Journal of Classification 2:63-76 (1985)

Journal of
Classification

Statistical Theory in Clustering

J. A. Hartigan

Yale University

Abstract: A number of statistical models for forming and evaluating clusters are reviewed. Hierarchical algorithms are evaluated by their ability to discover high density regions in a population, and complete linkage hopelessly fails; the others don't do too well either. Single linkage is at least of mathematical interest because it is related to the minimum spanning tree and percolation. Mixture methods are examined, related to k-means, and the failure of likelihood tests for the number of components is noted. The DIP test for estimating the number of modes in a univariate population measures the distance between the empirical distribution function and the closest unimodal distribution function (or k-modal distribution function when testing for k modes). Its properties are examined and multivariate extensions are proposed. Ultrametric and evolutionary distances on trees are considered briefly.

Keywords: Theory of clustering; High density clusters; Tests of unimodality.

1. Introduction

Classification, placing sets of objects in similar classes, is necessary for language and thought and is the foundation of statistical data collection and of probability judgments. This coin gives heads with probability ½ because you classify it with other remembered coins half giving heads and half giving tails. You predict rain after thunder because you classify the thunder with other thunders followed by rain.

The statistician is pleased to inform the biologist that his fossil shellfish divide distinctly into three clusters evidenced by a trimodal distribution of the measurements on number of whorls and relative diameter of the innermost and outermost chambers of the shell. The biologist is not surprised because they looked like three different species and he made those measurements that he thought would best distinguish them. Classification precedes measurement.

Research supported by the National Science Foundation Grant No. MCS-8102280.
Author's Address: J.A. Hartigan, Yale University, Department of Statistics, Box 2179 Yale Station, New Haven, Connecticut 06520, USA.

Statistical theory cannot provide a complete theory of classification. We cannot say how similarities should be judged, although we can give technical assistance in constructing distances. Different classifications are right for different purposes, so we cannot say any one classification is best. Statistical theory in clustering provides a testing ground for various clustering methods -- we discover how well the methods work for various idealized forms of data, and reject those methods that fail, at least for application to similar types of real data.

One general model is that the data form a random sample $X_1, X_2, \ldots, X_n$ from some population with a probability distribution P. A technique produces some clusters in the sample. A theoretical model generates some clusters in the population with the distribution P. We evaluate the technique by asking how well the sample clusters agree with the population clusters.

Many classifications apply to complete sets of objects, not samples of them. For example, classify the 50 United States by their agricultural products. Nevertheless you might not want to use a technique such as complete linkage that is asymptotically inconsistent. Frequently, you have available a sample that cannot be regarded as random. You collect all the specimens available at the site, but you wish to form a taxonomy of general utility. Some species will be represented much more highly than others; if you treat the whole as a random sample, the overrepresented species will receive too much attention. In survey sample theory much attention is paid to probability sampling, where the probability that each population individual enters the sample is known; perhaps this theory can be adapted to clustering problems. We do not usually know selection probabilities, but we might be able to progress by assuming the selection probabilities are the same within each cluster. For example, in the normal mixture model we could estimate the normal parameters for each component in the population, but not the mixing proportions, because these would be confounded with unknown selection probabilities.

2. High Density Clusters

A model suggested by clusters of stars is that a cluster corresponds to a high density region in p-dimensional space, Hartigan (1975).

Let P be the population distribution, let x be a typical point in p dimensional euclidean space, let f be the density of P with respect to Lebesque measure. The population clusters are the maximal connected subsets of the high density region $\{x \mid f(x) \geqslant c\}$ for each c. The family of population clusters forms a tree, in that if two clusters overlap, one includes the other. This model is thus suitable for examining hierarchical techniques. Taking the density of P with respect to Lebesque measure rather

than some other measure ensures that the population clusters are the same if a non-singular linear transformation is performed on the space.

