Estimation of a Convex Density Contour in Two Dimensions

J. A. Hartigan

Journal of the American Statistical Association, Vol. 82, No. 397 (Mar., 1987), 267-270.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28198703%2982%3A397%3C267%3AEOACDC%3E2.0.CO%3B2-G

Journal of the American Statistical Association is currently published by American Statistical Association.

Estimation of a Convex Density Contour in Two Dimensions

J. A. HARTIGAN*

If a density in two dimensions has a convex contour containing probability α, the contour may be estimated from a sample by finding the convex polygon of smallest area containing a proportion α of the sample points. An algorithm for finding a particular contour is given that takes $O(n^2)$ space and $O(n^3)$ time for n sample points.

1. INTRODUCTION

Chernoff (1964) proposed estimating the mode of a univariate distribution by the midpoint of the interval of length $2a$ containing a maximum number of sample points, where the interval length decreases with n. Venter (1967) varied the procedure slightly by using the midpoint of the interval of minimum length containing a specified fraction of sample points. Robertson and Cryer (1974) obtained a nested sequence of intervals, each one being the shortest subinterval containing a certain proportion of the sample points in its predecessor.

Sager (1979) generalized this approach to higher dimensions by constructing a nested sequence of convex sets, each of which is the smallest volume convex set containing a certain proportion of the sample points in its predecessor. Eddy and Hartigan (1977) proposed using the boundary of the convex set of smallest volume containing probability α as an estimate of the population contour containing probability α when the population contour was a convex boundary.

Convex contour estimation is an alternative to kernel density estimation that avoids the mysterious problem of specifying kernel size and shape; indeed the minimal convex set containing, say, $n^{5/7}$ points might itself be used as a kernel. Convex contour estimation does require that the population contours be convex, an assumption that might be violated in interesting ways—for example, the distribution might be bimodal. If it is assumed only that the population contours are *unions* of one or two convex boundaries, allowing bimodality, then the contours might be estimated by the boundaries of unions of two convex sets of minimum total volume, containing a given proportion of the sample points.

A test for bimodality may then be constructed by considering the number of points in the smaller of the two convex sets at each proportion of sample points, comparing it with the number of points to be expected in a unimodal population.

This program requires an efficient method of finding the minimal volume convex set containing a given number of points in more dimensions than one. The convex hull of n points may be found in $O(n \log n)$ operations (see, e.g., Eddy 1977). The number of convex subsets of n points in two dimensions can be as many as 2^n (points on a circle), so algorithms for searching over the convex subsets must avoid looking at them all.

The c contour of a density f is the boundary of the set $S_0 = \{x \mid f(x) \geq c\}$. This set is the largest set S maximizing $(P - \lambda)S$, where

$$PS = \int_S f(x)\,dx, \qquad \lambda S = c \int_S dx.$$

If P_n denotes the empirical distribution of a sample $X_1, \ldots, X_n$ from f, S_0 is not estimated by the set S_n maximizing $(P_n - \lambda)S$, because the optimal S_n is just the set of all sample points. It is necessary to reduce the size of the class of possible sets S by requiring it to consist of convex sets, or unions of a few convex sets, or otherwise. In this article, S_n will be chosen to maximize $(P_n - \lambda)S$ over the class of closed convex sets. An algorithm is given for finding S_n for a particular level c in $O(n^3)$ steps. The boundary of S_n is shown consistent for the boundary of S_0; it is speculated that the maximum error then is $O_p(\log n/n)^{2/7}$.

2. OPTIMAL CONVEX CONTOURS FOR P DISCRETE

Assume that P is carried by n points and that it is desired to find a closed convex polygon S_0 maximizing $QS = PS - c \int_S dx$. The only possible vertices of S_0 are the atoms of P, since otherwise the area of S_0 may be reduced without changing PS.

First find the optimal polygon for each choice of a particular atom as its leftmost vertex. Let this atom be numbered 1, and let the atoms not to its left be numbered 2, 3, . . . , m. The coordinates of the ith atom are denoted by x_i, and the line segment $\alpha x_i + (1 - \alpha)x_j$ $(0 \leq \alpha \leq 1)$ is written as $[i, j]$. Assume that the atoms 1, . . . , m are ordered so that the segments $[1, i]$ move counterclockwise as i increases and so that $i \leq j$ if $i \in [1, j]$.

