The Annals of Statistics
1982, Vol. 10, No. 3, 690–707

MAXIMUM LIKELIHOOD ESTIMATION OF ISOTONIC MODAL REGRESSION

BY THOMAS W. SAGER[1] AND RONALD A. THISTED[2]

University of Texas at Austin and University of Chicago

For each t in an index set T, let P_t be a probability measure with mode $M(t)$. In this paper we consider a maximum likelihood nonparametric estimator $\hat{M}_n(t)$ of $M(t)$ subject to the constraint that $\hat{M}_n(\cdot)$ be isotonic with respect to an order on T. The estimator is a solution to a minimization problem with zero-one loss. The estimator is not a max-min or min-max representation of "naive" estimators. Naive modal estimators are used but they are not linear in the sense of Robertson and Wright (1975) nor do they have the Cauchy mean value property. Consistency results are given for the cases of T finite, P_t discrete; T infinite, P_t continuous; T finite, P_t continuous. An efficient and economical quadratic-time dynamic programming algorithm is presented for computations. In the case of T finite, P_t discrete, the algorithm operates on a matrix of frequency counts, "backing up" one column at a time as the optimal completion cells for the modal estimates are searched for. Illustrative simulations suggest that the estimator performs well in small samples and is robust to certain kinds of contamination perturbations.

1. Introduction. The estimation of a regression function $M(t)$, $t \in T$, which is known to be monotonic (isotonic) with respect to a partial order $\ll$ on T is a problem which has attracted some interest in the literature. The following general model provides a framework for this problem: Let X be a stochastic process indexed by T. For each $t \in T$, let P_t be the marginal probability measure of $X(t, \cdot)$. Suppose Θ is a functional on the space of distributions $\{P_t; t \in T\}$ such that $\Theta(P_t) = M(t)$, a real-valued function of t, where $M(t)$ is known to be isotonic (i.e., $s, t \in T$ and $s \ll t \Rightarrow M(s) \leq M(t)$). The problem is to estimate $M(\cdot)$ from a finite segment of a sequence $\{(t_n, X_n); n = 1, 2, \cdots\}$ such that $t_n \in T$ and X_n is a random variable distributed as P_{t_n}. Usually, $\{X_n = 1, 2, \cdots\}$ are also assumed to be independent.

All estimators of $M(\cdot)$ are themselves required to be isotonic and many of the estimators proposed in the literature are solutions to the following kind of minimization problem: to minimize $\int_T L[g(t), m(t)]\, d\mu_n(t)$ within the class of isotonic, real-valued $m(t)$. Here, $g(t)$ is a given target function—possibly a known, nonisotonic, data-dependent "naive" estimator of $M(\cdot)$. $L(\cdot, \cdot)$ is a nonnegative loss function and $\mu_n(\cdot)$ is a nonnegative measure on T which may depend on observational data. A solution $\hat{M}_n(\cdot)$ to such a minimization problem is called an isotonic regression of $g(\cdot)$ with weight function $\mu_n(\cdot)$ and loss function $L(\cdot, \cdot)$.

EXAMPLES. (i) If T is finite and totally ordered, loss is squared-error,

$$n_i = \sum_{j=1}^{n} I_{\{t_i\}}(t_j), \quad g(t_i) = \sum_{j=1}^{n} X_j I_{\{t_i\}}(t_j)/n_i, \quad \mu_n(t_i) = n_i,$$

we have the classic initial isotonic problem considered by Brunk (1955). In addition, if the

Received June 1979; revised February 1982.

[1] Support for this research was provided by National Science Foundation Grant No. MPS 75-09450.

[2] Support for this research was provided by National Science Foundation Grant No. MCS 76-81435 and by U.S. Department of Energy Contract No. EY-76-S-02-2751.*000.

AMS 1970 *subject classifications.* Primary 62G05; secondary 60F15, 90C35, 62M10.

Key words and phrases. Regression; isotonic regression, mode, estimation of mode; nonparametric estimation; maximum likelihood estimation; consistency; Monte Carlo; algorithm; directed graph; dynamic programming.

X_i have independent normal distribution with mean $M(t_i)$ and variance σ^2, the solution $\hat{M}_n(\cdot)$ is also a restricted maximum likelihood estimate of $M(\cdot)$.

(ii) If T is finite and totally ordered, loss is absolute error, n_i and $\mu_n(\cdot)$ as in (i) and $g(t_i)$ is the median of the observations taken at t_i, we have the median regression case of Robertson and Waltman (1968). In this case, if the X_i have independent bilateral exponential distribution with median $M(t_i)$, then $\hat{M}_n(\cdot)$ is a restricted maximum likelihood estimate of $M(\cdot)$.

(iii) If T is the real line with usual order and n_i, $g(\cdot)$, $\mu_n(\cdot)$ are as in example (i), we have the regression setting considered by Brunk (1970).

(iv) If T is the plane with usual partial order, loss is squared-error, and n_i, $g(\cdot)$, $\mu(\cdot)$ are the multivariate analogs of (i), we have the problem of Hanson, Pledger, Wright (1973).

(v) If (iv) holds except that loss is absolute error, we have the problem of Cryer, et al (1972) or Robertson and Wright (1973).

In each of the cited examples, there is an explicit representation of the minimizing solution:

$$\hat{M}_n(t_i) = \max_{\{L;\, t_i \in L\}} \min_{\{K;\, t_i \in L \cap K\}} M_n\{X_j;\, j \leq n,\, t_j \in L \cap K\} \tag{1}$$

where $L \in \mathscr{L}$ (the class of all lower sets L defined by the conditions that $L \subset T$ and that $t \ll s$ and $t \in \mathscr{L} \Rightarrow s \in L$), $K \in \mathscr{L}^c$ (the class of upper sets, which are the complements of L-sets), and M_n is an estimator computed from the random variables in its argument. When T is totally ordered, the minimum lower sets algorithm (1) is equivalent to the simpler Pool-Adjacent-Violators algorithm (see Barlow, et al, 1972). There has been interest in the statistical properties of (1) independent of its relationship to any minimization problem (Hanson, Pledger, Wright, 1973; Robertson and Wright, 1974, 1975, 1980.) Since (1) is an isotonic function, it may receive consideration as an estimator of any constrained parameter for which M_n is a naive unconstrained estimator. This use of (1) has appeal if it is difficult to formulate or solve an appropriate optimality criterion. Moreover, the form of (1) suggests ways to apply properties of M_n to prove properties of the isotonic regression. However, it is not clear that independent of a minimization criterion (1) will have even the most basic desirable properties. An example given by Robertson and Wright (1974) provides the initial motivation for this paper: Let $T = \{1, 2\}$, X_{ij} be independent, $i = 1, 2; j = 1, 2, \cdots$, $P(X_{1j} = -\frac{1}{2}) = P(X_{1j} = -\frac{3}{2}) = \frac{2}{7}$, $P(X_{1j} = 0) = \frac{3}{7}$, X_{2j} distributed as $X_{1j} + 1$. With probability one, the sample modes converge to the true modes $M(1) = 0$ and $M(2) = 1$. But with probability one, (1) produces estimates converging to 0 and $-\frac{1}{2}$ if M_n is taken as the sample mode.

Conditions under which consistency of M_n implies consistency of (1) were investigated by Robertson and Wright (1975). These authors impose four general conditions on M_n: (i) measurability, (ii) symmetry in its arguments, (iii) linearity to location shift, and (iv) monotonicity in its arguments. It seems clear that they have location estimators in mind for M_n, yet the mode fails their condition (iv). In this paper, we let $M(t)$ be the mode of P_t and study the problem of estimating $M(\cdot)$ given that it is isotonic. In Section 2 we consider the case of T finite and P_t discrete; in Section 4, T is infinite and P_t continuous; in Section 5, T is finite and P_t continuous. Our estimator does not have a max-min representation as in (1), and we believe that it is the first such isotonic estimator to appear in the literature. The sample mode as a naive estimator is not linear in the sense of Robertson and Wright (1975) nor does it have the Cauchy mean value property (Robertson and Wright, 1974, 1980) characteristic of most isotonic estimators considered in the literature. Yet our isotonized estimator is the solution to a minimization problem with a very natural loss function and it is almost surely consistent. Moreover, it is easy to calculate with our algorithm (Section 3). We also believe that the estimator has practical utility as well as theoretical interest. Because the estimator seems to be fairly resistant to outliers, we believe that it deserves a place among robust methods for "bump-hunting" (e.g., in applied particle physics), particularly when there is a modest amount of contamination in the distributions P_t. Additionally, it provides a reasonable method for doing regression on ordinal data. Section 6 contains a few simulation studies which suggest that the estimator

may perform well even with quite modest sample sizes and in competition with other, more traditional estimators.

