The Annals of Statistics
1980, Vol. 8, No. 5, 1023–1035

ISOTONIC, CONVEX AND RELATED SPLINES[1]

BY IAN W. WRIGHT AND EDWARD J. WEGMAN

University of North Carolina

In this paper, we consider the estimation of isotonic, convex or related functions by means of splines. It is shown that certain classes of isotone or convex functions can be represented as a positive cone embedded in a Hilbert space. Using this representation, we give an existence and characterization theorem for isotonic or convex splines. Two special cases are examined showing the existence of a globally monotone cubic smoothing spline and a globally convex quintic smoothing spline. Finally, we examine a regression problem and show that the isotonic-type of spline provides a strongly consistent solution. We also point out several other statistical applications.

1. Introduction. A good deal of literature concerning statistical inference under order restrictions has appeared in the last 15 years. Barlow et al (1972) presents a rather convenient source book for this material. The main body of this work can be conveniently divided into two areas: inferences concerning functions on a finite or countable set and inferences concerning functions on the real line or intervals in the real line. It is with this latter group of inferences that we are concerned.

Examples of functions of interest would include regression functions, probability distribution and density functions, failure rate functions and functions related to spectral analysis such as spectral densities, gain and transfer functions and so on. It is often known that these functions are *isotone* (i.e., order-preserving) and, hence, should be estimated with a function that preserves the order. See Robertson (1967), Wegman (1970a, b), Barlow, Marshall and Proschan (1963) and Marshall and Proschan (1965) for examples of such estimation problems. A characteristic feature of these isotonic estimators is that they are step functions. In most of these situations, smoothness is frequently just as desirable as isotonicity so that while the step function may be isotone, its lack of continuity prevents it from being widely accepted as a satisfactory estimator.

More recently, there has appeared a good deal of literature concerning the use of splines in statistical estimation problems. See Wright and Wegman (1980) for a review of some of these efforts. While a spline fit satisfies the requisite smoothness properties, it may not be isotone as desired. In this present paper, we present a combined approach. For technical convenience, the class of partial orders is

Received February 1978; revised April 1979.

[1]This work was supported by the Air Force Office of Scientific Research under Grant No. AFOSR-75-2840. Dr. Wright's work was done while on leave from Department of Mathematics, Papua New Guinea University of Technology, Lae, Papua New Guina. Dr. Wegman is now Director of Statistics and Probability Program, Office of Naval Research.

AMS 1970 *subject classifications.* 62G05, 62M15, 65D05, 65D10, 06A10.

Key words and phrases. Isotone, isotonic, convex, partial order, spline, isotonic spline, smoothing spline, interpolating spline, regression, strong consistency.

restricted to those compatible with the additive group of real functions. Our framework is general enough to encompass estimation problems for monotone, or convex or positive functions and other related families of functions.

2. Partial orders and isotonic splines. In this paper we shall be dealing only with abelian groups of real functions which can be added pointwise.

DEFINITION 2.1. The relation, $\gg$, on the group, G, of functions is required to satisfy the following conditions:

(i) $f \gg f$
(ii) $f \gg g$, $g \gg h$ implies $f \gg h$
(iii) $f \gg g$ implies $f + h \gg g + h$ for all $h \in G$.
(iv) $f \gg 0$, $g \gg 0$ implies $f + g \gg 0$
(v) $f \gg g$ and $g \gg f$ implies $f = g$.

A group satisfying (i) to (v) is called a *partially ordered group*. If condition (v) is dropped, G is called a *preordered group*. The set, $P = \{g \in G: g \gg 0\}$ is called the *positive cone* of the order, $\gg$. The set, P, defines and is defined by the partial order, $\gg$.

The partial order, $\gg$, defined here is essentially distinct from the partial order which usually occurs in papers on isotonic methods—(cf. Barlow et al (1972, Chapter 7)). In the standard treatment, a partial order, say, $\succ$, is induced on the real line. A function f is said to be *isotone* if it preserves the order. That is, f is isotone iff $f(x_1) \geqslant f(x_2)$ whenever $x_1 \succ x_2$. Two functions may both be isotone and yet not comparable. What is true, however, is that the set of isotone functions forms a positive cone as just defined. We have in mind an essentially different use of the partial order, $\gg$.

We let L_2 represent the set of functions on [0, 1] which are Lebesgue measurable and square integrable with the usual Banach space norm. We let W_m, $m \geqslant 1$, be the set of functions on [0, 1] for which $f^{(j)}, j = 0, 1, \cdots, m-1$ are absolutely continuous and $f^{(m)}$ is in L_2. This is a Hilbert space with inner product $\langle f, g \rangle = \sum_{j=0}^{m} \int_0^1 f^{(j)}(t) g^{(j)}(t)\, dt$. Let C^k, $k = 1, 2, \cdots, \infty$, be the set of all functions on [0, 1] which are k-times continuously differentiable, and finally we let D be the ordinary differentiation operator, i.e., $Df(t) = df/dt$.

