*The Annals of Statistics*
1987, Vol. 16, No. 1, 161–173

# CHOICE OF KERNEL ORDER IN DENSITY ESTIMATION

BY PETER HALL AND J. S. MARRON[1,2]

*Australian National University*

The selection of the order, i.e., number of vanishing moments, of the kernel in a kernel density estimator is considered from two points of view. First, theoretical properties are investigated by a mean integrated squared error analysis of the problem. Second, and perhaps more importantly, cross validation is proposed as a practical method of choice, and theoretical backing for this is provided through an asymptotic optimality result.

**1.. Introduction.** In kernel density estimation, as well as in other curve-estimation settings, it has long been known that improved asymptotic rates of convergence can be attained through use of a "higher-order" kernel function; see Parzen (1962), Watson and Leadbetter (1963) and Bartlett (1963). Higher-order kernels are usually thought of as functions whose early moments are equal to zero. It has been demonstrated through simulation studies that the asymptotically indicated benefits of higher-order kernels can also be very significant in finite-sample situations; see, for example, Schucany and Sommers (1977) and Gasser, Müller and Mammitzsch (1985). However, higher-order kernels are virtually never used in practice for two reasons. First, they have the intuitively unappealing feature of taking negative values. Second, whereas a good deal is known about choosing the smoothing parameter, essentially nothing is known about how to choose the order of the kernel, both from a theoretical viewpoint and in practical situations.

It is the intention of this paper to address both parts of the second issue. We hope that, through clear resolution of this matter, the concerns of the first issue might be outweighed, in many statisticians' minds, by the advantages of higher-order kernels.

Two approaches to choosing the kernel order are considered here. First, intuitive insight is gained through an asymptotic mean integrated squared error (MISE) analysis. Second, with deeper practical implications, cross validation is proposed as a method of choosing both the smoothing parameter and the kernel order, and theoretical backing for this idea is provided.

Our MISE analysis simultaneously takes into account the problems of both smoothing parameter and kernel-order selection, instead of arbitrarily fixing the kernel order and concentrating on the smoothing parameter as has typically been done in the past. Asymptotic representations are found, in two interesting cases, for the optimal choices of the smoothing parameter and of the kernel order, and

Received October 1986; revised June 1987.

[1] On leave from University of North Carolina, Chapel Hill.

[2] Research partially supported by NSF Grant DMS-84-00602.

*AMS* 1980 *subject classifications*. Primary 62G05; secondary 62E20.

*Key words and phrases*. Cross validation, nonparametric density estimation, smoothness, variable-order kernel.

also for the minimum MISE. It is seen that if $f$ is not too "smooth," then there is a finite optimal order, whereas if $f$ is very "smooth," then, at least asymptotically, the optimal order is infinite.

The theoretical backing for cross validation takes the form of an asymptotic optimality result in the spirit of the results of Hall (1983, 1985) and Stone (1984). Intuitively, the result says that choosing the smoothing parameter and kernel order by cross validation is, in the limit, as good as using the optimal values.

Notation is introduced in the second section. The third and fourth sections treat the MISE analysis and cross validation, respectively. The rest of the paper is devoted to proofs.

**2. The setting.** Given a random sample $X_1, \ldots, X_n$ from an unknown univariate density $f$, the kernel estimator of $f$ is given by

$$\hat{f}(x) \equiv (nh)^{-1} \sum_{j=1}^{n} K\{(x - X_j)/h\}, \tag{2.1}$$

where $K$ is the kernel function and $h$ is the smoothing parameter or bandwidth. The kernel $K$ is typically said to be of order $p \in \mathbb{Z}$ if

$$\int x^j K(x)\, dx = \begin{cases} 1, & j = 0, \\ 0, & j = 1, \ldots, p-1, \\ C, & j = p, \end{cases}$$

where $C \neq 0$.

A number of authors, including Parzen (1962), Watson and Leadbetter (1963), and Davis (1975, 1977), have analyzed the estimator at (2.1) by Fourier transform analysis. By these methods, the notion of order of a kernel can be extended to real values of $p$ by requiring that the Fourier transform $\kappa$ of $K$ satisfy

$$\kappa(t) = 1 + C|t|^p + O(|t|^p),$$

as $t \to 0$, where $C$ is a nonzero constant.

To simplify presentation, the results are stated and proved for the particular family of kernels

$$K_p(x) \equiv \pi^{-1} \int_0^\infty \cos(tx)\exp(-t^p)\, dt, \tag{2.2}$$

for $p \geq 2$. Observe that $K_p$ has Fourier transform

$$\kappa_p(t) = \exp(-|t|^p),$$

and so $K_p$ is of order $p$. Also note that $K_2$ is a Gaussian probability density. Many alternatives to (2.2) are possible; for example, $\exp(-t^p)$ may be replaced by $(1 + t^p)^{-1}$ without affecting our conclusions. A more common method of constructing higher-order kernels is through piecewise polynomials; see Singh (1979), Müller (1984) and Gasser, Müller and Mammitzsch (1985). Although the idea of this paper extends in a straightforward fashion to such kernels, proofs in the case of unbounded order result in severe technical complications, which seem to obscure the statistical issues.

