the smoothing spline with weighted roughness measure

Carl de Boor

The smoothing spline, $f = f_\rho$, of Schoenberg [S] and Reinsch [R1], [R2], uniquely minimizes

$$\rho \sum_j w_j (y_j - f(x_j))^2 + \int_a^b \lambda (D^m f)^2$$

over all f in

$$X := L_2^{(m)}[a \mathinner{..} b],$$

given the points $x_1 < \cdots < x_N$ in $[a \mathinner{..} b]$, the data $y = (y_j)$, the weight vector $w = (w_j)$ of positive weights (usually equal to 1), the smoothing parameter $\rho \in [0+ \mathinner{..} \infty -]$, and the natural number m, in the special case that $\lambda = 1$. Over the years, this smoothing spline, particularly after the introduction of generalized cross validation by Wahba and ??? [citWK] for an automatic choice of the **smoothing parameter**, ρ, and for $m = 2$, has become the spline most often used in practical problems of data fitting and analysis. However, use of a nonconstant weight λ in the **roughness measure** $\int \lambda (D^m f)^2$ provides additional, very useful, flexibility in the shaping of the smoothing spline. It is the purpose of this note to provide a simple derivation of the numerical algorithm needed to construct such a more complex smoothing spline, for given ρ and λ. This derivation shows that, for a piecewise constant λ with breaks only at the x_j, the algorithms for the case $\lambda = 1$ only need minor adjustments to provide this potentially very useful added capability. The derivation is given in full, in a somewhat nonstandard way.

It seems simplest to me (and to some others, see, e.g., [A] and the references there) to view this minimization problem as a special case of best approximation in an inner product space, as follows: Use the linear maps

$$\alpha : X \to Y := \mathbb{R}^N : f \mapsto f|_x := \big(f(x_j)\big)_{j=1}^N, \qquad \beta : X \to Z := L_2[a \mathinner{..} b] : f \mapsto D^m f,$$

to embed X in the Hilbert space

$$H := Y \times Z$$

with natural inner product

$$\langle (f, g), (h, k) \rangle := \rho \langle f, h \rangle_Y + \langle g, k \rangle_Z$$

with

$$\langle f, g \rangle_Y := \sum_j w_j f_j g_j, \quad f, g \in \mathbb{R}^N,$$

$$\langle f, g \rangle_Z = \int_a^b \lambda f g, \quad f, g \in L_2[a \mathinner{..} b].$$

Assuming as we do that $0 < \rho < \infty$, the only issue here is whether

$$X \to H : f \mapsto (\alpha(f), \beta(f))$$

is an embedding and whether, with this embedding, X becomes a closed subspace of H. For the former, it is necessary and sufficient that

$$\ker\alpha \cap \ker\beta = \{0\},$$

and, since

$$\ker\alpha = \{f \in X : f|_x = 0\}, \quad \ker\beta = \Pi_{<m}$$

(the space of polynomials of degree $< m$), this will be so iff $N \ge m$, an assumption we make from now on. As to the latter, it is, in essence, the claim that D, hence D^m, is a closed linear map. Explicitly, I take for granted the standard representation theorem

“taylor (1) $$f = T_{a,m}f + R\beta(f), \qquad \forall f \in X,$$

with $T_{a,m}f$ the Taylor polynomial of order m for f at a and with

$$R : Z \to X : g \mapsto \int_a^b (\cdot - s)_+^{m-1} g(s)\,\mathrm{d}s/(m-1)!.$$

This identifies X as the sum $\Pi_{<m} + R(Z)$ of a finite-dimensional linear subspace (which therefore is closed) and the subspace $R(Z)$ which is closed, hence X itself is closed.

Thus, the smoothing spline f_ρ is the unique best approximation from $X \subset H$ to the element $(y, 0) \in H$, hence is characterized by the fact that the error, $(y,0) - (\alpha(f_\rho), \beta(f_\rho))$, is perpendicular to $X \subset H$, i.e.,

“orthog (2) $$\rho\langle y - \alpha(f_\rho), \alpha(f)\rangle_Y + \langle -\beta(f_\rho), \beta(f)\rangle_Z = 0 \qquad \forall f \in X.$$

Since $\ker\beta = \Pi_{<m}$, (2) implies that

“orthogY (3) $$\langle y - \alpha(f_\rho), \alpha(f)\rangle_Y = 0 \qquad \forall f \in \Pi_{<m}$$

and, with this and (1), (2) implies that

“orthogZ (4) $$\rho\langle y - \alpha(f_\rho), \alpha(Rg)\rangle_Y = \langle \beta(f_\rho), g\rangle_Z \qquad \forall g \in Z.$$

Conversely, (3)–(4) imply (2).

