


Consistency of Single Linkage for High-Density Clusters

J. A. HARTIGAN*

High-density clusters are defined on a population with density f in r dimensions to be the maximal connected sets of form $\{x \mid f(x) \geq c\}$. Single-linkage clustering is evaluated for consistency in detecting such high-density clusters—other standard hierarchical techniques, such as average and complete linkage, are hopelessly inconsistent for these clusters. The asymptotic consistency of single linkage closely depends on the percolation problem of Broadbent and Hammersley—if small spheres are removed at random from a solid, at which density of spheres will water begin to flow through the solid? If there is a single critical density such that no flow takes place below a certain density, and flow occurs through a single connected set above that density, then single linkage is consistent in separating high-density clusters (by disjoint single-linkage clusters that include a positive fraction of sample points in the respective clusters and pass arbitrarily close to all points in the respective clusters). The existence of a single critical point remains a conjecture. A weaker result is proved that shows that single-linkage clusters detect high-density clusters if there is a low enough valley separating them.

KEY WORDS: Single linkage; High-density clusters; Percolation processes.

1. HIERARCHICAL CLUSTERING

Given a set of points 1, 2, . . . , n with pairwise distances $d(i, j)$, $1 \leq i, j \leq n$, single-linkage clusters are defined as follows: Let i and j be the closest pair of points; amalgamate them to form a cluster ℓ and define the distance between that cluster and any point k by $d(\ell, k) = \min \{d(i, k), d(j, k)\}$; repeat the process, treating ℓ as a point and ignoring i and j. The amalgamation continues until all points are grouped in one large cluster. All clusters obtained in the course of the algorithm are *single-linkage* clusters. The first use of this technique is in Florek et al. (1951), although McQuitty (1957) and Sneath (1957) discovered it independently. It seems arbitrary that distance between clusters should be the minimum distance of pairs of points in the two clusters. Indeed, Sorensen (1948) suggests *complete linkage,* in which distance between clusters is the maximum over pairs of points, Sokal and Michener (1958) suggest *average linkage*, in which distance between clusters is an average of all pairs of distances, and Lance and Williams (1967) present a continuum of distance definitions including the above.

Single linkage tends to "chaining," producing long straggly clusters that are difficult to interpret; it is regarded with disfavor for this reason. Thus Kuiper and Fisher (1975) say that "single linkage is not appropriate unless one anticipates long chained clusters" and Sneath (1969) concedes that "complete linkage and average linkage will demonstrate mainly spherical clusters, while straggly ones can be found by single linkage." Wishart (1969), however, reports that single linkage failed to detect obvious straggly clusters in the Hirschsprung-Russell diagram, because of chaining. Lance and Williams (1967) "submit that nearest neighbour sorting should be regarded as obsolete."

Yet single linkage has some attractive properties: it is computationally simple; it is related to the minimum spanning tree (MST) by Gower and Ross (1969) (the MST is the graph of minimum length connecting all data points; single-linkage clusters are the connected subgraphs obtained by successively deleting links in the MST, largest first); single-linkage clusters are the maximal connected sets if points i and j are connected whenever $d(i, j) \leq d_0$, letting d_0 change to get clusters at various levels of the hierarchy; the two largest single-linkage clusters are the two sets dividing the points such that the minimum distance between the two sets is a maximum.

In this paper, the n points are treated as a sample from a population with density f with respect to Lebesgue measure on R^r. The *population clusters* are the maximal connected sets of density $\geq f_0$, with different levels of clusters obtained by varying f_0. These are called *high-density clusters* in Hartigan (1975). High-density concepts of clustering have been previously put forward by Carmichael, George, and Julius (1968), who require for clusters "(1) that there are continuous, relatively densely populated regions of the space, and (2) that these are surrounded by continuous relatively empty regions of the space." See also Wishart (1969), Ihm (1965), and Katz and Rohlf (1973).

If our task is to identify high-density clusters, it is necessary to estimate the density f or at least to estimate order relationships between the density at different

* J.A. Hartigan is Professor, Yale University, Department of Statistics, New Haven, CT 06520. The author is indebted to I.R. Savage, B. Spencer, and M. Steele for helpful discussions, and to P.A.P. Moran for some important references. This research was supported in part by NSF Grant MCS 75-08374.

