The Annals of Statistics
1991, Vol. 19, No. 4, 1883–1905

BANDWIDTH SELECTION FOR KERNEL DENSITY ESTIMATION[1]

BY SHEAN-TSONG CHIU

Colorado State University

The problem of automatic bandwidth selection for a kernel density estimator is considered. It is well recognized that the bandwidth estimate selected by the least squares cross-validation is subject to large sample variation. This difficulty limits the application of the cross-validation estimate. Based on characteristic functions, an important expression for the cross-validation bandwidth estimate is obtained. The expression clearly points out the source of variation. To stabilize the variation, a simple bandwidth selection procedure is proposed. It is shown that the stabilized bandwidth selector gives a strongly consistent estimate of the optimal bandwidth. Under commonly used smoothness conditions, the stabilized bandwidth estimate has a faster convergence rate than the convergence rate of the cross-validation estimate. For sufficiently smooth density functions, it is shown that the stabilized bandwidth estimate is asymptotically normal with a relative convergence rate $n^{-1/2}$ instead of the rate $n^{-1/10}$ of the cross-validation estimate. A plug-in estimate and an adjusted plug-in estimate are also proposed, and their asymptotic distributions are obtained. It is noted that the plug-in estimate is asymptotically efficient. The adjusted plug-in bandwidth estimate and the stabilized bandwidth estimate are shown to be asymptotically equivalent. The simulation results verify that the proposed procedures perform much better than the cross-validation for finite samples.

1. Introduction. Given a random sample $X_1, \ldots, X_n$ from a distribution with the density function $f(x)$, one is often interested in estimating $f(x)$. Silverman (1986) discussed many important applications of density estimates. The most commonly used nonparametric method is the kernel estimator

$$\hat{f}_\beta(x) = (n\beta)^{-1} \sum_{j=1}^{n} w\{(x - X_j)/\beta\}$$

[Rosenblatt (1956)], where the kernel function $w(x)$ is assumed to be a symmetric probability density function and β is the bandwidth. The bandwidth controls the smoothness of the resulting curve estimate. Selecting a proper β is a very critical step in estimating $f(x)$. Although in practice, one can choose the bandwidth subjectively, there is a great demand for automatic

Received April 1990; revised November 1990.
[1]Research supported in part by NSF Grant DMS-90-01734.
AMS 1980 *subject classifications*. Primary 62G99; secondary 62F10, 62E20.
Key words and phrases. Kernel density estimation, bandwidth selection, cross-validation, characteristic function, plug-in method.

(data-driven) bandwidth selection procedures. Some reasons of using automatic procedures were given in Silverman (1985).

The most studied automatic selector is the least squares cross-validation score function proposed by Rudemo (1982) and Bowman (1984). It was shown in Hall (1983) and Stone (1984) that the minimizer of the cross-validation score function is a consistent estimate of the optimal bandwidth, the minimizer of the mean integrated squared error. The asymptotic normality of the bandwidth estimate was established in Hall and Marron (1987a) and Scott and Terrell (1987). From the asymptotic results, it is well recognized that the bandwidth estimate is subject to large sample variation. In simulation studies, it is observed that the selector tends to choose smaller bandwidths more frequently than predicted by the asymptotic theorems; see Scott and Terrell (1987) and Section 7. A density estimate with a smaller bandwidth tends to show too many false features (structures) of the data. This difficulty is clearly demonstrated in Figure 1. The solid curve is the kernel estimate with the bandwidth selected by the cross-validation. More details about Figure 1 are given in Section 7. The difficulty limits the usefulness of the cross-validation in practice. Some studies about improving the cross-validation can be found in Scott and Terrell (1987), Park and Marron (1990) and Hall, Marron and Park (1989).

Since the only unknown quantity in the asymptotic mean integrated squared error is $\int\{f''(x)\}^2\,dx$ [Silverman (1986), page 40], another approach attempts to plug an estimate of $\int\{f''(x)\}^2\,dx$ into the approximation to get a bandwidth estimate. This approach was studied in Woodroofe (1970) and Scott and Factor

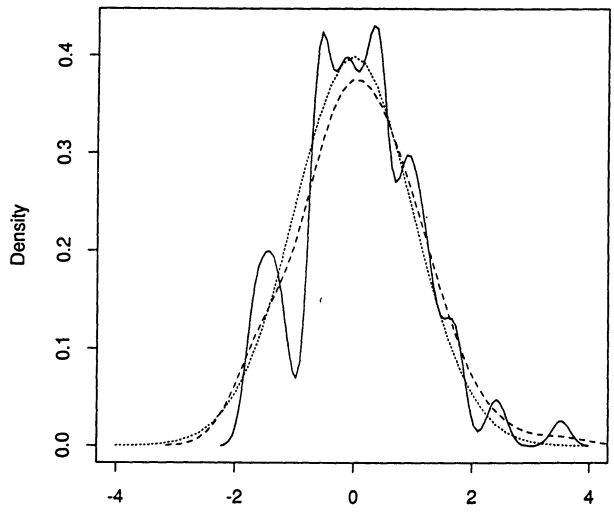

FIG. 1. *A simulated example with* $n = 100$. *The dotted curve is the standard normal density. The solid and dashed curves are, respectively, the density estimates with the bandwidths selected by the cross-validation* (0.156) *and the stabilized selector* (0.464).

(1981). The plug-in method has an apparent advantage that the method does not need an optimization program.

The main object of this work is to introduce several simple bandwidth selectors which give much more stable bandwidth estimates. In Section 2, we briefly discuss the automatic bandwidth selection problem and the cross-validation. Based on characteristic functions, an approximate expression for the cross-validation bandwidth estimate is obtained in Section 3. The expression clearly points out the source of variation. In Section 4, a stabilized bandwidth selector is proposed. Under commonly assumed smoothness conditions, the convergence rate of the stabilized bandwidth estimate is faster than the convergence rate of the cross-validation estimate. For sufficiently smooth f, we get a quite remarkable result that the relative convergence rate of the stabilized estimate is $n^{-1/2}$ instead of the rate $n^{-1/10}$ of the cross-validation estimate. It is also shown that the stabilized estimate is asymptotically normal and unbiased.

A plug-in estimate and an adjusted plug-in estimate are proposed in Sections 5 and 6. It is shown that the adjusted plug-in estimate is asymptotically equivalent to the stabilized estimate. It is noted that the asymptotic variances of the plug-in estimates attain the lower bound given in Bickel and Ritov (1988).

The simulation results in Section 7 agree well with the theoretic ones. The simulation study verifies that the proposed selectors perform much better than the cross-validation for finite samples. The procedures developed here could be useful in estimating the intensity function of a nonstationary Poisson process [Diggle and Marron (1988)].

2. The cross-validation. A commonly used measure of the performance of $\hat{f}_\beta(x)$ is the integrated squared error

$$\mathrm{ISE}_n(\beta) = \int \left\{ \hat{f}_\beta(x) - f(x) \right\}^2 dx$$

[Rosenblatt (1971)]. The integration above is over the whole real line. Unless indicated otherwise, this convention is used throughout the paper. Let $\mathrm{MISE}_n(\beta) = E\{\mathrm{ISE}_n(\beta)\}$ and $A_n(\theta) = n^{4/5}\,\mathrm{MISE}(n^{-1/5}\theta)$. Under the assumptions of Theorem 1 in Section 3 [also see Scott and Terrell (1987) and Hall and Marron (1987a) for some other smoothness conditions], $A_n(\theta)$ converges to

$$A(\theta) = \theta^{-1} \int w^2(x)\,dx + 4^{-1}\theta^4 \left\{ \int x^2 w(x)\,dx \right\}^2 \int \{ f''(x) \}^2 dx. \tag{2.1}$$

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

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

In the following discussion, $\beta_0 = n^{1/5}\theta_0$ and $\theta_{0n} = n^{1/5}\beta_{0n}$, where β_{0n} is the

minimizer of $\mathrm{MISE}_n(\beta)$. The optimal bandwidth θ_{0n} is approximately equal to θ_0.

A common approach in automatic bandwidth selection is to obtain an estimate of $\mathrm{MISE}_n(\beta)$ and to use the minimizer of the estimated risk function as an estimate of β_{0n}. Rudemo (1982) and Bowman (1984) proposed the least squares cross-validation

$$\mathrm{CV}_n(\beta) = \int \hat{f}_\beta^2(x)\,dx - 2n^{-1}\sum_{j=1}^{n} \hat{f}_{\beta,j}(X_j),$$

where $\hat{f}_{\beta,j}(x)$ is the kernel density estimate with the jth observation deleted from the sample. Up to a constant shift, $\mathrm{CV}_n(\beta)$ is an unbiased estimate of $\mathrm{MISE}_n(\beta)$ [Scott and Terrell (1987)]. The statistical properties of $\hat{\beta}_{\mathrm{CV}}$, the minimizer of $\mathrm{CV}_n(\beta)$, have been studied extensively; see Hall (1983), Stone (1984), Hall and Marron (1987a), Scott and Terrell (1987) and references given therein. It was established that $\hat{\theta}_{\mathrm{CV}} = n^{1/5}\hat{\beta}_{\mathrm{CV}}$ converges to θ_0 in probability and that $n^{1/10}(\hat{\theta}_{\mathrm{CV}} - \theta_0)$ is asymptotically normal. Note that the convergence rate of $\hat{\theta}_{\mathrm{CV}}$ is extremely slow.

3. The source of variation. Based on characteristic functions, Theorem 1 below gives an important expression for $\hat{\beta}_{\mathrm{CV}}$. The expression clearly points out the source of variation and leads to the consideration of the proposed procedures. The sample characteristic function is defined as

$$\tilde{\phi}(\lambda) = n^{-1}\sum_{j=1}^{n} \exp(i\lambda X_j), \qquad -\infty < \lambda < \infty.$$

Also let $\phi(\lambda) = \int \exp(i\lambda x) f(x)\,dx$ denote the characteristic function of $f(x)$. To make the discussion easier, we borrow the terminology "frequency" from time series analysis for λ. Using the methods in Brillinger (1981), pages 19–21, the cumulants of $\tilde{\phi}(\lambda)$ can be obtained by straightforward computation.

