THE MONOTONE CONVEX METHOD OF INTERPOLATION

GRAEME WEST, FINANCIAL MODELLING AGENCY

Contents

This method of interpolation was introduced in Hagan and West [2006], Hagan and West [2008].

1. Constructing yield curves

1.1. **Curve fitting.** There is a need to value all instruments consistently within a single valuation framework. For this we need a risk free yield curve which will be a continuous zero curve (because this is the standard format, for all option pricing formulae). Thus, a yield curve is a function $r = r(\tau)$, where a single payment investment for time τ will earn a continuous rate $r = r(\tau)$, that is, a payment of 1 at initiation will be redeemed by a payment of $\exp(r(\tau)\tau)$ at time τ.

As explained in Zangari [1997], Lin [2002] term structure estimation methods can be classified into two groups: theoretical and empirical. Theoretical term structure methods typically posit an explicit structure for a variable known as the short rate of interest, whose value depends on a set of parameters that might be determined using statistical analysis of market variables. Early examples of theoretical methods include Vasicek [1977] and Cox et al. [1985]. From such a method the yield curve can be derived. Because the theoretical method is parsimonious, the yield curve will fall into one of a few basic categories in terms of shape. In some circumstances, negative rates are possible.

Date: May 7, 2011.

Empirical methods are available to compute spot interest rates. Unlike the theoretical methods, the empirical methods are independent of any model or theory of the term structure. Whereas the theoretical methods attempt to explain typical features of the term structure, which may include how the term structure evolves through time, the empirical methods merely try to find a close representation of the term structure at any point in time, given some observed interest rate data.

Later developments, in particular the approach of Hull and White [1990], allowed the use of a empirically determined yield curve in a theoretical model. Furthermore, the classification scheme of Heath et al. [1990] takes as input the same empirically determined yield curve. Thus, while the practitioner has several choices for the theoretical model that will govern their evolution of the yield curve, and hence govern their pricing of derivative products, they will almost certainly have as starting point an empirically determined yield curve. This document is concerned with that task of determining the yield curve, a process typically called bootstrapping. In fact, our treatment is slightly more general, as it covers the construction of spread curves, forward curves, etc. as well.

As explained in several sources, for example Ron [2000], there is no single correct way to complete the term structure of a yield curve from a set of rates. It is desired that the derived yield curve should be smooth, but there must not be over-smoothing, as this might cause the elimination of valuable market pricing information.

It may or may not be a criterion that all inputs to the yield curve should price back exactly after the construction of the curve. If it is not a criterion, then almost surely one will favour the approaches of Nelson and Siegel [1987] and Svensson [1994].

We consider the situation where it is required to price exactly. Certainly this approach is completely feasible when bootstrapping a swap curve, it may or may not be feasible when bootstrapping a bond curve, this will depend on the number of liquid bonds available in the market. Even when we require that the curve perfectly replicates the price of the input instruments, the yield curve is not constructed uniquely; we need to select an interpolation method with which to build the curve.

1.2. **The yield curve.** Much of what is said here is a reprise of the excellent introduction in [Rebonato, 1998, §1.2].

Time is measured in years, with the current time typically being denoted t.

We have two basic functions: the capitalisation function and the discount function. $C(t,T)$ and called the capitalisation factor: it is the redemption amount earned at time T from an investment at time t of 1 unit of currency. Of course $C(t,T) > 1$ for $T > t$ as the owner of funds charges a fee, known as interest, for the usage of their funds by the counterparty.

The price of an instrument which pays 1 unit of currency at time T - this is called a discount or zero coupon bond - is denoted $Z(t,T)$. This is the present value function. A fundamental result in Mathematics of Finance is (intuitively) that the value of any instrument is the present value of its expected cash flows. So the present value function is important.

We say that 1 grows to $C(t,T)$ and $Z(t,T)$ grows to 1. If $Z(t,T)$ grows to 1, then $Z(t,T)C(t,T)$ grows to $C(t,T)$, and so

$$Z(t,T)C(t,T) = 1. \tag{1}$$

Note also that $Z(t,t) = 1 = C(t,t)$. The next most obvious fact is that $Z(t,T)$ is decreasing in T (equivalently, $C(t,T)$ is increasing).

Suppose $Z(t,T_1) < Z(t,T_2)$ for some $T_1 < T_2$. Then the arbitrageur will buy a zero coupon bond for time T_1, and sell one for time T_2, for an immediate income of $Z(t,T_2) - Z(t,T_1) > t$. At time T_1 they will receive 1 unit of currency from the bond they have bought, which they could keep under their bed for all we care until time T_2, when they deliver 1 in the bond they have sold.

FIGURE 1. The arbitrage argument that shows that $Z(t,T)$ must be decreasing.

What we have said so far assumes that such bonds do trade, with sufficient liquidity, and as a continuum i.e. a zero coupon bond exists for every redemption date T. In fact, such bonds rarely trade in the market. Rather what we need to do is impute such a continuum via a process known as bootstrapping.

The term structure of interest rates is defined as the relationship between the yield-to-maturity on a zero coupon bond and the bond's maturity. If we are going to price derivatives which have been modelled in continuous-time off of the curve, it makes sense to commit ourselves to using continuously-compounded rates from the outset. The time t continuously compounded risk free rate for maturity T, denoted $r(t,T)$, is given by the relationships

$$C(t,T) = \exp(r(t,T)(T-t)) \tag{2}$$

$$r(t,T) = \frac{1}{T-t}\ln C(t,T) \tag{3}$$

$$Z(t,T) = \exp(-r(t,T)(T-t)) \tag{4}$$

$$r(t,T) = -\frac{1}{T-t}\ln Z(t,T) \tag{5}$$

The rates will be known, or derived from a gentle model, for a set of times $t_1,\ t_2,\ \ldots,\ t_n$; let us abbreviate these rates as $r_i = r(t_i)$ for $1 \le i \le n$. Suppose that the rates $r_1,\ r_2,\ \ldots,\ r_n$ are known at the ordered times $t_1,\ t_2,\ \ldots,\ t_n$. Any interpolation method of the yield curve function $r(t)$ will construct a continuous function $r(t)$ satisfying $r(t_i) = r_i$ for $i = 1,\ 2,\ \ldots,\ n$. Various interpolation methods are reviewed, and a couple of new ones are introduced, in Hagan and West [2006], Hagan and West [2008]. In so-called normal markets, yield curves are upwardly sloping, with longer term interest rates being higher than short term. A yield curve which is downward sloping is called inverted. A yield curve with one or more turning points is called mixed. It is often stated that such mixed yield curves are signs of market illiquidity or instability. This is not the case. Supply and demand for the instruments that are used to bootstrap the curve may simply imply such shapes. One can, in a stable market with reasonable liquidity, observe a consistent mixed shape over long periods of time.