Our decision to work with Fourier transforms, rather than directly with estimators, is in keeping with efficient practical methods for numerical implementation of density estimation; see Silverman (1982). Minimal computation is achieved by taking fast Fourier transforms and constructing estimators in the Fourier domain. From that viewpoint, variable kernels such as $K_p$ are much better suited to computation than are piecewise polynomials. Silverman (1982) also gives a method for computing the cross-validatory criterion.

As $p \to \infty$, $K_p$ converges to the "sync kernel" $K_\infty(X) \equiv (\pi x)^{-1}\sin x$, which is of interest because of certain $L^2$ optimality properties; see Davis (1975, 1977) and Ibragimov and Khasminskii (1982). Unfortunately, density estimates based on $K_\infty$ exhibit erratic $L^1$ behavior, due to the fact that $|K_\infty|$ is not integrable. However, the kernel

$$L(x) \equiv (4/\pi x^2)\left[\{\sin(x/2)\}^2 - \{\sin(x/4)\}^2\right],$$

for $x \in R$, is integrable. Since $L$ has Fourier transform

$$\lambda(t) \equiv 1_{\{|t| \le 1/2\}} + 2(1 - |t|)1_{\{1/2 \le |t| \le 1\}},$$

which satisfies $\lambda(t) \equiv 1$ in a neighborhood of the origin, then $L$ (like $K_\infty$) is a kernel whose nominal order is infinite. Devroye and Györfi (1985), page 135, describe a family of kernels similar to $L$. The theory for any member of this family is similar to that which we shall give for $K_\infty$ and for $L$.

**3. MISE analysis.** The integrated squared error (ISE) and MISE are defined by

$$\text{ISE} \equiv \int (\hat{f}(x) - f(x))^2\, dx,$$

$$\text{MISE} \equiv E(\text{ISE}).$$

Bias influences the MISE behavior of $\hat{f}$ [defined at (2.1)] through a characteristic often termed "smoothness." Some descriptions of smoothness are based on various derivatives of $f$. The more sophisticated of these include Sobolev-type conditions [Wahba (1972)], Taylor remainder conditions [Stone (1980)] and Lipschitz bounds on derivatives [Stone (1982)]. Other descriptions of smoothness are based on the characteristic function $\phi$ of $f$ [Parzen (1962), Watson and Leadbetter (1963) and Davis (1975, 1977)].

A version of the second approach is to assume either

$$|\phi(t)| \sim \beta|t|^{-\alpha}, \quad \text{as } |t| \to \infty, \tag{3.1}$$

for some $\alpha > \frac{1}{2}$ and $\beta > 0$, or

$$|\phi(t)| \sim \exp(-\beta|t|^\alpha), \quad \text{as } t \to \infty, \tag{3.2}$$

for some $\alpha > 0$ and $\beta > 0$. These assumptions are much stronger than is mathematically necessary (in particular, they imply that $\phi$ cannot take on the value zero infinitely often), but are made for clarity of presentation. There are many ways to weaken (3.1) and (3.2), for example, by working with "integral averages" of those conditions, but these involve introduction of technicalities, which tend to obscure the main issues.

Condition (3.1) is closest to the derivative-based methods of describing smoothness. It may be very loosely interpreted by saying that "$f$ has only $\alpha - 1$ derivatives at some point." For example, it holds if $f$ is a gamma density with shape parameter $\alpha$, in which case only $\alpha - 1$ "derivatives" exist at the origin.

Condition (3.2) provides a way of describing smoothness when $f$ is infinitely differentiable, although the term "smoothness" loses much of its intuitive content in this case.

The asymptotic behavior of MISE, and properties of MISE-optimal $h$ and $p$ under the first type of smoothness assumption, are described by

THEOREM 3.1. *Assume $\phi$ satisfies* (3.1).

(a) *If $K \equiv K_p$, then*

$$\inf_{h,p} \text{MISE} \sim n^{-(2\alpha-1)/(2\alpha)}\pi^{-1}\{\beta^2(2\alpha-1)\}^{1/(2\alpha)} \times \{1 + (2\alpha-1)^{-1}\}\Big(\inf_{2\le p\le\infty} C_p\Big)^{1/(2\alpha)},$$

*where*

$$C_p \equiv \{2^{-1/p}\Gamma(1+p^{-1})\}^{2\alpha-1}\int_0^\infty t^{-2\alpha}\{1-\exp(-t^p)\}^2\,dt,$$

*for $p < \infty$ and $C_\infty \equiv (2\alpha-1)^{-1}$. The value $p_0$ of $p$, which minimizes $C_p$, is finite, and the values of $h$, $p$, which minimize MISE, satisfy $p \to p_0$ and*

$$n^{1/(2\alpha)}h \to \left[2^{1/p_0}\{\Gamma(1+p_0^{-1})^{-1}\}(2\alpha-1)\beta^2\int_0^\infty t^{-2\alpha}\{1-\exp(-t^{p_0})\}^2\,dt\right]^{-1/2\alpha}.$$

(b) *If $K \equiv K_\infty$, then*

$$\inf_h \text{MISE} \sim n^{-(2\alpha-1)/(2\alpha)}\pi^{-1}\{\beta^2(2\alpha-1)\}^{1/(2\alpha)}\{1+(2\alpha-1)^{-1}\}C_\infty^{1/(2\alpha)},$$