2. T finite, P_t discrete. First suppose that T is linearly ordered and identify T with the set of integers $\{1, 2, \cdots, k\}$. Let $\mathscr{U}$ be the σ-lattice of right subintervals of T, i.e., $\mathscr{U} = \{\phi, \{k\}, \{k-1, k\}, \cdots, T\}$, and let $R(\mathscr{U})$ be the class of all real-valued functions $g(t)$ which are measurable with respect to $\mathscr{U}$, i.e., $g^{-1}([\alpha, \infty]) \in \mathscr{U}$ for each real α. Note that $g \in R(\mathscr{U})$ if and only if g is isotonic on T. For each $t \in T$, let $X_{1t}, \cdots, X_{n_t,t}$ be a random sample from the discrete distribution P_t and let the collection $\{X_{ij};\ i = 1, \cdots, n_j, j = 1, \cdots, k\}$ be independent. Suppose each distribution P_t has a unique mode $M(t)$, which is unknown. If no order restrictions are imposed, then the class of all real-valued functions g on T is the parameter set for the problem of estimating the true modal function $\{M(t);\ t \in T\}$. If g is such a function and we observe $X_{ij} = x_{ij}$, $i = 1, \cdots, n_j$, $j = 1, \cdots, k$, it is reasonable to define the nonparametric likelihood of g as

$$\lambda(g) = \prod_{j=1}^{k} \{\sum_{i=1}^{n_j} I_{\{g(j)\}}(x_{ij})/n_j\}.$$

Clearly, $\lambda(g)$ is maximized by setting $g(t) = m(t)$ where, for each t, $m(t)$ is a sample mode of $\{x_{it};\ i = 1, \cdots, n_t\}$ (if the sample mode is not unique, we are indifferent at this point as to which one is chosen). However, $m(t)$ may not be $\mathscr{U}$-measurable. If the statistician has exogenous knowledge that the true modal path $M(t)$ is nondecreasing, either from physical necessity or strong personal belief, it seems objectionable in principle to estimate $M(t)$ by the sample mode. We propose instead an estimator $\hat{M}_n(t)$ which is always isotonic.

For a given function $g(\cdot)$, let $f(t, g(t))$ be the sample frequency of $g(t)$ in the tth sample. Let $F(t, g(t))$ be the probability of $g(t)$ in the tth population. That is,

$$f(t, g(t)) = \sum_{i=1}^{n_t} I_{\{g(j)\}}(x_{it}), \qquad F(t, g(t)) = P_t[X_{1t} = g(t)]. \tag{2}$$

In computational work, it is sometimes convenient to replace frequencies of 0 by some small positive number like ½; but this is by no means necessary for the theory and in practice makes a difference only for very small sample sizes. Let

$$L(a, b) = \begin{cases} 1 & \text{if } a \neq b \\ 0 & \text{if } a = b \end{cases} \tag{3}$$

denote zero-one loss. Let

$$\mu_n(t) = \log \frac{f(t, m(t))}{f(t, g(t))}, \tag{4}$$

where $M(t)$ is the sample mode of the tth sample and $n = n_1 + \cdots + n_k$. Let

$$\Delta(m, g) = \int_T L(m(t), g(t))\, d\mu_n(t) = \sum_{t=1}^{k} [1 - I_{\{m(t)\}}(g(t))] \log \frac{f(t, m(t))}{f(t, g(t))}. \tag{5}$$

We shall call any function $\hat{M}_n(\cdot)$ which is $\mathscr{U}$-measurable and satisfies $\Delta(m, \hat{M}_n) = \min\{\Delta(m, g);\ g \in R(\mathscr{U})\}$ an isotonic modal regression of $m(\cdot)$. Since the total number of observations is finite, it is clear that there is at least one such minimizing function. Moreover, if the sample mode $m(\cdot)$ is isotonic, then $\hat{M}_n = m$ because (5) achieves its minimum value of 0 by setting $g = m$. The main reason for our choice of weight measure (4) is the connection it provides between (5) and the likelihood: Since

$$\Delta(m, g) = \text{constant} - \log\{\prod_{j=1}^{k} n_j \cdot \lambda(g)\},$$

it is apparent that any isotonic modal regression $\hat{M}_n(\cdot)$ will also be a restricted maximum likelihood estimate of $M(\cdot)$. In this regard, one might also note that with μ_n and L as defined, $\Delta(m, g)$ is proportional to the log likelihood of the parameters $m(\cdot)$ and $g(\cdot)$. Δ would be precisely the log likelihood if we had used relative frequencies in defining

$f(t, g(t))$ in (2) (our procedure is invariant to multiplicative scaling of any of the k sample frequency sets). It is important to observe that the loss function L is unusual in that it does not depend at all on the magnitude of $m(t) - g(t)$. Instead, we judge closeness on the basis of frequencies. This point is the key to the entire paper.

Simple properties of the estimator.

DEFINITION. Let $\mathcal{F}$ denote the class of all real-valued functions on T. For any $g, h \in \mathcal{F}$ define

$$d(g, h) = \left| \frac{1}{k} \sum_{i=1}^{k} \log \frac{f(i, g(i))}{f(i, h(i))} \right| \quad \text{and} \quad D(g, h) = \left| \frac{1}{k} \sum_{i=1}^{k} \log \frac{F(i, g(i))}{F(i, h(i))} \right|.$$

Here and in (4), $\frac{a}{0} \equiv \log \frac{a}{0} \equiv -\log \frac{0}{a} \equiv \infty$ for any $a > 0$.

With the equivalence definition in (b) below, the functions in F form a one-dimensional set. To see this, consider the mapping $g \leftrightarrow (1/k) \sum_i \log f(i, g(i))$. Then $d(g, h)$ is simply the absolute difference of the images of g and h under this mapping. For other uses of such mappings in isotonic regression, see Goldstein and Kruskal (1976).

The following properties are easily verified:

PROPOSITION 1. (a) *$\hat{M}_n$ is a restricted maximum likelihood estimator of M within the class of discrete distributions.*

(b) *d and D are metrics on $\mathcal{F}$ provided we identify functions with the same log-product: g is d- [D-] equivalent to h if and only if $d(g, h) = 0$ [$D(g, h) = 0$].*

(c) *$d(m, \hat{M}_n) \le d(m, g)$ for all $g \in R(\mathcal{U})$*

(d) *$d(m, g) = d(m, r) + d(r, g)$ for all $g \in R(\mathcal{U})$ if and only if $r = \hat{M}_n$ (up to d-equivalence).*

(e) *$d(M, \hat{M}_n) \le d(M, m)$*

(f) *$\hat{M}_n \equiv M$ for all $\min\{n_j; j = 1, \cdots, k\}$ sufficiently large with probability one.*

(g) *$d(g, h) \to D(g, h)$ as $\min\{n_j; j = 1, \cdots, k\} \to \infty$ for all $g, h \in \mathcal{F}$.*

(h) *$d(\hat{M}_n, h) \to D(M, h)$ almost surely as $\min\{n_j; j = 1, \cdots, k\} \to \infty$ for all $h \in \mathcal{F}$.*

It should be noted that convergence in the sup metric on $\mathcal{F}$ is not equivalent to convergence in either the d- or D-metric. Indeed, it is possible to choose g and g_n such that $f(t, g_n(t)) = 0$ [$F(t, g_n(t)) = 0$] for all t but $g_n(t) \to g(t)$ uniformly in t, which yields $d(g_n, g) = \infty$ [$D(g_n, g) = \infty$].

In view of property (f), the asymptotic distribution of $\hat{M}_n$ is uninteresting: $a_n \cdot \sup_{t \in T} |\hat{M}_n(t) - M(t)|$ is degenerate at 0 in the limit for any a_n. Thus, asymptotically, $\hat{M}_n$ is at least no worse than any other estimator, including an isotonic mean regression, even when the population mean regression function and mode regression function $M(t)$ are identical. The asymptotic distribution of the metric distances $d(\hat{M}_n, M)$ and $D(\hat{M}_n, M)$ are similarly degenerate at 0. Thus, comparisons of estimators in terms of the relevant metrics may be meaningful only in small samples. Some simulation comparisons are presented in Section 7.

Use of our zero-one loss function in practice implies an all-out effort to maximize the probability of a correct decision and an indifference to alternatives among erroneous choices. Since the mode maximizes likelihood, zero-one loss, therefore, implies interest in the mode as the objective of statistical inquiry (e.g., see related problems in Ferguson, 1967, page 51, problem 5; page 174, problem 1; page 180, problem 1, 2; page 190, problem 2). Conversely, interest in estimating the mode leads directly to the criterion of maximum likelihood and thereby to the metric (5). Thus zero-one loss is the natural criterion for modal regression.

But when may modal regression be preferred to conventional alternatives? First, one may really want to know the mode rather than the mean or median. For example, one may

suppose that patients are randomly assigned to K different treatments, ordered according to their presumed efficacy. The response may be an ordinal categorical variable with values ranging from "very negative reaction" to "very positive reaction." Without arbitrary coding, the mode is the only location parameter which makes sense here. Moreover, it is of medical interest to know what the most frequent response to a treatment is. One expects a trend in the modes of the K treatments which modal regression can estimate, even though reversals in the trend of the sample modes may occur due to variability in treatment results.

