On The Asymptotics Of Penalised Splines

YINGXING LI
Department of Statistical Science, Malott Hall, Cornell University, NY 14853
yl377@cornell.edu

DAVID RUPPERT
School of Operational Research and Information Engineering
Rhodes Hall, Cornell University, NY 14853, USA
dr24@cornell.edu

February 15, 2007

Summary

The asymptotic behavior of penalised spline estimators is studied in the univariate case. B-splines are used and a penalty is placed on $m$th order differences of the coefficients. The number of knots is assumed to converge to $\infty$ as the sample size increases. We show that penalised splines behave similarly to Nadaraya-Watson kernel estimators with an “equivalent” kernels depending upon $m$. The equivalent kernels we obtain for penalised splines are the same as those found by Silverman for smoothing splines. The asymptotic distribution of the penalised spline estimate is Gaussian and we give simple expression for the asymptotic mean and variance. Providing that it is fast enough, the rate at which the number of knots converges to $\infty$ does not affect the asymptotic distribution. The optimal rate of convergence of the penalty parameter is given. Penalised splines are not design-adaptive.

Some key words: Asymptotic bias; Binning; B-splines; 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 \( x_t \), \( \epsilon_t \) 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 penalised spline estimators of \( f \).

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

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

where \( \Delta \) 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 \( \hat{f}(x) = \sum_{k=1}^{K+p} \hat{b}_k B_k^{[p]}(x) \) was introduced by Eilers and Marx (1996) and called a P-spline.

Let \( X^{[p]} \) be the \( n \times (K + p) \) 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 \((K + p - m) \times (K + p)\) differencing matrix satisfying

\[
D^m b = \begin{pmatrix} \Delta^m(b_{m+1}) \\ \vdots \\ \Delta^m(b_{K+p}) \end{pmatrix}.
\]

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

\[
\Omega^{[p,m]}_{K,\lambda^*} \hat{b} = (X^{[p]}'Y).
\]

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 Section 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 \( \hat{f} \) to \( f \) depends upon \( m \) but not upon \( p \) and \( K \).
provided only that $K \to \infty$ fast enough with the minimum rate depending on $p$. See Theorems 1 and 2 where $K$ is of order $n^\gamma$ and only a lower bound for $\gamma$ 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 penalised spline has two components: 1) modeling bias due to approximating the regression function by a spline, and 2) shrinkage bias due to estimation by penalised rather than ordinary least squares. In the theory presented here, $K$ grows sufficiently rapidly with $n$ that the modeling bias is asymptotically negligible compared to the shrinkage bias. This result agrees with finite-sample examples in Ruppert (2002) where the modeling 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 $\kappa_0 = 0, \kappa_1 = 1/K, \kappa_2 = 2/K, \ldots, \kappa_K = 1$. In Section 4, these results are generalized to unequally spaced design points and knots using a technique of Stute (1984). An interesting finding is that penalised 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 that in the interior.

Penalised splines use less 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 penalised splines are discussed extensively in Ruppert, Wand and Carroll (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 compared penalised splines.
with other nonparametric regression techniques. One exception are papers such as Yu and 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 and Opsomer (2005). The results of Hall and Opsomer differ in several major ways from the results presented here. First, Hall and Opsomer use an approximation where knots are placed continuously, that is, there is a knot at every value of $x$ in some interval. Moreover, the results in Hall and Opsomer 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 penalised splines are not design-adaptive. Another advantage of our approach over that in Hall and Opsomer is that we can find the minimum rate at which the number of knots must converge to $\infty$ in order for the modeling bias to be negligible.

Sections 2 and 3 cover piecewise constant and linear splines, respectively, with, in both cases, equally-spaced data and knots and $m = 1$ or 2. Section 4 discusses the case of unequally-spaced data with knots a sample quantiles. Section 5 contains extensions to higher-order penalties and penalties on derivatives instead of derivatives.

2. Zero-degree Splines

2.1. Overview

Zero-degree splines are piecewise constant. The $k$th zero-degree B-spline is $B_k^{[0]}(x) = I\{\kappa_{k-1} < x \leq \kappa_k\}, 1 \leq k \leq K$. For simplicity, assume that $n/K$ is an integer, which will be denoted by $M$. This assumption implies that every $M$th $x_i$ is a knot, that is $\kappa_j = x_{jM}$ for $j = 1, \ldots, K$. If $n/K$ is not integer, we could define $M = \lfloor n/K \rfloor$, the integer part of $n/K$, and place a knot at every $M$th $x_i$ and at $x_n$. This would intro-
duce an asymptotically negligible boundary effect in that the number of data points in the last “bin” would be less than 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 \(n \times K\) matrix with \(t, j\)th entry equal to \(B_j^{[0]}(x_t)\). Then \((X^{[0]})'X^{[0]} = MI_K\) where \(I_K\) is the \(K \times K\) identity matrix. Therefore, by (2), the penalised least-squares estimator solves

\[
\{I_K + \lambda(D^m)'D^m\}\hat{b} = \bar{y}. \tag{3}
\]

where \(\lambda = \lambda^*/M\) and \((X'Y/M) = \bar{y} = (\bar{y}_1, \ldots, \bar{y}_K)'\) where \(\bar{y}_k\) is the average of all \(y_t\) such that \(\kappa_{k-1} < x_t \leq \kappa_k\).

2.2. Solving banded linear equations

The matrix \(I_K + \lambda(D^m)'D^m\) in (3) and, more generally, \(\Omega_{K,\lambda^*}^{[p,m]}\) in (2) have a pattern that we will use to study the asymptotic behavior, as \(K \to \infty\), 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 \(\Omega_{K,\lambda^*}^{[p,m]}\) has the form

\[
(0 \cdots 0 \omega_q \cdots \omega_1 \omega_0 \omega_1 \cdots \omega_q \cdots 0)'
\tag{4}
\]

where \(\omega_0\) is the diagonal entry and \(\omega_q \neq 0\) by definition of \(q\). We will solve (2) by finding a vector \(T_t\) that is orthogonal to all columns of \(\Omega_{K,\lambda^*}^{[p,m]}\) except the \(t\)th and the first and last \(q\) columns. Moreover, for estimation in the interior, that is, for \(t/K \to x \in (0, 1)\), \(T_t\) will also be asymptotically orthogonal to the first and last \(q\) columns. Assume that there is root \(\rho\), possibly complex, of modulus less than 1 of the polynomial

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

Define \(T_t(\rho) = (\rho^{t-1} \cdots \rho^2 \rho^1 \rho \rho^2 \cdots \rho^{K+p-t})\). Then \(T_t(\rho)\) is orthogonal to all columns of \(\Omega_{K,\lambda^*}^{[p,m]}\) except the first and last \(q\) columns and columns \(t - q +\)
1, \cdots, t + q - 1. If we can find \( q \) distinct roots of \( P \), say \( \rho_1, \ldots, \rho_q \), all less than 1 in modulus, then we can find a linear combination, say \( S_t \), of \( T(\rho_1), \ldots, T(\rho_q) \) that is orthogonal to columns \( t - q + 1, \cdots, t - 1 \) and columns \( t + 1, \cdots, t + q - 1 \). Moreover, since \( |\rho_j| < 1 \) for \( j = 1, \ldots, q \), \( S_t \) is asymptotically orthogonal to all columns except the \( t \)th, assuming that \( t/K \rightarrow x \in (0, 1) \). As we will see, the boundary case 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 non-zero since \( q \neq 0 \). By the symmetry of the coefficients of \( P \), if \( \rho \) is a root, then \( 1/\rho \) is also a root. Thus, if the roots of \( p \) are distinct and none of them have modulus 1, then there will be exactly \( q \) roots less than 1 in modulus. Numerical experimental suggest that this is always the case with the matrix \( \Omega^{[p,m]}_{K,\lambda} \) in (2). In certain situation, we have a proof that no roots have modulus 1; see Proposition 1.