*and the minimizing value of $h$ satisfies*

$$h \sim \beta^{-1/\alpha}n^{-1/(2\alpha)}.$$

(c) *If $K \equiv L$, then*

$$\inf_h \text{MISE} \sim n^{-(2\alpha-1)/(2\alpha)}\pi^{-1}\tfrac{2}{3}D_0\{1+(2\alpha-1)^{-1}\},$$

*and the minimizing value of $h$ satisfies $h \sim D_0^{-1}n^{-1/(2\alpha)}$, where*

$$D_0 \equiv \left\{\tfrac{3}{2}(2\alpha-1)\beta^2\int_{1/2}^1 t^{-2\alpha}(2t-1)^2\,dt\right\}^{1/(2\alpha)}.$$

REMARK 3.1. Since $p_0 < \infty$, there is something to be gained in using the finite order kernel $K_{p_0}$ rather than the sync kernel $K_\infty$. However, the gain is not in the rate of convergence, but only in the size of the multiplicative constant.

REMARK 3.2. A comparison of the kernels $K_\infty$ and $L$ can be made by considering the ratio of their minimum asymptotic MISEs. For large $\alpha$ this ratio is close to $\frac{3}{4}$, which fits in with optimality results for $K_\infty$ [Davis (1977), Section 3], and which can be thought of as the price paid for using the integrable $L$ instead of $K_\infty$.

The asymptotic behavior of MISE, and properties of the MISE-optimal $h$ and $p$ under the second type of smoothness assumption, are contained in

THEOREM 3.2. *Assume $\phi$ satisfies* (3.2).

(a) *If $K \equiv K_p$, then*

$$\inf_{h,p} \text{MISE} \sim n^{-1}(\log n)^{1/\alpha}\pi^{-1}(2\beta)^{-1/\alpha},$$

*the minimizing value of $p$ diverges to $+\infty$, and the minimizing value of $h \sim (2\beta/\log n)^{1/\alpha}$.*

(b) *The same holds true if $K \equiv K_\infty$, except that $p \equiv \infty$.*

(c) *If $K \equiv L$, then*

$$\inf_{h} \text{MISE} \sim n^{-1}(\log n)^{1/\alpha}(3\pi)^{-1}2^{2-(1/\alpha)}\beta^{-1/\alpha},$$

*and the minimizing value of $h$ satisfies $h \sim (2^{1-\alpha}\beta/\log n)^{1/\alpha}$.*

REMARK 3.3. The other side of Remark 3.1 is that if $f$ is very smooth, then there is no first-order asymptotic gain in using any finite-order kernel $K_p$, even if $p$ increases with $n$. When one uses $K_\infty$ or $L$, there is no extra parameter such as $p$ to determine kernel order. Window size $h$ implicitly determines order. For related work on how $h$ should be chosen in this type of setting, see Cline and Hart (1986).

REMARK 3.4. Note that the ratio of smallest asymptotic MISEs in parts (b) and (c) of Theorem 3.2 is $\frac{3}{4}$, which fits in with the optimality results of Section 4 of Davis (1977) and also agrees with Remark 3.2.

REMARK 3.5. The arguments used to establish Theorem 3.2 may be employed to show that if

$$\liminf_{|t|\to\infty} \exp(|t|^\alpha)|\phi(t)| > 0,$$

for some $\alpha > 0$, then for each $\varepsilon > 0$, the bandwidth $h$ minimizing MISE satisfies $h = O\{(\log n)^{-(1/\alpha)+\varepsilon}\}$ as $n \to \infty$, no matter whether $K \equiv K_p$, $K_\infty$ or $L$. This fact will be used in the next section.

**4. Cross validation.** The results of the previous section provide insight into the theoretical issues of choosing kernel order, but are not directly useful in practice because their application would require detailed knowledge of the

unknown $f$. The analogous difficulty for smoothing parameter selection has been overcome through cross validation, and we propose doing the same here.

The squared-error cross-validatory criterion is an unbiased estimate of MISE $- \int f^2$, given by

$$\mathrm{CV} \equiv \int \hat{f}^2 - 2n^{-1} \sum_{i=1}^{n} \hat{f}_i(X_i),$$

where $\hat{f}_i$ denotes the leave-one-out version of $\hat{f}$ given by

$$\hat{f}_i(x) \equiv \{(n-1)h\}^{-1} \sum_{j \neq i} K\{(x - X_j)/h\}.$$

See Rudemo (1982), Hall (1983, 1985), Bowman (1984) and Stone (1984) for discussions of the idea of selecting the smoothing parameter $h$ to be the minimizer of CV.

In the case $K \equiv K_p$, let $(\hat{h}, \hat{p})$ denote that value of $(h, p)$, with

$$0 < h \leq (\log n)^{-4(1+\varepsilon)} \tag{4.1}$$

(any fixed $\varepsilon > 0$), which minimizes CV. If $K \equiv K_\infty$ or $K \equiv L$, let $\hat{h}$ denote the value of $h$ within the range (4.1), which minimizes CV. These adaptive choices of $h$ and $p$ work as well as "optimal" choices of $h$ and $p$ in the following asymptotic sense. Let ISE* and MISE* denote the minimum values of MISE and ISE, respectively, minimized over values of $p > 0$ and $h$ satisfying (4.1). Let ISE$^+$ denote the value of ISE evaluated at $(\hat{h}, \hat{p})$ (if $K \equiv K_p$) or at $\hat{h}$ (if $K \equiv K_\infty$ or $K \equiv L$).

