The Annals of Statistics
1984, Vol. 12, No. 4, 1215–1230

BANDWIDTH CHOICE FOR NONPARAMETRIC REGRESSION

BY JOHN RICE[1]

University of California, San Diego

This paper is concerned with the problem of choosing a bandwidth parameter for nonparametric regression. We analyze a tapered Fourier series estimate and discuss the relationship of this estimate to a kernel estimate. We first consider a method based on an unbiased estimate of mean square error, and show that the bandwidth thus chosen is asymptotically optimal. Other methods are examined as well and are shown to be asymptotically equivalent. A small simulation shows, however, that for small or moderate sample size, the methods perform quite differently.

1. Introduction. Nonparametric probability density and regression estimation have become quite popular in recent years, both as theoretical and practical enterprises. The application of such techniques always requires a crucial choice of a smoothing parameter, and various proposals have been made for data-driven procedures for choosing this parameter. Varieties of cross-validation have been especially prominent. Although there have been significant theoretical developments—for example, Craven and Wahba (1979), Chow, et al (1982) and Speckman (1982)—many questions remain open.

Given data

$$y_i = f(x_i) + \varepsilon_i$$

where $x_i = i/n$, $i = 0, \cdots, n-1$, and the ε_i's are independent random variables with mean 0 and variance σ^2, a kernel estimate of f is

$$f_n(x) = \sum_{i=0}^{n-1} (\lambda/n)\, w(\lambda(x_i - x))y_i. \tag{1.1}$$

Here w is the kernel which is symmetric about zero, and λ is the reciprocal of the bandwidth, which is to be chosen from the data. The estimate (1.1) as it stands should be modified for x near the boundary—see Gasser and Muller (1979) or Rice (1984)—unless f is smoothly periodic.

In this paper we will assume that f is smoothly periodic and will consider a circular version of (1.1). This makes possible the use of Fourier analysis (note that (1.1) is a convolution). As a further simplification, we will analyze a modified version of (1.1), to be described in detail below. Basically, this modification consists of discarding "aliased" Fourier coefficients of w, and thus replacing the kernel estimate by a tapered Fourier series estimate. This simplifies certain technical arguments and makes the essential arguments much easier to follow, as was pointed out by a referee. However, we wish to make it clear that our results apply to the modified estimator.

Received September 1982; revised April 1984.
[1] Research sponsored by NSF Grant MCS-7901800.
AMS 1980 *subject classification.* 62G99, 62J99.
Key words and phrases. Nonparametric regression, kernel regression, cross-validation, smoothing.

We consider attempting to choose λ so as to minimize the risk function

$$R_n(\lambda) = E(1/n) \sum_{i=0}^{n-1} (f(x_i) - f_n(x_i))^2$$

a discrete approximation to the integrated mean square error.

If we let

$$Y = (y_0, \cdots, y_{n-1})^T, \quad f = (f(x_0), \cdots, f(x_{n-1}))^T,$$

and

$$W(\lambda) = [(\lambda/n)w(\lambda(x_i - x_j))],$$

$R_n(\lambda)$ may be expressed as

$$R_n(\lambda) = E(1/n) \| f - W(\lambda)Y \|^2.$$

The residual sum of squares is

$$\mathrm{RSS}_n(\lambda) = \| Y - W(\lambda)Y \|^2.$$

A simple calculation shows that

$$\begin{aligned} \frac{1}{n} \mathrm{ERSS}_n(\lambda) &= R_n(\lambda) + \sigma^2 - \frac{2\sigma^2}{n} \operatorname{trace}(W(\lambda)) \\ &= R_n(\lambda) + \sigma^2 - \frac{2\sigma^2 \lambda w(0)}{n}. \end{aligned} \tag{1.2}$$

Thus, if σ^2 were known, one could form an unbiased estimate of the risk:

$$\hat{R}_n(\lambda) = (1/n)\mathrm{RSS}_n(\lambda) - \sigma^2 + 2\sigma^2\lambda w(0)/n \tag{1.3}$$

and choose λ to minimize $\hat{R}_n(\lambda)$. This sort of scheme was suggested by Mallows (1973) in the context of variable selection in regression, and by Craven and Wahba (1979) in the context of choosing a smoothing parameter for a smoothing spline. Since σ^2 will not typically be known, it is tempting to form an estimate of σ^2, $\hat{\sigma}^2$, from successive differences $y_{k+1} - y_k$ or by pooling single degree of freedom estimates formed from the residuals of straight lines fit to successive triples of points and choose λ to minimize

$$\tilde{R}_n(\lambda) = (1/n)\mathrm{RSS}_n(\lambda) - \hat{\sigma}^2 + 2\hat{\sigma}^2\lambda w(0)/n. \tag{1.4}$$

Crossvalidation is intended to accomplish the same aim. Letting $f_n^{(k)}(x)$ denote the estimated regression function with the kth point deleted, λ is chosen to minimize

$$CV(\lambda) = (1/n) \sum_{k=0}^{n-1} (y_k - f_n^{(k)}(x_k))^2. \tag{1.5}$$

Since this paper was first written, the author has seen manuscripts on crossvalidation by Hall (1982), Li (1983a, b) and Wong (1982). Hall shows asymptotic efficiency of a form of crossvalidated estimation for density estimation. Wong shows consistency of crossvalidated kernel regression estimates for equispaced data, using methods that are quite different from those used in this

paper. Li (1983a) shows consistency for crossvalidated nearest-neighbor regression estimates. A very interesting paper of Li (1983b) relates crossvalidation, risk estimation, and Stein estimates to nonparametric estimation problems. Breiman and Freedman (1983) discuss some related problems in variable selection for regression. However, as far as we know, this paper is the first to establish rates of convergence for an estimated bandwidth (the bandwidth of the modified estimator).

