DEPARTMENT OF STATISTICS

TECHNICAL REPORT NO. 509

February 1978

DATA-BASED OPTIMAL SMOOTHING OF ORTHOGONAL SERIES DENSITY ESTIMATES

by

Grace Wahba

UNIVERSITY OF WISCONSIN-MADISON

Data-Based Optimal Smoothing of Orthogonal Series Density Estimates

by

Grace Wahba

Abstract

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 $\hat{f}_\nu = \frac{1}{n}\sum_{j=1}^{n} \phi_\nu(X_j)$ 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 <u>objective</u> manner, to minimize integrated mean square error.

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

1. Introduction and statement of results.

Let f be a square integrable density on [0,1] possessing a Fourier series expansion

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

where $\phi_\nu(t) = e^{2\pi i\nu t}$. It is desired to estimate f from n independent observations $X_1,\ldots,X_n$ from the density f. Given a sequence $\underset{\sim}{b} = (\ldots, b_{-2}, b_{-1}, b_1, b_2, \ldots)$ of nonnegative real numbers with $b_\nu = b_{-\nu}$, and $\sum|b_\nu| < \infty$ the orthogonal series estimate $\hat{f}_{n,\underset{\sim}{b}}$ of f as considered by Whittle [26], Kronmal and Tarter [10], Brunk [1], Fellner [4] and others can be written in the form (n even)

$$\hat{f}_{n,\underset{\sim}{b}}(t) = 1 + \sum_{*} \frac{b_\nu}{b_\nu + \frac{1}{n}} \hat{f}_\nu \phi_\nu(t) , \tag{1.2}$$

where

$$\sum_{*} = \sum_{\substack{\nu=-n/2 \\ \nu\neq 0}}^{n/2}$$

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,

$$f_\nu \sim N_c(0,b_\nu), \text{ independent}, \qquad \nu = 1,2,\ldots \tag{1.3}$$

(N_c is the complex normal distribution, see [7]). Then, since

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

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

if one approximates $\frac{1}{n}(1-|f_\nu|^2)$ by $\frac{1}{n}$ 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 $\hat{f}_{n,\underset{\sim}{b}}(t)$ 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 non-negative, although they do integrate to 1. Motivation as a smoothing spline estimate will be discussed later.

Various specifications have been proposed in [1,10,17,26] for the $\underset{\sim}{b} = (\ldots,b_{-2},b_{-1},b_1,b_2,\ldots)$ which determines the prior. In this note we propose a two parameter family of $\underset{\sim}{b}$'s, namely, $\{b_\nu = (\lambda(2\pi\nu)^{2m})^{-1} \quad \nu = \pm1,2,\ldots\}$, $\lambda \geq 0$, $m > 1/2$. We will write the resulting estimate as $f_{n,\lambda,m}$,

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

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 $\underset{\sim}{b}$'s have distinct values of (λ,m), their associated (infinite-dimensional) prior distributions are perpendicular, further-more, 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, thus supporting 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

$$W_2^{(m)}\text{ (per)} = \{f\colon f,f',\ldots,f^{(m-1)} \text{ abs. cont.}, f^{(m)} \in L_2[0,1] ,$$

$$f^{(\nu)}(0) = f^{(\nu)}(1),\ \nu = 0,1,\ldots,m-1\} ,$$

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 $E\int_0^1[f_{n,m,\lambda}(t)-f(t)]^2dt$ satisfies

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

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 [20]), 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 nice, as side issues, as a major problem in density estimation is to choose the smoothing parameter(s), which are a part of every density estimate (see [20]), <u>objectively from the data</u>, to approximately <u>minimize some optimality criteria</u>. 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 [27] 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 [6] suggest a method for choosing the smoothing parameter in a penalized maximum likelihood method, based on a goodness-of-fit criteria. Scott, Tapia and Thompson [16], Leonard [11], Fellner [4], Brunk [1] and Tarter and Kronmal [17] discuss procedures for choosing the degree of smoothness in various estimates, which involve varying degrees of subjectivity, only Woodroofe attempts explicitly to minimize m.s.e., however. Two readily computable completely objective methods using cross-validation to determine the degree of smoothing (Hermans and Habbema [9], and Wahba [18]) will be discussed at the conclusion of this note.

