INFORMATION TO USERS

This material was produced from a microfilm copy of the original document. While the most advanced technological means to photograph and reproduce this document have been used, the quality is heavily dependent upon the quality of the original submitted.

The following explanation of techniques is provided to help you understand markings or patterns which may appear on this reproduction.

- 1. The sign or "target" for pages apparently lacking from the document photographed is "Missing Page(s)". If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting thru an image and duplicating adjacent pages to insure you complete continuity.**
- 2. When an image on the film is obliterated with a large round black mark, it is an indication that the photographer suspected that the copy may have moved during exposure and thus cause a blurred image. You will find a good image of the page in the adjacent frame.**
- 3. When a map, drawing or chart, etc., was part of the material being photographed the photographer followed a definite method in "sectioning" the material. It is customary to begin photoing at the upper left hand corner of a large sheet and to continue photoing from left to right in equal sections with a small overlap. If necessary, sectioning is continued again -- beginning below the first row and continuing on until complete.**
- 4. The majority of users indicate that the textual content is of greatest value, however, a somewhat higher quality reproduction could be made from "photographs" if essential to the understanding of the dissertation. Silver prints of "photographs" may be ordered at additional charge by writing the Order Department, giving the catalog number, title, author and specific pages you wish reproduced.**
- 5. PLEASE NOTE: Some pages may have indistinct print. Filmed as received.**

University Microfilms International

300 North Zeeb Road
Ann Arbor, Michigan 48106 USA
St. John's Road, Tyler's Green
High Wycombe, Bucks, England HP10 8HR

7814739

ANDERSON, G. LEIGH
AN ADAPTIVE ORTHOGONAL-SERIES ESTIMATOR FOR
PROBABILITY DENSITY FUNCTIONS.

RICE UNIVERSITY, PH.D., 1978

University
Microfilms
International

300 N. ZEEB ROAD, ANN ARBOR, MI 48106

RICE UNIVERSITY

AN ADAPTIVE ORTHOGONAL-SERIES ESTIMATOR
FOR PROBABILITY DENSITY FUNCTIONS

by

GEORGE LEIGH ANDERSON

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

THESIS DIRECTOR'S SIGNATURE:

Rui J. P. de Figueiredo

HOUSTON, TEXAS

(FEBRUARY 1978)

A C K N O W L E D G M E N T S

It is a pleasure to acknowledge the many helpful discussions with Professors Rui J. P. de Figueiredo and James R. Thompson; I greatly appreciated their direction, encouragement, and constructive criticism during the development of this thesis. Furthermore, I wish to thank Professors Paul E. Pfeiffer, David A. Schum, and Panayota P. Kazakos, who have served on several examination committees during my tenure at Rice University.

Next, I gratefully acknowledge the patience and help of my wife Bettie during the development and typing of this thesis.

Finally, I would like to recognize the financial support of the NSF grant ENG 74-17955 and the AFOSR grant 75-2777.

T A B L E O F C O N T E N T S

I. A. Introduction	1
B. Summary of Previous Approaches	2
C. Summary of the Thesis	3
D. Notation and Conventions	4
II. A. Series-Type Estimators	5
B. Fourier Series Estimators	8
III. The Data-Adaptive Estimator	
A. Motivation	22
B. Construction of the Estimator	23
C. Asymptotic Error Analysis	25
D. An Optimality Property of the Fourier Basis	41
IV. Computer Simulations	45
V. Summary and Conclusions	51
Figures and Tables	
References	

I. A. Introduction

A real random variable (r.v.) X is characterized by the associated cumulative distribution function (c.d.f.)

$$1) F(x) \triangleq \Pr \{X \leq x\}.$$

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 (p.d.f.) $f(\cdot)$ as

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

the Radon-Nykodym derivative of F .

In many statistical situations, the p.d.f. is not known a priori, and the investigator must estimate f from a sample set $\{X_1, \dots, X_N\}$ where each X_k is independent with density $f(\cdot)$.

In many cases, mathematical analysis or physical theory leads to the conclusion that f belongs to some class of functions which are characterized by some parameters $p_1, \dots, p_r$. Then the investigator must only determine the values of the r parameters. This is called "parametric estimation." An example is the frequently occurring case where X is a Gaussian r.v.; then only the mean

$$\mu = E[X]$$

and variance

$$\sigma^2(X) = E[X - \mu]^2$$

are required to characterize X .

However, in many situations, the p.d.f. f belongs to no known parametric class. This situation may arise when the underlying physical mechanism generating X is unknown or extremely complicated. In this case the investigator must estimate the entire function $f(\cdot)$

rather than a vector of parameters. This task is known as "non-parametric" estimation.

Several techniques of non-parametric estimation have been proposed by a number of researchers. These will be reviewed in the next section.

The current thesis concerns 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 transformation is to reduce the bias of the estimator for a sample set of a given size N . The transformation is obtained from a pre-processing step wherein we examine the sample set before applying the estimator.

I.B. Summary of Previous Approaches

One of the earliest and most widely studied non-parametric density function estimator was introduced by M. Rosenblatt [1] in 1955. He proposed the kernel-type estimator

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

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 [2] in 1961.

G.S. Watson and M.R. Leadbetter [3] 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 [4] with the use of B-splines. K.B. Davis [5] studied a kernel which is not L_1 and demonstrated

superior asymptotic properties; numerical trials with small sample sizes show poor performance, however [6]. Convergence conditions for kernel estimators [7] 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 [9], Cencov [10], van Ryzin [11], and Schwatz [12]; they developed error estimates and optimal series approximations. The optimal results require knowledge of the unknown density f . H.D. Brunk [13] 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 [14]. In this theoretical paper, the density estimate is the one which maximizes a penalized likelihood function. A discretized numerical implementation by D. Scott [21], gave excellent small-sample performance. An earlier effort along these lines is that of I.J. Good and R.A. Gaskins [15].

A. Wragg and D.C. Dowson [16] use the information-theoretic concept of entropy to fit density functions to a truncated moment sequence. Grace Wahba [17] and P. Whittle [18] employ notions from stochastic processes to obtain "optimally-smoothed" density estimates.

