The Annals of Statistics
1980, Vol. 8, No. 2, 347–376

AN ADAPTIVE ORTHOGONAL-SERIES ESTIMATOR FOR PROBABILITY DENSITY FUNCTIONS[1]

BY G. LEIGH ANDERSON[2] AND RUI J. P. DE FIGUEIREDO

Rice University

Given a sample set $X_1, \cdots, X_N$ of independent identically distributed real-valued random variables, each with the unknown probability density function $f(\cdot)$, the problem considered is to estimate f from the sample set. The function f is assumed to be in $L_2(a, b)$; f is not assumed to be in any parametric family. This paper constructs an adaptive "two-pass" solution to the problem: in a preprocessing step (the first pass), a preliminary rough estimate of f is obtained by means of a standard orthogonal-series estimator. In the second pass, the preliminary estimate is used to transform the orthogonal series. The new, transformed orthogonal series is then used to obtain the final estimate. The paper establishes consistency of the estimator and derives asymptotic (large sample set) estimates of the bias and variance. It is shown that the adaptive estimator offers reduced bias (better resolution) in comparison to the conventional orthogonal series estimator. Computer simulations are presented which demonstrate the small sample set behavior. A case study of a bimodal density confirms the theoretical conclusions.

1. Introduction

A. Background. A real random variable (rv) X is characterized by the associated cumulative distribution function (cdf)

$$F(x) \equiv \Pr\{X \leqslant x\}. \tag{1}$$

If the measure induced on $\mathbb{R}$ by F is absolutely continuous with respect to Lebesgue measure, then we may define the probability density function (pdf) $f(\cdot)$ as

$$f(x) \equiv \frac{d}{dx} F(x) \tag{2}$$

the Radon-Nykodym derivative of F.

In the present paper, we wish to estimate $f(\cdot)$ from a given sample $\{X_1, \cdots, X_N\}$ of X, without assuming that f belongs to a specified parametric family. This task falls in the category of "nonparametric" estimation. Several techniques of nonparametric estimation have been proposed by a number of researchers. These will be reviewed below.

The current work is concerned with a modification to one of these techniques, namely the orthogonal-series estimator. We propose a prior transformation of the orthogonal series which "tunes" the series to the given sample set. The effect of the

Received March 1978; revised September 1978.

[1] This work was supported by AFOSR grant 75-2777 and NSF grant ENG 74-17955, and ONR contract N000 14-79-C-0442.

[2] Now with EXXON Production Research, Houston, Texas.

AMS 1970 *subject classifications.* G2G05, 42A36, 93E10.

Key words and phrases. Probability density functions, adaptive estimation, orthogonal series, nonparametric estimation, Fourier methods.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Statistics.* www.jstor.org

transformation is to reduce the bias of the estimator for a sample set of a given size N. The transformation is obtained from a preprocessing step wherein we examine the sample set before applying the estimator.

One of the earliest and most widely studied nonparametric density function estimators was introduced by M. Rosenblatt [13] in 1955. He proposed the kernel-type estimator

$$f(x) = \frac{1}{hN}\Sigma_{j=1}^{N} K\left(\frac{x - X_j}{h}\right)$$

where $K(\cdot)$ is a given kernel function and $h = h(N)$ is a scaling factor depending on the sample size N. The estimator was further studied by E. Parzen [12] in 1961. G. S. Watson and M. R. Leadbetter [23] investigated optimal choices for the kernel shape $K(\cdot)$. A particular kernel shape offering attractive theoretical and practical properties was obtained by J. O. Bennett, R. J. P. de Figueiredo, and J. R. Thompson [2] with the use of B-splines. K. B. Davis [6] studied a kernel which is not L_1 and demonstrated superior asymptotic properties; numerical trials with small sample sizes show poor performance, however [17]. Convergence conditions for kernel estimators [19] and related nearest neighbor estimators [8] were studied by L. P. Devroye and T. J. Wagner.

Another type of estimator, using an orthogonal series expansion, was introduced by R. Kronmal and M. Tarter [10], Cencov [5], van Ryzin [18], and Schwartz [15]; they developed error estimates and optimal series approximations. The optimal results require knowledge of the unknown density f. H. D. Brunk [4] considered ways of extracting the needed knowledge from the sample itself.

A totally different approach was taken by G. F. de Montricher, R. A. Tapia, and J. R. Thompson [7]. In this theoretical paper the density estimate is the one which maximizes a penalized likelihood function. A discretized numerical implementation by D. Scott [16], gave excellent small-sample performance. An earlier effort along these lines is that of I. J. Good and R. A. Gaskins [9].

A. Wragg and D. C. Dowson [25] use the information-theoretic concept of entropy to fit density functions to a truncated moment sequence. Grace Wahba [20] and P. Whittle [24] employ notions from stochastic processes to obtain "optimally-smoothed" density estimates, Wahba's [20] result being, in addition, data adaptive.

B. Summary of results. In Section 2 we take a close look at the orthogonal series-type estimator, and develop asymptotic error analysis for the special case of the Fourier series estimator. In Section 3 we introduce a new data-adaptive modification of the Fourier series estimator. The series is modified with a transformation derived from a preprocessing step. The modified series reduces the bias of the estimator for a sample set of given size N. We develop the asymptotic error analysis of the estimator and produce consistency results. Finally, in Section 4 we examine some computer simulations to study the behavior of the estimator on small sample sets.

C. Notation and conventions. Throughout this paper we will assume the following notation and conventions.

(1) X is a real-valued random variable with probability density function (pdf) $f(\cdot)$.
(2) We are given a sample set of size $N\{X_1, X_2, \cdots, X_N\}$ where each X_k is an independent realization of X.
(3) The expected value of X is denoted by $E[X]$ and the square of $E[X]$ by $(E[X])^2$. The notation $E[X]^2$ is the same as $E[X^2]$.
(4) The asterisk z^* denotes complex conjugate.

2. Series-type estimators

A. Preliminary Considerations. Consider a (Lebesgue) integrable function g defined on the interval (a, b). Let g satisfy $g(x) > 0$ almost everywhere for x in (a, b) and $\int_a^b g(x)\,dx = 1$. We can define $L_2(g)$, the class of square-integrable functions weighted by g.

$$L_2(g) = \{s : (a, b) \to \mathbb{R} | \int_a^b s(x)^2 g(x)\,dx < \infty\}. \tag{1}$$

Furthermore, let there be given $\{u_k(\cdot)\}_{k=0}^{\infty}$, a complete orthonormal family in $L_2(g)$.

Suppose that $f(\cdot)$, the pdf of the random variable X, is such that f/g is in $L_2(g)$. Then f may be expanded as

$$f(x) = g(x)\Sigma_{k=0}^{\infty} b_k u_k(x). \tag{2}$$

By orthogonality, we can see

$$\begin{aligned} E[u_j(X)] &= \int_a^b u_j(x) f(x)\,dx \\ &= \int_a^b u_j(x) g(x) \Sigma_{k=0}^{\infty} b_k u_k(x)\,dx \\ &= b_j. \end{aligned}$$

Now an estimator for b_k is

$$\hat{b}_k \equiv \frac{1}{N}\Sigma_{j=1}^{N} u_k(X_j). \tag{3}$$

Thus we can construct an estimate of f by

$$\hat{f}(x) \equiv g(x)\Sigma_{k=0}^{n} \hat{b}_k u_k(x) \tag{4}$$

for some $n < N$.

It is easy to derive error expressions for (4) in terms of the coefficients in the expansion (2). A convenient error measure is

$$\begin{aligned} \int_a^b \frac{E[\hat{f}(x) - f(x)]^2}{g(x)}\,dx &= \int_a^b E\left[\frac{\hat{f}(x) - f(x)}{g(x)}\right]^2 g(x)\,dx \\ &= E\int_a^b\left[\Sigma_{k=0}^{n}(\hat{b}_k - b_k)u_k(x) - \Sigma_{k=n+1}^{\infty} b_k u_k(x)\right]^2 g(x)\,dx \qquad (5) \\ &= E\left\{\Sigma_{k=0}^{n}(\hat{b}_k - b_k)^2 + \Sigma_{k=n+1}^{\infty} b_k^2\right\}. \end{aligned}$$

This last expression is just

$$\Sigma_{k=0}^{n} \operatorname{Var}\frac{[u_k(X)]}{N} + \Sigma_{k=n+1}^{\infty} b_k^2. \tag{6}$$

In (6) the first term is the variance term and the second term is the bias term

A desirable property of any estimator is asymptotic consistency, which, loosely speaking, means that as the size of the sample set increases, the error decreases. To sharpen this notion, we define several types of asymptotic consistency.

(7) *Definition.* Let $\hat{f}_N$ be an estimator for f given a sample set of size N. Let x_0 be in (a, b).

If $E[\hat{f}_N(x_0) - f(x_0)]^2 \to 0$ as $N \to \infty$ then $\hat{f}_N$ is "asymptotically consistent in the mean square sense at x_0."

If $\int_a^b E[\hat{f}_N(x) - f(x)]^2\,dx \to 0$ as $N \to \infty$ then $\hat{f}_N$ is "asymptotically consistent in the integrated mean square sense."

If for every $c > 0$ there is an N_c such that for $N > N_c$ we have $P_r\{|\hat{f}_N(x_0) - f(x_0)| > c\} < c$, then $\hat{f}_N$ is asymptotically consistent in probability at x_0.

The definition of the estimator (4) is not complete, since we have not specified the choice of n. Let us choose $n = n(N)$ as a function of N in such a way that

$$n(N) \to \infty \quad \text{as} \quad N \to \infty; \tag{8.1}$$

$$\frac{n(N)}{N} \to 0 \quad \text{as} \quad N \to \infty. \tag{8.2}$$

If we assume that there is a uniform bound B such that

$$\operatorname{Var}[u_k(X)] \leqslant B, \qquad k = 0, 1, 2, \cdots,$$

then a simple argument shows that with choice (8), the estimator (4) is asymptotically consistent in the integrated mean square sense. The precise dependence of $n(N)$ is here left deliberately vague. Optimal choices are investigated in [10].

An often-studied extension of (4), first described by Watson [22], is

$$\hat{f}(x) \equiv g(x)\Sigma_{k=0}^{\infty} w_k(h)\hat{b}_k u_k(x) \tag{9}$$

where $\{w_k(\cdot)\}_{k=0}^{\infty}$ is a sequence of weights parameterized by a positive parameter h. We choose the weights so that

$$w_k(h) \to 0 \quad \text{as} \quad k \to \infty; \tag{10.1}$$

$$w_k(h) \to 1 \quad \text{as} \quad h \to 0. \tag{10.2}$$

Optimal choices of the weight sequence $\{w_k(h)\}_{k=0}^{\infty}$ have been studied in [4]. Briefly, the optimal functional form of $w_k(\cdot)$ depends on f, and the choice $h = h(N)$ depends on the sample set size.