The remainder of this paper is organized as follows: In Section 2 the modified estimate is defined and the properties of λ_n, the minimizer of $\hat{R}_n(\lambda)$, are investigated as $n \to \infty$. It is first shown that the minimizer of $\hat{R}_n$ yields a consistent estimate of f. It is next shown that there is a (possible local) minimizer λ_n such that

$$(\lambda_n - \lambda_n^*)/\lambda_n^* \to_p 0$$

where λ_n^* is the minimizer of $R_n(\lambda)$. We then deduce an asymptotically normal limiting distribution for an appropriately normalized version of λ_n. In Section 3 we show that for the original (unmodified) problem, the minimizers of $\tilde{R}_n(\lambda)$ and some other criteria have the same limiting behavior as the minimizer of $\hat{R}_n(\lambda)$. In Section 4 we report the results of a small simulation.

2. Unbiased risk estimation. In this section we derive the main results of the paper. We consider the estimate λ_n which minimizes $\hat{R}_n(\lambda)$. In Theorem 2.1 we show that $R_n(\lambda_n) \to_p 0$ as $n \to \infty$. Theorem 2.2 shows that $(\lambda_n - \lambda_n^*)/\lambda_n^* \to_p 0$. A limiting normal distribution for λ_n is given in Theorem 2.3, which shows that the standard deviation of $(\lambda_n - \lambda_n^*)/\lambda_n^* \sim 1/\sqrt{\lambda_n^*}$.

Throughout this section we will assume that w is a symmetric probability density with $\int x\, w(x)\, dx = 0$. In order to establish some of the results, it will be necessary to assume further smoothness properties of w.

The estimate 1.1 is a convolution, and to investigate its properties and those of the minimizer of $\hat{R}_n$ it is convenient to use Fourier analysis. We will assume that the model is circular—f will be regarded as a periodic function. This assumption of circularity is clearly something of an artifice; however, if the model is noncircular, an estimate of the form (1.1) will not be satisfactory for x near the boundary.

We first introduce some notation. Let

$$y_{jn} = f(j/n) + \varepsilon_{jn}, \quad j = 0, \cdots, n-1$$

where $E\,\varepsilon_{jn} = 0$ and the ε_{jn}'s are independent random variables with common variance σ^2. Let

$$\hat{y}_{kn} = (1/\sqrt{n}) \sum_{j=0}^{n-1} y_{jn} e^{-2\pi ijk/n} \tag{2.1}$$

be the finite Fourier transform of the sequence $y_{jn}/\sqrt{n}$ (Cooley, et al., 1977). It follows that

$$E\hat{y}_{kn} = (1/\sqrt{n}) \sum_{j=0}^{n-1} f(j/n) e^{-2\pi ijk/n} = \sqrt{n} f_{kn} \tag{2.2}$$

where

$$f_{kn} = (1/n) \sum_{j=0}^{n-1} f(j/n)e^{-2\pi ijk/n} \tag{2.3}$$

is the kth finite Fourier coefficient of the sequence $(1/n)f(j/n)$. Note that if

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

then

$$f_{kn} = f_k + \sum_{s\neq 0} f_{k+sn}. \tag{2.4}$$

We will assume that w is nonnegative and has support on $[-\frac{1}{2}, \frac{1}{2}]$. For each λ the function $\lambda w(\lambda x)$ is extended periodically in order that the convolution (1.1) be well defined. The finite Fourier coefficients of the sequence $(\lambda/n)w(\lambda j/n)$ will be denoted by $w_{kn}(\lambda)$. As was true for f

$$w_{kn}(\lambda) = w_k(\lambda) + \sum_{s\neq 0} w_{k+sn}(\lambda). \tag{2.5}$$

If w has support on $[-\frac{1}{2}, \frac{1}{2}]$,

$$\begin{aligned} w_k(\lambda) &= \lambda \int_{-1/2}^{1/2} w(\lambda x)e^{-2\pi ikx}\,dx \\ &= \int w(x)e^{-2\pi ikx/\lambda}\,dx = \tilde{w}(k/\lambda), \quad \text{say,} \end{aligned} \tag{2.6}$$

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

We now discuss our modified estimate. Since the estimate (1.1) computed at points j/n is a convolution, its finite Fourier coefficients are $\sqrt{n}\hat{y}_{kn}w_{kn}(\lambda)$. It is well known that in this circular case the effect of a kernel smoothing is to taper off the higher order Fourier coefficients. Now the modified estimate that we will actually analyze below is defined to be a trigonometric polynomial with Fourier coefficients $\sqrt{n}\hat{y}_{kn}\tilde{w}(k/\lambda)$. Here $a_n \leq k \leq b_n$ where

$$a_n = \begin{cases} -n/2 & n \text{ even} \\ -(n-1)/2 & n \text{ odd} \end{cases} \qquad b_n = \begin{cases} n/2 - 1 & n \text{ even} \\ (n-1)/2 & n \text{ odd.} \end{cases}$$

The estimate is thus a tapered Fourier series. Discarding the aliased Fourier coefficients eliminates many tedious technical details in the proofs that follow. Although we have not done so, we conjecture that a proof for the unmodified estimate can be obtained in a similar manner.

Now using Parseval's relation

$$\begin{aligned} (1/n)\mathrm{RSS}(\lambda) &= (1/n) \sum_{k=a_n}^{b_n} |\hat{y}_{kn} - \hat{y}_{kn}\tilde{w}(k/\lambda)|^2 \\ &= (1/n) \sum_{k=a_n}^{b_n} |\hat{y}_{kn}|^2 \, |1 - \tilde{w}(k/\lambda)|^2. \end{aligned} \tag{2.7}$$

(Had we retained the original estimate, $w_{kn}(\lambda)$ would have appeared in place of $\tilde{w}(k/\lambda)$). The reason for using Fourier analysis is to obtain this simple diagonalization of RSS.

In order to avoid degenerate cases, we will assume throughout that f is not a trigonometric polynomial of finite degree.

We will need the moments of $|\hat{y}_{kn}|^2$. Clearly,

$$E|\hat{y}_{kn}|^2 = \sigma^2 + n|f_{kn}|^2. \tag{2.8}$$

We also have

LEMMA 2.1. *If the ε_j's have fourth order cumulant, κ_4, for $k \neq 0$, $\ell \neq 0$,*

$$\operatorname{Cov}(|\hat{y}_{kn}|^2, |\hat{y}_{\ell n}|^2) = \kappa_4/n + (2\sigma^4 + 2n\sigma^2 |f_{kn}|^2)\delta_{k\ell}.$$

PROOF. The proof is straightforward but rather laborious, so details will be omitted. First, express the desired covariance in terms of the cumulants of the $\hat{y}$'s, and then compute these cumulants in terms of the cumulants of the y's from (2.1) using orthogonality properties of the sequence $\{e^{-2\pi ikj/n}\}$.

