Kernel Function Estimation and Cubic Smoothing Splines

Abstract

Two classical approaches to estimating an unknown regression function from noisy observations are developed. For a fixed-design kernel smoother, the bias and variance are calculated for a signed kernel of order r, the bandwidth is chosen by balancing the two terms, and the one-dimensional Stone rate and its d-dimensional analogue are obtained. Cubic smoothing splines are then derived by minimizing a residual sum of squares plus the roughness penalty $\int (f'')^2$. A variational argument first gives the piecewise-cubic form. The interval coefficients are then counted, the value and slope matching equations are imposed, and the remaining first-order equations are derived in terms of the knot values and slopes. The resulting block-tridiagonal system is shown to have a unique solution, after which its natural boundary and residual properties are identified. Practical computation is also described through a reduced natural-spline system and through the normal equations for a cubic B-spline basis.

Contents

1 Kernel estimators for density and regression

Suppose that

$$Y_i = f(x_i) + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{1}$$

where the design points are fixed and equally spaced on $[0, 1]$, and

$$\mathbb{E}\varepsilon_i = 0, \qquad \mathrm{Var}(\varepsilon_i) = \sigma^2, \qquad \varepsilon_1, \ldots, \varepsilon_n \text{ independent.} \tag{2}$$

For definiteness, take

$$x_i = \frac{i-1}{n-1}, \qquad \Delta_n = x_{i+1} - x_i = \frac{1}{n-1}.$$

All kernel calculations concern an interior point x that remains a fixed positive distance from 0 and 1. We study the customary normalization

$$\widehat{f}_b(x) = \frac{1}{nb} \sum_{i=1}^{n} Y_i K\left(\frac{x - x_i}{b}\right), \tag{3}$$

where $b > 0$ is the bandwidth and K is a kernel. It is convenient to write

$$K_b(t) = \frac{1}{b} K\left(\frac{t}{b}\right), \qquad \widehat{f}_b(x) = \frac{1}{n} \sum_{i=1}^{n} Y_i K_b(x - x_i).$$

The exact mesh width enters naturally in the corresponding Riemann sum. Define

$$\widetilde{f}_b(x) = \Delta_n \sum_{i=1}^{n} Y_i K_b(x - x_i). \tag{4}$$

Then

$$\widehat{f}_b(x) = \frac{n-1}{n}\, \widetilde{f}_b(x). \tag{5}$$

Thus the exact mesh-weighted form can be used for the quadrature calculation, after which (5) gives the result for (3). This is a fixed-design kernel regression estimator of the type studied by Priestley and Chao [12].

There are three superficially similar formulas that estimate three different objects.

1.1 Kernel density estimation

If $X_1, \ldots, X_n$ are random observations from a density p, then

$$\widehat{p}_b(x) = \frac{1}{nb} \sum_{i=1}^{n} K\left(\frac{x - X_i}{b}\right) \tag{6}$$

is the classical kernel density estimator, often called the Parzen–Rosenblatt estimator after Rosenblatt [14] and Parzen [11]. Thus, if the Y_i factors are removed from (3) and the locations are themselves random observations, one obtains a density estimator.

In the fixed-design problem of (1), the x_i are deterministic. The same expression without the Y_i smooths the empirical design measure, whose interior limit for the uniform grid is the known constant 1.

1.2 The Nadaraya–Watson regression estimator

Suppose instead that (X_i, Y_i) are iid, and let

$$m(x) = \mathbb{E}(Y \mid X = x), \qquad p(x) = \text{the density of } X.$$

The Nadaraya–Watson regression estimator is

$$\widehat{m}_{\mathrm{NW}}(x) = \frac{\sum_{i=1}^n Y_i K_b(x - X_i)}{\sum_{i=1}^n K_b(x - X_i)}. \tag{7}$$

Its numerator and denominator arise from the factorization

$$q(x) = m(x)p(x).$$

Indeed, define

$$\widehat{q}_b(x) = \frac{1}{n}\sum_{i=1}^n Y_i K_b(x - X_i). \tag{8}$$

Then

$$\begin{aligned}\mathbb{E}\widehat{q}_b(x) &= \mathbb{E}\{Y K_b(x - X)\} \\ &= \int m(u)p(u)K_b(x-u)\,\mathrm{d}u \longrightarrow m(x)p(x). \end{aligned} \tag{9}$$

The design-density factor is estimated by

$$\widehat{p}_b(x) = \frac{1}{n}\sum_{i=1}^n K_b(x - X_i),$$

which converges to $p(x)$. Consequently,

$$\widehat{m}_{\mathrm{NW}}(x) = \frac{\widehat{q}_b(x)}{\widehat{p}_b(x)}$$

uses the denominator to cancel the multiplicative design-density factor in the numerator. This construction was introduced independently by Nadaraya [10] and Watson [26].

The ratio also has a direct least-squares interpretation. For a nonnegative kernel, the value in (7) is the local constant fit

$$\widehat{m}_{\mathrm{NW}}(x) = \arg\min_{a \in \mathbb{R}} \sum_{i=1}^n K_b(x - X_i)(Y_i - a)^2. \tag{10}$$

The denominator in (7) simply normalizes the local weights to sum to one. In particular, if $Y_i \equiv c$, then $\widehat{m}_{\mathrm{NW}}(x) = c$ exactly, regardless of how unevenly the X_i are distributed.

1.3 Fixed, equally spaced design

For the deterministic uniform grid, the exact mesh-weighted sum satisfies

$$\Delta_n \sum_{i=1}^n K_b(x - x_i) \approx \int_0^1 K_b(x-u)\,\mathrm{d}u \approx 1 \tag{11}$$

at interior points. By (5), the same limit holds with the factor $1/n$ in place of Δ_n. Thus the design-density factor is asymptotically known in the fixed uniform design. Dividing (3) by its corresponding kernel-weight sum would enforce exact reproduction of constants at finite n, without changing the leading bias–variance calculation below.

Remark 1.1 (Terminology). The conventional terminology is:

problem	common name
density estimation from random X_i	Parzen–Rosenblatt estimator
regression with random (X_i, Y_i)	Nadaraya–Watson estimator
regression on a fixed regular design	Priestley–Chao-type kernel estimator.

Rosenblatt also made broader contributions to conditional-density and regression estimation [15]; his name is attached most routinely to the density estimator (6).

2 Higher order Kernels

Definition 2.1 (Kernel of order r). Let $r \geq 2$ be an integer. A function $K : \mathbb{R} \to \mathbb{R}$ is a kernel of order r if

$$\int_{\mathbb{R}} K(u)\,\mathrm{d}u = 1, \qquad \int_{\mathbb{R}} u^j K(u)\,\mathrm{d}u = 0, \quad 1 \leq j \leq r-1, \tag{12}$$

and

$$\mu_r(K) := \int_{\mathbb{R}} u^r K(u)\,\mathrm{d}u \neq 0. \tag{13}$$

We also write

$$R(K) := \int_{\mathbb{R}} K(u)^2\,\mathrm{d}u. \tag{14}$$

Throughout this section, r is even and K is symmetric. Symmetry gives the vanishing odd moments automatically, while the remaining even moments through order $r-2$ are imposed through the moment conditions.

A nonnegative integrable kernel cannot have order greater than two. For example, an order-four kernel would have to satisfy

$$\int u^2 K(u)\,\mathrm{d}u = 0.$$

If $K \geq 0$, the integrand is nonnegative, so this equality would force K to be concentrated at $u = 0$, which is impossible for an ordinary integrable kernel with integral one. Thus higher-order kernels must take both positive and negative values.

