Ann. Inst. Statist. Math.
Vol. 42, No. 2, 287–303 (1990)

OPTIMAL KERNEL ESTIMATION OF DENSITIES*

DAREN B. H. CLINE

Department of Statistics, Texas A&M University, College Station, TX 77843, U.S.A.

(Received November 21, 1988; revised April 24, 1989)

Abstract. Precise asymptotic behavior for mean integrated squared error (MISE) is determined for sequences of kernel estimators of a density in a broad class, including discontinuous and possibly unbounded densities. The paper shows that the sequence using the kernel optimal at each fixed sample size is asymptotically more efficient than a sequence generated by changing the bandwidth of a fixed kernel shape, regardless of the kernel shape. The class of densities considered are those whose characteristic functions behave at large arguments like the product of a Fourier series and a regularly varying function. This condition may be related to the smoothness of an m-th derivative of the density.

Key words and phrases: Kernel density estimation, mean integrated squared error, optimal kernel, regular variation.

1. Introduction and theorem statements

The kernel estimate of a density is given by

$$\hat{f}_n(x) = \frac{1}{n} \sum_{j=1}^{n} \kappa_n(x - X_j) ,$$

where $X_1, \ldots, X_n$ is a random sample from f and the kernel κ_n integrates to one. We are here concerned with the efficiency of $\hat{f}_n$ and the choice of an efficient sequence of kernels. This is a matter of practical concern and has seen frequent discussion (c.f., e.g., Epanechnikov (1969), Wahba (1975), Sacks and Ylvisacker (1981) and Müller (1984)). Generally, the discussion has been limited to sequences of estimators *generated* by a fixed kernel shape, that is, to estimators using kernels of the form

$$\kappa_n(x) = a_n \kappa(a_n x) . \tag{1.1}$$

*Partially supported by National Science Foundation Grant DMS-8711924.

The bandwidth, $h_n = 1/a_n$, is often chosen by cross-validation in order to ensure optimal convergence (Stone (1984)). (We restrict our attention to kernel estimators, but analogous results are possible for others, such as Fourier series estimators.)

Using the mean integrated squared error

$$\text{MISE}\,(\hat{f}_n) = E\left[\int_{-\infty}^{\infty} (\hat{f}_n(x) - f(x))^2 dx\right]$$

as our measure of fit, we set $J_n = \min_{\kappa_n} \text{MISE}\,(\hat{f}_n)$. If we specify that κ_n be generated by κ (as in (1.1)), we set $M_{n,\kappa} = \text{MISE}\,(\hat{f}_n)$. Our objective is to provide the precise asymptotic behavior of J_n and $M_{n,\kappa}$ for each density in a broad class and to show that even for the best choice of κ, one usually has $\lim_{n\to\infty} J_n/M_{n,\kappa} < 1$.

We emphasize that the efficiency $\lambda = \lim_{n\to\infty} J_n/M_{n,\kappa}$ depends on the individual density f. This contrasts with the work of others (Bretagnolle and Huber (1979), Ibragimov and Khasminskii (1983), Stone (1983) and Efroimovich (1985)) who define *minimax* efficiency criteria for classes of densities and show that full efficiency is possible for those classes. Similar minimax results for the class we define below can also be argued; we intend to do so in separate work.

The approach here (first used by Parzen (1958, 1962), and by Watson and Leadbetter (1963)) will use Fourier analysis techniques having the advantage that precise results are obtained. It applies to densities which are discontinuous or which have discontinuous first derivatives, cases not ordinarily dealt with. Their discussion, however, is limited to densities with algebraic characteristic functions, a restriction that excludes many densities having characteristic functions with oscillatory behavior. In this paper, we extend the results to allow such densities (see Definition 1).

Our results also extend and simplify the work of van Eeden (1985) and Cline and Hart (1987) who determined the asymptotic behavior of $\hat{f}_n$ in the case, for some m, $f^{(m)}$ has unknown simple discontinuities. The class of densities studied allows for countably many discontinuities at *unknown* locations. Other papers have considered the efficacy of boundary kernels in cases of known discontinuities (e.g., Gasser *et al.* (1985)).

The early paper by Watson and Leadbetter (1963) determined the kernel which optimizes MISE. This kernel is not of the form (1.1) but has Fourier transform

$$\psi_n(t) = \frac{n\phi_0(t)}{1 + (n-1)\phi_0(t)}, \tag{1.2}$$

where $\phi_0(t) = |\phi(t)|^2$ and ϕ is the characteristic function of the density. The minimum MISE is

$$J_n = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{\phi_0(t)(1-\phi_0(t))}{1+(n-1)\phi_0(t)}\,dt\,. \tag{1.3}$$

Watson and Leadbetter (c.f. also Parzen (1958)) demonstrate that when ϕ_0 is *algebraic of order* p (i.e., $\lim_{t\to\infty} t^p\phi_0(t) = c > 0$), then for $a_n = (cn)^{1/p}$,

$$\lim_{n\to\infty} \frac{n}{a_n} J_n = \frac{1}{2\pi}\int_{-\infty}^{\infty} (1+|t|^p)^{-1}dt\,. \tag{1.4}$$

Under the same condition, they also show, one may choose the generating kernel's transform to be $\psi(t;p) = (1+|t|^p)^{-1}$, so that the sequence of generated kernels is fully efficient relative to J_n. That is,

$$\lim_{n\to\infty} \frac{n}{a_n} M_{n,\kappa} = \frac{1}{2\pi}\int_{-\infty}^{\infty} (1+|t|^p)^{-1}dt\,.$$

(To see heuristically that $\psi(t;p)$ is efficient, note that for large n,

$$\psi_n(t) = \frac{n\phi_0(t)}{1+(n-1)\phi_0(t)} \approx (1+|t/a_n|^p)^{-1} = \psi(t/a_n;p)\,.)$$

The result (1.4) cannot be extended without modification, however and indeed it is not generally true that $M_{n,\kappa}/J_n$ will converge to 1.