In Section 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 \( \Omega^{[p,m]}_{K,\lambda} \) is a Toeplitz matrices 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 behavior of solutions to (2) in a direct manner. Also, since \( \rho \) is a root of \( P(\rho) \), we see that \( G(n) = \rho^n \) is a solution to the homogeneous difference equation 
\[
\omega_q G(n) + \omega_{q-1} G(n-1) + \cdots + \omega_1 G(n-q) + \cdots + \omega_0 G(n-2q) = 0.
\]
We used this fact when constructing \( T_t \). We had hoped to exploit the theory of difference equations, e.g., Elaydi (2005), in this research but we were unable to find an approach simpler than the one just given.

\section{First-order difference penalties: overview}

6
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 \((1 + 2\lambda)\) so that all diagonal elements, except the first and last, equal 1. This gives us

\[
\Lambda \hat{b} = z, \tag{5}
\]

where \( z = (z_1, \ldots, z_K)' = \bar{y}/(1 + 2\lambda) \) and \( \Lambda \) is the \( K \times K \) matrix with \( \Lambda_{ij} = 1 \) if \( 1 < t = j < K \), \( \Lambda_{ij} = \eta = -\lambda/(1+2\lambda) \) if \(|t-j| = 1\), \( \Lambda_{11} = \Lambda_{KK} = \theta = (1+\lambda)/(1+2\lambda) \), and \( \Lambda_{ij} = 0 \) if \(|t-j| > 1\).

To solve (5) we apply the methodology described in Section 2.2. In our case, \( m = 1 \) and \( p = 0 \), so \( q = 1 \). Let \( \rho \) be the solution between 0 and 1 of

\[
P(\rho) = \eta + \rho + \eta \rho^2 = 0. \tag{6}
\]

\((P(0) < 0 \text{ and } P(1) > 0\) so such a solution must exist.) Then

\[
\rho = \frac{1 - (1 - 4\eta^2)^{1/2}}{2 \eta} = \frac{1 + 2\lambda - (1 + 4\lambda)^{1/2}}{2\lambda}. \tag{7}
\]

We now solve for \( \hat{b}_1 \) and \( \hat{b}_K \). Let \( T_t = T_t(\rho) = (\rho^{t-1}, \rho^{t-2}, \ldots, 1, \rho, \rho^2, \ldots, \rho^{K-t})' \). By (6), \( T_t \) is orthogonal to all columns of \( \Lambda \) except the first, last, and \( t \)th. In particular, \( T_1 \) and \( T_K \) are orthogonal to all columns except the first and last, which makes it easy to solve for \( \hat{b}_1 \) and \( \hat{b}_K \). Multiplying both sides of (5) by \( T_t' \) we obtain \( \{\theta + \eta \rho\} \hat{b}_1 + \rho^{K-2}(\eta + \theta \rho) \hat{b}_K = \sum_{k=1}^{K} \rho^{k-1} z_k \) and then multiplying both sides of (5) by \( T_K' \) we get \( \rho^{K-2}(\eta + \theta \rho) \hat{b}_1 + (\theta + \eta \rho) \hat{b}_K = \sum_{k=1}^{K} \rho^{K-k} z_k \). Therefore,

\[
\hat{b}_1 = \frac{(\theta + \eta \rho) \left( \sum_{k=1}^{K} \rho^{k-1} z_k \right) - \rho^{K-2}(\eta + \theta \rho) \left( \sum_{k=1}^{K} \rho^{K-k} z_k \right)}{(\theta + \eta \rho)^2 - \rho^{2(K-2)}(\eta + \theta \rho)^2}. \tag{8}
\]

We will choose \( \lambda \) so that \( \rho \), which is a function of \( \lambda \), satisfies \( \rho^K = \exp(-n^{1/5} h^{-1}) \) for some positive constant \( h \); see (17) below. Then \( \hat{b}_1 \sim (\sum_{k=1}^{K} \rho^{k-1} z_k)/(\theta + \eta \rho) = (\sum_{k=1}^{K} \rho^{k-1} \bar{y}_k)/(\theta + \eta \rho)(1 + 2\lambda) \), where \( a_n \sim c_n \) means that \( a_n/c_n \rightarrow 1 \). Also

\[
\hat{b}_K = \frac{-\rho^{K-2}(\eta + \theta \rho) \left( \sum_{k=1}^{K} \rho^{k-1} z_k \right) + (\theta + \eta \rho) \left( \sum_{k=1}^{K} \rho^{K-k} z_k \right)}{(\theta + \eta \rho)^2 - \rho^{2(K-2)}(\eta + \theta \rho)^2}, \tag{9}
\]
so that \( \hat{b}_K \sim (\sum_{k=1}^{K} \rho^{K-k} \tilde{y}_k) / \{(\theta + \eta \rho)(1 + 2\lambda)\} \).

After some algebra, one can show that \((\theta + \eta \rho)(1 + 2\lambda) = 1 + \lambda - \rho \lambda = \{1 + (1 + 4\lambda)^{1/2}\}/2\) since \(\rho \lambda = 1/2 + \lambda - (1 + 4\lambda)^{1/2}/2\). Also, \(1/(1 - \rho) = (2\lambda)/\{(1 + 4\lambda)^{1/2} - 1\} = \{1 + (1 + 4\lambda)^{1/2}\}/2\). Thus, \((\theta + \eta \rho)(1 + 2\lambda) = 1/(1 - \rho) \sim \sum_{k=1}^{K} \rho^{k-1}\), so that \(\hat{b}_1 \sim (\sum_{k=1}^{K} \rho^{k-1} \tilde{y}_k) / (\sum_{k=1}^{K} \rho^{k-1})\), with a similar result for \(\hat{b}_K\).

Multiplying both side of (5) by \(T_t\), \(1 < t < K\), one obtains

\[
(1 + 2\rho \eta) \hat{b}_t = \sum_{j=1}^{K} \rho^{|t-j|} z_j - \left\{ (\rho^{t-1} \theta + \rho^{t-2} \eta) \hat{b}_1 + (\rho^{K-t-1} \eta + \rho^{K-t} \theta) \hat{b}_K \right\}. \tag{10}
\]

Substituting (8) and (9) into (10) gives an exact expression for \(\hat{b}_t\).