B. Fourier series estimators. The Fourier series estimator, a special case of (2.A.4), has been studied extensively by Kronmal and Tarter [10]. They were interested primarily in *integrated* mean square error and optimal truncation point n

for the estimator. We shall be concerned here and later with the *pointwise* mean square error, $E[\hat{f}(x_0) - f(x_0)]^2$. The following development in this section is new, although it follows somewhat in the spirit of [13] and [12].

From now on we will assume that f takes its support on a finite interval $[a, b]$. The error introduced by this assumption is small in comparison to the bias and variance components to be analyzed later. Furthermore, we will take $a = 0$, $b = 1$. This is done for technical convenience, since a simple linear scaling and translation will return us to the general case $[a, b]$.

Let $\{w_k(\cdot)\}_{k=-\infty}^{\infty}$ be a sequence of (complex) functions of a real positive variable h. Consider the estimator given by

$$(1.1) \qquad \hat{f}(x) = \Sigma_{k=-\infty}^{\infty} w_k(h)\hat{b}_k \exp(2\pi ikx),$$

$$(1.2) \qquad \hat{b}_k \equiv \frac{1}{N}\Sigma_{j=1}^{N} \exp(-2\pi ikX_j).$$

We are interested in the behavior of this estimator for large N. In particular, we will derive asymptotic estimates of $\mathrm{Var}[\hat{f}(x_0)]$ and bias $[\hat{f}(x_0)]$ for $x_0 \in [0, 1]$.

It is clear that the behavior of $\hat{f}$ depends greatly on the choice of $\{w_k(\cdot)\}_{k=-\infty}^{\infty}$ and of h. We will now take a digression to study some properties of $\{w_k(\cdot)\}_{k=-\infty}^{\infty}$ which we will then use to answer questions about $\hat{f}$.

(2) LEMMA. *Let $\{w_k(\cdot)\}_{k=-\infty}^{\infty}$ be a weight sequence. Suppose for each $h > 0$ $\Sigma_{k=-\infty}^{\infty}|w_k(h)|^2 < \infty$ and for each k, $w_k(h) = w_{-k}(h)^*$. Then the kernel K_h defined by*

$$(2.1) \qquad K_h(x) \equiv \Sigma_{k=-\infty}^{\infty} w_k(h) \exp(2\pi ikx)$$

is a real periodic function in $L_2[0, 1]$ with period 1. *Moreover, the estimator* (1) *may be written as*

$$(2.2) \qquad \hat{f}(x) = \frac{1}{N}\Sigma_{j=1}^{N} K_h(x - X_j).$$

PROOF. Straightforward and hence omitted.

Expression (2.2) has a form similar to that of the Parzen kernel estimator (see [12]). However, in the present case $K_h(\cdot)$ is a periodic kernel and does not depend on h as a simple scale factor. The dependence on h is more complicated, and this dependence must be conditioned for the estimator to behave properly.

Henceforth we will assume that the weight sequence satisfies the following:

(3) *Conditions.*

(3.1) $\{w_k(\cdot)\}_{k=-\infty}^{\infty}$ satisfies the hypotheses of Lemma (2). Moreover, $K_h(x) \equiv \Sigma_{k=-\infty}^{\infty} w_k(h) \exp(2\pi ikx)$ satisfies

$$(3.2) \qquad K_h(x) \geqslant 0;$$

$$(3.3) \qquad K_h(-x) = K_h(x);$$

(3.4) $$\int_{-\frac{1}{2}}^{\frac{1}{2}} K_h(x)\,dx = 1;$$

(3.5) $K_h(x)$ is pointwise continuous in $h > 0$ and x;

(3.6) $$\int_{-\frac{1}{2}}^{\frac{1}{2}} K_h(x)x^2\,dx \to 0 \text{ as } h \to 0;$$

(3.7) Let $\frac{1}{2} > \varepsilon > 0$. Then

$$\frac{\int_{\varepsilon}^{\frac{1}{2}} K_h(x)x^2\,dx}{\int_{0}^{\frac{1}{2}} K_h(x)x^2\,dx} \to 0 \quad \text{as} \quad h \to 0;$$

(3.8) Let $\frac{1}{2} > \varepsilon > 0$. Then there exists $B_\varepsilon > 0$ such that $\int_{\varepsilon}^{\frac{1}{2}} K_h(x)^2\,dx < B_\varepsilon$ as $h \to 0$.

Conditions (3.2)–(3.6) have ready analogues in terms of the weight sequence as follows:

(3′) *Conditions*

(3.2′) For any square-summable sequence $\{C_k\}$,

$$\Sigma_{k,l} w_{k-l} C_k C_l^* \geqslant 0;$$

(3.3′) $w_k = w_{-k}$;

(3.4′) $w_0(h) = 1$ for all $h > 0$;

(3.5′) $\{k w_k(h)\}$ is a square-summable sequence;

(3.6′) $w_k(h) \to 1$ as $h \to 0$.

Conditions (3.7) and (3.8) are not so easily expressed in terms of the weights. However, (3.7) and (3.8) are used in the sequel *only* to establish (4.2) and (4.3) below.

Through term-by-term integration we can show that (3.7′) and (3.8′) are equivalent to (4.2) and (4.3):

(3.7′) $$\frac{\Sigma_{k\neq 0}(-1)^k \frac{3}{2}\frac{w_k(h)}{(2\pi k)^2} + \Sigma_{k\ \text{odd}}\frac{24}{(2\pi k)^4} w_k(h) + \frac{1}{32}}{\Sigma_{k\neq 0}\frac{(-1)^k}{2\pi^2 k^2} w_k(h) + \frac{1}{12}}$$

goes to zero as $h \to 0$;

(3.8′) $$\frac{\Sigma_k\left(\Sigma_{j\neq 0} w_{k-j}(h)\frac{(-1)^j}{2\pi j}\right)^2}{\Sigma_k w_k^2} \to 0 \quad \text{as} \quad h \to 0.$$

Evidently conditions (3.7′) and (3.8′) are difficult to verify for a given sequence $\{w_k(\cdot)\}$. If a closed-form expression for the kernel $K_h(x)$ can be found, (3.7) and (3.8) are much easier to check.

As an example, consider the weight sequence $w_k(h) \equiv 1/(1 + (hk)^2)$. Using the fact that

$$\Sigma_{k=1}^{\infty} \frac{\cos kx}{k^2 + a^2} = \frac{\pi}{2a} \frac{\cosh a(\pi - x)}{\sinh a\pi} - \frac{1}{2a^2}$$

we get

$$K_h(x) = \Sigma_k \frac{1}{1 + (hk)^2} e^{2\pi ikx} = \frac{\pi}{h} \frac{\cosh \pi(1 - 2|x|)/h}{\sinh \pi/h}.$$

Conditions (3.1)–(3.6) are readily verifiable. For (3.7), we obtain, after some manipulation,

$$\frac{\int_{\varepsilon}^{\frac{1}{2}} K_h(x)x^2\,dx}{\int_{0}^{\frac{1}{2}} K_h(x)x^2\,dx} = \frac{\left(\sinh\frac{\pi}{h}(1 - 2\varepsilon)\right)\left(1 + \frac{h}{4\pi^3}\right) + \frac{h}{4\pi^2}\left(2\varepsilon \cosh\frac{\pi}{h}(1 - 2\varepsilon) - 1\right)}{\frac{h^2}{4\pi^3}\sinh\frac{\pi}{h} - \frac{h}{\pi^2}}.$$

From the fact that $x^p \sinh rx/\sinh x \to 0$ as $x \to \infty$ for $0 < r < 1$, we see that the above expression goes to zero as $h \to 0$. The same fact shows that condition (3.8) is satisfied.

Under assumptions (3) (or (3′)) it is possible to establish the following limits which will arise shortly in the asymptotic error analysis. The proof is a consequence of a straightforward though lengthy analysis and is omitted. (The omitted proofs may be found in [1].)

(4) Lemma. *Under the assumptions of conditions* (3), *we have*

(4.1) $$\int_{-\frac{1}{2}}^{\frac{1}{2}} K_h(x)^2\,dx \to \infty \quad \textit{as} \quad h \to 0;$$

(4.2) $$\frac{\int_{-\frac{1}{2}}^{\frac{1}{2}} K_h(x)|x|^3\,dx}{\int_{-\frac{1}{2}}^{\frac{1}{2}} K_h(x)x^2\,dx} \to 0 \quad \textit{as} \quad h \to 0;$$

(4.3) $$\frac{\int_{-\frac{1}{2}}^{\frac{1}{2}} K_h(x)^2 x^2\,dx}{\int_{-\frac{1}{2}}^{\frac{1}{2}} K_h(x)^2\,dx} \to 0 \quad \textit{as} \quad h \to 0.$$

Two of the quantities are important enough to merit specific notation which will be used extensively.

(5) DEFINITION. For a kernel $K_h(\cdot)$, let

$$c(h) \equiv \tfrac{1}{2}\textstyle\int_{-\frac{1}{2}}^{\frac{1}{2}} K_h(x) x^2\, dx;$$

$$v(h) \equiv \textstyle\int_{-\frac{1}{2}}^{\frac{1}{2}} K_h(x)^2\, dx.$$

We require one further lemma about these quantities.

(6) LEMMA.

(6.1) *$v(h)$ and $c(h)$ are continuous in $h > 0$.*

(6.2) *For every N sufficiently large, there is an h_N such that*

$$\frac{v(h_N)}{c(h_N)^2} = N.$$

(6.3) *If h_N is chosen by* (6.2), *then*

$$\frac{v(h_N)}{N} + c(h_N)^2 \to 0 \qquad \text{as} \quad N \to \infty.$$

PROOF. The first statement follows from condition (3.5) and the compactness of the interval of integration.

Since $v(h) \to \infty$ and $c(h) \to 0$ as $h \to 0$, it is clear that $v(h)/c(h)^2 \to \infty$ and is a continuous function. Hence (6.2) follows.

With h_N chosen by (6.2),

$$\frac{v(h_N)}{N} + c(h_N)^2 = 2c(h_N)^2 \to 0 \qquad \text{as} \quad h_N \to 0. \tag{7}$$ □

Now we are ready to state the main theorem of this section. Although the proof follows the spirit of Rosenblatt [13], the result is original for Fourier series estimators. Before now, all error estimates for series estimators were of the integral type $\int_0^1 E[\hat{f}(x) - f(x)]^2\, dx$. The following result gives estimates of local type $E[\hat{f}(x_0) - f(x_0)]^2$. It is an important step in the later construction of the modified estimator which adapts to the local properties of f.

To aid in the proof we introduce $\mathring{f}$, the periodic extension of f, defined by

$$\mathring{f}(x + k) = f(x)$$

where $x \in [0, 1]$ and k is an integer.