Second, one may set out to learn the location of a primary distribution, but there may be present an unknown amount of contamination from a secondary distribution. Robust methods are called for, and Section 7 suggests that modal regression performs well in these circumstances.

Finally, even with continuous distributions, one may really be primarily interested in the mode. For example, one may wish to know whether man's "natural" lifespan has increased since the Industrial Revolution. Available data suggest that mean and median lifespan have increased—the former more than the latter. However, much of the increase may be attributed to prevention of infantile death and cure of many of the former diseases of middle age, without necessarily having lengthened the "natural" lifespan. Although the latter concept is not entirely clear, it is evident that it is neither mean nor median lifespan. If it is taken to be maximum lifespan, estimation will be subject to the large variability of extreme values. We think a case could be made for thinking of it as the most frequent, or typical, age at death. Then modal regression could examine the evidence for monotonicity over time by means of the procedure of Section 4.

3. The algorithm. The algorithm for computing an isotonic modal regression that we present below is first developed for the case of finite T and discrete P_t. Subsequent sections describe the modifications to the algorithm needed to accomodate the cases in which T is infinite and/or in which P_t is continuous.

Assume, then, that T is finite and P_t is discrete. We have identified T with the set of integers $\{1, 2, \cdots, K\}$. Suppose that in the K samples that have been observed, N distinct values have been observed, say $C_1 < C_2 < \cdots < C_N$. The data can then be represented as the matrix of frequencies B, where

$$B_{ij} = f(j, C_i) = \text{the number of occurrences of } C_i \text{ in the } j\text{th sample.}$$

Thus, each column of B represents one sample, and within each sample larger observations are represented by frequency counts in lower rows. Finding an isotonic regression, then, is equivalent to selecting one element from each column such that (a) the values associated with these elements are nondecreasing (which is the same as requiring the row number of each selected element to be at least as great as that of the selected element in the previous column) *and* (b) the product of the selected elements is maximized over all such choices of cells. An example is given in Table 1. In the example, the isotonic regression does not coincide with the sample modes from the five populations.

A sequence of cells (one from each column) with nondecreasing row numbers is called an isotonic path. In an $N \times K$ frequency matrix, there are $\binom{N+K-1}{N-1}$ possible isotonic paths. Nevertheless, our algorithm requires computation time that grows only as $O(KN)$. This computational efficiency is achieved through dynamic programming.

Suppose that elements from columns 1 through $K-1$ of B have been selected, and that this partial path ends at position $(i, K-1)$. It is a simple matter to complete the path optimally—just select that cell in column K and in row i or below which has the largest value. Thus, from any cell in column $K-1$ we can easily determine both the optimal path *completion* and the contribution of each such path to the maximum product. In particular, the contribution of columns $K-1$ and K to the overall product is simply the product of the two elements selected from the two columns. The (partial) product of element $(i, K$

TABLE 1
Reading scores at each of five grade levels. (The data are artificial). The isotonic modal regression is indicated by the boldface array elements.

Reading level	Grade level				
	8	9	10	11	12
Remedial	**8**	5	4	1	3
Jr. High School	3	4	2	2	4
High School	6	**6**	**4**	**14**	5
College	3	5	10	4	**8**

TABLE 2
A possible frequency matrix

1	0	0	0	0
0	0	1	0	0
3	2	3	2	0
3	3	8	1	0
0	1	5	10	2
0	0	1	1	3
0	0	0	0	5
0	0	1	0	1

$- 1$) and its optimal completion in column K might be called the *potential* for the (overall) optimal path to pass through element $(i, K - 1)$. At this point we can "back up" to column $K - 2$ and consider optimal path completions from each element in that column. For each element i in column $K - 2$ we select that element i' in column $K - 1$ whose optimal completion (previously computed) has largest product frequencies, that is, whose potential is greatest. This process is then repeated until the potentials for the first column have been computed, at which point the maximal product is simply the largest potential from elements in column 1, and the path which corresponds is obtained by retracing the steps outlined above.

The following algorithm implements these ideas, with a few refinements to speed computation. One of these refinements implements the observation that all N completion potentials for column c may be found with a single pass through column $c + 1$: start in row N of column $c + 1$ and, working upwards, keep track of the partial maxima of the column (Steps 2 and 3). Moreover, the location of the maximizing potential may be stored in the *path* matrix (Step 4). Not until Step 5 is it necessary to multiply the completion potentials by the B_{ic}'s to obtain the potentials for column c.

ALGORITHM ISOTONIC.

Step 1. [Initialize.] For $i = 1, 2, \cdots, N$ set *potential* $(i) \leftarrow B_{i,k}$. Set $c \leftarrow K - 1$.

Step 2. [Initialize current column.] Set $r \leftarrow N - 1$, index $\leftarrow N$, $path(N, c) \leftarrow N$, and maxpot $\leftarrow potential(N)$. Maxpot contains the largest potential from column $c + 1$ at or above r, the current row; the corresponding row is contained in index.

Step 3. [Is current row potential larger than those above?] If $potential(r) >$ maxpot then set index $\leftarrow r$ and maxpot $\leftarrow potential(r)$.

Step 4. [Record optimal path completion]. Set $path(r, c) \leftarrow$ index, $r \leftarrow r - 1$. If $r > 0$ then go to Step 3.

Step 5. [Record column c potentials.] For $i = 1, 2, \cdots, N$, set $potential(i) \leftarrow B_{ic} \cdot potential(path(i, c))$. Set $c \leftarrow c - 1$. If $c > 0$ then go to Step 2.

Step 6. [Recover isotonic paths $g(\cdot)$.] Select i so that $potential(i) \geq potential(i')$ for $i' = 1, 2, \cdots, N$. Set $g(1) \leftarrow i$, $g(t) \leftarrow path(g(t - 1), t - 1)$ for $t = 2, \cdots, K$.

The proof of the correctness of the algorithm simply involves backward induction to show that after each time Step 5 is performed, the partial paths (pointed to by the *path* matrix) are indeed optimal paths from column c through the matrix of the subproblem consisting only of columns c through k. The details are omitted.

A further refinement, not implemented in ISOTONIC, might be considered for frequency matrices B when large upper right and lower left corners of B are zero. An example is Table 2. There, it is clear that the isotonic modal regression must be contained within the dark upper and lower boundaries. The preprocessing necessary to find the lower boundary is quite simple: if the boundary is to contain row r of column c, then it should also contain the first nonzero row $r' \geq r$ of column $c + 1$. Analogously for the upper

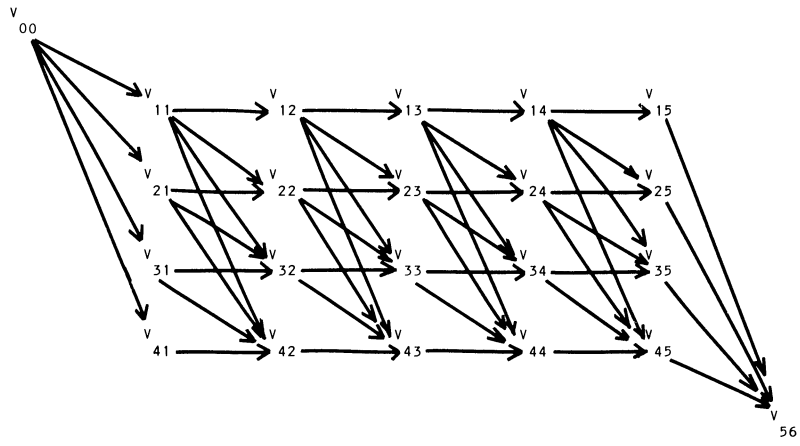

FIG. 1. *Directed graph corresponding to the isotonic modal regression problem of Table* 1. *The cost of traversing the edge from* $V_{i,J}$ *to* $V_{i',J+1}$ *is* $-\log B_{i',J+1}$, *except that the cost to travel from* V_{i5} *to* V_{56} *is zero for all* i.

boundary. We have found this refinement useful in applying the algorithm to continuous P_t.

The analysis of algorithm ISOTONIC is also straightforward. Step 1 is executed once and has $N + 1$ assignments. Step 2 is executed $K - 1$ times and has 4 assignments and one multiplication. Step 3 is executed $(K - 1)(N - 1)$ times, has one comparison and at most two assignments. Step 4 also is executed $(K - 1)(N - 1)$ times, has two assignments, one multiplication, one decrement, and one comparison. Step 5 occurs $K - 1$ times and involves $3N$ multiplications, $N + 1$ assignments, one decrement, and one compare. Step 6 is done once and implies $N - 1$ compares, at most $K + N$ assignments, and $K - 1$ multiplications and decrements. Multiplication counts above include those needed for subscript computations. If multiplication is the dominant expense, the procedure requires $(4N + 1)(K - 1)$, however all of the computations are $O(KN)$.