The kernel for $\psi(t;p)$ is easily expressed when p is an even integer; otherwise one might use the equivalent Fourier series estimator (Wahba (1981)). Wahba (1981) and Hall and Marron (1988) provide results similar to (1.4) and discuss a cross-validation approach to choosing p. Hart (1988) also considers algebraic tail behavior and its effect on the MISE of ARMA-type estimators of a density.

We begin by extending the class of Watson and Leadbetter (1963). For example, let f_0 be the standard gamma (α) density and

$$f(x) = \sum_{j=1}^{N} |\beta_j| f_0(\beta_j(x-x_j))\,.$$

Then one may easily determine that for large t

$$\phi(t) \approx \left[\sum_{j=1}^{N} (i\beta_j)^{\alpha} \exp(itx_j)\right] t^{-\alpha}\,,$$

which factors into an algebraic component and an oscillating component. This suggests the following definition.

DEFINITION 1. A probability density f is a member of $\mathscr{F}_{\rho/2}$ if

(i)
$$\lim_{t\to\infty} |\phi(t)/\zeta(t) - \chi(t)| = 0 ,$$

where (ii) ζ is nonincreasing on $[0,\infty)$, $\rho \ge 0$ and for all $y > 0$,

$$\lim_{t\to\infty} \zeta(yt)/\zeta(t) = y^{-\rho/2} , \tag{1.5}$$

and (iii) χ is the Fourier series for a bounded summable complex function g over a countable set $\mathscr{D}$ of real numbers.

When (1.5) holds one says that ζ is *regularly varying at* ∞ *with exponent* $-\rho/2$. Algebraic variation is a special case.

Although the definition of $\mathscr{F}_{\rho/2}$ is in terms of ϕ, it actually represents a smoothness class for f. The class is more specified than, say, the Lipschitz class considered by Ibragimov and Khasminskii (1983) and by Stone (1983) but with it we obtain precise asymptotic results and not just bounds. A density f will satisfy Definition 1 (see Cline (1988*b*)) when, for some $m \ge 0$, $f^{(j)}$ is absolutely continuous for $j < m$, $f^{(m)}$ is discontinuous and there exists a set of singularities $\mathscr{D}$ at which $f^{(m-1)}$ (or the distribution function if $m = 0$) varies regularly in a uniform sense. Specifically, for each $x \in \mathscr{D}$,

$$\lim_{\tau\downarrow 0} \tau^{-m}[f^{(m-1)}(x+\tau) - f^{(m-1)}(x)]/\zeta(\tau^{-1}) = g_+(x) ,$$
$$\lim_{\tau\downarrow 0} \tau^{-m}[f^{(m-1)}(x) - f^{(m-1)}(x-\tau)]/\zeta(\tau^{-1}) = g_-(x) .$$

In this case, $\alpha = \rho/2 - m$ and

$$g(x) = \Gamma(1+\alpha) i^m [e^{i\alpha\pi/2} g_+(x) + e^{-i\alpha\pi/2} g_-(x)] .$$

The function g thus measures the *magnitude* of the roughness at the singularities of $f^{(m-1)}$ while ρ measures the *index* of smoothness. When $f^{(m)}$ is simply discontinuous at $x \in \mathscr{D}$, then we may choose g to be the function of jumps, α to be 1 and $\zeta(t)$ to be asymptotic to ct^{-m-1}, $c > 0$.

The case that g is zero at all but one point x demands special attention. This occurs when $f^{(m-1)}$ is relatively smooth at all points but x. Then ϕ is regularly varying itself and has no periodic oscillations. Watson and Leadbetter's (1963) results apply only to such an example.

We are now ready to describe the asymptotic behavior of J_n.

PROPOSITION 1.1. *Suppose $f \in \mathscr{F}_{\rho/2}$, $\rho > 1$. Let χ, the Fourier series for g, be as in Definition* 1, *let $\chi_0 = |\chi|^2$, $g_0(0) = \sum_{\mathscr{D}} |g(x)|^2$ and let a_n satisfy*

(1.6) $$\lim_{n\to\infty} n g_0(0)|\zeta(a_n)|^2 = 1 .$$

Then there exists $\lambda = \lim_{t\to\infty} \int_0^1 |\chi_0(tu)/g_0(0)|^{1/\rho} du \le 1$ and

$$\lim_{n\to\infty} \frac{n}{a_n} J_n = \frac{\lambda}{2\pi} \int_{-\infty}^{\infty} (1 + |t|^\rho)^{-1} dt .$$

Furthermore, $\lambda = 1$ if and only if g is nonzero at exactly one point.

Watson and Leadbetter's result is an example of the special case $\lambda = 1$. The definition (1.6) for a_n effectively inverts the function ζ, which will be necessary for any convergence to hold. Condition (1.5) entails that $a_n = l(n)(g_0(0)n)^{1/\rho}$, where $l(yn)/l(n) \to 1$ for $y > 0$, and thus the optimal convergence rate will be $l(n)n^{(1-\rho)/\rho}$. (If $f^{(m)}$ is simply discontinuous, then we may choose $l(n) = 1$.) We will further discuss the convergence rate in the second section.