Polygons will be built up from triangles for $1 < i < j \leq m$; $\triangle_{ij}$ is the convex hull of $(1, i, j)$ excluding $[1, i]$. Note that the segment $[1, i]$ is excluded from $\triangle_{ij}$ in order to combine triangles without overlap.

The quadrilateral with vertices at 1, i, j, k for $i < j < k \leq m$ is convex if

$$D_{ijk} = \begin{vmatrix} x_i' & 1 \\ x_j' & 1 \\ x_k' & 1 \end{vmatrix} \geq 0 .$$

Theorem. Let Q_{jk} $(1 < j < k \leq m)$ be the maximum value of Q among closed convex polygons with successive

* J. A. Hartigan is Professor, Department of Statistics, Yale University, New Haven, CT 06520. Research for this article was supported in part by National Science Foundation Grant DCR-8401636. David Pollard helped with the asymptotics.

counterclockwise vertices j, k, 1. Let Q_{1j} be the value of Q on the line segment $[1, j]$. Then $Q_{jk} = Q_{ij} + Q(\triangle_{jk})$, where $i = N(k, j)$ is chosen to maximize Q_{ij} over vertices i with $i < j$, $D_{ijk} \geq 0$. One optimal polygon with leftmost vertex 1 has vertices $i_1, i_2, \ldots, i_r = 1$, where either $r = 1$ or $Q_{i_2 i_1} = \max_{1 \leq j < k} Q_{jk}$, $i_3 = N(i_1, i_2)$, $i_4 = N(i_2, i_3)$, $\ldots, 1 = i_r = N(i_{r-2}, i_{r-1})$.

Proof. Let C_{jk} be an arbitrary closed convex polygon with successive vertices j, k, 1.

Let $C_{ij} = C_{jk} - \triangle_{jk}$. If C_{jk} has only three distinct vertices, $i = 1$ and $C_{ij} = [1, j]$. Otherwise C_{jk} has successive vertices i, j, k, 1, say, and C_{ij} is a closed convex polygon with successive vertices i, j, 1, with $i < j$, $D_{ijk} \geq 0$. See Figure 1.

Conversely, if C_{ij} is a closed convex polygon with successive vertices i, j, 1 ($i < j$, $D_{ijk} \geq 0$), then $C_{ij} \cup \triangle_{jk}$ is a closed convex polygon with successive vertices i, j, k, 1. Thus

$$\begin{aligned} Q_{jk} &= \max Q(C_{jk}) \\ &= \max Q(C_{ij}) + Q(\triangle_{jk}), \\ &\qquad \text{over } C_{ij} \text{ with vertices } i, j, 1 \end{aligned}$$

satisfying $i < j$, $D_{ijk} \geq 0$. (The case $i = 1$ is allowed so that C^* can be the line segment $[1, j]$ when C has only three vertices.) Then $Q_{jk} = Q_{ij} + Q(\triangle_{jk})$, where $i = N(k, j)$ is chosen to maximize Q_{ij} subject to $i \leq i < j$, $D_{ijk} \geq 0$.

The optimal polygon C_0 with leftmost vertex 1 is possibly concentrated at 1 or on a line segment $[1, j]$, or it has at least three vertices. In that case suppose it has successive counterclockwise vertices i_2, i_1, 1. Since the polygon is optimal,