(8) THEOREM. *Suppose*

(8.1) *$f \in C^3[0, 1]$ and vanishes in a neighborhood of the end points;*

(8.2) *$\hat{f}$ is defined for $x \in [0, 1]$ and $h > 0$ by*

$$\hat{f}(x) \equiv \textstyle\sum_{k=-\infty}^{\infty} w_k(h) \hat{b}_k \exp(2\pi i k x)$$

$$\hat{b}_k \equiv \frac{1}{N} \textstyle\sum_{j=1}^{N} \exp(-2\pi i k X_j);$$

(8.3) *The sequence* $\{w_k(\cdot)\}_{k=-\infty}^{\infty}$ *satisfies conditions* (3).

Then for $x_0 \in [0, 1]$,

$$(8.4) \qquad \lim_{h\to 0}\frac{E[\hat{f}(x_0)] - f(x_0)}{c(h)} = f''(x_0).$$

If, furthermore, we choose $h = h_N$ *as a function of* N *in such a way that* $h_N \to 0$ *as* $N \to \infty$, *then*

$$\lim_{N\to\infty}\frac{N\operatorname{Var}[\hat{f}(x_0)]}{v(h_N)} = f(x_0).$$

PROOF. We can write $\hat{f}(x) = (1/N)\sum_{j=1}^{N} K_h(x - X_j)$ where $K_h(\cdot)$ is the kernel associated with $\{w_k(\cdot)\}_{k=-\infty}^{\infty}$. By independence of the samples,

$$\begin{aligned} E[\hat{f}(x_0)] &= E[K_h(x_0 - X)] \\ &= \int_0^1 K_h(x_0 - y)f(y)\,dy = \int_{-\frac{1}{2}}^{\frac{1}{2}} K_h(y)\mathring{f}(x_0 + y)\,dy, \end{aligned}$$

where $\mathring{f}$ is the periodic extension of f. Since f vanishes in a neighborhood of the end points of [0, 1], $\mathring{f}$ also has three continuous derivatives. Hence we can invoke Taylor's theorem with remainder and expand

$$E[\hat{f}(x_0)] = \int_{-\frac{1}{2}}^{\frac{1}{2}} K_h(y)\left[\mathring{f}(x_0) + \mathring{f}'(x_0)y + \tfrac{1}{2}\mathring{f}''(x_0)y^2 + \frac{y^3}{3!}\mathring{f}'''(z(y))\right]dy$$

where $x_0 < z(y) < y$ or $y < z(y) < x_0$. By conditions (3.3) and (3.4), this reduces to

$$E[\hat{f}(x_0)] = f(x_0) + f''(x_0)c(h) + \int_{-\frac{1}{2}}^{\frac{1}{2}} K_h(y)\frac{y^3}{3!}\mathring{f}'''(z(y))\,dy.$$

Now

$$\begin{aligned} \left| f''(x_0) - \frac{E[\hat{f}(x_0)] - f(x_0)}{c(h)} \right| &= \left| \frac{\int_{-\frac{1}{2}}^{\frac{1}{2}} K_h(y)\frac{y^3}{3!}\mathring{f}'''(z(y))\,dy}{c(h)} \right| \\ &\leqslant \frac{1}{3!}\sup_{x\in[0,1]}|f'''(x)| \cdot \left| \frac{\int_{-\frac{1}{2}}^{\frac{1}{2}} K_h(y)|y|^3\,dy}{c(h)} \right| \end{aligned}$$

and this → 0 as $h \to 0$ by Lemma (4). This establishes (8.4).

Again by independence of the samples,

$$\begin{aligned} \operatorname{Var}[\hat{f}(x_0)] &= \frac{1}{N}\operatorname{Var}[K_h(x_0 - X)] \\ &= \frac{1}{N}\left\{\int_0^1 K_h(x_0 - y)^2 f(y)\,dy - \left(E[\hat{f}(x_0)]\right)^2\right\}. \end{aligned}$$

Using the same extension and expansion, we have (for appropriately redefined $z(y)$)

$$\textstyle\int_0^1 K_h(x_0 - y)^2 f(y)\,dy = \int_{-\frac{1}{2}}^{\frac{1}{2}} K_h(y)^2\left[\mathring{f}(x_0) + \mathring{f}'(x_0)y + \frac{1}{2}\mathring{f}''(z(y))y^2\right] dy$$

$$\textstyle = v(h)f(x_0) + \frac{1}{2}\int_{-\frac{1}{2}}^{\frac{1}{2}} K_h(y)^2 y^2 \mathring{f}''(z(y))\,dy.$$

Thus

$$\left|\frac{N\,\mathrm{Var}[\hat{f}(x_0)]}{v(h)} - f(x_0)\right| = \left|\frac{\frac{1}{2}\int_{-\frac{1}{2}}^{\frac{1}{2}} K_h(y)^2 y^2 \mathring{f}''(z(y))\,dy}{v(h)} - \frac{\left(E[\hat{f}(x_0)]\right)^2}{v(h)}\right|$$

$$\leqslant \tfrac{1}{2}\sup\nolimits_{x\in[0,1]}|f''(x)|\left|\frac{\int_{-\frac{1}{2}}^{\frac{1}{2}} K_h(y)^2 y^2\,dy}{v(h)}\right| + \frac{\left(E[\hat{f}(x_0)]\right)^2}{v(h)}.$$

Now if $h = h_N \to 0$ as $N \to \infty$, then these two terms go to zero by Lemma (4). This completes the proof. □

Thus we have approximately for large N,

$$E[\hat{f}(x_0) - f(x_0)]^2 \approx \frac{f(x_0)}{N}v(h_N) + f''(x_0)^2 c(h_N)^2.$$

An obvious consequence is the following:

(9) Corollary. *Under the hypotheses of Theorem* (8), *suppose we choose* h_N *to solve*

$$\frac{v(h_N)}{N} = c(h_N)^2.$$

Then $\hat{f}$ *is asymptotically consistent in the mean square sense at* x_0. *That is,*

$$E[\hat{f}(x_0) - f(x_0)]^2 \to 0 \qquad \text{as} \quad N \to \infty.$$

Proof. By Lemma (6), $v(h_N)/(N) + c(h_N)^2 \to 0$. Thus, asymptotically,

$$E[\hat{f}(x_0) - f(x_0)]^2 \leqslant \left(f(x_0) + f''(x_0)^2\right)\left(\frac{v(h_N)}{N} + c(h_N)^2\right)$$

also goes to zero. □

3. A data-adaptive estimator.

A. *Motivation.* Recall the simple form of the estimator (2.A.4)

$$\hat{f}(x) \equiv g(x)\textstyle\sum_{k=0}^n \hat{b}_k u_k(x)$$

with the integrated variance

$$\int_a^b \frac{\mathrm{Var}[\hat{f}(x)]}{g(x)}\,dx = \sum_{k=0}^n \frac{\mathrm{Var}[u_k(X)]}{N}$$

and integrated bias squared

$$\int_a^b \frac{(E[\hat{f}(x)] - f(x))^2}{g(x)} dx = \Sigma_{k=n+1}^{\infty} b_k^2.$$

We see that for fixed N and increasing n, the bias decreases but the variance increases. For samples of moderate size (say $N = 100$), we may not take more than a few terms in the series before the variance overwhelms us. Thus we must hope that f may be well approximated by the first few terms in the expansion. Ideally, we would like to choose a family $\{u_k\}_{k=0}^{\infty}$ for which this occurs.

It is impossible to select a fixed family $\{u_k\}_{k=0}^{\infty}$ which works well for all functions f. So let us consider the following adaptive strategy. From the sample set $\{X_1, \cdots, X_N\}$ we will extract certain information about f. We use this information to fashion a family $\{u_k\}_{k=0}^{\infty}$ adapted to f. We will then use this family to obtain an estimate of f.

B. *Construction of the estimator.* Let us consider a way of transforming a given orthogonal family into a new orthogonal family. We start with the Fourier functions $\{\exp(2\pi ikx)\}_{k=-\infty}^{\infty}$ orthonormal on $[0, 1]$. Suppose that we have a transformation G satisfying

(1.1) $$G : [0, 1] \rightarrow [0, 1];$$

(1.2) G is one-to-one, onto, strictly increasing;

(1.3) $g(x) \equiv (d/dx)G(x)$ is continuous.

We can then define

(2) $$u_k(x) \equiv \exp(2\pi ikG(x))$$

for $-\infty < k < \infty$.

It is easily seen by a change of variable $t = G(x)$

$$\int_0^1 u_j(x)u_k(x)^* g(x)\, dx = \int_0^1 \exp(2\pi i(jG(x) - kG(x)))g(x)\, dx$$
$$= \int_0^1 \exp(2\pi i(jt - kt))\, dt = \delta_{jk}$$

that the family $\{u_k\}_{k=-\infty}^{\infty}$ is orthonormal with respect to g on $[0, 1]$. This immediately yields a series-type estimator considered earlier:

(3.1) $$\hat{f}(x) \equiv g(x)\Sigma_{k=-\infty}^{\infty} w_k(h)\hat{b}_k u_k(x)$$

(3.2) $$\hat{b}_k \equiv \frac{1}{N}\Sigma_{j=1}^{N} u_k(X_j)^*.$$

Thus a transformation G provides us with a new estimator. We will show later that if $G(x) \approx \int_0^x f(y)\, dy$ (that is, if $g \approx f$), then the new family $\{u_k\}_{k=-\infty}^{\infty}$ provides an improved estimate. We cannot choose G *a-priori*, of course, since knowledge of

G is equivalent to knowledge of f. However, we can estimate G from the sample. We propose the following algorithm.

(4) *Adaptive (or two-pass) estimator.* Choose $h_1 > 0$, $h_2 > 0$, N_1, and N_2 so that $N_1 + N_2 = N$. Let

$$\begin{aligned} &(4.1) && \hat{g}(x) \equiv \Sigma_{k=-\infty}^{\infty} w_k(h_1)\hat{a}_k \exp(2\pi ikx) \\ &&& \hat{a}_k \equiv \frac{1}{N_1}\Sigma_{j=1}^{N_1} \exp(-2\pi ikX_j) \\ &&& \hat{G}(x) \equiv \textstyle\int_0^x \hat{g}(y)\, dy \end{aligned}$$

$$\begin{aligned} &(4.2) && \hat{f}(x) \equiv \hat{g}(x)\Sigma_{k=-\infty}^{\infty} w_k(h_2)\hat{b}_k \exp(2\pi ik\hat{G}(x)) \\ &&& \hat{b}_k = \frac{1}{N_2}\Sigma_{j=N_1+1}^{N} \exp(-2\pi ik\hat{G}(X_j)). \end{aligned}$$

Remark. The choice of the parameters $\dot{N}_1$, N_2 and h_1, h_2 is not specified above. For theoretical analysis, h_1, and h_2 will be chosen as functions of N_1, N_2 (discussed below in Section 3.C). In practical application of .the estimator, we will choose $N_1 < N_2$, $h_1 > h_2$ so that $\hat{g}(x)$ is a low-resolution estimate of f and $\hat{f}$ in the second pass is a high resolution estimate. The reason we split the sample in two parts is to "decouple" the random functions $\hat{g}$ and $\hat{f}$, in order to simplify the analysis. In an actual application (Section 4), we will use the entire sample in both passes. Finally, we wish to note that from the proofs which will follow, it is clear that we need not choose the same sequence of weight functions $w_k(\cdot)$ for $\hat{g}$ and $\hat{f}$ (appearing in (4.1) and (4.2)) provided we require that, in each case, conditions (3) (or (3′)) hold.