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

\[
B_t^{[0]}(x) = 1 \text{ for all } n, \tag{11}
\]

so that \(x\) is in the \(t\)th bin and \(t/K \to x\). Then

\[
\hat{b}_t \sim \sum_{j=1}^{K} \rho^{|t-j|} \tilde{y}_j / \{(1 + 2\rho \eta)(1 + 2\lambda)\}. \tag{12}
\]

Also by (11), \(t \to \infty\) so that

\[
\sum_{j=1}^{K} \rho^{|t-j|} \sim \sum_{j=-\infty}^{\infty} \rho^{|t-j|} = \frac{1}{1 - \rho} + \frac{\rho}{1 - \rho} = \frac{1 + \rho}{1 - \rho} = \frac{(1 + 4\lambda) - (1 + 4\lambda)^{1/2}}{-1 + (1 + 4\lambda)^{1/2}}. \tag{13}
\]

Multiplying numerator and denominator in the right hand side of (13) by \((1 + 4\lambda) + (1 + 4\lambda)^{1/2}\), this expression simplifies to \((1 + 4\lambda)^{1/2}\). Also, by (7), \(1 + 2\rho \eta = (1 - 4\eta^2)^{1/2}\) and some algebra shows that \(1 - 4\eta^2 = 1 - 4\lambda^2/(1 + 2\lambda)^2 = (1 + 4\lambda)/(1 + 2\lambda)^2\), so that

\[
(1 + 2\rho \eta)(1 + 2\lambda) = (1 + 4\lambda)^{1/2} \sim \sum_{j=1}^{K} \rho^{|t-j|}. \tag{14}
\]

8
Thus, in the non-boundary case, we have by (12) and (14) 

$$
\beta_t \sim \frac{\sum_{j=1}^{K} \rho^{|t-j|} \bar{y}_j}{\sum_{j=1}^{K} \rho^{|t-j|}}. 
$$

(15)

This result shows that, in the non-boundary case, the penalised spline with \( p = 0 \) and \( m = 1 \) is asymptotically equivalent to a binned Nadaraya-Watson (NW) 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

$$
K = Cn^\gamma \text{ with } C > 0 \text{ and } \gamma > 2/5,
$$

(16)

and that \( \lambda \) is chosen so that

$$
\rho = \exp\{- (Ch)^{-1} n^{1/5 - \gamma}\}.
$$

(17)

Let \( \hat{f}_n(x) \) be the first order penalised 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

$$
n^{2/5} \{ \hat{f}_n(x) - f(x) \} \Rightarrow N\{B(x), V(x)\}, \text{ as } n \to \infty,
$$

where \( B(x) = h^2 f^{(2)}(x) \), \( V(x) = 2^{-1} h^{-1} \sigma^2(x) \), and \( \Rightarrow \) denotes the convergence in distribution. The equivalent kernel is the double exponential or Laplace kernel

$$
H(x) = (1/2) \exp(-|x|).
$$

(18)

**Proof:** Let \( \overline{x}_t = (2t-1)/2K \) be the midpoint of the \( t \)th bin, i.e., of \(( (t-1)K^{-1}, tK^{-1} \). Since \( \gamma > 2/5 \), the effect of binning is asymptotically negligible when the bandwidth is of the optimal order, \( n^{-1/5} \). Specifically, we have \( \overline{y}_t = f(\overline{x}_t) + \epsilon' + o(n^{-2/5}) \) with \( \epsilon' \) distributed \( N\{0, (K/n)\sigma^2(\overline{x}_t)\} \). Since \( \overline{x}_t - \overline{x}_j = (t-j)/K \),

$$
\rho^{|t-j|} = \exp\{-h^{-1} n^{1/5} (C^{-1} n^{-\gamma} |t-j|)\} = \exp\{-|\overline{x}_t - \overline{x}_j| (hn^{-1/5})^{-1}\},
$$

(19)
by (16) and (17). Thus, by (15), \( \hat{f}_n \) is asymptotically equivalent to the NW estimator with kernel (18) and bandwidth \( hn^{-1/5} \). Therefore, one can derive the asymptotic distribution of \( \hat{f}_n(x) \) using well-known techniques, e.g., in Wand and Jones (1995).

\[ \square \]

Note that (17) will hold for some choice of \( \lambda \) such that

\[ \lambda \sim C^2h^2n^{2\gamma-2/5} \sim (Kh_n^{-1/5})^2, \quad (20) \]

where \( h > 0 \) is a constant. To show (20), combine equations (7) and (17) to obtain

\[ \frac{1+2\lambda-(1+4\lambda)^{1/2}}{(2\lambda)} = \rho = \exp\{-C^{-1}h^{-1/5}n^{1/5-\gamma}\}. \]

By (20), \( \lambda \to \infty \) so we have

\[ -\log(\rho) = \log[2\lambda/(1+2\lambda-(1+4\lambda)^{1/2})] = \log(1+\lambda^{-1/2})+o(\lambda^{-1}) = \lambda^{-1/2}+o(\lambda^{-1/2}) \sim C^{-1}h^{-1}n^{1/5-\gamma}. \]

Hence \( \lambda \) should be chosen as \( \lambda \sim C^2h^2n^{2\gamma-2/5} \sim (Kh_n^{-1/5})^2, \) by (16).

**Example 1:** To study how quickly the finite-sample kernel converges to (18), we consider four sample sizes, \( n = 40, 80, 160, \) and \( 320 \). In each case, \( K = Cn^{\gamma} \), rounded to the nearest integer, with \( C = 2 \) and \( \gamma = 0.45 > 2/5 \). Also, \( \lambda = (Kh_n^{-1/5})^2 \) with \( h = 0.4 \). Fig. 1 shows the finite-sample kernels and the double exponential kernel for all four values of \( n \). The double exponential kernel for fitting at the \( j \)th bin has value \( \rho^{t-j}/(\sum \rho^{t-j}) \) at the \( t \)th bin with \( \rho \) a root of (6). The kernels are for estimation at the center of the design; \( j = K/2 \) rounded to the nearest integer. We see good agreement between the finite-sample and asymptotic kernels for \( n = 40 \) and excellent agreement for \( n = 320 \).

**2.5. First-order penalties: estimation at the boundary**

The boundary case is slightly more complex than the non-boundary case. The bias is of order \( n^{-1/5} \) but it is not the same as the bias of the NW estimator, though the NW bias is also of order \( n^{-1/5} \).

To find the equivalent kernel at the left boundary, we suppose that \( t/K \to 0 \) as
\( n \to \infty \) at look at the \( t \)th bin. Then from (10), we have
\[
(1 + 2\rho \eta)\hat{b}_t \sim \sum_{j=1}^{K} \rho^{t-j}z_j - \rho^t(\rho^{-1}\theta + \rho^{-2}\eta)\hat{b}_1
\]
\[
\sim \sum_{j=1}^{K} \rho^{t-j}z_j - \rho^t(\rho^{-1}\theta + \rho^{-2}\eta)\sum_{j=1}^{K} \rho^{-1}z_j/(\theta + \eta \rho)
\]
so that
\[
\hat{b}_t \sim \sum_{j=1}^{K} (a_1 \rho^{t-j} + a_2 \rho^{t+j})z_j,
\]
where \( a_1 = (1 + 2\rho \eta)^{-1} \) and \( a_2 = (\theta + \eta \rho^{-1})/\{\rho^2(\theta + \eta \rho)\} \). By (19) and (21),
\[
\hat{b}_t \sim \sum_{j=1}^{K} H(\bar{x}_t, \bar{x}_j; h n^{-1/5})z_j,
\]
where \( H \) is the equivalent boundary kernel such that
\[
H(\bar{x}_t, \bar{x}_j; h) \propto [a_1 \exp(-|\bar{x}_t - \bar{x}_j|/h) + a_1 a_2 \exp(-(\bar{x}_t + \bar{x}_j)/h)].
\]
The constant of proportionality is determined by \( \sum_{j=1}^{K} H(\bar{x}_t, \bar{x}_j; h n^{-1/5}) = 1 \).

