DEPARTMENT OF STATISTICS

University of Wisconsin
Department of Statistics

TECHNICAL REPORT NO. 324

December 1972

OPTIMAL CONVERGENCE PROPERTIES OF VARIABLE KNOT, KERNEL, AND ORTHOGONAL SERIES METHODS FOR DENSITY ESTIMATION

By

Grace Wahba

Typist: Jacquelyn R. Forer

This research was supported by the Air Force Office of Scientific Research under grant AFOSR 72-2363.

Optimal Convergence Properties of Variable Knot, Kernel, and Orthogonal Series Methods for Density Estimation

by

Grace Wahba

Abstract

Let f be a density function satisfying

$$\left[\int_{-\infty}^{\infty}(f^{(m)}(\xi))^p\right]^{1/p} \leq M < \infty, \quad m=1,2,\ldots, \quad p\geq 1,$$

and let $\hat{f}_n(x)$ be an estimate of $f(x)$ based on n independent observations from the density f. If $\hat{f}_n$ is the polynomial algorithm for density estimation then it is known that

$$E(f(x)-\hat{f}_n(x))^2 = 0(n^{-(2m-2/p)/(2m+1-2/p)})$$

for $m = 1, 2, \ldots$ and $p=2$, $p=\infty$. If $\hat{f}_n$ is an appropriate Parzen kernel type estimate, the above result is known to hold for $m = 1, 2, \ldots,$ and $p=\infty$. By applying a theorem of Farrell, it is shown that these are the best obtainable rates for $m = 1, 2, \ldots,$ and $p\geq 1$. These optimal convergence rates are then shown to hold for the polynomial and kernel estimates for all $p\geq 1$. The optimal rates are also shown to hold for the Kronmal-Tartar orthogonal series estimate for $m = 1, 2, \ldots,$ and $1<p\leq 2$, and for the ordinary histogram estimate with variable "bins" for $m = 1$, $p\geq 1$. Upper bounds for the constants covered by the "0" are exhibited.

1. Introduction

In a recent paper [6] estimates of a density at a point were studied, where the exact form of the estimate is chosen on the basis of assumed smoothness properties of the true, but unknown density. The choice of estimate depends m, p, and M, where it is assumed that $f^{(m)} \varepsilon \mathcal{L}_p$, m=1,2,... and

$$||f^{(m)}||_p = \left| \int_{-\infty}^{\infty} (f^{(m)}(\xi))^p d\xi \right|^{1/p} \leq M , \quad p \geq 1 ,$$

$$||f^{(m)}||_\infty = \sup_\xi f(\xi) \leq M, \quad p=\infty$$

If $\hat{f}_n(x)$ is the estimate of $f(x)$ based on n independent observations from the density f, it was shown, for the estimates in [6][1], that

$$E(f(x)-\hat{f}_n(x))^2 = O(n^{-\phi}) \qquad (1.1)$$

where

---

[1] Only p=2 and p=∞ are considered in [6].

$$\phi = \phi(m,p) = (2m-2/p)/(2m+1-2/p).$$

Subsequently, the author was led to conjecture that this is the best convergence rate possible for $f\varepsilon W_p^{(m)}$ ($W_p^{(m)}$ is the Sobolev space of functions whose first m-1 derivatives are absolutely continuous and whose mth derivative is in $\mathcal{L}_p$.) Indeed, it is known that the Parzen kernel estimates [4] achieve this rate for $m=1,2,\ldots,$ and $p=\infty$. Serendipitously, a paper by Farrell [1] appeared shortly thereafter with a theorem concerning the best available rates, which allows the question to be answered.

The purpose of this paper is twofold. First, it is shown as a consequence of Farrell's theorem that, if $f\varepsilon W_p^{(m)}$, then the best obtainable rate, good for all f with $||f^{(m)}||_p \le M$, is, in fact $n^{-\phi}$.

Secondly, several types of density estimates achieving the optimal rate are compared on the basis of mean square error. The result (1.1) in [6] for the polynomial algorithm for density estimation is extended to all $p \geq 1$. Next, it is shown that the Parzen kernel estimates achieve the rate $n^{-\phi}$, for $m=1,2,\ldots,$ $p \geq 1$. Then, it is shown that the Kronmal-Tartar orthogonal series method [3] achieves this rate for $m=1,2,\ldots,$ and $1<p\leq 2$. (The result is probably not true for arbitrary orthogonal series, however.) Finally, it is shown that, for $m=1$, $p \geq 1$, the ordinary histogram method achieves the best obtainable rate if the size of the "bins" is allowed to vary appropriately with n.

In each method we will have

$$E(f(x)-\hat{f}_n(x))^2 = Dn^{-\phi} \text{ plus negligible terms,}$$

where

$$D = \theta(||f^{(m)}||_p^2 \ AB^{2m-2/p})^{1/(2m+1-2/p)}$$

where

$$\theta = \theta(m,p) = \frac{(2m+1-2/p)}{(2m-2/p)^{(2m-2/p)}}$$

$$A = A(m,p,\lambda,\Lambda)$$

$$B = B(m,p,\lambda,\Lambda)$$

where λ and Λ satisfy

$$\lambda \leq f(x)$$

$$\sup_{\xi} f(\xi) \leq \Lambda$$

and A and B depend on the method. Thus a comparison between methods on the basis of mean square error consists of looking at $AB^{2m-2/p}$.

For a recent survey of density estimation methods, and Monte Carlo comparisons of methods, see [7] and [8].

2. Farrell's theorem for $f\varepsilon W_p^{(m)}$

The function f is said to be in Farrell's class $C_{k\eta}$ if

1. $f^{(\nu)}$ continuous, $\nu=0,1,\ldots,k$
2. there exists a polynomial s of degree k such that, for all x, $|f(x)-s(x)| \leq 2(k!)^{-1} x^k \eta'(x)$,

where, for our purposes we take $\eta(x) = Kx^\tau$ for some positive constants K and τ. (See [1] p. 172).