C. Asymptotic error analysis. We will now develop asymptotic error estimates for the estimator (3.B.4). The development will be in two steps. First we will derive estimates based on the assumption that $\hat{g} = g$, a deterministic function satisfying certain inequalities. Second, we will determine bounds on the probability that $\hat{g}$ satisfies these inequalities. Thus the final estimates will hold "in probability."

Let $G(\cdot)$ be some deterministic function satisfying (3.B.1), and let $\hat{f}$ be defined by (3.B.3). We can rewrite the expression (3.B.3.2) for $\hat{b}_k$ as

$$(1) \qquad \hat{b}_k \equiv \frac{1}{N}\Sigma_{j=1}^{N} \exp(-2\pi ikT_j)$$

where

$$(2) \qquad T_j = G(X_j).$$

We know that the pdf of the transformed random variable $T = G(X)$ is just (see [14]) $r(\cdot)$ defined by

$$(3) \qquad r(t) = r(G(x)) \equiv f(x)/g(x).$$

We may consider $\hat{r}$, a simple Fourier series estimator for r, defined by

$$\hat{r}(t) \equiv \Sigma_{k=-\infty}^{\infty} w_k(h)\hat{b}_k \exp(2\pi ikt) \tag{4.1}$$

$$\hat{b}_k \equiv \frac{1}{N}\Sigma_{j=1}^{N} \exp(-2\pi ikT_j). \tag{4.2}$$

Since we clearly have

$$\hat{f}(x) \equiv g(x)\hat{r}(G(x)), \tag{5}$$

it follows that

$$\operatorname{Var}[\hat{f}(x)] = g(x)^2 \operatorname{Var}[\hat{r}(G(x))] \tag{6.1}$$

$$\operatorname{bias}[\hat{f}(x)] = g(x) \operatorname{bias}[\hat{r}(G(x))]. \tag{6.2}$$

Putting this together, we have the following

(7) THEOREM. *Suppose f and* $\{w_k\}_{k=-\infty}^{\infty}$ *satisfy the hypotheses of Theorem* (2.B.8). *Let* $G \in C^3[0, 1]$ *satisfy* (3.B.1) *and* $\hat{f}$ *be defined by* (3.B.3), r *by* (3), *and* $\hat{r}$ *by* (4). *Then for* $x_0 \in [0, 1]$ *such that* $g(x_0) \neq 0$,

$$\lim_{h\to 0} \frac{E[\hat{f}(x_0)] - f(x_0)}{c(h)} = g(x_0)r''(t_0)$$

where $t_0 = G(x_0)$. *Further, if* $h_N \to 0$ *as* $N \to \infty$, *then*

$$\lim_{N\to\infty} \frac{N \operatorname{Var}[\hat{f}(x_0)]}{v(h_N)} = f(x_0)g(x_0).$$

The proof of the theorem is immediate on applying Theorem (2.8) to $\hat{r}(\cdot)$.

We can see by the preceding theorem that the quantity $r''(t_0)$ is of interest in the asymptotic error of $\hat{f}(x_0)$. We will spend some time examining r'' and its dependence on the transformation G.

(8) LEMMA. *Let* $f, g \in C^2[0, 1]$ *be* pdf's. *Define*

$$G(x) \equiv \int_0^x g(y)\, dy$$

and for $x \in [0, 1]$ *such that* $g(x) > 0$

$$r(G(x)) \equiv f(x)/g(x).$$

Let $x_0 \in (0, 1)$ *with* $g(x_0) > 0$, *and* $t_0 = G(x_0)$. *Then*

$$r''(t_0) = \frac{d^2}{dt^2} r(t)\bigg|_{t=t_0} = \frac{1}{g(x_0)^5}\{g(x_0)^2 f''(x_0) - g(x_0)f(x_0)g''(x_0)$$
$$+ 3f(x_0)[g'(x_0)]^2 - 3g(x_0)f'(x_0)g'(x_0)\}.$$

The proof of this lemma, a straightforward calculation, is omitted. We now establish a bound on $r''(t_0)$ under the assumption that $g \approx f$.

(9) LEMMA. *With the same hypotheses of Lemma* (8), *suppose further that we have*

$$|g^{(k)}(x_0) - f^{(k)}(x_0)| \leqslant A \leqslant 1 \qquad \text{for} \quad k = 0, 1, 2.$$

Let $B(f, x_0) = \max\{1, f(x_0), |f'(x_0)|, |f''(x_0)|\}$. *Then at* $t_0 = G(x_0)$ *we have*

$$|r''(t_0)| \leqslant \frac{24AB(f, x_0)^2}{g(x_0)^5}$$

PROOF. For convenience, we will write f for $f(x_0)$, etc. We have by Lemma (8),

$$r''(t_0) = \frac{1}{g^5}\{ g^2f'' - gg''f + 3g'^2f - 3gg'f'\}$$

$$= \frac{1}{g^5}\{ g[gf'' - fg''] + 3g'[fg' - gf']\}.$$

We will make use of the easily verified inequality

$$|pq - rs| \leqslant \tfrac{1}{2}|p - r| \cdot |q + s| + \tfrac{1}{2}|p + r| \cdot |q - s|$$

First,

$$|gf'' - fg''| \leqslant \tfrac{1}{2}|g - f||f'' + g''| + \tfrac{1}{2}|g + f||f'' - g''|$$

$$\leqslant \tfrac{1}{2}A(2B + A) + \tfrac{1}{2}(2B + A)A \leqslant 3AB.$$

Similarly,

$$|fg' - gf'| \leqslant 3AB.$$

Moreover,

$$g = f + g - f \leqslant |f| + |g - f| \leqslant B + A \leqslant 2B$$

$$g' = f' + g' - f' \leqslant B + A \leqslant 2B.$$

Thus

$$|r''(t_0)| \leqslant \frac{2B \cdot 3AB + 3 \cdot 2B \cdot 3AB}{g^5}$$

$$\leqslant \frac{24AB^2}{g^5}. \qquad \square$$

We now collect what we have so far into a theorem giving asymptotic error estimates.

(10) THEOREM. *Suppose*

(10.1) $f \in C^3[0, 1]$ *and vanishes in a neighborhood of the endpoints*;

(10.2) $\{w_k\}_{k=-\infty}^{\infty}$ *satisfies conditions* (2.B.3);

(10.3) $G \in C^3[0, 1]$ *satisfies* (3.B.1).

Let $g(x) \equiv (d/dx)G(x), \hat{f}$ *be defined by* (3.B.3), *and* $x_0 \in (0, 1)$ *such that* $f(x_0) \neq 0$. *Choose numbers* $0 < p < 1$ *and* $0 < A < pf(x_0)$. *Suppose, moreover, that*

$$|g^{(k)}(x_0) - f^{(k)}(x_0)| \leqslant A \qquad \textit{for} \quad k = 0, 1, 2.$$

Then we have

$$\lim_{h \to 0} \frac{|E[\hat{f}(x_0)] - f(x_0)|}{c(h)} \leqslant 24 \frac{AB(f, x_0)^2}{f(x_0)^4 (1-p)^4}, \tag{10.4}$$

where

$$B(f, x_0) = \max\{1, |f^{(k)}(x_0)|, \ (k = 0, 1, 2)\}.$$

Furthermore, if $h_N \to 0$ *as* $N \to \infty$, *then*

$$\lim_{N \to \infty} \left| \frac{N \operatorname{Var}[\hat{f}(x_0)]}{v(h_N)} - f(x_0)^2 \right| \leqslant Af(x_0). \tag{10.5}$$

PROOF. By Theorem (7) we have

$$\lim_{h \to 0} \frac{E[\hat{f}(x_0)] - f(x_0)}{c(h)} = g(x_0) r''(t_0)$$

and

$$\lim_{N \to \infty} \frac{N \operatorname{Var}[\hat{f}(x_0)]}{v(h_N)} = f(x_0) g(x_0).$$

By Lemma (9) we have

$$|r''(t_0)| \leqslant \frac{24AB^2}{g(x_0)^5}.$$

Thus

$$\lim_{h \to 0} \left| \frac{E\hat{f}(x_0) - f(x_0)}{c(h)} \right| \leqslant \frac{24AB^2}{g(x_0)^4}.$$

Since

$$\frac{f(x_0)}{g(x_0)} \leqslant \frac{f(x_0)}{f(x_0) - A} \leqslant \frac{f(x_0)}{f(x_0) - pf(x_0)} = \frac{1}{1-p},$$

we obtain

$$\lim_{h \to 0} \left| \frac{E[\hat{f}(x_0)] - f(x_0)}{c(h)} \right| \leqslant \frac{24AB^2}{f(x_0)^4} \frac{1}{(1-p)^4}$$

which is (10.4). (10.5) follows immediately since

$$|f(x_0) g(x_0) - f(x_0)^2| \leqslant Af(x_0). \qquad \square$$

Now let us return to the adaptive estimator (3.B.4). We know that $\hat{g}(x_0)$ is a consistent estimator for $f(x_0)$, by Theorem 2.B.8 (with proper choice of $h_1 = h_{N_1}$). The next theorem extends consistency to the first and second derivative. First, however, we define

$$v_k(h) \equiv \textstyle\int_{-\frac{1}{2}}^{\frac{1}{2}} K_h^{(k)}(x)^2\,dx, \qquad \text{for } k = 0, 1, 2. \tag{11.1}$$

where $K_h^{(k)}(x) \equiv (d^k/dx^k)K_h(x)$. (Note $v_0(h) \equiv v(h)$.)

$$V(h) \equiv \max\{v_0(h), v_1(h), v_2(h)\}. \tag{11.2}$$

(12) THEOREM. *Let $\hat{g}$ be defined by* (3.B.4). *Suppose that the kernel K_h associated with $\{w_k\}_{k=-\infty}^{\infty}$ is in $C^2[0, 1]$, and $f \in C^5[0, 1]$ vanishes in a neighborhood of the endpoints. Define for $x \in (0, 1)$ and $k = 0, 1, 2$*

$$\hat{g}^{(k)}(x) \equiv \frac{d^k}{dx^k}[\hat{g}(x)].$$

Choose $h_1 = h_{N_1}$ to satisfy

$$\frac{V(h_{N_1})}{N_1} = c(h_{N_1})^2.$$

Then for $x_0 \in (0, 1)$, $E[\hat{g}^{(k)}(x_0) - f^{(k)}(x_0)]^2 \to 0$ as $N_1 \to \infty$.

