Statistica Sinica **10**(2000), 457-473

INFORMATION BOUND FOR BANDWIDTH SELECTION IN KERNEL ESTIMATION OF DENSITY DERIVATIVES

Tiee-Jian Wu*† and Yue Lin†

**National Cheng-Kung University and †University of Houston*

Abstract: Based on a random sample of size n from an unknown density f on the real line, several data-driven methods for selecting the bandwidth in kernel estimation of $f^{(k)}$, $k = 0, 1, \ldots$, have recently been proposed which have a very fast asymptotic rate of convergence to the optimal bandwidth, where $f^{(k)}$ denotes the kth derivative of f. In particular, for all k and sufficiently smooth f, the best possible relative rate of convergence is $O_p(n^{-1/2})$. For $k = 0$, Fan and Marron (1992) employed semiparametric arguments to obtain the best possible constant coefficient, that is, an analog of the usual Fisher information bound, in this convergence. The purpose of this paper is to show that their arguments can be extended to establish information bounds for all k. The extension from the special case $k = 0$ to the case of general k requires some nontrival work and gives a significant benchmark as to how well a bandwidth selector can hope to perform in kernel estimation of $f^{(k)}$.

Key words and phrases: Bandwidth selection, density derivatives, kernel estimates, nonparametric information bounds, semiparametric methods.

1. Introduction

Let $X_1, \ldots, X_n$ be a random sample from an unknown density f. Let $g^{(k)}$ denote the kth derivative of any function g. The kernel estimate of $f^{(k)}(x)$ takes the form

$$\hat{f}_h^{(k)}(x) = (nh^{k+1})^{-1} \sum_{j=1}^{n} w^{(k)}\{(x - X_j)/h\}, \qquad -\infty < x < \infty, \tag{1.1}$$

where $w(\cdot)$ is a symmetric probability density (called the kernel function) and $h = h_{n,k}$ is a global bandwidth (i.e., smoothing parameter) satisfying $h \to 0$ and $nh^{k+1} \to \infty$ as $n \to \infty$. Practical application of (1.1) for any k is crucially dependent on the choice of h. If h is too small, the resulting (1.1) is subject to too much sample variation, and a curve which is too rough. In contrast, if h is too large, the resulting (1.1) will have an unacceptably large bias; important features of the underlying curve are smoothed away. Although subjective choice of h is sufficient for many cases, the usefulness of (1.1) would be greatly enhanced if an efficient and data-based method of selecting h could be agreed upon (see Silverman (1986) for further discussion).

The kernel density estimate $\hat{f}_h$ (the special case $k = 0$ here) is very useful in the exploration and presentation of data; see, e.g., Silverman (1986). The estimates $\hat{f}_h^{(k)}, k \geq 1$, can be used to evaluate modes and inflection points and to obtain plug-in bandwidth selectors for kernel estimation of $f(x)$. They can also be applied to the estimation of scores in certain additive models; see Härdle, Hart, Marron and Tsybakov (1992) and Härdle and Stoker (1989). Another application is to the empirical verification of uniqueness of equilibria of market demand in econometrics, where the estimation of derivatives of densities enters through so-called income effects; see Hildenbrand and Hildenbrand (1986). Here we remark that the kernel estimate (1.1) is popular because it is simple and easy to implement in practice. However, there are situations where more robust estimates merit consideration. For example, the local polynomial estimate of $f^{(k)}$ may be used if f has compact support; see Fan, Gijbels, Hu and Huang (1996) and Cheng, Fan and Marron (1997). The wavelet-based estimate of $f^{(k)}$ is more suitable if $f^{(k)}$ is smooth in some piecewise manner; see, e.g., Hall and Patil (1995) and Donoho, Johnstone, Kerkyacharian and Picard (1996).

The performance of (1.1) is commonly measured by the mean integrated squared error

$$MISE_k(h)\ (= M_k(h))\ = E\int_{-\infty}^{\infty}\{\hat{f}_h^{(k)}(x) - f^{(k)}(x)\}^2 dx. \tag{1.2}$$

Here we take the optimal bandwidth $h_k(f)$ to be the minimizer of M_k. For simplicity of notation the dependency of M_k, $h_k(f)$, etc. on n is suppressed throughout the paper. The readers are referred to Hall and Marron (1991), Jones (1991) and Grund, Hall and Marron (1994) for discussion of other measures of error, including reasons why our present goodness criterion (1.2) is sensible. In practice, the $h_k(f)$ is unavailable and needs to be estimated. However, $h_k(f)$ is of order $O(n^{-1/(2k+5)})$, which increases in k; see Stone (1980).

Many data-based bandwidth selectors for kernel estimation of f have been proposed over the past decade. See the survey papers by Jones, Marron and Sheather (1996), Loader (1995) and Marron (1988). Hall and Marron (1991) proved that for a data-driven bandwidth selector, the best possible relative convergence rate is $O_p(n^{-1/2})$, in a minimax sense. Bandwidth selectors that achieve this rate include the selectors of Hall, Marron and Park (1992), Jones, Marron and Park (1991), Hall, Sheather, Jones and Marron (1991) and Chiu (1991, 1992). Motivated by the fact that there are several competing selectors, Fan and Marron (1992) employed semiparametric arguments and calculated the best possible constant coefficient $B_0^2(f)$ (see (1.3) herein) in this convergence. This is an extension of the classical Fisher information ideas (see Bickel, Klassen, Ritov and Wellner (1991) and van der Vaart (1988) for details), which gives a benchmark as to how well a bandwidth selector can hope to perform in the kernel estimation of f. As noted by Fan and Marron (1992), among the preceding selectors, only the selectors by Chiu (1991, 1992) and by Hall, Sheather, Jones and Marron (1991) achieve the best possible constant coefficient $B_0^2(f)$, and the others have larger constants, and hence are not optimal in this sense.

For $k \geq 1$, Härdle, Marron and Wand (1990) proposed a cross-validated bandwidth selector for (1.1) that has a slow convergence rate. Recently, Wu (1997) proposed two types of data-based bandwidth selectors for (1.1) that achieve the best $O_p(n^{-1/2})$ relative convergence rate to the optimal $h_k(f)$, and, moreover, the asymptotic variance of the relative error of the selectors is the same as $n^{-1}B_k^2(f)$, where

$$B_k^2(f) = 4\text{Var}\,\{f^{(2k+4)}(X_1)\}/\{(2k+5)\theta_{k+2}(f)\}^2, \tag{1.3}$$

with

$$\theta_j(g) = \int_{-\infty}^{\infty} \{g^{(j)}(x)\}^2 dx, \qquad j \geq 0, \tag{1.4}$$

for any function g. It is conjectured in Wu (1997) that $B_k^2(f)$ is the best possible constant coefficient. Thus for all k, the selectors of Wu (1997) are optimal in this sense.

