# Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample

Werner Stuetzle *
Department of Statistics
University of Washington
wxs@stat.washington.edu

February 12, 2003

**Abstract**

We present *runt pruning*, a new clustering method that attempts to find modes of a density by analyzing the minimal spanning tree of a sample. The method exploits the connection between the minimal spanning tree and nearest neighbor density estimation. It does not rely on assumptions about the specific form of the data density (e.g., normal mixture) or about the geometric shapes of the clusters, and is computationally feasible for large data sets.

Keywords: Two-way, two-mode data; nearest neighbor density estimation; single linkage clustering; runt test; mixture models.

---

*Supported by NSF grant DMS-9803226 and NSA grant 62-1942. Work partially performed while on sabbatical at AT&T Labs - Research. Author's address: Department of Statistics, Box 354322, University of Washington, Seattle, WA 98195-4322; email: wxs-stat.washington.edu

# 1 Introduction

The goal of clustering is to identify distinct groups in a two-mode, two-way dataset $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset R^m$. For example, when presented with (a typically higher dimensional version of) a data set like the one in Figure 1 we would like to detect that there appear to be (perhaps) five or six distinct groups, and assign a group label to each observation.

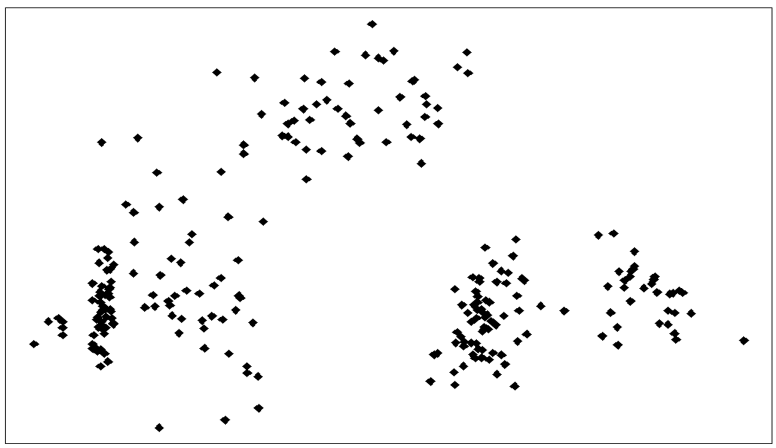

Figure 1: Data set with 5–6 apparent groups.

To cast clustering as a statistical problem we regard the data $\mathbf{x}_1, \ldots, \mathbf{x}_n$ as an iid sample form some unknown probability density $p(\mathbf{x})$. There are two statistical approaches to clustering. The parametric approach (Fraley and Raftery 1998, 1999; McLachlan and Peel 2000) is based on the assumption that each group $g$ is represented by a density $p_g(\mathbf{x})$ that is a member of some parametric family, such as the multivariate Gaussian distributions. The density $p(\mathbf{x})$ then is a mixture of the group densities, and the number of mixture components and their parameters are estimated from the data. Observations can be labeled using Bayes's rule.

In contrast, the nonparametric approach adopted in this paper is based on the premise that groups correspond to modes of the density $p(\mathbf{x})$. The goal then is to find the modes and assign each observation to the "domain of attraction" of a mode. Searching for modes as a manifestation of the presence of groups was first advocated in D. Wishart's (1969) paper on *Mode Analysis*. According to Wishart, clustering methods should be able to detect and "resolve distinct data modes, independently of their shape and variance".

Hartigan (1975, Section 11; 1981) expanded on Wishart's idea and made it more precise by introducing the notion of *high density clusters*. Define the level set $L(\lambda; p)$ of a density $p$ at level $\lambda$ as the subset of the feature space for which the density exceeds $\lambda$:

$$L(\lambda; p) = \{\mathbf{x} \mid p(\mathbf{x}) > \lambda\}.$$

The high density clusters at level $\lambda$ are the maximally connected subsets of $L(\lambda; p)$.

