The Annals of Statistics
1975, Vol. 3, No. 4, 1025-1030

MEAN SQUARE ERROR PROPERTIES OF DENSITY ESTIMATES[1]

BY KATHRYN BULLOCK DAVIS

University of Washington

The rate at which the mean 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 function. The estimates are then compared on the basis of these rates.

1. Introduction. Estimates of probability density functions by kernel series methods are now common in the literature. Both the theoretic and the Monte Carlo results have led some to wonder whether indeed the choice of kernel makes much difference. While the rate of decrease of the bias depends on the particular kernel chosen, in this paper it will be shown that within certain classes of kernels, the rates are the same. For L^1 kernels, if $f^{(m)}$ exists, the rate is at most λ^{-m}, where λ is the scaling parameter. If the L^1 kernels are restricted to be nonnegative, the rate is at most λ^{-2}, regardless of the smoothness of f. For the Fourier integral estimate, the rate of decrease of the bias depends on the smoothness of f, so that for sufficiently smooth functions, the rate is much faster than the rate for L^1 kernels.

2. Kernel estimates. A kernel estimate is an estimate $f^\lambda(x)$ of the probability density f of the form

$$f^\lambda(x) = \frac{1}{n} \sum_1^n K_\lambda(x - X_i)$$

where $X_1, \cdots, X_n$ are independent identically distributed random variables with probability density f. In this paper K_λ will be a kernel satisfying $K_\lambda(x) = \lambda K(\lambda x)$ and $\int K(y)\, dy = 1$, where $\lambda(n)$, the scaling parameter, is nonnegative increasing function such that

$$\lim_{n\to\infty} \lambda(n) = \infty \qquad \text{and} \qquad \lim_{n\to\infty} \lambda(n)/n = 0\,.$$

Integrals where no limits are written are to be taken over the entire real line. Important examples of kernels are given in the following table:

$K(x) = \pi^{-1}(1 + x^2)^{-1}$	(Cauchy)
$K(x) = \pi^{-1}(\sin x/x)^2$	(Fejér–de la Vallée Poussin)
$K(x) = \pi^{-\frac{1}{2}} e^{-x^2}$	(Weierstrass)

Received June 17, 1974; revised January 1975.

[1] This paper is based on part of the author's doctoral dissertation, which was supported by the National Institutes of Health under Public Health Service Grant 5-T01-GM-1269-10.

AMS 1970 *subject classification.* Primary 62G05.

Key words and phrases. Nonparametric estimation, density estimation, kernel estimates, Fourier integral estimate.

$K(x) = \frac{1}{2}e^{-|x|}$ (Picard)

$K(x) = 3\pi^{-1}(\sin x/x)^4$ (Jackson–de la Vallée Poussin)

$K(x) = \frac{1}{2}$ if $|x| \leqq 1$

$= 0$ if $|x| > 1$ (the "moving average")

$K(x) = (\pi x)^{-1} \sin x$ (the Fourier integral estimate kernel)

Estimates formed using these kernels are asymptotically unbiased, consistent, and asymptotically normal. At continuity points of $f(x)$, the variance of the estimate converges at the rate $\lambda(n)/n$ (Parzen (1962), Konakov (1973)).

The expected value of the estimate is given by

$$E(f^\lambda(x)) = \int K_\lambda(x - y)f(y)\,dy = (K_\lambda * f)(x) .$$

The bias b then has the simple expression

$$b(f^\lambda(x)) = (f - K_\lambda * f)(x) .$$

3. Rate of convergence of the bias for L^1 kernels. Shapiro (1969) has many results with direct bearing on density estimation (see also Butzer and Nessel (1971)). Shapiro's (1969) general theme is the relation between a given function f and its "smoothed" version obtained by forming its convolution with K_λ. The kernels considered are quite general kernels $K \in L^1(-\infty, \infty)$. (The kernel for the Fourier integral estimate is not in this class.)

A first result, true for all densities, is: if $K \in L^1(-\infty, \infty)$ then $\int |b(f^{\lambda(n)}(x))|\,dx \to 0$ as $n \to \infty$ (Shapiro, page 11). If in addition f is uniformly continuous and bounded on $(-\infty, \infty)$ then $b(f^{\lambda(n)}(x)) \to 0$ uniformly as $n \to \infty$ (Shapiro, page 13). With additional restrictions on K, pointwise convergence of the bias for general densities may be shown. Parzen (1962) gives criteria for "weighting functions" which are an example of such restrictions; Shapiro (page 14) gives slightly less restrictive conditions for the same results.