I. C. Summary of the Thesis

In chapter II, 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 chapter III, we introduce a new data-adaptive modification of the Fourier series estimator.

The series is modified with a transformation derived from a pre-processing 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. Also, we discuss a choice made in constructing the estimator; we argue that the Fourier series is a good starting point for the construction of the adaptive estimator. Finally, in chapter IV we examine some computer simulations to study the behavior of the estimator on small sample sets.

I. D. Notation and Conventions

Throughout this thesis we will assume the following notation and conventions.

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

II.A. Series-type Estimators

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 .

$$1) \quad L_2(g) = \left\{ s: (a,b) \rightarrow \mathbb{R} \mid \int_a^b s(x)^2 g(x) dx < \infty \right\}.$$

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

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

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

By orthogonality, we can see

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

Now an estimator for b_k is

$$3) \quad \hat{b}_k \triangleq \frac{1}{N} \sum_{j=1}^N u_k(X_j).$$

Thus we can construct an estimate of f by

$$4) \quad \hat{f}(x) \triangleq g(x) \sum_{k=0}^n \hat{b}_k u_k(x)$$

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

6.

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

This last expression is just

$$6) \sum_{k=0}^n \frac{\text{var}[u_k(X)]}{N} + \sum_{k=n+1}^{\infty} b_k^2.$$

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) .

$$\text{If } E[\hat{f}_N(x_0) - f(x_0)]^2 \xrightarrow[N \rightarrow \infty]{} 0 \quad \text{then } \hat{f}_N \text{ is}$$

"asymptotically consistent in the mean square sense at x_0 ."

$$\text{If } \int_b^a E[\hat{f}_N(x) - f(x)]^2 dx \xrightarrow[N \rightarrow \infty]{} 0 \quad \text{then } \hat{f}_N \text{ is}$$

"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

$$8.1) \quad n(N) \rightarrow \infty \text{ as } N \rightarrow \infty,$$

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

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

$$\text{var} [u_k(x)] \leq B, \quad k=0,1,2,\dots$$

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 [9].

An often-studied extension of (4) is

$$9) \quad \hat{f}(x) \triangleq g(x) \sum_{k=0}^{\infty} w_k(h) \hat{b}_k u_k(x)$$

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

$$10.1) \quad w_k(h) \rightarrow 0 \quad \text{as } k \rightarrow \infty$$

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

Optimal choices of the weight sequence $\{w_k(h)\}_{k=0}^{\infty}$ have been studied in [13]. Briefly, the optimal functional form of $w_k(\cdot)$ depends on f , and the choice $h = h(N)$ depends on the sample set size. For a fixed N , the integrated mean square error of $\hat{f}$ varies with h .

As h decreases, the bias squared decreases but the variance increases.

Since optimal choices for $\{w_k(h)\}_{k=0}^{\infty}$ depend on the unknown f , in practice the optimal theory can not be applied. Brunk [13] investigates heuristic ways of choosing the weight sequence.

The particular choices for $\{w_k(\cdot)\}_{k=0}^{\infty}$ are

$$11.1) \quad w_k(h) = \begin{cases} 1, & \text{for } k \leq 1/h \\ 0, & \text{for } k > 1/h \end{cases}$$

$$11.2) \quad w_k(h) = (1 - h)^k$$

(11.1) corresponds to simple truncation as in (4). We will use (11.2) somewhat later.

II.B. Fourier Series Estimators

We will now examine in some detail a special case of the orthogonal series estimator. This will be the starting point for the modified estimator of chapter III.

The Fourier series estimator (II.A.4) has been studied extensively by Kronmal and Tarter [9]. 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 [1] and [2].

From now on we will assume that f takes its support on a finite interval $[a, b]$. This assumption does not introduce a large

source of error, since we may reason as follows:

Suppose f is supported on $(-\infty, \infty)$, and our sample set satisfies $a < X_k < b$, $k = 1, \dots, N$, and a, b finite.

$$\text{Let } p = \int_{-\infty}^a f(x)dx + \int_b^{\infty} f(x)dx.$$

By constructing an estimator on $[a, b]$, we will actually be estimating f "raised on a pedestal"

$$\tilde{f}(x) \triangleq \begin{cases} \frac{1}{1-p} f(x), & x \in [a, b] \\ 0 & , x \notin [a, b]. \end{cases}$$

Thus truncation to $[a, b]$ introduces an error roughly of magnitude p .

But we can construct a hypothesis test for the magnitude of p . Pick significance $0 < \alpha < 1$, and let p_0 be such that $(1 - p_0)^N = \alpha$.

Then we can test the hypothesis

$$H : p \leq p_0.$$

$$\text{Clearly, } \Pr \{ X_1, \dots, X_N \in [a, b] ; p \} = (1-p)^N$$

Hence if $p > p_0$, the probability that we observe $X_1, \dots, X_N \in [a, b]$ is less than α . Thus we may accept H with significance α .

Hence we can assume that f is supported in $[a, b]$. 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 $[a, b]$.

Consider now the special case of (II.A.4) where

$$[a,b] = [0,1]$$

$$g \equiv 1$$

$$u_k(x) \triangleq \exp(2\pi i k x), \quad x \in [0,1]$$

$\{w_k(\cdot)\}_{k=-\infty}^{\infty}$ is a sequence of (complex) functions of a real positive variable h .

The estimator is given by

$$1.1) \quad \hat{f}(x) \triangleq \sum_{k=-\infty}^{\infty} w_k(h) \hat{b}_k \exp(2\pi i k x).$$

$$1.2) \quad \hat{b}_k \triangleq \frac{1}{N} \sum_{j=1}^N \exp(-2\pi i k X_j)$$

We are interested in the behavior of this estimator for large N . In particular, we will derive asymptotic estimates of $\text{var}[\hat{f}(x_0)]$ 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 $h > 0$

$$\sum_{k=-\infty}^{\infty} |w_k(h)|^2 < \infty \quad \text{and}$$

for each k , $w_k(h) = w_{-k}(h)^*$.