We provide the exact limiting behavior of $M_{n,\kappa}$ when the kernel is chosen from an appropriate class, defined next. As tradition suggests and Cline (1988*a*) demonstrates, we may limit the discussion to symmetric kernels.

DEFINITION 2. A bounded symmetric kernel κ, with Fourier transform ψ, is a member of $\mathscr{K}_{\rho/2}$, $\rho > 1$, if

$$V(\kappa) = \int_{-\infty}^{\infty} \kappa^2(x)\,dx = \frac{1}{2\pi} \int_{-\infty}^{\infty} \psi^2(t)\,dt < \infty$$

and $|t|^{-\rho/2}(1 - \psi(t))$ is bounded. In this case, we define (see also Lemma 3.5)

(1.7) $$B_\rho(\kappa) = \frac{1}{2\pi} \int_{-\infty}^{\infty} |t|^{-\rho}(1 - \psi(t))^2 dt .$$

For $\kappa \in \mathscr{K}_{\rho/2}$, $\rho/2$ is called the *order* of the kernel. If κ is bounded, integrates to 1, has moments up to order $\rho/2$ and the integer moments with order less than $\rho/2$ vanish, then $\kappa \in \mathscr{K}_{\rho/2}$ (see, e.g., Lukacs (1983), p. 23). However, some proposed kernels, while in $K_{\rho/2}$ for all ρ, are not even integrable (e.g., the kernel for the Fourier integral estimator investigated by Davis (1977)).

The definition for $B_p(\kappa)$ relies on knowledge of ψ. In Lemma 3.5, however, we will provide an alternative definition involving integrals in κ. We now state the asymptotic behavior of $M_{n,\kappa}$.

PROPOSITION 1.2. *Suppose* $f \in \mathscr{F}_{p/2}$ *and* $\kappa \in \mathscr{K}_{p/2}$, $p > 1$. *Let* a_n *be defined by* (1.6), $c_n = ca_n$ *and* $\kappa_n(x) = c_n\kappa(c_n x)$. *Define accordingly,* $M_{n,\kappa,c} =$ MISE $(\hat{f}_n)$. *Then, locally uniformly in* c,

$$\lim_{n\to\infty} \frac{n}{a_n} M_{n,\kappa,c} = cV(\kappa) + c^{1-p}B_p(\kappa)\,. \tag{1.8}$$

Furthermore, the limit in (1.8) *is minimized if and only if* κ *has Fourier transform* $\psi(t;p) = (1 + |t|^p)^{-1}$ *and* $c = 1$. *In this case,*

$$\lim_{n\to\infty} \frac{n}{a_n} M_{n,\kappa} = \frac{1}{2\pi}\int_{-\infty}^{\infty} (1 + |t|^p)^{-1} dt\,.$$

The parameter λ in Proposition 1.1, therefore, represents the asymptotic efficiency of a sequence generated by the optimal kernel of Watson and Leadbetter (1963) relative to the optimal sequence. In fact, a similar argument verifies that λ also represents the asymptotic ratios of the integrated variances and of the integrated squared biases of the two sequences. Interestingly, λ depends only on g and p. In the next section, we will give several examples and show that λ can be arbitrarily small.

When kernels other than that of Watson and Leadbetter are used, an optimal choice for c minimizes (1.8). A corollary to Proposition 1.2 describes how this choice is related to the bandwidth which optimizes MISE for a fixed sample size n.

COROLLARY 1.1. *Let* $c^0 = [(p-1)B_p(\kappa)/V(\kappa)]^{1/p}$ *and let* M_{n,κ,c^0} *be as in Proposition* 1.2. *Let* $M^0_{n,\kappa} =$ MISE $(\hat{f}_n)$ *when the kernel sequence is given by* $\kappa^0_n(x) = \kappa(x/h^0_n)/h^0_n$ *and* h^0_n *is chosen so that* MISE $(\hat{f}_n)$ *is minimized. Then*

$$\lim_{n\to\infty} c^0 a_n h^0_n = 1 \qquad \text{and} \qquad \lim_{n\to\infty} \frac{n}{a_n} M^0_{n,\kappa} = \frac{pc^0V(\kappa)}{(p-1)}\,.$$

As Cline and Hart (1987) show, using results by Stone (1984) and Marron and Härdle (1986), the limiting behavior of the MISE given in Corollary 1.1 also applies to the integrated squared error (ISE) when cross-validation is used to choose the norming constants. Indeed, the cross-validated bandwidths will be asymptotic to h^0_n.

The remainder of the paper is organized as follows. The next section presents a few comments on the convergence rate, on examples of densities and on efficient practical estimators. The third section covers preliminary results which are needed for the propositions, especially results involving the tail behavior of ϕ_0. The final section demonstrates the proofs of the two propositions and the corollary.

2. Comments and examples

In this section, we will discuss the convergence rate for J_n and $M_{n,\kappa}$ and how it relates to the optimal convergence rates described by other authors. We also present several examples and calculate the efficiency parameter λ given by Proposition 1.1 for each. In addition, we briefly consider possible data-based methods and the possibility of full efficiency.

2.1 *Rate of convergence*

When $\phi_0 = |\phi|^2$ is algebraic of order ρ, Watson and Leadbetter (1963) have shown that the optimal convergence rate of MISE is $n^{(1-\rho)/\rho}$. Our results show that if ϕ_0 damps according to a regularly varying function, the rate is $n^{(1-\rho)/\rho}l(n)$ with slowly varying $l(n)$. If m is the smallest integer strictly greater than $\rho/2$, the density f therefore has $m-1$ continuous derivatives with $f^{(m-1)}$ being Lipschitz of order $\alpha = \rho/2 - m - \varepsilon$, for every small positive ε. (This follows from Cline (1988*b*).) If it exists, $f^{(m)}$ is discontinuous and possibly unbounded.

