NASA CR-
144384

(NASA-CR-144384) NONPARAMETRIC MAXIMUM LIKELIHOOD ESTIMATION OF PROBABILITY DENSITIES BY PENALTY FUNCTION METHODS (Rice Univ.) 37 p HC $3.75 CSCL 12A

N75-30861

Unclas
G3/65 33993

ICSA

INSTITUTE FOR COMPUTER SERVICES AND APPLICATIONS

RICE UNIVERSITY

275-025-016

Nonparametric Maximum Likelihood Estimation of Probability Densities by Penalty Function Methods

by

G. F. de Montricher, R. A. Tapia
and J. R. Thompson
Dept. of Mathematical Sciences
Rice University

ABSTRACT

Except in the extreme case when it is known a priori exactly to which finite dimensional manifold the probability density function which gave rise to a set of samples belongs, the parametric maximum likelihood estimation procedure leads to poor estimates and is unstable ; while the nonparametric maximum likelihood procedure is undefined.

In this paper, we develop a very general theory of maximum penalized likelihood estimation which should avoid many of these present difficulties. We also demonstrate that each reproducing kernel Hilbert space leads, in a very natural way, to a maximum penalized likelihood estimator and that a well-known class of reproducing kernel Hilbert spaces gives polynomial splines as the nonparametric maximum penalized likelihood estimates.

Institute for Computer Services & Applications
Rice University
Houston, Texas

August, 1974

Research supported in part by NASA contract NAS 9-12776

Nonparametric Maximum Likelihood Estimation of Probability Densities by Penalty Function Methods[1]

by

G.F. de Montricher[2], R.A. Tapia[3] and J.R. Thompson[3]

ABSTRACT

Except in the extreme case when it is known a priori exactly to which finite dimensional manifold the probability density function which gave rise to a set of samples belongs, the parametric maximum likelihood estimation procedure leads to poor estimates and is unstable; while the nonparametric maximum likelihood procedure is undefined. Good and Gaskins have recently suggested replacing the nonparametric maximum likelihood estimate with a nonparametric maximum penalized likelihood estimate; however they did not show that these estimates existed. In this paper we develop a very general theory of maximum penalized likelihood estimation which should avoid many of these present difficulties. We also demonstrate that each reproducing kernel Hilbert space leads, in a very natural way, to a maximum penalized likelihood estimator and that a well-known class of reproducing kernel Hilbert spaces gives polynomial splines as the nonparametric maximum penalized likelihood estimates. In addition

---

(1) Invited paper presented to the 37th annual meeting of the Institute of Mathematical Statistics in Edmonton, Alberta Canada on August 15, 1974. This research was sponsored by the Office of Naval Research under Contract NR 042-283 and NASA MSC under Contract NAS 9-12776.

(2) Department of Statistics, Princeton University, Princeton, New Jersey 08540.

(3) Department of Mathematical Sciences, Rice University, Houston, Texas 77001.

our general theory is used to show that Good's and Gaskins' non-parametric maximum penalized likelihood estimators are well defined and that one of their estimators gives exponential splines as the estimates. Finally we show that Good's and Gaskins' method of implementation does not in general lead to their estimators.

1. Introduction. Let Ω be a subset of R^n. In this study we consider the problem of estimating the probability density function $\varphi \in L^1(\Omega)$ which gave rise to the random samples $x_1,\ldots,x_N \in \Omega$. The set Ω may be either bounded or unbounded.

As usual we define $L(v)$, the likelihood that $v \in L^1(\Omega)$ gave rise to the samples $x_1,\ldots,x_N$ by

$$L(v) = \prod_{i=1}^{N} v(x_i) \quad . \tag{1.1}$$

Let $H(\Omega)$ be a manifold in $L^1(\Omega)$ and consider the following optimization problem:

$$\begin{gathered} \text{maximize } L(v); \text{ subject to} \\ v \in H(\Omega), \int_\Omega v\,d\mu = 1 \text{ and } v(t) \geq 0 \ \forall\ t \in \Omega \ . \end{gathered} \tag{1.2}$$

We let $d\mu$ denote the Lebesgue measure on Ω. By the maximum likelihood estimator (corresponding to $H(\Omega)$) we mean the functional

$$L^*\colon \Omega^N \to 2^{L^1(a,b)}$$

(A^N denotes the Nth Cartesian product of A with itself and 2^A denotes the subsets of A) which assigns to each $\{x_1,\ldots,x_N\} \in \Omega^N$ the solutions of problem (1.2). Any $v \in L^*(x_1,\ldots,x_N)$ is said to be a maximum likelihood estimate (of the probability density φ) for the samples $\{x_1,\ldots,x_N\}$. The maximum likelihood estimator L^* is said to be well defined if $L^*(x_1,\ldots,x_N)$ consists of exactly one function (equivalently problem (1.2) possesses a unique solution). It is also usual to say that L^* is a parametric estimator if the manifold $H(\Omega)$ is finite dimensional and a nonparametric estimator otherwise.

It is well known and part of the folklore that the standard histogram estimates are parametric maximum likelihood estimates and that when $H(\Omega)$ is a finite dimensional linear manifold the corresponding maximum likelihood estimator is well defined. Except in the case when it is known a priori that $\varphi \in H(\Omega)$, it is generally true that the parametric maximum likelihood estimates are far from satisfactory. Moreover the nonparametric maximum likelihood estimator is essentially undefined. Some justification for these latter two statements follows.

Clearly if the manifold $H(\Omega)$ can approximate the Dirac delta function, i.e., contains nonnegative functions whose support is a given small sphere centered at $x \in \Omega$, integrate to one and have arbitrarily large values at x , then problem (1.1) has no solution. Moreover this approximation property is enjoyed by most infinite dimensional manifolds of $L^1(\Omega)$; hence we should not expect the nonparametric maximum likelihood estimation problem to have a solution. The situation is actually worse for it is often the case that in the parametric case we choose $H(\Omega)$ from a sequence of spaces $\{S_m\}$ where the dimension of S_m is m , $S_{m+1} \supset S_m$ and $\bigcup_{m=1}^{\infty} S_m$ is dense in $L^1(\Omega)$; hence the problem is definitely unstable and somewhat ill defined. Namely we are motivated to choose m large so that we can better approximate the probability density giving r[illegible] to the samples $x_1,\ldots,x_N$; however for large m our problem approximates a problem which has no solution.

Rosenblatt [7] in 1956 performed the first analytical study of the theoretical properties of histograms. In 1962 Parzen constructed a class of estimators which properly included the histogram estimators and examined the consistency properties of the estimators in this class. These results have been improved upon recently by Wahba [10] (1971). Kimeldorf and Wahba [3] in 1970 introduced the application of spline techniques in contemporary statistics. Boneva, Kendall and Stefanov [1] in 1971 and Schnenberg [8] in 1972 examined the use of spline techniques for obtaining from histograms smooth estimates of a probability density function. It is of interest to us that essentially all previous authors seem to either ignore the nonnegativity constraint or attempt to handle it with the seemingly clever trick of working with a function whose square is to be used as the estimate of the probability density; however in the case of maximum likelihood estimation this trick tacitly ignores the nonnegativity constraint. More will be said about this in Sections 3 and 4.

In 1971 Good and Gaskins [2] suggest adjoining a penalty term to the likelihood functional (1.1). They actually suggested two nonparametric maximum penalized likelihood estimators; however they did not show that these estimators were meaningful, i.e., well defined. Moreover in dealing with the nonnegativity constraint in problem (1.2), Good and Gaskins also fell into the trap described above of obtaining the estimate as the square of the solution of an optimization problem; hence Good's and Gaskins' implementation does not, in general, give their estimator.

In Section 2 we give a rigorous definition of the maximum penalized likelihood estimator. We also propose a very natural penalty term in the case when the underlying manifold is a reproducing kernel Hilbert space and show that a very important and well-known class of reproducing kernel Hilbert spaces gives rise to maximum penalized likelihood estimates which are polynomial splines with knots at the sample points.

Sections 3 and 4 contain a rigorous analysis and proof of the fact that the Good and Gaskins maximum penalized likelihood estimators and their pseudo maximum penalized likelihood estimators obtained by their incorrect method of implementation are well defined and in the first case identical, but in the second case distinct. It is also of interest that in Section 3 we show that Good's and Gaskins' first nonparametric maximum penalized likelihood estimator leads to estimates which are exponential splines with knots at the sample points.