1.3. **The shape of the curve.** The shape of the graph for $Z(0,t)$ does not reflect the shape of the yield curve in any obvious way. As already mentioned, the discount factor curve must

be monotonically decreasing whether the yield curve is normal, mixed or inverted. Nevertheless, many bootstrapping and interpolation algorithms for constructing yield curves miss this absolutely fundamental point.

Of the well known methods, only raw (linear rt) takes this point into consideration, almost by construction. Any variation of Hermite interpolation - this is very popular - miss this point. The monotone convex methods introduced in Hagan and West [2006], Hagan and West [2008] take this point into account explicitly. In Figure 3 we find some rather odd looking curves; these curves

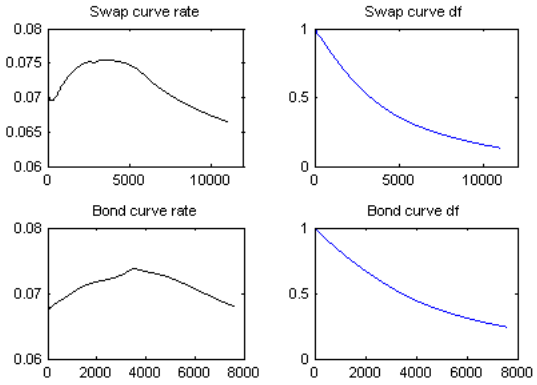


FIGURE 2. Some yield curves and their discount functions

were found in the mark to market system of a bank. In the SB curve, the bootstrap has failed to guarantee that the $Z(0,t)$ function is decreasing.

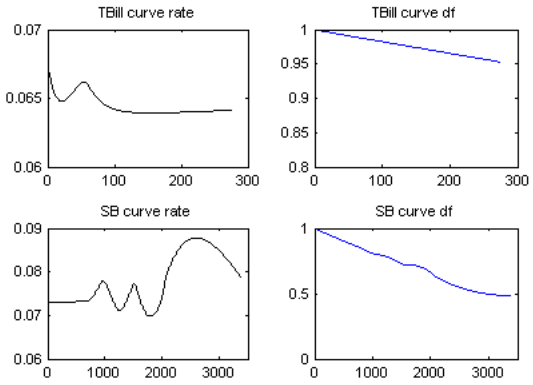


FIGURE 3. Some more yield curves and their discount functions: the second is illegal, because the $Z(\cdot)$ curve is not decreasing.

Interestingly, there will be at least one class of yield curve where the above argument for a decreasing Z function does not hold true - a real (inflation linked) curve. Because the actual size of the cash payments that will occur are unknown (as they are determined by the evolution of a price index, which is unknown) the usual arbitrage argument does not hold. Thus, for a real curve the Z function is not necessarily decreasing (and empirically this phenomenon does on occasion occur).

1.4. **Instantaneous Forward rates.** Using continuous rates, the forward rate governing the period from t_1 to t_2, denoted $f(0;t_1,t_2)$ satisfies

$$\exp(-f(0;t_1,t_2)(t_2-t_1)) = Z(0;t_1,t_2) := \frac{Z(0,t_2)}{Z(0,t_1)}$$

Immediately, we see that forward rates are positive (this is equivalent to the discount function decreasing). We have either of

$$f(0;t_1,t_2) = -\frac{\ln(Z(0,t_2)) - \ln(Z(0,t_1))}{t_2-t_1} \tag{6}$$

$$= \frac{r_2t_2 - r_1t_1}{t_2-t_1} \tag{7}$$

Let the instantaneous forward rate for a tenor of t be denoted $f(t)$, that is, $f(t) = \lim_{\epsilon\downarrow 0} f(0;t,t+\epsilon)$, for whichever t this limit exists. Clearly then

$$f(t) = -\frac{d}{dt}\ln(Z(t)) \tag{8}$$

$$= \frac{d}{dt}r(t)t \tag{9}$$

So $f(t) = r(t) + r'(t)t$, so the forward rates will lie above the yield curve when the yield curve is normal, and below the yield curve when it is inverted. By integrating,[1]

$$r(t)t = \int_0^t f(s)\,ds \tag{10}$$

$$= r(t_{i-1})t_{i-1} + \int_{t_{i-1}}^t f(s)\,ds \tag{11}$$

$$Z(t) = \exp\left(-\int_0^t f(s)\,ds\right) \tag{12}$$

Also

$$f_i^d := \frac{r_it_i - r_{i-1}t_{i-1}}{t_i - t_{i-1}} = \frac{1}{t_i-t_{i-1}}\int_{t_{i-1}}^{t_i} f(s)\,ds \tag{13}$$

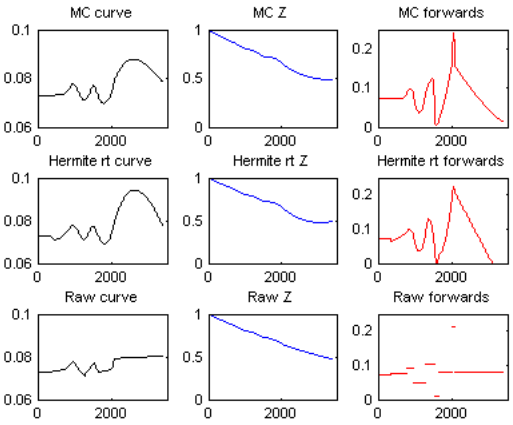
which shows that the average of the instantaneous forward rate over any interval $[t_{i-1},t_i]$ is equal to the discrete forward rate for that interval. Also, if we have a functional form for f, presumably we then have a functional form for the integral of f, and hence for r. This will be the approach taken in the monotone convex algorithm.

The discount factor curve being decreasing is equivalent to instantaneous forwards being positive. So it is required that forwards be positive to avoid arbitrage, but we should also whenever possible ask for continuity of instantaneous forwards. The pricing of interest sensitive instruments is sensitive to the stability of forward rates. As pointed out in McCulloch and Kochin [2000], 'a discontinuous forward curve implies either implausible expectations about future short-term interest rates, or implausible expectations about holding period returns'.