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

Theorem 4.1. Let

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

If $f \in W_2^{(\tilde{m})}$ (per) for some $\tilde{m} \geq 1/2$, then $\hat{T}_{n,m}(\lambda)$, defined by

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

$$- \frac{1}{n-1} \sum_{*} \{(\frac{\lambda}{\lambda_\nu+\lambda})^2 - (\frac{\lambda_\nu}{\lambda_\nu+\lambda})^2\} ,$$

where

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

satisfies

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

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 it does not seem to have been recognized before in this context, and is exceedingly useful. It should, however be considered to be in the spirit of Mallow's C_L method [12] for choosing the ridge parameter in ridge regression. (See [5], eqn. (1.8) for further details).

In the final section of this note we generalize these 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 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 $f_1, f_2, \ldots$ be an infinite sequence of independent, zero mean normally distributed random variables with probability measure denoted by $P_{m,\lambda}$ if $E\,f_\nu^2 = [\lambda(2\pi\nu)^{2m}]^{-1}$, $\nu = 1,2,\ldots$.

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

ii) Let $P_{\underset{\sim}{b}}$ be the probability measure corresponding to $E f_\nu^2 = b_\nu$, $\nu = 1,2,\ldots$, where

$$b_\nu = \left(\left|\sum_{j=0}^{q} \beta_j \nu^j\right|^2 \Big/ \left|\sum_{j=0}^{p} \alpha_j \nu^j\right|^2\right)(1 + O(\nu^{-1})), \quad \nu = 1,2,\ldots$$

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

Proof: This is a consequence of Hajek [8] who proves that, for any $\underset{\sim}{b}(j) = (b_1(j), b_2(j), \ldots)$, $j = 1,2$, $P_{\underset{\sim}{b}(1)} \equiv P_{\underset{\sim}{b}(2)}$ if

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

and $P_{\underset{\sim}{b}(1)} \perp P_{\underset{\sim}{b}(2)}$ 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

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

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 $E\,T_{m,n}(\lambda)$ of $f_{n,m,\lambda}$ satisfies

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

where

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

and

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

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

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

Proof: By Parseval's Theorem

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

$$= \sum_* \left|\frac{\lambda_\nu}{\lambda_\nu+\lambda}\hat{f}_\nu - f_\nu\right|^2 + \sum_{|\nu|>n/2} |f_\nu|^2 , \tag{3.1}$$

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

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

$$= \sum_*\left\{\left(\frac{\lambda}{\lambda_\nu+\lambda}\right)^2 - \frac{1}{n}\left(\frac{\lambda_\nu}{\lambda_\nu+\lambda}\right)^2\right\}|f_\nu|^2 + \frac{1}{n}\sum_*\left(\frac{\lambda_\nu}{\lambda_\nu+\lambda}\right)^2 + \sum_{|\nu|>n/2}|f_\nu|^2 . \tag{3.2}$$

The theorem follows upon noting that

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

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

and

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

We remark that this estimate (and the integrated mean square error convergence rate) essentially appear in Cogburn and Davis [3] and Wahba and Wold [22] 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 $\frac{1}{n}\sum_{j=1}^{n}(f(\frac{j}{n})-Y_j)^2 + \lambda\int_0^1 (f^{(m)}(t))^2dt$, where $Y_j \equiv f_{n,0}(\frac{j}{n})$, and so $f_{n,m,\lambda}$ is (approximately) a periodic spline function (see [3,22]). The method for choosing λ and m of this note also can be applied to the log spectral density estimate described in [3,22], see [19].

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

Theorem 4.1. Let

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

$$+ \frac{1}{n-1}\sum_{*}\left\{\left(\frac{\lambda_\nu}{\lambda_\nu+\lambda}\right)^2 - \left(\frac{\lambda}{\lambda_\nu+\lambda}\right)^2\right\}\ . \tag{4.1}$$

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

$$E\,\hat{T}_{n,m}(\lambda) = E\,T_{n,m}(\lambda) + O\left(\frac{1}{n^{2\tilde{m}}}\right)\ . \qquad \begin{array}{l}\lambda \ge 0\\ m > 1/2\end{array}$$

Proof:

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

$$= \frac{n}{n^2}\int_0^1 f(u)du + \frac{n^2-n}{n^2}\int_0^1 \phi_\nu(u)f(u)du \int_0^1 \phi_\nu^*(u)f(u)du$$

$$= \frac{1}{n} + \frac{n-1}{n}\,|f_\nu|^2\ . \tag{4.2}$$

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. Remarks.