PROOF. We can write $\hat{g}(x) = \frac{1}{N_1}\sum_{j=1}^{N_1} K_{h_1}(x - X_j)$. Since $K_{h_1} \in C^2[0, 1]$,

$$\hat{g}^{(k)}(x_0) = \frac{1}{N}\sum_{j=1}^{N} K_{h_1}^{(k)}(x_0 - X_j)$$

exists. Now by integration by parts, we get

$$\begin{aligned} E[\hat{g}^{(1)}(x_0)] &= \textstyle\int_0^1 K_{h_1}^{(1)}(x_0 - y)f(y)\,dy \\ &= -K_{h_1}(x_0 - y)f(y)|_0^1 - \textstyle\int_0^1[-K_{h_1}(x_0 - y)]f^{(1)}(y)\,dy \\ &= \textstyle\int_0^1 K_{h_1}(x_0 - y)f^{(1)}(y)\,dy. \end{aligned}$$

A similar result holds for $E[\hat{g}^{(2)}(x_0)]$.

Thus for $k = 0, 1, 2$, we obtain by previous methods

$$\begin{aligned} E[\hat{g}^{(k)}(x_0)] &= \textstyle\int_0^1 K_{h_1}(x_0 - y)f^{(k)}(y)\,dy \\ &= \textstyle\int_{-\frac{1}{2}}^{\frac{1}{2}} K_{h_1}(y)\Big[\mathring{f}^{(k)}(x_0) + \mathring{f}^{(k+1)}(x_0)y + \mathring{f}^{(k+2)}(x_0)\frac{y^2}{2} \\ &\quad + \mathring{f}^{(k+3)}(z(y))\frac{y^3}{3!}\Big]\,dy \\ &= f^{(k)}(x_0) + f^{(k+2)}(x_0)c(h_1) + \textstyle\int_{-\frac{1}{2}}^{\frac{1}{2}} K_{h_1}(y)\frac{y^3}{3!}\mathring{f}^{(k+3)}(z(y))\,dy. \end{aligned}$$

Thus we have an estimate for the bias

$$\begin{aligned}|E[\hat{g}^{(k)}(x_0)] - f^{(k)}(x_0)| &\leqslant c(h_1)\big[|f^{(k+2)}(x_0)| \\ &+ \sup_{x\in(0,1)}|f^{(k+3)}(x)|\big] \equiv c(h_1)A_k\end{aligned}$$

since

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} K_{h_1}(y)\frac{|y|^3}{3!}\,dy \leqslant \tfrac{1}{2}\int_{-\frac{1}{2}}^{\frac{1}{2}} K_{h_1}(y)y^2\,dy = c(h_1).$$

For the variance we have

$$\begin{aligned}\operatorname{Var}[\hat{g}^{(k)}(x_0)] &= \frac{1}{N_1}\operatorname{Var}[K_{h_1}^{(k)}(x_0 - X)] \\ &\leqslant \frac{1}{N_1}\int_0^1 K_{h_1}^{(k)}(x_0 - y)^2 f(y)\,dy.\end{aligned}$$

Again the Taylor expansion with remainder yields

$$\int_0^1 K_{h_1}^{(k)}(x_0 - y)^2 f(y)\,dy = v_k(h_1)f(x_0) + \tfrac{1}{2}\int_{-\frac{1}{2}}^{\frac{1}{2}} K_{h_1}^{(k)}(y)^2 y^2 \mathring{f}''(z(y))\,dy.$$

So

$$\begin{aligned}\operatorname{Var}[\hat{g}^{(k)}(x_0)] &\leqslant \frac{1}{N_1}\left\{v_k(h_1)f(x_0) + \sup_{x\in(0,1)}|f''(x)|\int_{-\frac{1}{2}}^{\frac{1}{2}} K_{h_1}^{(k)}(y)^2 y^2\,dy\right\} \\ &\leqslant \frac{1}{N_1}\{v_k(h_1)f(x_0) + \sup_{x\in(0,1)}|f''(x)|v_k(h_1)\} \\ &\equiv \frac{v_k(h_1)}{N_1}B_k.\end{aligned}$$

Hence, by the indicated choice $h_1 = h_{N_1}$,

$$\begin{aligned}E[\hat{g}^{(k)}(x_0) - f^{(k)}(x_0)]^2 &= |E[\hat{g}^{(k)}(x_0)] - f^{(k)}(x_0)|^2 + \operatorname{Var}[\hat{g}^{(k)}(x_0)] \\ &\leqslant c(h_{N_1})^2 A_k^2 + \frac{v_k(h_{N_1})}{N_1}B_k \to 0\end{aligned}$$

as $N_1 \to \infty$. □

We can now state the final and chief result on the asymptotic error of the adaptive estimator.

(13) THEOREM. *Suppose*

(13.1) $f \in C^5[0, 1]$ *and vanishes in a neighborhood of the endpoints*;

(13.2) $K_h(\cdot)$ *associated with* $\{w_k(h)\}_{k=-\infty}^{\infty}$ *is in* $C^2[0, 1]$;

(13.3) $\hat{f}$, $\hat{g}$ *are defined as in* (3.B.4);

(13.4) $\{h_{N_1}\}$ *is chosen to satisfy* $V(h_{N_1})/N_1 = c(h_{N_1})^2$;

$\{h_{N_2}\}$ *is chosen to satisfy* $v(h_{N_2})/N_2 = c(h_{N_2})^2$;

(13.5) $x_0 \in (0, 1)$ *such that* $f(x_0) \neq 0$.

Choose $\varepsilon > 0$, $1 > \delta > 0$. *Then there exists* N_1 *such that*

$$\Pr\left\{\lim_{N_2\to\infty}\left|\frac{E[\hat{f}(x_0)] - f(x_0)}{c(h_{N_2})}\right| \geqslant \varepsilon\right\} \leqslant \delta$$

and

$$\Pr\left\{\lim_{N_2\to\infty}\left|\frac{N_2 \operatorname{Var}[\hat{f}(x_0)]}{v(h_{N_2})} - f(x_0)^2\right| \geqslant \varepsilon\right\} \leqslant \delta.$$

PROOF. Recalling the notation of Theorem (10), let us pick A so that $0 < A < \frac{1}{2}f(x_0)$, $0 < A < \varepsilon/f(x_0)$, and

$$\frac{24AB(f, x_0)^2}{f(x_0)^4(1 - \frac{1}{2})^4} < \varepsilon.$$

Then by Theorem (10), if values $\{x_1, \cdots x_{N_1}\}$ are observed such that

(13.6) $$|\hat{g}^{(k)}(x_0) - f^{(k)}(x_0)| \leqslant A$$

for $k = 0, 1, 2$ then

(13.7) $$\lim_{N_2\to\infty}\left|\frac{E_1[\hat{f}(x_0)] - f(x_0)}{c(h_{N_2})}\right| \leqslant \varepsilon$$

and

(13.8) $$\lim_{N_2\to\infty}\left|\frac{N_2 \operatorname{Var}_1[\hat{f}(x_0)]}{v(h_{N_2})} - f(x_0)^2\right| \leqslant \varepsilon$$

where E_1 and Var_1 denote the conditional expectation and variance given $\{X_1 = x_1, \cdots, X_{N_1} = x_{N_1}\}$. Recall that by Tchebichev's inequality for a random variable Y we have

$$\Pr\{|\mathrm{Y}| \geqslant A\} \leqslant E[\mathrm{Y}]^2/A^2.$$

Now by Theorem (12) we have

$$E[\hat{g}^{(k)}(x_0) - f^{(k)}(x_0)]^2 \to 0 \quad \text{as} \quad N_1 \to \infty.$$

Thus there is some N_1 such that

$$E[\hat{g}^{(k)}(x_0) - f^{(k)}(x_0)]^2/A^2 \leqslant \delta.$$

Thus, for this N_1, bounds (13.7) and (3.8) fail to hold with probability $\leqslant \delta$. □

Discussion. We now consider an intuitive interpretation of Theorem (13). For this purpose, let us denote by $\hat{f}_1$ the simple Fourier series estimator defined in (2.B.1) and by $\hat{f}_2$ the adaptive estimator (3.B.4).

We have seen from Theorem (2.B.8) that for large N, the bias $|E[\hat{f}_1(x_0)] - f(x_0)| \approx |f''(x_0)|c(h_N)$. Theorem (13) gives the analogous result

$$|E[\hat{f}_2(x_0)] - f(x_0)| \leqslant \varepsilon c(h_{N_2}).$$

The factor of proportionality ε can be made as small as desired, such as $\varepsilon \ll |f''(x_0)|$, by reserving enough samples $X_1, \cdots, X_{N_1}$ in the first pass. Now if the ratio $c(h_N)/c(h_{N_2}) = c(h_N)/c(h_{N-N_1}) \to 1$ as $N \to \infty$, N_1 fixed, then the asymptotic bias of $\hat{f}_2(x_0)$ is smaller than that of $\hat{f}_1(X_0)$.

D. An optimality property of the Fourier basis. The choice of the Fourier basis simplifies analysis and implementation of the estimation (3.B.4). Moreover, one can argue that in a certain sense, the Fourier functions are a "good" choice.

It is necessary at this point to introduce further constraints on the class of densities we wish to estimate. This is necessary, because a particular basis is "good" only with respect to some particular class.

We begin by generalizing the estimator (3.B.3) by replacing the Fourier family with $\{v_k(\cdot)\}_{k=0}^{\infty}$, an arbitrary family which is orthonormal in $L_2[0, 1]$. Then the estimator (3.B.3) becomes

$$\hat{f}(x) \equiv g(x)\Sigma_{k=0}^{\infty}w_k(h)\hat{b}_k v_k(G(x)) \tag{1.1}$$

$$\hat{b}_k = \frac{1}{N}\Sigma_{j=1}^{N}v_k(G(X_j)). \tag{1.2}$$

We have seen that the estimator (1) may be viewed as an estimator for the transformed density

$$r(t) = r(G(x)) = f(x)/g(x) \tag{2}$$

of the random variable $T = G(X)$. The corresponding expression is

$$\hat{r}(t) = \Sigma_{k=0}^{\infty}w_k(h)\hat{b}_k v_k(t) \tag{3}$$

$$\hat{b}_k = \frac{1}{N}\Sigma_{j=1}^{N}v_k(T_j).$$

Now the error of $\hat{r}$ is related to the second derivative r''. Furthermore, the intent of the transformation $G(\cdot)$ is to reduce the magnitude of r''. In this spirit, we will place a constraint on the densities to be estimated by placing a bound on the magnitude of r''.

(4) Definition. The class $W_p[0, 1]$ consists of all functions $r \in C[0, 1]$ which have an absolutely continuous derivative $(p - 1)$ (thus $r^{(p)}(t)$ exists almost everywhere) and which satisfy $|r(t)| \leqslant 1$ a.e.