We first show that $f\varepsilon W_p^{(m)}$ with $||f^{(m)}||_p \leq M$ implies that $f\varepsilon C_{m-1,\eta}$ with $\eta(x) = K_0 Mx^\tau$, where $\tau = 2-1/p$, $K_0 = \{2\tau[(m-1)q+1]^{1/q}\}^{-1}$, $\frac{1}{q}+\frac{1}{p}=1$ and $[(m-1)\infty+1]^{\frac{1}{\infty}}$ is interpreted as 1. This follows upon taking $s(x) = \sum_{\nu=0}^{m-1} f^{(\nu)}(0)\frac{x^\nu}{\nu!}$ since, with $\frac{1}{p}+\frac{1}{q}=1$, using a Hölder inequality on Taylor's formula with remainder,

$$
\begin{aligned}
|f(x) - \sum_{\nu=0}^{m-1} f^{(\nu)}(0)\frac{x^\nu}{\nu!}| &\leq |\int_0^x \frac{(x-u)^{m-1}}{(m-1)!} f^{(m)}(u)du| \\
&\leq \frac{1}{(m-1)!}\left[\int_0^{|x|} (|x|-u)^{(m-1)q} du\right]^{1/q} \\
&\quad \cdot \left[\int_0^{|x|} |f^{(m)}(u)|^p du\right]^{1/p} \\
&\leq 2\frac{|x|^{m-1}}{(m-1)!} \cdot K\tau|x|^{\tau-1}
\end{aligned}
$$

with

$$K = \left\{ 2\ [(m-1)q + 1]^{1/q} \right\}^{-1} \cdot \left[\int_{-\infty}^{\infty} |f^{(m)}(u)|^p du \right]^{1/p}, \quad p \neq 1$$

$$= \frac{1}{2} \sup_u |f^{(m)}(u)|, \qquad p = 1.$$

Let $t_1, t_2, \ldots, t_n$ be the order statistics from a random sample of size n from a population with unknown density f. We will now apply the following

Theorem 2.1 (Farrell, [1], Thm. 1.1). Suppose $\{a_n, n \geq 1\}$ is a sequence of non-negative real numbers such that

$$(2.1) \quad \liminf_{n\to\infty} \inf_{f\varepsilon C_{m-1,\eta}} P_f(|\gamma(t_1,t_2,\ldots,t_n) - f(0)| \leq a_n) = 1.$$

with $\eta(x) = Kx^{(2-1/p)}$, (and where γ_n is an estimate of $f(0)$ based on $t_1,t_2,\ldots,t_n$). Then

$$(2.2) \qquad \liminf_{n\to\infty} n^{(2m-2/p)/(2m+1-2/p)} a_n^2 = \infty.$$

Let $Y_n = |\gamma_n(t_1,t_2,\ldots,t_n) - f(0)|$ and let $\phi = (2m-2/p)/(2m+1-2/p)$. By Tchebycheff's inequality,

$$P(Y_n \leq a_n) \geq 1 - \frac{EY_n^2}{a_n^2}.$$

Thus, if $EY_n^2 = b_n O(n^{-\phi})$ for any sequence b_n tending to 0, then, upon taking $a_n = O(n^{-\phi/2})$, we have that (2.1) is satisfied but (2.2) is not. Thus, no sequence of estimates of $f(0)$ can be found, that has a better convergence rate than that of (1.1) for all $f\varepsilon W_p^{(m)}$, with $||f^{(m)}||_p \leq M$.

3. Convergence Properties of the Polynomial Algorithm for Density Estimation

Let $F_n(x)$ be $n/(n+1)$ times the sample cumulative distribution function, based on $t_1, t_2, \ldots, t_n$, and let $k_n << n$ be an appropriately chosen sequence depending on n. An estimate for $f(x)$ may be obtained by interpolating F_n at t_{ik_n}, $i=1,2,\ldots,[\frac{n}{k_n}]$, by a smooth function, call it $\hat{F}_n$, and letting the density estimate $\hat{f}_n$ be given by

$$\hat{f}_n(x) = \frac{d}{dx}\hat{F}_n(x).$$

We call this class of methods "variable knot interpolating methods". The "knots" are the points of interpolation. The only examples of this method that we know of in the literature are the polynomial algorithm [6] and Van Ryzin's histogram method [5], of which [6] is a generalization. The method described in [6] uses local polynomial interpolation and is as follows:

Suppose $f \varepsilon W_p^{(m)}$. Let ℓ be the greatest integer in $(n-1)/k_n$. Let

$$\begin{aligned}
\hat{f}_n(x) &= 0, \quad x < t_{2k_n} \\
&= \frac{d}{dx}\hat{F}_n(x), \quad t_{2k_n} \leq x < t_{(\ell-m+1)k_n} \\
&= 0, \quad t_{(\ell-m+1)k_n} \leq x
\end{aligned}$$

where $\hat{F}_n(x)$ is defined as follows:

For $m=1$,

$$\hat{F}_n(x) = F_n(t_{ik_n}) + x \frac{F_n(t_{(i+1)k_n}) - F_n(t_{ik_n})}{t_{(i+1)k_n} - t_{ik_n}} ,$$

$$t_{ik_n} \leqq x < t_{(i+1)k_n}; \quad i=2,3,\ldots,\ell-1.$$

For $m \geqq 2$, let $\hat{F}_{n,i}(x)$, $i=1,2,\ldots,\ell-m-1$, be the mth degree polynomial which interpolates to $F_n(x)$ at the $m+1$ points $x = t_{ik_n}, t_{(i+1)k_n}, \ldots, t_{(i+m)k_n}$. For $x \varepsilon [t_{(i+1)k_n}, t_{(i+2)k_n})$, define $\hat{F}_n(x)$ to coincide with $\hat{F}_{n,i}(x)$, $i=1,2,\ldots,\ell-m-1$.

