The Annals of Statistics
1990, Vol. 18, No. 4, 1696–1711

ON THE ASYMPTOTIC DISTRIBUTIONS OF BANDWIDTH ESTIMATES[1]

BY SHEAN-TSONG CHIU

Colorado State University

The problem of automatic bandwidth selection for a kernel regression estimate is studied. Since the bandwidth considerably affects the features of the estimated curve, it is important to understand the behavior of bandwidth selection procedures. The bandwidth estimate considered here is the minimizer of Mallows' criterion. Though it was established that the bandwidth estimate is asymptotically normal, it is well recognized that the rate of convergence is extremely slow. In simulation studies, it is often observed that the normal distribution does not provide a satisfactory approximation. In this paper, the bandwidth estimate is shown to be approximately equal to a constant plus a linear combination of independent exponential random variables. In practice, the distribution of a weighted sum of chi-squared random variables can be approximated by a multiple of a chi-squared distribution. Simulation results indicate that this provides a very good approximation even for a modest sample size. It is shown that the degrees of freedom of the chi-squared distribution goes to infinity rather slowly (at the rate $T^{1/5}$ for a nonnegative kernel). This explains why the distributions of the bandwidth estimates converge to the normal distribution so slowly.

1. Introduction.

Given data from the model

$$Y(t) = m(x_t) + \varepsilon(t), \qquad x_t = t/T, t = 0, \ldots, T-1,$$

where $\varepsilon(t)$ is a sequence of independent random variables with mean zero and variance σ^2, one is often interested in recovering the regression function $m(x)$. In practice, we usually do not have a proper parametric model and have to use nonparametric regression methods to obtain an estimate of $m(x)$, for example, see Priestley and Chao (1972) and Reinsch (1967). We consider a kernel estimator,

$$\hat{m}_\beta(x) = (T\beta)^{-1} \sum_{t=0}^{T-1} w\{(x_t - x)/\beta\} Y(t),$$

where β is the bandwidth. The kernel $w(x)$ is a symmetric probability density function. The estimate of $m(x)$ is a weighted average of the observations with x_t close to x. The kernel assigns the weights. To eliminate boundary effects, we use a "circular design," that is, the estimate is obtained by applying the kernel on the extended series $\hat{Y}(t)$, where $\hat{Y}(t + kT) = Y(t)$, $k = 0, \pm 1, \ldots$.

Received April 1988; revised March 1989.

[1]Research completed while the author was at Rice University and supported in part by ONR Grant N 00014-85-K-0100 and ARO Grant DAAG 29-85-K-0212.

AMS 1980 *subject classifications*. Primary 62G99; secondary 62M10, 62F12.

Key words and phrases. Nonparametric regression, kernel estimate, bandwidth selection, Fourier transform, periodogram.

An important step in nonparametric regression is to choose a proper smoothing parameter (bandwidth), which controls the smoothness of the resulting estimate. The smoothing parameter considerably affects the features of the estimated curve. The problem of automatic (data-driven) bandwidth selection has been studied extensively and many automatic selectors have been proposed [see Rice (1984), Härdle, Hall and Marron (1988) and references given therein]. Since the choice of the smoothing parameter is very crucial, it is important to understand the behavior of the automatic selectors.

Most bandwidth selectors attempt to choose β so as to minimize the risk function

$$R_T(\beta) = E\sum_{t=0}^{T-1}\{m(x_t) - \hat{m}_\beta(x_t)\}^2. \tag{1.1}$$

Rice (1984) showed that Mallows' criterion gives a weakly consistent estimate of the minimizer of $R_T(\beta)$ and the estimate is asymptotically normal. Similar results for other selectors were also established in Rice (1984) and Härdle, Hall and Marron (1988). Though the bandwidth estimates are asymptotically normal, it is well known in simulation studies that the normal distribution does not provide a satisfactory approximation [see Figures 2 and 4 and Härdle, Hall and Marron (1988)]. The disagreement between the asymptotic and empirical results suggests that one needs a more precise approximation. To obtain a better approximation is the main objective of this work. We show that the bandwidth estimate is approximately equal to a constant plus a weighted sum of independent exponential random variables. In practice, the distribution of a weighted sum of chi-squared random variables can be approximated by a multiple of a chi-squared distribution. An asymptotic distribution can be used to construct an approximate confidence interval which provides some suggestion for the range of bandwidths to choose from. The results here might be useful in comparing some selectors which, though asymptotically equivalent, perform quite differently for finite samples [Rice (1984) and Härdle, Hall and Marron (1988)].

2. Asymptotic distribution. The estimate $\hat{\beta}$ considered here is the minimizer of Mallows' criterion

$$\hat{R}_T(\beta) = \mathrm{RSS}_T(\beta) - T\sigma^2 + 2\sigma^2 w(0)/\beta, \tag{2.1}$$

where

$$\mathrm{RSS}_T(\beta) = \sum_{t=0}^{T-1}\{Y(t) - \hat{m}_\beta(x_t)\}^2$$

is the residual sum of squares and $\hat{R}_T(\beta)$ is an asymptotically unbiased estimate of $R_T(\beta)$ [cf. Mallows (1973), Craven and Wahba (1979) and Rice (1984)]. In practice σ^2 in (2.1) is replaced by an estimate $\hat{\sigma}^2$. Rice (1984) argued that the error caused by substituting a $\sqrt{T}$ consistent estimate for σ^2 is negligible. The true σ^2 will be used in the theory and in the simulations

except at the end of Section 3 where the effects of the replacement for finite samples are assessed by a Monte Carlo study.

Under the conditions given in Section 4, $R_T(\beta)$ is equal to

$$\beta^{-1}\sigma^2\int w^2(x)\,dx + 4^{-1}T\beta^4\left\{\int x^2w(x)\,dx\right\}^2\int\{m''(x)\}^2\,dx + O(T^{-1}\beta^{-2}) + o(T\beta^{9/2}).$$

Defining $A_T(\theta) = T^{-1/5}R_T(T^{-1/5}\theta)$, we see that $A_T(\theta)$ converges to

$$A(\theta) = \theta^{-1}\sigma^2\int w^2(x)\,dx + 4^{-1}\theta^4\left\{\int x^2w(x)\,dx\right\}^2\int\{m''(x)\}^2\,dx.$$

$A(\theta)$ has a unique minimum at θ_0, where

$$\theta_0^5 = \sigma^2\int w^2(x)\,dx\Big/\left[\left\{\int x^2w(x)\,dx\right\}^2\int\{m''(x)\}^2\,dx\right].$$

Rice (1984) showed that the estimate $\hat{\theta} = T^{1/5}\hat{\beta}$ converges to θ_0 in probability.

