The Annals of Statistics
1977, Vol. 5, No. 1, 136–142

BOUNDS FOR ESTIMATION OF DENSITY FUNCTIONS AND THEIR DERIVATIVES

BY TERRY G. MEYER

Texas A and M University

Lower bounds on the radius of various confidence sets for density functions and their derivatives are determined. In each case investigated, the smallest radius obtainable and the radius actually obtained by using known estimates exhibit the same dependency on the fixed sample size n.

The lower bounds are derived using methods in Meyer (1976). Although only a few combinations of pseudometric and density class are considered here, the techniques illustrated can be used elsewhere with little conceptual difficulty.

1. Introduction. In estimating a density function, much attention has centered on kernel estimates, first proposed by Rosenblatt (1956) and later investigated extensively by Parzen (1962), Watson and Leadbetter (1963), and Schuster (1969), among others. Although the idea behind a kernel type estimate is conceptually appealing, rigorous arguments recommending these estimates over competing procedures were lacking until Farrell's (1972) paper. In that paper, Farrell considered a sequence of i.i.d. random variables, $X_1, X_2 \cdots, X_n$ with density $f(x)$ in a particular class C, and a sequence of estimators $\gamma_n(x_1, x_2, \cdots, x_n)$, one estimator for each sample size. Defining λ_n as any sequence for which

$$\liminf_n \inf_{f \in C} P_f\{|\gamma_n - f(0)| \leqq \lambda_n\} > 0 , \tag{1.1}$$

Farrell obtained "fastest" rates of convergence for λ_n. Since kernel estimates achieved the fastest rate for the classes of densities considered, an optimality property of kernel estimates was proved. Wahba (1975) recently used a slight variation of Farrell's proof to show that several types of density estimates including kernel estimates are optimal in this sense for a slightly different class of densities. The present paper shows that an improvement and a generalization of Farrell's procedure (Meyer (1976)) can be used to determine bounds for particular sequences $\{\lambda_n\}$ in a variety of cases not considered previously. Kernel estimates achieve the fastest rate allowable for λ_n in the cases considered.

Notation in this paper and references to theorems correspond to Meyer (1976). In accordance, we shall redefine $\{\lambda_n\}$ from (1.1) above to particular sequences as follows:

$$\lambda_n = \lambda_n(b) = \inf\{u : \inf_{\theta \in \Theta} P_\theta(d(\hat{t}_n, t(\theta)) \leqq u) \geqq 1 - b\} \tag{1.2}$$

where $\hat{t}_n = \hat{t}_n(x_1, x_2, \cdots, x_n)$ is an estimate of $t(\theta)$, and $d(\cdot, \cdot)$, Θ, P_θ and $t(\theta)$

Received February 1975; revised April 1976.

AMS 1970 *subject classifications*. Primary 62G05; Secondary 62G15.

Key words and phrases. Density estimation, minimal radius confidence set, probability convergence of density estimates.

are defined in Meyer (1976). Bounding λ_n is conceptually simple though computationally complex. Theorem 2.2, which in its statement includes the definition of $H_2(b)$, from Meyer (1976) is the basic tool used throughout this paper. For a given problem specified by a pseudometric d and a class $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$, the technique involves finding, in a suitably restricted subclass, a pair of measures which are furthest apart relative to d. Attention is restricted here to only a few of the more interesting combinations of d and $\mathscr{P}$.

2. Definition of the functions $e(h, k, \delta, x)$. In this section, a sequence of functions, $e(h, k, \delta, x)$, are defined in a fashion slightly more complicated than but similar to Farrell (1972). The constructions and properties although technically involved are needed to apply Theorem 2.2 in later sections. Let

$$\begin{aligned} e(h, 0, \delta, x) &= h && -\delta < x < 0 \\ &= -h && 0 < x < \delta \\ &= 0 && \text{elsewhere} \end{aligned}$$

and define recursively for $k \geqq 1$

$$\begin{aligned} e(h, k, \delta, x) &= \textstyle\int_{-2k-1\delta}^{x+2k-1\delta} e(h, k-1, \delta, t)\, dt && \text{for} \quad x \leqq 0 \\ &= -e(h, k, \delta, -x) && \text{for} \quad x > 0\,. \end{aligned}$$

