ESTIMATION OF PROBABILITY DENSITY BY AN ORTHOGONAL SERIES[1]

By Stuart C. Schwartz

Princeton University

1. Introduction and summary. Let $X_1 \cdots X_n$ represent a sequence of independent random variables with a common (unknown) density function $f(x)$. In this paper, an estimate of $f(x)$ of the form $\hat{f}_n(x) = \sum_{j=0}^{q(n)} \hat{a}_{jn}\varphi_j(x)$ is considered, where $\hat{a}_{jn} = (1/n)\sum_{i=1}^{n} \varphi_j(X_i)$, $\varphi_j(\ )$ is the jth Hermite function and $q(n)$ is an integer dependent on n. Assuming $f(x)$ is L_2, it is shown that the sequence of estimates is consistent in the sense of mean integrated square error, $\lim_{n\to\infty} E \cdot \int (f(x) - \hat{f}_n(x))^2\, dx = 0$ and, under additional conditions on $f(x)$, the sequence of estimates is also consistent in mean square error, $\lim_{n\to\infty} E(f(x) - \hat{f}_n(x))^2 = 0$, uniformly in x. For both error criteria, bounds on the rate of convergence of the estimate are obtained. The rate of convergence is seen to depend on the smoothness and integrability properties of $f(x)$—the maximum rate being bounded by $1/n$.

In order for the series method to achieve the same rate of convergence as an estimate which uses the "kernel" technique [4], [6], more assumptions on $f(x)$ are required. However, in estimating a multivariate density, with the same type of conditions as in the univariate case, the rate of convergence remains the same for the multivariate series estimate. With the "kernel" method, the rate depends on the dimension of the density being estimated; the rate of convergence of the estimate decreases as the dimension increases.

In the next section, we introduce notation and give some preliminary results. Conditions for consistency and rates of convergence are established in Section 3. These results are then compared in Section 4 to previous work in the area.

2. Preliminaries and notation. Let

$$H_j(x) = (-1)^j e^{x^2}(d^j/dx^j)(e^{-x^2}) \tag{2.1}$$

be the jth Hermite polynomial. It is known[7] that the normalized Hermite functions

$$\varphi_j(x) = (2^j j!\, \pi^{\frac{1}{2}})^{-\frac{1}{2}} e^{-x^2/2} H_j(x), \qquad j = 0, 1, \cdots, \tag{2.2}$$

are $L_1 \cap L_2$ and form a complete orthonormal set over the real line.

A useful bound for the Hermite functions has been given by Cramér (see[3], p. 208):

$$|\varphi_j(x)| < c_1/\pi^{\frac{1}{4}} = c_2, \tag{2.3}$$

where the constant c_1 (and hence c_2) is independent of x and j.

Received 12 November 1966; revised 10 April 1967.

[1] This paper is based on a part of the author's doctoral dissertation submitted to the University of Michigan (1966). The research was partially supported by NASA Research Grant NSG-2-59.

Throughout this paper, we shall assume that $f(x)$ is square integrable (e.g. it is sufficient that $f(x)$ be bounded). $f(x)$ can then be expanded in the orthogonal series

$$f(x) = \sum_{j=0}^{\infty} a_j \varphi_j(x) \tag{2.4}$$

with the coefficients defined by

$$a_j = \int f(x)\varphi_j(x)\, dx. \tag{2.5}$$

For the discussion of the mean square error, we need

LEMMA 1. *Let $f(x)$ be continuous, of bounded variation, L_1 and L_2 in $(-\infty, \infty)$. Then, the series in (2.4) converges uniformly in any interval interior to $(-\infty, \infty)$.*

The proof can be found in Sansone [5], Section 4.10. An alternate statement is given by Wiener [7].

To specify a rate of convergence for both mean square and mean integrated square error, we need a simple lemma.

LEMMA 2. *Assume the derivative of $f(x)$ exists and that the function $(xf(x) - f'(x))$ is square integrable. Then the coefficients a_j, $j = 1, 2, \cdots$, satisfy the bound*

$$|a_j| < c_3/(2j)^{\frac{1}{4}} \tag{2.6}$$

where c_3 is the L_2 norm of $xf(x) - f'(x)$.