In order to define a suitable positive cone, we let F be a continuous linear map of W_m into W_1 which commutes with the differentiation operator, i.e., $D(Ff) = F(Df)$ for all $f \in W_m \subset W_{m-1}$. We define a partial order, $\gg$, on W_m by $f \gg 0$ if and only if $(Ff)(t) \geqslant 0$ for every $t \in [0, 1]$.

There are several operators, F, of particular interest.

EXAMPLE 2.1. If F is the identity map, the set, $P = \{g \in W_m: g \gg 0\}$ is just the set of positive functions in W_m.

EXAMPLE 2.2. If $F = D$, P is just the set of monotone nondecreasing functions in W_m. Similarly, if $F = -D$, P is the set of nonincreasing functions.

EXAMPLE 2.3. If $F = D^2$, P is the set of convex functions in W_m while $F = -D^2$ yields the set of concave functions.

EXAMPLE 2.4. If F is defined by

$$\begin{aligned} Ff(t) &= Df(t) & 0 \leqslant t < M \\ &= -Df(t) & M < t \leqslant 1, \end{aligned}$$

then P is the set of unimodal functions with mode M.

EXAMPLE 2.5. If F is defined by

$$\begin{aligned} Ff(t) &= -D^2 f(t) & t_1 < t < t_2 \\ &= D^2 f(t) & 0 \leqslant t < t_1 \text{ or } t_2 \leqslant t < 1, \end{aligned}$$

then P is the set of functions which are concave on $[t_1, t_2]$ and convex elsewhere. The applications of these sample characterizations should be abundantly clear.

In principle we desire to find an estimating function which belongs to P. Clearly, in general, there will be many possibilities. In order to ensure that our estimate is also as smooth as reasonably possible, however, we select a function satisfying the following criterion:

$$\text{Minimize } \int_0^1 (f^{(m)}(t))^2\, dt \text{ subject to}$$

$$\text{(2.1)} \quad \begin{aligned} &\text{(a) } f \in W_m,\ Ff(t) \geqslant 0 \quad \text{for every } t \in [0, 1] \\ &\text{(b) } \alpha_i \leqslant f(t_i) \leqslant \beta_i, \qquad i = 1, 2, \cdots, n. \end{aligned}$$

Constraints (2.1a) are, of course, simply that $f \in P$. The second constraint (2.1b) deserves a bit more comment.

Suppose our data points are $\{(t_i, y_i);\ i = 1, \cdots, n\}$. A function f in W_m which coincides exactly with a polynomial of degree $2m - 1$ on each interval $[t_i, t_{i+1}]$ and which minimizes $\int_0^1 (f^{(m)}(t))^2\, dt$ is called a polynomial *spline* of degree $2m - 1$ and the t_i, $i = 1, 2, \cdots, n$ are the *knots* of this spline.

A well-studied problem in approximation theory is to find a solution of the following optimization problem:

$$\text{Minimize } \int_0^1 (f^{(m)}(t))^2\, dt \text{ subject to}$$

$$\text{(2.2)} \quad \begin{aligned} &\text{(a) } f \in W_m \\ &\text{(b) } f(t_i) = y_i, \qquad i = 1, 2, \cdots, n. \end{aligned}$$

The solution f, called an *interpolating spline* is a spline as just defined.

In contrast the problem:

$$\text{Minimize } \Sigma_{i=1}^n (y_i - f(t_i))^2 + \lambda \int_0^1 (f^{(m)}(t))^2\, dt$$

$$\text{(2.3)} \quad \text{with } \lambda > 0 \text{ fixed and subject to (a) } f \in W_m$$

also has a spline function solution called a *smoothing spline*. See Kimeldorf and Wahba (1970) or Cogburn and Davis (1974) for details.

A smoothing spline intermediate between (2.2) and (2.3) solves the following problem:

$$\text{Minimize } \int_0^1 (f^{(m)}(t))^2\, dt \text{ subject to}$$

$$\alpha_i \leqslant f(t_i) \leqslant \beta_i, \qquad i = 1, 2, \cdots, n. \tag{2.4}$$

See Attéia (1968) for more details.

We note here several closely related papers. The first by Copley and Schumaker (1978) (see also Daniel and Schumaker (1974) and Mangasarian and Schumaker (1969)) considers the problem, cast in our notation

$$\text{Minimize } \int_0^1 (Lf(t))^2\, dt \text{ subject to}$$

$$\begin{aligned} &\text{(a) } f \in W_m \\ &\text{(b) } \alpha_i \leqslant Ff(t_i) \leqslant \beta_i, \qquad i = 1, 2, \cdots, n, \end{aligned} \tag{2.5}$$

where L is a linear differential operator of degree m and F is now a bounded linear functional on W_m (rather than the continuous linear map we defined). Copley and Schumaker give existence and characterization results. Our approach, while having some similarity to theirs has several distinctive features including our characterization of sets of isotonic functions by F and, of course, that we are solving a different problem.