© Journal of the American Statistical Association
June 1981, Volume 76, Number 374
Theory and Methods Section

points. The vast literature on density estimation (for example, Wegman 1972a,b) now becomes relevant. The difficulties with determining a suitably shaped kernel in density estimation are analogous to the difficulties of determining a suitable distance measure in clustering. Consider the "nearest neighbor" density estimate in which the density at a point x_0 is inversely proportional to the volume of the smallest sphere centered at x_0 and containing a data point. The high-density clusters corresponding to this density estimate are precisely the single-linkage clusters (Hartigan 1977b). It is known that kth nearest neighbor density estimates are consistent for the true population density only if $k \to \infty$ as $n \to \infty$. The density estimate corresponding to single linkage is thus inconsistent, which may explain its tendency to excessive chaining. Wishart (1969), Carmichael, George, and Julius (1968), Jardine and Sibson (1971), and Ling (1973) suggest more stable extensions of single linkage that correspond, roughly, to kth nearest neighbor density estimation. It is important to notice that the inconsistency of density estimation by "nearest neighbor" does not translate immediately into inconsistency of clustering. For example, in one dimension, if A and B are disjoint compact high-density clusters on the population, there exist disjoint compact single-linkage clusters A_n and B_n such that $A_n \supset A$, $B_n \supset B$ in probability as $n \to \infty$. Thus single linkage is consistent in one dimension because the location of the largest interval between observations lying in the set $[a, b]$ converges to the minimum of the density f in $[a, b]$ when f is continuous (Hartigan 1977a).

In this paper the consistency properties of single linkage in r dimensions are studied. For $r \geq 2$ it is found that single linkage is *not consistent,* in the sense that two disjoint compact high-density clusters in the population will not be detected asymptotically by two disjoint single-linkage clusters, which include, respectively, the sample points in the two population clusters. However, single linkage is *fractionally consistent,* under certain conditions, in that there will be asymptotically two disjoint single-linkage clusters that include, respectively, a positive fraction of the sample points in the two population clusters. The method of proof uses concepts from percolation processes (Broadbent and Hammersley 1957), reviewed recently in Smythe and Wierman (1978); single linkage may be studied by percolation theory because single-linkage clusters are the maximal connected sets when each sample point is replaced by a sphere of radius d. (Varying d gives clusters of various size.) The probability theory given here uses cubes rather than spheres because the calculations are easier and the asymptotic results follow equally well. Behavior is first examined for samples from the uniform distribution on the unit cube and then generalized to samples from general densities.

2. CONNECTIVITY OF SAMPLES FROM A CUBE

A sample of size n is drawn from the unit cube in R^r. Each point in the sample is the center of a closed cube of volume ρ/n, which will be called a *box*. The total volume of all boxes is ρ. This sampling scheme is obtained from the Poisson process of Gilbert (1961) and Roberts and Storey (1968) by considering the behavior of the process in a cube. A *cluster* is a maximal connected set of boxes.

It is mathematically elegant to consider the whole Poisson process in examining questions such as the existence and uniqueness of infinite clusters, but for many practical purposes results must be obtained within a cube. Difficulties arise in specializing infinite results within a cube. For this reason, results have been stated asymptotically as $n \to \infty$ within the cube, rather than for infinite clusters on a Poisson process.

Lemma 1. A box is singleton if it overlaps no other box. The proportion of singleton boxes approaches $\exp(-2^r\rho)$ in probability as $n \to \infty$.

Proof. The first box is singleton if no other point lies in a volume $2^r\rho/n$ about it, which occurs with probability $[1 - 2^r\rho/n]^{n-1}$ if the first point lies in the cube of edge $1 - (2^r\rho/n)^{1/r}$, which occurs with limiting probability one. Thus the first box is singleton with limiting probability $\exp(-2^r\rho)$. The probability that the first two boxes are singleton converges to $\exp(-2 \cdot 2^r\rho)$, so that the events that two boxes are singleton are asymptotically independent. When Chebyshev's inequality is used, the limiting proportion of singleton boxes converges to $\exp(-2^r\rho)$ in probability, as required.

Note. Similar arguments show that the proportions of doubleton, triple, and so on, clusters of boxes converge to some limiting value.

Lemma 2. If $\rho 2^r < 1$, the maximum diameter of a cluster approaches zero in probability as $n \to \infty$.

Proof. The points $1, 2, \ldots, k$ form a chain of length k if each point lies in a cube of volume $2^r\rho/n$ about the previous point, which occurs with probability $\leq (2^r\rho/n)^{k-1}$; there are $n!/(n - k)!$ ordered sets of k points, so a chain of length k exists with probability

$$\leq \frac{n!}{(n-k)!}(2^r\rho/n)^{k-1} \leq n(2^r\rho)^{k-1}.$$

If a cluster exists with diameter ϵ, there must be at least $\epsilon/\sqrt{r}(\rho/n)^{1/r}$ boxes chained together in the cluster, since the boxes have diagonal $\sqrt{r}(\rho/n)^{1/r}$. Thus such a cluster exists with probability

$$\leq n(2^r\rho)^{\epsilon/\sqrt{r}(\rho/n)^{1/r}-1},$$

which approaches zero as $n \to \infty$ whenever $\rho 2^r < 1$.