Much of our analysis uses the notions of the Fréchet gradient, the Fréchet derivative and the second Fréchet derivative in an abstract Hilbert space. The reader not familiar with these notions is referred to Tapia [9].

2. Maximum Penalized Likelihood Estimators. In order to avoid the pitfalls and numerical instabilities attributed to the presently used maximum likelihood estimation procedures we suggest adjoining a penalty term to the likelihood functional.

Let $H(\Omega)$ be a manifold of real-valued functions defined and integrable on a set $\Omega \subset R^n$, i.e., $H(\Omega) \subset L^1(\Omega)$. Consider a functional $\Phi : H(\Omega) \to R$. Given the samples $x_1, \ldots, x_N \in \Omega$ we define the Φ-penalized likelihood of $v \in H(\Omega)$ by

$$(2.1) \qquad \hat{L}(v) = \prod_{i=1}^{N} v(x_i) \exp(-\Phi(v)) .$$

Consider the constrained optimization problem:

$$(2.2) \qquad \text{maximize } \hat{L}(v); \text{ subject to}$$

$$v \in H(\Omega), \ \int_{\Omega} v \, d\mu = 1 \ \text{ and } \ v(t) \geq 0, \ \forall \ t \in \Omega .$$

The maximum penalized likelihood estimator $\hat{L}^*$ corresponding to the set $H(\Omega)$ and the penalty function Φ is defined in a manner analogous to the definition of the maximum likelihood estimator given in Section 1, using the solutions of problem (2.2). The term parametric, the term nonparametric and the term well defined have the same meaning in this context as in Section 1. For the remainder of the paper we consider the nonparametric case of the maximum penalized likelihood estimator; specifically we will choose $H(\Omega)$ to be either an infinite dimensional Hilbert space or an infinite dimensional manifold in a Hilbert space. In the case when $H(\Omega)$ is a Hilbert space a very natural penalty function to use is $\Phi(v) = \|v\|^2$ where $\|\cdot\|$ denotes the norm on $H(\Omega)$. Consequently when $H(\Omega)$ is a

Hilbert space and we refer to the penalized likelihood functional on $H(\Omega)$ or to the maximum penalized likelihood estimator corresponding to $H(\Omega)$ with no reference to the penalty functional Φ we are assuming that Φ is the square of the norm in $H(\Omega)$. Recall that when $H(\Omega)$ is a Hilbert space it is said to be a reproducing kernel space if point evaluation is a continuous operation, i.e., $v_n \to v$ in $H(\Omega)$ implies $v_n(x) \to v(x)\ \forall\ x \in \Omega$.

In order for problem (2.2) to make sense we would like $H(\Omega)$ to have the property that for each $(x_1,\ldots,x_N) \in \Omega^N$ there exists at least one $v \in H(\Omega)$ such that

$$(2.3) \qquad \int_\Omega v\,d\mu = 1,\ v(t) \geq 0\ \forall\ t \in \Omega \quad \text{and} \quad v(x_i) > 0 \quad i = 1,\ldots,N.$$

<u>Proposition 2.1</u>. Suppose that $H(\Omega)$ is a reproducing kernel space and D is a closed convex subset of $\{v \in H(\Omega): v(x_i) \geq 0\}$ with the property that D contains at least one function which is positive at the points $x_1,\ldots,x_N$. Then the penalized likelihood functional on $H(\Omega)$ has a unique maximizer in D.

<u>Proof</u>. Since $H(\Omega)$ is a reproducing kernel space we have $|v(x_i)| \leq K_i\|v\|$ for $i = 1,\ldots,N$. It follows that

$$(2.4) \qquad |\hat{L}(v)| \leq C_1\|v\|^N \exp(-\|v\|^2) .$$

The function $\theta(\lambda) = \lambda^N \exp(-\lambda^2)$ is bounded above by $(N/2)^{\frac{N}{2}} \exp(-N/2)$; hence $|\hat{L}(v)| \leq C_2$. If $M = \sup\{\hat{L}(v): v \in D\}$, then there exists $\{v_j\} \subset D$ such that $\hat{L}(v_j) \to M$. From our hypothesis $M > 0$. Notice that $\theta(\lambda) \to 0$ as $\lambda \to +\infty$. Hence from (2.4) $\|v_j\| \leq C_3\ \forall\ j$. The ball $\{v \in H(\Omega): \|v\| \leq C_3\}$ is weakly compact. Hence $\{v_j\}$ contains a weakly convergent subsequence which we also denote by $\{v_j\}$. Let v^* denote the weak limit of $\{v_j\}$. We have that $v_j(x_i) \to v^*(x_i)$ as $j \to \infty$ for each $i = 1,\ldots,N$. The norm is a convex functional; hence weakly lower semicontinuous so that $\underline{\lim}\, \|v_j\| \geq \|v^*\|$. It follows that

$$(2.5) \qquad \lim_j \prod_{i=1}^{N} v_j(x_i) \exp(-\|v_j\|^2) \leq \prod_{i=1}^{N} v^*(x_i) \exp(-\|v^*\|^2) .$$

However the left-hand side of (2.5) is equal to M and the right-hand side is equal to $\hat{L}(v^*)$; so $M \leq \hat{L}(v^*)$. Now since D is closed and convex it is weakly closed; hence $v^* \in D$. This establishes the existence of a maximizer.

Since $M > 0$, maximizing $\hat{L}$ over D is equivalent to maximizing $J = \log \hat{L}$ over D. A straightforward calculation gives the second Fréchet derivative of J as

$$J''(v)(\mu,\eta) = - \sum_{i=1}^{N} \frac{\mu(x_i)\eta(x_i)}{v(x_i)^2} - 2 < \mu,\eta > .$$

Now since $J''(v)$ is negative definite J is strictly concave and can therefore have at most one maximizer on a convex set. This proves the proposition.

<u>Proposition 2.2</u>. Suppose $H(\Omega)$ is a reproducing kernel space, integration over Ω is a continuous functional and there exists at least one $v \in H(\Omega)$ satisfying (2.3). Then the maximum penalized likelihood estimator corresponding to $H(\Omega)$ is well defined.

<u>Proof</u>. The proof follows from Proposition 2.1 since the constraints in (2.2) give a closed convex subset of $\{v \in H(\Omega): v(x_i) \geq 0,\ i = 1,\ldots,N\}$.

Recall that by the Sobolev space of order s on the real line we mean

$$(2.6) \qquad H^s(-\infty,\infty) = \{\mu \in S':(1+\omega^2)^{\frac{s}{2}}F[\mu](\omega) \in L^2(-\infty,\infty)\}$$

where S' is the space of distributions with polynomial increase at infinity and $F[u]$ denotes the Fourier transform of u. The norm of $u \in H^s(-\infty,\infty)$ is given by

$$(2.7) \qquad \|u\|_{H^s(-\infty,\infty)} = \|(1+\omega^2)^{\frac{s}{2}} F[u](\omega)\|_{L^2(-\infty,\infty)} .$$

If s is an integer, then $u \in H^s(-\infty,\infty)$ if and only if $u, u^{(1)}, \ldots, u^{(s)} \in L^2(-\infty,\infty)$ and an equivalent norm is given by

$$(2.8) \qquad [\sum_{i=0}^{s} w_i \|u^{(i)}\|^2_{L_2(-\infty,\infty)}]^{\frac{1}{2}}$$

where $w_i \geq 0$ and $w_0, w_s > 0$. We have the analogous definitions in the case of the finite interval; however when considering the Fourier transform we must extend the function to the entire interval $(-\infty,\infty)$. As in the previous section the notation $H^s(a,b)$ does not preclude the possibility that either a or b (or both) may be infinite. The reader interested in more detail is refered to Lions and Magenes [5].

Lemma 2.3. The Sobolev space $H^s(a,b)$ is a reproducing kernel space if and only if $s > \frac{1}{2}$. Moreover for $s > \frac{1}{2}$ the linear functional $I: H^s(a,b) \to R$ defined by

$$I(v) = \int_a^b v(t)dt$$

is continuous if and only if $[a,b]$ is a finite interval.

Proof. The proof follows in a reasonably straightforward manner using results in Lions [3].

Proposition 2.4. The maximum penalized likelihood estimator corresponding to the Hilbert space $H^s(a,b)$ where $s > \frac{1}{2}$ and $[a,b]$ is a finite interval containing the sample points is well defined.