**Example 2:** Fig. 2 compares the finite-sample kernel with the asymptotic boundary kernel given by (21), both for estimation at the 3rd bin. The sample sizes are 40, 80, 160, and 320, and \( K \) and \( \lambda \) are functions of \( n \) suggested by the asymptotics. We see that the agreement is very good for \( n = 320 \). For smaller sample sizes, the finite-sample kernel is well above zero at both the left and right 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 extremely well.

2.6. Second-order penalties: overview

Now suppose that \( m = 2 \). Then \( D^m = D^2 \) is \( (K - 2) \times K \) and the \( t \)th row of \( D^2 \) has 1 in coordinates \( t \) and \( t + 2, -2 \) in coordinate \( t + 1 \), and 0 elsewhere. Except for \( t = 1, 2, K - 1, \) and \( K \), the \( t \)th column of \((D^2)^TD^2\) has entries, 1, \(-4, 6, -4, \) and 1 in rows \( t - 2 \) to \( t + 2 \) and 0 elsewhere. Now \( \hat{b} \) 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 \( K \sim Cn^\gamma \) with \( C > 0 \) and \( \gamma > 4/9 \), and that there exists a constant \( h > 0 \) such that \( \lambda \sim 4C^4h^4n^{4\gamma-4/9} \). Let \( \hat{f}_n(x) \) be the penalised estimator with \( p = 0 \), \( m = 2 \), and equally spaced knots. Then for any \( x \in (0, 1) \), when \( n \to \infty \), we have

\[
n^{4/9} \{ \hat{f}_n(x) - f(x) \} \Rightarrow N \{ B_1(x), V_1(x) \},
\]

where \( B_1(x) = (1/24)f^{(4)}(x)h^4 \int x^4 T(x)dx, \) \( V_1(x) = h^{-1} \left\{ \int T^2(x)dx \right\} \sigma^2(x), \) and \( T(x) \) is a fourth order kernel given by

\[
L^{-1} \left[ d' \exp(-|x|) \cos(x) - c' \exp(-|x|) \sin(|x|) \right],
\]

where \( L \) is a normalizing constant and \( c' \) and \( d' \) are determined by the requirement that the second moment of the kernel is zero.

Before proving the theorem, we need some preliminary results. Define

\[
w(\xi) = \lambda(1 - 4\xi + 6\xi^2 - 4\xi^3 + \xi^4) + \xi^2 = \lambda(1 - \xi)^4 + \xi^2, \quad \lambda > 0.
\]

As discussed in Section 2.2, the roots of \( w \) will be used to find a vector orthogonal to all columns of \( \Lambda \) except the first and last two and the \( t \)th. Clearly, \( w \) has no real roots. Also, if \( r \) is a root of \( w \), then so also is \( r^{-1} \). Thus, for some complex \( r \), the four roots of \( w \) are \( r, \text{conj}(r), r^{-1}, \) and \( \text{conj}(r)^{-1} \), where \( \text{conj}(r) \) is the complex conjugate of \( r \).

By the following proposition, Tone of the roots \( r \) and \( r^{-1} \) is less than 1 in magnitude and without loss of generality we can assume that it is \( r \).

**Proposition 1** No root of \( w \) has modulus equal to 1.
Proof of Proposition 1: Suppose there is a $\xi$ such that $w(\xi) = 0$ and $\xi = \exp(i\theta)$. Here $i = (-1)^{1/2}$. Notice that $\xi - 1 = 2 \sin(\theta/2) \exp\{(\pi/2 + \theta/2)i\}$, so $-\lambda(\xi - 1)^4 = \xi^2$ implies that

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

Comparing the real and image part on both sides, we have, $16\lambda \sin^4(\theta/2) = -1$. For any positive $\lambda$, this is impossible. So there will be no root with norm 1. \qed

Since $|\rho| < 1$, $r = \rho \exp(i\alpha)$ for some $\rho$ in $(0, 1)$. Therefore, we have the following proposition.

**Proposition 2** Let $c$ and $d$ be the real and imaginary parts of $r - 4 + (6 + \lambda^{-1})r - 4r^2 + r^3$ where $r$ is defined as above. Let $T_t = (T_{t,1}, \ldots, T_{t,K})$ be defined by

$$T_t = d \Re(r^{t-1}, r^{t-2}, \ldots, r, 1, r, \ldots, r^{K-t}) - c \Im(r^{t-1}, r^{t-2}, \ldots, r, 1, r, \ldots, r^{K-t}),$$

(25)

where $\Re$ and $\Im$ are the real and imaginary parts, respectively, and each “1” is in the $t$th position. Then

1) $T_t$ is orthogonal to all columns of $\{I_K + \lambda(D^2)'D^2\}$ except the first two, last two and $t$th.

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

Proof of Proposition 2: The definition of $T_t$ guarantees that it is orthogonal to the $(t - 1)$th and $(t + 1)$th column of $\Lambda$ for any $x$. When $r$ satisfies $(1 - 4r + 6r^2 - 4r^3 + r^4) + r^2/\lambda = 0$, any linear combination of $\Re(r^{t-1}, r^{t-2}, \ldots, r, 1, r, r^2, \ldots, r^{K-t})$ and $\Im(r^{t-1}, r^{t-2}, \ldots, r, 1, r, r^2, \ldots, r^{K-t})$ is orthogonal to columns of $\Lambda$ except the first two, last two, $(t - 1)$th, $t$th and $(t + 1)$th. Combining these two results, we obtain 1).

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

$$\sum_{j=t+1}^K r^{j-t}(j - t)^2 = \sum_{j=1}^{K-t} r^j j^2 \sim -\frac{2r}{(r - 1)^2} \frac{r + 1}{r - 1}.$$
and
\[
(c + d \cdot i) \frac{r}{(r - 1)^2} = (r - 4 + (6 + \lambda^{-1})r - 4r^2 + r^3) \frac{r}{(r - 1)^2} = \frac{r^2 - 1 + (1 - 4r + (6 + \lambda^{-1})r^2 - 4r^3 + r^4)}{(r - 1)^2} = \frac{r + 1}{r - 1}.
\]

Hence
\[
\sum_{j=t+1}^{K} r^{j-t}(j - t)^2 \sim -\frac{2r}{(r - 1)^2} \left( (c + d \cdot i) \frac{r}{(r - 1)^2} \right) = (c + d \cdot i) \frac{-2r^2}{(r - 1)^4}.
\]

