The Annals of Statistics
1974, Vol. 2, No. 6, 1302–1307

A NORM REDUCING PROPERTY FOR ISOTONIZED CAUCHY MEAN VALUE FUNCTIONS[1]

By Tim Robertson and F. T. Wright

The University of Iowa

We consider functions $\alpha(\cdot)$ and $\hat{\alpha}(\cdot)$ on a finite set S which correspond to a function $M(\cdot)$ on the nonempty subsets of S which has the Cauchy mean value property (i.e., $M(A + B)$ is between $M(A)$ and $M(B)$ whenever A and B are nonempty disjoint subsets of S). $\hat{\alpha}(\cdot)$ is isotone with respect to a partial ordering on S and is equal to $\alpha(\cdot)$ when $\alpha(\cdot)$ is isotone. It is shown that $\hat{\alpha}(\cdot)$ has the following norm reducing property: $\max_{s \in S} |\hat{\alpha}(s) - \theta(s)| \leqq \max_{s \in S} |\alpha(s) - \theta(s)|$ for all isotone $\theta(\cdot)$. Computation algorithms for $\hat{\alpha}(\cdot)$ are discussed and the norm reducing property is shown to give consistency results in several isotonic regression problems.

Consider a finite set $S = \{s_1, s_2, \cdots, s_k\}$. A function $M(\cdot)$ whose domain is the collection of nonempty subsets of S is said to have the "Cauchy mean value property" provided $M(A + B)$ is between (not necessarily strictly) $M(A)$ and $M(B)$ whenever A and B are nonempty disjoint subsets of S. Suppose $\ll$ is a partial order on S and define the complete lattice, $\mathscr{L}$, of subsets of S by: $L \in \mathscr{L}$ if and only if $s_i \ll s_j$ and $s_i \in L$ imply that $s_j \in L$. We shall refer to members of $\mathscr{L}$ as upper layers and, in order to simplify some of the notation we use the symbol L exclusively to denote upper layers.

Let $R(\mathscr{L})$ denote the collection of $\mathscr{L}$ measurable functions on S (a function θ which maps S into the real numbers is $\mathscr{L}$ measurable if and only if $\{s_i : \theta(s_i) \geqq a\} \in \mathscr{L}$ for all real a, or equivalently, θ is isotone with respect to $\ll$). Define the functions α and $\hat{\alpha}$ on S by $\alpha(s_i) = M(\{s_i\})$ and $\hat{\alpha}(s_i) = \max_{L \ni s_i} \min_{L' \not\ni s_i} M(L - L')$. It is easy to see that $\hat{\alpha}(\cdot)$ is $\mathscr{L}$-measurable.

Several estimates which have been proposed in isotonic regression problems fall within the framework described above. Generally, S denotes a set indexing a collection of distributions and we have a random sample from each of these distributions. Suppose the sample items from the distribution associated with s_i are denoted by X_{ij}; $j = 1, 2, \cdots, n_i$. Typically, we wish to estimate a real-valued function $\theta(\cdot)$ defined on S such that the value of $\theta(\cdot)$ at a point s_i of S is a characteristic of the distribution at s_i. If one has no prior knowledge about $\theta(\cdot)$ then one might use an $\alpha(\cdot)$ which corresponds to a mean value function M, which, in turn, would generally be a function of the sample items (recall that $\alpha(s_i) = M(\{s_i\})$). However, if it is known that $\theta \in R(\mathscr{L})$ then we might want to modify our estimate to ensure that it is also in $R(\mathscr{L})$. $\hat{\alpha}(\cdot)$ is one

Received July 1973.

[1] This research was supported by the National Science Foundation under Grant GP-30966X.

AMS 1970 *subject classifications*. Primary 62G05; Secondary 60F15.

Key words and phrases. Cauchy mean value functions, isotonic estimation, norm reducing, extrema.

possible candidate. ($M(A)$ is usually a function of the sample items from distributions associated with points in A.)

The following theorem can be interpreted as saying that isotonizing Cauchy mean value functions reduces the L_∞ distance to $R(\mathscr{L})$.