In Section 4, we show that the estimate $\hat{\beta}$ is approximately equal to a constant plus a linear combination of independent exponential random variables. This result suggests the approximate distribution described in Theorem 1.

THEOREM 1. *Under Assumptions* 1–3 *in Section* 4, *the distribution of* $T^{1/10}(\hat{\theta} - \theta_0)$ *can be approximated by the distribution of*

$$-T^{-3/10}\{A''(\theta_0)\}^{-1}Z(\beta_0),$$

where $\beta_0 = T^{-1/5}\theta_0$ *and*

$$Z(\beta_0) = 2\sigma^2\sum_{j=1}^{N}(X_j - 2)V_{\beta_0}(2\pi j/T),$$

where X_j, $j = 1,\ldots,N = [(T-1)/2]$, *are independent* χ_2^2 *random variables and* $V_\beta(\lambda)$ *is defined in* (4.8).

Here $[x]$ means the largest integer which is less than or equal to x. From Theorem 1, we get the following corollary.

COROLLARY 1. *Under the conditions in Theorem* 1, $T^{1/10}(\hat{\theta} - \theta_0)$ *is asymptotically normal with mean zero and variance* $8\{A''(\theta_0)\}^{-2}\theta_0^{-3}\sigma^4\int\{\int w(x-y)v(y)\,dy - v(x)\}^2\,dx$, *where* $-v(x) = w(x) + xw'(x)$.

Though $\hat{\theta}$ is asymptotically normal, the simulation results in the next section indicate that Corollary 1 is not very useful in practice.

REMARK 1. Corollary 1 is similar to Theorem 2.3 of Rice (1984). Härdle, Hall and Marron (1988) also gave a similar result. For an easy comparison

with the result of Härdle, Hall and Marron (1988), we list the corresponding notation: $n = T$; $\hat{h} = \hat{\beta}$; $K(x) = w(x)$; and $L(x) = w(x) + v(x)$ or $-v(x) = K(x) - L(x)$. h_0 is the minimizer of $R_T(h)$ which is approximately equal to β_0. Corollary 1 can be established by applying Lemma 4 of Härdle, Hall and Marron (1988) and noting that

$$n^{3/10}(\hat{h} - h_0) = n^{3/10}(\hat{h} - \hat{h}_0) + n^{3/10}(\hat{h}_0 - h_0),$$

where $\hat{h}_0$ is the minimizer of the sum of squared error.

To use Theorem 1, we need the distribution of $\Sigma X_j V_{\beta_0}(2\pi j/T)$. It might be difficult to obtain the distribution of a weighted sum of chi-squared random variables. For practical purposes, it has been suggested [see Brillinger (1981), page 145 and references given therein] to approximate the distribution by a multiple, $\eta\chi^2_\nu$, of a chi-squared distribution whose mean and degrees of freedom are determined by equating the first- and second-order moments. For our case, we set

$$\eta\nu = 2\sum_{j=1}^{N} V_{\beta_0}(2\pi j/T), \tag{2.2}$$

$$2\eta^2\nu = 4\sum_{j=1}^{N} V^2_{\beta_0}(2\pi j/T) \tag{2.3}$$

or

$$\nu = 2\left\{\sum_{j=1}^{N} V_{\beta_0}(2\pi j/T)\right\}^2 \Big/ \sum_{j=1}^{N} V^2_{\beta_0}(2\pi j/T), \tag{2.4}$$

which is approximately equal to

$$(4\beta_0)^{-1}\left\{2w(0) - \int w^2(x)\,dx\right\}^2 \Big/ \int\left\{\int w(x-y)v(y)\,dy - v(x)\right\}^2 dx. \tag{2.5}$$

The expression (2.5) indicates that the degrees of freedom goes to infinity rather slowly (at the rate $T^{1/5}$). This explains why the distribution of $\hat{\beta}$ converges to the normal distribution so slowly. For the example considered in the next section, one would need a sample size $T = 109{,}000$ in order to have an approximate χ^2_{25} distribution. It is also interesting to note that η and ν depend upon $m(x)$ only through β_0. Estimates of η and ν can be obtained by substituting $\hat{\beta}$ for β_0 in (2.2) and (2.3).

3. Simulation results. To evaluate the approximate distributions for finite sample sizes, we carried out some simulations and report the results here. The observations $Y(t)$, $t = 1, \ldots, T$, are obtained by adding independent Gaussian random variables with mean zero and variance $\sigma^2 = 0.003^2$ to the function

$$m(x_t) = x_t^3(1 - x_t)^3, \qquad x_t = t/T.$$

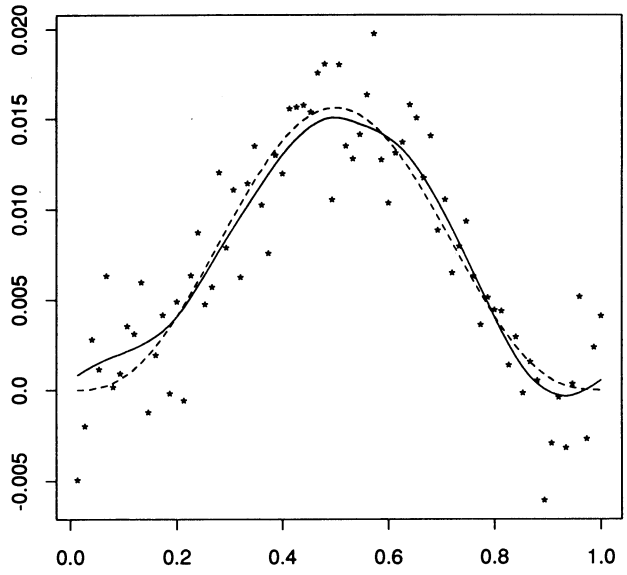

FIG. 1. *A realization of the simulation* ($T = 75$). *The dashed curve is* $m(x)$ *and the solid curve is the kernel estimate with bandwidth* 0.281.