For k or $\ell = 0$ some additional terms appear, but these play no role in the development below.

Now let $\zeta_{kn} = |\hat{y}_{kn}|^2 - E|\hat{y}_{kn}|^2$. Since $\hat{R}_n(\lambda)$ is an unbiased estimate of $R_n(\lambda)$

$$\hat{R}_n(\lambda) = R_n(\lambda) + \Delta_n(\lambda) \tag{2.9}$$

where

$$\Delta_n(\lambda) = (1/n) \sum_{k=a_n}^{b_n} \zeta_{kn} |1 - \tilde{w}(k/\lambda)|^2 \tag{2.10}$$

and $E\Delta_n = 0$.

We will now argue that the estimated bandwidth gives a consistent estimate of f.

THEOREM 2.1. *Assume that $\tilde{w}(t)$ is of bounded variation and that the ε_j's have fourth moments. Then*

$$P\{\sup_{1<\lambda<\infty} |1/n \sum_{k=a_n}^{b_n} \zeta_{kn} |1 - \tilde{w}(k/\lambda)|^2| \geq \varepsilon\} \leq (c/\varepsilon^2) n^{-1} (\log 4n)^2$$

where c does not depend on n.

PROOF.

$$\begin{aligned}(1/n) \sum_{k=a_n}^{b_n} \zeta_{kn} |1 - \tilde{w}(k/\lambda)|^2 &= (1/n) \sum{}' \zeta_{kn} - (2/n) \sum{}' \zeta_{kn} \tilde{w}(k/\lambda) + (1/n) \sum{}' \zeta_{kn} |\tilde{w}(k/\lambda)|^2 \\ &= T_1 + T_2 + T_3, \quad \text{say.}\end{aligned}$$

Here we have used the fact that since w is symmetric, $\tilde{w}$ is real. Since $\tilde{w}(0) = 1$, the term $k = 0$ vanishes. We have used $\sum'$ to denote a sum excluding 0. We will verify the bound for each term. Chebyshev's inequality gives the result for T_1. Using summation by parts, T_2 can be expressed as

$$T_2 = (2/n) \sum_{k=a_n}^{b_n - 1\prime} \Delta\tilde{w}(k/\lambda) \sum_{j=a_n}^{k\prime} \zeta_{jn} - (2/n)\tilde{w}(b_n/\lambda) \sum_{k=a_n}^{b_n\prime} \zeta_{kn}.$$

The bound on the second term of the equation above follows from Chebyshev's

inequality and the fact that $\tilde{w}$ is bounded. As for the first term

$$\sup_\lambda (2/n) \sum_{k=a_n}^{b_n-1\prime} \Delta\tilde{w}(k/\lambda) \sum_{j=a_n}^{k\prime} \zeta_{jn}$$

$$\leq (2/n)\sup_k \left| \sum_{j=a_n}^{k\prime} \zeta_{jn} \right| \sup_\lambda \sum_{k=a_n}^{b_n-1\prime} |\Delta\tilde{w}(k/\lambda)|.$$

The second sup is finite since $\tilde{w}$ is of bounded variation. For the first sup we can appeal to Lemma 4.1 of Chapter IV of Doob (1953). As stated, Doob's lemma applies to mean zero uncorrelated random variable $Y_1, \cdots, Y_n$ with variances σ_i^2 and states that

$$E\{\max_{j\leq n} |\sum_{i=1}^{j} Y_j|^2\} \leq \left(\frac{\log 4n}{\log 2}\right)^2 \sum_{i=1}^{n} \sigma_i^2.$$

Doob uses the fact that the Y_i's are uncorrelated in concluding that the expectation of sums of the form $|Y_1 + Y_2 + Y_3|^2 + \cdots + |Y_{n-2} + Y_{n-1} + Y_n|^2$, for example, is $\sigma_1^2 + \cdots + \sigma_n^2$. If the Y_i's are correlated, covariance terms appear, but if all the covariance terms are positive the expectation of any such sum is less than or equal to $\operatorname{var} \sum_{i=1}^{n} Y_i$. In our case the covariances are given by Lemma 2.1. If we define $\operatorname{var}^* \sum \zeta_{in}$ to be the variance computed with κ_4 replaced by $|\kappa_4|$, then

$$E\{\max_{k\leq m} |\sum_{j=a_n}^{k\prime} \zeta_{jn}|^2\} \leq \left(\frac{\log 4n}{\log 2}\right)^2 \operatorname{var}^* \sum_{j=a_n}^{b_n\prime} \zeta_{jn}.$$

It may be noted that the contribution from the covariance terms is, in any case, of smaller order than the contribution from the variance terms. T_3 can be analyzed similarly by writing

$$|\tilde{w}((k+1)/\lambda)|^2 - |\tilde{w}(k/\lambda)|^2$$

$$= \tilde{w}((k+1)/\lambda)[\tilde{w}((k+1)/\lambda) - \tilde{w}(k/\lambda)] + \tilde{w}(k/\lambda)[\tilde{w}((k+1)/\lambda) - \tilde{w}(k/\lambda)]$$

and using that $\tilde{w}$ is bounded. This completes the proof of the theorem.

