

Locally Adaptive Smoothing Splines

Joan G. Staniswalis²

Brian S. Yandell³

Virginia Commonwealth U.

U. of Wisconsin-Madison

November 30, 1989

¹Keywords: nonparametric estimators, kernel estimators

²Supported by NSF grant DMS 8717560 while visiting the University of Wisconsin-Madison

³Supported by NSF grant DMS 8704341 and USDA-CSRS grant 511-100

Abstract

Locally adaptive smoothing splines combine features of variable kernels and smoothing splines to allow for local adaptive fitting and for a minimization of integrated mean squared error. Basically, one first adaptively fits a function with a local bandwidth kernel estimator, followed by a global fit to the presmoothed data using a penalized likelihood. We present some properties of the estimator and demonstrate its practical use through simulations and data analysis.

1 Introduction

Suppose one obtains observations Y_0, \dots, Y_{n-1} of the form

$$Y_i = f(x_i) + \epsilon_i ,$$

with $x_i = i/n, i = 0, \dots, n-1$. It is assumed that the errors ϵ_i contaminating the observations of $f(x)$ are independent random variates with mean 0 and variance σ^2 . Of interest is the nonparametric estimation of the function $f \in C^k[0,1], k \geq 4$, such that $f^{(k)} \in Lip_\gamma[0,1]$, by a function from the Sobolev space of order two,

$$W_2^2 = \{g|g, g' \in C[0,1] \text{ and } g^{(2)} \in L_2[0,1]\}.$$

We propose a nonparametric estimate of a curve which is a hybrid of kernel smoothing [1] and spline smoothing [2,3].

We combine the ideas of the computationally simple kernel estimator,

$$f_n(x, b) = \sum_{i=0}^{n-1} w\left(\frac{x-x_i}{b(x)}\right) Y_i / \sum_{i=0}^{n-1} w\left(\frac{x-x_i}{b(x)}\right), \quad 0 < b(x) < 1/2 , \quad (1)$$

with the cubic smoothing spline [4]. The kernel estimator minimizes the following weighted least squares criterion [5,6] at fixed x ,

$$(nb)^{-1} \sum_{i=0}^{n-1} w\left(\frac{x-x_i}{b}\right) [Y_i - g(x)]^2 \quad (2)$$

while the cubic smoothing spline minimizes the penalized least squares [7]

$$n^{-1} \sum_{i=0}^{n-1} (Y_i - g(x_i))^2 + \lambda \int_0^1 [g^{(2)}(x)]^2 dx \quad (3)$$

among all $g \in W_2^2$, with $\lambda \geq 0$ the smoothness constant. Our purpose is to show that the minimizer of

$$U(g) = n^{-1} \sum_{j=0}^{n-1} \left[[W_j^{-1} \sum_{i=0}^{n-1} w\left(\frac{x_j - x_i}{b(x_j)}\right) (Y_i - g(x_j))^2 \right] + \lambda \int_0^1 [g^{(2)}(x)]^2 dx , \quad (4)$$

where $W_j = \sum_{i=0}^{n-1} w\left(\frac{x_j - x_i}{b(x_j)}\right)$, has smaller IMSE than that of the spline estimator of (3). Thus a locally adaptive smoothing spline is proposed, which improves on the IMSE of the global smoothing spline, using ideas and methods from adaptive kernel estimators.

2 Kernels and Splines

We consider kernel estimators of the form (1) where $w(v)$ with k continuous derivatives is a symmetric kernel of order k with compact support on $[-1,1]$, satisfying the moment conditions

$$\int_{-1}^1 v^j w(v) dv = \begin{cases} 1 & j = 0 \\ 0 & j = 1, \dots, k-1 \\ W_k \neq 0 & j = k \end{cases} .$$

The bandwidth $b \in (0, 1/2]$ is a function of n and x although the notation does not reflect this. If $b \rightarrow 0$ and $nb \rightarrow \infty$ as $n \rightarrow \infty$ then $f_n(x, b)$ is a consistent estimator of $f(x)$ [8]. The bias of the kernel estimator is proportional to $b^k f^{(k)}(x)$ and the variance is inversely proportional to nb . Therefore, the bias of $f_n(x, b)$ can only be reduced at the cost of increasing the variance.

[9] proposed a method for estimating the optimal finite sample global bandwidth b^g which minimizes the integrated squared error $IMSE(b)$ of $f_n(x, b)$. The choice of the global bandwidth b^g is governed largely by the peaks and troughs of f . This global bandwidth results in a kernel estimate which tracks the observations Y_i in the flat regions of f , rather than averaging out the contaminating noise. [10] proposed a method for estimating the optimal finite sample local bandwidth $b^\ell(x)$ which minimizes the mean squared error $MSE(x; b)$ of $f_n(x, b)$. The local bandwidth $b^\ell(x)$ results in a kernel estimate with a small bandwidth near

peaks of f (reducing bias) and a larger bandwidth in the flat regions of f (reducing variance). Kernel estimators using data adaptive global and local bandwidth selection procedures have been shown to exhibit these properties as well [10]. However, the local bandwidth can be difficult to estimate in places where $f^{(k)}$ has high curvature. Consequently, the greatest incentive to using a kernel estimator with a local bandwidth over a global bandwidth selection procedure is in the reduction in variance realized in places where $f^{(k)}$ is very smooth.

The smoothing spline $\mu_{n,\lambda}$ which minimizes (3) is a piecewise cubic polynomial with knots at x_1, \dots, x_n and two continuous derivatives satisfying the boundary conditions

$$\mu_{n,\lambda}^{(i)}(0) = \mu_{n,\lambda}^{(i)}(1) = 0 \text{ for } i = 2, 3.$$

The smoothing parameter (λ) can be chosen from the data by either cross validation or maximum likelihood methods [11]. [12] and [13] showed that in the interior of $(0,1)$, the cubic smoothing spline is asymptotically equivalent to a kernel estimator with a global bandwidth of $h(\lambda) = \lambda^{1/4}$ and kernel of order 4 given by

$$S(u) = 2^{-1} \exp(-|u|/\sqrt{2}) \sin(|u|/\sqrt{2} + \pi/4).$$

[14] point out that it is desirable for $h(\lambda)$ to depend on the local curvature of f .