Since the left side of (4) is a continuous linear functional as a function of g, it is expressible in the form $\langle z, g\rangle_Z$ for some $z \in Z$. Explicitly, with $h := y - \alpha(f_\rho)$,

$$\begin{aligned}\langle y - \alpha(f_\rho), \alpha(Rg)\rangle_Y &= \sum_j w_j h_j \int_a^b (x_j - s)_+^{m-1} g(s)\,\mathrm{d}s/(m-1)! \\ &= \int_a^b \Big(\sum_j w_j h_j\,(x_j - s)_+^{m-1}/(m-1)!\Big) g(s)\,\mathrm{d}s \\ &= \langle \frac{1}{\lambda}\sum_j w_j h_j (x_j - \cdot)_+^{m-1}/(m-1)!, g\rangle_Z.\end{aligned}$$

It follows that (4) is equivalent to

“orthogZZ (5)
$$\rho \sum_j w_j(y_j - f_\rho(x_j))(x_j - \cdot)_+^{m-1}/(m-1)! \;=\; \lambda D^m f_\rho.$$

This shows that $D^m f_\rho$ is a spline of order m with knot sequence x, and vanishes to the right of x_N. However, by (3),

$$\sum_j w_j(y_j - f_\rho(x_j))(x_j - \cdot)^{m-1}/(m-1)! \;=\; 0,$$

hence, with $(x_j - t)_+^{m-1} = (x_j - t)^{m-1} - (-1)^{m-1}(t - x_j)_+^{m-1}$, also

“orthogZZZ (6)
$$\rho(-1)^m \sum_j w_j(y_j - f_\rho(x_j))(\cdot - x_j)_+^{m-1}/(m-1)! \;=\; \lambda D^m f_\rho,$$

showing that $D^m f_\rho$ also vanishes to the left of x_1. Consequently, $\lambda D^m f_\rho$ is an element of $S_{m,x}$ (i.e., for $\lambda = 1$, f_ρ is a 'natural' spline of order $2m$ with break sequence x). In particular, we may write

“defc (7)
$$\lambda D^m f_\rho =: \sum_k B_{k,m,x} c_k,$$

with $B_{k,m,x}$ the normalized B-spline with knots $x_k, \ldots, x_{k+m}$, i.e.,

$$\begin{aligned} B_{k,m,x}(t) &= (x_{k+m} - x_k)[x_k, \ldots, x_{k+m}](\cdot - t)_+^{m-1} \\ &= \sum_j (x_j - t)_+^{m-1} c_{j,k}, \end{aligned}$$

with

$$c_{j,k} := \begin{cases} (x_{k+m} - x_k)/\prod\{(x_j - x_i) : i \in \{k, \ldots, k+m\}\backslash\{j\}\}, & j = k, \ldots, k+m; \\ 0, & \text{otherwise.} \end{cases}$$

Thus, on expressing $\lambda D^m f_\rho$ in (5) in this way and comparing coefficients of $(x_j - \cdot)_+^{m-1}$, we obtain

“relationone (8)
$$\rho W(y - \alpha(f_\rho)) = Cc,$$

with

$$W := \operatorname{diag}(w), \qquad C := (m-1)!(c_{j,k} : j = 1, \ldots, N; k = 1, \ldots, N-m),$$

and c the B-spline coefficient sequence for $\lambda D^m f_\rho$, as defined in (7). Now note that, for any $f \in X$,

“Cta (9)
$$(C^t \alpha(f))_j \;=\; (m-1)!\,(x_{j+m} - x_j)[x_j, \ldots, x_{j+m}]f = \int_a^b B_{j,m,x} D^m f.$$

Therefore, any $\alpha(f_\rho)$ satisfying (8) automatically satisfies (3), since, for any such $\alpha(f_\rho)$ and any $d \in X$,

$$\langle y-\alpha(f_\rho), \alpha(f)\rangle_Y \;=\; \alpha(f)^t W(y-\alpha(f_\rho)) \;=\; \alpha(f)^t Cc/\rho \;=\; (C^t\alpha(f))^t c/\rho,$$

while $C^t\alpha(f)=0$ for $f\in\Pi_{<m}$ by (9). Since (8) with (7) implies (5), hence (4), it follows that, with (7), (8) is equivalent to (2). We therefore now concentrate on (8).

