JOURNAL OF MULTIVARIATE ANALYSIS 19, 251–264 (1986)

Bandwidth Choice for Differentiation

JOHN A. RICE*

University of California, San Diego

Communicated by M. Rosenblatt

We propose a class of procedures for choosing the bandwidth, or smoothing parameter, for linear nonparametric estimates of the rth derivative of a smooth function observed with error on a discrete set of points. These procedures are based on minimizing a nearly unbiased estimate of the integrated mean square error. Theoretical justification is provided in the special case of a tapered Fourier series estimate.

1. INTRODUCTION

All methods of nonparametric regression and differentiation entail the choice of a smoothing or bandwidth parameter which controls the size of the region over which local averaging is done. In many applications, it is sensible to try several different choices: small bandwidths may preserve local features of the data that are obscured by larger bandwidths which, however, may be globally more effective. However, if only for preliminary analysis, it can be convenient to use methods which automatically determine the bandwidth from the data.

Of methods for bandwidth choice, cross-validation (e.g., Craven and Wahba [2]) has probably received the greatest attention. Other criteria, related to Akaike's information criterion [1], have also been applied to regression (Shibata [8]; Rice [6]). Bandwidth choice for regression is also considered by Speckman [9], Wong [10], Li [5], and Hardle and Marron [4]. There has, however, been very little work on data-driven bandwidth choice for differentation.

We will consider attempting to choose a smoothing parameter to minimize expected integrated squared error, or a discrete approximation

Received November 17, 1984; revised June 27, 1985.

AMS 1980 subject classification: primary 62G05.

Key words and phrases: bandwidth choice, nonparametric estimation, tapered Fourier series estimate.

* Research partially supported by NSF MCS-7901800.

0047-259X/86 $3.00

thereof. Letting $f_n^{(r)}(x;\lambda)$ be a nonparametric estimate of the rth derivative of f with smoothing parameter λ and sample size n, the risk function we will consider is

$$MISE_n(\lambda) = E\int [f^{(r)}(x) - f_n^{(r)}(x;\lambda)]^2\,dx. \tag{1.1}$$

For motivation, we consider kernel estimation. Suppose that

$$y_i = f(x_i) + \varepsilon_i, \qquad i = 0,\ldots, n \tag{1.2}$$

where $x_i = i/n$ and the ε_i are independent mean zero random variables with variance σ^2. Let $w(x)$ be a symmetric, sufficiently smooth, probability density, and λ a smoothing parameter. The estimate of the rth derivative of f is

$$f_n^{(r)}(x;\lambda) = \frac{\lambda}{n}\sum_{i=1}^{n}\frac{d^r}{dx^r}w[\lambda(x-x_i)]\,y_i \tag{1.3}$$

$$= \frac{\lambda^{r+1}}{n}\sum_{i=1}^{n} w^{(r)}[\lambda(x-x_i)]\,y_i. \tag{1.4}$$

In Section 2 we consider linear estimates of $f^{(r)}$, develop a nearly unbiased estimate of $\mathrm{MISE}_n(\lambda)$ and consider choosing λ to minimize this estimate. In Section 3 we introduce a tapered Fourier series estimate closely related to the kernel estimate above in the case when the x_i are equally spaced. We assume there that f is smoothly periodic and that w satisfies certain smoothness conditions. In Section 4 we provide theoretical justification for the procedure suggested in Section 2 for this estimate; we show that the minimizer chosen from the data, and restricted to an interval I_n, is a consistent estimate of the optimal bandwidth and deduce an asymptotic normal distribution for the estimated bandwidth. Some final remarks are contained in Section 5.

2. Risk Estimation

We will assume that the estimate of the rth derivative of f is linear in the observations (as is the case for the kernel estimate and smoothing spline estimates, for example) and that the domain of f is $[0, 1]$ for simplicity. We construct an estimate of $MISE_n(\lambda)$ in the following way:

- Let $\zeta_1,\ldots, \zeta_m$ be a set of points in $[0, 1]$; these points will depend upon n although the dependence is notationally suppressed. The ζ_i will be quadrature points in $[0, 1]$; there is some freedom in choosing these points, but in order to be concrete, we will assume that m is of order n and that the points are equally spaced.