The penalized weighted least squares criterion (4) combines features of the weighted least squares criterion (2) and the penalized least squares criterion (3). When $b(x_j) \leq n^{-1}$, $j = 0, \dots, n-1$, $U(g)$ reduces to (3) whose minimizer is the cubic smoothing spline. On the other hand, if $\lambda = 0$ then $U(g)$ reduces to (2) whose minimizer is the kernel estimator. The following lemma shows that the minimizer of $U(g)$ is simply the minimizer of the penalized least squares criterion applied to the kernel-smoothed data. Its proof is a straightforward algebraic manipulation which is not given here.

Lemma 1 *The unique minimizer of $U(g)$ among $g \in W_2^2$ is also the unique minimizer of*

$$n^{-1} \sum_{i=0}^{n-1} [f(x_i) - \tilde{Y}_i]^2 + \lambda \int_0^1 [f^{(2)}(x)]^2 dx$$

among all $g \in W_2^2$, where $\tilde{Y}_i = W_i^{-1} \sum_{j=0}^{n-1} w \left(\frac{x_j - x_i}{b(x_i)} \right) Y_j$.

Let $\hat{f}_n(x; \lambda, b)$ denote the minimizer of $U(g)$, then \hat{f}_n is a smoothing spline fit to the presmoothed data (x_i, \tilde{Y}_i) , $i = 0, \dots, n - 1$.

3 Asymptotic Bias and Variance

Of interest are conditions under which the asymptotic IMSE is smaller for $\hat{f}_n(x; \lambda, b)$ than for $\mu_{n,\lambda}(x)$. In order to get some insight into how to select the bandwidth $b(x)$ to achieve this, asymptotic expressions are derived for

$$B^2(\lambda) = \int [E\hat{f}_n(x; \lambda, b) - f(x)]^2 dx$$

and

$$V(\lambda) = \int E[\hat{f}_n(x; \lambda, b) - E\hat{f}_n(x; \lambda, b)]^2 dx ,$$

where $b = b(x)$. Set

$$B_0^2(\lambda) = \int [E\mu_{n,\lambda}(x) - f(x)]^2 dx$$

and

$$V_0(\lambda) = \int E[\mu_{n,\lambda}(x) - E\mu_{n,\lambda}(x)]^2 dx .$$

It is of interest to select $b(x)$ in such a way that

$$V(\lambda) \leq V_0(\lambda) \text{ for all } \lambda > 0$$

without suffering a large increase in $B^2(\lambda)$ relative to $B_0^2(\lambda)$.

Lemma 1 of [15] showed that if $f(0) = f(1)$ and $f^{(1)}(0) = f^{(1)}(1)$, then

$$V_0(\lambda) \approx n^{-1} \sigma^2 \sum_{j=0}^{n-1} \lambda_j^2,$$

where

$$\lambda_j^2 = \begin{cases} \sum_{s=-\infty}^{\infty} (j + sn)^{-8} (\lambda' + r_j)^{-2} & ; j = 1, \dots, n-1 \\ 1 & ; j = 0 \end{cases}$$

with $\lambda' = \lambda(2\pi)^4$ and $r_j = \sum_{s=-\infty}^{\infty} (j + sn)^{-4}$. The following theorem provides an asymptotic expression for $V(\lambda)$.

Theorem 1 . *If $f(0) = f(1)$ and $f^{(1)}(0) = f^{(1)}(1)$, then $V(\lambda) = n^{-1} \sigma^2 \sum_{j=0}^{n-1} q_j \lambda_j^2$, where*

$$q_j = n^{-1} \sum_{i=0}^{n-1} \sum_{m=0}^{n-1} \sum_{l=0}^{n-1} W_m^{-1} W_i^{-1} w \left(\frac{x_m - x_l}{b(x_m)} \right) w \left(\frac{x_l - x_i}{b(x_i)} \right) \cos[2\pi j(m - i)/n],$$

for $j = 0, \dots, n-1$.

The q_j are converging to $\cos(0) = 1$ as $n \rightarrow \infty$. Furthermore, from the bias properties of kernel estimators [16], for large n , $0 < q_{n-1} \leq \dots \leq q_0 \leq 1$. Therefore, for large n the expected result that $V(\lambda) \leq V_0(\lambda)$ follows.

The following theorem allows us to compare $B^2(\lambda)$ with $B_0^2(\lambda)$. The proof is in the Appendix.

Theorem 2

$$B^2(\lambda) = B_0^2(\lambda) + \int [\nu_{n,\lambda}(x)]^2 dx + 2 \int E[\mu_{n,\lambda}(x) - f(x)] \nu_{n,\lambda}(x) dx$$

where $\nu_{n,\lambda}(x)$ is the smoothing spline with smoothing parameter λ which is fit to $(x_i, \text{Bias}(x_i))$,

$i = 0, \dots, n-1$. Here $\text{Bias}(x) = E[f_n(x, b)] - f(x)$.

The optimal rate of convergence for IMSE of a kernel estimator with a kernel of order $k = 4$ is $O(n^{-8/9})$. For general $f \in C^k[0,1]$, i.e., f that does not necessarily satisfy the boundary conditions

$$f^{(i)}(0) = f^{(i)}(1) = 0 \text{ for } i = 2, 3, \quad (5)$$

the optimal rate of convergence for the IMSE of $\mu_{n,\lambda}(x)$ is slower than $n^{-8/9}$ [17]. If the above boundary conditions are satisfied, then the IMSE of $\mu_{n,\lambda}(x)$ can attain the optimal rate $n^{-8/9}$.

If $b(x)$ is $O[n^{-1/(2k+1)}]$, the optimal rate for minimizing MSE of the kernel estimator, then $\nu_{n,\lambda}(x) = O[n^{-k/(2k+1)}]$. If f satisfies the boundary conditions (5), then $k > 4$ ensures that $B^2(\lambda) \approx B_0^2(\lambda)$. If f does not satisfy (5), then $k \geq 4$ is sufficient for $B^2(\lambda) \approx B_0^2(\lambda)$.