$$Q(C_0) = Q_{i_2 i_1} = \max_{1<j<k} Q_{jk}.$$

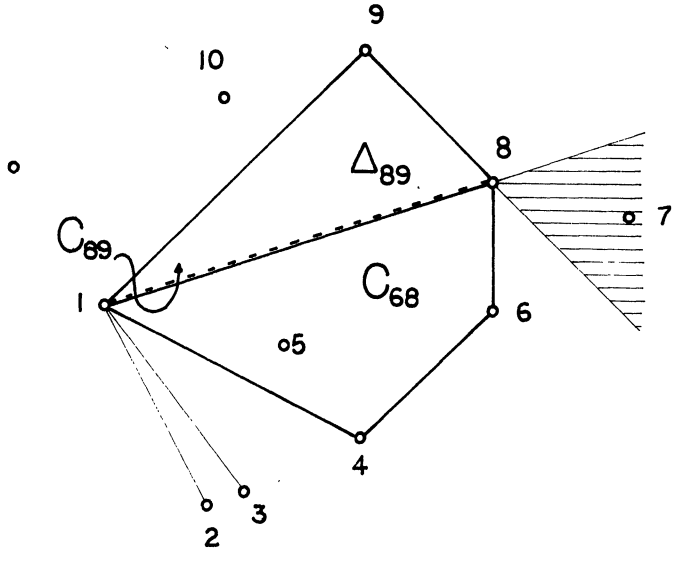

Figure 1. Notation for Constructing Optimal Polygons. An arbitrary vertex is numbered 1, and those not to its left are numbered 2, 3, . . . , 10 so that the rays [1, 2], [1, 3], [1, 4], . . . move counterclockwise. The triangle $\triangle_{89}$ excludes the segment [1, 8]. The polygon C_{89} has successive vertices 8, 9, 1, and it is composed of $\triangle_{89}$ and C_{68}, where C_{68} has successive vertices 6, 8, 1 (6 < 8), excluding those vertices such as 7 for which the quadrilateral with vertices 1, 7, 8, 9 is not convex.

(Any vertices i_2, i_1, 1 satisfying this equation will give an optimal polygon.)

If C_0 has a fourth vertex i_3, then $Q(C_0 - \triangle_{i_2 i_1})$ must be maximal over all polygons with vertices i, i_2, i_1, 1 ($i < i_2$, $D_{ii_2i_1} \geq 0$), which is ensured by setting $i_3 = N(i_1, i_2)$. (There may be more than one optimal i_3.) The other vertices are obtained similarly.

3. SIZE OF COMPUTATION

There are two parts to the computation, first determining $Q(\triangle_{ij})$ for the triangles with vertices 1, i, j and then finding Q_{ij} by recursion. Since there are n^3 triangles and n atoms, it appears that n^4 computations may be required for all $Q(\triangle_{ij})$, It is possible to reduce these calculations, however, by considering for each i, j the quadrilateral with vertices x_i, x_j, z_i, z_j, where z_i and z_j denote the projections of x_i and x_j on the x axis. Let F_{ij} denote the points in this quadrilateral excluding the segments $[x_i, z_i]$ and $[z_i, z_j]$, where x_i is to the left of x_j.

Now consider a triangle $\triangle_{ij}$ consisting of points in the convex hull of 1, i, j excluding $[1, i]$. See Figure 2.

If x_j is to the left of x_i, $\triangle_{ij} = F_{1j} + F_{ji} - F_{i1}$. If x_j is not to the left of x_i, $\triangle_{ij} = F_{1j} - F_{ij} - F_{i1} + (i, j]$, where $(i, j]$ is the set of points in $[i, j]$ excluding i. Here the sets are being treated as 0 − 1 functions. Thus Q is determined on $\triangle_{ij}$ by its values on F_{ij} and on $(i, j]$, and this requires only n^3 computations.

The recursive computation of $Q(C_{ij})$ at first sight requires $O(n^4)$ computations—n for the number of leftmost vertices, n^2 for the number of C_{ij}, and n for the number of C's to be checked at each i, j. This computation may be reduced to n^3 as follows: For each fixed j,

$$Q_{jk} = \sup_{i<j, D_{ijk} \geq 0} Q_{ij} + Q(\triangle_{jk}).$$

Generate a new ordering of atoms so that segments $[j, i]$ move counterclockwise as i increases from 1 through m. For each $i < j$ define

$$Q_i = \sup_{1 \leq k \leq i} Q_{kj}.$$

Then $Q_{jk} = Q_i + Q(\triangle_{jk})$, where i is the largest index with $D_{ijk} \geq 0$. As k increases from $j + 1$ to m, i increases from 1 to $j - 1$.

For each fixed j, the updating requires $O(n)$ computa-

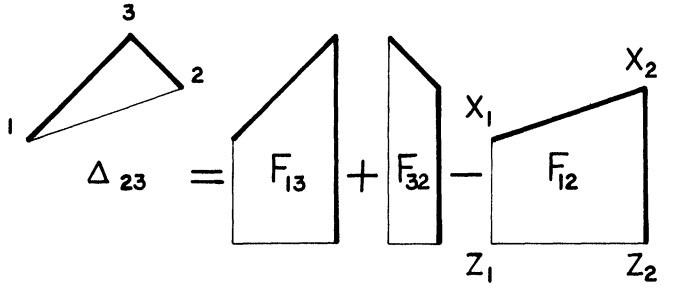

Figure 2. Evaluating $Q(\triangle_{ij})$. Each triangle $\triangle_{ij}$ may be expressed as a linear function of the trapezoids F_{1i}, F_{1j}, F_{ij}. All trapezoids may be evaluated in n^3 operations, so all triangles may be evaluated in n^3 operations.