Of course, with the inclusion of OISs in the bootstrap, the forward curve could (or should) have discontinuities at known MPC announcement dates.

Let us return to that rather nasty looking SB curve, and compare the bootstrap under several interpolation methods.

[1] We have $r(s)s + C = \int f(s)\,ds$, so $r(t)t = [r(s)s]_0^t = \int_0^t f(s)\,ds$.



The forwards graphed are the one day continuous forwards i.e. $T_2 = T_1 + 1d$, to all intents and purposes this is the same as the instantaneous forwards.

The Hermite rt method allows for a Z curve which is increasing in places; this is equivalent to the existence of negative forward rates. Raw interpolation is a very simple and very reliable method which posits piecewise constant forward rates. The MC method is an attempted blend of the best features: forward rates are positive, but they are also continuous.

2. Interpolation and bootstrap of yield curves - not separate processes

As has been mentioned, many interpolation methods for curve construction are available. What needs to be stressed is that in the case of bootstrapping yield curves, the interpolation method is intimately connected to the bootstrap, as the bootstrap proceeds with incomplete information. This information is 'completed' (in a non unique way) using the interpolation scheme.

It is possible to develop iterative schemes for bootstrapping yield curves given some market information. This method completes information using an iterative process. The yield curve is found as a fixed point of this iterative process; the iteration certainly has a fixed point as we may invoke a theorem of mathematics such as Schauder's fixed point theorem.

Using this fixed point algorithm is far superior to using something like a multi-dimensional Newton's method. The computation burden is almost trivial and the time taken to converge is at least an order of magnitude less.

The method was illustrated for swap curves in Hagan and West [2006] and for bond curves in Hagan and West [2008], in both instances in Sections entitled 'Interpolation and Bootstrap of Yield Curves not two Separate Processes'. We have implemented this technique for other curves such as inflation curves and OIS curves.

3. The Monotone Convex Method of Interpolation

3.1. **The monotone convex algorithm.** Many of the ideas of the method of Hyman [1983] have a natural development with the introduction of the monotone convex method. This method was developed to resolve the only remaining financial deficiency of Hyman [1983]: very simply, none of the methods commonly in use are aware that they are trying to solve a financial problem - indeed,

the breeding ground for these methods is typically engineering or physics. As such, there is no mechanism which ensures that the forward rates generated by the method are positive (equivalently, that the discount factor curve is decreasing), and some simple experimentation will uncover a set of inputs to a yield curve which give some negative forward rates under all of the methods mentioned here, as seen in Hagan and West [2006]. Thus, in introducing the monotone convex method, we use the ideas of Hyman [1983], but explicitly ensure that the continuous forward rates are positive (whenever the discrete forward rates are themselves positive).

The point of view taken in the monotone convex method is that the inputs are (or can be manipulated to be) discrete forwards belonging to intervals; the interpolation is not performed on the interest rate curve itself. We may have actual discrete forwards - FRA rates. On the other hand if we have interest rates $r_1,\ r_2,\ \ldots,\ r_n$ for periods $\tau_1,\ \tau_2,\ \ldots,\ \tau_n$ then the first thing we do is calculate

$$f_i^d = \frac{r_i\tau_i - r_{i-1}\tau_{i-1}}{\tau_i - \tau_{i-1}} \quad 1 \leq i \leq n \tag{14}$$

Here we also check that these are all positive, and so conclude that the curve is legal i.e. arbitrage free - except in those few cases where forward rates may be negative. As an interpolation algorithm the monotone convex method will now bootstrap a forward curve, and then if required recover the continuum of risk free rates using

$$r(\tau)\tau = \int_0^\tau f(s)\,ds \tag{15}$$

One rather simple observation is that all of the spline methods - quadratic, cubic or quartic splines - fail in forward extrapolation beyond the interval $[\tau_1, \tau_n]$. Clearly if the interpolation is on rates then we will apply horizonal extrapolation to the rate outside of that interval: $r(\tau) = r_1$ for $\tau < \tau_1$ and $r(\tau) = r_n$ for $\tau > \tau_n$. So far so good. What happens to the forward rates? Perhaps surprisingly we cannot apply the same extrapolation rule to the forwards, in fact, we now need to set $f(\tau) = r_1$ for $\tau < \tau_1$ and $f(\tau) = r_n$ for $\tau > \tau_n$ - consider (9). This makes it almost certain that the forward curve has a material discontinuity at τ_1, and probably one at τ_n too (the latter will be less severe as the curve, either by design or by nature, probably has a horizontal asymptote as $\tau \uparrow \tau_n$). But this problem is only properly resolved if the forward curve has a horizonal asymptote at τ_n.

In order to avoid this pathology, we now have terms $0 = \tau_0,\ \tau_1,\ \ldots,\ \tau_n$ and the generic interval for consideration is $[\tau_{i-1}, \tau_i]$. A 'short rate' (instantaneous) rate may be provided, if not, the algorithm will model one. Usually the shortest rate that might be input will be an overnight rate, if it is provided, the algorithm here simply has some 'overkill' - there will be an overnight rate and an instantaneous short rate - but it need not be modified.

f_i^d is the discrete rate which 'belongs' to the entire interval $[\tau_{i-1}, \tau_i]$; it would be a mistake to model that rate as being the instantaneous rate at τ_i. Rather, we begin by assigning it to the midpoint of the interval, and then modelling the instantaneous rate at τ_i as being on the straight line that joins the adjacent midpoints. Let this be denoted f_i. This explains (16). (17) and (18) will ensure, with

the functional form we choose, that we have $f'(0) = 0 = f'(\tau_n)$.[2]

$$f_i = \frac{\tau_i - \tau_{i-1}}{\tau_{i+1} - \tau_{i-1}} f^d_{i+1} + \frac{\tau_{i+1} - \tau_i}{\tau_{i+1} - \tau_{i-1}} f^d_i, \qquad \text{for } i = 1,\ 2,\ \ldots,\ n-1 \tag{16}$$

$$f_0 = f^d_1 - \tfrac{1}{2}(f_1 - f^d_1) \tag{17}$$

$$f_n = f^d_n - \tfrac{1}{2}(f_{n-1} - f^d_n) \tag{18}$$

Note that if the discrete forward rates are positive then so are the f_i for $i = 1,\ 2,\ \ldots,\ n-1$.