Hartigan also pointed out that the collection of high density clusters has a hierarchical structure: for any two clusters $A$ and $B$ (possibly at different levels) we have either $A \subset B$ or $B \subset A$ or $A \cap B = \emptyset$. This hierarchical structure is summarized by the *cluster tree* of $p$. Each node $N$ of the tree represents a subset $D(N)$ of the support $L(0; p)$ of $p$ — a high density cluster of $p$ — and is associated with a density level $\lambda(N)$. The cluster tree is easiest to define recursively. The root node represents the entire support of $p$, and has associated density level $\lambda(N) = 0$. To determine the descendents of a node $N$ we find the lowest level $\lambda_d$ for which $L(\lambda; p) \cap D(N)$ has two or more connected components. If there is no such $\lambda_d$ then $p$ has only one mode in $D(N)$, and $N$ is a leaf of the tree. Otherwise, let $C_1, \ldots, C_k$ be the connected components of $L(\lambda_d; p) \cap D(N)$. If $k = 2$ (the usual case) we create daughter nodes representing the connected components $C_1$ and $C_2$, both with associated level $\lambda_d$, and apply the definition recursively to the daughters. If $k > 2$ we create daughter nodes representing $C_1$ and $C_2 \cup \cdots \cup C_k$ and recurse.

Figure 2 shows a density and the corresponding cluster tree. Estimating the cluster tree is a fundamental goal of nonparametric cluster analysis.

## 1.1 Previous work

Several previously suggested clustering methods can be described in terms of levels sets and high density clusters.

Probably the earliest such method is Wishart's (1969) *one level mode analysis*. The goal of one level mode analysis is to find the high density clus-

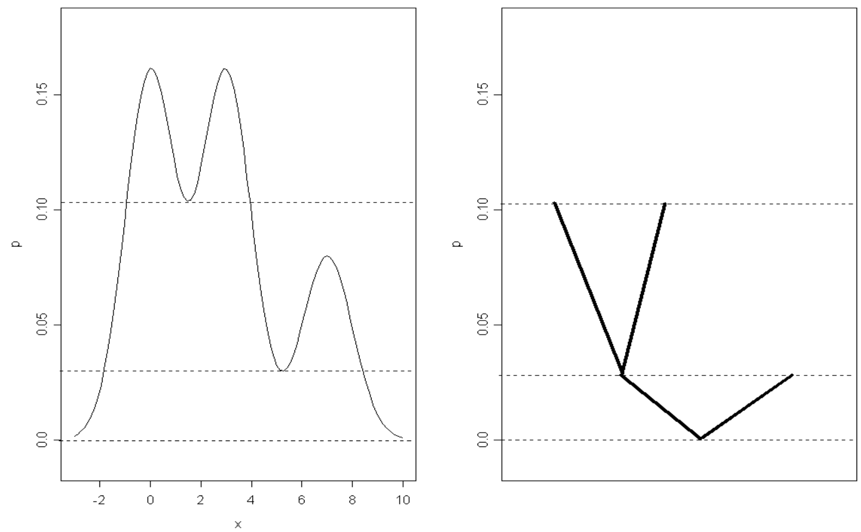


Figure 2: Density and corresponding tree of high density clusters.

ters at a given density level $\lambda$ chosen by the user. The idea is to first compute a kernel density estimate $\hat{p}$ (Silverman 1986, Chapter 4) and set aside all observations with $\hat{p}(\mathbf{x}_i) \leq \lambda$, i.e., all observations not in the level set $L(\lambda; \hat{p})$. If the population density has several well separated high density clusters at level $\lambda$ then the remaining high density observations should fall into clearly separated groups. Wishart suggests using single linkage clustering of the high density observations to identify the groups. One level mode analysis anticipates some of the "sharpening" ideas later put forth by P.A. Tukey and J.W. Tukey (1981).

A reincarnation of one level mode analysis is the DBScan algorithm of Ester, Kriegel, Sander, and Xu (1996). DBScan consists of four steps: (a) for each data point calculate a kernel density estimate using a spherical uniform kernel with radius $r$; (b) choose a density threshold $\lambda$ and find the observations with $\hat{p}(\mathbf{x}_i) > \lambda$; (c) construct a graph connecting each high density observation to all other observations within distance $r$; (d) define the clusters to be the connected components of this graph. All observations not within distance $r$ of a high density observation are considered "noise".