LEMMA 1. *Suppose $X_1, \ldots, X_n$ is a random sample from the distribution $F(x)$ and let $\phi(x)$ be the characteristic function of $F(x)$, then*

$$E\tilde{\phi}(\lambda) = \phi(\lambda),$$

$$\mathrm{cum}\{\tilde{\phi}(\lambda_1), \tilde{\phi}(\lambda_2)\} = \{\phi(\lambda_1 + \lambda_2) - \phi(\lambda_1)\phi(\lambda_2)\}/n$$

and

$$\mathrm{cum}\{\tilde{\phi}(\lambda_1), \ldots, \tilde{\phi}(\lambda_k)\} = O(n^{-k+1}).$$

From Lemma 1, $\tilde{\phi}(\lambda)$ is approximately complex normal [see Brillinger (1981) for the definition], and so $|\tilde{\phi}(\lambda) - \phi(\lambda)|^2$ is, approximately, an exponential random variable with mean $\{1 - |\phi(\lambda)|^2\}/n$.

Applying Parseval's formula yields

$$\mathrm{ISE}_n(\beta) = (2\pi)^{-1}\int \left|\hat{\phi}_\beta(\lambda) - \phi(\lambda)\right|^2 d\lambda$$

[Rudin (1974), page 202], where $\hat{\phi}_\beta(\lambda)$ is the characteristic function of the estimated density $\hat{f}_\beta(x)$. Letting $W(\lambda) = \int \exp(i\lambda x)w(x)\,dx$ and $\tilde{\phi}_d(\lambda) = \tilde{\phi}(\lambda) - \phi(\lambda)$ and noting that $\hat{\phi}_\beta(\lambda) = \tilde{\phi}(\lambda)W(\beta\lambda)$, we have

$$\left|\hat{\phi}_\beta(\lambda) - \phi(\lambda)\right|^2 = |\phi(\lambda)|^2\{1 - W(\beta\lambda)\}^2 + W^2(\beta\lambda)\left|\tilde{\phi}_d(\lambda)\right|^2 - 2\,\mathrm{Re}\left\{\phi(\lambda)\tilde{\phi}_d(-\lambda)\right\}\{1 - W(\beta\lambda)\}W(\beta\lambda).$$

Since $E\{\tilde{\phi}_d(\lambda)\} = 0$ and $E\{|\tilde{\phi}_d(\lambda)|^2\} = n^{-1}\{1 - |\phi(\lambda)|^2\}$,

$$\pi\,\mathrm{MISE}_n(\beta) = \int_0^\infty |\phi(\lambda)|^2\{1 - W(\beta\lambda)\}^2\,d\lambda + n^{-1}\int_0^\infty W^2(\beta\lambda)\left\{1 - |\phi(\lambda)|^2\right\}d\lambda. \tag{3.1}$$

Silverman (1986), pages 62–63, showed that $\pi\,\mathrm{CV}_n(\beta)$ is approximately equal to

$$\pi\,\widetilde{\mathrm{CV}}_n(\beta) = \int_0^\infty \left|\tilde{\phi}(\lambda)\right|^2\{W^2(\beta\lambda) - 2W(\beta\lambda)\}\,d\lambda + 2\pi w(0)/(n\beta). \tag{3.2}$$

It is interesting to compare (3.2) with Mallows' criterion for nonparametric regression in Rice (1984). Comparing (3.1) and (3.2) yields

$$\pi\,\widetilde{\mathrm{CV}}_n(\beta) - \pi\,\mathrm{MISE}_n(\beta) = B_1 + B_2(\beta) + B_3(\beta) + B_4(\beta),$$

where

$$B_1 = -\int_0^\infty |\phi(\lambda)|^2\,d\lambda,$$

$$B_2(\beta) = \int_0^\infty \left[\left|\tilde{\phi}_d(\lambda)\right|^2 - \left\{1 - |\phi(\lambda)|^2\right\}/n\right]\{W^2(\beta\lambda) - 2W(\beta\lambda)\}\,d\lambda,$$

$$B_3(\beta) = 2\,\mathrm{Re}\int_0^\infty \phi(\lambda)\tilde{\phi}_d(-\lambda)\{W^2(\beta\lambda) - 2W(\beta\lambda)\}\,d\lambda$$

and

$$B_4(\beta) = 2n^{-1}\int_0^\infty W(\beta\lambda)|\phi(\lambda)|^2\,d\lambda.$$

Theorem 1 shows that the behavior of $\hat{\beta}_{\mathrm{CV}}$ is dominated by $B_2'(\beta_{0n})$. The proofs of this and other results are given in Section 8. The assumptions are also described in that section.

THEOREM 1. *Under Assumptions* 1 *to* 3 *given in Section* 8 *with* $K_1 \geq 6$,

$$\left(\hat{\beta}_{\mathrm{CV}} - \beta_{0n}\right) = -B_2'(\beta_{0n})/\{\pi\,\mathrm{MISE}_n''(\beta_{0n})\} + o_p(n^{-3/10}).$$

From the definition of $B_2(\beta)$, we see that

$$B_2'(\beta) = \int_0^\infty \left[\left| \tilde{\phi}_d(\lambda) \right|^2 - \left\{ 1 - |\phi(\lambda)|^2 \right\}/n \right] \beta^{-1} V(\beta\lambda)\, d\lambda,$$

where

$$V(\lambda) = -2\{1 - W(\lambda)\} W'(\lambda)\lambda.$$

Therefore $\hat{\beta}_{\mathrm{CV}}$ is approximately equal to a constant plus a weighted integral of the process $|\tilde{\phi}_d(\lambda)|^2$. A closer look at $V(\lambda)$ reveals that $V(\beta_{0n}\lambda)$ has significant amplitude at $\lambda = \mathrm{O}(n^{1/5})$ and so $|\tilde{\phi}_d(\lambda)|^2$ at $\lambda = O(n^{1/5})$ make the major contribution to the variation of $B_2'(\beta_{0n})$. However, for smooth $f(x)$, $\phi(\lambda)$ at $\lambda = O(n^{1/5})$ has negligible effects on $\tilde{\phi}(\lambda)$ [relative to the noise level of $\tilde{\phi}(\lambda)$]. This observation suggests that we can reduce the variation in $\hat{\beta}_{\mathrm{CV}}$ by modifying these $|\tilde{\phi}(\lambda)|^2$ which do not contain much information about $f(x)$.

4. The stabilized bandwidth selector. In this section, we propose a bandwidth selection procedure and describe its asymptotic properties. We first find the random variable Λ which is the first λ such that $|\tilde{\phi}(\lambda)|^2 \le c/n$ for some constant $c > 1$. As shown in Theorem 4 and the simulation studies in Section 7, the choice of c is not important when f is sufficiently smooth. Some comments about the choice of c are given in Section 7. To reduce the variation, $|\tilde{\phi}(\lambda)|^2$ at $\lambda > \Lambda$ in (3.2) is substituted by $1/n$ and we get the stabilized bandwidth estimate by minimizing

$$\begin{aligned} S_n(\beta) = {} & \int_0^\Lambda \left| \tilde{\phi}(\lambda) \right|^2 \{W^2(\beta\lambda) - 2W(\beta\lambda)\}\, d\lambda \\ & + n^{-1} \int_\Lambda^\infty \{W^2(\beta\lambda) - 2W(\beta\lambda)\}\, d\lambda + 2\pi w(0)/(n\beta). \end{aligned} \tag{4.1}$$

The criterion $S_n(\beta)$ has an equivalent representation

$$\begin{aligned} S_n(\beta) = {} & \pi(n\beta)^{-1} \int w^2(x)\, dx \\ & + \int_0^\Lambda \left\{ \left| \tilde{\phi}(\lambda) \right|^2 - n^{-1} \right\} \{W^2(\beta\lambda) - 2W(\beta\lambda)\}\, d\lambda, \end{aligned} \tag{4.2}$$

which is much easier and faster to evaluate than (4.1). Comparing (4.2) with (2.1) or (3.1), we see that $S_n(\beta)$ use the second part of (4.2) to estimate the bias term in $\pi\,\mathrm{MISE}_n(\beta)$.

Before describing the theoretic properties of the estimate, we give the motivation and some remarks about the procedure. As noted in the previous section, the noise in the cross-validation estimate is mainly contributed by $|\tilde{\phi}(\lambda)|^2$ at high frequency, which does not contain much information about $f(x)$. In order to reduce the variation, we have to identify the part which

contains most information about $f(x)$. We note that when $|\phi(\lambda)|$ is negligible, $|\tilde{\phi}(\lambda)|^2 \approx |\tilde{\phi}_d(\lambda)|^2$, which is approximately exponentially distributed with mean $1/n$. Therefore we compare $n|\tilde{\phi}(\lambda)|^2$ with a critical value, say $3 \approx -\log_e(0.05)$, to decide when $|\phi(\lambda)|^2$ becomes negligible.

In the statements of the theorems below, let $\theta_{0n} = n^{1/5}\beta_{0n}$ and $\hat{\theta}_S = n^{1/5}\hat{\beta}_S$, where $\hat{\beta}_S$ is the minimizer of $S_n(\beta)$. Theorem 2 establishes the strong consistency of $\hat{\theta}_S$.

THEOREM 2. *Under Assumptions* 1 *to* 3 *given in Section* 8 *with* $K_1 > 5$, $\hat{\theta}_S$ *converges to* θ_0 *almost surely.*

Under assumptions a little bit stronger than the conditions of Theorem 1, Theorem 3 indicates that $\hat{\theta}_S$ has a faster convergence rate than the convergence rate of $\hat{\theta}_{\mathrm{CV}}$.