This estimate would be an ideal, easily computable, all purpose density estimate for 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 $f_{n,m,\lambda}^{(\nu)}(0) = f_{n,m,\lambda}^{(\nu)}(1)$, $\nu = 0,1,\ldots,m-1$. 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 be avoided by other choices of orthogonal series, see section 6. Monte Carlo experiments with $f_{n,m,\lambda}$ with m fixed at 2 and the method of generalized cross validation (GCV) used to estimate

λ, were reported in [18], with excellent results. Since the GCV method is also estimating the minimizer of $T_{n,m}(\lambda)$ (but via a further approximation, see [2,5,22]) results minimizing $\hat{T}_{n,m}(\lambda)$ should be as good, if not better than those reported for GCV in [22]. Monte Carlo experiments in progress in the context of smoothing splines indicate that, if m is fixed, the integrated mean square error is sensitive to λ, but as m varies over small integers (carrying along the optimum λ in each case) the actual integrated mean square error changes only slightly ($n \sim 100$). $\hat{T}_{n,m}(\lambda)$ can be negative at its minimum, but numerical results indicate that the minimizer is still a good estimate of the minimizer of $T_{n,m}(\lambda)$.

Hermans and Habbema [9] choose h in a kernel estimate (see [13,14]) of the form

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

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

$$V(h) = \prod_{k=1}^{n} f_h^{(k)}(X_k)$$

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

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

They use the normal kernel $K(\tau) = \frac{1}{\sqrt{2\pi}} e^{-\tau^2/2}$. The method is adapted easily to certain multidimensional kernel estimates. It remains to determine the properties of this method, and how it compares with the present approach in practice.

6. Abstract orthogonal series density estimates with optimal smoothing.

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 $L_2(T)$, and assume

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

for all $m \geq m_0$ where $m_0 \geq \frac{1}{2}$. Let H_m be the collection of functions $\{h\}$ in $L_2[T]$ which further satisfy

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

where

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

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 [24])

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 $f \in H_{\tilde{m}}$. Then, for any m with $m_0 \le m \le \tilde{m}$, the expected integrated mean square error $E\,T_{n,m}(\lambda)$ of $f_{n,m,\lambda}$ satisfies

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

where

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

$$\theta_m = \sum_\nu \nu^{2m} f_\nu^2 \;(=\|f\|^2_{H_m})$$

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

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

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 $0 \le g_\nu \le \sqrt{\theta_{\tilde{m}} M_0}$ 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

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

$$= \left\{\sum_{\nu=1}^{n}\left\{\left(\frac{\lambda}{\lambda_\nu+\lambda}\right)^2 - \frac{1}{n}\left(\frac{\lambda_\nu}{\lambda_\nu+\lambda}\right)^2\right\} f_\nu^2 + \frac{1}{n}\sum_{\nu=1}^{n}\left(\frac{\lambda_\nu}{\lambda_\nu+\lambda}\right)^2 g_\nu\right\} + \sum_{\nu=n+1}^{\infty} |f_\nu|^2 \tag{6.1}$$

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

Theorem 6.2. Let $f \in H_{\tilde{m}}$. Then $\hat{T}_{n,m}(\lambda)$ defined by

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

where

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

satisfies

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

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

These results can be used e.g. 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 [15,25].

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 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.

Some of the ideas of this section appeared in the Abstract [21].

References

[1] Brunk, H.D., Univariate density estimation by orthogonal series, with application to estimation of wildlife populations by line transect surveys. Manuscript (1977).

[2] Craven, P., and Wahba, G., Smoothing noisy data with spline functions, estimating the correct degree of smoothing by the method of generalized cross-validation. Dept. of Statistics, University of Wisconsin-Madison Technical Report #445, October 1977.

[3] Cogburn, R., and Davis, H.T., Periodic splines and spectra estimation, Ann. Statist., 2, 6, 1108-1126 (1974).