COROLLARY 2.1. *Let λ_n be the minimizer of $\hat{R}_n(\lambda)$. Then under the assumption of the previous theorem, $R_n(\lambda_n) \to_p 0$.*

PROOF. There exists a deterministic sequence $\overline{\lambda}_n$ such that $R_n(\overline{\lambda}_n) \to 0$. Now

$$R_n(\lambda_n) = \hat{R}_n(\lambda_n) - \Delta_n(\lambda_n) \leq \hat{R}_n(\overline{\lambda}_n) - \Delta_n(\lambda_n)$$

$$= R_n(\overline{\lambda}_n) + \Delta_n(\overline{\lambda}_n) + \Delta_n(\lambda_n)$$

$$\leq R_n(\overline{\lambda}_n) + 2 \sup_\lambda |\Delta_n(\lambda)| \to 0.$$

If $f'' \in C^2$ then the asymptotically optimal value of λ_n is $\lambda_n^* = c_0 n^{1/5}$. (Rosenblatt (1971) gives results of this character for kernel density estimation. Similar reasoning can be applied in our case. Since such arguments are by now routine in this field, we will omit them.) We would like to show that λ_n, the minimizer

of $\hat{R}_n(\lambda)$, is a consistent estimate of λ_n^* in the sense that $n^{-1/5}(\lambda_n - \lambda_n^*) \to 0$, but we have not been able to do this in full generality. We will, however, show that for any constants a and b $(a < c_0 < b)$ there is a consistent local minimizer $an^{1/5} \le \lambda_n \le bn^{1/5}$. The question of whether this is a global minimizer remains open.

First, we will rewrite $\Delta_n(\lambda)$

$$\begin{aligned}\Delta_n(\lambda) &= (1/n) \textstyle\sum_{|k|\le an^{1/5}} \zeta_{kn} |1 - \tilde{w}(k/\lambda)|^2 \\ &\quad + (1/n) \textstyle\sum_{|k|>an^{1/5}} \zeta_{kn} |\tilde{w}(k/\lambda)|^2 - (2/n) \textstyle\sum_{|k|>an^{1/5}} \zeta_{kn}\tilde{w}(k/\lambda) \\ &\quad + (1/n) \textstyle\sum_{|k|>an^{1/5}} \zeta_{kn}.\end{aligned} \tag{2.11}$$

Here, $\alpha > 0$, is to be determined later. Let

$$\hat{Q}_n(\lambda) = R_n(\lambda) + \Delta_n(\lambda) - (1/n) \textstyle\sum_{|k|>\alpha n^{1/5}} \xi_{kn}. \tag{2.12}$$

Since the last term does not depend on λ, $\hat{Q}_n$ and $\hat{R}_n$ have the same minimizer.

We will first show that

$$\sup_{an^{1/5}\le\lambda\le bn^{1/5}} n^{4/5} |R_n - \hat{Q}_n| \to_p 0 \tag{2.13}$$

and then see that this implies that $n^{-1/5}(\lambda_n - \lambda_n^*) \to_p 0$. The reason for the normalization by $n^{4/5}$ is that $n^{4/5}R_n(\lambda_n^*) \to c$. The reason for considering $\hat{Q}_n$ rather than $\hat{R}_n$ is to discard the variability contributed by the extra term. (In the following theorem and throughout the paper we will let c be a generic constant the precise meaning of which may change from usage to usage.)

THEOREM 2.2. *Assume that* $|f_k|^2 = o(k^{-5})$ *(which implies that* $f'' \in C^2$*). Assume 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 a derivative* v*, and* $\tilde{v}$ *is of bounded variation. Then*

$$P\{\sup_{an^{1/5}\le\lambda\le bn^{1/5}} n^{4/5} |R_n - \hat{Q}_n| \ge \varepsilon\} \le \frac{cn^{-1/5}}{\varepsilon^2} (\log 4n)^2.$$

PROOF. $n^{4/5}(R_n - \hat{Q}_n)$ consists of the first three terms of (2.11). We will estimate each term, the first one being

$$T_1 = n^{-1/5} \textstyle\sum'_{|k|\le an^{1/5}} \zeta_{kn} |1 - \tilde{w}(k/\lambda)|^2.$$

Expanding $\tilde{w}$ about 0

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

We claim that for α sufficiently small, the sequence $\{\rho_k\}$ is increasing. To see this first note that $\tilde{w}''(0) = -(2\pi)^2 \int x^2 w(x)\, dx < 0$ implies that $(1 - \tilde{w}(t)/t^2$ is decreasing in a neighborhood of 0. Secondly, $\tilde{w}''$ is increasing in a neighborhood of 0 since $\tilde{w}^{(3)}(0) = 0$ and $\tilde{w}^{(4)}(0) = (2\pi)^4 \int x^4 w(x)\, dx > 0$. Now

$$T_1 = n^{-1/5}\lambda^{-4} \textstyle\sum'_{|k|<an^{1/5}} \zeta_{kn} k^4 |\tilde{w}''(\rho_k)|^2.$$

Summation by parts may be used as in Theorem 1.1. Note that

$$\text{var}^*(\sum'_{|k|\le\alpha n^{1/5}} k^4\zeta_{kn}) \simeq \sum'_{|k|\le\alpha n^{1/5}} (k^8\sigma^4 + n|f_{kn}|^2k^8)$$

$$= O(n^{9/5}) + o(n^{9/5})$$

by the assumption on the rate of decrease of the Fourier coefficients of f.

The second term in $n^{4/5}(R_n - Q_n)$ is

$$T_2 = -2n^{-1/5} \sum_{|k|>\alpha n^{1/5}} \zeta_{kn}\tilde{w}(k/\lambda).$$

Consider the sum T_2^+ over $k > \alpha n^{1/5}$ (the other part of the sum may be analyzed similarly). Writing $\tilde{w}(t) = \tilde{v}(t)/(2\pi it)$ and using summation by parts

$$T_2^+ \le c\lambda n^{-1/5}\sup_k |\sum_{j>\alpha n^{1/5}}^k (1/j)\zeta_{jn}| \sum_{k>\alpha n^{1/5}} |\Delta\tilde{v}(k/\lambda)|$$

and

$$P\{\sup_\lambda |T_2^+| \ge \varepsilon\} \le (c/\varepsilon^2)\text{var}^* \sum_{k>\alpha n^{1/5}} (1/k)\zeta_{kn}(\log 4n)^2.$$

Furthermore

$$\text{var}^*(\sum_{k>\alpha n^{1/5}} (1/k)\zeta_{kn}) \sim \sum_{k>\alpha n^{1/5}} (1/k^2) + n \sum_{k>\alpha n^{1/5}} |f_{kn}|^2(1/k^2)$$

$$= O(n^{-1/5}) + o(n^{-1/5})$$

which completes the analysis of T_2. The analysis of

$$T_3 = n^{-1/5} \sum_{|k|>\alpha n^{1/5}} \zeta_{kn}|\tilde{w}(k/\lambda)|^2$$

is similar to that of T_2. This concludes the proof of the theorem.