Since \( T_{i,j} = \Re \left\{ r^{j-t}(d + c \cdot i) \right\} \) for \( j > t \),
\[
\sum_{j=t+1}^{K} T_{i,j}(j - t)^2 = \Re \left\{ \sum_{j=t+1}^{K} r^{j-t}(d + c \cdot i)(j - t)^2 \right\} \sim \Re \left\{ (c + d \cdot i)(d + c \cdot i) \frac{-2r^2}{(r - 1)^4} \right\} = 0.
\]

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

For some \( h > 0 \) and \( h' > 0 \), \( \rho = \exp\{-Ch^{-1}n^{1/9-\gamma}\} \), and \( \alpha = h'Kh^{-1}n^{1/9} \). Then, since \( \bar{x}_t - \bar{x}_j = (t - j)/K \), \( \rho^{j-t} = \exp\{-h^{-1}n^{1/9}(C^{-1}n^{-\gamma}|t - j|)\} \sim \exp\{-|\bar{x}_t - \bar{x}_j|\ (hn^{-1/9})^{-1}\} \), and \( \alpha|t - j| = h'|\bar{x}_t - \bar{x}_j|hn^{-1/9} \). Hence \( \hat{f}_n(x) \) is equivalent to the NW estimator with the kernel \( T(x) = L^{-1}\{d'e^{-|x|} \cos(h'x) - c'e^{-|x|} \sin(h'x)\} \). Here \( L \) is a normalized factor. The constants \( d' = \int x^2 e^{-|x|} \sin(h'x) \) and \( c' = \int x^2 e^{-|x|} \cos(h'x) \) are determined by the vanishing second moment of the kernel; see point 2) of Proposition 2.

Since the kernel \( T(x) \) is of the fourth order, we have
\[
E \hat{\phi}_{\text{num}}(x) = \frac{1}{hn^{-4/9}} \int_0^1 T \left( \frac{x - s}{hn^{-1/9}} \right) f(s) ds + O(K^{-1}) = f(x) + \frac{h^4n^{-4/9}}{24} f^{(4)}(x) \int u^4 T(u) du + o(n^{-4/9}).
\]
By standard arguments for kernel estimators, e.g., in Wand and Jones (1995), for any $x \in (0, 1)$, we have $n^{4/9} \{ \hat{f}_n(x) - f(x) \} \Rightarrow N \{ B_1(x), V_1(x) \}$, where $B_1(x) = 24^{-1} h^4 f^{(4)}(x) \int x^4 T(x) dx$, and $V_1(x) = h^{-1} \sigma^2(x) \int T^2(x) dx$.

To derive the optimal rate for $\lambda$, first notice that $r$ satisfies $r^4 - 4r^3 + (6 + 1/\lambda)r^2 - 4r + 1 = 0$. One possible solution for $r$ is

$$r = 1 - \frac{1}{2}(\lambda)^{-1/2} - \frac{1}{2} \{ 4(\lambda)^{-1/2} - \lambda^{-1} \}^{1/2} = 1 - \frac{1}{2} \mathcal{E}_1(\lambda) + \frac{1}{2} \mathcal{E}_2(\lambda),$$

where

$$\mathcal{E}_1(\lambda) = \left\{ \frac{-\lambda^{-1} + (\lambda^{-2} + 16\lambda^{-1})^{1/2}}{2} \right\}^{1/2} \quad \text{and} \quad \mathcal{E}_2(\lambda) = \left\{ \frac{\lambda^{-1} + (\lambda^{-2} + 16\lambda^{-1})^{1/2}}{2} \right\}^{1/2}.$$

Hence

$$\rho^2 = \left\{ \left\{ 1 - \frac{1}{2} \mathcal{E}_1(\lambda) \right\}^2 + \left\{ \frac{1}{2} \lambda^{-1/2} - \mathcal{E}_2(\lambda) \right\}^2 \right\}^2 = 1 + \frac{(\lambda^{-2} + 16\lambda^{-1})^{1/2}}{4} + \frac{\lambda^{-1}}{4} - \mathcal{E}_1(\lambda) - \lambda^{-1} \{ \mathcal{E}_2(\lambda) \}^{-1} = 1 - 2^{1/2} \lambda^{-1/4} + o(\lambda^{-1/4}),$$

and $2^{-1} \log(\rho^2) = -2^{-1/2} \lambda^{-1/4} + o(\lambda^{-1/4}) = -(Ch)^{-1} n^{1/9 - \gamma}$. Provided that $\lambda \to \infty$, the optimal penalty $\lambda$ is $\lambda \sim 4C^4 h^4 n^{4\gamma - 4/9}$.

Notice that $\rho \to 1$ and $\alpha \to 0$, when $\lambda \to 1$. Thus $\rho \sin(\alpha) = (2)^{-1/2} \lambda^{-1/4} + o(\lambda^{-1/4})$. So $\alpha = (Ch)^{-1} n^{1/9 - \gamma}$ and $h' = \alpha h C n^{-\gamma - 1/9} = 1$. Therefore, the kernel can be simplified to (23).

Example 3: Fig. 3 compares finite-sample and asymptotic kernels for four sample sizes, 40, 80, 160, and 320. For each sample size, $K$ was $2n^{1/2}$ rounded to the nearest integer and $\lambda = 4(Khn^{-1/9})^4$ where $h = 0.1$. There is reasonably good agreement between the finite-sample and asymptotic kernels, especially for $n = 160$ and larger.

2.8. Second-order penalties: estimation at the boundary

We now consider the boundary case where $x \to 0$ or 1 at a suitable rate.

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

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

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

Before proving this theorem, we need the following result.

**Proposition 3** Let $\Lambda = \{I_K + \lambda(D^2)'D^2\}$ and let be $T_t$ be defined as in (25). Denote the $t, j$th element of matrix $\Lambda$ by $\Lambda_{tj}$ and the $j$th element of $T_t$ by $T_{t,j}$. Let

$$u_1t = \sum_{j=1}^K T_{1,j} \Lambda_{2j}, \quad v_1t = \sum_{j=1}^K T_{2,j} \Lambda_{2j}, \quad s_1t = \sum_{j=1}^K T_{t,j} \Lambda_{2j};$$

$$u_2t = \sum_{j=1}^K T_{1,j} \Lambda_{1j}, \quad v_2t = \sum_{j=1}^K T_{2,j} \Lambda_{1j}, \quad s_2t = \sum_{j=1}^K T_{t,j} \Lambda_{1j}.$$

Let

$$\alpha_t = \begin{vmatrix} v_1t & s_1t \\ v_2t & s_2t \end{vmatrix}, \quad \beta_t = \begin{vmatrix} s_1t & u_1t \\ s_2t & u_2t \end{vmatrix}, \quad \gamma_t = \begin{vmatrix} u_1t & v_1t \\ u_2t & v_2t \end{vmatrix}.$$

Define $S_t = (S_{t,1}, \ldots, S_{t,K})' = \alpha_t T_1 + \beta_t T_2 + \gamma_t T_t$. Then

1. $S_t$ is orthogonal to the columns of $\Lambda$ except the last two and the $t$th, and

2. $\sum_{j=1}^K (j-t)S_{t,j} = 0$.