For that, from (9) with (7),

“relationtwo (10)
$$C^t\alpha(f_\rho) = Ac,$$

with

$$A := (\int_a^b B_{j,m,x}B_{k,m,x} : j,k=1,\ldots,N-m).$$

Substitution of (8), in the form

“equationfora (11)
$$\alpha(f_\rho) \;=\; y - \; W^{-1}C(c/\rho),$$

into (10) gives the equation

“equationforc (12)
$$C^t y = (C^tW^{-1}C + \rho\, A)(c/\rho)$$

which may be solved stably for $u := c/\rho$, for given data y (since its coefficient matrix is symmetric positive definite). From this, we obtain the smoothed values $\alpha(f_\rho)$ directly from (11). To obtain f_ρ, integrate the resulting $D^m f_\rho = (1/\lambda)\sum_j B_{j,m,x}c_j$ m times , to obtain $f := D^{-m}(D^m f_\rho)$, which differs from f_ρ only by some $q\in\Pi_{<m}$. Determine this q as the unique $q\in\Pi_{<m}$ for which $\alpha(q+f)=\alpha(f_\rho)$, i.e., for which $\alpha(q)=\alpha(f_\rho)-\alpha(f)$, with the vector $\alpha(f_\rho)$ computed from (11).

It is only at this point, of m-fold integration, that the choice of the weight λ in the roughness measure begins to matter (other than an assumption that λ be measurable and essentially positive, to ensure that $\langle\,,\,\rangle_Z$ is an inner product). For the special case $\lambda=1$, one would use the standard formula, see, e.g., [pgs: p.150], to carry out the integration, obtaining, in that case, f_ρ as a natural spline of order $2m$ with simple interior knots $(x_i : i=2,\ldots,N-1)$. While the integration can be carried out in closed form for a somewhat larger class, we will, at this point, restrict attention to those λ for which f_ρ is still piecewise polynomial and, specifically, on the simplest of these, namely the piecewise constants with breaks only at the x_i, i.e.,

“choicegl (13)
$$\lambda \in \Pi_{1,x}.$$

Others (e.g.,[cit???],[cit???]) have considered λ that are reciprocals of continuous piecewise linears with breaks only at the x_i), presumably in order to avoid the jumps in $D^m f_\rho$ introduced when λ is piecewise constant.

The behavior of the error as a function of ρ

According to (12), as $\rho \to 0$, $u = c/\rho$ converges to $(C^t W^{-1} C)^{-1} C^t y$, hence $c = \rho u \to 0$, therefore f_{0+} is the unique polynomial $q \in \Pi_{<m}$ that minimizes $\|y - \alpha(q)\|$. At the other extreme, as $\rho \to \infty$, (12) approaches the equation $C^t y = Ac$, hence, with (10), $f_{\infty -}$ is the unique natural spline of order $2m$ with knots x that interpolates to the given data.

As a function of ρ, the error

$$E_\rho := \|y - \alpha(f_\rho)\|^2$$

decreases with increasing ρ, as can be seen as follows: For each $f \in X$,

$$\mathbb{R}_+ \to \mathbb{R} : \rho \mapsto \rho\|y - \alpha(f)\|^2 + \|\beta(f)\|^2$$

is a straight line, hence the function

$$F : \mathbb{R}_+ \to \mathbb{R} : \rho \mapsto \min_{f \in X} \bigl(\rho\|y - \alpha(f)\|^2 + \|\beta(f)\|^2\bigr),$$

as the pointwise minimum of a collection of straight lines with nonnegative y-intercepts and nonnegative slopes, is continuous, nondecreasing, concave downward and is bounded (above) by its asymptote at infinity, the constant line of height $\|\beta(f_{\infty -})\|^2$, while the asymptote at the other extreme (i.e., the tangent at the origin) is the line through the origin with slope $\|y - \alpha(f_{0+})\|^2$. Since the straight line

$$\rho \mapsto \rho\|y - \alpha(f_\rho)\|^2 + \|\beta(f_\rho)\|^2$$

is the tangent to F at ρ, we have

$$DF(\rho) = E_\rho,$$

showing that E_ρ decreases with increasing ρ and that, correspondingly, $\|\beta(f_\rho)\|^2$ (the y-intercept of the tangent) increases with increasing ρ.

