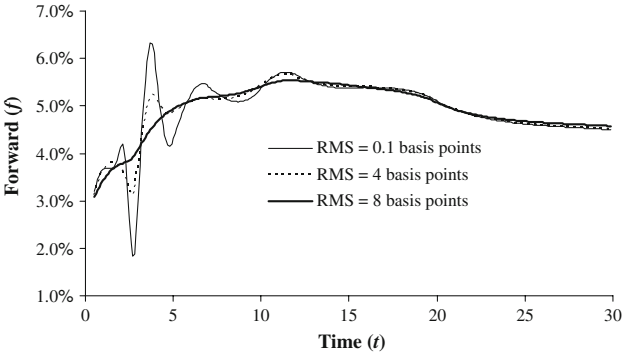

Fig. 6 Forward curve. *Notes*: The figure shows instantaneous forward curve $f(t)$, as computed from an algorithm similar to that used in Fig. 2, but with the weighted RMS norms constrained to the values indicated in the figure. The tension parameter was set to $\sigma = 0.5$ in all graphs

At low levels of the tension parameters, using a reduced set of knots makes little difference relative to the full non-parametric approach. For high tension parameters both approaches produce forward curves that are close to piecewise flat on their respective knot sets, but the appearances of the curves are quite different, with the non-parametric approach producing a forward curve that roughly takes the form of a "reverse saw-tooth" (compare to Fig. 2). We note that the curves for $\varphi = ty$ itself, rather than for $f(t)$, turn out to be quite similar for the two cases (graphs are omitted for brevity), with the non-parametric curve φ being about 1% shorter than the curve produced; the choice of φ in our algorithms can apparently have considerable impact on the resulting yield and forward curves, and must be treated as an important design parameter. This is consistent with earlier discussion in Sect. 2.2.

6.4 Precision of fit

We revert back to the specification $\varphi(t) = y(t)$ and the fully non-parametric setting of Sect. 6.2, and now wish to briefly demonstrate the effects of changing the weight λ between curve regularity and precision of fit. Figure 6 shows the effect of increasing the RMS fit precision from 0.1 basis points to 8 basis points. As one would expect, areas of rapid change in the yield curve get increasingly smoothed out as the required RMS fit precision is decreased. We point out that if we were only interested in a localized smoothing of the forward curve in the region around $t = 3$, this could be accomplished by lowering the RMS weights (W_i) for the specific benchmark securities maturing around $t = 3$. The resulting curve shape is easily imagined, and we omit it for brevity.

6.5 Input perturbation

Let us now examine how a perturbation in market data for a single benchmark security will affect the forward curve. As discussed earlier, this exercise is important in

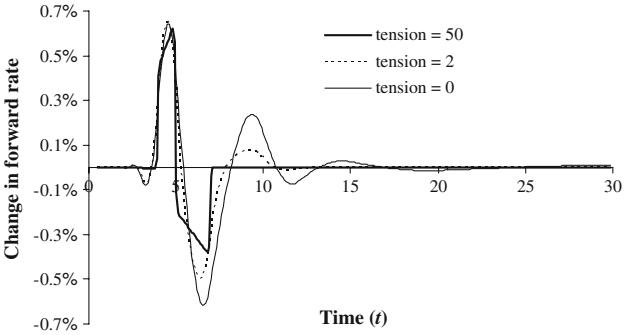

Fig. 7 Changes in forward curve. *Notes*: The figure shows the change in the instantaneous forward curve $f(t)$ associated with moving the 5-year par rate from 3.95% to 4.05%. The curve construction algorithm is similar to that used in Fig. 2. The tension parameter σ is uniform and equal to the values indicated in the figure

risk-management and hedging of interest rate sensitive securities. For our tests we arbitrarily pick the 5-year swap and bump its par coupon by 10 basis points, from 3.95% to 4.05%. As we can see in Fig. 7, using too low a value for the tension parameter results in strongly non-local perturbation effects, with the change in the 5-year par rate causing ripples in the forward curve that extend all the way up to the 30-year maturity bucket. For hedging purposes this is, as described earlier, quite inconvenient. As tension is increased, however, the perturbation effects become increasing local, and in the high-tension limit affect only the forward rate bucket $t \in [4, 7]$.

The ability to control perturbation locality through tension parameters is quite important in applications, and does away with one of the primary drawbacks of cubic splines. As before, by using non-uniform tension parameters, we can control such locality in a security-specific way, allowing us great flexibility to adapt to our specific application requirements curve smoothness, precision of fit, and perturbation locality.

