*The Annals of Statisitcs*
1977, Vol. 5, No. 3, 530-535

# MEAN INTEGRATED SQUARE ERROR PROPERTIES OF DENSITY ESTIMATES

BY KATHRYN BULLOCK DAVIS

*University of Washington*

The rate at which the mean integrated square error decreases as sample size increases is evaluated for general $L^1$ kernel estimates and for the Fourier integral estimate for a probability density. The rates are compared to that of the minimum M.I.S.E.; the Fourier integral estimate is found to be asymptotically optimal.

**1. Introduction.** Define the estimate $f^{\lambda(n)}$ of a density $f$ by

$$f^{\lambda(n)}(x) = \frac{1}{n} \sum_{i=1}^{n} K_{\lambda(n)}(x - X_i)$$

where $X_1, \cdots, X_n$ are independent identically distributed random variables with probability density $f$ and the kernel $K_{\lambda(n)}$ is square integrable as is $f$. Then the mean integrated square error (M.I.S.E.) given by $J(f^{\lambda(n)}) = E(\int (f^{\lambda(n)}(x) - f(x))^2\, dx)$ is well defined. (Integrals where no limits are written are taken to be over the entire real line.) Watson and Leadbetter (1963) showed that the minimum M.I.S.E. within this class of estimates is

$$J_n^* = \frac{1}{2\pi} \int \frac{|\Phi_f(t)|^2(1 - |\Phi_f(t)|^2)}{1 + (n-1)|\Phi_f(t)|^2}\, dt$$

where $\Phi_f$ is the characteristic function of $f$. The Fourier transform of the kernel which gives this M.I.S.E. was also derived; however, the kernel depends on the unknown density $f$ and is often difficult to evaluate.

A large subclass of such estimates often considered (e.g., Parzen, 1962) is the class in which $K_\lambda$ is a kernel satisfying $K_\lambda(x) = \lambda K(\lambda x)$ and $\int K(y)\, dy = 1$, where $\lambda(n)$, the scaling parameter, is a nonnegative increasing function such that $\lambda(n)/n \to 0$ as $n \to \infty$ and $\lambda(n) \to R$ as $n \to \infty$ where $R = \inf\{r\colon \Phi_f(t) = 0 \text{ a.e. for } t > r > 0\}$. This is the class of estimates considered in this paper. The rate of decrease of the M.I.S.E. for estimates of three large classes of densities, those whose characteristic functions decrease algebraically, exponentially, or have compact support, is given and compared with the rate of decrease of $J_n^*$. For $L^1$ kernels, the rate of decrease of the M.I.S.E. is shown to be generally less than the rate of $J_n^*$. For the kernel $(\pi x)^{-1} \sin x$, however, the rate of decrease of the M.I.S.E. is shown to be of the same order as $J_n^*$.

**2. Order of consistency.** An estimator $f^{\lambda(n)}$ is said to be integratedly consistent of order $H(n)$, where $H(n)$ is a nonnegative increasing function such that

Received February 1976.

*AMS* 1970 *subject classification.* Primary 62G05.

*Key words and phrases.* Nonparametric estimation, density estimation, kernel estimates, Fourier integral estimate, order of consistency.

$H(n) \to \infty$ as $n \to \infty$, if $H(n)J(f^{\lambda(n)})$ tends to a limit which is finite and nonzero as $n \to \infty$. An estimate $f^{\lambda(n)}$ is asymptotically optimal if $J_n^*/J(f^{\lambda(n)}) \to 1$ as $n \to \infty$. Watson and Leadbetter (1963) showed that the order of consistency $H_n^*$ of $J_n^*$ is always less than or equal to $n$ and, for densities whose characteristic function has compact support, is equal to $n$. In fact, this is the only case for which $H_n^*$ is equal to $n$. To show this, suppose $H_n^*/n \to A$ and $H_n^* J_n^* \to L$ as $n \to \infty$ where $A$ and $L$ are finite and nonzero. By Fatou's lemma,

$$\frac{1}{2\pi} \int \lim_{n\to\infty} \left( \frac{n|\Phi_f(t)|^2(1 - |\Phi_f(t)|^2)}{1 + (n-1)|\Phi_f(t)|^2} \right) dt \leqq \lim_{n\to\infty} nJ_n^* = \frac{L}{A} < \infty .$$

But the limit on the left is

$$\int (1 - |\Phi_f(t)|^2)\chi_f(t)\, dt$$

(where $\chi_f(t) = 0$ if $|\Phi_f(t)| = 0$ and $\chi_f(t) = 1$ if $|\Phi_f(t)| > 0$) and this integral exists if and only if $\Phi_f$ has compact support.