Note. A similar bound is given by Gilbert (1961) for random circles on the plane, and by Roberts and Storey (1968) for random spheres in three dimensions. A branching process is constructed beginning at, say, the first box; its descendants are any boxes overlapping it; their descendants are any boxes overlapping them, and so on. The expected number of descendants of each box is $\rho 2^r$, and the expected number of descendants over all generations

is finite if $\rho 2^r < 1$. Since many boxes are counted more than once, the expected cluster size is finite if $\rho 2^r < 1$.

Lemma 3. If $\rho > 4 \log 3$, there exists a "big" cluster with the following properties:

I The maximum distance between a point in the cube and the big cluster approaches zero in probability as $n \to \infty$.
II The maximum diameter of all other clusters approaches zero in probability as $n \to \infty$.
III A sample point belongs to the big cluster with probability exceeding $\alpha > 0$ as $n \to \infty$.
IV The fraction of sample points in the big cluster exceeds $\alpha > 0$ in probability as $n \to \infty$.

Proof. Consider first $r = 2$. Partition the square into K_n square "cells," where $\sqrt{K_n}$ is the largest integer smaller than $\sqrt{4n/\rho}$; obviously, $K_n\rho/4n \to 1$ as $n \to \infty$. If a sample point lies in one of the cells, the corresponding box covers the cell. A chain of k cells is a sequence of k cells such that neighboring members of the sequence have an edge in common. The number of chains of k cells beginning at a given cell $\leq 3^{k-1}$, since chains of length k may be generated (with some repetition) by adding one of the three adjoining cells to the end of a chain of length $(k - 1)$. The probability that a particular chain of k cells is empty is $(1 - k/K_n)^n$, so the probability that some chain of k cells, beginning from a certain point, is empty is less than $3^{k-1} \exp(-kn/K_n)$.

A cluster connects a point in the corner of the square to an edge opposite the corner, unless there is a polygonal line joining the edges adjacent to the corner and intersecting no boxes. Such a polygonal line passes through a chain of empty cells connecting the adjacent edges; the chain begins at one of $\sqrt{K_n}$ cells on an adjacent edge, and if it begins at cell i it is at least i long. The edges are connected by an empty chain with probability $\leq \sum_{k=1}^{\infty} 3^{k-1} \exp(-kn/K_n)$. This is a geometric series that converges for large n when $\rho > 4 \log 3$, to $(\exp(\rho/4) - 3)^{-1}$. Thus the probability of reaching from a corner to an opposite edge exceeds $1 - (\exp(\rho/4) - 3)^{-1}$ asymptotically.

A rectangle in the square is connected between opposite edges unless there is a polygonal line between opposite edges that intersects no box. Such a polygonal line passes through a chain of k connected cells where $k \geq A\sqrt{n}$, beginning on one side of the rectangle. The probability of such a chain $\leq \sqrt{K_n} 3^{A\sqrt{n-1}} \exp(-AN^{3/2}/K_n) \to 0$ as $n \to \infty$. Let B be the maximal connected set of boxes that contains the boxes connecting opposite sides of the square. For each $\epsilon > 0$, divide the square into $1/\epsilon$ rectangles that are 1 by ϵ; the connected set that crosses between edges of each rectangle must intersect the set crossing between edges of the square, so each such connected set is included in B. Every point in the rectangle is within ϵ of some point in B, in probability as $n \to \infty$, which establishes property I.

Any connected set of diameter ϵ must have range $\epsilon/\sqrt{2}$ in an east-west or north-south direction, say east-west. It crosses one of the $2\sqrt{2}/\epsilon$ rectangles that are 1 by $\epsilon/2\sqrt{2}$ and so intersects the "big set" B in probability. Thus every cluster other than the big set has diameter approaching zero, in probability, proving property II.

The corner of the square is connected to an opposite edge with positive probability. For any sample point, divide the square into four rectangles with the sample point at the corner. The sample point is connected to an opposite edge of the largest of these rectangles with positive probability, and this rectangle has opposite edges connected by the big set. The sample point therefore lies in the big set with positive probability, proving III.