COROLLARY 2.2. *Let* $\theta_n = n^{-1/5}\lambda_n$, $\theta_n^* = n^{-1/5}\lambda_n^*$ *and* $\theta^* = \lim_{n\to\infty}\theta_n^*$. *Then under the assumptions of Theorem* 2.2, $\theta_n \to_p \theta^*$.

PROOF. Let $T(\theta) = \lim_{n\to\infty} n^{4/5}R_n(\theta)$ for $a < \theta < b$. $T(\theta)$ is a continuous and convex function and the convergence of $n^{4/5}R_n$ to T is uniform in $[a, b]$. (Again, these claims may be established by arguments similar to those in Rosenblatt, 1971.) θ^* is the minimizer of T. Now

$$\sup_{a\le\theta\le b}|n^{4/5}\hat{Q}_n(\theta) - T(\theta)|$$

$$\le \sup_{a\le\theta\le b}|n^{4/5}\hat{Q}_n(\theta) - n^{4/5}R_n(\theta)| + \sup_{a\le\theta\le b}|n^{4/5}R_n(\theta) - T(\theta)| \to 0.$$

To prove that the minimizer of $\hat{T}_n(\theta) = n^{4/5}\hat{Q}_n(\theta)$ tends in probability to the minimizer of T, the following argument, due to Ian Abramson, will be used. For any $\delta > 0$ define

$$D(\delta) = \inf_{|\theta-\theta^*|>\delta}(T(\theta) - T(\theta^*)).$$

Then

$$P[|\theta_n - \theta^*| > \delta]$$
$$\leq P[T(\theta_n) - T(\theta^*) > D(\delta)]$$
$$\leq P[T(\theta_n) - \hat{T}_n(\theta_n) + \hat{T}_n(\theta^*) - T(\theta^*) > D(\delta)]$$
$$\leq P[T(\theta_n) - \hat{T}_n(\theta_n) \geq D(\delta)/2] + P[\hat{T}_n(\theta^*) - T(\theta^*) \geq D(\delta)/2]$$
$$\to_p 0$$

since $\hat{T}_n$ converges to T uniformly in probability.

This argument can be used to obtain bounds on $|\theta_n - \theta^*|$, since $D(\delta) \sim \delta^2$.

The next goal is to analyze the asymptotic behavior of $\lambda_n - \lambda_n^*$ by means of the Taylor expansion

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

We will do this term by term:

LEMMA 2.2. *Assume that* $|f_k|^2 = o(k^{-5})$ *and that* $\int t^2 |\tilde{w}'(t)|^2 \, dt < \infty$. *Then*

$$\text{var}[\hat{R}_n'(\lambda_n^*)] = D\, n^{-11/5} + o(n^{-11/5})$$

where D *is a constant given in the proof.*

PROOF.

$$\hat{R}_n'(\lambda) = (-2/n)\lambda^{-2} \sum \zeta_{kn} k\, \tilde{w}'(k/\lambda)(1 - \tilde{w}(k/\lambda))$$

and

$$\text{var}\, \hat{R}_n'(\lambda) = (8/n^2)\, \lambda^{-4} \sum (\sigma^4 + n\sigma^2 |f_{kn}|^2)k^2 |\tilde{w}'(k/\lambda)|^2 |1 - \tilde{w}(k/\lambda)|^2$$
$$+ \text{covariance terms.}$$

The covariance terms are of smaller order. We first consider

$$(8\lambda^{-4}/n^2)\, \sigma^4 \sum k^2 |\tilde{w}'(k/\lambda)|^2 |1 - \tilde{w}(k/\lambda)|^2$$
$$\cong (8\lambda^{-1}/n^2)\sigma^4 \int t^2 |\tilde{w}'(t)|^2 |1 - \tilde{w}(t)|^2 \, dt.$$

At $\lambda = \lambda_n^*$, this term is of order $n^{-11/5}$; let D be the coefficient of $n^{-11/5}$ as $n \to \infty$ (and $n^{-1/5}\lambda_n^* \to \theta^*$). The remaining term may be analyzed by splitting the range of summation as before. Over the range $|k| < \alpha n^{1/5}$

$$1 - \tilde{w}(k/\lambda) \simeq k^2\lambda^{-2}\tilde{w}''(0).$$

Since $\tilde{w}'(t)$ is bounded, the magnitude of the sum is

$$(\lambda^{-8}/n) \sum_{|k|<\alpha n^{1/5}} k^6 |f_{kn}|^2 = o(n^{-11/5})$$

by the assumption on f_k.

Over the range $|k| > \alpha n^{1/5}$, we have

$$(8\lambda^{-4}/n) \sum_{|k|>\alpha n^{1/5}} k^2 |f_{kn}|^2 |\tilde{w}'(k/\lambda)|^2 |1 - \tilde{w}(k/\lambda)|^2$$

$k^2 |f_{kn}|^2 = o(k^{-3})$; pulling this term out and bounding the sum by an integral shows that the expression above is also $o(n^{-11/5})$.