• Denote the vector of observations y_i by Y_n. Let $A_n(\lambda)$ be the $m \times n$ matrix such that $A_n(\lambda)\, Y_n$ is the vector of estimates

$$[f_n^{(r)}(\zeta_1; \lambda),..., f_n^{(r)}(\zeta_m; \lambda)]^T. \tag{2.1}$$

• Let

$$f_n^{(r)} = [f^{(r)}(\zeta_1),..., f^{(r)}(\zeta_m)]^T. \tag{2.2}$$

• Let

$$f_n = [f(x_1),..., f(x_n)]^T. \tag{2.3}$$

• Let D_n be an $m \times n$ matrix such that $D_n f_n$ is a good discrete approximation to $f_n^{(r)}$; D_n is a differencing operator.

• Let

$$\begin{aligned} R_n(\lambda) &= E\frac{1}{m}\| f_n^{(r)} - A_n(\lambda)\, Y_n\|^2 \\ &= \frac{1}{m}\| f_n^{(r)} - A_n(\lambda) f_n\|^2 + \frac{\sigma^2}{m}\operatorname{tr}(A_n(\lambda)^T A_n(\lambda)). \end{aligned} \tag{2.4}$$

• Let

$$SSD_n(\lambda) = \frac{1}{m}\| D_n Y_n - A_n(\lambda)\, Y_n\|^2. \tag{2.5}$$

A simple calculation shows that

$$\begin{aligned} ESSD_n(\lambda) = \frac{1}{m}\| D_n f_n - A_n(\lambda) f_n\|^2 &+ \frac{\sigma^2}{m}\operatorname{tr}(D_n^T D_n) \\ &- 2\frac{\sigma^2}{m}\operatorname{tr}(D_n^T A_n(\lambda)) + \frac{\sigma^2}{m}\operatorname{tr}(A_n(\lambda)^T A_n(\lambda)). \end{aligned} \tag{2.6}$$

Comparing $ESSD_n(\lambda)$ to $R_n(\lambda)$, we are led to estimate the latter by

$$\hat{R}_n(\lambda) = SSD_n(\lambda) - \frac{\sigma^2}{m}\operatorname{tr}(D_n^T D_n) + \frac{2\sigma^2}{m}\operatorname{tr}(D_n^T A_n(\lambda)). \tag{2.7}$$

Providing that σ^2 is known (we will consider the case on unknown σ^2 in Section 5), this estimate may be computed from the data since neither D_n

nor $A_n(\lambda)$ depend on the unknown f. We thus propose choosing λ to minimize $\hat{R}_n(\lambda)$.

The second term on the right of the expression above does not depend on λ. In the case of a kernel estimate, the last term on the right can be seen to be approximately $(2\sigma^2/m)\,\lambda^{2r+1}(-1)^r\,w^{(2r)}(0)$. Minimizing $\hat{R}_n(\lambda)$ is thus approximately equivalent to minimizing

$$SSD_n(\lambda)+\frac{2\sigma^2}{m}\lambda^{2r+1}(-1)^r\,w^{(2r)}(0) \tag{2.8}$$

which penalyzes for large values of λ. In order for this procedure to make sense we clearly need $w^{(2r)}(0)\neq 0$; note that in this case $w^{(2r)}(0)$ is positive or negative according as r is even or odd.

We further note that

$$E\hat{R}_n(\lambda)-R_n(\lambda)=\frac{1}{m}\|D_n f_n-A_n(\lambda)f_n\|^2 - \frac{1}{m}\|f_n^{(r)}-A_n(\lambda)f_n\|^2, \tag{2.9}$$

which should be close to 0 if $D_n f_n$ is a good approximation to $f_n^{(r)}$. Furthermore, $R_n(\lambda)$ should be close to $MISE_n(\lambda)$ if the sum is a good approximation to the integral. If $m\approx n$, we would expect the difference between $E\hat{R}_n(\lambda)$ and $MISE_n(\lambda)$ to be of order n^{-1}.

We thus consider choosing λ to minimize $\hat{R}_n(\lambda)$ in the hope that this minimizer tends to the minimizer of $MISE_n(\lambda)$. In the following sections we provide theoretical justification for this procedure for a particular linear estimation scheme—a tapered Fourier series estimate of the rth derivative of a smooth periodic function.

