Biometrika (2008), **95**, 2, *pp.* 415–436
doi: 10.1093/biomet/asn010

Printed in Great Britain

On the asymptotics of penalized splines

By YINGXING LI

Department of Statistical Science, Malott Hall, Cornell University, New York 14853, U.S.A.
yl377@cornell.edu

AND DAVID RUPPERT

School of Operational Research and Information Engineering,
Rhodes Hall, Cornell University, New York 14853, U.S.A.
dr24@cornell.edu

Summary

We study the asymptotic behaviour of penalized spline estimators in the univariate case. We use B-splines and a penalty is placed on mth-order differences of the coefficients. The number of knots is assumed to converge to infinity as the sample size increases. We show that penalized splines behave similarly to Nadaraya–Watson kernel estimators with 'equivalent' kernels depending upon m. The equivalent kernels we obtain for penalized splines are the same as those found by Silverman for smoothing splines. The asymptotic distribution of the penalized spline estimator is Gaussian and we give simple expressions for the asymptotic mean and variance. Provided that it is fast enough, the rate at which the number of knots converges to infinity does not affect the asymptotic distribution. The optimal rate of convergence of the penalty parameter is given. Penalized splines are not design-adaptive.

Some key words: Asymptotic bias; Binning; B-spline; Difference penalty; Equivalent kernel; Increasing number of knots; P-spline.

1. Introduction

Suppose we have a univariate regression model $y_t = f(x_t) + \epsilon_t$, $t = 1, \ldots, n$, where, conditionally given x_t, ϵ_t has mean zero and variance $\sigma^2(x_t)$. For simplicity, we assume that the x_t are in [0, 1]. This paper presents an asymptotic theory of penalized spline estimators of f.

The model is $f(x) = \sum_{k=1}^{K(n)+p} b_k B_k^{[p]}(x)$, where $\{B_k^{[p]} : k = 1, \ldots, K(n) + p\}$ is the pth-degree B-spline basis with knots $0 = \kappa_0 < \kappa_1 < \cdots < \kappa_{K(n)} = 1$. The value of $K(n)$ will depend upon n as discussed below. The penalized least-squares estimator $\widehat{b} = (\widehat{b}_1, \ldots, \widehat{b}_{K(n)+p})'$ minimizes

$$\sum_{t=1}^{n} \left\{ y_t - \sum_{j=1}^{K(n)+p} b_j B_j^{[p]}(x_t) \right\}^2 + \lambda_n^* \sum_{k=m+1}^{K(n)+p} \{\Delta^m(b_k)\}^2, \quad \lambda_n^* \geqslant 0, \tag{1}$$

where Δ is the difference operator, that is, $\Delta b_k = b_k - b_{k-1}$, m is a positive integer, and $\Delta^m = \Delta(\Delta^{m-1})$. The nonparametric regression estimator $\widehat{f}(x) = \sum_{k=1}^{K(n)+p} \widehat{b}_k B_k^{[p]}(x)$ was introduced by Eilers & Marx (1996) and is called a P-spline.

Let $X^{[p]}$ be the $n \times \{K(n)+p\}$ matrix with (t, j)th entry equal to $B_j^{[p]}(x_t)$ and let $Y = (Y_1, \ldots, Y_n)'$. Define D^m as the $\{K(n)+p-m\} \times \{K(n)+p\}$ differencing matrix satisfying

$$D^m b = \begin{pmatrix} \Delta^m(b_{m+1}) \\ \vdots \\ \Delta^m\left(b_{K(n)+p}\right) \end{pmatrix}.$$

For simplicity of notation, the dependence of D^m on p will not be made explicit. Let $\Omega_n^{[p,m]} = (X^{[p]})'X^{[p]} + \lambda_n^*(D^m)'D^m$. Then, by (1), $\widehat{b}$ solves

$$\Omega_n^{[p,m]}\widehat{b} = \left(X^{[p]}\right)'Y. \tag{2}$$

This paper develops an asymptotic theory of P-splines for the cases $p = 0$ and 1 and $m = 1$ and 2, that is, piecewise-constant or linear splines, with a first- or second-order difference penalty. In § 5·1 we discuss possible extensions to higher-degree splines and higher-order penalties. One interesting, and perhaps surprising, result is that the rate of convergence of $\widehat{f}$ to f depends upon m but not upon p and $K(n)$, provided only that $K(n) \to \infty$ fast enough with the minimum rate depending on p; see Theorems 1 and 2 where $K(n)$ is of order n^γ and only a lower bound for γ is assumed, though the lower bound depends on p. The minimum number of knots grows more slowly with n as p increases. The asymptotic results presented here provide theoretical justification for the conventional wisdom that the number of knots is not important, provided only that the number is above some minimum depending upon the degree of the spline. Previously there was empirical support for this assertion (Ruppert, 2002) but no theoretical support. The bias of a penalized spline has two components, namely modelling bias due to approximating the regression function by a spline, and shrinkage bias due to estimation by penalized rather than ordinary least squares. In the theory presented here, $K(n)$ grows sufficiently rapidly with n that the modelling bias is asymptotically negligible compared to the shrinkage bias. This result agrees with finite-sample examples in Ruppert (2002), where the modelling bias is quite small compared to the shrinkage bias.

For simplicity, most of our results are for the case of equally spaced design points and knots, so that $x_1 = 1/n, x_2 = 2/n, \ldots, x_n = 1$ and $\kappa_0 = 0, \kappa_1 = 1/K(n), \kappa_2 = 2/K(n), \ldots, \kappa_{K(n)} = 1$. In § 4, these results are generalized to unequally spaced design points and knots. An interesting finding is that penalized splines are not design-adaptive as defined by Fan (1992) because their asymptotic bias depends on the design density and the bias converges to zero at a slower rate at the boundary than in the interior.

Penalized splines use fewer knots than smoothing splines, which use a knot at each data-point. Reducing the number of knots, which goes back at least to O'Sullivan (1986), makes computations easier. The methodology and applications of penalized splines are discussed extensively in Ruppert et al. (2003) and in papers by many authors, too many to review here. What, with a few exceptions, has been largely absent is an asymptotic theory that can be used to compare penalized splines with other nonparametric regression techniques. Exceptions are papers such as Yu & Ruppert (2002) and Wand (1999), where the number of knots is held fixed as the sample size increases. Another exception is the paper by Hall & Opsomer (2005). The results of Hall and Opsomer differ in several major ways from the results presented here. First, they use an approximation where knots are placed continuously; that is, there is a knot at every value of x in some interval. Moreover, their results are expressed as infinite series involving the eigenvalues of a certain operator; for example, see their expression (25) for the mean integrated squared error. In contrast our results are expressed in a form similar to results for kernel estimators, making comparisons with other nonparametric regression estimators easier. For example, we obtain explicit expressions for bias which, as just mentioned, show that penalized splines are not design-adaptive. Another advantage of our approach over that in Hall & Opsomer (2005) is that we can find the minimum rate at which the number of knots must converge to infinity in order for the modelling bias to be negligible.

2. Zero-degree splines

2·1. *Overview*

Zero-degree splines are piecewise-constant. The kth zero-degree B-spline is $B_k^{[0]}(x) = I\{\kappa_{k-1} < x \leqslant \kappa_k\}$, $1 \leqslant k \leqslant K(n)$. For simplicity, assume that $n/K(n)$ is an integer, which will be denoted by M. This assumption implies that every Mth x_t is a knot; that is $\kappa_j = x_{jM}$ for $j = 1, \ldots, K(n)$. If $n/K(n)$ is not an integer, we could define $M = \lfloor n/K(n) \rfloor$, the integer part of $n/K(n)$, and place a knot at every Mth x_t and at x_n. This would introduce an asymptotically negligible boundary effect in that the number of data points in the last 'bin' would be less than that in other bins. Here the kth 'bin' is $(\kappa_{k-1}, \kappa_k]$ and equals the support of $B_k^{[0]}$.

Recall that $X^{[0]}$ is the $n \times K(n)$ matrix with (t, j)th entry equal to $B_j^{[0]}(x_t)$. Then $(X^{[0]})' X^{[0]} = M I_{K(n)}$ where $I_{K(n)}$ is the $K(n) \times K(n)$ identity matrix. Therefore, by (2), the penalized least-squares estimator solves

$$\{I_{K(n)} + \lambda_n (D^m)' D^m\} \widehat{b} = \overline{y}, \tag{3}$$

where $\lambda_n = \lambda_n^*/M$ and $(X'Y/M) = \overline{y} = (\overline{y}_1, \ldots, \overline{y}_{K(n)})'$ where $\overline{y}_k$ is the average of all y_t such that $\kappa_{k-1} < x_t \leqslant \kappa_k$.

2·2. *Solving banded linear equations*

The matrix $I_{K(n)} + \lambda_n (D^m)' D^m$ in (3) and, more generally, $\Omega_n^{[p,m]}$ in (2) have a pattern that we will use to study the asymptotic behaviour, as $K(n) \to \infty$, of the solutions to equations such as (2). Define $q = \max(p-1, m)$. Except for the first q and last q columns, every column of $\Omega_n^{[p,m]}$ has the form

$$(0 \cdots 0 \quad \omega_q \cdots \omega_1 \quad \omega_0 \quad \omega_1 \cdots \omega_q \cdots 0)',$$

where ω_0 is the diagonal entry and $\omega_q \neq 0$ by definition of q. We will approximate the solution to (2) by finding a vector T_t that is orthogonal to all columns of $\Omega_n^{[p,m]}$ except the tth and the first and last q columns. Moreover, for estimation in the interior, that is, for $t/K(n) \to x \in (0, 1)$, T_t will also be asymptotically orthogonal to the first and last q columns.

We will say that $H_{x,n}(\cdot)$ is the 'equivalent kernel' for an estimator $\widehat{f}$ at $x \in [0, 1]$ if $\widehat{f}(x)$ has the same asymptotics as $\sum_{t=1}^n H_{x,n}(x_t) y_t$. In the common case where $H_{x,n}(\cdot) = H\{(\cdot - x)/b_n\}$ for some function H independent of x and n, we also call H the equivalent kernel and b_n is called the 'equivalent bandwidth'. Here T_t determines the form of the equivalent kernel at x_t for the penalized spline estimator.

Assume that there is root ρ_n, possibly complex, of modulus less than 1 of the polynomial

$$P(\rho) = \omega_q + \omega_{q-1}\rho + \cdots + \omega_0 {\rho_n}^q + \cdots + \omega_q {\rho_n}^{2q}.$$

Define $T_t(\rho_n) = (\rho_n^{t-1} \cdots \rho_n^2 \rho_n 1 \rho_n \rho_n^2 \cdots \rho_n^{K(n)+p-t})$. Then $T_t(\rho_n)$ is orthogonal to all columns of $\Omega_n^{[p,m]}$ except the first and last q columns and columns $t - q + 1, \ldots, t + q - 1$. If we can find q distinct roots of P, $\rho_{n1}, \ldots, \rho_{nq}$, say, all less than 1 in modulus, then we can find a linear combination, S_t say, of $T(\rho_{n1}), \ldots, T(\rho_{nq})$ that is orthogonal to columns $t - q + 1, \ldots, t - 1$ and columns $t + 1, \ldots, t + q - 1$. Moreover, since $|\rho_{nj}| < 1$ for $j = 1, \ldots, q$, S_t is asymptotically orthogonal to all columns except the tth, assuming that $t/K(n) \to x \in (0, 1)$. As we will see, the boundary case where $x \to 0$ or 1 at a suitable rate can be handled similarly.

We see that our technique requires P to have q distinct roots of modulus less than 1. Will this happen? Note that P has $2q$ roots, and all are nonzero since $\omega_q \neq 0$. By the symmetry of the coefficients of P, if ρ_n is a root, then $1/\rho_n$ is also a root. Thus, if the roots of p are distinct and none of them has modulus 1, then there will be exactly q roots less than 1 in modulus. Numerical experiments suggest that this is always the case with the matrix $\Omega_n^{[p,m]}$ in (2). In certain situations, we have a proof that no root has modulus 1; see Proposition 1.