THEOREM 3. *Under Assumptions* 1 *to* 3 *with* $K_1 \geq 6$ *and* $1/10 < (K_1 - 5)/K_2$, *then* $(\hat{\theta}_S - \theta_{0n}) = o_p(\hat{\theta}_{\mathrm{CV}} - \theta_{0n})$.

When f is sufficiently smooth, Theorem 4 shows that $\hat{\theta}_S$ is asymptotically normal and that the convergence rate is $n^{-1/2}$, which is much faster than the rate $n^{-1/10}$ of $\hat{\theta}_{\mathrm{CV}}$. Hall and Marron (1990) showed that the convergence rates of the estimates of θ_{0n} have a lower bound $n^{-1/2}$.

THEOREM 4. *Under Assumptions* 1 *to* 3 *with* $10 < K_1 \leq K_2 < 2K_1 - 10$, $n^{1/2}(\hat{\theta}_S - \theta_{0n})$ *is asymptotically normal with mean* 0 *and variance*

$$\sigma_\theta^2 = 4\theta_0^6\{A''(\theta_0)\}^{-2}\left\{\int x^2 w(x)\,dx\right\}^4 \times\left[\int\{f^{(4)}(x)\}^2 f(x)\,dx - \left\{\int f^{(4)}(x) f(x)\,dx\right\}^2\right].$$

REMARK 1. Following the arguments of the proofs in Section 8, similar results can be established under some other smoothness conditions of f. For example, Theorem 4 holds for Gaussian and Cauchy distributions.

REMARK 2. Let $\tilde{\phi}_s(\lambda)$ be the sample characteristic function of the data $sX_1, \ldots, sX_n$ for some constant $s > 0$. Also let Λ_s and $S_s(\beta)$ be the frequency Λ and the stabilized criterion for the scale transformed data. Since $\tilde{\phi}_s(\lambda) = \tilde{\phi}(s\lambda)$, $\Lambda_s = s\Lambda$ and $S_s(\beta) = S_n(\beta/s)/s$. Therefore the stabilized bandwidth estimate is scale equivariant, and there is no need to rescale the data or to adjust c when the scale is changed.

REMARK 3. The smoothed cross-validation, proposed in Hall, Marron and Park (1989), can be written as

$$\mathrm{SCV}(\beta) = (n\beta)^{-1}\int w^2(x)\,dx + \pi^{-1}\int_0^\infty \left\{\left|\tilde{\phi}(\lambda)\right|^2 - n^{-1}\right\}\{1 - W(\beta\lambda)\}^2 U^2(\alpha\lambda)\,d\lambda,$$

where $U(\lambda)$ is the characteristic function of a kernel $u(x)$, and α is the bandwidth for the kernel $u(x)$. They showed that one can obtain a $\sqrt{n}$ bandwidth estimate when $u(x)$ is a sixth-order kernel. The best choice of α depends on the unknown density, and is unavailable in practice. They suggested replacing the unknown bandwidth with the best α for some "reference density." The approach of using a reference density is also suggested in Park and Marron (1990), Hall, Sheather, Jones and Marron (1991) and Jones, Marron and Park (1991). Using the standard normal density as the reference density is recommended in these papers. In order to make the procedure scale equivariant, the data are rescaled according to some (estimated) measure of the scale, such as the standard deviation or the interquartile range.

REMARK 4. The indicator function $1_{[-\Lambda,\Lambda]}(\lambda)$ can be viewed as the Fourier transform of an infinite-order kernel with the bandwidth proportional to $1/\Lambda$. In fact, $1_{[-\Lambda,\Lambda]}(\lambda)$ is the transfer function of an ideal low-pass filter; see Brillinger (1981) for more discussion. The proposed estimation procedure effectively solves the bandwidth selection problem in estimating the bias term in $\mathrm{MISE}(\beta)$.

REMARK 5. Jones, Marron and Park (1991) considered the criterion

$$(4.3)\quad \pi(n\beta)^{-1}\int w^2(x)\,dx + \int_0^\infty \left|\tilde{\phi}(\lambda)\right|^2\{1 - W(\beta\lambda)\}^2 W^2\{\alpha(\beta)\lambda\}\,d\lambda,$$

where $\alpha(\beta) = C_{\mathrm{JMP}} n^{-23/45}\beta^{-2}$ [$= O(n^{-1/9})$ when $\beta = O(n^{-1/5})$]. Note that $E\{|\tilde{\phi}_d(\lambda)|^2\} \approx 1/n$ is not subtracted from $|\tilde{\phi}(\lambda)|^2$ in the second term of (4.3). They showed that with a proper C_{JMP} [depending on $f(x)$], the minimizer of (4.3) is $\sqrt{n}$ consistent and is asymptotically unbiased. The nice theoretic property follows the fact that asymptotically, the leading term of the bias

$$\int_0^\infty |\phi(\lambda)|^2\{1 - W(\beta\lambda)\}^2\left[1 - W^2\{\alpha(\beta)\lambda\}\right]d\lambda + n^{-1}\int_0^\infty \{1 - W(\beta\lambda)\}^2 W^2\{\alpha(\beta)\lambda\}\,d\lambda$$

does not depend on β. Let $\hat{\beta}_{\mathrm{JMP}}$ be the minimizer of (4.3) and $\hat{\theta}_{\mathrm{JMP}} = n^{1/5}\hat{\beta}_{\mathrm{JMP}}$. Following the arguments for Theorems 1 to 4, it can be seen that $\hat{\theta}_{\mathrm{JMP}} \approx \hat{\theta}_S + D$, where $\mathrm{Var}(D) = O(n^{-1})$ and D is asymptotically uncorrelated with $\hat{\theta}$. Therefore $\hat{\theta}_S$ uniformly dominates $\hat{\theta}_{\mathrm{JMP}}$. The extra variation comes from a

term analogous to $D_3(\beta)$ in (8.3), which is negligible only when the bandwidth $\alpha = o(n^{-1/9})$.

5. The proposed plug-in estimate. Note that $G = \int\{f''(x)\}^2\,dx$ is the only unknown quantity in the right-hand side of (2.1). The observation leads to the consideration of the plug-in method, which obtains a bandwidth estimate by replacing G in (2.1) with an estimate of G. This approach was considered in Woodroofe (1970), Sheather (1986) and Scott and Factor (1981). The problem of estimating G was studied in Hall and Marron (1987b) and Bickel and Ritov (1988).

Since

$$G = \int\{f''(x)\}^2\,dx = (2\pi)^{-1}\int\lambda^4|\phi(\lambda)|^2\,d\lambda$$

and $E|\tilde{\phi}(\lambda)|^2 = n^{-1}(n-1)|\phi(\lambda)|^2 + n^{-1}$, one might attempt to estimate $\int\{f''(x)\}^2\,dx$ by

$$(2\pi)^{-1}\int\lambda^4\left\{\left|\tilde{\phi}(\lambda)\right|^2 - 1/n\right\}d\lambda. \tag{5.1}$$

However, the integral above does not exist since $\operatorname{var}\{|\tilde{\phi}(\lambda)|^2\} \approx n^{-2}$ at high frequency. The difficulty here is that (5.1) includes too much $\tilde{\phi}(\lambda)$ which is dominated by the sample variation.

As suggested in Sections 3 and 4, we should modify $\tilde{\phi}(\lambda)$ at high frequency when estimating functionals of $\phi(\lambda)$. Following the idea, we propose the estimate

$$\hat{G} = \pi^{-1}\int_0^{\Lambda}\lambda^4\left\{\left|\tilde{\phi}(\lambda)\right|^2 - 1/n\right\}d\lambda, \tag{5.2}$$

where Λ is defined in Section 4. The asymptotic properties of $\hat{G}$ are given in Theorems 5 and 6.

THEOREM 5. *Under Assumptions* 1 *and* 2 *with* $K_1 > 5$, $\hat{G}$ *converges to* G *almost surely.*

THEOREM 6. *Under Assumptions* 1 *and* 2 *with* $10 < K_1 \le K_2 < 2K_1 - 10$, $n^{1/2}(\hat{G} - G)$ *is asymptotically normal with mean* 0 *and variance* $\sigma_G^2 = 4[\int\{f^{(4)}(x)\}^2 f(x)\,dx - \{\int f^{(4)}(x)f(x)\,dx\}^2]$.

From Theorems 5 and 6, we get the following corollaries about the asymptotic properties of the plug-in bandwidth estimate

$$\hat{\theta}_P = \left\{\int w^2(x)\,dx\right\}^{1/5}\left\{\int x^2 w(x)\,dx\right\}^{-2/5}\hat{G}^{-1/5}.$$

COROLLARY 1. *Under the assumption of Theorem* 6, $\hat{\theta}_P$ *converges to* θ_0 *almost surely.*

COROLLARY 2. *Under the assumption of Theorem 7, $n^{1/2}(\hat{\theta}_P - \theta_0)$ is asymptotically normal with mean 0 and variance σ_θ^2 given in Theorem 4.*

REMARK 6. Since $\hat{G}$ does not depend on $w(x)$, Theorems 5 and 6 and Corollaries 1 and 2 do not need any assumption about $w(x)$, except that $\int w^2(x)\,dx < \infty$ and $0 < \int x^2 w(x)\,dx < \infty$. Corollary 2 might be useful in selecting the bandwidth for less smooth $w(x)$, for example, the rectangle kernel or the Epanechnikov kernel [Epanechnikov (1969)]. However, there is no guarantee that $n^{1/5}\beta_{0n}$ will converge to θ_0 without proper smoothness assumptions.

REMARK 7. Although the idea of the plug-in method existed for a long time, the problem of selecting proper β and kernel in estimating G deters people from using the plug-in method.

REMARK 8. Bickel and Ritov (1988) gave an information bound for nonparametric estimates of G. The asymptotic variance of the proposed estimate attains the bound and thus is asymptotically efficient. Based on the results, we conjecture that σ_θ^2 is the lower bound for the asymptotic variance of any nonparametric bandwidth estimate.