3. The Tapered Fourier Series Estimate

We will assume henceforth that

$$y_{jn}=f(j/n)+e_j, \qquad j=0,\ldots,n-1 \tag{3.1}$$

where f is a smooth periodic function, and that the e_j are independent random variables with mean 0 and variance σ^2. Let

$$\hat{y}_{kn}=\frac{1}{\sqrt{n}}\sum_{j=0}^{n-1} y_{jn}\,e^{-2\pi ijk/n} \tag{3.2}$$

be the finite Fourier transform of the y_{jn}. Then

$$E\hat{y}_{kn} = \frac{1}{\sqrt{n}} \sum f(j/n)\, e^{-2\pi ijk/n} \tag{3.3}$$

$$= \sqrt{n}\, f_{kn} \tag{3.4}$$

where

$$f_{kn} = \frac{1}{n} \sum f(j/n)\, e^{-2\pi ijk/n} \tag{3.5}$$

$$= f_k + \sum_{s \neq 0} f_{k+sn} \tag{3.6}$$

and where

$$f_k = \int_0^1 f(x)\, e^{-2\pi ikx}\, dx. \tag{3.7}$$

We refer to the coefficients f_{k+sn}, $s \neq 0$ as "aliased" coefficients. The orthogonality of the complex exponential sequence implies that

$$\text{Var}\, \hat{y}_{kn} = \sigma^2. \tag{3.8}$$

We will assume that w has support on $[-\frac{1}{2}, \frac{1}{2}]$. The discrete Fourier coefficients of $\lambda w(\lambda x)$ are

$$w_{kn}(\lambda) = w_k(\lambda) + \sum_{s \neq 0} w_{k+sn}(\lambda) \tag{3.9}$$

where

$$w_k(\lambda) = \lambda \int_{-1/2}^{1/2} w(\lambda x)\, e^{-2\pi ikx}\, dx$$

$$= \int_{-\infty}^{\infty} w(x)\, e^{-2\pi ikx/\lambda}\, dx$$

$$= \tilde{w}(k/\lambda) \tag{3.10}$$

for $\lambda > 1$. The function $\tilde{w}$ is the Fourier transform of w.

The estimate of $f^{(r)}$ that we will consider has Fourier coefficients $(2\pi ik)^r\, \hat{y}_{kn}\, \tilde{w}(k/\lambda)/\sqrt{n}$ for $|k| \leq n/2$, and 0 for $|k| > n/2$. This estimate is very similar to the kernel estimate in the circular case—the difference is that we have discarded aliased Fourier coefficients. To avoid degeneracies we will assume throughout that f is not a trigonometric polynomial of finite degree.

For a discrete approximation to $MISE_n(\lambda)$ we take a Fourier representation via Parseval's theorem:

$$R_n(\lambda) = (2\pi)^{2r} E \sum_{|k| \le n/2} k^{2r} \left| f_{kn} - \tilde{w}(k/\lambda)\, \hat{y}_{kn}/\sqrt{n} \right|^2$$

$$= (2\pi)^{2r} \sum_{|k| \le n/2} k^{2r} \left| f_{kn} - f_{kn}\, \tilde{w}(k/\lambda) \right|^2$$

$$+ (2\pi)^{2r} \frac{\sigma^2}{n} \sum_{|k| \le n/2} k^{2r} \left| \tilde{w}(k/\lambda) \right|^2. \tag{3.11}$$

Our estimate of R_n is based on

$$SSD_n(\lambda) = (2\pi)^{2r} \sum_{|k| \le n/2} k^{2r} \left| \hat{y}_{kn}/\sqrt{n} - \tilde{w}(k/\lambda)\, \hat{y}_{kn}/\sqrt{n} \right|^2$$

$$= \frac{(2\pi)^{2r}}{n} \sum_{|k| \le n/2} k^{2r} \left| \hat{y}_{kn} \right|^2 \left| 1 - \tilde{w}(k/\lambda) \right|^2. \tag{3.12}$$

Now since $E \left| \hat{y}_{kn} \right|^2 = \sigma^2 + n \left| f_{kn} \right|^2$

$$ESSD_n(\lambda) = \frac{(2\pi)^{2r} \sigma^2}{n} \sum_{|k| \le n/2} k^{2r} \left| 1 - \tilde{w}(k/\lambda) \right|^2$$

$$+ (2\pi)^{2r} \sum_{|k| \le n/2} k^{2r} \left| f_{kn} - f_{kn}\, \tilde{w}(k/\lambda) \right|^2. \tag{3.13}$$