For discrete data, we might assume that P is supported by the vertices of a cube in p-dimensional space, and take f to be the density with respect to the uniform distribution on the vertices. A set is connected if any two vertices in the set may be connected by a chain of cube edges between vertices in the set. The same definition of high density clustering may then be used.

Methods of density estimation produce clusters on the sample, namely the high density clusters corresponding to the estimate. This is to be expected because density estimates at a point depend on nearby sample points, and the definition of "nearby" corresponds to the similarity assessments in clustering methods. (Probability rests on similarity between what we know and what we are guessing!) Conversely hierarchical clustering methods may be interpreted as estimates of density contours, although the density itself is only specified by the clustering up to a monotone transformation.

3. Agglomerative Methods for High Density Clusters

Agglomerative methods define a distance between any two possible clusters, and the clusters are constructed by beginning with n singleton clusters, one for each point, and successively joining the closest pairs of clusters to form new clusters. These methods are poor estimates of high density clusters. See Hartigan (1977, 1982).

Complete linkage, in which the distance between clusters is the maximum distance between points in the two clusters, is the worst of all standard methods for high density clustering. If the distribution P is carried by a compact set C on which the minimum density is positive, I conjecture that the asymptotic behavior of the complete linkage clustering depends only on C, not on P. To be more precise, fix three points x, y, z and let the closest sample points to them, in a sample of size n, be x_n, y_n, z_n. I conjecture that as $n \to \infty$, the probability that x_n and y_n are joined together before x_n and z_n depends only on C.

Complete linkage remains a popular method because it gives nice evenly bifurcating trees for almost all data sets -- the real world, not so nice, does not show through. If P is supported by disconnected sets, then complete linkage will discover those sets, which depend only on C, the supporting set. What upsets complete linkage is the little fuzz of observations between the high density regions.

Why does complete linkage fail? After we have joined the small clusters together, all clusters have roughly the same diameter (if the maximum diameter of the clusters is d, no neighboring pair of clusters can amalgamate to a cluster of diameter less than d, so at least one of the pair must have

diameter $d/2$). Later decisions are made entirely on the pairwise distances between clusters, which do not depend on the number of points in the clusters; thus at this stage information about the distribution of points is already lost.

Average linkage defines distance between clusters as the average distance between pairs of points in the two clusters. It is known in the numerical taxonomy literature as the unweighted pair group method. It behaves somewhat better than complete linkage in sensitivity to the population distribution because the distance measure is affected by the number of points in the clusters. If two neighboring clusters are formed that cut across a high density region, the distance between clusters will be smaller than usual because of the many close pairs of points, and so these neighboring clusters will be quickly joined identifying the high density region. See Figure 1, where the clusters (2,3) and 4 are joined before (2,3) and 1 so that the high density cluster (3,4) is separated from the high density cluster 1. The weighted pair group method, in which distance between two clusters is just the average distance between component *clusters* (rather than points), should be no better than complete linkage, since after a small amount of joining the numbers of points in the various clusters becomes irrelevant.

The centroid method measures distance between clusters 1 and 2 by $n_1\, n_2\, \rho^2\, (\bar{x}_1,\bar{x}_2)/\, (n_1+ n_2)$ where n_i is the number of points and $\bar{x}_i$ the mean point in the ith cluster, and p is the euclidean distance. This ensures that the two clusters are joined to least increase the within cluster sum of squares; the method is the hierarchical analogue of the k-means algorithm. The resulting clusters are sensitive to the population distribution; the intermediate clusters (those obtained by a moderate amount of joining) are smaller in diameter in high density regions. Nevertheless these clusters are not consistent for high density clusters -- it is easy to have the edge of a large cluster join with a neighboring small cluster rather than with the other parts of the large cluster.