4 Simulations and Data Analysis

4.1 Simulations

The simulations were performed on the Statistics Research VAX at the University of Wisconsin-Madison. The purpose was to convincingly demonstrate that the locally adaptive smoothing spline has smaller IMSE than the global smoothing spline. A rescaled version of the function used by [3] was selected for the simulations

$$f(x) = 4.26[e^{-3.25x} - 4e^{-6.5x} + 3e^{-9.75x}].$$

Independent identically distributed $N(0, \sigma^2)$, $\sigma = .1f(0)$, contaminating errors for $n = 50$ were generated with the public domain random number generator RNOR. Noisy observations of f on $[-1,2]$ were used by the kernel smoother in order to avoid boundary modifications to

the kernel [18]. The spline fit to the presmoothed data (the LASS) and the global spline fit used only the region $[0,1]$.

One hundred independent realizations of size 50 of the locally adaptive smoothing spline (LASS) and the global spline smoother were generated. The LASS were created by generating raw data, presmoothing with the local bandwidth kernel smoother of [10], and then applying a cubic spline smoother. The LASS was applied with the kernels of [19].

As described earlier, the advantage to using a locally adaptive spline smoother over a global spline smoother is the reduction in variance where the underlying curve f is very smooth. The function used in the simulation is a paradigm of the undesirable 'wiggleness' which can result from locally undersmoothing the noisy data. Figure 1 is a realization of the two locally adaptive spline smoothers and the global spline smoother for this mixture of exponentials.

The mean squared error of the locally adaptive spline smoother and the global spline smoother were estimated from these one hundred realizations. Figure 2 presents the ratio of local to global MSE. Note the reduction in MSE achieved by the local spline smoother over the global spline smoother, particularly for the kernel of order 6. The average (over x) estimated MSE for the locally adaptive spline smoother are .0063 and .0050 for $k = 4$ and 6, respectively. The average estimated MSE for the global spline smoother is .0097. Smoothing the data with higher order kernels allows the locally adaptive smoothing spline to enjoy a large decrease in variance without a subsequent increase in bias.

4.2 Data Analysis

The voltage drop data in Ch. 3 (ex. 14) of [20] was analysed. Figure 3 is a plot of the locally adaptive spline smoother for $k = 4, 6$ and 8 and the global spline smoother superimposed on the data. The curve f was assumed to be periodic on $[0,20]$ in order to avoid boundary modifications to the kernel [18]. Again, it is evident that the higher order kernels relieve the bias problem of the locally adaptive spline smoother while allowing for a decrease in variance over the global spline smoother.

5 Conclusion

In practice, the LASS smoother is indistinguishable from the locally adaptive kernel estimator. This is due to the fact that the cross-validated λ which is estimated from the kernel-smoothed data is very close to zero. A natural problem to consider next is a "weighted" least squares cubic spline with $k \ll n$ knots which minimize equation (2) with respect to g . The TURBO smoother of [21] would probably benefit from this presmoothing of the data, allowing for a numerically stable algorithm.

6 Appendix

6.1 Proof of Theorem 1

Represent both the spline \hat{f}_n fit to the presmoothed data and the smooth function f in terms of a Fourier series expansion:

$$\hat{f}_n(x; \lambda, b) = \sum_{s=-\infty}^{\infty} c_s e^{2\pi i s x}$$

and

$$f(x) = \sum_{s=-\infty}^{\infty} a_s e^{2\pi i s x} .$$

For computational convenience, let $\hat{f}_n(x; \lambda, b)$ be the minimizer of

$$n^{-1} \sum_{i=0}^{n-1} (\tilde{Y}_i - f(x_i))^2 + \lambda' (2\pi)^{-4} \int_0^1 (f^{(2)}(x))^2 dx , \quad \lambda' = \lambda(2\pi)^4 .$$

Referring back to the results and methods of [15], it can be shown that

$$c_{sn} = \begin{cases} 0 & \text{when } s \neq 0 \\ n^{-1/2} \hat{Y}_0 & \text{when } s = 0 \end{cases}$$

and

$$c_{j+sn} = (j + sn)^{-4} (\lambda' + r_j)^{-1} n^{-1/2} \hat{Y}_j , \quad j = 1, \dots, n-1 .$$

Here $\hat{Y}_0, \dots, \hat{Y}_{n-1}$ are the discrete Fourier coefficients of $\tilde{Y}_0, \dots, \tilde{Y}_{n-1}$; i.e.,

$$\begin{aligned} \hat{Y}_j &= n^{-1/2} \sum_{t=0}^{n-1} \tilde{Y}_t \exp(-2\pi i j t / n) \\ &= U_j^*(WY) , \end{aligned}$$

where $U_j = \{\exp(2\pi ijt/n)\}_{t=0}^{n-1}$, $W_{ij} = W_i^{-1}w\left(\frac{x_i-x_j}{b(x_i)}\right)$, $i, j = 0, \dots, n-1$, and $Y = (Y_0, \dots, Y_{n-1})^T$. It follows that $\text{var}(\hat{Y}_j) = \sigma^2 U_j^* W W^T U_j$ and $V = \sum_{s=-\infty}^{\infty} \text{var}(c_s)$ by Parseval's Theorem. Therefore

$$\begin{aligned} V &= \sigma^2 n^{-1} \text{trace}(\Lambda U^* W W^T U \Lambda) \\ &= \sigma^2 n^{-1} \text{trace}(W W^T U \Lambda^2 U^*) \end{aligned}$$

where $\Lambda = \text{diag}(\lambda_0, \dots, \lambda_{n-1})$ and $U = [U_0 \dots U_{n-1}]$.