Ties may occur in Step 6 and any method of breaking ties is acceptable; the most easily implemented is simply to select the first path encountered which satisfies the conditions. Indeed, ISOTONIC incorporates this method explicitly in its Step 3.

When K is large (it does not matter that N be), or if the elements of B are very large, overflow or loss of precision may be a problem. Replacing the data by their logarithms and changing the multiplication in Step 3 of the algorithm to addition remedies the problem. Of course, the logarithm of the maximal product is obtained instead of the product itself. In this format our algorithm bears a family resemblance to "sequence comparison" algorithms based on Levenshtein distance (Levenshtein, 1965). Sequences of letters, protein molecules, computer codes, phonemes, etc., may be compared with similar sequences in terms of the minimum weighted number of certain elementary transformations necessary to change one sequence into another (the Levenshtein distance). "Back up" dynamic programming algorithms exist which permit the efficient computation of this minimum distance. In this context, our algorithm could be thought of as transforming the sample modal sequence into the minimum-distance isotonic sequence via row substitutions. Here, of course, only the beginning sequence (the sample modal path) is known and the target sequence must be searched for.

The problem of computing an isotonic modal regression can be reformulated as that of finding the shortest path from a single-source through a directed graph. The corresponding graph has $NK + 2$ vertices, all but two of which correspond to elements of B. The two extra vertices represent a starting vertex and a terminal vertex shared by all paths. The edges of the graph connect points $v_{i,J}$ to $v_{i',J+1}$ where $i' \geq i$. The starting vertex is $v_{0,0}$ and the terminal vertex is $v_{N+1,K+1}$. The cost associated with the edge from $v_{i,J}$ to $v_{i',J+1}$ is simply

$-\log F_{i',j+1}$, except all edges to $v_{N+1,K+1}$ have cost zero. Such a graph is displayed in Figure 1. The graphs which correspond to isotonic modal regression problems display a special structure which was exploited in the construction of our algorithms—all of the edges proceed downward and to the right. General procedures for finding shortest paths from single-sources have time complexity of $O(KN^2 \log KN)$; our algorithm is $O(KN)$. For a general discussion of shortest path algorithms, see Aho, Hopcroft, and Ullman (1974).

4. T infinite, X_t continuous. In this section, we shall consider estimation of an isotonic modal regression function $\theta(t)$ when t lies in the closed interval $T = [0, 1]$ and the distributions of X_t are continuous. The main problems we encounter here arise from the difficulty of estimating the mode of a continuous random variable and the manner in which the data points become available to us. By assuming continuity of the modal regression $\theta(t)$, we may hope that observations taken near t will behave somewhat like observations taken at t. If so, we may estimate $\theta(\cdot)$ by adapting versions of modal estimators previously considered for a single random variable (e.g., Parzen, 1962; Chernoff, 1964; Venter, 1967). Of course, a good estimate will require that the observations near t become available to us fairly rapidly. And because we are estimating the entire curve $\theta(\cdot)$, observations must be fairly dense everywhere in T.

The estimate. The essential idea in constructing the estimate for the continuous case of this section is to reduce the problem to the case of Section 2 by discretization. We construct a gridwork of cells in the plane and record the number of data points lying in each cell. Applying the algorithm of Section 3 to the matrix of counts just derived, we obtain an optimal path of cells. The estimate is then obtained by constructing a monotone nondecreasing curve which passes through each cell of the optimal path of cells. Details follow.

Choose an integer $m = m(n)$. Let $\mathscr{S} = \mathscr{S}(m(n))$ be the set of all squares of the form $S = (i/m, (i+1)/m] \times (j/m, (j+1)/m]$, $i = 1, 2, \cdots, m-1$; $j = 0, \pm 1, \pm 2, \cdots$; or $S = [0, 1/m] \times (j/m, (j+1)/m]$. When plotted in the plane, the squares in $\mathscr{S}$ partition $[0, 1] \times (-\infty, \infty)$ into a finite number of columns $\mathscr{C}_1, \cdots, \mathscr{C}_m$ of squares and an infinite number of rows of squares. Given observations $(t_1, x_{t_1}), \cdots, (t_n, x_{t_n})$, we record in each square the number of observations in the square. The algorithm of Section 3 is then applied to the matrix of recorded counts to select a path $\hat{M}_1, \cdots, \hat{M}_m$ of m squares—one square from each column of the matrix—having the greatest product of recorded counts among all nondecreasing paths. (It is not necessary that $\hat{M}_1, \cdots, \hat{M}_m$ be unique.) For our estimated modal regression, we choose a function $\hat{\theta}_n(t)$, $0 \le t \le 1$, which is continuous and monotone nondecreasing and whose graph contains at least one point from each of the m squares $\hat{M}_1, \cdots, \hat{M}_m$ constituting our optimal path. For our purposes (and in particular, for consistency), it does not matter which of the possible functions $\theta_n(\cdot)$ satisfying these conditions is chosen, except that for consistency we require that $(0, \theta_n(0)) \in \hat{M}_1$ and $(1, \theta_n(1)) \in \hat{M}_m$. This requirement will allow us to assert uniform consistency of $\hat{\theta}_n(t)$ for $0 \le t \le 1$. Without some kind of restriction on $\theta_n(t)$ at $t = 0$ and $t = 1$, we could assert uniform consistency only for some subinterval $[a, b]$, $0 < a < b < 1$ as in Hanson, Pledger, and Wright (1973) and Robertson and Wright (1973), (1975). (For example if the range of X_t is $[0, 1]$ for all t and the population modal regression is $\theta(t) = 1/4 + 1/2\, t$, then the estimate $\hat{\theta}_n(t)$ which connects centers of $\hat{M}_1, \cdots, \hat{M}_m$ but is tied down at (0, 0) and (1, 1) will fail of consistency at $t = 0$ and $t = 1$.) An example of an estimate $\hat{\theta}_n(\cdot)$ meeting our conditions is $\hat{\theta}_n(t) \equiv \hat{\theta}_n(1/2\, m)$ for $0 \le t < 1/2\, m$, $\hat{\theta}_n(t) \equiv \hat{\theta}_n(1 - 1/2\, m)$ for $1 - 1/2\, m < t \le 1$, and otherwise $\hat{\theta}_n(t)$ is the polygonal path connecting centers of the optimal squares $\hat{M}_1, \cdots, \hat{M}_m$.

Of the three approaches to modal estimation mentioned earlier (Parzen, Chernoff, Venter), our approach here is closest in spirit to that of Chernoff, who estimates a univariate mode by a point from an interval of given length containing the most observations. In principle, there is no reason why we could not adapt the methods of Parzen and Venter. For example, to use the Venter-type estimator, one could determine the columns $\mathscr{C}_1, \cdots, \mathscr{C}_m$ by partitioning the t-axis $[0, 1]$ at every (n/m)th observation; and the horizontal

sides of boxes within each column would be drawn at every kth observation (from bottom to top of the column). Since all boxes would contain the same number of observations, the algorithm of Section 3 could then be applied to select a nondecreasing path of boxes which *minimizes* the product of the *areas* of the boxes. To adapt the Parzen estimator, choose a bivariate kernal $K(t, x)$ and sequences $h_1 = h_1(n)$, $h_2 = h_2(n)$ and define

$$F(t, x) = (n\, h_1 h_2)^{-1} \sum_{j=1}^{n} K\left(\frac{t - t_j}{h_1}, \frac{x - X_{t_j}}{h_2}\right)$$

in the spirit of Cacoullos' (1966) multivariate density estimator. Then choose the continuous, nondecreasing path $\{(t, F(t, x)); 0 \le t \le 1\}$ which maximizes

$$\exp\left\{\int_0^1 \log F(t, x)\, dt\right\}$$

(cf. discrete analog in Section 2).

It may also be possible to combine the three methods of modal estimation. For example, we might use the Chernoff fixed-interval method to divide [0, 1] into columns and then use the Venter variable-box method on the data within each column. The method of fixed squares we have chosen for this section has the advantages of simplicity and easy applicability.

We now formalize the mathematical assumptions for this section.

ASSUMPTIONS.

1. $T = [0, 1]$; the finite dimensional distributions of the stochastic process X are independent. This is called an independent observations model by Brunk (1970).
2. For each t, X_t has a distribution function $P_t(\cdot)$ which is absolutely continuous with respect to Lebesque measure on the real line and has a uniquely defined density $p_t(\cdot)$.
3. For each t, X_t has a unique mode $\theta(t) : p_t(\theta(t)) > p_t(x)$ for all $x \neq \theta(t)$; and $\forall\, \varepsilon > 0$, $\exists$ positive δ (independent of t) such that $|\theta(t) - x| > \varepsilon \Rightarrow p_t(\theta(t)) > p_t(x) + \delta$.
4. The family $\{p_t(\theta(t) - x); 0 \le t \le 1\}$ is equicontinuous at $x = 0$.
5. $\theta(t)$ is a monotone, nondecreasing, continuous function of t.