We now seek an interpolating function f, which will be the instantaneous forward curve, defined on $[0, \tau_n]$ that has values $f(\tau_i) = f_i$ for $i = 0,\ 1,\ \ldots,\ n$ that satisfies the following conditions (in some sense, they are arranged in decreasing order of necessity):

(i) $\dfrac{1}{\tau_i - \tau_{i-1}} \displaystyle\int_{\tau_{i-1}}^{\tau_i} f(t)dt = f^d_i$, so the discrete forward is recovered by the curve.
(ii) f is positive.
(iii) f is continuous.
(iv) If $f^d_{i-1} < f^d_i < f^d_{i+1}$ then $f(\tau)$ is increasing on $[\tau_{i-1}, \tau_i]$, and if $f^d_{i-1} > f^d_i > f^d_{i+1}$ then $f(\tau)$ is decreasing on $[\tau_{i-1}, \tau_i]$.

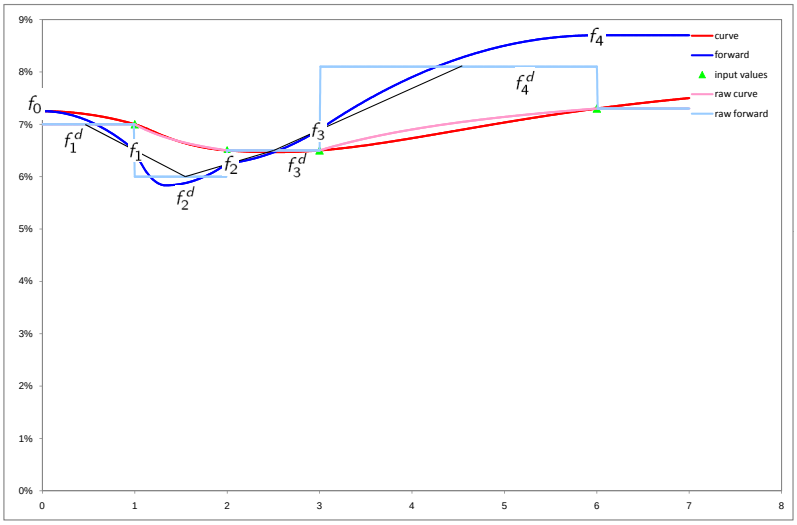


FIGURE 4. The requirements of the monotone convex algorithm

Let us first normalise things, so we seek a function g defined on $[0, 1]$ such that[3]

$$g(x) = f(\tau_{i-1} + (\tau_i - \tau_{i-1})x) - f^d_i. \tag{19}$$

[2]This is a little anticipatory. We anticipate the functional form (19) and the functional form $g(x) = K + Lx + Mx^2$. If $f'(0) = 0$ then $g'(0) = 0$ and so $L = 0$. Also, because $g'(0) = 0$ we will have $g(1) = -2g(0)$ - we will see this later - and so $M = -3K$. Now suppose $f_0 = f^d_1 - \lambda(f_1 - f^d_1)$. Then $g(0) = -\lambda(f_1 - f^d_1)$ and $g(1) = f_1 - f^d_1$. Thus $\lambda = \frac{1}{2}$.

[3]Strictly speaking, we are defining functions g_i, each corresponding to the interval $[\tau_{i-1}, \tau_i]$. As the g_i are constructed one at a time, we suppress the subscript.

Let us give a sketch of how we will proceed. We will choose g to be piecewise quadratic in such a way that (i) is satisfied by construction. Of course, g is continuous, so (iii) is satisfied. As a quadratic, it is easy to perform an analysis of where the minimum or maximum occurs, and we thereby are able to apply some modifications to g to ensure that (iv) is satisfied, while ensuring (i) and (iii) are still satisfied.

Also, we see a posteriori that if the values of f_i had satisfied certain constraints, then (ii) would have been satisfied. So, the algorithm will be to construct (16), (17) and (18), then modify the f_i to satisfy those constraints, then construct the quadratics, and then modify those quadratics. Penultimately,

$$f(\tau) = g\left(\frac{\tau - \tau_{i-1}}{\tau_i - \tau_{i-1}}\right) + f_i^d. \tag{20}$$

Finally, we use (10) to find the risk free rates.

Thus, the current choices of f_i are provisional; we might make some adjustments in order to guarantee the positivity of the interpolating function f.

Here follow the details. We have only three pieces of information about g: $g(0) = f_{i-1} - f_i^d$, $g(1) = f_i - f_i^d$, and $\int_0^1 g(x)\,dx = 0$. We postulate a functional form $g(x) = K + Lx + Mx^2$, having 3 equations in 3 unknowns we get $\begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & \frac{1}{2} & \frac{1}{3} \end{bmatrix} \begin{bmatrix} K \\ L \\ M \end{bmatrix} = \begin{bmatrix} g(0) \\ g(1) \\ 0 \end{bmatrix}$, and easily solve to find that

$$g(x) = g(0)[1 - 4x + 3x^2] + g(1)[-2x + 3x^2] \tag{21}$$

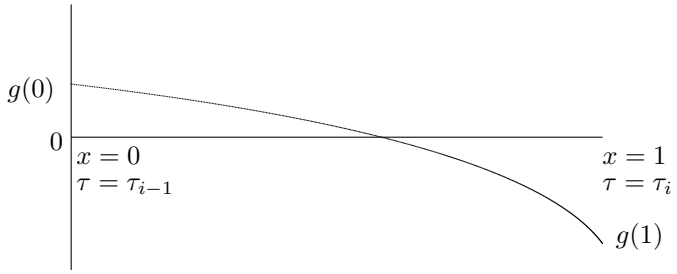


FIGURE 5. The function g

Note that by (16) that (iv) is equivalent to requiring that if $f_{i-1} < f_i^d < f_i$ then $f(\tau)$ is increasing on $[\tau_{i-1}, \tau_i]$, while if $f_{i-1} > f_i^d > f_i$ then $f(\tau)$ is decreasing on $[\tau_{i-1}, \tau_i]$. This is equivalent to requiring that if $g(0)$ and $g(1)$ are of opposite sign then g is monotone.