THEOREM 4.1. *If $f$ is bounded, then ISE$^+$/ISE* and ISE*/MISE* both converge to unity with probability* 1.

REMARK 4.1. Theorem 4.1 is most meaningful when the value of $h$ which minimizes MISE lies in the interval $(0, (\log n)^{-4(1+\varepsilon)})$. From Remark 3.6, a simple sufficient condition for this is

$$\liminf_{|t| \to \infty} \exp(|t|^{(1/4)-\varepsilon}) |\phi(t)| > 0,$$

for some $\varepsilon > 0$. This restriction can be weakened considerably, for example, by asking that it hold in integral average form. Conditions of this type ask that $f$ be not too smooth. It is unreasonable to expect that cross validation will produce an optimal estimator regardless of the smoothness of $f$. If $f$ is so smooth that $\phi$ vanishes outside a compact interval, then the minimum MISE is of order $n^{-1}$ [Ibragimov and Khasminskii (1982) and Devroye and Györfi (1985), page 133ff.]. But the cross-validatory criterion is equivalent to a quantity that estimates ISE with an error of order $n^{-1}$, so cross validation is doomed to failure.

REMARK 4.2. The idea of using cross validation as a method for choosing between density estimators (as opposed to simply selecting the smoothing parameter) has been discussed before by Rudemo (1982) and Marron (1987).

REMARK 4.3. Cross validation also provides a practical solution to the issues raised in Remarks 3.1 and 3.3, in that choice can be made among $K_p$, $K_\infty$ and $L$ (together with any other density estimator) by taking the one with smallest minimum CV.

REMARK 4.4. It should now be quite clear that the results of this section, and also those of Section 3, would be very cumbersome if $\{K_p\colon\ p \geq 2\}$ were a family of piecewise polynomials.

REMARK 4.5. An interesting feature of the proof of Theorem 4.1, is that the methods used (based on martingale methods and Burkholder's inequality) could be used to give much simpler proofs of the results of Hall (1983, 1985), Stone (1984), Marron (1985, 1987) and Marron and Härdle (1986). The Fourier transform aspect of our proof also provides a new viewpoint into the structure of such results.

REMARK 4.6. It is important to keep in mind that there can be significant differences between the effects described by the theorems in this paper and what happens in a practical situation. In the case of Theorems 3.1 and 3.2, there are two levels of approximation. First, the concept of MISE demands that instead of optimizing ISE for the data set at hand, we minimize the average of ISE over all possible data sets. Second, Theorems 3.1 and 3.2 are only asymptotic in character. Their validity depends on things like $h$ being "small," which can require very large $n$ when $h \sim (\log n)^{-1/\alpha}$. In the case of Theorem 4.1, the results of Hall and Marron (1987a) indicate that the rates of convergence in Theorem 4.1 can be very slow, particularly when $f$ is quite smooth. In general, the performance of cross validation as a device for minimizing ISE becomes poorer as smoothness of the density increases. The difficulty is a feature of the problem, not a deficiency of cross validation. Indeed, there is a sense in which cross validation copes with this difficulty best of all possible smoothing parameter selection methods [Hall and Marron (1987b)].

## 5. Proofs.

PROOF OF THEOREMS 3.1 AND 3.2. We give only an outline. If the symmetric kernel $K$ has Fourier transform $\kappa$, then the estimator $\hat{f}$ defined at (2.1) has MISE given by

$$\pi \text{ MISE} = (nh)^{-1}\int_0^\infty \kappa^2 + \int_0^\infty |\phi|^2(1-\kappa_h)^2 - n^{-1}\int_0^\infty |\phi|^2\kappa_h^2,$$

where $\kappa_h \equiv \kappa(h\,\cdot\,)$ denotes the Fourier transform of $h^{-1}K(\cdot/h)$. Here we have used Parseval's identity.

Minimum MISE is achieved only under conditions of consistency, and that entails $h \to 0$ as $n \to \infty$. In this circumstance, if $\phi$ satisfies (3.1) and if $p > \alpha - \frac{1}{2}$,

$$\int_0^\infty |\phi|^2(1 - \kappa_h)^2 \sim \begin{cases} \beta^2 h^{2\alpha-1} \int_0^\infty t^{-2\alpha}\{1 - \exp(-t^p)\}^2\, dt, & \text{if } K \equiv K_p, 2 \le p \le \infty, \\ \frac{2}{3} h^{2\alpha-1}(2\alpha - 1)^{-1} D_0^{2\alpha}, & \text{if } K \equiv L, \end{cases}$$

and

$$\int_0^\infty \kappa^2 = \begin{cases} 2^{-1/p}\Gamma(1 + p^{-1}), & \text{if } K \equiv K_p, \\ \frac{2}{3}, & \text{if } K \equiv L. \end{cases}$$