Graphs of $e(h, k, \delta, x)$ for $k = 1, 2$, and 3 are shown in Figure 2.1. Some of

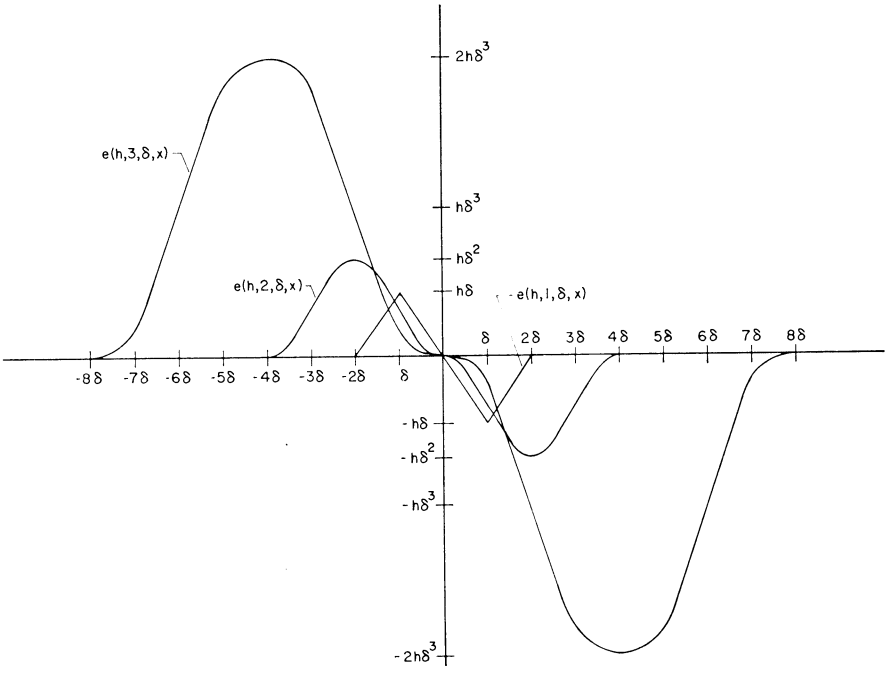

FIG. 2.1.

the basic properties of these functions are catalogued here for later use:

(2.1) $e(h, k, \delta, x)$ is odd

(2.2) $e(h, k, \delta, x)$ is zero on the complement of $(-2^k\delta, 2^k\delta)$

(2.3) $|e(h, k, \delta, x)| \leqq h\delta^k 2^{(k-1)(k-2)/2} = e(h, k, \delta, -2^{k-1}\delta)$ for $k \geqq 1$

(2.4) $e^{(i)}(h, k, \delta, x)$ exists everywhere if $1 \leqq i \leqq k-1$, ($e^{(i)}$ denotes as usual the ith derivative) and $|e^{(i)}(h, k, \delta, x)| \leqq h\delta^{k-i}2^{(k-1-i)(k-2-i)/2} = e^{(i)}(h, k, \delta, x_0)$, some x_0, for $1 \leqq i \leqq k-1$

(2.5) $e^{(k)}(h, k, \delta, x)$ exists except at a finite number of points, $|e^{(k)}(h, k, \delta, x)| \leqq h$, and there exists a point x_0 such that $e^{(k)}(h, k, \delta, x_0) = h$

(2.6) $\int_{-\delta \cdot 2^k}^{0} e(h, k, \delta, x)\, dx = h\delta^{k+1}2^{k(k-1)/2}$

(2.7) the first $k-1$ derivatives of $e(h, k, \delta, x)$ are zero at $2^k\delta$, $-2^{k-1}\delta$, 0, $2^{k-1}\delta$, and $2^k\delta$

(2.8) $e^{k-1}(h, k, \delta, x)$ is a sum of translated replicas of $e(h, 1, \delta, x)$.