The class of densities we will try to estimate is $W_2[0, 1]$. We thus seek a family $\{v_k(\cdot)\}_{k=0}^{\infty}$ which will provide a "good" estimator for densities $r \in W_2[0, 1]$.

If we simplify the form of the estimator (3) to

$$\hat{r}(t) = \Sigma_{k=0}^{n}\hat{b}_k v_k(t) \tag{5}$$

then it is possible to pose the problem in such a way that it has a ready solution. Recall that the integrated bias-squared is

$$(6) \qquad \textstyle\int_0^1 \{E[\hat{r}(t)] - r(t)\}^2 \, dt = \sum_{k=n+1}^{\infty} b_k^2$$

where we assume r may be expanded

$$(7) \qquad r(t) = \textstyle\sum_{k=0}^{\infty} b_k v_k(t).$$

One approach to selecting $\{v_k\}$ is, for each n, to pick v_n so that the maximum (for r in W_2) of (6) is minimized. To make this precise, we must introduce some definitions.

(8) Definition.

(8.1) Let $C \subset L_2[0, 1]$ be a class of functions and $S_n \subset L_2[0, 1]$ be an n-dimensional subspace. The "degree of approximation" of C by S_n is

$$E_{S_n}(C) \equiv \sup_{u \in C} \inf_{v \in S_n} \|u - v\|_{L_2}.$$

(8.2) The n-width of C is

$$d_n(C) = \inf_{S_n} E_{S_n}(C)$$

where the infimum ranges over all n-dimensional subspaces in $L_2[0, 1]$

(8.3) If, for a particular subspace S_n^*, we have $d_n(C) = E_{S_n^*}(C)$ then S_n^* is called an optimal approximating n-dimensional subspace for C.

Now for our estimation problem, it is clear that the integrated bias-squared (7) is just the L_2 approximation error of r in $S(v_0, v_1, \cdots, v_n)$, the space spanned by $v_0, v_1, \cdots, v_n$. The maximum error for r in W_2 is the quantity $E_{S(v_0, v_1, \cdots, v_n)}(C)$. We seek the family $\{v_k\}$ which minimizes the maximum error.

We quote the following result, which may be found in [20].

(9) Theorem. *The functions* $\{1, \sin 2\pi t, \cos 2\pi t, \cdots, \sin 2\pi nt, \cos 2\pi nt\}$ *span a* $(2n + 1)$*-dimensional optimal approximating subspace for the class* W_p.

Thus if we take $\{v_k\}$ to be the Fourier functions, we minimize the maximum integrated bias squared for densities r in W_2.

Remark. The class W_2 contains many functions which are not densities, since there is no nonnegativity constraint in the definition of W_p. Thus the Fourier functions may not be strictly optimal when this constraint is included. However, it is reasonable to suppose that the Fourier functions are "good", if not strictly optimal.

4. Computer simulations. In Section 3 we have developed an asymptotic error analysis for the adaptive estimator which describes large-sample behavior. The asymptotic approximations made are not valid for small samples. Yet it is the case of small samples which is most important in practice. Hence we must turn to computer simulations to demonstrate the behavior for small samples.

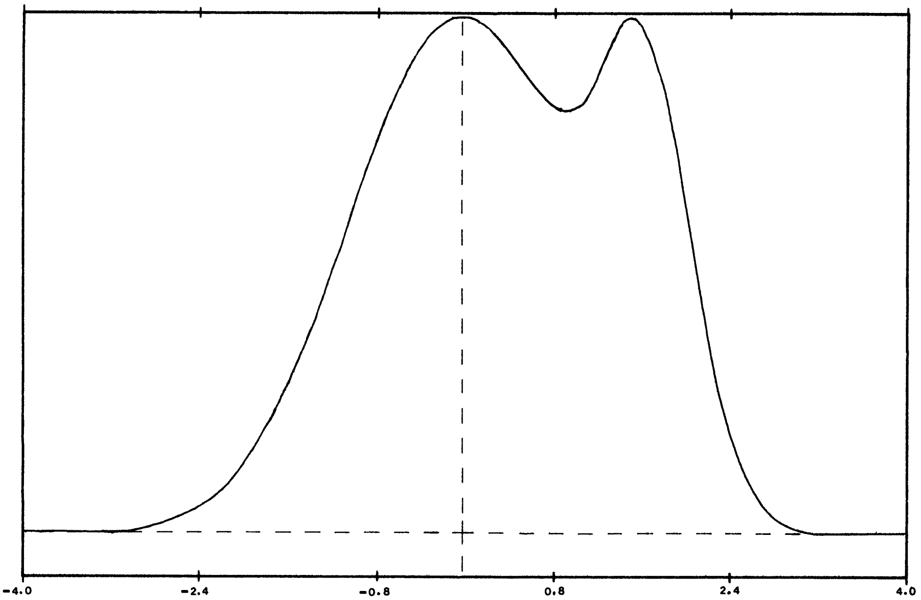

FIG. 4.1. *True density f.*

In the following simulations we consider a mixture of two Gaussians

$$(1) \qquad f(x) = 0.78 f_1(x) + 0.22 f_2(x)$$

where f_1 is $N(0, 1)$ and f_2 is $N(1.6, 0.4)$. The sample set consists of $N = 100$ independent variates drawn from this density, generated by a standard (polar method) pseudorandom number generator.

This pdf was chosen as a test case because it has two closely spaced modes separated by a shallow valley (see Figure 4.1). The adaptive estimator promises reduced bias, and hence it should be able to resolve the modes better than the conventional Fourier series estimator.

In the theoretical (asymptotic) analysis in Section 3, we partitioned the sample set $\{X_1 \cdots, X_N\}$ into two parts $\{X_1, \cdots X_{N_1}\}$, $\{X_{N_1+1}, \cdots, X_N\}$. The first part was used in the first pass, and the second part was used in the second pass. The partitioning greatly simplified the theoretical analysis. However, in small-sample-set numerical trials, it was found that performance of the estimator improved if the *entire* sample was used in both passes. The numerical trials reported below were thus conducted.

Specifically, for a sample set $\{X_1, \cdots X_N\}$ $(N = 100)$, the estimator was implemented as follows:

$$(2.1) \qquad \hat{g}(x) \equiv \Sigma_{k=0}^{20}(1 - h)^k \hat{a}_k \cos 2\pi kx;$$

$$(2.2) \qquad \hat{a}_k \equiv \frac{2}{N}\Sigma_{j=1}^{N} \cos 2\pi kX_j \qquad (k \geqslant 1) \qquad a_0 \equiv 1;$$

$$\hat{G}(x) \equiv \textstyle\int_0^x \hat{g}(y)\,dy; \tag{2.3}$$

$$\hat{f}_2(x) \equiv \hat{g}(x)\Sigma_{k=0}^{5}\hat{b}_k \cos(2\pi k\hat{G}(x)); \tag{2.4}$$

$$\hat{b}_k \equiv \frac{2}{N}\Sigma_{j=1}^{N} \cos(2\pi k\hat{G}(X_j)) \qquad (k \geqslant 1) \qquad \hat{b}_0 \equiv 1. \tag{2.5}$$

(The expansions employ only cosines in order to simplify the computer program.)

The adaptive estimator $\hat{f}_2$ will be compared to the simple Kronmal-Tarter type defined by

$$\hat{f}_1(x) \equiv \Sigma_{k=0}^{n}\hat{c}_k \cos 2\pi kx; \tag{3.1}$$

$$\hat{c}_k \equiv \frac{2}{N}\Sigma_{j=1}^{N} \cos 2\pi kX_j \qquad (k \geqslant 1) \qquad \hat{c}_0 \equiv 1. \tag{3.2}$$

To make this comparison more direct, in (2.4) we have chosen a weight sequence corresponding to simple truncation. (The truncation point 5 was chosen by trial and error.) Note that for $h = 1$, the estimator $\hat{f}_2$ is identical to $\hat{f}_1$ for $n = 5$. Below we will observe the effect of varying h and n.

The results of the trials will be presented in two ways. First, we will examine the estimates obtained from one fixed sample set as h varies for $\hat{f}_2$ and n varies for $\hat{f}_1$. These estimates are shown in graphical form in Figures 4.2 through 4.7. Second, the

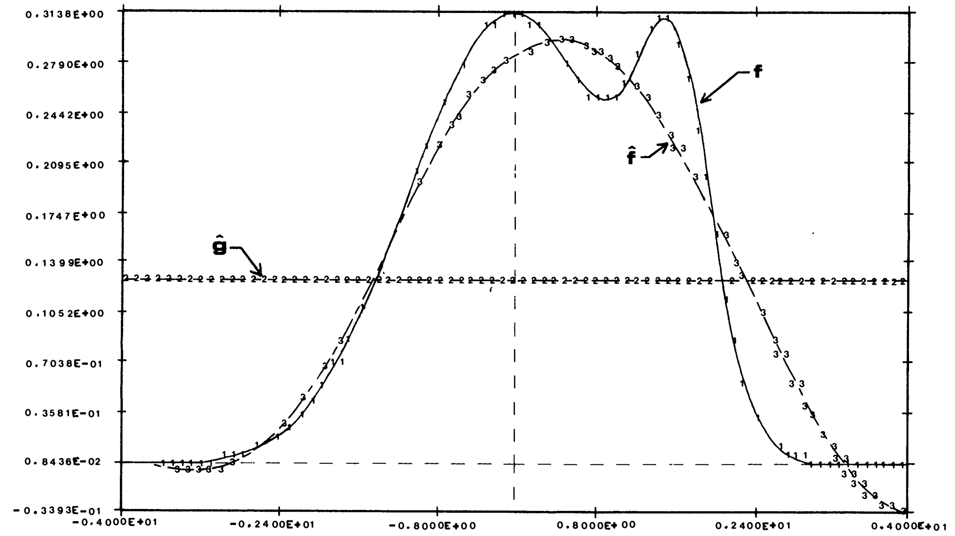

FIG. 4.2. *Adaptive estimator* (h = 1.0).

integrated square error

$$\int_0^1(\hat{f}(x) - f(x))^2\,dx$$

will be computed for 25 sample sets, and statistically reliable conclusions will be drawn.