The full force of equicontinuity (Assumption 4) is not used in the sequel. We require that the P_t probability content of intervals close to $\theta(t)$ be uniformly (in t) greater than the P_t probability content of equally long intervals far from $\theta(t)$. In conjunction with the uniformity condition on δ in Assumption 3, equicontinuity yields this but eliminates many pathologies that Assumption 3 alone permits.

We suppose that the statistician has available the first n observations in the sequence $\{(t_i, X_i); i = 1, 2, \cdots\}$ where $0 \le t_i \le 1$, and $X_1, X_2, \cdots$ are independent with X_i distributed as P_{t_i}. At times the statistician will be able to control the points $t_1, t_2, \cdots$ at which he collects data; at other times he will be at the mercy of chance. For example, the data may become available according to a jointly distributed random variable (T, X). In this case, the observations are thought of as a realization of N i.i.d. copies of (T, X) where $T = t$ is observed according to the marginal distribution of T and then $X_t = x_t$ is observed according to the conditional distribution of X given $T = t$ (see Brunk, 1970). Regardless of the mechanism, we must have "enough" observations in each subinterval of [0, 1] to estimate $\theta(t)$ over that subinterval. At least three distinct definitions of "enough" have been considered in the literature in other regression contexts:

(i) $\{t_1, t_2, \cdots\}$ is dense in [0, 1].

(ii) For every nondegenerate interval $J \subset [0, 1]$,

$$\liminf_{n\to\infty} n^{-1} \sum_{i=1}^{n} I_J(t_i) > 0.$$

(iii) $\{t_1, t_2 \cdots\}$ is a realization of a sequence of i.i.d. variables distributed as T which assigns positive probability to every nondegenerate interval of [0, 1].

In most regressions, the t's are considered fixed, nonrandom points and the analysis is

conditional on the t's. We now assume explicitly that $t_1, \cdots, t_n$ are fixed and known at this stage of the analysis, even though they may have been generated earlier as observations on a random T. Before we present our definition of "enough," which is more convenient for our purposes than (i), (ii), or (iii) above, we introduce some useful notation.

DEFINITION 1. Let A be any set in the plane. Given observations $(t_1, x_1), \cdots, (t_n, x_n)$, define

$$N(A) \equiv N_n(A) = \sum_{i=1}^n I_A(t_i, X_i), \quad Q(A) \equiv Q_n(A) = \sum_{i=1}^n P_{t_i}(A_{t_i})$$

where $A_{t_i} = \{x;\ (t_i, x) \in A\}$ is the t_i-section of A,

$$n_j = \sum_{i=1}^n I_{[j/m,(j+1)m]}(t_i), \quad j = 1, \cdots, m.$$

$$n_{\min} = \min\{n_1, \cdots, n_m\}.$$

Note that $N(A)$ is the sum of non-identically distributed Bernoulli variables and so is not quite binomial, but that $E\{N(A)\} = Q(A)$.

We shall suppose that the following two additional assumptions hold.

ASSUMPTIONS.

6. $$\sum_{n=1}^\infty [m(n)]\, \lambda^{n_{\min}/[m(n)]} < \infty \quad \text{for all} \quad 0 < \lambda < 1.$$

7. $$m(n) \to \infty.$$

Since $n_{\min} \le n/m(n)$ and Assumption 6 requires $m(n) = o(n_{\min})$, then $m(n)$ is at most $o(n^{1/2})$ under our assumptions. In fact, if the points $\{t_i\}$ become available at a "uniform" rate, then $n_{\min} \approx n/m(n)$, so a choice of $m(n) = O(n^{1/2}/\log n)$ but $m(n) \to \infty$ would satisfy Assumptions 6 and 7. It should be noted that the latter are less general than (i) above, more general than (iii) (if $m(n) = O(n^\alpha)$ for $0 < \alpha < 1/2$), but neither more nor less general than (ii).

THEOREM 1. *Under Assumptions* 1 − 7, $\sup_{0 \le t \le 1} |\theta_n(t) - \theta(t)| \to 0$ *almost surely.*

The proof of the theorem proceeds through a series of lemmas, but the idea is relatively simple. Associated with each nondecreasing path of squares—one square from each column of squares—is the value which is the product of the numbers of observations in the constituent squares of the path. Roughly, the greater the probability content of the path, the greater its product-count should be. Thus we expect to find the largest product-counts associated with paths lying near the true modal regression, where the concentration of probability is greatest. However, if square size decreases too rapidly as n grows, observed product-counts will be too unstable. Assumption 6 ensures that square size decreases at an appropriate rate. The idea of the proof, therefore, is to show that a path deviating from the true modal regression by more than ε at any point has a lower expected product-count (thus ultimately a lower observed product-count) than a reference path which we shall choose uniformly within ε of the modal regression.

The first step is to reduce the infinite collection of squares to a manageable finite number of rectangles by selection and amalgamation. Fix j and focus on column $\mathscr{C}_j$. Choose $r \equiv r_j$, $s \equiv s_j$, and $u_k \equiv u_{kj}$, $k = 0, 1, \cdots, r_j$ such that

(6)
$$\begin{cases} \mathscr{A}_j = \{S \in \mathscr{S};\ (t, \theta(t)) \in S \quad \text{for some } t,\ j/m < t \le (j+1)/m\} \\ p_j = n_j^{-1} \min\{Q(S);\ S \in \mathscr{A}_j\} \\ -\infty = u_0 < u_1 < \cdots < u_{r-1} < u_r = \infty \\ R_{kj} = (j/m,\ (j+1)/m] \times (u_{kj},\ u_{k+1,j}] \\ \delta \text{ corresponds to } \varepsilon \text{ as in Assumption 3} \\ Q(R_{kj}) = n_j p_j \left(1 - \dfrac{\delta}{2}\right)(s-2)^{-1} = n_j r^{-1}, \quad k = 0, 1, \cdots, r-1. \end{cases}$$

LEMMA 1. *Let $\nu > 0$, $r \equiv r_j$ and choose an integer $s \equiv s_j < r_j$.*

$$P\left[\min_{0\le i\le r-s}\frac{N(\cup_{k=i}^{i+s-1}R_{kj})}{Q(\cup_{k=i}^{i+s-1}R_{kj})} < e^{-\nu}\right] \le r\exp\left[n_jp_j\left(1-\frac{\delta}{2}\right)s(s-2)^{-1}\{-\nu^2/2+O(\nu^3)\}\right]$$

PROOF. For $\lambda > 0$, the probability in question is less than or equal to

$$\begin{aligned}
&\textstyle\sum_{i=0}^{r-s} P[N(\cup_{k=i}^{i+s-1}R_{kj}) < e^{-\nu}Q(\cup_{k=i}^{i+s-1}R_{kj})] \\
&\quad\le \textstyle\sum_{i=0}^{r-s}\exp\{-\lambda e^{-\nu}Q(\cup_{k=i}^{i+s-1}R_{kj})\}E[\exp\{-\lambda N(\cup_{k=i}^{i+s-1}R_{kj})\}] \\
&\quad= \exp\left\{-\lambda e^{-\nu}n_jp_j\left(1-\frac{\delta}{2}\right)s(s-2)^{-1}\right\} \\
&\qquad\textstyle\sum_{i=0}^{r-s}\prod_{t_l=1}^{n_j}[1-P_{t_l}((\cup_{k=i}^{i+s-1}R_{kj})_{t_l})+P_{t_l}((\cup_{k=i}^{i+s-1}R_{kj})_{t_l})e^{-\lambda}] \\
&\quad\le \exp\left\{-\lambda e^{-\nu}n_jp_j\left(1-\frac{\delta}{2}\right)s(s-2)^{-1}\right\}\textstyle\sum_{i=0}^{r-s}\exp\{(e^{-\lambda}-1)Q(\cup_{k=i}^{i+s-1}R_{kj})\}.
\end{aligned}$$

In obtaining this result, we have used the familiar inequality $P(Z < a) \le \exp(\lambda a)\cdot E[\exp(-\lambda Z)]$, $\lambda > 0$, the Bernoulli moment-generating function, and the inequality $1 + z \le \exp(z)$. Setting $\lambda = \nu$ in the final expression and simplifying, we obtain the upper bound.

LEMMA 2. *Let $\nu > 0$.*

$$P\left[\min_{\{S\in\mathscr{A}_j;\,Q(S)\ge n_jp_j(1-\delta)\}}\frac{N(S)}{Q(S)} < e^{-\nu}\right] < p_j^{-1}\exp\left[n_jp_j\left(1-\frac{\delta}{2}\right)s(s-2)^{-1}\left\{\frac{-\nu^2}{2}+O(\nu^3)\right\}\right].$$

PROOF. To obtain this result, observe that $\{S \in \mathscr{A}_j;\ Q(S) \ge n_jp_j(1-\delta)\}$ contains at most p_j^{-1} member squares S and apply the argument of Lemma 1.