If one assumes, as is common, that $f^{(m)}$ be square integrable, then one must have $\alpha \geq 1/2$. Taking the least smooth case, $\alpha = 1/2$, the optimal convergence rate has exponent $(1-\rho)/\rho = -2m/(2m+1)$, and this agrees with the optimal convergence rate obtained by Bretagnolle and Huber (1979) and by Müller and Gasser (1979).

Ibragimov and Khasminskii (1983) and Stone (1983) have determined the uniform optimal rate of $\| \hat{f}_n - f \|_q$ for classes of densities with $f^{(m-1)}$ being Lipschitz (α), $\alpha = \rho/2 - m$. The class used by Efroimovich (1985) assumes a generalized derivative $f^{((\rho-2)/2)}$ which is square integrable. In either case, the optimal rate for MISE is $n^{(2-\rho)/(\rho-1)}$, slower than the Watson-Leadbetter rate. The distinction, apparently, is in the function class. For $f^{(m-1)} \in \text{Lip}\,(\alpha)$ with bounded support, we possibly may have

$$\lim_{\tau \downarrow 0} \tau^{-2\alpha} \| f^{(m-1)}(x+\tau) - f^{(m-1)}(x) \|_2^2 < \infty ,$$

but for f also in $F_{\rho/2}$ we will have

$$\lim_{\tau \downarrow 0} \tau^{-2\alpha-1} \| f^{(m-1)}(x+\tau) - f^{(m-1)}(x) \|_2^2 < \infty .$$

Thus the rates of decrease for the characteristic functions can differ

(Titchmarsh (1948), Theorem 85) and so may the convergence rates for MISE.

2.2 *Examples*

Watson and Leadbetter consider the gamma (ν) density. The squared modulus of the characteristic function is $\phi_0(t) = (1 + t^2)^{-\nu}$, which is clearly algebraic of order $\rho = 2\nu$. Because ϕ_0 is regularly varying, $\lambda = 1$. More generally, $\lambda = 1$ anytime the density is algebraic near one point and smooth everywhere else.

To obtain an example which oscillates, let ζ be a characteristic function such that $|\zeta|^2$ varies regularly with exponent $-\rho$. Let ϕ be the characteristic function for the location-scale-reflection mixture

$$\phi(t) = \sum_{j=1}^{N} g_j \exp(itx_j)\zeta(t/\beta_j) .$$

For this example one has $\lim_{t\to\infty} |\phi(t)/\zeta(t) - \chi(t)| = 0$, where

$$\chi(t) = \sum_{j=1}^{N} g_j\beta_j^{\rho/2} \exp(itx_j) .$$

The value of λ depends on ζ only through the value of ρ. The density for ζ has just one singularity contributing to the rate ρ while f is a mixture and has N such singularities. In fact, f needs only to behave like such a mixture near the singularities and to be smooth elsewhere.

In a particular case, take $N = 2$ and $\beta_1 = \beta_2 = 1$. Then f has exactly two singularities of the same order and

$$\chi_0(t) = (g_1^2 + g_2^2) - 2g_1g_2 \cos(|x_1 - x_2|t) .$$

One can easily show that only the value of ρ and $\gamma = 2g_1g_2/(g_1^2 + g_2^2)$ are relevant in determining λ. That is,

$$\lambda = \frac{1}{\pi}\int_0^{\pi} (1 - \gamma \cos u)^{1/\rho} du .$$

This achieves its maximum (1.00) at $\gamma = 0$ (i.e., when $f^{(m)}$ has only one discontinuity) and its minimum (.900) at $\gamma = 1$.

For another example, assume f has N equally spaced and identical singularities. Thus, $g_j = N^{-1}$ and (without loss) $x_j = 2j$. Then $\chi_0(t) = [\sin(Nt)/(N \sin t)]^2$, $g_0(0) = N^{-1}$ and for some finite A,

$$\lambda = \frac{2}{\pi}\int_0^{\pi/2} \left[\frac{\sin^2(Nt)}{N \sin^2 t} \right]^{1/\rho} dt \le AN^{(1-\rho)/\rho} .$$

This example demonstrates that the efficiency parameter can be arbitrarily small for any value of $\rho > 1$.

As a final example, let f be the $(m+1)$-fold convolution of a uniform density. Then $\phi_0(t) = (\sin t/t)^{2m+2}$ which satisfies Definition 1 with $\rho = 2m+2$. Therefore,

$$g_0(0) = \frac{1}{2\pi}\int_0^{2\pi} |\sin u|^{2m+2} du = \frac{\Gamma(3/2)\Gamma(m+3/2)}{\Gamma(m+2)},$$

$$\lambda = g_0(0)^{-1/\rho}\frac{1}{2\pi}\int_0^{2\pi} |\sin u| du = \frac{2}{\pi}g_0(0)^{-1/\rho}.$$

This value is least when m is large ($\lambda \approx .63662$).