The elements of $A = U \Lambda^2 U^*$ are of the form

$$A_{jl} = n^{-1} \sum_{t=0}^{n-1} \lambda_t^2 \exp(2\pi it(j-l)/n) .$$

Thus

$$\begin{aligned} n\sigma^{-2}V &= \sum_{l=0}^{n-1} (W W^T A)_{ll} = \sum_{l=0}^{n-1} \sum_{j=0}^{n-1} (W W^T)_{lj} A_{jl} \\ &= \sum_{l=0}^{n-1} \sum_{j=0}^{n-1} \left[\sum_{q=0}^{n-1} (W_l W_j)^{-1} w\left(\frac{x_l-x_q}{b_l}\right) w\left(\frac{x_q-x_j}{b_j}\right) \right] \left(n^{-1} \sum_{t=0}^{n-1} \lambda_t^2 \exp(2\pi it(j-l)/n) \right) \\ &\equiv \sum_{t=0}^{n-1} \lambda_t^2 q_t , \end{aligned}$$

with q_t defined accordingly. Note that q_t is real valued since w is symmetric and q_t is symmetric in x_j, x_l .

6.2 Proof of Theorem 2

Using Parseval's theorem, we can express

$$B^2(\lambda) = \sum_{s=-\infty}^{\infty} \text{Bias}^2(c_s) .$$

As in Lemma 1 of [15],

$$B^2(\lambda) = \sum_{j=0}^{n-1} \sum_{s=-\infty}^{\infty} |\lambda_{js}(\tilde{a}_j + \tilde{d}_j) - a_{j+sn}|^2 ,$$

where

$$\tilde{d}_j = n^{-1/2} \sum_{l=0}^{n-1} \text{Bias}(x_l) \exp(-2\pi ijl/n) ,$$

$$\tilde{a}_j = \sum_{s=-\infty}^{\infty} a_{j+sn} , \quad j = 0, \dots, n-1 ,$$

$\lambda_{js} = (j + sn)^{-4}(\lambda' + r_j)^{-1}$, $j = 1, \dots, n-1$, $\lambda_{00} = 1$ and $\lambda_{0s} = 0$ for $s \neq 0$. Here $\text{Bias}(x)$ is the bias of the kernel estimator of $f(x)$ which uses the kernel w and the bandwidth $b(x)$.

From Parseval's theorem, we recognize that

$$B^2(\lambda) = \int \{[E\mu_{n,\lambda}(x) - f(x)] + \nu_{n,\lambda}(x)\}^2 dx$$

where $\nu_{n,\lambda}$ is the smoothing spline fit to $(x_i, \text{Bias}(x_i))$ and where $\mu_{n,\lambda}$ is the smoothing spline fit to (x_i, Y_i) , $i = 0, \dots, n-1$. Therefore,

$$B^2(\lambda) = B_0^2(\lambda) + \int [\nu_{n,\lambda}(x)]^2 dx + 2 \int [E\mu_{n,\lambda}(x) - f(x)] \nu_{n,\lambda}(x) dx .$$

7 References

- [1] M. B. Priestly & M. T. Chao, "Nonparametric function fitting," *Journal of the Royal Statistical Society, Series B* 34 (1972), 385–392.
- [2] C. H. Reinsch, "Smoothing by spline functions," *Numerische Mathematik* 10 (1967), 177–183.
- [3] Grace Wahba & S. Wold, "A completely automatic French curve: fitting spline functions by cross-validation," *Communications in Statistics* (1975).
- [4] Grace Wahba, "Smoothing noisy data by spline functions," *Numerische Mathematik* 24 (1975), 383–393.

- [5] Joan G. Staniswalis, "The kernel estimate of a regression function in likelihood-based models," *Journal of the American Statistical Association* 84 (1989), 276–283.
- [6] W. S. Cleveland, "Robust locally weighted regression and smoothing scatterplots," *Journal of the American Statistical Association* 74 (1979), 829–836.
- [7] I. J. Good & R. A. Gaskins, "Non-parametric roughness penalties for probability densities," *Biometrika* 58 (1971), 255–277.
- [8] T. Gasser & H. G. Muller, "Kernel estimation of regression functions," in *Smoothing Techniques for Curve Estimation*, T. Gasser & M. Rosenblatt, eds., Lect. Notes in Math. #757, Springer-Verlag, New York–Heidelberg–Berlin, 1979, 23–68.
- [9] John A. Rice, "Bandwidth Choice for nonparametric regression," *Annals of Statistics* 12 (1984), 1215–1230.
- [10] Joan G. Staniswalis, "Local bandwidth selection for kernel estimates," *Journal of the American Statistical Association* 84 (1989), 284–288.
- [11] Grace Wahba, "A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem," *Annals of Statistics* 13 (1985), 1378–1402.
- [12] R. Cogburn & H.T. Davies, "Periodic splines and spectral density estimation," *Annals of Statistics* 2 (1974), 1108–1126.
- [13] B. W. Silverman, "Spline smoothing: the equivalent variable kernel method," *Annals of Statistics* 12 (1984), 898–916.
- [14] M. Tanner & W-H. Wong, "Contribution to the discussion of the paper by Silverman," *Journal of the Royal Statistical Society, Series B* 47 (1985), 44–45.

- [15] John A. Rice & M. Rosenblatt, "Integrated mean squared error of a smoothing spline," *Journal of Approximation Theory* 33 (1981), 353–369.
- [16] H. G. Muller, "Kernel estimators of zeros and of location and size of extrema of regression functions," *Scandinavian Journal of Statistics* 12 (1985), 221–232.
- [17] Utreras, "Convergence rates for multivariate smoothing spline functions," *Journal of Approximation Theory* 39 (1987).
- [18] John A. Rice, "Boundary modification for kernel regression," *Communications in Statistics, Series A* 13 (1984), 893–900.
- [19] H. G. Muller, "Smooth optimum kernel estimators of densities, regression curves and modes," *Annals of Statistics* 12 (1984), 766–774.
- [20] R. L. Eubank, in *Spline Smoothing and Nonparametric Regression*, Marcell Dekker, Inc., New York, 1988.

- [21] Jerome Friedman & Bernard W. Silverman, "Flexible parsimonious smoothing and additive modeling," *Technometrics* 31 (1989), 3–39.

8 Figure Captions

1. Realizations of the LASS and Global Spline Smoothers for Data Simulated from a Mixture of Exponentials. Solid line is true f ; dotted line is LASS with $k=4$; short dashed line is LASS with $k=6$; long dashed line is global spline smoother.