The order of consistency $H(n)$ and the scaling parameter $\lambda(n)$ are related. The M.I.S.E. for the estimate $f^{\lambda(n)}$ may be written

$$2\pi J(f^{\lambda(n)}) = n^{-1}\lambda(n) \int |\Phi_K(t)|^2(1 - |\Phi_f(\lambda(n)t)|^2)\, dt + \lambda(n) \int |\Phi_f(\lambda(n)t)|^2 |1 - \Phi_K(t)|^2\, dt . \tag{2.1}$$

If $H(n)$ is the order of consistency of $f^{\lambda(n)}$, then, from the first integral on the right, $H(n)\lambda(n)/n \to L$ as $n \to \infty$ where $L$ is positive and finite. This is consistent with the properties of $\lambda(n)$ given in Section 1.

The remainder of this section will show the order of consistency for the estimate using the kernel $(\pi x)^{-1} \sin x$. The resulting estimate is called the Fourier integral estimate (F.I.E.) since it is derived by evaluating the Fourier integral over $(-\lambda(n), \lambda(n))$ with the sample characteristic function substituted for $\Phi_f$. The M.I.S.E. for the F.I.E. is given by

$$J(f^\lambda) = (2\pi)^{-1}||\Phi_f||_2^2 + (\pi n)^{-1}(\lambda - (n+1)\int_0^\lambda |\Phi_f(t)|^2\, dt) . \tag{2.2}$$

By differentiating this expression, it may be shown the M.I.S.E. is minimized when $|\Phi_f(\lambda(n))|^2 = (n+1)^{-1}$. $\lambda(n)$ is uniquely defined by this expression and satisfies the requirements for the scaling parameter given in Section 1 when $|\Phi_f(t)|$ is monotone decreasing as $|t|$ increases. This is the case with all common densities except the uniform. The expression also suggests a sample based method of estimating the optimal $\lambda(n)$. An unbiased estimate for $|\Phi_f(t)|^2$ may be easily constructed using the sample characteristic function and substituted into the formula; the expression is then solved for the smallest such $\lambda(n)$. The following theorem shows the F.I.E. (with optimal $\lambda(n)$) has the same order of consistency as $J_n^*$.

THEOREM 2.1. *Let $f \in L^2$ and suppose $|\Phi_f(t)|$ is monotone decreasing as $|t|$ increases. Let $f^{\lambda(n)}$ be the* F.I.E. *where $|\Phi_f(\lambda(n))|^{-2} = n + 1$; then*

$$\tfrac{1}{2} \leqq \lim_{n\to\infty} \inf J_n^*/J(f^{\lambda(n)}) \leqq \lim_{n\to\infty} \sup J_n^*/J(f^{\lambda(n)}) \leqq 1 .$$

PROOF. Let

$$g(x) = \frac{x(1-x)}{1+(n-1)x}, \qquad 0 < x < 1,$$

and

$$\begin{aligned} h(x) &= \tfrac{1}{2}x && \text{if } 0 < x < (n+1)^{-1}, \\ &= (2n)^{-1}(1-x) && \text{if } (n+1)^{-1} < x < 1. \end{aligned}$$

It is easily verified that

$$h(x) \leqq g(x) \leqq 2h(x), \qquad 0 < x < 1.$$

Using this relation with the monotonicity of $|\Phi_f(t)|$ and (2.2),

$$\begin{aligned} \tfrac{1}{2}J(f^{\lambda(n)}) &= (2\pi)^{-1}\textstyle\int_{-\lambda(n)}^{\lambda(n)} (2n)^{-1}(1-|\Phi_f(t)|^2)\,dt + (2\pi)^{-1}\int_{|t|>\lambda(n)} \tfrac{1}{2}|\Phi_f(t)|^2\,dt \\ &< (2\pi)^{-1}\int \frac{|\Phi_f(t)|^2(1-|\Phi_f(t)|^2)}{1+(n-1)|\Phi_f(t)|^2}\,dt = J_n^* \\ &< (2\pi)^{-1}\textstyle\int_{-\lambda(n)}^{\lambda(n)} n^{-1}(1-|\Phi_f(t)|^2)\,dt + (2\pi)^{-1}\int_{|t|>\lambda(n)} |\Phi_f(t)|^2\,dt \\ &= J(f^{\lambda(n)}). \end{aligned}$$

