
DEPARTMENT OF STATISTICS

University of Wisconsin
Madison, Wisconsin 53706

TECHNICAL REPORT NO. 469

October 1976

OPTIMAL SMOOTHING OF DENSITY ESTIMATES

by

Grace Wahba
University of Wisconsin-Madison

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

OPTIMAL SMOOTHING OF DENSITY ESTIMATES

by

Grace Wahba
University of Wisconsin

1. Introduction.

1.1. Density estimation in classification.

The Neyman-Pearson lemma tells us that, if we want to classify an object as coming from one of two possible populations with associated densities f_1 and f_2, then we should base the classification on the likelihood ratio f_1/f_2. Frequently this leads to simple and effective algorithms. For example if each of the two densities can be assumed to be multivariate normal then the likelihood ratio is constant along hyperplanes or hyperquadratics in Euclidean p-space. Then the optimal classification rule consists of determining which side of a certain hyperplane or hyperquadratic of constant likelihood the measurement vector to be classified lies. If the means and covariances are not known a priori, the problem of "learning" the optimal classification rule from the data reduces to "learning" an appropriate hyperplane or hyperquadratic.

This procedure is in common use, and gives satisfactory results for a wide class of non-normal densities. However, if the underlying densities have, for example, C or S shaped ridges or multiple modes, then one would like to estimate the likelihood f_1/f_2. (See Chi and Van Ryzin [5] who present a simple consistent non-parametric procedure for estimating f_1/f_2 in several dimensions.)

In this paper we approach the problem of estimating the likelihood ratio from the point of view of estimating the individual densities. For the purpose of classification there is no obvious advantage to estimating the densities as an intermediate step, other than that the state of the art

---

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

of density estimation seems to be further along than that of likelihood ratio estimation. In fact we would be pleased if the methods described here could be applied directly to the estimation of the likelihood ratio.

1.2. The major types of density estimation.

There is a very extensive literature on one dimensional and multidimensional density estimation, we mention only a few of the many papers. The early papers, on kernel methods, are Whittle [47], Rosenblatt [27], Parzen [24]. Orthogonal series estimates have been discussed by Watson [44], Kronmal and Tarter [20], and recently by Brunk [4] and Crain [8]. Boneva, Kendall and Stefanov [3] introduced spline methods, see also Lii and Rosenblatt [21], Wahba [35,39]. The k-nearest neighbor methods were introduced by Loftsgarten and Quesenberry [22], see also Van Ryzin [31], Wahba [32,39]. Penalized maximum likelihood methods were treated by Good and Gaskins [15], and de Montricher, Tapia and Thompson, [10] showed their relationship to spline methods. Recently, Walter and Blum [42] have provided a framework which unifies some of these methods, see also Parzen [25]. A survey and some Monte Carlo studies appear in Wegman [45,46]. For an extensive bibliography and survey, see Cover and Wagner [7]. The above list has no pretentions to completeness.

1.3. The smoothing parameter in kernel, orthogonal series, k-nearest neighbor, histospline, and penalized maximum likelihood methods.

The (univariate) kernel estimate $\hat{f}$ is of the form

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)$$

where $X_1, X_2, \ldots, X_n$ are independent, identically distributed observations from the density f, and $K(\cdot)$ is a "hill" function integrating to one and satisfying some regularity conditions. See Parzen [24]. The experimenter must choose the parameter h, which will control the visual smoothness of the resulting density. Mathematically, h controls the tradeoff between the squared bias and the variance. The optimal h from the point of view of mean square error depends on both the sample

size and on the unknown density. Woodroofe [48] has given a theoretical approach to choosing h from the data, but it appears that the method is not practical. In practice, if f is a univariate density, h can frequently be chosen visually in a satisfactory manner.

The orthogonal series estimate for densities supported on [0,1] is given by

$$\hat{f}(x) = \sum_{\nu=1}^{r} \hat{f}_\nu \phi_\nu(x)$$

where $r << n$, the $\hat{f}_\nu$ are the sample Fourier coefficients

$$\hat{f}_\nu = \frac{1}{n} \sum_{i=1}^{n} \phi_\nu(X_i)$$

and the $\{\phi_\nu\}$ are a set of $\mathcal{L}_2[0,1]$ orthonormal functions. See Kronmal and Tarter [20]. The parameter r, which must be chosen, controls the tradeoff between the square bias and the variance, small r means small variance but possibly large bias, large r means small bias at the expense of large variance. Again the optimal r depends on the sample size and the unknown density. Similarly every non-parametric density estimate (or estimate of the likelihood ratio) has a parameter (sometimes hidden!) which must be chosen by the experimenter, and which controls the tradeoff between the square bias and the variance. In the k-nearest neighbor methods, the parameter is k, and in the Boneva-Kendall-Stefanov histospline (or the histogram, for that matter) the parameter is the "bin size". In penalty function methods the parameter is a multiplier on the penalty.

Probably one major reason why multidimensional density estimates are not commonly used for classification in practice despite the fact that there is so much theoretical interest in the subject, is that the resulting classification methods cannot be made to "work" unless the smoothing parameter is chosen reasonably well, and this is hard to do visually in more than one dimension.

In this paper we present a "new" class of density estimates.†

†The example we treat in detail in Section 4 has previously been proposed by Cogburn and Davis [6] in connection with spectral density estimation.

In truth, a new class of density estimates is not really needed at the present state of the art of density estimation. There are plenty of perfectly good old ones around. Furthermore, if the density to be estimated is known to be smooth, then there is an upper bound to convergence rates for mean square error and most of the methods mentioned above are known to essentially attain it. So no one is likely to come up with a startlingly better nonparametric method. To be specific, let $\mathcal{C}_{m,p}$ be the class of densities defined by

$$\mathcal{C}_{m,p} = \{f : f \text{ a density}, \ f, f', \ldots, f^{(m-1)} \text{ abs. cont.}, \ f^{(m)} \in \mathcal{L}_p, \ \|f^{(m)}\|_p \leq M\},$$

where $\|\cdot\|_p$ is the norm in $\mathcal{L}_p$ and M is a fixed constant. Let f_n be any density estimate based on n independent observations from $f \in \mathcal{C}_{m,p}$. Then the rate of convergence of $\sup_{f \in \mathcal{C}_{m,p}} E(f_n(x) - f(x))^2$ cannot be better than $n^{-(2m-2/p)/(2m+1-2/p)-\varepsilon}$ for arbitrarily small ε. See Farrell [11], Wahba [34]. Furthermore, if the smoothing parameter is chosen correctly, then kernel methods, orthogonal series methods, histospline methods and certain examples of k-nearest neighbor methods are all known to achieve the rate $n^{-(2m-2/p)/(2m+1-2/p)}$. See Wahba [32], [35], [39]). Parzen [24] gave the rate $n^{-4/5}$ in 1963 for the kernel method for the case $m = 2$, $p = \infty$.

Why, then, do we test the reader's patience with yet another class of estimates? The method presented here appears to be as good as some of the others floating around, for medium sample sizes (on the basis of convergence rate calculations and some very preliminary Monte Carlo results). The difference between this class of methods and the others we know of is that it comes with a viable algorithm for estimating the optimal (integrated mean square error) smoothing parameter from the data.†

Denote by $f_{n,\lambda}$ the density estimate to be proposed in this paper, where n is the sample size, and λ is the smoothing parameter to

†I. J. Good has a procedure for selecting the parameter that goes with a penalized maximum likelihood method [15], and Tarter and Kronmal [49] have done the same for orthogonal series methods.

be chosen. Let $X_1, X_2, \ldots, X_n$ be n independent observations from a density f. We provide an estimate $\hat{\lambda} = \hat{\lambda}(X_1, X_2, \ldots, X_n)$ for λ^* where $\lambda^* = \lambda^*(n;f)$ is defined as the minimizer of the "integrated" mean square error,

$$E \frac{1}{n} \sum_{i=1}^{n} (f_{n,\lambda}(\tfrac{i}{n}) - f(\tfrac{i}{n}))^2 .$$

Some (relatively weak) theorems are presented on the properties of this estimator. In Monte Carlo studies the estimate works better than the Theorems indicate. Let $x_1, x_2, \ldots, x_n$ be realizations of $X_1, X_2, \ldots, X_n$, and $f_{n,\lambda}(t;z_1, z_2, \ldots, z_n)$ indicate the dependence of $f_{n,\lambda}$ on $z_1, z_2, \ldots, z_n$, where $z_i = X_i$ or x_i. We have observed in related studies [9][14] that $\hat{\lambda}(x_1, x_2, \ldots, x_n)$ approximates the minimizer of

$$(1.1) \qquad \frac{1}{n} \sum_{i=1}^{n} (f_{n,\lambda}(\tfrac{i}{n}; x_1, x_2, \ldots, x_n) - f(\tfrac{i}{n}))^2$$

<u>better</u> than it estimates the minimizer of

$$E \frac{1}{n} \sum_{i=1}^{m} (f_{n,\lambda}(\tfrac{i}{n}; X_1, X_2, \ldots, X_n) - f(\tfrac{i}{n}))^2 !$$

In the experiments reported on here we found that $\hat{\lambda}(x_1, \ldots, x_n)$ came almost as close to the minimizer of (1.1) as it is possible to get, we typically observed that the relative inefficiency of $\hat{\lambda}$, defined as

$$(1.2) \qquad \frac{\frac{1}{n} \sum_{i=1}^{n} (f_{n,\hat{\lambda}(x_1,\ldots,x_n)}(\tfrac{i}{n}; x_1, x_2, \ldots, x_n) - f(\tfrac{i}{n}))^2}{\inf_{\lambda} \frac{1}{n} \sum_{i=1}^{n} (f_{n,\lambda}(\tfrac{i}{n}; x_1, x_2, \ldots, x_n) - f(\tfrac{i}{n}))^2}$$

was between 1.0 and 1.1! (In these cases $n = 170$ and f is a mixture of Beta densities.)

The estimate $\hat{\lambda}$ of λ is based on the method of generalized cross validation (GCV), and this method has general applicability in other contexts, including curve smoothing (Wahba and Wold [40][41], surface smoothing (Wahba [36]), ridge regression (Golub, Heath, and Wahba [14]), and the approximate solution of linear operator equations when the data are noisy (Wahba [37]). These problems are all related to the problems of estimating the mean of a multivariate normal vector, which has

received much attention by Stein and co-workers, see Hudson [18] for a bibliography. Feinberg and Holland consider a similar problem for estimating cell probabilities [12].