4. Single Linkage, the Minimum Spanning Tree and Percolation

Single linkage clustering measures the distance between clusters as the minimum distance between pairs of points in the two clusters. Single linkage clustering is consistent for high density clusters in one dimension in the sense that two fixed disjoint population clusters will eventually lie within some two disjoint sample clusters with probability one. Only approximate consistency holds in more than one dimension: let A and B be two disjoint population clusters, and define the distance between two sets C and D,

$$\rho(C,D) = \sup_{x\in C} \inf_{y\in D} \rho(x,y) \quad .$$

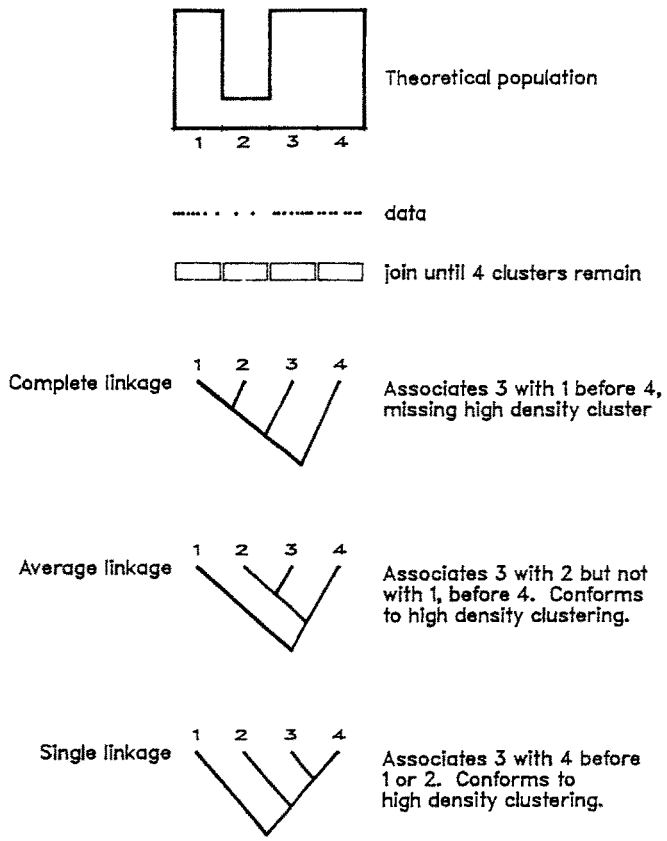

Figure 1. Comparative behavior of complete, average and single linkage.

If $\rho(C,D)$ is small, every point of C has some point of D close to it, so that D approximately includes C. As $n\to\infty$, with probability one, there exist disjoint single linkage clusters A_n and B_n such that $\rho(A,A_n)\to 0$, $\rho(B,B_n)\to 0$, Hartigan (1981). The single linkage clusters are straggly affairs whose contours by no means approximate the population density contours, but each of the two single linkage clusters has a point near each point in the two population clusters.

Single linkage clusters have a number of equivalent characterizations that make single linkage attractive for theoretical study. For example, divide the points into two clusters so that the minimum distance between the two clusters is as large as possible, and continue dividing the clusters obtained, in the same way. This produces single linkage clusters; the other agglomerative methods have no simple divisive characterization, and so it is to be expected that the large clusters they produce have no known desirable properties.

Replace each point by a sphere of radius d and consider the maximal connected subsets of the union of spheres, for all d. These are the single linkage clusters. Clusters of this type are studied in percolation theory (Broadbent and Hammersley 1957, Smythe and Wierman 1978) and so asymptotic results about single linkage clusters may be obtained from percolation asymptotics.

The nearest neighbor density estimate is

$$f_n(x) = C/\inf_i \rho^p(x,x_i) \quad .$$

The high density clusters of f_n are the maximal connected subsets of unions of spheres of the previous paragraph, the single linkage clusters. The nearest neighbor density estimate is a poor estimate, not consistent for the true density; it is remarkable that single linkage clusters retain approximate consistency. Following Wishart (1969) and Ling (1973) we should use high density clusters corresponding to some form of kth nearest neighbor density estimation, where for consistency $k \to \infty$ as $n \to \infty$ but $k/n \to 0$. For example, we define distance between clusters in a joining algorithm by the kth smallest among distances between pairs of points in the two clusters. This clustering method is the analogue of kth nearest neighbor discriminant procedures, in which a new point is allocated to the class that appears most frequently in its k nearest neighbors. See, for example, Wong (1982).

