The Annals of Statistics
1981, Vol 9, No. 1, 146–156

DATA-BASED OPTIMAL SMOOTHING OF ORTHOGONAL SERIES DENSITY ESTIMATES[1]

BY GRACE WAHBA

University of Wisconsin-Madison

Let f be a density possessing some smoothness properties and let $X_1, \ldots, X_n$ be independent observations from f. Some desirable properties of orthogonal series density estimates $f_{n,m,\lambda}$ of f of the form

$$f_{n,m,\lambda}(t) = \sum_{\nu=1}^{n} \frac{\hat{f}_\nu}{(1+\lambda\nu^{2m})} \phi_\nu(t)$$

where $\{\phi_\nu\}$ is an orthonormal sequence and $\hat{f}_\nu = (1/n)\sum_{j=1}^{n} \phi_\nu(X_j)$ is an estimate of $f_\nu = \int \phi_\nu(t)f(t)\,dt$, are discussed. The parameter λ plays the role of a bandwidth or "smoothing" parameter and m controls a "shape" factor. The major novel result of this note is a simple method for estimating λ (and m) from the data in an *objective* manner, to minimize integrated mean square error. The results extend to multivariate estimates.

1. Introduction and statement of results. Let f be a square integrable density on [0, 1] possessing a Fourier series expansion

$$f(t) \sim 1 + \sum_{\nu=-\infty;\,\nu\neq 0}^{\infty} f_\nu\phi_\nu(t) \tag{1.1}$$

where $\phi_\nu(t) = e^{2\pi i\nu t}$. It is desired to estimate f from n independent observations $X_1, \cdots, X_n$ from the density f. Given a sequence $\mathbf{b} = (\cdots, b_{-2}, b_{-1}, b_1, b_2, \cdots)$ of nonnegative real numbers with $b_\nu = b_{-\nu}$, and $\sum |b_\nu| < \infty$ the orthogonal series estimate $\hat{f}_{n,\mathbf{b}}$ of f as considered by Whittle [38], Kronmal and Tartar [15], Brunk [2], Fellner [7] and others can be written in the form (n even)

$$\hat{f}_{n,\mathbf{b}}(t) = 1 + \sum\nolimits_* \frac{b_\nu}{b_\nu + \frac{1}{n}} \hat{f}_\nu\phi_\nu(t), \tag{1.2}$$

where

$$\sum\nolimits_* = \sum_{\nu=-n/2,\,\nu\neq 0}^{n/2}$$

and

$$\hat{f}_\nu = \frac{1}{n}\sum_{j=1}^{n} \phi_\nu^*(X_j).$$

This type of estimate can be motivated in several ways. Suppose the Fourier coefficients $\{f_\nu\}$ of f have the "phoney" prior,

$$f_\nu \sim \mathcal{N}_c(0, b_\nu), \quad \text{independent}, \qquad \nu = 1, 2, \cdots \tag{1.3}$$

($\mathcal{N}_c$ is the complex normal distribution, see [12]). Then,

$$E\hat{f}_\nu = \int_0^1 \phi_\nu(t)f(t)\,dt = f_\nu \tag{1.4}$$

Received February 1978; revised September 1979.

[1] Research supported by U.S. Army under Contract No. DAAG29-77-G-0207.

AMS 1970 *subject classification.* Primary 62G05.

Key words and phrases. Optimal smoothing, orthogonal series, density estimation.

and (by (4.2) below),

$$E|\hat{f}_\nu - E\hat{f}_\nu|^2 = \frac{1}{n}(1 - |f_\nu|^2). \tag{1.5}$$

If one approximates $(1/n)(1 - |f_\nu|^2)$ by $1/n$ and the distribution of $\hat{f}_\nu$ by a (complex) normal distribution, the posterior mean of f_ν given $\hat{f}_\nu$ is

$$\frac{b_\nu}{b_\nu + \frac{1}{n}}\hat{f}_\nu.$$

Then $\hat{f}_{n,\mathbf{b}}(t)$ of (1.2) can be viewed as a Bayesian estimate of $f(t)$ with the phoney prior on $f(t)$ induced by (1.1) and (1.3). The prior is phoney because the sample functions are not required to be nonnegative, although they do integrate to 1. Motivation as a smoothing spline estimate will be discussed later.

Various specifications have been proposed in [2, 15, 26, 38] for the $\mathbf{b} = (\cdots, b_{-2}, b_{-1}, b_1, b_2, \cdots)$ which determines the prior. In this note we propose a two parameter family of $\mathbf{b}$'s, namely, $\{b_\nu = (\lambda(2\pi\nu)^{2m})^{-1}\nu = \pm 1, 2, \cdots\}$, $\lambda \geq 0$, $m > \frac{1}{2}$. We will write the resulting estimate as $f_{n,\lambda,m}$,

$$f_{n,\lambda,m}(t) = 1 + \sum_* \frac{\hat{f}_\nu}{(1 + \lambda(2\pi\nu)^{2m})}\phi_\nu(t), \tag{1.6}$$

where the factor $1/n$ has been absorbed into λ. This family of estimates possess a wealth of nice properties, which we shall demonstrate.