The second closely related paper is that of Wahba (1973) in which she considers not a finite number of constraints of the form of (2.4), but a continuous constraint of the form $\alpha(t) \leqslant \langle \eta_t, f \rangle$, $t \in [0, 1]$ where the $\{\eta_t\}$ are elements in W_m. Again, Wahba's problem is somewhat different in character from the one we solve here.

In closing this section, we comment that Problem 2.1 is the isotonic version of Problem 2.4. Our results show that a solution of (2.1) exists and is a polynomial spline of degree $2m - 1$. In general, we would also like to solve an isotonic version of (2.3). Problem 2.1 is solved by present methods because the objective function is a norm of a Hilbert space. While Anselone and Laurent (1968) have shown that the objective function in (2.3) can be cast into the form of a Hilbert space norm, the carryover of our present methods is not obvious and we do not attempt to solve an isotonic version of (2.3) in the present paper.

3. A restricted isotonic spline. We consider first a restricted problem. Let F be a continuous linear map of W_m into W_1 which commutes with differentiation. Denote the partial order defined by F on W_m by $\gg$, and assume there exists a function $g \in W_m$ satisfying $\alpha_i < g(t_i) < \beta_i$ for $i = 1, 2, \cdots, n$ with $(Fg)(t) \geqslant \varepsilon > 0$ for every $t \in [0, 1]$. The continuity of DF implies there exists a $d > 0$ such that

$$\|DFf\|_{L_2} < d\|D^m f\|_{L_2} \text{ for every } f \in W_m.$$

Take $N > d^2\|D^m g\|^2 L_2/\varepsilon^2$. In what follows, we shall assume $n \geqslant m$.

THEOREM 3.1. *Under the conditions just given, the problem:*

$$\text{minimize } \int_0^1 (f^{(m)}(t))^2\,dt \text{ subject to}$$

$$(3.1)\qquad \begin{aligned} &(a)\, f \in W_m \\ &(b)\, f \gg 0 \\ &(c)\, \alpha_i \leqslant f(t_i) \leqslant \beta_i, && i = 1, 2, \cdots, n \\ &(d)\, Ff(j/N) \geqslant \varepsilon, && j = 0, 1, 2, \cdots, N, \end{aligned}$$

has a solution which is a polynomial spline of degree $2m - 1$ with possible knots at t_i, $i = 1, 2, \cdots, n$ and at j/N, $j = 0, 1, \cdots, N$.

In order to establish Theorem 3.1, we need to establish some preliminary results. The straightforward proof of Lemma 3.2 is omitted.

LEMMA 3.2. *$\|D^m f\|_{L_2} \leqslant \|D^m g\|_{L_2}$ and $Ff(j/N) \geqslant \varepsilon$, $j = 0, 1, \cdots, N$ together imply that $f \gg 0$.*

We next define

$$\Lambda_1 = \{u_1 \in W_m, i = 1, 2, \cdots, n: \langle u_i, f\rangle = -f(t_i) \quad \text{for every } f \in W_m\}$$

$$\Lambda_2 = \{v_i \in W_m, i = 1, 2, \cdots, n: \langle v_i, f\rangle = f(t_i) \quad \text{for every } f \in W_m\}$$

and finally

$$\Lambda_3 = \{w_i \in W_m, i = 0, 1, \cdots, N: \langle w_i, f\rangle = -Ff(i/N) \quad \text{for every } f \in W_m\}.$$

Take $L = \Lambda_1 \cup \Lambda_2 \cup \Lambda_3$ and define $p: L \to R$ by

$$\begin{aligned} p(u_i) &= -\alpha_i && u_i \in \Lambda_1 \\ p(v_i) &= \beta_i && v_i \in \Lambda_2 \\ p(w_i) &= -\varepsilon && w_i \in \Lambda_3. \end{aligned}$$

We define a convex subset C of W_m by

$$C = \{f \in W_m: \text{for every } l \in L, \langle l, f\rangle \leqslant p(l)\}.$$

That is to say, C is just the set of functions in W_m which satisfy the constraints. Let T be a continuous linear map of W_m onto L_2 and suppose Ker T is of finite dimension. (In the case of the application we have in mind, $T = D^m$ so that both conditions are trivially satisfied.) We thus have in mind finding $s \in C$ satisfying $\|Ts\| = \min_{f \in C} \|Tf\|$. Such an element s will be called the generalized spline function relative to T in the set C.

Next let $\tilde{C} = \{f \in W_m: \langle l, f\rangle \leqslant 0$ for every $l \in L\}$. We posit four structure axioms and, in turn state two theorems.

