The Annals of Statistics
1992, Vol. 20, No. 4, 2057–2070

BEST POSSIBLE CONSTANT FOR BANDWIDTH SELECTION[1]

By Jianqing Fan and James S. Marron

University of North Carolina

For the data based choice of the bandwidth of a kernel density estimator, several methods have recently been proposed which have a very fast asymptotic rate of convergence to the optimal bandwidth. In particular the relative rate of convergence is the square root of the sample size, which is known to be the best possible. The point of this paper is to show how semiparametric arguments can be employed to calculate the best possible constant coefficient, that is, an analog of the usual Fisher information, in this convergence. This establishes an important benchmark as to how well bandwidth selection methods can ever hope to perform. It is seen that some existing methods attain the bound, others do not.

1. Introduction. Nonparametric curve estimation provides a useful tool for understanding the structure of a data set. See Silverman (1986), Eubank (1988), Müller (1988), Härdle (1990) and Wahba (1990) for many examples of this and good introductions to the general subject area. The most important practical hurdle, in applications of this methodology, is choice of the smoothing parameter.

A large amount of recent progress has been made on data based smoothing parameter selection. See the survey paper by Marron (1988). Because it provides a simple context in which to study the problem (hence allowing deeper results), much of this progress has come in the case of kernel density estimation. Hence that setting is discussed here as well.

A useful asymptotic means of assessing performance of a data driven smoothing parameter, that is, bandwidth, is through the relative rate of convergence to the bandwidth that minimizes the mean integrated squared error (MISE).

Hall, Sheather, Jones and Marron (1991), Jones, Marron and Park (1991) and Chiu (1991) have all proposed methods for which this rate of convergence is extremely fast. In particular, it goes down as $O(n^{-1/2})$, where n denotes sample size, which is unusually fast in nonparametric settings. This rate of convergence has been shown to be the best possible, in an important minimax sense, by Hall and Marron (1991). But the fact that there are competing selectors motivates deeper analysis.

Received October 1990; revised February 1992.

[1] Manuscript prepared using computer equipment supported in part by NSF Grant DMS-90-05905. The first author was supported by a grant of the University of North Carolina at Chapel Hill. The second author was supported by NSF Grant DMS-89-02973.

AMS 1980 *subject classifications*. Primary 62G07; secondary 62B10, 62C20.

Key words and phrases. Bandwidth selection, efficient bounds, kernel density estimator, semiparametric methods.

A natural step in this direction is to consider not only the exponent in the rate of convergence, but also the constant coefficient. This type of question is frequently addressed in semiparametric analysis, which is an extension of the classical Fisher information ideas. See Bickel, Klassen, Ritov and Wellner (1991) and van de Vaart (1988) for details. In this paper a straightforward application of these methods is used to calculate the best possible constant in our setting of bandwidth selection for kernel density estimation. It turns out that the problem of bandwidth selection is closely related to the problem of estimating some specific kinds of quadratic functionals, which are studied by Hall and Marron (1987), Bickel and Ritov (1988) and Jones and Sheather (1991) in density estimation models, and by Donoho and Nussbaum (1990) and Fan (1991) in Gaussian white noise models. The knowledge gained there is also very useful to bandwidth selection.

Chiu (1991) proposes two $n^{-1/2}$ bandwidth selectors, and shows that for both, the relative error is asymptotically normal. It is a simple calculation to show that his asymptotic variance is the same as the best possible constant coefficient derived here. This provides a strong sense in which our lower bound is informative. With more work, the selector of Hall, Sheather, Jones and Marron (1991) can be shown to have the same limiting distribution. However the $n^{-1/2}$ method of Jones, Marron and Park (1991) has a larger constant, and thus is not optimal in this sense.

Section 2 gives a precise formulation and discussion of the main results. Proofs are in Section 3.

2. Main results. To mathematically formulate the problem of bandwidth selection, assume that $X_1, \dots, X_n$ are i.i.d. from an unknown density f. Let $K(\cdot)$ denote a kernel function and h_n be a bandwidth. A kernel density estimator is defined by

$$\hat{f}_n(x) = \frac{1}{nh_n}\sum_1^n K\left(\frac{x - X_j}{h_n}\right). \tag{2.1}$$

Its performance is typically measured by the MISE

$$M(h_n) = E\int_{-\infty}^{\infty}\left(\hat{f}_n(x) - f(x)\right)^2 dx. \tag{2.2}$$

Here we take the optimal bandwidth $h_n(f)$ to be the minimizer (with ties broken arbitrarily) of MISE. See Hall and Marron (1991) for discussion of other viewpoints on assessing performance of bandwidth selectors, including reasons why the present approach is sensible.

The practical implementation of estimator (2.1) involves selecting a suitable amount of smoothing. The optimal bandwidth $h_n(f)$ naturally would be used, if it were known. In applications, $h_n(f)$ needs to be estimated. Several promising methods have been proposed, as indicated in the Introduction. Which methods are optimal? Discussions on best possible bandwidth selectors form the core of the paper.