2.3 *Practical and efficient estimators?*

Several authors have suggested data-based methods to achieve the efficiency of the optimal generated sequence. Wahba (1981) offers a Fourier series estimator which essentially uses the optimal transform $\psi(t;\rho)$ and which adaptively estimates a_n and ρ (see also Hall and Marron (1988)). This would be fully efficient if ϕ_0 is regularly varying. The Fourier integral estimator of Davis (1977) is another which does not require prior knowledge of ρ, but it would sustain some loss of efficiency relative to Wahba's estimator. Efroimovich (1985) provides a minimax efficient estimator (for a different density class) which likewise requires no prior knowledge of the underlying class parameters.

More generally, the optimal kernel given by (1.2) depends on f and therefore must itself be estimated. Thaler ((1974), p. 91) suggested an adaptive Fourier series method, but found that as tail estimation of ϕ was very poor, efficiency was not assured.

Consider, however, the kernel whose Fourier transform is

$$\tilde{\psi}_n(t) = \frac{n\chi_0(t)\zeta_0(t)}{1 + n\chi_0(t)\zeta_0(t)}$$

and the corresponding estimator $\tilde{f}_n(x)$. Using the methods described in the proof of Proposition 1.1, one may show $\lim_{n\to\infty} \text{MISE}(\tilde{f}_n)/J_n = 1$. Of course, $\tilde{\psi}_n(t)$ still requires knowing the tail behavior of ϕ_0 and for that reason, practical versions may not be possible.

3. Preliminary results

The method of proof used by Watson and Leadbetter (1963) to obtain (1.4), relies on the assumption that ϕ_0 is algebraic. In fact, all they really require is that ϕ_0 is *regularly varying*, i.e.

$$\lim_{n\to\infty}\frac{\phi_0(yt)}{\phi_0(t)}=y^{-\rho},$$

for all $y>0$. Since $\rho>0$, this requires that ϕ_0 oscillate very little: a strong assumption. However, we will demonstrate that, whenever $f\in\mathscr{F}_{\rho/2}$, the behavior for $M_{n,\kappa}$ depends on a sort of average behavior of ϕ_0. In addition, if U has a bounded density σ, then $n(tU)^\rho\phi_0(a_ntU)$ will be shown to have a limiting distribution for each t and this will be exploited to determine the behavior of J_n.

For the following discussion, let $\zeta_0=|\zeta|^2$, $\chi_0=|\chi|^2$ and $g_0(y)=\sum_{\mathscr{D}} g(x)\bar{g}(x-y)$ with support $\mathscr{D}_0$ ($\bar{g}$ denotes the complex conjugate of g).

LEMMA 3.1. *Suppose ζ satisfies condition* (i) *of Definition* 1, *with $\rho>0$, and a_n is chosen according to* (1.6).

(i) *Uniformly on compact subsets of* $(0,\infty)$, $\lim_{n\to\infty} ng_0(0)\zeta_0(a_nt)=t^{-\rho}$.

(ii) *For each $\varepsilon\in(0,\rho)$, there exists n_0 such that for all $n\geq n_0$ and all* $t>0$, $ng_0(0)\zeta_0(a_nt)\leq(1+\varepsilon)(t^{\varepsilon-\rho}+t^{-\varepsilon-\rho})$.

PROOF. The first statement follows directly from (1.5) and (1.6) (c.f., Seneta (1976), p. 2) and the second from Cline ((1989), Lemma 2.1). □

LEMMA 3.2. *Suppose χ satisfies condition* (iii) *in Definition* 1 *and σ is a bounded probability density, symmetric about zero. Then*

$$\lim_{t\to\infty}\int_{-\infty}^{\infty}\sigma(u)\chi_0(tu)\,du=g_0(0)\,.$$

PROOF. Let s be the characteristic function of σ. As $t\to\infty$, $s(t)$ vanishes. Therefore, by Parseval's relation and Lebesgue dominated convergence,

$$\lim_{t\to\infty}\int_{-\infty}^{\infty}\sigma(u)\chi_0(tu)\,du=\lim_{t\to\infty}\sum_{x\in\mathscr{D}_0}s(tx)g_0(x)=g_0(0)\,. \qquad \square$$

LEMMA 3.3. *Suppose χ is as in Definition* 1. *Let U be distributed with bounded probability density σ, symmetric about zero. Then $\chi_0(tU)/g_0(0)$ is uniformly integrable and there exists a probability distribution H, independently of σ and with mean* 1, *such that $\chi_0(tU)/g_0(0)$ converges in distribution to H, as $t\to\infty$. Furthermore, H is degenerate if and only if χ_0 is constant* (*i.e., if and only if g is nonzero at exactly one point*).

PROOF. Let g_0^{*j} be the j-fold convolution of g_0 with itself. Since

$$g_0^{*j}(0) \le \Big[\sum_{x \in \mathscr{D}_0} |g_0(x)| \Big]^j ,$$

it follows that $g_0^{*j}(0)/(g_0(0))^j$ is the j-th moment of a unique distribution H. In particular, H has mean 1. By Lemma 3.2,

$$\lim_{t\to\infty} \int_{-\infty}^{\infty} \sigma(u)\chi_0^j(tu)\,du = g_0^{*j}(0) .$$

And by the Frechet-Shohat theorem, we conclude that $\chi_0(tU)/g_0(0)$ converges in distribution to H and the sequence is uniformly integrable. Finally, H is degenerate if and only if $g_0^{*2}(0) = \sum_{\mathscr{D}_0} |g_0(x)|^2 = g_0^2(0)$, which requires g_0 (and hence g) to be nonzero at exactly one point. □

LEMMA 3.4. *Suppose $f \in \mathscr{F}_{p/2}$ and define a_n by* (1.6). *Let H and U be as described in Lemma* 3.3. *Then, for each $t > 0$, $\xi_n(tU) = n(tU)^p\phi_0(a_n tU)$ converges in distribution to H, as $n \to \infty$, and is uniformly integrable.*