LEMMA 2.3. *Assume that* $|f_k|^2 = o(k^{-5})$. *Also assume that*

(1) $xw(x)$ *is differentiable and the Fourier transform of its derivative is of bounded variation,*

(2) $x^2w(x)$ *is three times differentiable and the Fourier transform of its third derivative is of bounded variation.*

Then

$$P\{\sup_{\alpha n^{1/5} \le \lambda \le b n^{1/5}} |\Delta_n''(\lambda)| \ge \varepsilon\} \le (c/\varepsilon^2) n^{-13/5} (\log 4n)^2.$$

PROOF.

$$\Delta_n(\lambda) = (1/n) \sum \zeta_{kn} (d^2/d\lambda^2)(1 - \tilde{w}(k/\lambda))^2$$

and

$$(d^2/d\lambda^2)(1 - \tilde{w}(k/\lambda))^2 = -(4k/\lambda^3)[\tilde{w}'(k/\lambda)(1 - \tilde{w}(k/\lambda))] - (2k^2/\lambda^4)[\tilde{w}''(k/\lambda)[1 - \tilde{w}(k/\lambda)] + |\tilde{w}'(k/\lambda)|^2].$$

We will break this up into several terms. Let

$$T_{11} = (-4/\lambda^3 n) \sum_{|k|<\alpha n^{1/5}} k \zeta_{kn} \tilde{w}'(k/\lambda)(1 - \tilde{w}(k/\lambda)).$$

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

$$T_{11} = (4/\lambda^5 n) \sum_{|k| \le \alpha n^{1/5}} k^3 \zeta_{kn} \tilde{w}'(k/\lambda) \tilde{w}''(\rho_k).$$

We use the summation by parts argument again, noting that if functions h and g are bounded and of bounded variation, so is their product. Furthermore,

$$\text{var}^*((\lambda^{-5}/n) \sum_{|k| \le \alpha n^{1/5}} k^3 \zeta_{kn}) \simeq (\lambda^{-10}/n^2) \sum_{|k| \le \alpha n^{1/5}} (k^6 + k^6 n |f_{kn}|^2) = O(n^{-13/5}) + o(n^{-13/5}).$$

Secondly, let

$$T_{12} = -(4\lambda^{-3}/n) \sum_{k > \alpha n^{1/5}} k \zeta_{kn} \tilde{w}'(k/\lambda)(1 - \tilde{w}(k/\lambda))^2.$$

The assumption on $xw(x)$ allows us to write $\tilde{w}'(t) = \tilde{r}(t)/t^2$ where $\tilde{r}$ is of bounded variation. Substituting this, and using the summation by parts argument, we have to estimate

$$\text{var}^*\left(\frac{\lambda^{-1}}{n} \sum_{k > \alpha n^{1/5}} \frac{1}{k} \zeta_{kn}\right) \sim \frac{\lambda^{-2}}{n^2} \sum_{|k| > \alpha n^{1/5}} (\sigma^4 + n\sigma^2 |f_{kn}|^2) \frac{1}{k^2} = O(n^{-13/5}).$$

The term

$$T_2 = -(2\lambda^{-4}/n) \sum k^2 \zeta_{kn} \tilde{w}''(k/\lambda)(1 - \tilde{w}(k/\lambda))$$

can be analyzed similarly. Finally, there is the term

$$T_3 = -(2\lambda^{-4}/n) \sum k^2 \zeta_{kn} |\tilde{w}'(k/\lambda)|^2.$$

Over the range $|k| \leq \alpha n^{1/5}$ we expand $\tilde{w}'$ about zero and use the fact that $\tilde{w}'(0) = 0$. Over the range $k > \alpha n^{1/5}$ the argument proceeds as before.

LEMMA 2.4. *Assume the conditions of Lemma* 2.2 *and also assume that the* ε's *have finite moments of all orders. Then*

$$n^{11/10}\hat{R}_n'(\lambda_n^*) \to_{\mathcal{D}} N(0, D)$$

where D is given in Lemma 2.2.

PROOF. The proof follows from showing that the standardized cumulants of order greater than two of $\hat{R}_n'(\lambda_n^*)$ tend to zero. These cumulants can be calculated from the cumulants of the y's via the rules of Leonov and Shiryaev.

We now have all the pieces needed to show that λ_n has a limiting normal distribution. The Taylor series expansion is

$$\begin{aligned} O &= R_n'(\lambda_n^*) + (\lambda_n - \lambda_n^*)\hat{R}_n''(\overline{\lambda}_n) \\ &= \Delta_n'(\lambda_n^*) + (\lambda_n - \lambda_n^*)[R_n''(\overline{\lambda}_n) + \Delta_n''(\overline{\lambda}_n)] \end{aligned}$$

or

$$n^{-1/10}(\lambda_n - \lambda_n^*) = \frac{n^{11/10}\Delta_n'(\lambda_n^*)}{n^{6/5}R_n''(\overline{\lambda}_n) + n^{6/5}\Delta_n''(\overline{\lambda}_n)}.$$

From Lemma 2.2

$$\operatorname{var}(n^{11/10}\Delta_n'(\lambda_n^*)) = D.$$

From Lemma 2.3

$$P\{\sup_{an^{1/5}<\lambda<bn^{1/5}} n^{6/5} |\Delta_n''(\lambda)| \geq \varepsilon\} \leq c\, n^{-1/5}(\log 4n)^2.$$

Also, R_n'' is a continuous function and $R_n''(\lambda_n^*) \simeq c\, n^{-6/5}$. We thus have

THEOREM 2.3. *Assume the conditions of Lemmas* 2.2, 2.3, *and* 2.4. *Then*

$$n^{-1/10}(\lambda_n - \lambda_n^*) \to_{\mathcal{D}} N(0, \rho)$$

where $\rho = D/c^2$.

Since λ_n and λ_n^* are of order $n^{1/5}$, the theorem says that the relative fluctuations, $(\lambda_n - \lambda_n^*)/\lambda_n^*$ are of order $n^{-1/10}$ or $1/\sqrt{\lambda_n^*}$. This result has been derived under the assumption that $f'' \in L_2$ and w is "matched" to f in the sense that $\int xw(x)\,dx = 0$ and $\int x^2 w(x)\,dx > 0$. For the more general case, $f^{(m)} \in L_2$, $\int x^k w(x)\,dx = 0$, $k = 1, \cdots, m-1$, and $\int x^m w(x)\,dx > 0$, the optimal λ is $\lambda_n^* \sim n^{1/(2m+1)}$. The arguments above can be traced through and it can be seen that the relative fluctuations are of order $1/\sqrt{\lambda_n^*}$ again.

Lemma 2.3 assumes that w is smoother than kernels generally used in practice. For example w might be three times differentiable with support on $[-\frac{1}{2}, \frac{1}{2}]$ and

$w^{(3)}(\pm 1/2) = 0$. This assumption enabled us to argue that a Fourier transform was of bounded variation. It might well be that more delicate arguments could use weaker assumptions. The real practical importance of the smoothness of w is not clear to us.