Firstly, from a Bayesian point of view, we shall show (trivially) that if two $\mathbf{b}$'s have distinct values of (λ, m), their associated (infinite-dimensional) prior distributions are perpendicular; furthermore, as λ and m range over their permissible values, the class of all priors equivalent to some member in the family of associated priors is exceedingly large. This supports the argument that there is no need to go outside this family.

Secondly, leaving the Bayesian point of view and supposing f is a fixed density in the space $W_2^{(m)}$ (per) of periodic functions

$$W_2^{(m)}\text{ (per)} = \{f{:}f, f', \cdots, f^{(m-1)} \text{ abs. cont.}, f^{(m)} \in \mathscr{L}_2[0, 1],$$
$$f^{(\nu)}(0) = f^{(\nu)}(1), \nu = 0, 1, \cdots, m - 1\},$$

where now m is a given fixed integer, it will follow easily that, if $\lambda = \text{const. } n^{-2m/(2m+1)}$ then the integrated mean square error $E\int_0^1 [f_{n,m,\lambda}(t) - f(t)]^2\, dt$ satisfies

$$E\int_0^1 [f_{n,m,\lambda}(t) - f(t)]^2\, dt = O(n^{-2m/(2m+1)}).$$

This integrated mean square error convergence rate is slightly better than the optimal achieveable mean square error at a point convergence rate for densities possessing the same continuity conditions, (see [30]), and it appears that it cannot be substantially improved upon uniformly for densities in $W_2^{(m)}$ (per). (If m is not an integer, $f \in W_2^{(m)}$ (per) if $\sum_{\nu=-\infty}^{\infty} (2\pi\nu)^{2m}|f_\nu| < \infty$.)

We view the above two properties, although important, as side issues. A major problem in density estimation is to choose the smoothing parameter(s), which are a part of every density estimate (see [30]), *objectively from the data*, to approximately *minimize some optimality criterion*. Here the major smoothing parameter is λ, and m is a secondary "shape" parameter—we amplify this remark: let $f_{n,0}(t)$, be the "raw" orthogonal series estimate of f,

$$f_{n,0}(t) = 1 + \sum_* \hat{f}_\nu\phi_\nu(t).$$

Then $f_{n,m,\lambda}$ may be viewed as the result of passing $f_{n,0}$ through a low pass filter with frequency response function $\Psi(\nu) = 1/[1 + \lambda(2\pi\nu)^{2m})]$. The parameter λ controls the half power point of the filter (large λ corresponds to "low-pass"), and m controls the "shape" (large m corresponds to a steep roll off). The primary original contribution of this note is a simple objective method for estimating from the data, the λ and m which minimize integrated mean square error.

Woodroofe [39] provides an objective, iterative procedure for choosing the smoothing parameter in a kernel estimate to minimize mean square error at a point; however, this technique appears to be slow to converge and computationally impractical. Good and Gaskins [10] suggest a method for choosing the smoothing parameter in a penalized maximum likelihood method, based on a goodness-of-fit criterion. Leonard [16], Fellner [7], Brunk [2] and Tarter and Kronmal [26] discuss procedures for choosing the degree of smoothness in various estimates, which involve varying degrees of subjectivity. Scott, Tapia and Thompson [24] provide an objective, iterative method for choosing the smoothing parameter in a kernel estimate to minimize integrated mean square error (IMSE). Their method is based on estimating $\int (f''(t))^2\,dt$, which appears in a theoretical expression for the IMSE, and is different from the method proposed here.

Two readily computable completely objective methods using cross-validation to determine the degree of smoothing (Hermans and Habbema [14], and Wahba [31] will be discussed in Section 5.

The objective determination of λ and m is based on the following.

THEOREM 4.1. *Let*

$$T_{n,m}(\lambda) = \int_0^1 (f_{n,m,\lambda}(t) - f(t))^2\,dt.$$

If $f \in W_2^{(\tilde m)}$ (per) *for some* $\tilde m > ½$, *then* $\hat T_{n,m}(\lambda)$, *defined by*

$$\hat T_{n,m}(\lambda) = \frac{n}{n-1}\sum\nolimits_* \left\{\left(\frac{\lambda}{\lambda_\nu+\lambda}\right)^2 - \frac{1}{n}\left(\frac{\lambda_\nu}{\lambda_\nu+\lambda}\right)^2\right\}|\hat f_\nu|^2 - \frac{1}{n-1}\sum\nolimits_* \left\{\left(\frac{\lambda}{\lambda_\nu+\lambda}\right)^2 - \left(\frac{\lambda_\nu}{\lambda_\nu+\lambda}\right)^2\right\},$$

where

$$\lambda_\nu = 1/(2\pi\nu)^{2m}$$

satisfies

$$E\hat T_{n,m}(\lambda) = ET_{n,m}(\lambda) + O\left(\frac{1}{n^{2\tilde m}}\right).$$

The procedure is to compute $\hat T_{n,m}(\lambda)$, and to choose λ and m as the minimizers of $\hat T_{n,m}(\lambda)$.

Although the proof of this result is trivial, at the time it first appeared in 1975 (Wahba [29]) it apparently had not been recognized in this context, and is exceedingly useful. It should, however, be considered to be in the spirit of Mallows' C_L method [17] for choosing the ridge parameter in ridge regression. See [9], equation (1.8), for further details.