The purpose of this paper is to prove the validity of the above conjecture. This extends the information bound $n^{-1}B_0^2(f)$ obtained by Fan and Marron (1992) to a general k. The method of proof involves heavy semiparametric arguments, and is an extension of the arguments in Fan and Marron (1992). Here we remark that although perhaps only $k = 0,\ 1$ and 2 will come close to what will be considered "practically important" cases, there is no harm in having this machinery available for all k, and giving a unified approach. Section 2 contains the main results (Theorems 1 and 2) and a numerical example. This example shows that $B_k(f)$ usually (but not always) increases with k, although probably not excessively so. Proofs are deferred to the Appendix.

2. Main Results

The problem of estimating $h_k(f)$ is closely related to that of estimating quadratic functionals $\theta_{k+2}(f), \ldots, \theta_{k+[(k+6)/2]}(f)$. Here and below, $[x]$ denotes the greatest integer $\leq x$. Indeed, from Section 1.5 of Wu (1997) we see that the optimal bandwidth $h_k(f)$ can be approximated by

$$\phi_k(f) = h_{k,S}(f) + Q_k(f), \tag{2.1}$$

where

$$h_{k,S}(f)=c_1\{\theta_{k+2}(f)\}^{-1/(2k+5)}n^{-1/(2k+5)},\ Q_k(f)=c_1 \sum_{l=1}^{[(k+2)/2]} \delta_l(f)n^{-(2l+1)/(2k+5)},$$

$$(2.2)$$

$$c_1 = \{(2k+1)\theta_k(w)\mu_2^{-2}\}^{1/(2k+5)}, \qquad \mu_t = \int_{-\infty}^{\infty} |x|^t w(x)dx, \ t \geq 0, \tag{2.3}$$

and $\delta_1(f), \ldots, \delta_{[(k+2)/2]}(f)$ are constants depending on the unknown quadratic functionals $\theta_{k+2}(f), \ldots, \theta_{k+[(k+6)/2]}(f)$. (For explicit expressions of the $\delta_l(f)$'s, see Hall, Sheather, Jones and Marron (1991) for $k = 0$, and Remark 1 herein for $k \geq 1$.) This reduces the problem of estimating the optimal bandwidth to that of estimating these quadratic functionals.

Recent work on estimating $\theta_k(f)$ includes Hall and Marron (1987a, 1991), Bickel and Ritov (1988), Jones and Sheather (1991), Aldershof (1991), Cheng (1997) and Wu (1995), among others. Results on estimating $\theta_k(f)$ with f being supported in $[0, 1]$ are given by Fan (1991), who dealt with a white noise model (see also Donoho and Nussbaum (1990)) and by Goldstein and Messer (1992).

Let us denote a class of densities having $j + \alpha \geq 2k + 4$ derivatives by

$$\mathcal{F}_{j+\alpha} = \{g : |g^{(j)}(x) - g^{(j)}(y)| \leq M|x - y|^{\alpha}, \ |g^{(2k+4)}(x)| \leq g_0(x)\},$$

where $0 \leq \alpha \leq 1$, $M > 0$ is a constant, and $g_0(x)$ is a bounded, continuous, and integrable function on $(-\infty, \infty)$ (this ensures that $\mathrm{Var}\,\{f^{(2k+4)}(X_1)\} < \infty$ and, consequently, $B_k(f) < \infty$ if $f \in \mathcal{F}_{j+\alpha}$). Let

$$H_n(f, C) = \{g \in \mathcal{F}_{j+\alpha} : \| g^{1/2} - f^{1/2} \|_2 \leq C/n^{1/2}\}, \qquad C > 0, \tag{2.4}$$

be a Hellinger ball in the neighborhood of f, where $|| \cdot ||_2$ denotes the usual L_2-norm. The following theorem shows that, for all k, the asymptotic relative error of any bandwidth selector for kernel estimation of $f^{(k)}$ cannot be smaller than $n^{-1/2}B_k(f)$.

Theorem 1. *Assume $\mu_{2[(k+6)/2]} < \infty$, $\theta_k(w) < \infty$, and $f \in \mathcal{F}_{j+\alpha}$ with $j + \alpha > 2k + 4$. Then, for any bandwidth selector $\hat{h}_k$,*

$$\lim_{C\to\infty} \liminf_{n\to\infty} \inf_{\hat{h}_k} \sup_{g \in H_n(f,C)} nE_g\Big\{\frac{\hat{h}_k - h_k(g)}{h_k(g)}\Big\}^2 \geq B_k^2(f). \tag{2.5}$$

Theorem 1 generalizes Theorem 1 of Fan and Marron (1992). Note that $B_k(f)$ does not depend on the kernel function w, even though the optimal bandwidth $h_k(f)$ does. This indicates $B_k(f)$ is a measure of the intrinsic difficulty of bandwidth selection in kernel estimation of $f^{(k)}$.

The following theorem establishes the close relationship between the relative error of $MISE$ (1.2) and that of a bandwidth selector. Let $\mathcal{F}^*_{j+\alpha}$ denote the class of densities having $j + \alpha \geq k + 4$ derivatives that results from replacing the condition $|g^{(2k+4)}(x)| \leq g_0(x)$ by the condition $|g^{(k+4)}(x)| \leq g_0(x)$ in the definition of $\mathcal{F}_{j+\alpha}$.

Theorem 2. *Assume $\mu_4 < \infty$, $\theta_k(w) < \infty$, and $f \in \mathcal{F}^*_{j+\alpha}$. Then, for any $0 \leq \delta \leq 1$,*

$$n^\delta \Big\{ \frac{M_k(\hat{h}_k) - M_k(h_k(f))}{M_k(h_k(f))} \Big\} = 2(2k+1) n^\delta \Big\{ \frac{\hat{h}_k - h_k(f)}{h_k(f)} \Big\}^2 + o_p(1) \tag{2.6}$$

provided that the bandwidth selector $\hat{h}_k$ is relatively consistent at rate $o_p(n^{-\rho})$, i.e., $\hat{h}_k/h_k(f) = 1 + o_p(n^{-\rho})$, where $\rho = \max\{\delta/3,\ \delta/2 - 1/(2k+5),\ \delta/2 - (k+0.5)/(2k+5)\}$.

The result (2.6) is well known when $k = 0$ (see e.g., Hall and Marron (1987b) or Jones (1991), among others). In typical situations, if f is sufficiently smooth, there exist $0 < \delta \leq 1$ and $\sigma_k^2(f)$ (both depend on $\hat{h}_k$) such that $n^{\delta/2}\{\hat{h}_k/h_k(f) - 1\}$ has a limiting $N(0, \sigma_k^2(f))$ distribution and, consequently, $n^\delta\{M_k(\hat{h}_k)/M_k(h_k(f)) - 1\}$ has a limiting $\{2(2k+1)\sigma_k^2(f)\}\cdot\chi_1^2$ distribution, as implied by (2.6). Therefore, the asymptotic relative error (or the asymptotic relative bias) of $MISE$ is measured by the constant coefficient $2(2k+1)n^{-\delta}\sigma_k^2(f)$, which, under the conditions of Theorem 1, can not be smaller than $2(2k+1)n^{-1}B_k^2(f)$, as implied by (2.5) (see Remark 3 herein for further discussion).