The kernel is

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

The functions $m(x)$ and $w(x)$ were the ones considered in Rice (1984). The same functions were also used in Härdle, Hall and Marron (1988). All random variables were generated by the function RAND in Fortran 77 on a SUN 3 computer. Figure 1 shows a realization of the simulation ($T = 75$) and the regression functions. The bandwidth of the kernel estimate is selected by $\hat{R}_T(\beta)$. Simulations for several sample sizes between $T = 75$ and 300 were carried out. Since the results are quite similar, we only report the results for the sample sizes $T = 75$ and 300.

For each sample size, 2000 series were generated. It is well known that $\hat{R}_T(\beta)$ sometimes contains multiple local minima and most numerical methods might fail to find the global minimum. The global minimizer $\hat{\beta}$ is obtained by searching over 401 equally spaced points in the interval $[0.03, 0.45]$. The sample means and standard deviations of the estimates are summarized in Table 1. The empirical values agree well with the approximate ones. The values inside the parentheses are the estimated standard errors.

For the sample size $T = 75$, the approximate distributions are compared in Figures 2–4 and Tables 2 and 3. We obtained an estimated density of $\hat{\beta}$ by using the *S* function "density" with a Gaussian kernel and a width 0.04 (in

TABLE 1
Comparison of the sample means and standard deviations of the estimates $\hat{\beta}$, $\hat{\beta}(\hat{\sigma}^2)$ *and* $\hat{\beta}(\tilde{\sigma}^2)$

	$T = 75$				$T = 300$			
	Approx.	$\hat{\beta}$	$\hat{\beta}(\hat{\sigma}^2)$	$\hat{\beta}(\tilde{\sigma}^2)$	**Approx.**	$\hat{\beta}$	$\hat{\beta}(\hat{\sigma}^2)$	$\hat{\beta}(\tilde{\sigma}^2)$
Mean	0.298	0.291 (0.002)	0.286	0.291	0.226	0.217 (0.001)	0.214	0.219
SD	0.056	0.068 (0.001)	0.078	0.066	0.037	0.045 (0.0009)	0.049	0.043

The sample sizes are 2000. The values inside the parentheses are the estimated standard errors.

S, width/4 = standard deviation of the Gaussian kernel). The width is chosen subjectively. Figure 2 compares the estimated density (solid curve) with the approximate normal (dotted curve) and χ^2 (dashed curve) densities. The location and scale parameters of the approximate densities are the asymptotic ones given in Theorem 1 and Corollary 1. The degrees of freedom of the χ^2 distribution, obtained from (2.4), is 5.94. It can be seen that the χ^2 approximation is quite accurate. The quantile–quantile plot of the empirical distribution of $-\hat{\beta}$ against the χ^2 distribution is shown in Figure 3. Except at the tail, the chi-squared distribution provides an excellent approximation. Figure 4 shows the quantile–quantile plot against the standard normal distribution. It is clear that the normal distribution does not provide a satisfactory approximation.

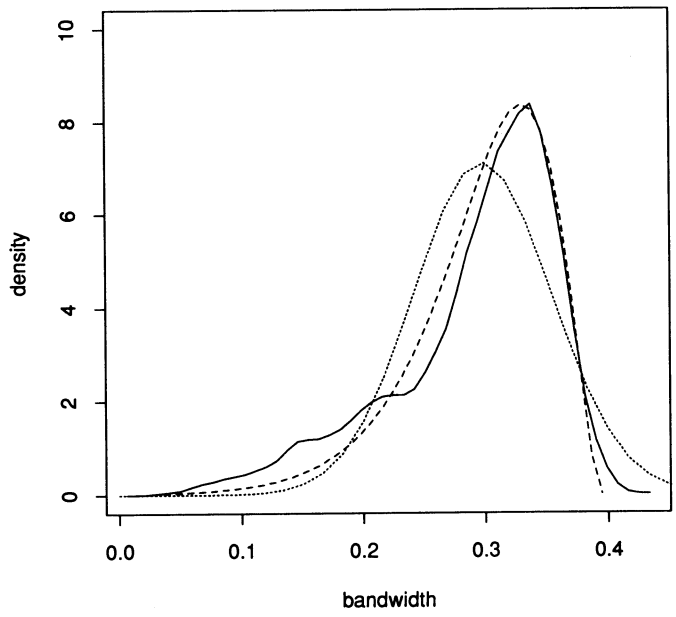

FIG. 2. *The estimated density (solid curve) of* $\hat{\beta}_{75}$ *and the approximate normal (dotted curve) and chi-squared (dashed curve) densities. The sample size is* 2000.

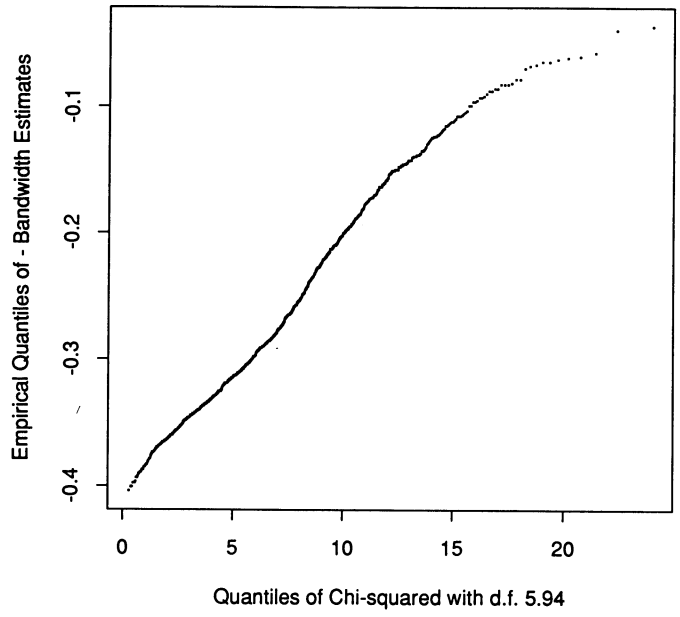

FIG. 3. *The quantile–quantile plot of* $2000 - \hat{\beta}_{75}$ *against* $\chi^2_{5.94}$ *distribution.*