7 Conclusion

This article has introduced a new methodology to construct yield curves from quoted security prices. Working in a general setup capable of accommodating most real bond and swap markets, our construction methodology uses generalized tension splines to address a number of key issues, including control of convexity and "locality", as well as providing explicit trade-off between curve regularity properties and (weighted) errors in the fit to benchmark bonds. We have demonstrated that the hyperbolic tension spline yield curve, in particular, can be constructed non-parametrically from a natural penalized least-squares problem. Our specific algorithms are based on local GB splines applied to appropriate transformations of the discount curve, allowing for numerical efficiency and great flexibility in the choice of tension-parameters and/or tension spline type.

The numerical results of this article demonstrate the usage and scope of our methodology by constructing a Libor benchmark curve, likely the most important

application of yield curve building routines. Another, slightly different, application of our methodology would be in constructing best-fit curves on sets of noisy corporate or Treasury bonds, to aid in empirical analysis of these markets. While Appendix C contain some indicative results, we leave a detailed examination of this application to future research. Another use of our algorithm is in the construction of term structures of survival probabilities and hazard rates, as needed in the pricing of credit default swaps and other default-risky securities. It can be shown (see, e.g., Andersen 2003) that the mathematical formulation of hazard rate curve construction is identical to that used in Sect. 2.1, allowing for direct application of our methodology. Other topics of future research in the area of term structure estimation include the application of routines to automatically select tension parameters from user-specified curve criteria, as well as a comparative analysis of various tension splines (e.g. hyperbolic versus rational). Consideration of optimization norms different from that used in this paper could also be undertaken.

Finally, let us note that applications of tension spline technology to problems in finance are not limited to yield curve construction. Indeed, the problem of constructing parameter surfaces or curves from discrete market observations is an extremely pervasive one in virtually all branches of finance. Examples from contingent claims pricing applications would include maturity interpolation of option pricing model parameters; construction of smooth volatility surfaces (and cubes) from noisy observations of put and call option prices; construction of implied correlation curves from observed synthetic CDO prices; and maturity interpolation of short-dated quadratic variation in FX options markets; just to name a few. Many of these applications require some element of shape-preservation in the resulting curves and surfaces–monotonicity and non-negative convexity of call prices in strike,[19] monotonicity of quadratic variation in maturity, monotonicity of expected portfolio credit losses in maturity, monotonicity and non-positive convexity of synthetic equity CDO tranche prices in detachment point, and so forth—which makes application of standard cubic splines fraught with danger. On the other hand, tension splines with their ability to eliminate wiggles and preserve shape are much more suitable for curve construction in these situations. Indeed, as we mentioned earlier, there are well-established techniques and algorithms to ensure shape preservation for all the types of tension splines we discussed in this article. Kvasov (2000), Chapter 8, contains references and a sample algorithm with automatic tension parameter selection; his algorithm preserves monotonicity and convexity of input data and can accept tolerances (a bid-offer, say) on the fitting precision. Kvasov (2000) also discusses techniques to build splines of more than one variable–so-called *tensor product splines*—as needed to construct surfaces and cubes, rather than curves.

While we do not want to overstate the case for tension splines—their effectiveness in outright probability density estimation, for instance, is largely unknown and

[19] From Breeden and Litzenberger (1978) it is known the risk-neutral density of an asset is proportional to the second derivative of call prices with respect to strike, hence the need for non-negative convexity. Note that the density implied by a tension spline fitted directly to call option prices (rather than to some transformation or, say, to derivatives of the call options with respect to strike) would depend on the type of tension spline used, but would range from a completely discrete distribution (infinite tension) to a piecewise linear function (zero tension).

most likely outperformed by special-purpose kernel methods (but see Monteiro et al. 2007)—it is safe to say that elements of the theory outlined in this paper could find a number of relevant applications elsewhere in finance. Needless to say, these are left for future work.

Acknowledgements Helpful comments from Tom Lyche, Robert Renka and, especially, Carsten Tanggaard are gratefully acknowledged.

A Appendix: norm minimization

We are interested in establishing

$$\hat{\varphi} = \arg\min_{\varphi \in \mathscr{A}} \frac{1}{N} \sum_{i=1}^{N} W_i^2 \left(V_i - \sum_{j=1}^{M} c_{ij} P\left(t_j, \varphi(t_j)\right) \right)^2 + \lambda \left(\int_{t_1}^{t_M} \left[\varphi''(t)^2 + \sigma^2 \varphi'(t)^2 \right] dt \right) \tag{A-1}$$

where λ and σ^2 are parameters.