The event that the first sample point $\mathbf{x}_1$ belongs to B is determined by the sample points in the neighborhood of $\mathbf{x}_1$, since $\mathbf{x}_1$ belongs to the big set if and only if it belongs to a cluster of diameter ϵ, some $\epsilon > 0$, in probability. The events that $\mathbf{x}_1$ and $\mathbf{x}_2$ belong to the big set are thus asymptotically independent, being determined by independent sample points in the neighborhood of $\mathbf{x}_1$ and $\mathbf{x}_2$. It follows from the law of large numbers that the proportion of sample points belonging to the big set is asymptotically positive, proving IV. (This argument may be made rigorous by dividing the square into cells of area ϵ, noting that $\mathbf{x}_1$ and $\mathbf{x}_2$ nearly always fall in different cells and that asymptotic behavior occurs independently in the different cells.) Turning now to the three-dimensional case, consider $1 + (n/\rho)^{1/3}$ planes distant $(\rho/n)^{1/3}$ apart; each box intersects exactly one of the planes in a two-dimensional box of area $(\rho/n)^{2/3}$; each plane contains approximately $n/(n/\rho)^{1/3}$ such boxes, so the total area of all two-dimensional boxes in each plane is approximately ρ. It will be shown that a two-dimensional big set exists in each plane and that the big sets for different planes are connected.

It is necessary to consider the simultaneous behavior of all big sets in the planes. Divide each plane into $n^{1/6} = 1/\epsilon_n$ rectangles $1 \times \epsilon_n$. A rectangle is connected between opposite edges with probability $\geq$

$$1 - \sqrt{K_m} \exp(-km/K_m + k \log 3)$$

where m is the number of points in the plane, $4K_m(\rho/n)^{2/3} \to 1$ as $n \to \infty$, and $k \geq \sqrt{K_m}/\epsilon_n$. All rectangles in all planes are connected across opposite edges with probability $\geq$

$$1 - \left(\frac{n}{\rho}\right)^{1/3} n^{1/6} \sqrt{K_m} \exp(-km/K_m + k \log 3)$$

where m denotes the minimum number of points in a plane, over the $1 + (n/\rho)^{1/3}$ planes. It follows that $m = \rho^{1/3} n^{2/3} - 0(n^{1/3} \log n)$ and that $4K_m\rho/m \to 1$ as $n \to \infty$. Thus all rectangles in all planes are connected across opposite edges in probability as $n \to \infty$.

Now consider corresponding rectangles in neighboring planes; there are two big plane sets corresponding to the two planes, and these big sets must intersect when put on the same plane since they both connect opposite sides of the rectangle. There is a box belonging to the first big set and a box belonging to the second big set, which overlap when the two big sets are put on the same plane.

The distribution of a box, in the third dimension, given that it lies in a plane, is uniform over an interval of length $(\rho/n)^{1/3}$ and independent of its location in the plane. Two boxes in neighboring planes, whose two-dimensional projections overlap, overlap with probability $\frac{1}{2}$. Thus two neighboring big sets are connected if any big sets in $n^{1/6}$ neighboring rectangles are connected, which occurs with probability at least $1 - 2^{-n^{1/6}}$, and all neighboring big sets are connected with probability at least

$$1 - \left(\frac{n}{\rho}\right)^{1/3} 2^{-n^{1/6}} \to 1 \quad \text{as } n \to \infty.$$

Thus I is established for cubes. Also, III and IV may be established by considering big sets on the planes—given that a sample point lies in a certain plane, it has positive probability of belonging to the big set for that plane, and a positive fraction of points in each plane belong to the big set in probability as $n \to \infty$.

To prove II, name the parallel planes $H_1, H_2, \ldots$ and let B_k denote the big set generated from squares in plane H_k. Beginning with a box b_1 intersecting H_1, let b_{k-1}, be the box H_{k-1}, connected to b_0, and closest to H_k. There is a probability $\geq \rho_1 > 0$ for all n that b_{k-1} overlaps a box that intersects H_k. Given b_{k-1}, there is a probability $\geq \rho_2 > 0$ for all n that b_{k-1} is connected to B_k (since B_k intersects a prespecified point in H_k with positive limiting probability, independent of behavior in $H_1, \ldots H_{k-1}$). Therefore, b_1 is connected to a box in H_k without belonging to $B_1, B_2, \ldots,$ or B_k with probability $\leq (1 - \rho_2)^k$.

A cluster of length ϵ must reach at least $\epsilon/\sqrt{3}$ in one of three directions, say the direction perpendicular to the planes, and must cross at least $k_n = \epsilon/\sqrt{3}(\rho/n)^{1/3}$ planes. Since there are n boxes to start the cluster, it exists with probability $\leq n(1 - \rho_2)^{k_n} \to 0$ as $n \to \infty$, proving II.