In § 2·6, $m = 2$ and $q = 2$ and P has two complex roots of modulus less than 1. The complex roots cause the effective kernel to be an exponentially damped linear combination of $\cos(x)$ and $\sin(|x|)$.

We remark that $\Omega_n^{[p,m]}$ is a Toeplitz matrix with modified upper left and lower right corners. The inverses of such matrices have been much studied; see Dow (2003) for a review. We have tried, but without success, to find results in the literature that would give the asymptotic behaviour of solutions to (2) in a direct manner. Also, since ρ_n is a root of $P(\rho)$, we see that $G(n) = \rho_n^n$ is a solution to the homogeneous difference equation $\omega_q G(n) + \omega_{q-1} G(n-1) + \cdots + \omega_0 G(n-q) + \cdots + \omega_q G(n-2q) = 0$. We used this fact when constructing T_t. We had hoped to exploit the theory of difference equations—see, for example, Elaydi (2005)—in this research but we were unable to find an approach simpler than the one just described.

2·3. *First-order difference penalties: overview*

Now we specialize to the case where $m = 1$. To find the solution to (3) it is convenient to divide both sides of (3) by $(1 + 2\lambda_n)$ so that all diagonal elements, except the first and last, equal 1. This gives us

$$\Lambda \widehat{b} = z, \tag{4}$$

where $z = (z_1, \ldots, z_{K(n)})' = \overline{y}/(1 + 2\lambda_n)$ and Λ is the $K(n) \times K(n)$ matrix with $\Lambda_{tj} = 1$ if $1 < t = j < K(n)$, $\Lambda_{tj} = \eta_n = -\lambda_n/(1 + 2\lambda_n)$ if $|t - j| = 1$, $\Lambda_{11} = \Lambda_{K(n)K(n)} = \theta_n = (1 + \lambda_n)/(1 + 2\lambda_n)$, and $\Lambda_{tj} = 0$ if $|t - j| > 1$.

To solve (4) we apply the methodology described in § 2·2. In our case, $m = 1$ and $p = 0$, so $q = 1$. Let ρ_n be the solution between 0 and 1 of

$$P(\rho) = \eta_n + \rho + \eta_n \rho^2 = 0; \tag{5}$$

since $P(0) < 0$ and $P(1) > 0$ such a solution must exist. Then

$$\rho_n = \frac{1 - (1 - 4\eta_n^2)^{1/2}}{-2\eta_n} = \frac{1 + 2\lambda_n - (1 + 4\lambda_n)^{1/2}}{2\lambda_n}. \tag{6}$$

We now solve for $\widehat{b}_1$ and $\widehat{b}_{K(n)}$. Let $T_t = T_t(\rho_n) = (\rho_n^{t-1}, \rho_n^{t-2}, \ldots, \rho_n, 1, \rho_n, \rho_n^2, \ldots, \rho_n^{K(n)-t})'$. By (5), T_t is orthogonal to all columns of Λ except the first, last and tth. In particular, T_1 and $T_{K(n)}$ are orthogonal to all columns except the first and last, which makes it easy to solve for $\widehat{b}_1$ and $\widehat{b}_{K(n)}$. Multiplying both sides of (4) by T_1', we obtain $\{\theta_n + \eta_n\rho_n\}\widehat{b}_1 + \rho_n^{K(n)-2}(\eta_n + \theta_n\rho_n)\widehat{b}_{K(n)} = \sum_{k=1}^{K(n)} \rho_n^{k-1} z_k$, and then multiplying both sides of (4) by $T_{K(n)}'$ we obtain $\rho_n^{K(n)-2}(\eta_n + \theta_n\rho_n)\widehat{b}_1 + (\theta_n + \eta_n\rho_n)\widehat{b}_{K(n)} = \sum_{k=1}^{K(n)} \rho_n^{K(n)-k} z_k$. Therefore,

$$\widehat{b}_1 = \frac{(\theta_n + \eta_n\rho_n)\left(\sum_{k=1}^{K(n)} \rho_n^{k-1} z_k\right) - \rho_n^{K(n)-2}(\eta_n + \theta_n\rho_n)\left(\sum_{k=1}^{K(n)} \rho_n^{K(n)-k} z_k\right)}{(\theta_n + \eta_n\rho_n)^2 - \rho_n^{2(K(n)-2)}(\eta_n + \theta_n\rho_n)^2}. \tag{7}$$

We will choose λ_n so that ρ_n, which is a function of λ_n, satisfies $\rho_n^{K(n)} = \exp(-n^{1/5}h^{-1})$ for some positive constant h; equations (16) and (18) below show that λ_n can be chosen to achieve this result. Then $\widehat{b}_1 \sim (\sum_{k=1}^{K(n)} \rho_n^{k-1} z_k)/(\theta_n + \eta_n \rho_n) = (\sum_{k=1}^{K(n)} \rho_n^{k-1} \overline{y}_k)/\{(\theta_n + \eta_n \rho_n)(1 + 2\lambda_n)\}$, where $a_n \sim c_n$ means that $a_n/c_n \to 1$. Also,

$$\widehat{b}_{K(n)} = \frac{-\rho_n^{K(n)-2}(\eta_n + \theta_n \rho_n)\left(\sum_{k=1}^{K(n)} \rho_n^{k-1} z_k\right) + (\theta_n + \eta_n \rho_n)\left(\sum_{k=1}^{K(n)} \rho_n^{K(n)-k} z_k\right)}{(\theta_n + \eta_n \rho_n)^2 - \rho_n^{2(K(n)-2)}(\eta_n + \theta_n \rho_n)^2}, \tag{8}$$

so that $\widehat{b}_{K(n)} \sim (\sum_{k=1}^{K(n)} \rho_n^{K(n)-k} \overline{y}_k)/\{(\theta_n + \eta_n \rho_n)(1 + 2\lambda_n)\}$.

After some algebra, one can show that

$$(\theta_n + \eta_n \rho_n)(1 + 2\lambda_n) = 1 + \lambda_n - \rho_n \lambda_n = \{1 + (1 + 4\lambda_n)^{1/2}\}/2,$$

since $\rho_n \lambda_n = 1/2 + \lambda_n - (1 + 4\lambda_n)^{1/2}/2$. Also,

$$1/(1 - \rho_n) = (2\lambda_n)/\{(1 + 4\lambda_n)^{1/2} - 1\} = \{1 + (1 + 4\lambda_n)^{1/2}\}/2.$$

Thus,

$$(\theta_n + \eta_n \rho_n)(1 + 2\lambda_n) = 1/(1 - \rho_n) \sim \sum_{k=1}^{K(n)} \rho_n^{k-1},$$

so that $\widehat{b}_1 \overset{\frown}{=} (\sum_{k=1}^{K(n)} \rho_n^{k-1} \overline{y}_k)/(\sum_{k=1}^{K(n)} \rho_n^{k-1})$, with a similar result for $\widehat{b}_{K(n)}$.

Multiplying both sides of (4) by T_t, $1 < t < K(n)$, one obtains

$$(1 + 2\rho_n \eta_n)\widehat{b}_t = \sum_{j=1}^{K(n)} \rho_n^{|t-j|} z_j - \left\{\left(\rho_n^{t-1}\theta_n + \rho_n^{t-2}\eta_n\right)\widehat{b}_1 + \left(\rho_n^{K(n)-t-1}\eta_n + \rho_n^{K(n)-t}\theta_n\right)\widehat{b}_{K(n)}\right\}. \tag{9}$$

Substituting (7) and (8) into (9) gives an exact expression for $\widehat{b}_t$.

2·4. *First-order penalties: estimation at interior points*

Consider the non-boundary case where we fix $x \in (0, 1)$ and let $t = t_n(x)$ be such that

$$t/K(n) \to x. \tag{10}$$

Then

$$\widehat{b}_t \sim \sum_{j=1}^{K(n)} \rho_n^{|t-j|} \overline{y}_j / \{(1 + 2\rho_n \eta_n)(1 + 2\lambda_n)\}. \tag{11}$$

Also, by (10), $t \to \infty$ so that

$$\sum_{j=1}^{K(n)} \rho_n^{|t-j|} \sim \sum_{j=-\infty}^{\infty} \rho_n^{|t-j|} = \frac{1}{1 - \rho_n} + \frac{\rho_n}{1 - \rho_n} = \frac{1 + \rho_n}{1 - \rho_n} = \frac{(1 + 4\lambda_n) - (1 + 4\lambda_n)^{1/2}}{-1 + (1 + 4\lambda_n)^{1/2}}. \tag{12}$$

If we multiply the numerator and denominator on the right-hand side of (12) by $(1 + 4\lambda_n) + (1 + 4\lambda_n)^{1/2}$, this expression simplifies to $(1 + 4\lambda_n)^{1/2}$. Also, by (6), $1 + 2\rho_n \eta_n = (1 - 4\eta_n^2)^{1/2}$ and some algebra show that $1 - 4\eta_n^2 = 1 - 4\lambda_n^2/(1 + 2\lambda_n)^2 = (1 + 4\lambda_n)/(1 + 2\lambda_n)^2$, so that

$$(1 + 2\rho_n \eta_n)(1 + 2\lambda_n) = (1 + 4\lambda_n)^{1/2} \sim \sum_{j=1}^{K(n)} \rho_n^{|t-j|}. \tag{13}$$

Thus, in the nonboundary case, we have by (11) and (13), and since $\widehat{f}$ is piecewise constant, that, for any $x \in (\kappa_{t-a}, \kappa_t]$,

$$\widehat{f}(x) = b_t \sim \frac{\sum_{j=1}^{K(n)} \rho_n^{|t-j|} \overline{y}_j}{\sum_{j=1}^{K(n)} \rho_n^{|t-j|}}. \tag{14}$$

This result shows that, in the nonboundary case, the penalized spline with $p = 0$ and $m = 1$ is asymptotically equivalent to a binned Nadaraya–Watson kernel estimator. More precisely, we have the following result.

THEOREM 1. *Suppose there exists $\delta > 0$ such that $E(Y^{2+\delta}) < \infty$, that the regression function $f(x)$ has a continuous second derivative, that the conditional variance function $\sigma^2(x)$ is continuous, that*

$$K(n) = Cn^{\gamma} \text{ with } C > 0 \quad \text{and} \quad \gamma > 2/5, \tag{15}$$

and that λ_n is chosen so that

$$\rho_n = \exp\left\{-(Ch)^{-1} n^{1/5-\gamma}\right\} = \exp\left\{-n^{1/5} h^{-1} K(n)^{-1}\right\}. \tag{16}$$

Let $\widehat{f}_n(x)$ be the first-order penalized estimator using zero-degree splines, that is $m = 1$ and $p = 0$, with equally spaced knots. Then, for any $x \in (0, 1)$, we have that

$$n^{2/5}\{\hat{f}_n(x) - f(x)\} \rightarrow N\{\mathcal{B}(x), \mathcal{V}(x)\},$$

in distribution as $n \rightarrow \infty$, where $\mathcal{B}(x) = h^2 f^{(2)}(x)$ and $\mathcal{V}(x) = 4^{-1}h^{-1}\sigma^2(x)$. The equivalent kernel is the double-exponential or Laplace kernel

$$H(x) = \tfrac{1}{2}\exp(-|x|), \tag{17}$$

and the equivalent bandwidth is $h_n = hn^{-1/5}$.

Proofs of all theorems in this paper can be found in the technical appendix.