More explicitly, for any given numbers $x_0 < x_1 < \ldots < x_m$, let $\ell_\nu(x) = \ell_\nu(x;x_0,x_1,\ldots,x_m)$ be the mth degree polynomial with $\ell_\nu(x_\mu) = 1$, $\nu=\mu=0,1,\ldots,m$, $\ell_\nu(x_\mu) = 0$, $\mu \neq \nu$. Let $\ell_{i,\nu}(x) = \ell_\nu(x;t_{ik_n}, t_{(i+1)k_n}, \ldots, t_{(i+m)k_n})$. Then

$$(3.1) \quad f_n(x) = \frac{d}{dx} \sum_{\nu=0}^{m} \ell_{i,\nu}(x) \frac{k_n}{(n+1)} , \quad i = i(x), \quad x \varepsilon [t_{2k_n}, t_{(\ell-m+1)k_n})$$

$$= 0 \quad \text{otherwise}$$

where $i(x)$ is defined for $x \varepsilon [t_{2k_n}, t_{(\ell-m+1)k_n})$ as that value i which satisfies

$$t_{(i+1)k_n} \leqq x < t_{(i+2)k_n}$$

for $m \geqq 2$, and by that value i which satisfies

$$t_{ik_n} \leqq x < t_{(i+1)k_n}$$

when $m=1$.

Thus,

$$f(x) - \hat{f}_n(x) = \left\{ f(x) - \sum_{\nu=1}^{m} \frac{d}{dx} \ell_{i,\nu}(x) \int_{t_{ik_n}}^{t_{(i+\nu)k_n}} f(\xi)d\xi \right\}$$

$$+ \left\{ \sum_{\nu=1}^{m} \frac{d}{dx} \ell_{i,\nu}(x) \left(F(t_{(i+\nu)k_n}) - F(t_{ik_n}) - \frac{\nu k_n}{n+1} \right) \right\}, \quad x\varepsilon[t_{2k_n}, t_{(\ell-m+1)k_n}),$$

$$= f(x) \qquad x \notin [t_{2k_n}, t_{(\ell-m+1)k_n}) .$$

It is shown in [6] that

$$\text{(3.2)} \quad E(f(x)-\hat{f}_n(x))^2 \leq 2E\left\{ f(x) - \sum_{\nu=1}^{m} \frac{d}{dx} \ell_{i,\nu}(x) \int_{t_{ik_n}}^{t_{(i+\nu)k_n}} f(\xi)d\xi \right\}^2$$

$$+ 2E\left\{ \sum_{\nu=1}^{m} \frac{d}{dx} \ell_{i,\nu}(x) \left(F(t_{(i+\nu)k_n}) - F(t_{ik_n}) - \frac{\nu k_n}{n+1} \right) \right\}^2$$

$$+ \text{ negligible terms}$$

where $i = i(x)$ is a random integer, and, if $x \notin [t_{2k_n}, t_{(\ell-m+1)k_n})$, then $\ell_{i,\nu}(x)$ is defined as 0. The first term on the right is the bias term, the second, the variance.

Letting $F(x) = \int_{-\infty}^{x} f(u)du$, the variance part is due to the error in approximating $F(t_{ik_n})$ by $\hat{F}(t_{ik_n}) = \frac{ik_n}{n+1}$. Under some additional tail conditions to be stated later, it is shown in [6], that the variance term has the bound

$$(3.3a)\quad 2E\left\{\sum_{\nu=1}^{m} \frac{d}{dx}\ell_{i,\nu}(x)\left(F(t_{(i+\nu)k_n}) - F(t_{ik_n}) - \frac{\nu k_n}{n+1}\right)\right\}^2 \le B_1 \frac{1}{k_n}\left(1+O(\frac{1}{k_n}) + O(\frac{k_n}{n})\right)$$

where

$$(3.3b)\qquad B_1 = 2m^{2m+3\frac{1}{4}} \frac{\Lambda^{2m}}{\lambda^{2(m-1)}} 3^{\frac{1}{2}} \;[2]$$

(We remark that B_1 is probably not the best constant.)

the data).

The bias part is due to the error committed in approximating $f(x)$ from values of $F(t_{ik_n})$, $i=1,2,\ldots,[\frac{n}{k_n}]$. The mth degree polynomial $\tilde{F}(x)$ interpolating to $F(x)$ at $x_0, x_1, \ldots, x_m$ is given by

$$\tilde{F}(x) = \sum_{\nu=0}^{m} \ell_\nu(x) \int_{-\infty}^{\nu} f(\xi)d\xi$$

and its derivative $\tilde{f}(x) = \frac{d}{dx}\tilde{F}(x)$ is given by

$$\tilde{f}(x) = \sum_{\nu=0}^{m} \frac{d}{dx}\ell_\nu(x) \int_{-\infty}^{x_\nu} f(\xi)d\xi = \sum_{\nu=1}^{m} \frac{d}{dx}\ell_\nu(x) \int_{x_0}^{x_\nu} f(\xi)d\xi.$$

---

[2] The factor 2 was erroneously omitted in [6], Eqn. (2.26b).

To analyze $f(x) - \tilde{f}(x)$, the following lemma was given in ([6], Theorem 3), for $p=2$.