Our results are formulated essentially in terms of a nonnegative kernel K. We assume rather strong smoothness, so strong that one could have a faster asymptotic rate of MISE convergence through the use of higher order kernels. However, we explicitly treat only nonnegative kernels because they are used almost exclusively in practice. One reason is that nonnegative kernels give a more interpretable result, since the intuition behind a local average is obvious, while it takes far greater insight to understand at an intuitive level how negative weights can benefit the averaging process. Another reason, as shown in Marron and Wand (1992), is that large practical gains for higher order kernels are often absent, or insignificant, in terms of MISE, for realistic sample sizes.

The problem of estimating $h_n(f)$ is closely related with that of estimating quadratic functionals

$$\theta_j(f) = \int_{-\infty}^{\infty} \left(f^{(j)}(x)\right)^2 dx, \qquad j = 2, 3. \tag{2.3}$$

Indeed, it will be shown (see Lemma 1 in Section 3) that the optimal bandwidth $h_n(f)$ can be approximated by

$$\phi_n(f) = c_1 \theta_2^{-1/5} n^{-1/5} + c_2 \theta_3 \theta_2^{-8/5} n^{-3/5}, \tag{2.4}$$

where

$$c_1 = \left(\int_{-\infty}^{\infty} K^2(x)\, dx \left(\int_{-\infty}^{\infty} z^2 K(z)\, dz \right)^{-2} \right)^{1/5}$$

and

$$c_2 = \frac{1}{20} \int_{-\infty}^{\infty} x^4 K(x)\, dx \left(\int_{-\infty}^{\infty} K^2(x)\, dx \right)^{3/5} \left(\int_{-\infty}^{\infty} z^2 K(z)\, dz \right)^{-11/5}.$$

This reduces the problem of estimating the optimal bandwidth to that of estimating the two quadratic functionals $\theta_2(f)$ and $\theta_3(f)$.

For convenience, denote a class of densities having $(k + \alpha)$ derivatives:

$$\mathscr{F}_{k+\alpha} = \left\{ g \colon \left| g^{(k)}(x) - g^{(k)}(y) \right| \le M |x - y|^{\alpha}, \left| g^{(4)}(x) \right| \le g_0(x) \right\},$$

where $g_0(x)$ is bounded continuous and integrable and $0 \le \alpha < 1$. Let $\| \cdot \|_2$ denote the usual L_2-norm, and let

$$H_n(f, C) = \left\{ g \in \mathscr{F}_{k+\alpha} \colon \left\| \sqrt{g} - \sqrt{f} \right\|_2 \le C/\sqrt{n} \right\} \tag{2.5}$$

be a Hellinger ball in the neighborhood of f.

The following theorem shows that the asymptotic relative error of any bandwidth selection procedure cannot be smaller than $B(f) n^{-1/2}$, where

$$B^2(f) = \frac{4}{25} \left(\frac{\int_{-\infty}^{\infty} \left(f^{(4)}(x) \right)^2 f(x)\, dx}{\left(\int_{-\infty}^{\infty} \left(f''(x) \right)^2 dx \right)^2} - 1 \right). \tag{2.6}$$

THEOREM 1. *Let K be a continuous second order kernel with $\int_{-\infty}^{\infty}|x|^6 K(x)\,dx < \infty$. Assume that $f \in \mathscr{F}_{k+\alpha}$ and $k + \alpha > 4$. Then, for any bandwidth selection procedure $\hat{h}_n$,*

$$\lim_{C\to\infty} \liminf_{n\to\infty} \inf_{\hat{h}_n} \sup_{g\in H_n(f,C)} nE_g\left(\frac{\hat{h}_n - h_n(g)}{h_n(g)}\right)^2 \geq B^2(f). \tag{2.7}$$

As discussed in the Introduction, the lower bound in (2.7) is the best attainable when $k + \alpha \geq 4.25$. It is worthwhile to note that (2.7) does not depend on the kernel function K, even though the optimal bandwidth $h_n(f)$ does. In other words, $B(f)$ measures the intrinsic difficulty of bandwidth selection.

The following theorem gives an analogous lower bound for the relative error of MISE. See, for example, Hall and Marron (1987) for discussion of the close relationship between Theorems 1 and 2.

THEOREM 2. *Under the assumption of Theorem 1, for any bandwidth selector $\hat{h}_n$,*

$$\lim_{C\to\infty} \liminf_{n\to\infty} \inf_{\hat{h}_n} \sup_{g\in H_n(f,C)} n^2E_g\left(\frac{M(\hat{h}_n) - M(h_n(g))}{M(h_n(g))}\right)^2 \geq 4B^4(f),$$

where $M(\cdot)$ was defined by (2.2).

The last result indicates that for any bandwidth selector, the relative error of MISE cannot be smaller than $2n^{-1}B^2(f)$. Thus, the quantity $B(f)$ plays an important role to the relative error of bandwidth selection, measured in either way: the larger $B(f)$, the harder the problem. In other words, $B(f)$ measures the difficulty of bandwidth selection problems.

Note that $B(f)$ is both location and scale invariant for any $\sigma > 0$ and μ, $B(f_{\mu,\sigma}) = B(f)$, where

$$f_{\mu,\sigma}(x) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right).$$

This is expected because, for example, estimating the $N(0,1)$ density is as difficult as estimating the $N(2,4)$ density: Plots of two estimates should look the same except the scales on x axis and y axis are labeled differently. For the normal case,

$$B(f) = \frac{2}{5}\sqrt{\frac{4864}{3^{5.5}} - 1} = 1.300.$$

Table 1 shows the values of $B(f)$ for the 15 normal mixture densities in Figure 1. See Marron and Wand (1992) for the parameters and for discussion of these densities.