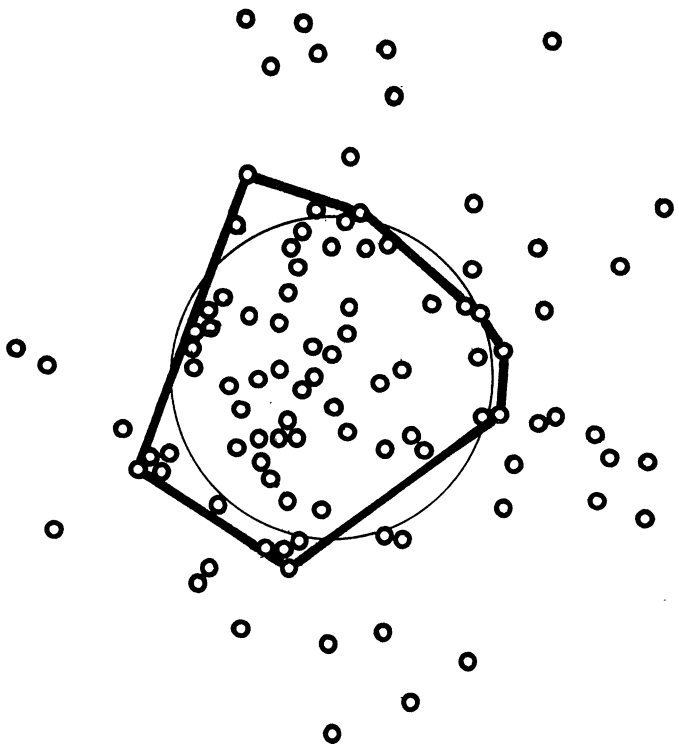

Figure 3. Convex Contour Estimates for the Bivariate Normal. The 100 points are sampled from a unit circular normal. The circle is the density contour containing 50% of the population; the density along the contour is $1/4\pi = .08$. The polygon maximizes the proportion of points inside less .08 times its area, over all convex polygons.

tions; for each k, it is only necessary to examine $O(1)$ Q_i's. The ordering of atoms about j is also taken as $O(n)$—sort the angles of $[j, i]$ into n equal-sized cells and bubble sort the atoms classified into the cells. There are n leftmost vertices and $O(n)$ selections of j, making $O(n^3)$ altogether.

On an IBM-AT computer with a mathematics coprocessor, the times to estimate the contour of a bivariate normal containing half the probability are as follows:

Sample size n	5	10	20	50	100
Time in seconds	.68	1.72	7.86	118.23	967.78
$(n/10)^3$	.13	1.0	8	125	1,000

An example for $n = 100$ is given in Figure 3.

4. ASYMPTOTICS

The Hausdorff distance between sets S and T in the plane is

$$\rho(S, T) = \sup_{y \in S} \inf_{x \in T} |x - y| + \sup_{y \in T} \inf_{y \in S} |x - y|,$$

where $|x - y|$ denotes Euclidean distance.

Theorem. Let P be a probability on the plane having density f with $f(x) \to 0$ as $|x| \to \infty$. Let P_n be the empirical probability for a sample of size n from P. Let S_0 be a closed convex set maximizing $(P - \lambda)S$, where $\lambda S = c \int_{x \in S} dx$, and let S_n be a closed convex set maximizing $(P_n - \lambda)S$. Suppose that S_0 is unique. Then $\rho(S_n, S_0) \to 0$ with probability 1.