Expanding this and using that w is an even function,

$$ESSD_n(\lambda) = R_n(\lambda) - 2 \frac{(2\pi)^{2r} \sigma^2}{n} \sum_{|k| \le n/2} k^{2r}\, \tilde{w}(k/\lambda) \tag{3.14}$$

$$+ \frac{(2\pi)^{2r} \sigma^2}{n} \sum_{|k| \le n/2} k^{2r} \tag{3.15}$$

which allows us to construct an unbiased estimate $\hat{R}_n(\lambda)$ of $R_n(\lambda)$:

$$\hat{R}_n(\lambda) = SSD_n(\lambda) + 2 \frac{(2\pi)^{2r} \sigma^2}{n} \sum_{|k| \le n/2} k^{2r}\, \tilde{w}(k/\lambda)$$

$$- \frac{(2\pi)^{2r} \sigma^2}{n} \sum_{|k| \le n/2} k^{2r}. \tag{3.16}$$

Now the last term above does not depend on λ, and approximating the

sum in the second term on the right by an integral, we see that minimizing $\hat{R}_n(\lambda)$ is nearly equivalent to minimizing

$$SSD_n(\lambda) + \frac{2\sigma^2}{n}(-1)^r \lambda^{2r+1} w^{(2r)}(0). \tag{3.17}$$

(cf. (2.8)).

Now let

$$\zeta_{kn} = |\hat{y}_{kn}|^2 - E|\hat{y}_{kn}|^2. \tag{3.18}$$

Since $\hat{R}_n(\lambda)$ is an unbiased estimate of $R_n(\lambda)$,

$$\hat{R}_n(\lambda) = R_n(\lambda) + \Delta_n(\lambda) \tag{3.19}$$

where

$$\Delta_n(\lambda) = \frac{(2\pi)^{2r}\sigma^2}{n} \sum_{|k| \leq n/2} \zeta_{kn} k^{2r} |1 - \tilde{w}(k/\lambda)|^2 \tag{3.20}$$

has mean 0. From Lemma 2.1 of Rice [6] we have

$$\mathrm{Cov}(\zeta_{kn}, \zeta_{ln}) = \frac{\kappa_4}{n} + (2\sigma^4 + 2n\sigma^2 |f_{kn}|^2)\delta_{kl} \tag{3.21}$$

for $k, l, \neq 0$. Here κ_4 is the fourth order cumulant of the errors.

We close this section with some remarks on $MISE_n$. We will assume that f has $r+2$ derivatives and that $f^{(r+2)}$ is in L_2. By decomposing $MISE_n$ into integrated variance and integrated squared bias, and following the lines of Rosenblatt [7], it can be shown that the integrated variance is of order λ^{2r+1}/n and that the integrated squared bias is of order λ^{-4}. The asymptotically optimal value of λ is $\lambda_n^* = c_0 n^{-1/(2r+5)}$ and $MISE_n(\lambda_n^*)$ is of order $n^{-4/(2r+5)}$.

4. Theoretical Results

In this section we prove several results about $\hat{R}_n(\lambda)$ and its minimizer λ_n. Theorem 1 and Lemma 1 show that $\hat{R}_n(\lambda)$ is uniformly close to $R_n(\lambda)$ and $MISE_n(\lambda)$ for λ in a neighborhood of λ_n^*. This enables us to conclude (Corollary 1) that $(\lambda_n - \lambda_n^*)/\lambda_n^* \to 0$ in probability. Following some preliminary lemmas, Theorem 2 presents an asymptotic normal limiting distribution for λ_n. These results and their proofs closely parallel those of Rice [6]. Throughout the proofs, c denotes a generic constant, and q denotes $1/(2r+5)$.