Theorems 1 and 2 indicate that, for any k, the quantity $B_k(f)$ measures the difficulty of the bandwidth selection problem in kernel estimation of $f^{(k)}$. The larger the $B_k(f)$, the harder the problem. On the other hand, for any f one wonders whether the quantity $B_k(f)$ increases rapidly with k or not. This is probably quite hard to work out in any generality, but is easy when restricting attention to numerical examples. Table 1 provides such an example. It shows the values of $B_0(f)$, $B_1(f)$, $B_2(f)$, $B_1^2(f)/B_0^2(f)$, $B_2^2(f)/B_1^2(f)$, $3B_1^2(f)/B_0^2(f)$ and $5B_2^2(f)/3B_1^2(f)$ for the 15 normal mixture densities in Marron and Wand (1992). These densities typify many different types of challenges to curve estimators (see their paper for details of these densities). The $B_0(f)$, obtained by Fan and Marron (1992), is included for ease of comparison. The $B_i(f)$, $i = 1, 2$, for densities (#1) − (#2) and (#4) − (#7), obtained by Wu (1997), are also included for the sake of completeness.

We point out that both $D_{k+j,k} = B_{k+j}^2(f)/B_k^2(f)$ (according to Theorem 1) and $D^*_{k+j,k} = \{(2k+2j+1)/(2k+1)\}D_{k+j,k}$ (according to Theorems 2 and 1) can be used to measure the asymptotic relative difficulty of optimal bandwidth selection in kernel estimation of $f^{(k+j)}$ to $f^{(k)}$, $k \geq 0$, $j \geq 0$. Evidently, $D^*_{k+j,k}$ is more sensitive to the value of j than $D_{k+j,k}$ is, because of the factor $\{(2k+2j+1)/(2k+1)\}$ contained in $D^*_{k+j,k}$.

Table 1 not only gives us an idea as to how difficult it is to select a bandwidth for a variety of densities and derivatives, but also implies that this problem usually (but not always) gets worse with k, although probably not exceedingly so. For example, if $f = N(0,1)$ (density (#1)), then asymptotically f' is $D_{1,0} = 2.16$ (or $D^*_{1,0} = 6.48$) times as difficult as f, and f'' is $D_{2,1} = 2.45$ (or $D^*_{2,1} = 4.08$) times as difficult as f' in terms of bandwidth selection. The best selector for f'' with sample size $n = 1059$ would have approximately the same accuracy of estimating optimal bandwidth as for f' with $n = 432$, and f with $n = 200$. On the other hand, as $n \to \infty$, the density itself, the first derivative and the second derivative of density (#4) (a kurtotic unimodal density) are respectively $(2.638/1.300)^2 \approx 4.12$, $(3.782/1.911)^2 \approx 3.92$ and $(5.885/2.993)^2 \approx 3.87$ times as difficult as those of the $N(0,1)$ density in bandwidth selection terms. In contrast, the corresponding values for density (#13) (an asymmetric double claw density), upon comparing with the $N(0,1)$ density, are respectively $(25.59/1.300)^2 \approx 387.48$, $(36.48/1.911)^2 \approx 364.41$ and $(56.61/2.993)^2 \approx 357.74$.

Table 1. Constant factors in the lower bounds.

Density	$B_0(f)$	$B_1(f)$	$B_2(f)$	$\frac{B_1^2(f)}{B_0^2(f)}$	$\frac{B_2^2(f)}{B_1^2(f)}$	$\frac{3B_1^2(f)}{B_0^2(f)}$	$\frac{5B_2^2(f)}{3B_1^2(f)}$
#1	1.300	1.911	2.993	2.16	2.45	6.48	4.08
#2	1.771	2.842	4.728	2.58	2.77	7.74	4.62
#3	4.973	8.097	12.84	2.65	2.51	7.95	4.18
#4	2.638	3.782	5.885	2.06	2.42	6.18	4.03
#5	1.388	2.033	3.181	2.15	2.45	6.45	4.08
#6	1.868	1.841	2.105	0.97	1.31	2.91	2.18
#7	1.286	1.942	2.982	2.28	2.36	6.84	3.93
#8	3.390	4.751	7.337	1.96	2.38	5.88	3.97
#9	4.742	7.762	12.67	2.68	2.66	8.04	4.43
#10	2.125	3.591	5.800	2.86	2.61	8.58	4.35
#11	19.39	27.57	42.79	2.02	2.41	6.06	4.02
#12	9.635	15.10	24.01	2.46	2.53	7.38	4.22
#13	25.59	36.48	56.61	2.03	2.41	6.09	4.02
#14	9.408	14.91	23.65	2.51	2.52	7.53	4.20
#15	3.515	5.214	7.903	2.20	2.30	6.60	3.83

Remark 1. From Section 1.5 of Wu (1997), we have

$$\delta_l(f) = \{\theta_{k+2}(f)\}^{-1/(2k+5)} A_l.$$

with

$$A_1 = c_1^2 \{\theta_{k+2}(f)\}^{-2/(2k+5)} J_1, \quad A_2 = \{k + 5 - (J_2/J_1^2)\} A_1^2,$$

and, in general,

$$A_l = \sum_{j=0}^{l}(-1)^{j+1}(J_j/J_1^j)A_1^j\sum_{R_j}(2k+5+2j)!\{r_1!\cdots r_{l-1}!(2k+5+2j \\ -\sum_{m=1}^{l-1} r_m)!\}^{-1}A_1^{r_1}\cdots A_{l-1}^{r_{l-1}}$$

where $R_j = \{(r_1,\ldots,r_{l-1}) : 0 \le r_1,\ldots,r_{l-1} \le l, \sum_{m=1}^{l-1} mr_m = l-j\}$,

$$J_j\ (= J_j(f)) = 2(2+j)b_{2+j}\theta_{k+2+j}(f)/\{(2k+5)\mu_2^2\theta_{k+2}(f)\},$$

$$b_m = \sum_{i=1}^{m-1}\mu_{2i}\mu_{2m-2i}/\{(2i)!(2m-2i)!\}, \quad m = 2,3,\ldots. \tag{2.7}$$

Furthermore, under the conditions of Theorem 1, it is evident that

$$\{h_{k,S}(f) - \phi_k(f)\}/\phi_k(f) = O(n^{-2/(2k+5)}). \tag{2.8}$$

Remark 2. We note that in Theorem 1, the smoothness condition that $f \in \mathcal{F}_{j+\alpha}$ with $j+\alpha > 2k+4$ may be replaced by the condition $p > 2k+5$, where p denotes the decay rate of the characteristic function $\psi_f(\lambda)$ of f, that is, $|\lambda|^p|\psi_f(\lambda)| = O(1)$ as $|\lambda| \to \infty$. These two conditions are analogous and compatible with each other, but neither one is weaker or stronger than the other.