Proof. For some closed sphere T, $f(x) < c$ for $x \notin T$. Let $Q = P - \lambda$, $Q_n = P_n - \lambda$. Thus

$$QS_0 = Q(S_0 \cap T) + Q(S_0 \cap T^c) \leq Q(S_0 \cap T),$$

since $Q(S_0 \cap T^c) \leq 0$. $S_0 \cap T$ is a closed convex set and S_0 is unique. Thus $S_0 = S_0 \cap T$; that is, $S_0 \subseteq T$. Since S_0 is maximal for Q, $Q(S_0 - S_n) \geq 0$. Since S_n is maximal for Q_n, $Q_n(S_n - S_0) \geq 0$. Since $\sup_{S \text{ convex}} |(P_n - P)S| \to 0$ with probability 1 (Rao 1962), $(P_n - P)(S_0 - S_n) \to 0$ with probability 1, $(Q_n - Q)(S_0 - S_n) \to 0$ with probability 1, $Q(S_0 - S_n) \to 0$ with probability 1, and $QS_n \to QS_0$ with probability 1. Since $Q(S_n \cap T^c) \leq 0$,

$$QS_n \leq Q(S_n \cap T) \leq QS_0$$

by S_0 optimality. Thus $Q(S_n \cap T) \to QS_0$ with probability 1. Let K be the nonempty closed convex subsets of T. The function QS from K to $\mathbf{R}$ is continuous in the Hausdorff topology. The set K is compact in the Hausdorff topology (see, e.g., Dieudonné 1966, p. 58). Since QS achieves its unique maximum at S_0, for each $\delta > 0$ there exists $\varepsilon > 0$ such that $S \in T$; $QS \leq QS_0 - \varepsilon$ implies that $\rho(S, S_0) \geq \delta$. Otherwise, compactness of K implies the existence of S^* with $QS^* = S_0$, $\rho(S^*, S_0) \geq \delta$, and that contradicts uniqueness of S_0. Thus $Q(S_n \cap T) \to QS_0$ with probability 1 implies that $\rho(S_n \cap T, S_0) \to 0$ with probability 1. Since this T may be chosen so that S_0 lies in its interior, $S_n \cap T$ eventually lies in the interior of T and $S_n \cap T = S_n$ for n large enough with probability 1. Thus $\rho(S_n, S_0) \to 0$ with probability 1, concluding the proof.

If P has a density f that has continuous nonzero derivatives near the convex contour $f = c$, it will be argued that

$$\rho(S_n, S) = 0\left(\frac{\log n}{n}\right)^{2/7}.$$

To see why this might be, consider f circular normal with means 0 and variances 1, and let $c = 1/4\pi$. Consider the convex set $S(x)$ obtained by the convex hull of x and the set $\{y \mid f(y) \leq c\} = S_0$.

Note that $f(y) = c$ iff $|y| = (2 \ln 2)^{1/2}$. Set $|x| = (2 \ln 2)^{1/2} + \delta$ $(\delta > 0)$. Then $S(x) - S_0$ has area $O(\delta^{3/2})$ and average density $1/4\pi - O(\delta)$. Thus

$$(P - \lambda)(S(x) - S_0) = O(\delta^{5/2}),$$

and $(P_n - \lambda)(S(x) - S_0)$ is linearly transformed binomial with mean $O(-\delta^{5/2})$ and variance $O(\delta^{3/2}/n)$.

The maximum of $(P_n - \lambda)(S(x) - S_0)$ over n^ε points x will be

$$O(-\delta^{5/2}) + O\left(\frac{\delta^{3/4}}{n^{1/2}} \sqrt{\log n}\right).$$

This maximum will be positive if

$$\delta^{5/2} = O(\delta^{3/4} \sqrt{\log n}/n^{1/2})$$

$$\delta^{7/2} = O((\log n)/n)$$

$$\delta = O((\log n)/n)^{2/7}.$$

Thus we can beat the region S_0 by forming a convex hull with some point x, $|x| = \sqrt{2 \ln 2} + O((\log n)/n)^{2/7}$.

A similar argument shows that we can beat S_0 by removing all points outside some chord distant $O((\log n)/n)^{2/7}$ from the circumference of S_0. This does not show that the optimal S_n is $O((\log n)/n)^{2/7}$ away from S_0, only that such distances arise in convex sets beating S_0; we speculate that if S_n' is any polygon within $n^{-2/7}$ of S_0, then it will be improved by adding a point $O((\log n)/n)^{2/7}$ away from S_n'.

5. CONDENSED DATA

For even moderate sample sizes, the algorithm is time consuming. Calculations will be reduced by limiting the number of edges that may appear in the optimal polygon. For example, only allow edges of small length.