Example 2.2 (An explicit symmetric fourth-order kernel). Define

$$K_4(u) = \frac{15}{32}(3 - 7u^2)(1 - u^2)\,\mathbf{1}\{|u| \leq 1\}. \tag{15}$$

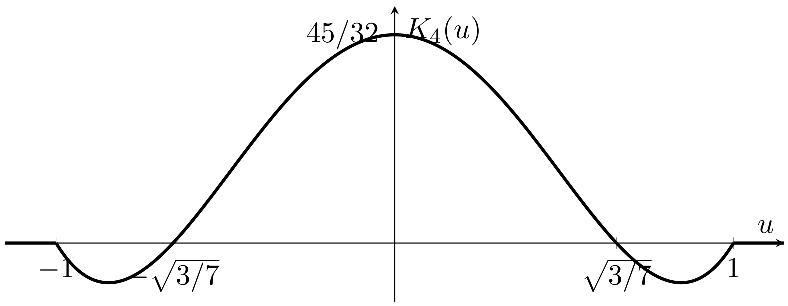
A direct calculation gives

$$\int K_4(u)\,\mathrm{d}u = 1, \qquad \int u^2 K_4(u)\,\mathrm{d}u = 0,$$

while symmetry gives the vanishing first and third moments. Moreover,

$$\mu_4(K_4) = -\frac{1}{21}, \qquad R(K_4) = \frac{5}{4}.$$

The kernel is negative when $\sqrt{3/7} < |u| < 1$. Since the factor $1 - u^2$ makes it vanish at the endpoints of its support, its extension by zero is absolutely continuous.



Proposition 2.3 (Existence of compactly supported kernels of every even order)**.** *Let $r = 2m$. Let w be a positive symmetric integrable function on $[-1, 1]$ with all moments finite. There are unique coefficients $a_0, \dots, a_{m-1}$ such that*

$$K(u) = w(u) \sum_{\ell=0}^{m-1} a_\ell u^{2\ell} \tag{16}$$

has integral one and vanishing moments of orders $2, 4, \dots, 2m - 2$.

Proof. The coefficients must solve

$$\sum_{\ell=0}^{m-1} a_\ell \int_{-1}^{1} u^{2(j+\ell)} w(u) \,\mathrm{d}u = \begin{cases} 1, & j = 0, \\ 0, & 1 \le j \le m-1. \end{cases}$$

The coefficient matrix is the Gram matrix of the functions $1, u^2, \dots, u^{2m-2}$ in $L^2(w)$, and is therefore positive definite. Hence it is invertible. Symmetry supplies all the odd-moment conditions. □

Higher-order signed kernels are useful for transparent rate calculations. In practice, local polynomial estimators are often preferred because they obtain comparable bias cancellation without requiring one to specify a globally signed weighting function; see Fan and Gijbels [7].

3 Bias of the fixed-design kernel estimator

Assume that K is compactly supported and absolutely continuous, with $K' \in L^1(\mathbb{R})$, and that f has r continuous derivatives in a neighborhood of the interior point x. Since the design is fixed,

$$\mathbb{E}\widehat{f}_b(x) = \frac{1}{n} \sum_{i=1}^{n} f(x_i) K_b(x - x_i). \tag{17}$$

The sum is a Riemann approximation to a convolution. The following bound will be used for both the bias and the variance calculations.

Lemma 3.1 (Regular-grid quadrature bound)**.** *Let*

$$0 = x_1 < \cdots < x_n = 1, \qquad x_i = \frac{i-1}{n-1}.$$

If h is absolutely continuous on $[0, 1]$, then

$$\left| \frac{1}{n-1} \sum_{i=1}^{n-1} h(x_i) - \int_0^1 h(u) \,\mathrm{d}u \right| \le \frac{1}{n-1} \int_0^1 |h'(u)| \,\mathrm{d}u. \tag{18}$$

Proof. Since $x_{i+1} - x_i = 1/(n-1)$,

$$\frac{h(x_i)}{n-1} - \int_{x_i}^{x_{i+1}} h(u)\,\mathrm{d}u = \int_{x_i}^{x_{i+1}} \{h(x_i) - h(u)\}\,\mathrm{d}u.$$

Absolute continuity gives the exact identity

$$h(x_i) - h(u) = -\int_{x_i}^{u} h'(v)\,\mathrm{d}v.$$

Therefore, by the triangle inequality and a change in the order of integration,

$$\begin{aligned}
&\left|\frac{1}{n-1}\sum_{i=1}^{n-1} h(x_i) - \int_0^1 h(u)\,\mathrm{d}u\right| \\
&\quad \le \sum_{i=1}^{n-1}\int_{x_i}^{x_{i+1}}\int_{x_i}^{u} |h'(v)|\,\mathrm{d}v\,\mathrm{d}u \\
&\quad = \sum_{i=1}^{n-1}\int_{x_i}^{x_{i+1}} (x_{i+1} - v)|h'(v)|\,\mathrm{d}v \\
&\quad \le \frac{1}{n-1}\int_0^1 |h'(v)|\,\mathrm{d}v.
\end{aligned}$$

□

Proposition 3.2 (Pointwise bias)**.** *As $b \to 0$ and $n \to \infty$,*

$$\mathbb{E}\widehat{f}_b(x) = f(x) + B_r(x)b^r + o(b^r) + O\left(\frac{1}{nb}\right), \tag{19}$$

$$\mathrm{Bias}\{\widehat{f}_b(x)\} = B_r(x)b^r + o(b^r) + O\left(\frac{1}{nb}\right), \tag{20}$$

where

$$B_r(x) = \frac{(-1)^r \mu_r(K)}{r!} f^{(r)}(x). \tag{21}$$

For the even orders used here, $(-1)^r = 1$.

Proof. Set

$$h_b(u) = f(u)K_b(x-u).$$

Because x is an interior point and K has compact support, $h_b(0) = h_b(1) = 0$ for all sufficiently small b. Hence

$$\frac{1}{n-1}\sum_{i=1}^{n} h_b(x_i) = \frac{1}{n-1}\sum_{i=1}^{n-1} h_b(x_i).$$

Moreover,

$$h_b'(u) = f'(u)K_b(x-u) - f(u)K_b'(x-u),$$

where

$$K_b'(t) = \frac{1}{b^2}K'\left(\frac{t}{b}\right).$$

On a fixed neighborhood of x containing the support of h_b for all sufficiently small b,

$$\int_0^1 |f'(u)K_b(x-u)|\,\mathrm{d}u \le \|f'\|_\infty \int_{\mathbb{R}} |K(v)|\,\mathrm{d}v = O(1),$$
$$\int_0^1 |f(u)K_b'(x-u)|\,\mathrm{d}u \le \frac{\|f\|_\infty}{b} \int_{\mathbb{R}} |K'(v)|\,\mathrm{d}v = O(b^{-1}).$$

Thus

$$\int_0^1 |h_b'(u)|\,\mathrm{d}u = O(1) + O(b^{-1}) = O(b^{-1}).$$

In the quadrature bound, the first product-rule term contributes $O(n^{-1})$, while the second contributes $O((nb)^{-1})$. For $b \le 1$, their total is $O((nb)^{-1})$. The quadrature bound therefore gives

$$\frac{1}{n-1}\sum_{i=1}^{n} h_b(x_i) = \int_0^1 h_b(u)\,\mathrm{d}u + O\left(\frac{1}{(n-1)b}\right). \tag{22}$$

Since

$$\frac{1}{n}\sum_{i=1}^{n} h_b(x_i) = \frac{n-1}{n}\left\{\frac{1}{n-1}\sum_{i=1}^{n} h_b(x_i)\right\}$$

and $\int_0^1 |h_b(u)|\,\mathrm{d}u = O(1)$, it follows that

$$\frac{1}{n}\sum_{i=1}^{n} h_b(x_i) = \int_0^1 h_b(u)\,\mathrm{d}u + O\left(\frac{1}{n} + \frac{1}{nb}\right) = \int_0^1 h_b(u)\,\mathrm{d}u + O\left(\frac{1}{nb}\right), \tag{23}$$

where the final equality holds for $b \le 1$.