Proof. The proof follows from Proposition 2.2 and Lemma 2.3.

Recall that if s is an integer, then

$$H_0^s(a,b) = \{u \in H^s(a,b): u^{(k)}(a) = u^{(k)}(b) = 0,\ k = 0,\dots,s-1\}$$

Let $\bar{H}_0^s(a,b)$ be the collection of functions in $H_0^s(a,b)$ with the Hilbert space structure induced instead by the inner product

$$(2.9) \qquad < u,v > = \int_a^b u^{(s)} v^{(s)} \quad .$$

It can be shown that $H_0^s(a,b)$ and $\bar{H}_0^s(a,b)$ are equivalent, i.e., have the same topology, in a manner similar to that which shows that (2.7) and (2.8) are equivalent. Clearly $H_0^s(a,b)$ and $\bar{H}_0^s(a,b)$ do not have the same inner product.

<u>Theorem 2.5</u>. Suppose (a,b) is a finite interval properly containing the sample points $x_1,\dots,x_N$. Let s be a positive integer. Then the maximum penalized likelihood estimator corresponding to $\bar{H}_0^s(a,b)$ is well defined and gives as an estimate a polynomial spline of degree $2s$. Moreover, if the estimate is positive in the interior of an interval, then in this interval it is a polynomial spline of degree $2s$ and of continuity class $2s-2$ with knots exactly at the sample points.

<u>Proof</u>. Clearly $\bar{H}_0^s(a,b)$ is a reproducing kernel Hilbert space since $H_0^s(a,b)$ is such a space. It follows that the maximum penalized likelihood estimator corresponding to $\bar{H}_0^s(a,b)$ is well defined from Proposition 2.2.

Consider an interval $I_+ = [\alpha,\beta] \subset [a,b]$. Let $I_- = \{t \in [a,b]: t \notin [\alpha,\beta]\}$. Define the two functionals J_+ and J_- on $\bar{H}_0^s(a,b)$ by

$$J_+(v) = \sum_i \log v(x_i) - \int_{I_+} v(t)^2 dt$$

and

$$J_-(v) = \sum_i \log v(x_i) - \int_{I_-} v(t)^2 dt \quad ,$$

where the summation in the first formula is taken over all i such that $x_i \in I_+$ and the summation in the second formula is taken over all i such that $x_i \in I_-$. It should be clear that

$$J(v) = J_+(v) + J_-(v)$$

where as before $J(v) = \log \hat{L}(v)$ and $\hat{L}$ is the penalized likelihood in $\bar{H}_0^s(a,b)$. Let V_* denote the maximum penalized likelihood estimate for the samples $x_1,\ldots,x_N$. Suppose V_* is positive on the interval I_+. We claim that V_* restricted to this interval solves the following constrained optimization problem:

$$\text{maximize} \quad J_+(v); \text{ subject to}$$

$$(2.10) \qquad v \in H^s(a,b)\ ,\ v^{(m)}(\alpha) = v_*^{(m)}(\alpha)\ ,\ v^{(m)}(\beta) = v_*^{(m)}(\beta)\ , \quad m = 0,\ldots,s-1\ ,$$

$$\int_{I_+} v(t)dt = \int_{I_+} v_*(t)dt \text{ and } v(t) \le 0\ ,\ t \in I_+\ .$$

To see this observe that if v_+ satisfies the constraints of problem (2.10) and $J_+(v_*) < J_+(v_+)$, then the function v^* defined by

$$v^*(t) = \begin{cases} v_+(t)\ , & t \in I_+ \\ v_*(t)\ , & t \in I_- \end{cases}$$

satisfies the constraints of problem (2.2) with $\bar{H}_0^s(a,b)$ playing the role of $H(\Omega)$ and $J(v_*) = J_+(v_*) + J_-(v_*) < J_+(v_+) + J_-(v_+) = J(v^*)$, which in turn implies that $L(v_*) < L(v^*)$; however this contradicts the optimality of v^*. Now define the functional G on $\bar{H}_0^s(\alpha,\beta)$ by

$$G(v) = J_+(v_+ + v) \quad \text{for} \quad v \in \overline{H}_0^s(\alpha,\beta) \ .$$

Consider the constrained optimization problem

$$\text{maximize} \quad G(v) \ ; \text{ subject to}$$

$$(2.11) \qquad v \in \overline{H}_0^s(\alpha,\beta) \quad \text{and} \quad \int_{I_+} v = 0 \ .$$

If v satisfies the constraints of problem (2.11), then $v_* + tv$ satisfies the constraints of problem (2.10) for t sufficiently small, since v_* is positive in I_+. It follows that the zero function is the unique solution of problem (2.11). From the theory of Lagrange multipliers we therefore must have

$$(2.12) \qquad \nabla G(0) + \lambda v_0 = 0 \ ,$$

where λ is a real number, $\nabla G(0)$ is the Frechet gradient of G at 0 and v_0 is the Fréchet gradient of the functional $v \to \int_{I_+} v$ in the space $\overline{H}_0^s(\alpha,\beta)$. Clearly in this case v_0 is merely the Riesz representer of the functional $v \to \int_{I_+} v$.

Specifically

$$\int_{I_+} v_0^{(s)} v^{(s)} = \int_{I_+} v \ .$$

Integrating by parts in the distribution sense we see that $v_0^{(2s)} = 1$; hence v_0 is a polynomial of degree $2s$ in $[\alpha,\beta]$. A straightforward calculation shows that

$$(2.13) \qquad \nabla G(0) = J_+(v_*)\Big(\sum_i \frac{v_i}{v_*(x_i)} - 2v_*\Big)$$

where the summation is taken over i such that $x_i \in I_+$ and v_i is the Riesz representer of the functional $v \to v(x_i)$ in $\overline{H}_0^{(s)}(\alpha,\beta)$, i.e.,

$$\int_{I_+} v_i^{(s)} v^{(s)} = v(x_i) \ .$$

As before integrating by parts in the distribution sense we see that $v_i^{(2s)} = \delta_i$ where δ_i is the Dirac mass at the point x_i . It follows that v_i is a polynomial spline of degree 2s-1 and of continuity class 2s-2 with a knot exactly at the sample point x_i . From (2.12) and (2.13) we have that v_* restricted to the interval $[\alpha,\beta]$ is a polynomial spline of degree 2s and of continuity class 2s-2 with knots exactly at the sample points in $[\alpha,\beta]$. A simple continuity argument takes care of the case when v_* is only positive on the interior of $[\alpha,\beta]$. This proves the theorem.

Remark. Observe that Theorem 2.5 says that the spline estimate is necessarily zero at knots which are not sample points.

In the case when $s=1$ we can say substantially more about the distribution of the knots and zeros of the spline estimate.

Theorem 2.6. Suppose (a,b) is a finite interval properly containing the sample points $x_1, \ldots, x_N$. Then the maximum penalized likelihood estimator corresponding to $\overline{H}_0^1(a,b)$ is well defined and gives as an estimate a continuous quadratic spline with knots at the sample points and at most two knots in the interior of each interval $[x_i,x_{i+1}]$, $i = 0,\ldots,N+1$ ($x_0 = a$ and $x_{N+1} = b$) . Moreover in each such interval the spline is either zero at no points, zero at one point (which must be a knot) or zero on a proper subinterval whose end-points are necessarily knots.

Proof. Suppose the estimate v_* is zero at α and β where $x_i \leq \alpha < \beta \leq x_{i+1}$ and not identically zero in $[\alpha,\beta]$. Consider the function

$$v^*(t) = \begin{cases} v_*(t) \ , & t \notin [\alpha,\beta] \\ 0 \ , & t \in [\alpha,\beta] \ . \end{cases}$$

Clearly $\gamma = 1/\int_{\alpha}^{\beta} v^* > 1$. We also have that

$$J(\alpha v^*) > J(v^*) > J(v_*)$$

and that $\gamma v^* \in \overline{H}_0^1(a,b)$, $\gamma v^*(t) \geq 0$ for $t \in [a,b]$ and $\int_{\alpha}^{\beta} \gamma v^* = 1$. This, however, contradicts the optimality of v_* and shows that v_* must be identically zero in the interval $[\alpha,\beta]$. The remainder of the theorem follows from Theorem 2.5 and the remark following it.