Now

$$\begin{aligned} g'(x) &= g(0)(-4 + 6x) + g(1)(-2 + 6x) \\ g'(0) &= -4g(0) - 2g(1) \\ g'(1) &= 2g(0) + 4g(1) \end{aligned}$$

g being a quadratic it is now easy to determine, simply by inspecting $g'(0)$ and $g'(1)$, the behaviour of g on $[0, 1]$. The cases where $g'(0) = 0$ and $g'(1) = 0$ are crucial; these correspond to $g(1) = -2g(0)$ and $g(0) = -2g(1)$ respectively. These two lines divide the $g(0)/g(1)$ plane into eight sectors. We

seek to modify the definition of g on each sector, taking care that on the boundary of any two sectors, the formulae from those two sectors actually coincide (to preserve continuity). In actual fact the treatment for every diametrically opposite pair of sectors is the same, so we really have four cases to consider, as follows (refer Figure 6):

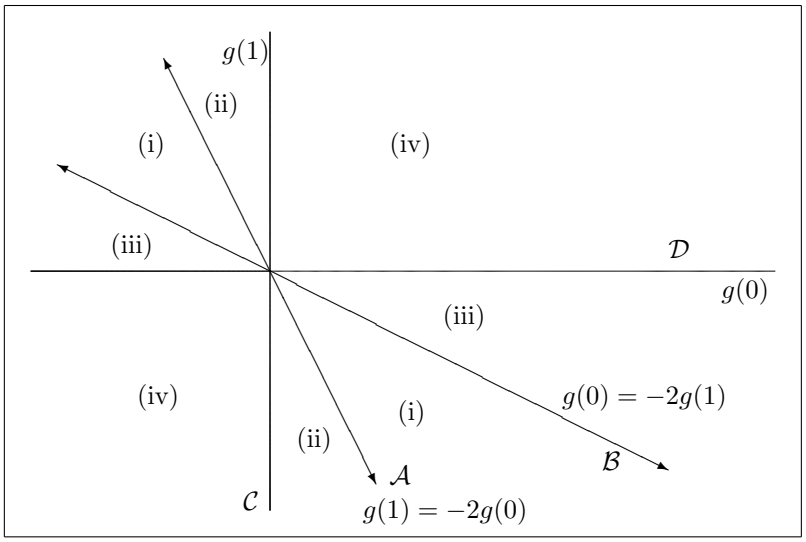


FIGURE 6. The reformulated possibilities for g

(i) In these sectors $g(0)$ and $g(1)$ are of opposite signs and $g'(0)$ and $g'(1)$ are of the same sign, so g is monotone, and does not need to be modified.

(ii) In these sectors $g(0)$ and $g(1)$ are also of opposite sign, but $g'(0)$ and $g'(1)$ are of opposite sign, so g is currently not monotone, but needs to be adjusted to be so. Furthermore, the formula for (i) and for (ii) need to agree on the boundary $\mathcal{A}$ to ensure continuity.

(iii) The situation here is the same as in the previous case. Now the formula for (i) and for (iii) need to agree on the boundary $\mathcal{B}$ to ensure continuity.

(iv) In these sectors $g(0)$ and $g(1)$ are of the same sign so at first it appears that g does not need to be modified. Unfortunately this is not the case: modification will be needed to ensure that the formula for (ii) and (iv) agree on $\mathcal{C}$ and (iii) and (iv) agree on $\mathcal{D}$.

The origin is a special case: if $g'(0) = 0 = g'(1)$ then $g(x) = 0$ for all x, and $f^d_{i-1} = f^d_i = f^d_{i+1}$, and we put $f(\tau) = f^d_i$ for $\tau \in [\tau_{i-1}, \tau_i]$.

So we proceed as follows:

(i) As already mentioned g does not need to be modified. Note that on $\mathcal{A}$ we have $g(x) = g(0)(1 - 3x^2)$ and on $\mathcal{B}$ we have $g(x) = g(0)(1 - 3x + \frac{3}{2}x^2)$.

(ii) A simple solution is to insert a flat segment, which changes to a quadratic at exactly the right moment to ensure that $\int_0^1 g(x)\,dx = 0$. So we take

$$g(x) = \begin{cases} g(0) & \text{for} \quad 0 \le x \le \eta \\ g(0) + (g(1) - g(0)) \left(\dfrac{x - \eta}{1 - \eta} \right)^2 & \text{for} \quad \eta < x \le 1 \end{cases} \tag{22}$$

$$\eta = 1 + 3\frac{g(0)}{g(1) - g(0)} = \frac{g(1) + 2g(0)}{g(1) - g(0)} \tag{23}$$

Note that $\eta \longrightarrow 0$ as $g(1) \longrightarrow -2g(0)$, so the interpolation formula reduces to $g(x) = g(0)(1 - 3x^2)$ at $\mathcal{A}$, as required.

(iii) Here again we insert a flat segment. So we take

$$g(x) = \begin{cases} g(1) + (g(0) - g(1)) \left(\frac{\eta - x}{\eta}\right)^2 & \text{for} \quad 0 < x < \eta \\ g(1) & \text{for} \quad \eta \leq x < 1 \end{cases} \tag{24}$$

$$\eta = 3\frac{g(1)}{g(1) - g(0)} \tag{25}$$

Note that $\eta \longrightarrow 1$ as $g(1) \longrightarrow -\frac{1}{2}g(0)$, so the interpolation formula reduces to $g(x) = g(0)(1 - 3x + \frac{3}{2}x^2)$ at $\mathcal{B}$, as required.

(iv) We want a formula that reduces in form to that defined in (ii) as we approach $\mathcal{C}$, and to that defined in (iii) as we approach $\mathcal{D}$. This suggests

$$g(x) = \begin{cases} A + (g(0) - A) \left(\frac{\eta - x}{\eta}\right)^2 & \text{for} \quad 0 < x < \eta \\ A + (g(1) - A) \left(\frac{x - \eta}{1 - \eta}\right)^2 & \text{for} \quad \eta < x < 1 \end{cases} \tag{26}$$

where $A = 0$ when $g(1) = 0$ - so the first line satisfies (iii)) and $A = 0$ when $g(0) = 0$ (so the second line satisfies (ii). Straightforward calculus gives

$$\int_0^1 g(x)\, dx = \frac{2}{3}A + \frac{\eta}{3}g(0) + \frac{1 - \eta}{3}g(1)$$

and so

$$A = -\tfrac{1}{2}\left[\eta g(0) + (1 - \eta)\, g(1)\right]$$

A simple choice satisfying the various requirements is

$$\eta = \frac{g(1)}{g(1) + g(0)} \tag{27}$$

$$A = -\frac{g(0)g(1)}{g(0) + g(1)} \tag{28}$$

In this case (and this case only) we might need to be concerned about the possibility that $\eta = 0$ or $\eta = 1$, else (in some languages and some implementations) there is the possibility of a divide by 0. This

- If $g(1) = 0$, then $\eta = 0 = A$, and $g(x) = 0$ except at $x = 0$.
- If $g(0) = 0$, then $\eta = 1$, $A = 0$ and $g(x) = 0$ except at $x = 1$.