At an interior point, a change of variables gives

$$\int_0^1 f(u)K_b(x-u)\,\mathrm{d}u = \int_{\mathbb{R}} f(x-bv)K(v)\,\mathrm{d}v. \tag{24}$$

Taylor expansion gives

$$f(x-bv) = \sum_{j=0}^{r} \frac{(-bv)^j}{j!} f^{(j)}(x) + o(b^r) \tag{25}$$

uniformly for v in the support of K. The moment conditions then give

$$\int_{\mathbb{R}} f(x-bv)K(v)\,\mathrm{d}v = f(x) + \frac{(-1)^r \mu_r(K)}{r!} f^{(r)}(x) b^r + o(b^r). \tag{26}$$

Combining (17), (23), and (26) proves the result. □

The two bias contributions have relative size

$$\frac{(nb)^{-1}}{b^r} = \frac{1}{nb^{r+1}}.$$

Thus, if

$$nb^{r+1} \longrightarrow \infty, \tag{27}$$

the grid approximation term is $o(b^r)$ and (20) simplifies to

$$\mathrm{Bias}\{\widehat{f}_b(x)\} = B_r(x) b^r + o(b^r). \tag{28}$$

The bandwidth obtained below satisfies (27).

4 Variance, mean squared error, and bandwidth balance

By independence of the errors,

$$\mathrm{Var}\{\widehat{f}_b(x)\} = \frac{\sigma^2}{n^2 b^2} \sum_{i=1}^{n} K^2\left(\frac{x - x_i}{b}\right). \tag{29}$$

To approximate this sum, apply the regular-grid quadrature bound to

$$g_b(u) = K^2\left(\frac{x-u}{b}\right).$$

For all sufficiently small b, the support of g_b lies inside $(0, 1)$, so $g_b(0) = g_b(1) = 0$. Its derivative is

$$g_b'(u) = -\frac{2}{b} K\left(\frac{x-u}{b}\right) K'\left(\frac{x-u}{b}\right),$$

and the change of variables $v = (x - u)/b$ gives

$$\int_0^1 |g_b'(u)| \, \mathrm{d}u = 2 \int_{\mathbb{R}} |K(v)K'(v)| \, \mathrm{d}v = O(1). \tag{30}$$

The quadrature bound therefore yields

$$\frac{1}{n-1} \sum_{i=1}^{n} g_b(x_i) = \int_0^1 g_b(u) \, \mathrm{d}u + O\left(\frac{1}{n}\right). \tag{31}$$

The integral on the right is

$$\int_0^1 g_b(u) \, \mathrm{d}u = b \int_{\mathbb{R}} K(v)^2 \, \mathrm{d}v \tag{32}$$

$$= bR(K), \tag{33}$$

where

$$R(K) := \int_{\mathbb{R}} K(v)^2 \, \mathrm{d}v \tag{34}$$

is the squared L^2 norm of the kernel. Combining (31) and (33), and then changing from the exact mesh weight to $1/n$, gives

$$\frac{1}{n} \sum_{i=1}^{n} K^2\left(\frac{x - x_i}{b}\right) = \frac{n-1}{n} \left\{ \frac{1}{n-1} \sum_{i=1}^{n} g_b(x_i) \right\} \tag{35}$$

$$= \frac{n-1}{n} \left\{ bR(K) + O\left(\frac{1}{n}\right) \right\} \tag{36}$$

$$= bR(K) + O\left(\frac{1}{n}\right). \tag{37}$$

Equivalently,

$$\sum_{i=1}^{n} K^2\left(\frac{x - x_i}{b}\right) = nbR(K) + O(1). \tag{38}$$

Substitution into (29) gives

$$\mathrm{Var}\{\widehat{f}_b(x)\} = \frac{\sigma^2 R(K)}{nb} + O\left(\frac{1}{n^2 b^2}\right). \tag{39}$$

If $nb \to \infty$, the remainder is $o((nb)^{-1})$, and hence

$$\mathrm{Var}\{\widehat{f}_b(x)\} = \frac{\sigma^2 R(K)}{nb} + o\left(\frac{1}{nb}\right). \tag{40}$$

Approximately nb design points fall inside a window of width b, so the local averaging variance is of order $(nb)^{-1}$.

Combining (28) and (40),

$$\mathrm{MSE}\{\widehat{f}_b(x)\} = B_r(x)^2 b^{2r} + \frac{\sigma^2 R(K)}{nb} + o\left(b^{2r} + \frac{1}{nb}\right). \tag{41}$$

The first term decreases as b decreases, whereas the second increases. The leading terms are balanced by

$$b^{2r} \asymp \frac{1}{nb}, \qquad \text{hence} \qquad b \asymp n^{-1/(2r+1)}. \tag{42}$$

At a point where $B_r(x) \neq 0$, direct differentiation gives the asymptotically optimal bandwidth

$$b_{\mathrm{opt}}(x) = \left\{\frac{\sigma^2 R(K)}{2rB_r(x)^2 n}\right\}^{1/(2r+1)}. \tag{43}$$

Substitution into (41) gives

$$\mathrm{MSE}\{\widehat{f}_{b_{\mathrm{opt}}}(x)\} \asymp n^{-2r/(2r+1)}. \tag{44}$$

Because mean squared error has the square of the physical units of f, it is often more interpretable to state the result in root mean squared error:

$$\mathrm{RMSE}\{\widehat{f}_{b_{\mathrm{opt}}}(x)\} \asymp n^{-r/(2r+1)}. \tag{45}$$

The RMSE has the same units as the estimand $f(x)$.

Remark 4.1 (Integrated risk). The same calculation can be integrated over an interior subinterval. If

$$\mathrm{MISE}(b) = \mathbb{E}\int (\widehat{f}_b(x) - f(x))^2 \,\mathrm{d}x,$$

then

$$\mathrm{MISE}(b) = \frac{\mu_r(K)^2}{(r!)^2} b^{2r} \int \{f^{(r)}(x)\}^2 \,\mathrm{d}x + \frac{\sigma^2 R(K)}{nb} + o\left(b^{2r} + \frac{1}{nb}\right). \tag{46}$$

Thus the integrated squared risk has the same rate as the pointwise MSE.

5 The effect of dimension

The one-dimensional bias–variance calculation extends directly to d dimensions. For a regular design in $[0,1]^d$, consider

$$\widehat{f}_b(x) = \frac{1}{nb^d}\sum_{i=1}^{n} Y_i K\left(\frac{x - x_i}{b}\right), \tag{47}$$

where K is a multivariate kernel. A ball or cube of radius b has volume of order b^d, so a neighborhood of this size contains about nb^d observations. The bias and variance therefore have the orders

$$\text{squared bias} \asymp b^{2r}, \qquad \text{variance} \asymp \frac{1}{nb^d}. \tag{48}$$

Balancing these terms gives

$$b \asymp n^{-1/(2r+d)}, \qquad \mathrm{MSE} \asymp n^{-2r/(2r+d)}, \qquad \mathrm{RMSE} \asymp n^{-r/(2r+d)}. \tag{49}$$

The exponent decreases as d increases, expressing the classical curse of dimensionality.

6 Stone's minimax result

The rates in (49) are the optimal minimax rates over standard smoothness classes.

Definition 6.1 (Hölder continuity and Hölder classes)**.** Let $0 < \alpha \leq 1$, and write $\|g\|_\infty = \sup_{x\in[0,1]^d} |g(x)|$. A function $g : [0,1]^d \to \mathbb{R}$ is *Hölder continuous of order α with constant C* if

$$|g(x) - g(y)| \leq C\|x - y\|_2^\alpha \qquad \text{for all } x, y \in [0,1]^d. \tag{50}$$

For $r > 0$, write $r = s + \alpha$, where s is a nonnegative integer and $0 < \alpha \leq 1$; when r is an integer, take $s = r - 1$ and $\alpha = 1$. For a multi-index $\beta = (\beta_1, \ldots, \beta_d)$, write $|\beta| = \beta_1 + \cdots + \beta_d$ and $D^\beta f$ for the corresponding mixed partial derivative. Define

$$\begin{aligned} \|f\|_{\mathcal{H}^r} = \max_{|\beta|\leq s} \|D^\beta f\|_\infty& \\ + \max_{|\beta|=s} \sup_{x\neq y} \frac{|D^\beta f(x) - D^\beta f(y)|}{\|x - y\|_2^\alpha}&. \end{aligned} \tag{51}$$

The Hölder class $\mathcal{H}^r(L)$ consists of the functions for which all derivatives in (51) exist and are continuous and for which $\|f\|_{\mathcal{H}^r} \leq L$. In particular, every derivative of total order s is Hölder continuous of order α.