3. The class $C_{k\alpha}(\varepsilon, y)$. Define $C_{k\alpha}(\varepsilon, y)$ as the class of density functions f satisfying:

(a) for $k = 1$, f in $C_{k\alpha}(\varepsilon, y)$ iff $f(x) \leqq \alpha$ and $|f(x) - f(u)| \leqq \alpha \cdot |x - u|$ for all x, u in $(y - \varepsilon, y + \varepsilon)$;

(b) for $k > 1$, f in $C_{k\alpha}(\varepsilon, y)$ iff $f(x)$ and its first $k-1$ derivatives exist and are bounded by α in $(y - \varepsilon, y + \varepsilon)$ and $|f^{(k-1)}(x) - f^{(k-1)}(u)| \leqq \alpha \cdot |x - u|$ for all x, u in $(y - \varepsilon, y + \varepsilon)$.

Also define a function f for $k \geqq 1$ as

$$\begin{aligned} f(x) &= e(l, k, \sigma, x - y + 3\sigma 2^{k-1}) && x < y - 2^{k+1}\sigma \\ &= \beta && y - 2^{k+1}\sigma \leqq x \leqq y \\ &= f(2y - x) && x > y\,. \end{aligned}$$

Note that f is symmetric about y and vanishes outside $(y - 5 \cdot 2^{k+1}\sigma, y + 5 \cdot 2^{k-1}\sigma)$. Now define

$$g(x) = f(x) + e(h, k, \delta, x - y - 2^{k-1}\delta) \qquad \text{where} \quad \delta \leqq \sigma\,.$$

In order that f and g be densities,

(3.1) $$2^{k+2}\sigma\beta + l\sigma^{k+1}2^{k(k-1)/2} = 1$$

by property (2.6) and $\int_R f = 1$;

(3.2) $$\beta \geqq h\delta^k 2^{(k-1)(k-2)/2}$$

by property (2.3) and $g \geqq 0$. Note $\int_R g = 1$ follows from (3.1) and the fact that e is odd.

Further if we wish f and g to be in $C_{k\alpha}(\varepsilon, y)$, we must have

$$\beta = l\sigma^k 2^{(k-1)(k-2)/2} \tag{3.3}$$

by (2.3) and $f(y - 2^{k+1}\sigma-) = f(y - 2^{k+1}\sigma+)$,

$$h \leqq \alpha\,, \qquad l \leqq \alpha \tag{3.4}$$

by (2.8) and the Lipschitz condition, and

$$\begin{aligned} h\delta^{k-i}2^{(k-1-i)(k-2-i)/2} &\leqq \alpha \\ l\sigma^{k-i}2^{(k-1-i)(k-2-i)/2} &\leqq \alpha \qquad \text{for} \quad 0 \leqq i \leqq k-1 \end{aligned} \tag{3.5}$$

by (2.4) and the bound on derivatives.

Note by property (2.7), (2.4) and (3.3) above, the first $k - 1$ derivatives of f and g exist everywhere. Graphs of f and g when $k = 2$ are shown in Figure 3.1.

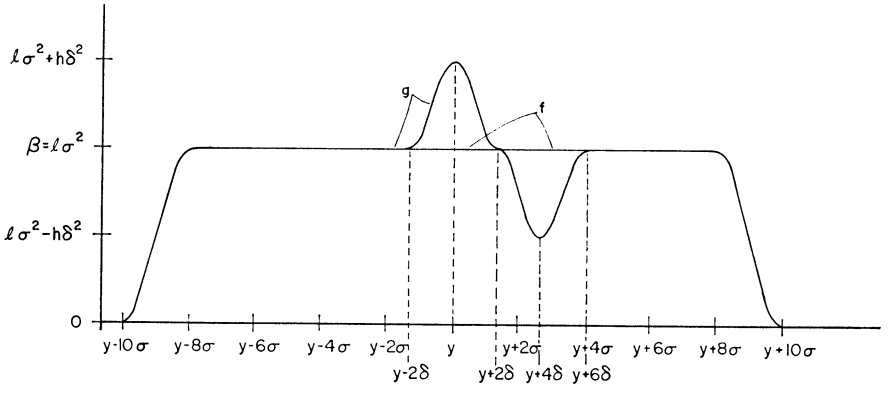

FIG. 3.1. Graph of f and $g = f + e$ when $k = 2$.