3.2. **Integrating the g function.** We will have need to calculate $\int_\alpha^\beta f(s)\, ds$ for some values of α, β. Suppose $\tau_{i-1} \leq \alpha < \beta \leq \tau_i$, let g apply to the interval $[\tau_{i-1}, \tau_i]$, and suppose $G' = g$. Then using (20) we have

$$\int_\alpha^\beta f(s)\, ds = (\beta - \alpha) f_i^d + (\tau_i - \tau_{i-1}) \left[G\left(\frac{\beta - \tau_{i-1}}{\tau_i - \tau_{i-1}}\right) - G\left(\frac{\alpha - \tau_{i-1}}{\tau_i - \tau_{i-1}}\right)\right] \tag{29}$$

Our typical integral will be as in (11), so the argument of the second evaluation of G will be 0. Thus, we arrange that $G(0) = 0$. This means that each calculation of r will require only one and not two evaluations of G. The formula for G clearly depends as before on which region applies, and if g is defined piecewise then the definition of G on $[\tau, 1]$ must be modified appropriately to ensure that at η its value coincides at η with the definition on $[0, \eta]$. So we proceed as follows:

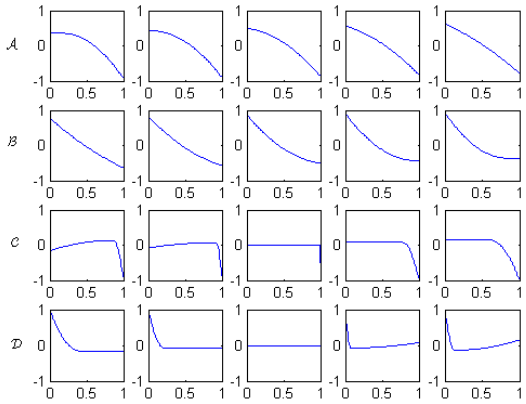


FIGURE 7. The g function as we cross the boundaries.

(i) Via (21) we have

$$G(x) = g(0)(x - 2x^2 + x^3) + g(1)(-x^2 + x^3)$$

(ii)

$$G(x) = \begin{cases} g(0)x & \text{for } 0 < x < \eta \\ g(0)x + \frac{1}{3}(g(1) - g(0))\frac{(x-\eta)^3}{(1-\eta)^2} & \text{for } \eta \le x < 1 \end{cases} \tag{30}$$

(iii)

$$G(x) = \begin{cases} g(1)x + \frac{1}{3}(g(0) - g(1))\left[\eta - \frac{(\eta-x)^3}{\eta^2}\right] & \text{for } 0 < x < \eta \\ g(1)x + \frac{1}{3}(g(0) - g(1))\eta & \text{for } \eta \le x < 1 \end{cases} \tag{31}$$

(iv)

$$G(x) = \begin{cases} Ax + \frac{1}{3}(g(0) - A)\left[\eta - \frac{(\eta-x)^3}{\eta^2}\right] & \text{for } 0 < x < \eta \\ Ax + \frac{1}{3}(g(0) - A)\eta + \frac{1}{3}(g(1) - A)\frac{(x-\eta)^3}{(1-\eta)^2} & \text{for } \eta \le x < 1 \end{cases} \tag{32}$$

In the special cases mentioned, $G = 0$.

If $\tau_{i-1} \le \alpha < \tau_i < \beta$ then $\int_\alpha^\beta g(s)\,ds = \int_\alpha^{\tau_i} g(s)\,ds + \int_{\tau_i}^\beta g(s)\,ds$ and we recurse down until the above condition is satisfied.

3.3. **Extrapolation.** The methodology has constructed the f function on the interval $[0, \tau_1]$. We integrate to find the r function on that interval.

For $\tau > \tau_n$ note that $f'(\tau_n) = 0$. Thus for $\tau > \tau_n$ we can put $f(\tau) = f(\tau_n)$. Now we use (10):

$$\begin{aligned} r(\tau)\tau &= r(\tau_n)\tau_n + \int_{\tau_n}^{\tau} f(s)\,ds \\ &= r(\tau_n)\tau_n + \int_{\tau_n}^{\tau} f(\tau_n)\,ds \\ &= r(\tau_n)\tau_n + (\tau - \tau_n)f(\tau_n) \end{aligned}$$

3.4. **Ensuring positivity.** Suppose we wish to guarantee that the interpolatory function f is everywhere positive.

Clearly from the formula (20) it suffices to ensure that $g(x) > -f_i^d$ for $x \in [0,1]$. Now $g(0) = f_{i-1} - f_i^d > -f_i^d$ and $g(1) = f_i - f_i^d > -f_i^d$ since $f_{i-1},\ f_i$ are positive. Thus the inequality is satisfied at the endpoints of the interval. Now, in regions (i), (ii) and (iii), g is monotone, so those regions are fine.

In region (iv) g is not monotone. g is positive at the endpoints and has a minimum of A (as in (28)) at the x-value η (as in (27)). So, it now suffices to prove that $\frac{g(0)g(1)}{g(0)+g(1)} < f_i^d$. This is the case if $f_{i-1},\ f_i < 3f_i^d$. To see this, note that then $0 < g(0),\ g(1) < 2f_i^d$ and the result follows, since if $0 < y,\ z < 2a$ then $\frac{y+z}{yz} = \frac{1}{z} + \frac{1}{y} > \frac{1}{2a} + \frac{1}{2a} = \frac{1}{a}$ and so $\frac{yz}{y+z} < a$.

We might choose the slightly stricter condition $f_{i-1},\ f_i < 2f_i^d$.