Table 1 gives us an idea as to how difficult it is to select a bandwidth for a variety of densities. For example, density 4 is asymptotically $(2.638/1.300)^2 \approx$

TABLE 1
Constant factors in the lower bounds

Density number	$B(f)$	Density number	$B(f)$	Density number	$B(f)$
1	1.300	2	1.771	3	4.973
4	2.638	5	1.388	6	1.868
7	1.286	8	3.390	9	4.742
10	2.125	11	19.394	12	9.635
13	25.587	14	9.408	15	3.515

4.1 times as difficult as a normal density in terms of bandwidth selection: The best selector with sample size 820 would have roughly the same accuracy of estimating optimal bandwidth as in the normal case with sample size 200. Similarly, density 11 would be asymptotically $(19.394/1.3000)^2 \approx 222.6$ times as difficult as normal density in bandwidth selection terms. These are compatible in an intuitive sense.

REMARK 1. A direct consequence of Theorem 1 is that for any open neighborhood V of f (in L_2 topology), we have

$$\liminf_{n\to\infty} \inf_{\hat{h}_n} \sup_{g\in V\cap\mathscr{F}_{k+\alpha}} nE_g\left(\frac{\hat{h}_n - h_n(g)}{h_n(g)}\right)^2 \geq B^2(f).$$

A similar formula holds for MISE. Alternate formulations are possible in terms of balls, in various metrics, centered at f.

REMARK 2. Note that $B^2(f)$ plays a role analogous to the classical Fisher information. Thus, given any bandwidth selector (either already existing, or that may be constructed in the future) $\hat{h}_n$, its efficiency can be defined by

$$\frac{B^2(f)}{nE_f\big((\hat{h}_n - h_n(f))/h_n(f)\big)^2}.$$

REMARK 3. On the Hellinger ball $H_n(f, C)$, we have

$$\lim_{n\to\infty} \sup_{g\in H_n(f,C)} \left|\frac{h_n(g)}{h_n(f)} - 1\right| = 0.$$

Moreover,

$$\lim_{n\to\infty} \sup_{g\in H_n(f,C)} |g^{(4)}(x) - f^{(4)}(x)| = 0, \quad \forall x$$

and

$$\lim_{n\to\infty} \sup_{g\in H_n(f,C)} \left|\int_{-\infty}^{\infty}(g''(x))^2\,dx - \int_{-\infty}^{\infty}(f''(x))^2\,dx\right| = 0. \tag{2.8}$$