PROOF. By assumption, and since $|\phi_0/\zeta_0 - \chi_0| \le |\phi/\zeta - \chi|^2 + 2|\chi|\,|\phi/\zeta - \chi|$,

$$\lim_{t\to\infty} |\phi_0(tU)/\zeta_0(tU) - \chi_0(tU)| = 0 ,$$

almost surely. However, from Lemma 3.3, $\chi_0(a_n tU)/g_0(0)$ converges to H, as $n \to \infty$. Thus $\phi_0(a_n tU)/(g_0(0)\zeta_0(a_n tU))$ also converges to H. In addition, by Lemma 3.1(i),

$$\lim_{n\to\infty} ng_0(0)\zeta_0(a_n tU) = (tU)^p ,$$

almost surely. We thus have that $\xi_n(tU) = n(tU)^p\phi_0(a_n tU)$ converges in distribution to H.

Furthermore, ϕ_0/ζ_0 is bounded so that, from Lemma 3.1(ii), there exists n_0 and S such that

$$\sup_{n \ge n_0} n[(tU)^{p+\varepsilon} + (tU)^{p-\varepsilon}]\phi_0(a_n tU) < S ,$$

almost surely. Since $u^{-\varepsilon}\sigma(u)$ is integrable for small enough ε, it follows that $\xi_n(tU)$ is uniformly integrable. □

Our final lemma provides an alternative definition for $B_p(\kappa)$.

LEMMA 3.5. *Suppose $\kappa \in \mathscr{K}_{\rho/2}$, $\rho > 1$, with Fourier transform ψ, and $B_\rho(\kappa)$ is defined by* (1.7). *Let $q_\rho = 2\Gamma(\rho)\cos(\rho\pi/2)$. Assume κ is integrable and $\int_{-\infty}^{\infty} |x^{\rho-1}\kappa(x)|dx < \infty$. If ρ is not an odd integer, then,*

$$B_\rho(\kappa) = q_\rho^{-1}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(|x-y|^{\rho-1} - |x|^{\rho-1} - |y|^{\rho-1})\kappa(x)\kappa(y)dxdy\,.$$

If ρ is an odd integer, then $B_\rho(\kappa) = (2/\pi)B_{\rho-1}(\kappa)$.

PROOF. By assumption, $(1-\psi(t))^2$ is the Fourier-Stieltjes transform of the finite signed measure $K_2 = (\delta - K)*(\delta - K)$, where K has density κ and δ has mass 1 at 0. Calculating with Lebesgue convergence and Parseval's relation,

$$\begin{aligned}
&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(|x-y|^{\rho-1} - |x|^{\rho-1} - |y|^{\rho-1})\kappa(x)\kappa(y)dxdy \\
&\quad = \int_{-\infty}^{\infty}|x|^{\rho-1}K_2(dx) \\
&\quad = \lim_{\varepsilon\downarrow 0}\int_{-\infty}^{\infty}|x|^{\rho-1}e^{-\varepsilon|x|}K_2(dx) \\
&\quad = \lim_{\varepsilon\downarrow 0}\frac{\Gamma(\rho)}{\pi}\int_{-\infty}^{\infty}\cos(\rho\arctan(t/\varepsilon))(\varepsilon^2+t^2)^{-\rho/2}(1-\psi(t))^2dt \\
&\quad = 2\Gamma(\rho)\cos(\rho\pi/2)\frac{1}{2\pi}\int_{-\infty}^{\infty}|t|^{-\rho}(1-\psi(t))^2dt\,.
\end{aligned}$$

When ρ is not odd, the result follows from the definition of $B_\rho(\kappa)$. When ρ is odd, we note that the above implies $\int_{-\infty}^{\infty}|x|^{\rho-1}K_2(dx) = 0$, so that

$$\begin{aligned}
B_\rho(\kappa) &= \lim_{\varepsilon\downarrow 0}\varepsilon q_{\rho-\varepsilon}^{-1}\int_{-\infty}^{\infty}(|x|^{\rho-\varepsilon-1} - |x|^{\rho-1})/\varepsilon K_2(dx) \\
&= [\pi\Gamma(\rho)\sin(\rho\pi/2)]^{-1}(1-\rho)\int_{-\infty}^{\infty}|x|^{\rho-2}\kappa_2(dx) = \frac{2}{\pi}B_{\rho-1}(\kappa)\,. \quad \square
\end{aligned}$$

In case, ρ is an even integer, the above result also follows from the work of Cline and Hart (1987) combined with Proposition 1.2.

4. Principle results

In this section, we prove the two propositions and the corollary. As Proposition 1.1 is more involved, we leave its proof until last.

PROOF OF PROPOSITION 1.2. Assume $\kappa \in \mathcal{K}_{p/2}$ with Fourier transform ψ. Following Watson and Leadbetter (1963), $M_{n,\kappa}$ may be written as the sum of two parts, the integrated variance,

$$V_n = \frac{1}{2\pi}\frac{1}{n}\int_{-\infty}^{\infty} \psi^2(t/c_n)(1 - \phi_0(t))\,dt\,,$$

and the integrated squared bias,

$$B_n = \frac{1}{2\pi}\int_{-\infty}^{\infty} (1 - \psi(t/c_n))^2 \phi_0(t)\,dt\,.$$

Clearly, since $c_n \to \infty$,

$$\lim_{n\to\infty} \frac{n}{a_n} V_n = \lim_{n\to\infty} \frac{c}{2\pi}\int_{-\infty}^{\infty} \psi^2(t)(1 - \phi_0(c_n t))\,dt = cV(\kappa)\,. \tag{4.1}$$

Furthermore, this clearly holds uniformly for c in a compact subset of $(0, \infty)$.