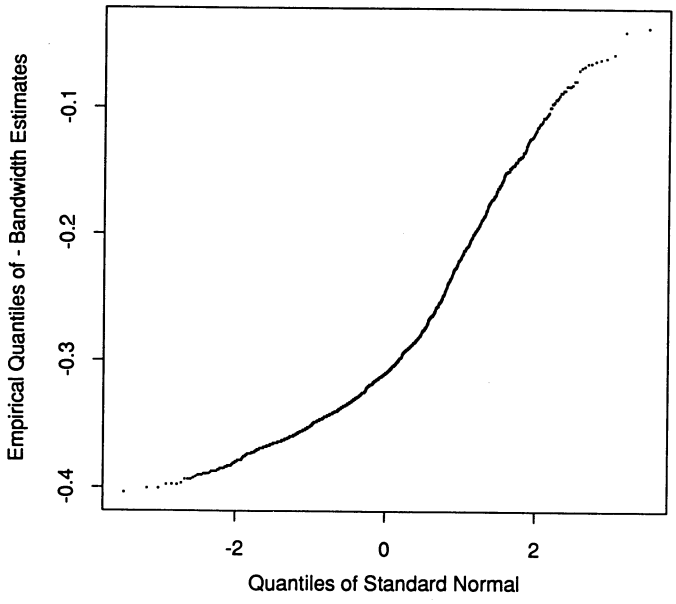

FIG. 4. *The quantile–quantile plot of* $2000 - \hat{\beta}_{75}$ *against the standard normal distribution.*

TABLE 2
The percentages of $\hat{\beta}_{75}$ which are smaller than the percentiles of the approximate $\chi^2_{5.94}$ and normal distributions

%	5	10	20	30	40	50	60	70	80	90	95
Chi-squared	7	13	21	28	37	45	57	68	81	92	97
Normal	13	17	23	28	34	43	52	64	80	95	99

Table 2 gives the percentages of $\hat{\beta}$ which are bigger than the percentiles of the approximate distributions. It is clear that the chi-squared distribution gives a better fit. We also use the Kolmogorov test statistics to compare the asymptotic distributions. Since any small difference can be detected as the sample size (of $\hat{\beta}$) increases, we feel that it is more proper to compare the distributions at a moderate sample size. The 2000 simulated estimates were divided into 10 samples of size 200 each. Table 3 gives the Kolmogorov distances to the best fitted Gaussian and chi-squared distributions for each sample. The p-values (observed significance levels) are also given in the table. All p-values against the chi-squared hypothesis are bigger than 0.15. On the contrary, only one of the p-values against the normal hypothesis is bigger than 0.005. Although the chi-squared distribution provides a good approximation, we should point out that $\hat{R}_T(\beta)$ gives small bandwidth estimates more often than predicted by Theorem 1. This can be seen from Figure 2. Table 1 also suggests that the estimate is biased toward undersmoothing.

In practice, σ^2 in (2.1) has to be replaced by an estimate. The estimate

$$\hat{\sigma}^2 = \sum_{t=1}^{T-1} \{Y(t) - Y(t-1)\}^2 \Big/ (2T-2) \tag{3.1}$$

was suggested in Rice (1984). We also consider a more efficient estimate $\tilde{\sigma}^2$, which is described in the Appendix. The sample means and variances of the variance estimates are compared in Table 4. The estimate $\tilde{\sigma}^2$ has a smaller variance. The estimates $\hat{\beta}(\hat{\sigma}^2)$ and $\hat{\beta}(\tilde{\sigma}^2)$ are obtained by substituting the corresponding variance estimates for σ^2 in (2.1). The sample means and the

TABLE 3
The Kolmogorov test statistics of $-\hat{\beta}_{75}$ against the $\chi^2_{5.94}$ and normal distributions and their corresponding p-values

	1	2	3	4	5	6	7	8	9	10
χ^2 distribution	0.046	0.078	0.054	0.057	0.054	0.073	0.070	0.051	0.070	0.069
p-value	0.79	0.18	0.59	0.54	0.61	0.23	0.28	0.68	0.28	0.29
Normal	0.113	0.156	0.131	0.132	0.131	0.148	0.127	0.124	0.143	0.145
p-value	0.012	*	*	*	*	*	*	*	*	*

The p-values represented by "*" are smaller than 0.005. The location and scale parameters are estimated from each of the 10 samples of size 200.

TABLE 4
The sample means and variances of $\hat{\sigma}^2$ and $\tilde{\sigma}^2$

	$T = 75$		$T = 300$	
	$\hat{\sigma}^2$	$\tilde{\sigma}^2$	$\hat{\sigma}^2$	$\tilde{\sigma}^2$
Mean ($\times 10^{-6}$)	9.08 (0.04)	8.93 (0.036)	8.98 (0.02)	9.15 (0.02)
Var ($\times 10^{-12}$)	3.39 (0.12)	2.55 (0.09)	0.816 (0.026)	0.655 (0.023)

The sample sizes are 2000. The values inside the parentheses are the estimated standard errors.

standard deviations of the estimates $\hat{\beta}(\hat{\sigma}^2)$ and $\hat{\beta}(\tilde{\sigma}^2)$ are also given in Table 1. For $T = 75$, the variance of $\hat{\beta}(\tilde{\sigma}^2)$ is about 30% smaller than the variance of $\hat{\beta}(\hat{\sigma}^2)$. The estimated densities of the estimates are plotted in Figure 5. The same width (0.04 for $T = 75$ and 0.03 for $T = 300$) was used in estimating the densities in each plot. For $T = 75$, the densities of $\hat{\beta}(\hat{\sigma}^2)$ and $\hat{\beta}(\tilde{\sigma}^2)$ have modes lower than the mode of $\hat{\beta}$. However, there are only small differences between the densities when $T = 300$. This confirms the argument of Rice (1984) that the effects of replacing σ^2 with estimated ones are asymptotically negligible.