A weakness of one level mode analysis is apparent from Figure 2. The degree of separation between connected components of $L(\lambda; p)$, and therefore of $L(\lambda; \hat{p})$, depends critically on the choice of the cut level $\lambda$, which is left to the user. Moreover, there might not be a single value of $\lambda$ that reveals all the modes.

Citing the difficulty in choosing a cut level, Wishart (1969) proposed *hierarchical mode analysis*, which can be regarded as a heuristic for computing the cluster tree of a kernel density estimate $\hat{p}$, although it appears that Wishart did not view it thus. (The word "tree" does not occur in the section of his paper on hierarchical mode analysis.) We use the term "heuristic" because there is no guarantee that hierarchical mode analysis will indeed correctly compute the cluster tree of $\hat{p}$ as defined above. Wishart's (1969) algorithm constructs the tree by iterative merging (i.e., is an agglomerative algorithm). It is quite complex, probably because its iterative approach is not well matched to the tree structure it is trying to generate.

The basic weakness of one level mode analysis was also noted by Ankerst, Breuning, Kriegel, and Sander (1999) who proposed OPTICS, an algorithm for "Ordering Points to Identify the Clustering Structure". OPTICS generates a data structure that allows one to calculate efficiently the result of DBScan for any desired density threshold $\lambda$. It also produces a graphical summary of the cluster structure. The idea behind their algorithm is hard to understand.

## 1.2 Outline of runt pruning

An obvious way of estimating the cluster tree of a density $p$ from a sample is to first compute a density estimate $\hat{p}$ and then use the cluster tree of $\hat{p}$ as an estimate for the cluster tree of $p$. A difficulty with this approach is that for most density estimates computing the cluster tree seems computationally intractable. To determine the number of connected components of a level set $L(\lambda; \hat{p})$ one would have to rely on heuristics, like the ones suggested by Wishart (1969) and Ester *et al.* (1996), which is at the very least an esthetic drawback. A notable exception is the nearest neighbor density estimate

$$\hat{p}_1(\mathbf{y}) = \frac{1}{n\, V\, d(\mathbf{y}, \mathcal{X})^p}\,,$$

where $V$ is the volume of the unit sphere in $R^m$ and $d(\mathbf{y}, \mathcal{X}) = \min_i d(\mathbf{y}, \mathbf{x}_i)$. In Section 2 we show that the cluster tree of the nearest neighbor density estimate is isomorphic to the single linkage dendogram. The argument exploits a connection between the minimal spanning tree (MST) and nearest neighbor density estimation first pointed out by Hartigan (1985).

The nearest neighbor density estimate has some undesirable properties. For example, it has a high variance and it cannot be normalized. As we are not interested in estimating the density itself but rather its cluster tree, these flaws are not necessarily fatal. However, it also has a singularity at every data point, leading to a cluster tree with as many leaves as there are observations. Therefore the cluster tree has to be pruned.

Our pruning method, *runt pruning*, is based on the *runt test* for multimodality proposed by Hartigan and Mohanty (1992). In Section 3 we describe runt pruning, provide a heuristic justification for the method, and present an algorithm.

In Section 4 we compare runt pruning to the standard single linkage method for extracting clusters from a MST. In Section 5 we show runt pruning in action, illustrate diagnostic tools that can be helpful in choosing the runt size threshold determining tree size, and compare its performance to other clustering methods. In Section 6 we discuss some general issues such as the underlying assumptions and the relative merits of parametric and nonparametric clustering methods. Section 7 concludes the paper with a summary and ideas for future work.