Therefore

$$\pi \text{ MISE} \sim \begin{cases} (nh)^{-1} 2^{-1/p}\Gamma(1 + p^{-1}) + \beta^2 h^{2\alpha-1} \int_0^\infty t^{-2\alpha}\{1 - \exp(-t^p)\}^2\, dt, \\ \qquad\qquad \text{if } K \equiv K_p, \\ (nh)^{-1}\frac{2}{3} + \frac{2}{3} h^{2\alpha-1}(2\alpha - 1)^{-1} D_0^{2\alpha}, \\ \qquad\qquad \text{if } K \equiv L, \end{cases}$$

whence the result. [The fact that $p_0 < \infty$ may be deduced from an expansion of $C_p$ in powers of $p^{-1}$; note that $\Gamma(1 + \delta) \equiv 1 - C\delta + O(\delta^2)$ as $\delta \to 0$, where $C \equiv -0.5772\ldots$ is minus the Euler constant.]

If $\phi$ satisfies (3.2) and $K \equiv K_p$, then it can be shown that the first-order asymptotics do not change if we take $p \equiv \infty$. It may be proved that

$$\int_0^\infty |\phi|^2(1 - \kappa_h)^2 \sim \begin{cases} (2\alpha\beta)^{-1} h^{\alpha-1} \exp(-2\beta h^{-\alpha}), & \text{if } K \equiv K_\infty, \\ (2^{1-\alpha}\alpha\beta)^{-3} h^{3\alpha-1} \exp(-2^{1-\alpha}\beta h^{-\alpha}), & \text{if } K \equiv L. \end{cases}$$

Therefore

$$\pi \text{ MISE} \sim \begin{cases} (nh)^{-1} + (2\alpha\beta)^{-1} h^{\alpha-1} \exp(-2\beta h^{-\alpha}), & \text{if } K \equiv K_\infty, \\ (nh)^{-1}\frac{2}{3} + (2^{1-\alpha}\alpha\beta)^{-3} h^{3\alpha-1} \exp(-2^{1-\alpha}\beta h^{-\alpha}), & \text{if } K \equiv L, \end{cases}$$

whence the result.

PROOF OF THEOREM 4.1. We treat only the case $K \equiv K_p$, $2 \le p \le \infty$. Other cases are similar. Write $\kappa_{h,p}$ for the Fourier transform of the function $h^{-1}K_p(\cdot/h)$, let $\hat{\phi}(t) \equiv n^{-1}\sum \exp(itX_j)$, $\eta = \eta(h, p) \equiv \int_0^\infty |\phi|^2(1 - \kappa_{h,p})^2$,

$$\hat{\psi}(t) \equiv \{n(n-1)\}^{-1} \sum\sum_{j \ne k} \exp\{it(X_j - X_k)\}$$

and

$$\text{CV}_0 \equiv \text{CV} + \int f^2 + 2n^{-1} \sum_{i=1}^\infty \left\{ f(X_i) - \int f^2 \right\},$$

and note that for a constant $C > 0$,

$$\pi \,\mathrm{MISE} = n^{-1}\int_0^\infty (1 - |\phi|^2)\kappa_{h,p}^2 + \int_0^\infty |\phi|^2(1 - \kappa_{h,p})^2 \geq C\{(nh)^{-1} + \eta\},$$

$$\pi \,\mathrm{ISE} = \int_0^\infty |\hat\phi|^2\kappa_{h,p}^2 + \int_0^\infty |\phi|^2 - 2\int_0^\infty (\mathrm{Re}\,\hat\phi\bar\phi)\kappa_{h,p}, \tag{5.1}$$

$$\pi \,\mathrm{CV}_0 = \int_0^\infty |\hat\phi|^2\kappa_{h,p}^2 + \int_0^\infty |\phi|^2 - 2\int_0^\infty (\mathrm{Re}\,\hat\psi)\kappa_{h,p} + 2\int_0^\infty \mathrm{Re}\{(\hat\phi - \phi)\bar\phi\}. \tag{5.2}$$

Minimizing CV is equivalent to minimizing $\mathrm{CV}_0$.

Fix $\varepsilon, a, b > 0$, let $\mathscr{H}$ denote the set of values of $h$ satisfying $n^{-a} \leq h \leq (\log n)^{-4(1+\varepsilon)}$ and $\mathscr{P}$ the set of values $p$ satisfying $2 \leq p \leq n^b$. A little algebra shows that if $c = c(a, b)$ and $C = C(a, b)$ are chosen sufficiently large, then

$$\left|\kappa_{h_1,p_1}(t) - \kappa_{h_2,p_2}(t)\right| \leq Cn^{-(a+2)}\exp\left\{-\tfrac{1}{2}(n^{-a}t)^2\right\},$$

uniformly in $t > 0$, values $h_1, h_2 \in \mathscr{H}$ satisfying $|h_1 - h_2| \leq n^{-c}$, and values $p_1, p_2 \in \mathscr{P}$ satisfying $|p_1 - p_2| \leq n^{-c}$. We may now deduce from (5.1) and (5.2) that for some $C_1 > 0$,