Remark 3. For both the case $\hat{h}_k = \hat{h}_{k,ST}$ and the case $\hat{h}_k = \hat{h}_{k,EP}$, where $\hat{h}_{k,ST}$ and $\hat{h}_{k,EP}$ are respectively the stabilized- and the extended plug-in bandwidth selectors of Wu (1997), the limiting distribution of $n\{M_k(\hat{h}_k)/M_k(h_k(f)) - 1\}$ is $\{2(2k+1)B_k^2(f)\}\cdot\chi_1^2$, as can be seen from Theorems 1 to 4 of Wu (1997) and (2.6) above. Consequently, the asymptotic relative error (or the asymptotic relative bias) of $M_k(\hat{h}_k)$ is the same as $2(2k+1)n^{-1}B_k^2(f)$, which is the best possible constant coefficient implied by Theorems 2 and 1. Similarly, for estimating f (when $k = 0$), the $MISE$ evaluated at the selectors of Chiu (1991, 1992) and of Hall, Sheather, Jones and Marron (1991) can be shown to achieve the best possible constant $2n^{-1}B_0^2(f)$. Furthermore, based on the optimality of $\hat{h}_{k,ST}$ and the fact that the second moment of a χ_1^2 distribution is 3, we conjecture that under the assumption of Theorem 1, for any $k \ge 0$ and bandwidth selector $\tilde{h}_k$, the inequality

$$\lim_{C\to\infty}\liminf_{n\to\infty}\inf_{\tilde{h}_k}\sup_{g\in H_n(f,C)} n^2E_g\Big\{\frac{M_k(\tilde{h}_k) - M_k(h_k(g))}{M_k(h_k(g))}\Big\}^2 \ge 12(2k+1)^2B_k^4(f) \tag{2.9}$$

holds and the lower bound $12(2k+1)^2B_k^4(f)$ is the sharpest possible.

Remark 4. Since the special cases $k = 1$ and $k = 2$ have pratical applications, we write the implications of Theorems 1 and 2 explicitly. For kernel estimation of f', the asymptotic relative error of any bandwidth selector and that of its $MISE$ can not be smaller than $n^{-1/2}B_1(f)$ and $6n^{-1}B_1^2(f)$, respectively, where

$$B_1^2(f) = \frac{4}{49}\left\{\frac{\int_{-\infty}^{\infty}(f^{(6)}(x))^2 f(x)dx}{(\int_{-\infty}^{\infty}(f^{(3)}(x))^2 dx)^2} - 1\right\}.$$

Likewise, for kernel estimation of f'', the asymptotic relative error of any bandwidth selector and that of its $MISE$ can not be smaller than $n^{-1/2}B_2(f)$ and $10n^{-1}B_2^2(f)$, respectively, where

$$B_2^2(f) = \frac{4}{81}\left\{\frac{\int_{-\infty}^{\infty}(f^{(8)}(x))^2 f(x)dx}{(\int_{-\infty}^{\infty}(f^{(4)}(x))^2 dx)^2} - 1\right\}.$$

3. Concluding Remarks

In this article we have established the information bound for bandwidth selection in kernel estimation of $f^{(k)}$. Our results are formulated only in terms of nonnegative kernel functions because they are used almost exclusively in practice. Other reasons for using nonnegative kernels can be found in Fan and Marron (1992) and Marron and Wand (1992), among others. Furthermore, Theorem 1 suggests that the constant $nB_k^{-2}(f)$ plays a role similar to the classical Fisher information number contained in a sample of size n, so one can define the efficiency of any bandwidth selector $\hat{h}_k$ by

$$eff(\hat{h}_k) = n^{-1}B_k^2(f)/E_f\{\hat{h}_k/h_k(f) - 1\}^2.$$

The root n bandwidth selectors proposed by Wu (1997) are optimal since the asymptotic variance of the relative error of his selectors is the same as $n^{-1}B_k^2(f)$. This provides a strong sense in which the lower bound established in Theorem 1 is informative.

Acknowledgement

The authors thank the referee for many valuable comments and suggestions that significantly improved the presentation of the article. The research of T.-J. Wu was supported by Grant 86-2115-M-259-003 from the National Science Council of Taiwan.

Appendix. Proofs

We first state five lemmas that will be used in the proof of Theorem 1. These lemmas show that the minimax lower bound for estimating $h_k(f)$ is equivalent to and completely determined by that for estimating $h_{k,S}(f)$ (see (2.2)), and the latter bound involves $B_k^2(f)$. Lemmas 1 to 5 are proved under the conditions of Theorem 1. We will not state the conditions explicitly in the following lemmas. Here we point out that Lemmas 1 to 5 herein generalize Lemmas 1 to 5, respectively, of Fan and Marron (1992) (i.e., our lemmas reduce to their lemmas when $k = 0$).

The following lemma indicates that the problem of estimating $h_k(f)$ is equivalent to that of estimating $\phi_k(f)$.

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

$$\sup_{g\in H_n(f,C)} \{h_k(g) - \phi_k(g)\}/\phi_k(g) = o(n^{-1/2}), \tag{A.1}$$

where $\phi_k(g)$ was defined by (2.1).

The next lemma gives a lower bound for estimating $\{\theta_{k+2}(f)\}^{-1/(2k+5)}$.

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

$$R_{n,C,1}(f) = \inf_{\hat{h}_k} \sup_{g\in H_n(f,C)} E_g\Big\{\hat{h}_k - \{\theta_{k+2}(g)\}^{-1/(2k+5)}\Big\}^2.$$

Then

$$\lim_{C\to\infty} \liminf_{n\to\infty}\ nR_{n,C,1}(f) \geq \{\theta_{k+2}(f)\}^{-2/(2k+5)} B_k^2(f),$$

where $B_k(f)$ was defined by (1.3).

In order to show that the second term $Q_k(f)$ of $\phi_k(f)$ (see (2.1) and (2.2)) is negligible, the next lemma gives an estimate of $\delta_l(f)$.

Lemma 3. *For any $l = 1, \ldots, [(k+2)/2]$, there exists an estimator $\hat{\delta}_l$ such that*

$$\sup_{g\in H_n(f,C)} E_g\{\hat{\delta}_l - \delta_l(g)\}^2 = O(n^{-16(k+2)(k+2-l)/\{(8k+17)(2k+5)\}}).$$

Note that $l = 1$ when $k = 0$, and the convergence rate reduces to $O(n^{-32/85})$, which is the rate obtained in Lemma 3 *of Fan and Marron* (1992).

The next lemma indicates that the minimax lower bound for $\phi_k(f)$ is equivalent to that of $h_{k,S}(f)$ (see (2.2)), i.e., the second term $Q_k(f)$ of $\phi_k(f)$ is indeed negligible.