# 2 Nearest neighbor density estimation and the Euclidean minimal spanning tree

In this section we show that the cluster tree of the nearest neighbor density estimate can be obtained from the MST of the data, and that it is isomorphic to the single linkage dendogram. For a given density level $\lambda$, define

$$r(\lambda) = \left( \frac{1}{n\,V\,\lambda} \right)^{\frac{1}{p}} .$$

By definition, $\hat{p}_1(\mathbf{y}) > \lambda$ iff $d(\mathbf{y}, \mathcal{X}) < r(\lambda)$, and therefore $L(\lambda; \hat{p}_1)$ is the union of (open) spheres of radius $r(\lambda)$, centered at the observations:

$$L(\lambda; \hat{p}_1) = \bigcup_i \overset{\circ}{S} \left( \mathbf{x}_i, r(\lambda) \right) .$$

Let $T$ be the Euclidean MST of $\mathcal{X}$, that is, the graph with shortest total edge length connecting all the observations. Breaking all MST edges with length $\geq 2r(\lambda)$ defines a partition of the MST into $k$ subtrees $T_1, \ldots, T_k$ (possible $k = 1$) and a corresponding partition of the observations into subsets $\mathcal{X}_1, \ldots, \mathcal{X}_k$.

**Proposition 1:** (Hartigan 1985): The sets $L_i = \bigcup_{i \in \mathcal{X}_j} \overset{\circ}{S}(\mathbf{x}_i, r(\lambda))$ are the connected components of $L(\lambda; \hat{p}_1)$.

**Proof:** Each of the sets $L_i$ is connected, because by construction the maximum edge length of the corresponding MST fragment $T_i$ is smaller than $2r(\lambda)$, and therefore the MST fragment is a subset of $L_i$.

On the other hand, $L_i$ and $L_j$ are disconnected for $i \neq j$. Otherwise there would have to be observations $\mathbf{x}^*$ and $\mathbf{x}^{**}$ in $\mathcal{X}_i$ and $\mathcal{X}_j$, respectively, with $d(\mathbf{x}^*, \mathbf{x}^{**}) < 2r(\lambda)$. We could then break an edge of length $\geq 2r(\lambda)$ in the MST path connecting fragments $T_i$ and $T_j$ and insert an edge connecting $\mathbf{x}^*$ and $\mathbf{x}^{**}$, thereby obtaining a spanning tree of smaller total edge length. This contradicts the assumption that $T$ was the MST.

Proposition 1 implies that we can compute the cluster tree of the nearest neighbor density estimate by breaking the longest edge of the MST, thereby splitting the MST into two subtrees, and then applying the splitting operation recursively to the subtrees. Gower and Ross (1969) show that this algorithm finds the single linkage dendogram, which demonstrates that the cluster tree of the nearest neighbor density estimate and the single linkage dendogram are isomorphic.

# 3 Runt pruning

The nearest neighbor density estimate has a singularity at every observation, and consequently its cluster tree — the single linkage dendogram — has as many leaves as there are observations and is a poor estimate for the cluster tree of the underlying density. It has to be pruned.

Runt pruning is based on the *runt test* for multimodality proposed by Hartigan and Mohanty (1992). They define the *runt size* of a dendogram node $N$ as the smaller of the number of leaves of the two subtrees rooted at $N$. If we interpret the single linkage dendogram as the cluster tree of the nearest neighbor density estimate $\hat{p}_1$, then a node $N$ and its daughters represent high density clusters of $\hat{p}_1$. The runt size of $N$ can therefore also be regarded as the smaller of the number of observations falling into the two daughter clusters. As each node of the single linkage dendogram corresponds to an edge of the MST, we can also define the runt size for an MST edge $e$: Break all MST edges that are as long or longer than $e$. The two MST nodes originally joined by $e$ are the roots of two subtrees of the MST, and the runt size of $e$ is the smaller of the number of nodes of those subtrees.

The idea of runt pruning is to consider a split of a high density cluster of $\hat{p}_1$ into two connected components to be "real" or "significant" if both daughters contain a sufficiently large number of observations, i.e., if the runt size of the corresponding dendogram node is larger than some threshold. The runt size threshold controls the size of the estimated cluster tree.

## 3.1 Heuristic justification

MST edges with large runt size indicate the presence of multiple modes, as was first observed by Hartigan and Mohanty (1992). We can verify this assertion by considering a simple algorithm for computing a MST: Define the distance between two groups of observations $G_1$ and $G_2$ as the minimum distance between observations:

$$d(G_1, G_2) = \min_{\mathbf{x} \in G_1} \min_{\mathbf{y} \in G_2} d(\mathbf{x}, \mathbf{y}) \ .$$

Initialize each observation to form its own group. Find the two closest groups, add the shortest edge connecting them to the graph, and merge the two groups. Repeat this merge step until only one group remains. The runt size of an edge is the size of the smaller of the two groups connected by the edge.