More recently Davis [5] and Tarter [25] have observed that approximately unbiassed estimates of IMSE can be obtained for the density estimators they suggest. Good and Gaskins [11] have recently continued the work in [10]. Habbema and Hermans' method has also been proposed by Duin [6]. A Monte Carlo study comparing the estimators in [6], [24], and [31] has been performed by Scott and Factor [23]. Scott [22] also compares the methods in [6, 11, 24, 31], on some of the data in [11]. Parzen [19] has proposed the autoregressive density estimation with CAT and/or graphical methods for choosing the smoothing parameter.

The density estimate proposed here is related to the periodic smoothing polynomial spline and we briefly describe this relationship. We also describe the bivariate orthogonal series estimate that is related analogously to the bivariate thin plate smoothing spline.

In Section 6 of this note we generalize the results to rather arbitrary orthogonal series estimates for densities with support on an arbitrary index set T. It is noted that if the density being estimated is assumed to be in some reproducing kernel Hilbert space, then the IMSE convergence rate $O(n^{-2m/(2m+1)})$ is achievable whenever the eigenvalues of the reproducing kernel tend to 0 at the rate n^{-2m}.

2. Equivalence and perpendicularity of priors.

THEOREM 2.1. *Let $f_1, f_2, \cdots$ be an infinite sequence of independent, zero mean normally distributed random variables with probability measure denoted $P_{m,\lambda}$ if $Ef_\nu^2 = [\lambda(2\pi\nu)^{2m}]^{-1}$, $\nu = 1, 2, \cdots$.*

(i) *For $0 < \lambda$, $0 \le m < \infty$, $P_{m_1,\lambda_1} \perp P_{m_2,\lambda_2}$ unless $m_1 = m_2$ and $\lambda_1 = \lambda_2$.*

(ii) *Let $P_{\mathbf{b}}$ be the probability measure corresponding to $Ef_\nu^2 = b_\nu$, $\nu = 1, 2, \cdots$, where*

$$b_\nu = (|\textstyle\sum_{j=0}^{q} \beta_j \nu^j|^2 / |\sum_{j=0}^{p} \alpha_j \nu^j|^2)(1 + o(\nu^{-1})), \qquad \nu = 1, 2, \cdots$$

where the α's and β's are such that $0 < b_\nu < \infty$. Then $P_{\mathbf{b}} \equiv P_{m,\lambda}$ with $m = p - q$ and $\lambda = (2\pi)^{-2m}\alpha_p/\beta_q$.

PROOF. This is a consequence of Hajek [13] who proves that, for any $\mathbf{b}(j) = (b_1(j), b_2(j), \cdots)$, $j = 1, 2$, $P_{\mathbf{b}(1)} \equiv P_{\mathbf{b}(2)}$ if

$$\textstyle\sum_{\nu=1}^{\infty} \left| \dfrac{b_\nu(1)}{b_\nu(2)} - 1 \right|^2 < \infty$$

and $P_{\mathbf{b}(1)} \perp P_{\mathbf{b}(2)}$ otherwise.

We take this opportunity to remark that sample functions from $P_{m,\lambda}$ are, with probability 1, not in $W_2^{(m)}$ (per), since

$$E_{P_{m,\lambda}} \textstyle\sum_{\nu=-\infty}^{\infty} (2\pi\nu)^{2m} |f_\nu|^2 = \infty.$$

3. Convergence properties of $f_{n,m,\lambda}$.

THEOREM 3.1. *Let $f \in W_2^{(\tilde m)}$ (per). Then, for any m with $\frac{1}{2} < m \le \tilde m$, the expected integrated mean square error $ET_{m,n}(\lambda)$ of $f_{n,m,\lambda}$ satisfies*

$$ET_{m,n}(\lambda) \le \lambda\theta_m + \frac{k_m}{n\lambda^{1/2m}} + \frac{\theta_{\tilde m}}{n^{2\tilde m}}$$

where

$$k_m = \frac{1}{\pi}\int_0^\infty \frac{dx}{(1 + x^{2m})^2}$$

and

$$\theta_m = \textstyle\sum_{\nu=-\infty}^{\infty} (2\pi\nu)^{2m} |f_\nu|^2.$$

($\theta_m = \int_0^1 (f^{(m)}(t))^2\, dt$ if m is an integer). Thus, if $\lambda = O(n^{-2m/(2m+1)})$ then

$$ET_{m,n} = O(n^{-2m/(2m+1)}).$$

PROOF. By Parseval's theorem

$$T_{n,m}(\lambda) = \int_0^1 (f(t) - f_{n,m,\lambda}(t))^2\,dt \tag{3.1}$$
$$= \textstyle\sum_* \left| \frac{\lambda_\nu}{\lambda_\nu + \lambda}\hat{f}_\nu - f_\nu \right|^2 + \sum_{|\nu|>n/2} |f_\nu|^2,$$

where $\lambda_\nu = 1/(2\pi\nu)^{2m}$. Since

$$E(\hat{f}_\nu - f_\nu) = 0$$
$$E|\hat{f}_\nu - f_\nu|^2 = \frac{1}{n}(1 - |f_\nu|^2)$$

we have

$$ET_{n,m}(\lambda) = \textstyle\sum_* \left(\frac{\lambda_\nu}{\lambda_\nu + \lambda}\right)^2 E|\hat{f}_\nu - f_\nu|^2 + \sum_* \left(\frac{\lambda}{\lambda_\nu + \lambda}\right)^2 |f_\nu|^2 + \sum_{|\nu|>n/2} |f_\nu|^2 \tag{3.2}$$
$$= \textstyle\sum_* \left\{\left(\frac{\lambda}{\lambda_\nu + \lambda}\right)^2 - \frac{1}{n}\left(\frac{\lambda_\nu}{\lambda_\nu + \lambda}\right)^2\right\}|f_\nu|^2 + \frac{1}{n}\sum_* \left(\frac{\lambda_\nu}{\lambda_\nu + \lambda}\right)^2 + \sum_{|\nu|>n/2} |f_\nu|^2.$$