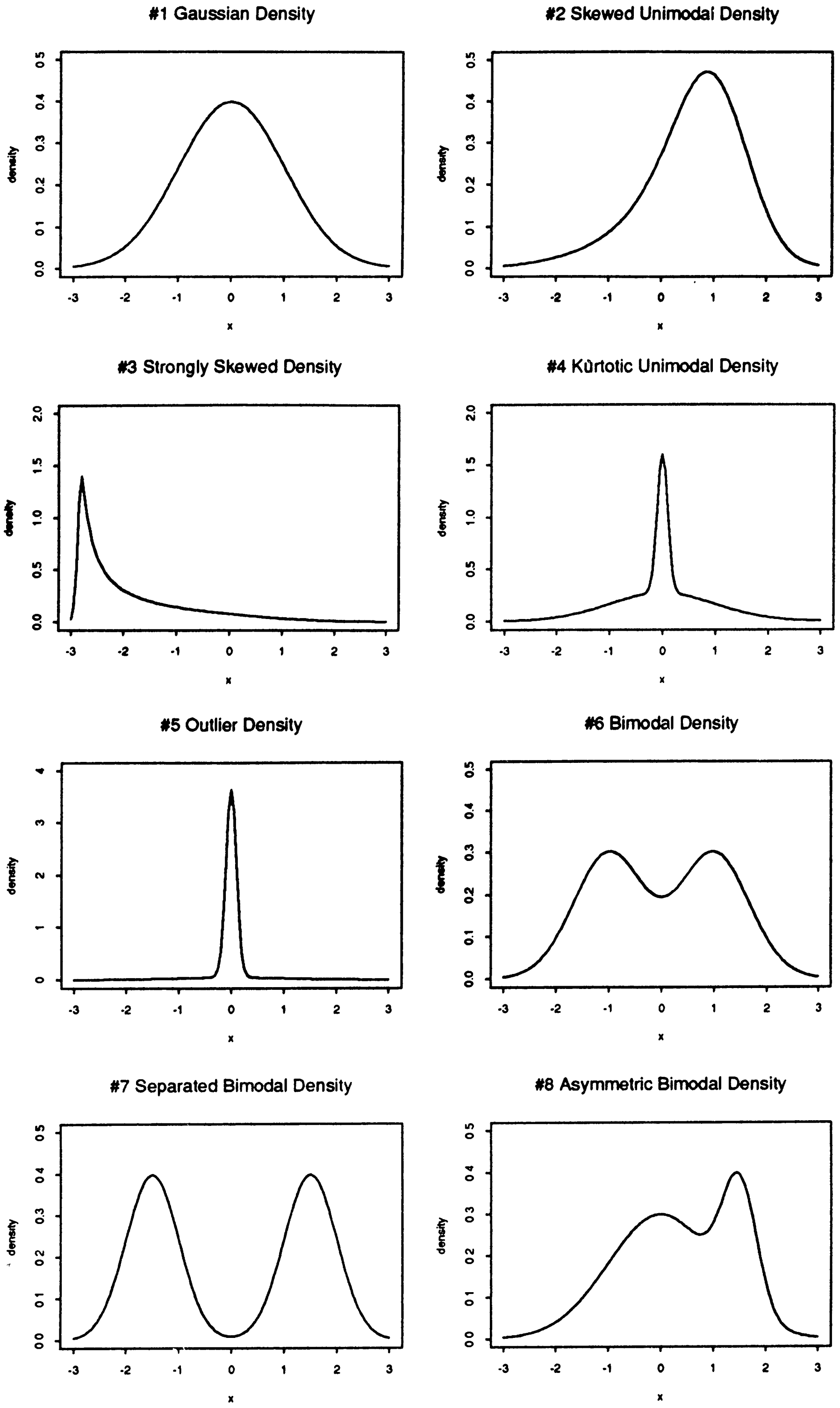

FIG. 1. *Normal mixture densities.*

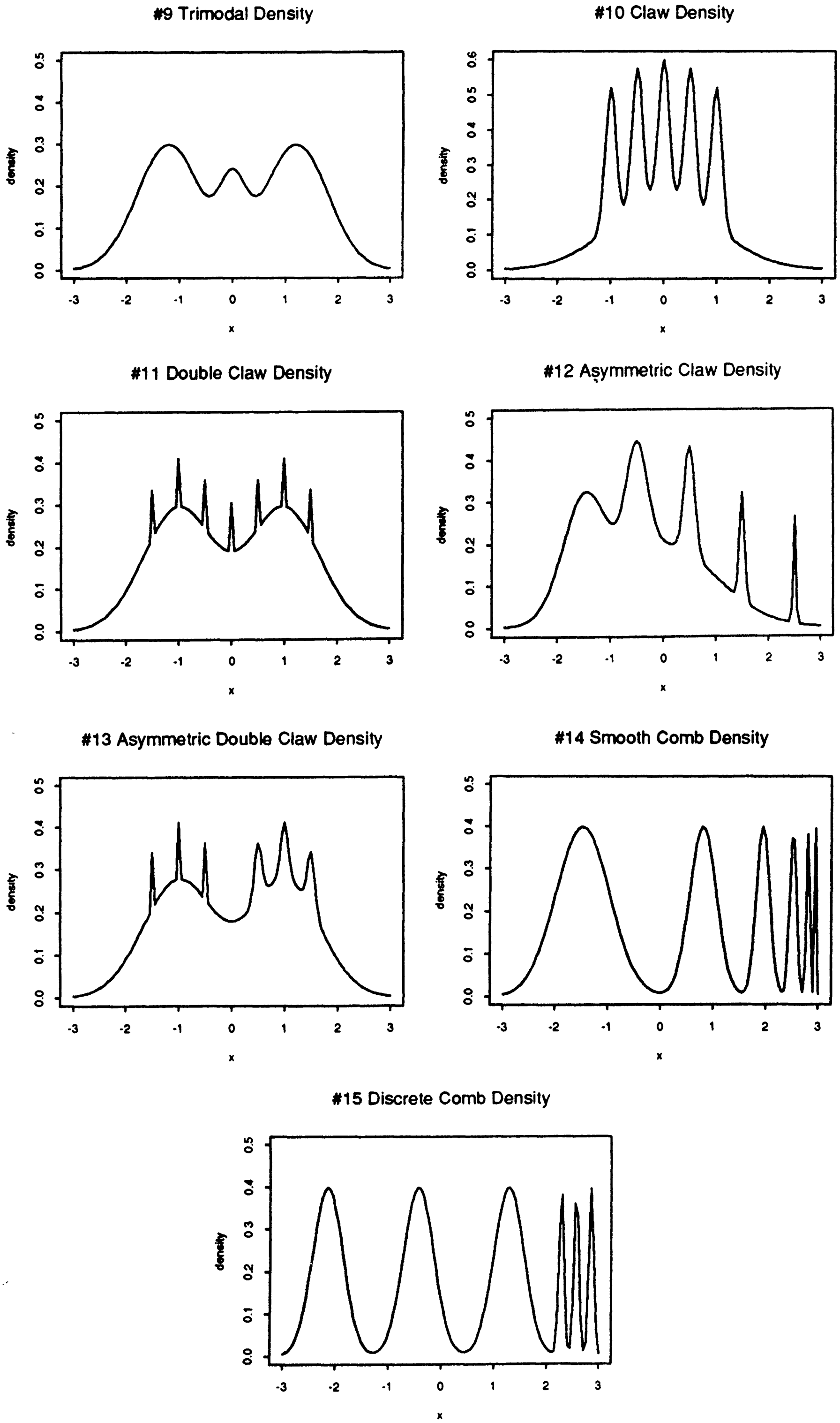

FIG. 1 (continued). *Normal mixture densities.*

In other words, the Hellinger neighborhood is so small that the important characteristics of g are very close to those of f. These conclusions are proved in Lemma 5 of Section 3, by using statistical ideas, because they cannot easily be shown by conventional mathematical methods.