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

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

Then, as $n \to \infty$ and $C \to \infty$ we have

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

where c_1 was defined in (2.3).

The next lemma indicates that the Hellinger neighborhood is so small that $h_k(g)$ is asymptotically equivalent to $h_k(f)$. Consequently, important characteristics of g are very close to those of f for large n (see the proof of Lemma 5 for details).

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

$$\lim_{n\to\infty} \sup_{g \in H_n(f,C)} |\{h_k(g)/h_k(f)\} - 1| = 0.$$

In the sequel, we write θ_j for $\theta_j(f)$ and suppress the subscript k in M_k, $\phi_k(g)$, $h_k(f)$, $h_k(g)$, $\hat{h}_k$, $h_{k,S}(f)$, etc., whenever it causes no confusion. Throughout the rest of the paper, all the suprema are taken over $g \in H_n(f, C)$ (see (2.4)), and this range will not be specified explicitly.

The proofs of all the lemmas and theorems, except Lemma 3 and Theorem 2, are straightforward extensions to general k of counterparts in Fan and Marron (1992). For ease of comparison and understanding, throughout we follow closely the organization and arguments (with suitable adaptations and necessary generalizations) in their proofs.

Proof of Lemma 1. For $k = 0$, Fan and Marron (1992) proved (A.1) using calculations as in Section 2 of Hall, Sheather, Jones and Marron (1991). For a general k, we can use results from Section 1.5 of Wu (1997) (see, also, Scott (1992)) to establish (A.1).

Proof of Lemma 2. From the proof of Theorem 2(i) of Bickel and Ritov (1988), we know that θ_{k+2} is (Frèchet) pathwise differentiable along paths

$$\{f_v : \| {f_v}^{1/2} - f^{1/2} \|_2 \to 0, \textit{ and } \| (f_v^{(2k+4)} - f^{(2k+4)}) f^{1/2} \|_2 \to 0\}$$

with derivative $4\{(-1)^{k+2} f^{(2k+4)}(x) - \theta_{k+2}\} f^{1/2}(x)$. Hence $\theta_{k+2}^{-1/(2k+5)}$ is also (Frèchet) pathwise differentiable along such paths with derivative

$$\{-(2k+5)^{-1} \theta_{k+2}^{-(2k+6)/(2k+5)}\} 4\{(-1)^{k+2} f^{(2k+4)}(x) - \theta_{k+2}\} f^{1/2}(x).$$

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

$$\begin{aligned}
&\| -2(2k+5)^{-1} \theta_{k+2}^{-(2k+6)/(2k+5)} \{(-1)^{k+2} f^{(2k+4)}(x) - \theta_{k+2}\} f^{1/2}(x) \|_2^2 \\
&= 4(2k+5)^{-2} \theta_{k+2}^{-2(2k+6)/(2k+5)} \int_{-\infty}^{\infty} \{f^{(2k+4)}(x) - (-1)^{k+2}\theta_{k+2}\}^2 f(x) dx \\
&= 4(2k+5)^{-2} \theta_{k+2}^{-2-\{2/(2k+5)\}} \mathrm{Var}\,\{f^{(2k+4)}(X_1)\} = \theta_{k+2}^{-2/(2k+5)} B_k^2(f)
\end{aligned}$$

by using the fact that $\theta_{k+2} = (-1)^{k+2}\int_{-\infty}^{\infty} f^{(2k+4)}(x)f(x)dx$. The result follows from arguments at the end of the proof of Lemma 2 of Fan and Marron (1992), where standard semiparametric theory (cf., Theorem 2.10 of van der Vaart (1988)) was used.

Proof of Lemma 3. We note that for $g \in \mathcal{F}_{j+\alpha}$, $g^{(2k+4)}$ is bounded by $g_0 \in L_1 \cap L_\infty$. Since $k+2+j < 2k+4 < 2(k+2+j)+1/4$ for all $j = 0, \ldots, [(k+2)/2]$, it follows by the construction of Bickel and Ritov (1988) (using their notations, if $k+2+j < m+\alpha \leq 2(k+2+j)+1/4$ then the bound $n^{4\gamma}E\{\hat{\theta}_{k+2+j}-\theta_{k+2+j}(F_n)\}^4 = O(1)$, where $\gamma = 4\{m+\alpha-(k+2+j)\}/(1+4m+4\alpha)$, can be established by arguments analogous to those in the proof of Theorem 1 (ii) of their paper; and then setting $m+\alpha = 2k+4$) that there exist estimators $\hat{\theta}_{k+2} \geq 0$ and $\hat{\theta}_{k+3}, \ldots, \hat{\theta}_{k+2+[(k+2)/2]}$ such that

$$\sup E_g\{\hat{\theta}_{k+2+j}-\theta_{k+2+j}(g)\}^4 = O(n^{-16(k+2-j)/(8k+17)}),\ j=0,\ldots,[(k+2)/2], \quad \text{(A.2)}$$

$$\sup E_g\hat{\theta}_{k+2}^{(4k+4)} = O(1),\ \sup E_g|\hat{\theta}_{k+2+j}|^{(4k+4)/j} = O(1),\ j=1,\ldots,[(k+2)/2] \quad \text{(A.3)}$$

(cf., Hall and Marron (1987a) and Jones and Sheather (1991) for a different estimator which can also be used here). Next, by (1.6) of Wu (1995) and by Remark 2, we have for all $j = 0, \ldots, [(k+2)/2]$,

$$\sup \pi\theta_{k+2+j}(g) = \int_0^\infty \lambda^{2k+4+2j}|\psi_g(\lambda)|^2 d\lambda = O\Big(1+\int_1^\infty \lambda^{2k+4+2j}/\lambda^{4k+8}d\lambda\Big) < \infty. \quad \text{(A.4)}$$

Furthermore, from Remark 1 it is not difficult to see that for $l = 1, \ldots, [(k+2)/2]$,

$$\theta_{k+2}^{(2l+1)/(2k+5)}(g)\delta_l(g) = \sum_{r \in B_l} c_r J_1^{r_1}(g)\cdots J_l^{r_l}(g) = \sum_{r\in B_l} c_r^*\theta(g)/\theta_{k+2}^{r_1+r_2+\cdots+r_l}(g), \quad \text{(A.5)}$$

where $\theta(g)\ (= \theta_{l,r}(g)) = \theta_{k+3}^{r_1}(g)\cdots\theta_{k+2+l}^{r_l}(g)$, $B_l = \{r = (r_1, \ldots, r_l) : 0 \leq r_1, \ldots, r_l \leq l,\ \sum_{m=1}^l mr_m = l\}$ and c_r and c_r^* are constants not depending on n and the $\theta_j(g)$'s. Put $\delta_{l,r}(g) = \theta(g)/\theta_{k+2}^\beta(g)$, where $\beta = \{(2l+1)/(2k+5)\} + \sum_{i=1}^l r_i$ (note that $\sum_{i=1}^l r_i \leq l$). In view of (A.5), we have $\delta_l(g) = \sum_{r\in B_l} c_r^*\delta_{l,r}(g)$, and the lemma will be proved if we show that for each $l = 1, \ldots, [(k+2)/2]$ and $r \in B_l$, there exists an estimator $\hat{\delta}_{l,r}$ such that

$$\sup E_g\{\hat{\delta}_{l,r} - \delta_{l,r}(g)\}^2 = O(n^{-16(k+2)(k+2-l)/\{(2k+5)(8k+17)\}}). \quad \text{(A.6)}$$