Then the kernel K_h defined by

$$2.1) \quad K_h(x) \triangleq \sum_{k=-\infty}^{\infty} w_k(h) \exp(2\pi i k x)$$

is a real periodic function in $L_2 [0,1]$ with period 1.

Moreover, the estimator (1) may be written as

$$2.2) \quad \hat{f}(x) = \frac{1}{N} \sum_{j=1}^N K_h(x - X_j)$$

Proof

Statement (2.1) is immediate.

For (2.2), notice

$$\begin{aligned} f(x) &= \sum_{k=-\infty}^{\infty} w_k(h) \hat{b}_k \exp(2\pi i k x) \\ &= \sum_{k=-\infty}^{\infty} \left[\frac{1}{N} \sum_{j=1}^N \exp(-2\pi i k X_j) \right] w_k(h) \exp(2\pi i k x) \\ &= \frac{1}{N} \sum_{j=1}^N \sum_{k=-\infty}^{\infty} w_k(h) \exp(2\pi i k x - 2\pi i k X_j) \\ &= \frac{1}{N} \sum_{j=1}^N K_h(x - X_j) \end{aligned}$$

Expression (2.2) has a form similar to that of the Parzen kernel estimator (see [2]). 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 hypothesis of lemma (2).

Moreover, $K_h(x) \triangleq \sum_{k=-\infty}^{\infty} w_k(h) \exp(2\pi i k x)$

satisfies

$$3.2) \quad K_h(x) \geq 0$$

$$3.3) \quad K_h(-x) = K_h(x)$$

$$3.4) \quad \int_{-1/2}^{1/2} K_h(x) dx = 1$$

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

$$3.6) \quad \int_{-1/2}^{1/2} K_h(x) x^2 dx \rightarrow 0 \quad \text{as } h \rightarrow 0.$$

3.7) Let $1/2 > \epsilon > 0$.

then

$$\frac{\int_{\epsilon}^{1/2} K_h(x) x^2 dx}{\int_0^{1/2} K_h(x) x^2 dx} \rightarrow 0 \quad \text{as } h \rightarrow 0$$

3.8) Let $1/2 > \epsilon > 0$. Then there exists $B_{\epsilon} > 0$ such that

$$\int_{\epsilon}^{\frac{1}{2}} K_h(x)^2 dx < B_{\epsilon}$$

as $h \rightarrow 0$.

We next establish some limits which will arise shortly in the asymptotic error analysis.

4) Lemma

Under the assumptions of conditions (3), we have

$$4.1) \int_{-\frac{1}{2}}^{\frac{1}{2}} K_h(x)^2 dx \rightarrow \infty \text{ as } h \rightarrow 0$$

$$4.2) \frac{\int_{-\frac{1}{2}}^{\frac{1}{2}} K_h(x) |x|^3 dx}{\int_{-\frac{1}{2}}^{\frac{1}{2}} K_h(x) x^2 dx} \rightarrow 0 \text{ as } h \rightarrow 0$$

$$4.3) \frac{\int_{-\frac{1}{2}}^{\frac{1}{2}} K_h(x)^2 x^2 dx}{\int_{-\frac{1}{2}}^{\frac{1}{2}} K_h(x)^2 dx} \rightarrow 0 \text{ as } h \rightarrow 0$$

Proof

To prove (4.1), let $\epsilon > 0$.

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} K_h(x)^2 dx = 2 \int_0^{\frac{1}{2}} K_h(x)^2 dx \geq 2 \int_0^{\epsilon} K_h(x)^2 dx.$$

It follows from the Schwartz inequality that

$$\int_0^{\epsilon} K_h(x)^2 dx \geq \frac{1}{\epsilon} \left[\int_0^{\epsilon} K_h(x) dx \right]^2.$$

Also

$$2 \int_0^\epsilon K_h(x) dx = 1 - 2 \int_\epsilon^{1/2} K_h(x) dx$$

and

$$2 \int_\epsilon^{1/2} K_h(x) dx \leq \frac{2}{\epsilon^2} \int_\epsilon^{1/2} K_h(x) x^2 dx \leq \frac{1}{\epsilon^2} \int_{-1/2}^{1/2} K_h(x) x^2 dx.$$

$$\text{So } 2 \int_0^\epsilon K_h(x) dx \geq 1 - \frac{1}{\epsilon^2} \int_{-1/2}^{1/2} K_h(x) x^2 dx$$

and finally

$$\int_{-1/2}^{1/2} K_h(x)^2 dx \geq \frac{1}{2\epsilon} \left[1 - \frac{1}{\epsilon^2} \int_{-1/2}^{1/2} K_h(x) x^2 dx \right]^2.$$

$$\text{Since } \int_{-1/2}^{1/2} K_h(x) x^2 dx \rightarrow 0 \text{ as } h \rightarrow 0,$$

we can make

$$\int_{-1/2}^{1/2} K_h(x)^2 dx \geq \frac{1}{\epsilon}. \text{ Since } \epsilon \text{ was arbitrary, this proves (4.1).}$$

For (4.2), let $\epsilon > 0$.

$$\int_{-1/2}^{1/2} K_h(x) |x|^3 dx = 2 \int_0^\epsilon K_h(x) |x|^3 dx + 2 \int_\epsilon^{1/2} K_h(x) |x|^3 dx$$

$$\leq 2\epsilon \int_0^\epsilon K_h(x) x^2 dx + \int_\epsilon^{1/2} K_h(x) x^2 dx$$

$$\leq \epsilon \int_{-1/2}^{1/2} K_h(x) x^2 dx + \int_\epsilon^{1/2} K_h(x) x^2 dx$$

$$\text{Thus } \frac{\int_{-1/2}^{1/2} K_h(x) |x|^3 dx}{\int_{-1/2}^{1/2} K_h(x) x^2 dx} \leq \frac{\epsilon \int_{-1/2}^{1/2} K_h(x) x^2 dx + \int_\epsilon^{1/2} K_h(x) x^2 dx}{\int_{-1/2}^{1/2} K_h(x) x^2 dx} =$$

$$= \epsilon + \frac{\int_{\epsilon}^{\frac{1}{2}} K_h(x) x^2 dx}{\int_{-\frac{1}{2}}^{\frac{1}{2}} K_h(x) x^2 dx}.$$

The second term $\rightarrow 0$ as $h \rightarrow 0$ by (3.7). Since ϵ is arbitrary, the result follows.