4. Assumptions and proofs. We use the technique of Fourier analysis of time series to derive the asymptotic distribution of the bandwidth estimate. Most notation and terminology used here follows Brillinger (1981). In the following discussion, we let $S(t) = m(t/T)$ and $\hat{S}_\beta(t) = \hat{m}_\beta(t/T)$. The periodogram of the series $Y(t)$, $t = 0, \ldots, T-1$, is defined by

$$I_Y(\lambda) = |d_Y(\lambda)|^2/(2\pi T),$$

where

$$d_Y(\lambda) = \sum_{t=0}^{T-1} Y(t)\exp(-i\lambda t), \qquad -\infty < \lambda < \infty,$$

is the (finite) Fourier transform of the series $Y(t)$. The periodograms and Fourier transforms of the series $\varepsilon(t)$, $S(t)$ and $\hat{S}_\beta(t)$ are defined similarly. By Parseval's formula, we have

$$T\sum_{t=0}^{T-1}\{S(t) - \hat{S}_\beta(t)\}^2 = \sum_{j=0}^{T-1}|d_S(\lambda_j) - d_{\hat{S}}(\lambda_j)|^2, \tag{4.1}$$

where $\lambda_j = 2\pi j/T$, $j = 0, \ldots, T-1$, are the Fourier frequencies. Similarly,

$$T\,\mathrm{RSS}_T(\beta) = \sum_{j=0}^{T-1}|d_S(\lambda_j) + d_\varepsilon(\lambda_j) - d_{\hat{S}}(\lambda_j)|^2. \tag{4.2}$$

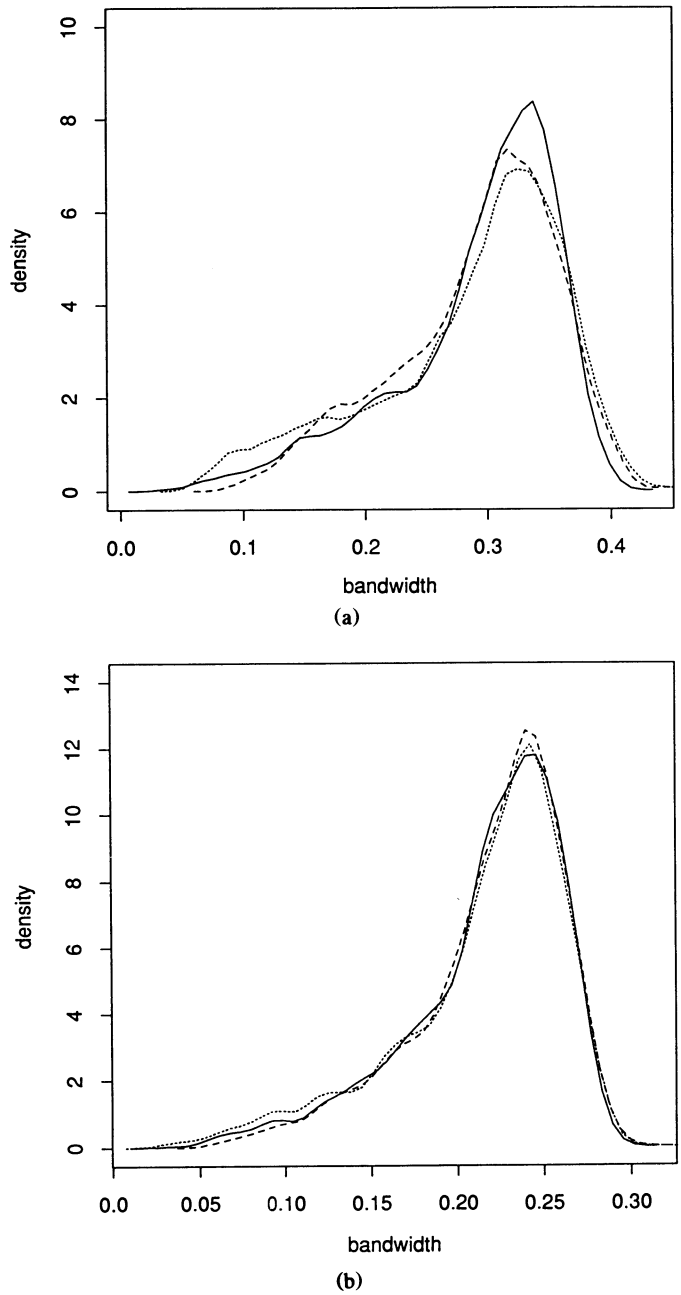

FIG. 5. *The estimated densities of the bandwidth estimates* $\hat{\beta}$ (*solid curves*), $\hat{\beta}(\hat{\sigma}^2)$ (*dotted curves*) *and* $\hat{\beta}(\tilde{\sigma}^2)$ (*dashed curves*) ($T = 75$ *in Figure* 5a *and* $T = 300$ *in Figure* 5b).

Since

$$\hat{S}_\beta(t) = (T\beta)^{-1} \sum_{u=0}^{T-1} w\{(u-t)/(T\beta)\}S(u) + (T\beta)^{-1} \sum_{u=0}^{T-1} w\{(u-t)/(T\beta)\}\varepsilon(u),$$

we have

$$d_{\hat{S}}(\lambda) = W_\beta(\lambda)\, d_S(\lambda) W_\beta(\lambda)\, d_\varepsilon(\lambda), \tag{4.3}$$

where

$$W_\beta(\lambda) = (T\beta)^{-1} \sum_{u=-T/2}^{T/2} w\{u/(T\beta)\}\exp(-i\lambda u)$$

is the transfer function of the filter $\{(T\beta)^{-1}w[u/(T\beta)]\}$. The function $W_\beta(\lambda)$ is real when $w(x)$ is symmetric. From (4.1) to (4.3), we get

$$\sum_{j=0}^{T-1} \left|d_S(\lambda_j) - d_{\hat{S}}(\lambda_j)\right|^2 = \sum_{j=0}^{T-1} \left|d_S(\lambda_j)\{1 - W_\beta(\lambda_j)\} - W_\beta(\lambda_j)\, d_\varepsilon(\lambda_j)\right|^2$$

and

$$\begin{aligned}\mathrm{RSS}_T(\beta) &= \frac{1}{T} \sum_{j=0}^{T-1} \left|d_S(\lambda_j) + d_\varepsilon(\lambda_j) - W_\beta(\lambda_j)\{d_S(\lambda_j) + d_\varepsilon(\lambda_j)\}\right|^2 \\ &= \frac{1}{T} \sum_{j=0}^{T-1} \left|d_Y(\lambda_j)\right|^2 \{1 - W_\beta(\lambda_j)\}^2.\end{aligned}$$