Fix an interior point x_0. Under standard regular-design and noise assumptions, Stone's pointwise result [21] gives

$$\inf_{\widetilde{f}} \sup_{f\in\mathcal{H}^r(L)} \mathbb{E}_f\left[\{\widetilde{f}(x_0) - f(x_0)\}^2\right] \asymp n^{-2r/(2r+d)}. \tag{52}$$

Equivalently, the optimal pointwise root mean squared error satisfies

$$\inf_{\widetilde{f}} \sup_{f\in\mathcal{H}^r(L)} \left(\mathbb{E}_f\left[\{\widetilde{f}(x_0) - f(x_0)\}^2\right]\right)^{1/2} \asymp n^{-r/(2r+d)}. \tag{53}$$

Stone's global minimax theorem [22] gives the same algebraic rate for integrated squared error:

$$\inf_{\widetilde{f}} \sup_{f\in\mathcal{H}^r(L)} \mathbb{E}_f\|\widetilde{f} - f\|_{L^2}^2 \asymp n^{-2r/(2r+d)}. \tag{54}$$

The corresponding root integrated mean squared error satisfies

$$\inf_{\widetilde{f}} \sup_{f\in\mathcal{H}^r(L)} \left\{\mathbb{E}_f\|\widetilde{f} - f\|_{L^2}^2\right\}^{1/2} \asymp n^{-r/(2r+d)}. \tag{55}$$

The explicit kernel calculation assumes an even integer r and a kernel of order r. Stone's results also cover noninteger smoothness and supply minimax lower bounds together with achievable upper bounds. Modern accounts include Tsybakov [23]; kernel treatments include Wand and Jones [25].

7 The smoothing-spline optimization problem

We now return to one dimension and assume

$$0 = x_1 < x_2 < \cdots < x_n = 1. \tag{56}$$

For $\lambda > 0$, consider

$$J_\lambda(f) = \sum_{i=1}^{n} \{Y_i - f(x_i)\}^2 + \lambda \int_0^1 \{f''(x)\}^2 \,\mathrm{d}x. \tag{57}$$

The natural domain is the Sobolev space

$$H^2[0,1] = \{f : f' \text{ is absolutely continuous and } f'' \in L^2[0,1]\}. \tag{58}$$

In one dimension, functions in H^2 have continuous f and f', so the point evaluations $f(x_i)$ are well defined.

The two terms in (57) pull in opposite directions:

- the residual sum of squares rewards agreement with the data;
- the roughness penalty discourages curvature.

The null space of the penalty consists of affine functions, because

$$\int_0^1 (f'')^2 = 0 \quad \iff \quad f'' = 0 \text{ a.e.} \quad \iff \quad f(x) = a + bx.$$

Thus $\lambda \to \infty$ drives the solution toward the least-squares affine fit. As $\lambda \downarrow 0$, the solution tends to the natural cubic interpolating spline, the least-rough interpolant of the data.

8 First variation and the piecewise-cubic form

Let $g \in H^2[0,1]$ and perturb f to $f + \epsilon g$. Differentiating (57) gives

$$\frac{1}{2}\,\frac{\mathrm{d}}{\mathrm{d}\epsilon} J_\lambda(f + \epsilon g)\Big|_{\epsilon=0} = \sum_{i=1}^{n} \{f(x_i) - Y_i\} g(x_i) + \lambda \int_0^1 f''(x) g''(x) \,\mathrm{d}x. \tag{59}$$

At a minimizer this expression is zero for every admissible perturbation g.

8.1 Reduction to a finite dimensional optimization problem

We reduce the problem to finite dimension by using the general form of the perturbation to infer the cubic spline nature of the solution, which now depends on only finitely many parameters, and we then optimize over any that are not otherwise constrained.

To start, let the variables a_0, a_1, b_0, b_1 correspond to the specification of the value of the function and its derivative at the endpoints of the interval.

$$f(0) = a_0, \qquad f'(0) = a_1, \qquad f(1) = b_0, \qquad f'(1) = b_1, \tag{60}$$

With the quantities in (60) held fixed, admissible perturbations satisfy

$$g(0) = g'(0) = g(1) = g'(1) = 0. \tag{61}$$

To determine the form of the minimizer on one knot interval, fix i and restrict the first-variation identity to perturbations

$$g \in C_c^\infty((x_i, x_{i+1})).$$

Every such g satisfies (61). It also satisfies

$$g(x_j) = 0, \qquad j = 1, \ldots, n,$$

so the data-fit sum in (59) vanishes. Moreover, g'' is supported in (x_i, x_{i+1}), so the roughness integral is confined to that interval. Since (59) is zero for this class of perturbations and $\lambda > 0$, it follows that

$$\int_{x_i}^{x_{i+1}} f''(x)g''(x)\,\mathrm{d}x = 0, \qquad g \in C_c^\infty((x_i, x_{i+1})). \tag{62}$$

By the definition of the distributional fourth derivative, (62) says that

$$f^{(4)} = 0 \qquad \text{on } (x_i, x_{i+1}). \tag{63}$$

Therefore the minimizer is a polynomial of degree at most three on every interval between consecutive design points. The next step is to write these polynomial pieces explicitly and minimize the original functional over their coefficients.

9 Determining the piecewise-cubic coefficients

Let

$$h_i = x_{i+1} - x_i, \qquad i = 1, \ldots, n-1. \tag{64}$$

On the interval $[x_i, x_{i+1}]$, write the cubic piece as

$$p_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3, \qquad i = 1, \ldots, n-1. \tag{65}$$

There are $n - 1$ intervals and four coefficients on each interval, so there are

$$4(n-1) \tag{66}$$

coefficients to determine.

Since $f \in H^2[0,1]$, both f and f' are continuous. At each interior knot x_{i+1}, the pieces p_i and p_{i+1} must therefore satisfy

$$a_{i+1} = a_i + b_i h_i + c_i h_i^2 + d_i h_i^3, \tag{67}$$
$$b_{i+1} = b_i + 2c_i h_i + 3d_i h_i^2, \tag{68}$$

for $i = 1, \ldots, n-2$. There are $n - 2$ interior knots and two equations at each one, giving $2(n-2)$ matching equations. The number of remaining degrees of freedom is therefore

$$4(n-1) - 2(n-2) = 2n. \tag{69}$$

9.1 Knot values and slopes

Use the common value and slope at each knot as the remaining $2n$ unknowns:

$$z_i = f(x_i), \qquad v_i = f'(x_i), \qquad i = 1, \dots, n. \tag{70}$$

For the interval $[x_i, x_{i+1}]$, the four conditions

$$p_i(x_i) = z_i, \qquad p_i'(x_i) = v_i, \qquad p_i(x_{i+1}) = z_{i+1}, \qquad p_i'(x_{i+1}) = v_{i+1}$$

determine its four coefficients uniquely. In the representation (65),

$$\begin{aligned} a_i &= z_i, & b_i &= v_i, \\ c_i &= \frac{3(z_{i+1} - z_i)}{h_i^2} - \frac{2v_i + v_{i+1}}{h_i}, & d_i &= \frac{2(z_i - z_{i+1})}{h_i^3} + \frac{v_i + v_{i+1}}{h_i^2}. \end{aligned} \tag{71}$$

Equivalently, with $u = (x - x_i)/h_i$,

$$p_i(x) = H_{00}(u) z_i + h_i H_{10}(u) v_i + H_{01}(u) z_{i+1} + h_i H_{11}(u) v_{i+1}, \tag{72}$$

where

$$\begin{aligned} H_{00}(u) &= 2u^3 - 3u^2 + 1, & H_{10}(u) &= u^3 - 2u^2 + u, \\ H_{01}(u) &= -2u^3 + 3u^2, & H_{11}(u) &= u^3 - u^2. \end{aligned}$$

Thus every choice of $(z_1, v_1, \dots, z_n, v_n)$ produces a C^1 piecewise-cubic function satisfying all $2(n-2)$ matching equations.