[4] Fellner, W.H., Heuristic estimation of probability densities, Biometrika; 61, 3, 485-492 (1974).

[5] Golub, G., Health, M., and Wahba, G., Generalized cross-validation as a method for choosing a good ridge parameter. University of Wisconsin-Madison Statistics Department Technical Report #491 (1977).

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

[7] Goodman, N.R., Statistical analysis based on a certain multivariate complex Gaussian distribution, Ann. Math. Statist. 34, 1, 152-177 (1963).

[8] Hajek, J., On a property of normal distributions of any stochastic process, Czech Math J. 8, 610-618 (1958), translated in Selected Translations in Mathematical Statistics and Probability, Vol. 1, 245-251, IMS and AMS Translation Series S.

[9] Hermans, J., and Habbema, J.D.F., Manual for the ALLOC discriminant analysis programs, Department of Medical Statistics University of Leiden, Netherlands (1976).

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

[11] Leonard, T, Density estimation, stochastic processes and prior information. J. Roy. Stat. Soc. B 40, 1, (1978), to appear.

[12] Mallows, C.L., Some comments on C_p. Technometrics 15, 661-675 (1973).

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

[14] Rosenblatt, M., Remarks on some non-parametric estimates of a density function, Ann. Math. Statist. 27, 832-837 (1956).

[15] Schwartz, S.C., Estimation of a probability density by an orthogonal series, Ann. Math. Statist. 38, 4, 1261-1265 (1968).

[16] Scott, D.W., Tapia, R.A., and Thompson, J.R., Kernel density estimation revisited. Nonlinear Analysis, Theory, Methods and Applications, 1, 4, 339-372 (1977).

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

[18] Wahba, G., Optimal smoothing of density estimates, in Classification and Clustering, J. Van Ryzin, Ed., 423-458 Academic Press (1977).

[19] Wahba, G., Optimal smoothing of log spectral density estimates, Abstract in Preprints of the Fifth Conference on Probability and Statistics in Atmospheric Sciences, Nov. 15-18, 1977, Las Vegas, Nevada. Technical report in preparation.

[20] Wahba, G., Optimal convergence properties of variable knot, kernel and orthogonal series methods for density estimation, Ann. Statist. 3, 1, 15-29 (1975).

[21] Wahba, G., Cross-validation and empirical Bayes estimates for optimal rate density estimation (preliminary report). Abstract, Bull. IMS, 4, 6, #23, p. 257 (1975).

[22] Wahba, G., and Wold, S., Periodic splines for spectral density estimation, the use of cross-validation for determining the degree of smoothing. Communications in Statistics, 4, 2 125-141 (1975).

[23] Wahba, G., Regression design for some equivalence classes of kernels, Ann. Statist. 2, 5, 925-934 (1974).

[24] Wahba, G., Convergence rates for certain approximate solutions to Fredholm integral equations of the first kind, J. Approximation Theory 7, 167-185 (1973).

[25] Walter, G.G., Properties of Hermite series estimation of probability density, Ann. Statist. 5, 6, 1258-1264 (1977).

[26] Whittle, P., On the smoothing of probability density functions, J. Roy. Stat. Soc. B 20, 334-343 (1958).

[27] Woodroofe, M., On choosing a Delta-sequence, Ann. Math. Statist. 41, 5, 1665-1671 (1970).

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER Technical Report No. 509	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) DATA-BASED OPTIMAL SMOOTHING OF ORTHOGONAL SERIES DENSITY ESTIMATES		5. TYPE OF REPORT & PERIOD COVERED 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Grace Wahba		8. CONTRACT OR GRANT NUMBER(s) DAAG29-77-G-0207
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics University of Wisconsin Madison, Wisconsin		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office Research Triangle Park, N.C.		12. REPORT DATE February 1978 13. NUMBER OF PAGES 17
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Distribution of this document is unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

None

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

optimal smoothing of density estimates
orthogonal series density estimates

20. ABSTRACT (Continue on reverse side if necessary and identify by block number) 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 orthogonal sequence and $\hat{f}_\nu = \frac{1}{n}\sum_{j=1}^{n} \phi_\nu(X_j)$ is an estimate of

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20 continued:

$f_\nu = \int \phi_\nu(t) f(t) dt$, 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.