LEMMA 3. *Let $\{T_1, \cdots, T_m\}$ be a path of sets such that either $Q(T_j) \ge n_jp_j(1-\delta)$ and $T_j \in S$ or $T_j = \cup_{k=i}^{i+s-1}R_{kj}$ for some i. Suppose* $\max r_j = O(m(n))$. *Then*

$$\{\textstyle\prod_{j=1}^m N(T_j)/(Q(T_j))\}^{1/m} \to 1 \quad \textit{almost surely.}$$

PROOF. Let $\nu > 0$. By Lemmas 1 and 2,

$$\begin{aligned}
P\left[\left\{\prod_{j=1}^m\frac{N(T_j)}{Q(T_j)}\right\}^{1/m} < e^{-\nu}\right] &< \sum_{j=1}^m r_j\exp\left[n_jp_j\left(1-\frac{\delta}{2}\right)s(s-2)^{-1}\left\{\frac{-\nu^2}{2}+O(\nu^3)\right\}\right] \\
&\quad+\sum_{j=1}^m p_j^{-1}\exp\left[n_jp_j(1-\delta)s(s-2)^{-1}\left\{\frac{-\nu^2}{2}+O(\nu^3)\right\}\right].
\end{aligned}$$

From Assumptions 3 and 4 it follows that for large n there is a constant $c > 0$ such that $p_j > c/m$ for all j. Replace p_j by c/m, n_j by $n_{\min}$, and set $\lambda = \exp\{-c(1-\delta)\nu^2/2\}$ in the above expression. Assumption 6 and the Borel-Cantelli lemma now yield $\liminf_{n\to\infty}[\Pi N(T_j)/Q(T_j)]^{1/m} \le 1$ a.s. To get the lim sup ≥ 1, observe that analogues of Lemmas 1 and 2 may be obtained for $P[\max(\text{expression}) > e^{\nu}]$. The proofs use $P[Z > a] \le \exp(-\lambda a)E[\exp(\lambda Z)]$ but are otherwise the same as for Lemmas 1 and 2, and the same probability bounds are obtained.

We now choose a particular reference path of squares. For each $j = 1, \cdots, m$, let E_j be any member $S \in \mathscr{A}_j$. We shall use the path $\{E_1, \cdots, E_m\}$ to prove consistency by showing that (for large n) any other path deviating from the modal regression by more than ε has smaller product-count. First we note some elementary properties:

PROPOSITION 1. *$\{E_1, \cdots, E_m\}$ is a monotonically nondecreasing path.*

PROPOSITION 2. *$\{E_1, \cdots, E_m\}$ lies entirely within the band $\{(\theta(t) - \varepsilon, \theta(t) + \varepsilon); 0 \le t \le 1\}$ for all large n.*

PROPOSITION 3. *$\{\prod_{j=1}^m N(E_j)/Q(E_j)\}^{1/m} \to 1$ almost surely.*

The last of these propositions follows from Lemma 3.

We are now ready to prove Theorem 1.

PROOF OF THEOREM 1. Consider a monotonic nondecreasing random function $\hat{\xi}_n(t)$, $0 \le t \le 1$, with associated path of squares $\{S_1, \cdots, S_m; S_j \in \mathscr{C}_j\}$, for which

$$\sup_{0\le t\le 1} | \theta(t) - \hat{\xi}_n(t) | > 3\varepsilon \tag{7}$$

for a particular n and realization of $\hat{\xi}_n(t)$. *For* $j = 1, \cdots, m$, associate T_j with S_j in the following manner:

$$T_j = \begin{cases} S_j, & \text{if } Q(S_j) \ge n_j p_j (1 - \delta) \\ \cup_{k=i}^{i+s-1} R_{kj}, & \text{if } Q(S_j) < n_j p_j (1 - \delta) \text{ for some } i, \\ & \text{where } R_{ij}, \cdots, R_{i+s-1,j} \text{ cover } S_j. \end{cases} \tag{8}$$

If $Q(S_j) < n_j p_j (1 - \delta)$, then since $n_j p_j (1 - \delta/2) = (s - 2) Q(R_{kj})$ we may always find s consecutive R_{kj}'s to cover S_j.

Provided $\max_j r_j = O(m)$, we have

$$\{\prod_{j=1}^m N(T_j)/Q(T_j)\}^{1/m} \to 1 \quad \text{almost surely} \tag{9}$$

by Lemma 3. But $p_j > c/m$ uniformly in j for some $c > 0$ by Assumptions 3 and 4. And from (6) we have $r_j/(s_j - 2) = p_j^{-1}(1 - \delta/2)^{-1} = O(m)$ uniformly in j. Thus by choosing s_j to be a large but universally bounded number (in j and n), we obtain (9).

Since $\theta(\cdot)$ is uniformly continuous on [0, 1], there is some interval on which the minimum (Euclidean) distance between $\theta(\cdot)$ and $\hat{\xi}_n(\cdot)$ is no less than 2ε. The length of this interval may be bounded below independently of n and of the realization of $\hat{\xi}_n(\cdot)$. If n is sufficiently large, the path of sets $\{T_1, \cdots, T_m\}$ associated with $\hat{\xi}_n(\cdot)$ will then contain a block $\{T_\alpha, \cdots, T_\beta\}$ of consecutive sets whose minimum distance from $\theta(\cdot)$ is at least ε. Thus for each realization of $\hat{\xi}_n(\cdot)$ satisfying (7) the proportion of $\{T_1, \cdots, T_m\}$ lying at least ε from $\theta(\cdot)$ is at least $(\beta - \alpha)/m$, which may be bounded below by $a > 0$, where a is free of n and the realization.

It follows from Assumptions 3 and 4 that, for large n,

$$Q(T_j) < \left(1 - \frac{\delta}{2}\right) Q(E_j), \quad j = \alpha, \cdots, \beta.$$

Hence,

$$\prod_{j=1}^m Q(T_j) < \left(1 - \frac{\delta}{2}\right)^{\beta-\alpha} \prod_{j=1}^m Q(E_j).$$

If (7) holds infinitely often, then

$$\liminf_{n\to\infty} \{\prod_{j=1}^m N(T_j)/Q(E_j)\}^{1/m} \le \left(1 - \frac{\delta}{2}\right)^a < 1. \tag{10}$$

Using Proposition 3 with (10), we have

$$\liminf_{n\to\infty} \{\prod_{j=1}^m N(T_j)/N(E_j)\}^{1/m} < 1. \tag{11}$$

Recall that $\{\hat{M}_1, \cdots, \hat{M}_m\}$ is the monotonic path of squares with maximum product-count.

So $\prod_{j=1}^{m} N(E_j) \leq \prod_{j=1}^{m} N(\hat{M}_j)$ and $N(S_j) \leq N(T_j), j = 1, \cdots, m$ because T_j covers S_j. Thus from (11),

$$\liminf_{n\to\infty} \{\prod_{j=1}^{m} N(S_j)/N(\hat{M}_j)\}^{1/m} < 1. \tag{12}$$

So the path $\{S_1, \cdots, S_m\}$ associated with $\hat{\xi}_n(\cdot)$ is nonoptimal. Hence $\hat{\theta}(\cdot)$ may satisfy (7) only on a set of probability zero.

5. T finite, X_t continuous. The case of T finite and X_t continuous may be treated by combining the methods of Sections 2 and 4. Let $T = \{1, 2, \cdots, k\}$. Mirroring Section 4, we divide the real line into intervals of the forms $S_i = [i/m, (i + 1)/m], i = 0, \pm 1, \pm 2, \cdots$. Let $\{X_{j1}, \cdots, X_{jn}\}$ be a random sample from the jth population and suppose the samples are independent for $j = 1, \cdots, k$. Let

$$B_{ij} = \sum_{k=1}^{n_j} I_{S_i}(X_{jk})$$

be the frequency of interval S_i in the jth sample. Then apply the algorithm of Section 3 to the matrix of frequencies $\{B_{ij}\}$. The estimator $\theta_n(j)$ is constructed by choosing a point from each interval of the algorithmic path of intervals so that $\theta_n(j) \leq \theta_n(j + 1), j = 1, \cdots, k - 1$.

Consistency is easier to establish than in Section 4. Because the details are elementary modifications of those of Section 4, we omit proofs and merely state the result.

THEOREM 2. *Let $p_j(\cdot)$ be the density of the jth population with respect to Lebesque measure. Suppose*

(a) *For each j there is a unique mode $\theta(j)$: $\forall \varepsilon > 0$, $\exists$ positive δ such that $|\theta(j) - x| > \varepsilon \Rightarrow p_j(\theta(j)) > p_t(x) + \delta$,*

(b) *$p_j(\cdot)$ is continuous at $\theta(j)$ for each j,*

(c) *$\theta(j)$ is a monotone nondecreasing function of j,*

(d) *$\sum_{n=1}^{\infty} n\lambda^{\min\{n_1,\cdots,n_k\}} < \infty$ for all $0 < \lambda < 1$.*

Then $\max_{j=1,\ldots,k} |\theta_n(j) - \theta(j)| \to 0$ almost surely.