REMARK 9. The plug-in method discussed above is different from the conventional plug-in methods referred to in Scott and Terrell (1987) and Park and Marron (1990). Let $\hat{G}_\alpha = \pi^{-1}\int_0^\infty \lambda^4\{|\tilde{\phi}(\lambda)|^2 - 1/n\}W^2(\alpha\lambda)\,d\lambda$; Scott and Terrell (1987) proposed the bandwidth estimate, which minimizes the biased cross-validation

$$\mathrm{BCV}(\beta) = 4^{-1}\beta^4\left\{\int x^2 w(x)\,dx\right\}^2 \hat{G}_\beta + \int w^2(x)\,dx/(n\beta).$$

The bandwidth estimate proposed in Park and Marron (1990) is the root of

$$\beta^5 = \frac{n^{-1}\int w^2(x)\,dx}{\left\{\int x^2 w(x)\,dx\right\}^2 \hat{G}_\alpha},$$

where $\alpha(\beta) = C_{\mathrm{PM}} s^{3/13}\beta^{10/13}$ and s is an estimate of some measure of the scale of $F(x)$. The constant C_{PM} depends on the unknown f, and the approach of using a reference density was suggested. An interesting modification of Park and Marron (1990) was given in Sheather and Jones (1990).

6. Adjusted plug-in estimate. As pointed out in Hall, Sheather, Jones and Marron (1991), $\theta_0 - \theta_{0n} = O(n^{-2/5})$ and so $\hat{\theta}_P$ is not a $\sqrt{n}$ consistent estimate of θ_{0n}. A method for adjusting the difference is suggested below. The method is similar to the one considered in Hall, Sheather, Jones and Marron (1991).

From (3.1), we can write $A_n(\theta) = A(\theta) - R_n(\theta)$, where

$$R_n(\theta) = (24\pi)^{-1} n^{-2/5} \theta^6 \int_0^\infty \lambda^6 |\phi(\lambda)|^2 \, d\lambda \int x^2 w(x) \, dx \int x^4 w(x) \, dx + n^{-1/5} \int f^2(x) \, dx + O(n^{-3/5}). \tag{6.1}$$

By a Taylor series expansion, we have

$$R_n'(\theta_0) = -(\theta_0 - \theta_{0n}) A_n''(\tilde{\theta}) \tag{6.2}$$

for some $\tilde{\theta}$ lies in between θ_0 and θ_{0n}. (6.1) and (6.2) suggest the adjusted plug-in estimate

$$\hat{\theta}_{\mathrm{AP}} = \hat{\theta}_P + \tilde{R}_n'(\hat{\theta}_P) / \tilde{A}_n''(\hat{\theta}_P), \tag{6.3}$$

where

$$\tilde{R}_n(\theta) = (24\pi)^{-1} n^{-2/5} \theta^6 \int_0^\Lambda \lambda^6 \left\{ |\tilde{\phi}(\lambda)|^2 - 1/n \right\} d\lambda \times \int x^2 w(x) \, dx \int x^4 w(x) \, dx \tag{6.4}$$

and

$$\tilde{A}_n(\theta) = 4^{-1} \theta^4 \hat{G} \left\{ \int x^2 w(x) \, dx \right\}^2 + \theta^{-1} \int w^2(x) \, dx - \tilde{R}_n(\theta). \tag{6.5}$$

The estimates $\hat{\theta}_S$ and $\hat{\theta}_{\mathrm{AP}}$ have the same limiting distribution. In fact, they are asymptotically equivalent when the kernel function also satisfies some proper conditions. This is the first result that establishes the asymptotic equivalence of a plug-in estimate and the minimizer of an estimated $\mathrm{MISE}_n(\beta)$.

THEOREM 7. *In addition to the conditions in Theorem* 4, *also assuming* $\int |x|^5 w(x) \, dx < \infty$, *then* $\hat{\theta}_S - \hat{\theta}_{\mathrm{AP}} = o(n^{-1/2})$.

7. Simulation results and an example. We carried out some simulations to compare the performance of the stabilized bandwidth estimate, the cross-validation bandwidth estimate and the plug-in estimates. We first considered the case of the standard normal distribution. Two hundred replications were generated for each of the sample sizes $n = 25$, 100, 400 and 1600. All random variables were generated by the function RAND in Fortran 77 on a SUN 4/60 computer. The Gaussian kernel is used throughout this section. The bandwidth is the standard deviation of the Gaussian kernel.

For each sample, $\tilde{\phi}(\lambda)$ was evaluated by applying the fast Fourier transform to the series

$$Y(t) = F_n(x_t) - F_n(x_{t-1}), \qquad t = 1, \ldots, 2048,$$

where $F_n(x)$ is the empirical distribution function, $x_0 = -\infty$ and $x_t = -16 +$

$32t/2048$. Three values of $c = 1$, 2 and $3 \approx -\log_e(0.05)$ were used in defining Λ and the stabilized criterion. Since $CV_n(\beta)$ often has multiple minima, the global minimizers were obtained by searching over 301 equally spaced points in proper intervals.

The kernel estimates shown in Figure 1 are based on a sample of size $n = 100$. The dotted curve is the standard normal density. The solid and dashed curves are, respectively, the density estimates with the bandwidths selected by the cross-validation (0.156) and stabilizer selector (0.464). It is clear that the density estimate with the bandwidth selected by the cross-validation does not provide a satisfactory estimate. This example is not an extreme case; the tenth sample percentile of $\hat{\beta}_{CV}$ is about 0.16.

The simulation results are summarized in Tables 1, 3 and 4. The rows with $\hat{\beta}_{S,c}$ are the results for $\hat{\beta}_S$ with $c = 1$, 2 or 3. The values inside the parentheses are the estimated standard errors of the sample averages. The standard errors of the averages of the bandwidths can be obtained from the sample standard deviations of the bandwidths. The optimal bandwidths, $MISE_n(\beta_{0n})$ and the asymptotic standard deviations of the bandwidth estimates are given in Tables 2 and 5. Figures 2 and 3 plot the estimated densities of the bandwidth estimates for the normal density with $n = 100$ and 400. The bandwidths of the density estimates are selected by the stabilized selector. The stars indicate the locations of the optimal bandwidths. It is clear that $\hat{\beta}_S$ is superior to $\hat{\beta}_{CV}$. The sample standard deviations of $\hat{\beta}_S$ agree very well with the asymptotic standard deviations given in Theorem 4. The sample means of $\hat{\beta}_S$ are very close to the optimal bandwidths for $n = 400$ and 1600. The average $ISE_n(\hat{\beta})$ are also greatly reduced. We also constructed the normal probability plots for $\hat{\beta}_S$, which suggest that the normal distribution provides an excellent approximation for the distributions of $\hat{\beta}_S$.

Although it was established that $\hat{\beta}_{CV}$ is asymptotically normal, the density estimates of $\hat{\beta}_{CV}$ in Figures 2 and 3 suggest that the distributions of $\hat{\beta}_{CV}$ are skew to the left and that the normal distribution does not provide a good approximation. For the regression case, a similar phenomenon was observed and an explanation for it was given in Chiu (1990). Based on Theorem 1, we can have an analogous explanation for the density case.

We apply the stabilized selector (with $c = 3$) to the data set of the eruption lengths of Old Faithful geyser in Yellowstone National Park, given in Weisberg (1980). The data set is also available in Silverman (1990). Figure 4 shows the estimated densities with the bandwidths selected by the cross-validation (0.101) and the stabilized selector (0.215). While the density estimate with the bandwidth 0.101 is too rough, it seems that the bandwidth 0.215 gives the right amount of smoothness. In fact, the picture looks much like the "best visual fit," chosen by a very experienced data analyst, Silverman (1986), Figure 2.8.

We next checked the performance of the stabilized selector when the true density is not very smooth. We consider the χ^2 distributions with degrees of freedom $k = 4, 6, 8, 10, 12, 14$. For each case, 200 replications were generated. To make it easier to see the effect of smoothness of f on Λ and the bandwidths β_{0n}, $\hat{\beta}_{CV}$ and $\hat{\beta}_S$, we normalized each random variable by its