In Section 2 of this paper we review the general problem of recovering a smooth curve from noisy data, without any reference whatever to density estimation. A smoothing parameter must be chosen, and we present the GCV method for choosing it from the data, and report some of the known theoretical properties of the method. In Section 3 we show how density estimation can be cast in the context of Section 2, that of recovering a smooth curve from noisy data, and in certain special cases we are able to obtain the corresponding theoretical properties of the estimate for λ in the context of density estimation. In Section 4 we present a few typical examples from a very modest Monte Carlo study. In Section 5 we outline future work remaining to be done on the method.

2. Recovering a Smooth Curve from Noisy Data.

In this section we discuss the problem of recovering a smooth curve from noisy data, without reference to density estimation. In Section 3 we show how the density estimation problem can be put in the context of this section and the results of this section used to smooth density estimates optimally.

2.1. The models.

Consider the model

$$y(t) = f(t) + \varepsilon(t), \qquad t \in [0,1] \tag{2.1}$$

where f and ε are independent zero mean Gaussian stochastic processes with

$$E\, f(s)\, f(t) = b\, Q(s,t), \qquad s,t \in [0,1], \tag{2.2}$$

b is an unknown positive constant and Q is a known strictly positive definite continuous covariance, and

$$E\, \varepsilon(s)\varepsilon(t) = \sigma^2, \quad s = t; \qquad = 0, \quad s \neq t. \tag{2.3}$$

Suppose y of (2.1) is observed for $t = t_1, t_2, \ldots, t_n$ and it is desired to recover f. Then it follows from elementary principles that

$$E\{f(t) \mid y(t_i) = y_i,\ i = 1,2,\ldots,n\}$$
$$= (Q_{t_1}(t), Q_{t_2}(t), \ldots, Q_{t_n}(t))(Q_n + n\lambda I)^{-1} y \tag{2.4a}$$

where

$$Q_{t_i}(t) = Q(t_i, t),$$

Q_n is the $n \times n$ matrix with ijth entry

$$[Q_n]_{ij} = Q(t_i, t_j), \tag{2.4b}$$
$$\lambda = \sigma^2/nb,$$

and

$$y = (y_1, y_2, \ldots, y_n)'.$$

Next, suppose that instead of being a stochastic process, f is an element of $\mathscr{H}_Q$, where $\mathscr{H}_Q$ is the reproducing kernel Hilbert space (rkhs) with reproducing kernel (rk) Q. Define

$$f_{n,\lambda}(t) = (Q_{t_1}(t), \ldots, Q_{t_n}(t))(Q_n + n\lambda I)^{-1} y. \tag{2.5}$$

Then (see Kimeldorf and Wahba [19]) $f_{n,\lambda}$ is the solution to the minimization problem: Find $f \in \mathscr{H}_Q$ to minimize

$$\frac{1}{n} \sum_{j=1}^{n} (f(t_j) - y_j)^2 + \lambda \|f\|_Q^2, \qquad (y_j = y(t_j)) \tag{2.6}$$

where $\|\cdot\|_Q$ is the norm in $\mathscr{H}_Q$. Here λ may be thought of as controlling the tradeoff between smoothness as measured by $\|f\|_Q^2$ and infidelity to the data, as measured by

$$\frac{1}{n} \sum_{j=1}^{n} (f(t_j) - y_j)^2.$$

Thus, the "f is a stochastic process" point of view" and "f is an element of $\mathscr{H}_Q$" point of view, lead to the same form of algorithm for recovering f from noisy data, only the meaning given to λ is different. (We remark that it is well known that sample functions f from the stochastic model are, with probability 1, not in $\mathscr{H}_Q$).

Assume that Q is given and σ^2 and b are unknown. If f is a stochastic process then one must estimate $\lambda = \sigma^2/nb$ from the data in order to estimate f. On the other hand, if f is known to be in $\mathcal{H}_Q$, then we would like to estimate λ which minimizes some measure of the error. We adopt as our measure the average square error $T(\lambda)$ given by

$$T(\lambda) = \frac{1}{n}\sum_{j=1}^{n}(f_{n,\lambda}(t_j) - f(t_j))^2 . \tag{2.7}$$

We have obtained elsewhere (Wahba and Wold [40,41], Wahba [37], Craven and Wahba [9], and Golub, Heath and Wahba [14]) an estimate $\hat{\lambda} = \hat{\lambda}(y_1, y_2, \ldots, y_n)$ called the generalized cross validation (GCV) estimate of λ which has the remarkably nice property that if f is the stochastic process of (2.2) then $\hat{\lambda}$ estimates σ^2/nb, and if $f \in \mathcal{H}_Q$ then $\hat{\lambda}$ estimates the minimizer of $T(\lambda)$!

2.2. The Generalized Cross Validation (GCV) estimate $\hat{\lambda}$ for the smoothing parameter λ, and some of its properties.

In this subsection we give the GCV estimate $\hat{\lambda}$ for λ, and report some of its properties. We defer until the end of this section a discussion of where the estimate came from.

To define the estimate, let $A(\lambda)$ be the $n \times n$ matrix defined by

$$A(\lambda) = Q_n(Q_n + n\lambda I)^{-1} \tag{2.8}$$

and define the function $V(\lambda)$ as

$$V(\lambda) = \frac{\frac{1}{n}\|(I - A(\lambda))y\|_n^2}{[\frac{1}{n}\mathrm{Tr}\,(I - A(\lambda))]^2} \tag{2.9}$$

where $\|\cdot\|_n$ is the Euclidean n-space norm, and Q_n has been defined in (2.4b).

Definition of $\hat{\lambda}$: The GCV estimate $\hat{\lambda}$ of λ is defined as the value of λ which minimizes $V(\lambda)$ defined in (2.8) and (2.9).

The first property of $\hat{\lambda}$ is easy to verify, and we state the results as a theorem.

Theorem 1. Let f be the stochastic process given by (2.2). Let E_N denote expectation with respect to the noise random variables

$\varepsilon(t_1), \varepsilon(t_2), \ldots, \varepsilon(t_n))$ and E_S expectation with respect to the "signal" random variables $\{f(t_1), f(t_2), \ldots, f(t_n))\}$. Then $E_S E_N V(\lambda)$ and $E_S E_N T(\lambda)$ are both minimized for $\lambda = \sigma^2/nb$.

Proof: Let $\lambda_{\nu n}$, $\nu = 1, 2, \ldots, n$ be the eigenvalues of Q_n. Then

$$E_N V(\lambda) = \frac{\frac{1}{n}\|(I - A)f\|^2 + \sigma^2 \operatorname{Tr}(I - A)^2}{(\frac{1}{n}\operatorname{Tr}(I - A))^2} \tag{2.10}$$

$$E_S E_N V(\lambda) = \frac{\frac{b}{n}\operatorname{Tr}(I - A)Q_n(I - A) + \frac{\sigma^2}{n}\operatorname{Tr}(I - A)^2}{(\frac{1}{n}\operatorname{Tr}(I - A))^2} \tag{2.11}$$

$$= b\Big(\sum_{\nu=1}^{n} \frac{(\lambda_{\nu n}/n) + (\sigma^2/nb)}{((\lambda_{\nu n}/n) + \lambda)^2}\Big) \Big/ \Big(\sum_{\nu=1}^{n} \frac{1}{((\lambda_{\nu n}/n) + \lambda)}\Big)^2.$$

The minimum is achieved when $U_1 U_2' = U_2 U_1'$ where U_1 and U_2 are the numerator and denominator, respectively, in (2.11), or

$$\Big(\sum{}' \frac{(\lambda_{\nu n}/n) + (\sigma^2/nb)}{((\lambda_{\nu n}/n) + \lambda)^2}\Big)\Big(\sum \frac{1}{((\lambda_{\nu n}/n) + \lambda)}\Big)\Big(\sum \frac{1}{((\lambda_{\nu n}/n) + \lambda)^2}\Big) =$$

(2.12)

$$\Big(\sum \frac{1}{((\lambda_{\nu n}/n) + \lambda)}\Big)^2 \Big(\sum{}' \frac{(\lambda_{\nu n}/n) + (\sigma^2/nb)}{((\lambda_{\nu n}/n) + \lambda)^3}\Big)$$

and it is easily seen by setting $\lambda = \sigma^2/nb$ in (2.12) that $\lambda = \sigma^2/nb$ is a solution, and it can be verified that it is a minimizer. We can write

$$T(\lambda) = \frac{1}{n}\|A(\lambda)y - f\|_n^2$$

and

$$E_N T(\lambda) = \frac{1}{n}\|(I - A)f\|_n^2 + \frac{\sigma^2}{n}\operatorname{Tr} A^2 \tag{2.13}$$

$$E_S E_N T(\lambda) = b\{\frac{1}{n}\,\mathrm{Tr}\,(I-A)Q_n(I-A) + \frac{\sigma^2}{nb}\,\mathrm{Tr}\,A^2\}$$

(2.14)

$$= b\Big(\sum{}' \frac{(\lambda_{\nu n}/n)\lambda^2}{((\lambda_{\nu n}/n)+\lambda)^2} + \frac{\sigma^2}{nb}\sum\frac{(\lambda_{\nu n}/n)^2}{((\lambda_{\nu n}/n)+\lambda)^2}$$

which is also minimized for $\lambda = \sigma^2/nb$. ∎

We remark that with the stochastic model (2.1-2.3) the maximum likelihood estimate for λ is given by the minimizer of

$$M(\lambda) = \frac{y'(I - A(\lambda))y}{[\mathrm{Det}(I - A(\lambda))]^{1/n}} \tag{2.15}$$

and it can be verified that Theorem 1 is also true for this estimate. That is, the minimizer of $E_S E_N M(\lambda)$ is $\lambda = \sigma^2/nb$. This is the approach taken by Anderssen and Bloomfield in their pioneering papers [1][2].

What happens if $f \in \mathscr{N}_Q$? We have the property that the minimizer of $E_N V(\lambda)$ tends to the minimizer of $E_N T(\lambda)$ as $n \to \infty$ <u>for any</u> $f \in \mathscr{N}_Q$ under very general conditions on the mesh $\{t_i\}$, and on Q.