Thus $J_n^* < J(f^{\lambda(n)}) < 2J_n^*$, and the theorem follows.

**3. Characteristic functions which decrease algebraically.** A characteristic function $\Phi_f$ is said to decrease algebraically of degree $p > 0$ if

$$\lim_{t\to\infty} |t|^p|\Phi_f(t)| = B^{\frac{1}{2}}, \qquad 0 < B < \infty.$$

This class includes the gamma, chi-square ($2p$ = degrees of freedom), exponential ($p = 1$), and double exponential ($p = 1$) probability densities. Since it is assumed $f \in L^1 \cap L^2$, then $\Phi_f \in L^2$ and necessarily $p > \frac{1}{2}$. Watson and Leadbetter (1963) showed that

$$\lim_{n\to\infty} n^{1-1/2p}J_n^* = \frac{1}{\pi}B^{1/2p}\int_0^\infty (1+t^{2p})^{-1}\,dt.$$

Since estimates which are of the same order of consistency as $J_n^*$ are the estimates of interest, in this section $\lambda(n)$ will be defined to be $(Bn)^{1/2p}$ (so that $H_n^*\lambda(n)/n = B$). Watson and Leadbetter (1963) showed that kernel estimates with this $\lambda(n)$ for which $\int_0^\infty (1-\Phi_K(t))^2t^{-2p}\,dt$ exists satisfy

$$\lim_{n\to\infty} n^{1-1/2p}J(f^{\lambda(n)}) = \pi^{-1}B^{1/2p}\Big(\int_0^\infty \Phi_K{}^2(t)\,dt + \int_0^\infty (1-\Phi_K(t))^2t^{-2p}\,dt\Big). \tag{3.1}$$

For what kernels does $\int_0^\infty (1-\Phi_K(t))^2t^{-2p}\,dt$ exist? For $K \in L^1$, $\Phi_K(t)$ is continuous and bounded so $\int_1^\infty (1-\Phi_K(t))^2t^{-2p}\,dt$ exists for $p > \frac{1}{2}$. Suppose there exists an integer $r$ such that $\int x^mK(x)\,dx = 0$, $m = 1, 2, \cdots, r-1$ and $\int x^rK(x)\,dx \neq 0$. Then, for $K \in L^1$,

$$\begin{aligned} t^{-r}(1-\Phi_K(t)) &= \int (1-e^{itx})t^{-r}K(x)\,dx \\ &= -(r!)^{-1}i^r\int x^rK(x)\,dx \\ &\quad - \int \left(e^{itx} - 1 - itx - \frac{(itx)^2}{2!} - \cdots - \frac{(itx)^r}{r!}\right)t^{-r}K(x)\,dx \\ &\to -(r!)^{-1}i^r\int x^rK(x)\,dx \qquad \text{as } t \to 0. \end{aligned}$$

(The value $r$ is called the characteristic exponent of the transform $\Phi_K$.) Comparing the integrand with $t^{2(r-p)}$, $\int_0^1 (1 - \Phi_K(t))^2 t^{-2p}\,dt$ converges for $r \geqq p$ and diverges for $r < p$. Thus for $L^1$ kernels the convergence of $\int_0^\infty (1 - \Phi_K(t))^2 t^{-2p}\,dt$ and hence the rate of convergence of the M.I.S.E. depend on the characteristic exponent $r$. The most important group of these kernels are the weighting functions considered by Parzen (1962). These kernels are even and positive, and thus have characteristic exponent $\leqq 2$. Therefore these kernels have a M.I.S.E. which decreases at the same rate as $J_n^*$ only when the density being estimated has $p \leqq 2$. Examples of such densities are the exponential, double exponential, and chi-square with 1 degree of freedom. Densities which have more than two derivatives are too smooth for these kernels.