For any $\mu \in \mathbb{R}$ and for any $\upsilon \in C^2[t_1, t_M]$, define

$$F(\varphi, \mu, \upsilon) = \frac{1}{N} \sum_{i=1}^{N} W_i^2 \left(V_i - \sum_{j=1}^{M} c_{ij} P\left(t_j, \varphi(t_j) + \mu \upsilon(t_j)\right) \right)^2 + \lambda \int_{t_1}^{t_M} \left[\left(\varphi''(t) + \mu \upsilon''(t)\right)^2 + \sigma^2 \left(\varphi'(t) + \mu \upsilon'(t)\right)^2 \right] dt.$$

For a minimizer $\hat{\varphi}$ of (A-1), we must have, for any υ, $F(\hat{\varphi}, \mu, \upsilon) \geq F(\hat{\varphi}, 0, \upsilon)$. By standard variational calculus,

$$\left. \frac{\partial F(\hat{\varphi}, \mu, \upsilon)}{\partial \mu} \right|_{\mu=0} = 0 \tag{A-2}$$

is the condition for a local optimum. One may recognize the left-hand side as the Gateaux variation in direction υ.

Given our definition of F, (A-2) can be written as

$$-\frac{2}{N} \sum_{j=1}^{M} \sum_{i=1}^{N} \upsilon(t_j) c_{ij} P'\left(t_j, \hat{\varphi}(t_j)\right) W_i^2 \left(V_i - \sum_{k=1}^{M} c_{ik} P\left(t_k, \hat{\varphi}(t_k)\right) \right) + 2\lambda \int_{t_1}^{t_M} \left(\hat{\varphi}''(t) \upsilon''(t) + \sigma^2 \hat{\varphi}'(t) \upsilon'(t) \right) dt = 0$$

where

$$P'\left(t_j, \hat{\varphi}(t_j)\right) \equiv \left.\frac{\partial P\left(t_j, \hat{\varphi}(t_j) + x\right)}{\partial x}\right|_{x=0}.$$

That is,

$$\frac{1}{N\lambda}\sum_{j=1}^{M}\sum_{i=1}^{N} \upsilon(t_j) c_{ij} P'\left(t_j, \hat{\varphi}(t_j)\right) W_i^2 \left(V_i - \sum_{k=1}^{M} c_{ik} P\left(t_k, \hat{\varphi}(t_k)\right)\right)$$
$$= \int_{t_1}^{t_M} \left(\hat{\varphi}''(t)\upsilon''(t) + \sigma^2 \hat{\varphi}'(t)\upsilon'(t)\right) dt, \tag{A-3}$$

for any function $\upsilon \in C^2[t_1, t_M]$. We notice that the left-hand side of (A-3) only depends on υ in the knots t_j, so the right-hand side must be of this form as well. This naturally leads us to guess that $\hat{\varphi}(t)$ is a hyperbolic tension spline, as in this case $\hat{\varphi}''(t) - \sigma^2\hat{\varphi}(t)$ is linear on each interval $[t_j, t_{j+1}]$. Specifically, with constants d_j, $j = 1, \ldots, M-1$, defined as

$$d_j = \hat{\varphi}^{(3)}(t) - \sigma^2\hat{\varphi}'(t), \quad t \in (t_j, t_{j+1}),$$

integration by parts (as in Sect. 3.5) shows that

$$\int_{t_1}^{t_M} \left(\hat{\varphi}''(t)\upsilon''(t) + \sigma^2 \hat{\varphi}'(t)\upsilon'(t)\right) dt = \hat{\varphi}''(t_M)\upsilon'(t_M) - \hat{\varphi}''(t_1)\upsilon'(t_1)$$
$$+ \sum_{j=1}^{M} \upsilon(t_j)(d_j - d_{j-1}) \tag{A-4}$$

with the convention $d_0 = d_M = 0$. If the tension spline has natural boundary conditions, the terms $\hat{\varphi}''(t_M)\upsilon'(t_M)$ and $\hat{\varphi}''(t_1)\upsilon'(t_1)$ vanish, and we can write our first-order condition (A-3) as

$$\frac{1}{N\lambda}\sum_{i=1}^{N} c_{ij} P'\left(t_j, \hat{\varphi}(t_j)\right) W_i^2 \left(V_i - \sum_{k=1}^{M} c_{ik} P\left(t_k, \hat{\varphi}(t_k)\right)\right) = d_j - d_{j-1}, \tag{A-5}$$