3. Proof.

The idea of the proof of Theorem 1 is to relate the problem of estimating $h_n(f)$ with that of estimating $\theta_2^{-1/5}(f)$ defined by (2.3), via a series of lemmas. These lemmas are established under the conditions of Theorem 1. We will not state them explicitly in the following lemmas.

In the following discussions, we will suppress the dependence of $\theta_j(g)$ on the argument g, whenever the density g is in the Hellinger neighborhood of f. Recall that f is fixed throughout the following proofs.

Lemma 1. *The optimal bandwidth $h_n(f)$ satisfies*

$$\sup_{g \in H_n(f,C)} \frac{h_n(g) - \phi_n(g)}{\phi_n(g)} = o(n^{-1/2}), \tag{3.1}$$

where $\phi_n(g)$ is defined by (2.4).

Proof. Straightforward from calculations as in Section 2 of Hall, Sheather, Jones and Marron (1991). □

Thus, it is intuitively clear that the problem of estimating $h_n(f)$ is equivalent to that of estimating $\phi_n(f)$. The following lemma gives a lower bound for estimating $\theta_2^{-1/5}(f)$.

Lemma 2. *Let $R_{n,C,1}(f)$ be the minimax risk for estimating $\theta_2^{-1/5}(f)$:*

$$R_{n,C,1}(f) = \inf_{\hat{h}_n} \sup_{g \in H_n(f,C)} E_g\left(\hat{h}_n - \theta_2^{-1/5}(g)\right)^2.$$

Then,

$$\lim_{C \to \infty} \liminf_{n \to \infty} nR_{n,C,1}(f) \ge \theta_2^{-2/5}(f) B^2(f), \tag{3.2}$$

where $B(f)$ was defined by (2.6).

Proof. It was shown in the proof of Theorem 2(i) of Bickel and Ritov (1988) that $\theta_2(f)$ is pathwise differentiable along paths

$$\left\{ f_\nu: \left\| \sqrt{f_\nu} - \sqrt{f} \right\|_2 \to 0, \text{ and } \left\| \left(f_\nu^{(4)} - f^{(4)} \right) \sqrt{f} \right\|_2 \to 0 \right\}$$

with the derivative function

$$4\left(f^{(4)}(x) - \theta_2(f) \right) \sqrt{f(x)}\,.$$

Thus, $\theta_2^{-1/5}(f)$ is also pathwise differentiable along such paths with derivative function

$$\left(-\tfrac{1}{5}\theta_2^{-6/5}(f)\right)4\left(f^{(4)}(x) - \theta_2(f)\right)\sqrt{f(x)}\,.$$

As at the end of the proof of Theorem 2(i) of Bickel and Ritov (1988), the information bound for $\theta_2^{-1/5}(f)$ is

$$\begin{aligned}&\left\|-\tfrac{2}{5}\theta_2^{-6/5}(f)\left(f^{(4)}(x) - \theta_2(f)\right)\sqrt{f(x)}\right\|_2^2\\ &\quad= \tfrac{4}{25}\theta_2^{-12/5}\int_{-\infty}^{\infty}\left(f^{(4)}(x) - \theta_2\right)^2 f(x)\,dx\\ &\quad= \tfrac{4}{25}\theta_2^{-12/5}\left(\int_{-\infty}^{\infty}\left(f^{(4)}(x)\right)^2 f(x)\,dx - \theta_2^2\right),\end{aligned}$$

by using the fact that $\theta_2 = \int_{-\infty}^{\infty} f^{(4)}(x) f(x)\,dx$. The result follows by standard semiparametric theory [see Theorem 2.10 of van der Vaart (1988)]. □

In order to show that the second term of $\phi_n(f)$ is negligible, the following lemma gives an estimate of $\theta(f) \equiv \theta_3(f)\theta_2^{-8/5}(f)$.

Lemma 3. *There exists an estimator $\hat{\delta}_n$ such that*

$$\sup_{g\in H_n(f,C)} E_g\left(\hat{\delta}_n - \theta(g)\right)^2 = O\left(n^{-32/85}\right). \tag{3.3}$$

Proof. Note that for $g \in \mathscr{F}_{k+\alpha}$, $g^{(4)}(x)$ is bounded by $g_0(x) \in L_1 \cap L_\infty$. By the construction of Bickel and Ritov (1988) [see Hall and Marron (1991) and Jones and Sheather (1991) for a simpler estimator which can also be used], there exist estimators $\hat{\theta}_2 \geq 0$ and $\hat{\theta}_3$ such that

$$\sup_{g\in H_n(f,C)} E\left(\hat{\theta}_3 - \theta_3\right)^4 = O\left(n^{-4\times 4/17}\right) \tag{3.4}$$

and

$$\sup_{g\in H_n(f,C)} E\left(\hat{\theta}_2 - \theta_2\right)^4 = O\left(n^{-4\times 8/17}\right). \tag{3.5}$$