For (4.3), Let $\epsilon > 0$.

$$\begin{aligned} \int_{-\frac{1}{2}}^{\frac{1}{2}} K_h(x)^2 x^2 dx &= 2 \int_0^{\epsilon} K_h(x)^2 x^2 dx + 2 \int_{\epsilon}^{\frac{1}{2}} K_h(x)^2 x^2 dx \\ &\leq 2 \epsilon^2 \int_0^{\epsilon} K_h(x)^2 dx + 2 \int_{\epsilon}^{\frac{1}{2}} K_h(x)^2 dx \\ &\leq \epsilon^2 \int_{-\frac{1}{2}}^{\frac{1}{2}} K_h(x)^2 dx + B_{\epsilon} \quad \text{by (3.8).} \end{aligned}$$

$$\text{Thus } \frac{\int_{-\frac{1}{2}}^{\frac{1}{2}} K_h(x)^2 x^2 dx}{\int_{-\frac{1}{2}}^{\frac{1}{2}} K_h(x)^2 dx} \leq \epsilon^2 + \frac{B_{\epsilon}}{\int_{-\frac{1}{2}}^{\frac{1}{2}} K_h(x)^2 dx}.$$

Since $\int_{-\frac{1}{2}}^{\frac{1}{2}} K_h(x)^2 dx \rightarrow \infty$ as $h \rightarrow 0$ and ϵ

was arbitrary, the result follows.

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

$$c(h) \triangleq \frac{1}{2} \int_{-\frac{1}{2}}^{\frac{1}{2}} K_h(x) x^2 dx$$

$$v(h) \triangleq \int_{-\frac{1}{2}}^{\frac{1}{2}} K_h(x)^2 dx$$

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

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

Proof

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

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

With h_N chosen by (6.2),

$$\frac{v(h_N)}{N} + c(h_N)^2 = 2c(h_N)^2 \rightarrow 0 \text{ as } h_N \rightarrow 0.$$

Now we are ready to state the main theorem of this section. Although the proof follows the spirit of Rosenblatt[1], the result is original for Fourier series estimators. Before now, all error estimates for series estimators were of the integral type

$$\int_0^1 E [\hat{f}(x) - f(x)]^2 dx. \quad \text{The following result gives estimates}$$

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 $\hat{f}$, the periodic extension of f , defined by

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

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

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

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

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

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

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

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

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

Proof

$$\text{We can write } \hat{f}(x) = \frac{1}{N} \sum_{j=1}^N K_h(x - X_j)$$

where $K_h(\cdot)$ is the kernel associated with $\{w_k(\cdot)\}_{k=-\infty}^{\infty}$.

By independence of the samples,

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

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

$$\begin{aligned} E[\hat{f}(x_0)] &= \int_{-\frac{1}{2}}^{\frac{1}{2}} K_h(y) \left[\overset{\circ}{f}(x_0) + \overset{\circ}{f}'(x_0) y + \frac{1}{2} \overset{\circ}{f}''(x_0) y^2 \right. \\ &\quad \left. + \frac{y^3}{3!} \overset{\circ}{f}'''(z(y)) \right] dy \end{aligned}$$

where $x_0 < z(y) < y$ or $y < z(y) < x_0$.

By conditions (3.3) and (3.4), this reduces to

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

Now

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

and this $\rightarrow 0$ as $h \rightarrow 0$ by lemma (4). This establishes (8.4).

Again by independence of the samples,

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

Using the same extension and expansion, we have

$$\begin{aligned} \int_0^1 K_h(x_0 - y)^2 f(y) dy &= \int_{-\frac{1}{2}}^{\frac{1}{2}} K_h(y)^2 \left[f(x_0) + f'(x_0)y + \frac{1}{2} f''(z(y))y^2 \right] dy \\ &= v(h) f(x_0) + \frac{1}{2} \int_{-\frac{1}{2}}^{\frac{1}{2}} K_h(y)^2 y^2 f''(z(y)) dy. \end{aligned}$$

Thus

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

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

Now if $h = h_N \rightarrow 0$ as $N \rightarrow \infty$, 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 hypothesis 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,

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

Proof

$$\text{By lemma (6), } \frac{v(h_N)}{N} + c(h_N)^2 \rightarrow 0.$$

Thus asymptotically,

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

also goes to zero.

III. A Data-Adaptive Estimator

A. Motivation

Recall the simple form of the estimator (II.A.4)

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

with the integrated variance

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

and integrated bias squared

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

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 $\{x_1, \dots, x_N\}$ 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 $\{\exp(2\pi i k x)\}_{k=-\infty}^{\infty}$ orthonormal on $[0,1]$. Suppose that we have a transformation G satisfying

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

$$1.2) \quad G \text{ is one-to-one, onto, strictly increasing}$$

$$1.3) \quad g(x) \triangleq \frac{d}{dx} G(x) \text{ is continuous.}$$

We can then define

$$2) \quad u_k(x) \triangleq \exp(2\pi i k G(x))$$

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

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

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

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:

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

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

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 $N_1 + N_2 = N$.

4.1) Let

$$\hat{g}(x) \triangleq \sum_{k=-\infty}^{\infty} w_k(h_1) \hat{a}_k \exp(2\pi i k x)$$

$$\hat{a}_k \triangleq \frac{1}{N_1} \sum_{j=1}^{N_1} \exp(-2\pi i k x_j)$$

$$\hat{G}(x) \triangleq \int_0^x \hat{g}(y) dy$$

$$4.2) \quad \hat{f}(x) \triangleq \hat{g}(x) \sum_{k=-\infty}^{\infty} w_k(h_2) \hat{b}_k \exp(2\pi i k \hat{G}(x))$$

$$\hat{b}_k = \frac{1}{N_2} \sum_{j=N_1+1}^{N_2} \exp(-2\pi i k \hat{G}(x_j))$$

Remark

The choice of the parameters 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 III.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. There is no way to apply theory in practical choice of the parameters. As in the case of all other p.d.f. estimators, we must resort to setting the values by heuristic means.