In r dimensions, consider $(n/\rho)^{1-2/r}$ parallel planes separated by increments of length $(n/\rho)^{1/r}$ in the remaining $(r - 2)$ dimensions; each plane contains about $n^{2/r}\rho^{1-2/r}$ boxes of area $(\rho/n)^{2/r}$ and total area ρ. The theorem applies to all planes simultaneously, and neighboring big sets (in each of the remaining $(r - 2)$ dimensions), are connected asymptotically. The results I through II are proved as for $r = 3$.

Conjecture. There exists a critical density ρ_r such that

I For $\rho < \rho_r$, the maximum diameter of clusters approaches zero asymptotically;

II For $\rho > \rho_r$, the maximum diameter of all clusters but one, the "big" cluster, approaches zero asymptotically. The big cluster contains a positive fraction of sample points asymptotically and the maximum distance of a point in the cube from the big cluster approaches zero as $n \to \infty$.

The existence of such a critical density has been assumed in much of the empirical work on percolation processes. However, the existence of ρ_r has not been established even in the closely studied case of the square lattice (Smythe and Wierman 1978). Pike and Seager (1974) assume that if a cluster exists connecting all opposite pairs of faces of the cube, then Hammersley's critical density has been reached in which each point has a finite probability of belonging to an infinite cluster. However, it is quite possible that many different clusters exist with diameters not asymptotically zero, and yet each point has limiting probability zero of belonging to such a cluster. It must also be demonstrated that many big clusters containing positive fractions of sample points cannot coexist. (The uniqueness of an infinite cluster has been shown by Harris (1960) for the square lattice.)

In the square, the conjecture implies that for $\rho < \rho_r$ no clusters cross the square, but for $\rho > \rho_r$ a big cluster exists that crosses the square asymptotically; the case $\rho = \rho_r$ is indeterminate—the square is crossed with probability not asymptotically zero or one. If the conjecture were not true, there would be a range of ρ values where the crossing probability is asymptotically nondegenerate. The Pike and Seager (1974) experiment shows that for circles in the square, the smallest value of ρ for which both pairs of opposite edges are connected has an average value of 1.12 and a standard deviation of .027, when 4,000 points are taken in the square. This suggests that the interval of ρ values where crossing is probable but not certain is very small; according to the conjecture it consists of a single point.

Theorem 1. A density f is bounded away from 0 and ∞ in a compact set S in R^r having a connected interior S^0. A distance measure d on R^r is such that $0 < a < d(x, y)/d_0(x, y) < b < \infty$ for all x, y in R^r, where d_0 is euclidean distance. A sample of n points is taken from f and each point x_i is the center of a sphere $\{y \mid d^r(x_i, y) \leq \rho/n\}$.

I For ρ large enough there is a big cluster of spheres that includes a positive fraction of sample points and passes within ϵ_n of each point of S^0, where $\epsilon_n \to 0$ in probability as $n \to \infty$. Further, every other cluster has diameter less than ϵ_n, where $\epsilon_n \to 0$ in probability as $n \to \infty$.

II For ρ small enough, every cluster has diameter less than ϵ_n, where $\epsilon_n \to 0$ in probability as $n \to \infty$.

Proof. If a big cluster exists for a distance measure d at density ρ, it will exist for d_0 at some density ρ_0, because it is possible to choose the density ρ_0 so that d_0 spheres include d spheres. Similarly, if the maximum diameter of d clusters approaches zero for some ρ, the maximum diameter of d_0 clusters approaches zero for some ρ_0. The same statements hold if d and d_0 are interchanged.

It is more difficult to show that the maximum diameter of other clusters converges to zero, when a big cluster exists for ρ large enough.

Divide the cube into cubical cells of volume $A\rho/n$, where A is chosen so that if a sample point falls in a cell, the corresponding d box covers the cell. Then if clusters (other than the big cluster) exist having diameter ϵ, there

is a connected line of length ϵ that intersects no box, and so there is a connected set of $\epsilon/(A\rho/n)^{1/r}$ cells that are empty. For ρ sufficiently large the probability of such a connected set approaches zero as $n \to \infty$.

Thus I and II hold for the uniform distribution over the cube. Obtain a sample from an arbitrary density over the cube by taking a uniform sample and accepting points with probability $f(x)/\sup f(x)$; from n points accept approximately $n/\sup f(x)$. Choose ρ so that II is satisfied for the uniform—then II is satisfied for f at density $\rho/\sup f(x)$, since the accepted points will certainly satisfy II if the complete set does. Similarly, construct a sample from the uniform by accepting points in a sample from f with probability $\inf f/f(x)$; from n points accept about $n \inf f(x)$; choose ρ so that I is satisfied for the uniform; then a big set exists for samples from f at density $\rho/\inf f(x)$, since a big set will exist on a subset of those points. Also, no line of length ϵ exists that does not touch one of the boxes in the uniform sample, so all other clusters have a diameter that approaches zero in probability as $n \to \infty$.