On the other hand, applying Lemma 3.4 with $\sigma(t)$ proportional to $|t|^{-p}(1 - \psi(t))^2$, which is bounded and integrable,

$$\begin{aligned} \lim_{n\to\infty} \frac{n}{a_n} B_n &= \lim_{n\to\infty} \frac{c}{2\pi}\int_{-\infty}^{\infty} (1 - \psi(t))^2 n\phi_0(ca_n t)\,dt \\ &= \frac{c^{1-p}}{2\pi}\int_{-\infty}^{\infty} |t|^{-p}(1 - \psi(t))^2\,dt \int_0^{\infty} yH(dy) \\ &= c^{1-p}B_p(\kappa)\,. \end{aligned} \tag{4.2}$$

That (4.2) holds locally uniformly is a consequence of the uniform integrability of $nt^p\phi_0(ca_n t)$ with respect to $\sigma(t)$. Equation (1.8) and its local uniformity follows from (4.1) and (4.2).

Parzen (1958) and Watson and Leadbetter (1963) demonstrate that $\psi(t; p) = (1 + |t|^p)^{-1}$ is the Fourier transform of a minimizing kernel. It is also easy to show it is the unique minimizer. □

PROOF OF COROLLARY 1.1. From Proposition 1.2, we have locally uniformly in c,

$$\lim_{n\to\infty} \frac{n}{a_n} M_{n,\kappa,c} = cV(\kappa) + c^{1-p}B_p(\kappa)\,. \tag{4.3}$$

The limit is minimized by c^0 and therefore

$$\lim_{n\to\infty} \frac{n}{a_n} M_{n,\kappa,c^0} = [\rho c^0 V(\kappa)]/(\rho - 1) .$$

Note that $M^0_{n,\kappa} = M_{n,\kappa,a_n/h^0_n}$. Let j_n be increasing integers such that

$$\lim_{n\to\infty} h^0_{j_n}/a_{j_n} = lc^0 \in [0,\infty] . \tag{4.4}$$

Since (4.3) holds locally uniformly,

$$\lim_{n\to\infty} M^0_{j_n,\kappa}/M_{j_n,\kappa,c^0} = \frac{lc^0 V(\kappa) + (lc^0)^{1-\rho} B(\kappa)}{[\rho c^0 V(\kappa)]/(\rho-1)}$$
$$= [(\rho-1)l + l^{1-\rho}]/\rho \ge 1 .$$

Since in fact the limit must be no more than 1, we conclude $l = 1$. This being true for any convergent subsequence satisfying (4.4), the result holds for the sequence itself. Both limits in the corollary statement follow. □

We next demonstrate the asymptotic behavior of the minimum MISE, J_n.

PROOF OF PROPOSITION 1.1. Let H be as in Lemma 3.3. Since $\rho > 1$, it readily follows from Lemma 3.3 and Jensen's inequality that

$$\lambda = \lim_{t\to\infty} \int_0^1 [\chi_0(tu)/g_0(0)]^{1/\rho} du = \int_0^\infty y^{1/\rho} H(dy) \le 1 .$$

Indeed, $\lambda < 1$ except when H is degenerate at 1 or, equivalently, when g is nonzero at exactly one point.

We have from (1.3),

$$\frac{n}{a_n} J_n = \frac{n}{2\pi a_n} \int_{-\infty}^{\infty} \frac{\phi_0(t)}{1 + n\phi_0(t)} \left[1 - \frac{n\phi_0^2(t)}{1 + (n-1)\phi_0(t)} \right] dt .$$

Since $a_n \to \infty$,

$$\limsup_{n\to\infty} \frac{n}{2\pi a_n} \int_{-\infty}^{\infty} \frac{n\phi_0^3(t)}{(1 + n\phi_0(t))(1 + (n-1)\phi_0(t))} dt$$
$$\le \lim_{n\to\infty} \frac{1}{2\pi a_n} \int_{-\infty}^{\infty} \phi_0(t) dt = 0 .$$

Thus we need only consider the limiting behavior of

$$Q_n = \frac{n}{2\pi a_n}\int_{-\infty}^{\infty} \frac{\phi_0(t)}{1 + n\phi_0(t)}\, dt = \frac{1}{\pi}\int_0^{\infty} \frac{\xi_n(t)}{\xi_n(t) + t^\rho}\, dt ,$$

where $\xi_n(t) = nt^\rho \phi_0(a_n t)$. We have

$$Q_n \le \frac{(1+\varepsilon)}{\varepsilon\pi}\int_0^{\infty}\int_1^{1+\varepsilon} \frac{\xi_n(t)/u}{\xi_n(t) + t^\rho}\, dudt \tag{4.5}$$

$$\le \frac{(1+\varepsilon)}{\varepsilon\pi}\int_0^{\infty}\int_1^{1+\varepsilon} \frac{\xi_n(tu)}{\xi_n(tu) + t^\rho}\, dudt .$$

Consider the inner integral in this last expression. By Lemma 3.4 and the fact that ξ_n is symmetric,

$$\lim_{n\to\infty} \frac{1}{\varepsilon}\int_1^{1+\varepsilon} \frac{\xi_n(tu)}{\xi_n(tu) + t^\rho}\, du = \int_0^{\infty} \frac{y}{y + t^\rho}\, H(dy) . \tag{4.6}$$