3. The First Maximum Penalized Likelihood Estimator of Good and Gaskins.

In [2] Good and Gaskins consider the maximum penalized likelihood estimator corresponding to the penalty function

$$\Phi_1(v) = \alpha \int_{-\infty}^{\infty} \frac{v'(t)^2}{v(t)}\, dt \qquad (\alpha > 0) .$$

They do not define the manifold $H(\Omega)$; but it is obvious from the constraints that must be satisfied and the fact that

$$\frac{1}{4}\Phi_1(v) = \alpha \int_{-\infty}^{\infty} \left(\frac{d\sqrt{v}}{dt}\right)^2 dt$$

that the underlying manifold $H(\Omega)$ should be

$$\sqrt{H^1(-\infty,\infty)} = \{v \in L^1(-\infty,\infty): \sqrt{v} \in H^1(-\infty,\infty)\} .$$

This leads us to analyzing the following constrained optimization problem:

$$\text{(3.1)} \qquad \text{maximize} \quad \hat{L}_1(v) = \prod_{i=1}^{N} v(x_i)\exp(-\Phi_1(v)); \quad \text{subject to}$$

$$v \in \sqrt{H^1(-\infty,\infty)}, \ \int_{-\infty}^{\infty} v(t)dt = 1 \ \text{ and } \ v(t) \geq 0 \ \ \forall\ t \in (-\infty,\infty) .$$

In an effort to avoid the nonnegativity constraint in problem (3.1) Good and Gaskins considered working with the $\sqrt{v}$ instead of v. Specifically if we let $u = \sqrt{v}$, then restating problem (3.1) in terms of u we obtain

$$\text{(3.2)} \qquad \text{maximize} \quad \prod_{i=1}^{N} u(x_i)^2 \exp(-4\alpha\int_{-\infty}^{\infty} u'(t)^2 dt) \ ; \ \text{subject to}$$

$$u \in H^1(-\infty,\infty),\ \int_{-\infty}^{\infty} u(t)^2 dt = 1 \quad \text{and} \quad u(t)^2 \geq 0,\ t \in (-\infty,\infty) .$$

Since the constraint $u(t)^2 \geq 0$ is redundant they suggest solving problem (3.2) for u_* and then accepting $v_* = u_*^2$ as the solution of problem (3.1). On first impressions everything looks fine; however a moments reflection should convince the reader that what tacitly has been assumed is that the unique solution of problem (3.2) is actually nonnegative. Hence adding the nonnegativity constraint to problem (3.2) and restating in the equivalent form obtained by taking the square root of the objective functional (since it is nonnegative) we arrive at the following constrained optimization problem:

$$\text{(3.3)} \qquad \text{maximize} \quad \hat{L}(v) = \prod_{i=1}^{N} v(x_i)\exp(-\Phi(v)); \quad \text{subject to}$$

$$v \in H^1(-\infty,\infty)\ ,\ \int_{-\infty}^{\infty} v(t)^2 dt = 1 \quad \text{and} \quad v(t) \geq 0,\ \forall\ t \in (-\infty,\infty)$$

where

$$\Phi(v) = 2\alpha\int_{-\infty}^{\infty} v'(t)^2 dt$$

and α is given in problem (3.1).

<u>Proposition 3.1</u>. Problems (3.1) and (3.3) are equivalent in the sense that if v_* is a solution of problem (3.1), then $\sqrt{v_*}$ is a solution of problem (3.3) and if v_* is a solution of problem (3.3), then v_*^2 is a solution of problem (3.1).

<u>Proof</u>. The proof follows from the fact that if $v \geq 0$, then

$$\Phi(\sqrt{v}) = \frac{1}{2}\Phi_1(v)$$

and

$$\hat{L}_1(v) = \hat{L}(\sqrt{v})^2 .$$

It is very surprising and quite fortunate that Good's and Gaskins' omission does not really effect this estimator; since we will presently show that the nonnegativity constraint in problem (3.3) is not active at the solution, i.e., problems (3.2) and (3.3) actually have the same solutions. Unfortunately this will not be the case for the second maximum penalized likelihood estimator Good and Gaskins propose. Good and Gaskins did not show that their estimators are well defined; hence this is our first task. Along with problem (3.3) we will consider the constrained optimization problem obtained by only requiring nonnegativity at the sample points:

(3.4) maximize $\hat{L}(v)$; subject to

$$v \in H^1(-\infty,\infty) \ , \int_{-\infty}^{\infty} v(t)^2 dt = 1 \quad \text{and} \quad v(x_i) \geq 0, \ i = 1,\ldots,N \ .$$

Given $\lambda > 0$ and α in problem (3.3) we may also consider the constrained optimization problem:

(3.5) maximize $\hat{L}_\lambda(v) = \prod_{i=1}^{N} v(x_i)\exp(-\Phi_\lambda(v))$; subject to

$$v \in H^1(-\infty,\infty) \ , \int_{-\infty}^{\infty} v(t)^2 dt = 1 \quad \text{and} \quad v(x_i) \geq 0, \ i = 1,\ldots,N$$

where

$$\Phi_\lambda(v) = 2\alpha\int_{-\infty}^{\infty} v'(t)^2 dt + \lambda\int_{-\infty}^{\infty} v(t)^2 dt \ .$$

Our study of problem (3.5) will begin with the study of the following constrained optimization problem:

(3.6) maximize $\hat{L}_\lambda(v)$; subject to

$$v \in H^1(-\infty,\infty) \quad \text{and} \quad v(x_i) \geq 0 \ , \quad i = 1,\ldots,N$$

where $\hat{L}_\lambda$ is given by problem (3.5). Let $L^2 = L^2(-\infty,\infty)$.

Proposition 3.2. Problem (3.6) has a unique solution. Moreover if v_λ denotes this solution, then

(i) v_λ is an exponential spline with knots at the sample points $x_1,\ldots, x_N$;

(ii) $v_\lambda(t) > 0$, $\forall\, t \in (-\infty,\infty)$; and

(iii) $\|v_\lambda\|_{L^2} \geq \sqrt{\pi/(4\lambda)}$.

Proof. From Lemma 2.3 $H^1(-\infty,\infty)$ is a reproducing kernel space. Also $\|v\|_\lambda^2 = \Phi_\lambda(v)$ gives a norm equivalent to the original norm on $H^1(-\infty,\infty)$. The existence of v_λ now follows from Proposition 2.1 with $D = \{v \in H^1(-\infty,\infty)$: $v(x_i) \geq 0,\ i = 1,\ldots,N\}$. We will denote the Φ_λ inner product by $<\,,\,>_\lambda$. Let v_i be the representer in the Φ_λ inner product of the continuous linear functional given by point evaluation at the point x_i, $i = 1,\ldots,N$, i.e.

$$< v_i,\eta >_\lambda = \eta(x_i) \ , \ \forall\, \eta \in H^1(-\infty,\infty) \ .$$

Equivalently

$$2\alpha\int_{-\infty}^{\infty} v_i'(t)\eta'(t)dt + \lambda\int_{-\infty}^{\infty} v_i(t)\eta(t)dt = \eta(x_i) \ , \ \forall\, \eta \in H^1(-\infty,\infty) \ .$$

Integrating by parts in the distribution sense gives

$$\int_{-\infty}^{\infty}[-2\alpha v_i''(t) + \lambda v_i(t)]\ \eta(t)dt = \eta(x_i) \ , \ \forall\, \eta \in H^1(-\infty,\infty) \ ;$$

hence

$$(3.7) \qquad -2\alpha v_i'' + \lambda v_i = \delta_i \quad , \qquad i = 1,\ldots,N$$

where $\delta_i(t) = \delta_0(t-x_i)$ and δ_0 denotes the Dirac distribution, i.e., $\int_{-\infty}^{\infty}\delta_0(t)\eta(t)dt = \eta(0)$. If we let v_0 be the solution of (3.7) for $i = 0$, then

$$v_0(t) = \begin{cases} \frac{1}{2\sqrt{2\alpha\lambda}} \exp(\sqrt{\lambda}/(2\alpha) t) & , \quad t < 0 \\ \frac{1}{2\sqrt{2\alpha\lambda}} \exp(-\sqrt{\lambda}/(2\alpha) t) & , \quad t > 0 \end{cases}$$

and $v_i(t) = v_0(t-x_i)$ for $i = 1,\dots,N$. Since v_λ is the maximizer we have that $v_\lambda(x_i) > 0$, $i = 1,\dots,N$ we necessarily have that the Frechet derivative of $\hat{L}_\lambda$ at v_λ must be the zero functional; equivalently the gradient of $\hat{L}_\lambda$ or for that matter the gradient of log $\hat{L}_\lambda$ must vanish at v_λ since $\hat{L}_\lambda$ and log $\hat{L}_\lambda$ have the same maxima. A calculation similar to that used in the proof of Proposition 2.1 gives

$$\nabla_\lambda \log \hat{L}_\lambda(v) = 2v - \sum_{i=1}^{N} \frac{v_i}{v(x_i)} \tag{3.8}$$

where ∇_λ denotes the gradient. It follows from (3.8) that

$$v_\lambda = \frac{1}{2} \sum_{i=1}^{N} \frac{v_i}{v_\lambda(x_i)} \quad . \tag{3.9}$$