Another characterization of single linkage clusters is through the *ultrametric*

$$\rho^*(x,y) = \inf_{x=x_1,x_2,\ldots,x_k=y} \sup_i \rho(x_i,x_{i+1}) \quad .$$

The ultrametric satisfies $\rho^*(x,y) \leqslant \sup[\rho^*(x,z), \rho^*(y,z)]$ and determines a family of clusters $\{x \mid \rho^*(x,y) \leqslant C\}$ for various C,y, that turn out to be the single linkage clusters. See Wong and Lane (1983).

Finally, the minimum spanning tree is the graph of minimum total length connecting the sample points. Gower and Ross (1969) show that single linkage clusters are the connected sets obtained by successively deleting, largest to smallest, the links of the minimum spanning tree. Thus single linkage computation and asymptotics are intimately related to minimum spanning tree computation and asymptotics.

5. Mixtures

The population density

$$f = \sum_{i=1}^{k} p_i f_i$$

is a mixture of *components* f_i in proportions p_i. This may be viewed as a model for k clusters. The f_i and p_i are unidentified without some further constraints. The usual assumption is that each f_i is a member of the same parametric family, the multivariate normal (for example Wolfe 1970, Day 1969, Dick and Bowden 1973); however the mixture model may also be applied to general sample spaces, not only to points in p dimensional euclidean space. In discriminant analysis, a random observaion X is associated with a classification I into one of k classes. Suppose that I takes the value i with probability p_i, and that X given $I = i$ has density f_i. Then the marginal density of X is just $f = \sum p_i f_i$. In discriminant analysis we know X and I; in clustering we know only X; thus the mixture model for clustering corresponds to the marginal probability model for discriminant analysis. Lack of knowledge of the classification variable makes the general mixture model unidentifiable however, so further constraints are needed for clustering.

Let $p(i \mid x) = p_i f_i(x) / \sum p_i f_i(x)$ denote the posterior probability that the observation x belongs to class i. These posterior probabilities are useful in maximum likelihood estimation for the model

$$f(x) = \sum p_i f_i(x, \theta_i)$$

where the f_i are known up to the parameter θ_i taking values in r dimensional euclidean space. Assume the f_i are differentiable with respect to θ_i. Then the maximum likelihood estimates for p_i, θ_i based on observations $X_1, X_2, \ldots, X_n$ satisfy

$$(i) \qquad \sum_{j=1}^{n} p(i \mid X_j)\, d/d\theta_i \{f_i(X_j, \theta_i)\} = 0$$

$$(ii) \qquad p_i = \sum_{j=1}^{n} p(i \mid X_j)/n$$

The estimation proceeds in alternating steps: given $p(i \mid X_j)$, estimate θ_i weighting the observation X_j by its probability of belonging to the ith component; then given these estimates θ_i, estimate p_i and $p(i \mid X_j)$; then repeat the first step.

Rao (1948) appears to have been the first to use maximum likelihood for normal mixtures. See also Day (1969), Wolfe (1970), Hosmer (1973),

Everitt and Hand (1981). We can't show that the alternating procedure leads to the true maximum likelihood estimates, or even that the likelihood increases after each step, even in the simple case of normal mixtures in one dimension. The standard maximum likelihood regularity conditions do not hold in the normal case, and the usual asymptotic consistency and distribution results do not always hold.