2. Estimated MSE of LASS Relative to Estimated MSE of Global Spline Smoother. Solid line is $k=4$; dotted line is $k=6$.

3. Raw Data, LASS, and the Global Spline Smoother for Voltage Data. Solid line is LASS with $k=4$; dotted line is LASS with $k=6$; short dashed line is LASS with $k=8$; long dashed line is global spline smoother.

Figure 1

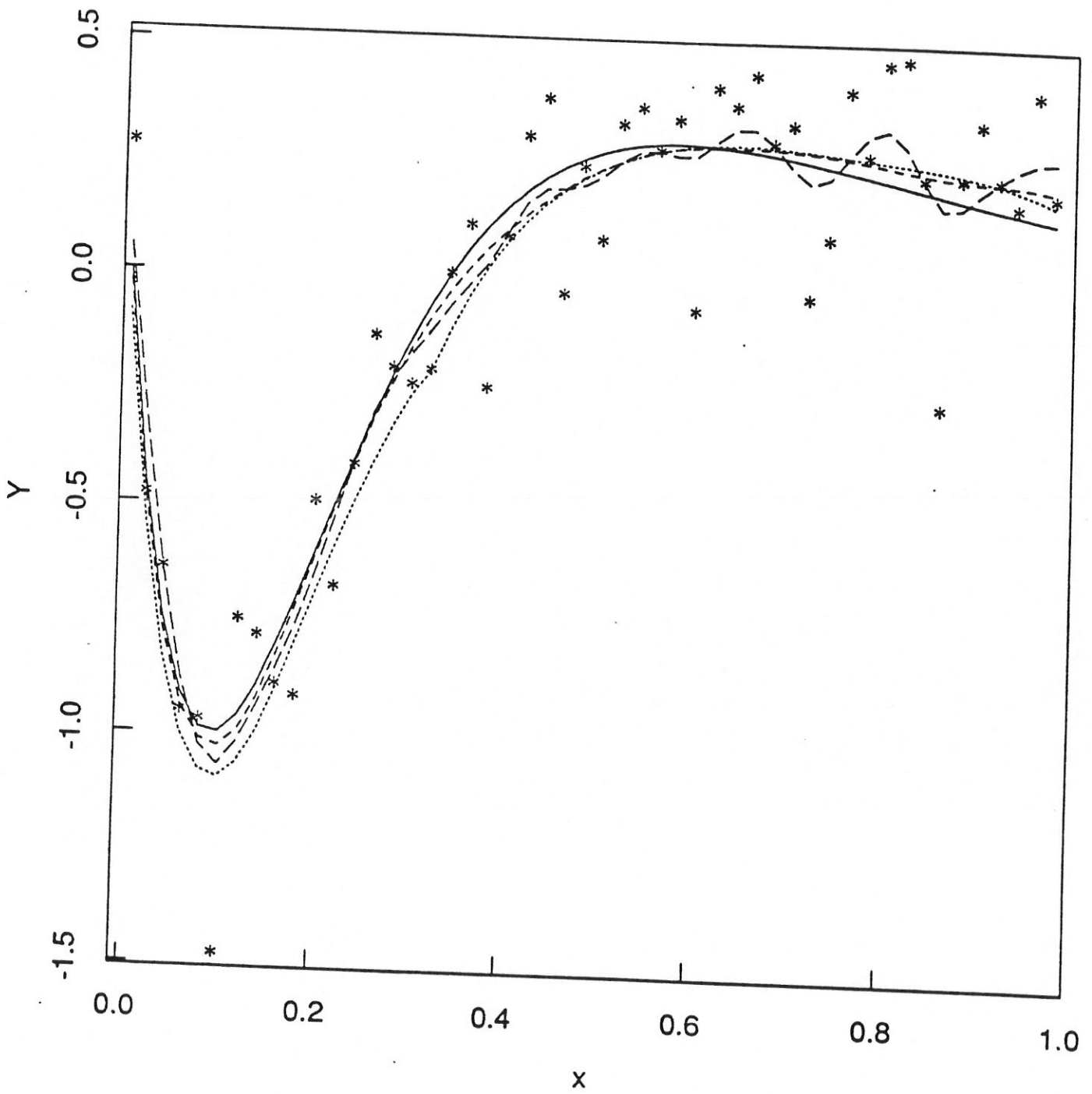


Figure 2.