There is no loss in carrying out the minimization over this finite-dimensional family. To see this, take any $h \in H^2[0,1]$ and let p be its piecewise Hermite interpolant: on each interval $[x_i, x_{i+1}]$, the cubic p_i matches the values and slopes of h at both endpoints. Then

$$p(x_i) = h(x_i), \qquad i = 1, \dots, n,$$

so p and h have the same residual sum of squares. Write $w = h - p$. On each interval, w and w' vanish at both endpoints, while $p_i^{(4)} = 0$. Two integrations by parts therefore give

$$\int_{x_i}^{x_{i+1}} p_i''(x) w''(x)\,\mathrm{d}x = [p_i'' w' - p_i''' w]_{x_i}^{x_{i+1}} + \int_{x_i}^{x_{i+1}} p_i^{(4)}(x) w(x)\,\mathrm{d}x = 0. \tag{73}$$

Summing over the intervals yields

$$\int_0^1 \{h''(x)\}^2\,\mathrm{d}x = \int_0^1 \{p''(x)\}^2\,\mathrm{d}x + \int_0^1 \{w''(x)\}^2\,\mathrm{d}x \ge \int_0^1 \{p''(x)\}^2\,\mathrm{d}x. \tag{74}$$

Consequently, $J_\lambda(p) \le J_\lambda(h)$. It is therefore sufficient to minimize J_λ over the finite-dimensional C^1 piecewise-cubic family.

For later use, define the second derivatives at the two ends of the ith interval and its constant third derivative by

$$L_i = p_i''(x_i), \qquad R_i = p_i''(x_{i+1}), \qquad T_i = p_i'''(x), \quad x_i < x < x_{i+1}. \tag{75}$$

Using (71), these quantities are

$$L_i = \frac{6(z_{i+1} - z_i)}{h_i^2} - \frac{4v_i + 2v_{i+1}}{h_i}, \tag{76}$$

$$R_i = \frac{6(z_i - z_{i+1})}{h_i^2} + \frac{2v_i + 4v_{i+1}}{h_i}, \tag{77}$$

$$T_i = \frac{12(z_i - z_{i+1})}{h_i^3} + \frac{6(v_i + v_{i+1})}{h_i^2}. \tag{78}$$

9.2 The stationarity equations

The $2n$ variables in (70) are determined by minimizing J_λ over the piecewise-cubic family. Let δz_i and δv_i be arbitrary changes in the knot values and slopes, and let q_i be the corresponding change in p_i. Thus

$$q_i(x_i) = \delta z_i, \qquad q_i'(x_i) = \delta v_i, \qquad q_i(x_{i+1}) = \delta z_{i+1}, \qquad q_i'(x_{i+1}) = \delta v_{i+1}.$$

Since $p_i^{(4)} = 0$, integration by parts on one interval gives

$$\begin{aligned} \int_{x_i}^{x_{i+1}} p_i''(x) q_i''(x)\,\mathrm{d}x &= \left[p_i'' q_i' - p_i''' q_i\right]_{x_i}^{x_{i+1}} \\ &= R_i\,\delta v_{i+1} - T_i\,\delta z_{i+1} - L_i\,\delta v_i + T_i\,\delta z_i. \end{aligned} \tag{79}$$

Consequently, the first variation of the finite-dimensional objective is

$$\begin{aligned} \frac{1}{2}\,\delta J_\lambda = &\sum_{i=1}^{n} (z_i - Y_i)\,\delta z_i \\ &+ \lambda \sum_{i=1}^{n-1} \left\{R_i\,\delta v_{i+1} - T_i\,\delta z_{i+1} - L_i\,\delta v_i + T_i\,\delta z_i\right\}. \end{aligned} \tag{80}$$

The coefficients of the n slope variations give

$$L_1 = 0, \qquad R_{i-1} = L_i \quad (2 \le i \le n-1), \qquad R_{n-1} = 0. \tag{81}$$

The coefficients of the n value variations give

$$\begin{aligned} z_1 - Y_1 + \lambda T_1 &= 0, \\ z_i - Y_i + \lambda(T_i - T_{i-1}) &= 0, \quad 2 \le i \le n-1, \\ z_n - Y_n - \lambda T_{n-1} &= 0. \end{aligned} \tag{82}$$

Equations (81) and (82) supply the remaining $n + n = 2n$ equations. Together with the $2(n-2)$ matching equations, they give

$$2(n-2) + 2n = 4(n-1)$$

equations for the $4(n-1)$ interval coefficients.

9.3 Uniqueness and computation

Order the remaining unknowns as

$$\theta = (z_1, v_1, z_2, v_2, \dots, z_n, v_n)^{\mathsf{T}}, \qquad y_H = (Y_1, 0, Y_2, 0, \dots, Y_n, 0)^{\mathsf{T}}. \tag{83}$$

Because the quantities L_i, R_i, T_i are linear in the four variables at the ends of interval i, the stationarity equations form a linear system

$$A_\lambda \theta = y_H, \tag{84}$$

where A_λ is one half of the Hessian of J_λ with respect to θ.

Theorem 9.1 (Existence and uniqueness)**.** *For $\lambda > 0$ and distinct design points, A_λ is symmetric positive definite. Hence* (84) *has a unique solution.*

Proof. Let $\eta = (\alpha_1, \beta_1, \dots, \alpha_n, \beta_n)^{\mathsf{T}}$ be a perturbation of the knot values and slopes, and let g be the corresponding C^1 piecewise-cubic function. The quadratic form associated with A_λ is

$$\eta^{\mathsf{T}} A_\lambda \eta = \sum_{i=1}^{n} \alpha_i^2 + \lambda \int_0^1 \{g''(x)\}^2 \, \mathrm{d}x. \tag{85}$$

If this quantity is zero, then every $\alpha_i = 0$ and $g'' = 0$ on every interval. Hence g is affine on each interval. Continuity of g' makes these affine pieces a single affine function on $[0, 1]$. Since it vanishes at the two distinct points x_1 and x_n, it is identically zero, so every $\beta_i = 0$ as well. Thus (85) is positive for every nonzero η. □

The functional is therefore strictly convex in the $2n$ remaining variables. The unique solution of (84) is the unique piecewise-cubic minimizer. The reduction in (73)–(74) shows that it is also the unique minimizer of (57) over $H^2[0, 1]$.

Each equation at an interior knot involves only the variables at that knot and its two neighbors. With the ordering in (83), A_λ is block tridiagonal with 2×2 blocks. A banded Cholesky factorization therefore finds the solution in $O(n)$ operations. Once the knot values and slopes are known, the fitted curve is evaluated interval by interval using (65), (71), or equivalently (72).

10 Natural cubic-spline properties

The equations used to determine the coefficients also identify the matching and endpoint properties of the fitted curve.

10.1 Second derivatives and the natural endpoints

By definition, L_i and R_i are the second derivatives of the ith cubic piece at its left and right endpoints. The slope equations (81) therefore give

$$f''(0) = 0, \qquad f''(x_i-) = f''(x_i+) \quad (2 \le i \le n-1), \qquad f''(1) = 0. \tag{86}$$

Thus the fitted function is C^2 and satisfies the natural boundary conditions

$$f''(0) = f''(1) = 0. \tag{87}$$

It is therefore a natural cubic spline with knots among the design points.