It is convenient to work with a modification of Δ_n; express Δ_n as

$$\begin{aligned}\Delta_n(\lambda) = &\frac{(2\pi)^{2r}}{n} \sum_{k < \alpha n^q} \zeta_{kn} k^{2r} |1 - \tilde{w}(k/\lambda)|^2 \\ &+ \frac{(2\pi)^{2r}}{n} \sum_{k \geq \alpha n^q} \zeta_{kn} k^{2r} |\tilde{w}(k/\lambda)|^2 \\ &- 2\frac{(2\pi)^{2r}}{n} \sum_{k \geq \alpha n^q} \zeta_{kn} k^{2r} \tilde{w}(k/\lambda) + \frac{(2\pi)^{2r}}{n} \sum_{k \geq \alpha n^q} k^{2r} \zeta_{kn} \end{aligned} \tag{4.1}$$

where α is to be determined later. Let

$$Q_n(\lambda) = R_n(\lambda) + \Delta_n(\lambda) - \frac{(2\pi)^{2r}}{n} \sum_{k \geq \alpha n^q} k^{2r} \zeta_{kn}. \tag{4.2}$$

Since the extra term does not depend on λ, $Q_n(\lambda)$ and $\hat{R}_n(\lambda)$ have the same minimizer.

THEOREM 1. *Assume that* $|f_k|^2 = o(k^{2r-5})$ (*which implies that* $f^{(r+2)}$ *is in* L_2). *Assume that* w *is nonnegative and even with support on* $[-\frac{1}{2}, \frac{1}{2}]$ *and that*

(1) $\tilde{w}''(t) = -(2\pi)^2 \int e^{-2\pi i x t} x^2 w(x)\, dx$ *is of bounded variation.*

(2) w *has* $2r+1$ *continuous derivatives with* $w^{(k)}(\pm\frac{1}{2}) = 0$, $k = 0, 1, \ldots, 2r+1$, *so that from integration by parts*

$$\begin{aligned}\tilde{w}(t) &= \frac{1}{(2\pi i t)^{2r+1}} \int_{-1/2}^{1/2} e^{2\pi i t x} w^{(2r+1)}(x)\, dx \\ &= \frac{1}{(2\pi i t)^{2r+1}} \tilde{v}(t).\end{aligned}$$

(3) $\tilde{v}(t)$ *is of bounded variation.*

Let $I_n = [an^q, bn^q]$, *where* $a < c_0 < b$. *Then*

$$P\left(\sup_{\lambda \in I_n} n^{4q} |R_n(\lambda) - Q_n(\lambda)| \geq \varepsilon\right) \leq \frac{C}{\varepsilon^2} n^{-q} (\log 4n)^2.$$

Proof. The proof follows that of Theorem 2.2 of Rice [6], with a few modifications, and is therefore sketched. Omitting the factor $(2\pi)^{2r}$ for notational simplicity,

$$n^{4q}(R_n(\lambda)-Q_n(\lambda)) = n^{4q-1}\sum_{|k|<\alpha n^q}\zeta_{kn}k^{2r}\,|1-\tilde{w}(k/\lambda)|^2$$

$$+ n^{4q-1}\sum_{|k|\geqslant\alpha n^q}\zeta_{kn}k^{2r}\,|\tilde{w}(k/\lambda)|^2$$

$$- 2n^{4q-1}\sum_{|k|\geqslant\alpha n^q}\zeta_{kn}k^{2r}\tilde{w}(k/\lambda)$$

$$= T_1 + T_2 + T_3.$$

First consider T_1; expanding $\tilde{w}$ about 0

$$1-\tilde{w}(k/\lambda) = -\lambda^2k^2\tilde{w}''(\rho_k)$$

where for α sufficiently small, $\{\rho_k\}$ is an increasing sequence. We thus have

$$T_1 = n^{4q+1}\lambda^{-4}\sum_{|k|<\alpha n^q}\zeta_{kn}k^{2r+4}\,|\tilde{w}''(\rho_k)|^2.$$

Since $\tilde{w}''$ is of bounded variation so is $|\tilde{w}''|^2$. Using summation by parts and this fact

$$T_1 \leqslant cn^{4q+1}\lambda^{-4}\sup_{j<\alpha n^q}\left|\sum_{k=-\alpha n^q}^{j}k^{2r+4}\zeta_{kn}\right|.$$