Alternatively, set up an $N \times N$ array of square cells centered at $\{(i, j), 1 \le i \le N, 1 \le j \le N\}$ after transformation of the data, and allow edges that are line segments between centers. The value of $P_n - \lambda$ is computed for each cell, and the whole cell is assumed to lie in the polygon when its center does. The algorithm previously described may be used, with the centers playing the role of points, but N^6 computations are required for the N^2 centers.

It is possible to gain a little by considering (without loss of optimality) only edges that pass through precisely two centers. The proportion of such edges to the total is

N	2	5	10	20	50	100	200
proportion	1.000	.6667	.6275	.6132	.6089	.6080	.6080

No very great gain is made. The limiting fraction .608 might be of interest to number theorists. But it seems that $N = 20$ is a practical limit to the side of the grid.

Let us consider restrictions to small edges. If the only edges allowed are of length 1, the resulting convex set must be rectangular. The optimum rectangle may be computed in N^3 steps: Suppose the horizontal edges are fixed (k_1, k_2); let $V(j)$ be the value of $P_n - \lambda$ computed over all cells (i, k), $i \le j$, $k_1 \le k \le k_2$; then the best vertical edges i_1, i_2 maximize $V(i_2) - V(i_1)$, $i_1 \le i_2$.

If edges of length no greater than $\sqrt{2}$ are allowed, the resulting polygon has eight or fewer edges. It now becomes profitable to build the optimal polygon a line at a time. For the jth line, for each $i_1 \le i_2$, we compute the value of the optimal polygon with lower boundary the line from (i_1, j) to (i_2, j) and with edges running (i_r, j) to $(i_r, j - 1)$ or to $(i_r - 1, j - 1)$ or to $(i_r + 1, j - 1)$, $r = 1, 2$. There are nine such values for each i_1, i_2, j, and each may be computed by looking at a few of the optimal values for $j - 1$. The overall optimum is the largest value over all i_1, i_2, j requiring $O(N^3)$ computations, just like the rectangle.

For edges not exceeding $\sqrt{5}$, the optimal polygon has 16 or fewer edges and, as before, can be constructed in $O(N^3)$, although it is a big N^3. There are 49 optimal polygons to be considered having a particular line segment as the lower boundary. I suspect that this case is frequently a satisfactory compromise between having too many edges to search over and having too crude an approximation to the optimal contour.

6. BIMODALITY

My main motivation in putting forward this estimate is to develop a test of bimodality in two dimensions. If the population has two modes, there will be a density level c at which the contours become disconnected. To detect two modes, compute the empirical convex contour for each c and compute the c-contour $S_{n,c}^*$ among those points excluded by the empirical convex contour at c. The quantity

$$\sup_c \left[P_n(S_{n,c}^*) - c \int_{S_{n,c}^*} dx\right]$$

is suggested as a test statistic for the presence of two modes. It estimates the excess probability in a secondary mode over the amount of probability to be expected in $S_{n,c}^*$ if the density were unimodal.

[*Received August 1985. Revised March 1986.*]

REFERENCES

Chernoff, H. (1964), "Estimation of the Mode," *Annals of the Institute of Statistical Mathematics,* 16, 31–41.

Dieudonné, J. (1966), *Foundations of Modern Analysis,* New York: Academic Press.

Eddy, W. F. (1977), "A New Convex Hull Algorithm for Planar Sets," *ACM Transactions on Mathematical Software,* 3, 398–403.

Eddy, W. F., and Hartigan, J. A. (1977), "Uniform Convergence of the Empirical Distribution Function Over Convex Sets," *The Annals of Statistics,* 5, 370–374.

Rao, Ranga (1962), "Relations Between Weak and Uniform Convergence of Measures With Applications," *Annals of Mathematical Statistics,* 33, 659–680.

Robertson, Tim, and Cryer, J. D. (1974), "An Iterative Procedure for Estimating the Mode," *Journal of the American Statistical Association,* 69, 1012–1016.

Sager, T. W. (1979), "An Iterative Method for Estimating a Multivariate Mode and Isopleth," *Journal of the American Statistical Association,* 74, 329–339.

Venter, J. H. (1967), "On Estimation of the Mode," *Annals of Mathematical Statistics,* 38, 1446–1455.