H1. $\tilde{C} \cap (\text{Ker}\, T) = \{0\}$

H2. The subset $L \subset W_m$ is weakly compact and the map p is a continuous map of L into R.

H3. $C \cap (\text{Ker } T)$ is empty.

H4. $I = \{f \in W_m$: for every l, $\langle l, f\rangle < p(l)\}$ is not empty.

Under H1, Attéia (1968) proves Theorem 3.3.

THEOREM 3.3. *Under* H1, *there is at least one spline function relative to T in the convex set C.*

We refer the reader to Attéia for proof. If, in addition, we consider the remaining axioms, we have the following result due to Laurent (1969).

THEOREM 3.4. *Under* H1 − 4, *the element* $s \in C$ *is a spline function* (*relative to T in C*) *if and only if*

$$-T^* Ts \in \overline{CC}(F_s)$$

where

$$F_s = \{l \in L : \langle l, s\rangle = p(l)\}$$

and $\overline{CC}(F_s)$ *is the smallest closed convex cone with vertex* 0 *containing* F_s *and* T^* *is the adjoint of T.*

We are now in a position to apply Lemma 3.2 and Theorems 3.3 and 3.4 to the present case.

PROOF OF THEOREM 3.1. At the optimum s, $\|D^m s\|_{L_2} \leqslant \|D^m g\|_{L_2}$ and so by Lemma 3.2, the spline function must satisfy all the constraints of L, and hence $s \gg 0$. We need only to verify axioms H1 − 4 to guarantee the existence of s via Attéia's existence theorem and its characterization via Laurent's characterization theorem.

In our case , $T = D^m$, so that Ker T is the set of polynomials of degree $m - 1$. Every function in $\tilde{C}$ is zero at all points t_i, $i = 1, 2, \cdots, n$. Hence as soon as $n > m - 1$, $\tilde{C} \cap \text{Ker } T = \{0\}$. Hence H1 holds for $n \geqslant m$. For the L described, L is finite hence compact in any topology. This implies p is continuous so that H2 is satisfied.

Next we consider H3. C is the set of functions which satisfy the constraints, while Ker T is the set of polynomials of degree $m - 1$. Hence if $C \cap (\text{Ker } T)$ is not empty, then an optimal solution, s, independent of the theorems of Attéia and Laurent, will be a polynomial of degree $m - 1$. A polynomial of degree $m - 1$ is trivially a polynomial spline of degree $2m - 1$. Hence, we may assume $C \cap (\text{Ker } T)$ is empty.

Finally, $I = \{f \in W_m$: $\langle l, f\rangle < p(l)$ for every $l\}$ contains the function g of the conditions to Theorem 3.1 so that I is not empty. Thus provided $C \cap \text{Ker } T$ is empty, the characterizing Theorem 3.4 applies. Thus, $-T^*Ts$ belongs to the closed convex cone generated by all $l \in L$ corresponding to active constraints. In other words,

$$T^* Ts = -\Sigma_{l \in L} d_l l$$

with $d_l \geqslant 0$ and $d_l = 0$ when l is not an active constraint. Recall that all $l \in L$ are point evaluation functionals. Thus T^*Ts is zero at all points except those corresponding to active constraints (which become knots). Between knot points, $T^*Ts = 0$, i.e., $D^{2m}s = 0$. Hence, s is a polynomial of degree $2m - 1$ between knot points. ▯

4. The general isotonic spline. We have already mentioned the general isotonic spline with possibly an infinite number of knots able to occur anywhere. The following result makes the details precise.

THEOREM 4.1. *Let $\gg$ be the partial order defined by F: $W_m \to W_1$. If there exists $g \in W_m$ satisfying*

$$Fg(t) > 0 \qquad t \in [0, 1]$$

and

$$\alpha_i \leq g(t_i) \leq \beta_i \qquad i = 1, 2, \cdots, n,$$

then the problem:

Minimize $\int_0^1 (f^{(m)}(t))^2 \, dt$ *subject to*

$$\text{(a) } f \gg 0; \text{ (b) } \alpha_i \leq f(t_i) \leq \beta_i \qquad i = 1, 2, \cdots, n \tag{4.1}$$

has a solution which is a polynomial spline of degree $2m - 1$. Knots are located (potentially) at data points, t_i, and, possibly at a countable number of points elsewhere.

PROOF. The proof follows the lines of argument for Theorem 3.1. We take $T = D^m$ and define

$$\Lambda_3 = \{ w_t \in W_m, \, t \in [0, 1] : \langle w_t, f \rangle = Ff(t) \text{ for every } f \in W_m \}.$$

Let $p(w) = 0$ for all $w \in \Lambda_3$. The key requirement that Λ_3 be weakly compact holds in this case and p is continuous. The remainder follows as before. ▯

A remark on the potentially infinite number of additional knot points in order. Intuitively, the additional knot points are needed to force the spline to satisfy the inequality $Ff(t) > 0$ or more colorfully to prop the spline up where it sags. Theorem 4.1 characterizes the general isotonic spline, but does not specify where the knots are to be placed. We make further comment on computations in Section 7. Intuitively, however, we would not expect to need the full complement of potential knot points. We would only need a finite number where the spline would tend to go wrong, the others being inactive. Hence in actually computing an isotonic spline, it would only be in the most pathological cases that more than this finite number are needed.