Figure 4.2 shows the result for $\hat{f}_2$ and $h = 1$. This is the trivial case, since for this choice of h, $\hat{g}(x) \equiv 1$; it is identical to a simple Fourier series estimate. Note that the estimate $\hat{f}_2$ does not resolve the two modes of f. Also we see a substantial

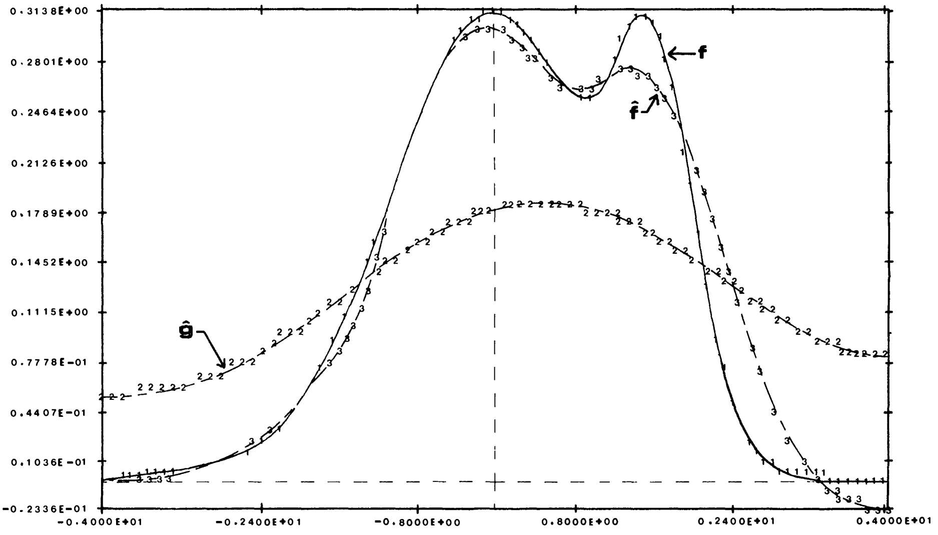

FIG. 4.3. *Adaptive estimator* (h = .04).

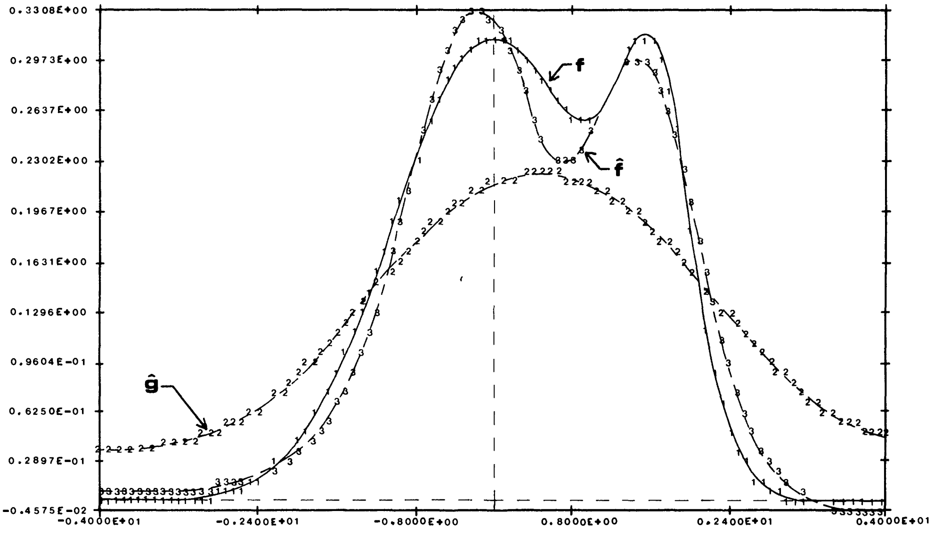

FIG. 4.4. *Adaptive estimator* (h = 0.25).

negative tail at the right of the graph. The negativity is a result of truncating rather than tapering the series terms in (2.4).

Figure 4.3 shows the results for $h = 0.4$. Now $\hat{g}$ begins to concentrate mass near the modes of f. We see that $\hat{f}_2$ begins to resolve the modes and that the negative tail is somewhat reduced.

In Figure 4.4, h equals 0.25. Now $\hat{f}_2$ does a very good job of resolving the modes, and the negative tail is almost eliminated.

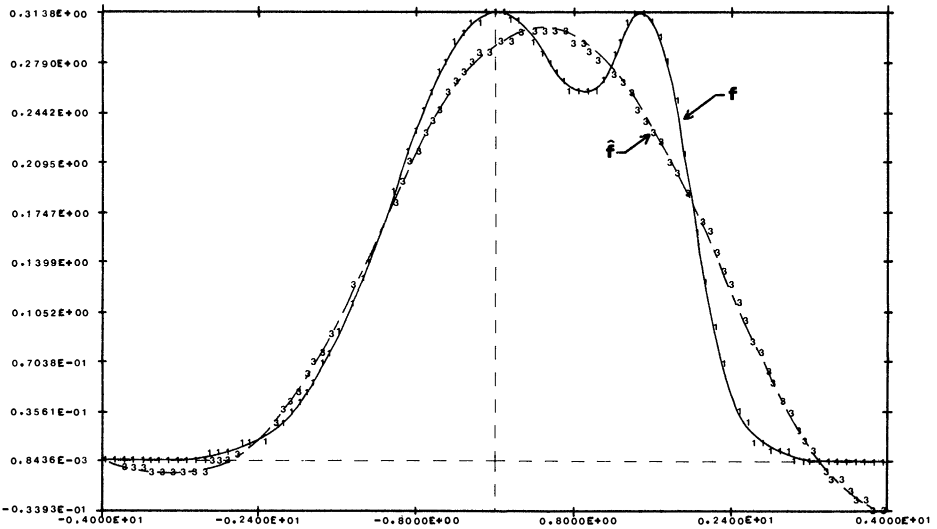

FIG. 4.5. *Fourier series estimator* ($n = 5$).

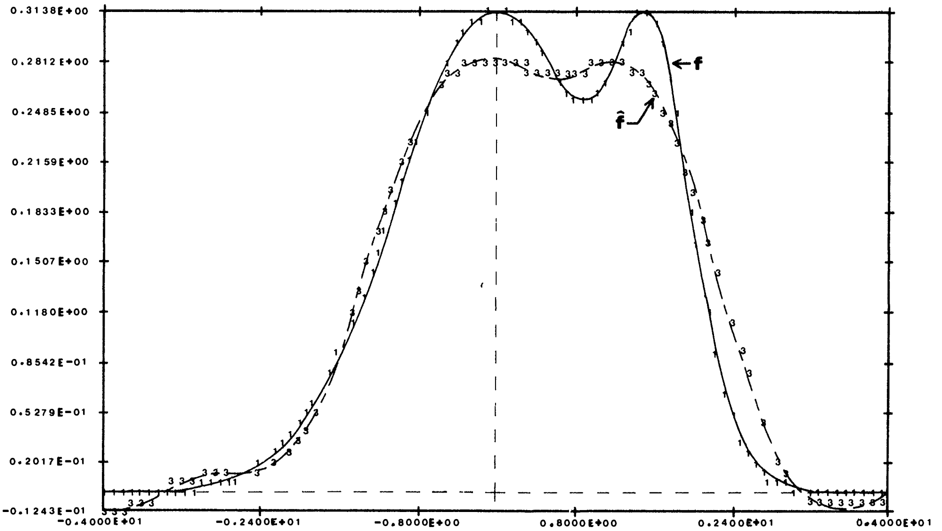

FIG. 4.6. *Fourier series estimator* ($n = 7$).

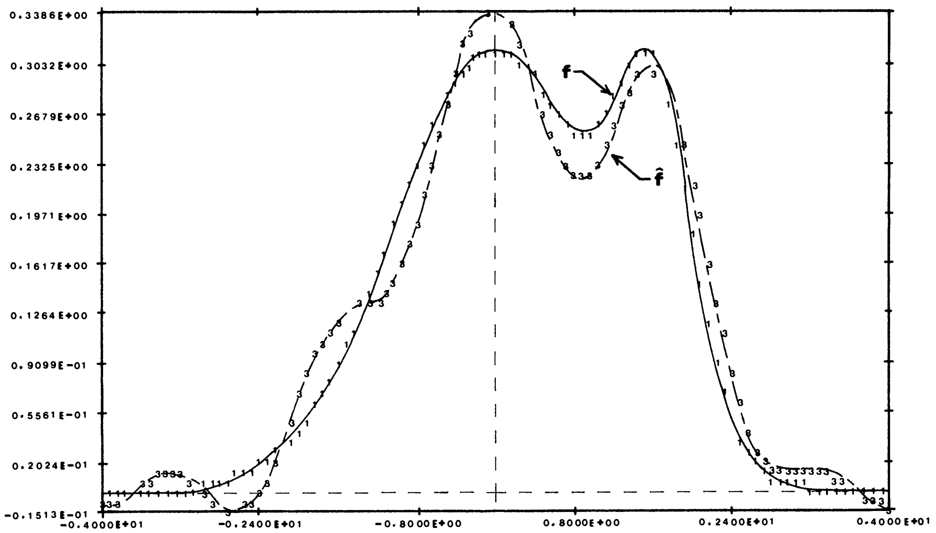

FIG. 4.7. *Fourier series estimator* (*n* = 10).

Clearly, Figure 4.4 is a much better estimate than Figure 4.2. By allowing the estimator to adapt (as h varies) we have greatly reduced the bias.

One may wonder how well the simple Fourier estimator (3) would perform if we vary n. The case of $n = 5$ is shown in Figure 4.5. (This is in fact the same estimate as in Figure 4.1.) Now as we increase to $n = 7$ (Figure 4.6) and to $n = 10$ (Figure 4.7), the performance is improved. However, even in the best case ($n = 10$), the simple Fourier series estimator is inferior to the adaptive estimator. Note in particular that the simple estimator is able to resolve the modes in Figure 4.7 only at the expense of introducing spurious modes (and negative values) in the tails. This behavior is characteristic, since the simple series estimator provides a constant amount of resolution over the entire interval $[a, b]$. The adaptive estimator, on the other hand, tunes its resolution to the data; it provides higher resolution where the density of the data is higher.

Next, we examine some Monte Carlo estimates of the integrated mean square error of $\hat{f}_1$ and $\hat{f}_2$. Twenty-five sample sets, each set consisting of one hundred variates, were independently generated. For the ith sample set ($i = 1, \cdots, 25$), estimates $\hat{f}_{1,i}$ and $\hat{f}_{2,i}$ were obtained. Fór each estimate, the integrated square error

$$(4) \qquad e_{k,i} = \int_0^1 \left(\hat{f}_{k,i}(x) - f(x)\right)^2 dx \quad (k = 1, 2;\ i = 1, \cdots 25)$$

was computed by numerical integration. These errors are tabulated in Table 4.1.

Column A is the result for the adaptive estimator $\hat{f}_2$ with $h = 0.25$. The average $\bar{e}_2$ is 0.0078 with standard deviation 0.0043. Compare this with column B, the result for the simple Fourier series estimator $\hat{f}_1$ with $n = 5$. For the latter, $\bar{e}_1 = 0.0099$ with standard deviation 0.0028.