As in [6] we make use of Lemma 4.1 of Chapter IV of [3] to bound the sup of the partial sums by

$$P\left(\sup_{j<\alpha n^q}\left|\sum_{k=-\alpha n^q}^{j}k^{2r+1}\zeta_{kn}\right|\geqslant\varepsilon\right)\leqslant\frac{1}{\varepsilon^2}(\log 4n)^2\,\mathrm{Var}^*\left(\sum_{|k|<\alpha n^q}k^{2r+4}\zeta_{kn}\right)$$

where Var* denotes the variance of the sum with $|\kappa_4|$ rather then κ_4 used in the expression for covariances. The term involving κ_4 is of smaller order of magnitude in any case, and

$$\mathrm{Var}^*\left(\sum_{|k|<\alpha n^q}\zeta_{kn}k^{2r+4}\right) = O\left(\sum_{|k|<\alpha n^q}k^{4r+8}(\sigma^4+n\,|f_{kn}|^2\sigma^2\right)$$

$$= O(n^{q(4r+9)}).$$

Putting all this together, we find

$$P(\sup_{\lambda}T_1\geqslant\varepsilon)\leqslant cn^{-q}(\log 4n)^2.$$

We now turn to T_3. Replacing $\tilde{w}(t)$ by $\tilde{v}(t)/(2\pi it)^{2r+1}$ and neglecting constants, we have

$$T_3 = n^{4q-1}\lambda^{2r+1}\sum_{|k|>\alpha n^q}\zeta_{kn}k^{-1}\tilde{v}(k/\lambda).$$

Noting that $n^{4q-1}\lambda^{2r+1}$ is bounded above and below for λ in I_n and using the summation by parts argument,

$$P(\sup_{\lambda \in I_n} T_3 \geqslant \varepsilon) \leqslant \frac{1}{\varepsilon^2} \operatorname{Var}^* \sum_{|k| > \alpha n^q} \zeta_{kn} k^{-1}.$$

Now

$$\operatorname{Var}^* \sum_{|k| > \alpha n^q} \zeta_{kn} k^{-1} \approx \sum_{|k| > \alpha n^q} k^{-2} + n \sum_{|k| > \alpha n^q} |f_{kn}|^2 k^{-2}.$$

The first term in $O(n^{-q})$ and the second term is $o(n^{-q})$ from the assumption on the rate of decay of the Fourier coefficients of f.

The analysis of T_2 may be done similarly. This completes the proof of the theorem.

LEMMA 1. *If* $|f_k|^2 = o(k^{-2r-5})$, *then*

$$\sup_{\lambda} |n^{4q}[R_n(\lambda) - MISE_n(\lambda)]| = O(n^{-r-3/2}).$$

Proof. Expressing $MISE_n(\lambda)$ by Parseval's relation in terms of Fourier coefficients, decomposing into integrated variance and integrated bias squared, and comparing with $R_n(\lambda)$ (Eq. (3.11)), it is seen that they differ slightly in the bias term. For $MISE_n(\lambda)$ and $R_n(\lambda)$, these are, respectively,

$$\sum_{|k| < n/2} |f_{kn}\, \tilde{w}(k/\lambda) - f_k|^2 + \sum_{|k| > n/2} |f_k|^2$$

and

$$\sum_{|k| < n/2} |f_{kn}\, \tilde{w}(k/\lambda) - f_{kn}|^2.$$

The result follows on noting that $\sum_{s \neq 0} |f_{k+sn}| = o(n^{-r-3/2})$ for $|k| < n/2$.

COROLLARY 1. *Let* $\theta_n = n^{-q}\lambda_n$, *where* λ_n *is the minimizer of* $\hat{R}_n(\lambda)$ *over* I_n, *and let* $\theta^* = \lim_{n \to \infty} n^{-q}\lambda_n^*$. *Then under the assumptions of Theorem* 1, $\theta_n \to \theta^*$ *in probability.*

Proof. Follows from Theorem 1, Lemma 1 and the method of proof of Corollary 2.2 of Rice [6].

The theorem and corollary thus show that the method of risk estimation produces a bandwidth which converges to the optimal bandwidth. We investigate the rate of convergence by means of the Taylor expansion

$$0 = \hat{R}_n'(\lambda_n^*) + (\lambda_n - \lambda_n^*)\, \hat{R}_n''(\bar{\lambda}_n).$$

We will investigate the terms in this expansion in the following lemmas.