The theorem follows upon noting that

$$\textstyle\sum_* \left(\frac{\lambda}{\lambda_\nu + \lambda}\right)^2 |f_\nu|^2 \le \lambda \sum_{-\infty}^{\infty} \frac{|f_\nu|^2}{\lambda_\nu} = \lambda\theta_m$$

$$\frac{1}{n}\textstyle\sum_* \left(\frac{\lambda_\nu}{\lambda_\nu + \lambda}\right)^2 = \frac{1}{n}\sum_* \frac{1}{(1 + \lambda(2\pi\nu)^{2m})^2} \le \frac{2}{n}\int_0^\infty \frac{dx}{(1 + \lambda(2\pi x)^{2m})^2} = \frac{k_m}{n\lambda^{1/2m}}$$

and

$$\textstyle\sum_{|\nu|>n/2} |f_\nu|^2 \le \frac{1}{(\pi n)^{2m}} \sum_{|\nu|>n/2} (2\pi\nu)^{2\tilde{m}} |f_\nu|^2 \le \frac{\theta_{\tilde{m}}}{(\pi n)^{2m}}. \tag{3.3}$$

We remark that this estimate (and the integrated mean square error convergence rate) essentially appear in Cogburn and Davis [3] and Wahba and Wold [36] as a spectral density and log spectral density estimate, respectively. For m an integer $f_{n,m,\lambda}$ is, to a good approximation, the solution to the minimization problem: find $f \in W_2^{(m)}$ (per) to minimize $(1/n)\sum_{j=1}^n (f(j/n) - Y_j)^2 + \lambda \int_0^1 (f^{(m)}(t))^2\,dt$, where $Y_j \equiv f_{n,0}(j/n)$, and so $f_{n,m,\lambda}$ is (approximately) a periodic spline function (see [3, 36]). The method for choosing λ and m of this note also can be applied to the log spectral density estimate described in [3, 36], see [34].

4. Unbiased estimates of the expected integrated mean square error, when λ and m are used.

THEOREM 4.1. *Let*

$$\hat{T}_{n,m}(\lambda) = \frac{n}{n-1}\textstyle\sum_* \left\{\left(\frac{\lambda}{\lambda_\nu + \lambda}\right)^2 - \frac{1}{n}\left(\frac{\lambda_\nu}{\lambda_\nu + \lambda}\right)^2\right\}|\hat{f}_\nu|^2 \tag{4.1}$$
$$+ \frac{1}{n-1}\textstyle\sum_* \left\{\left(\frac{\lambda_\nu}{\lambda_\nu + \lambda}\right)^2 - \left(\frac{\lambda}{\lambda_\nu + \lambda}\right)^2\right\}$$

Then, if $f \in W_2^{(\tilde{m})}$ (per),

$$E\hat{T}_{n,m}(\lambda) = ET_{n,m}(\lambda) + O\left(\frac{1}{n^{2\tilde{m}}}\right) \qquad \lambda \ge 0,\ m > ½.$$

PROOF.

$$
\begin{aligned}
E|\hat{f}_\nu|^2 &= E\frac{1}{n^2}\textstyle\sum_{j=1}^n \sum_{k=1}^n \phi_\nu(X_j)\phi_\nu^*(X_k) \\
&= \frac{n}{n^2}\int_0^1 f(u)\,du + \frac{n^2-n}{n^2}\int_0^1 \phi_\nu(u)f(u)\,du \int_0^1 \phi_\nu^*(u)f(u)\,du \qquad (4.2)\\
&= \frac{1}{n} + \frac{n-1}{n}|f_\nu|^2.
\end{aligned}
$$

Taking the expectation of the right-hand side of (4.1) substituting in (4.2) and comparing with (3.2) and (3.3) gives the result.

5. Cross-validation and generalized cross-validation estimates. The generalized cross-validation (GCV) method also provides an objective estimate for the minimizer of $T_{n,m}(\lambda)$, but via a further approximation, see [31], Monte Carlo experiments with $f_{n,m,\lambda}$ with m fixed at 2 and the use of GCV to estimate λ, were reported in [31], with excellent results. Results obtained by minimizing $\hat{T}_{n,m}(\lambda)$ here should be at least as good, if not better than those reported for GCV in [31]. Experiments comparing the GCV estimates and estimates obtained by minimizing the expression analogous to $\hat{T}_{n,m}(\lambda)$ in the context of smoothing splines for nonparametric regression with $m = 2$ support this latter statement, see [4].