<u>Theorem 2.</u>

Let $t_i = t_{in}$, $i = 1,2,\ldots,n$, $n = 1,2,\ldots$, satisfy $\int_0^{t_{in}} w(u)du = \frac{i}{n}$, where $w(u)$ is a strictly positive continuous function with $\int_0^1 w(u)du = 1$, and suppose that the probability measure associated with the covariance Q is equivalent (see Root [28], Hajek [17], Wahba [33] for definitions and examples) to that corresponding to some continuous time stochastic process $\{X(t),\ t \in [0,1]\}$, satisfying an mth order linear differential equation

$$(L_m X)(t) \sum_{j=0}^{m} a_j(t)X^{(j)}(t) = \frac{dW(t)}{dt}, \qquad t \in [0,1], \tag{2.16}$$

for some m, where W is the Wiener process, $a_m(t) > 0$, and the a_j's satisfy some regularity conditions. Then, for any fixed $f \in \mathscr{N}_Q$, the minimizer, λ^*, say of $E_N V(\lambda)$, and the minimizer, $\tilde{\lambda}$, say of $E_N T(\lambda)$ satisfy

$$\lambda^* = \tilde{\lambda}\,(1 + o(1)) \tag{2.17}$$

where $o(1) \to 0$ as $n \to \infty$.

Proof:

A proof for the case $t_{in} = i/n$ and $f \in \mathscr{W}_Q$ satisfies further "very smooth" conditions roughly equivalent to $f^{(\nu)}$ abs. cont., $\nu = 0,1,\ldots,$ $2m-1$, $L_m^* L_m f \in \mathcal{L}_2$, may be found in Wahba [37]. The case of general t_{in} and general $f \in \mathscr{W}_Q$ with $L_m = D^m$ will appear in Craven and Wahba [9] ($D = \frac{d}{dt}$). A sketch of a proof in the special case $\mathscr{W}_Q$ a space of periodic functions, $t_{in} = i/n$, and $L_m = D^m$ appears in Wahba and Wold [41] and in more detail in Section 4 .

We remark that the equivalence condition on Q can be reformulated, roughly, that $\mathscr{W}_Q$ is topologically equivalent to a Hilbert space of functions $\{f: f^{(\nu)}$ abs. cont., $\nu = 0,1,\ldots.m-1$, $L_m f \in \mathcal{L}_2[0,1]\}$. Since topological equivalence means "same convergent sequences" it is not important which Q from a particular equivalence class is used in practice, when n is large. For computational reasons, then, one chooses the simplest Q which in most applications will turn out to be one corresponding to $L_m = D^m$. Then $f_{n,\lambda}$ will be a polynomial spline of degree $2m-1$ (See [19]). The practical minded statistician or numerical analyst who has small or medium sized data sets would probably choose the lowest order method which is nontrivial - i.e., $m = 2$. We recommend $m = 2$ for small to medium sized data sets no matter how many derivatives one can assume f has.

It can be shown (See Section 3) that if f satisfies the "very smooth" conditions given above plus some boundary conditions, then λ^* and $\tilde{\lambda}$ go to zero at the rate $n^{-2m/(4m+1)}$. However, the maximum likelihood estimate can be shown to go to zero at a faster rate than $n^{-2m/(4m+1)}$, and for this reason we use the GCV estimate instead.

We pause to indicate where the mysterious function $V(\lambda)$ with the magical properties came from. The fundamental idea began with cross-validation (also known as predictive sample reuse) as discussed in Geisser [13], Stone [29]. Suppose a particular λ is a good choice for the smoothing parameter. Let $f_{n,\lambda}^{(k)}$ be the solution of the

minimization problem of (2.6) with the k^{th} data point left out. Then $f^{(k)}_{n,\lambda}(t_k)$ should be a good predictor of the missing data point $y(t_k)$, and we measure this by

$$(2.18) \qquad V_0(\lambda) = \sum_{k=1}^{n} (f^{(k)}_{n,\lambda}(t_k) - y(t_k))^2 w_k(\lambda)$$

where the weights $w_k(\lambda)$ will be discussed shortly. If Q is periodic and stationary, that is, $Q(s,t) = q((s-t) \bmod 1)$ for some q; and $t_j = j/n$ then Q_n is a circulant matrix and using this fact it can be shown that $V_0(\lambda) \equiv V(\lambda)$ with the weights $w_k(\lambda) \equiv 1$. (See Wahba and Wold [41]). In general, if the $w_k(\lambda)$ are chosen so that

$$w_k(\lambda) = \left[\frac{(1 - a_{kk}(\lambda))}{1 - \frac{1}{n} \sum_{j=1}^{n} a_{jj}(\lambda)} \right]^2$$

where the $a_{kk}(\lambda)$ are the kk^{th} entries of $A(\lambda)$, then it can be shown that $V_0(\lambda) = V(\lambda)$. See Craven and Wahba [9]. The weights $w_k(\lambda)$ were chosen because they are just those weights required to give the result of Theorem 2. We challenge the reader to find $V(\lambda)$ as the result of applying some reasonable optimality principle, we don't know what this principle is, but think it exists!

3. Density Estimation as a Problem in Recovering a Smooth Curve from Noisy Data.

3.1. The density estimate.

Let Q be a r.k. satisfying the hypotheses of Theorem 2. Let the Mercer-Hilbert-Schmidt expansion (see Riesz-Nagy [26]) of Q be

$$(3.1) \qquad Q(s,t) = \sum_{\nu=0}^{\infty} \lambda_\nu \phi_\nu(s) \phi_\nu(t) .$$

(That is, $\{\lambda_\nu\}$ and $\{\phi_\nu\}$ are the eigenvalues and eigenfunctions of the Hilbert-Schmidt operator with Hilbert-Schmidt kernel Q). We will always assume here that $\phi_0(t) \equiv 1$, and this assures that $\int_0^1 \phi_\nu(s)ds \equiv 0$, $\nu = 1,2,\ldots,$. If Q is associated with L_m, of Theorem 2, then Q behaves like a Green's function for a $2m^{th}$ order linear differential operator, and then it is known [23] that the eigenvalues $\{\lambda_\nu\}$ of Q go to zero at the

rate ν^{-2m}. The $\{\phi_\nu\}$ are infinitely differentiable [26].

Let $X_1, X_2, \ldots, X_n$ be n independent observations from a density f supported on $[0,1]$ and let the sample (generalized) Fourier coefficients $\hat{f}_\nu$ of f be defined by

$$\hat{f}_\nu = \frac{1}{n} \sum_{j=1}^{n} \phi_\nu(X_j), \qquad \nu = 1, 2, \ldots, n,$$

$$\hat{f}_0 = f_0 = 1.$$

It is easy to see that $\hat{f}_\nu$ is an unbiased estimate of $f_\nu = \int_0^1 \phi_\nu(t) f(t)\,dt$, with variance going to zero at the rate $1/n$. We define $f_n(t)$, the sample inverse Fourier transform of f by

$$f_n(t) = \sum_{\nu=0}^{n} \hat{f}_\nu \phi_\nu(t) \tag{3.2}$$

$$= \frac{1}{n} \sum_{j=0}^{n} K_n(t, X_j), \tag{3.3}$$

where

$$K_n(s,t) = \sum_{\nu=0}^{n} \phi_\nu(s)\, \phi_\nu(t). \tag{3.4}$$

If $f \in \mathcal{H}_Q$, it can be shown that $\|f\|_Q^2 = \sum_{\nu=0}^{\infty} \frac{f_\nu^2}{\lambda_\nu} < \infty$, and then $f_n(t)$ is very nearly unbiased for $f(t)$ if n is large since

$$E f_n(t) = \int_0^1 K_n(t,u) f(u)\,du = \sum_{\nu=0}^{n} f_\nu \phi_\nu(t) = f(t) + o(1), \tag{3.5}$$

where

$$|o(1)| = \Big| \sum_{\nu=n+1}^{\infty} f_\nu \phi_\nu(t) \Big| \le \sum_{\nu=n+1}^{\infty} \frac{f_\nu^2}{\lambda_\nu} \sum_{\nu=n+1}^{\infty} \lambda_\nu \phi_\nu^2(t) \le \sum_{\nu=n+1}^{\infty} \frac{f_\nu^2}{\lambda_\nu} \cdot Q(t,t). \tag{3.6}$$

The estimate $f_n(t)$ is not consistent for $f(t)$, however. Letting Z be a random variable with density f, we have

$$E f_n(s) f_n(t) = \frac{1}{n} E K_n(s,Z) K_n(t,Z) + (1 - \tfrac{1}{n}) E K_n(s,Z) E K_n(t,Z), \tag{3.7}$$

$$\operatorname{cov} f_n(s) f_n(t) = \tag{3.8}$$

$$\frac{1}{n} \Big[\int_0^1 K_n(s,u) K_n(t,u) f(u)\,du - \int_0^1 K_n(s,u) f(u)\,du \int_0^1 K_n(t,u) f(u)\,du \Big].$$

Now since the eigenfunctions ϕ_ν of Q are analytic, and f is "nice" we may write

$$(3.9)\quad \int_0^1 K_n(\tfrac{i}{n},u)\, K_n(\tfrac{k}{n},u)\, f(u)du$$

$$\simeq \frac{1}{n}\sum_{\ell=1}^{n} K_n(\tfrac{i}{n},\tfrac{\ell}{n})\, K_n(\tfrac{k}{n},\tfrac{\ell}{n})\, f(\tfrac{\ell}{n})$$

$$= \frac{1}{n}\sum_{\ell=0}^{n} f(\tfrac{\ell}{n}) \sum_{\mu=0}^{n}{}' \phi_\mu(\tfrac{i}{n})\,\phi_\mu(\tfrac{\ell}{n}) \sum_{\nu=0}^{n} \phi_\nu(\tfrac{k}{n})\,\phi_\nu(\tfrac{\ell}{n}) .$$

Now, since

$$(3.10)\quad \frac{1}{n+1}\sum_{j=0}^{n} \phi_\mu(\tfrac{j}{n})\,\phi_\nu(\tfrac{j}{n}) \simeq \int_0^1 \phi_\mu(s)\,\phi_\nu(s)ds = \begin{cases} 1 & \mu=\nu \\ 0, & \mu\neq\nu, \end{cases}$$

$$\mu,\nu = 0,1,\ldots,n ,$$

we also must have

$$(3.11)\quad \frac{1}{n+1}\sum_{\mu=0}^{n} \phi_\mu(\tfrac{j}{n})\,\phi_\mu(\tfrac{\ell}{n}) \simeq \begin{cases} 1, & j=\ell \\ 0, & j\neq\ell \end{cases}$$

Thus, we observe from the right hand side of (3.9)

$$\frac{1}{n}\int_0^1 K_n(s,u)\, K_n(t,u)\, f(u)du \simeq \begin{cases} f(t), & s=t \\ 0, & |s-t| = \frac{j}{n},\ j=1,2,\ldots,n \end{cases}$$

which results in

$$(3.12)\quad E\, f_n(t) \simeq f(t)$$

$$\operatorname{cov} f_n(\tfrac{j}{n})\, f_n(\tfrac{k}{n}) \simeq \begin{cases} f(\tfrac{j}{n}), & j=k \\ 0, & j\neq k . \end{cases}$$

Thus we have the approximate model for $f_n(t)$:

$$(3.13)\quad f_n(t) \simeq f(t) + \varepsilon(t)$$

where

$$(3.14)\quad E\,\varepsilon(\tfrac{i}{n}) \simeq 0$$

$$E\,\varepsilon(\tfrac{i}{n})\,\varepsilon(\tfrac{j}{n}) \simeq \begin{cases} f(\tfrac{i}{n}), & i=j \\ 0, & i\neq j . \end{cases}$$

We will take as "data" the "observations"

$$(3.15) \qquad y(\tfrac{i}{n}) \equiv y_i = f_n(\tfrac{i}{n}), \qquad i = 1,2,\ldots,n$$

where $f_n(t)$ is defined by (3.2), and proceed to look at the problem of obtaining an estimate of f as that of recovering f given the noisy "ordinate" data y_i.