Note that (16) will hold for some choice of λ_n such that

$$\lambda_n \sim C^2 h^2 n^{2\gamma-2/5} \sim \left\{K(n) h n^{-1/5}\right\}^2, \tag{18}$$

where $h > 0$ is a constant. To show (18), combine equations (6) and (16) to obtain $\{1 + 2\lambda_n - (1 + 4\lambda_n)^{1/2}\}/(2\lambda_n) = \rho_n = \exp\{-C^{-1}h^{-1}n^{1/5-\gamma}\}$. By (18), $\lambda_n \rightarrow \infty$ so we have

$$\begin{aligned} -\log(\rho_n) &= \log[2\lambda_n/\{1 + 2\lambda_n - (1 + 4\lambda_n)^{1/2}\}] = \log(1 + \lambda_n^{-1/2}) + o(\lambda_n^{-1}) \\ &= \lambda_n^{-1/2} + o(\lambda_n^{-1/2}) \sim C^{-1}h^{-1}n^{1/5-\gamma}. \end{aligned}$$

Hence λ_n should be chosen as $\lambda_n \sim C^2h^2n^{2\gamma-2/5} \sim \{K(n)hn^{-1/5}\}^2$, by (15).

Example 1. This example studies how quickly the finite-sample kernel converges to (17). Figure 1 shows the finite-sample kernels and the double-exponential kernel for all four values, 40, 80, 160 and 320, of n. Also, $K(n)$ and λ_n are functions of n suggested by the asymptotics; see the caption of Fig. 1. The double-exponential kernel for fitting at the jth bin has value $\rho_n^{|t-j|}/(\sum_\ell \rho_n^{|\ell-j|})$ at the tth bin with ρ_n a root of (5). The kernels are for estimation at the centre of the design. We see good agreement between the finite-sample and asymptotic kernels for $n = 40$ and excellent agreement for $n = 320$.

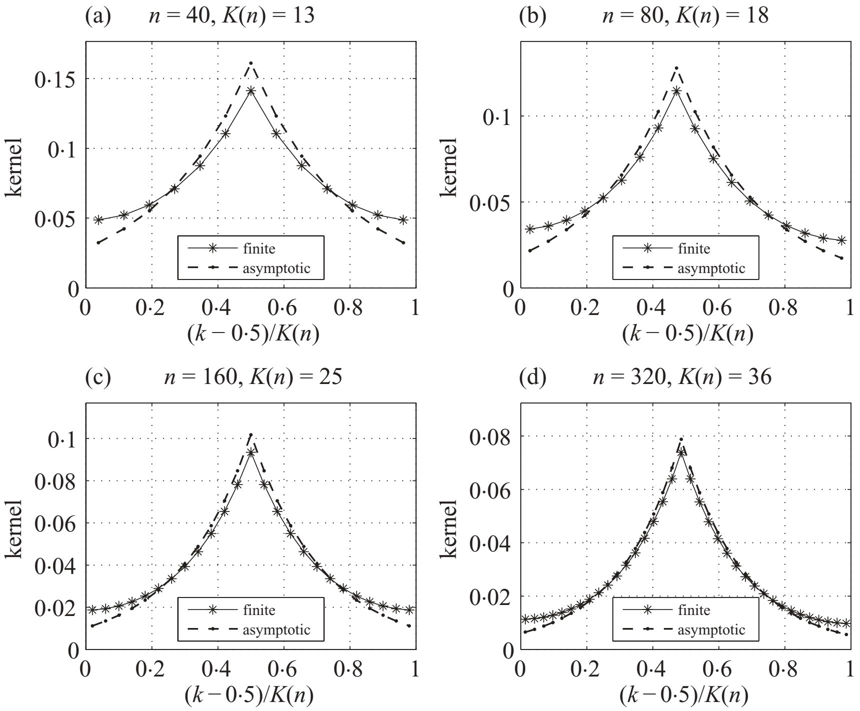

Fig. 1. The finite-sample kernel (solid) and asymptotic double-exponential kernel (dashed) for $m = 1$ and $p = 0$ and for (a) $n = 40$, (b) $n = 80$, (c) $n = 160$, (d) $n = 320$. In each case $K(n)$ and λ_n are functions of n suggested by the asymptotics: $K(n) = 2n^{1/2}$, rounded to the nearest integer, and $\lambda_n = \{K(n)hn^{-1/5}\}^2$ with $h = 0{\cdot}6$. Here k is the bin number, $(k - 0{\cdot}5)/K(n)$ is the midpoint of the kth bin, and the kernels are for estimation at midpoint of the bin containing 0·5.

2·5. *First-order penalties: estimation at the boundary*

The boundary case is slightly more complex than the nonboundary case. The bias is of order $n^{-1/5}$ but it is not the same as the bias of the Nadaraya–Watson estimator, though the Nadaraya–Watson bias is also of order $n^{-1/5}$.

To find the equivalent kernel at the left-hand boundary, we suppose that $t/K(n) \to 0$ as $n \to \infty$ at look at the tth bin. Then, from (9), we have

$$\begin{aligned}(1 + 2\rho_n\eta_n)\widehat{b}_t &\sim \sum_{j=1}^{K(n)} \rho_n^{|t-j|} z_j - \rho_n^t(\rho_n^{-1}\theta_n + \rho_n^{-2}\eta_n)\widehat{b}_1 \\ &\sim \sum_{j=1}^{K(n)} \rho_n^{|t-j|} z_j - \rho_n^t(\rho_n^{-1}\theta_n + \rho_n^{-2}\eta_n)\sum_{j=1}^{K(n)} \rho_n^{j-1} z_j/(\theta_n + \eta_n\rho_n)\end{aligned}$$

so that

$$\widehat{b}_t \sim \sum_{j=1}^{K(n)} (a_1\rho_n^{|t-j|} + a_1a_2\rho_n^{t+j})z_j, \tag{19}$$

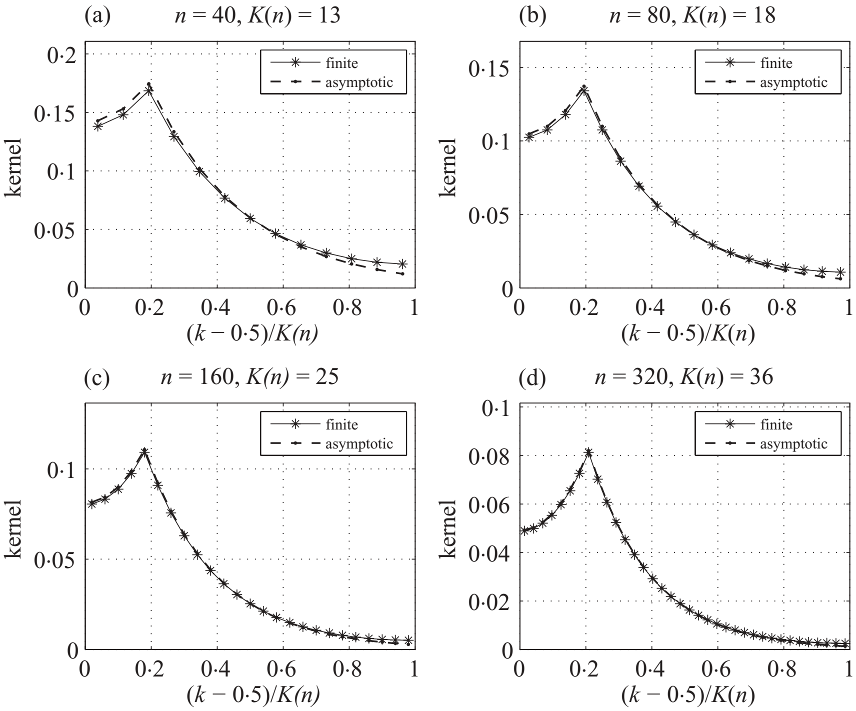

Fig. 2. The finite-sample kernel (solid) and asymptotic boundary kernel (dashed) given by (19) for $m = 1$ and $p = 0$ and for (a) $n = 40$, (b) $n = 80$, (c) $n = 160$, (d) $n = 320$. In each case $K(n)$ and λ_n are functions of n suggested by the asymptotics: $K(n) = 2n^{1/2}$, rounded to the nearest integer, and $\lambda_n = \{K(n)hn^{-1/5}\}^2$ with $h = 0{\cdot}6$. Here k is the bin number, $(k - 0{\cdot}5)/K(n)$ is the midpoint of the kth bin, and the kernels are for estimation at the midpoint of the bin containing 0·2. One can see that, in addition to the truncation at the boundary, the kernels are asymmetric, in agreement with (19).

where $a_1 = (1 + 2\rho_n\eta_n)^{-1}$ and $a_2 = (\theta_n + \eta_n\rho_n^{-1})/\{\rho_n^2(\theta_n + \eta_n\rho_n)\}$. By (A1) and (19),

$$\widehat{b}_t \sim \sum_{j=1}^{K(n)} H(\overline{x}_t, \overline{x}_j; hn^{-1/5})z_j$$

where H is the equivalent boundary kernel such that

$$H(\overline{x}_t, \overline{x}_j; h) \propto [a_1 \exp(-|\overline{x}_t - \overline{x}_j|/h) + a_1a_2 \exp\{-(\overline{x}_t + \overline{x}_j)/h\}]. \tag{20}$$

The constant of proportionality is determined by $\sum_{j=1}^{K(n)} H(\overline{x}_t, \overline{x}_j; hn^{-1/5}) = 1$. Since the equivalent bandwidth is of order $n^{-1/5}$ by (A2), the second term in (20) is asymptotically negligible in the nonboundary case where, for some $x \in (0, 1)$, $\overline{x}_t$ is chosen so that $\overline{x}_t \to x$. However, in the left-hand boundary case under consideration where $\overline{x}_t = cn^{-1/5}$ for some $c \geqslant 0$, the contribution of the second term persists as $n \to \infty$.

Example 2. Figure 2 compares the finite-sample kernel with the asymptotic boundary kernel given by (19), both for estimation at $x = 0{\cdot}2$. The sample sizes are 40, 80, 160 and 320, and $K(n)$ and λ_n are functions of n suggested by the asymptotics; see the caption of Fig. 2. We see that the agreement between the finite-sample and asymptotic kernels is very good for $n = 320$. For smaller sample sizes, the finite-sample kernel is well above zero at both the left-hand and right-hand boundaries, so there are effects from both boundaries; in this situation, the asymptotic boundary kernel should not be expected to approximate the finite-sample kernel (20) extremely well.

2·6. *Second-order penalties: overview*

Now suppose that $m = 2$. Then $D^m = D^2$ is $\{K(n) - 2\} \times K(n)$ and the tth row of D^2 has 1 in coordinates t and $t + 2$, -2 in coordinate $t + 1$, and 0 elsewhere. Except for $t = 1, 2, K(n) - 1$ and $K(n)$, the tth column of $(D^2)'D^2$ has entries, 1, -4, 6, -4 and 1 in rows $t - 2$ to $t + 2$ and 0 elsewhere. Now $\widehat{b}$ solves (3) with $m = 2$.

2·7. *Second-order penalties: estimation at interior points*

The next theorem treats the interior case where $x \in (0, 1)$. Theorem 3 below covers the boundary case.

THEOREM 2. *Suppose that there exists $\delta > 0$ such that $E(Y^{2+\delta}) < \infty$, that $f(x)$ has a continuous fourth derivative, that $\sigma^2(x)$ is continuous, that $K(n) \sim Cn^\gamma$ with $C > 0$ and $\gamma > 4/9$, and that there exists a constant $h > 0$ such that*

$$\lambda_n \sim 4^{-1}C^4h^4n^{4\gamma-4/9}. \tag{21}$$

Let $\widehat{f}_n(x)$ be the penalized estimator with $p = 0$, $m = 2$, and equally spaced knots. Then, for any $x \in (0, 1)$, when $n \to \infty$, we have

$$n^{4/9}\{\widehat{f}_n(x) - f(x)\} \to N\{\mathcal{B}_1(x), \mathcal{V}_1(x)\},$$

in distribution, where $\mathcal{B}_1(x) = (1/24)f^{(4)}(x)h^4 \int x^4T(x)dx$, $\mathcal{V}_1(x) = h^{-1}\{\int T^2(x)dx\}\sigma^2(x)$, and $T(x)$ is a fourth-order kernel given by

$$L^{-1}\{\exp(-|x|)\cos(x) + \exp(-|x|)\sin(|x|)\}, \tag{22}$$

where L is a normalizing constant. The equivalent bandwidth is

$$h_n = hn^{-1/9}, \tag{23}$$

where h is given by (21), so that $\lambda_n \sim 4^{-1}K(n)^4h_n^4$.