The question of the practical benefits of varying m has also been addressed in Monte Carlo experiments in the context of smoothing splines for nonparametric regression [8, 32, 35]. These experiments tend to indicate typical reduction in true IMSE of a few percent if m is estimated as opposed to being fixed at 2.

$\hat{T}_{n,m}(\lambda)$ was occasionally found to be negative at its minimum in [8] but the numerical results indicate that the minimizer is still a good estimate of the minimizer of $T_{n,m}(\lambda)$.

Hermans and Habbema [14] choose h in a kernel estimate (see [18, 20]) of the form

$$f_{n,h}(x) = \frac{1}{nh}\textstyle\sum_{j=1}^n K\left(\frac{x-X_j}{h}\right)$$

by choosing h to maximize what might be called the "cross-validation likelihood function" $V(h)$,

$$V(h) = \textstyle\prod_{k=1}^n f_{n,h}^{(k)}(X_k)$$

where $f_{n,h}^{(k)}(X_k)$ is a cross-validatory estimate of $f(X_k)$,

$$f_{n,h}^{(k)}(x) = \frac{1}{(n-1)h}\textstyle\sum_{j=1;j\neq k}^n K\left(\frac{x-X_j}{h}\right).$$

They use the normal kernel $K(\tau) = 2\pi^{-1/2}e^{-\tau^2/2}$. The method is adapted easily to certain multidimensional kernel estimates. The properties of this method remain to be determined. Some numerical comparisons appear in [22, 23].

6. Abstract orthogonal series density estimates with optimal smoothing. The Fourier series density estimate of this paper would be an ideal, easily computable, all purpose density estimate for smooth densities with compact support, if it were not for the fact that the density estimate is periodic: $f_{n,m,\lambda}$ for m an integer satisfies the periodic boundary conditions $f_{n,m,\lambda}^{(\nu)}(0) = f_{n,m,\lambda}^{(\nu)}(1)$, $\nu = 0, 1, \cdots, m-1$. If the true underlying density does not satisfy appropriate periodic conditions, then an unpleasant Gibbs phenomena can result. If, on the other hand f goes smoothly to zero at the boundaries, then the estimate should be quite satisfactory. It is, of course, a natural estimate for densities on a circle. This problem can in theory be avoided by other choices of orthogonal series.

We now consider the estimation of densities with support on some arbitrary index set T, for example, the real line, or a subset of Euclidean d-space, and eliminate the periodicity requirements on f.

Let $\{\phi_\nu\}_{\nu=1}^{\infty}$ be a complete orthonormal sequence of functions in $\mathscr{L}_2(T)$, and assume

$$\sum_{\nu=1}^{\infty} \frac{\phi_\nu^2(t)}{\nu^{2m}} < M_0 < \infty, \qquad t \in T$$

for all $m \geq m_0$ where $m_0 > ½$. Let $\mathscr{H}_m$ be the collection of functions $\{h\}$ in $\mathscr{L}_2[T]$ which further satisfy

$$\sum_{\nu=1}^{\infty} \nu^{2m} h_\nu^2 = \theta_m < \infty$$

where

$$h_\nu = \int_T \phi_\nu(t) h(t)\, dt.$$

$\mathscr{H}_m$ is the reproducing kernel Hilbert space with reproducing kernel

$$R(s, t) = \sum_{\nu=1}^{\infty} \frac{\phi_\nu(s)\phi_\nu(t)}{\nu^{2m}}.$$

(See [27]).

The orthogonal series density estimate $f_{n,m,\lambda}$ is

$$f_{n,m,\lambda}(t) = \sum_{\nu=1}^{n} \frac{\lambda_\nu}{\lambda_\nu + \lambda} \hat{f}_\nu \phi_\nu(t)$$

where

$$\lambda_\nu = \nu^{-2m},$$

$$\hat{f}_\nu = \frac{1}{n} \sum_{j=1}^{n} \phi_\nu(X_j),$$

and the integrated mean square error is

$$T_{n,m}(\lambda) = \int_T (f_{n,m,\lambda}(t) - f(t))^2\, dt$$

$$= \sum_{\nu=1}^{n} \left(\frac{\lambda_\nu}{\lambda_\nu + \lambda} (\hat{f}_\nu - f_\nu) - \frac{\lambda}{\lambda_\nu + \lambda} f_\nu \right)^2 + \sum_{\nu=n+1}^{\infty} f_\nu^2.$$

THEOREM 6.1. *Let $f \in \mathscr{H}_{\tilde{m}}$. Then, for any m with $m_0 \leq m \leq \tilde{m}$, the expected integrated mean square error $ET_{n,m}(\lambda)$ of $f_{n,m,\lambda}$ satisfies*

$$ET_{m,n}(\lambda) \leq \lambda\theta_m + \frac{k_m}{n\lambda^{1/2m}} (\theta_{\tilde{m}} M_0)^{1/2} + \frac{\theta_{\tilde{m}}}{n^{2\tilde{m}}}$$

where

$$k_m = \frac{1}{\pi} \int_0^{\infty} \frac{dx}{(1 + x^{2m})^2}$$

$$\theta_m = \sum_\nu^{\infty} \nu^{2m} f_\nu^2 \;(= \|f\|_{\mathscr{H}_m}^2)$$

and $(\theta_{\tilde{m}} M_0)^{1/2}$ is a bound on $\sup_t f(t)$.