For well separated clusters, the posterior probabilities $p(i \mid X_j)$ are all near 0 or 1, and the maximum likelihood solution is approximated by dividing the observations into k clusters, estimating θ_i by maximum likelihood separately within the clusters, and finding that division into k clusters that maximizes the product of the likelihoods. This procedure is the strict maximum likelihood estimate for the model in which the observations $\{X_j\}$ are supposed drawn from components $\{I_j\}$, and the components are regarded as unknown *parameters*. The likelihood is

$$L[X_1, \ldots, X_n, I_1, \ldots, I_n, \theta_1, \theta_2, \ldots, \theta_k] = \Pi\, f_{I_j}(X_j, \theta_{I_j}) \quad .$$

In practice we can rarely afford to search all partitions, but we use an alternating step algorithm:

(i)′ : Given I_j, select θ_i to maximize $\prod_{I_j=i} f_i\,(X_j, \theta_i)$

(ii)′ : Given θ_i, select I_j to maximize $f_{I_j}\,(X_j, \theta_{I_j})$.

Thus given the clusters we estimate θ_i by maximum likelihood, and given the θ_i we specify cluster membership to make X_j most likely. This is the alternating method for mixture maximum likelihood when the $P(i \mid X_j)$ are all zero or one.

In the particular case when $f_i(X) = (2\pi)^{-p/2} \exp[-\frac{1}{2}(X-\mu_i)'(X-\mu_i)]$, the above algorithm has a simplified form known in the clustering literature as k-means. (See for example, MacQueen 1967 and Hartigan 1975.) We select μ_i to be the mean of the observations in the ith cluster, and we allocate the observation X_j to that cluster i that minimizes the distance between X_j and μ_i.

6. The Number of Clusters: Modes

In the high density clustering model, we associate a family of clusters with each mode of the density f: f has a mode at m if there is a neighborhood M of m such that $f(x) \leqslant f(m)$ for $x \in M$, and $f(x) < f(m)$ for x in the boundary of M. There are disjoint high density clusters only if f is

multimodal. Thus we can test for the presence of clusters by testing for multimodality.

For X one dimensional, unimodal and bimodal densities may be fit by maximum likelihood giving a likelihood ratio test for bimodality, but it is difficult to handle the large contributions to the likelihood made by small intervals between neighboring observations. A better test is the *dip* test, which measures the maximum difference between the empirical distribution function, and the unimodal distribution function chosen to minimize that maximum difference. The dip approaches zero for unimodal distributions, and some non-zero value for multimodal distributions, as the sample size increases. It is therefore consistent for distinguishing unimodal from multimodal distributions. It is argued in Hartigan and Hartigan (1984) that the uniform is the appropriate null unimodal distribution, because the dip is asymptotically stochastically larger for the uniform than for other unimodal distributions; the asymptotic distribution of the dip and some empirically determined distributions for finite sample sizes are given in that paper.

The dip does not generalize simply to many dimensions. The minimum spanning tree provides a kind of ordering of the n sample points that may be used to generate an analogue of the dip statistic: select a particular sample point x_0 to be the mode or root, and consider probability distributions P supported by the links of the minimum spanning tree. A unimodal P has a density, with respect to the uniform distribution on the tree, that is a non-decreasing function of x as x moves toward the root. At each point x on the tree define a distribution function value $F(x)$ to be the probability that a random point X is such that x lies between X and x_0. Let F_n be the empirical distribution function corresponding to the empirical distribution which gives each sample point probability $1/n$. Define $d(F,F_n) = \sup_x | F_n(x) - F(x) |$,

$$D(F_n,x_0) = \inf_F d(F,F_n) \text{ where } F \text{ is unimodal with mode } x_0 \ ,$$

$$DIP(F_n) = \inf_{x_0} D(F_n,x_0) \ .$$

This procedure locates an optimal mode x_0, and states how well the data fit the unimodal hypothesis, $DIP(F_n)$. In the one dimensional case the usual definition of dip gives the same value. The asymptotic behavior of the multivariate version is unknown.