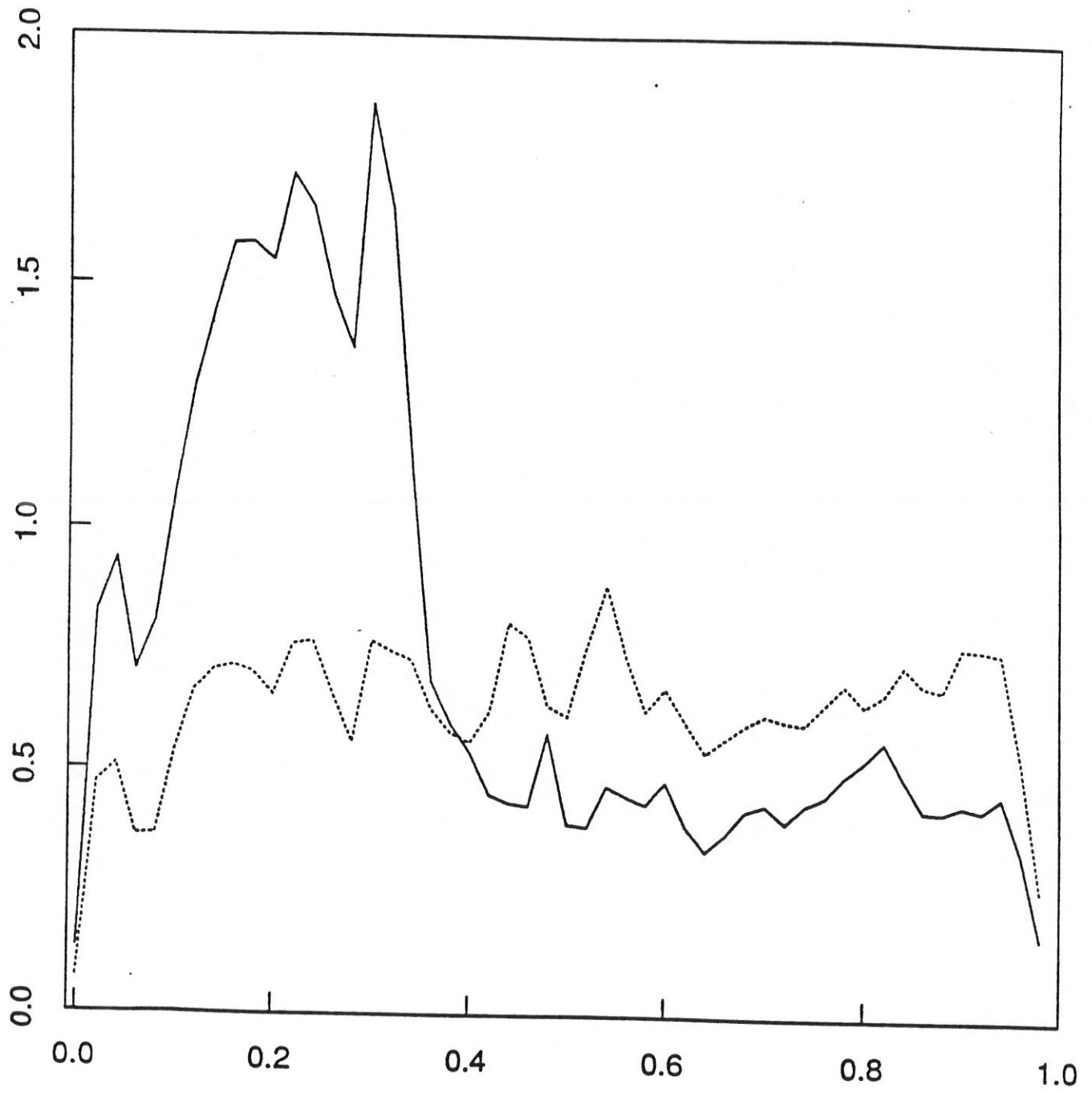
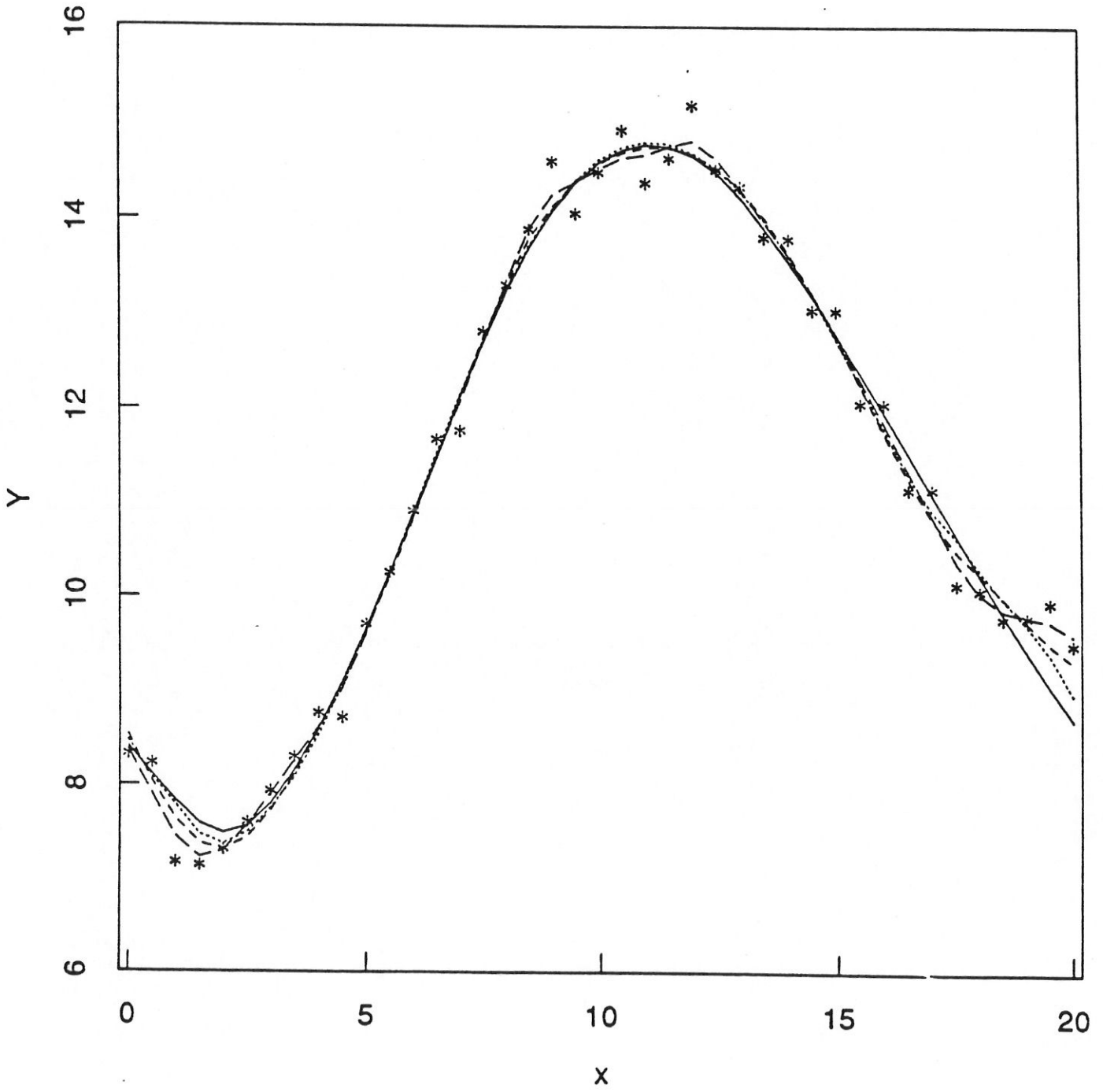


Figure 3



Reply to:

EDITOR-IN-CHIEF:
Stanley P. Azen
Dept. of Family and
Preventive Medicine
USC School of Medicine
2025 Zonal Ave., PMB B 101
Los Angeles, CA 90033

 **COMPUTATIONAL
STATISTICS &
DATA ANALYSIS**

April 10, 1990

Joan G. Staniswalis, Ph.D.
Department of Biostatistics
Medical College of Virginia
Virginia Commonwealth University
P.O. Box 32
Richmond, Virginia 23298-0032

Dear Dr. Staniswalis: RE: CSDA #89-465

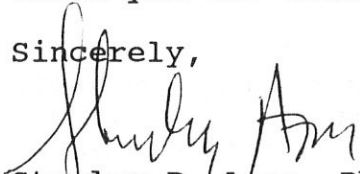
The reviewing process for your manuscript is now complete.