PROOF.[2]

$$a_j = \int_{-\infty}^{+\infty} f(x)e^{-x^2/2}H_j(x)/(2^j j!\,\pi^{\frac{1}{2}})^{\frac{1}{2}}\, dx. \tag{2.7}$$

Use the relationship $(d/dx)H_{j+1}(x) = 2(j+1)H_j(x)$ ([3], p. 193) to substitute for $H_j(x)$ in (2.7) and integrate by parts:

$$a_j = (2j+2)^{-\frac{1}{2}} \int_{-\infty}^{+\infty} [xf(x) - f'(x)]\varphi_{j+1}(x)\, dx. \tag{2.8}$$

Hence, by the Schwarz inequality, a_j is bounded as above.

By repeated application of the method of Lemma 2 we obtain:

LEMMA 3. *Assume that the function*

$$e^{x^2/2}(d^r/dx^r)[e^{-x^2/2}f(x)] = \sum_{i=0}^{r} [r!/i!(r-i)!](-1)^i 2^{-i/2} H_i(x/2^{\frac{1}{2}})(d^{r-i}/dx^{r-i})f(x) \tag{2.9}$$

exists and is square integrable. Then, the coefficients a_j, $j = 1, 2, \cdots$, are bounded by $|a_j| < c_3(r)/(2j)^{r/2}$, where $c_3(r)$ is the L_2 norm of (2.9).

COMMENT. The L_2 assumption on the function in (2.9) can be replaced by an L_1 requirement. This follows from the boundedness of the Hermite functions.

3. Consistency and rates of convergence of the estimate. Let $X_1, X_2, \cdots, X_n$ be independent random variables with a common probability density function $f(x)$. As an estimate of $f(x)$, we form

$$\hat{f}_n(X_1, \cdots, X_n; x) = \hat{f}_n(x) = \sum_{j=0}^{q(n)} \hat{a}_{jn}\varphi_j(x); \tag{3.1}$$

[2] The essential idea of the lemma can be found in Sansone [5], pp. 368-369.

$$\hat{a}_{jn} = (1/n) \sum_{k=1}^{n} \varphi_j(X_k). \tag{3.2}$$

It is easy to see that the $\hat{a}_{jn}$ are unbiased estimates of a_j :

$$E\hat{a}_{jn} = (1/n) \sum_{k=1}^{n} E\varphi_j(X_k) = \int_{-\infty}^{+\infty} f(x)\varphi_j(x)\, dx = a_j . \tag{3.3}$$

The variance of the estimate is bounded by

$$\begin{aligned} E(\hat{a}_{jn} - a_j)^2 &= (1/n^2)E\{\sum_{k=0}^{n} \varphi_j^2(X_k) \\ &\quad + \sum_{k_1,k_2,k_1\neq k_2}^{n} \varphi_j(X_{k_1})\varphi_j(X_{k_2})\} - a_j^2 \\ &\leqq (1/n)(c_2^2 - a_j^2) \leqq c_4/n. \end{aligned} \tag{3.4}$$

The MISE (mean integrated square error) in the estimate of $f(x)$ is:

$$\begin{aligned} E\int (\hat{f}_n(x) - f(x))^2\, dx &= \sum_{j=q+1}^{\infty} a_j^2 + \sum_{j=0}^{q(n)} E(\hat{a}_{jn} - a_j)^2 \\ &\leqq \sum_{j=q(n)+1}^{\infty} a_j^2 + (q(n)/n)c_4 . \end{aligned} \tag{3.5}$$

THEOREM 1. *Assume that $f(x)$ is square integrable and that the sequence of positive integers $q(n)$ is chosen so that $q(n)/n \to 0$ as $q(n) \to \infty$. Under these conditions, the sequence of estimates defined by* (3.1) *and* (3.2) *is consistent in the sense of* MISE. *Furthermore, assume that the function $f(x)$ satisfies the hypotheses of Lemma* 3 *with $r \geqq 2$. Then with $q(n) = O(n^{1/r})$ the* MISE *satisfies*

$$E\int (f(x) - \hat{f}_n(x))^2\, dx = O(1/n^{(r-1)/r}). \tag{3.6}$$

PROOF. The first part of the theorem follows immediately from (3.5). Using Lemma 3, (3.5) is dominated by

$$E\int (f(x) - \hat{f}_n(x))^2\, dx \leqq c_3^2(r) \sum_{j=q+1}^{\infty} 1/(2j)^r + c_4 q(n)/n. \tag{3.7}$$

A convenient integral upper bound for the first term of (3.7) is $1/(2^r(r-1)q^{r-1})$. Upon choosing $q(n)$ as the largest integer less than or equal to $(nc_3^2(r)/c_4)^{1/r}$, the desired result (equation (3.6)) follows.