The conditions in Theorem 2 are freer than those of Theorem 1 of Section 4. Although equicontinuity and the uniformity of δ follow from (b) and (a) by the finiteness of T, (d) allows considerably more latitude than Assumptions 6 and 7 of Section 4.

It is interesting to consider an alternative scheme based on the modal estimator of Venter rather than that of Chernoff. Choose $r_j(n) \equiv r_j$, $j = 1, \cdots, k$, and divide the jth sample into $n_j - r_j + 1$ intervals at every r_jth order statistic. Apply the algorithm of Section 3 to the matrix of inverse interval lengths. (Counts are no longer used because they are all the same for the cells within a column. Obvious modifications to the algorithm will be necessary because the "cells" in adjacent columns may overlap as well as those in the same column.) The isotonic modal regression $\theta_n(j)$ is chosen from the algorithmic solution path of cells. This method is advantageous with moderate sample sizes because it automatically "conditions" the matrix for the algorithm by forcing the cell entries to be based on a reasonable number of data. Also, consistency and rates of convergence follow immediately for $\sup_{j=1,\ldots,k} |\theta_n(j) - \theta(j)|$ from Sager (1975).

6. Examples and computational results. In order to illustrate our procedures, we performed several small scale Monte Carlo experiments for both continuous and discrete cases. The simulations that we present are for illustration only; although the results that we present here are typical of other runs that we have made, we base no claims concerning the small sample behavior of these estimators on them.

In the case of T finite, P_t discrete, we chose $k = 9$ and four different distributions for P_t:

$$A\colon\ X_t \sim B(10, t/10),$$

$$B\colon\ X_t \sim .90\ B(10, t/10) + .10\ B(10, 1 - t/10),$$

$$C\colon\ X_t \sim .75\ B(10, t/10) + .25\ B(10, 1 - t/10),$$

$$D\colon\ X_t \sim .33\ \delta_t + .67\ B(10, 1 - t/10),$$

where $B(n, p)$ represents the binomial distribution with probability of success p and n trials, and where δ_x represents the probability measure degenerate at x. A single trial consisted of generating $9m$ independent observations: m observations on each of $X_1, \cdots, X_9$. 100 trials were performed for each value of m that we used ($m = 5$, 12, and 25) and each distribution A, B, C, D. For each trial we compared the isotonic modal regression estimator to three competitors in terms of squared-error loss and zero-one loss. The competitors were the isotonic mean (see Barlow, et al, 1972), the nonisotonic vector of sample means, and the estimator which rounds the isotonic mean to the nearest integer (that is the least squares estimator under the constraints of ordered and integer-valued solutions—see Goldstein and Kruskal, 1976).

The results of the experiments are reported in Tables 3 and 4. Note that for each distribution A, B, C, D, the true modal path is (1, 2, 3, 4, 5, 6, 7, 8, 9). In case A the modal path and the path of means coincide, so we expected the modal estimator to do poorly in small samples. In cases B, C, D, the primary mode is contaminated by a secondary mode running in the opposite direction. This kind of contamination may occur when failures are erroneously tallied as successes in the underlying binomial experiment. In case D, the path of means is actually *decreasing*, so the mean estimators should perform poorly. Tables 3 and 4 bear out our expectations. Note that even in the presence of a small amount of contamination, as in experiment B, and even with a sample of size $m = 5$, the isotonic modal estimator is doing quite well compared to the mean estimators, particularly with respect to zero-one loss. This situation becomes more pronounced as the amount of contamination and/or sample size increases. The fact that the average losses incurred by the modal estimator were affected only slightly by the amount of contamination in A, B, C whereas the losses of the mean estimators were highly unstable suggests that the isotonic modal regression may enjoy some robustness.

For the case in which P_t is continuous, seven distributions were studied:

$$E\colon\ X_t \sim N(t, 1),$$

$$F\colon\ X_t \sim .9\ N(t, 1) + .1\ N(10 - t, .5^2),$$

$$G\colon\ X_t \sim .75\ N(t, 1) + .25\ N(10 - t, .5^2),$$

$$H\colon\ X_t \sim .8\ N(t, 1) + .2\ N(t + 3, .5^2),$$

$$I\colon\ X_t \sim .8\ N(t, 1) + .2\ N(t + 3, .75^2),$$

$$J\colon\ X_t \sim .75\ N(t, 1) + .25\ N(t + 3, .75^2), \quad \text{and}$$

$$K\colon\ X_t \sim .75\ N(t, 1) + .25\ N(t + 2, .75^2),$$

where $N(\mu, \sigma^2)$ is the normal distribution with mean μ and variance σ^2. We again chose $k = 9$ ($t = 1, 2, \cdots, 9$); a single trial consisted of $9m$ independent observations—m observations for each t; 100 trials were performed for each value of m (5, 25, 100) and each distribution $E - K$. Three estimators were compared: the isotonic modal regression, the isotonic means, and the (nonisotonic) vector of sample means. To compute the isotonic modal regression, it is necessary to discretize the problem as in Section 4 by constructing

TABLE 3

Squared-error loss, frequency weights. Discrete distributions. Average losses and standard errors based upon 100 repetitions with k = 9 populations and m observations per population.

	$m = 5$	$m = 12$	$m = 20$
A. Isotonic mode	8.75 (.43)	6.26 (.33)	4.82 (.20)
Isotonic mean	2.52 (.11)	1.32 (.06)	.81 (.04)
Least squares	3.11 (.15)	1.78 (.10)	.74 (.09)
Mean	3.19 (.15)	1.43 (.07)	.83 (.04)
B. Isotonic mode	10.04 (.61)	6.31 (.30)	5.26 (.22)
Isotonic mean	6.36 (.36)	5.07 (.30)	3.97 (.19)
Least squares	7.19 (.41)	5.98 (.35)	4.62 (.23)
Mean	9.54 (.51)	6.14 (.41)	4.32 (.21)
C. Isotonic mode	10.63 (.70)	6.77 (.34)	5.79 (.30)
Isotonic mean	18.60 (1.02)	17.04 (.66)	17.00 (.51)
Least squares	19.33 (1.08)	18.10 (.78)	17.99 (.57)
Mean	26.75 (1.65)	19.92 (.81)	18.74 (.60)
D. Isotonic mode	3.84 (.60)	.94 (.34)	.22 (.08)
Isotonic mean	57.02 (.61)	59.85 (.22)	60.14 (.10)
Least squares	57.34 (.71)	59.97 (.35)	59.88 (.15)
Mean	115.09 (2.93)	115.23 (2.04)	108.46 (1.50)
Number of times out of 100 isotonic mode had smaller loss than isotonic mean.			
Experiment A	2	3	0
B	26	32	33
C	78	95	98
D	100	100	100

TABLE 4

Zero-one loss, log frequency weights. Discrete distributions. Average losses and standard errors based upon 100 repetitions with k = 9 populations and m observations per population.

	$m = 5$	$m = 12$	$m = 20$
A. Isotonic mode	.087 (.0046)	.060 (.0031)	.042 (.0021)
Isotonic mean	.041 (.0013)	.030 (.0010)	.022 (.0007)
Least squares	.033 (.0018)	.020 (.0014)	.008 (.0010)
Mean	.048 (.0018)	.031 (.0011)	.023 (.0008)
B. Isotonic mode	.102 (.0067)	.058 (.0031)	.046 (.0024)
Isotonic mean	.113 (.0061)	.110 (.0053)	.094 (.0033)
Least squares	.116 (.0072)	.114 (.0063)	.094 (.0042)
Mean	.149 (.0079)	.124 (.0067)	.099 (.0036)
C. Isotonic mode	.096 (.0056)	.054 (.0028)	.048 (.0026)
Isotonic mean	.270 (.0132)	.273 (.0089)	.284 (.0070)
Least squares	.281 (.0146)	.291 (.0105)	.304 (.0083)
Mean	.300 (.0125)	.297 (.0095)	.304 (.0079)
D. Isotonic mode	.384 (.0413)	.061 (.0174)	.009 (.0023)
Isotonic mean	1.116 (.0128)	1.071 (.0040)	1.070 (.0025)
Least squares	1.177 (.0139)	1.124 (.0042)	1.111 (.0021)
Mean	.780 (.0228)	.684 (.0143)	.702 (.0108)
Number of times out of 100 isotonic mode had smaller loss than isotonic mean.			
Experiment A	17	10	19
B	58	82	91
C	92	100	100
D	91	98	100

TABLE 5

Squared-error loss. Continuous distributions. Average losses and standard errors based upon 100 repetitions with $k = 9$ populations and m observations per population.