The kernel for the F.I.E. is not $L^1$ and does not have these limitations. From (3.1),

$$\lim_{n\to\infty} n^{1-1/2p} J(f^{\lambda(n)}) = \pi^{-1} B^{(1/2p)} 2p(2p-1)^{-1}$$

so the F.I.E. has the same order of consistency as $J_n^*$ and

$$\lim_{n\to\infty} J_n^*/J(f^{\lambda(n)}) = (1 - (2p)^{-1}) \int_0^\infty (1 + t^{2p})^{-1}\,dt\,.$$

Since

$$\int_0^\infty (1 + t^{2p})^{-1}\,dt > \Gamma(1 + (2p)^{-1}) \geqq 1 \qquad \text{for} \quad p > \tfrac{1}{2}\,,$$

it follows that

$$\max(\tfrac{1}{2},\, 1 - (2p)^{-1}) \leqq \lim_{n\to\infty} \inf J_n^*/J(f^{\lambda(n)}) \leqq \lim_{n\to\infty} \sup J_n^*/J(f^{\lambda(n)}) \leqq 1$$

and

$$\lim_{p\to\infty} \lim_{n\to\infty} J_n^*/J(f^{\lambda(n)}) = 1\,.$$

Thus the F.I.E. is closer to the asymptotic optimality property as $p$ increases; that is, as the smoothness of the underlying density increases.

**4. Characteristic functions which decrease exponentially.** A characteristic function $\Phi_f$ is said to decrease exponentially with degree $r$ and coefficient $\rho$ if

(i) $|\Phi_f(t)| \leqq Ae^{-\rho|t|^r}$ for some constants $A > 0$, $\rho > 0$, $0 < r \leqq 2$

and

(4.1) (ii) $\lim_{t\to\infty} \int_0^1 (1 + \exp(2\rho t^r)|\Phi_f(tx)|^2)^{-1}\,dx = 0\,.$

This class includes the normal probability density ($A = 1$, $\rho = \frac{1}{2}\sigma^2$, $r = 2$) and the Cauchy density ($A = 1$, $\rho = 1$, $r = 1$). Watson and Leadbetter (1963) showed

$$\lim_{n\to\infty} n(\log n)^{-1/r} J_n^* = \pi^{-1}(2\rho)^{-1/r}\,.$$

Accordingly, $\lambda(n)$ such that

$$\lim_{n\to\infty} \lambda(n)(\log n)^{-1/r} = (2\rho)^{-1/r}$$

will be used in this section.

THEOREM 4.1. *Let $f \in L^2$ and $\Phi_f(t)$ decrease exponentially with coefficient $\rho$ and degree $r$. Suppose*

(4.2) $$\lim_{n\to\infty} \lambda(n)(\log n)^{-1/r} = (2\rho)^{-1/r}\,.$$

*Then*

$$\limsup_{n\to\infty} n(\log n)^{-1/r} J(f^{\lambda(n)}) < \infty$$

*if and only if* $\Phi_K(t) = 1$ a.e., $0 < t < 1$.

PROOF. Rewriting the expression (2.1) for $J(f^{\lambda(n)})$,

$$2\pi n(\log n)^{-1/r} J(f^{\lambda(n)}) = \lambda(n)(\log n)^{-1/r} \int_{-\infty}^{\infty} |\Phi_K(t)|^2(1 - |\Phi_f(t\lambda(n))|^2)\, dt$$
$$n\lambda(n)(\log n)^{-1/r} \int_{-\infty}^{\infty} |\Phi_f(\lambda(n)t)|^2 |1 - \Phi_K(t)|^2\, dt\,.$$

Since $\Phi_K \in L^2$, by Riemann–Lebesgue the first integral on the right has limit $(2\rho)^{-1/r}||\Phi_K||_2^2$ as $n \to \infty$. The second integral may be broken into two parts. First,

$$n\lambda(n)(\log n)^{-1/r} \int_1^{\infty} |\Phi_f(\lambda(n)t)|^2 |1 - \Phi_K(t)|^2\, dt$$
$$\leqq A^2 n(2\rho \log n)^{-1/r} \int_{(2\rho)^{1/r}\lambda(n)}^{\infty} e^{-t^r}\, dt \to 0 \qquad \text{as} \quad n \to \infty\,.$$