Properties (i) and (ii) are now immediate. Since $< v_i, v_\lambda >_\lambda = v_\lambda(x_i)$ from (3.9) we have

$$\|v_\lambda\|_\lambda^2 = N/2 \quad . \tag{3.10}$$

A straightforward calculation shows that

$$v_i'(t)v_j'(t) \le \frac{\lambda}{2\alpha} v_i(t)v_j(t) \quad , \text{ for } \quad i,j = 1,\dots,N \quad .$$

So

$$v_\lambda'(t)^2 = \frac{1}{4} [\sum_i (\frac{v_i'(t)}{v_\lambda(x_i)})^2 + \sum_{i,j} \frac{v_i'(t)v_j'(t)}{v_\lambda(x_i)v_\lambda(x_j)}]$$

$$\le \frac{\lambda}{8\alpha} [\sum_i (\frac{v_i(t)}{v_\lambda(x_i)})^2 + \sum_{i,j} \frac{v_i(t)v_j(t)}{v_\lambda(x_i)v_\lambda(x_j)}] = \frac{\lambda}{2\alpha} v_\lambda(t)^2 \quad .$$

Integrating in t gives

$$2\alpha\|v_\lambda'\|^2_{L^2(-\infty,\infty)} \leq \lambda\|v_\lambda\|^2_{L^2(-\infty,\infty)} \quad .$$

By definition of the $\mathcal{E}_\lambda$-norm and (3.10) we have property (iii). This proves the proposition.

Proposition 3.3. Problem (3.4) has a unique solution.

Proof. Let $B = \{v \in H^1(-\infty,\infty): \int_{-\infty}^{\infty} v(t)^2 dt \leq 1$ and $v(x_i) \geq 0$, $i = 1,\ldots,N\}$. Clearly B is closed and convex. If $\hat{L}_\lambda$ is given by (3.5), then by Proposition 2.1 the functional has a unique maximizer in B; say u_λ. Now by property (iii) of Proposition 3.2 if we choose $0 < \lambda < \frac{1}{4}$, then v_λ the unique solution of problem (3.6) will be such that $\|v_\lambda\|_{L^2(-\infty,\infty)} > 1$. We will show that for this range of $\lambda, \|u_\lambda\|_{L^2(-\infty,\infty)} = 1$. Consider $v_\theta = \theta v_\lambda + (1-\theta)u_\lambda$. We know that $\log \hat{L}_\lambda$ is a strictly concave functional (see the proof of proposition 2.1). Moreover $\log \hat{L}_\lambda(v_\lambda) > \log \hat{L}_\lambda(u_\lambda)$; hence $\log \hat{L}_\lambda(v_\theta) > \log \hat{L}_\lambda(u_\lambda)$ for $0 < \theta < 1$. Now suppose $\|u_\lambda\|_{L^2(-\infty,\infty)} < 1$ and consider

$$g(\theta) = \|v_\theta\|_{L^2(-\infty,\infty)} \quad .$$

We have $g(0) < 1$ and $g(1) > 1$. So for some $0 < \theta_0 < 1$, $g(\theta_0) = 1$ and $\log \hat{L}_\lambda(u_\lambda) < \log \hat{L}_\lambda(v_{\theta_0})$. This is a contradiction since u_λ is the unique maximizer of $\hat{L}_\lambda$ in B; hence $\|u_\lambda\|_{L^2(-\infty,\infty)} = 1$. This shows that u_λ is the unique solution of problem (3.5) for $0 < \lambda < \frac{1}{4}$. However, the term $\lambda \int_{-\infty}^{\infty} v(t)^2 dt$ is constant over the constraint set in problems (3.4) and (3.5); hence problems (3.4) and (3.5) have the same solutions for any $\lambda > 0$. This proves the proposition since we have demonstrated that problem (3.3) has a unique solution for at least one λ.

Proposition 3.4. Problem (3.3) has a unique solution which is positive and an exponential spline with knots at the points $x_1,\ldots,x_N$.

Proof. If we can demonstrate that $\tilde{v}$ the unique solution of problem (3.4) has these properties we will be through. Let $G(v) = \log \hat{L}(v)$ where $\hat{L}$ is given in problem (3.3) and let

$$g(v) = \int_{-\infty}^{\infty} v(t)^2 dt$$

for $v \in H^1(-\infty,\infty)$. Clearly $\tilde{v}(x_i) > 0$ for $i = 1,\ldots,N$; hence from the theory of Lagrange multipliers there exist λ such that $\tilde{v}$ satisfies the equations

$$G'(v) - \lambda g'(v) = 0 \quad \text{and} \quad g(v) = 1. \tag{3.11}$$

Using $L^2(-\infty,\infty)$ gradients in the sense of distributions (3.11) is equivalent to

$$-4\alpha v'' + 2\lambda v = \sum_{i=1}^{N} \frac{\delta_i}{v(x_i)} \quad \text{and} \quad g(v) = 1 \tag{3.12}$$

where δ_i is the distribution such that $\int_{-\infty}^{\infty} v(t)\delta_i(t)dt = v(x_i)$, $i = 1,\ldots,N$. Since we have already established that problem (3.4) has a unique solution it follows that (3.12) must have a unique solution in $H^1(-\infty,\infty)$; namely $\tilde{v}$. If $\lambda \leq 0$, then any solution of the first equation in (3.12) would be a sum of trigonemetric functions and could not possibly satisfy the constraint $g(v) = 1$, i.e., can not be contained in $L^2(-\infty,\infty)$. It follows that $\lambda > 0$. Now observe that

$$G - \lambda g = \log \hat{L}_\lambda$$

where $\hat{L}_\lambda$ is given by problem (3.5); hence if $\tilde{v}$ satisfies (3.11) (from the first equation alone) it must also be a solution of problem (3.6) for this λ and therefore has the desired properties according to Proposition 3.2. This proves the proposition.

Proposition 3.5. The first nonparametric maximum penalized likelihood estimator of Good and Gaskins is well defined; specifically the maximum penalized likelihood estimator corresponding to the penalty function

$$\Phi(v) = \alpha \int_{-\infty}^{\infty} \frac{v'(t)^2}{v(t)}\, dt \qquad (\alpha > 0)$$

and the manifold

$$H(\Omega) = \{v \in L^1(-\infty,\infty) : \sqrt{v} \in H^1(-\infty,\infty)\}$$

is well defined. Moreover the estimate for the sample points $x_1,\ldots,x_N$ given by this estimator is positive and an exponential spline with knots at the sample points.

Proof. From Proposition 3.1 this estimate is $\tilde{v}^2$ where $\tilde{v}$ solves problem (3.3). By Proposition 3.4 $\tilde{v}$ is positive and an exponential spline with knots at $x_1,\ldots,x_N$; hence so is $\tilde{v}^2$. This proves the proposition.