LEMMA 2. *Under the assumptions of Theorem* 1

$$\operatorname{Var}(\hat{R}_n'(\lambda)) = Dn^{-11q} + o(n^{-11q}).$$

Proof. Neglecting the covariance terms, which are of smaller order,

$$\operatorname{Var}(\hat{R}_n'(\lambda)) \approx 2\frac{(2\pi)^{4r+2}\lambda^{-2}}{n^2}\sum k^{4r}(\sigma^4 + n\sigma^2 |f_{kn}|^2)$$
$$\times |1 - \tilde{w}(k/\lambda)|^2 |k\lambda^{-1}|^2 |\tilde{u}(k/\lambda)|^2$$

where $\tilde{u}(t) = \int xw(x)\, e^{-2\pi itx}\, dx$. We split the variance into two parts, the first of which turns out to be dominant and

$$= \frac{2(2\pi)^{4r+2}\sigma^4\lambda^{-2}}{n^2}\sum k^{4r} |1 - \tilde{w}(k/\lambda)|^2 |k\lambda^{-1}|^2 |u(k/\lambda)|^2$$

$$\approx 2(2\pi)^{4r+2}\sigma^4\frac{4^{4r-1}}{n^2}\int |t|^{4r+2} |\tilde{u}(t)|^2 |1 - \tilde{w}(t)|^2\, dt.$$

Evaluating at λ_n^* this term is seen to be of order n^{-11q}. Let D be the coefficient of n^{-11q}.

The other term involves $|f_{kn}|^2$. The range of summation is broken up into $|k| < \alpha n^q$ and $|k| \geq \alpha n^q$ each of which can be shown to be $o(n^{-11q})$ by techniques similar to those used in the proof of Theorem 1.

LEMMA 3. *Under the assumptions of Theorem* 1 *and under the additional assumption the errors have finite moments of all orders,*

$$n^{11q/2}\Delta_n'(\lambda_n^*) \to N(0, D).$$

Proof. See Lemma 2.3 of Rice [6].

LEMMA 4. *Assume the conditions of Theorem* 1 *and also that*

(1) *The function* $xw(x)$ *is* $2r+1$ *times differentiable, these derivatives vanish at* ± 1, *and that the Fourier transform of the* $2r+1^{st}$ *derivative is of bounded variation.*

(2) *The function* $x^2w(x)$ *is* $2r+3$ *times differentiable, these derivatives vanish at* ± 1, *and the Fourier transform of the* $2r+3$*rd derivative is of bounded variation. Then*

$$P(\sup_{\lambda} |\Delta_n''(\lambda)| \geq \varepsilon) \leq \frac{C}{\varepsilon^2} n^{-13q}(\log 4n)^2.$$

Proof. The proof is very similar to that of Theorem 1 and to Lemma

2.2 of Rice [6]. The functions $xw(x)$ and $x^2w(x)$ occur as Fourier transform pairs of the first and second derivatives of $\tilde{w}$. The assumptions on those functions are used in same way as were assumptions (2) and (3) in the proof of Theorem 1.

We now have all the pieces necessary to prove

THEOREM 2. *Under the assumptions of the previous lemmas,*

$$n^{-q/2}(\lambda_n - \lambda_n^*) \to N(0, \rho^2)$$

where ρ is a constant.

Proof. From the Taylor series,

$$0 = n^{11q/2}\Delta_n'(\lambda_n^*) + n^{-q/2}(\lambda_n - \lambda_n^*)\, n^{6q}[R_n''(\bar{\lambda}_n) + \Delta_n''(\bar{\lambda}_n)].$$

From Corollary 1 we know that λ_n is a consistent estimate of λ_n^* in the sense that $(\lambda_n - \lambda_n^*)/\lambda_n^* \to 0$ in probability. R_n'' is a continuous function of λ and if λ is of order n^q, R_n'' is of order n^{-6q}. Lemma 3 controls $\Delta_n''(\bar{\lambda}_n)$ and by Lemma $2\Delta_n'(\lambda_n^*)$ is asymptotically normally distributed.

It is more meaningful to consider $(\lambda_n - \lambda_n^*)/\lambda_n^*$ than $(\lambda_n - \lambda_n^*)$. The theorem says that the standard deviation of the limiting distribution of the former quantity is of order $n^{-q/2}$, so that the relative precision decreases as r increases.

The assumptions under which these results have been derived are quite strong. The assumption that f is smoothly periodic and that the design points are equally spaced allows a diagonal representation for SSD_n. The kernel w has also been assumed to be very smooth. Although these assumptions are necessary for our proofs, we conjecture that similar results hold under more general assumptions, and we believe that the importance and interest of the results lie in their character rather than in the specific assumptions or techniques of proof.