To guard against a zero denominator, take $\hat{\delta}_{l,r} = \hat{\theta}/\{\hat{\theta}_{k+2}^\beta + n^{-\tau}\}$, where $\hat{\theta} = \hat{\theta}_{k+3}^{r_1}\cdots\hat{\theta}_{k+2+l}^{r_l}$ and $\tau = 4(k+2-l)/(8k+17)$. For simplicity of notation we denote $\tilde{\theta} = \hat{\theta}-\theta(g)$ and $\tilde{\theta}_{k+2+j} = \hat{\theta}_{k+2+j}-\theta_{k+2+j}(g)$. Also, for ease of writing we define the non-negative r.v. $\xi_{k+2+j} = |\hat{\theta}_{k+2+j}| \vee \theta_{k+2+j}(g)$. Put $p = (2k+5)/(k+3+l)$ (note that $1 < p < 2$). Using (A.2) to (A.4), we have

$$\begin{aligned}
\sup E_g|\tilde{\theta}|^{2p} &= \sup E_g\Big|\sum_j \Big(\hat{\theta}^{r_j}_{k+2+j} - \theta^{r_j}_{k+2+j}(g)\Big) \prod_{m=1}^{j-1} \theta^{r_m}_{k+2+m}(g) \prod_{m=j+1}^{l} \hat{\theta}^{r_m}_{k+2+m}\Big|^{2p} \\
&= \sup E_g\Big|\sum_j \tilde{\theta}_{k+2+j} \sum_i \hat{\theta}^{r_j-i-1}_{k+2+j}\,\theta^{i}_{k+2+j}(g) \prod_{m=1}^{j-1} \theta^{r_m}_{k+2+m}(g) \prod_{m=j+1}^{l} \hat{\theta}^{r_m}_{k+2+m}\Big|^{2p} \\
&= O\Big(\sum_j\sum_i \sup E_g\Big\{|\tilde{\theta}_{k+2+j}|^{2p}|\hat{\theta}_{k+2+j}|^{2p(r_j-i-1)}\theta^{2pi}_{k+2+j}(g) \prod_{m\neq j} \xi^{2pr_m}_{k+2+m}\Big\}\Big) \\
&= O\Big(\sum_j\sum_i \sup E_g\Big\{|\tilde{\theta}_{k+2+j}|^{2p}\xi^{2p(r_j-1)}_{k+2+j} \prod_{m\neq j} \xi^{2pr_m}_{k+2+m}\Big\}\Big) \\
&= O\Big(\sum_j \{\sup E_g\tilde{\theta}^4_{k+2+j}\}^{p/2} \prod_{m=1}^{l} \{\sup E_g\xi^{4p(l-j)/((2-p)m)}_{k+2+m}\}^{1/p_m}\Big) \\
&= O\Big(\sum_j n^{-8p(k+2-j)/(8k+17)} \prod_{m=1}^{l} \{\sup E_g(\xi_{k+2+m}\vee 1)^{(4k+4)/m}\}^{1/p_m}\Big) \\
&= O(n^{-2p\tau}), \qquad\qquad (A.7)
\end{aligned}$$

where the summation is taken over $\{j : \ 1 \leq j \leq l, \ r_j \geq 1\}$ and $\{i : \ 0 \leq i \leq r_j - 1\}$. The third equality is obtained by applying the Generalized C_r-inequality (i.e., $|\sum_{i=1}^t a_i|^p \leq C_{p,t}\sum_{i=1}^t |a_i|^p$, $C_{p,t}$ is a constant); the fourth equality is obtained by noting that $|\hat{\theta}_{k+2+j}|^{2p(r_j-i-1)}\theta^{2pi}_{k+2+j}(g) \leq \xi^{2p(r_j-1)}_{k+2+j}$; the fifth equality is obtained by applying the Generalized Hölder's inequality with $p_0^{-1} = p/2$, $p_j^{-1} = j(r_j-1)(2-p)/\{2(l-j)\}$ and $p_m^{-1} = mr_m(2-p)/\{2(l-j)\}$, $m \neq j$; and the sixth equality follows from the fact $4p(l-j)/((2-p)m) \leq (4k+4)/m$ and the inequality $|x|^s \leq (|x| \vee 1)^t$ if $s \leq t$. Put $I_1 = E_g\{(\hat{\delta}_{l,r} - \delta_{l,r}(g))^2 I_S\}$ and $I_2 = E_g\{(\hat{\delta}_{l,r} - \delta_{l,r}(g))^2 I_{S'}\}$, where $I(\cdot)$ is the indicator function, $S = \{|\tilde{\theta}_{k+2}| \geq \theta_{k+2}(g)/2\}$, and S' is the complement of S. Then

$$\begin{aligned}
E_g(\hat{\delta}_{l,r} - \delta_{l,r}(g))^2 &= I_1 + I_2 \\
&= E_g\Big\{\{\theta^{\beta}_{k+2}(g)\tilde{\theta} - (\hat{\theta}^{\beta*}_{k+2} + n^{-\tau})\theta(g)\}/\{(\hat{\theta}^{\beta}_{k+2} + n^{-\tau})\theta^{\beta}_{k+2}(g)\}\Big\}^2,
\end{aligned}$$

where $\hat{\theta}^{\beta*}_{k+2} = \hat{\theta}^{\beta}_{k+2} - \theta^{\beta}_{k+2}(g)$. Evidently, $S' \subset \{\hat{\theta}_{k+2} > \theta_{k+2}(g)/2\}$. This implies that the denominator in I_2 is bounded away from 0. Hence, using (A.2) to (A.4) and (A.7), we get

$$\sup I_2 = O\Big(\sup E_g\tilde{\theta}^2 + \sup E_g\{\hat{\theta}^{\beta*}_{k+2}\}^2 + n^{-2\tau}\Big) = O(n^{-2\tau}).$$

Next, let us consider I_1. The fact that $\hat{\theta}_{k+2} \geq 0$ ensures that

$$I_1 = O\Big(n^{2\tau} E_g\Big\{\{\theta_{k+2}^{\beta}(g)\tilde{\theta} - (\hat{\theta}_{k+2}^{\beta *} + n^{-\tau})\theta(g)\}^2 I_S\Big\}\Big).$$