We have assumed that the data are equally spaced and that the model is circular. This avoids the problem of modifying the estimate at the boundary, where it would no longer be of convolution type, and makes possible the use of finite Fourier analysis to obtain a nice expression for $\hat{R}$. An analysis for the more general situation would have to take account of unequally spaced data and of the boundary and would proceed along different lines, but we conjecture that the results would be comparable. We believe that the importance and interest of our results lie in their character rather than in the specific assumptions or techniques of proof.

3. Estimation of σ^2 and cross-validation. In the previous section we have considered properties of the bandwidth which minimizes (1.3), which entails knowing the error variance σ^2. In this section we consider minimizing (1.4), which uses an estimate of σ^2, and minimizing some other criteria, including cross-validation.

From (1.3) and (1.4) we may write

$$\tilde{R}_n(\lambda) = \hat{R}_n(\lambda) + \lambda(\hat{\sigma}^2 - \sigma^2)w(0)/n \tag{3.1}$$

where $\hat{\sigma}^2$ is estimated from the residuals of straight lines fit to successive triples of points or from successive differences. Under the assumption that $f \in C^2$ the variance of $\hat{\sigma}^2$ is of order n^{-1} and the bias is of order n^{-4} or n^{-2}, and is relatively negligible. Thus

$$\sup_{an^{1/5} \le \lambda \le bn^{1/5}} | n^{4/5}[\tilde{R}_n(\lambda) - \hat{R}_n(\lambda)] | \to 0 \tag{3.2}$$

from which it follows that the asymptotically efficient minimizer of $\hat{R}_n$ is asymptotically a minimizer of $\tilde{R}_n$.

A variety of other criteria used in model selection may be considered as well. These include:

1. Akaike's Information Criterion (Akaike, 1974)

$$\exp \mathrm{AIC}(\lambda) = \mathrm{RSS}(\lambda)\exp[2\lambda w(0)/n].$$

2. Generalized Cross Validation (Craven and Wahba, 1979)

$$\mathrm{GCV}(\lambda) = \frac{(1/n)\mathrm{RSS}(\lambda)}{(1 - \lambda w(0)/n)^2}.$$

3. Finite Prediction Error (Akaike, 1970):

$$\mathrm{FPE}(\lambda) = \frac{1 + \lambda w(0)/n}{1 - \lambda w(0)/n} \frac{1}{n} \mathrm{RSS}(\lambda).$$

4. A criterion mentioned by Shibata (1981):

$$S(\lambda) = (1/n)\mathrm{RSS}(\lambda)(1 + 2\lambda w(0)/n).$$

5. Another possible criterion is

$$T(\lambda) = \frac{(1/n)\mathrm{RSS}(\lambda)}{1 - 2\lambda w(0)/n},$$

which tends to guard against undersmoothing. $T(\lambda)$ may be expressed as

$$T(\lambda) = \frac{\hat{R}(\lambda) + \sigma^2 - 2\sigma^2\lambda w(0)/n}{1 - 2\lambda w(0)/n} = \frac{\hat{R}(\lambda)}{1 - 2\lambda w(0)/n} + \sigma^2.$$

The minimizer of T should thus be biased toward oversmoothing.

To see that these criteria are all asymptotically equivalent to $\hat{R}(\lambda)$ over the range $an^{1/5} \le \lambda \le bn^{1/5}$, consider first

$$\begin{aligned} S(\lambda) &= (\hat{R}_n(\lambda) + \sigma^2 - 2\lambda\sigma^2 w(0)/n)(1 + 2\lambda w(0)/n) \\ &= \hat{R}_n(\lambda) + \sigma^2 + \frac{2\lambda w(0)}{n}\hat{R}_n(\lambda) - 4\sigma^2\left(\frac{\lambda w(0)}{n}\right)^2. \end{aligned} \tag{3.3}$$

The second term does not depend on λ and the third and fourth terms are of smaller order than the first. The other four criteria may be expanded in Taylor series and in each case the leading term is $S(\lambda)$. As above we may conclude that each of these criteria has an asymptotically efficient minimizer.

The best known method for bandwidth choice is probably crossvalidation (CV). To define CV we use a regression estimate for unequally spaced data (Benedetti, 1977):

$$f_n(x) = \sum (x_j - x_{j-1})y_j\lambda w(\lambda(x - x_j)). \tag{3.4}$$

When this formula is used in (1.5) we find a closed form expression:

$$\begin{aligned} CV(\lambda) &= \frac{1}{n}\sum_{k=1}^{n}\left[y_k\left(1 + \frac{\lambda w(0)}{n}\right) - \frac{\lambda}{n}w\left(\frac{\lambda}{n}\right)y_{k+1} - f_n(x_k)\right]^2 \\ &= \frac{1}{n}\sum [y_k - f_n(x_k)]^2 + \frac{1}{n}\sum\left[\frac{\lambda w(0)}{n}y_k - \frac{\lambda}{n}w\left(\frac{\lambda}{n}\right)y_{k+1}\right]^2 \\ &\quad + \frac{2}{n}\sum [y_k - f_n(x_k)]\left[\frac{\lambda w(0)}{n}y_k - \frac{\lambda}{n}w\left(\frac{\lambda}{n}\right)y_{k+1}\right]. \end{aligned} \tag{3.5}$$

The first term is $(1/n)\mathrm{RSS}(\lambda)$. The second term is of smaller order. The principal contribution from the third term comes from

$$(2\lambda w(0)/n)(1/n)\sum (y_k - f_n(x_k))y_k, \tag{3.6}$$

the expectation of which is approximately $2\sigma^2\lambda w(0)/n$.

4. An example. We report here the results of a modest simulation. Kernel estimates of the function

$$f(x) = x^3(1 - x)^3 \tag{4.1}$$

were constructed from a sample of 75 equispaced points. Gaussian noise with $\sigma = .0015$ was added; this value of σ gives a signal-to-noise ratio of 10 at the peak, $x = \frac{1}{2}$. The kernel was

$$w(x) = \begin{cases} (15/8)(1 - 4x^2)^2 & |x| \leq \frac{1}{2} \\ 0 & |x| > \frac{1}{2}. \end{cases} \tag{4.2}$$

Computations were done in double precision on a VAX 11/780 computer using the IMSL routine GGMNL to generate the Gaussian deviates. For each of 100 runs, bandwidths were chosen to minimize the criteria of the previous section. The IMSL routine ZXLSF was used to find the minima.