$$\begin{aligned}&|\mathrm{ISE}(h_1, p_1) - \mathrm{ISE}(h_2, p_2)| \\ &\quad + |\mathrm{CV}_0(h_1, p_1) - \mathrm{CV}_0(h_2, p_2)| \leq C_1 n^{-2},\end{aligned} \tag{5.3}$$

uniformly in such values of $h_1$, $h_2$, $p_1$, $p_2$ and all $n$-samples $\{X_i\}$. A similar argument shows that for sufficiently large $b$, the left-hand side of (5.3), with $p_1 \equiv p$ and $p_2 \equiv \infty$, is dominated by $C_2 n^{-2}$ uniformly in $h_1, h_2 \in \mathscr{H}$ and $n^b \leq p \leq \infty$; use $\kappa_{h,\infty}$ to approximate $\kappa_{h,p}$, and note that for $t > 0$,

$$|\kappa_{h,p}(t) - \kappa_{h,\infty}(t)| \leq C_3\{(ht)^p I(ht \leq 1) + (ht)^{-p} I(ht > 1)\}$$

and that the integral over $(0, \infty)$ of the right-hand side is dominated by $C_1(hp)^{-1}$. If we now show that for any $\delta > 0$ and $r > 0$,

$$\begin{aligned}&\sup_{h\in\mathscr{H},\, 2\leq p\leq\infty} \Big(P\big[|\mathrm{ISE}(h, p) - \mathrm{MISE}(h, p)| > \delta\{(nh)^{-1} + \eta\}\big] \\ &\qquad + P\big[|\mathrm{CV}_0(h, p) - \mathrm{ISE}(h, p)| > \delta\{(nh)^{-1} + \eta\}\big]\Big) \\ &\quad = O(n^{-r}),\end{aligned} \tag{5.4}$$

then we shall have Theorem (4.1) for $h$ restricted to $n^{-a} \leq h \leq (\log n)^{-2(1+\varepsilon)}$. [See, e.g., Marron (1985) for the argument.] The case $h \leq n^{-a}$ for large $a$ is easily treated separately.

To prove (5.4), note from (5.1) and (5.2) that

$$\begin{aligned}\pi(\mathrm{CV}_0 - \mathrm{ISE}) &= 2\int_0^\infty \mathrm{Re}\{(\hat\phi - \phi)\bar\phi\}(1 - \kappa_{h,p}) \\ &\quad + 2\int_0^\infty \mathrm{Re}(2\hat\phi\bar\phi - \hat\psi - |\phi|^2)\kappa_{h,p},\end{aligned} \tag{5.5}$$

$$\begin{aligned}\pi(\mathrm{ISE} - \mathrm{MISE}) &= 2\int_0^\infty \mathrm{Re}\{(\hat\phi - \phi)\bar\phi\}\{(1 - n^{-1})\kappa_{h,p} - 1\}\kappa_{h,p} \\ &\quad + \int_0^\infty \big[|\hat\phi|^2 - E|\hat\phi|^2 - 2(1 - n^{-1})\mathrm{Re}\{(\hat\phi - \phi)\bar\phi\}\big]\kappa_{h,p}.\end{aligned} \tag{5.6}$$

Let $f_{h,p}(x) \equiv E\{\hat{f}(x|h)\}$ and $Y_i \equiv f_{h,p}(X_i) - f(X_i) - E\{f_{h,p}(X_i) - f(X_i)\}$, and notice that, uniformly in $h$ and $p$,

$$|f_{h,p}(x)| = \left|\int K_p(z) f(x - hz)\, dz\right| \le \left\{\int K_p^2(z)\, dz \int f^2(hz)\, dz\right\}^{1/2} \le C_5 h^{-1/2},$$

$$E(Y_i^2) \le C_6 \int (f_{h,p} - f)^2 - C_6 \pi^{-1} \int_0^\infty |\phi|^2 (1 - \kappa_{h,p})^2 = C_6 \pi^{-1} \eta,$$

and $E|Y_i|^l \le C_7^l h^{-l/2}$. Therefore by Rosenthal's (1970) inequality [see also Hall and Heyde (1980)],

$$\begin{aligned} E\left|\pi^{-1} \int_0^\infty \mathrm{Re}\{(\hat{\phi} - \phi)\bar{\phi}\}(1 - \kappa_{h,p})\right|^l &= E\left|n^{-1} \sum Y_i\right|^l \\ &\le B_l\left[\{n^{-1}E(Y_i^2)\}^{l/2} + n^{1-l}E|Y_i|^l\right] \\ &\le B_l C_8^l\left\{(n^{-1}\eta)^{l/2} + n^{1-l}h^{-l/2}\right\}, \end{aligned}$$

where $B_l$ depends only on $l$. Since $h^{-1}\eta \le h\{(nh)^{-1} + \eta\}^2$, then, by Markov's inequality,

$$\begin{aligned} (5.7) \qquad P\left[\left|\int_0^\infty \mathrm{Re}\{(\hat{\phi} - \phi)\bar{\phi}\}(1 - \kappa_{h,p})\right| > \delta\{(nh)^{-1} + \eta\}\right] \\ \le B_l n\left(C_9 \delta^{-1} h^{1/2}\right)^l. \end{aligned}$$