TABLE 1
Summary of simulation results for normal density

n	$\hat{\beta}$	$E(\hat{\beta})$	Sample SD	$E(\hat{\beta} - \beta_{0n})^2$	$E\{\mathrm{ISE}_n(\hat{\beta})\}$	$E(\Lambda)$
25	$\hat{\beta}_{\mathrm{CV}}$	0.632	0.246	$6.10(0.47) \times 10^{-2}$	$3.00(0.24) \times 10^{-2}$	
	$\hat{\beta}_{S,1}$	0.680	0.161	$3.08(0.30) \times 10^{-2}$	$2.07(0.11) \times 10^{-2}$	2.597(0.069)
	$\hat{\beta}_{S,2}$	0.682	0.144	$2.58(0.26) \times 10^{-2}$	$1.96(0.11) \times 10^{-2}$	2.099(0.042)
	$\hat{\beta}_{S,3}$	0.690	0.132	$2.40(0.25) \times 10^{-2}$	$1.90(0.10) \times 10^{-2}$	1.871(0.028)
	$\hat{\beta}_{\mathrm{AP}}$	0.723	0.126	$2.86(0.24) \times 10^{-2}$	$1.91(0.10) \times 10^{-2}$	
	$\hat{\beta}_P$	0.624	0.129	$1.68(0.21) \times 10^{-2}$	$1.89(0.11) \times 10^{-2}$	
100	$\hat{\beta}_{\mathrm{CV}}$	0.432	0.144	$2.09(0.22) \times 10^{-2}$	$1.03(0.08) \times 10^{-2}$	
	$\hat{\beta}_{S,1}$	0.454	0.070	$4.90(0.56) \times 10^{-3}$	$7.01(0.35) \times 10^{-3}$	3.025(0.077)
	$\hat{\beta}_{S,2}$	0.461	0.059	$3.72(0.36) \times 10^{-3}$	$6.71(0.33) \times 10^{-3}$	2.512(0.040)
	$\hat{\beta}_{S,3}$	0.464	0.054	$3.27(0.33) \times 10^{-3}$	$6.62(0.33) \times 10^{-3}$	2.286(0.030)
	$\hat{\beta}_{\mathrm{AP}}$	0.473	0.051	$3.41(0.31) \times 10^{-3}$	$6.58(0.32) \times 10^{-3}$	
	$\hat{\beta}_P$	0.435	0.055	$3.09(0.41) \times 10^{-3}$	$6.65(0.34) \times 10^{-3}$	
400	$\hat{\beta}_{\mathrm{CV}}$	0.319	0.090	$8.20(0.98) \times 10^{-3}$	$3.32(0.20) \times 10^{-3}$	
	$\hat{\beta}_{S,1}$	0.333	0.027	$7.16(1.10) \times 10^{-4}$	$2.46(0.11) \times 10^{-3}$	3.222(0.062)
	$\hat{\beta}_{S,2}$	0.335	0.021	$4.76(0.56) \times 10^{-4}$	$2.42(0.11) \times 10^{-3}$	2.770(0.032)
	$\hat{\beta}_{S,3}$	0.336	0.020	$4.24(0.37) \times 10^{-4}$	$2.41(0.11) \times 10^{-3}$	2.589(0.018)
	$\hat{\beta}_{\mathrm{AP}}$	0.338	0.019	$4.38(0.38) \times 10^{-4}$	$2.41(0.11) \times 10^{-3}$	
	$\hat{\beta}_P$	0.324	0.020	$4.40(0.40) \times 10^{-4}$	$2.41(0.11) \times 10^{-3}$	
1600	$\hat{\beta}_{\mathrm{CV}}$	0.240	0.055	$3.03(0.47) \times 10^{-3}$	$9.56(0.56) \times 10^{-4}$	
	$\hat{\beta}_{S,1}$	0.245	0.011	$1.29(0.27) \times 10^{-4}$	$7.49(0.32) \times 10^{-4}$	3.616(0.075)
	$\hat{\beta}_{S,2}$	0.247	0.008	$6.47(0.66) \times 10^{-5}$	$7.43(0.32) \times 10^{-4}$	3.091(0.028)
	$\hat{\beta}_{S,3}$	0.247	0.008	$6.29(0.74) \times 10^{-5}$	$7.42(0.32) \times 10^{-4}$	2.931(0.021)
	$\hat{\beta}_{\mathrm{AP}}$	0.247	0.008	$5.77(0.54) \times 10^{-5}$	$7.42(0.32) \times 10^{-4}$	
	$\hat{\beta}_P$	0.241	0.008	$9.98(1.14) \times 10^{-5}$	$7.44(0.32) \times 10^{-4}$	

The number of replications for each sample size is 200. The estimates $\hat{\beta}_{S,c}$ are the stabilized estimates with $c = 1$, 2 or 3. The values inside the parentheses are the estimated standard errors of the sample averages.

TABLE 2
Optimal bandwidths, MISE(β) *and asymptotic standard deviations of* $\hat{\beta}_{\mathrm{CV}}$ *and stabilized bandwidth estimate for normal density*

n	β_{0n}	β_0	$\mathrm{MISE}_n(\beta_{0n})$	$\sigma(\hat{\beta}_{\mathrm{CV}})$	$\sigma(\hat{\beta}_S)$
25	0.610	0.556	1.37×10^{-2}	0.198	0.147
100	0.445	0.422	5.41×10^{-3}	0.131	0.056
400	0.330	0.320	2.02×10^{-3}	0.086	0.021
1600	0.247	0.242	7.25×10^{-4}	0.057	0.008

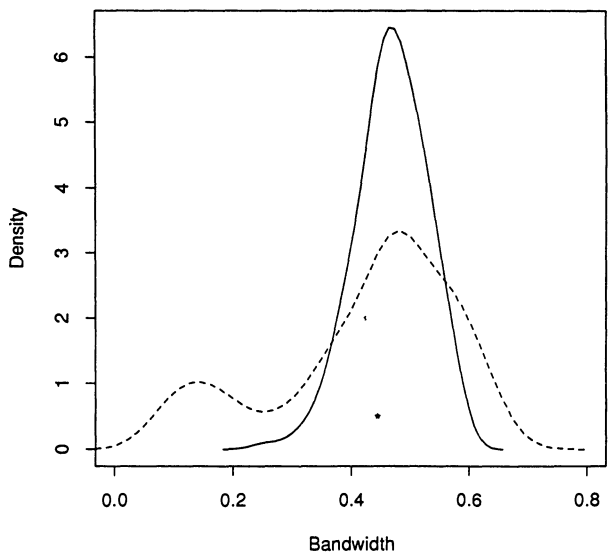

FIG. 2. *The estimated densities of* $\hat{\beta}_S$ *(solid curve) and* $\hat{\beta}_{CV}$ *(dashed curve) for the normal density with* $n = 100$. *The sample size is* 200 *and the bandwidths selected by the stabilized selector are* 0.024 ($\hat{\beta}_S$) *and* 0.04 ($\hat{\beta}_{CV}$). *The star indicates the location of* $\beta_{0n} = 0.445$.

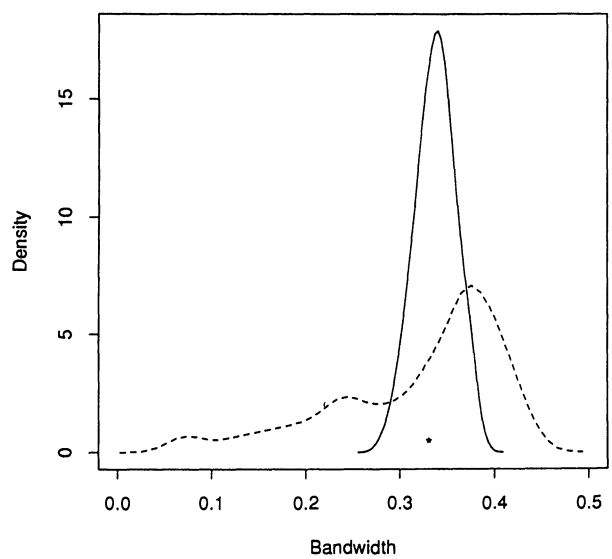

FIG. 3. *The estimated densities of* $\hat{\beta}_S$ *(solid curve) and* $\hat{\beta}_{CV}$ *(dashed curve) for the normal density with* $n = 400$. *The sample size is* 200 *and the bandwidths selected by the stabilized selector are* 0.0085 ($\hat{\beta}_S$) *and* 0.017 ($\hat{\beta}_{CV}$). *The star indicates the location of* $\beta_{0n} = 0.331$.

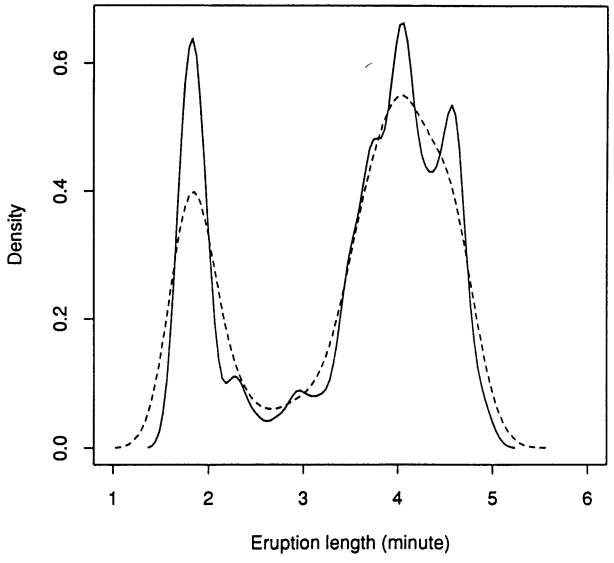

FIG. 4. *Density estimates of the eruption length of Old Faithful geyser. The solid and dashed curves are, respectively, the density estimates with the bandwidths selected by the cross-validation* (0.101) *and the stabilized selector* (0.215), *respectively.*

standard deviation. Note that the characteristic function of a χ^2 distribution with k degrees of freedom is $\phi(\lambda) = (1 - 2i\lambda)^{-k/2}$, and so $|\phi(\lambda)|^2 = O(\lambda^{-k})$.

The results for $k = 4$, 8 and 12 are summarized in Tables 3 and 4. From the tables, we see that $E(\Lambda)$ decreases as the degrees of freedom (smoothness) of the χ^2 density increases. This demonstrates an important property of the stabilized procedure that it is adaptive to the smoothness of $f(x)$. As expected, when the true density is not smooth enough, the stabilized procedure is more biased toward oversmoothing than $\hat{\beta}_{\mathrm{CV}}$. However, in terms of the mean squared error or the averaged $\mathrm{ISE}_n(\hat{\beta})$, the stabilized estimates still perform much better.

From Tables 1, 3 and 4, we see that there is only little difference between the stabilized estimates based on different c. For the cases of the normal density, setting $c = 1$ slightly reduces the bias but increases the standard deviation by about 20%. For the cases of normalized χ^2 densities, setting $c = 2$ or 3 yields slightly better results when $k \geq 6$. Reducing c from 3 to 2 increases only slightly the value of Λ. An intuitive reason here is that when $|\phi(\lambda)|$ is negligible, the chance that $\min_{\lambda \leq \mu} |\tilde{\phi}(\lambda)|^2 > c/n$ decreases exponentially as $\mu \to \infty$. These empirical results are consistent with Theorems 4 and 6 that the selection of c is not important. This should relieve us from being overly concerned with the choice of c. For most practical purposes, setting $-\log_e(0.15) \leq c \leq -\log_e(0.05)$ should yield good results. When $|\phi(\lambda)|$ decays slowly [this can be seen from the plot of $|\tilde{\phi}(\lambda)|^2$], a smaller c can be used to reduce the bias in the bandwidth and density estimates.