Figure 1 shows the risk as a function of the bandwidth $b = \lambda^{-1}$. $w(0)/nb \simeq 1$ or $b \simeq .025$ corresponds to no smoothing.

Table 1 summarizes the results. Define $E(\lambda) = R(\lambda)/R(\lambda^*)$ where λ^* is the optimal value of λ. The table shows for various values of E the number of times out of 100 that E was exceeded by each estimator.

The estimators in the table are listed roughly in order of efficiency. T and CV were the most effective, followed by GCV and $\tilde{R}$. Successive differences were used to estimate σ^2 for $\tilde{R}$ and in this case $\tilde{R}$ can be written simply as

$$\tilde{R}(\lambda) = \frac{1}{n} \sum (y_k - f_n(x_k))^2 + \frac{\lambda w(0)}{n} \frac{1}{n} \sum (y_{k+1} - y_k)^2. \tag{4.3}$$

This was found to be more effective than using residuals from lines fit to successive triples of points. AIC, FPE, and S frequently undersmoothed and performed poorly. FPE and S in particular show a strange tendency to choose λ so that $\lambda w(0)/n \simeq 1$ which amounts to almost no smoothing at all. In fact, if histograms of the bandwidths chosen by FPE and S are examined, the distributions look much like a mixture of a continuous distribution and a discrete mass at $\lambda = n/w(0)$.

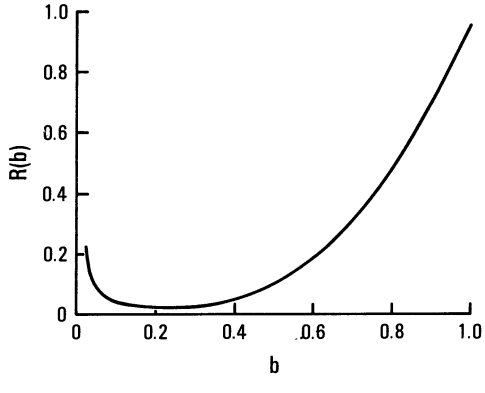

FIG. 1. *Risk as a function of bandwidth.*

TABLE 1

	1.05	1.1	1.2	1.4	1.6	1.8	2	4	6	8
T	28	17	4	1	0	0	0	0	0	0
CV	33	22	7	1	0	0	0	0	0	0
$\tilde{R}$	36	21	6	3	1	0	0	0	0	0
GCV	33	21	8	4	1	1	0	0	0	0
AIC	46	27	18	16	14	13	13	11	4	4
FPE	46	38	28	25	22	21	21	21	18	18
S	66	57	50	43	42	42	41	41	19	19

Thus, despite their asymptotic equivalence, there are real differences in the behavior of the estimators. Estimators that penalize undersmoothing heavily perform much better. In terms of mean square error, there appears to be much less danger of oversmoothing. These results are suggestive but hardly conclusive. Various functions, different error distributions, and different sample sizes should be tried. Unequally spaced data and noncircular models (with appropriate boundary modification) should be run as well. It would also be interesting to consider the effect of serially correlated errors. $\tilde{R}$, for example, is clearly sensitive to serial correlation.

Acknowledgements. I gratefully acknowledge the diligence of Charles Stone and two referees who caught errors in earlier versions of the manuscript.

REFERENCES

AKAIKE, H. (1970). Statistical predictor identification. *Ann. Inst. Statist. Math.* **22** 203–217.

AKAIKE, H. (1974). A new look at statistical model identification. *IEEE Trans. Auto. Cont.* **19** 716–723.

BENEDETTI, J. (1977). On the non-parametric estimation of regression functions. *J. Roy. Statist. Soc. B* **39** 248–253.

BREIMAN, L. and FREEDMAN, D. (1983). How many variables should be entered in a regression equation. *J. Amer. Statist. Assoc.* **78** 131–136.

CHOW, Y. S., GEMAN, S. and WU, L.-D. (1983). Consistent cross-validated density estimation. *Ann. Statist.* **11** 25–38.

COOLEY, J. W., LEWIS, P. A. W. and WELCH, P. D. (1977). The fast Fourier transform and its application to time series analysis. *Statistical Methods for Digital Computers* 377–423 (Einsein, Ralston, Wilf, eds.). Wiley, New York.

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

DOOB, J. L. (1953). *Stochastic Processes.* Wiley, New York.

GASSER, T., and MULLER, H.-G. (1979). Kernel estimation of regression functions. *Smoothing Techniques for Curve Estimation* 23–68. *Lecture Notes in Math.* **757.** Springer-Verlag, Berlin.

HALL, P. (1982). Large sample optimality of least squares cross-validation in density estimation. *Ann. Statist.* **11** 1156–1174.

LEONOV, V., and SHIRYAEV, A. (1959). On a method of calculation of semi-invariants. *Theor. Probab. Appl.* **4** 319–329.

LI, K.-C. (1983a). Cross validated nearest neighbor estimates. *Ann. Statist.* **12** 230–240.

LI, K.-C (1983b). From Stein's unbiased risk estimates to the method of generalized cross-validation. Unpublished manuscript.

MALLOWS, C. (1973). Some comments on C_p. *Technometrics* **15** 661–675.
RICE, J. (1984). Boundary modification for kernel regression. *Comm. Statist. A-Theory Methods* **13** 893–900.
ROSENBLATT, M. (1971). Curve estimates. *Ann. Math. Statist.* **42** 1815–1842.
SHIBATA, R. (1981). An optimal selection of regression variables. *Biometrika* **68** 45–54.
SPECKMAN, P. (1982). Spline smoothing and optimal rates of convergence in nonparametric regression models. Manuscript.
WONG, W. (1982). On the consistency of cross-validation in kernel nonparametric regression. *Ann. Statist.* **11** 1136–1141.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CALIFORNIA, SAN DIGEO
LA JOLLA, CALIFORNIA 92093