To guard against zero denominator, we choose

$$\hat{\delta}_n = \frac{\hat{\theta}_3}{\hat{\theta}_2^{8/5} + n^{-4/17}}\,.$$

Then,

$$E\left(\hat{\delta}_n - \theta\right)^2 = E\frac{\left(\theta_2^{8/5}\hat{\theta}_3 - \hat{\theta}_2^{8/5}\theta_3 - n^{-4/17}\theta_3\right)^2}{\left(\hat{\theta}_2^{8/5} + n^{-4/17}\right)^2\theta_2^{16/5}} = I_1 + I_2, \tag{3.6}$$

where

$$I_1 = E\frac{\left(\theta_2^{8/5}\hat{\theta}_3 - \hat{\theta}_2^{8/5}\theta_3 - n^{-4/17}\theta_3\right)^2}{\left(\hat{\theta}_2^{8/5} + n^{-4/17}\right)^2\theta_2^{16/5}} 1_{\{|\hat{\theta}_2 - \theta_2| > \theta_2/2\}}$$

and I_2 is defined with the complementary indicators.

Since I_2 is integrated over the range $|\hat{\theta}_2 - \theta_2| \le \theta_2/2$ and $\hat{\theta}_2 \ge 0$, we have $\hat{\theta}_2 \ge \theta_2/2$. Hence, the denominator in I_2 is bounded away from zero. This leads to

$$\begin{aligned} I_2 &= O\left(E\left(\theta_2^{8/5}\hat{\theta}_3 - \hat{\theta}_2^{8/5}\theta_3 - n^{-4/17}\theta_3\right)^2\right) \\ &= O\left(E\left(\hat{\theta}_3 - \theta_3\right)^2 + E\left(\hat{\theta}_2^{8/5} - \theta_2^{8/5}\right)^2 + n^{-8/17}\right) = O(n^{-8/17}). \end{aligned}$$

Now, let us consider I_1. The fact that $\hat{\theta}_2 \ge 0$ entails that

$$I_1 = O\left(n^{8/17}E\left(\theta_2^{8/5}\hat{\theta}_3 - \hat{\theta}_2^{8/5}\theta_3 - n^{-4/17}\theta_3\right)^2 1_{\{|\hat{\theta}_2 - \theta_2| > \theta_2/2\}}\right).$$

By Hölder's inequality with $p = 5/4$ and $q = 5$,

$$\begin{aligned} I_1 &= O\left(n^{8/17}\left(E\left(\theta_2^{8/5}\hat{\theta}_3 - \hat{\theta}_2^{8/5}\theta_3 - n^{-4/17}\theta_3\right)^{10/4}\right)^{4/5}\left(E1_{|\hat{\theta}_2 - \theta_2| > \theta_2/2}\right)^{1/5}\right) \\ &= O(n^{8/17}n^{-8/17}n^{-32/85}) = O(n^{-32/85}), \end{aligned}$$

where the inequality

$$P\left(\left|\hat{\theta}_2 - \theta_2\right| > \frac{\theta_2}{2}\right) \le \left(\frac{2}{\theta_2}\right)^4 E\left|\hat{\theta}_2 - \theta_2\right|^4 = O(n^{-32/17})$$

was used. This completes the proof. □

The following lemma shows that the minimax lower bound for $\phi_n(f)$ is equivalent to that of $n^{-1/5}c_1\theta_2^{-1/5}$, that is, the second term of $\phi_n(f)$ is indeed negligible.

Lemma 4. *Let $R_{n,C,2}(f)$ be the minimax risk for estimating $\phi_n(f)$:*

$$R_{n,C,2}(f) = \inf_{\hat{h}_n} \sup_{g \in H_n(f,C)} E_g\left(\hat{h}_n - \phi_n(g)\right)^2.$$

Then, we have the lower bound

$$R_{n,C,2}(f) \ge n^{-2/5}c_1^2 R_{n,C,1}(f)(1 + o(1)),$$

where $\xi_{n,C} = o(1)$ means that $\lim_{C\to\infty}\lim_{n\to\infty}\xi_{n,C} = 0$, and c_1 was defined by (2.4).

Proof. Recall that $\theta = \theta_3\theta_2^{-8/5}$ and that $\phi_n(g)$ is given by (2.4). Let $\hat{\delta}_n$ be the estimator defined by Lemma 3 and $c_3 = c_2/c_1$. Then by making the change of variable $\hat{h}_n \to n^{-1/5}c_1(\hat{h}_n + n^{-2/5}c_3\hat{\delta})$,

$$\begin{aligned} R_{n,C,2}(f) &= n^{-2/5}c_1^2 \inf_{\hat{h}_n} \sup_{g\in H_n(f,C)} E\big(\hat{h}_n - \theta_2^{-1/5} + n^{-2/5}c_3(\hat{\delta} - \theta)\big)^2 \\ &\geq n^{-2/5}c_1^2 \inf_{\hat{h}_n} \sup_{g\in H_n(f,C)} \left(E\big(\hat{h}_n - \theta_2^{-1/5}\big)^2 - a_n\sqrt{E\big(\hat{h}_n - \theta_2^{-1/5}\big)^2}\right), \end{aligned}$$

where $a_n = 2c_3n^{-2/5}(E(\hat{\delta}_n - \theta)^2)^{1/2}$. By Lemma 3, we have

$$a_n = O(n^{-2/5-16/85}) = o(n^{-1/2}).$$

Thus,

$$R_{n,C,2}(f) \geq n^{-2/5}c_1^2 \inf_{\hat{h}_n}\big(q^2(\hat{h}_n) - a_n q(\hat{h}_n)\big), \tag{3.7}$$

where

$$q(\hat{h}_n) = \left(\sup_{g\in H_n(f,C)} E\big(\hat{h}_n - \theta_2^{-1/5}\big)^2\right)^{1/2}.$$