Example 3. We plotted finite-sample and asymptotic kernels for estimation at $x = 0{\cdot}5$ using four sample sizes: 40, 80, 160 and 320. For each sample size, $K(n) = 2n^{1/2}$, rounded to the nearest integer, and $\lambda_n = 4^{-1}\{K(n)hn^{-1/9}\}^4$. To save space, we only describe the plots. There is reasonably good agreement between the finite-sample and asymptotic kernels, especially for $n = 160$ and larger where the two kernels were difficult to distinguish visually.

2·8. *Second-order penalties: estimation at the boundary*

We now consider the boundary case where $x \to 0$ or 1 at the same rate at which the equivalent bandwidth converges to 0.

THEOREM 3. *Suppose that there exists $\delta > 0$ such that $E(Y^{2+\delta}) < \infty$, that $f(x)$ has a continuous second derivative, that $\sigma^2(x)$ is continuous, that $K(n) \sim Cn^\gamma$ with $\gamma > 2/5$, and that there exists a constant $h > 0$ such that $\lambda_n \sim 4^{-1}C^4h^4n^{4\gamma-4/5}$. Let $\widehat{f}_n$ be the penalized estimator using zero-degree splines with a second-order penalty and equally spaced knots. Assume that we are in the boundary case so that either $x = cn^{-1/5}$ or $x = 1 - cn^{-1/5}$ for some $c \geqslant 0$. Then, when $n \to \infty$, we have*

$$n^{2/5}\{\widehat{f}_n(x) - f(x)\} \to N\{\mathcal{B}(x), \mathcal{V}(x)\},$$

in distribution, where $\mathcal{V}(x) = h^{-1}\{\int T'^2(x', x)dx'\}\sigma^2(x)$*,* $T'(\cdot, x)$ *is a second-order boundary kernel described in* (A10)*, and* $\mathcal{B}(x) = 2^{-1} f^{(2)}(x)h^2 \int x^2 T'(x', x)dx'$*.*

Example 4. We plotted finite-sample and asymptotic boundary kernels for estimation at $x = 0{\cdot}2$ using four sample sizes: 40, 80, 160 and 320. For each sample size, $K(n)$ is a function of n specified in Example 3. There was extremely good agreement between the finite-sample and asymptotic kernels, even for the smaller sample sizes.

3. LINEAR SPLINES

3·1. *Overview*

The linear B-spline basis with knots $\kappa_{-1} < 0 = \kappa_0, \ldots, \kappa_{K(n)} = 1 < \kappa_{K(n)+1}$ is $\{B_0^{[1]}, \ldots, B_{K(n)}^{[1]}\}$, where

$$\begin{aligned} B_k^{[1]}(x) &= 0, && x < \kappa_{k-1} \\ &= K(n)(x - \kappa_{k-1}), && \kappa_{k-1} \leqslant x \leqslant \kappa_k \\ &= K(n)(\kappa_{k+1} - x), && \kappa_k \leqslant x \leqslant \kappa_{k+1} \\ &= 0, && x > \kappa_{k+1}. \end{aligned} \tag{24}$$

Thus, $B_k^{[1]}(x)$ increases linearly from 0 to 1 as x increases from κ_{k-1} to κ_k and then decreases linearly to 0 as x increases from κ_k to κ_{k+1}; the actual values of the knots κ_{-1} and $\kappa_{K(n)+1}$ are immaterial, since the B-splines will be evaluated only on [0, 1].

Note that $\int_0^1 \{B_k^{[1]}(x)\}^2 dx$ equals $2/3\, K(n)^{-1}$ for $k = 1, \ldots, K(n) - 1$ and equals $1/3\, K(n)^{-1}$ for $k = 0$ or $K(n)$. Also, $\int_0^1 B_k^{[1]}(x) B_{k+1}^{[1]}(x)dx = 1/6\, K(n)^{-1}$ for $k = 0, \ldots, K(n) - 1$. Therefore, $X'X \doteq M\,\Sigma$, where

$$\Sigma = \begin{pmatrix} 1/3 & 1/6 & 0 & 0 & \cdots & 0 & 0 \\ 1/6 & 2/3 & 1/6 & 0 & \cdots & 0 & 0 \\ 0 & 1/6 & 2/3 & 1/6 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 2/3 & 1/6 \\ 0 & 0 & 0 & 0 & \cdots & 1/6 & 1/3 \end{pmatrix}. \tag{25}$$

Equation (2) is solved by

$$\widehat{b} = \{M\Sigma + \lambda_n^*(D^m)'D^m\}^{-1}X'Y = \{\Sigma + \lambda_n(D^m)'D^m\}^{-1}(X'Y/M). \tag{26}$$

Equation (26) has a banded matrix, $\{\Sigma + \lambda_n(D^m)'D^m\}$, with the same number of nonzero diagonals as the matrix $\{I_k + \lambda_n(D^m)'D^m\}$ in equation (3).

Also, in (26), $(X'Y/M)$ can again be regarded as a vector of bin averages of the Y_t using linear binning (Hall & Wand, 1996). To appreciate this, let the kth bin be $[\kappa_{k-1}, \kappa_{k+1}]$, $k = 0, \ldots, K(n)$. Thus, if $\kappa_{k-1} \leqslant x_t \leqslant \kappa_k$, then x_t is in bins $k - 1$ and k. A fraction $K(n)(x - \kappa_{k-1})$ of Y_t is placed in the kth bin and the remaining fraction $K(n)(\kappa_k - x)$ goes into bin $k - 1$. It follows that the analysis for linear splines can be done in the same way as for piecewise-constant splines.

3·2. *First-order difference penalty*

If a first-order difference penalty is used, then, in the nonboundary region, the penalized spline behaves asymptotically as an exponential kernel-weighted average of the bin averages, just as with zero-degree splines. The only differences are that the bin counts are from linear binning and the bandwidth is different because of the nonzero off-diagonal terms in Σ.

Theorem 4. *Suppose there exists $\delta > 0$ such that $E(Y^{2+\delta}) < \infty$, that $f(x)$ has a continuous second derivative, that $\sigma^2(x)$ is continuous, that $K(n) \sim Cn^{\gamma}$ with $C > 0$ and $\gamma > 1/5$, and that there exists a constant $h > 0$ such that the penalty $\lambda_n \sim C^2h^2n^{2\gamma-2/5} = \{K(n)hn^{-1/5}\}^2$. Let $\widehat{f}_n$ be the first-order penalized estimator using linear splines with first-order penalty and equally spaced knots. Then, for any $x \in (0, 1)$, when $n \to \infty$, we have that*

$$n^{2/5}\{\widehat{f}_n(x) - f(x)\} \to N\{\mathcal{B}(x), \mathcal{V}(x)\},$$

in distribution, where $\mathcal{B}(x) = h^2 f^{(2)}(x)$ and $\mathcal{V}(x) = 4^{-1}h^{-1}\sigma^2(x)$.

The equivalent kernel is double-exponential and the equivalent bandwidth is $hn^{-1/5}$.

Although their asymptotic behaviour is similar to that of zero-order splines, linear splines are different in two ways. There is no significant difference in asymptotic behaviour between zero-degree and linear splines when $\gamma > 2/5$. However, if $K(n) \sim Cn^{\gamma}$ with $1/5 < \gamma \leqslant 2/5$, then only linear splines obtain a $O(n^{-2/5})$ rate of convergence since zero-order splines have an infinite bias at this rate. Also, for linear splines we require that $\rho_n > 0$. This always holds for zero-degree splines according to equation (6). However, with linear splines, $\rho_n = \{6\lambda_n + 2 - (3 + 36\lambda_n)^{1/2}\}(6\lambda_n - 1)^{-1}$. We can rewrite this as $\rho_n = 1 + \{3 - (3 + 36\lambda_n)^{1/2}\}(6\lambda_n - 1)^{-1}$. Thus, $\rho_n > 0$ implies that $\lambda_n > 1/6$. The assumptions of Theorem 4 imply that $\gamma \to \infty$, so $\lambda_n > 1/6$ will hold eventually.

4. Unequally spaced data and knots

So far, equally spaced x_t and knots have been assumed. This assumption can be relaxed using an idea of Stute (1984). Assume that the x_t are in some finite interval (a, b) and that, for all t and n, $G(x_t) = u_t = t/n$ for some smooth function G from (a, b) to $(0, 1)$. If we fit a penalized spline to (Y_t, u_t), then the regression function is $f \circ G^{-1}$. Equally spaced knots for the (Y_t, u_t) data translate for the (Y_t, x_t) data into placing knots at sample quantiles so there are equal numbers of data points between pairs of consecutive knots. Therefore, our theory does not cover the situation where the x_t are unequally spaced but the knots are equally spaced. The asymptotics for this case would be interesting but require a different approach.

The following theorem follows from the application of Theorem 1 to (Y_t, u_t) and translation of the results back to (Y_t, x_t). Similar results can be obtained corresponding to Theorems 2, 3 and 4.

Theorem 5. *Assume that there is a twice-differentiable strictly increasing function G such that $G(x_t) = t/n$ for all t and n and that $f \circ G^{-1}$ is twice continuously differentiable on $(0, 1)$. Assume also that $\sigma^2(x)$ is continuous, that there exists $\delta > 0$ such that $E(Y^{2+\delta}) < \infty$, that $K(n) \sim Cn^{\gamma}$ with $C > 0$ and $\gamma > 2/5$, and that there exists a constant $h > 0$ such that $\lambda_n \sim C^2h^2n^{2\gamma-2/5}$. Let $\widehat{f}_n$ be the penalized spline estimator with $p = 0$ and $m = 1$ and with knots at equally spaced sample quantiles. Then, for any $x \in (0, 1)$, when $n \to \infty$, we have $n^{2/5}\{\hat{f}_n(x) - f(x)\} \to N\{\mathcal{B}(x), \mathcal{V}(x)\}$, in distribution, where, with $g = G'$,*

$$\mathcal{B}(x) = h^2(f \circ G^{-1})^{(2)}\{G(x)\} = \frac{h^2}{g^2(x)}\left\{f^{(2)}(x) - \frac{f'(x)g'(x)}{g(x)}\right\} \tag{27}$$

and $\mathcal{V}(x) = 4^{-1}h^{-1}\sigma^2(x)$.

Thus, the bias of the penalized spline differs in several ways from that of the Nadaraya–Watson estimator, which is

$$\mathcal{B}(x) = h^2 \left\{ f^{(2)}(x) + \frac{2f'(x)g'(x)}{g(x)} \right\}.$$

Interestingly, the second term inside the curly brackets in (27) appears in the bias of the Nadaraya–Watson estimator, though with a plus sign. The term $g^2(x)$ in the denominator of (27) is a spatially varying local bandwidth induced by the transformation of the x_t to the u_t.

Nonparametric regression estimators whose bias does not involve the design density g are called 'design-adaptive' by Fan (1992). Theorem 5 shows that penalized splines with $p = 0$, $m = 1$ and knots at sample quantiles are not design-adaptive. An open question is the behaviour of penalized splines when the knots are equally spaced or higher-order B-splines or penalties are used. This will be investigated in another paper.