The model given by (3.12) - (3.15) is not exactly the same as that considered in Section 2 since here $E\,\varepsilon^2(\frac{i}{n})$ depends on i in general whereas in (2.3) $E\,\varepsilon^2(\frac{i}{n}) = \sigma^2$. However, at least in the stationary periodic case (see immediately following (3.18)) and below, we will be able to obtain the density estimate version of Theorems 1 and 2.

Following (2.4) and (2.5) we begin by considering as a density estimate

$$(3.16) \qquad f_{n,\lambda}(t) = (Q_{t_i}(t),\ldots,Q_{t_n}(t))(Q_n + n\lambda I)^{-1} y$$

where $t_i = i/n$, and

$$y = (f_n(\tfrac{1}{n}), f_n(\tfrac{2}{n}), \ldots, f_n(\tfrac{n}{n})),$$

and λ is going to be chosen by the method of generalized cross-validation. We want to obtain an approximate form for (3.16) which is easier to compute, and then modify the result to use the information that $f_0 = \int_0^1 f(u)du \equiv 1$. By observing that since

$$Q(s,t) = \sum_{\nu=0}^{\infty} \lambda_\nu \phi_\nu(s)\,\phi_\nu(t)$$

then

$$(3.17) \qquad Q_n \approx \Gamma D \Gamma'$$

where here Γ is the $n\times n$ matrix with νjth entry $\frac{1}{\sqrt{n}}\phi_{\nu-1}(\frac{j}{n})$, $\nu, j = 0,1,2,\ldots,n-1$, and D is the diagonal matrix with $\nu\nu^{th}$ entry $n\lambda_{\nu-1}$, $\nu = 1,2,\ldots,n$. Substituting (3.17) into (3.16) gives the approximate expression for (3.16) evaluated at $t = \frac{k}{n}$, $k = 1,2,\ldots,n$,

$$(3.18) \qquad f_{n,\lambda}(\tfrac{k}{n}) \simeq \sum_{\nu=0}^{n-1} \frac{\hat{f}_\nu\,\phi_\nu(\frac{k}{n})}{1 + \lambda/\lambda_\nu}$$

where

$$\hat{f}_\nu = \frac{1}{n}\sum_{i=0}^{n-1} \phi_\nu(\tfrac{i}{n})y_i \,. \tag{3.19}$$

The expression we actually want to use is

$$f_{n,\lambda}(t) = 1 + \sum_{\nu=1}^{n} \frac{\hat{f}_\nu}{(1+\lambda/\lambda_\nu)}\,\phi_\nu(t) \,. \tag{3.20}$$

This is an approximation to

$$E\,\{f(t)\,|\; y(t_i) = y_i\,,\ i = 1,2,\ldots,n,\ \int_0^1 f(u)du = 1\}$$

for the stochastic model of Section 2. For the model $f \in \mathscr{W}_Q$, this is an approximation to the solution of a constrained version of the minimization problem of (2.6), namely, find $f \in \mathscr{W}_Q$ to minimize

$$\frac{1}{n}\sum_{j=1}^{n}(f(t_j) - y_j)^2 + \lambda\|f\|_Q^2$$

subject to

$$\int_0^1 f(u)du = 1 \tag{3.22}$$

(See [19] for details).

Our density estimate will be $f_{n,\lambda}(t)$ given by (3.20), where the crucial smoothing parameter λ will be chosen according to the method of generalized cross-validation for densities, to be described in the next section. The estimate is a form of an orthogonal series estimate; instead of truncating the series we "taper" it. The function $f_{n,\lambda}$ may be thought of as the result of putting f_n through a "low-pass" filter, where the shape of the filter is determined by Q and the "half power point", ν_0, of the filter is determined by $\lambda_{\nu_0} = \hat{\lambda}$. On the other hand, letting

$$K_{n,\lambda}(s,t) = 1 + \sum_{j=1}^{n} \frac{\phi_\nu(s)\,\phi_\nu(t)}{1 + \lambda/\lambda_\nu}\,, \tag{3.23}$$

we have

$$f_{n,\hat{\lambda}} \simeq \frac{1}{n}\sum_{j=1}^{n} K_{n,\hat{\lambda}}(t,X_i) \tag{3.24}$$

which is a window estimate in the case $K_{n,\lambda}(s,t) = K_{n,\lambda}(s-t)$, and $\hat{\lambda}$ controls the width of the window. (An example will appear later.) In general, $f_{n,\lambda}$ is a δ-function estimate of the type considered recently

by Walter and Blum [42].

3.2. Generalized cross-validation for estimating λ in the density estimation case.

Returning to the expression for $V(\lambda)$ in Section 2 for the curve smoothing problem we can write (2.9) as

$$V(\lambda)= \frac{\sum_{\nu} \frac{\lambda^2 |\tilde{f}_\nu|^2}{((\lambda_{\nu n}/n)+\lambda)^2}}{[\frac{1}{n} \sum_{\nu} \frac{\lambda}{((\lambda_{\nu n}/n)+\lambda)}]^2} \tag{3.25}$$

where the $\{\lambda_{\nu n}\}$ are the eigenvalues of Q_n as in (2.10) and (2.11), and

$$\begin{pmatrix} \tilde{f}_1 \\ \tilde{f}_2 \\ \vdots \\ \tilde{f}_n \end{pmatrix} = \tilde{\Gamma}^T y$$

where $\tilde{\Gamma}$ is the orthogonalizing matrix for Q_n, $\tilde{\Gamma}^T Q_n \tilde{\Gamma} = \text{diag}\{\lambda_{\nu n}\}$. In the density estimation context the expression we want to use to approximate (3.25) is

$$V(\lambda) \overset{=}{=} \frac{\sum_{\nu=1}^{n} \frac{\lambda^2 |\hat{f}_\nu|^2}{(\lambda_\nu + \lambda)^2}}{[\frac{1}{n} \sum_{\nu=1}^{n} \frac{\lambda}{(\lambda_\nu+\lambda)}]^2} \tag{3.26}$$

where

$$\hat{f}_\nu = \frac{1}{n} \sum_{i=1}^{n} \phi_\nu(X_i)$$

is an estimate for f_ν, as before. (It can be shown that the $\{\lambda_{\nu n}\}$ "behave like" $n\lambda_\nu$, $\nu = 1,2,\ldots,n$.)

Since the "noise model" of (3.12)-(3.15) is not exactly the same as that of Section 2 it does not necessarily follow that the minimizer of $V(\lambda)$ has the properties described in Section 2. It turns out however,

that if $\mathcal{H}_Q$ is a space of periodic functions in $[0,1]$, then we can get the same properties for the minimizer of $V(\lambda)$ in the density estimation case as in Section 2. For the remainder of this section we let

$$(3.27) \qquad Q(s,t) = \sum_{\nu=-\infty}^{\infty} \lambda_\nu \phi_\nu(s) \phi_\nu^*(t)$$

where $\phi_\nu(s) = e^{2\pi i \nu s}$, $\nu = 0, \pm 1, 2, \ldots$, and

$$\lambda_0 = 1$$

$$\lambda_\nu = \lambda_{-\nu} = |P_m(2\pi i \nu)|^{-2}, \qquad \nu = 1, 2, \ldots$$

where $P_m(z)$ is an mth degree polynomial with all its zeros inside the unit circle. Q satisfies the hypotheses of Theorem 2, with

$$L_m X = P_m(D)X$$

where $D = \frac{d}{dt}$ and X satisfies the periodic boundary conditions $X^{(\nu)}(1) \equiv X^{(\nu)}(0)$, $\nu = 0, 1, \ldots, m-1$. Furthermore, it can be verified that

$$\|f\|_Q^2 = \left[\int_0^1 f(u)du\right]^2 + \sum_{\substack{\nu=-\infty \\ \nu \neq 0}}^{\infty} |P_m(2\pi i \nu)|^2 |f_\nu|^2$$

$$(3.28) \qquad = \left[\int_0^1 f(u)du\right]^2 + \int_0^1 [(L_m f)(u)]^2 du ,$$

where

$$(3.29) \qquad f_\nu = \int_0^1 e^{-2\pi i \nu t} f(t)dt, \qquad \nu = \pm 1, 2, \ldots .$$

Also, the solution to the constrained minimization problem of (3.22) in this space is just the solution to: Find $f \in \mathcal{H}_Q$ to minimize

$$(3.30) \qquad \frac{1}{n} \sum_{i=1}^{n} (f(t_i) - y_i)^2 + \lambda \int_0^1 [(L_m f)(u)]^2 du$$

subject to

$$\int_0^1 f(u)du = 1 .$$

Here

$$(3.31) \qquad \hat{f}_\nu = \frac{1}{n} \sum_{j=1}^{n} e^{-2\pi i \nu X_j}$$

and it is appropriate to take $V(\lambda)$ as

$$V(\lambda) = \frac{\sum_{\nu=1}^{n/2} \frac{\lambda^2 |\hat{f}_\nu|^2}{(\lambda_\nu + \lambda)^2}}{\left[\frac{1}{n}\sum_{\nu=1}^{n/2} \frac{\lambda}{(\lambda_\nu + \lambda)}\right]^2}, \tag{3.32}$$

where we are supposing n even for simplicity. Then

$$E|\hat{f}_\nu|^2 = \frac{1}{n^2} E \sum_{j=1}^{n}\sum_{k=1}^{n} e^{2\pi i \nu (X_j - X_k)} = (1 - \frac{1}{n})|f_\nu|^2 + \frac{1}{n} \tag{3.33}$$

and

$$EV(\lambda) = \frac{(1-\frac{1}{n})\sum_{\nu=1}^{n/2} \frac{\lambda^2 |f_\nu|^2}{(\lambda_\nu + \lambda)^2} + \frac{1}{n}\sum_{\nu=1}^{n/2} \frac{\lambda^2}{(\lambda_\nu + \lambda)^2}}{\left[\frac{1}{n}\sum_{\nu=1}^{n/2} \frac{\lambda}{(\lambda_\nu + \lambda)}\right]^2} \tag{3.34}$$