5. Two examples. Passow (1974) and Passow and Roulier (1977) presented some results concerning monotone and convex splines. If (t_i, y_i), $i = 1, 2, \cdots, n$ represents the data points, they were interested in obtaining piecewise monotone or convex interpolating splines. Theorems 3.1 and 4.1 may be used to extend these results. We assume $t_1 < t_2 < \cdots < t_n$.

THEOREM 5.1. *Let* $y_1 < y_2 < \cdots < y_n$. (a) *With arbitrarily small error bounds, there is a globally monotone cubic smoothing spline satisfying* (3.1) *for* $m = 2$ *with possible knot points at* t_i, $i = 1, 2, \cdots, n$ *and at* $j/N, j = 0, 1, \cdots, N$ *where* N *is as defined in Section* 3.

(b) *With arbitrarily small error bounds, there is a globally monotone cubic smoothing spline satisfying* (4.1) *for* $m = 2$ *with possible knot points at* t_i, $i = 1, 2, \cdots, n$ *and at a countable number of points elsewhere.*

PROOF. We need establish that there exists a suitable g satisfying the assumptions of Theorem 3.1. Let

$$\begin{aligned} \phi(t) &= \exp(1/(t^2 - 1)) \qquad &|t| < 1 \\ &= 0 \qquad &|t| \geq 1. \end{aligned}$$

$\phi(t)$ is known to be infinitely differentiable on the real line and hence $\phi \in W_m$, $m \geq 1$. Let $\phi_1(t) = \int_{-\infty}^{t}\phi(u)\,du$. ϕ_1 is in C^∞ and it is clear that

$$\begin{aligned} \phi_1(t) &= 0 \qquad &t \leq -1 \\ &= k - \int_{-1}^{1}\phi(u)\,du \qquad &t \geq 1 \end{aligned}$$

and, of course, $D\phi_1(t) \geq 0$ for all t. When the conditions of this proposition hold, adding together suitably scaled translates of ϕ_1 will give a function $f_1 \in C^\infty$ with $Df_1(t) \geq 0$ for every t which satisfies $f_1(t_i) = y_i$, $i = 1, 2, \cdots, n$. If ε is the error bound on y_i, the function $g_1(t) = f_1(t) + \varepsilon t$, $t \in [0, 1]$ satisfies the hypotheses of Theorem 3.1 and (4.1) with $F = D$ and $m = 2$. □

THEOREM 5.2. *Let*

$$\frac{y_{i+1} - y_i}{t_{i+1} - t_i} < \frac{y_{i+2} - y_{i+1}}{t_{i+2} - t_{i+1}} \textit{ for } i = 1, 2, \cdots, n-2.$$

(a) *With arbitrarily small error bounds, there is a globally convex quintic smoothing spline satisfying* (3.1) *for* $m = 3$ *with possible knot points at* t_i, $i = 1, 2, \cdots, n$ *and at* $j/N, j = 0, 1, \cdots, N$ *where* N *is as defined in Section* 3.

(b) *With arbitrarily small error bounds, there is a globally convex quintic smoothing spline satisfying* (4.1) *for* $m = 3$ *with possible knot points at* t_i, $i = 1, 2, \cdots, n$ *and at a countable number of points elsewhere.*

PROOF. From the function $\phi(t)$ of Theorem 5.1, we define

$$\phi_2(t) = \int_{-\infty}^{t}\left[\int_{-\infty}^{x}\phi(u)\,du\right]dx.$$

When the conditions of the theorem hold, adding together suitably scaled translates of $\phi_2(t)$ will give a function $f_2 \in C^\infty$ with $D^2 f_2(t) \geq 0$ for all t which satisfies $f_2(t_i) = y_i$, $i = 1, 2, \cdots, n$. If ε is the maximum error bound on the y_i, the function $g_2(t) = f_2(t) + (\varepsilon t^2/2)$, $t \in [0, 1]$ meets the assumptions of Theorems 3.1 and 4.1 with $F = D^2$ and $m = 3$. □

6. Statistical interpretation. In the foregoing sections, we describe the isotonic spline purely in terms of its mathematical character. We now turn to its statistical nature. We consider a statistical model of the form

$$Y(t) = f(t) + e(t), \qquad t \in [0, 1], \tag{6.1}$$

where $Y(t)$ is an observation at location t, $f(t)$ is a function to be estimated and $e(t)$ is an error. Of course, [0, 1] can be replaced with any finite interval. We have in mind to discuss several possible error structures.