to hold for all $j = 1, \ldots, M$. We note that the same expression will hold if instead of enforcing natural boundary conditions, we prescribe first-order derivative conditions $\varphi'(t_1) = a$, $\varphi'(t_M) = b$ and optimize in the set $\mathscr{A} = \{\varphi \in C^2[t_1, t_M] : \varphi'(t_1) = a, \varphi'(t_M) = b\}$. For $\varphi + \mu\upsilon$ to stay inside $\mathscr{A}$, we necessarily must have $\upsilon'(t_1) = \upsilon'(t_M) = 0$, which again ensures that the terms $\hat{\varphi}''(t_M)\upsilon'(t_M)$ and $\hat{\varphi}''(t_1)\upsilon'(t_1)$ vanish in (A-4).

Equation (A-5) constitutes a necessary condition only and, in principle, it should be tested whether convexity properties are such that we indeed have found a minimum.

This can easily be done for linear functions, but can be quite onerous to do for general specifications of $P\left(t_k, \varphi(t_k)\right)$.

B Appendix: forward curve overlays

Many of the curve construction algorithms so far have been designed around the implicit idea that the forward curve should ideally be *smooth*. While this is generally a sound principle, exceptions do exist. For instance, it may be reasonable to expect instantaneous forward rates to jump on or around meetings of monetary authorities, such as the Federal Reserve in the US. In additional, other "special" situations may exist that might warrant introduction of discontinuities into the forward curve. A well-known example is the turn-of-year effect where short-dated loan premiums spike for loans between the last business day of the year and the first business day of the next year.

One possible way of incorporating forward rate jumps of known magnitude into the discount curve machinery is to exogenously specify an *overlay curve* $\epsilon_f(t)$. The forward curve $f(t)$ is then written as

$$f(t) = \epsilon_f(t) + f^*(t), \tag{A-6}$$

where $\epsilon_f(t)$ is user-specified—and most likely contains discontinuities around special events dates—and $f^*(t)$ is unknown. The discount curve algorithm is then subsequently applied to the construction of $f^*(t)$. That is, rather than solving $\mathbf{cP} = \mathbf{V}$ (see Eq. 3), we instead write

$$P(t) = e^{-\int_0^T \epsilon_f(t)dt} e^{-\int_0^T f^*(t)dt} \equiv P_\epsilon(t) P^*(t) \tag{A-7}$$

and solve

$$\mathbf{c}_\epsilon \mathbf{P}^* = \mathbf{V}, \tag{A-8}$$

where $\mathbf{P}^* = \left(P^*(t_1), \ldots, P^*(t_M)\right)^\top$ and $\mathbf{c}_\epsilon$ is a modified $N \times M$ coupon matrix, with elements

$$(c_\epsilon)_{ij} = c_{ij} P_\epsilon(t_j). \tag{A-9}$$

Construction of $\mathbf{c}_\epsilon$ can be done as a pre-processing step, after which any of the algorithms discussed earlier in this paper can be applied to attack (A-8). Once the curve $P^*(t)$ (or, equivalently, some transformation such as the yield curve $y^*(t) = -t^{-1} \ln P^*(t)$) has been constructed, any subsequent use of the curve for cash-flow discounting requires, according to (A-7), a multiplicative adjustment of time t discount factors by the quantity $P_\epsilon(t)$.

C Appendix: noisy or Ill-posed benchmark data

To, very briefly, exemplify the use of our curve construction algorithm on noisy and/or ill-posed benchmark sets, we construct a little representative experiment. Specifically, we start out with a smooth discount curve, generated by a suitably parametrized version of the function in (Nelson and Siegel 1987); see the "Pre-Noise Data" graph in Fig. 9 below for the curve used. We then pick 10 \$1-notional coupon bonds with maturities 1,2, . . . ,10 years with coupons set such that all bonds price at par (i.e. \$1). We double the set of securities to generate 20 bonds—two identical 1-year bonds, two identical 2-year bonds, and so forth. Finally, we create our benchmark set by adding to all 20 coupons random zero-mean Gaussian noise with a standard deviation of 25 basis points (0.25%); the prices of all bonds are kept at \$1.

It should be clear that the resulting benchmark set is both noisy and severely ill-posed; in particular, notice that all securities in the set have a "twin" that has the same payment schedule and price, yet pays a different coupon. In noisy illiquid markets, bond prices may exhibit similar characteristics (albeit rarely as blatantly as our example, of course). Figure 8 below shows the L-curve (see Sect. 4.4.3) for our benchmark set; for comparison we have also shown the L-curve for the (well-posed) benchmark set used in the Libor curve example in Sect. 6. Notice the fundamental difference in the shape of the two curves: for the ill-posed set, when λ is lowered beyond a certain threshold, the smoothness penalty term increases dramatically in magnitude, as the yield curve starts getting increasingly irregular.