Letting

$$y = (f_n(\tfrac{1}{n}), f_n(\tfrac{2}{n}); \ \ldots \ , f_n(\tfrac{n}{n}))'$$

$$f = (f(\tfrac{1}{n}), f(\tfrac{2}{n}), \ \ldots \ , f(\tfrac{n}{n}))'$$

$$\varepsilon = (\varepsilon(\tfrac{1}{n}), \varepsilon(\tfrac{2}{n}), \ \ldots \ , \varepsilon(\tfrac{n}{n}))'$$

where $f_n(\cdot)$ and $\varepsilon(\cdot)$ are defined by (3.2) and (3.13), we have

$$T(\lambda) \stackrel{def}{=} \frac{1}{n}\sum_{j=1}^{n}(f_{n,\lambda}(\tfrac{i}{n}) - f(\tfrac{i}{n}))^2 \simeq \frac{1}{n}\|A(\lambda)y - f\|_n^2$$

$$E\,T(\lambda) \simeq \frac{1}{n}\|(I-A)f\|_n^2 + E\,\varepsilon' A^2 \varepsilon \tag{3.35}$$

$$\simeq \frac{1}{n}\|(I-A)f\|_n^2 + \frac{1}{n}\sum^{n} f(\tfrac{i}{n}))\,\tilde{a}_{ii}(\lambda)\,,$$

where $A(\lambda) = Q_n(Q_n + n\lambda I)^{-1}$ as in (2.8), and $\tilde{a}_{ii}(\lambda)$ is the ii^{th} entry of $A^2(\lambda)$. Since Q_n is circulant A is circulant, A^2 is circulant and $\tilde{a}_{ii}(\lambda) \equiv \frac{1}{n}\mathrm{Tr}\,A^2(\lambda)$. Since $\frac{1}{n}\sum_{i=1}^{n} f(\frac{i}{n}) \approx 1$, we have

$$E\,T(\lambda) \simeq \frac{1}{n}\|(I-A)f\|_n^2 + \frac{1}{n}\mathrm{Tr}\,A^2(\lambda)\,. \tag{3.36}$$

Inspection of (3.34) and (3.36) with the aid of (3.17) reveals that (3.34) and (3.36) are essentially the same as (2.10) and (2.13), the

expressions for $E_N V(\lambda)$ and $E_N T(\lambda)$, with $\sigma^2 = 1$.

Let us put a "phoney prior" on $f(s)$ induced by

$$(3.37) \qquad f(s) = 1 + \sum_{\substack{\nu=-\infty \\ \nu \neq 0}}^{\infty} f_\nu \phi_\nu(s)$$

where $f_\nu = f_{-\nu}^*$ and otherwise the f_ν are independent zero mean complex normal random variables (See Goodman [14]) with $E|f_\nu|^2 = b\lambda_\nu$. We have in effect introduced the prior convariance $bQ(s,t)$ on f and then conditioned on $\int_0^1 f(u)du = 1$. The prior is "phoney" of course, because the "sample functions" are not necessarily positive even though they do integrate to 1. Proceeding boldly despite the phoniness of the prior, and letting E_S be expectation with respect to this prior, we have

$$(3.38) \qquad E_S\, E\, V(\lambda) \cong b \Big(\sum_{\nu=1}^{n/2} \frac{\lambda_\nu + 1/nb}{(\lambda_\nu + \lambda)^2} \Big) \Big/ \Big(\sum_{\nu=1}^{n/2} \frac{1}{(\lambda_\nu + \lambda)} \Big)^2$$

$$(3.39) \qquad E_S\, E\, T(\lambda) \cong b \Big(\sum_{\nu=1}^{n/2} \frac{\lambda^2 \lambda_\nu}{(\lambda_\nu + \lambda)^2} + \frac{1}{nb} \sum_{\nu=1}^{n/2} \frac{\lambda_\nu^2}{(\lambda_\nu + \lambda)^2}$$

and the right hand sides of these expressions are minimized for the same λ, namely $\lambda = 1/nb$.

We now return to a more defensible assumption, namely, f is a density in $\mathcal{V}_Q$.

Theorem 2'.

Let Q be given by (3.27). Let λ^* and $\tilde{\lambda}$ be the minimizers of $EV(\lambda)$ and $ET(\lambda)$ as given by the right hand sides of (3.34) and (3.36). Let f be any density in $\mathcal{V}_Q$. Then as $n \to \infty$

$$(3.40) \qquad \lambda^* = \tilde{\lambda}(1 + o(1)).$$

Outline of Proof.

Set $(1 - \frac{1}{n}) \simeq 1$, in the expression for $V(\lambda)$ in (3.34). It can be shown that if $n \to \infty$, $\lambda \to 0$ in such a way that $n\lambda^{1/2m} \to \infty$, then

$$(3.41) \qquad 1 - \frac{1}{n} \sum \frac{\lambda}{\lambda_\nu + \lambda} \equiv \frac{1}{n} \sum \frac{\lambda_\nu}{\lambda_\nu + \lambda} \equiv \frac{1}{n} \sum \frac{1}{(1 + |P_m(2\pi\nu)|^2 \lambda)} \simeq \frac{1}{n\lambda^{1/2m}} \tilde{k}_m (1 + o(1)),$$

$$(3.42)\quad 1 - \frac{2}{n}\sum\frac{\lambda}{\lambda_\nu+\lambda} + \frac{1}{n}\sum\Big(\frac{\lambda^2}{(\lambda_\nu+\lambda)^2} = \frac{1}{n}\sum\Big(\frac{\lambda_\nu^2}{(\lambda_\nu+\lambda)^2} \equiv \frac{1}{n}\sum\frac{1}{(1+|P_m(2\pi\nu)|^2\lambda)^2}$$

$$= \frac{1}{n\lambda^{1/2m}}\, k_m(1 + o(1))$$

where

$$\tilde{k}_m = \frac{1}{2\pi}\int_0^\infty \frac{dx}{(1+x^{2m})}, \quad k_m = \frac{1}{2\pi}\int_0^\infty \frac{dx}{(1+x^{2m})^2},$$

and $o(1)\to 0$ as $n\to\infty$. Setting

$$(3.43)\quad \psi_f(\lambda) = \lambda^2\sum\frac{f_\nu^2}{(\lambda_\nu+\lambda)^2}$$

and substituting (3.41) and (3.42) into (3.34) and (3.36) one obtains,

$$(3.44)\quad E\,V(\lambda) \simeq \{\psi_f(\lambda) + (1 - \frac{2\tilde{k}_m}{n\lambda^{1/2m}} + \frac{k_m}{n\lambda^{1/2m}})\}/[1 - \frac{\tilde{k}_m}{n\lambda^{1/2m}}]^2$$

$$(3.45)\quad E\,T(\lambda) \simeq \psi_f(\lambda) + \frac{k_m}{n\lambda^{1/2m}}$$

The minimum $\tilde{\lambda}$ of $E\,T(\lambda)$ given by the right hand side of (3.45) occurs for the (smallest) solution of

$$(3.46)\quad \psi_f'(\lambda) = \frac{k_m}{2mn\lambda^{(2m+1)/2m}}$$

Differentiating $E\,V(\lambda)$ given by the right hand side of (3.44) gives

$$\frac{d}{d\lambda}V(\lambda) = \frac{1}{U_1}\{\psi_f'(\lambda) - \frac{k_m}{2mn\lambda^{(2m+1)/2m}}[1 - \frac{2\tilde{k}_m}{k_m}(1 - \frac{U_2}{U_1^{1/2}})]\}$$

where

$$U_1 = (1 - \frac{\tilde{k}_m}{n\lambda^{1/2m}})^2$$

$$U_2 = \psi_f(\lambda) + (1 - \frac{1}{n\lambda^{1/2m}}(2\tilde{k}_m - k_m)).$$

Provided $n\lambda\to\infty$ in such a way that $n\lambda^{1/2m}\to\infty$, we have

$$U_1 = 1 + o(1)$$

$$U_2 = 1 + o(1)$$

and

$$\frac{U_2}{U_1^{1/2}} = 1 + o(1) .$$

It can be shown that the minimizers of $EV(\lambda)$ and $ET(\lambda)$ have the property that $\lambda \to 0$, $n\lambda^{1/2m} \to \infty$, and so

$$\frac{d}{d\lambda} EV(\lambda) = [1 + o(1)]\{\psi_f'(\lambda) - \frac{k_m}{2mn\lambda^{1/2m}}[1 + o(1)]\},$$

so that for large n, $EV(\lambda)$ and $ET(\lambda)$ have asymptotically the same minimizer, that is, $\lambda^* = \tilde{\lambda}(1+o(1))$.

If $\sum \frac{f_\nu^2}{\lambda_\nu^2} < \infty$, (which in this case entails that $f^{(\nu)}$ abs. cont., $\nu = 0,1,\ldots 2m-1$, and

$$\int((L_m^* L_m f)(t))^2 dt \equiv C_f < \infty ,)$$

then

$$\psi_f(\lambda) = (\lambda^2 \sum \frac{f_\nu^2}{\lambda_\nu^2})(1 + o(1)) = \lambda^2 C_f(1 + o(1))$$

and it can be shown (See [38]) that λ^* and $\tilde{\lambda}$ satisfy

$$\lambda^2 C_f = \frac{k_m}{2mn\lambda^{(2m+1)/2m}}(1 + o(1))$$

or

$$\lambda^* = \tilde{\lambda}(1 + o(1)) = (\frac{k_m}{2mnC_f})^{2m/4m+1}(1 + o(1))$$

and then

$$(3.47)\quad ET(\lambda^*) = \frac{k_m^{4m/(4m+1)}}{n^{4m/(4m+1)}}[\frac{1}{(2mC_f)^{4m/(4m+1)}} + (2mC_f)^{1/(4m+1)}](1 + o(1)).$$

We note that the above above assumptions on f say that $f \in C_{2m,2}$ ($C_{2m,2}$ was defined in the introduction) and so if f_{n,λ^*} is to be in the class of "good" estimates we should have, for mean square error at a point

$$E(f_{n,\lambda^*}(t) - f(t))^2 \simeq C\, n^{-(4m-1)/(4m)} .$$

The rate $n^{-4m/(4m+1)}$ of (3.47) reflects the fact that slightly better convergence rates obtain for integrated mean square error over a finite interval than mean square error at a point.