In our first type of error structure, we consider a data set $(t_i, Y(t_i))$, $i = 1, 2, \cdots$ such that the t_i are dense on the interval [0, 1] and such that $e(t_i)$ form an i.i.d. sequence of independent errors. We assume, first of all, that the support of the common density of $e(t)$ is a finite interval, say $[-e_1, e_2]$ containing 0. There is no a priori need to consider the density to be symmetric or with mean 0. We note that $Y(t_i) + e_1 = \beta_i$ and $Y(t_i) - e_2 = \alpha_i$ forms a 100% confidence interval for $f(t_i)$. We have the following lemma.

LEMMA 6.1. *Suppose the model* (6.1) *holds with* $e(t_i)$, $i = 1, 2, \cdots$ *a sequence of* i.i.d. *random variables with support as described above. Suppose further that* $\{t_i: i = 1, 2, \cdots\}$ *is dense in* [0, 1]. *Then for* $\eta > 0$ *and any interval* (t, t'), *there is a* t_i *and* t_j *in* (t, t') *such that*

$$\beta_i - f(t_i) < \eta \text{ almost surely}$$

and

$$f(t_j) - \alpha_j < \eta \text{ almost surely.}$$

PROOF. We consider the subsequence of t_i's falling in the interval (t, t'). We relabel this sequence t_i and we note that the subsequence, $e(t_i)$, of random errors corresponding to the $t_i \in (t, t')$ again forms an i.i.d. sequence. Consider the interval $(-e_1, -e_1 + \eta)$. Since the support of $e(t)$ is $(-e_1, e_2)$, the set $(-e_1, -e_1 + \eta)$ has positive probability. Hence with probability one, for sufficiently large i, $e(t_i) \in (-e_1, -e_1 + \eta)$. Thus

$$\beta_i = Y(t_i) + e_1 = f(t_i) + e(t_i) + e_1 < f(t_i) - e_1 + \eta + e_1 = f(t_i) + \eta.$$

Similarly for α_j. □

We may now state a consistency theorem.

THEOREM 6.2. *Let* $F = D$ *be the operator, so that we have a monotone nondecreasing spline. Suppose further that* α_i, β_i *are as defined above and that* $\{t_i\}$ *are dense in* [0, 1]. *Finally let* $f \in W_m$ *be nondecreasing, then* s_n, *the isotonic spline based on* $(t_1, Y(t_1)), \cdots, (t_n, Y(t_n))$, *exists and*

$$s_n \to f \text{ almost uniformly with probability one.}$$

PROOF. Since $\alpha_i < f(t_i) < \beta_i$ w.p. 1, f is the g needed to guarantee the existence of s_n in Theorems 3.1 and 4.1.

Let $\varepsilon > 0$ be given. Since f is continuous on [0, 1], f is uniformly continuous. Corresponding to $\varepsilon/3$ there is a $\delta > 0$ such that $|f(x) - f(y)| < \varepsilon/3$ whenever

$|x - y| \leqslant \delta$. We divide [0, 1] into consecutive intervals, I_j, of length δ except for possibly the last which may be shorter than δ.

Consider the interval I_j and let its endpoints be $x_j < x_{j+1}$. Since both f and s_n are nondecreasing on $[x_j, x_{j+1}]$,

$$|f(t) - s_n(t)| \leqslant \max\{f(x_{j+1}) - s_n(x_j), s_n(x_{j+1}) - f(x_j)\}, \qquad t \in I_j.$$

Consider first $f(x_{j+1}) - s_n(x_j)$.

$$f(x_{j+1}) - s_n(x_j) \leqslant f(x_{j+1}) - f(x_j) + f(x_j) - s_n(x_j) \leqslant \frac{\varepsilon}{3} + f(x_j) - s_n(x_j),$$

since $|x_{j+1} - x_j| = \delta$. Now $s_n(x_j) \geqslant s_n(t_i) \geqslant \alpha_i$ for every $t_i \leqslant x_j$. For the interval, $(x_{j-1}, x_j]$, we may choose an i such that $t_i \in (x_{j-1}, x_j]$ and $f(t_i) - \alpha_i < \varepsilon/3$. This happens with probability one by Lemma 6.1. Thus, since $s_n(x_j) \geqslant s_n(t_i) \geqslant \alpha_i$

$$f(x_j) - s_n(x_j) \leqslant f(x_j) - s_n(t_i) \leqslant f(x_j) - \alpha_i \leqslant f(x_j) - f(t_i) + f(t_i) - \alpha_i.$$

But $f(x_j) - f(t_i) \leqslant \varepsilon/3$ since $|x_j - t_i| < |x_j - x_{j-1}| = \delta$ and $f(t_i) - \alpha_i < \varepsilon/3$ with probability one by above. Hence $f(x_{j+1}) - s_n(x_j) \leqslant 3\varepsilon/3 = \varepsilon$ with probability one. Similarly $s_n(x_{j+1}) - f(x_j) \leqslant \varepsilon$ with probability one so that for n sufficiently large

$$|f(t) - s_n(t)| \leqslant \varepsilon, \qquad t \in I_j, \text{ w.p. } 1.$$

Clearly this convergence holds for all I_j except possibly the first and last, hence, except on a set of measure less than 2δ. The almost uniform convergence is clear. □

This result may be extended by the following corollary.