7. The Number of Clusters: Components

If the components of a multivariate normal mixture are sufficiently well separated, there will be one mode for each component. In this case the

number of clusters is the number of components or the number of modes, but in general the number of modes is fewer than the number of components, so testing for the presence of more than one component is less conservative than testing for the presence of more than one mode.

Wolfe (1971) considers the likelihood ratio test for say one component against two, but notes that the regularity conditions which are usually required for the log likelihood ratio to be chi square are not fulfilled. See also Binder (1978) and Hartigan (1977).

Consider the simplest case: $X_1, \ldots, X_n$ sampled from $N(0,1)$ under the null hypothesis, and from $(1-p)N(0,1) + pN(\mu,1)$ for some $0 \leqslant p \leqslant 1$, $-\infty < \mu < \infty$ under the alternative hypothesis. Let $Z_i = \exp(X_i\mu - \frac{1}{2}\mu^2) - 1$. Then the likelihood $\propto L(p,\mu) = \Pi_{i=1}^n (1+pZ_i)$. Note that Z_i has mean 0 and variance $e^{\mu^2}-1$; if $z_i(\mu)$ and $Z_i(\mu')$ denote the Z-values computed for μ and μ', then $cov(Z_i(\mu), Z_i(\mu')) = e^{\mu\mu'}-1$.

For each fixed μ, $ln\ L(p,\mu)$ is a concave function of p that has maximum value 0 if $\sum Z_i < 0$ but maximum value approximately $\frac{1}{2}(\sum Z_i)^2/\sum Z_i^2)$ otherwise. Thus asymptotically, $L(\mu) = \sup_p ln\ L(p,\mu)$ is equal to zero with probability $\frac{1}{2}$ and to $\frac{1}{2}\chi_1^2$ with probability $\frac{1}{2}$. The $\frac{1}{2}\chi_1^2$ would be expected from usual likelihood asymptotics.

If μ and μ' are widely separated, $Z_i(\mu)$ and $Z_i(\mu')$ are nearly uncorrelated, and so asymptotically $L(\mu)$ and $L(\mu')$ are nearly independent. Thus $\sup_\mu L(\mu)$ is greater than the maximum of k nearly independent $L(\mu)$ for each k. Thus $\sup_\mu L(\mu)$ is asymptotically infinite.

The likelihood ratio test does not therefore follow the usual asymptotics, and is not conservative: the usual significance test will (with probability 1 asymptotically) reject the hypothesis of a single component when only a single component is present. For each μ, $\sup_p L(p,\mu)$ has asymptotically the same distribution, and these distributions are nearly independent for well separated μ; maximum likelihood computations will therefore be difficult; we can expect to see local maxima of $\sup_p L(p,\mu)$ near every value of μ.

If μ has prior density normal with mean 0 and variance 1, large values of μ are inhibited and the maximum posterior density will occur only with μ moderate. The corresponding ratio test may have better asymptotic behavior than the likelihood ratio test. More generally, if the mixture model has components with means $\mu_1, \mu_2, \ldots, \mu_k$ we might assume the μ_k to be a priori a sample from a normal; this prevents the artificially large separation of the μ's that occurs in likelihood estimation and testing.

The behavior of the likelihood ratio statistic in the k-means case has been examined in one dimension by Hartigan (1978) and in higher dimensions by Pollard (1982).

8. Ultrametric and Evolutionary Distances

Assume that there are N objects, and $N(N-1)/2$ distances between pairs of objects. From these distances we wish to form clusters of close objects. One way to go about constructing the clusters is to require that the distances satisfy certain properties in the final clustering. For example, all distances within two disjoint clusters must be smaller than all distances between clusters. Or, each pair of points in the same cluster must be connected by a chain of points such that neighboring points in the chain are closer than some neighboring points in a chain connecting points in different clusters. (This definition leads to single linkage.) Another way is to suppose that the clusters correspond to some ideal distance matrix, and to attempt to approximate the given distance matrix d with a best fitting cluster distance D. For example, hierarchical clustering might correspond to an *ultrametric* D, a distance satisfying $D(i,j) \leqslant sup\,[D(i,k), D(j,k)]$ and we would find the ultrametric D closest to d. See Hartigan (1967), Johnson (1967) and Jardine, Jardine and Sibson (1967). Another plausible definition, the *evolutionary* model from Fitch and Margoliash (1967), is based on an evolutionary tree generating the objects. The distance between any pair of objects is the sum of links on the unique path connecting them in the tree. If there exists an ancestor in the tree such that all points are equidistant (in sum of links) from the ancestor, then this evolutionary distance or ultrametric can be fitted by regression methods; the hard part is searching for the best tree.