4. Theorem 2.2 applied to $d = |\hat{t} - f(y)|$. Consider the distribution of n i.i.d. random variables with density in $C_{k\alpha}(\varepsilon, y)$, and let $\mathscr{P}$ be the class of all such distributions for all densities in $C_{k\alpha}(\varepsilon, y)$. In order to apply Theorem 2.2 in Meyer (1976), an element in $H_2(b)$ must be found. Consider $(\prod_{i=1}^n g(x_i), \prod_{i=1}^n f(x_i))$ where g and f are defined in Section 3. If this pair is to be in $H_2(b)$, f and g must first be in $C_{k\alpha}(\varepsilon, y)$, so (3.1) through (3.5) must hold, and in addition

$$[\int_R (g/f) f\,dx]^n \leqq (1 - b)^2/b\,. \tag{4.1}$$

Since $e(h, k, \delta, x)$ is odd, (4.1) becomes

$$\left[1 + \int_U \frac{e^2(h, k, \delta, t)}{f(t)}\,dt\right]^n \leqq (1 - b)^2/b \tag{4.2}$$

where U is $(y - 2^{k-1}\delta, y + 3 \cdot 2^{k-1}\delta)$. (4.2) is implied by

$$\beta^{-1} \int_R e^2(h, k, \delta, t)\,dt \leqq \frac{\ln((1 - b)^2/b)}{n} \tag{4.3}$$

which by (2.3) and (2.6) is implied by

$$2\beta^{-1}h^2 2^{(k-1)(k-1)}\delta^{2k+1} \leqq \frac{\ln((1-b)^2/b)}{n}. \tag{4.4}$$

Thus if $d = |\hat{t} - f(y)|$, y fixed, f in $C_{k\alpha}(\varepsilon, y)$ by Theorem 2.2 and property (2.3),

$$2\lambda_n \geqq |f(y) - g(y)| = h\delta^k 2^{(k-1)(k-2)/2} \tag{4.5}$$

where δ and h are any numbers subject only to constraints (3.1) through (3.5) and (4.4).

If δ is taken proportional to $n^{-1/(2k+1)}$, l is taken proportional to σ^{-k}, and h, β, and σ are constants, the R.H.S. of (4.5) is maximized and all constraints are satisfied. The rate achieved is thus $O(n^{-k/(2k+1)})$. Since λ_n for the kernel estimate is proportional to $n^{-k/(2k+1)}$ by a slight variation of Parzen's (1962) analysis, this estimate achieves the best possible rate.

5. Different pseudometrics for $C_{k\alpha}(\varepsilon, y)$; extensions to $C_{k\alpha}(\infty, 0)$. As explained, the above argument essentially is in Farrell (1972). (Actually Farrell proved that $\liminf_n \lambda_n \cdot n^{k/(2k+1)} > 0$, without allowing calculation of a lower bound for fixed finite n.) By Theorem 2.2, though, the constructions in Sections 2 and 3 can provide many similar new results. For example, if the class $C_{k\alpha}(\infty, 0)$ and the pseudometric $\sup_x |\hat{t}(x) - f(x)|$ are considered, the above argument unchanged gives a best rate of $O(n^{-k/(2k+1)})$ again. Schuster (1969) has proved that the rate for the kernel estimate with the "sup norm" is $o(n^{k/(2k+2)})$, in excellent agreement.

Or consider $C_{k\alpha}(\varepsilon, y)$ again and the pseudometric $|f^{(i)}(y) - \hat{t}|$ for fixed y and $1 \leqq i \leqq k-1$ ($f^{(i)}$ again denotes the ith derivative of f at y). By translating the function $e(h, k, \delta, x)$ which is added to f to make g, (4.5) becomes $2\lambda_n \geqq h\delta^{k-1}2^{(k-i-1)(k-i-2)/2}$ subject to (3.1) through (3.5) and (4.4). Making δ, σ, h, β, and l the same as before yields an $O(n^{-(k-i)/(2k+1)})$ rate. It is relatively easy to show this is precisely the rate achieved by the ith derivative of the kernel estimate.

Using instead $\sup_x |\hat{t}(x) - f^{(i)}(x)|$ as the pseudometric and $C_{k\alpha}(\infty, 0)$ as the class, the $O(n^{-(k-i)/(2k+1)})$ rate still obtains by identical arguments. However, in this problem, only the case $i = k-1$ was considered in Schuster (1969) where it is proved that the kernel estimate has a λ_n which is $o(n^{-1/(2k+2)})$, in good agreement again. No one, to this author's knowledge, has considered the cases $i < k-1$ for the "sup norm".