THEOREM 1. *For each* $\theta \in R(\mathscr{L})$,

$$\max_{1\leq i\leq k} |\hat{\alpha}(s_i) - \theta(s_i)| \leqq \max_{1\leq i\leq k} |\alpha(s_i) - \theta(s_i)|\,.$$

PROOF. Fix s_i and $\theta \in R(\mathscr{L})$ and let $L_0 = \{s_j;\, \theta(s_j) \geqq \theta(s_i)\}$ and $L_0' = \{s_j;\, \theta(s_j) > \theta(s_i)\}$. Using the mean value property and obvious properties of maxima, we can write:

$$\begin{aligned}\hat{\alpha}(s_i) - \theta(s_i) &\leqq \max_{L\ni s_i} M(L - L_0') - \theta(s_i)\\ &\leqq \max_{L\ni s_i} \max_{s_j\in L-L_0'} [\alpha(s_j) - \theta(s_i)]\\ &\leqq \max_{L\ni s_i} \max_{s_j\in L-L_0'} [\alpha(s_j) - \theta(s_j)]\\ &\leqq \max_{1\leq j\leq k} |\alpha(s_j) - \theta(s_j)|\,.\end{aligned}$$

Similarly, $\hat{\alpha}(s_i) - \theta(s_i) \geqq \min_{L'\not\ni s_i} \min_{s_j\in L_0-L'} [\alpha(s_j) - \theta(s_j)] \geqq -\max_{1\leq j\leq k} |\alpha(s_j) - \theta(s_j)|$. Since s_i was arbitrary the proof is completed.

It is argued in Robertson and Waltman (1968) that if $\ll$ is a linear order then there exists a measure $\mu(\cdot)$ on 2^S such that $\mu(\{s_i\}) > 0$ for each i and $\hat{\alpha} = E_\mu(\alpha \mid \mathscr{L})$. It would be of interest to know if such a result holds when $\ll$ is simply a partial order. For if so, Theorem 1 could be obtained by the following, less direct, approach. If $\hat{\alpha} = E_\mu(\alpha \mid \mathscr{L})$ then Remark 3.3 of Brunk (1965) gives

$$[\textstyle\sum_{i=1}^k |\hat{\alpha}(s_i) - \theta(s_i)|^p \mu(\{s_i\})]^{1/p} \leqq [\sum_{i=1}^k |\alpha(s_i) - \theta(s_i)|^p \mu(\{s_i\})]^{1/p} \tag{1}$$

and Theorem 1 follows by letting $p \to \infty$.

Malmgren (1972) points out that (1) does not hold for all $\theta \in R(\mathscr{L})$ when $p = 1$ and $M(\{s_i\})$ is the median of the observations from the distribution at s_i.

Ubhaya (1971) studied isotonic functions which approximate a given function and which are closest in the L_∞ distance to the given function. He shows (see Theorem 2 of Chapter 3) that a particular member of this class of approximating functions is closer in the L_∞ distance to any isotonic function than is the given function. Theorem 1 is more general in the sense that it applies to all Cauchy mean value functions.

Since versions of Cauchy mean value functions arise in estimation problems, computation algorithms for $\hat{\alpha}$ are of interest. We mention three of these which can be argued in this general setting. Define $\bar{\alpha}(s_i) = \min_{L'\not\ni s_i} \max_{L\ni s_i} M(L - L')$ and $\alpha^*(s_i) = \max_{L\ni s_i} \min_{L'\in\mathscr{V}(L)} M(L - L')$ where $\mathscr{V}(L)$ denotes the collection of proper subsets of L which are upper layers. Let $L_1 = S$ and choose L_2 to be the smallest member of $\mathscr{V}(L_1)$ for which $M(L_1 - L_2) = \min_{L'\in\mathscr{V}(L_1)} M(L_1 - L')$. This can be seen to be possible by considering the argument given for Remark 2.3 of Robertson and Wright (1973) and the comments after the remark. Continuing, we obtain $L_1 \supset L_2 \supset \cdots \supset L_H$ with $M(L_i - L_{i+1}) = \min_{L'\in\mathscr{V}(L_i)} M(L_i - L')$

for $i = 1, 2, \cdots, H$ and $L_{H+1} = \emptyset$ (since S is finite and $L_{i+1} \in \mathscr{V}(L_i)$ requires that $L_i - L_{i+1} = \emptyset$ this process must terminate). Modifying slightly the arguments given for Theorem 2.4 and Corollary 2.5 of the paper mentioned above, it can be shown that $\hat{\alpha} = \tilde{\alpha} = \alpha^*$ and if $s_j \in L_i - L_{i+1}$ then $\hat{\alpha}(s_j) = M(L_i - L_{i+1})$.

We conclude with some examples to indicate the generality of these results and some applications of Theorem 1.