5. Concluding Remarks

5·1. *Higher-order difference penalties*

We intend to study higher-order penalties, where $m > 2$, in the future. Here we merely make a few remarks about the case $p = 0$, i.e., piecewise-constant splines. The effective kernel will depend on the roots of modulus less than 1 of the polynomial

$$P(\rho_n) = (1 - \rho_n)^{2m}(-1)^m + C_n \rho_n^m,$$

where $C_n > 0$ and $C_n \to 0$ as $K(n) \to \infty$. We have seen that, for $m = 1$, P has one real root with modulus less than 1, and, for $m = 2$, there is a conjugate pair of roots with modulus less than 1. Since $C_n \to 0$ as $n \to \infty$ and $K(n) \to \infty$, all roots of P converge to 1. This ensures that, at each x, the effective bandwidth is of the optimal order and $\widehat{f}(x)$ is an average over an increasing number of bins.

In the case $m = 3$, our numerical experimentation has always found that there is one real root and one conjugate pair of roots with modulus less than 1. Therefore, the effective kernel is a linear combination of a double-exponential kernel, $\cos(ax)$ for some $a > 0$, and $\sin(b|x|)$ for some $b > 0$. The effective kernel for smoothing splines with a penalty on the third derivative is of this form; see equation (4·20) of Silverman (1984).

For $m = 4$, we have found that there are two conjugate pairs of roots with modulus less than 1. Therefore, the effective kernel will be a linear combination of the effective kernel for $m = 2$ with one bandwidth and the same kernel with a second bandwidth.

Typically, the bias of a smoother has an expansion

$$E\{\widehat{f}(x)\} - f(x) = \sum_{\ell=1}^{L} c_\ell h^\ell f^{(\ell)}(x) + o(h^\ell), \tag{28}$$

where h is the 'effective bandwidth'. If, in (28), $c_\ell = 0$ for $\ell < L$ and $c_\ell \neq 0$, then the smoother is of order L at x.

For $m = 1$ and 2, we have found that the effective kernel is of the order $2m$ in the interior and order m at the boundary. Some numerical experiments suggest that this pattern continues for larger values of m. In fact, we have a heuristic justification for believing that the pattern continues for all m. The heuristic that works for $p = 0$, we believe, can be extended to all p, with some care. Let Z be a $K(n)$-dimensional vector such that $Z_t = Q(t)$, $t = 1, \ldots, K(n)$, for some polynomial Q of degree d_Q. Then $D^m Z = 0$ if $m > d_Q$. Therefore, if we modify the data by subtracting

$\sum_{\ell=1}^{m-1} f^{(\ell)}(x)(x_t - x)^{\ell}/(\ell!)$ from y_t for all t, then the value of $\widehat{f}(x)$ is unchanged because $p = 0$; if $p > 0$ the estimator will change because $(X^{[p]})' X^{[p]}$ in $\Omega^{[p,m]}_{K(n),\lambda^*_n}$ is not a scalar multiple of the identity matrix, but the change should be asymptotically negligible. With this modification, the bias at x is of order m. Thus, since $\widehat{f}(x)$ is unchanged by the modification, the bias must have been of order m even without this modification. The penalized spline behaves as if an oracle told us the value of $f^{\ell}(x)$ for $\ell = 1, \ldots, m-1$. Moreover, except for the first and last m columns, all columns of D^{2m} are orthogonal to a polynomial of degree less than $2m - 1$. This suggests that, in the interior, penalized splines are of order $2m$ rather than m.

5·2. *Comparisons with other spline smoothers*

Silverman (1984) found equivalent kernels for smoothing splines using Laplace transform techniques. For a cubic smoothing spline with an integral penalty of the squared second derivative, the equivalent kernel given by his equation (1·3) is $1/2 \exp(-2^{-1/2}|u|) \sin(2^{-1/2}|u| + \pi/4)$, which can be rewritten as an equally weighted linear combination of $\exp(-2^{-1/2}|u|) \sin(2^{-1/2}|u|)$ and $\exp(-2^{-1/2}|u|) \cos(2^{-1/2}u)$. This is a rescaled version of the equivalent kernel for second-order difference penalties given by (22), which we have found for piecewise-constant penalized splines.

This result is not too surprising, since the penalty in (1) is a rescaled discrete approximation to a smoothing spline penalty. More precisely,

$$\sum_{k=m+1}^{K(n)+p} \{\Delta^m(b_k)\}^2 \simeq K(n)^{-2m} \int_0^1 \{f^{(m)}(x)\}^2 dx.$$

Moreover, we have found that the behaviour of a spline estimator depends on the penalty, not the degree of the spline. Silverman also found that the Laplace density is the equivalent kernel when the penalty is on the first derivative, a result in agreement with (17).

Agarwal & Studden (1980) discuss ordinary least-squares estimation of spline models. Since they do not use a penalty, overfitting is controlled by knot selection. In this context, there is no shrinkage bias and only model bias, a situation opposite to ours. Thus, it is not surprising that the results they obtain differ substantially from ours. In particular, Agarwal and Studden's optimal estimator uses fewer knots than ours. From their equation (3·12), their optimal rate for $K(n)$ is $K(n) \sim n^{-1/(2p+3)}$; note that their d is our $p + 1$. Thus, for piecewise-constant splines, their optimal rate is $K(n) \sim n^{-1/3}$ while ours is $K(n) \sim n^{-\gamma}$ for any $\gamma > 2/5$. For linear splines, their optimal rate is $K(n) \sim n^{-1/5}$ while ours is $K(n) \sim n^{-\gamma}$ for any $\gamma > 1/5$.

An asymptotic theory intermediate between ours and that in Agarwal & Studden (1980) would select $K(n)$ so that modelling and shrinkage biases are of the same order. For the case $p = 0$ and $d = 1$, this would require $K(n) \sim Cn^{\gamma}$ with $\gamma = 2/5$ rather than $\gamma > 2/5$ as assumed in Theorem 1. Asymptotics of this type would require new research and will not be pursued here. It is not clear to us how valuable they would be from a practical standpoint.

It is interesting to compare penalized splines with local polynomial estimators. Local zero-degree polynomials are Nadaraya–Watson estimators. Therefore, penalized splines with a penalty of order 1 coincide with local polynomials with degree 0 and double-exponential kernels.

Penalized splines with $m > 1$ have different bias-order properties from those of local polynomial estimators. As shown in Ruppert & Wand (1994), local polynomial smoothers of degree p behave differently for p odd compared to p even. For p odd, they are of order $p + 1$ for all x. If p is even, then the order is again $p + 1$ at the boundary but is of order $p + 2$ in the interior. Thus, their bias orders at the interior and boundary are either identical or differ by 1. In contrast, the bias-orders at the interior and boundary of a penalized spline differ by m, at least if the heuristics in § 5·1 are correct.

5·3. *Choice of basis*

We have worked with the B-spline bases advocated by Eilers & Marx (1996). However, other bases are often used for penalized splines; for example, the truncated polynomials are used extensively in Ruppert et al. (2003). Our results apply, of course, to an estimator defined with other bases provided that this estimator is identical to one of the P-splines, penalized B-spline, estimators studied here. This is often the case. As discussed in § 3·7·1 of Ruppert et al. (2003), a penalized spline in one basis will be algebraically identical to a penalized spline in a second basis, if the two bases span the same vector space of functions and if they use identical penalties. For example, suppose we use a basis consisting of a constant function and the functions $I(x > \kappa_k)$, that is, step functions that jump from 0 to 1 at the knots. Then the spline model is $\beta_0 + \sum_{k=1}^{K(n)} a_k I(x > \kappa_k)$. Suppose as well that we use the penalty

$$\lambda_n \sum_{k=1}^{K(n)} a_k^2, \tag{29}$$

that is, the sum of squared jumps of the spline at the knots is penalized. Then this estimator is the same as the P-spline with $p = 0$ and $m = 1$. Similarly, the truncated line model $\beta_0 + \beta_1 x + \sum_{k=1}^{K(n)} a_k(x - \kappa_k)_+$ with penalty (29) is identical to the P-spline model with $p = 1$ and $m = 2$. In both cases, the model is piecewise linear and the penalty is on the sum of squared jumps in the first derivative.

5·4. *Penalizing derivatives*

Smoothing splines put a penalty on the integral of the squared mth derivative of the regression function, with $m = 2$ being the most common choice. Such penalties can be used on a penalized spline, if $p \geqslant m$, by replacing the penalty in (1) by $\lambda_n^* \int_0^1 \{\sum_{j=1}^{K(n)+p} b_j (B_j^{[p]})^{(m)}(x)\}^2 dx$, where $(B_j^{[p]})^{(m)}(x)$ is the mth derivative of $B_j^{[p]}(x)$. If one changes to the derivative penalty, then the only change in $\widehat{b}$ is that the matrix $(D^m)' D^m$ in $\Omega_n^{[p,m]}$ is replaced by M where $M_{ij} = \int_0^1 (B_i^{[p]})^{(m)}(x)(B_j^{[p]})^{(m)}(x)dx$. Since M is a banded matrix with modified corners having the same structure as $(D^m)' D^m$, a penalized spline with penalty on the mth derivative has the same asymptotic behaviour as penalized splines with an mth-order difference penalty. In fact, for some choices of p and m, such as $m = p = 1$, M is proportional to $(D^m)' D^m$, so, if the constant of proportionality is absorbed into the penalty parameter, then the spline with the derivative penalty is identical to the spline with the difference penalty.

Acknowledgement

The research of D. Ruppert was supported by grants from the U.S. National Science Foundation and National Institutes of Health.

Appendix

Technical details

Proof of Theorem 1. Let $\bar{x}_t = (2t - 1)/2K(n)$ be the midpoint of the tth bin, i.e., of $[(t-1)K(n)^{-1}, tK(n)^{-1}]$. Since $\gamma > 2/5$, the effect of binning is asymptotically negligible when the bandwidth is of the optimal order, $n^{-1/5}$. To be specific, we have $\bar{y}_t = f(\bar{x}_t) + \epsilon' + o(n^{-2/5})$ with ϵ' distributed as $N[0, \{K(n)/n\}\sigma^2(\bar{x}_t)]$. Since $\bar{x}_t - \bar{x}_j = (t-j)/K(n)$,

$$\rho_n^{|t-j|} = \exp\left\{-h^{-1}n^{1/5}\left(C^{-1}n^{-\gamma}|t-j|\right)\right\} = \exp\left\{-|\bar{x}_t - \bar{x}_j|\left(hn^{-1/5}\right)^{-1}\right\}, \tag{A1}$$

by (15) and (16). Thus, by (14), $\widehat{f}_n$ is asymptotically equivalent to the Nadaraya–Watson estimator with kernel (17) and bandwidth

$$h_n = hn^{-1/5}. \tag{A2}$$

Therefore, one can derive the asymptotic distribution of $\widehat{f}_n(x)$ using well-known techniques, for example as in Wand & Jones (1995). □

Before proving Theorem 2, we need some preliminary results. Define

$$w(\xi) = \lambda_n(1 - 4\xi + 6\xi^2 - 4\xi^3 + \xi^4) + \xi^2 = \lambda_n(1-\xi)^4 + \xi^2, \quad \lambda_n > 0. \tag{A3}$$

As discussed in § 2·2, the roots of w will be used to find a vector orthogonal to all columns of Λ except the first and last two and the tth. Clearly, w has no real root. Also, if r is a root of w, so is r^{-1}. Thus, for some complex r, the four roots of w are r, conj(r), r^{-1} and conj$(r)^{-1}$, where conj(r) is the complex conjugate of r. By the following proposition, one of the roots r and r^{-1} is less than 1 in magnitude and we will denote it by r_n.

PROPOSITION A1. *No root of w has modulus equal to* 1.

Proof. Suppose there is a ξ such that $w(\xi) = 0$ and $\xi = \exp(i\theta_n)$. Here $i = (-1)^{1/2}$. Note that $\xi - 1 = 2\sin(\theta_n/2)\exp\{(\pi/2 + \theta_n/2)i\}$, so $-\lambda_n(\xi-1)^4 = \xi^2$ implies that

$$16\lambda_n \sin^4(\theta_n/2)\exp\{(3\pi + 2\theta_n)i\} = \exp(2i\theta_n).$$

Comparing the real and imaginary parts on both sides, we have that $16\lambda_n\sin^4(\theta_n/2) = -1$. For any positive λ_n, this is impossible, so that there will be no root with norm 1. □

Since $|\rho_n| < 1$, $r_n = \rho_n\exp(i\alpha_n)$ for some ρ_n in $(0, 1)$. Therefore, we have the following proposition.