4. The Second Maximum Penalized Likelihood Estimator of Good and Gaskins.

Consider the functional $\Phi : H^2(-\infty,\infty) \to R$ defined by

$$(4.1) \qquad \Phi(v) = \alpha \int_{-\infty}^{\infty} v'(t)^2 dt + \beta \int_{-\infty}^{\infty} v''(t)^2 dt$$

for some $\alpha \geq 0$ and $\beta > 0$. Also consider the functional Φ_1 defined on $\sqrt{H^2(-\infty,\infty)} = \{v \in L^1(-\infty,\infty) : \sqrt{v} \in H^2(-\infty,\infty)\}$ by

$$(4.2) \qquad \Phi_1(v) = \Phi(\sqrt{v})$$

where Φ is given by (4.1). By the second maximum penalized likelihood estimator of Good and Gaskins we mean the estimator corresponding to the manifold $\sqrt{H^2(-\infty,\infty)}$ and the penalty function Φ_1. Hence we must consider the following constrained optimization problem:

$$(4.3) \qquad \text{maximize } \hat{L}_1(v) = \prod_{i=1}^{N} v(x_i)\exp(-\Phi_1(v)); \text{ subject to}$$
$$v \in \sqrt{H^2(-\infty,\infty)},\ \int_{-\infty}^{\infty} v(t)dt = 1 \text{ and } v(t) \geq 0\ \forall\ t \in (-\infty,\infty)\ .$$

As in the first case (described in the previous section) Good and Gaskins suggest avoiding the nonnegativity constraint by calculating the solution of problem (4.3) from the following constrained optimization problem:

$$(4.4) \qquad \text{maximize } \hat{L}(v) = \prod_{i=1}^{N} v(x_i)\exp(-\tfrac{1}{2}\Phi(v)); \text{ subject to}$$
$$v \in H^2(-\infty,\infty),\ \int_{-\infty}^{\infty} v(t)^2 dt = 1 \text{ and } v(x_i) \geq 0,\ i = 1,\ldots,N$$

where Φ is given by (4.1).

Clearly problems (4.3) and (4.4) are equivalent in the sense that the solution of one can be obtained from the solution of the other by either taking the square or square root if and only if the solutions of problem (4.4) are nonnegative. Moreover we will presently demonstrate that the solutions of problem (4.4) are <u>not</u> necessarily nonnegative. It will then follow that we can not obtain the second estimator by considering problem (4.4). If we naively use v_*^2, where v_* solves problem (4.4), as an estimate for the probability density function giving rise to the samples $x_1,\ldots,x_N$, then clearly v_*^2 will be nonnegative and integrate to 1 and is therefore a probability density; however the estimator obtained in this manner will not in the strict sense of our definition be a maximum penalized likelihood estimator. For this reason we will refer to this latter estimator as the pseudo maximum penalized likelihood estimator of Good and Gaskins.

The next six propositions are needed to show that the second maximum penalized likelihood estimator and the pseudo maximum penalized likelihood estimator of Good and Gaskins are distinct and well defined.

<u>Proposition 4.1.</u> The second maximum penalized likelihood estimator and the pseudo maximum likelihood estimator of Good and Gaskins are distinct.

<u>Proof.</u> We will show that it is possible for problem (4.4) to have solutions which are not nonnegative. Toward this end let $N = 1$, $x_1 = 0$, $\alpha = 0$ and $\beta = 2$. Let $G(v) = \log \hat{L}(v)$, i.e.,

$$G(v) = \log v(0) - \int_{-\infty}^{\infty} v''(t)^2 dt$$

and let

$$g(v) = \int_{-\infty}^{\infty} v(t)^2 dt \; .$$

As in the proof of Proposition 3.4 using the theory of distributions and the theory of Lagrange multipliers we see that the solutions of problem (4.4) in this case are exactly the solutions of

$$v^{(iv)} + \lambda v = \frac{\delta_1}{2v(0)} \quad \text{and} \quad g(v) = 1 \tag{4.5}$$

where δ_1 is defined in the proof of Proposition 3.4. If we let $\tilde{v}$ denote the Fourier transform of v, then taking the Fourier transform of the first expression in (4.5) gives

$$\tilde{v}(\omega) = [2v(0)(\lambda + 16\Pi^4\omega^4)]^{-1} \; .$$

Since $\|\tilde{v}\|_{L^2(-\infty,\infty)} = \|v\|_{L^2(-\infty,\infty)} = 1$ we must have

$$\int_{-\infty}^{\infty} \frac{d\omega}{(\lambda+16\Pi^4\omega^4)^2} = 4v(0)^2 \; . \tag{4.6}$$

For the integral in (4.6) to exist we must have $\lambda > 0$. Now the inverse Fourier transform of $(\lambda+16\Pi^4\omega^4)^{-1}$ is given by v where

$$v(t) = \begin{cases} \dfrac{e^{bt}}{8b^3}[\cos bt - \sin bt] \; , & t \leq 0 \\[2ex] \dfrac{e^{-bt}}{8b^3}[\cos bt + \sin bt] \; , & t > 0 \; . \end{cases} \tag{4.7}$$

with $b = \lambda^{\frac{1}{4}}/\sqrt{2}$. From (4.7) $v(0) = (8b^3)^{-1}$ and from (4.6) $v(0)^2 = \frac{1}{4}\lambda^{-\frac{7}{4}} K$

where $K = \|(1+16\pi^4\omega^4)^{-1}\|^2_{L^2(-\infty,\infty)}$. Hence $\lambda^{\frac{1}{4}} = 2K$ and $b = \sqrt{2}K$. It follows that the unique solution of problem (4.4) is given by (4.7) with $b = \sqrt{2}K$ which is clearly not nonnegative. This proves the proposition.

We will devote the remainder of this section to showing that both the second estimator and the pseudo estimator are well defined. The approach taken will be very similar to that used in Section 3 to show that the first estimator of Good and Gaskins is well defined.

Given $\lambda > 0$ consider the constrained optimization problem:

(4.8) $$\text{maximize } \hat{L}_\lambda(v) = \prod_{i=1}^{N} v(x_i)\exp(-\Phi_\lambda(v)); \text{ subject to}$$

$$v \in H^2(-\infty,\infty), \int_{-\infty}^{\infty} v(t)^2 dt = 1 \text{ and } v(x_i) \geq 0, \quad i = 1,\ldots,N$$

where

$$\Phi_\lambda(v) = \frac{1}{2}\Phi(v) + \lambda\int_{-\infty}^{\infty} v(t)^2 dt$$

with $\Phi(v)$ given by (4.1).

As before we also consider the constrained optimization problem obtained by dropping the integral constraint:

(4.9) $$\text{maximize } \hat{L}_\lambda(v) \text{ ; subject to}$$

$$v \in H^2(-\infty,\infty) \text{ and } v(x_i) \geq 0, \qquad i = 1,\ldots,N.$$

Proposition 4.2. Problem (4.9) has a unique solution. Moreover if v_λ denotes this solution, then

$$\|v_\lambda\|_{L^2(-\infty,\infty)} \to +\infty \text{ as } \lambda \to 0.$$

Proof. By Lemma 2.3 the Sobolev space $H^2(-\infty,\infty)$ is a reproducing kernel space. Moreover, if

$$\|v\|_\lambda^2 = \delta_\lambda(v) \quad ,$$

then an integration by parts gives

$$(4.10) \qquad \begin{aligned} \|v'\|_{L^2}^2 &= |\langle v,v''\rangle_{L^2}| \le \|v\|_{L^2}\|v''\|_{L^2} \\ &\le \tfrac{1}{2}[\|v\|_{L^2}^2 + \|v''\|_{L^2}^2] \end{aligned}$$

where L^2 denotes $L^2(-\infty,\infty)$; hence $\|\cdot\|_\lambda$ is equivalent to the original norm on $H^2(-\infty,\infty)$. The existence and uniqueness of v_λ now follows from Proposition 2.1.