(The limit follows from (4.2) and the fact $\lim_{\lambda\to\infty} \lambda^{-1}e^{\lambda^r} \int_{\lambda}^{\infty} e^{-t^r}\, dt = 0$, $r > 0$.) For the other part of the integral, note from (4.1) that

$$\lim_{n\to\infty} e^{2\rho\lambda(n)^r} |\Phi_f(\lambda(n)t)|^2 = \infty \qquad \text{for} \quad 0 < t < 1\,.$$

Using this and (4.2),

$$\lim_{n\to\infty} \lambda(n)(\log n)^{-1/r} \int_0^1 n|\Phi_f(\lambda(n)t)|^2 |1 - \Phi_K(t)|\, dt$$

is finite (and equal to zero) if and only if $\Phi_K(t) = 1$, a.e., $0 < t < 1$, and the theorem is proved.

Since $\Phi_K(t) = 1$, $0 < t < 1$, is the Fourier transform for the kernel of the F.I.E., for this special kernel estimate the limit may be evaluated:

$$\lim_{n\to\infty} n(\log n)^{-1/r} J(f^{\lambda(n)}) = \pi^{-1}(2\rho)^{-1/r}\,,$$

and it follows that

$$\lim_{n\to\infty} J_n^*/J(f^{\lambda(n)}) = 1$$

if and only if $f^{\lambda(n)}(x)$ is the F.I.E. (a.e.).

Watson and Leadbetter derive an entirely different type of kernel estimate based on kernels with

$$\Phi_{K_\lambda}(t) = \Phi_K(\lambda e^{\alpha|t|})\,, \quad \alpha > 0$$

which also have the asymptotic optimum property with $\lambda(n) = cn^{-b}$ where $\alpha = 2\rho b$. No examples of such kernels are given.

**5. Characteristic functions with compact support.** For densities in this class, $H_n^* = n$ so $\lambda(n)$ is chosen to satisfy $\lambda(n) \to R$ as $n \to \infty$. From (2.1),

$$2\pi n J(f^{\lambda(n)}) = \lambda(n) \int |\Phi_K(t)|^2(1 - |\Phi_f(\lambda(n)t)|^2)\, dt$$
$$+ n\lambda(n) \int_{-R/\lambda}^{R/\lambda} |\Phi_f(\lambda(n)t)|^2 |1 - \Phi_K(t)|^2\, dt\,.$$

The limit of the first integral is $R \int |\Phi_K(t)|^2(1 - |\Phi_f(Rt)|^2)\, dt$. The limit of the second expression, however, exists (and is zero) if and only if $\Phi_K(t) = 1$ a.e. for

$|t| < 1$. Thus again the F.I.E. is the only estimate with the same order of consistency as $J_n^*$. The F.I.E. is also easily shown to be asymptotically optimal.

**6. Summary.** The rate at which the mean integrated square error decreases as the sample size increases for the kernel estimates considered here depends on the smoothness of the probability density being estimated. For densities with only two derivatives, estimates using nonnegative $L^1$ kernels have mean integrated square errors which decrease at the same rate as the minimum mean integrated square error $J_n^*$. For densities with higher order derivatives, however, $L^1$ kernels do not perform as well; in general $J_n^*/J(f^{\lambda(n)}) \to 0$ as $n \to \infty$. In contrast, for the Fourier integral estimate the rate of decrease of the mean integrated square error improves with the smoothness of the density under consideration. Under minimal conditions $\frac{1}{2} \leqq J_n^*/J(f^{\lambda(n)}) \leqq 1$, so the rate of decrease is of the same order as $J_n^*$. These results agree with the earlier conclusions (Davis, 1975) for the mean square error. The Fourier integral estimate has good asymptotic error properties for a wider class of densities than does an estimate formed using an $L^1$ kernel.

## REFERENCES

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

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

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

DEPARTMENT OF BIOSTATISTICS JD-30
UNIVERSITY OF WASHINGTON
SEATTLE, WASHINGTON 98195