PROPOSITION A2. *Let c_n and d_n be the real and imaginary parts of $r_n - 4 + (6 + \lambda_n^{-1})r_n - 4r_n^2 + r_n^3$, where r is defined as above. Let $T_t = (T_{t,1}, \ldots, T_{t,K(n)})$ be defined by*

$$T_t = d_n\,\Re\left(r_n^{t-1}, r_n^{t-2}, \ldots, r_n, 1, r_n, \ldots, r_n^{K(n)-t}\right) - c_n\,\Im\left(r_n^{t-1}, r_n^{t-2}, \ldots, r_n, 1, r_n, \ldots, r_n^{K(n)-t}\right), \tag{A4}$$

where $\Re$ and $\Im$ are the real and imaginary parts, respectively, and each '1' is in the tth position. Then

(i) *T_t is orthogonal to all columns of $\{I_{K(n)} + \lambda_n(D^2)'D^2\}$ except the first two, the last two and the tth, and*

(ii) $\lim_{K(n)\to\infty}\sum_{j=1}^{K(n)} T_{t,j}(j-t)^k = 0$, *for* $k = 1, 2, 3$.

Proof. The definition of T_t guarantees that it is orthogonal to the $(t-1)$th and $(t+1)$th columns of Λ for any x. Since r_n satisfies $(1 - 4r_n + 6r_n^2 - 4r_n^3 + r_n^4) + r_n^2/\lambda_n = 0$, any linear combination of $\Re(r_n^{t-1}, r_n^{t-2}, \ldots, r_n, 1, r_n, r_n^2, \ldots, r_n^{K(n)-t})$ and $\Im(r_n^{t-1}, r_n^{t-2}, \ldots, r_n, 1, r_n, r_n^2, \ldots, r_n^{K(n)-t})$ is orthogonal to columns of Λ except for the first two, the last two, the $(t-1)$th, the tth and the $(t+1)$th. Combining these two results, we obtain (i).

Note that the jth element of T_t is equal to the $(2t-j)$th. As a result of the symmetry, when $K(n)$ is large enough, result (ii) of the proposition holds for the cases $k = 1, 3$. It remains to prove this result for $k = 2$. Note that

$$\sum_{j=t+1}^{K(n)} r_n^{j-t}(j-t)^2 = \sum_{j=1}^{K(n)-t} r_n^j j^2 \sim -\frac{2r_n}{(r_n-1)^2}\frac{r_n+1}{r_n-1},$$

$$(c_n + d_n i)\frac{r_n}{(r_n-1)^2} = \left\{r_n - 4 + \left(6 + \lambda_n^{-1}\right)r_n - 4r_n^2 + r_n^3\right\}\frac{r_n}{(r_n-1)^2}$$

$$= \frac{r_n^2 - 1 + \left\{1 - 4r_n + \left(6 + \lambda_n^{-1}\right)r_n^2 - 4r_n^3 + r_n^4\right\}}{(r_n - 1)^2} = \frac{r_n + 1}{r_n - 1}.$$

Hence

$$\sum_{j=t+1}^{K(n)} r_n^{j-t}(j-t)^2 \sim -\frac{2r_n}{(r_n-1)^2}\left\{(c_n + d_n\, i)\frac{r_n}{(r_n-1)^2}\right\} = (c_n + d_n\, i)\frac{-2r_n^2}{(r_n-1)^4}.$$

Since $T_{t,j} = \Re\{r_n^{j-t}(d_n + c_n\, i)\}$ for $j > t$,

$$\sum_{j=t+1}^{K(n)} T_{t,j}(j-t)^2 = \Re\left\{\sum_{j=t+1}^{K(n)} r_n^{j-t}(d_n + c_n\, i)(j-t)^2\right\}$$
$$\sim \Re\left\{(c_n + d_n\, i)(d_n + c_n\, i)\frac{-2r_n^2}{(r_n-1)^4}\right\} = 0. \qquad \square$$

Proof of Theorem 2. Since $r = \rho_n \exp(i\alpha_n)$, the equivalent kernel is proportional to the linear combination of $\rho_n^{|j-t|}\cos\{(j-t)\alpha_n\}$ and $\rho_n^{|j-t|}\sin(|j-t|\alpha_n)$. Since $K(n) \sim Cn^{\gamma}$ and $\gamma > 4/9$, we have $\bar{y}_t = f(\bar{x}_t) + \epsilon' + O(n^{-\gamma}) = f(\bar{x}_t) + \epsilon' + o(n^{-4/9})$, where ϵ' is distributed $N\{0, (K(n)/n)\sigma^2(\bar{x}_t)\}$.

At the end of this proof, we show that, for $h > 0$ given in (21) and for some $h' > 0$,

$$\rho_n = \exp\left\{-(Ch)^{-1}n^{1/9-\gamma}\right\}, \quad \alpha_n = h'K(n)^{-1}h^{-1}n^{1/9}, \tag{A5}$$

provided that λ_n satisfies (21).

Then, since $\bar{x}_t - \bar{x}_j = (t-j)/K(n)$, $\rho_n^{|t-j|} = \exp\{-h^{-1}n^{1/9}(C^{-1}n^{-\gamma}|t-j|)\} \sim \exp\{-|\bar{x}_t - \bar{x}_j|(hn^{-1/9})^{-1}\}$, and $\alpha_n|t-j| = h'|\bar{x}_t - \bar{x}_j|h^{-1}n^{1/9}$. Hence $\widehat{f}_n(x)$ is equivalent to the Nadaraya–Watson estimator with the kernel $T(x) = L^{-1}\{d'\,e^{-|x|}\cos(h'x) - c'\,e^{-|x|}\sin(h'|x|)\}$. Here L a normalizing factor. The constants $d' = \int_0^\infty x^2\,e^{-|x|}\sin(h'x)\,dx$ and $c' = \int_0^\infty x^2\,e^{-|x|}\cos(h'x)\,dx$ are determined by the vanishing second moment of the kernel; see point (ii) of Proposition 2. Using the indefinite integrals given by results 7 and 8 on p. 198 of Gradshteyn & Ryzhik (1980), one can show that $d' = -c'$. This proves (22), because at the end of this proof we show that $h' = 1$.

Since the kernel $T(x)$ is of the fourth order, we have

$$E\widehat{f}_{\text{num}}(x) = \frac{1}{hn^{-4/9}}\int_0^1 T\left(\frac{x-s}{hn^{-1/9}}\right)f(s)ds + O(K(n)^{-1})$$
$$= f(x) + \frac{h^4 n^{-4/9}}{24}f^{(4)}(x)\int u^4 T(u)du + o\left(n^{-4/9}\right).$$

By standard arguments for kernel estimators, e.g. in Wand & Jones (1995), for any $x \in (0,1)$, we have that $n^{4/9}\{\widehat{f}_n(x) - f(x)\} \to N\{\mathcal{B}_1(x), \mathcal{V}_1(x)\}$, in distribution, where $\mathcal{B}_1(x) = 24^{-1}h^4 f^{(4)}(x)\int x^4 T(x)dx$ and $\mathcal{V}_1(x) = h^{-1}\sigma^2(x)\int T^2(x)dx$.

We now show that (A5) holds if λ_n satisfies (21). First, note that r_n satisfies $r_n^4 - 4r_n^3 + (6 + 1/\lambda_n)r_n^2 - 4r_n + 1 = 0$. One possible solution for r_n is

$$r_n = 1 - \tfrac{1}{2}(-\lambda_n)^{-1/2} - \tfrac{1}{2}\left\{-4(-\lambda_n)^{-1/2} - \lambda_n^{-1}\right\}^{1/2}$$
$$= 1 - \tfrac{1}{2}\mathcal{E}_1(\lambda_n) + i\left\{-\tfrac{1}{2}(\lambda_n)^{-1/2} + \tfrac{1}{2}\mathcal{E}_2(\lambda_n)\right\},$$

where

$$\mathcal{E}_1(\lambda_n) = \left\{\frac{-\lambda_n^{-1} + \left(\lambda_n^{-2} + 16\lambda_n^{-1}\right)^{1/2}}{2}\right\}^{1/2},$$

$$\mathcal{E}_2(\lambda_n) = \left\{\frac{\lambda_n^{-1} + \left(\lambda_n^{-2} + 16\lambda_n^{-1}\right)^{1/2}}{2}\right\}^{1/2}.$$

We assume that ρ_n is this solution. Hence

$$\begin{aligned}\rho_n^2 &= \left\{1-\frac{1}{2}\mathcal{E}_1(\lambda_n)\right\}^2+\left\{-\frac{1}{2}\lambda_n^{-1/2}+\frac{1}{2}\mathcal{E}_2(\lambda_n)\right\}^2\\ &= 1+\frac{\left(\lambda_n^{-2}+16\lambda_n^{-1}\right)^{1/2}}{4}+\frac{\lambda_n^{-1}}{4}-\mathcal{E}_1(\lambda_n)-\lambda_n^{-1}\left\{\mathcal{E}_2(\lambda_n)\right\}^{-1}\\ &= 1-2^{1/2}\lambda_n^{-1/4}+o\left(\lambda_n^{-1/4}\right),\\ 2^{-1}\log(\rho_n^2) &= -2^{-1/2}\lambda_n^{-1/4}+o\left(\lambda_n^{-1/4}\right)=-(Ch)^{-1}n^{1/9-\gamma}. \qquad \text{(A6)}\end{aligned}$$

Note that $\rho_n \to 1$ and $\alpha_n \to 0$, when $\lambda_n \to \infty$. Thus $\rho_n \sin(\alpha_n) = (2)^{-1/2}\lambda_n^{-1/4} + o(\lambda_n^{-1/4})$. Moreover, $\alpha_n \sim (Ch)^{-1}n^{1/9-\gamma}$ and $h' \sim \alpha_n h C n^{\gamma-1/9} = 1$, so the kernel can be simplified to (22). □

In order to prove Theorem 3, we need the following result. We will consider the case of the left-hand boundary only, since the right-hand boundary is similar.

Proposition A3. *Let $\Lambda = \{I_{K(n)} + \lambda_n (D^2)' D^2\}$. Suppose that t depends on n in such a way that $t/K(n) = x = cn^{-1/5}$ for some $c \geqslant 0$. As before, T_t is defined by (A4) where r_n is a root of (A3) that has magnitude less than 1. Also, $\gamma > 2/5$. Let $S_t = (r_n^t, r_n^{t+1}, \ldots, r_n^{K(n)+t-1})$. Denote the tth element of T_t and S_t by $T_{t,j}$ and $S_{t,j}$ respectively. Denote the tth column of the matrix Λ by $\Lambda_{\cdot,t}$. Define $u_{t,\ell'} = \sum_{j=1}^{K(n)} T_{t,j}\Lambda_{j,\ell'}$, $v_{t,\ell'} = \sum_{j=1}^{K(n)} S_{t,j}\Lambda_{j,\ell'}$. Let*

$$\beta_t^I = \begin{vmatrix}\Im(v_{t,2}) & u_{t,2}\\ \Im(v_{t,1}) & u_{t,1}\end{vmatrix}, \qquad \beta_t^{II} = \begin{vmatrix}u_{t,2} & \Re(v_{t,2})\\ u_{t,1} & \Re(v_{t,1})\end{vmatrix}, \qquad \beta_t^{III} = \begin{vmatrix}\Re(v_{t,2}) & \Im(v_{t,2})\\ \Re(v_{t,1}) & \Im(v_{t,1})\end{vmatrix}.$$

Define $T_t' = \beta_t^I \Re(S_t) + \beta_t^{II}\Im(S_t) + \beta_t^{III} T_t$. Then

(i) *T_t' is orthogonal to the columns of Λ except the last two and the tth, and*

(ii) $\lim_{K(n)\to\infty} \sum_{j=1}^{K(n)} (j-t)T_{t,j}' = 0$.