The Associate Editor received two detailed reviews which I am enclosing for your review. Generally speaking, the notion of adaptive smoothing splines is appealing, but it was not obvious that the proposed method is really an improvement on either spline smoothing or adaptive kernel estimation. One referee was not sure of the "bottom line" recommendation. A second referee was concerned with the overly restrictive nature of your design (e.g., equally spaced data), and felt that the general case could be easily derived. Additional detailed comments are enclosed.

On the basis of the reviews, and the recommendation of the Associate Editor, we are unable to publish your manuscript in its current form. We do, however, encourage you to resubmit a greatly revised manuscript, taking into account the referees' comments. If you choose to revise, please indicate in a covering letter the nature of your changes. This will expedite the re-reviewing process.

Thank you for considering CSDA.

Sincerely,


Stanley P. Azen, Ph.D.
Editor-in-Chief

Comments for the authors of
“Locally adaptive smoothing splines”

This is an interesting manuscript that seems to leave some questions unanswered. The notion of adaptive smoothing splines is appealing. Somehow, though, I’m not sure that the proposed method is really an improvement on either spline smoothing or adaptive kernel estimation. After reading the entire article, I get the impression that adaptive kernel estimation by itself may be the best method. (Perhaps this is what the authors have in mind!) I think the paper would be strengthened by making a convincing case for LASS over adaptive kernel estimation alone. I also had difficulty understanding the application of the Theorems. My apologies if I missed something important and obvious. I have outlined several questions and suggestions below, and I look forward to further comment and development.

1. In the argument following Theorem 2, I had difficulty with the statement “... for large $n, 0 < q_{n-1} \leq \dots \leq q_0 \leq 1$.” It’s not clear to me how this follows from bias calculations. This point needs clarification. One possible way to demonstrate the desired result would be to note that $q_j = U_j^* W W^T U_j$, so the q_j are bounded above by the eigenvalues of $W W^T$. Can you prove that these eigenvalues are all bounded above by one? Suppose you let S denote the smoothing spline matrix satisfying

$$(\mu_{n,\lambda}(x_1), \dots, \mu_{n,\lambda}(x_1))^T = S y.$$

Does the argument now depend on the specific form of the U_j or would it apply to any smoother matrix S ? Can one instead extend to arbitrary S via the principal value decomposition $S = P \Lambda Q^T$ where P and Q are both orthogonal?

2. This is probably my slowness, but I didn’t understand the application of Theorem 2 on bias either. In particular, what does the last paragraph of section 3 mean? Does “ $B^2(\lambda) \approx B_0^2(\lambda)$ ” mean “ $B^2(\lambda)/B_0^2(\lambda) \rightarrow 1$?” If so, wouldn’t I want $B^2(\lambda)/B_0^2(\lambda) < 1$? Or are you arguing that λ could be decreased some without increasing the variance too much. Although I understand the difficulties, some of the confusion is perhaps caused

by the ambiguous use of the $O(n^{-k/(2k+1)})$ notation. Do you sometimes mean exact order? Presumably the spline smoother in the two-stage estimator should have a different smoothing parameter than when used by itself. How is this fact used in the estimates of mse? Any help you could give in showing how the bias would actually be decreased would be helpful.

3. As I mentioned above, I have difficulty understanding the benefit of the second smoothing. Intuition suggests that further smoothing will reduce the variance, so I believe the application of Theorem 1 even if the proof is not entirely clear. However, smoothing induces bias by filling in valleys and cutting off peaks. My intuition suggests that a second global smoother will make the bias worse. If the presmoothing underfits, there could be some benefit to additional smoothing. But in this case would shouldn't you simply use a better presmoothing? Why perform the smoothing spline component at all? The comments in the conclusion appear to hint at this as I read "... the LASS smoother is indistinguishable from the locally adaptive kernel estimator." It seems clear that there will be hardly any roughness left after the presmoothing. Perhaps you can clarify the matter.

Minor points

1. You might mention that bias and variance calculations assume fixed b and λ throughout.
2. (p. 5 ↓2) What is the role of periodicity in the variance of the smoothing spline? I thought it only affected the bias.
3. Do you assume equally spaced points throughout?
4. (p. 10 ↓1) Should U_j be multiplied by $n^{-1/2}$?
5. This is only a matter of taste, but I think Theorem 2 is clearer with matrix notation (the L_2 norm can be replaced by the Euclidean norm in R_n):

$$B^2(\lambda) = \|SWf - f\|^2$$

$$\begin{aligned} &= \|Sf - f + SWf - Sf\|^2 \\ &= \|Sf - f\|^2 + \|S(Wf - f)\|^2 \\ &\quad + 2(Sf - f)^t(S(Wf - f)). \end{aligned}$$

This is also more general.

6. (Figure 1) It would be interesting to see the presmoother plotted along with the LASS and global smoothing spline.

Referee's report on "Locally Adaptive Smoothing Splines" by Staniswalis and Yandell

The authors propose a hybrid of kernel and spline smoothing to create a computationally simple way of computing smoothing splines with locally adaptive levels of smoothing. The proposed technique is quite interesting and the paper is well written overall.