Noting that $W_\beta(0) = 1$ and the periodograms and $W_\beta(\lambda)$ are symmetric functions with period 2π, we see that, for odd T, (1.1) and (2.1) are equal to, respectively,

$$R_T(\beta) = 4\pi \sum_{j=1}^{N} \left[I_S(\lambda_j)\{1 - W_\beta(\lambda_j)\}^2 + (2\pi)^{-1}\sigma^2 W_\beta(\lambda_j)^2\right] + \sigma^2 \tag{4.4}$$

and

$$\hat{R}_T(\beta) = 4\pi \sum_{j=1}^{N} I_Y(\lambda_j)\{1 - W_\beta(\lambda_j)\}^2 - T\sigma^2 + 2\sigma^2 w(0)/\beta. \tag{4.5}$$

For even T, (4.4) and (4.5) drop the terms at frequency π, which have negligible effects. Let $G(\theta) = \hat{A}_T(\theta) - A_T(\theta)$, where $\hat{A}_T(\theta) = T^{-1/5}\hat{R}_T(T^{-1/5}\theta)$. By a Taylor series expansion, we have

$$-G'(\hat{\theta}) - \{A_T'(\hat{\theta}) - A'(\theta)\} = (\hat{\theta} - \theta_0)A''(\tilde{\theta}), \tag{4.6}$$

for some $\tilde{\theta}$ which lies in between θ_0 and $\hat{\theta}$. We need the following assumptions in deriving an approximate distribution of $G'(\theta)$.

ASSUMPTION 1. The noise $\varepsilon(t)$ is a sequence of independent random variables with mean zero, variance σ^2 and finite cumulants κ_k of all orders.

ASSUMPTION 2. The function $m(x)$ satisfies $m(0) = m(1)$, $m'(0) = m'(1)$ and the second-order derivative of $m(x)$ satisfies a Lipschitz condition of order $\alpha > \frac{1}{2}$.

Under Assumption 2, it can be shown that $|d_S(\lambda_j)| = O(T/j^3)$ for $j \le T$ [cf. Zygmund (1959), page 241].

ASSUMPTION 3. The kernel $w(x)$ is a symmetric probability density function with support on $[-\frac{1}{2}, \frac{1}{2}]$ and the second-order derivative of $w(x)$ is of bounded variation.

Letting $\tilde{W}(\beta j) = \int w(x)\exp(-i2\pi\beta jx)\,dx$, Assumption 3 implies that $\tilde{W}(\beta j) = O(\beta^{-3}j^{-3})$ and $(\partial/\partial\beta)\tilde{W}(\beta j) = O(\beta^{-3}j^{-2})$ [cf. 2.3.6 of Edwards (1979) and note that $w(\frac{1}{2}) = w'(\frac{1}{2}) = 0$]. Since

$$W_\beta(\lambda_j) = \sum_{n=-\infty}^{\infty} \tilde{W}\{\beta(nT + j)\},$$

we have, for some constant $M > 0$,

$$\left|W_\beta(\lambda_j) - \tilde{W}(\beta j)\right| \le M \sum_{n=1}^{\infty} \{\beta(nT + j)\}^{-3} = O(T^{-3}\beta^{-3}).$$

Therefore, $W_\beta(\lambda_j)$ can be replaced by $\tilde{W}(\beta j)$. Similarly, the derivatives $(\partial/\partial\beta)W_\beta(\lambda_j)$ in the following discussion can be replaced by $(\partial/\partial\beta)\tilde{W}(\beta j)$. Under these assumptions, the estimate $\hat{\theta}$ is consistent [Rice (1984)] and $A_T'(\hat{\theta}) - A'(\hat{\theta})$ is of order $T^{-4/5} + T^{-\alpha/5} = o(T^{-1/10})$.

We proceed to evaluate $G'(\theta)$. Write the difference between (4.4) and (4.5) as

$$\hat{R}_T(\beta) - R_T(\beta) = B_1 + B_2(\beta) + B_3(\beta) + B_4(\beta) + O(T^{-3}\beta^{-3}), \tag{4.7}$$

where

$$B_1 = 2\pi \sum_{j=0}^{T-1} \{I_\varepsilon(\lambda_j) - \sigma^2/(2\pi)\},$$

$$B_2(\beta) = 4\pi \sum_{j=1}^{N} \{I_\varepsilon(\lambda_j) - \sigma^2/(2\pi)\}W_\beta(\lambda_j)^2,$$

$$B_3(\beta) = -8\pi \sum_{j=1}^{N} \{I_\varepsilon(\lambda_j) - \sigma^2/(2\pi)\}W_\beta(\lambda_j)$$

and

$$B_4(\beta) = 4T^{-1}\,\mathrm{Re} \sum_{j=1}^{N} d_S(\lambda_j)\, d_\varepsilon(-\lambda_j)\{1 - W_\beta(\lambda_j)\}^2.$$

The error term in (4.7) is caused by dropping the terms at frequency π for even T. From (4.7), we see

$$T^{2/5}G'(T^{1/5}\beta) = B_2'(\beta) + B_3'(\beta) + B_4'(\beta) + O(T^{-2}\beta^{-3}).$$

Lemma 1 gives the asymptotic variance of $B_2'(\beta) + B_3'(\beta)$.