We must now show that $\|v_\lambda\|_{L^2} \to +\infty$ as $\lambda \to 0$. From the fundamental theorem of calculus we have

$$(4.11) \qquad \begin{aligned} v(x)^2 &= \int_{-\infty}^{x} \frac{dv(t)^2}{dt}\, dt = 2\int_{-\infty}^{x} v(t)v'(t)dt \\ &\le 2\|v\|_{L^2}\|v'\|_{L^2} \quad . \end{aligned}$$

Also, $\|v''\|_{L^2} \le \|v\|_\lambda/\sqrt{\beta}$ so that from (4.10) and (4.11)

$$(4.12) \qquad v(x)^2 \le 2\, \|v\|_{L^2}^{\frac{3}{2}} \sqrt{\|v\|_\lambda/\sqrt{\beta}} \,.$$

Evaluating (4.12) at x_i, taking logs (since $v(x_i) \ge 0$) and summing over i gives

$$(4.13) \qquad \sum_{i=1}^{N} \log v(x_i) \le \frac{N}{4} \log\left(\frac{4}{\sqrt{\beta}}\|v\|_\lambda\right) + \frac{3N}{4} \log(\|v\|_{L^2}) \,.$$

Hence from (4.13) we see that

$$(4.14) \qquad \log \hat{L}_\lambda(v) \le \frac{3N}{4} \log(\|v\|_{L^2}) + \frac{N}{4} \log\left(\frac{4}{\sqrt{\beta}}\, \|v\|_\lambda\right) - \|v\|_\lambda^2 \,.$$

In a manner exactly the same as that used to establish (3.10) we have that

$\|v_\lambda\|_\lambda^2 = \frac{N}{2}$. Hence from (4.14) and the fact that $\log \hat{L}_\lambda(v) \le \log \hat{L}_\lambda(v_\lambda)$ we obtain

$$\text{(4.15)} \qquad \log \hat{L}_\lambda(v) \le \frac{3N}{4} \log(\|v_\lambda\|_{L^2}) + \frac{N}{8} \log(8N/\beta) - \frac{N}{2} ,$$

for any $v \in \{u \in H^2(-\infty,\infty): u(x_i) \ge 0, \ i = 1,\ldots,N\}$.

Let a and b be such that

$$a < \min_i(x_i) \quad \text{and} \quad \max_i(x_i) < b .$$

Given $\lambda > 0$ and ϵ and δ define the function θ_λ in the following piece-wise fashion:

$$\theta_\lambda(t) = \begin{cases} \lambda^\epsilon \exp(-(t-a)^2/2\sigma^2) & \text{for } t \in (-\infty,a) \\ \lambda^\epsilon & \text{for } t \in [a,b] \\ \lambda^\epsilon \exp(-(t-b)^2/2\sigma^2) & \text{for } t \in (b,+\infty) \end{cases}$$

where $\sigma = \lambda^\delta$. Straightforward calculations can be used to show

$$\log(\prod_{i=1}^N \theta_\lambda(x_i)) = \epsilon N \log(\lambda) ,$$

$$\|\theta_\lambda\|_{L^2}^2 = (b-a)\lambda^{2\epsilon} + \sqrt{\Pi}\lambda^{2\epsilon+\delta} ,$$

$$\|\theta_\lambda'\|_{L^2}^2 = \sqrt{2\Pi}\lambda^{2\epsilon-\delta} ,$$

$$\|\theta_\lambda''\|_{L^2}^2 = 2\sqrt{2\Pi}\lambda^{2\epsilon-3\delta} ,$$

and

$$\text{(4.16)} \qquad \|\theta_\lambda\|_\lambda^2 = (b-a)\lambda^{2\epsilon+1} + \sqrt{\Pi}\lambda^{2\epsilon+\delta+1} + 4\alpha\sqrt{2\Pi}\lambda^{2\epsilon-\delta} + 2\beta\sqrt{2\Pi}\lambda^{2\epsilon-3\delta} .$$

If we want $\|\theta_\lambda\|_\lambda^2 \to 0$ as $\lambda \to 0$ it is sufficient to choose all exponents of λ in (4.16) positive. If we also want

$$\log\left(\prod_{i=1}^{N} \theta_\lambda(x_i)\right) \to +\infty \quad \text{as} \quad \lambda \to 0$$

we should choose $\epsilon < 0$. This leads to the inequalities

$$\begin{aligned} &2\epsilon + 1 > 0 \\ &2\epsilon + \delta + 1 > 0 \\ (4.17) \qquad &2\epsilon - \delta > 0 \\ &2\epsilon - 3\delta > 0 \\ &\epsilon < 0 . \end{aligned}$$

The system of inequalities (4.17) has solutions; specifically $\epsilon = -\frac{1}{32}$ and $\delta = -\frac{1}{8}$ is one such solution. With this choice of ϵ and δ we see that $\log \hat{L}_\lambda(\theta_\lambda) \to +\infty$ as $\lambda \to 0$. It follows from (4.15) by choosing $v = \theta_\lambda$ that $\|v_\lambda\|_{L^2} \to +\infty$ as $\lambda \to 0$. This proves the proposition.

Proposition 4.3. Problem (4.8) has a unique solution.

Proof. By Proposition 4.2 there exists $\lambda > 0$ such that if v_λ is the unique solution of problem (4.9), then $\|v_\lambda\|_{L^2} > 1$. Now, if $B = \{v \in H^2(-\infty,\infty): \int_{-\infty}^{\infty} v(t)^2 dt \leq 1\}$ and $v(x_i) \geq 0$, $i = 1,\ldots,N\}$, then B is closed and convex. The proof of the proposition is now exactly the same as the proof of Proposition 3.3.

Proposition 4.4. The pseudo maximum penalized likelihood estimator of Good and Gaskins is well defined.

Proof. Since problems (4.4) and (4.8) have the same solutions the proposition follows from Proposition 4.3.

By the change of unknown function $v \to \sqrt{v}$ we see that problem (4.3) is equivalent to the following constrained optimization problem:

$$(4.18) \qquad \text{maximize} \quad \hat{L}(v) = \prod_{i=1}^{N} v(x_i)\exp(-\tfrac{1}{2}\Phi(v)); \quad \text{subject to}$$

$$v \in H^2(-\infty,\infty), \int_{-\infty}^{\infty} v(t)^2 dt = 1 \text{ and } v(t) \geq 0 \;\; \forall\, t \in (-\infty,\infty)$$

where $\Phi(v)$ is given by (4.1).

In turn for $\lambda > 0$ problem (4.18) is equivalent to

(4.19) maximize $\hat{L}_\lambda(v)$; subject to

$$v \in H^2(-\infty,\infty) , \int_{-\infty}^{\infty} v(t)^2 dt = 1 \text{ and } v(t) \geq 0 \;\forall\, t \in (-\infty,\infty)$$

where $\hat{L}_\lambda$ is defined in problem (4.8).

As in the previous two cases we also consider the constrained optimization problem:

(4.20) maximize $\hat{L}_\lambda(v)$; subject to

$$v \in H^2(-\infty,\infty) \text{ and } v(t) \geq 0 \;\; \forall\, t \in (-\infty,\infty)$$

where $\hat{L}_\lambda(v)$ is defined in problem (4.8).

Proposition 4.5. Problem (4.20) has a unique solution. Moreover if v_λ^+ denotes this solution, then

$$\|v_\lambda^+\|_{L^2} \to +\infty \text{ as } \lambda \to 0 .$$

Proof. The existence of v_λ^+ follows from Proposition 2.1 as in the proof of Proposition 4.2. Let us first show that

(4.21) $$\|v_\lambda^+\|_\lambda \leq \sqrt{N/2} .$$

From Lions [4, p.9] we see that

(4.22) $$\hat{L}_\lambda'(v_\lambda^+)(\eta - v_\lambda^+) \leq 0$$

for all nonnegative η in $H^2(-\infty,\infty)$. We have

$$\hat{L}'_\lambda(v)(\eta) = \sum_{i=1}^{N} \frac{\eta(x_i)}{v(x_i)} - 2 < v,\eta >_\lambda ;$$

hence

$$\hat{L}'_\lambda(v^+_\lambda)(v^+_\lambda) = N - 2\|v_\lambda\|^2_\lambda . \tag{4.23}$$

Now choosing $\eta = 0$ in (4.22) and using (4.23) we arrive at (4.21). The functions θ_λ defined in the proof of Proposition 4.2 satisfy the constraints of this problem; hence

$$\log \hat{L}_\lambda(\theta_\lambda) \leq \log \hat{L}_\lambda(v^+_\lambda) .$$

From (4.14) and (4.21) we have

$$\log \hat{L}_\lambda(\theta_\lambda) \leq \frac{3N}{4} \log(\|v^+_\lambda\|_{L^2}) + \frac{N}{8} \log(8N/\beta) + \frac{N}{2} . \tag{4.24}$$

The proof now follows from (4.24) since $\log L_\lambda(\theta_\lambda) \to +\infty$ as $\lambda \to 0$.

Proposition 4.6. The second maximum penalized estimator of Good and Gaskins is well defined.

Proof. Using Proposition 4.5 the argument used to prove Proposition 4.3 shows that problem (4.19) has a unique solution which is also the unique solution of problem (4.18). This proves the proposition.

The authors would like to thank Professors B.F. Jones, P.E. Pfeiffer and W.A. Veech for helpful discussions. They would also like to thank the referee for helpful suggestions.