Lemma 3.1. Let $f \varepsilon W_p^{(m)}$ for $p=2$. Then

$$(f(x) - \tilde{f}(x))^2 \leq a(m)\left(\int_{x_0}^{x_m} |f^{(m)}(u)|^p\right)^{2/p} (x_m - x_0)^{2m-2/p},$$

$$x \varepsilon [x_0, x_m], \quad m=1,2$$

$$x \varepsilon [x_1, x_{m-1}], \quad m \geq 3$$

where

$$a(1) = 1, \quad a(2) = (5/2)^2, \quad a(m) = \left[\frac{2(m+3)}{(m-1)!}\right]^2, \quad m \geq 3.$$

Lemma 3.1 is immediately extended to $p \geq 1$ by replacing the Cauchy-Schwartz inequality in (3.9) of [6] by a Hölder inequality with 2 replaced by p.

The entire argument of [6] now goes through exactly for $p \geq 1$, simply by replacing 2 by p in Theorem 3 of [6]. The result (from [6]) is then

$$(3.4a) \qquad 2E\left(f(x) = \sum_{\nu=1}^{m} \frac{d}{dx} \ell_{i,\nu}(x) \int_{t_{ik_n}}^{t_{(i+\nu)k_n}} f(\xi)d\xi\right)^2 \leq ||f^{(m)}||_p^2 \, A_1\left(\frac{k_n}{n+1}\right)^{2m-2/p} \left(1+O(\tfrac{1}{k_n})\right), \quad p \geq 1$$

where

$$(3.4b) \qquad A_1 = 2a(m) \cdot m(\tfrac{m}{\lambda})^{2m-2/p}.$$

Thus, ignoring negligible terms in (3.2),

$$(3.5)\quad E\left(f(x) - \hat{f}_n(x)\right)^2 \le ||f^{(m)}||_p^2 \, A_1\left(\frac{k_n}{n+1}\right)^{2m-2/p} + B_1 \frac{1}{k_n} .$$

The right hand side of (3.5) is minimized (see lemma 4a of [4]) by taking

$$(3.6)\quad k_n = \left[\frac{1}{(2m-2/p)} \frac{B_1}{||f^{(m)}||_p^2 A_1}\right]^{1/(2m+1-2/p)} (n+1)^{(2m-2/p)/(2m+1-2/p)} .$$

in which case

$$E\left(f(x) - \hat{f}_n(x)\right)^2 \le D_1 \, n^{-\phi}$$

where

$$D_1 = \theta\left(||f^{(m)}||_p^2 \, A_1 \, B_1^{2m-2/p}\right)^{1/(2m+1-2/p)}$$

For completeness we state the extended version of Theorems 1 and 2 of [6], as now obtains for $p \ge 1$.

Theorem 3.1.

Let $f(u) \le \Lambda$, all u, let $f(u) \ge \lambda$ for u in a neighborhood of x, let $|u(1 - F(u))|$ and $|uF(u)|$ be bounded respectively for $u \ge x$ and $u \le x$. Let $f \varepsilon W_p^{(m)}$ for $m=1,2,\ldots,$ $p \ge 1$. Let $\hat{f}_n(x)$ be given by (3.1) with k_n given by (3.6). Then

$$E\left(f(x)-\hat{f}_n(x)\right)^2 \leq D_1\, n^{-(2m-2/p)/(2m+1-2/p)} + \text{lower order terms}$$

where

$$D_1 = \theta\left(||f^{(m)}||_p^2\, A_1\, B_1^{2m-2/p}\right)^{1/(2m+1-2/p)}$$

and

$$A_1\, B_1^{2m-2/p} = \left[2a(m)\cdot m\cdot \left(\frac{m}{\lambda}\right)^{2m-2/p}\right]\left[2m^{2m+3\frac{1}{4}}\, \frac{\Lambda^{2m}}{\lambda^{2(m-1)}}\, 3^{\frac{1}{2}}\right]^{2m-2/p}$$

4. Convergence Properties of the Parzen Kernel-Type Density Estimates

The argument of this section was graciously suggested to the author by Professor Farrell. Suppose $f \varepsilon W_p^{(m)}$. Let $K(y)$ be a real valued function on $(-\infty,\infty)$ satisfying

i) $\sup\limits_{-\infty<y<\infty} |K(y)| < \infty$

ii) $\int_{-\infty}^{\infty} |K(y)| < \infty$

iii) $\lim\limits_{y\to\infty} |yK(y)| = 0$

iv) $\int_{-\infty}^{\infty} K(y)\,dy = 1$

$$\text{v)} \quad \int_{-\infty}^{\infty} y^i K(y) = 0 \qquad i=1,2,\ldots,m-1$$

$$\text{vi)} \quad \int_{-\infty}^{\infty} |y|^{m-1/p} |K(y)|\,dy < \infty \qquad p \geq 1.$$

Parzen's density estimate $\hat{f}_n(x)$ is then given by

$$(4.1) \qquad \hat{f}_n(x) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{x-t_j}{h}\right)$$

where $h>0$ is to be chosen so that $h \to 0$, $nh \to \infty$.