COROLLARY 6.3. *Suppose the operator F in Theorem 6.2 is replaced by any of those in Examples 2.2, 2.3, 2.4 or 2.5, and suppose $Ff(t) \geqslant 0$. If the remaining conditions of Theorem 6.2 hold, then s_n exists and*

$$s_n \to f \quad \textit{almost uniformly with probability one}.$$

PROOF. If $F = -D$, the result follows in an obvious parallel to Theorem 6.2. If $F = D^2$, the result follows since f and s_n will be first nonincreasing and then nondecreasing. The results of Theorem 6.2 can be applied to each part individually. Similarly for the remaining cases. □

In the foregoing results, the convergence is uniform except possibly near the end points 0 and 1. We remark here that by suitably regularizing the behavior of s_n outside the interval (min t_i, max t_i) we can extend the uniformity of convergence to the entire interval, [0, 1]. This can be done in several obvious ways which we shall not detail here.

The regression model with $e(t)$ having bounded support may be viewed with suspicion by those used to conventional models with normal errors. However, much data these days are collected with digital instrumentation which almost implies bounded range on the errors. We also point out that at any given data point, $(t_i, Y(t_i))$, the error bounds α_i, β_i do not change as a function of n, i.e., we are not supposing increasingly accurate bounds.

Of course, the fact that the support is bounded allows us to give 100% error bounds which implies that $f(t_i)$ always falls in (α_i, β_i), which, in turn, guarantees the existence of s_n. If the support is unbounded, as with normal errors, finite 100% error bounds are impossible. A natural suggestion is to replace the 100% error bounds with $\alpha\%$ bounds for some $\alpha < 100$. The problem is that, with probability one, $f(t_i)$ will fall outside the error bounds for some i, and, hence, we are not even guaranteed the existence of a sequence of splines much less their consistency. Hence, the consistency in the style of Theorem 6.2 for unbounded support is a moot point.

Not all is hopeless, of course, in this case. For large sets of data, about $\alpha\%$ of the intervals are not going to contain $f(t)$ and hence may even be inconsistent with isotonicity. A practical procedure may be to discard a percentage of the bounds (say no more than $\alpha\%$) which prevent us from fitting an isotone spline, and use the remainder to fit the spline. Alternatively, we could borrow a leaf from Barlow, et al, (1972) and simply "pool adjacent violators". Hence extend the length of interval for a violating interval by "averaging" with an adjacent nonviolating interval.

We may also note that the situation where there are many data points is not really the optimal situation for use of the isotonic spline anyway. If there is much noisy data, a conventional smoothing spline is appropriate. And of course in a noise-free data situation, an ordinary interpolating spline is appropriate. It is in the context of a relatively sparse data set that the added knowledge of isotonicity will allow for the relatively largest improvement in efficiency.

7. Concluding remarks. Sections 2 through 6 give existence, characterization and statistical properties of isotonic splines. Questions of the computation of such splines can be approached through a quadratic programming approach. The work of prime interest here is the work of Kimeldorf and Wahba (1971); but also related are the papers of Ritter (1969), Anselone and Laurent (1968) and Wahba (1978). The thrust of Kimeldorf and Wahba (1971) is to give explicit, although rather complicated, algorithms for constructing both interpolating and smoothing spline. In fact, they give not only algorithms, but a general approach, based on reproducing kernel Hilbert spaces, for developing such algorithms. In Section 6 of their paper, they show that problems of type of (2.1) with linear inequality constraints may be solved as a quadratic programming problem. (They give a set of basis functions in that paper which is not particularly easy to use computationally. More favorable basis functions are given in Wahba (1978).) So long as there is a finite number of linear inequality constraints, their method is applicable and so restricted isotonic splines as in Section 3 of this paper may be computed directly. The key to being able to do this is characterization of isotonicity by the function, F, so that isotonicity becomes simply an additional finite set of linear inequality constraints.

The generalized isotonic spline has possibly an infinite number of knot points, that is, of linear inequality constraints. Moreover, Theorem 4.1 in contrast to

Theorem 3.1 does not specify the location of the constraints. Thus the results of Section 6 of Kimeldorf and Wahba (1971) do not apply directly to the computation of the generalized isotonic spline. As pointed out in Section 4, the role of the additional knot points is to force the spline to be isotonic where it may not naturally be so without the additional knots. In general, unless there is a very pathological situation, a finite number of constraints will do to make the spline isotone.