III.C Asymptotic Error Analysis

We will now develop asymptotic error estimates for the estimator (III.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 (III.B.1), and let $\hat{f}$ be defined by (III.B.3). We can rewrite the expression (III.B.3.2) for $\hat{b}_k$ as

$$1) \quad \hat{b}_k \triangleq \frac{1}{N} \sum_{j=1}^N \exp(-2\pi i k T_j)$$

where

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

We know that the p.d.f. of the transformed random variable $T=G(X)$ is just (see [19]) $r(\cdot)$ defined by

$$3) \quad r(t) = r(G(x)) \triangleq f(x)/g(x).$$

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

$$4.1) \quad \hat{r}(t) \triangleq \sum_{k=-\infty}^{\infty} w_k(h) \hat{b}_k \exp(2\pi i k t)$$

$$4.2) \quad \hat{b}_k \triangleq \frac{1}{N} \sum_{j=1}^N \exp(-2\pi i k T_j) .$$

Since we clearly have

$$5) \quad \hat{f}(x) \triangleq g(x) \hat{r}(G(x)),$$

it follows that

$$6.1) \quad \text{var} [\hat{f}(x)] = g(x)^2 \text{var} [\hat{r}(G(x))]$$

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

Putting this together, we have the following

7) Theorem

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

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

where $t_0 = G(x_0)$.

Further, if $h_N \rightarrow 0$ as $N \rightarrow \infty$, then

$$\lim_{N \rightarrow \infty} \frac{N \text{ var}[\hat{f}(x_0)]}{V(h_N)} = f(x_0) g(x_0).$$

Proof

Applying theorem (II.B.8) to the estimator $\hat{r}$ we have

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

and

$$\lim_{N \rightarrow \infty} \frac{N \text{ var}[\hat{r}(t_0)]}{v(h_N)} = r(t_0).$$

By (3) and (5) we have

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

Thus

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

Also,

$$\frac{N \operatorname{var} [\hat{f}(x_0)]}{v(h_N)} = g(x_0)^2 \frac{N \operatorname{var} [\hat{r}(t_0)]}{v(h_N)}$$

$$\text{Thus } \lim_{N \rightarrow \infty} \frac{N \operatorname{var} [\hat{f}(x_0)]}{v(h_N)} = g(x_0)^2 r(t_0) = f(x_0)g(x_0) .$$

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 p.d.f.'s.

Define

$$G(x) \triangleq \int_0^x g(y)dy$$

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

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

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

Then

$$\begin{aligned} r''(t_0) = \frac{d^2}{dt^2} \Big|_{t=t_0} r(t) &= \frac{1}{g(x_0)^5} \left\{ g(x_0)^2 f''(x_0) - g(x_0) f(x_0) g''(x_0) \right. \\ &\quad \left. + 3f(x_0) [g'(x_0)]^2 - 3g(x_0) f'(x_0) g'(x_0) \right\} \end{aligned}$$

Proof

Define momentarily $h(x) \triangleq f(x)/g(x)$;

$$\text{so } r(G(x_0)) = h(x_0).$$

$$f(x_0) = g(x_0)h(x_0).$$

$$f'(x_0) = g'(x_0)h(x_0) + g(x_0)h'(x_0)$$

$$f''(x_0) = g''(x_0)h(x_0) + 2g'(x_0)h'(x_0) + g(x_0)h''(x_0)$$

Solving for $h'(x_0)$ and $h''(x_0)$ gives

$$h'(x_0) = [f'(x_0) - g'(x_0)h(x_0)] / g(x_0)$$

$$h''(x_0) = [f''(x_0) - g''(x_0)h(x_0) - 2g'(x_0)h'(x_0)] / g(x_0)$$

$$= \frac{f''(x_0) - g''(x_0)h(x_0) - 2g'(x_0)[f'(x_0) - g'(x_0)h(x_0)]}{g(x_0)^2}$$

$$= \frac{g(x_0)f''(x_0) - g''(x_0)f(x_0) - 2g'(x_0)f'(x_0) + 2g'(x_0)^2h(x_0)}{g(x_0)^2} .$$

$$\text{Now } h(x_0) = r(G(x_0))$$

$$h'(x_0) = g(x_0)r'(G(x_0))$$

$$h''(x_0) = g'(x_0)r'(G(x_0)) + g(x_0)^2r''(G(x_0)) .$$

$$\text{Thus } r''(t_0) = r''(G(x_0))$$

$$= \frac{h''(x_0) - g'(x_0)r'(G(x_0))}{g(x_0)^2}$$

$$= \frac{h''(x_0) - g'(x_0)h'(x_0) / g(x_0)}{g(x_0)^2}$$

$$= \frac{g(x_0)h''(x_0) - g'(x_0)h'(x_0)}{g(x_0)^3}$$

$$= \frac{g(x_0)h''(x_0) - g'(x_0)[f'(x_0) - g'(x_0)h(x_0)]}{g(x_0)^3} / g(x_0)$$

$$= \frac{g(x_0)^2 h''(x_0) - g'(x_0)f'(x_0) + g'(x_0)^2 h(x_0)}{g(x_0)^4}$$

And finally,

$$r''(t_0) = \frac{g(x_0)f''(x_0) - g''(x_0)f(x_0) - 3g'(x_0)f'(x_0) + 3g'(x_0)^2 h(x_0)}{g(x_0)^4}$$

$$= \frac{g(x_0)^2 f''(x_0) - g''(x_0)g(x_0)f(x_0) - 3g'(x_0)f'(x_0)g(x_0) + 3g'(x_0)^2 f(x_0)}{g(x_0)^5}$$

Next we establish a bound on $r''(t_0)$ under the assumption that $g \approx f$.