The version of $B_l$ given by Rosenthal (1970) is unduly large. Following the proof of Burkholder's (1973) Theorem 21.1, we see that we may take $B_l$ to be the infimum of $\gamma\eta/(1 - \gamma\varepsilon)$, where (in Burkholder's notation) $\varepsilon \equiv \delta^2(\beta - \delta - 1)^{-2}$, $\gamma \equiv \beta^l$, $\eta \equiv \delta^{-l}$, and the infimum is over values of $\beta, \delta$ such that $0 < \delta < \beta - 1$ and $\gamma\varepsilon < 1$. For our purposes we may take $\beta = 1 + sl^{-1}\log\log l$ for any $s > 0$, and $\delta \sim l^{-1}\{2s^{-2}(\log l)^s(\log\log l)^{-2}\}^{-1/2}$ for large $l$, which gives $B_l \le \{C_{10} l (\log l)^{s/2}\}^l$. Therefore if $l \equiv \log n$ and if $h \le (\log n)^{-4(1+\varepsilon)}$, then for large $n$ the right-hand side of (5.7) is dominated by

$$n\left\{C_9 C_{10}\, \delta^{-1} l\, (\log l)^{s/2} h^{1/2}\right\}^l \le n\left\{l^{1+\varepsilon}(\log n)^{-2(1+\varepsilon)}\right\}^l = O(n^{-r}),$$

for all $r > 0$. Therefore

$$P\left[\left|\int_0^\infty \mathrm{Re}\{(\hat{\phi} - \phi)\bar{\phi}\}(1 - \kappa_{h,p})\right| > \delta\{(nh)^{-1} + \eta\}\right] = O(n^{-r}),$$

for all $r > 0$. The analogous result, with $1 - \kappa$ replaced by $\{(1 - n^{-1})\kappa - 1\}\kappa$ [see (5.6)], also holds.

To treat the remaining terms in (5.5) and (5.6), notice that

$$T_1 \equiv -\int_0^\infty \mathrm{Re}(2\hat\phi\bar\phi - \hat\psi - |\phi|^2)\kappa_{h,p} = \{n(n-1)\}^{-1}\sum\sum_{j\neq k} H_1(X_j, X_k),$$

$$T_2 \equiv (1-n^{-1})^{-1}\int_0^\infty \big[|\hat\phi|^2 - E|\hat\phi|^2 - 2(1-n^{-1})\mathrm{Re}\{(\hat\phi-\phi)\bar\phi\}\big]\kappa_{h,p}^2$$

$$= \{n(n-1)\}^{-1}\sum\sum_{j\neq k} H_2(X_j, X_k),$$

where $H_j(x, y) = H_{j1}(x, y) + H_{j2}(x, y)$, $u(t) \equiv E(\cos tX)$, $v(t) \equiv E(\sin tX)$,

$$H_{j1}(x, y) \equiv \int_0^\infty \{\cos(tx) - u(t)\}\{\cos(ty) - u(t)\}\kappa_{h,p}^j(t)\,dt, \tag{5.8}$$

and $H_{j2}$ is defined similarly on replacing cos by sin and $u$ by $v$. Let $H$ stand for any one of the four $H_{ij}$'s, and put $Z_j \equiv \sum_{1\le k\le j-1} H(X_j, X_k)$ and $S_n \equiv \sum_{1\le j\le n} Z_j$. Since $E(Z_j|X_1, \ldots, X_{j-1}) = 0$, then $\{S_n\}$ is a martingale, and so by Burkholder's inequality and Hölder's inequality [Hall and Heyde (1980), formula (3.67), page 87],

$$\begin{aligned} s_n &\equiv E\Big|\sum\sum_{1\le j<k\le n} H(X_j, X_k)\Big|^l \\ &= E|S_n|^l \le \big\{18l(1-l^{-1})^{-1/2}\big\}^l n^{(l/2)-1}\sum_{j=1}^n E|Z_j|^l. \end{aligned} \tag{5.9}$$

Conditional on $X_j$, $Z_j$ is a sum of $j-1$ independent and identically distributed random variables, and so by Rosenthal's (1970) inequality,

$$E\big(|Z_j|^l|X_j\big) \le B_l\Big(\big[(j-1)E\big\{H^2(X_j, X_1)|X_j\big\}\big]^{l/2} + (j-1)E\big\{|H(X_j, X_1)|^l|X_j\big\}\Big).$$

Notice that $|H(x, y)| \le C_{11}h^{-1}$ uniformly in $h$, $2 \le p \le \infty$ and real $x$, $y$; and also, in the case $H = H_{j1}$ defined at (5.8),

$$\begin{aligned} &E\{H^2(x, X_1)\} \\ &\quad= \int_0^\infty\int_0^\infty E\big[\{\cos(sX_1) - u(s)\}\{\cos(tX_1) - u(t)\}\big]\{\cos(sx) - u(s)\} \\ &\qquad\times\{\cos(tx) - u(t)\}\kappa_{h,p}^j(s)\kappa_{h,p}^j(t)\,ds\,dt \\ &\quad= \tfrac12\int_0^\infty\int_0^\infty \{u(s+t) + u(s-t) - 2u(s)u(t)\}\{\cos(sx) - u(s)\} \\ &\qquad\times\{\cos(tx) - u(t)\}\kappa_{h,p}^j(s)\kappa_{h,p}^j(t)\,ds\,dt \\ &\quad\le C_{12}h^{-1}, \end{aligned}$$