**Proof of Proposition 3:** The definition of $\alpha_t$, $\beta_t$ and $\gamma_t$ imply (1). To prove (2), we want to show

$$\begin{vmatrix} \sum_{j=1}^K (j-t)T_{1,j} & \sum_{j=1}^K (j-t)T_{1,j} & \sum_{j=1}^K (j-t)T_{1,j} \\ u_1t & v_1t & s_1t \\ u_2t & v_2t & s_2t \end{vmatrix} = 0.$$
The above argument is true if we can prove

\[
\frac{1}{\lambda} \left( \sum_{j=1}^{K} (j - t)T_{1,j} \right) = -(t - 1) \begin{pmatrix} u_{1,t} \\ v_{1,t} \\ s_{1,t} \end{pmatrix} - (t - 2) \begin{pmatrix} u_{2,t} \\ v_{2,t} \\ s_{2,t} \end{pmatrix}.
\]

Take

\[
\frac{1}{\lambda} \sum_{j=1}^{K} (j - t)T_{1,j} = -(t - 1)u_{1,t} - (t - 2)u_{2,t},
\] as an example, the others can be proved similarly. Notice that

\[
\Lambda (t + 1 \ t \ t - 1 \ \cdots \ 1) = \frac{1}{\lambda} (t + 1 \ t \ t - 1 \ \cdots \ 1).
\]

Hence

\[
T_{1}' \Lambda \left( \begin{array}{c}
-t + 1 \\
-t + 2 \\
\vdots \\
-t + K
\end{array} \right) = T_{1}' \frac{1}{\lambda} \left( \begin{array}{c}
-t + 1 \\
-t + 2 \\
\vdots \\
-t + K
\end{array} \right) = \frac{1}{\lambda} \sum_{j=1}^{K} (j - t)T_{1,j}.
\] 

The left hand side of (27) can also be rewritten as

\[
T_{1}' \left[ (-t + 1) (1 + \lambda^{-1} \ -2 \ 1 \ 0 \ 0 \ \cdots \ 0) \\
+ (-t + 2) (-2 \ \lambda^{-1} + 5 \ -4 \ 1 \ 0 \ \cdots \ 0) \\
+ (-t + 3) (1 \ -4 \ \lambda^{-1} + 6 \ -4 \ 1 \ \cdots \ 0) + \ldots \right]',
\]

\[
= \left( (T_{1})_{1}, \ldots, (T_{1})_{K} \right) \left[ (-t + 1) (1 + \lambda^{-1} \ -2 \ 1 \ 0 \ 0 \ \cdots \ 0) \\
+ (-t + 2) (-2 \ \lambda^{-1} + 5 \ -4 \ 1 \ 0 \ \cdots \ 0) \right]',
\]

\[
= -(t - 1)u_{1,t} - (t - 2)u_{2,t}.
\]

From the proposition above, we see that the equivalent kernel is proportional to a linear combination of $T_{1}$, $T_{2}$ and $T_{t}$ and is of the second order. Denote the boundary kernel by $T'(x)$.

By using Proposition 2 to calculate the bias and variance, we obtain Theorem 3. The derivation of $\lambda$ is exactly the same as Section 2.3.
3. LINEAR SPLINES

3.1. Overview

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

\[
B_k^{[1]}(x) = \begin{cases} 
0, & x < \kappa_{k-1} \\
K(x - \kappa_{k-1}), & \kappa_{k-1} \leq x \leq \kappa_k \\
K(\kappa_{k+1} - x), & \kappa_k \leq x \leq \kappa_{k+1} \\
0, & x > \kappa_{k+1} 
\end{cases}
\]  

Thus, \( B_k^{[1]}(x) \) increases linearly from 0 to 1 as \( x \) increases from \( \kappa_{k-1} \) to \( \kappa_k \) and then decreases linearly to 0 as \( x \) increases from \( \kappa_k \) to \( \kappa_{k+1} \). (The actual values of the knots \( \kappa_{-1} \) and \( \kappa_{K+1} \) 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 \( 2/3 \) \( K^{-1} \) for \( k = 1, \ldots, K - 1 \) and equal \( 1/3 \) \( K^{-1} \) for \( k = 0 \) or \( K \). Also, \( \int_0^1 B_k^{[1]}(x)B_{k+1}^{[1]}(x) dx = 1/6 K^{-1} \) for \( k = 0, \ldots, K - 1 \). Therefore, \( X'X \approx M \Omega \) where

\[
\Omega = \begin{pmatrix}
1/3 & 1/6 & 0 & 0 & \cdots & 0 & 0 \\
1/6 & 2/3 & 1/6 & 0 & \cdots & 0 & 0 \\
0 & 1/6 & 2/3 & 1/6 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & 0 & \cdots & 2/3 & 1/6 \\
0 & 0 & 0 & 0 & \cdots & 1/6 & 1/3
\end{pmatrix}
\]  

Equation (2) is solved by

\[
\hat{b} = \{ M \Omega + \lambda (D^m)'D^m \}^{-1} X'Y = \{ \Omega + \lambda (D^m)'D^m \}^{-1} (X'Y/M),
\]

Equation (30) has a banded matrix, \( \{ \Omega + \lambda (D^m)'D^m \} \), with the same number of non-zero diagonals as matrix, \( \{ I_k + \lambda (D^m)'D^m \} \), in equation (3).

Also, in (30), \( (X'Y/M) \) can again to regarded as a vector of bin averages of the \( Y_t \) using linear binning (Hall and Wand, 1996). To appreciate this, let the \( k \)th bin be \([\kappa_{k-1}, \kappa_{k+1}]\), \( k = 0, \ldots, K \). Thus, if \( \kappa_{k-1} \leq x_t \leq \kappa_k \), then \( x_t \) is in bins \( k - 1 \) and \( k \). A fraction \( K(x - \kappa_{k-1}) \) of \( Y_t \) is placed in the \( k \)th bin and the remaining fraction \( K(\kappa_k - x) \) 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 non-boundary region, the penalised 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 due to the non-zero off diagonal terms in $\Omega$.

**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 $K \sim Cn^\gamma$ with $C > 0$ and $\gamma > 1/5$, and that there exists a constant $h > 0$ such that the penalty $\lambda \sim C^2h^2n^{2\gamma-2/5}$.

Let $\hat{f}_n$ be the first order penalised 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

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

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

**Proof of Theorem 4:** Let $\rho$ be the root of the equation $(6^{-1} - \lambda)x^2 + (2/3 + \lambda)x + (6^{-1} - \lambda) = 0$, and $T_t = (\rho^{t-1}, \rho^{t-2}, \ldots, \rho, 1, \rho, \rho^2, \ldots, \rho^{K-t})'$, with the “1” in the $t$th coordinate.

Similar as above, $T_t$ is orthogonal to all columns of $\Lambda$ except the first, last, and $t$th. Consider the non-boundary case where $t$ not too close to 1 or $K$. It also has $b_t \sim \{\sum_{j=1}^K \rho^{t-j}\tilde{y}_j\} \{\sum_{j=1}^K \rho^{t-j}\}$, and $b_t \sim \{\sum_{j=1}^K \rho^{t-j}\tilde{y}_j\} \{\sum_{j=1}^K \rho^{t-j}\}^{-1}$. According to Hall and Wand (1996), we have $\tilde{y}_t = f(x_t) + \epsilon' + O(K^{-2})$ with $\epsilon' \sim N(0, K/N\sigma^2)$. Compared to the zero-degree spline case, the bias due to binning is reduced from $K^{-1}$ to $K^{-2}$. Hence if $K \sim Cn^\gamma$ for some $C > 0$, then we only require $\gamma > 1/5$ instead of $2/5$.