Baker (1974) has considered probability models in which an observed distance matrix d varies by some amount from an ultrametric D, and has investigated empirically how well the various hierarchical techniques recover the true ultrametric D. The results are opposite to those obtained using the high density model: complete linkage does well, and single linkage poorly.

Euclidean distances in p dimensional space will form an ultrametric distance matrix on at most $(p+1)$ points. For a density f, we can construct an ultrametric by

$$D(x,y) = \min_{C} \max_{\mu \in C} 1/f(u)$$

where C is any path connecting x and y. Thus x and y are close if they can be connected by a path of high density, or equivalently if they lie together in a high density cluster. In fitting such an ultrametric to objects in p dimensional space we would use only the small distances between objects to obtain an estimate of the density f. Single linkage works only with the small distances, whereas complete linkage depends on the large distances. This may be the explanation for Baker's results favoring complete linkage, in that he requires the fitted ultrametric to be close to the true ultrametric when

averaged over *all* distances, and the large distances are neglected by single linkage. In practice, the large distances deviate most from the fitted ultrametric (however fitted) and it seems correct to downweight their contribution. Theoretically, if we wish to allow clusters of arbitrary shape and size, it also seems impossible to give large distances much weight. Perhaps we should fit an ultrametric D to minimize

$$\sum w(D)\ [d(i,j) - D(i,j)]^2 / \sum w(D)$$

where $w(D)$ was small or zero for D large. This moves single linkage a little way towards average linkage. More weight should be given to the large distances in high dimensional spaces.

Let the objects $1, \ldots, n$ be generated by an evolutionary tree, beginning at some ancestor, 0. For a particular measurement X taking values X_i on the objects, assume that X changes in time t, $t + \Delta\ t$ on a particular link of the tree, by an amount that has mean 0, variance $\sigma^2 \Delta\ t$, and is uncorrelated with changes in different intervals or links.

Then, letting PY denote the average value of the random variable Y,

$$P(X_i - X_j)^2 = 2\sigma^2\ t_{ij}$$

where t_{ij} is the time since i and j evolved from their most recent ancestor, so $P(X_i - X_j)^2$ is an ultrametric!

If we had used different rates of evolution in the different links of the tree, so that the changes in X had variance $\sigma_l^2 \Delta\ t$ for link l, then $P(X_i - X_j)^2$ would be an evolutionary distance.

Suppose that X is normal, and there are p independent samples of X. (Here the number of objects is fixed, and the measurements are assumed sampled from an infinite population of possible measurements; it will require careful standardization to achieve something like this in practice.) Then

$$\sum_{r=1}^{p} (X_i^r - X_j^r)^2 \sim 2\sigma^2\ t_{ij}\ \chi_p^2$$

where $\sim$ means "is distributed as."

$$d(i,j) \sim \sqrt{2\sigma^2\ t_{ij}\ \chi_p^2}$$

Let $D(i,j) = \sqrt{2\sigma^2\ t_{ij}\ p}$, an ultrametric. For large p

$$d(i,j) \approx D(i,j)[1+N(0, \tfrac{1}{2}p)]$$

This suggests fitting the ultrametric D by minimizing

$$\sum [D(i,j)-d(i,j)]^2/D^2(i,j)$$

which downweights the large distances nicely, but probably not enough. We have to take note also of the high correlation between the large distances, caused by them sharing a high fraction of their paths through the tree. We can compute these, but the criterion to be minimized is then a complex quadratic in $D-d$.