9) Lemma

With the same hypothesis of lemma (8), suppose further that we have

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

$$\text{Let } B(f, x_0) = \max \{1, f(x_0), |f'(x_0)|, |f''(x_0)|\}.$$

Then at $t_0 = G(x_0)$ we have

$$|r''(t_0)| \leq \frac{24 A B(f, x_0)^2}{g(x_0)^5}$$

Proof

For convenience, we will write f for $f(x_0)$, etc.

We have by lemma (8),

$$\begin{aligned} r''(t_0) &= \frac{1}{g^5} \left\{ g^2 f'' - g g'' f + 3g'^2 f - 3g g' f' \right\} \\ &= \frac{1}{g^5} \left\{ g [g f'' - f g''] + 3g' [f g' - g f'] \right\}. \end{aligned}$$

We will make use of the easily verified inequality

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

First,

$$\begin{aligned} |g f'' - f g''| &\leq \frac{1}{2} |g-f| |f''+g''| + \frac{1}{2} |g+f| |f''-g''| \\ &\leq \frac{1}{2} A (2B+A) + \frac{1}{2} (2B+A) A \leq 3AB \end{aligned}$$

Second,

$$\begin{aligned} |f g' - g f'| &\leq \frac{1}{2} |f-g| |f'+g'| + \frac{1}{2} |f+g| |f'-g'| \\ &\leq \frac{1}{2} A (2B+A) + \frac{1}{2} (2B+A) A \leq 3AB \end{aligned}$$

Moreover,

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

$$g' = f' + g' - f' \leq B + A \leq 2B$$

Thus

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

$$\leq \frac{24AB^2}{g^5} .$$

We now collect what we have so far into a theorem giving asymptotic error estimates under the assumption $g \approx f$.

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 (II.B.3).

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

Let $g(x) \triangleq \frac{d}{dx} G(x)$, $\hat{f}$ be defined by (III.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

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

Then we have

$$10.4) \quad \lim_{h \rightarrow 0} \frac{E[\hat{f}(x_0)] - f(x_0)}{c(h)} \leq 24 \frac{AB(f, x_0)^2}{f(x_0)^4(1-p)^4},$$

where

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

Furthermore, if $h_N \rightarrow 0$ as $N \rightarrow \infty$, then

$$10.5) \lim_{N \rightarrow \infty} \left| \frac{N \operatorname{var} [\hat{f}(x_0)]}{v(h_N)} - f(x_0)^2 \right| \leq A f(x_0) .$$

Proof

By theorem (7) we have

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

and

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

By lemma (9) we have

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

Thus

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

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

we obtain

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

which is (10.4).

(10.5) follows immediately since

$$|f(x_0)g(x_0) - f(x_0)^2| \leq A f(x_0) .$$

Now let us return to the adaptive estimator (III.B.4). We know that $\hat{g}(x_0)$ is a consistent estimator for $f(x_0)$, by theorem II.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

11.1) For $k=0,1,2$,

$$v_k(h) \triangleq \int_{-\frac{1}{2}}^{\frac{1}{2}} K_h^{(k)}(x)^2 dx ,$$

where $K_h^{(k)}(x) \triangleq \frac{d^k}{dx^k} K_h(x)$.

Note $v_0(h) \equiv v(h)$.

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

12) Theorem

Let $\hat{g}$ be defined by (III.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 $k = 0,1,2$

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

Choose $h_1 = h_{N_1}$ to satisfy

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

Then for $x_0 \in (0,1)$, $E[\hat{g}^{(k)}(x_0) - f^{(k)}(x_0)]^2 \rightarrow 0$

as $N_1 \rightarrow \infty$.

Proof

We can write $\hat{g}(x) = \frac{1}{N_1} \sum_{j=1}^{N_1} K_{h_1}(x - \bar{X}_j)$.

Since $K_{h_1} \in C^2[0,1]$,

$$\hat{g}^{(k)}(x_0) = \frac{1}{N} \sum_{j=1}^N K_{h_1}^{(k)}(x_0 - \bar{X}_j) \text{ exists.}$$

Now by integration by parts, we get

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

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

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

$$E[\hat{g}^{(k)}(x_0)] = \int_0^1 K_{h_1}(x_0 - y) f^{(k)}(y) dy =$$

$$\begin{aligned}
&= \int_{-\frac{1}{2}}^{\frac{1}{2}} K_{h_1}(y) \left[f^{(k)}(x_0) + f^{(k+1)}(x_0)y + \frac{f^{(k+2)}(x_0)y^2}{2} \right. \\
&\quad \left. + \frac{f^{(k+3)}(z(y))y^3}{3!} \right] dy \\
&= f^{(k)}(x_0) + f^{(k+2)}(x_0) c(h_1) + \int_{-\frac{1}{2}}^{\frac{1}{2}} K_{h_1}(y) \frac{y^3}{3!} f^{(k+3)}(z(y)) dy.
\end{aligned}$$

Thus we have an estimate for the bias

$$\begin{aligned}
|E[\hat{g}^{(k)}(x_0)] - f^{(k)}(x_0)| &\leq c(h_1) \left[|f^{(k+2)}(x_0)| \right. \\
&\quad \left. + \sup_{x \in (0,1)} |f^{(k+3)}(x)| \right] \triangleq c(h_1) a_k
\end{aligned}$$

$$\text{since } \int_{-\frac{1}{2}}^{\frac{1}{2}} K_{h_1}(y) \frac{y^3}{3!} dy \leq \int_{-\frac{1}{2}}^{\frac{1}{2}} K_{h_1}(y) y^2 dy = c(h_1).$$

For the variance we have

$$\begin{aligned}
\text{var} [\hat{g}^{(k)}(x_0)] &= \frac{1}{N_1} \text{var} [K_{h_1}^{(k)}(x_0 - \bar{X})] \\
&\leq \frac{1}{N_1} \int_0^1 K_{h_1}^{(k)}(x_0 - y)^2 f(y) dy.
\end{aligned}$$

Again the Taylor expansion with remainder yields

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

So

$$\begin{aligned} \text{var} [\hat{g}^{(k)}(x_0)] &\leq \frac{1}{N_1} \left\{ v_k(h_1) f(x_0) + \sup_{x \in (0,1)} |f''(x)| \int_{-\frac{1}{2}}^{\frac{1}{2}} K_{h_1}^{(k)}(y) y^2 dy \right\} \\ &\leq \frac{1}{N_1} \left\{ v_k(h_1) f(x_0) + \sup_{x \in (0,1)} |f''(x)| v_k(h_1) \right\} \\ &\cong \frac{v_k(h_1)}{N_1} b \end{aligned}$$

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

$$\begin{aligned} E [\hat{g}^{(k)}(x_0) - f^{(k)}(x_0)]^2 &= \left(E [\hat{g}^{(k)}(x_0)] - f^{(k)}(x_0) \right)^2 + \text{var} [\hat{g}^{(k)}(x_0)] \\ &\leq c(h_{N_1})^2 a_k^2 + \frac{v_k(h_{N_1})}{N_1} b \rightarrow 0 \end{aligned}$$

as $N_1 \rightarrow \infty$.