Similarly to before, suppose $\rho = \exp\{- (Ch)^{-1}n^{1/5-\gamma}\}$, for some $h > 0$, where $\rho$ satisfies $(6^{-1} - \lambda)x^2 + (2/3 + 2\lambda)x + 6^{-1} - \lambda = 0$. Since $\tilde{x}_t - \tilde{x}_j = (t-j)/K$, $\rho^{t-j} = \exp\{-h^{-1}n^{1/5}(C^{-1}n^{-\gamma}|t-j|)\} \sim \exp\{-|\tilde{x}_t - \tilde{x}_j|/(hn^{-1/5})\}$, and hence we
can use the exponential kernel as the zero-degree case. Then the conclusion can be obtained by following the argument in Section 2.3.

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

Although their asymptotic behavior is similar to zero order splines, linear splines are different in two ways. There is no significant difference in asymptotic behavior between zero degree and linear splines in asymptotic behavior when $\gamma > 2/5$. However, if $K \sim Cn^\gamma$ with $1/5 < \gamma \leq 2/5$, 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 $\rho > 0$. This always holds for zero degree splines according to equation (7). But with linear splines, $\rho = \{6\lambda + 2 - (3 + 36\lambda)^{1/2}\}(6\lambda - 1)^{-1}$. We can rewrite this as $\rho = 1 + \{3 - (3 + 36\lambda)^{1/2}\}(6\lambda - 1)^{-1}$. Thus, $\rho > 0$ implies that $\lambda > 1/6$.

The assumptions of Theorem 4 imply that $\gamma \to \infty$, so $\lambda > 1/6$ 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 penalised 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 correspond-
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 $K \sim C n^\gamma$ with $C > 0$ and $\gamma > 2/5$, and that there exists a constant $h > 0$ such that $\lambda \sim C^2 h^2 n^{2\gamma-2/5}$. Let $\hat{f}_n$ be the penalised 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 $n^{2/5}\{\hat{f}_n(x) - f(x)\} \Rightarrow N\{B(x), V(x)\}$, where, with $g = G'$,

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

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

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

$$\hat{f}(\bar{x}_t) = \frac{(Khn^{-1/5})^{-1} \sum_{j=1}^{K} Y_i H\{(G_n(\bar{x}_t) - G_n(\bar{x}_j))/(hn^{-1/5})\}}{(Khn^{-1/5})^{-1} \sum_{j=1}^{K} H\{(G_n(\bar{x}_t) - G_n(\bar{x}_j))/(hn^{-1/5})\}}.$$  

The given choice of penalty yields to a bandwidth $hn^{-1/5}$ and we still have $(Khn^{-1/5})^{-1} \sum_{j=1}^{K} H\{(G_n(\bar{x}_t) - G_n(\bar{x}_j))/(hn^{-1/5})\} \to 1$ in probability. Hence we only need to consider the numerator of (32) which will be denoted $\hat{f}_{\text{num}}(\bar{x}_t)$.

Since $H$ is twice differentiable, Taylor expansion yields

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

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

$$+ (hn^{-1/5})^{-3} \sum_{j=1}^{K} \bar{y}_j \{G_n(\bar{x}_t) - G_n(\bar{x}_t) - G(\bar{x}_j) + G(\bar{x}_j)\}^2 H''(\triangle(\bar{x}))/2 := I_1 + I_2 + I_3,$$

where $\triangle(\bar{x})$ is between $\{G_n(\bar{x}_t) - G_n(\bar{x}_j)\}/(hn^{-1/5})$ and $\{G(\bar{x}_t) - G(\bar{x}_j)\}/(hn^{-1/5})$. 

21
First, \( n^{2/5}I_3 \to 0 \) in probability as \( n \to \infty \). Since \( H(x) \) has a much faster rate of decrease, we only need to consider \( x \) that satisfies \( |G(\bar{x}_i) - G(x)| \leq C_1(hn^{-1/5}) \) for some \( C_1 < \infty \). Stute (1984) proves that

\[
\sup_{x:|G(\bar{x}_i) - G(x)| \leq C_1(hn^{-1/5})} n(hn^{-1/5})^{-1}|G_n(\bar{x}_i) - G_n(x) - G(\bar{x}_i) + G(x)|^2H''(\Delta(\bar{x}))
\]

is stochastically bounded. Hence \( n^{2/5}I_3 = O_p(n^{-1/5}) \) and the argument is valid.

Second, \( n^{2/5}I_2 \) is asymptotically equivalent to

\[
-n^{2/5}(hn^{-1/5})^{-1} f(\bar{x}_i) \int H \left\{ \frac{G(\bar{x}_i) - G(x)}{hn^{-1/5}} \right\} \{G_n(dx) - G(dx)\}. \tag{33}
\]

Denote \( Z_n^1 = K^{-1}h^{-2}(n^{-1/5})^{-3/2} \sum_{i=1}^{K} (\bar{y}_i - f(\bar{x}_i))\{\tau_n(\bar{x}_i) - \tau_n(x)\}H'[hn^{-1/5}]\{G(\bar{x}_i) - G(\bar{x}_j)\} \), where \( \tau_n(x) = n^{1/2}[G_n(x) - G(x)] \). \( E\{Z_n^1\} \to 0 \). Since \( Z_n^1 \to 0 \) in probability, \( n^{2/5}I_2 \) is asymptotically equivalent to

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

Further denote \( Z_n^2 \) as

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

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

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

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

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

Hence (33) is valid.

Third, let \( I_4 \) denote \( n^{2/5}\{I_1 - f(\bar{x}_i) + I_2\} \). \( I_4 \) is a standardized sum of i.i.d random variables with Since \( E(Y^{2+\delta}) < \infty \), the Lindeberg condition is satisfied. Therefore, \( n^{2/5}\{\hat{f}_{num}(\bar{x}_i) - E\hat{f}_{num}(\bar{x}_i)\} \Rightarrow N\{0, \mathcal{V}(\bar{x}_i)\} \). For the bias, we have

\[
E\hat{f}_{num}(\bar{x}_i) - f(\bar{x}_i) = hn^{-1/5} \int \{f(x) - f(\bar{x}_i)\}H' \left\{ \frac{G(x) - G(\bar{x}_i)}{hn^{-1/5}} \right\} G(dx)
\]

\[= (hn^{-1/5})^2 (f \circ G^{-1})^{(2)}(G(\bar{x}_i)) \int u^2 H(u)du/2 = n^{-2/5}\mathcal{B}(\bar{x}_i).
\]
Hence, we can conclude that \( n^{2/3} \{ \hat{f}(\bar{x}_t) - f(\bar{x}_t) \} \Rightarrow N(\mathcal{B}(\bar{x}_t), \mathcal{V}(\bar{x}_t)) \). \qed