For "nice" densities, then, the estimate f^λ will be asymptotically unbiased for any kernel $K_\lambda \in L^1$. In order to compare estimates using different kernels, the rates of convergence of the bias may be compared. In general, the rate of convergence improves with the smoothness of the density f being estimated; however, for some kernels there may be a limit beyond which even if greater smoothness of f is assumed, the rate of convergence does not increase. This phenomenon is called "saturation."

THEOREM 3.1. (*Pointwise saturation theorem*, [12], page 27). *Let* $K \in L^1$, $x^2K \in L^1$, $\int xK(x)\,dx = 0$, *and* $A = \int x^2K(x)\,dx$. *If* f *is a bounded measurable function and if* $f''(x)$ *exists, then*

$$\lim_{n\to\infty} \lambda^2(n)b(f^\lambda(x)) = -\tfrac{1}{2}Af''(x) .$$

Kernels which satisfy the theorem include all nonnegative even kernels with $x^2K \in L^1$, in particular, those of Weirstrass, Picard, Jackson–de la Vallée Poussin, and the moving average. Note that for nonnegative kernels, A is always nonzero

so that this is the best possible result, that is, the bias cannot decrease any faster than $1/\lambda^2$. For the Cauchy and Fejér–de la Vallée Poussin kernels, the asymptotic decrease is slower, of the order $1/\lambda$:

THEOREM 3.2. ([12], page 33). *Let f be bounded and measurable on $(-\infty, \infty)$ and suppose*

$$A(f; x) = \pi^{-1} \int_0^\infty t^{-2}(f(x + t) - 2f(x) + f(x - t))\, dt$$

exists as a Lebesgue integral for some particular value of x. If K is the Cauchy kernel, then $\lim_{n\to\infty} \lambda(n)b(f^\lambda(x)) = -A(f; x)$. If K is the Fejér–de le Vallée Poussin kernel, then

$$\lim_{n\to\infty} \lambda(n)b(f^\lambda(x)) = -\tfrac{1}{2}A(f; x)\,.$$

All of Parzen's (1962) examples of weighting functions are included in these two theorems. Intuitively one might think nonnegative kernels might provide the best estimates since they are themselves densities. The following theorem shows the converse is true. If the kernels are not restricted to be nonnegative, the degree of approximation may actually improve, although the resulting density estimate may be negative at some points. (A similar theorem with different restrictions on K appears in Parzen (1958) in the context of estimation of spectral density functions.)

THEOREM 3.3. ([12], page 31). *Let $K \in L^1$, $x^m K \in L^1$ where m is a positive integer $\geqq 1$, $\int x^r K(x)\, dx = 0$ for $r = 1, 2, \cdots, m - 1$, and $\int x^m K(x)\, dx = A \neq 0$. If f is a bounded measurable function and if $f^{(m)}(x)$ exists, then*

$$\lim_{n\to\infty} \lambda(n)^m b(f^\lambda(x)) = -(Af^{(m)}(x)/m!\,.$$

4. The Fourier integral estimate. Shapiro's theorems do not apply to the Fourier integral estimate since $K \notin L^1$. The bias for the Fourier integral estimate is given by

$$b(f^\lambda(x)) = -\tfrac{1}{2}\pi \int_{|t|>\lambda} \Phi_f(t)e^{-itx}\, dt \tag{4.1}$$

(Parzen (1967), Davis (1974b)). The rate of decrease of the bias depends on the smoothness of f as reflected in the rate of decrease of the characteristic function $\Phi_f(t)$. This relationship may be seen in expression

$$f^{(m)}(x) = \int (it)^m \Phi_f(t)e^{itx}\, dx\,,$$

which holds if both the derivative $f^{(m)}(x)$ and the integral exist. The rate of convergence for certain classes of characteristic functions, those which decrease exponentially and those which decrease algebraically, as defined by Watson and Leadbetter (1963) and similarly by Parzen (1958), will be investigated below.