Let

$$f_n(x) = E\hat{f}_n(x) = \frac{1}{h} \int_{-\infty}^{\infty} K\left(\frac{x-\xi}{h}\right) f(\xi)\,d\xi$$

From [4], Theorem 2A, the variance term is

$$(4.2a) \quad E\left(\hat{f}_n(x) - f_n(x)\right)^2 = B_2 \frac{1}{nh} + \text{lower order terms}$$

where

$$(4.2b) \qquad B_2 = f(x) \int_{-\infty}^{\infty} K^2(y)\,dy\left(1 + o\left(\frac{1}{nh}\right)\right)$$

The bias term may be established for $m=1,2,\ldots,p \geq 1$, by noting that

$$E\left(f_n(x) - f(x)\right) = \int_{-\infty}^{\infty} K\left(\frac{x-\xi}{h}\right) f(\xi) \frac{d\xi}{h} - f(x)$$

$$= \int_{-\infty}^{\infty} K(-\xi) f(x+\xi h)\,d\xi - f(x).$$

Now

$$(4.3)\quad f(x+\xi h) = f(x)+\sum_{j=1}^{m-1} \frac{(\xi h)^j}{j!} f^{(j)}(x) + \int_x^{x+\xi h} \frac{(x+\xi h-u)^{m-1}}{(m-1)!} f^{(m)}(u)du.$$

Using (iv)-(vi) in (4.3) gives

$$E\left(f_n(x) - f(x)\right) = \int_{-\infty}^{\infty} K(-\xi) \int_x^{x+\xi h} \frac{(x+\xi h-u)^{m-1}}{(m-1)!} f^{(m)}(u)du.$$

Since

$$\left|\int_x^{x+\xi h} \frac{(x+\xi h-u)^{m-1}}{(m-1)!} f^{(m)}(u)du\right| \leq \frac{1}{(m-1)!} \left|\int_x^{x+\xi h} (x+\xi h-u)^{(m-1)q}\right|^{1/q} \cdot \left[\int_{-\infty}^{\infty} |f^{(m)}(u)|^p du\right]^{1/p}$$

$$= \frac{1}{(m-1)!} \frac{|\xi h|^{m-1+1/q}}{[(m-1)q+1]^{1/q}} ||f^{(m)}||_p;$$

$$q = \frac{1}{1-\frac{1}{p}}$$

we have

$$(4.4a)\qquad \left[E\left(f_n(x) - f(x)\right)\right]^2 \leq ||f^{(m)}||_p^2 A_2 h^{2m-2/p}$$

where

$$(4.4b)\quad A_2 = \frac{1}{(m-1)!} \frac{1}{[(m-1)q+1]^{2/q}} \left[\int_{-\infty}^{\infty} |K(\xi)||\xi|^{m-1/p} d\xi\right]^2$$

Thus

$$(4.5)\qquad E\left(f(x) - \hat{f}_n(x)\right)^2 \le ||f^{(m)}||_p^2\, A_2\, h^{2m-2/p} + B_2 \frac{1}{nn}$$

Define $k_n = nh$, and choose

$$(4.6)\quad k_n = \left[\frac{1}{(2m-2/p)} \frac{B_2}{||f^{(m)}||_p^2 A_2}\right]^{1/(2m+1-2/p)} \cdot n^{(2m-2/p)/(2m+1-2/p)},$$

which minimizes the right hand side of (4.5).

We have the following

Theorem 4.1.

Let $f \varepsilon W_p^{(m)}$ for $m=1,2,\ldots,$ $p \ge 1$. Let $\hat{f}_n(x)$ be given by (4.1) where K satisfies (i)-(vi) and $h = k_n/n$ with k_n given by (4.6). Then

$$E\left(f(x) - \hat{f}_n(x)\right)^2 \le D_2\, n^{-(2m-2/p)/(2m+1-2/p)} + \text{lower order terms}$$

with

$$D_2 = \theta(||f^{(m)}||_p^2\, A_2\, B_2^{2m-2/p})^{1/(2m+1-2/p)}$$

and

$$A_2 B_2^{(2m-2/p)} = \frac{1}{\left[(m-1)!\,[((m-1)/(1-1/p))+1]^{1-1/p}\right]^2}$$

$$\cdot \left[\int_{-\infty}^{\infty} |K(\xi)|\,|\xi|^{m-1/p} d\xi\right]^2 \cdot \left[f(x) \cdot \int_{-\infty}^{\infty} K^2(y)\,dy\right]^{(2m-2/p)}.$$

From the point of view of minimizing mean square error here, to optimize the choice of kernel, one should choose K subject to (i)-(vi) to minimize

$$\int_{-\infty}^{\infty} |K(\xi)|\,|\xi|^{m-1/p} d\xi \left[\int_{-\infty}^{\infty} K^2(\xi)\,d\xi\right]^{m-1/p}.$$

5. Convergence Properties of the Kronmal-Tartar Orthogonal Series Density Estimate

Suppose that $f \epsilon W_p^{(m)}$ and $f(\xi) = 0$ for $\xi \notin [0,1]$. Let $\psi_k(x) = \cos \pi kx$, $k=0,1,2,\ldots$ Then the Kronmal-Tartar orthogonal series density estimate [3] is given by

(5.1) $$\hat{f}_n(x) = \sum_{k=0}^{r} \hat{a}_k \psi_k(x)$$

where r is to be chosen, and

(5.2) $$\hat{a}_k = \frac{2}{n} \sum_{j=1}^{n} \psi_k(t_j), \qquad k=0,1,2,\ldots$$

$\hat{a}_k$ is an unbiassed estimator of a_k, where

$$a_k = 2\int_0^1 f(\xi)\psi_k(\xi)d\xi, \qquad k=0,1,2,\ldots .$$

Since $\{\psi_k\}_{k=0}^{\infty}$ are complete on $\mathcal{L}_2[-1,1]$ with respect to even functions on $[-1,1]$ and we can define $f(-\xi) = f(\xi)$, f has the Fourier expansion

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos\pi kx$$

Thus

$$f(x) - \hat{f}_n(x) = \sum_{k=1}^{r}(a_k-\hat{a}_k)\psi_k(x) + \sum_{k=r+1}^{\infty} a_k\psi_k(x)$$

$(a_0 = \hat{a}_0 = 2)$.