TABLE 3
Summary of simulation results for normalized χ^2 densities, $n = 100$

d.f.	$\hat{\beta}$	$E(\hat{\beta})$	Sample SD	$E(\hat{\beta} - \beta_{0n})^2$	$E\{\mathrm{ISE}_n(\hat{\beta})\}$	$E(\Lambda)$
4	$\hat{\beta}_{\mathrm{CV}}$	0.264	0.083	$6.84(0.65) \times 10^{-3}$	$1.68(0.08) \times 10^{-2}$	
	$\hat{\beta}_{S,1}$	0.293	0.048	$3.40(0.28) \times 10^{-3}$	$1.36(0.04) \times 10^{-2}$	4.781(0.089)
	$\hat{\beta}_{S,2}$	0.298	0.046	$3.67(0.31) \times 10^{-3}$	$1.37(0.04) \times 10^{-2}$	4.063(0.057)
	$\hat{\beta}_{S,3}$	0.308	0.048	$4.66(0.38) \times 10^{-3}$	$1.39(0.04) \times 10^{-2}$	3.633(0.050)
	$\hat{\beta}_{\mathrm{AP}}$	0.319	0.047	$5.86(0.44) \times 10^{-3}$	$1.41(0.04) \times 10^{-2}$	
	$\hat{\beta}_P$	0.282	0.046	$2.66(0.24) \times 10^{-3}$	$1.35(0.04) \times 10^{-2}$	
8	$\hat{\beta}_{\mathrm{CV}}$	0.352	0.107	$1.15(0.12) \times 10^{-2}$	$1.10(0.08) \times 10^{-2}$	
	$\hat{\beta}_{S,1}$	0.372	0.061	$3.96(0.44) \times 10^{-3}$	$8.25(0.41) \times 10^{-3}$	3.747(0.083)
	$\hat{\beta}_{S,2}$	0.379	0.051	$3.18(0.28) \times 10^{-3}$	$7.89(0.38) \times 10^{-3}$	3.111(0.048)
	$\hat{\beta}_{S,3}$	0.383	0.049	$3.17(0.27) \times 10^{-3}$	$7.83(0.37) \times 10^{-3}$	2.834(0.036)
	$\hat{\beta}_{\mathrm{AP}}$	0.394	0.047	$3.70(0.30) \times 10^{-3}$	$7.87(0.37) \times 10^{-3}$	
	$\hat{\beta}_P$	0.355	0.049	$2.37(0.23) \times 10^{-3}$	$7.77(0.37) \times 10^{-3}$	
12	$\hat{\beta}_{\mathrm{CV}}$	0.405	0.101	$1.06(0.10) \times 10^{-2}$	$8.94(0.49) \times 10^{-3}$	
	$\hat{\beta}_{S,1}$	0.409	0.055	$3.51(0.31) \times 10^{-3}$	$7.51(0.37) \times 10^{-3}$	3.290(0.047)
	$\hat{\beta}_{S,2}$	0.407	0.052	$3.21(0.30) \times 10^{-3}$	$7.45(0.37) \times 10^{-3}$	2.869(0.046)
	$\hat{\beta}_{S,3}$	0.412	0.048	$2.99(0.28) \times 10^{-3}$	$7.35(0.36) \times 10^{-3}$	2.604(0.033)
	$\hat{\beta}_{\mathrm{AP}}$	0.422	0.046	$3.39(0.30) \times 10^{-3}$	$7.36(0.35) \times 10^{-3}$	
	$\hat{\beta}_P$	0.384	0.048	$2.35(0.30) \times 10^{-3}$	$7.31(0.36) \times 10^{-3}$	

The sample size n is 100. The number of replications for each sample size is 200. The estimates $\hat{\beta}_{S,c}$ are the stabilized estimates with $c = 1$, 2 or 3. The values inside the parentheses are the estimated standard errors of the sample averages.

Finally, we study the performance of the plug-in estimates defined in Sections 5 and 6. We considered the same cases and used the same data sets. The estimates were obtained by setting $c = 3$. The results are also given in Tables 1, 3 and 4. It is interesting to see that for smaller sample sizes or rougher densities, $\hat{\beta}_P = n^{-1/5}\hat{\theta}_P$ is a much more accurate estimate of β_{0n} than the estimates $\hat{\beta}_S$ and $\hat{\beta}_{\mathrm{AP}} = n^{-1/5}\hat{\theta}_{\mathrm{AP}}$ are. We note that the proposed bandwidth estimates are slightly biased toward oversmoothing—caused by dropping $|\phi(\lambda)|^2$ at high frequency. Although $\hat{\beta}_P$ is designed for estimating β_0, the bias in $\hat{\beta}_P$ somewhat offsets the difference between β_{0n} and β_0. However, in terms of the averaged $\mathrm{ISE}_n(\hat{\beta})$, the proposed estimates have essentially the same performance.

8. Assumptions and proofs. The conditions about $f(x)$ and $w(x)$ are summarized in Assumptions 1 to 3. Since the bandwidth selection procedures and the proofs are based on characteristic functions, it is natural to describe the conditions in terms of $\phi(\lambda)$ and $W(\lambda)$.

TABLE 4
Summary of simulation results for normalized χ^2 densities, $n = 400$

d.f.	$\hat{\beta}$	$E(\hat{\beta})$	Sample SD	$E(\hat{\beta} - \beta_{0n})^2$	$E\{\mathrm{ISE}_n(\hat{\beta})\}$	$E(\Lambda)$
4	$\hat{\beta}_{\mathrm{CV}}$	0.165	0.046	$2.17(0.24) \times 10^{-3}$	$6.33(0.27) \times 10^{-3}$	
	$\hat{\beta}_{S,1}$	0.190	0.026	$0.94(0.08) \times 10^{-3}$	$5.35(0.18) \times 10^{-3}$	6.492(0.129)
	$\hat{\beta}_{S,2}$	0.196	0.024	$1.08(0.09) \times 10^{-3}$	$5.37(0.18) \times 10^{-3}$	5.550(0.080)
	$\hat{\beta}_{S,3}$	0.200	0.023	$1.24(0.11) \times 10^{-3}$	$5.38(0.18) \times 10^{-3}$	5.082(0.065)
	$\hat{\beta}_{\mathrm{AP}}$	0.204	0.022	$1.45(0.12) \times 10^{-3}$	$5.42(0.18) \times 10^{-3}$	
	$\hat{\beta}_P$	0.186	0.024	$0.78(0.08) \times 10^{-3}$	$5.29(0.18) \times 10^{-3}$	
8	$\hat{\beta}_{\mathrm{CV}}$	0.256	0.056	$3.18(0.39) \times 10^{-3}$	$3.40(0.16) \times 10^{-3}$	
	$\hat{\beta}_{S,1}$	0.261	0.026	$6.74(0.76) \times 10^{-4}$	$2.91(0.12) \times 10^{-3}$	4.494(0.084)
	$\hat{\beta}_{S,2}$	0.264	0.021	$5.25(0.56) \times 10^{-4}$	$2.87(0.12) \times 10^{-3}$	3.847(0.046)
	$\hat{\beta}_{S,3}$	0.266	0.021	$5.40(0.29) \times 10^{-4}$	$2.87(0.12) \times 10^{-3}$	3.567(0.037)
	$\hat{\beta}_{\mathrm{AP}}$	0.270	0.020	$5.83(0.54) \times 10^{-4}$	$2.87(0.12) \times 10^{-3}$	
	$\hat{\beta}_P$	0.252	0.022	$4.94(0.64) \times 10^{-4}$	$2.87(0.12) \times 10^{-3}$	
12	$\hat{\beta}_{\mathrm{CV}}$	0.278	0.067	$4.57(0.62) \times 10^{-3}$	$3.31(0.21) \times 10^{-3}$	
	$\hat{\beta}_{S,1}$	0.284	0.026	$6.66(0.71) \times 10^{-4}$	$2.60(0.11) \times 10^{-3}$	4.052(0.076)
	$\hat{\beta}_{S,2}$	0.285	0.024	$5.64(0.62) \times 10^{-4}$	$2.57(0.11) \times 10^{-3}$	3.494(0.048)
	$\hat{\beta}_{S,3}$	0.287	0.022	$5.01(0.55) \times 10^{-4}$	$2.56(0.11) \times 10^{-3}$	3.253(0.035)
	$\hat{\beta}_{\mathrm{AP}}$	0.290	0.021	$4.88(0.45) \times 10^{-4}$	$2.55(0.11) \times 10^{-3}$	
	$\hat{\beta}_P$	0.273	0.023	$6.36(0.89) \times 10^{-4}$	$2.57(0.11) \times 10^{-3}$	

The sample size n is 400. The number of replications for each sample size is 200. The estimates $\hat{\beta}_{S,c}$ are the stabilized estimates with $c = 1$, 2 or 3. The values inside the parentheses are the estimated standard errors of the sample averages.

TABLE 5
Optimal bandwidths, $\mathrm{MISE}(\beta)$ and asymptotic standard deviations of $\hat{\beta}_{\mathrm{CV}}$ and stabilized bandwidth estimate for normalized χ^2 densities

n	d.f.	β_{0n}	β_0	$\mathrm{MISE}_n(\beta_{0n})$	$\sigma(\hat{\beta}_{\mathrm{CV}})$	$\sigma(\hat{\beta}_S)$
100	4	0.259	0.000	1.19×10^{-2}		
	8	0.355	0.309	7.40×10^{-3}	0.051	
	12	0.386	0.353	6.56×10^{-3}	0.076	0.047
400	4	0.173	0.000	4.71×10^{-3}		
	8	0.256	0.234	2.78×10^{-3}	0.034	
	12	0.283	0.267	2.45×10^{-3}	0.050	0.018

ASSUMPTION 1. There exist positive constants M_1, M_2, K_1 and K_2 such that $M_1|\lambda|^{-K_1} \geq |\phi(\lambda)|^2 \geq M_2|\lambda|^{-K_2}$ as $|\lambda| \to \infty$. Also assume $|\phi(\lambda)|^2 > 0$ for all λ.