6. Algebraic classes and integrated square error. Another type of global loss function besides the "sup norm" is integrated square error (ISE), examined by Watson and Leadbetter (1963):

$$[\int (\hat{t}(x) - f(x))^2\, dx]^{\frac{1}{2}}. \tag{6.1}$$

They proved that $\lambda_n \sim n^{-\frac{1}{2}+1/4k}$ for a kernel estimate when distributions are

restricted to $C_{k\alpha}$, an "algebraic class of degree k." That is, $C_{k\alpha}$ is the class of all probability measures with characteristic functions $\chi(t)$ which satisfy

$$\lim_{|t|\to\infty} |t|^k|\chi(t)| \leqq \alpha . \tag{6.2}$$

To prove this rate is in fact the fastest possible rate, the same functions f and g are used (the dependency on β, δ, h, σ, and l is suppressed as before). Conditions (3.1) and (3.2) are imposed to make f and g densities again. Integrating f and g by parts shows that only condition (3.3) need be imposed to insure the limit in (6.2) is less than or equal to α. (In fact, it is 0.) Thus (3.1) through (3.3) insure f and g are in $C_{k\alpha}$.

Theorem 2.2 states

$$2\lambda_n \geqq [\textstyle\int_R e^2(h, k, \delta, x)\, dx]^{\frac{1}{2}} \tag{6.3}$$

where f and g obey (4.4) in addition to (3.1) through (3.3). Since the R.H.S. is proportional to $h\delta^{k+\frac{1}{2}}$, it might appear from (4.4) that the fastest rate is $O(n^{-\frac{1}{2}})$. To see this is not the case, choose

β proportional to $n^{1/2k}$
δ proportional to $n^{-1/2k}$
h proportional to $n^{1/2k}$
l proportional to $n^{\frac{1}{2}+1\ 2k}$
σ proportional to $n^{-1/2k}$.

Then $h\delta^{k+\frac{1}{2}}$ becomes proportional to $n^{-\frac{1}{2}+1.4k}$, the kernel estimate rate.

7. Other classes of distributions and other pseudometrics. Other classes of distributions and pseudometrics besides those mentioned have been studied by various authors (e.g., Leadbetter and Watson (1963) studied other forms of asymptotic behavior of the characteristic function besides algebraic for ISE loss; Schwartz (1967) and Wahba (1971) studied various combinations of square integrability conditions on derivatives of the density while investigating estimates that were not of the kernel type). Each new variation can be similarly analyzed using the general theorems in Meyer (1976) without proving special cases of these theorems for every problem. The cases in this paper were selected as examples only because they have been extensively studied in the past.

Acknowledgments. The author wishes to thank Dr. Robert H. Berk and Dr. Lawrence D. Brown for their technical assistance, guidance and encouragement. This paper is a portion of the author's dissertation submitted in partial fulfillment of the requirements for the Ph. D. degree at Rutgers, written under the supervision of Professor Robert H. Berk.

REFERENCES

FARRELL, R. H. (1972). On the best obtainable asymptotic rates of convergence in estimation of a density at a point. *Ann. Math. Statist.* **43** 170–180.

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

MEYER, T. G. (1976). On fixed or scaled radii confidence sets: the fixed sample size case. *Ann. Statist.* **5** 65–78.

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

ROSENBLATT, M. (1956). Remarks on some non-parametric estimations of a density function. *Ann. Math. Statist.* **27** 832–837.

SCHUSTER, E. F. (1969). Estimation of a probability density function and its derivatives. *Ann. Math. Statist.* **49** 1187–1195.

SCHWARTZ, S. C. (1967). Estimation of probability density by an orthogonal series. *Ann. Math. Statist.* **47** 1261–1265.

WAHBA, G. (1971). A polynomial algorithm for density estimation. *Ann. Math. Statist.* **42** 1870–1886.

WAHBA, G. (1975). Optimal convergence properties of variable knot, kernel, and orthogonal series methods for density estimation. *Ann. Statist.* **3** 15–30.

INSTITUTE OF STATISTICS
TEXAS A AND M UNIVERSITY
COLLEGE STATION, TEXAS 77843