LEMMA 1. *Under Assumptions* 1 *and* 3 *and assuming* $T\beta \to \infty$ *as* $T \to \infty$, *the asymptotic variance of* $\beta^{3/2}\{B_2'(\beta) + B_3'(\beta)\}$ *is equal to* $8\sigma^4 \int\{\int w(x - y)v(y)\,dy - v(x)\}^2\,dx$, *where* $-v(x) = w(x) + xw'(x)$.

PROOF. Letting

$$V_\beta(\lambda) = -\{1 - W_\beta(\lambda)\}(\partial/\partial\beta)W_\beta(\lambda), \tag{4.8}$$

we have an important expression

$$B_2'(\beta) + B_3'(\beta) = 8\pi \sum_{j=1}^{N} \{I_\varepsilon(\lambda_j) - \sigma^2/(2\pi)\}V_\beta(\lambda_j). \tag{4.9}$$

Since

$$\operatorname{cum}\left\{\left|d_\varepsilon(\lambda_j)\right|^2, \left|d_\varepsilon(\lambda_k)\right|^2\right\} = \begin{cases} T\kappa_4 + T^2\sigma^4 & \text{if } j = k, \\ T\kappa_4 & \text{otherwise} \end{cases}$$

[cf. Brillinger (1981)], the variance of $B_2'(\beta) + B_3'(\beta)$ is equal to

$$16\sigma^4 \sum_{j=1}^{N} \left[\{1 - W_\beta(\lambda_j)\}(\partial/\partial\beta)W_\beta(\lambda_j)\right]^2$$

plus a negligible term. Define

$$a(u) = \beta^{-1}(T\beta)^{-2} \sum_t w\{(u - t)/(T\beta)\}v\{t/(T\beta)\} \tag{4.10}$$

and

$$b(u) = \beta^{-1}(T\beta)^{-1}v\{u/(T\beta)\}. \tag{4.11}$$

The filters $\{a(u)\}$ and $\{b(u)\}$ have the transfer functions $W_\beta(\lambda)(\partial/\partial\beta)W_\beta(\lambda)$ and $(\partial/\partial\beta)W_\beta(\lambda)$, respectively. By Parseval's formula, we get

$$\sum_{j=0}^{T-1} \left[\{1 - W_\beta(\lambda_j)\}(\partial/\partial\beta)W_\beta(\lambda_j)\right]^2 = T \sum_{t=-T/2}^{T/2} \{a(t) - b(t)\}^2 \tag{4.12}$$

when $\beta < \frac{1}{2}$. It can be seen that β^3 times (4.12) converges to $\int\{\int w(x - y)v(y)\,dy - v(x)\}^2\,dx$ and the proof is finished. □

A bound for the asymptotic variance of $B_4'(\beta)$ is given next.

LEMMA 2. *Under Assumptions* 1–3 *and assuming* $\beta = b_T c$, *where* $b_T = o(1)$ *and* $c > 0$ *is a constant, the variance of* $B_4'(\beta)$ *is of order* Tb_T^3.

PROOF. The variance of $B_4'(\beta)$ is equal to

$$16\pi\sigma^2 \sum_{j=0}^{N} I_S(\lambda_j)\left[(\partial/\partial\beta)\{1 - W_\beta(\lambda_j)\}^2\right]^2$$

plus a negligible term. Since $(\partial/\partial\beta)\{1 - W_\beta(\lambda_j)\}^2 = O(\beta^3 j^4)$ when $\beta j = O(1)$, we see

$$\sum_{j \le b_T^{-1}} I_S(\lambda_j)\left[(\partial/\partial\beta)\{1 - W_\beta(\lambda_j)\}^2\right]^2$$

is of order $T\beta^6 b_T^{-3} = O(Tb_T^3)$. Because $(\partial/\partial\beta)W_\beta(\lambda_j) = O(\beta^{-3}j^{-2})$, we also have

$$\sum_{j \ge b_T^{-1}} I_S(\lambda_j)\left[(\partial/\partial\beta)\{1 - W_\beta(\lambda_j)\}^2\right]^2 \le MT\beta^{-6} \sum_{j \ge b_T^{-1}} j^{-10},$$

for some constant $M > 0$. Noting that the right-hand side is of order Tb_T^3 finishes the proof. □

From Lemmas 1 and 2, we see that $B_4'(\beta)$ is negligible when $b_T T^{1/6} = o(1)$. For the example considered in Section 3, the variance of $B_4'(\beta_0)$ is about 4% of the variance of $B_2'(\beta_0) + B_3'(\beta_0)$ for the sample size $T = 75$.

For a Gaussian series $\varepsilon(t)$, $I_\varepsilon(\lambda_j)$, $j = 1, \ldots, N$, are true independent exponential random variables. Therefore $G'(\theta)$ is approximately equal to a constant plus a weighted sum of independent χ_2^2 random variables. As shown in Lemma 3, this is still true for a non-Gaussian series.

LEMMA 3. *Under the conditions in Lemma* 1, *the distribution of* $\{B_2'(\beta) + B_3'(\beta)\}$ *is approximately equal to the distribution of* $Z(\beta) = 2\sigma^2\Sigma(X_j - 2)V_\beta(\lambda_j)$, *where* X_j, $j = 1, \ldots, N$, *are independent* χ_2^2 *random variables.*

PROOF. We need to show that the cumulants of $\beta^{3/2}\{B_2'(\beta) + B_3'(\beta)\}$ and $\beta^{3/2}Z(\beta)$ have the same limits. The expected values are equal to zero. From Lemma 1, the variances are asymptotically equal when $\beta T \to \infty$ as $T \to \infty$. The kth ($k \ge 3$) cumulant of $\beta^{3/2}\{B_2'(\beta) + B_3'(\beta)\}$ is equal to the kth cumulant of $8\pi\beta^{3/2}\Sigma I_\varepsilon(\lambda_j)V_\beta(\lambda_j)$, which is

$$(4.13)\quad 4^k T^{-k}\beta^{3k/2} \sum_{j_1, \ldots, j_k} \operatorname{cum}\left\{\left|d_\varepsilon(\lambda_{j_1})\right|^2, \ldots, \left|d_\varepsilon(\lambda_{j_k})\right|^2\right\} \prod_{l=1}^{k} V_\beta(\lambda_{j_l}).$$