Thus, if $\lambda = O(n^{-2m/(2m+1)})$, then

$$ET_{m,n} = O(n^{-2m/(2m+1)}).$$

PROOF. The proof follows that of Theorem 3.1 and we only give details that are different.

We have

$$E\hat{f}_\nu = \int_T \phi_\nu(x) f(x)\, dx = f_\nu$$

$$E(\hat{f}_\nu - f_\nu)^2 = \frac{1}{n^2} E \sum_{j=1}^{n} \sum_{k=1}^{n} \phi_\nu(X_j)\phi_\nu(X_k) - f_\nu^2$$

$$= \frac{1}{n}(g_\nu - f_\nu^2)$$

where

$$g_\nu = \int_T \phi_\nu^2(x) f(x)\, dx.$$

(g_ν was always 1 in the Fourier series estimate). Now $0 \le g_\nu \le (\theta_{\tilde{m}} M_0)^{1/2}$ since

$$g_\nu \le \sup_x f(x) \int_T \phi_\nu^2(x)\, dx = \sup_x f(x)$$

and

$$f(x) = \sum_{\nu=1}^{\infty} f_\nu \phi_\nu(x) \le \left[\sum_{\nu=1}^{\infty} \nu^{2m} f_\nu^2 \sum_{\nu=1}^{\infty} \frac{\phi_\nu^2(x)}{\nu^{2m}} \right]^{1/2} \le [\theta_{\tilde{m}} M_0]^{1/2}.$$

Thus

$$ET_{n,m}(\lambda) = \sum_{\nu=1}^{n} \left(\frac{\lambda_\nu}{\lambda_\nu + \lambda} \right)^2 E(\hat{f}_\nu - f_\nu)^2$$

$$+ \sum_{\nu=1}^{n} \left(\frac{\lambda}{\lambda_\nu + \lambda} \right)^2 f_\nu^2 + \sum_{\nu=n+1}^{\infty} |f_\nu|^2$$

$$(6.1) \qquad = \left\{ \sum_{\nu=1}^{n} \left\{ \left(\frac{\lambda}{\lambda_\nu + \lambda} \right)^2 - \frac{1}{n} \left(\frac{\lambda_\nu}{\lambda_\nu + \lambda} \right)^2 \right\} f_\nu^2 \right.$$

$$\left. + \frac{1}{n} \sum_{\nu=1}^{n} \left(\frac{\lambda_\nu}{\lambda_\nu + \lambda} \right)^2 g_\nu \right\} + \sum_{\nu=n+1}^{\infty} |f_\nu|^2$$

$$\le \lambda \theta_{\tilde{m}} + \frac{k_m}{n\lambda^{1/2m}} (\theta_{\tilde{m}} M_0)^{1/2} + \frac{\theta_{\tilde{m}}}{n^{2\tilde{m}}}.$$

THEOREM 6.2. *Let $f \in \mathscr{H}_{\tilde{m}}$. Then $\hat{T}_{n,m}(\lambda)$ defined by*

$$(6.2) \qquad \hat{T}_{n,m}(\lambda) = \frac{n}{n-1} \sum_{\nu=1}^{n} \left\{ \left(\frac{\lambda}{\lambda_\nu + \lambda} \right)^2 - \frac{1}{n} \left(\frac{\lambda_\nu}{\lambda_\nu + \lambda} \right)^2 \right\} \hat{f}_\nu^2$$

$$+ \frac{1}{n-1} \sum_{\nu=1}^{n} \left\{ \left(\frac{\lambda_\nu}{\lambda_\nu + \lambda} \right)^2 - \left(\frac{\lambda}{\lambda_\nu + \lambda} \right)^2 \right\} \hat{g}_\nu,$$

where

$$\hat{g}_\nu = \frac{1}{n} \sum_{j=1}^{n} \phi_\nu^2(X_j)$$

satisfies

$$E\hat{T}_{n,m}(\lambda) = ET_{n,m}(\lambda) + O\left(\frac{1}{n^{2\tilde{m}}} \right).$$

PROOF. Take the expectation of (6.2), substitute in $E\hat{f}_\nu^2 = (1/n)g_\nu + (1 - (1/n))f_\nu^2$ and $E\hat{g}_\nu = g_\nu$ and compare with (6.1).

These results can be used, for example, to estimate densities supported on the real line with the $\{\phi_\nu\}$ the Hermite functions, and Theorems 6.1 and 6.2 can be applied by using properties of the Hermite functions given in [21, 37].

We leave as an open question whether $O(n^{-2m/(2m+1)})$ is the best or nearly best possible integrated mean square convergence rate for densities in $\mathscr{H}_m$. Note that this convergence rate depends only on the rate of decay of the eigenvalues λ_ν $(=\nu^{-2m})$ independent of the nature of T.

We remark that Good and Gaskins [11] have used both Fourier series and Hermite series in the computation of their estimates and found Fourier series far preferable. I. Chang (unpublished) has done a few rather nondefinitive Monte Carlo experiments with Hermite series, but the same conclusion was suggested. Good and Gaskins give an explanation, another possible explanation is that the problem is the unfavorable L_∞ convergence properties of Hermite series on $[-\infty, \infty]$ see, e.g., Askey and Wainger [1]. Thus, in many applications it might be preferable to assume the true density has compact support and to scale the data to the interior of [0, 1].