EXAMPLE 1. Ayer, Brunk, Ewing, Reid and Silverman (1955) and independently, van Eeden (1956, 1957) considered estimating the function $\theta(s_i) = p_i$ when the distribution at s_i was taken to be Bernoulli with parameter p_i and $\theta \in R(\mathscr{L})$. If, in this case, one sets $M(A) = \sum_A \sum_{j=1}^{n_i} X_{ij} / \sum_A n_i$, where $\sum_A$ is interpreted as $\sum_{\{i : s_i \in A\}}$, it is clear that M is a Cauchy mean value function and the corresponding $\hat{\alpha}$ is the estimate studied in the above-mentioned papers.

Consistency results are immediate. The Borel strong law of large numbers and Theorem 1 show that $\max_{1 \leq i \leq k} |\hat{\alpha}(s_i) - p_i|$ converges almost surely to zero as $\min_{1 \leq i \leq k} n_i \to \infty$. In many cases Theorem 1 can be used to obtain rates of weak convergence for $\max_{1 \leq i \leq k} |\hat{\alpha}(s_i) - \theta(s_i)|$ to zero, that is, one could obtain rates at which $P[\max_{1 \leq i \leq k} |\hat{\alpha}(s_i) - \theta(s_i)| \geqq \varepsilon]$ converges to zero by combining Theorem 1 with known rates of weak convergence for the estimators $\alpha(s_i)$ to the true $\theta(s_i)$. Also Theorem 1 would provide almost sure rates of convergence in situations where such rates are known for the estimators $\alpha(s_i)$. We have chosen to illustrate the latter statement in this example because the results seem to be nicest here.

THEOREM 2. *If* $0 < p_i < 1$ *for* $i = 1, 2, \cdots, k$ *and* $n_1 = n_2 = \cdots = n_k = n$ *then with* $\emptyset(n) = (n / \log \log n)^{\frac{1}{2}}$

$$P[\limsup_{n \to \infty} \emptyset(n) \max_{1 \leq i \leq k} |\hat{\alpha}(s_i) - p_i| = \max_{1 \leq i \leq k} (2p_i(1 - p_i))^{\frac{1}{2}}] = 1 .$$

PROOF. Kolmogorov's law of the iterated logarithm and Theorem 1 yield

$$P[\limsup_{n \to \infty} \emptyset(n) \max_{1 \leq i \leq k} |\hat{\alpha}(s_i) - p_i| \leqq \max_{1 \leq i \leq k} (2p_i(1 - p_i))^{\frac{1}{2}}] = 1 . \tag{2}$$

Let p be chosen so that $(p(1-p))^{\frac{1}{2}} = \max_{1 \leq i \leq k} (p_i(1-p_i))^{\frac{1}{2}}$ and consider the upper layers $L_1(p) = \{s_i : p_i \geqq p\}$ and $L_2(p) = \{s_i : p_i > p\}$. Clearly, $\{s_i : p_i = p\} = L_1(p) - L_2(p)$ and it can be shown by induction that there is a point $s_{i_0} \in L_1(p) - L_2(p)$ and an upper layer $L_0 \subset L_1(p)$ such that $L_0 \cap (L_1(p) - L_2(p)) = \{s_{i_0}\}$. Hence $\hat{\alpha}(s_{i_0}) \geqq \min_{L' \not\ni s_{i_0}} M(L_0 - L')$. However, by the strong law of large numbers there is a set of ω's with probability one such that for $n \geqq N(\omega)$, $\alpha(s_i) > \alpha(s_{i_0})$ for all i for which $s_i \in L_2(p)$, and so using the averaging property, $\hat{\alpha}(s_{i_0}) \geqq \alpha(s_{i_0})$ for $n \geqq N(\omega)$ and ω in the given sure event. Hence

$$\begin{aligned} P[\limsup_{n \to \infty} \emptyset(n) \max_{1 \leq i \leq k} |\hat{\alpha}(s_i) - p_i| &\geqq (2p(1 - p))^{\frac{1}{2}}] \\ &\geqq P[\limsup_{n \to \infty} \emptyset(n)(\hat{\alpha}(s_{i_0}) - p_{i_0}) \geqq (2p(1 - p)^{\frac{1}{2}}] \\ &\geqq P[\limsup_{n \to \infty} \emptyset(n)(\alpha(s_{i_0}) - p_{i_0}) \geqq (2p(1 - p)^{\frac{1}{2}}] = 1 . \end{aligned}$$

The proof is completed.