By Lemma 2, for any estimator $\hat{h}_n$,

$$q(\hat{h}_n) \geq R_{n,C,1}^{1/2} \geq \tfrac{1}{2}\theta_2^{-1/5}(f)B(f)n^{-1/2},$$

where n and C are large. This entails that $q(\hat{h}_n) > a_n$ for large n and C. Since the quadratic function $x^2 - a_n x$ is increasing for $x > a_n/2$ and $R_{n,C,1}^{1/2} = \inf_{\hat{h}_n} q(\hat{h}_n)$, we obtain that

$$\begin{aligned} \inf_{\hat{h}_n}\big(g^2(\hat{h}_n) - a_n g(\hat{h}_n)\big) &= \inf_{\hat{h}_n} g^2(\hat{h}_n) - a_n \inf_{\hat{h}_n} g(\hat{h}_n) \\ &= R_{n,C,1} - a_n\sqrt{R_{n,C,1}} \\ &= R_{n,C,1}(1 + o(1)). \end{aligned} \tag{3.8}$$

The conclusion follows from (3.7) and (3.8). □

LEMMA 5. *On the Hellinger ball $H_n(f,C)$, we have*

$$\lim_{n\to\infty} \sup_{g\in H_n(f,C)} \left|\frac{h_n(g)}{h_n(f)} - 1\right| = 0.$$

PROOF. By Lemma 1, $h_n(g)$ and $h_n(f)$ can be approximated by $\phi_n(g)$ and $\phi_n(f)$, respectively. Hence, we need only to prove (2.8). By a useful statistical lower bound [see for example page 18 of Fan (1989)], for any estimator $\hat{T}_n$, we have

$$\begin{aligned} &\sup_{g\in H_n(f,C)} E\left|\hat{T}_n - g^{(j)}(x)\right|^2 \\ &\quad\geq \frac{1 - \sqrt{1 - e^{-2C}}}{2} \sup_{g\in H_n(f,C)} \left|g^{(j)}(x) - f^{(j)}(x)\right|^2. \end{aligned} \tag{3.9}$$

Since g has more than four derivatives, there exist estimators [e.g., kernel density estimators (2.1)] such that $g^{(j)}(x)$ $(j = 0, \dots, 4)$ can be estimated consistently, that is, such that the left-hand side of (3.9) converges to 0. Thus,

$$\sup_{g \in H_n(f, C)} |g^{(j)}(x) - f^{(j)}(x)| \to 0, \quad \text{for } j = 0, \dots, 4.$$

Now, by the dominate convergence theorem,

$$\begin{aligned}
&\sup_{g \in H_n(f, C)} \left| \int_{-\infty}^{\infty} (g''(x))^2 \, dx - \int_{-\infty}^{\infty} (f''(x))^2 \, dx \right| \\
&\quad = \sup_{g \in H_n(f, C)} \left| \int_{-\infty}^{\infty} g^{(4)} g - \int_{-\infty}^{\infty} f^{(4)} f \right| \\
&\quad \le \int_{-\infty}^{\infty} f \sup_{g \in H_n(f, C)} |g^{(4)} - f^{(4)}| + \int_{-\infty}^{\infty} g_0 \sup_{g \in H_n(f, C)} |g - f|, \\
&\quad \to 0,
\end{aligned}$$

where $|g^{(4)}| \le g_0$ (see the definition of $\mathscr{F}_{k+\alpha}$) was used in the last inequality. This completes the proof. □

Proof of Theorem 1. Write $h_n(g) = \phi_n(g) + \xi_n(g)$, where by Lemma 1,

$$\sup_{g \in H_n(f, C)} \xi_n(g) = o(n^{-1/2 - 1/5}).$$

Lemma 5 entails that

$$\begin{aligned}
&\inf_{\hat{h}_n} \sup_{g \in H_n(f, C)} E_g \left(\frac{\hat{h}_n - h_n(g)}{h_n(g)} \right)^2 \\
&\quad \ge \inf_{\hat{h}_n} \left(\sup_{g \in H_n(f, C)} E_g \big(\hat{h}_n - h_n(g) \big)^2 \Big/ \sup_{g \in H_n(f, C)} h_n^2(g) \right) \\
&\quad = \inf_{\hat{h}_n} \sup_{g \in H_n(f, C)} E_g \big(\hat{h}_n - \phi_n(g) - \xi_n(g) \big)^2 n^{2/5} c_1^{-2} \theta_2^{2/5}(f)(1 + o(1)).
\end{aligned}$$

Using this together with the argument used at the end of the proof of Lemma 4, we can show that $\xi_n(g)$ is in deed negligible and conclude that

$$\inf_{\hat{h}_n} \sup_{g \in H_n(f, C)} E_g \left(\frac{\hat{h}_n - h_n(g)}{h_n(g)} \right)^2 \ge n^{2/5} c_1^{-2} \theta_2^{2/5}(f) R_{n, c, 2}(f)(1 + o(1)).$$