7. Bivariate orthogonal series density estimates analogous to bivariate thin plate splines. A theory of bivariate smoothing splines has recently been developed based on the minimization problem: find f in a suitable function space to minimize

$$\frac{1}{n}\sum_{j=1}^{n} (f(s_j, t_j) - Y_j)^2 + \lambda \int\int \sum_{k=0}^{m} \binom{m}{k} \left(\frac{\partial^m f}{\partial s^k \partial t^{m-k}}\right)^2 ds\, dt. \tag{7.1}$$

See Wahba [32], and reference cited there. The solution to this minimization problem is known as a thin plate spline. We briefly remark on the doubly periodic orthogonal series density estimate on the unit square which is, approximately, the related doubly periodic spline function. For a little more generality, replace the second term in (7.1) by

$$\int_0^1 \int_0^1 \sum_{k=0}^{m} \alpha^k \beta^{m-k} \binom{m}{k} \left(\frac{\partial^m f}{\partial s^k \partial t^{m-k}}\right)^2 ds\, dt.$$

Then one is led to the density estimate

$$f_{n,m,\alpha,\beta}(s, t) = 1 + \sum_{**} \frac{\hat{f}_{\mu\nu}}{1 + [\alpha(2\pi\mu)^2 + \beta(2\pi\nu)^2]^m} \phi_{\mu\nu}(s, t) \tag{7.2}$$

where $\phi_{\mu\nu}(s, t) = \phi_\mu(s)\phi_\nu(t)$, $\hat{f}_{\mu\nu} = (1/n)\sum_{j=1}^{n} \phi^*_{\mu\nu}(\mathbf{X}_j)$, $\sum_{**}$ indicates an appropriate sum over (approximately) n values of the pair (μ, ν) for which $\mu^2 + \nu^2$ is smallest, and $\mathbf{X}_j$ is the jth (bivariate) observation.

One is led to (7.2) by replacing $\lambda/\lambda_\nu = \lambda(2\pi\nu)^{2m}$ in the univariate estimate of (1.6) by

$$\sum_{k=0}^{m} \alpha^k \beta^{m-k} \binom{m}{k} (2\pi\mu)^{2k} (2\pi\nu)^{2m-2k} \equiv [\alpha(2\pi\mu)^2 + \beta(2\pi\nu)^2]^m.$$

Letting

$$\Theta_{\mu\nu} = [\alpha(2\pi\mu)^2 + \beta(2\pi\nu)^2]^m,$$

to estimate m, α and β one minimizes

$$\begin{aligned} T_{n,m}(\alpha, \beta) &= \frac{n}{n-1} \sum_{**} \left\{ \left(\frac{\Theta_{\mu\nu}}{1 + \Theta_{\mu\nu}}\right)^2 - \frac{1}{n}\left(\frac{1}{1 + \Theta_{\mu\nu}}\right)^2 \right\} |\hat{f}_{\mu\nu}|^2 \\ &\quad + \frac{1}{n-1} \sum_{**} \left\{ \left(\frac{1}{1 + \Theta_{\mu\nu}}\right)^2 - \left(\frac{\Theta_{\mu\nu}}{1 + \Theta_{\mu\nu}}\right)^2 \right\}. \end{aligned} \tag{7.3}$$

Equation (7.3) was obtained from (4.1) by replacing λ/λ_ν by $\Theta_{\mu\nu}$.

Monte Carlo experiments with $\alpha = \beta (\equiv \lambda^{1/m})$ in the nonparametric regression context of

(7.1) have been very successful, see [33, 35]. A d-dimensional generalization of (7.2) and (7.3) is obtained by replacing $\phi_{\mu\nu}$ by

$$\phi_{\mu_1\mu_2\cdots\mu_d} = \phi_{\mu_1}\phi_{\mu_2}\cdots\ \phi_{\mu_d}$$

and $\Theta_{\mu\nu}$ by

$$\Theta_{\mu_1\mu_2\cdots\mu_d} = [\textstyle\sum_{l=1}^{d}\alpha_l(2\pi\mu_l)^2]^m,$$

to obtain IMSE convergence rates one rearranges the eigenvalues (which behave as a constant times $\Theta_{\mu_1\mu_2\cdots\mu_d}$) in size place, to find that they decay at the rate $n^{-2m/d}$, provided each $\alpha_l > 0$, (see [33]). Then provided $2m/d > 1$, $2m/d$ can be substituted for $2m$ in Theorem 6.1, giving an IMSE convergence rate $O(n^{-2m/(2m+d)})$.

REFERENCES

[1] Askey, R. and Wainger, S. (1965). Mean convergence of expansions in Laguerre and Hermite series. *Amer. J. Math.* **87** 695–708.

[2] Brunk, H. D. (1978). Univariate density estimation by orthogonal series. *Biometrica* **65** 521–528.

[3] Cogburn, R. and Davis, H. T. (1974). Periodic splines and spectra estimation. *Ann. Statist.* **2** 1108–1126.

[4] Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions, estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.* **31** 377–403.

[5] Davis, K. B. (1977). Mean integrated square error properties of density estimates. *Ann. Statist.* **5** 530–535.