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 (III.B.4)

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

$\{h_{N_2}\}$ 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 $\epsilon > 0$, $1 > \delta > 0$.

Then there exists N_1 such that

$$P_r \left\{ \lim_{N_2 \rightarrow \infty} \left| \frac{E[\hat{f}(x_0)] - f(x_0)}{c(h_{N_2})} \right| \geq \epsilon \right\} \leq \delta$$

and

$$P_r \left\{ \lim_{N_2 \rightarrow \infty} \left| \frac{N_2 \text{var}[\hat{f}(x_0)]}{v(h_{N_2})} - f(x_0)^2 \right| \geq \epsilon \right\} \leq \delta.$$

Proof

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

$$\frac{24 A B(f, x_0)^2}{f(x_0)^4 (1 - \frac{1}{2})^4} \cdot$$

Then by theorem (10), if

$$13.6) \quad |\hat{g}^{(k)}(x_0) - f^{(k)}(x_0)| \leq A \quad \text{for } k=0,1,2 \text{ then}$$

$$13.7) \quad \lim_{N_2 \rightarrow \infty} \left| \frac{E[\hat{f}(x_0)] - f(x_0)}{c(h_{N_2})} \right| \leq \epsilon$$

and

$$13.8) \quad \lim_{N_2 \rightarrow \infty} \left| \frac{N_2 \text{ var}[\hat{f}(x_0)]}{v(h_{N_2})} - f(x_0)^2 \right| \leq \epsilon.$$

Recall that by Tchebichev's inequality for a random variable Y we have

$$P_r \{ |Y| \geq A \} \leq E[Y]^2 / A^2.$$

Now by theorem (12) we have

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

Thus there is some N_1 such that

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

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

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 (II.B.1) and by $\hat{f}_2$ the adaptive estimator (III.B.4).

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

Theorem (13) gives the analogous result

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

The factor of proportionality ϵ can be made as small as desired, such as $\epsilon \ll |f''(x_0)|$, by reserving enough samples $X_1, \dots, X_{N_1}$ in the

first pass. Now if the ratio $\frac{c(h_N)}{c(h_{N_2})} = \frac{c(h_N)}{c(h_{N-N_1})} \rightarrow 1$ as $N \rightarrow \infty$, N_1 fixed,

then the asymptotic bias of $\hat{f}_2(x_0)$ is smaller than that of

$$\hat{f}_1(x_0).$$

III.D. An Optimality Property of the Fourier Basis

The choice of the Fourier basis simplifies analysis and implementation of the estimation (III.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 (III.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 (III.B.3) becomes

$$1.1) \quad \hat{f}(x) \triangleq g(x) \sum_{k=0}^{\infty} w_k(h) \hat{b}_k v_k(G(x))$$

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

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

$$2) \quad r(t) = r(G(x)) = f(x)/g(x) \\ \text{of the random variable } T = G(X).$$

The corresponding expression is

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

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

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 $|r(t)| \leq 1$ 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

$$5) \quad \hat{r}(t) = \sum_{k=0}^n b_k v_k(t)$$

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

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

where we assume r may be expanded

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

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.

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

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

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 $S(v_0, v_1, \dots, v_n)$, the space spanned by $v_0, v_1, \dots, v_n$. The maximum error for r in W_2 is the quantity $E_{S(v_0, v_1, \dots, v_n)}(C)$.

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 $\{1, \sin 2\pi t, \cos 2\pi t, \dots, \sin 2\pi n t, \cos 2\pi n t\}$ 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 non-negativity 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.

IV. Computer Simulations

In chapter III 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.

In the following simulations we consider a mixture of two Gaussians

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

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) pseudo-random number generator.

This p.d.f. was chosen as a test case because it has two closely spaced modes separated by a shallow valley (see figure IV.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 chapter III, we partitioned the sample set $\{x_1, \dots, x_N\}$ into two parts $\{x_1, \dots, x_{N_1}\}$, $\{x_{N_1+1}, \dots, x_N\}$. 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 $\{X_1, \dots, X_N\}$ ($N=100$), the estimator was implemented as follows:

$$2.1) \quad \hat{g}(x) \triangleq \sum_{k=0}^{20} (1-h)^k \hat{a}_k \cos 2\pi kx$$

$$2.2) \quad \hat{a}_k \triangleq \frac{2}{N} \sum_{j=1}^N \cos 2\pi kX_j \quad (k \geq 1) \\ \hat{a}_0 \triangleq 1$$

$$2.3) \quad \hat{G}(x) \triangleq \int_0^x \hat{g}(y) dy$$

$$2.4) \quad \hat{f}_2(x) \triangleq \hat{g}(x) \sum_{k=0}^5 \hat{b}_k \cos(2\pi k\hat{G}(x))$$

$$2.5) \quad \hat{b}_k \triangleq \frac{2}{N} \sum_{j=1}^N \cos(2\pi k\hat{G}(X_j)) \quad (k \geq 1) \\ \hat{b}_0 \triangleq 1$$

The expansions employ only cosines in order to simplify the computer program. Also, the summation in (2.1) is limited to 20 terms.