Applying Hölder's inequality with the foregoing p and $q = (2k+5)/(k+2-l)$, and using (A.2) to (A.4) and arguments analogous to those in deriving (A.7), we get

$$\sup I_1 = O\Big(n^{2\tau}\Big\{\sup E_g|\tilde{\theta}|^{2p} + \sup E_g\{\hat{\theta}_{k+2}^{\beta *}\}^{2p} + n^{-2p\tau}\Big\}^{1/p}\Big\{\sup E_g I_S\Big\}^{1/q}\Big)$$
$$= O(n^{2\tau} n^{-2\tau}\{\sup E_g I_S\}^{1/q}) = O(n^{-2\tau(2k+4)/(2k+5)}),$$

where the last equality follows from

$$\sup E_g I_S \leq 16\{\inf \theta_{k+2}(g)\}^{-4} \sup E_g|\tilde{\theta}_{k+2}|^4 = O(n^{-16(k+2)/(8k+17)})$$

where the infimum is taken over $g \in H_n(f, C)$. This can be derived by applying Markov's inequality and using the fact that $\inf \theta_{k+2}(g) > 0$ eventually in n, as implied by (A.8) herein. This leads immediately to (A.6).

Proof of Lemma 4. Let $\hat{\delta} = \sum_{l=1}^{[(k+2)/2]} \hat{\delta}_l n^{-2l/(2k+5)}$ be the estimator of $\delta(f) = c_1^{-1} n^{1/(2k+5)} Q_k(f)$ (recalling (2.1)-(2.3)), where $\hat{\delta}_l$ is the estimator of $\delta_l(f)$ defined in Lemma 3. Then by making the change of variable $\hat{h} \rightarrow n^{-1/(2k+5)} c_1(\hat{h} + \hat{\delta})$, we have

$$R_{n,C,2}(f) = n^{-2/(2k+5)} c_1^2 \inf_{\hat{h}} \sup E_g\{\hat{h} - \theta_{k+2}^{-1/(2k+5)}(g) + \hat{\delta} - \delta(g)\}^2$$
$$\geq n^{-2/(2k+5)} c_1^2 \inf_{\hat{h}} \sup\Big(E_g\{\hat{h} - \theta_{k+2}^{-1/(2k+5)}(g)\}^2 - a_n\{E_g\{\hat{h} - \theta_{k+2}^{-1/(2k+5)}(g)\}^2\}^{1/2}\Big),$$

where $a_n = 2\{E_g\{\hat{\delta} - \delta(g)\}^2\}^{1/2} = o(n^{-1/2})$, as can be seen from Lemma 3. Thus,

$$R_{n,C,2}(f) \geq n^{-2/(2k+5)} c_1^2 \inf_{\hat{h}} \{q^2(\hat{h}) - a_n q(\hat{h})\},$$

where

$$q(\hat{h}) = \{\sup\ E_g\{\hat{h} - \theta_{k+2}^{-1/(2k+5)}(g)\}^2\}^{1/2}.$$

By Lemma 2, for any estimator $\hat{h}$ and all sufficiently large n and C, we have

$$q(\hat{h}) \geq \inf_{\hat{h}} q(\hat{h}) = R_{n,C,1}^{1/2}(f) \geq 2^{-1}\{\theta_{k+2}(f)\}^{-1/(2k+5)} B_k(f) n^{-1/2}.$$

This entails that $q(\hat{h}) > a_n$ for all sufficiently large n and C. The rest of the proof follows directly from arguments near the end of the proof of Lemma 4 of Fan and Marron (1992).

Proof of Lemma 5. The proof is a straightforward extension of the proof of Fan and Marron (1992). Indeed, by the arguments at the beginning of the proof of Lemma 5 of Fan and Marron (1992) and by our present Lemma 1, we see that it suffices to show that sup $|\theta_{k+2}(g)-\theta_{k+2}(f)|\to 0$ as $n\to\infty$. Now, using (3.9) of Fan and Marron (1992) and the fact that $g^{(j)}(x)$ $(j=0,\ldots,2k+4)$ can be estimated consistently, we get $\sup|g^{(j)}(x)-f^{(j)}(x)|^2\to 0$ for $j=0,\ldots,2k+4$. The desired result follows from

$$\sup|\theta_{k+2}(g)-\theta_{k+2}(f)|=\sup\Big|\int_{-\infty}^{\infty}g^{(2k+4)}g-\int_{-\infty}^{\infty}f^{(2k+4)}f\Big|$$
$$\leq\int_{-\infty}^{\infty}f\sup\ |g^{(2k+4)}-f^{(2k+4)}|+\int_{-\infty}^{\infty}g_0\sup\ |g-f|\to 0, \tag{A.8}$$

as can be seen by using Dominated Convergence Theorem and the fact that $|g^{(2k+4)}|\leq g_0$.

Proof of Theorem 1. Write $h(g)=\phi(g)+\eta(g)$. We get by Lemma 1, (2.1) and (2.2), that sup $\eta(g)=o(n^{-1/2-1/(2k+5)})$. In view of Lemmas 5 and 1, (2.2) and (2.8), we have

$$\begin{aligned}&\inf_{\hat h}\ \sup\ E_g\{(\hat h-h(g))/h(g)\}^2\geq\ \inf_{\hat h}\ \{\sup\ E_g(\hat h-h(g))^2/\sup\ h^2(g)\}\\ &=\inf_{\hat h}\ \sup\ E_g\{\hat h-\phi(g)-\eta(g)\}^2h_S^{-2}(f)(1+o(1)).\end{aligned}$$

Putting this together with the argument used at the end of the proof of Lemma 4, we can show that $\eta(g)$ is indeed negligible and conclude that

$$\inf_{\hat h}\ \sup\ E_g\{(\hat h-h(g))/h(g)\}^2\geq h_S^{-2}(f)R_{n,C,2}(f)(1+o(1)).$$

The proof follows directly from (2.2), Lemmas 4 and 2.