Proof. By part (i) of Proposition 2, T_t' is orthogonal to all columns of Λ, except possibly the first two, the last two and the tth. Moreover, β_t^I, β_t^{II} and β_t^{III} have been chosen such that T_t' is orthogonal to the first two columns of Λ. To see this, note that the inner product of T_t' and $\Lambda_{\cdot,j}$ is $\{\Re(v_{t,j})\,\beta_t^I + \Im(v_{t,j})\,\beta_t^{II} + u_{t,j}\,\beta_t^{III}\}$, and Cramer's rule shows that

$$\begin{pmatrix}\Re(v_{t,1}) & \Im(v_{t,1})\\ \Re(v_{t,2}) & \Im(v_{t,2})\end{pmatrix}\begin{pmatrix}\beta_t^I/\beta_t^{III}\\ \beta_t^{II}/\beta_t^{III}\end{pmatrix} = -\begin{pmatrix}u_{t,1}\\ u_{t,2}\end{pmatrix}.$$

Thus, (i) holds.

To save space, in the remainder of the proof '$\lim_{K(n)\to\infty}$' will be abbreviated to 'lim'. By the definition of β_t^I, β_t^{II} and β_t^{III}, to prove (ii) it is enough to show that

$$\begin{vmatrix}\lim\sum_{j=1}^{K(n)}(j-t)\Re(S_{t,j}) & \lim\sum_{j=1}^{K(n)}(j-t)\Im(S_{t,j}) & \lim\sum_{j=1}^{K(n)}(j-t)T_{t,j}\\ \Re(v_{t,2}) & \Im(v_{t,2}) & u_{t,2}\\ \Re(v_{t,1}) & \Im(v_{t,1}) & u_{t,1}\end{vmatrix} = 0. \qquad \text{(A7)}$$

Equation (A7) holds if we can prove that

$$\begin{pmatrix}\lim\sum_{j=1}^{K(n)}(j-t)S_{t,j}\\ \lim\sum_{j=1}^{K(n)}(j-t)T_{t,j}\end{pmatrix} = (-t+1)\begin{pmatrix}v_{t,1}\\ u_{t,1}\end{pmatrix} + (-t+2)\begin{pmatrix}v_{t,2}\\ u_{t,2}\end{pmatrix}.$$

Take

$$\lim\sum_{j=1}^{K(n)}(j-t)S_{t,j} = (-t+1)v_{t,1} + (-t+2)v_{t,2}, \qquad \text{(A8)}$$

as an example; the other equation can be proved similarly. Define $W = [(1-t) \quad (2-t) \cdots \{K(n)-t\}]'$, and note that $\Lambda W = W$. Hence

$$S_t \Lambda W = S_t W = \sum_{j=1}^{K} (j-t) S_{t,j}. \tag{A9}$$

The left-hand side of (A9) can also be rewritten as

$$\begin{aligned}\sum_{j=1}^{K(n)} (-t+j) S_t \Lambda_{\cdot,j} = {} & (-t+1)v_{1,1} + (-t+2)v_{1,2} + \sum_{j=3}^{K-2} (-t+j) S_t \Lambda_{\cdot,j} \\ & + \lambda_n \big[\{-t+K(n)-1\} S_t \big(0 \cdots 0 \quad 1 \quad -4 \quad 5+\lambda_n^{-1} \quad -2\big)' \\ & + \{-t+K(n)\} S_t \big(0 \quad \cdots \quad 0 \quad 1 \quad -2 \quad 1+\lambda_n^{-1}\big)'\big].\end{aligned}$$

Note that $\lim\{-t+K(n)-1\} S_t (0 \ \cdots \ 0 \ 1 \ -4 \ 5+\lambda_n^{-1} \ -2)' = 0$ and $\lim(-t+K) S_t (0 \ \cdots \ 0 \ 1 \ -2 \ 1+\lambda_n^{-1})' = 0$. Hence

$$\lim \sum_{j=1}^{K(n)} (j-t) S_{t,j} = \lim \sum_{j=1}^{K(n)} (-t+j) S_t \Lambda_{\cdot,j} = (-t+1)v_{1,1} + (-t+2)v_{1,2},$$

which proves (A8).

Let $\Lambda^{(t,\cdot)}$ be the tth row of Λ^{-1}. Since $S_t \Lambda_{\cdot,t} = 0$, when $t/K(n) \sim cn^{-1/5}$ for some $c \geqslant 0$ we have

$$\Lambda^{(t,\cdot)} \simeq T_t' / \big(\beta_t^{III} u_{t,t}\big) = \big(\beta_t^{I} \Re(S_t) + \beta_t^{II} \Im(S_t)\big) / \big(\beta_t^{III} u_{t,t}\big) + T_t / u_{t,t}$$

with the approximation errors converging to 0 exponentially fast.

When $t/K(n) \to x \in (0, 1)$, i.e., the nonboundary case, it follows from (i) of Proposition A2 that $T_t/u_{t,t}$ is approximately $\Lambda^{(t,\cdot)}$.

Now we want to derive the equivalent kernel. From Proposition A3, we see that the equivalent boundary kernel is proportional to a linear combination of $\Re(S_t)$, $\Im(S_t)$ and T_t. First, we can show that $T_t/u_{t,t}$ is still equivalent to the kernel defined in (22) with a new bandwidth $h_n = hn^{-1/5}$. From (A6), $\lambda_n \sim 4^{-1} K(n)^4 h_n^4$. Let ρ_n and α_n be chosen as

$$\begin{aligned}\rho_n &= \exp\big\{-(Ch)^{-1} n^{1/5-\gamma}\big\} \\ \alpha_n &= h' K(n)^{-1} h^{-1} n^{1/5},\end{aligned}$$

for some $h > 0$, $h' > 0$.

Note that $\rho_n^2 = 1 - 2^{1/2}\lambda_n^{-1/4} + o(\lambda_n^{-1/4})$ when $\lambda_n \to \infty$, which was shown at the end of the proof of Proposition A2, continues to hold. Hence h' is still 1. Recall that $\bar{x}_t - \bar{x}_j = (t-j)/K(n)$. We can conclude that $T_t/u_{t,t}$ is still equivalent to the kernel defined in (22) with a new bandwidth h_n.

Note that $-\bar{x}_t - \bar{x}_j = -(j+t-1)/K(n)$. Then $\Re(S_{t,j})$ is proportional to $\exp(|-\bar{x}_t - \bar{x}_j|/h_n) \cos(|-\bar{x}_t - \bar{x}_j|/h_n)$ and $\Im(S_{t,j})$ is proportional to $\exp(-|-\bar{x}_t - \bar{x}_j|/h_n) \sin(|-\bar{x}_t - \bar{x}_j|/h_n)$.

If we view $-\bar{x}_t$ as a reflection of $\bar{x}_t$, then we can view the boundary kernel $T'(\cdot, x)$ as a linear combination of the nonboundary kernel T centred at x and two other kernels centred at the reflection point $-x$:

$$T'(x', x) = T\left(\frac{x-x'}{\Delta}\right) + k_1 \exp\left(-\frac{x+x'}{\Delta}\right) \cos\left(\frac{x+x'}{\Delta}\right) + k_2 \exp\left(-\frac{x+x'}{\Delta}\right) \sin\left(\frac{x+x'}{\Delta}\right), \tag{A10}$$

where k_1 and k_2 are chosen such that $\int_{-x/\Delta}^{1} T'(x', x)dx' = 1$ and $\int_{-x/\Delta}^{1} x'T'(x', x)dx' = 0$. Note that the choice of k_1 and k_2 is unique because the three functions are linearly independent.

Furthermore, we can compare with Silverman's equivalent smoothing kernel. His kernel is also of second order, implying that it is the same as ours, except for a scaling difference, which can be subsumed into the equivalent bandwidth. Furthermore, when $x = 0$, the boundary kernel can also be simplified as $T'(x', 0) =$

$2\exp(-x')\cos(x')$. One proof is to show that $\Re(1, r, r^2, r^3, \ldots, r^{K(n)-1})$ is orthogonal to Λ except for the first and the last two columns. We only need to show that $\Re(1, r, r^2, r^3, \ldots, r^{K(n)-1})$ is orthogonal to $\Lambda_{\cdot,2}$, i.e. $\Re(-2\lambda + (5\lambda + 1)r - 4\lambda r^2 + \lambda r^3) = 0$. Recall that $\Re[r^{-1}\{\lambda - 4\lambda r + (6\lambda + 1)r^2 - 4\lambda r^3 + \lambda r^4\}] = 0$. Therefore, we only need to show that $\Re(-\lambda/r + 2\lambda - \lambda r) = 0$. Since $\lambda(r-1)^4 + r^2 = 0$, $\Re(-1/r - r + 2) = \Re\{(1 - 1/r)(1 - r)\} = \Re\{-(1 - r)^2/r\} = 0$. Note that $T_1'/u_{1,1}$ is approximately the first row of Λ^{-1}. From above,

$$\Re\left(1, r, r^2, r^3, \ldots, r^{K(n)-1}\right)/\left(\Re\left(1, r, r^2, r^3, \ldots, r^{K(n)-1}\right)\Lambda_{\cdot,1}\right),$$

with the approximation error converging exponentially fast to 0. □

Proof of Theorem 3. From Proposition A3, we see that the equivalent boundary kernel is proportional to a linear combination of $\Re(S_t)$, $\Im(S_t)$ and T_t and is, by (ii), of the second order. The kernel is given by (A10). By using standard calculations for second-order kernels to calculate the bias and variance, we obtain Theorem 3. □

Proof of Theorem 4. Let ρ_n be the root of the equation $(6^{-1} - \lambda_n)x^2 + (2/3 + \lambda_n)x + (6^{-1} - \lambda_n) = 0$, and $T_t = (\rho_n^{t-1}, \rho_n^{t-2}, \ldots, \rho_n, 1, \rho_n, \rho_n^2, \ldots, \rho_n^{K(n)-t})'$, with the '1' in the tth coordinate.

Similarly to the proof of Proposition A3, T_t is orthogonal to all columns of Λ except the first, the last, and the tth. Consider the nonboundary case where t is not too close to 1 or $K(n)$. This case also has $b_t \sim \{\sum_{j=1}^{K(n)} \rho_n^{|t-j|}\overline{y}_j\}\{\sum_{j=1}^{K(n)} \rho_n^{|t-j|}\}^{-1}$. According to Hall & Wand (1996), we have $\bar{y}_t = f(x_t) + \epsilon' + O\{K(n)^{-2}\}$ with $\epsilon' \sim N\{0, K(n)/N\sigma^2\}$. Compared to the zero-degree spline case, the bias due to binning is reduced from $K(n)^{-1}$ to $K(n)^{-2}$. Hence, if $K(n) \sim Cn^{\gamma}$ for some $C > 0$, then we only require $\gamma > 1/5$ instead of $2/5$.

Similarly to previous arguments, suppose $\rho_n = \exp\{-(Ch)^{-1}n^{1/5-\gamma}\}$, for some $h > 0$, where ρ_n satisfies $(6^{-1} - \lambda_n)x^2 + (2/3 + 2\lambda_n)x + 6^{-1} - \lambda_n = 0$. Since $\bar{x}_t - \bar{x}_j = (t - j)/K(n)$, $\rho_n^{|t-j|} = \exp\{-h^{-1}n^{1/5}(C^{-1}n^{-\gamma}|t - j|)\} \sim \exp\{-|\bar{x}_t - \bar{x}_j|/(hn^{-1/5})\}$, and hence we can use the exponential kernel as the zero-degree case. Then the conclusion can be obtained by following the argument in § 2·3.