ASSUMPTION 2. The density $f(x)$ has a uniformly bounded derivative and satisfies $\int_{|x|>M} f(x)\,dx \leq O(M^{-1})$ as $M \to \infty$.

ASSUMPTION 3. The kernel function $w(x)$ is a symmetric probability density and satisfies $\int |x|^3 w(x) < \infty$. The characteristic function of $w(x)$ satisfies $W(\lambda) = O(\lambda^{-3})$ and $W'(\lambda) = O(\lambda^{-3})$ as $\lambda \to \infty$.

Assumption 2 is a weak condition, which is satisfied by the Cauchy distribution. Assumption 3 is a standard one. Note that $\int |x|^3 w(x)\,dx < \infty$ implies $1 - W(\lambda) = 2^{-1}\lambda^2 \int x^2 w(x)\,dx + O(\lambda^3)$ as $\lambda \to 0$. The part of Assumption 1 that requires $|\phi(\lambda)|$ to decay fast is not strict. If $\int |f^{(l)}(x)|\,dx$ and $\int |f^{(l)}(x)|^2\,dx$ are bounded for $l = 1, \ldots, k-1$, and $f^{(k)}(x)$ is of bounded variation, then $|\phi(\lambda)| = O(|\lambda|^{k+1})$. The other part that requires $|\phi(\lambda)|$ to decay in some regular way is a crucial one. We need this condition to ensure that the bias of the proposed estimates caused by dropping $\phi(\lambda)$ at $\lambda > \Lambda$ is negligible.

We now proceed to prove the results in Sections 3 and 4.

PROOF OF THEOREM 1. By a Taylor expansion, we have

$$B_2'(\hat{\beta}_{\mathrm{CV}}) + B_3'(\hat{\beta}_{\mathrm{CV}}) + B_4'(\hat{\beta}_{\mathrm{CV}}) = -(\hat{\beta}_{\mathrm{CV}} - \beta_{0n})\pi\,\mathrm{MISE}_n''(\tilde{\beta})$$

for some $\tilde{\beta}$ in between $\hat{\beta}_{\mathrm{CV}}$ and β_{0n}. Since $\hat{\theta}_{\mathrm{CV}}$ converges to θ_0, it is sufficient to show that for $\beta = O(n^{-1/5})$, $B_2'(\beta)$ is the dominant term. This can be established by arguments analogous to the ones in Chiu (1988). □

We next establish the probability one bounds for Λ.

LEMMA 2. *Under Assumption* 1 *and for any* $\delta > 0$, $\Lambda \le n^{1/K_1+\delta}$ *with probability tending to* 1.

PROOF. For any constant $\delta > \delta_1 > 0$, let $a = n^{1/K_1+\delta_1}$ and $b = n^{1/K_1+\delta}$. Note that

$$P(\Lambda \ge b) \le P\left\{\min_{a \le \lambda \le b} |\tilde{\phi}(\lambda)|^2 > c/n\right\} \tag{8.1}$$

for any $\delta > \delta_1 > 0$. Since

$$|\tilde{\phi}_d(\lambda)| \le |\tilde{\phi}(\lambda)| + |\phi(\lambda)| \le |\tilde{\phi}(\lambda)| + \varepsilon/n$$

for any $\varepsilon > 0$, the right-hand side in (8.1) is less than

$$P\left\{\int_a^b (b-a)^{-1} |\tilde{\phi}_d(\lambda)|^2\,d\lambda > c' n^{-1}\right\} \tag{8.2}$$

for some $c > c' > 1$. By applying the Chebyshev inequality [Chung (1974), page 48], it can be shown that (8.2) is of order $(b-a)^{-k}$ for any even integer $k > 0$. Choosing k large enough to make $k(1/K_1 + \delta) > 1$ finishes the proof. □

LEMMA 3. *Under Assumptions* 1 *and* 3 *and for any* $\delta > 0$, $n^{1/K_2-\delta} \le \Lambda$ *with probability tending to* 1.

PROOF. It is sufficient to show that for any $\delta > 0$, with probability 1,

$$\max_{0\le\lambda\le n} \left|\tilde{\phi}_d(\lambda)\right| \le Mn^{-1/2+\delta}$$

for some constant $M > 0$. Let $\lambda_j = jn/N$, $j = 0, \ldots, N = n^3$. From Assumption 2 and the uniform continuity of $\phi(\lambda)$, we have that

$$n \max_{0\le\lambda\le n} \left|\hat{\phi}_d(\lambda)\right| - n \max_{0\le j\le N} \left|\tilde{\phi}_d(\lambda_j)\right|$$

converges to 0 almost surely. The lemma now follows from the fact that from Lemma 1,

$$P\left\{\max_{0\le j\le N} \left|\tilde{\phi}_d(\lambda_j)\right| > n^{-1/2+\delta}\right\} = O(Nn^{-2k\delta}) = O(n^{-2k\delta+3})$$

for any constant $k > 0$. □

We now proceed to prove Theorems 2 to 4. Comparing (3.1) and (4.1) yields

$$\begin{aligned} S_n(\beta) - \pi \,\mathrm{MISE}_n(\beta) = &-\int_0^\infty |\phi(\lambda)|^2\, d\lambda + D_1(\beta) + D_2(\beta) \\ &+ D_3(\beta) + D_4(\beta) + D_5(\beta), \end{aligned} \tag{8.3}$$

where

$$D_1(\beta) = -\int_\Lambda^\infty |\phi(\lambda)|^2 \{W^2(\beta\lambda) - 2W(\beta\lambda)\}\, d\lambda,$$

$$D_2(\beta) = 2\,\mathrm{Re}\int_0^\Lambda \phi(\lambda)\tilde{\phi}_d(-\lambda)\{W^2(\beta\lambda) - 2W(\beta\lambda)\}\, d\lambda,$$

$$D_3(\beta) = \int_0^\Lambda \left[\left|\tilde{\phi}_d(\lambda)\right|^2 - E\left\{\left|\tilde{\phi}_d(\lambda)\right|^2\right\}\right]\{W^2(\beta\lambda) - 2W(\beta\lambda)\}\, d\lambda,$$

$$D_4(\beta) = 2n^{-1}\int_0^\Lambda |\phi(\lambda)|^2 W(\beta\lambda)\, d\lambda$$

and

$$D_5(\beta) = n^{-1}\int_\Lambda^\infty |\phi(\lambda)|^2 W^2(\beta\lambda)\, d\lambda.$$

PROOF OF THEOREM 2. It is clear that $D_4(\beta) = O(n^{-1})$ and $D_5(\beta) = O(n^{-1})$. From Lemmas 2 and 3, and noting that $1 - W(\beta\lambda) = O(\beta^2\lambda^2)$, we have that, for any $0 < \tau < \min\{(K_1 - 5)/(4K_2), (K_1 - 5)/\{5(K_1 - 1)\}\}$,

$$\int_\Lambda^\infty |\phi(\lambda)|^2\{1 - W(\beta\lambda)\}^2\, d\lambda, \qquad \int_0^\Lambda \phi(\lambda)\tilde{\phi}_d(-\lambda)\{1 - W(\beta\lambda)\}^2\, d\lambda$$

and

$$\int_0^{\Lambda}\left\{|\phi_d(\lambda)|^2 - E|\tilde{\phi}_d(\lambda)|^2\right\}\{1 - W(\beta\lambda)\}^2$$

are of order $o(n^{-4/5})$, almost surely and uniformly on $n^{-1/5-\tau} < \beta < n^{-1/5+\tau}$. Since $\mathrm{MISE}_n(\beta) \geq Mn^{-4/5}$ for some constant $M > 0$,

$$\lim_{n\to\infty}\{\pi\,\mathrm{MISE}_n(\beta) - S_n(\beta) - C_n\}/\mathrm{MISE}_n(\beta) = 0$$

almost surely and uniformly on $n^{-1/5-\tau} < \beta < n^{-1/5+\tau}$, where C_n are some random variables which do not depend on β. Since $n^{-4/5}/\{S_n(\beta) + C_n\} = o(1)$ almost surely and uniformly for β outside $(n^{-1/5-\tau}, n^{-1/5+\tau})$, the minimizer of $S_n(\beta)$ must be inside $(n^{-1/5-\tau}, n^{-1/5+\tau})$ for large enough n. Applying the classical argument in Jennrich (1969) finishes the proof. □

PROOF OF THEOREM 3. By a Taylor expansion, we have

$$D_1'(\hat{\beta}_S) + D_2'(\hat{\beta}_S) + D_3'(\hat{\beta}_S) + D_4'(\hat{\beta}_S) = -(\hat{\beta}_S - \beta_{0n})\pi\,\mathrm{MISE}_n''(\tilde{\beta}) \tag{8.4}$$

for some $\tilde{\beta}$ in between $\hat{\beta}_S$ and β_{0n}. From Theorem 1, it is sufficient to show that $D_j'(\beta_{0n}) = o(n^{-7/10})$ for $j = 1, 2, 3, 4$, which can be established by straightforward computation. □

Since the $D_2'(\beta_{0n})$ is the dominant term, we give a precise description of its asymptotic distribution.