An approximate isotonic spline may be computed by adding a finite number of knot points, say j/N, $j = 1, 2, \cdots, N$ and requiring that $Ff(j/N) \geqslant 0$, $j = 1, 2, \cdots, N$. In case $F = I$, this requirement would become simply $f(j/N) \geqslant 0$, $j = 1, 2, \cdots, N$. If $F = D$, we simply have $f'(j/N) \geqslant 0$ which could be discretized as $f((j+1)/N) - f(j/n) \geqslant 0$, $j = 1, 2, \cdots, N-1$ and similarly for other F's. Then a numerical solution to the problem with discretized constraints is obtained by the quadratic programming approach of Kimeldorf and Wahba. In practice, if a solution with discretized constraints failed to satisfy isotonicity, one would add discretized constraints where the approximate solution was inadequate.

We note in closing, that if the j/N are sufficiently dense, then they will closely approximate the knot points specified by Theorem 4.1. This suggests that as the constraints are discretized more finely, the approximate solution will converge to the generalized isotonic spline (cf. Wahba (1978)).

Acknowledgment. The authors thank the referees for bringing to our attention several references we had overlooked and for very helpful discussions.

REFERENCES

[1] Anselone, P. M. and Laurent, P. J. (1968). A general method for the construction of interpolating or smoothing spline-functions. *Numer. Math.* **12** 66–82.

[2] Attéia, M. (1968). Fonctions (spline) definies sur un ensemble convexe. *Numer. Math.* **12** 192–210.

[3] Barlow, R. E., Bartholomew, D. J., Bremner, J. M. and Brunk, H. D. (1972). *Statistical Inference under Order Restrictions*. Wiley and Sons, New York.

[4] Barlow, R. E., Marshall, A. W. and Proschan, F. (1963). Properties of probability distributions with monotone hazard rate. *Ann. Math. Statist.* **34** 375–389.

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

[6] Copley, P. and Schumaker, L. L. (1978). On *pLg*-splines. *J. Approximation Theory* **23** 1–28.

[7] Daniel, J. W. and Schumaker, L. L. (1974). On the closedness of the linear image of a set with applications to generalized spline functions. *Applicable Anal.* **4** 191–205.

[8] Kimeldorf, G. S. and Wahba, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Statist.* **41** 495–502.

[9] Kimeldorf, G. S. and Wahba, G. (1971). Some results on Tchebycheffian spline functions. *J. Math. Anal. Appl.* **33** 82–95.

[10] Laurent, P. J. (1969). Construction of spline functions on a convex set. In *Approximation with Special Emphasis on Spline Functions*. (I. J. Schoenberg, ed.) 415–446. Academic Press, New York.

[11] Mangasarian, O. L. and Schumaker, L. L. (1969). Splines via optimal control. In *Approximation with Special Emphasis on Spline Functions*. (I. J. Schoenberg, ed.) 119–156. Academic Press, New York.

[12] Marshall, A. W. and Proschan, F. (1965). Maximum likelihood estimation for distributions with monotone failure rate. *Ann. Math. Statist.* **36**, 69–77.

[13] PASSOW, E. (1974). Piecewise monotone spline interpolation. *J. Approximation Theory* **12** 240–241.

[14] PASSOW, E. and ROULIER, J. A. (1977). Monotone and convex spline interpolation. *SIAM J. Numer. Anal.* **14** 904–909.

[15] RITTER, K. (1969). Splines and quadratic programming. Conference on Approximations, Univ. Wisconsin, Madison.

[16] ROBERTSON, T. (1967). On estimating a density which is measurable with respect to a σ-lattice. *Ann. Math. Statist.* **38** 482–493.

[17] WAHBA, G. (1973). On the minimization of a quadratic functional subject to a continuous family of linear inequality constraints. *SIAM J. Control* **11** 64–79.

[18] WAHBA, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. Tech. Report No. 508, Depart. Statist., Univ. Wisconsin, Madison.

[19] WEGMAN, E. J. (1970a). Maximum likelihood estimation of a unimodal density function. *Ann. Math. Statist.* **41** 457–471.

[20] WEGMAN, E. J. (1970b). Maximum likelihood estimation of a unimodal density, II. *Ann. Math. Statist.* **41** 2169–2174.

[21] WRIGHT, I. W. and WEGMAN, E. J. (1980). Splines in statistics. Unpublished manuscript.

DEPARTMENT OF MATHEMATICS
PAPUA NEW GUINEA UNIVERSITY OF TECHNOLOGY
LAE, PAPUA NEW GUINEA

DIRECTOR
STATISTICS AND PROBABILITY PROGRAM
DEPARTMENT OF THE NAVY
OFFICE OF NAVAL RESEARCH
ARLINGTON, VIRGINIA 22217