The variance term is studied by observing that

$$E(a_k-\hat{a}_k)(a_\ell-\hat{a}_\ell) = 4\int_0^1 \psi_k(\xi)\psi_\ell(\xi)f(\xi)d\xi - a_k a_\ell, \qquad k,\ell=1,2,\ldots,$$

thus

$$E\left(\sum_{k=1}^{r}(a_k-\hat{a}_k)\psi_k(x)\right)^2 = \frac{1}{n}\left\{\sum_{k,\ell=1}^{r}\psi_k(x)\psi_\ell(x)\left[4\int_0^1 \psi_k(\xi)\psi_\ell(\xi)d\xi - a_k a_\ell\right]\right\}$$

$$= \frac{1}{n}\left[4\int_0^1\left(\sum_{k=1}^{r}\psi_k(x)\psi_k(\xi)\right)^2 f(\xi)d\xi - \left(\sum_{k=1}^{r}a_k\psi_k(x)\right)^2\right].$$

Now

$$\sum_{k=1}^{r}\psi_k(x)\psi_k(\xi) = \sum_{k=1}^{r}\cos\pi kx\cos\pi k\xi = \frac{1}{2}\left[w_r(x+\xi) + w_r(x-\xi)\right]$$

where

$$w_r(\tau) = \cos\left(\frac{1}{2}(r+1)\pi\tau\right) \frac{\sin(\frac{r\pi\tau}{2})}{\sin(\frac{\pi\tau}{2})} .$$

Therefore, for large r, the variance term "behaves like" $\frac{r}{n}\frac{f(x)}{2}$. For concreteness, we note that since

$$\int_0^1 \left(\sum_{k=1}^{r} \psi_k(x)\psi_k(\xi)\right)^2 d\xi \leq \frac{r}{2}$$

(5.3a) $$E\left(\sum_{k=1}^{r}(a_k-\hat{a}_k)\psi_k(x)\right)^2 \leq B_3 \frac{r}{n}$$

where

(5.3b) $$B_3 = 2\Lambda$$

and Λ satisfies

$$\max_{\xi} f(\xi) \leq \Lambda.$$

To establish a bound on the bias term, we use the following

Lemma 5.1 (Young and Hausdorff).

Suppose $g(x)\varepsilon\mathcal{L}_2[-1,1]$ with Fourier series $\sum_{-\infty}^{\infty} g_k e^{i\pi kx}$, $g_k = \frac{1}{2}\int_{-1}^{1} g(x)e^{i\pi kx}dx$. If $1<p\leq 2$, and $\frac{1}{p}+\frac{1}{q} = 1$, then

$$\left(\sum_{-\infty}^{\infty} |g_k|^q\right)^{1/q} \leq \left(\frac{1}{2}\int_{-1}^{1} |g(x)|^p dx\right)^{1/p} .$$

This result is stated in Hardy, Littlewood, and Polya [2], equation (8.5.7), for the proof see [2], p. 221. The limitation on p is essential, indeed, if $p \geq 2$, the reverse inequality holds (see (8.5.6)).

Now,

$$(5.4) \qquad \left| \sum_{k=r+1}^{\infty} a_k \cos \pi kx \right| \leq \sum_{k=r+1}^{\infty} |a_k| k^m \cdot \frac{1}{k^m}$$

$$\leq \left(\sum_{r+1}^{\infty} |a_k k^m|^q \right)^{1/q} \left(\sum_{r+1}^{\infty} \frac{1}{k^{pm}} \right)^{1/p} .$$

Also,

$$(5.5) \qquad \sum_{r+1}^{\infty} \frac{1}{k^{pm}} \leq \int_r^{\infty} \frac{1}{\xi^{pm}} d\xi = \frac{1}{(pm-1)} \left(\frac{1}{r}\right)^{pm-1} .$$

Next, observe that

$$f^{(m)}(x) = \pi^m \sum_{k=1}^{\infty} a_k (-1)^{\frac{m}{2}} k^m \cos \pi kx, \qquad m \text{ even}$$

$$= \pi^m \sum_{k=1}^{\infty} a_k (-1)^{\frac{m+1}{2}} k^m \sin \pi kx, \qquad m \text{ odd}$$

Let $g(x) = f^{(m)}(x)$. Then the non zero Fourier coefficients of g have absolute values $\pi^m |a_k| k^m$, $k=1,2,\ldots,$ and by the Lemma of Young and Hausdorff, we have

$$(5.6) \qquad \pi^m \left[\sum_{k=1}^{\infty} |a_k k^m|^q \right]^{1/q} \leq \left(\int_0^1 |f^{(m)}(\xi)|^p \right)^{1/p}, \qquad 1<p\leq 2.$$

Putting together (5.4), (5.5) and (5.6) gives

$$(5.7a) \qquad \Big|\sum_{k=r+1}^{\infty} a_k \cos\pi kx\Big|^2 \le ||f^{(m)}||_p^2 \, A_3 \, (\tfrac{1}{r})^{2m-2/p}$$

where

$$(5.7b) \qquad A_3 = \frac{1}{\pi^{2m}} \frac{1}{(pm-1)^{2/p}} \, .$$

Thus

$$(5.8) \quad E\left(f(x) - \hat{f}_n(x)\right)^2 \le ||f^{(m)}||_p^2 \, A_3 (\tfrac{1}{r})^{2m-2/p} + B_3 \frac{r}{n}$$

where

$$A_3 = \frac{1}{\pi^{2m}} \frac{1}{(pm-1)^{2/p}}$$

$$B_3 = 2\Lambda .$$

Define k_n by $k_n = \frac{n}{r}$, and choose $r = n/k_n$ with

$$(5.9) \quad k_n = \left[\frac{1}{(2m-2/p)} \frac{B_3}{||f^{(m)}||_p^2 A_3}\right]^{1/(2m+1-2/p)} n^{(2m-2/p)/(2m+1-2/p)} \, .$$

Then the right hand side of (5.9) is minimized. We have the following

Theorem 5.1.