3.5. **Amelioration.** By shifting the f_i values we can make the interpolated curve smoother. The penalty is that the interpolated function will be less local; in some intervals $[\tau_{i-1}, \tau_i]$ the value of $f(\tau)$ might depend on f_j^d for $i-2 \le j \le i+2$. Thus in any particular application we must make a conscious decision as to whether we want the most locality or the best smoothness.

Let us consider the value $f_i \equiv f(\tau_i)$ between intervals i and $i+1$. Suppose first that $f_i^d > f_{i-1}^d$. If we also have $f_{i+1}^d > f_i^d$, then $f(\tau)$ is increasing in the interval i, and the smoothest results occur when f_i is in the range:

$$f_i^d + \tfrac{1}{2}(f_i^d - f_{i-1}) < f_i < f_i^d + 2(f_i^d - f_{i-1}) \tag{33}$$

This is our target range, the range in which we would prefer f_i to lie. Suppose now that $f_{i+1}^d < f_i^d$. Then $f(\tau)$ has a maximum in the interval. The maximum becomes steadily smaller as f_i increases towards f_i^d, but our interpolation function becomes increasingly asymmetric. In this case our target range is anything in

$$f_i^d - \tfrac{1}{2}\lambda(f_i^d - f_{i-1}) < f_i < f_i^d \tag{34}$$

where the parameter $0 \le \lambda \le 1$ determines the smoothness of the interpolated function. Experimentally $\lambda = 0.2$ seems to work well.

We cannot afford to have criteria for f_i which depend on values of $f(\tau)$ at other endpoints; this could lead to unpredictable non-locality and stability issues for marginal gains in smoothness. Instead we use the linear approximation to f_{i-1} as its proxy. Thus, to get good smoothness results for the interval i, we would like f_i to fall in the range

$$\begin{array}{ll} f_i^d + \frac{1}{2}\theta_i^- < f_i^- < f_i^d + \frac{1}{2}\theta_i^- & \text{if} \quad f_{i-1}^d < f_i^d < f_{i+1}^d \\ f_i^d - \frac{1}{2}\lambda\theta_i^- < f_i^- < f_i^d & \text{if} \quad f_{i-1}^d < f_i^d,\ f_i^d \ge f_{i+1}^d \end{array} \tag{35}$$

The targets for f_i if $f_i^d < f_{i-1}^d$ are obtained from similar considerations. Thus, considerations about the smoothness within interval i leads to the target range

$$f_{i,1}^{\min} \le f_i \le f_{i,1}^{\max} \tag{36}$$

$$\begin{array}{rll} f_{i,1}^{\min} = \min(f_i^d + \frac{1}{2}\theta_i^-, f_{i+1}^d), \quad f_{i,1}^{\max} = \min(f_i^d + 2\theta_i^-, f_{i+1}^d) & \text{if} & f_{i-1}^d < f_i^d \le f_{i+1}^d, \\ f_{i,1}^{\min} = \max(f_i^d - \frac{1}{2}\lambda\theta_i^-, f_{i+1}^d), \quad f_{i,1}^{\max} = f_i^d & \text{if} & f_{i-1}^d < f_i^d,\ f_i^d > f_{i+1}^d, \\ f_{i,1}^{\min} = f_i^d, \quad f_{i,1}^{\max} = \min(f_i^d - \frac{1}{2}\lambda\theta_i^-, f_{i+1}^d) & \text{if} & f_{i-1}^d \ge f_i^d,\ f_i^d \le f_{i+1}^d, \\ f_{i,1}^{\min} = \max(f_i^d + 2\theta_i^-, f_{i+1}^d), \quad f_{i,1}^{\max} = \max(f_i^d + \frac{1}{2}\theta_i^-, f_{i+1}^d) & \text{if} & f_{i-1}^d \ge f_i^d > f_{i+1}^d \end{array} \tag{37}$$

where

$$\theta_i^- = \frac{\tau_i - \tau_{i-1}}{\tau_i - \tau_{i-2}}(f_i^d - f_{i-1}^d) \tag{38}$$

Similar considerations about the smoothness of $f(\tau)$ in the interval $i+1$ lead to the target ranges

$$f_{i,2}^{\min} \le f_i \le f_{i,2}^{\max} \tag{39}$$

(40)

$$\begin{aligned}
f_{i,2}^{\min} &= \max(f_{i+1}^d - 2\theta_i^+, f_i^d), \quad f_{i,2}^{\max} = \max(f_{i+1}^d - \tfrac{1}{2}\theta_i^+, f_i^d) && \text{if} \quad f_i^d < f_{i+1}^d \le f_{i+2}^d, \\
f_{i,2}^{\min} &= \max(f_{i+1}^d + \tfrac{1}{2}\lambda\theta_i^+, f_i^d), \quad f_{i,2}^{\max} = f_{i+1}^d && \text{if} \quad f_i^d < f_{i+1}^d,\ f_{i+1}^d > f_{i+2}^d, \\
f_{i,2}^{\min} &= f_{i+1}^d, \quad f_{i,2}^{\max} = \min(f_{i+1}^d + \tfrac{1}{2}\lambda\theta_i^+, f_i^d) && \text{if} \quad f_i^d \ge f_{i+1}^d,\ f_{i+1}^d < f_{i+2}^d, \\
f_{i,2}^{\min} &= \min(f_{i+1}^d - \tfrac{1}{2}\theta_i^+, f_i^d), \quad f_{i,2}^{\max} = \min(f_{i+1}^d - 2\theta_i^+, f_i^d) && \text{if} \quad f_i^d \ge f_{i+1}^d \ge f_{i+2}^d
\end{aligned}$$

where

$$\theta_i^+ = \frac{\tau_{i+1} - \tau_i}{\tau_{i+2} - \tau_i}(f_{i+2}^d - f_{i+1}^d) \tag{41}$$

To ameliorate the max's, min's, and general ugliness of the interpolant, we use the following procedure:

(a) add an additional interval at the beginning and the end:

$$\tau_{-1} = \tau_0 - (\tau_1 - \tau_0) \quad , \quad f_0^d = f_1^d - \frac{\tau_1 - \tau_0}{\tau_2 - \tau_0}\left(f_2^d - f_1^d\right) \tag{42}$$

$$\tau_{n+1} = \tau_n + (\tau_n - \tau_{n-1}) \quad , \quad f_{n+1}^d = f_n^d + \frac{\tau_n - \tau_{n-1}}{\tau_n - \tau_{n-2}}\left(f_n^d - f_{n-1}^d\right). \tag{43}$$

(b) Select the f_i's by linearly interpolating on the midpoints of the intervals:

$$f_i = \frac{\tau_i - \tau_{i-1}}{\tau_{i+1} - \tau_{i-1}} f_{i+1}^d + \frac{\tau_{i+1} - \tau_i}{\tau_{i+1} - \tau_{i-1}} f_i^d, \qquad \text{for } i = 0,\ 1,\ \ldots,\ n. \tag{44}$$

Note that with the false intervals, this formula works for $i = 0$ and $i = n$.