We remark on the importance of the periodic assumption. In principle f can be estimated by (3.16) with the minimizer of $V(\lambda)$ given by (3.25) used as the estimate of λ. The practical problem occurs in computing the eigenfunctions. If the ϕ_ν are taken as complex exponentials, and the true density is not periodic, then the resulting $f_{n,\lambda}$ will not approximate f at the endpoints. Of course $f^{(\nu)}(1) = f^{(\nu)}(0) = 0$, $\nu = 0,1,\ldots,m-1$, are a perfectly good and quite reasonable set of periodic boundary conditions. With respect to the theoretical properties of the minimizer of $V(\lambda)$, we used $\mathrm{Var}\,\hat{f}_\nu = \text{const.}$ and $A(\lambda)$ circulant to get Theorem 2', these properties do not hold in general when $\mathscr{W}_Q$ is not a periodic space. Without this simplification the expressions for $EV(\lambda)$ and $ET(\lambda)$ are slightly more messy. We believe that in general the minimizer of $V(\lambda)$ estimates the minimizer of $\tilde{T}(\lambda)$ for $\tilde{T}$ some other quadratic form in the errors, but do not have a demonstration.

4. Preliminary Experimental Results.

The method was tried experimentally on five densities satisfying periodic boundary conditions, by generating a set of pseudo-random numbers distributed according to each density. We present details of the results from three of the examples. (The other two were not substantially different). We let Q be the periodic covariance of (3.27) with $m = 2$ and $\lambda_\nu^{-1} = (2\pi\nu)^{2m}$, $\nu = \pm 1, 2, \ldots$. This corresponds to the assumption that f and f' are continuous, $f'' \in \mathfrak{L}_2$, and $f^{(\nu)}(1) = f^{(\nu)}(0)$, $\nu = 0,1$. The density estimate we use is then an approximation to the solution of the problem: Find f such that f, f' continuous, $f'' \in \mathfrak{L}_2$, $f^{(\nu)}(1) = f^{(\nu)}(0)$, $\nu = 0,1$, $\int_0^1 f(u)du = 1$, to minimize

$$\frac{1}{n}\sum_{j=1}^{n}\left(f\left(\frac{j}{n}\right) - y_j\right)^2 + \lambda\int_0^1 (f''(u))^2 du, \tag{4.1}$$

where

$$y_i = f_n\left(\frac{i}{n}\right)$$

$$(4.2)\qquad f_n(t) = 1 + \sum_{\substack{\nu=-n/2 \\ \nu \neq 0}}^{n/2} \hat{f}_\nu e^{2\pi i \nu t}$$

$$\hat{f}_\nu = \frac{1}{n}\sum_{j=1}^{n} e^{-2\pi i \nu X_j}$$

The approximation to the solution of the minimization problem actually being computed is

$$(4.3)\qquad f_{n,\lambda}(t) = 1 + \sum_{\substack{\nu=-n/2 \\ \nu \neq 0}}^{n/2} \frac{\hat{f}_\nu}{(1+(2\pi\nu)^4\lambda)} e^{2\pi i \nu t}.$$

The GCV estimate $\hat{\lambda}$ for λ is found by computing

$$(4.4)\qquad V(\lambda) = \frac{\frac{1}{n}\sum_{\substack{\nu=-n/2 \\ \nu \neq 0}}^{n/2} \frac{|\hat{f}_\nu|^2}{(\lambda_\nu + \lambda)^2}}{\left[\frac{1}{n}\sum_{\substack{\nu=-n/2 \\ \nu \neq 0}}^{n/2} \frac{1}{(\lambda_\nu + \lambda)}\right]^2}, \quad \lambda_\nu = (2\pi\nu)^{-4}$$

at increments $\lambda = 10^{j/3}$ for $j = -21$ to 3 and determining the global minimum $\hat{\lambda}$ by inspection (to the nearest $10^{1/3}$).

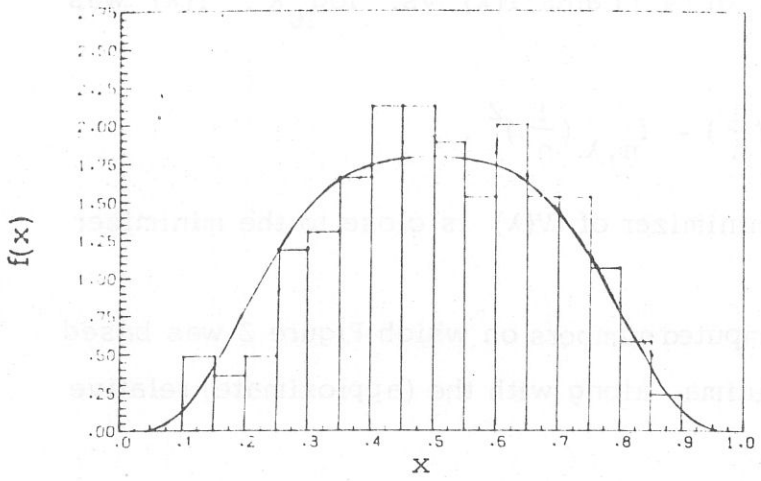
The densities tried were all mixtures of beta densities and the three cases presented are

Case I $\quad f \sim \frac{1}{3}\beta(10,5) + \frac{1}{3}\beta(7,7) + \frac{1}{3}\beta(5,10), \qquad n = 174$

Case II $\quad f \sim \frac{6}{10}\beta(12,7) + \frac{4}{10}\beta(3,11), \qquad n = 174$

Case IV $\quad f \sim \frac{1}{3}\beta(20,5) + \frac{1}{3}\beta(12,12) + \frac{1}{3}\beta(5,20) \qquad n = 170$

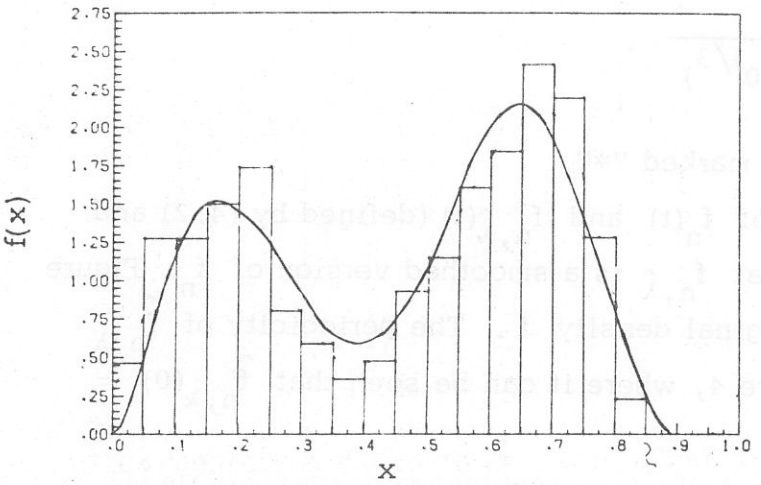
Figure 1 presents plots of the true densities along with histograms of the Monte Carlo realizations of n independent observations from each density. The "bin size" of the histogram was chosen by eye to give a pleasing picture.



Case I

$$f = \frac{1}{3}\beta(10,5) + \frac{1}{3}\beta(7,7) + \frac{1}{3}\beta(5,10)$$

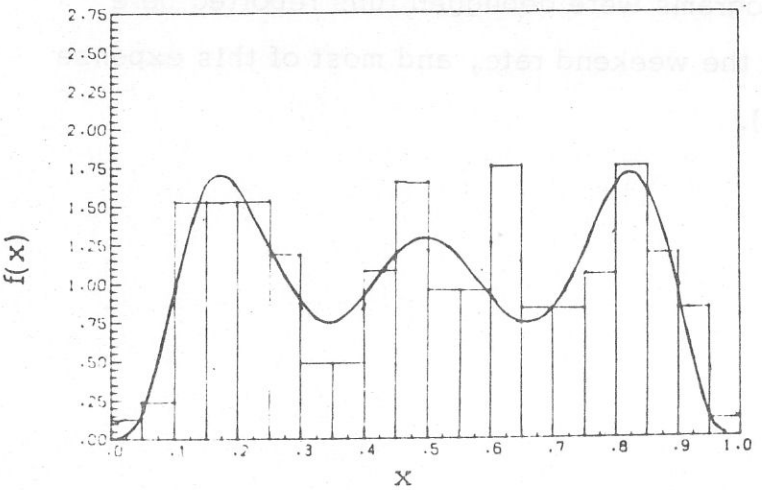
$n = 174$



Case II

$$f = \frac{6}{10}\beta(12,7) + \frac{4}{10}\beta(3,11)$$

$n = 174$



Case III

$$f = \frac{1}{3}\beta(20,5) + \frac{1}{3}\beta(12,12) + \frac{1}{3}\beta(5,20)$$

$n = 170$

True Density and Histogram of n Computer-Generated Observations

Figure 1

Figure 2 gives plots of $V(\lambda)$ of (4.4) and $T(\lambda)$ vs. $\log_{10}\lambda$. $T(\lambda)$ was computed by

$$T(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(f(\tfrac{i}{n}) - f_{n,\lambda}(\tfrac{i}{n})\right)^2 .$$

The reader can see that the minimizer of $V(\lambda)$ is close to the minimizer of $T(\lambda)$.