The conclusion follows directly from Lemmas 4 and 2. □

Proof of Theorem 2. Denote

$$r = \int_{-\infty}^{\infty} K^2(x) \, dx, \quad \text{and} \quad \mu = \int_{-\infty}^{\infty} x^2 K(x) \, dx.$$

By using the fact that $M'(h_n(g)) = 0$, we have

$$M(\hat{h}_n) - M(h_n(g)) = \tfrac{1}{2} M''(\tilde{h})\big(\hat{h}_n - h_n(g)\big)^2, \tag{3.10}$$

where $\tilde{h}$ lies between $\hat{h}_n$ and $h_n(g)$. Remark that [see Hall, Sheather, Jones and Marron (1991)]

$$\begin{aligned} M''(h) &= 2rn^{-1}h^{-3} + 3h^2\mu^2\theta_2 + O\big((nh)^{-1} + h^4\big) \\ &\geq 5r^{2/5}\mu^{6/5}\theta_2^{3/5}n^{-2/5}(1 + o(1)) \end{aligned} \tag{3.11}$$

and

$$M(h_n(g)) = \tfrac{5}{4} r^{4/5}\mu^{2/5}\theta_2^{1/5}n^{-4/5}(1 + o(1)). \tag{3.12}$$

Expression (3.11) entails that

$$M''(\tilde{h}) \geq 5r^{2/5}\mu^{6/5}\theta_2^{3/5}n^{-2/5}(1 + o(1)). \tag{3.13}$$

Combination of Lemma 1, (3.12) and (3.13) gives

$$\frac{M''(\tilde{h})h_n^2(g)}{2M(h_n(g))} \geq 2 + o(1).$$

This together with (3.10) leads to

$$\begin{aligned} &\inf_{\hat{h}_n} \sup_{g \in H_n(f,C)} n^2 E_g \left(\frac{M(\hat{h}_n) - M(h_n(g))}{M(h_n(g))} \right)^2 \\ &\quad = \inf_{\hat{h}_n} \sup_{g \in H_n(f,C)} n^2 E_g \left(\frac{M''(\tilde{h})h_n^2(g)}{2M(h_n(g))} \right)^2 \left(\frac{\hat{h}_n - h_n(g)}{h_n(g)} \right)^4 \\ &\quad \geq \inf_{\hat{h}_n} \sup_{g \in H_n(g,C)} n^2 E_g \left(\frac{\hat{h}_n - h_n(g)}{h_n(g)} \right)^4 (4 + o(1)) \\ &\quad \geq \left(\inf_{\hat{h}_n} \sup_{g \in H_n(f,C)} n E_g \left(\frac{\hat{h}_n - h_n(g)}{h_n(g)} \right)^2 \right)^2 (4 + o(1)). \end{aligned}$$

The conclusion follows directly from Theorem 1. □

REFERENCES

Bickel, P. J., Klassen, C. A. J., Ritov, Y. and Wellner, J. A. (1991). *Efficient and Adaptive Inference in Semi-parametric Models*. Johns Hopkins Univ. Press. To appear.

Bickel, P. J. and Ritov, Y. (1977). Estimating integrated squared density derivatives: Sharp best order of convergence estimates. *Sankhyā Ser. A* **50** 381–393.

Chiu, S.-T. (1991). Bandwidth selection for kernel density estimation. *Ann. Statist.* **19** 1883–1905.

Donoho, D. L. and Nussbaum, M. (1990). Minimax quadratic estimation of a quadratic functional. *J. Complexity* **6** 290–323.

Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression*. Dekker, New York.

Fan, J. (1989). Contributions to the estimation of nonregular functionals. Dissertation, Univ. California, Berkeley.

FAN, J. (1991). On the estimation of quadratic functionals. *Ann. Statist.* **19** 1273–1294.

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

HALL, P. and MARRON, J. S. (1991). Lower bounds for bandwidth selection in density estimation. *Probab. Theory Related Fields*. To appear.

HALL, P., SHEATHER, S. J., JONES, M. C. and MARRON, J. S. (1991). On optimal data-based bandwidth selection in kernel density estimation. *Biometrika* **78** 263–269.

HÄRDLE, W. (1990). *Applied Nonparametric Regression*. Cambridge Univ. Press.

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

JONES, M. C. and SHEATHER, S. J. (1991). Using nonstochastic terms to advantage in kernel based estimation of integrated squared density derivatives. *Statist. Probab. Lett.* **11** 511–514.

MARRON, J. S. (1988). Automatic smoothing parameter selection: a survey. *Empirical Economics* **13** 187–208.

MARRON, J. S. and WAND, M. P. (1992). Exact mean integrated squared error. *Ann. Statist.* **20** 712–736.

MÜLLER, H.-G. (1988). *Nonparametric Analysis of Longitudinal Data*. Springer, Berlin.

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

VAN DER VAART, A. W. (1988). *Statistical Estimation in Large Parameter Spaces*. *CWI Tract* **44**. Math. Centrum, Amsterdam.

WAHBA, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia.

DEPARTMENT OF STATISTICS
UNIVERSITY OF NORTH CAROLINA
CHAPEL HILL, NORTH CAROLINA 27599-3260