Applying our curve construction algorithm to the yield curve $y(t)$, Fig. 9 below show the curves that arise when the penalty multiplier λ in the norm (27) is auto-selected according to several of the methods in Sect. 4.4. The L-curve method here undersmoothes the data, whereas the Morozov discrepancy principle oversmoothes. The GCV method also seems to undersmoothe the data somewhat, but overall appears to do the best job for our specific example. We remark that Tanggaard (1997) finds that the GCV criterion has a tendency to work better when the discount curve algorithm is applied to the function $w(t)$ in Sect. 2.2, rather than $y(t)$.

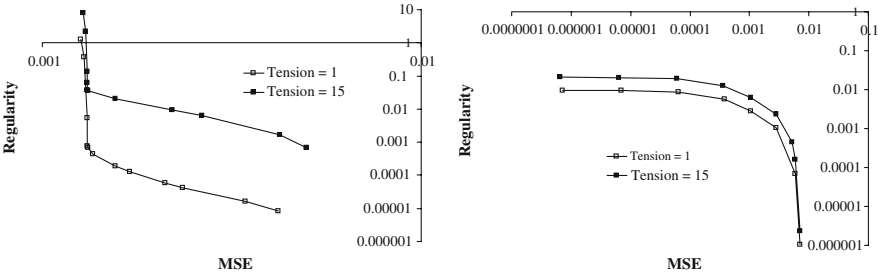

Fig. 8 L-curves. *Notes*: The figure shows the L-curves for our noisy bond data (left panel) and for the swap data in Table 1 (right panel). On the graph axes, "MSE" represents the mean-square-error $\mathscr{I}_{ls}$ in the regularity norm (27), and "Regularity" represents the penalty term $\mathscr{I}_{reg}$. The tension parameter σ was set as indicated in the graphs

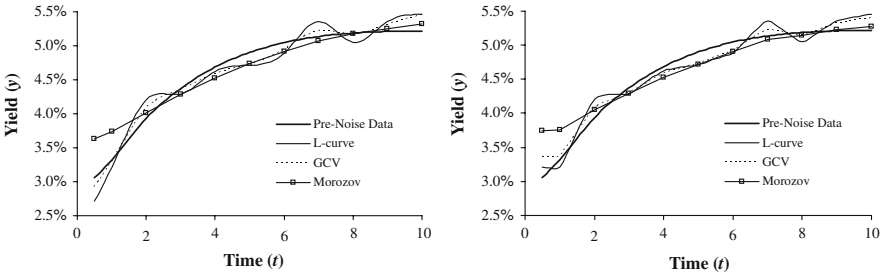

Fig. 9 Yield curves. *Notes*: Yield curves constructed with various method of auto-selecting the regularity weight λ. For the Morozov approach we use the Rice (1984) estimator to find the noise standard deviation; the estimator returned a standard deviation of 27.9 basis points. On the left panel, the tension parameter is $\sigma = 1$; on the right panel, the tension parameter is $\sigma = 15$

References

Adams, K. (2001). Smooth interpolation of yield curves. *Algo Research Quarterly, 4*, 11–22.

Adams, K., & van Deventer, D. R. (1994). Fitting yield curves and forward rate curves with maximum smoothness. *Journal of Fixed Income, 4*, 52–62.

Andersen, L. (2003). Reduced-form models: Curve construction and the pricing of credit swaps. In: *Credit Derivatives: The Definitive Guide*. Risk Books.

Barzanti, L., & Corradi, C. (1998). A note on interest rate term structure estimation using tension splines. *Insurance: Mathematics and Economics, 22*, 139–143.

Catmull, E., & Rom, R. (1974). A class of local interpolating spline. In R. E. Barnhill & Riesenfled (Eds.), *Computer Aided Geometric Design*. New York: Academic Press.

Chambers, D. R., Carleton, W. T., & Waldman, D. W. (1984). A new approach to estimation of the term structure of interest rates. *Journal of Financial and Quantitative Analysis, 19*, 233–269.

Cline, A. K. (1974). Scalar- and planar-valued curve fitting using splines under tension. *Communications of the ACM, 17*, 218–223.