	$m = 5$	$m = 12$	$m = 20$
E. Isotonic mode	5.93 (.24)	2.50 (.10)	1.44 (.07)
Isotonic mean	1.62 (.07)	.35 (.02)	.09 (.00)
Mean	1.75 (.08)	.35 (.02)	.09 (.00)
F. Isotonic mode	6.55 (.28)	2.60 (.12)	1.62 (.07)
Isotonic mean	6.13 (.41)	3.61 (.16)	2.66 (.07)
Mean	8.33 (.58)	3.75 (.18)	2.66 (.07)
G. Isotonic mode	7.58 (.37)	3.16 (.15)	1.90 (.09)
Isotonic mean	16.90 (.97)	16.38 (.49)	15.70 (.27)
Mean	22.93 (1.52)	17.22 (.54)	15.80 (.28)
H. Isotonic mode	7.83 (.44)	2.94 (.13)	1.81 (.08)
Isotonic mean	6.52 (.32)	3.98 (.12)	3.33 (.06)
Mean	7.63 (.37)	4.00 (.12)	3.33 (.06)
I. Isotonic mode	7.69 (.44)	3.24 (.23)	2.00 (.12)
Isotonic mean	6.65 (.30)	4.19 (.12)	3.41 (.05)
Mean	7.44 (.31)	4.21 (.13)	3.41 (.05)
J. Isotonic mode	10.63 (.82)	3.26 (.18)	2.14 (.10)
Isotonic mean	9.15 (.42)	5.92 (.16)	5.31 (.07)
Mean	10.19 (.46)	5.95 (.16)	5.31 (.07)
K. Isotonic mode	7.42 (.46)	3.59 (.19)	2.43 (.12)
Isotonic mean	4.42 (.18)	2.43 (.08)	1.97 (.03)
Mean	4.87 (.19)	2.43 (.08)	1.97 (.03)
Number of times out of 100 isotonic mode had smaller loss than isotonic mean.			
Experiment E	0	0	0
F	48	74	92
G	84	100	100
H	42	74	95
I	44	70	90
J	46	86	100
K	20	29	45

a grid. Rather arbitrarily, we spaced grid lines evenly 1.84, .62, and .28 units apart (beginning at -4) for $m = 5$, 25, 100, respectively. The modal path was estimated by the midpoints of the cells selected by our algorithm from the grid-matrix of frequency counts.

The results are presented in Tables 5 and 6. Note that for most of the distributions $E - K$, squared-error loss and log-likelihood ratio loss (the latter being the natural analog of our zero-one loss function) will perform similarly, since they are proportional to each other in case E and not vastly different in the other cases. Distributions E, F, G are the continuous analogues of A, B, C. We expected the isotonic mode to do poorly in E. But even with a small contamination and small sample size (case B, $m = 5$), the modal estimator is almost as good as the isotonic mean and improves rapidly thereafter. In distributions $H - K$ the contamination comes from a string of secondary modes running "parallel" to the primary modes rather than opposite the primary trend as in E, F, G. (Actually, K is unimodal with a large "shoulder" on the right.) Although this sort of contamination appears on the surface to be less severe than for E, F, G, the pattern is the same for H, I, J: for even the smallest sample sizes, the isotonic modal estimator does as well as the mean estimators and for moderate and large samples does very much better.

Copies of the programs used in the simulations are available from the authors.

TABLE 6

Log-probability ratio loss. Continuous distributions. Average losses and standard errors based upon 100 repetitions with k = 9 populations and m observations per population.

	$m = 5$	$m = 12$	$m = 20$
E. Isotonic mode	.329 (.0132)	.139 (.0053)	.080 (.0036)
Isotonic mean	.090 (.0039)	.019 (.0009)	.005 (.0002)
Mean	.097 (.0044)	.009 (.0009)	.005 (.0002)
F. Isotonic mode	.359 (.0138)	.148 (.0066)	.093 (.0040)
Isotonic mean	.342 (.0225)	.201 (.0090)	.148 (.0038)
Mean	.463 (.0317)	.209 (.0102)	.148 (.0039)
G. Isotonic mode	.386 (.0153)	.177 (.0072)	.110 (.0049)
Isotonic mean	.927 (.0532)	.908 (.0271)	.871 (.0150)
Mean	1.097 (.0598)	.954 (.0299)	.876 (.0153)
H. Isotonic mode	.388 (.0169)	.163 (.0070)	.100 (.0043)
Isotonic mean	.337 (.0146)	.221 (.0066)	.185 (.0032)
Mean	.360 (.0137)	.221 (.0067)	.221 (.0032)
I. Isotonic mode	.371 (.0143)	.166 (.0074)	.105 (.0044)
Isotonic mean	.328 (.0129)	.226 (.0064)	.187 (.0029)
Mean	.351 (.0125)	.227 (.0065)	.187 (.0029)
J. Isotonic mode	.394 (.0144)	.170 (.0081)	.116 (.0051)
Isotonic mean	.399 (.0136)	.312 (.0076)	.287 (.0036)
Mean	.407 (.0122)	.313 (.0076)	.287 (.0036)
K. Isotonic mode	.251 (.0013)	.125 (.0049)	.088 (.0036)
Isotonic mean	.129 (.0040)	.088 (.0026)	.077 (.0012)
Mean	.139 (.0041)	.088 (.0026)	.077 (.0012)
Number of times out of 100 isotonic mode had smaller loss than isotonic mean.			
Experiment E	0	0	0
F	46	72	90
G	88	100	100
H	46	74	95
I	43	70	92
J	48	87	100
K	8	23	49

REFERENCES

AHO, ALFRED V., HOPCROFT, JOHN E., and ULLMAN, JEFFREY D. (1974). *The Design and Analysis of Computer Algorithms.* Addison-Wesley, Reading.

AHRENS, J., DIETER, U., and GRUBE, A. (1970). Pseudo-random numbers: A new proposal for the choice of multiplicators. *Computing* **6** 121–138.

BARLOW, R. E., BARTHOLOMEW, D. J., BREMNER, J. M., BRUNK, H. D. (1972). *Statistical Inference under Order Restrictions.* Wiley, New York.

BRUNK, H. D. (1955). Maximum likelihood estimates of monotone parameters. *Ann. Math. Statist.* **26** 607–616.

BRUNK, H. D. (1970). Estimation of isotonic regression. *Nonparametric Techniques in Statistical Inference*, 177–195. Cambridge University Press.

CACOULLOS, THEOPHILES (1966). Estimation of a multivariate density. *Ann. Instit. Statist. Math.* **18** 178–189.

CHERNOFF, HERMAN (1964). Estimation of the mode. *Ann. Instit. Statist. Math.* **16** 31–41.

CRYER, J. D., ROBERTSON, TIM, WRIGHT, F. T. and CASADY, R. J. (1972). Monotone median regression. *Ann. Math. Statist.* **43** 1459–1469.

FERGUSON, THOMAS S. (1967). *Mathematical Statistics: A Decision Theoretic Approach.* Academic, New York.

GOLDSTEIN, A. J. and KRUSKAL, J. B. (1976). Least-squares fitting by monotonic functions having integer values. *J. Amer. Statist. Assoc.* **71** 370–373.

HANSON, D. L., PLEDGER, GORDON, and WRIGHT, F. T. (1973). On consistency in monotonic regression. *Ann. Statist.* **1** 401–421.

LEVENSHTEIN, V. I. (1965; English, 1966). Binary codes capable of correction deletions, insertions, and reversals. *Cybernetics and Control Theory* **10(8)** 707–710. *Doklady Akademii Nauk SSSR* **163(4)** 845–848.

PARZEN, E. (1962). On the estimation of a probability density function and the mode. *Ann. Math. Statist.* **33** 1065–1076.

ROBERTSON, TIM, and WRIGHT, F. T. (1973). Multiple isotonic median regression. *Ann. Statist.* **1** 422–432.

ROBERTSON, TIM, and WRIGHT, F. T. (1974). A norm reducing property for isotonized Cauchy mean value functions. *Ann. Statist.* **2** 1302–1307.

ROBERTSON, TIM, and WRIGHT, F. T. (1975). Consistency in generalized isotonic regression. *Ann. Statist.* **3** 350–362.

ROBERTSON, T. J., and WRIGHT, F. T. (1980). Algorithms in order restricted statistical inference and the Cauchy mean value property. *Ann. Statist.* **8** 645–651.

SAGER, THOMAS W. (1975). Consistency in nonparametric estimation of the mode. *Ann. Statist.* **3** 698–706.

VENTER, J. H. (1967). On estimation of the mode. *Ann. Math. Statist.* **38** 1446–1455.

DEPARTMENT OF GENERAL BUSINESS
UNIVERSITY OF TEXAS AT AUSTIN
AUSTIN, TEXAS 78712

DEPARTMENT OF STATISTICS
UNIVERSITY OF CHICAGO
CHICAGO, ILLINOIS 60637