In addition, for large enough n,

$$\frac{1}{\varepsilon}\int_1^{1+\varepsilon} \frac{\xi_n(tu)}{\xi_n(tu) + t^\rho}\, du \le \min\left[\, 1, t^{-\rho}\, \frac{1}{\varepsilon}\int_1^{1+\varepsilon} \xi_n(tu)\, du \right]$$

$$\le (1+\varepsilon)\, \min\,(1, t^{-\rho}) .$$

Applying (4.6) and Lebesgue convergence to (4.5),

$$\limsup_{n\to\infty} \frac{n}{a_n}\, J_n = \limsup_{n\to\infty}\, Q_n \le \frac{(1+\varepsilon)}{\pi}\int_0^{\infty}\int_0^{\infty} \frac{y}{y + t^\rho}\, H(dy)\, dt .$$

Since ε is arbitrary,

$$\limsup_{n\to\infty} \frac{n}{a_n}\, J_n \le \frac{1}{\pi}\int_0^{\infty}\int_0^{\infty} \frac{y}{y + t^\rho}\, H(dy)\, dt$$

$$= \frac{1}{\pi}\int_0^{\infty} y^{1/\rho} H(dy) \int_0^{\infty} (1 + t^\rho)^{-1} dt$$

$$= \frac{\lambda}{\pi}\int_0^{\infty} (1 + t^\rho)^{-1} dt .$$

By a similar argument, the limit infinum obtains. □

Acknowledgement

The author is indebted to Professor Jeffrey D. Hart for many helpful discussions.

REFERENCES

Bretagnolle, J. and Huber, C. (1979). Estimation des densités: risque minimax, *Z. Wahrsch. Verw. Gebiete*, **47**, 119–137.

Cline, D. B. H. (1988*a*). Admissible kernel estimators of a multivariate density, *Ann. Statist.*, **16**, 1421–1427.

Cline, D. B. H. (1988*b*). Abelian and Tauberian theorems relating the local behavior of an integrable function to the tail behavior of its Fourier transform, *J. Math. Anal. Appl.*, (to appear).

Cline, D. B. H. (1989). Consistency for least squares regression estimators with infinite variance data, *J. Statist. Plann. Inference*, **23**, 163–179.

Cline, D. B. H. and Hart, J. D. (1987). Kernel estimation of densities with discontinuities or discontinuous derivatives, Tech. Report, 87-9, Statistics Dept., Texas A&M Univ., Texas.

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

Efroimovich, S. Y. (1985). Nonparametric estimation of a density of unknown smoothness, *Theory Probab. Appl.*, **30**, 557–568.

Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density, *Theory Probab. Appl.*, **14**, 153–158.

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. and Marron, J. S. (1988). Choice of kernel order in density estimation, *Ann. Statist.*, **16**, 161–173.

Hart, J. D. (1988). An ARMA type probability density estimator, *Ann. Statist.*, **16**, 842–855.

Ibragimov, I. A. and Khasminskii, R. Z. (1983). Estimation of a distribution density, *J. Sov. Math.*, **21**, 40–57.

Lukacs, E. (1983). *Developments in Characteristic Function Theory*, Oxford Univ. Press, New York.

Marron, J. S. and Härdle, W. (1986). Random approximations to some measures of accuracy in nonparametric curve estimation, *J. Multivariate Anal.*, **20**, 91–113.

Müller, H.-G. and Gasser, Th. (1979). Optimal convergence properties of kernel estimates of derivatives of a destiny function, *Smoothing Techniques for Curve Estimation*, (eds. Th. Gasser and M. Rosenblatt), 144–154, Springer, Berlin (Lecture Notes in Mathematics, No. 757).

Müller, H.-G. (1984). Smooth optimum kernel estimators of densities, regression curves and modes, *Ann. Statist.*, **12**, 776–774.

Parzen, E. (1958). On asymptotically efficient consistent estimates of the spectral density of a stationary time series, *J. Roy. Statist. Soc. Ser. B*, **20**, 303–322.

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

Sacks, J. and Ylvisacker, D. (1981). Asymptotically optimum kernels for density estimation at a point, *Ann. Statist.*, **9**, 334–346.

Seneta, E. (1976). Regularly varying functions, *Lecture Notes in Mathematics* 508, Springer, New York-Berlin.

Stone, C. (1983). Optimal uniform rate of convergence for nonparametric estimators of a density function or its derivatives, *Recent Advances in Statistics: Papers in Honor of Herman Chernoff on His Sixtieth Birthday*, (eds. M. H. Rizvi, J. S. Rustagi and D. Siegmund), Academic Press, New York.

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

Thaler, H. (1974). Non-parametric probability density estimation and the empirical characteristic function, Tech. Report, 14, Statist. Sci. Div., State Univ. of New York at Buffalo, New York.

Titchmarsh, E. C. (1948). *Introduction to the Theory of Fourier Integrals*, 2nd ed., Oxford Univ. Press, New York.

van Eeden, C. (1985). Mean integrated squared error of kernel estimators when the density and its derivative are not necessarily continuous, *Ann. Inst. Statist. Math.*, **37**, 461–472.

Wahba, G. (1975). Interpolating spline methods for density estimates I: Equi-spaced knots, *Ann. Statist.*, **3**, 15–29.

Wahba, G. (1981). Data-based optimal smoothing of orthogonal series density estimates, *Ann. Statist.*, **9**, 146–156.

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