(c) For each $i = 1,\ 2,\ \ldots,\ n-1$,

(i) if the target ranges overlap, define the common range

$$\max(f_{i,1}^{\min}, f_{i,2}^{\min}) \le f_i \le \min(f_{i,1}^{\max}, f_{i,2}^{\max}). \tag{45}$$

If f_i is outside this common range, make the minimum adjustment to f_i to place it in the common range:

$$\begin{aligned}
&\text{if} \quad f_i < \max(f_{i,1}^{\min}, f_{i,2}^{\min}) \quad \text{set} \quad f_i = \max(f_{i,1}^{\min}, f_{i,2}^{\min}) \\
&\text{if} \quad f_i > \min(f_{i,1}^{\max}, f_{i,2}^{\max}) \quad \text{set} \quad f_i = \min(f_{i,1}^{\max}, f_{i,2}^{\max})
\end{aligned} \tag{46}$$

(ii) if the target ranges don't overlap, define the gap by

$$\min(f_{i,1}^{\max}, f_{i,2}^{\max}) \le f_i \le \max(f_{i,1}^{\min}, f_{i,2}^{\min}). \tag{47}$$

If f_i is below or above the gap, make the minimum adjustment to f_i to place it on the edge of the gap:

$$\begin{aligned}
&\text{if} \quad f_i < \min(f_{i,1}^{\max}, f_{i,2}^{\max}) \quad \text{set} \quad f_i = \min(f_{i,1}^{\max}, f_{i,2}^{\max}) \\
&\text{if} \quad f_i > \max(f_{i,1}^{\min}, f_{i,2}^{\min}) \quad \text{set} \quad f_i = \max(f_{i,1}^{\min}, f_{i,2}^{\min})
\end{aligned} \tag{48}$$

(d) if now $\left|f_0 - f_0^d\right| > \frac{1}{2}\left|f_1 - f_0^d\right|$, replace f_0 with

$$f_0 = f_1^d - \tfrac{1}{2}\left(f_1 - f_0^d\right), \tag{49}$$

provided we don't know the value of f_0 (some markets explicitly quote f_0.)

(e) Similarly, if $\left|f_n - f_n^d\right| > \frac{1}{2}\left|f_{n-1} - f_n^d\right|$, replace f_n with

$$f_n = f_n^d + \tfrac{1}{2}\left(f_n^d - f_{n-1}\right). \tag{50}$$

(f) If the application requires $f(\tau) > 0$, apply the transformations of §3.4.

3.6. **Algorithm.** Our algorithm is

(1) Determine the f_i^d from the input data.
(2) Define f_i for $i = 0,\ 1,\ \ldots,\ n$ as in (16), (17) and (18).
(3) If f is required to be everywhere positive, then collar f_0 between 0 and $2f_1^d$, for $i = 1,\ 2,\ \ldots,\ n-1$ collar f_i between 0 and $2\min(f_i^d, f_{i+1}^d)$, and collar f_n between 0 and $2f_n^d$. If f is not required to be everywhere positive, simply omit this step.
(4) Construct g with regard to which of the four sectors we are in.
(5) Define f as in (20).
(6) If required recover r as in (11). Integration formulae are easily established as the functional forms of g are straightforward.

Bibliography

J. C. Cox, J. E. Ingersoll, and S. A. Ross. A theory of the term structure of interest rates. *Econometrica*, 53:385–407, 1985. 1

Patrick S. Hagan and Graeme West. Interpolation methods for curve construction. *Applied Mathematical Finance*, 13(2):89–129, 2006. 1, 3, 4, 6, 7

Patrick S. Hagan and Graeme West. Methods for constructing a yield curve. *WILMOTT Magazine*, May:70–81, 2008. URL http://www.finmod.co.za/interpreview.pdf. 1, 3, 4, 6

D. Heath, R. Jarrow, and A. Morton. Bond pricing and the term structure of interest rates: A discrete time approximation. *Journal of Financial and Quantitative Analysis*, 25:419–440, 1990. 2

J. Hull and A. White. Pricing interest rate derivative securities. *Review of Financial Studies*, 3, 1990. 2

James M. Hyman. Accurate monotonicity preserving cubic interpolation. *SIAM Journal on Scientific and Statistical Computing*, 4(4):645–654, 1983. 6, 7

Bing-Huei Lin. Fitting term structure of interest rates using B-Splines: The case of Taiwanese government bonds. *Applied Financial Economics*, 12(1):57–75, 2002. 1

J. Huston McCulloch and Levis A. Kochin. The inflation premium implicit in the US real and nominal term structures of interest rates. Technical Report 12, Ohio State University Economics Department, 2000. URL http://www.econ.ohio-state.edu/jhm/jhm.html. 5

C.R. Nelson and A.F. Siegel. Parsimonious modelling of yield curves. *Journal of Business*, 60(4), 1987. 2

Ricardo Rebonato. *Interest-Rate Option Models*. John Wiley and Sons Ltd, second edition, 1998. 2

Uri Ron. A practical guide to swap curve construction. Technical Report 17, Bank of Canada, 2000. URL http://www.bankofcanada.ca/en/res/wp00-17.htm. 2

L.E. Svensson. Estimating and interpreting forward interest rates: Sweden 1992-1994, 1994. URL http://www.nber.org/papers/W4871. IMF working paper. 2

O. A. Vasicek. An equilibrium characterisation of the term structure. *Journal of Financial Economics*, 15, 1977. 1

Peter Zangari. An investigation into term structure estimation methods for RiskMetrics. *RiskMetrics Monitor*, Third Quarter:3–48, 1997. 1

Financial Modelling Agency, 19 First Ave East, Parktown North, 2193, South Africa.
E-mail address: graeme@finmod.co.za