If $\lambda_n \to \infty$, we can also obtain the optimal penalty. Since $-\log(\rho_n) = \{(36\lambda_n + 3)^{1/2} - 3\}(6\lambda_n - 1)^{-1} + o(\lambda_n^{-1/2}) = C^{-1}h^{-1}n^{1/5-\gamma}$. An optimal choice is $\lambda_n = C^2h^2n^{2\gamma-2/5}$. □

Proof of Theorem 5. The idea is similar to that in Stute (1984). We just give a brief outline here. First, note that the estimator is

$$\widehat{f}(\bar{x}_t) = \frac{(K(n)hn^{-1/5})^{-1}\sum_{j=1}^{K(n)} Y_t H\{(G_n(\bar{x}_t) - G_n(\bar{x}_j))/(hn^{-1/5})\}}{(K(n)hn^{-1/5})^{-1}\sum_{j=1}^{K(n)} H\{(G_n(\bar{x}_t) - G_n(\bar{x}_j))/(hn^{-1/5})\}}. \qquad \text{(A11)}$$

The given choice of penalty yields to a bandwidth $hn^{-1/5}$ and we still have that

$$\{K(n)hn^{-1/5}\}^{-1}\sum_{j=1}^{K(n)} H[\{G_n(\bar{x}_t) - G_n(\bar{x}_j)\}/(hn^{-1/5})] \to 1$$

in probability. Hence we only need to consider the numerator of (A11), which will be denoted $\widehat{f}_{\text{num}}(\bar{x}_t)$.

Since H is twice differentiable, Taylor expansion yields

$$\widehat{f}_{\text{num}}(\bar{x}_t) = \left(hn^{-1/5}\right)^{-1}\sum_{j=1}^{K} \bar{y}_j H\left[\{G(\bar{x}_t) - G(\bar{x}_j)\}/\left(hn^{-1/5}\right)\right]$$

$$+ \left(hn^{-1/5}\right)^{-2}\sum_{j=1}^{K} \bar{y}_j\{G_n(\bar{x}_t) - G_n(\bar{x}_t) - G(\bar{x}_j) + G(\bar{x}_j)\}H'\left[\{G(\bar{x}_t) - G(\bar{x}_j)\}/\left(hn^{-1/5}\right)\right]$$

$$+\left(hn^{-1/5}\right)^{-3}\sum_{j=1}^{K}\bar{y}_j\{G_n(\bar{x}_t)-G_n(\bar{x}_t)-G(\bar{x}_j)+G(\bar{x}_j)\}^2H''\left[\{G(\bar{x}_t)-G(\bar{x}_j)\}/\left(hn^{-1/5}\right)\right]/2$$

$$+\text{higher-order terms}=I_1+I_2+I_3+I_0.$$

First, we want to show that $n^{2/5}I_3\to 0$ in probability as $n\to\infty$.

Choose a constant $C_1>3/5$. Partition I_3 into two parts

$$I_3=I_3\chi_{|G(\bar{x}_t)-G(x)|\leqslant C_1hn^{-1/5}\log(n)}+I_3\chi_{|G(\bar{x}_t)-G(x)|\geqslant C_1hn^{-1/5}\log(n)}=:I_{31}+I_{32},$$

where χ is the indicator function. According to Stute (1982),

$$\sup_{x:|G(\bar{x}_t)-G(x)|\leqslant C_1(hn^{-1/5})\log(n)}\frac{n}{\left(hn^{-1/5}\right)\log(n)}|G_n(\bar{x}_t)-G_n(x)-G(\bar{x}_t)+G(x)|^2$$

is stochastically bounded. Notice that H'' is bounded,

$$\sum_{j=1}^{K}|\bar{y}_j|\chi_{|G(\bar{x}_t)-G(x)|\leqslant C_1hn^{-1/5}\log(n)}<\infty$$

in probability, so that $n^{2/5}I_{31}=O_p(n^{-1/5}\log(n))$.

Following the idea of Stute (1984), we can show that $n^{2/5}I_{32}$ is asymptotically equivalent to $-n^{2/5}(hn^{-1/5})^{-2}f(\bar{x}_t)W+O_p\{n^{2/5-C_1}\log^2(n)\}$, where W is

$$\int_{|G(\bar{x}_t)-G(x)|\geqslant C_1hn^{-1/5}\log(n)}|G_n(\bar{x}_t)-G_n(x)-G(\bar{x}_t)+G(x)|H'\left\{\frac{G(\bar{x}_t)-G(x)}{hn^{-1/5}}\right\}\{G_n(dx)-G(dx)\}.$$

Choose $C_1>3/5$. Since $W=O_p(n^{-C_1-1/5})$, we can also obtain $n^{2/5}I_{32}=o_p(1)$. Therefore $n^{2/5}I_3\to 0$ in probability as $n\to\infty$ and $n^{2/5}I_0$ is also negligible.

Secondly, $n^{2/5}I_2$ is asymptotically equivalent to

$$-n^{2/5}\left(hn^{-1/5}\right)^{-1}f(\bar{x}_t)\int H\left\{\frac{G(\bar{x}_t)-G(x)}{hn^{-1/5}}\right\}\{G_n(dx)-G(dx)\}. \tag{A12}$$

Let

$$Z_n^1=K(n)^{-1}h^{-2}\left(n^{-1/5}\right)^{-3/2}\sum_{i=1}^{K(n)}\{\bar{y}_j-f(\bar{x}_j)\}\{\tau_n(\bar{x}_t)-\tau_n(\bar{x}_j)\}H'\left[\frac{G(\bar{x}_t)-G(\bar{x}_j)}{hn^{-1/5}}\right],$$

where $\tau_n(x)=n^{1/2}[G_n(x)-G(x)]$. Then $E\{(Z_n^1)^2\}\to 0$. Since $Z_n^1\to 0$ in probability, $n^{2/5}I_2$ is asymptotically equivalent to

$$h^{-2}\left(n^{-1/5}\right)^{-3/2}\int f(x)[\tau_n(\bar{x}_t)-\tau_n(x)]H'\left\{\frac{G(\bar{x}_t)-G(x)}{hn^{-1/5}}\right\}G(dx).$$

Furthermore, let

$$Z_n^2=h^{-2}\left(n^{-1/5}\right)^{-3/2}\int|f(\bar{x}_t)-f(x)||\tau_n(\bar{x}_t)-\tau_n(x)|\left|H'\left\{\frac{G(\bar{x}_t)-G(x)}{hn^{-1/5}}\right\}\right|G(dx).$$

Then $Z_n^2 \to 0$ in probability and hence $n^{2/5} I_2$ is also asymptotically equivalent to

$$\begin{aligned} h^{-2}\left(n^{-1/5}\right)^{-3/2} f(\bar{x}_t) \int \{\tau_n(\bar{x}_t) - \tau_n(x)\} H' \left\{ \frac{G(\bar{x}_t) - G(x)}{hn^{-1/5}} \right\} G(dx) \\ = -h^{-2}\left(n^{-1/5}\right)^{-3/2} f(\bar{x}_t) \int \tau_n(x) H' \left\{ \frac{G(\bar{x}_t) - G(x)}{hn^{-1/5}} \right\} G(dx) \\ = -h^{-1}\left(n^{-1/5}\right)^{-1/2} f(\bar{x}_t) \int H \left\{ \frac{G(\bar{x}_t) - G(x)}{hn^{-1/5}} \right\} \tau_n(dx). \end{aligned}$$

Hence (A12) is valid.

Thirdly, let I_4 denote $n^{2/5}\{I_1 - E\hat{f}_{\text{num}}(\bar{x}_t) + I_2\}$. Then I_4 is a standardized sum of independent and identically distributed random variables, with

$$\begin{aligned} \text{var}(I_4) &= h^{-1}\left(hn^{-1/5}\right)^{-1} \left(\int E\left[\{Y - f(\bar{x}_t)\}^2 | x\right] H^2 \left\{ \frac{G(\bar{x}_t) - G(x)}{hn^{-1/5}} \right\} G(dx) \right. \\ &\quad \left. - \left[\int \{f(x) - f(\bar{x}_t)\} H \left\{ \frac{G(\bar{x}_t) - G(x)}{hn^{-1/5}} \right\} G(dx) \right]^2 \right) \\ &\to h^{-1}\sigma^2(\bar{x}_t) \int H^2(u) du = \mathcal{V}(\bar{x}_t). \end{aligned}$$

Since $E(Y^{2+\delta}) < \infty$, the Lindeberg condition is satisfied. Therefore, $n^{2/5}\{\widehat{f}_{\text{num}}(\bar{x}_t) - E\widehat{f}_{\text{num}}(\bar{x}_t)\} \to N\{0, \mathcal{V}(\bar{x}_t)\}$, in distribution.

For the bias, we have

$$\begin{aligned} E\widehat{f}_{\text{num}}(\bar{x}_t) - f(\bar{x}_t) &= hn^{-1/5} \int \{f(x) - f(\bar{x}_t)\} H \left\{ \frac{G(x) - G(\bar{x}_t)}{hn^{-1/5}} \right\} G(dx) \\ &= \left(hn^{-1/5}\right)^2 (f \circ G^{-1})^{(2)}\{G(\bar{x}_t)\} \int u^2 H(u) du/2 = n^{-2/5}\mathcal{B}(\bar{x}_t). \end{aligned}$$

Hence, we can conclude that $n^{2/5}\{\widehat{f}(\bar{x}_t) - f(\bar{x}_t)\} \to N\{\mathcal{B}(\bar{x}_t), \mathcal{V}(\bar{x}_t)\}$, in distribution. □

References

Agarwal, G. G. & Studden, W. J. (1980). Asymptotic integrated mean square error using least squares and bias minimizing splines. *Ann. Statist.* **8**, 1307–25.

Dow, M. (2003). Explicit inverses of Toeplitz and associated matrices. *ANZIAM J.* **44**, E185–E215.

Eilers, P. H. C. & Marx, B. D. (1996). Flexible smoothing with *B*-splines and penalties (with Discussion). *Statist. Sci.* **11**, 89–121.

Elaydi, S. (2005). *An Introduction to Difference Equations*. New York: Springer.

Fan, J. (1992). Design-adaptive nonparametric regression. *J. Am. Statist. Assoc.* **87**, 998–1004.

Gradshteyn, I. & Ryzhik, I. (1980). *Table of Integrals, Series, and Products.* New York: Academic Press.

Hall, P. & Opsomer, J. D. (2005). Theory for penalized spline regression. *Biometrika* **92**, 105–18.

Hall, P. & Wand, M. P. (1996). On the accuracy of binned kernel density estimators. *J. Mult. Anal.* **5**, 165–84.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems (with Discussion). *Statist. Sci.* **1**, 505–27.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *J. Comp. Graph. Statist.* **11**, 735–57.

Ruppert, D. & Wand, M. P. (1994). Multivariate locally weighted least squares regression. *Ann. Statist.* **22**, 1346–69.

Ruppert, D., Wand, M. P. & Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge: Cambridge University Press.

Silverman, B. W. (1984). Spline smoothing: the equivalent variable kernel method. *Ann. Statist.* **12**, 898–916.

Stute, W. (1982). The oscillation behavior of empirical processes. *Ann. Prob.* **10**, 86–107.

STUTE, W. (1984). Asymptotic normality of nearest neighbor regression function estimates. *Ann. Statist.* **12**, 917–26.

WAND, M. P. (1999). On the optimal amount of smoothing in penalized spline regression. *Biometrika* **86**, 936–40.

WAND, M. P. & JONES, M. C. (1995). *Kernel Smoothing*. London: Chapman & Hall/CRC.

YU, Y. & RUPPERT, D. (2002). Penalized spline estimation for partially linear single index model. *J. Am. Statist. Assoc.* **97**, 1042–54.

[*Received September* 2006. *Revised September* 2007]