This cumulant is equal to

$$\sum_{\nu} \operatorname{cum}\{d_\varepsilon(\omega_{jh}); jh \in \nu_1\} \cdots \operatorname{cum}\{d_\varepsilon(\omega_{jh}); jh \in \nu_p\},$$

where $\omega_{jh} = (-1)^h\lambda_j$ and the summation is over all indecomposable partitions of the table

$$\begin{matrix} (1,1) & (1,2) \\ \vdots & \vdots \\ (k,1) & (k,2) \end{matrix}$$

[cf. Brillinger (1981), pages 20 and 21]. By applying Theorem 4.3.2 of Brillinger (1981) and noting that $\Sigma V_\beta^2(\lambda_j) = O(\beta^{-3})$ from Lemma 1, it can be shown that all partitions, except those described below, give values converging to zero. The partitions which might give nonvanishing contributions are those partitions $\nu = \nu_1 \cup \cdots \cup \nu_k$, each of whose members ν_l contains exactly two elements from different columns. The number of such partitions is $k!$ (4.13) is approximately equal to

$$(4.14) \qquad (4\sigma^2)^k \beta^{3k/2} k! \sum_{j=1}^{N} V_\beta^k(\lambda_j),$$

which is equal to the kth-order cumulant of $\beta^{3/2} Z(\beta)$. □

Theorem 1 in Section 2 follows directly from (4.6) and Lemmas 1 to 3. The Proof of Corollary 1 is given next.

PROOF OF COROLLARY 1. It is sufficient to show that (4.14) converges to zero for $k \geq 3$. Define $\tilde{a}(u) = a(u) - b(u)$, where $a(u)$ and $b(u)$ are given in (4.10) and (4.11), respectively. From the proof of Lemma 1, $\{\tilde{a}(u)\}$ has the transfer function $V_\beta(\lambda)$. Now,

$$(4.15) \qquad \left| \sum_{j=1}^{T-1} V_\beta^k(\lambda_j) \right| = T \left| \sum_n \sum_{t_1 + \cdots + t_k = nT} \tilde{a}(t_1) \cdots \tilde{a}(t_k) \right|.$$

Since there are at most $2T$ nonzero $\tilde{a}(u)$'s, we have a finite number of n's with nonzero summations. Hence, (4.15) is less than

$$(4.16) \qquad M\beta^{-2} \sum_{t_1} |\tilde{a}(t_1)| \cdots \sum_{t_{k-1}} |\tilde{a}(t_{k-1})|,$$

for some constant $M > 0$. Since (4.16) is of order $\beta^{-(k+1)}$, (4.14) is of order $\beta^{-1+k/2}$. The proof is finished by noting that $\beta^{-1+k/2} = o(1)$ when $\beta = o(1)$ and $k \geq 3$. □

We remark that Lemmas 1–3 still hold for the more general case, $m^{(k)}(x) \in L_2$ and $\int x^l w(x)\,dx = 0$, for $l = 1, \ldots, k-1$, and $\int x^k w(x)\,dx > 0$. Results similar to Theorem 1 and Corollary 1 can also be obtained. For the general case, the bandwidth β_0 is of order $T^{-1/(2k+1)}$ and the number of degrees of freedom given in (2.6) is of order $T^{1/(2k+1)}$.

APPENDIX

Estimation of noise variance. We consider the estimate

$$\tilde{\sigma}^2 = 2\pi(N - j_0 + 1)^{-1} \sum_{j=j_0}^{N} \frac{I_{\tilde{Y}}(\lambda_j)}{\left|1 - \exp(-i\lambda_j)\right|^4},$$

where $\tilde{Y}(t) = Y(t) - Y(t-1)$, $t = 1, \ldots, T-1$, is the differenced series of

$Y(t)$. In the simulation study, we set $j_0 = 5$. $\tilde{\sigma}^2$ is a special case of the estimate

$$\bar{\sigma}^2 = \frac{2\pi\Sigma_{j=1}^{N} I_{\tilde{Y}}(\lambda_j)\left|1 - \exp(-i\lambda_j)\right|^2 \phi(\lambda_j)}{\Sigma_{j=1}^{N}\left|1 - \exp(-i\lambda_j)\right|^4 \phi(\lambda_j)}, \tag{A.1}$$

where $\phi(\lambda)$ is a weighting function. The estimate $\hat{\sigma}^2$ given in (3.1) is also a special case of (A.1) with $\phi(\lambda) = |1 - \exp(-i\lambda_j)|^{-2}$. More details about the estimation procedure (including the choice of j_0) can be found in Chiu (1989). For a Gaussian series $\varepsilon(t)$, the asymptotic variances of $\hat{\sigma}^2$ and $\tilde{\sigma}^2$ are $1.5(2\sigma^4/T)$ and $\{N/(N - j_0 + 1)\}(2\sigma^4/T)$, respectively.

Acknowledgments. I gratefully thank Professor David Scott, the referees and the Associate Editor for their valuable comments, which substantially improved the presentation.

REFERENCES

BRILLINGER, D. R. (1981). *Time Series Data Analysis and Theory*. Holt, Rinehart, and Winston, New York.

CHIU, S. T. (1989). Bandwidth selection for kernel estimate with correlated noise. *Statist. Probab. Lett.* **8** 347–354.

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

EDWARDS, R. E. (1979). *Fourier Series: A Modern Introduction* **1**, 2nd ed. Springer, New York.

HÄRDLE, W., HALL, P. and MARRON, J. S. (1988). How far are automatically chosen regression smoothing parameters from their optimum? *J. Amer. Statist. Assoc.* **83** 86–101.

MALLOWS, C. (1973). Some comments on C_p. *Technometrics* **15** 661–675.

PRIESTLEY, M. B. and CHAO, M. T. (1972). Non-parametric function fitting. *J. Roy. Statist. Soc. B* **34** 385–392.

REINSCH, C. (1967). Smoothing by spline functions. *Numer. Math.* **10** 177–183.

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

ZYGMUND, A. (1959). *Trigonometric Series* **1**. Cambridge Univ. Press, Cambridge.

DEPARTMENT OF STATISTICS
COLORADO STATE UNIVERSITY
FORT COLLINS, COLORADO 80523