Let $f \varepsilon W_p^{(m)}[0,1]$, for $m=1,2,\ldots$, $1<p\leq 2$. Let $\hat{f}_n(x)$ be given by (5.1) where $r = n/k_n$ with k_n given by (5.9). Then

$$E\left(f(x) - \hat{f}_n(x)\right)^2 \leq D_3\, n^{-(2m-2/p)/(2m+1-2/p)}$$

with

$$D_3 = \theta\left(||f^{(m)}||_p^2\, A_3\, B_3^{2m-2/p}\right)^{1/(2m+1-2/p)}$$

and

$$A_3\, B_3^{2m-2/p} = \frac{1}{\pi^{2m}(pm-1)^{2/p}}\,(2\Lambda)^{2m-2/p}\ .$$

We remark here that there is some doubt as to the truth of this result for $2<p\leq\infty$. Also, one cannot use an arbitrary orthonormal series and expect to obtain the same result, as $\sup_{x,k}|\cos\pi kx| \leq 1$ was needed in the proof in (5.4).

6. Convergence Properties of the Ordinary Histogram

Suppose that $f \varepsilon W_p^{(m)}$ for $m=1$, $p\geq 1$, and $f(\xi) = 0$ for $\xi \notin [0,1]$. Let h be chosen so that $1/h = \ell$, an integer. Let I_j be the interval $[jh,(j+1)h)$, $j=0,1,\ldots,\ell-1$. Let

$$(6.1) \qquad \hat{f}_n(x) = \frac{Y_j}{nh}, \quad x \varepsilon I_j, \quad j=0,1,\ldots,\ell-1,$$

where

$$Y_j = \text{number of } t_1, t_2, \ldots, t_n \text{ in } I_j.$$

Since Y_j is binomial $B(n,p_j)$ where $p_j = \int_{jh}^{(j+1)h} f(\xi)d\xi$,

$$E\hat{f}_n(x) = \frac{1}{h} p_j$$

$$(6.2) \qquad \operatorname{Var} \hat{f}_n(x) = \frac{p_j(1-p_j)}{nh^2} \le \frac{\Lambda}{nh}$$

Now,

$$E\left(f(x) - \hat{f}_n(x)\right) = f(x) - \frac{1}{h}\int_{jh}^{(j+1)h} f(\xi)d\xi = \frac{1}{h}\int_{jh}^{(j+1)h}\left(f(x) - f(\xi)\right)d\xi$$

For $f \varepsilon W_p^{(m)}$, $m=1$ and $x \varepsilon I_j$, $\xi \varepsilon I_j$,

$$|f(x) - f(\xi)| = \left|\int_\xi^x f^{(m)}(u)du\right| \le \int_{jh}^{(j+1)h} |f^{(m)}(u)|du$$

$$\le h^{1/q}\left[\int_{-\infty}^{\infty} |f^{(m)}(u)|^p du\right]^{1/p} = h^{1-1/p}||f^{(1)}||_p .$$

Thus

$$\left[E(f(x) - \hat{f}_n(x))\right]^2 \le ||f^{(m)}||_p^2 \, h^{2m-2/p}, \qquad (m=1)$$

and

$$(6.3)\quad E\left(f(x) - \hat{f}_n(x)\right)^2 \le ||f^{(1)}||_p\, A_4\, h^{2m-2/p} + B_4 \frac{1}{nh}$$

where

$$A_4 = 1$$

$$B_4 = \Lambda.$$

Define k_n by $k_n = nh$, and choose $h = \frac{k_n}{n}$ with

$$(6.4)\quad k_n = \left[\frac{1}{(2m-2/p)} \frac{B_4}{||f^{(m)}||_p^2 A_4}\right]^{1/(2m+1-2/p)} n^{(2m-2/p)/(2m+1-2/p)},$$

$$m=1.$$

Then the right hand side of (6.3) is minimized and we have the following

Theorem 6.1.

Let $f \varepsilon W_p^{(m)}[0,1]$ for $m=1$, $p \ge 1$. Let $\hat{f}_n(x)$ be given by (6.1) where $h = k_n/n$ with k_n chosen as in (6.4). Then

$$E\left(f(x) - \hat{f}_n(x)\right)^2 \le D_4\, n^{-(2m-2/p)/(2m+1-2/p)}$$

with

$$D_4 = \theta\left(A_4 B_4^{2m-2/p}\right)^{1/(2m+1-2/p)}$$

with

$$A_4 B_4^{2m-2/p} = A_4 B_4^{2-2/p} = \Lambda^{2-2/p}.$$

7. Summary and Concluding Remarks

We summarize the results in Table 1.

We conclude with a brief remark concerning the criteria we have been using, namely minimum mean square error at a point. Firstly, there is no asymptotic distribution theory here, and it probably doesn't exist. In order for asymptotic normality to obtain, it is apparent that the bias (squared) term must be asymptotically negligible compared to the variance. If the parameter (respectively k_n, h, r and h here) is chosen so that this happens, then the optimum rate will not obtain. Thus, it seems preferable to choose the parameter for the optimum rate, and use Tchebycheff's Theorem to construct confidence intervals. It remains an open problem to provide a lower bound D_0 on the constant

$$D = \theta\left(AB^{2m-2/p}\right)^{1/(2m+1-2/p)} .$$

$$E\left(f(x) - \hat{f}_n(x)\right)^2 \leq \frac{(2m+1-2/p)}{(2m-2/p)^{(2m-2/p)}} \left(||f^{(m)}||_p^2 \; AB^{2m-2/p} \right)^{1/(2m+1-2/p)}$$