5. CONCLUDING REMARKS

We have assumed that the error variance σ^2 is known, which is not the typical case in practice. If σ^2 is not known, it may be estimated from the data. For example, the estimate

$$\hat{\sigma}^2 = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (y_i - y_{i-1})^2 \tag{5.1}$$

is good enough so that substituting it for σ^2 will not perturb the asymptotics of Section 4.

Another class of methods arise in the following way: Let P be a function such that

$$P(x) = 1 + 2x + O(x^2) \tag{5.2}$$

and consider choosing λ to minimize

$$P_n(\lambda) = SSD_n(\lambda)\, P\left(\frac{\operatorname{tr}[D_n^T A_n(\lambda)]}{\operatorname{tr}(D_n^T D_n)}\right). \tag{5.3}$$

Several choices of P, including generalized cross-validation and Akaike's Information Criterion, are discussed by Shibata [8] and Rice [6]. Writing

$$SSD_n(\lambda) = \hat{R}_n(\lambda) + \frac{\sigma^2}{m}\operatorname{tr}(D_n^T D_n) - \frac{2\sigma^2}{m}\operatorname{tr}[D_n^T A_n(\lambda)] \tag{5.4}$$

and expanding P we have

$$P_n(\lambda) \approx \hat{R}_n(\lambda) + \sigma^2 \operatorname{tr}(D_n^T D_n) + 2\hat{R}_n(\lambda)\frac{\operatorname{tr}[D_n^T A_n(\lambda)]}{\operatorname{tr}(D_n^T D_n)} - \frac{4\sigma^2}{m}\frac{[\operatorname{tr}(D_n^T A_n(\lambda))]^2}{\operatorname{tr}(D_n^T D_n)}. \tag{5.5}$$

The first term on the right-hand side is what we would like to minimize, the second term does not depend on λ, and the third and fourth terms are typically of smaller order than the first for λ in I_n. In the case treated in Section 4 of this paper this expression reduces to

$$P_n(\lambda) \approx \hat{R}_n(\lambda) + \frac{(2\pi)^r \sigma^2}{m}\sum k^{2r} + \hat{R}_n(\lambda)\frac{\sum k^{2r}\tilde{w}(k/\lambda)}{\sum k^{2r}} - \frac{2(2\pi)^r \sigma^2}{m}\frac{(\sum k^{2r}\tilde{w}(k/\lambda))^2}{\sum k^{2r}}. \tag{5.6}$$

The third and fourth terms and the remainder term are really quite small, especially if $w^{(2r)}$ is in L_2. It is easy to see that apart from the second term, which does not depend on λ, $\hat{R}_n(\lambda)$ and $P_n(\lambda)$ are close enough uniformly for λ in I_n so that their minimizers are asymptotically equivalent.

In this paper we have proposed a procedure for bandwidth choice for differentiation and have provided theoretical support in a special case. There is clearly a need for a more general theoretical approach and for practical experience and simulation. It would be interesting to see various functions, equally and unequally spaced data, different signal to noise ratios, and modification of the estimates at the boundary incorporated in a simulation study.

REFERENCES

[1] AKAIKE, H. (1974). A new look at statistical model identification. *IEEE Trans. Automatic Control* **AC-19** 716-723.

[2] CRAVEN, P., AND WAHBA, G. (1979). Smoothing noisy data with spline functions. *Numer. Math.* **31** 377–403.

[3] DOOB, J. (1953). *Stochastic Processes.* Wiley, New York.

[4] HARDLE, W., AND MARRON, J. (1983). Optimal bandwidth selection in nonparametric regression function estimation. Manuscript.

[5] LI, K.-C. (1984). Cross-validated nearest neighbor estimates. *Ann. Statist.* **12** 230–240.

[6] RICE, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.* **12** 1215–1230.

[7] ROSENBLATT, M. (1971). Curve estimates. *Ann. Math. Statist.* **42** 1815–1842.

[8] SHIBATA, R. (1981). An optimal selection of regression variables. *Biometrika* **68** 45–54.

[9] SPECKMAN, P. (1982). Spline smoothing and optimal rates of convergence in nonparametric regression models. Manuscript.

[10] WONG, M. (1983). On the consistency of cross-validation in kernel nonparametric regression. *Ann. Statist.* **11** 1136–1141.