A characteristic function is said to decrease exponentially with degree r and coefficient ρ if

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

and

$$\text{(ii)} \quad \lim_{t\to\infty} \int_0^1 (1 + \exp(2\rho t^r)|\Phi_f(tx)|^2)^{-1}\, dx = 0\,. \tag{4.2}$$

This class includes the normal probability density ($A = 1$, $\rho = \frac{1}{2}\sigma^2$, $r = 2$) and the Cauchy density ($A = 1$, $\rho = 1$, $r = 1$). Using integration by parts, it is easily shown that

$$\lim_{\lambda\to\infty} \lambda^{-1}e^{\lambda^r} \int_\lambda^\infty e^{-t^r}\,dt = 0\,, \qquad r > 0\,.$$

From the definition of exponential decrease,

$$|b(f^\lambda(x))| \leqq \pi^{-1} \int_\lambda^\infty Ae^{-t^r}\,dt\,.$$

Using a change of variable $u^r = \rho t^r$, an immediate result is:

THEOREM 4.1. *Suppose $\Phi_f(t)$ decreases exponentially with degree r and coefficient ρ. Then the bias $b(f^\lambda(x))$ of the* F.I.E. *satisfies*

$$\lim_{n\to\infty} \lambda(n)^{-1}e^{\rho\lambda(n)^r}|b(f^{\lambda(n)}(x)| = 0\,. \tag{4.3}$$

The bias of the F.I.E. then decreases quite rapidly, at least as fast as $\lambda e^{-\rho\lambda^r}$.

A characteristic function $\Phi_f(t)$ is said to decrease algebraically of degree $p > 0$ if

$$\lim_{t\to\infty} |t|^p|\Phi_f(t)| = K^{\frac{1}{2}} > 0\,. \tag{4.4}$$

This class includes the gamma, chi-square ($2p$ = degrees of freedom), exponential ($p = 1$), and double exponential ($p = 1$) probability densities. The F.I.E. $f^{\lambda(n)}(x)$ has been shown to be asymptotically unbiased so

$$\lim_{n\to\infty} |b(f^{\lambda(n)}(x))| = 0 \qquad \text{for all} \quad p\,. \tag{4.5}$$

Using (4.4), if $p > 1$, then

$$\lim_{\lambda\to\infty} \lambda^{p-1} \int_{|t|>\lambda} |\Phi_f(t)|\,dt = 2K^{\frac{1}{2}}(p-1)^{-1}\,. \tag{4.6}$$

Using 4.1, 4.5, and 4.6, the theorem follows:

THEOREM 4.2. *Suppose $\Phi_f(t)$ decreases algebraically of degree $p > 0$. Then the bias $b(f^{\lambda(n)}(x))$ of the* F.I.E. *satisfies*

$$\lim_{n\to\infty} |b(f^{\lambda(n)}(x))| = 0\,, \qquad p > 0$$

and

$$\lim_{n\to\infty} \lambda(n)^{p-1}|b(f^{\lambda(n)}(x))| \leqq K^{\frac{1}{2}}\pi^{-1}(p-1)^{-1}\,, \qquad p > 1\,.$$

The bias of the F.I.E. then decreases at the rate $\lambda(n)^{1-p}$.

5. Comparison of the mean square errors. For the F.I.E. and for general L^1 kernels, the variance tends to zero at the rate $\lambda(n)/n$. The rate of decrease of the bias depends on the kernel, the smoothness of the underlying density, and the function λ. It is interesting to compare the estimates when the optimal (in some sense) λ's are used in each estimate. For kernels satisfying Theorem 3.3 with $m = 2$, $\lambda(n)$ of the order $cn^{\frac{1}{5}}$, where c is a constant depending on the density f and the kernel K, has been shown to be asymptotically optimal under the criteria of mean square error (Rosenblatt, 1956) and of mean integrated square error (Epanechnikov, 1969). (This includes all of the examples cited by Parzen (1962) except the Cauchy and Fejér–de la Vallée Poussin, for which $m = 1$).

Davis (1974a) showed that the optimal $\lambda(n)$ for the F.I.E. in terms of MISE is of the order $(\log n/2\rho)^{1/r}$ for exponential decrease of degree r and coefficient ρ, and $n^{\frac{1}{2}p}$ for algebraic decrease of degree $p > \frac{1}{2}$. In the following let $f^{\lambda(n)}(x)$ be the F.I.E. and let $g^{\nu(n)}(x)$ be an L^1 kernel estimate with $\nu(n) = cn^{\frac{1}{2}}$.

For the exponential case, let $\lambda(n) = (\ln n/2\rho)^{1/r}$. Since $\nu(n)^2\lambda(n)\exp(-\rho\lambda(n)^r) = c^2((\ln n)/2\rho)^{1/r}n^{-\frac{1}{2}}$, and this tends to zero as $n \to \infty$, using Theorems 3.1 and 4.1 it follows that

$$\lim_{n\to\infty} b(f^{\lambda(n)}(x))/b(g^{\nu(n)}(x)) = 0 .$$

Upon examining the ratio $\lambda(n)/\nu(n)$, it is also clear that

$$\lim_{n\to\infty} \operatorname{Var}(f^{\lambda(n)}(x))/\operatorname{Var}(g^{\nu(n)}(x)) = 0 .$$

(These results hold in general if $\nu(n) = cn^a$ where $0 < a < \frac{1}{4}$.) Thus for the class of functions whose characteristic functions decrease exponentially, the F.I.E. is better than general L^1 kernel estimates in terms of rate of decrease of the mean square error.