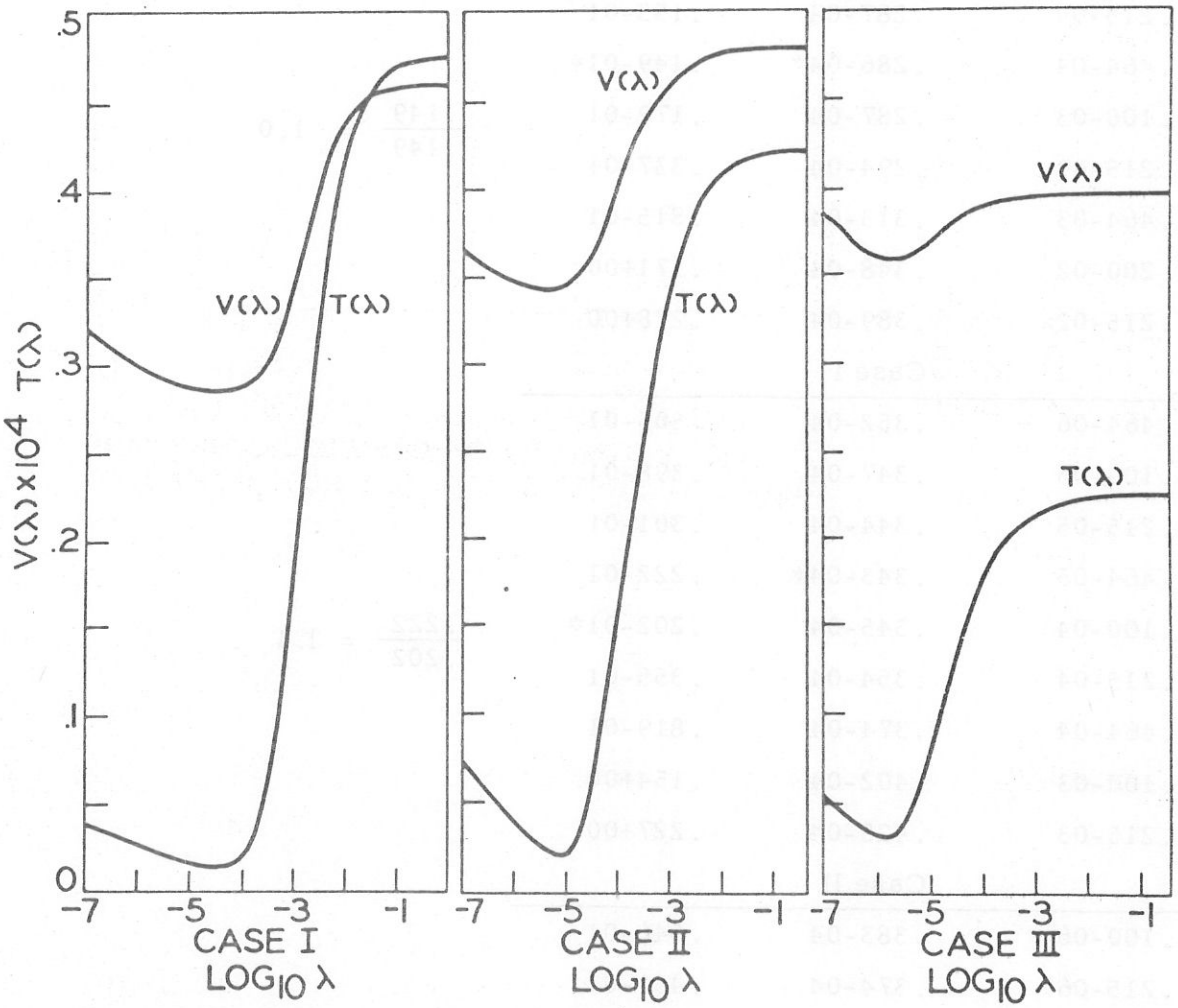
Table 1 gives the computed numbers on which Figure 2 was based in the neighborhood of the minima, along with the (approximate) relative inefficiency

$$\frac{T(\hat{\lambda})}{\inf_j T(\lambda = 10^{j/3})} .$$

The minima of V and T are marked "*".

Figure 3 gives plots of $f_n(t)$ and $f_{n,\hat{\lambda}}(t)$ (defined by (4.2) and (4.3) respectively). Note that $f_{n,\hat{\lambda}}$ is a smoothed version of f_n. Figure 4 compares $f_{n,\hat{\lambda}}$ and the original density f. The periodicity of $\hat{f}_{n,\lambda}$ is evident in Case II of Figure 4, where it can be seen that $\hat{f}'_{n,\lambda}(0) = f'_{n,\lambda}(1) \neq 0$.

Once the computer programs were debugged runs reported here cost less than $30 to run at the weekend rate, and most of this expense went toward determining $T(\lambda)$.



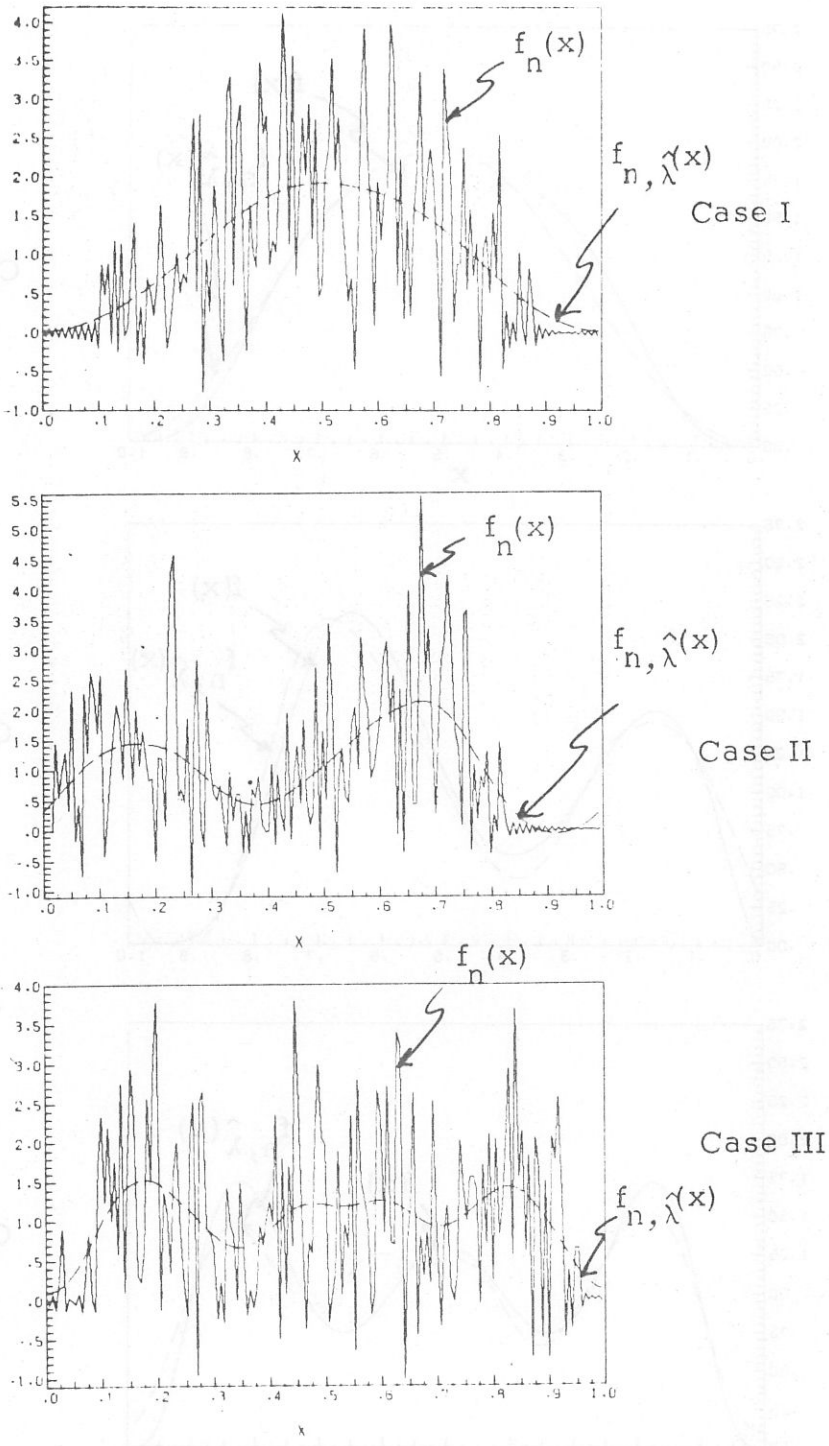
V(λ), THE CROSS VALIDATION FUNCTION AND
T(λ), THE INTEGRATED SQUARE ERROR, VS $LOG_{10}\lambda$

FIG. 2

λ	$V(\lambda)$	$T(\lambda)$	Relative Inefficiency (defined by(1.2))
.464-05	.291-04	.203-01	
.100-04	.288-04	.171-01	
.215-04	.287-04	.153-01	
.464-04	.286-04*	.149-01*	
.100-03	.287-04	.179-01	$\frac{.149}{.149} = 1.0$
.215-03	.294-04	.337-01	
.464-03	.313-04	.815-01	
.200-02	.348-04	.171+00	
.215-02	.389-04	.278+00	
	Case I		
.464-06	.352-04	.505-01	
.100-05	.347-04	.398-01	
.215-05	.344-04	.301-01	
.464-05	.343-04*	.222-01	
.100-04	.345-04	.202-01*	$\frac{.222}{.202} = 1.1$
.215-04	.354-04	.355-01	
.464-04	.374-04	.819-01	
.100-03	.402-04	.154+00	
.215-03	.428-04	.227+00	
	Case II		
.100-06	.383-04	.546-01	
.215-06	.374-04	.453-01	
.464-06	.367-04	.371-01	
.100-05	.361-04	.317-01*	$\frac{.330}{.317} = 1.04$
.215-05	.359-04*	.330-01	
.464-05	.360-04	.458-01	
.100-04	.366-04	.728-01	
.215-04	.375-04	.108+00	
.464-04	.383-04	.143+00	
	Case III		

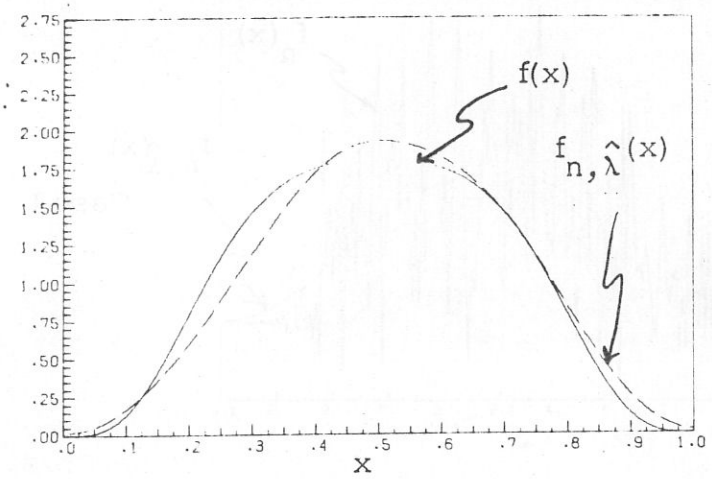
Table I

$V(\lambda)$ and $T(\lambda)$ vs. λ in the neighborhood of the minima, and the Relative Inefficiency of $\hat{\lambda}$.