Thus, the bias of the penalised spline differs in several ways from that of the NW 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 (31) appears in the bias of the NW estimator, though with a plus sign. The term \( g^2(x) \) in the denominator of (31) 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 penalised splines with \( p = 0, m = 1 \), and knots at sample quantiles are not design-adaptive. An open question is the behavior of penalised 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 \( p = 0 \) (piecewise constant splines). The effective kernel will depend on the roots of modulus less than 1 of the polynomial

\[
P(\rho) = (1 - \rho)^2m(-1)^m + C\rho^m,
\]

where \( C > 0 \) and \( C \to 0 \) as \( K \to \infty \). 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. Because \( C \to 0 \) as \( n \to \infty \) and \( K \to \infty \), all roots of \( P \) converge to 1. This ensures that at each \( x \), the effective bandwidth is of the optimal order and \( \hat{f}(x) \) 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 one. 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

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

where $h$ is the “effective bandwidth.” If, in (34), $c_{\ell} = 0$ 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 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 the pattern continues for all $m$. The heuristic works for $p = 0$, but we believe it can, with some care, be extended to all $p$. Let $Z$ be a $K$ dimensional vector such that $Z_{t} = Q(t)$, $t = 1, \ldots, K$, 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 $\sum_{\ell=1}^{m-1} f^{(\ell)}(x_{t}) (x_{t} - x)^{\ell}/(\ell!)$ from $y_{t}$ for all $t$, then the value of $\hat{f}(x)$ is unchanged because $p = 0$. (If $p > 0$ the estimator will change because $(X^{[p]})^{\prime}X^{[p]}$ in $\Omega^{[p,m]}_{K,\lambda}$ 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 $\hat{f}(x)$ is unchanged by the modification, the bias must have been of order $m$ even without this modification. The penalised spline behave 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, penalised splines are of order $2m$ rather than $m$.

5.2. Comparison’s 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 (23) which we have found for piecewise constant penalised splines.

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

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

Moreover, we have found that the behavior 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 (18).

Agarwal and 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 the opposite of ours. Thus, it is not surprising that the results they obtain differ substantially from ours. In particular, Agarwal and Studden’s optimal estimator use less knots than ours do. From their equation (3.12), their optimal rate for $K$ is $K \sim n^{-1/(2p+3)}$. (Note: their $d$ is our $p + 1$.) Thus, for piecewise constant splines, their optimal rate is $K \sim n^{-1/3}$ while ours is $K \sim n^{-\gamma}$ for any $\gamma > 2/5$. For linear splines, their optimal rate is $K \sim n^{-1/5}$ while ours is $K \sim n^{-\gamma}$ for any $\gamma > 1/5$. 

25
An asymptotic theory intermediate between ours and that in Agarwal and Studden would select $K$ so that modeling and shrinkage biases are of the same order. For the case $p = 0$ and $d = 1$, this would require $K \sim C n^{\gamma}$ 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 penalised splines with local polynomial estimators. Local 0-degree polynomials are Nadaraya-Watson estimators. Therefore, penalised splines with a penalty of order 1 coincidence with local polynomials with degree 0 and double exponential kernels.

Penalised splines with $m > 1$ have different bias-order properties than local polynomial estimators. As shown in Ruppert and Wand (1994), local polynomials 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 penalised spline differ by $m$, at least if the heuristics in Section 5.1 are correct.

5.3. Choice of basis

We have worked with the B-spline bases advocated by Eilers and Marx (1996). However, other bases are often used for penalised splines, e.g., the truncated polynomials are used extensively in Ruppert, Wand, and Carroll (2003). Our results apply, of course, to an estimator defined with another bases provided that this estimator is identical to one of the P-spline (penalised B-spline) estimators studied here. This is often the case. As discussed in Section 3.7.1, a penalised spline in one basis will be algebraically identical to a penalised spline in a second basis, if the two basis span
the same vector space of functions and if they use identical penalties. For example, suppose we use the basis functions \( I(x > \kappa_k) \), that is, step functions that jump from 0 to 1 at the knots. Then the spline model is \( \sum_{k=1}^{K} a_k I(x > \kappa_k) \). Suppose we use the penalty

\[
\lambda \sum_{k=1}^{K} a_k^2,
\]

(35)

that is, the sum of squared jumps of the spline at the knots is penalised. Then this estimator is the same as the P-spline with \( p = 0 \) and \( m = 1 \). Similarly, the truncated line model \( \sum_{k=1}^{K} a_k (x - \kappa_k)_+ \) with penalty (35) 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. Penalising derivatives

Smoothing splines put a penalty on the integral of the squared \( m \)th derivative of the regression function, with \( m = 2 \) being the most common choice. Such penalties can be used on a penalised spline, if \( p \geq m \), by replacing the penalty in (1) by \( \lambda^* \int_0^1 \{ \sum_{j=1}^{K+p} b_j (B_j^{[p]}(x))^{(m)}(x) \}^2 dx \), where \( (B_j^{[p]}(x))^{(m)}(x) \) is the \( m \)th derivative of \( B_j^{[p]}(x) \). If one changes to the derivative penalty, then the only change in \( \hat{b} \) is that the matrix \( (D^m)'D^m \) in \( \Omega_{K,\lambda}^{[p,m]} \) is replaced by \( M \) where \( M_{ij} = \int_0^1 (B_i^{[p]}(x))^{(m)}(x) (B_j^{[p]}(x))^{(m)}(x) dx \). Since \( M \) is a banded matrix with modified corners having the same structure as \( (D^m)'D^m \), a penalised spline with penalty on the \( m \)th derivative has the same asymptotic behavior as penalised splines with an \( m \)th order difference penalties. In fact, for some choices of \( p \) and \( m \), e.g., \( m = p = 1 \), \( M \) 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 NSF Grant DMS 04-0434390 and
NIH Grant CA57030.

REFERENCES


Fig. 1: The finite-sample kernel (solid) and double-exponential kernel (dashed) for $m = 1$ and $p = 0$ and four sample sizes. In each case $K$ and $\lambda$ are functions of $n$ suggested by the asymptotics, specifically, $K = Cn^\gamma$, rounded to the nearest integer, and $\lambda = (Khn^{-1/5})^2$ with $\gamma$ and $h$ given in the text. Here $k$ is bin number and the kernels are for estimation at the bin $K/2$, rounded to the nearest integer.
Fig. 2: The finite-sample kernel (solid) and asymptotic boundary kernel (dashed) given by (21) for $m = 1$ and $p = 0$ and four sample sizes. In each case $K$ and $\lambda$ are functions of $n$ suggested by the asymptotics, specifically, $K = Cn^\gamma$, rounded to the nearest integer, and $\lambda = (Kh^{-1/5})^2$ with $C = 2$, $\gamma = 0.45$ and $h = 0.6$. Here $k$ is bin number and the kernels are for estimation at the bin 3. One can see that, besides the truncation at the boundary, the kernels are asymmetric.
Fig. 3: The finite-sample kernel (solid) and asymptotic kernel (dashed) for $m = 2$ and $p = 0$ and four sample sizes. In each case $K$ and $\lambda$ are functions of $n$ suggested by the asymptotics, specifically, $K = Cn^\gamma$, rounded to the nearest integer, and $\lambda = (Kh^{-1/5})^2$ with $\gamma$ and $h$ given in the text. Here $k$ is bin number and the kernels are for estimation at the bin $K/2$, rounded to the nearest integer.