Proof of Theorem 2. Throughout the proof we denote $h(f)$ by h_f. By Taylor expansion and using the fact that $M'(h_f)=0$, we get

$$M(\hat h)-M(h_f)=2^{-1}M''(h_f)(\hat h-h_f)^2+6^{-1}M'''(\tilde h)(\hat h-h_f)^3, \tag{A.9}$$

where $\tilde h$ lies between $\hat h$ and h_f. Let us denote $d_1=(2k+1)^{2/(2k+5)}(2k+5)$, $d_2=d_1/\{4(2k+1)^{(2k+3)/(2k+5)}\}$, $r=\theta_k^{1/(2k+5)}(w)$ and $\nu=\mu_2^{1/(2k+5)}$. Then, from results in Section 1.5 of Wu (1997) (which extends results in Hall et al. 1991 to general k), we have

$$\begin{aligned}M(h)+n^{-1}\theta_k(f)&=N(h)+O(n^{-1}h^2+h^6),\\ N(h)&=(nh^{2k+1})^{-1}\theta_k(w)+4^{-1}\mu_2^2h^4\theta_{k+2}(f),\\ h_f/h_S&=1+O(n^{-2/(2k+5)}),\end{aligned} \tag{A.10}$$

where $h_S=h_S(f)$, as defined by (2.2), is the minimizer of $N(h)$, and, consequently,

$$\begin{aligned}M''(h_f)&=N''(h_f)+O(n^{-1}+h_f^4)=N''(h_S)(1+O(h_f^2))\\ &=d_1r^2\nu^{4k+6}\theta_{k+2}^{(2k+3)/(2k+5)}n^{-2/(2k+5)}(1+O(h_f^2))\end{aligned} \tag{A.11}$$

and

$$\begin{aligned} M(h_f) &= N(h_f) + O(h_f^6 + n^{-1}) = N(h_S)\{1 + O(h_f^2 + h_f^{2k+1})\} \\ &= d_2 r^4 \nu^{4k+2} \theta_{k+2}^{(2k+1)/(2k+5)} n^{-4/(2k+5)} \{1 + O(h_f^2 + h_f^{2k+1})\}. \quad \text{(A.12)} \end{aligned}$$

Combining (A.10)-(A.12) yields

$$M''(h_f) h_f^2 / \{2M(h_f)\} = (4k+2)\{1 + O(h_f^2 + h_f^{2k+1})\}, \qquad \text{(A.13)}$$

where $N''(h_S)h_S^2/\{2N(h_S)\} = (2k+1)^{2/(2k+5)} d_1/(2d_2) = 4k+2$ was used. By arguments analogous to those in deriving (A.13) we get

$$M'''(\tilde{h}) h_f^3 / \{6M(h_f)\} = O_p(\tilde{h}/h_f)\{1 + O_p(\tilde{h}^2) + O(h_f^2 + h_f^{2k+1})\}.$$

This, together with (A.9) and (A.13), leads to

$$\begin{aligned} & n^{\delta}\{M(\hat{h})/M(h_f) - 1\} \\ &= (4k+2) n^{\delta} (\hat{h}/h_f - 1)^2 \{1 + O(n^{-2/(2k+5)} + n^{-(2k+1)/(2k+5)})\} \\ &\quad + O_p(1)\{n^{\delta}(\hat{h}/h_f - 1)^3\}. \end{aligned}$$

This completes the proof.

References

Aldershof, B. (1991). Estimation of integrated squared density derivatives. *Mimeo Series* #2053, Institute of Statistics, University of North Carolina.

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

Bickel, P. J. and Ritov, Y. (1988). Estimating integrated squared density derivatives: sharp best order of convergence estimates. *Sankhyā Ser. A* **50**, 381-393.

Cheng, M.-Y. (1997). Boundary-aware estimators of integrated squared density derivatives. *J. Roy. Statist. Soc. Ser. B* **59**, 191-203.

Cheng, M.-Y., Fan, J. and Marron, J. S. (1997). On automatic boundary corrections. *Ann. Statist.* **25**, 1691-1708.

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

Chiu, S. T. (1992). An automatic bandwidth selector for kernel density estimation. *Biometrika* **79**, 771-782.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1996). Density estimation by wavelet thresholding. *Ann. Statist.* **24**, 508-539.

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

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

Fan, J., Gijbels, I., Hu, T.-C. and Huang, L.-S. (1996). A study of variable bandwidth selection for local polynomial regression. *Statist. Sinica* **6**, 113-127.

Fan, J. and Marron, J. S. (1992). Best possible constant for bandwidth selection. *Ann. Statist.* **20**, 2057-2070.

Goldstein, L. and Messer, K. (1992). Optimal plug-in estimators for nonparametric functional estimation. *Ann. Statist.* **20**, 1306-1328.

Grund, B., Hall, P. and Marron, J. S. (1994). Loss and risk in smoothing parameter selection. *J. Nonparametr. Statist.* **4**, 107-132.

Hall, P. and Marron, J. S. (1987a). Estimation of integrated squared density derivatives. *Statist. Probab. Lett.* **6**, 109-115.

Hall, P. and Marron, J. S. (1987b). 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. (1991). Lower bounds for bandwidth selection in density estimation. *Probab. Theory Related Fields* **90**, 149-173.

Hall, P., Marron, J. S. and Park, B. U. (1992). Smoothed cross-validation. *Probab. Theory Related Fields* **92**, 1-20.

Hall, P. and Patil, P. (1995). Formulae for mean integrated squared error of nonlinear wavelet-based density estimators. *Ann. Statist.* **23**, 905-928.

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**, 521-530.

Härdle W., Hart, J., Marron, J. S. and Tsybakov, A. B. (1992). Bandwidth choice for average derivative estimation. *J. Amer. Statist. Assoc.* **87**, 218-226.

Härdle, W., Marron, J. S. and Wand, M. P. (1990). Bandwidth choice for density derivatives. *J. Roy. Statist. Soc. Ser. B* **52**, 223-232.

Härdle, W. and Stoker, T. (1989). Investigating smooth multiple regression by the method of average derivatives. *J. Amer. Statist. Assoc.* **84**, 986-995.

Hewitt, E. and Stromberg, K. (1969). *Real and Abstract Analysis.* Springer-Verlag, New York.

Hildenbrand, K. and Hildenbrand, W. (1986). On the mean income effect: a data analysis of the U.K. family expenditure survey. In *Contributions to Mathematical Economics, in Honor of Gerard Debreu* (Edited by W. Hildenbrand and A. Mas-Colell), 247-268. North-Holland, Amsterdam.

Jones, M. C. (1991). The role of ISE and MISE in density estimation. *Statist. Probab. Lett.* **12**, 51-56.

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., Marron, J. S. and Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *J. Amer. Statist. Assoc.* **91**, 401-407.

Jones, M. C. and Sheather, S. J. (1991). Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives. *Statist. Probab. Lett.* **11**, 511-514.

Loader, C. R. (1995). Old faithful erupts: bandwidth selection reviewed. Manuscript, AT&T Bell Laboratories.

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.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization.* John Wiley, New York.

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

Stone, C. J. (1980). Optimal convergence rates for nonparametric estimators. *Ann. Statist.* **8**, 1348-1360.

van der Vaart, A. W. (1988). *Statistical Estimation in Large Parameter Spaces. CWI Tract* **44**. Mathematical Centrum, Amsterdam.

Wu, T.-J. (1995). Adaptive root n estimates of integrated squared density derivatives. *Ann. Statist.* **23**, 1474-1495.

Wu, T.-J. (1997). Root n bandwidth selectors for kernel estimation of density derivatives. *J. Amer. Statist. Assoc.* **92**, 536-547.

Department of Mathematics, University of Houston, Houston, TX 77204-3476, U.S.A.

E-mail: tjwu@math.uh.edu

(Received April 1998; accepted May 1999)