For this reason, Reinsch and others have proposed to choose the smoothing parameter ρ as small as possible subject to the constraint that E_ρ not exceed a given tolerance, *tol*. Further, Reinsch has pointed out that the function

$$G : \rho \mapsto 1/\|y - \alpha(f_\rho)\|$$

is concave upward and becomes ever more linear with growing ρ, hence Newton's method applied to the equation

"tosolve (14)
$$1/E_\rho^{1/2} - 1/(tol)^{1/2} = 0$$

for ρ and started at $\rho = 0$ is bound to converge, and to converge quite fast, particularly if the solution is 'large'. Further, since

$$E_\rho = u^t C^t W^{-1} C u$$

by (11), one gets $DE_\rho = 2u^t C^t W^{-1} CDu$, while, from (12), Du uniquely solves the equation

$$-Au = (C^t W^{-1} C + \rho A) Du.$$

In particular, for $\rho = 0$, this says that $DE_\rho = -2u^t Au$, with $u = (C^t W^{-1} C)^{-1} C^t y$ needed in any case for the calculation of $\alpha(f_\rho)$ via (11). This provides the slope needed for the starting step, at $\rho = 0$, of Newton's method applied to (14). For subsequent steps, I would avoid calculation of DE_ρ (which requires solution of a linear system) by using the Secant method instead.

Another limiting case of interest concerns the confluence of some of the x_j. If the data y come from a smooth function and the relevant weights behave appropriately, then confluence of $r \le m$ neighboring points leads to the smoothing problem in which α also involves all the derivatives of order $< r$ at the multiple point and, correspondingly, f_ρ has only $2m - 1 - r$ continuous derivatives across that multiple point. Of course, the relevant formulæ for such an α can be derived directly in the above way, using divided differences with repeated nodes and, correspondingly, B-splines with repeated knots, in the standard way. In particular, there is some practical use for the *complete cubic smoothing spline* for which $\alpha(f) = (Df(x_1), f|_x, Df(x_N))$.

Numerical construction of the B-spline Gramian There is one final hurdle to writing a program for the computation of f_ρ for general m, namely the construction of the matrix A of inner products of B-splines. This is the second point at which the choice of λ becomes important. With our choice of

$$\lambda =: \sum_{j=1}^{N-1} \lambda_j \chi_{(x_j \,..\, x_{j+1})} \in \Pi_{1,x}$$

instead of just $\lambda = 1$, the calculation of the entries of A is not at all complicated since, for small m, the integrals

$$A_{j,k} = \int_a^b B_{j,m,x} B_{k,m,x} / \lambda, \qquad j, k = 1, \ldots, N - m,$$

are most easily evaluated break interval by break interval anyway. To be sure, for $\lambda = 1$, there are stable recurrence relations for the integrals available in the literature, e.g., in [BLS]. For the first few values of m, though, it is easy to work out the matrix entries, as follows:

case $m = 1$**:** In this case, $B_{j,m,x} = \chi_{[x_j \,..\, x_{j+1})}$, hence A is the diagonal matrix with diagonal entries $\Delta x_j / \lambda_j$, $j = 1, \ldots, N$.

case $m = 2$**:** In this case, $B_{j,m,x}$ is the piecewise linear function that is zero at all its breaks x, except at x_{j+1}, where it is 1. Correspondingly,

$$D^2 f_\rho(t) = \begin{cases} c_j / \lambda_j & \text{for } t = x_j^+; \\ c_{j-1} / \lambda_{j-1} & \text{for } t = x_j^-. \end{cases}$$

Hence, with $\alpha(f_\rho)$ computed from (11), construction of the local cubic pieces is immediate once c (or c/ρ) is obtained from (12).