References

[1] L. Boneva, D. Kendall, I. Stefanov, Spline transformations: Three new diagnostic aids for the statistical data-analyst, Journal Royal Stat. Soc. B, 33(1971), 1-77.

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

[3] G.S. Kimeldorf and G. Wahba, A correspondence between Baysian estimation on stochastic processes and smoothing by splines, Annals of Math. Stat., 41(1970), 495-502.

[4] J.L. Lions, Contrôle optimal de systèms gouvernés par des équations aux dérivées partielles, Dunod, Gauthier-Villars, Paris, 1968.

[5] J.L. Lions and E. Magenes, Problèms aux limites non homogènes et applications Vol. 1, Dunod, Paris, 1968.

[6] E. Parzen, On estimation of a probability density function and mode, Annals of Math. Stat., 33(1962), 1065-1076.

[7] M. Rosenblatt, Remarks on some nonparametric estimates of a density function, Annals of Math. Stat., 27(1956), 832-837.

[8] I.J. Schoenberg, Splines and histograms, with an appendix by Carl de Boor, Mathematics Research Center Report 1273, University of Wisconsin, Madison, October 1972.

[9] R.A. Tapia, The differentiation and integration of nonlinear operators, in Nonlinear Functional Analysis and Applications, Ed. Louis B. Rall, Academic, New York, 1971.

[10] G. Wahba, A polynomial algorithm for density estimation. Annals of Math. Stat., 42(1971), 1870-1886.

APPENDIX

Numerical implementation

We wish to implement numerically the maximum penalized likelihood estimator corresponding to the reproducing kernel space $\overline{H}_0^1(a,b)$ discussed in Section 2. Toward this end we introduce a partition of the interval (a,b):

$$a = t_0 < t_1 < \ldots < \ldots < t_m < t_{m+1} = b \ ,$$

where the mesh spacing is equal to $h = (b-a)/m$ for some predetermined positive integer m. We let y_i denote the value of the discrete solution at the mesh point t_i. Clearly since we are approximating elements in $\overline{H}_0^1(a,b)$ we will require that $y_0 = y_{m+1} = 0$. We choose as a discrete approximation to the derivative at the mesh point t_i the first forward difference $(y_{i+1} - y_i)/h$. As the discrete form of the integral constraint we choose the trapezoidal rule, which in this case leads to $\sum_{i=1}^{m} y_i = h^{-1}$. Given the samples $x_1,\ldots,x_N \in [a,b]$ let α_i denote the number of samples in the interval $[t_i - \frac{h}{2}, t_i + \frac{h}{2})$ for $i = 2,\ldots,m-1$, let α_1 denote the number of samples in $[a, t_1 + \frac{h}{2})$ and finally let α_m denote the number of samples in the interval $[t_m - \frac{h}{2}, b]$. Our discrete maximum penalized likelihood estimate is obtained as the solution of the following constrained finite dimensional optimization problem:

$$\text{maximize} \quad J(y_1,\ldots,y_m) = \prod_{i=1}^{m} y_i^{\alpha_i} \exp\left(-h^{-2} \sum_{i=0}^{m} (y_{i+1} - y_i)^2\right) ;$$

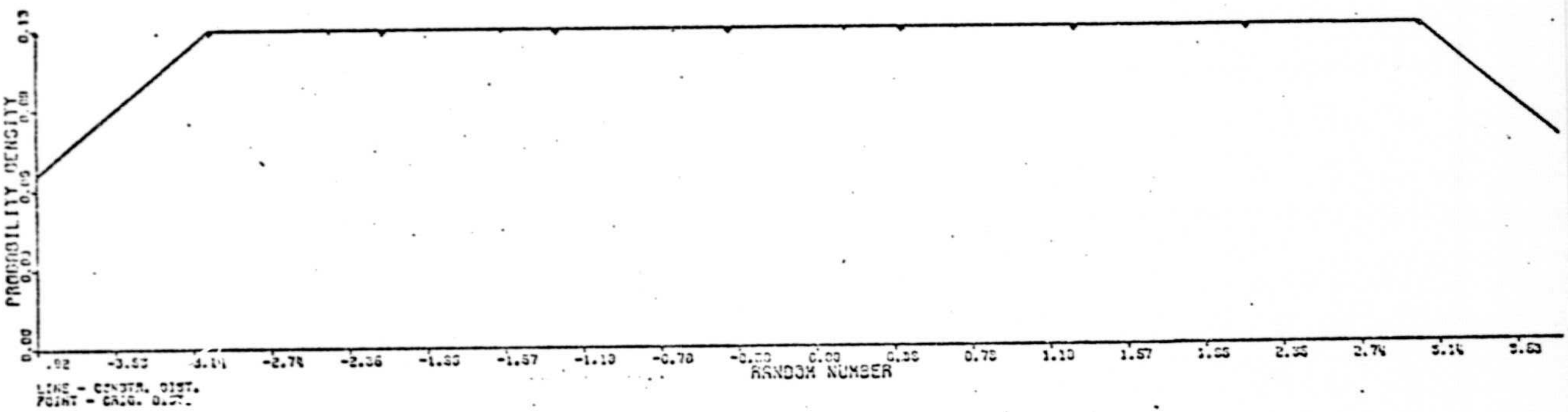
subject to

$$\sum_{i=1}^{m} y_i = h^{-1} \quad \text{and} \quad y_i \geq 0 \ , \ i = 1,\ldots,m \ .$$

The fact that this optimization problem has a unique solution follows as in the proof of Proposition 2.1. Figure 1 shows our numerical results when this procedure was applied to 100 samples obtained from the uniform distribution and Figure 2 shows the result obtained when this procedure was applied to 100 samples obtained from the Gaussian distribution. Since the curves are only described at the mesh points we have interpolated linearly between every two mesh points.

FIGURE 1

COMPARISON OF ORIGINAL AND CONSTRUCTED DISTRIBUTIONS -



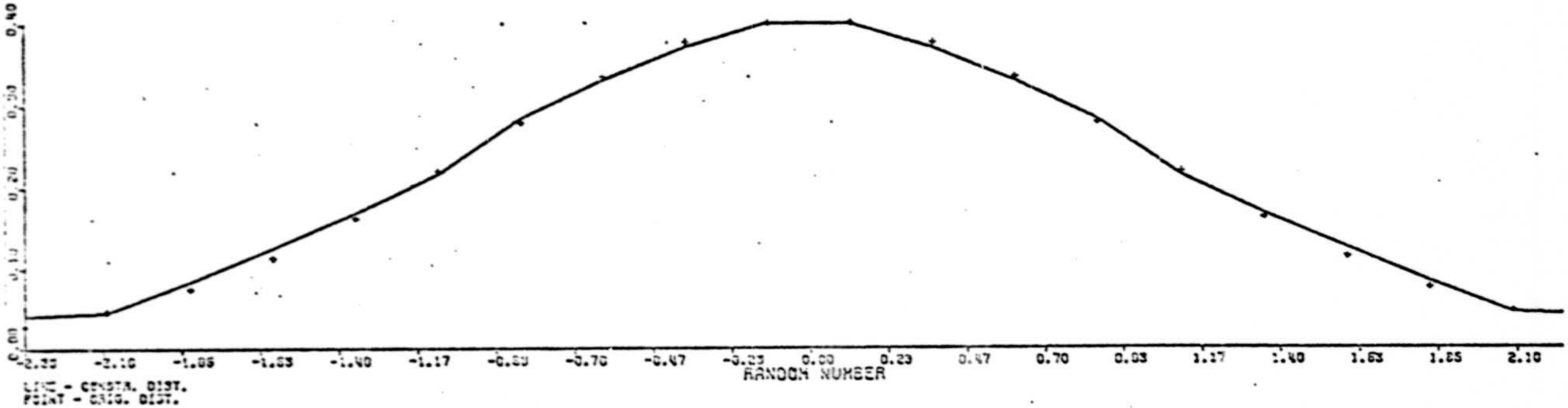
100 RANDOM NUMBERS

10 MESH POINTS

UNIFORM DISTRIBUTION

FIGURE 2

COMPARISON OF ORIGINAL AND CONSTRUCTED DISTRIBUTIONS -



100 RANDOM NUMBERS

20 MESH POINTS

GAUSSIAN DISTRIBUTION