We comment that results like (2) could be obtained in any situation where M is a Cauchy mean value function and laws of the iterated logarithm are known for the estimators $\alpha(s_i)$. (i.e., $P[\limsup_{n\to\infty} \varnothing(n)|\alpha(s_i) - \theta(s_i)| = k_i] = 1$ for $i = 1, 2, \cdots, k$.) However, the argument given to obtain the other inequality required $k_i = k_j$ if $\theta(s_i) = \theta(s_j)$. We also note that if, instead of assuming common sample sizes, we assume there is a positive constant A for which $n_i \geqq AN$ for $i = 1, 2, \cdots, k$ where $N = \sum_{i=1}^k n_i$ then we can obtain, reasoning as above,

$$P[c \leqq \limsup_{N\to\infty} \varnothing(N) \max_{1\leqq i\leqq k} |\hat{\alpha}(s_i) - p_i| \leqq A^{-\frac{1}{2}}c] = 1$$

where $c = \max_{1\leqq i\leqq k} (2p_i(1-p_i))^{\frac{1}{2}}$.

EXAMPLE 2. The isotonic regression problem that has received most attention in the literature is that of estimating an isotonic mean function. (i.e., $\theta(s_i) = \mu_i$ is the mean of the distribution at s_i and $\theta \in R(\mathscr{L})$). For a detailed bibliography of this and related problems see Barlow and Brunk (1972). The solutions presented in this situation are $\hat{\alpha}$'s corresponding to $M(A) = \sum_A w_i \sum_{j=1}^{n_i} X_{ij} / \sum_A w_i \cdot n_i$ where w_i are positive weights and $\sum_A$ is interpreted as before.

If $X_{ij} - \mu_i$ has the same distribution for $j = 1, \cdots, n_i$ and $i = 1, 2, \cdots, k$ and the law of the iterated logarithm holds for each $\alpha(s_i)$ then a law of the iterated logarithm like Theorem 2 could be established for $\max_{1\leqq i\leqq k} |\hat{\alpha}(s_i) - \mu_i|$ in this case.

EXAMPLE 3. Robertson and Waltman (1968), Cryer, Robertson, Wright and Casady (1972) and Robertson and Wright (1973) have considered the problem of estimating an isotonic regression function with $\hat{\alpha}(\cdot)$ when $M(A)$ is the median of the observations at distributions associated with points in A. For $A \subset S$ define the weighted empirical distribution function $F_A(x) = \sum_A w_i \sum_{j=1}^{n_i} I_{(-\infty,x]}(X_{ij})$ with $I(X_{ij=x}) / \sum_A w_i \cdot n_i$ where the w_i's are positive and $M(A) = \min\{x; F_A(x) \geqq \frac{1}{2}\}$. The argument given for Remark 4.1 of Robertson and Wright (1973) shows that $M(A)$ is a Cauchy mean value function and the corresponding $\hat{\alpha}$ is the estimator proposed in Section 4 of that paper when the w_i's are all taken to be 1. The estimator obtained when the weights are not equal is of interest because the resulting $\hat{\alpha}$ minimizes $\sum_{i=1}^k w_i \sum_{j=1}^{n_i} |X_{ij} - \theta(s_i)|$ among all $\theta \in R(\mathscr{L})$. This would appear to follow from Example 3.8 of Brunk and Johansen (1970) by choosing $\psi(i, a)$, γ and the μ_i's properly. Another approach is to observe that $f(\theta) = \sum_A w_i \sum_{j=1}^{n_i} |x_{ij} - \theta|$ is minimized by choosing θ to be the $M(A)$ discussed in this example (this is equivalent to minimizing $\int |x - \theta| \, dF_A(x)$); to note that $f(\theta)$ is non-increasing for $\theta \leqq M(A)$ and non-decreasing for $\theta \geqq M(A)$ and then to apply the arguments given for Lemma 2.6 and Theorem 2.7 of Robertson and Wright (1973). Since $\alpha(s_i)$ is an ordinary sample median and much is known about consistency properties for such estimators, Theorem 1 could be used to obtain consistency results for $\hat{\alpha}$ in this case.

Percentiles other than the median could be dealt with as above.

EXAMPLE 4. For our last example of a Cauchy mean value function we take

$M(A)$ to be the midrange of the set $\{X_{ij} : j = 1, 2, \cdots, n_i$ with i such that $s_i \in A\}$. (The midrange is the average of the smallest and largest items.) Barlow and Ubhaya (1972) and Ubhaya (1971) have characterized all $\theta \in R(\mathscr{L})$ which minimize $\max_{1 \leq j \leq k} \max_{1 \leq j \leq n_i} |X_{ij} - \theta(s_i)|$ when $n_i \equiv 1$; however, the arguments given in Chapter 3 of the latter reference for Lemma 2 and Theorem 3 of Section 4 show that the $\hat{\alpha}$ corresponding to the Cauchy mean value function considered in this example minimizes the above objective function when the n_i's are not necessarily one. The solution to this minimization problem need not be unique, but the solution presented in this example would seem to be of interest because Theorem 1 can be combined with known consistency properties of the midrange to obtain consistency results for $\hat{\alpha}$. Consistency results for the midrange are immediate if the underlying distribution is symmetric and bounded. Other cases are discussed in Barndorff-Nielsen (1963).