References

BAKER, F.B. (1974), "Stability of Two Hierarchical Grouping Techniques, Case I: Sensitivity to Data Errors," *Journal of the American Statistical Association, 69*, 440-445.

BINDER, D.A. (1978), Comment on 'Estimating Mixtures of Normal Distributions and Switching Regressions', *Journal of the American Statistical Association, 73*, 746-747.

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

DAY, N.E. (1969), "Estimating the Components of a Mixture of Normal Distributions," *Biometrika, 56*, 463-474.

DICK, N.P., and BOWDEN, D.C. (1973), "Maximum Likelihood Estimation for Mixture of Two Normal Distributions," *Biometrics, 29*, 781-790.

EVERITT, B.S., and HAND, D.J. (1981), *Finite Mixture Distributions*, London: Chapman and Hall.

FITCH, W.M., and MARGOLIASH, E. (1967), "Construction of Phylogenetic Trees," *Science N.Y., 155*, 279-284.

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

HARTIGAN, J.A. (1967), "Representation of Similarity Matrices by Trees," *Journal of the American Statistical Association, 62*, 1140-1158.

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

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

HARTIGAN, J.A. (1978), "Asymptotic Distributions for Clustering Criteria," *The Annals of Statistics, 6*, 117-131.

HARTIGAN, J.A. (1981), "Consistency of Single Linkage for High Density Clusters," *Journal of the American Statistical Association, 76*, 388-394.

HARTIGAN, J.A., and HARTIGAN, P.M. (1984), "The Dip Test of Multimodality," *The Annals of Statistics*, submitted.

HOSMER, D.W. (1973), "A Comparison of Iterative Maximum Likelihood Estimates of the Parameters of a Mixture of Two Normal Distributions under Three Different Types of Sample," *Biometrics, 29*, 761-770.

JARDINE, C.J., JARDINE, N., and SIBSON, R. (1967), "The Structure and Construction of Taxonomic Hierarchies," *Math. Biosciences, 1*, 173-179.

JOHNSON, S.C. (1967), "Hierarchical Clustering Schemes," *Psychometrika, 32*, 241-254.

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

MAC QUEEN, J. (1967), "Some Methods for Classification and Analysis of Multivariate Observations," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1*, 281-297.

POLLARD, D. (1982), "A Central Limit Theorem for k-means Clustering," *Annals of Probability, 10*, 919-926.

RAO, C.R. (1948), "The Utilization of Multiple Measurements in Problems of Biological Classification," *Journal of the Royal Statistical Society, Series B, 10*, 159-203.

SMYTHE, R.T., and WIERMAN, J.C. (1978), "First Passage Percolation on the Square Lattice," *Lecture Notes in Mathematics*, 671, Berlin: Springer-Verlag.

WISHART, D. (1969), "Mode Analysis: A Generalization of Nearest Neighbor Which Reduces Chaining Effects," in *Numerical Taxonomy*, ed. A. J. Cole, London: Academic Press.

WOLFE, J.H. (1970), "Pattern Clustering by Multivariate Analysis," *Multivariate Behavioral Research, 5*, 329-350.

WOLFE, J.H. (1971), "A Monte-Carlo Study of the Sampling Distribution of the Likelihood Ratio fro Mixtures of Multinormal Distributions," *Research Memorandum, 72-2*, Naval Personnel and Research Training Laboratory, San Diego.

WONG, M.A. (1982), "A Hybrid Clustering Algorithm for Identifying High Density Clusters," *Journal of the American Statistical Association, 77*, 841-847.

WONG, M.A., and LANE, T. (1983), "A kth Nearest Neighbor Clustering Procedure," *Journal of the Royal Statistical Society, Series B, 45*, 362-368.