10.2 Residual relations

The quantity T_i is the constant third derivative on (x_i, x_{i+1}). The value equations (82) can therefore be written as

$$f(x_i) - Y_i + \lambda\{f'''(x_i+) - f'''(x_i-)\} = 0, \qquad 2 \le i \le n-1, \tag{88}$$

and

$$f(0) - Y_1 + \lambda f'''(0+) = 0, \qquad f(1) - Y_n - \lambda f'''(1-) = 0. \tag{89}$$

The third derivative is constant on each knot interval and may change from one interval to the next; the change is determined by the residual at the intervening knot through (88).

Theorem 10.1 (Characterization of the smoothing-spline minimizer)**.** *For $\lambda > 0$ and distinct design points, the unique minimizer of* (57) *is the natural cubic spline satisfying* (88) *and* (89)*.*

The classical historical development includes Schoenberg [16, 17] and Reinsch [13]; a general Hilbert-space formulation is given by Kimeldorf and Wahba [9].

11 A reduced system for the natural spline

The system (84) solves directly for the knot values and slopes. The natural-spline conditions allow a smaller system based on the knot values

$$z_i = f(x_i)$$

and the second derivatives

$$M_i = f''(x_i), \qquad i = 1, \dots, n. \tag{90}$$

By (87),

$$M_1 = M_n = 0. \tag{91}$$

11.1 The cubic on one interval

Since f'' is linear on $[x_i, x_{i+1}]$,

$$f''(x) = M_i \frac{x_{i+1} - x}{h_i} + M_{i+1} \frac{x - x_i}{h_i}. \tag{92}$$

Integrating twice and imposing $f(x_i) = z_i$ and $f(x_{i+1}) = z_{i+1}$ gives

$$\begin{aligned} f(x) = {} & \frac{M_i (x_{i+1} - x)^3}{6h_i} + \frac{M_{i+1}(x - x_i)^3}{6h_i} \\ & + \left(z_i - \frac{h_i^2 M_i}{6} \right) \frac{x_{i+1} - x}{h_i} \\ & + \left(z_{i+1} - \frac{h_i^2 M_{i+1}}{6} \right) \frac{x - x_i}{h_i}, \qquad x_i \le x \le x_{i+1}. \end{aligned} \tag{93}$$

Thus the fitted curve is explicit once z and M are known.

11.2 Equations linking knot values and second derivatives

Continuity of f' at the interior knots gives

$$\frac{h_{i-1}}{6}M_{i-1} + \frac{h_{i-1}+h_i}{3}M_i + \frac{h_i}{6}M_{i+1} = \frac{z_{i+1}-z_i}{h_i} - \frac{z_i - z_{i-1}}{h_{i-1}}, \tag{94}$$

for $2 \le i \le n-1$.

Let $M = (M_2, \dots, M_{n-1})^\mathsf{T} \in \mathbb{R}^{n-2}$. Define the $n \times (n-2)$ matrix Q as follows. Its kth column corresponds to $i = k+1$ and has nonzero entries

$$Q_{i-1,k} = \frac{1}{h_{i-1}}, \qquad Q_{i,k} = -\left(\frac{1}{h_{i-1}} + \frac{1}{h_i}\right), \qquad Q_{i+1,k} = \frac{1}{h_i}. \tag{95}$$

Define the symmetric tridiagonal matrix $R \in \mathbb{R}^{(n-2)\times(n-2)}$ by

$$R_{kk} = \frac{h_k + h_{k+1}}{3}, \qquad R_{k,k+1} = R_{k+1,k} = \frac{h_{k+1}}{6}. \tag{96}$$

Then (94) is

$$RM = Q^\mathsf{T} z. \tag{97}$$

On $[x_i, x_{i+1}]$, the third derivative is

$$f'''(x) = \frac{M_{i+1} - M_i}{h_i}.$$

The definition of Q therefore gives

$$\begin{aligned} (QM)_1 &= f'''(0+), \\ (QM)_i &= f'''(x_i+) - f'''(x_i-), \qquad 2 \le i \le n-1, \\ (QM)_n &= -f'''(1-). \end{aligned}$$

Consequently, the residual conditions (88)–(89) become

$$z + \lambda QM = Y. \tag{98}$$

Equations (97) and (98) form a complete system for z and M.

11.3 Solution of the reduced system

Eliminating z gives

$$(R + \lambda Q^\mathsf{T} Q)M = Q^\mathsf{T} Y, \tag{99}$$

followed by

$$z = Y - \lambda QM. \tag{100}$$

To see that the coefficient matrix is positive definite, note that

$$M^\mathsf{T} RM = \sum_{i=1}^{n-1} \frac{h_i}{3}\{M_i^2 + M_i M_{i+1} + M_{i+1}^2\}, \tag{101}$$

where $M_1 = M_n = 0$. The right side is strictly positive unless $M = 0$. Therefore R and $R + \lambda Q^\mathsf{T} Q$ are positive definite, and (99) has a unique solution.

The matrix R is tridiagonal and $Q^\mathsf{T} Q$ is pentadiagonal. Banded Cholesky factorization solves (99) in $O(n)$ operations. After computing M and z, the spline is evaluated from (93).

The fitted values are linear in Y:

$$z = S_\lambda Y, \qquad S_\lambda = I - \lambda Q(R + \lambda Q^\mathsf{T} Q)^{-1} Q^\mathsf{T}. \tag{102}$$

Remark 11.1 (Equally spaced knots). If every $h_i = h$, then

$$R = \frac{h}{6} \begin{pmatrix} 4 & 1 & & \\ 1 & 4 & 1 & \\ & \ddots & \ddots & \ddots \\ & & 1 & 4 \end{pmatrix},$$

while $Q^\mathsf{T} z$ is the vector of second differences divided by h:

$$(Q^\mathsf{T} z)_k = \frac{z_k - 2z_{k+1} + z_{k+2}}{h}.$$

The roughness penalty is therefore a continuous analogue of a squared second-difference penalty.

12 Cubic B-splines

The reduced system above uses knot values and second derivatives. A second computational route represents the same natural cubic spline in a local B-spline basis, whose modern recursive evaluation is associated with Cox [1] and de Boor [4]; foundational basis results go back to Curry and Schoenberg [3]. Standard references are de Boor [5] and Schumaker [18].

12.1 The clamped knot vector

For the breakpoints $x_1 < \cdots < x_n$, form the clamped cubic knot sequence

$$\boldsymbol{\tau} = (x_1, x_1, x_1, x_1, x_2, \ldots, x_{n-1}, x_n, x_n, x_n, x_n). \tag{103}$$

There are $n + 6$ knots when multiplicity is counted. A degree-p B-spline basis associated with N knots has $N - p - 1$ members. For $p = 3$, there are therefore

$$n + 6 - 3 - 1 = n + 2 \tag{104}$$

cubic B-splines.

12.2 The Cox–de Boor recursion

Let the knot sequence be

$$\tau_1 \le \tau_2 \le \cdots \le \tau_N.$$

Define degree-zero B-splines by

$$B_{j,0}(x) = \begin{cases} 1, & \tau_j \le x < \tau_{j+1}, \\ 0, & \text{otherwise}, \end{cases} \tag{105}$$

with the conventional modification at the right endpoint so that the final basis function takes the value one at $x = \tau_N$. For $p \ge 1$, define recursively

$$\begin{aligned} B_{j,p}(x) = & \frac{x - \tau_j}{\tau_{j+p} - \tau_j} B_{j,p-1}(x) \\ & + \frac{\tau_{j+p+1} - x}{\tau_{j+p+1} - \tau_{j+1}} B_{j+1,p-1}(x), \end{aligned} \tag{106}$$

where a fraction with zero denominator is interpreted as zero. For the knot vector (103), the cubic basis is

$$B_j(x) = B_{j,3}(x), \qquad j = 1, \dots, n+2. \tag{107}$$

12.3 Support and nonnegativity

Proposition 12.1. *For every $p \ge 0$,*

$$\operatorname{supp}(B_{j,p}) \subseteq [\tau_j, \tau_{j+p+1}], \tag{108}$$

and $B_{j,p}(x) \ge 0$.