TABLE 4.1
Integrated squared error

Trial	A	B	C	D
	$e_{2,i}$ for $\hat{f}_2$ $h = 0.25$	$e_{1,i}$ for $\hat{f}_1$ $n = 5$	$e_{1,i}$ for $\hat{f}_1$ $n = 10$	$e_{1,i}$ for $\hat{f}_1$ $n = 7$
1	.0027	.0075	.0038	.0036
2	.0120	.0090	.0117	.0090
3	.0186	.0193	.0160	.0169
4	.0169	.0115	.0179	.0132
5	.0039	.0093	.0141	.0073
6	.0086	.0118	.0068	.0083
7	.0048	.0087	.0045	.0062
8	.0064	.0085	.0039	.0055
9	.0072	.0099	.0118	.0061
10	.0063	.0095	.0072	.0063
11	.0063	.0078	.0108	.0051
12	.0036	.0073	.0030	.0043
13	.0148	.0162	.0169	.0154
14	.0034	.0079	.0041	.0040
15	.0042	.0078	.0037	.0036
16	.0033	.0072	.0018	.0030
17	.0043	.0071	.0046	.0029
18	.0129	.0097	.0147	.0105
19	.0064	.0084	.0125	.0052
20	.0104	.0107	.0202	.0140
21	.0076	.0107	.0079	.0074
22	.0096	.0112	.0144	.0104
23	.0085	.0103	.0154	.0074
24	.0058	.0100	.0052	.0067
25	.0067	.0103	.0063	.0070
Mean	.0078	.0099	.0096	.0076
Standard Deviation	.0043	.0028	.0055	.0039

For these trials, the average integrated squared error for $\hat{f}_2$ is substantially less than that for $\hat{f}_1$. Since $n = 5$, the only difference between the two estimators is the preprocessing step (2.1–2.3). This clearly shows the improvement obtained by the prior transformation $\hat{G}$.

We would like to test the difference in the averages of $\bar{e}_1$ and $\bar{e}_2$ for statistical significance. Since the random variables $e_{k,i}$ have no readily identifiable distribution, we will employ a distribution-free sign test for the median difference (see [22]). Consider the null hypothesis

$$H : \text{median}(e_1 - e_2) = 0$$

against the alternative

$$A : \text{median}(e_1 - e_2) > 0.$$

Clearly if H is true then $e_2 > e_1$ is as likely as $e_2 < e_1$ and $\hat{f}_2$ is no better than $\hat{f}_1$. If A is true, however, then $e_2 < e_1$ is more likely.

Comparing columns A and B, we find $e_{2,i} < e_{1,i}$ occurs 22 times, with the reverse occuring three times. Referring to the one-tailed cumulative binomial distribution we see that H may be rejected with significance 0.001.

Next we compare $\hat{f}_2$ to $\hat{f}_1$ for $n = 10$ (column C). Here again the average $\bar{e}_2 < \bar{e}_1$. However, the sign test is not significant for 25 trials. Therefore, another 25 trials were run and the results are tabulated in Table 4.2. Applying the sign test for the 50 trials yields 34 occurrences of $e_{2i} < e_{1i}$ and 16 occurrences of $e_{2i} \geqslant e_{1i}$. Thus we may reject H with significance 0.01.

Column D tabulates the results of 25 trials for $\hat{f}_2$ with $n = 7$. Note that $\bar{e}_1 = 0.0076$, which is not significantly different from $\bar{e}_2$. Thus, in mean-square error alone, $\hat{f}_2$ is not better than $\hat{f}_1$ for $n = 7$. However, by another performance measure, $\hat{f}_2$ is substantially better. One important task of a pdf estimator is to resolve and estimate the location of the modes of the pdf. Thus, let us define another error measure m equal to the sum of the squared distances from the true modes (located

TABLE 4.2
Integrated squared error (continued)

Trial	A	C
26	.0032	.0025
27	.0045	.0048
28	.0074	.0076
29	.0051	.0060
30	.0172	.0184
31	.0095	.0112
32	.0102	.0119
33	.0088	.0121
34	.0047	.0071
35	.0091	.0145
36	.0064	.0101
37	.0034	.0060
38	.0084	.0087
39	.0166	.0154
40	.0083	.0120
41	.0070	.0105
42	.0105	.0096
43	.0065	.0065
44	.0094	.0099
45	.0052	.0080
46	.0097	.0153
47	.0055	.0083
48	.0031	.0027
49	.0088	.0092
50	.0063	.0053
Mean	.0078	.0093
Standard Deviation	.0035	.0040

TABLE 4.3
Error in location of modes

Trial	m_{2i} for $\hat{f}_2$ ($h = 0.25$)	m_{1i} for $\hat{f}_1$ ($n = 7$)
1	.06	.11
2	.08	2.85*
3	.39	4.23*
4	.42	.34
5	.32	1.24*
6	.22	.26
7	.16	1.31*
8	.39	1.16*
9	.03	.13
10	.01	.12
11	.03	1.70*
12	.03	.19
13	1.54*	1.41*
14	.03	.12
15	.32	.16
16	.26	.26
17	.33	2.32*
18	.62	.58
19	.08	2.57*
20	.34	.31
21	.26	.34
22	.05	.16
23	.32	1.54*
24	1.41*	1.18*
25	.01	1.48*
Mean	0.31	1.04

* Estimate was unimodal

at $x = 0$ and $x = 1.6$) to the nearest modes of the estimate. Thus if $\hat{f}$ has modes at $x = -0.2$ and 1.4, then $m = (-0.2 - 0)^2 + (1.4 - 1.6)^2 = 0.08$; if $\hat{f}$ is unimodal with mode at, say, $x = 1.0$, then $m = (1 - 0)^2 + (1 - 1.6)^2 = 1.36$. Errors m_{2i} for $\hat{f}_2$ and m_{1i} for $\hat{f}_1(n = 7)$ are tabulated in Table 4.3 for the 25 trials. The average $\overline{m}_2 = 0.31$ which is substantially less than $\overline{m}_1 = 1.04$. Note that $\hat{f}_1$ failed to resolve the modes (that is, $\hat{f}_1$ was unimodal) in 12 of the 25 trials; $\hat{f}_2$ failed to resolve in only 2 trials. Thus, although $\hat{f}_1$ with $n = 7$ performs as well as $\hat{f}_2$ in the "average" measure of integrated square error, $\hat{f}_2$ provides greatly enhanced resolution (that is, lower bias). Applying the median difference sign test to Table 4.3 yields a significance of 0.02.

5. Summary and conclusions. We have looked in detail at the orthogonal-series type of estimator and at its asymptotic error analysis. The main contribution of this paper is the proposal of a new estimator. This estimator is constructed by means of a prior data-dependent transformation of the basis in order to reduce the bias of the estimate. We have developed an asymptotic error analysis of the adaptive estimator; and to demonstrate the small-sample behavior of the estimator, we have considered some computer implementations.

As we see from both the error analysis and the computer simulations, there is an advantage to be gained from performing the data-dependent transformation. Resolution is improved (bias is reduced) in comparison to the conventional Fourier-series estimator. This improvement could be of significance in pattern-recognition applications. As shown in the computer simulations, the adaptive estimator was able to resolve closely-spaced modes without introducing spurious modes in the tails of the densities. In pattern recognition we are interested in ratios of probability density functions. The ability to detect the fine structure of densities from a limited set of samples can lead to improved discriminant functions (and hence a lower rate of misclassification).

Acknowledgments. It is a pleasure to acknowledge the many helpful discussions with Professor James R. Thompson on the topic of this research; we greatly appreciated his encouragement and constructive criticism. We are indebted to the referees for their careful reading of the original manuscript and constructive comments.

REFERENCES

[1] Anderson, G. L. (1978). An adaptive orthogonal-series estimator for probability density functions. Ph.D thesis, Dept. Math. Sciences, Rice Univ.

[2] Bennett, J. O., de Figueiredo, R. J. P. and Thompson, J. R. (1974). Classification by means of B-spline potential functions with application to remote sensing. (Invited paper) *Proc. Sixth Southeastern Symp. System Theory*. Baton Rouge, Louisiana. IEEE Publ. 74CH0872-2-SSST, *FA*-3:1–8.

[3] Bradley, J. V. (1968). *Distribution-free Statistical Tests*. Prentice-Hall.

[4] Brunk, H. D. (1976). Univariate density estimation by orthogonal series. *Biometrika*. To appear.

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

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

[7] De Montricher, G. F., Tapia, R. A. and Thompson, J. R. (1975). Nonparametric maximum likelihood estimation of the probability densities by penalty function methods. *Ann. Statist.* **3** 1329–1348.

[8] Devroye, L. P. and Wagner, T. J. (1977). The strong uniform consistency of nearest neighbor density estimates. *Ann. Statist.* **5** 536–540.

[9] Good, I. J. and Gaskins, R. A. (1972). Global nonparametric estimation of probability densities. *Virginia J. Science* **23** 171.

[10] Kronmal, R. and Tarter, M. (1968). The estimation of probability densities and cumulatives by Fourier series methods. *J. Amer. Statist. Assoc.* **63** 925–952.

[11] Lorentz, G. G. (1966). *Approximation of Functions*. Chapter 8. Holt, Rinehart and Winston.

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

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

[14] Roussas, G. G. (1973). *A First Course in Mathematical Statistics*. Addison-Wesley.

[15] Schwartz, S. C. (1967). Estimation of probability density by an orthogonal series. *Ann. Math. Statist.* **38** 1261–1265.

[16] SCOTT, D. W., TAPIA, R. A. and THOMPSON, J. R. (1976). An algorithm for nonparametric density estimation. *Proc. Ninth Interface Symp. Computer Sci. Statist.* 287–292. Prindel, Weber, & Schmidt, Boston.

[17] SCOTT, D. W., TAPIA, R. A. and THOMPSON, J. R. (1977). Kernel density estimation revisited. *J. Nonlinear Analysis, Theory, Methods and Applications.* **1** 339–372.

[18] VAN RYZIN, J. (1966). Bayes risk consistency of classification procedures using density estimation. *Sankhya, Ser. A* **28** 261–270.

[19] WAGNER, T. J. (1973). Strong consistency of a nonparametric estimate of a density function. *IEEE Trans. Systems Man Cybernet.* **3** 289–290.

[20] WAHBA, G. (1977). Optimal smoothing of density estimates. In *Classification and Clustering.* (J. van Ryzin, ed.). Academic Press.

[21] WAHBA, G. (1978). Data based optimal smoothing of orthogonal series density estimators. Dept. Statist., Univ. Wisconsin, Report no 509.

[22] WATSON, G. S. (1969). Density estimation of orthogonal series. *Ann. Math. Statist.* **40** 1496–1498.

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

[24] WHITTLE, P. (1958). On the smoothing of probability density functions. *J. Roy. Statist. Soc., Ser. B* **20** 334–343.

[25] WRAGG, A. and DOWSON, D. C. (1970). Fitting continuous probability density functions over (0, ∞) using information theory ideas. *IEEE Trans. Information Theory* **IT-16** 226–230.

DEPARTMENT OF MATHEMATICAL SCIENCES AND
DEPARTMENT OF ELECTRICAL ENGINEERING
RICE UNIVERSITY
HOUSTON, TEXAS 77001