LEMMA 4. *Under the conditions of Theorem* 4 *and assuming* $\beta = n^{-1/5}\theta$ *for some* $\theta > 0$, $n^{1/2}\beta^{-3}D_2'(\beta)$ *is asymptotically normal with mean* 0 *and variance*

$$\left\{\int x^2 w(x)\,dx\right\}^4\left[\int\{f^{(4)}(x)\}^2 f(x)\,dx - \left\{\int f^{(4)}(x)\,f(x)\,dx\right\}^2\right].$$

PROOF. We first note that $D_2'(\beta) = \tilde{D}_2'(\beta) + o(n^{-11/10})$, where

$$\tilde{D}_2'(\beta) = \beta^3\left\{\int x^2 w(x)\,dx\right\}^2\int\lambda^4\phi(\lambda)\tilde{\phi}_d(-\lambda)\,d\lambda.$$

It is clear that $E\{\tilde{D}_2'(\beta)\} = 0$. The variance of $\int\lambda^4\phi(\lambda)\tilde{\phi}_d(-\lambda)\,d\lambda$ is equal to

$$n^{-1}\int\int\lambda^4\mu^4\phi(-\lambda)\phi(\mu)\phi(\lambda-\mu)\,d\lambda\,d\mu - n^{-1}\left\{\int\lambda^4|\phi(\lambda)|^2\,d\lambda\right\}^2. \tag{8.5}$$

Noting that (8.5) is equal to $4\pi^2 n^{-1}\int\{f^{(4)}(x)\}^2 f(x)\,dx - \{\int f^{(4)}(x)\,f(x)\,dx\}^2$ gives the asymptotic variance of $D_2'(\beta)$.

To establish the asymptotic normality, it is sufficient to show that the cumulants of $n^{1/2}\beta^{-3}\tilde{D}_2'(\beta)$ of order $k \geq 3$ converge to 0. From Lemma 1, the kth-order cumulant of $n^{1/2}\beta^{-3}\tilde{D}_2'(\beta)$ is of order $(n^{1/2}\beta^{-3})^k\beta^{3k}n^{-k+1} = n^{1-k/2}$, which converges to 0 for $k \geq 3$. □

PROOF OF THEOREM 4. Under the conditions of the theorem, it can be seen that $D_j'(\beta_{0n}) = o(n^{-11/10})$, for $j = 1, 3, 4, 5$. Therefore $D_2'(\beta_{0n})$ is the dominant factor on the right-hand side of (8.4). The theorem now follows Lemma 4. □

In the proofs of Theorems 5 to 7, write

$$\pi\hat{G} - \pi\int\{f''(x)\}^2\,dx = D_1 + D_2 + D_3, \tag{8.6}$$

where

$$D_1 = -\int_\Lambda^\infty \lambda^4|\phi(\lambda)|^2\,d\lambda,$$

$$D_2 = 2\,\mathrm{Re}\int_0^\Lambda \lambda^4\phi(\lambda)\tilde{\phi}_d(-\lambda)\,d\lambda$$

and

$$D_3 = \int_0^\Lambda \lambda^4\left\{\left|\tilde{\phi}_d(\lambda)\right|^2 - 1/n\right\}d\lambda.$$

PROOF OF THEOREM 5. From Lemmas 3 and 4, we have that, for any $1/K_2 > \delta > 0$, there exists a constant $M > 0$ such that $|D_1| \le Mn^{-(K_1-5)(1/K_2-\delta)}$, $|D_3| \le Mn^{5/K_1-1+\delta}$ with probability 1. We also have, with probability 1, $|D_2| \le Mn^{-1/2}$ when $K_1 > 10$ and $|D_2| \le Mn^{-1/2}n^{-(K_1/2-5+\delta)}$ when $K_1 \le 10$. It can be seen that when $K_1 > 5$, D_j's converge to 0 almost surely and the proof is finished. □

PROOF OF THEOREM 6. From the proof of Theorem 5, we see that D_2 is the dominant term in the right-hand side of (8.6) when $10 < K_1 \le K_2 \le 2K_1 - 10$. The asymptotic distribution of D_2 can be obtained by following the argument in Lemma 4. □

PROOF OF COROLLARY 2. By a Taylor expansion, we have

$$\hat{\theta}_P - \theta_0 = -5^{-1}\left\{\int w^2(x)\,dx\right\}^{1/5}\left\{\int x^2w(x)\,dx\right\}^{-2/5}\tilde{G}^{-6/5}(\hat{G} - G)$$

for some $\tilde{G}$ in between $\hat{G}$ and G. The proof is finished by noting that

$$\theta_0^3\left\{\int x^2w(x)\,dx\right\}^2/A''(\theta_0) = 5^{-1}\left\{\int w^2(x)\,dx\right\}^{1/5}\left\{\int x^2w(x)\,dx\right\}^{-2/5}G^{-6/5}.$$

□

PROOF OF THEOREM 7. From (6.2) to (6.5), we have

$$\hat{\theta}_{\mathrm{AP}} - \theta_{0n} = \hat{\theta}_P - \theta_0 + R_n(\theta_0)/A_n''(\tilde{\theta}) - \tilde{R}_n(\hat{\theta}_P)/\tilde{A}_n''(\hat{\theta}_P)$$

for some $\tilde{\theta}$ lies in between θ_0 and θ_{0n}. The functions $R_n(\theta)$, $\tilde{R}_n(\theta)$ and $\tilde{A}_n(\theta)$

are defined in Section 6. Following the proof in Lemma 4, it can be shown that

$$\int_0^{\Lambda} \lambda^6 \left\{ \left| \tilde{\phi}(\lambda) \right|^2 - n^{-1} \right\} d\lambda = O\left(n^{(-K_1+7)/(K_2+\delta)}\right)$$

for any $\delta > 0$. Therefore $R_n(\theta_0)/A_n''(\tilde{\theta}) - \tilde{R}_n(\hat{\theta}_P)/\tilde{A}_n''(\hat{\theta}_P)$ is of order $o(n^{-1/2})$ under the conditions of the theorem.

It is then sufficient to show that $\hat{\theta}_P - \theta_0$ and $\hat{\theta}_S - \theta_{0n}$ are asymptotically equivalent. From Theorem 4, we have

$$n^{1/2}\left(\hat{\theta}_S - \theta_{0n}\right) \approx -n^{1/2} n^{-3/5} \tilde{D}_2\left(n^{1/5}\theta_0\right)/\{\pi A''(\theta_0)\},$$

where

$$\tilde{D}_2(\beta) = 2\beta^{-1} \operatorname{Re} \int_0^{\Lambda} \phi(\lambda)\tilde{\phi}_d(-\lambda) V(B\lambda)\, d\lambda,$$

and $V(\lambda) = -2\{1 - W(\lambda)\}W(\lambda)\lambda$. Under the assumptions of the theorem, $V(\lambda) = \lambda^4\{\int x^2 w(x)\, dx\}^2 + O(\lambda^5)$. Noting that $\tilde{D}_2(\beta) = \beta^3 D_2\{\int x^2 w(x)\, dx\}^2 + O(\beta^4)$ gives the result. □

Acknowledgments. The valuable comments, which significantly improved the paper, from the referee and an Associate Editor are gratefully appreciated. The author thanks Professors J. S. Marron, P. Hall, B. Park, M. C. Jones and S. J. Sheather for providing their preprints.

REFERENCES

BICKEL, P. J. and RITOV, Y. (1988). Estimating integrated squared density derivatives: Sharp best order of convergence estimates. *Sankhyā Ser. A* **50** 381–393.

BOWMAN, A. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* **71** 353–360.

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

CHIU, S.-T. (1990). On the asymptotic distribution of bandwidth estimates. *Ann. Statist.* **18** 1696–1711.

CHUNG, K. L. (1974). *A Course in Probability Theory*, 2nd ed. Academic, New York.

DIGGLE, P. J. and MARRON, J. S. (1988). Equivalence of smoothing parameter selectors in density and intensity estimation. *J. Amer. Statist. Assoc.* **83** 793–800.

EPANECHNIKOV, V. A. (1969). Nonparametric estimation of a multidimensional probability density. *Theory Probab. Appl.* **14** 153–158.

HALL, P. (1983). Large sample optimality of least squares cross-validation in density estimation. *Ann. Statist.* **11** 1156–1174.

HALL, P. and MARRON J. S. (1987a). Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation. *Probab. Theory Related Fields* **74** 567–581.

HALL, P. and MARRON, J. S. (1987b). Estimation of integrated squared density derivatives. *Statist. Probab. Lett.* **6** 109–115.

HALL, P. and MARRON, J. S. (1990). Lower bounds for bandwidth selection in density estimation. *Probab. Theory Related Fields*. To appear.

HALL, P., MARRON, J. S. and PARK, B. (1989). Smoothed cross-validation. Unpublished manuscript.

HALL, P., SHEATHER, S. J., JONES, M. C. and MARRON, J. S. (1991). On optimal data-based bandwidth selection in kernel density estimation. *Biometrika*. To appear.

JENNRICH, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *Ann. Math. Statist.* **40** 633–543.

JONES, M. C., MARRON, J. S. and PARK, B. U. (1991). A simple root n bandwidth selector. *Ann. Statist.* **19** 1919–1932.

PARK, B. and MARRON, J. S. (1990). Comparison of data-driven bandwidth selectors. *J. Amer. Statist. Assoc.* **85** 66–72.

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

ROSENBLATT, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **27** 832–837.

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

RUDEMO, M. (1982). Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.* **9** 65–78.

RUDIN, W. (1974). *Real and Complex Analysis*, 2nd ed. McGraw-Hill, New York.

SCOTT, D. W. and FACTOR, L. E. (1981). Monte Carlo study of three data-based nonparametric probability density estimators. *J. Amer. Statist. Assoc.* **76** 9–15.

SCOTT, D. W. and TERRELL, G. R. (1987). Biased and unbiased cross-validation in density estimation. *J. Amer. Statist. Assoc.* **82** 1131–1146.

SHEATHER, S. J. (1986). An improved data-based algorithm for choosing the window width when estimating the density at a point. *Comput. Statist. Data Anal.* **4** 61–65.

SHEATHER, S. J. and JONES, M. C. (1990). A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc. Ser. B.* To appear.

SILVERMAN, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *J. Roy. Statist. Soc. Ser. B* **47** 1–52.

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

STONE, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.* **12** 1285–1297.

WEISBERG, S. (1980). *Applied Linear Regression*. Wiley, New York.

WOODROOFE, M. (1970). On choosing a delta sequence. *Ann. Math. Statist.* **41** 1665–1671.

DEPARTMENT OF STATISTICS
COLORADO STATE UNIVERSITY
FORT COLLINS, COLORADO 80523