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

$$3.1) \quad \hat{f}_1(x) \triangleq \sum_{k=0}^n \hat{c}_k \cos 2\pi kx$$

$$3.2) \quad \hat{c}_k \triangleq \frac{2}{N} \sum_{j=1}^N \cos 2\pi kX_j \quad (k \geq 1) \\ \hat{c}_0 \triangleq 1$$

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 IV.2 through IV.7. Second, the 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 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 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 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, h equals 0.25. Now $\hat{f}_2$ does a very good job of resolving the modes, and the negative tail is almost eliminated.

Clearly, figure 4 is a much better estimate than figure 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 5. (This is in fact the same estimate as in figure 1.) Now as we increase to $n=7$ (figure 6) and to $n=10$ (figure 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 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 i^{th} sample set ($i=1, \dots, 25$), estimates $\hat{f}_{1,i}$ and $\hat{f}_{2,i}$ were obtained. For each estimate, the integrated square error

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

was computed by numerical integration. These errors are tabulated in table IV.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.

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 $e_{2i} < e_{1i}$ occurs 22 times, with the reverse occurring 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 IV.2. Applying the sign test for the 50 trials yields 34 occurrences of $e_{2i} < e_{1i}$ and 16 occurrences of $e_{2i} \geq e_{1i}$. 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 p.d.f. estimator is to resolve and estimate the location of the modes of the p.d.f. Thus, let us define another error measure m equal to the sum of the squared distances from the true modes (located 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 IV.3 for the 25 trials. The average $\bar{m}_2 = 0.31$ which is substantially less than $\bar{m}_1 = 1.04$. 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 IV.3 yields a significance of 0.02.

V. Summary and Conclusions

This thesis has reviewed the p.d.f. estimation problem and some approaches to its solution. We have looked in detail at the orthogonal-series type of estimator and at its asymptotic error analysis. The main contribution 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 pointed out an "optimality" property of the Fourier basis. 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.

Figure IV.1
True density f

Figure IV.2
Adaptive
Estimator
($h = 1.0$)

Figure IV.3
Adaptive Estimator
($h = 0.4$)

Figure IV.4
Adaptive Estimator
($h = 0.25$)

Figure IV. 5
Fourier Series
Estimator
($n = 5$)

Figure IV.6
Fourier Series
Estimator
($n = 7$)

Figure IV. 7
Fourier Series
Estimator
($n = 10$)

TABLE IV.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 Devia - tion	.0043	.0028	.0055	.0039

TABLE IV.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 IV.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

REFERENCES

- 1) M. Rosenblatt, "Remarks on some nonparametric estimates of a density function," Annals of Math. Stat., vol. 27, pp.832-837 (1956).
- 2) E. Parzen, "On estimation of a probability density function and mode," Annals of Math. Stat., vol. 33, no.3, pp 1065-1076 (Sept. 1962).
- 3) G.S. Watson, M.R. Leadbetter, "On the estimation of the probability density, I," Annals of Math. Stat., vol. 34, pp 480-491 (1963).
- 4) J.O. Bennett, R.J.P. de Figueiredo, J.R. Thompson, "Classification by means of B-spline potential functions with application to remote sensing," presented at Sixth Southeastern Symposium on System Theory, Baton Rouge, La. (Sponsored by I.E.E.E.).
- 5) K.B. Davis, "Mean square error properties of density estimates," Annals of Statistics, vol.3, no. 4, pp 1025-1030 (1975).
- 6) D.W. Scott, R.A. Tapia, J.R. Thompson, "Kernel density estimation revisited," J. Nonlinear Analysis, Theory, Methods, and Applications, vol. 1 no. 4, pp. 339-372 (1977).
- 7) T.J. Wagner, "Strong consistency of a nonparametric estimate of a density function," I.E.E.E. Trans. Systems, Man, and Cybernetics, vol.3, pp. 289-290 (1973).
- 8) L.P. Devroye, T.J. Wagner, "The strong uniform consistency of nearest neighbor density estimates," The Annals of Statistics, vol. 5, no. 3, pp 536-540 (1977).
- 9) R. Kronmal, M. Tarter, "The estimation of probability densities and cumulatives by Fourier series methods," J. American Stat. Association, vol. 63, pp. 925-952 (1968).
- 10) N.N. Cencov, "Evaluation of an unknown distribution density from observations," Soviet Math, vol. 3, 1559-1562 (1962).
- 11) J. van Ryzin, "Bayes risk consistency of classification procedures using density estimation," Sankhya Ser. A., vol. 28, pp261-270 (1966).
- 12) S.C. Schwartz, "Estimation of probability density by an orthogonal series," Annals of Math. Stat., vol. 38, pp. 1261-1265 (1967).

- 13) H.D. Brunk, "Univariate density estimation by orthogonal series," Tech. Report no. 51, Oregon State University (1976).
- 14) G.F. de Montricher, R.A. Tapia, J.R. Thompson, "Nonparametric maximum likelihood estimation of probability densities by penalty function methods," The Annals of Statistics, vol. 3, no. 6, pp 1329-1348, (1975).
- 15) I.J. Good, R.A. Gaskins, "Global nonparametric estimation of probability densities," Virginia J. Science, vol. 23, p.171 (1972).
- 16) A. Wragg, D.C. Dowson, "Fitting continuous probability density functions over $[0, \infty)$ using information theory ideas," I.E.E.E. Trans. on Information Theory, pp 226-230, (Mar. 1970)
- 17) G. Wahba, "Optimal smoothing of density estimates," Tech Report no. 469, University of Wisconsin - Madison (Oct., 1976).
- 18) P. Whittle, "On the smoothing of probability density functions," J.R. Statistic. Soc., (B) 20, pp 334-343 (1958).
- 19) G.G. Roussas, A First Course in Mathematical Statistics, Addison-Wesley, (1973).
- 20) G.G. Lorentz, Approximation of Functions, Holt, Rinehart, and Winston, (1966).
- 21) D.W. Scott, R.A. Tapia, J.R. Thompson, "An algorithm for nonparametric density estimation," Computer Science and Statistics: Ninth Annual Symposium on the Interface, (May 1976).
- 22) J.V. Bradley, Distribution-Free Statistical Tests, Prentice-Hall (1968).