Proof. For $p = 0$ both assertions follow from (105). Suppose they hold for degree $p - 1$. The first term in (106) is supported on $[\tau_j, \tau_{j+p}]$, and the second on $[\tau_{j+1}, \tau_{j+p+1}]$; their union is contained in $[\tau_j, \tau_{j+p+1}]$. On the support of each lower-degree B-spline, the corresponding coefficient in (106) lies between zero and one. Hence nonnegativity is preserved. □

A cubic B-spline is consequently supported on at most four consecutive knot intervals. This local support is the source of sparse matrices.

12.4 Partition of unity

Proposition 12.2. *On the interior of the knot range,*

$$\sum_j B_{j,p}(x) = 1. \tag{109}$$

Proof. The claim is immediate for $p = 0$, because exactly one interval indicator is one away from a knot. Sum (106) over j and reindex the second sum. For a fixed $B_{j,p-1}$, the two coefficients that arise are

$$\frac{x - \tau_j}{\tau_{j+p} - \tau_j} \quad \text{and} \quad \frac{\tau_{j+p} - x}{\tau_{j+p} - \tau_j},$$

which add to one. The result follows by induction. □

Thus a B-spline expansion with all coefficients equal to a constant c reproduces the constant function c.

12.5 Derivative formula

Differentiating the recursion and simplifying gives

$$\frac{\mathrm{d}}{\mathrm{d}x} B_{j,p}(x) = \frac{p}{\tau_{j+p} - \tau_j} B_{j,p-1}(x) - \frac{p}{\tau_{j+p+1} - \tau_{j+1}} B_{j+1,p-1}(x). \tag{110}$$

The same zero-denominator convention applies. The formula can be proved by induction: differentiate (106), apply the induction hypothesis to the lower-degree derivatives, and use the degree-$p-1$ recursion to cancel the remaining terms.

Repeated application gives $B''_{j,3}$ as a linear combination of degree-one B-splines. Hence $B''_{j,3}$ is piecewise linear, a fact used to compute the roughness matrix exactly.

12.6 Smoothness at knots

A degree-p B-spline is C^{p-m} at a knot of multiplicity m. This follows inductively from (110): differentiating once lowers the degree and the available continuity by one. In our clamped vector, every interior knot is simple, so each cubic B-spline is C^2 there. The fourfold endpoint repetitions allow the basis to represent arbitrary endpoint values and derivatives.

12.7 Local linear independence and the basis theorem

Proposition 12.3 (Local linear independence)**.** *Fix a nonempty elementary interval $(\tau_\ell, \tau_{\ell+1})$. If*

$$\sum_j c_j B_{j,p}(x) = 0 \quad \textit{for every } x \in (\tau_\ell, \tau_{\ell+1}), \tag{111}$$

then every coefficient c_j for which $B_{j,p}$ is nonzero on that interval is zero.

Proof. For $p = 0$, exactly one degree-zero B-spline is nonzero on the interval, so the result is immediate. Assume it for degree $p-1$. Differentiate (111) and use (110). After reindexing, the coefficient of an active $B_{j,p-1}$ is proportional to $c_j - c_{j-1}$. By the induction hypothesis, adjacent active coefficients are equal. Hence all coefficients of the active degree-p B-splines are equal to one common value c. By the partition of unity, the left side of (111) is then c, so $c = 0$. □

Let

$$\mathcal{S}_3 = \{s \in C^2[x_1, x_n] : s \text{ is cubic on each } [x_i, x_{i+1}]\}. \tag{112}$$

There are $n-1$ cubic pieces, giving $4(n-1)$ raw coefficients. At each of the $n-2$ interior knots, continuity of s, s', and s'' imposes three conditions. Therefore

$$\dim(\mathcal{S}_3) = 4(n-1) - 3(n-2) = n + 2. \tag{113}$$

The $n+2$ functions in (107) belong to $\mathcal{S}_3$ and are linearly independent by local linear independence. Thus they form a basis of $\mathcal{S}_3$.

13 The B-spline normal equations

Write

$$f(x) = \sum_{j=1}^{n+2} c_j B_j(x). \tag{114}$$

Define the $n \times (n+2)$ design matrix

$$A_{ij} = B_j(x_i) \tag{115}$$

and the $(n+2) \times (n+2)$ roughness matrix

$$\Omega_{jk} = \int_0^1 B_j''(x) B_k''(x)\,\mathrm{d}x. \tag{116}$$

Then

$$J_\lambda(c) = \|Y - Ac\|^2 + \lambda c^\mathsf{T} \Omega c. \tag{117}$$

Differentiating gives the normal equations

$$(A^\mathsf{T} A + \lambda\Omega)c = A^\mathsf{T} Y. \tag{118}$$

The coefficient matrix is positive definite. Indeed,

$$c^\mathsf{T}(A^\mathsf{T} A + \lambda\Omega)c = \sum_{i=1}^{n} f_c(x_i)^2 + \lambda \int_0^1 \{f_c''(x)\}^2\,\mathrm{d}x. \tag{119}$$

If this is zero, then f_c is affine and vanishes at all design points, so $f_c \equiv 0$; the basis property then gives $c = 0$.

The system is banded. From (108), B_j and B_k have disjoint supports when their indices are sufficiently far apart; for cubic splines, only a fixed number of neighboring basis functions overlap. Consequently, both $A^\mathsf{T} A$ and Ω have at most seven nonzero diagonals. Banded Cholesky factorization therefore solves (118) in $O(n)$ operations.

13.1 How the design and roughness matrices are computed

The Cox–de Boor recursion evaluates all nonzero $B_j(x_i)$ without forming global polynomials. At any fixed x, at most four cubic B-splines are nonzero, so every row of A contains at most four nonzero entries.

For Ω, use (110) twice. On each knot interval, B_j'' and B_k'' are linear, so their product is a polynomial of degree at most two. Therefore the contribution

$$\int_{x_i}^{x_{i+1}} B_j''(x) B_k''(x)\,\mathrm{d}x \tag{120}$$

can be evaluated analytically or exactly by two-point Gauss–Legendre quadrature. Only overlapping supports need be considered.

The basis (114) spans all C^2 piecewise cubics. Minimizing (117) over this space gives the same unique function constructed in section 9; by section 10, the resulting coefficients satisfy the natural boundary conditions. The endpoint values and slopes are components of the fitted spline and are determined by the normal equations.

13.2 Relationship with the reduced system

The B-spline normal equations and the Q, R equations are two coordinate descriptions of the same finite-dimensional minimization problem. The Q, R formulation uses the fitted values z_i and second derivatives M_i, while the B-spline formulation uses locally supported basis coefficients. Both lead to banded linear systems and the same fitted function.

14 Choosing the smoothing parameter in practice

Both formulations show that the spline estimate is linear in the response vector:

$$\widehat{Y} = S_\lambda Y. \tag{121}$$

In the reduced natural-spline formulation, S_λ is given by (102). In the B-spline formulation,

$$S_\lambda = A(A^\mathsf{T} A + \lambda\Omega)^{-1} A^\mathsf{T}. \tag{122}$$

The effective degrees of freedom are

$$\mathrm{df}(\lambda) = \mathrm{tr}(S_\lambda). \tag{123}$$

As $\lambda \downarrow 0$, $\mathrm{df}(\lambda) \to n$ and the fit approaches interpolation. As $\lambda \to \infty$, the effective degrees of freedom approach 2, the dimension of the affine null space.