For an arbitrary compact S with connected interior S^0, choose a cube $C \subset S^0$ and a cube $U \supset S$. Sample points with density proportional to f in S and proportional to one in U. For ρ small enough, clusters in U will be small asymptotically, and so necessarily clusters in S will have maximum diameter approaching zero in probability, proving II. For ρ large enough, there will be a big cluster in C and a big cluster B in U; these will intersect in probability as $n \to \infty$, since all clusters other than B must be asymptotically small. Let B_0 be the component of $S \cap B$ that contains the big cluster in C. If any other component of $S \cap B$ has diameter $> \epsilon$, since S has a connected interior, there exists a set of diameter $> \epsilon$ that intersects no box in U; for ρ large enough, this event occurs with limiting probability zero as $n \to \infty$. Thus all clusters in S other than the big cluster B_0 have maximum diameter approaching zero in probability as $n \to \infty$. Finally, every point in S distance at least ϵ from the boundary, and distance at most $\frac{1}{2}\epsilon$ from B, must be distance at most $\frac{1}{2}\epsilon$ from B_0, or else there exists a component of $S \cap B$, not B_0, of diameter at least $\frac{1}{2}\epsilon$. Thus every point in S distance at least ϵ from the boundary must be distance at most $\frac{1}{2}\epsilon$ from B_0 in probability as $n \to 0$. Therefore, every point in S is distance at most $\frac{3}{2}\epsilon$ from B_0 in probability as $n \to \infty$, for every choice of $\epsilon > 0$, proving I.

3. CONSISTENCY OF SINGLE LINKAGE

A *high-density cluster* is a maximal connected component of a set of form $\{x \mid f(x) \geq c\}$, where f is a probability density on some space. The family τ of high-density clusters has the tree property, that $A, B \in \tau$ implies $A \subset B$, or $B \subset A$, or $A \cap B = \phi$. The *value* of a cluster A is $v(A) = \inf_{x \in A} f(x)$. The *similarity* of any two sets A, B, $s(A, B) = \sup_S v(S)$ where S denotes a cluster including A and B.

Theorem 2. Let f denote a density in R^r such that $\{x \mid f(x) \geq \epsilon\}$ is the union of a finite number of compact sets with connected interiors, each $\epsilon > 0$. Let A and B be disjoint high-density clusters included in a cluster $S \neq R^r$. Single-linkage clusters are constructed from a sample of size n from f, using a distance d such that $a < d(x, y)/d_0(x, y) < b$ all x, y, where d_0 is euclidean distance.

I For $r = 1$, single linkage is *fully consistent* for separating A and B—if $x_1, \ldots, x_n$ is a random sample from f, there exists a single-linkage cluster A_n containing all sample points in A and a single-linkage cluster B_n containing all sample points in B, such that A_n and B_n are disjoint with probability approaching 1 as $n \to \infty$.

II For $r > 1$, single linkage is *not* fully consistent—any single-linkage cluster that contains all the sample points in A will also contain nearly all sample points in B, in probability as $n \to \infty$.

III For $r > 1$, single linkage is *fractionally consistent:* If $s(A, B) \leq kv(A)$, $s(A, B) \leq kv(B)$ for k sufficiently small, there exist disjoint single-linkage clusters A_n and B_n so that, in the limit, A_n contains a positive fraction of points in A, passes arbitrarily close to every point in A, and all other clusters in A are arbitrarily small; B_n behaves similarly in B; and for some $\epsilon > 0$, $\inf\{d(x, y) \mid x \in A_n, y \in B_n\} > \epsilon$; in probability as $n \to \infty$.

Proof. For I, see Hartigan (1977a). The proof uses the known distribution of the largest interval between uniform order statistics to show that for samples from a density f in $[a, b]$ with unique minimum, the location of the largest interval between sample points converges to the location of the minimum of f. Thus the two largest single-linkage clusters are separated by the location of the minimum.

For II, note that single-linkage clusters are formed by enclosing x_i in the sphere $\{y \mid d^r(x_i, y) \leq \rho_n/n\}$ and constructing connected sets of spheres, for all values of ρ_n. From Lemma 1, there will be singletons in A if ρ_n is bounded as $n \to \infty$. If ρ_n is not bounded, there will be a big cluster reaching over S, and including nearly all points of S, from Theorem 1; thus any cluster that contains all points of A must contain nearly all points of B.