[6] Duin, R. P. W. (1976). On the choice of smoothing parameters for Parzen estimators of probability density functions. *IEEE Trans. Computers* **C-25** 1175–1179.

[7] Fellner, W. H. (1974). Heuristic estimation of probability densities. *Biometrika* **61** 485–492.

[8] Gamber, H. A. (1979). Choice of an optimal shape parameter when smoothing noisy data. *Comm. Statist.* **A8** 1425–1436.

[9] Golub, G., Heath, M. and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21** 215–223.

[10] Good, I. J. and Gaskins, R. A. (1971). Nonparametric roughness penalties for probability densities. *Biometrika* **58** 255–277.

[11] Good, I. J. and Gaskins, R. A. (1980). Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. *J. Amer. Statist. Assoc.* **75** 42–55.

[12] Goodman, N. R. (1963). Statistical analysis based on a certain multivariate complex Gaussian distribution. *Ann. Math. Statist.* **34** 152–177.

[13] Hajek, J. (1958). On a property of normal distribution of any stochastic process. *Czechoslovak Math. J.* **8** 610–618. (English translation in *Selected Translations in Mathematical Statistics and Probability, Vol. 1.* 245–251. IMS and AMS Translation Series S.)

[14] Hermans, J. and Habbema, J. D. F. (1976). Manual for the ALLOC discriminant analysis programs. Depart. Medical Statist., Univ. Leiden, Netherlands.

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

[16] Leonard, T. (1978). Density estimation, stochastic processes and prior information. *J. Roy. Statist. Soc. Ser. B.* **40**. 113–146.

[17] Mallows, C. L. (1973). Some comments on C_p. *Technometrics* **15** 661–675.

[18] Parzen, E. (1961). On the estimation of a probability density function and mode. *Ann. Math. Statist.* **33** 1065–1076.

[19] Parzen, E. (1979). Nonparametric statistical data modeling. *J. Amer. Statist. Assoc.* **74** 105–121.

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

[21] Schwartz, S. C. (1968). Estimation of a probability density by an orthogonal series. *Ann. Math. Statist.* **38** 1261–1265.

[22] Scott, D. W. (1980). Comment on "Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data," by I. J. Good and R. A. Gaskins. *J. Amer. Statist. Assoc.* **75** 61–62.

[23] Scott, D. W. and Factor, L. E. (1979). Monte Carlo study of three data based nonparametric probability density estimators. To appear, *J. Amer. Statist. Assoc.*

[24] Scott, D. W., Tapia, R. A. and Thompson, J. R. (1977). Kernel density estimation revisited. In *Nonlinear Analysis, Theory, Methods and Applications* **1** 339–372.

[25] Tarter, M. E. (1979). Trigonometric maximum likelihood estimation and application to the analysis of incomplete survival information. *J. Amer. Statist. Assoc.* **74** 132–139.

[26] Tarter, M. E. and Kronmal, R. A. (1976). An introduction to the implementation and theory of nonparametric density estimation. *Amer. Statist.* **30** 105–112.

[27] Wahba, G. (1973). Convergence rates for certain approximate solutions to Fredholm integral equations of the first kind. *J. Approximation Theory* **6** 167–185.

[28] Wahba, G. (1974). Regression design for some equivalence classes of kernels. *Ann. Statist.* **2** 925–934.

[29] Wahba, G. (1975a). Cross-validation and empirical Bayes estimation for optimal rate density estimation (preliminary report). Abstract. *Bull. Inst. Math. Statist.* **4** 257.

[30] Wahba, G. (1975b). Optimal convergence properties of variable knot, kernel and orthogonal series methods for density estimation. *Ann. Statist.* **3** 15–29.

[31] Wahba, G. (1977). Optimal smoothing of density estimates. In *Classification and Clustering.* (ed, J. Van Ryzin) 423–458. Academic Press.

[32] Wahba, G. (1979a). How to smooth curves and surfaces with splines and cross-validation. Depart. Statist., Univ. Wisconsin-Madison Technical Report #555; also in *Proceedings of the 24th Conference on the Design of Experiments in Army Research Development and Testing.* Report No. 79-2, USARO, Research Triangle Park, North Carolina.

[33] Wahba, G. (1979b). Convergence rates of "thin plate" smoothing splines when the data are noisy. In *Smoothing Techniques for Curve Estimation. Lecture Notes in Math.* **757** 233–245. (Th. Gaser and M. Rosenblatt, eds.). Springer, Berlin.

[34] Wahba, G. (1980). Automatic smoothing of the log periodogram. *J. Amer. Statist. Soc.* **75** 122–131.

[35] Wahba, G. and Wendelberger, J. (1979). Some new mathematical methods for variational objective analysis. To appear, *Monthly Weather Review.*

[36] Wahba, G. and Wold, S. (1975). Periodic splines for spectral density estimation, the use of cross-validation for determining the degree of smoothing. *Commun. Statist.* **4** 125–141.

[37] Walter, G. G. (1977). Properties of Hermite series estimation of probability density. *Ann. Statist.* **5** 1258–1264.

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

[39] Woodroofe, M. (1970). On choosing a Delta sequence. *Ann. Math. Statist.* **41** 1655–1671.

Department of Statistics
University of Wisconsin
1210 W. Dayton St.
Madison, Wisconsin 53706