There are two basic problems with this paper. First, it is written for a journal concerned with computation but the computational side of the methodology is almost totally ignored. Secondly, the asymptotics deal only with the case of periodic smoothing splines and equally spaced data. This is overly restrictive and the same essential results can be derived in the general case. Some ways to correct these two problems are discussed below.

Computational aspects.

At the very least all the essential details of the computation of the estimator should be discussed. This includes the locally adaptive bandwidth selection method of Staniswalis and the method of selecting λ for the presmoothed data. I feel that the computations should be done using ordinary smoothing splines and boundary kernels as well. The authors should discuss the creation of code for their estimator. Presumably this can be done by combining code from Staniswalis' work with widely available code for spline smoothing.

Asymptotics.

The essential conclusions you reach in your asymptotics can be reached for general smoothing splines. In fact, the case of general designs and higher than second derivative penalties can also be handled. One can show

$$V(\lambda) \leq \left(\int_{-1}^1 |\omega(u)|^2 du \right) V_0(\lambda) (1 + o(1)) \quad (1)$$

and, if $k > 2$,

$$B^2(\lambda) \leq B_0(\lambda) (1 + o(1)), \quad (2)$$

under your same conditions, with all this pertaining to the general non-periodic smoothing spline. The details are actually simpler. I therefore suggest that the asymptotics for the general case be given and the space that is saved by doing this be used for elaboration on the computational aspects of the estimator.

Note that (1) is but one of several bounds that can be used here. It gives $V(\lambda) < V_0(\lambda)$ for n

sufficiently large and ω the Epanechnikov kernel. This is an improvement, in some sense, over the result presented in the present paper.

Some minor comments and corrections are given below.

pg. 1. i) Why not begin the paper with a non-technical introductory paragraph telling us the objective of the article?

ii) Why restrict attention to the equispaced design in the introduction? Even if you decide to use this in your asymptotics, there is no reason to restrict the applicability of your technique to this case.

iii) Why is b restricted to be $< 1/2$? (Note also that b is allowed to equal $1/2$ on pg. 2.)

iv) On line 5 \uparrow you need the word 'criterion'.

pg. 2. i) You show that the IMSE of your estimator is no larger, asymptotically, than that of the ordinary smoothing spline. You do not show that it is smaller.

ii) Strictly speaking the Rice reference [9] deals with a damped Fourier series estimator not a kernel estimator. There are many other references for global bandwidth selection methods for kernel estimators. Some of these should probably be mentioned here.

pg. 3. i) The second paragraph is a waste of space and should be eliminated.

ii) The proof should be given for the lemma.

pg. 4. 3 \downarrow $g \rightarrow f$

pg. 7. How was λ selected in the simulation and data analysis.

pg. 8. i) Unfortunately, the assumption of periodicity is not really valid for this data. One can show that there is a significant edge effect. This is all the more reason to use a regular smoothing spline and boundary kernels.

ii) There is no reference [21].

pg. 11. If I'm wrong on this please ignore this comment, but I have never seen a journal that does not require the references to be alphabetized.

Figure 2. i) The axes need to be labeled.

ii) Why not break this into two (or even three) plots of squared bias and variance (and maybe MSE). I

think this would make your point that the estimator decreases variance much better and show that the peak in the figure for $k=4$ is mostly due to bias.

Proof of (1) and (2). First we need some notation. Let H_λ be the matrix that transforms the data to fitted values for a smoothing spline and let W be the matrix that does this for your kernel estimator (this can include boundary corrections). If $f_{\lambda b}$ is the vector of fitted values for your estimator, then

$$n^{-1} \text{tr}(\text{Var}(f_{\lambda b})) = n^{-1} \sigma^2 \sum_{j=1}^n \lambda_j^2 t_j' W W' t_j \leq V_0(\lambda) \|W\|^2,$$

where the t_j are the vectors of the values for the Demmler-Reinsch basis functions, the λ_j are the eigenvalues of H_λ , $V_0(\lambda) = n^{-1} \sigma^2 \sum_{j=1}^n \lambda_j^2$ and $\|\cdot\|$ is the Euclidean matrix norm, i.e., $\|W\|^2 = \text{largest eigenvalue of } W W' = \lambda_{\max}(W W')$.

We require a bound for $\|W\|$. By the Gershgorin circle Theorem we know that

$$\lambda_{\max}(W W') \leq \max_i \sum_j \sum_r |\omega_i^{-1} \omega_r^{-1} \omega(\frac{x_i - x_r}{b(x_i)}) \omega(\frac{x_r - x_j}{b(x_r)})|.$$

By various quadrature arguments one can then show that this will be bounded asymptotically by

$$\int_{-1}^1 \int_{-1}^1 |\omega(y) \omega(u - y)| dy du \leq \int_{-1}^1 \omega(u)^2 du,$$

assuming boundary corrections have been made and that the data is equally spaced. Explicit evaluation of the first bound may give better results, although $\int \omega(u)^2 du$ is quite satisfactory for a kernel of order 2. This argument can be extended to unequally spaced data by assuming the x_i are the n -tiles of some positive, continuous density on $[0, 1]$.

Concerning the bias, we have

$$\begin{aligned} B^2(\lambda) &= n^{-1} \sum (f(x_i) - E f(x_i; \lambda, b))^2 = f'(I - H_\lambda W)'(I - H_\lambda W)f/n \\ &= B_0^2(\lambda) + 2f'(I - W)'H_\lambda(I - H_\lambda)f/n + f'(I - W)'H_\lambda^2(I - W)f/n \\ &\leq B_0^2(\lambda) + O(B_0(\lambda)B_\omega) + O(B_\omega^2), \end{aligned}$$

where $B_\omega^2 = f'(I - W)'(I - W)f/n$. Now B_ω^2 is the average squared bias of the kernel estimator and will be $O(n^{-2k/(2k+1)})$ so the bound (2) holds using the fact that B_0^2 will decay at the $n^{-4/5}$ rate if f is twice continuously differentiable but can attain the $n^{-8/9}$ rate only if f has four derivatives and satisfies the natural boundary conditions (cf. Speckman, 1981 and Rice and Rosenblatt, 1983, *Ann. Statist.*).