COMMENT. The referee has pointed out the paper by Čencov [2] who considers the problem of density estimation using other orthogonal systems. Conditions for MISE convergence rates are established ([2], Theorem 1a) and the convergence rate of the present theorem can be obtained from the inequality in Lemma 3 above and Corollary 3 of Čencov.

To investigate the mean square error for fixed x, define

$$f_{q(n)}(x) = \sum_{j=0}^{q(n)} a_j\varphi_j(x). \tag{3.8}$$

The mean square error can be expressed as

$$\begin{aligned} E(f(x) - \hat{f}_n(x))^2 &= (f(x) - f_q(x))^2 + 2(f(x) - f_q(x)) \\ &\quad \cdot E(f_q(x) - \hat{f}_n(x)) + E(f_q(x) - \hat{f}_n(x))^2. \end{aligned} \tag{3.9}$$

$\hat{f}_n(x)$ is an unbiased estimate of $f_q(x) = \sum_{j=1}^{q} a_j\varphi_j(x)$ (see equations (3.1) and (3.3)). Hence, the middle term in (3.9) is zero. We then use the boundedness of the Hermite functions (2.3), (3.4), and the Schwarz inequality to obtain the

bound $|E\{(a_j - \hat{a}_{jn})(a_k - \hat{a}_{kn})\varphi_j(x)\varphi_k(x)\}| \leqq c_2^2 c_4/n$. Using this in the last term of (3.9) gives

$$E(f(x) - \hat{f}_n(x))^2 \leqq (f(x) - f_q(x))^2 + c_2^2 c_4 q^2(n)/n. \tag{3.10}$$

In analogy to the proof of the previous theorem, we obtain the following result concerning the mean square error.

THEOREM 2. *Assume $f(x)$ is continuous, of bounded variation, L_1 and L_2 in $(-\infty, \infty)$. Choose the sequence of positive integers $q(n)$ so that $q^2(n)/n \to 0$ as $q(n) \to \infty$. Then, the sequence of estimates defined by (3.1) and (3.2) converges in mean square, uniformly in x. Furthermore, assume that $f(x)$ satisfies the hypotheses of Lemma 3 with $r \geqq 3$ (cf. Theorem 1). Then with $q(n) = O(n^{1/r})$, the mean square error satisfies $E(f(x) - \hat{f}_n(x))^2 = O(1/n^{(r-2)/r})$.*

PROOF. The first part of the theorem follows from Lemma 1 and (3.10). The second part follows from the bound

$$\begin{aligned} |f(x) - f_q(x)| &\leqq \textstyle\sum_{j=q+1}^{\infty} |a_j \varphi_j(x)| \leqq c_2 \sum_{j=q+1}^{\infty} |a_j| \\ &\leqq c_2 c_3(r)/(2^{r/2}(r/2 - 1)q^{r/2-1}) \end{aligned}$$

and upon choosing $q(n)$ as the largest integer less than or equal to $((r - 2)\, c_3^2 n)^{1/r}/(2^r(r/2 - 1)^2 c_4)^{1/r}$.

4. Discussion. A previously proposed method of estimating $f(x)$ is the estimate

$$\hat{f}_n(x) = 1/(nh(n)) \textstyle\sum_{i=1}^{n} K_n((x - X_i)/h(n)). \tag{4.1}$$

$K_n(x)$ is the kernel or weighting function and $h(n)$ is a sequence of positive numbers chosen so that $h(n) \to 0$ as $n \to \infty$.