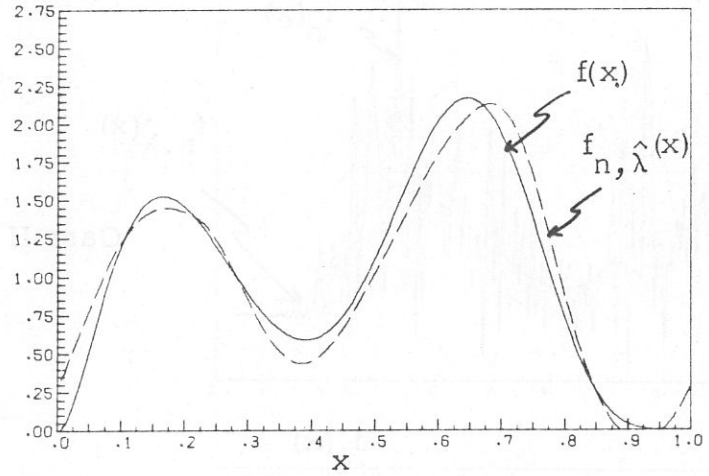


Sample Inverse Fourier Transform, $f_n(x)$, and the Density Estimate $f_{n,\hat{\lambda}}(x)$.

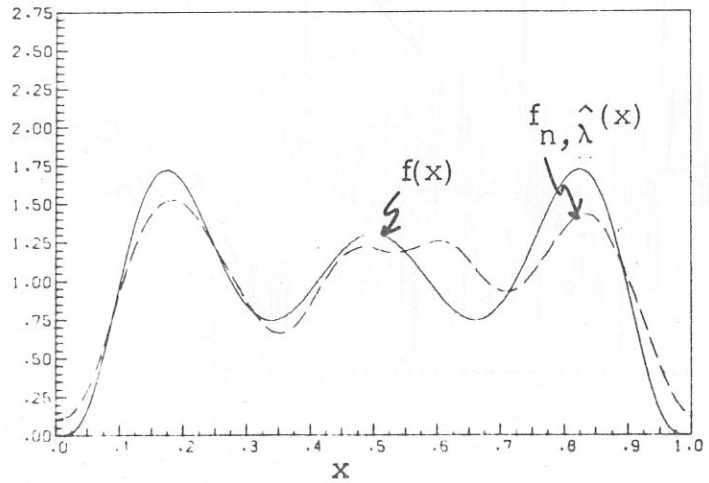
Figure 3



Case I



Case II



Case III

True and Estimated Densities

Figure 4

5. Further Work.

The first question that arises is: How does this method compare with other methods for estimating a density, under comparable assumptions on the unknown density. We believe that the method will give about the same results as a corresponding kernel or histospline method for large n, provided that the smoothing parameter in each method is chosen as well as possible (i.e. to minimize the average mean square error knowing the true density). A fair study of how two methods might compare in practice, however, requires that both methods have an objective procedure for choosing their respective smoothing parameters from the sample.

It remains to determine theoretically exactly what the minimizer of $V(\lambda)$ estimates in the non-periodic case.

It remains to develop cheap calculational procedures when $\mathcal{H}_Q$ is not a periodic space. Any function on $[0,1]$ with $f^{(\nu)}$ abs. cont., $\nu = 0,1,\ldots,m-1$, $f^{(m)} \in \mathcal{L}_2$, can be decomposed into a polynomial of degree m and an element of the periodic space $\mathcal{H}_Q$ in (3.27) with $\lambda_\nu = (2\pi\nu)^{-2m}$. We are attempting to use this to develop approximate computational procedures that can handle the non-periodic case easily.

When f is known to be periodic and furthermore, $f^{(\nu)}(0) = f^{(\nu)}(1) = 0$, $\nu = 0,1,\ldots,m-1$, then $f_{n,\lambda}$ can be taken as the solution to the minimization problem of (4.1) subject to these additional constraints. (The solution can be written down using results in e.g. Kimeldorf and Wahba [19]). $V(\lambda)$ would be unchanged, and Case II of Figure 4 would not have that annoying increase in $f_{n,\hat{\lambda}}$ near $x = 1.0$.

In principle the method extends immediately to f a (doubly periodic) density on the unit square, where f is in the tensor product space of 2 periodic rkhs on the unit interval. See Wahba [36]. It is apparent that all of the theoretical results will hold, since they depend only on the eigenvalues of the r.k. and the Fourier coefficients of f. We think the method may actually be computationally feasible, with very clever programming, in up to 3 or 4 dimensions, but of course at some

point an overwhelmingly large number of observations would be required to estimate a multi-dimensional density by this method. One approach to simplifying the computation in the multi-dimensional approach would be to use a simple approximation to the window of (3.24) to compute $f_{n,\hat{\lambda}}$.

Acknowledgment.

The computer programs for the study in Section 4 were ably written by Mr. Dick Jones. General assistance was provided by Mr. Michael Akritas.

References

1. Anderssen, R. S. and Bloomfield, P. (1974), A time series approach to numerical differentiation, Technometrics 16 (1), 69-75.

2. Anderssen, B. and Bloomfield, P. (1974), Numerical differentiation procedures for non-exact data, Numer. Math. 22, 157-182.

3. Boneva, L., Kendall, D. and Stefanov, I. (1971), Spline transformations: Three new diagnostic aids for the statistical data analyst. J. Roy. Statist. Soc. 33, 1-70.

4. Brunk, H.D. (1976), Univariate density estimation by orthogonal series, TR#51, Dept. of Statistics, Oregon State University, Corvallis, Oregon.

5. Chi, P.Y. and Van Ryzin, John, A histogram method for non-parametric classification, this volume.

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

7. Cover, T. M. and Wagner, I.T. (1976), Topics in statistical pattern recognition, in Digital Pattern Recognition, Vol. 10 of Communications in Statistics, eds. Fu, Keidel and Wolter, Springer Verlag, 15-46.

8. Crain, B.R. (1976), Matrix density estimation, Commun. Statist. A5(1), 89-96.

9. Craven, P. and Wahba, G. (1976), Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation, in preparation.

10. de Montricher, G.F., Tapia, R. A., and Thompson, J. R. (1975), Nonparametric maximum likelihood estimation of probability densities by penalty function methods, Ann. Statist. 6, 1329-1348.

11. 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.

12. Fienberg, S. and Holland, P. (1972), On the choice of flattening constants for estimating multinomial probabilities, J. Multivariate Anal. 2, 1, 127-134.

13. Geisser, S. (1975), The predictive sample reuse method with applications, JASA, 70, 350, 320-328.

14. Golub, G., Heath, M. and Wahba, G. (1975), Cross-validation and optimum ridge regression, abstract in Abstracts of Papers to be presented at the SIAM-SIGUM 1975 Fall meeting, December 3, 4, 5, 1975, San Francisco.

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

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

17. Hájek, Jaroslav (1962), On linear statistical problems in stochastic processes, Czech. Math. J., 12 (87), 404-444.

18. Hudson, H.M. (1974), Empirical Bayes estimation, Technical Report #58, Stanford University Dept. of Statistics, Stanford, California.

19. Kimeldorf, George and Wahba, Grace (1971), Some results on Tchebycheffian spline functions, J. Math. Anal. Appl. 33, 82-95.

20. Kronmal, R. and Tarter, M. (1968), The estimation of probability densities and cumulatives by Fourier series mehtods, J. Amer. Statist. Assoc. 63, 925-952.

21. Lii, K.-S. and Rosenblatt, M. (1974), Asymptotic behavior of a spline estimate of a density function, manuscript, University of California, San Diego.

22. Loftsgarten, D. O. and Quesenberry, C. P. (1965), A non-parametric estimate of a multivariate density function, Ann. Math. Statist. 36 , 1049-1051.

23. Naimark, M. A. (1968), Linear differential operators, Part II, Ungar, New York.

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

25. Parzen, E. (1973), Relations between methods of non parametric probability density estimation, State University of N.Y., Buffalo, manuscript.

26. Riesz, F. and Sz.-Nagy, B. (1955), Functional Analysis, Unger, New York.

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

28. Root, W.L. (1962), Singular Gaussian measures in detection theory. Time Series Analysis, Proceedings of a Symposium held at Brown University, ed. M. Rosenblatt, Wiley, New York, 292-314.

29. Stone, M. (1974), Cross-validatory choice and assessment of statistical prediction, JRSS, Series B, 36, 2, 111-147.

30. Van Ryzin, J. (1966), Bayes risk consistency of classification procedures using density estimation, Sankhya, Ser. A 28, 261-270.

31. Van Ryzin, J. (1973), A histogram method of density estimation, Commun. Statist, 12, 493-506.

32. Wahba, Grace (1971), A polynomial algorithm for density estimation, Am. Math. Statist. 42, 1870-1886.

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

34. Wahba, Grace (1975), Optimal convergence properties of variable knot kernel, and orthogonal series methods for density estimation, Ann. Statist. 3, 15-29

35. Wahba, G. (1975), Interpolating spline methods for density estimation I. Equi-spaced knots, Ann. Stat. 3, 1, 30-48.

36. Wahba, G. (1975), A canonical form for the problem of estimating smooth surfaces, Univ. of Wisconsin-Madison, Department of Statistics, Technical Report #420.

37. Wahba, G. (1975), Practical approximate solutions to linear operator equations when the data are noisy, University of Wisconsin-Madison, Department of Statistics, Technical Report #430, to appear, SIAM J. Num. Anal.

38. Wahba, G. (1975), Smoothing noisy data by spline functions, Numer. Math. 24, 303-394.

39. Wahba, G. (1976), Histosplines with knots which are order statistics, J. Roy. Stat. Soc. Series B. 38, 2, 140-151.

40. Wahba, G and Wold, S. (1975), A completely automatic French curve: Fitting spline functions by cross-validation. Comm. Statist. 4. (1), 1-17.

41. Wahba, G. and Wold, S. (1975), Periodic splines for spectral density estimation: The use of cross-validation for determining the degree of smoothing, Comm. Statist. 4, 2, (125-141.

42. Walter, G. and Blum, J. (1976), Probability density estimation using delta-sequences, manuscript, University of Wisconsin-Milwaukee.

43. Watson, G.S. and Leadbetter, M. R. (1963), On the estimation of the probability density I , Ann. Math. Statist. 34, 480-491.

44. Watson, G. S. (1969), Density estimation by orthogonal series, Ann. Math. Statist. 40, 1496-1498.

45. Wegman, E. (1972), Nonparametric probability density estimation I, a survey of available methods, Technometrics.

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

47. Whittle, P. (1958). On the smoothing of probability density functions, J. R. Statist. Soc., B. 20, 334-343.

48. Woodroofe, M. (1970). On choosing a delta-sequence, Ann. Math. Statist., 41, 166-171.

49. Tarter, M. E., and Kronmal, R. A. (1976), An introduction to the implementation and theory of nonparametric density estimation, The American Statistician, 30, 3, 105-112.

Supported by the United States Air Force under Grant No. AFOSR 72-2363-C.

Department of Statistics
University of Wisconsin-Madison
Madison, Wisconsin 53706