For any linear smoother, the leave-one-out cross-validation score can be computed without refitting n times:

$$\mathrm{CV}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{Y_i - \widehat{Y}_i}{1-(S_\lambda)_{ii}}\right\}^2. \tag{124}$$

Generalized cross-validation replaces the individual diagonal terms by their average:

$$\mathrm{GCV}(\lambda) = \frac{n^{-1}\|(I - S_\lambda)Y\|^2}{\{1 - n^{-1}\,\mathrm{tr}(S_\lambda)\}^2}. \tag{125}$$

One chooses λ to minimize (125). This method was developed for smoothing splines by Craven and Wahba [2]; see also Wahba [24] and Green and Silverman [8].

15 Practical algorithms

Algorithm A: the reduced natural-spline system

Given sorted $x_1 < \cdots < x_n$, responses Y, and $\lambda > 0$:

1. Compute $h_i = x_{i+1} - x_i$.
2. Form the sparse matrices Q and R from (95) and (96).
3. Solve
$$(R + \lambda Q^\mathsf{T} Q)M = Q^\mathsf{T} Y.$$
4. Compute fitted knot values $z = Y - \lambda QM$.
5. Set $M_1 = M_n = 0$ and use (93) to evaluate the fitted curve on each interval.

All matrices have fixed bandwidth, so storage and arithmetic are linear in n.

Algorithm B: a cubic B-spline basis

1. Form the clamped knot vector (103).
2. Use the Cox–de Boor recursion to evaluate the nonzero $B_j(x_i)$ and construct the sparse design matrix A.
3. Use the derivative recursion to construct the banded roughness matrix Ω interval by interval.
4. Solve the symmetric positive definite banded system

$$(A^\mathsf{T} A + \lambda\Omega)c = A^\mathsf{T} Y.$$

5. Evaluate $\widehat{f}(x) = \sum_j c_j B_j(x)$ by the same recursion.

This formulation extends directly to nonuniform knots, higher degree, other derivative penalties, and tensor-product constructions.

16 The equivalent-kernel connection

Kernel smoothers and smoothing splines are closely connected. At interior points on a dense uniform design, a cubic smoothing spline behaves like a kernel smoother with a bandwidth determined by the penalty parameter. This viewpoint was developed in detail by Silverman [19, 20].

A simple continuous idealization makes the connection visible. Replace the discrete criterion by

$$\int_{\mathbb{R}} \{y(x) - f(x)\}^2 \, \mathrm{d}x + \rho \int_{\mathbb{R}} \{f''(x)\}^2 \, \mathrm{d}x. \tag{126}$$

The Euler equation is

$$f(x) + \rho f^{(4)}(x) = y(x). \tag{127}$$

Taking Fourier transforms gives

$$\widehat{f}(\omega) = \frac{1}{1+\rho\omega^4}\widehat{y}(\omega). \tag{128}$$

Thus f is a convolution of y with a kernel. Let $h = \rho^{1/4}$. The inverse transform of $(1+\omega^4)^{-1}$ is

$$K_{\mathrm{eq}}(u) = \frac{1}{2\sqrt{2}} e^{-|u|/\sqrt{2}} \left\{ \cos\left(\frac{|u|}{\sqrt{2}}\right) + \sin\left(\frac{|u|}{\sqrt{2}}\right) \right\}. \tag{129}$$

Consequently, the local form of the smoother is approximately

$$\widehat{f}(x) \approx \frac{1}{nh} \sum_{i=1}^{n} K_{\mathrm{eq}}\left(\frac{x - x_i}{h}\right) Y_i. \tag{130}$$

The formula in (129) shows that the equivalent kernel takes both positive and negative values.

For a Sobolev class with two square-integrable derivatives, the integrated squared bias is of order h^4, while the variance is of order $(nh)^{-1}$. Balancing gives

$$h \asymp n^{-1/5}, \qquad \mathrm{MISE} \asymp n^{-4/5}, \qquad \sqrt{\mathrm{MISE}} \asymp n^{-2/5}, \tag{131}$$

which is the Stone rate with $r = 2$ and $d = 1$.

Bibliography

[1] M. G. Cox. The numerical evaluation of B-splines. *IMA Journal of Applied Mathematics*, 10(2):134–149, 1972. doi:10.1093/imamat/10.2.134.

[2] P. Craven and G. Wahba. Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31(4):377–403, 1978. doi:10.1007/BF01404567.

[3] H. B. Curry and I. J. Schoenberg. On Pólya frequency functions IV: The fundamental spline functions and their limits. *Journal d'Analyse Mathématique*, 17(1):71–107, 1966. doi:10.1007/BF02788653.

[4] C. de Boor. On calculating with B-splines. *Journal of Approximation Theory*, 6(1):50–62, 1972. doi:10.1016/0021-9045(72)90080-9.

[5] C. de Boor. *A Practical Guide to Splines*. Applied Mathematical Sciences, vol. 27, revised edition. Springer, New York, 2001. doi:10.1007/978-1-4612-6333-3.

[6] R. L. Eubank. *Nonparametric Regression and Spline Smoothing*, 2nd ed. Marcel Dekker, New York, 1999.

[7] J. Fan and I. Gijbels. *Local Polynomial Modelling and Its Applications*. Chapman & Hall, London, 1996. doi:10.1201/9780203748725.

[8] P. J. Green and B. W. Silverman. *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman & Hall, London, 1994. doi:10.1201/b15710.

[9] G. S. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33(1):82–95, 1971. doi:10.1016/0022-247X(71)90184-3.

[10] E. A. Nadaraya. On estimating regression. *Theory of Probability and Its Applications*, 9(1):141–142, 1964. doi:10.1137/1109020.

[11] E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962. doi:10.1214/aoms/1177704472.

[12] M. B. Priestley and M. T. Chao. Non-parametric function fitting. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(3):385–392, 1972.

[13] C. H. Reinsch. Smoothing by spline functions. *Numerische Mathematik*, 10:177–183, 1967. doi:10.1007/BF02162161.

[14] M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3):832–837, 1956. doi:10.1214/aoms/1177728190.

[15] M. Rosenblatt. Conditional probability density and regression estimators. In P. R. Krishnaiah, editor, *Multivariate Analysis II*, pp. 25–31. Academic Press, New York, 1969. Proceedings of the Second International Symposium on Multivariate Analysis, Dayton, Ohio, June 17–22, 1968.

[16] I. J. Schoenberg. Contributions to the problem of approximation of equidistant data by analytic functions. Part A: On the problem of smoothing or graduation. A first class of analytic approximation formulae. *Quarterly of Applied Mathematics*, 4(1):45–99, 1946. doi:10.1090/qam/15914.

[17] I. J. Schoenberg. Spline functions and the problem of graduation. *Proceedings of the National Academy of Sciences of the United States of America*, 52(4):947–950, 1964. doi:10.1073/pnas.52.4.947.

[18] L. L. Schumaker. *Spline Functions: Basic Theory*, 3rd ed. Cambridge University Press, Cambridge, 2007. doi:10.1017/CBO9780511618994.

[19] B. W. Silverman. Spline smoothing: The equivalent variable kernel method. *The Annals of Statistics*, 12(3):898–916, 1984. doi:10.1214/aos/1176346710.

[20] B. W. Silverman. Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society: Series B (Methodological)*, 47(1):1–52, 1985. doi:10.1111/j.2517-6161.1985.tb01327.x.

[21] C. J. Stone. Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, 8(6):1348–1360, 1980. doi:10.1214/aos/1176345206.

[22] C. J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053, 1982. doi:10.1214/aos/1176345969.

[23] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, New York, 2009. doi:10.1007/b13794.

[24] G. Wahba. *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 59. Society for Industrial and Applied Mathematics, Philadelphia, 1990. doi:10.1137/1.9781611970128.

[25] M. P. Wand and M. C. Jones. *Kernel Smoothing*. Chapman & Hall, London, 1995. doi:10.1201/b14876.

[26] G. S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 26(4):359–372, 1964.