Watson and Leadbetter [6] have considered the MISE of this estimate. They show that the MISE is a minimum if the Fourier transform of the kernel $K_n(x)$, which we denote by $\Psi_{K_n}(t)$ is

$$\Psi_{K_n}(t) = n\,|\Psi_f(t)|^2/(1 + (n - 1)|\Psi_f(t)|^2). \tag{4.2}$$

$\Psi_f(t)$ is the characteristic function of $f(x)$ which is assumed to be L_2. By specifying the asymptotic behavior of $\Psi_f(t)$, they are able to calculate the maximum rate of convergence. A typical result is: if $\Psi_f(t)$ satisfies $\lim_{|t|\to\infty} |t|^{r/2} |\Psi_f(t)| = c$, with $r > 1$, the MISE of the estimate is of order $O(1/n^{(r-1)/r})$. Hence, if the kth derivative of $f(x)$ exists and is absolutely integrable, the above condition is satisfied with $r = 2k$.

In order to guarantee the same rate of convergence with the series method, we need to require that the function (2.9) be integrable. For this it is sufficient that the functions $x^i(d^{r-i}/dx^{r-i})f(x)$, $i = 0, 1, \cdots, r$, be integrable. Consequently, Watson and Leadbetter require less restrictive conditions. This is at the expense of a more complicated estimator—the inverse transform of (4.2) as compared to (3.1)–(3.2).

Parzen [4] has considered an estimator of the form (4.1) with the kernel's functional form independent of n. With $\int K(x)\, dx = 1$ and $h(n)$ satisfying $\lim_{n\to\infty} nh(n) = 0$, he has shown that $\hat{f}_n(x)$ is a consistent estimate (in mean-

square) at every point of continuity of $f(x)$. Assuming further that the kernel satisfies

$$\int_{-\infty}^{+\infty} x^i K(x)\,dx = 0, \qquad i = 1, 2, \cdots r-1, \tag{4.3}$$

$$\int_{-\infty}^{+\infty} x^r |K(x)|\,dx < \infty,$$

and that $\int_{-\infty}^{+\infty} |t^r \Psi_f(t)|\,dt < \infty$, the bias in the estimate is shown to be $O(h^{2r})$. Since the variance of the estimate is $O(1/nh)$, choosing $h(n)$ proportional to $1/n^{1/(2r+1)}$ gives a mean square error with order of consistency $O(1/n^{2r/(1+2r)})$.

Inspection of the hypotheses of Theorem 2 shows that we require more restrictive conditions on $f(x)$ to achieve the same rate. Note, however, that if $r > 2$, from (4.3) the kernel $K(x)$ cannot be a non-negative function—the sequence of estimates of the density function $f(x)$, as given by

$$\hat{f}_n(x) = 1/(nh) \sum_{i=1}^{n} K((x - X_i)/h(n)),$$

will be negative over non-degenerate intervals of x for some values of n. Using the series method, the sequence of estimates may or may not take on negative values for some values of n.

It is in the problem of estimating a multivariate density function that the series method may prove to be the most advantageous. With the same type of assumptions as in Theorems 1 and 2, except now extended to the multivariate case, the rate of convergence of the multivariate estimate will remain the same as given above. (The density $f(x_1, x_2, \cdots, x_k)$ is expressed as

$$f(x_1, \cdots, x_k) = \sum_{j_1 \cdots j_k} a_{j_1 \cdots j_k} \varphi_{j_1}(x_1) \cdots \varphi_{j_k}(x_k),$$

and the $a_{j_1 \cdots j_k}$ are estimated in a manner analogous to (3.2).) Using a multidimensional kernel estimator, the rate of convergence will depend on the dimension of the density function $f(x_1 \cdots x_k)$. (See, for example, [1].) For this multivariate case, the variance of the estimate is now of order $O(1/nh^k)$ while the bias is still $O(h^{2r})$. This leads to a rate of convergence of order $O(1/n^{2r/(2r+k)})$—the rate of convergence decreases as the dimension of the density, k, increases.

REFERENCES

[1] Cacoullos, T. (1966). Estimation of a multivariate density. *Ann. Inst. Statist. Math.* **18** 179–190.

[2] Čencov, N. N. (1962). Evaluation of an unknown distribution density from observations. *Soviet Math.* **3** 1559–1562.

[3] Erdèlyi, A., Magnus, W., Oberhettinger, F. and Tricomi, F. G. (1953). *Higher Transcendental Functions.* **2** Bateman Manuscript Project, McGraw-Hill, New York.

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

[5] Sansone, G. (1959). *Orthogonal Functions.* Interscience, New York. (Translated from the Italian by A. H. Diamond.)

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

[7] Wiener, N. (1933). *The Fourier Integral and Certain of its Applications.* Dover, New York.