For III, let $S_\epsilon = \{x \mid x \in S^0, f(x) < s(A, B) + \epsilon\}$. From the definition of $s(A, B)$ every path between A and B meets S_ϵ. Since $\{x \mid f(x) \geq s(A, B) + \epsilon\}$ has finitely many components, S_ϵ is the union of a finite number of connected open sets, and on the closure of each of these sets, $f(x) \leq s(A, B) + \epsilon$. Note that $s(A, B)/k \leq v(A)$, $v(B) \leq f(x)$, $x \in A \cup B$. Choose spheres of volume ρ/n about the sample points; in S_ϵ the total volume of spheres per unit volume $\leq [S(A, B) + \epsilon]\rho$, and in $A \cup B$ the total volume per unit volume $\geq \inf[v(A), v(B)]\rho$. Thus for k large enough, Theorem 1 implies that there exists ρ for which big clusters exist in A and B, but only small clusters in each component of S_ϵ in probability as $n \to \infty$. If A

and B are connected by a cluster, that cluster will meet S_ϵ in a point x where $f(x) < s(A, B) + \epsilon/2$; such a point is contained in a cube of edge δ_ϵ lying entirely in S_ϵ, since S_ϵ is open. Thus a cluster of length $\frac{1}{2}\delta_\epsilon$ must occur in S_ϵ, which occurs with probability approaching zero as $n \to \infty$. Thus A and B are not connected by a cluster, but big clusters reach over each of A and B, in probability as $n \to \infty$.

Remark. If the conjecture is true (that all clusters are small for $\rho < \rho_0$ and a unique big cluster exists for $\rho > \rho_0$), then in uniform sampling from the cube, a stronger consistency result applies. For any two disjoint high-density clusters A and B, there exist disjoint big single-linkage clusters A_n and B_n such that $d(A_n, A) \to 0$, $d(B_n, B) \to 0$ in probability, where $d(C, D) = \sup_{x\epsilon C} \inf_{y\epsilon D} d(x, y) + \sup_{y\epsilon C} \inf_{x\epsilon D} d(x, y)$.

4. INTERPRETING SINGLE-LINKAGE CLUSTERS

How can we tell whether there are disjoint population clusters? We will assume a very large number of sample points; for n sample points there will be $(n - 1)$ single-linkage clusters, but it is only of interest to examine clusters containing a positive fraction (say 5 percent) of the sample points, since only these may be big sets for a population cluster.

If two disjoint population clusters exist, with a sufficiently deep valley separating them, asymptotically there will be big sets passing arbitrarily close to each point in the respective population clusters. Each of the big sets specified includes a positive fraction of the sample points lying in the clusters. Asymptotically, the distance between the big sets is bounded away from zero.

It follows that single linkage is conservative—it won't necessarily detect all clusters, but it will detect modes separated by a sufficiently deep valley. In contrast, complete and average linkage are consistently misleading (Hartigan 1977a)—the final clusters depend on the range of the data, not the density, and they do not identify modes.

If the conjecture is true, a stronger result holds. Let A be any high-density cluster 0, let A_n be the largest single-linkage cluster consisting of sample points in A. Then $d(A_n, A) \to 0$ in probability. Thus all high-density clusters will be identified.

A suggested test for the existence of more than one cluster is the second largest cluster size test; find two disjoint clusters such that the size of the smaller cluster is as large as possible. If this size is large enough, reject the null hypothesis of unimodality.

5. IMPROVING SINGLE LINKAGE

Improved estimates of clusters are obtained from improved density estimates, for which there exists a vast statistical literature. Much of it concerns kernel estimates, which are difficult to adapt to the task of determining density contours. An obvious extension of single linkage is kth nearest neighbor, in which the density at a point is a decreasing function of the distance to the kth nearest neighbor; analogous extensions in clustering have been considered by Ling (1973), and Wishart (1969). To obtain a consistent density estimate, it is necessary to have $k \to \infty$ as $n \to \infty$.

There is a simple operation on the minimum spanning tree that should give improved density estimates. (The minimum spanning tree is the network of minimal length that connects all the points; single-linkage clusters are the connected sets obtained by omitting all links in the tree exceeding a certain length.) Each link on the tree is replaced by the average of it and neighboring links; this operation is repeated several times, and clusters are computed from the remaining tree in the usual way.