EXAMPLE 5. In our last example we consider a function M which does not have the Cauchy mean value property to give some indication of the importance of this property to the results discussed here. Let $M(A)$ be the mode of the set of observations X_{ij}; $j = 1, 2, \cdots, n_i$ and i with $s_i \in A$.

Let S consist of two distinct points $s_1 \ll s_2$ and suppose we have observations $X_{ij} = 0$; $i, j = 1, 2$, $X_{1j} = 1$; $j = 3, 4, 5$, and $X_{2j} = 2$; $j = 3, 4, 5$, then $\alpha(s_i) = 1$, $\alpha(s_2) = 2$, $\hat{\alpha}(s_1) = 0$ and $\hat{\alpha}(s_2) = 2$. Taking $\theta = \alpha$, we have an example to show that Theorem 1 does not hold for this choice of M. Furthermore, examples can be constructed on a linearly ordered set $S = \{s_1, s_2, s_3\}$ to show that $\hat{\alpha} \neq \bar{\alpha}$ and $\hat{\alpha} \neq \alpha^*$.

With S as above let the distributions at s_1 and s_2 both be discrete; let X_{ij} be independent random variables for $i = 1, 2$ and $j = 1, 2, \cdots$; for each i, let X_{1i} have the distribution at s_1, say $P[X_{11} = -\frac{1}{2}] = P[X_{11} = -\frac{3}{2}] = \frac{2}{7}$ and $P[X_{11} = 0] = \frac{3}{7}$; and for each i let X_{2i} have the distribution at s_2 which we determine by letting $X_{21} - 1$ and X_{11} have the same distribution. The function $\theta(s_i)$ equal to the mode of the distribution at s_i is isotone and $X_{ij} - \theta(s_i)$ are independent and identically distributed. However, using the strong law of large numbers one can show that for "almost all" ω points in the underlying probability space $\hat{\alpha}(s_1) = -\frac{1}{2} \neq \theta(s_1)$ for n sufficiently large where $\hat{\alpha}$ is based on X_{ij}; $j = 1, 2, \cdots, n$ and $i = 1, 2$.

REFERENCES

[1] AYER, M., BRUNK, H. D., EWING, G. M., REID, W. T. and SILVERMAN, E. (1955). An empirical distribution function for sampling with incomplete information. *Ann. Math. Statist.* **26** 641-647.

[2] BARLOW, R. E. and BRUNK, H. D. (1972). The isotonic regression problem and its dual. *J. Amer. Statist. Assoc.* **67** 140-147.

[3] BARLOW, R. E. and UBHAYA, VASANT A. Isotonic approximation. *Proceedings of the Symposium on Optimizing Methods in Statistics*, Academic Press Inc., New York. To appear.

[4] BARNDORFF-NIELSON, OLE (1963). On the limit behaviour of extreme order statistics. *Ann. Math. Statist.* **34** 992-1002.

[5] BRUNK, H. D. (1965). Conditional expectations given a σ-lattice and applications. *Ann. Math. Statist.* **5** 1339–1349.

[6] BRUNK, H. D. and JOHANSEN, S. (1970). A generalized Radon–Nikodym derivative. *Pacific J. Math.* **34** 585–617.

[7] CASADY, ROBERT J. (1972). Monotone percentile regression estimation: large sample and other properties. Ph. D. thesis, Univ. of Iowa.

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

[9] MALMGREN, EDWARD (1972). Contributions to the estimation of ordered parameters. Ph. D. thesis, Univ. of Iowa.

[10] ROBERTSON, TIM and WALTMAN, PAUL (1968). On estimating monotone parameters. *Ann. Math. Statist.* **39** 1030–1039.

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

[12] UBHAYA, VASANT A. (1971). Isotonic and convex optimization. Technical Report, Operations Research Center, University of California, Berkeley.

[13] VAN EEDEN, C. (1956). Maximum likelihood estimation of ordered probabilities. *Indag. Math.* **18** 444–445.

[14] VAN EEDEN, C. (1957). Maximum likelihood estimation of partially or completely ordered parameters, I and II. *Indag. Math.* **19** 201–211.

DEPARTMENT OF STATISTICS
UNIVERSITY OF IOWA
IOWA CITY, IOWA 52242