Cox, J., Ingersoll, J., & Ross, S. (1985). A theory of the term structure of interest rates. *Econometrica, 53*, 385–407.

Craven, P., & Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Matematik, 31*, 377–403.

de Boor, C. (1978). *A practical guide to splines*. New York: Springer Verlag.

Diament, P. (1993). Semi-empirical smooth fit to the Treasury Yield Curve. *Journal of Fixed Income, 3*, 55–70.

Eubank, R. L. (1988). *Spline smoothing and nonparametric regression*. Marcel Dekker.

Golub, B., & Tilman, L. (2000). No room for nostalgia in fixed income. Risk Magazine, July, 44–48.

Hagan, P., & West, G. (2004). Interpolation methods for yield curve construction. Working Paper.

Hansen, P. (1992). Analysis of discrete ill-posed problems by means of the L-curve. *SIAM Rev., 34*, 561–580.

Heath, D., Jarrow, R., & Morton, A. (1992). Bond pricing and the term structure of interest rates: A new methodology for contingent claim valuation. *Econometrica, 66*, 77–105.

Hull, J. (2000). *Futures, options, and other derivative securities* (4th ed.). New Jersey: Prentice Hall.

Hull, J., & White, A. (1990). Pricing interest rate derivative securities. *Review of Financial Studies, 3*(4), 87–100.

Hyman, J. M. (1983). Accurate monotonicity preserving cubic interpolation. *SIAM Journal of Statistics and Computation, 4*, 645–654.

Jamishidian, F. (1991). Bond and option evaluation in the Gaussian interest rate model. *Research in Finance, 9*, 131–170.

Koch, P. E., & Lyche, T. (1989). Exponential B-splines in tension. In C. K. Chui et al. (Eds.), *Approximation Theory VI: Proceedings of the Sixth International Symposium on Approximation Theory* (Vol. II, pp. 361–364). Boston: Academic Press.

Koch, P. E., & Lyche, T. (1993). Interpolation with exponential splines in tension. In G. Farin et al. (Eds.), *Geometric Modeling, Computing/Supplementum 8* (pp. 173–190). Wien: Springer Verlag.

Kvasov, B. (2000). *Methods of shape-preserving spline approximation*. Singapore: World Scientific Publishing.
Breeden, D., & Litzenberger, R. (1978). Prices of state-contingent claims implicit in option prices. *Journal of Business, 51*, 621–651.
Lynch, R. W. (1982). A method for choosing a tension factor for spline under tension interpolation. Working Paper, University of Texas.
McCulloch, J. H. (1975). The tax-adjusted yield curve. *Journal of Finance, 30*, 811–830.
McCulloch, J. H., & Kochin, L. A. (2000). The inflation premium implicit in the US real and nominal term structures of interest rates. Technical Report 12, Ohio State University.
Monteiro, A., Tutuncu, R., & Vicente, L. (2007). Recovering risk-neutral probability density functions from options prices using cubic Splines and ensuring nonnegativity. Working Paper.
Morozov, V. (1966). On the solution of functional equations by the method of regularization. *Soviet Mathematics Doklayh, 7*, 414–417.
Nelson, C. R., & Siegel, A. F. (1987). Parsimonious modeling of yield curves. *Journal of Business, 60*, 473–489.
Press, W., Teukolsky, S., Vetterling, W., & Flannery, B. (1992). *Numerical Recipes in C*. Cambridge University Press.
Pruess, S. (1976). Properties of splines in tension. *Journal of Approximation Theory, 17*, 86–96.
Renka, R. J. (1987). Interpolatory tension splines with automatic selection of tension factors. *SIAM Journal of Scientific and Statistical Computing, 8*(3), 393–415.
Rentrop, P. (1980). An algorithm for the computation of exponential splines. *Numerische Matematik, 35*, 81–93.
Rice, J. A. (1984). Bandwidth choice for nonparametric regression. *Annals of Statistics, 12*, 1215–1230.
Schweikert, D. G. (1966). An interpolating curve using a spline in tension. *Journal of Mathematics and Physics, 45*, 312–317.
Sheah, G. S. (1984). Pitfalls in smoothing interest rate terms structure data: equilibrium models and spline approximation. *Journal of Financial and Quantitative Analysis, 19*, 253–269.
Svensson, L. (1994). Estimating and interpreting forward interest rates: Sweden 1992–1994. CEPR Discussion Paper 1051 (October).
Tanggaard, C. (1997). Nonparametric smoothing of yield curves. *Review of Quantitative Finance and Accounting, 9*, 251–267.