In constructing clusters as maximal connected high-density regions, it is necessary to have estimates of density and also measures of connectivity; thus it is not sufficient to have only estimates of density at each sample point—it is necessary also to specify which sample points may be connected. One "kth nearest neighbor" estimate that provides densities and connections is as follows: Estimate the minimum density on the line between any two points x and y as a monotone function of the minimum distance r_k such that k sample points are within r_k of x and y, and apply single linkage to r_k.

[*Received January 1979. Revised August 1980.*]

REFERENCES

BROADBENT, S.R., and HAMMERSLEY, J.M. (1957), "Percolation Processes, I.: Crystals and Mazes," *Proceedings of the Cambridge Philosophical Society,* 53, 629–641.

CARMICHAEL, J.W., GEORGE, J.A., and JULIUS, R.S. (1968), "Finding Natural Clusters," *Systematic Zoology,* 17, 144–150.

FLOREK, J., LUKASZEWICZ, J., PERKAL, J., STEINHAUS, H., and ZYBRZYCKI, S. (1951), "Sur la liaison et la division des points d'un ensemble fini," *Colloquia Mathematicae,* 2, 282–285, 319.

GILBERT, E.N. (1961), "Random Plane Networks," *Journal of the Society for Industrial and Applied Mathematics,* 9, 533–543.

GOWER, J.C., and ROSS, G.J.S. (1969), "Minimum Spanning Trees and Single Linkage Cluster Analysis," *Applied Statistics,* 18, 54–65.

HARRIS, T.E. (1960), "A Lower Bound for the Critical Probability in a Certain Percolation Process," *Proceedings of the Cambridge Philosophical Society,* 56, 13–20.

HARTIGAN, J.A. (1975), *Clustering Algorithms,* New York: John Wiley.

——— (1977a), "Distribution Problems in Clustering," in *Classification and Clustering,* ed. J.V. Ryzin, New York: Academic Press.

——— (1977b), "Clusters As Modes," *First International Symposium on Data Analysis and Informatics,* Versailles: IRIA.

IHM, P. (1965), "Automatic Classification in Anthropology," in *Use of Computers in Anthropology,* ed. Dell Humes, London: Horton, 358–376.

JARDINE N., and SIBSON, R. (1971), *Mathematical Taxonomy,* London: John Wiley.

KATZ, J.O., and ROHLF, F.J. (1973), "Function Point Cluster Analysis," *Systematic Zoology,* 22, 295–301.

KUIPER, F.K., and FISHER, L. (1975), "A Monte Carlo Comparison for Six Clustering Procedures," *Biometrics,* 31, 777–784.

LANCE, G.N., and WILLIAMS, W.T. (1967), "A General Theory of Classificatory Sorting Strategies: I. Hierarchical Systems," *Computer Journal,* 9, 373–380.

LING, R.F. (1973), "A Probability Theory of Cluster Analysis," *Journal of the American Statistical Association,* 68, 159–169.

McQUITTY, L.L. (1957), "Elementary Linkage Analysis for Isolating

Orthogonal and Oblique Types and Typal Relevancies," *Educational and Psychological Measurement,* 17, 207–229.
PIKE, G.E., and SEAGER, C.H. (1974), "Percolation and Conductivity: A Computer Study, I.," *Physical Review,* B, 10, 1421–1446.
ROBERTS, F.D.K., and STOREY, S.H. (1968), "A Three-Dimensional Cluster Problem," *Biometrika,* 55, 258–260.
SMYTHE, R.T., and WIERMAN, J.C. (1978), "First-Passage Percolation on the Square Lattice," *Lecture Notes in Mathematics 671,* Berlin: Springer-Verlag.
SNEATH, P.H.A. (1957), "The Applications of Computers to Taxonomy," *Journal of General Microbiology,* 17, 201–206.
——— (1969), "The Evaluation of Cluster Methods," in *Numerical Taxonomy,* ed. A.J. Cole, London: Academic Press.
SOKAL, R.R., and MICHENER, C.D. (1958), "A Statistical Method for Evaluation Systematic Relationships," *University of Kansas Science Bulletin,* 38, 1409–1438.
SORENSON, T. (1948), "A Method of Estimating Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content," *Biologiske Skrifter,* 5, 1–34.
WEGMAN, E. (1972a), "Non-parametric Probability Density Estimation I," *Technometrics,* 14, 533–546.
——— (1972b), "Non-parametric Probability Density Estimation II," *Journal of Statistical Computing and Simulation,* 1, 225–245.
WISHART, D. (1969), "Mode Analysis: A Generalisation of Nearest Neighbour Which Reduces Chaining Effects," in *Numerical Taxonomy,* A.J. Cole, London: Academic Press.