Further, with $t = x_j + s\Delta x_j$,

$$\int_{x_j}^{x_{j+1}} B_{j,2}(t)^2 \,\mathrm{d}t = \Delta x_j \int_0^1 s^2 \,\mathrm{d}s = \Delta x_j/3,$$

while

$$\int_{x_j}^{x_{j+1}} B_{j-1,2}(t)B_{j,2}(t)\,\mathrm{d}t = \int_{x_j}^{x_{j+1}} B_{j,2}(t)\,\mathrm{d}t - \int_{x_j}^{x_{j+1}} B_{j,2}(t)^2\,\mathrm{d}t = \Delta x_j - \Delta x_j/3 = \Delta x_j/6.$$

Consequently, A is the tridiagonal matrix with general row

$$(\frac{\Delta x_j}{\lambda_j}, 2(\frac{\Delta x_j}{\lambda_j} + \frac{\Delta x_{j+1}}{\lambda_{j+1}}), \frac{\Delta x_{j+1}}{\lambda_{j+1}})/6, \qquad j = 1, \ldots, N-2.$$

Note that, for $\lambda = 1$, the entries in such a row add up to $(x_{j+2} - x_j)/2 = \int B_{j,2,x}$, exactly as they should.

case $m = 3$**:** (In `spaps`, I used Gauss quadrature for this case, but should replace that by the formulas (to be) obtained here.) In this case, A is five-diagonal and

$$B_{j,m,x}(t) \;=\; ([x_{j+1}, x_{j+2}, x_{j+3}] - [x_j, x_{j+1}, x_{j+2}])(\cdot - t)_+^2.$$

In particular,

$$\int B_{j-2,3}B_{j,3} = \int_{x_j}^{x_{j+1}} (t - x_j)^2 (x_{j+1} - t)^2 \,\mathrm{d}t/a_j,$$

with

$$a_j := (x_{j+1} - x_j)(x_{j+2} - x_j)(x_j - x_{j+1})(x_{j-1} - x_{j+1}),$$

hence

$$A_{j-2,j} = (\Delta x_j)^3/(\lambda_j 30(x_{j+1} - x_{j-1})(x_{j+2} - x_j)).$$

Next, the slightly harder calculation of $\int B_{j-1,3}B_{j,3}$, for which it is, by symmetry, sufficient to calculate

$$\int_{x_j}^{x_{j+1}} B_{j-1,3}B_{j,3}.$$

etc. Finally, $A_{j,j}$ is obtained from this by symmetry and by the fact that, necessarily, $\sum_r A_{j,r} = \int B_{j,3} = (x_{j+3} - x_j)/3$.

Following S. Kersey's good advice, the calculation of $A_{j-1,j}$ in the case $\lambda = 1$ might be better accomplished by using the formula

$$\int B_{i,k}B_{j,k} = (-1)^k \frac{(2k-1)!}{(k!)^2}(t_{i+k} - t_i)(t_{j+k} - t_j)[t_i, \ldots, t_{i+k}]_x [t_j, \ldots, t_{j+h}]_y (x-y)_+^{2k-1},$$

which, in slightly different form, appears already for that purpose in [JS].

References

[A] M. Attéia, *Hilbertian Kernels and Spline Functions*, Elsevier Science Publishers, Amsterdam, 1992.

[pgs] C. de Boor, *A Practical Guide to Splines*, Springer Verlag, New York, 1978.

[BLS] C. de Boor, T. Lyche, and L. L. Schumaker, *On calculating with B-splines II. Integration*, Numerische Methoden der Approximationstheorie Vol. 3, ISNM 30 (L. Collatz, G. Meinardus and H. Werner, eds), Birkhäuser Verlag, Basel, 1976, pp. 123–146.

[JS] J. Jerome and L. L. Schumaker, *A note on obtaining natural spline functions by the abstract approach of Atteia and Laurent*, SIAM J. Numer. Anal., 5 (1968), pp. 657–663.

[R1] C. H. Reinsch, *Smoothing by spline functions*, Numer. Math., 10 (1967), pp. 177–183.

[R2] Christian H. Reinsch, *Smoothing by spline functions. II*, Numer. Math., 16 (1971), pp. 451–454.

[S] I. J. Schoenberg, *Spline functions and the problem of graduation*, Proc. Amer. Math. Soc., 52 (1964), pp. 947–950.