uniformly in $h$, $2 \le p \le \infty$ and real $x$, by the Cauchy–Schwarz inequality (since $u^2$ is integrable). Identical bounds are valid for other $H$'s. Substituting into (5.9) and using the bound given earlier for $B_l$, we have, for large $l$,

$$s_n \le l^{(2+\varepsilon)l}(n^l h^{-l/2} + n^{(l/2)+1}h^{-l}).$$

Take $l \equiv \log n$, assume $h \leq (\log n)^{-4(1+\varepsilon)}$ and observe that by Markov's inequality,

$$P\left\{\left|n^{-2}\sum\sum_{1\leq j<k\leq n} H(X_j, X_k)\right| > \delta(nh)^{-1}\right\} \leq \delta^{-l}n^{-l}h^l s_n$$
$$\leq \delta^{-l}l^{(2+\varepsilon)l}\left\{(\log n)^{-2(1+\varepsilon)l} + n^{-(l/2)+1}\right\}$$
$$\sim n^{-\varepsilon \log\log n - \log\delta} = O(n^{-r}),$$

for all $r > 0$. In consequence, $P\{|T_j| > \delta(nh)^{-1}\} = O(n^{-r})$ for all $r > 0$, uniformly in $h$ and $p$, for $j = 1, 2$.

**Acknowledgment.** We are grateful to the referee for helpful suggestions.

## REFERENCES

BARTLETT, M. S. (1963). Statistical estimation of density functions. *Sankhyā Ser. A* **25** 245–254.

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

BURKHOLDER, D. L. (1973). Distribution function inequalities for martingales. *Ann. Probab.* **1** 19–42.

CLINE, D. B. H. and HART, J. D. (1986). Kernel estimation of densities with discontinuities or discontinuous derivatives. Preprint.

DAVIS, K. B. (1975). Mean square error properties of density estimates. *Ann. Statist.* **3** 1025–1030.

DAVIS, K. B. (1977). Mean integrated square error properties of density estimates. *Ann. Statist.* **5** 530–535.

DEVROYE, L. and GYÖRFI, L. (1985). *Nonparametric Density Estimation: the $L^1$ View*. Wiley, New York.

GASSER, T., MÜLLER, H.-G. and MAMMITZSCH, V. (1985). Kernels for nonparametric curve estimation. *J. Roy. Statist. Soc. Ser. B* **47** 238–252.

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

HALL, P. (1985). Asymptotic theory of minimum integrated square error for multivariate density estimation. In *Proc. Sixth Internat. Symp. Multivariate Anal.* (P. R. Krishnaiah, ed.) 289–309. North-Holland, Amsterdam.

HALL, P. and HEYDE, C. C. (1980. *Martingale Limit Theory and Its Application*. Academic, New York.

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). On the amount of noise inherent in bandwidth selection for a kernel density estimator. *Ann. Statist.* **15** 163–181.

IBRAGIMOV, I. A. and KHASMINSKII, R. Z. (1982). Estimation of distribution density belonging to a class of entire functions. *Theory Probab. Appl.* **27** 551–562.

MARRON, J. S. (1985). An asymptotically efficient solution to the bandwidth problem of kernel density estimation. *Ann. Statist.* **13** 1011–1023.

MARRON, J. S. (1987). A comparison of cross-validation techniques in density estimation. *Ann. Statist.* **15** 152–162.

MARRON, J. S. AND HÄRDLE, W. (1986). Random approximations to an error criterion of nonparametric statistics. *J. Multivariate Anal.* **20** 91–113.

MÜLLER, H.-G. (1984). Smooth optimum kernel estimators of densities, regression curves and modes. *Ann. Statist.* **12** 766–774.

PARZEN, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.* **33** 1065–1076.

ROSENTHAL, H. P. (1970). On the subspaces of $L^p$ ($p > 2$) spanned by sequences of independent random variables. *Israel J. Math.* **8** 273–303.

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

SCHUCANY, W. R. and SOMMERS, J. P. (1977). Improvement of kernel type density estimators. *J. Amer. Statist. Assoc.* **72** 420–423.

SILVERMAN, B. W. (1982). Kernel density estimation using the fast Fourier transform. *Appl. Statist.* **31** 93–99.

SINGH, R. S. (1979). Mean squared errors of estimates of a density and its derivatives. *Biometrika* **66** 177–180.

STONE, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8** 1348–1360.

STONE, C. J. (1982). Optimal global rates of convergence of nonparametric regression. *Ann. Statist.* **10** 1040–1053.

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

WAHBA, G. (1975). Optimal convergence properties of variable knot, kernel and orthogonal series methods for density estimation. *Ann. Statist.* **3** 15–29.

WATSON, G. S. and LEADBETTER, M. R. (1963). On the estimation of the probability density. I. *Ann. Math. Statist.* **34** 480–491.

DEPARTMENT OF STATISTICS
AUSTRALIAN NATIONAL UNIVERSITY
GPO BOX 4
CANBERRA, ACT 2601
AUSTRALIA