$$\cdot \; n^{-(2m-2/p)/(2m+1-2/p)} + \text{lower order terms}$$

Method	$AB^{2m-2/p}$	range of validity
1. Polynomial	$\left[2a(m) \cdot m \cdot (\frac{m}{\lambda})^{2m-2/p}\right] \left[2m^{2m+3\frac{1}{4}} \; \frac{\Lambda^{2m}}{\lambda^{2(m-1)}} \; 3^{\frac{1}{2}}\right]^{2m-2/p}$	m=1,2,...,p$\geq$1
2. Kernel	$\left[\left((m-1)!\left(((m-1)/(1-1/p))+1\right)^{1-1/p}\right)^{-1} \int_{-\infty}^{\infty} \lvert K(\xi)\rvert \, \lvert\xi\rvert^{m-1/p} d\xi\right]^2 \cdot \left[\Lambda \int_{-\infty}^{\infty} K^2(y)dy\right]^{2m-2/p}$	m=1,2,...,p$\geq$1
3. Orthogonal Series	$\left[\pi^{2m}(pm-1)^{2/p}\right]^{-1} \left[2\Lambda\right]^{2m-2/p}$	m=1,2,...,2$\geq$p>1
4. Histogram	$\Lambda^{2m-2/p}$	m=1, p$\geq$1

$$\sup_{\xi} f(\xi) \leq \Lambda$$

$$f(u) \leq \lambda \text{ in a neighborhood of } x$$

Table 1. Summary of Results

To examine the relationship between mean square error at a point and integrated mean square error, suppose $0<\lambda\leq f(x)\leq\Lambda$ on a known bounded interval (say [0,1]) and $f(x) = 0$, $x \notin [0,1]$. Then

$$D_0 ||f^{(m)}||_p^{2/(2m+1-2/p)} n^{-(2m-2/p)/(2m+1-2/p)} + \text{negligible terms}$$

$$\leq \int_0^1 E\left(f(\xi) - \hat{f}_n(\xi)\right)^2 d\xi$$

$$\leq D_\nu ||f^{(m)}||_p^{2/(2m+1-2/p)} n^{-(2m-2/p)/(2m+1-2/p)}$$

$$+ \text{negligible terms}$$

for method ν, $\nu=1,2,3,4$. Thus, optimum constants may be different for integrated mean square error vs. mean square error at a point but the rates will not.

We have recently obtained two more entries in the table above, for fixed knot and for variable knot cubic spline density estimates, for $m=1,2,3$. This work will appear separately.

We wish to acknowledge a number of helpful discussions with Professor R. H. Farrell.

REFERENCES

[1] Farrell, R. H. (1972). On best obtainable asymptotic rates of convergence in estimation of a density function at a point. Ann. Math. Statist. 43, 170-180.

[2] Hardy, G. H., Littlewood, J. E., and Polya, G. (1934). Inequalities, Cambridge University Press, (second edition reprinted 1964).

[3] Kronmal, R. and Tartar, M. (1968). The estimation of probability densities and cumulatives by Fourier series methods. JASA 63, 925-952.

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

[5] Van Ryzin, J. (1970). On a histogram method of density estimation. University of Wisconsin, Department of Statistics Technical Report #226.

[6] Wahba, Grace (1971). A polynomial algorithm for density estimation. Ann. Math. Statist. 43, 1870-1886.

[7] Wegman, E. J. (1972). Nonparametric probability density estimation: I. A summary of available methods. Technometrics 14, 533-546.

[8] Wegman, E. J. (1972). Nonparametric probability density estimation: II. A comparison of density estimation methods. J. Statist. Comput. Simul., 1, 225-245.

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)* University of Wisconsin Department of Statistics Madison, Wisconsin 53706	2a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED 2b. GROUP

3. REPORT TITLE

OPTIMAL CONVERGENCE PROPERTIES OF VARIABLE KNOT, KERNEL, AND ORTHOGONAL SERIES METHODS FOR DENSITY ESTIMATION

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

Scientific Interim

5. AUTHOR(S) *(First name, middle initial, last name)*

Grace Wahba

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
December 1972	27	8

8a. CONTRACT OR GRANT NO. AFOSR 72-2363	9a. ORIGINATOR'S REPORT NUMBER(S) Technical Report No. 324
b. PROJECT NO. c. d.	9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*

10. DISTRIBUTION STATEMENT

This document has been approved for public release and sale; its distribution is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY Air Force Office of Scientific Research 1400 Wilson Boulevard Arlington, Virginia

13. ABSTRACT

Let f be a density function satisfying

$$\left[\int_{-\infty}^{\infty}(f^{(m)}(\xi))^p\right]^{1/p} \le M < \infty,\ m=1,2,\ldots,\ p\ge 1,$$

and let $\hat{f}_n(x)$ be an estimate of $f(x)$ based on n independent observations from the density f. If $\hat{f}_n$ is the polynomial algorithm for density estimation then it is known that

$$E(f(x)-\hat{f}_n(x))^2 = O(n^{-(2m-2/p)/(2m+1-2/p)})$$

for $m = 1, 2, \ldots$ and $p=2$, $p=\infty$. If $\hat{f}_n$ is an appropriate Parzen kernel type estimate, the above result is known to hold for $m = 1, 2, \ldots$, and $p=\infty$. By applying a theorem of Farrell, it is shown that these are the best obtainable rates for $m = 1, 2, \ldots$, and $p\ge 1$. These optimal convergence rates are then shown to hold for the polynomial and kernel estimates for all $p\ge 1$. The optimal rates are also shown to hold for the Kronmal-Tartar orthogonal series estimate for $m = 1, 2, \ldots$, and $1<p\le 2$, and for the ordinary histogram estimate with variable "bins" for $m = 1$, $p\ge 1$. Upper bounds for the constants covered by the "O" are exhibited.

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Density estimates						
Optimum convergence rates						

Security Classification