For the algebraic case, let $\lambda(n) = n^{\frac{1}{2}p}$. Now $\nu(n)\lambda(n)^{1-p} = c^2n^{(1/2p-\frac{1}{2})}$ and this tends to zero for $p > 5$ as $n \to \infty$. For $p = 5$, the ratio is c^2. Thus from Theorems 3.1 and 4.2,

$$\lim_{n\to\infty} b(f^{\lambda(n)}(x))/b(g^{\nu(n)}(x)) = 0$$

if $p > 5$. Theorem 4.2 is not strong enough to provide results for $p \leqq 5$, although for $p = 5$ the limit is bounded. The result holds in general if $\nu(n) = cn^a$, $0 < a < \frac{1}{4}(1 - 1/p)$. From inspection of $\lambda(n)/\nu(n)$, the ratio of the variances satisfies

$$\begin{aligned} \lim_{n\to\infty} \operatorname{Var}(f^{\lambda(n)}(x))/\operatorname{Var}(g^{\nu(n)}(x)) &= 0 && \text{if } p > \tfrac{5}{2} \\ &= \pi^{-1}\|cK\|_2^{-2} && \text{if } p = \tfrac{5}{2} \\ &= \infty && \text{if } p < \tfrac{5}{2} \end{aligned}$$

where K is the particular kernel used in the L^1 estimate. Once again, the F.I.E. is the better estimate in terms of mean square error for smooth density functions, that is, functions with $p > 5$.

EXAMPLES.

(i) Let f be a normal (μ, σ^2) density. Then

$$|\Phi_f(t)| = e^{-\frac{1}{2}t^2\sigma^2}$$

so $\Phi_f(t)$ decreases exponentially. Thus the mean square error decreases faster for the F.I.E. than for an L^1 kernel estimate.

(ii) Let f be the chi-square density with r degrees of freedom. The characteristic function Φ_f satisfies

$$|\Phi_f(t)| = (1 + 4t^2)^{-r/4}$$

so that $\Phi_f(t)$ decreases algebraically of degree $r/2$. The F.I.E. is then the better estimate if the degrees of freedom exceed 10.

(iii) Let f be the exponential density $f(x) = ae^{-ax}$, $a > 0$, $x > 0$. Then

$$|\Phi_f(t)| = (1 - t^2/a^2)^{-\frac{1}{2}},$$

so $\Phi_f(t)$ decreases algebraically of degree 1. In this case the F.I.E. is inferior to even the Cauchy and Fejér–de la Vallée Poussin kernels. Note that $f'(0)$ does not exist, that is, f is not smooth.

REFERENCES

[1] BUTZER, P. L. and NESSEL, R. J. (1971). *Fourier Analysis and Approximation*, **1**. Academic Press, New York.

[2] DAVIS, K. B. (1974a). A Fourier integral estimate for probability density functions. Unpublished doctoral dissertation, Univ. of Washington.

[3] DAVIS, K. B. (1974b). A Fourier integral estimate for probability density functions. Submitted for publication.

[4] EPANECHNIKOV, V. A. (1969). Nonparametric estimation of a multivariate probability density. *Theor. Probability Appl.* **14** 153–158.

[5] KONAKOV, V. D. (1973). Nonparametric estimation of density functions. *Theor. Probability Appl.* **17** 361–362.

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

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

[8] PARZEN, E. (1967). *Time Series Analysis Papers*. Holden-Day, San Francisco.

[9] PARZEN, E. (1972). Some recent advances in time series analysis. *Lecture Notes in Physics Vol.* **12**: *Statistical Models and Turbulence*. (M. Rosenblatt and C. Van Atta, eds.). 470–492. Springer-Verlag, New York.

[10] ROSENBLATT, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **27** 832–837.

[11] ROSENBLATT, M. (1971). Curve estimates. *Ann. Math. Statist.* **42** 1815–1842.

[12] SHAPIRO, J. S. (1969). *Smoothing and Approximation of Functions*. Von Nostrand Reinhold, New York.

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