Suppose now that the underlying density is multimodal. Initial merges tend to take place in high density regions where interpoint distances are small, and tree fragments will tend to grow in those regions. Eventually, those fragments will have to be joined by edges, and those edges will have large runt sizes, as illustrated in Figure 3. Panel (a) shows a sample from a bimodal density, and panel (b) shows the corresponding rootogram of runt sizes. (A rootogram is a version of a histogram where the square roots of the counts are plotted on the vertical axis.) There is one edge with runt size 75. Panel (c) shows the MST after removal of all edges with length greater than the length of the edge with largest runt size. Note the two large tree fragments in the two high density regions.

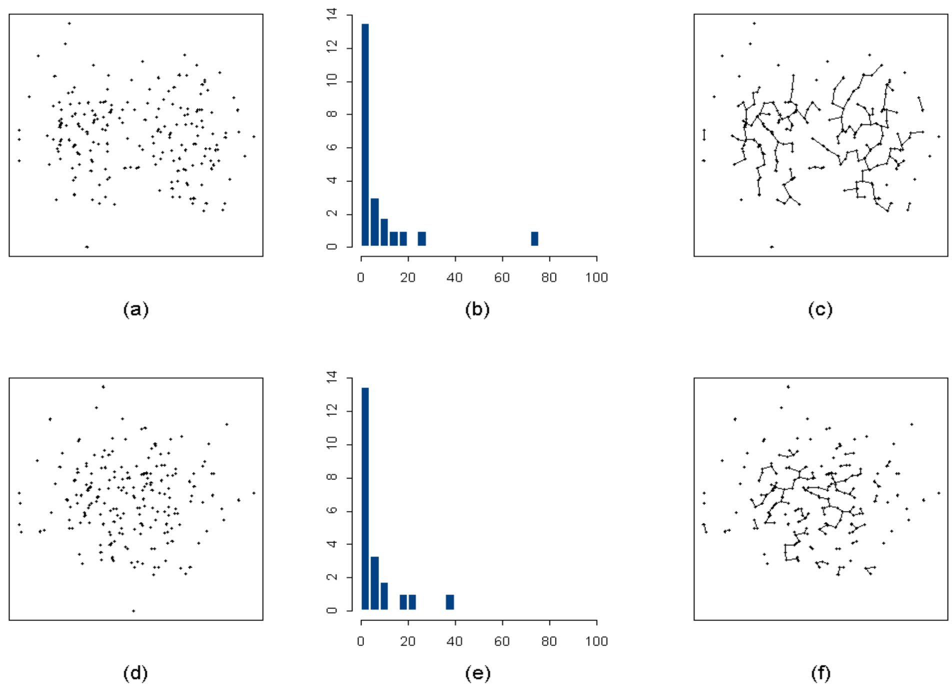


Figure 3: (a) Sample from bimodal density; (b) Rootogram of runt sizes; (c) MST with longest edges removed; (d)...(f) Corresponding plots for unimodal density.

If the density is unimodal, on the other hand, then a single fragment will start in the area of highest density and grow toward the lower density fringe, where interpoint distances tend to be higher. This result is illustrated in panels (d)...(f) of Figure 3. The largest runt size here is 37. When all longer edges are removed, there is a large fragment in the high density area and a number of smaller fragments towards the fringe.

## 3.2 Algorithm

Our algorithm for constructing a pruned cluster tree of the nearest neighbor density estimate parallels exactly the recursive definition of a cluster tree.

Each node $N$ represents a high density cluster $D(N)$ of $\hat{p}_1$, a *sample cluster core* $\mathcal{X}(N)$ consisting of the observations in $D(N)$, and a subtree $T(N)$ of the MST, and is associated with a density level $\lambda(N)$. The root node represents the entire feature space, sample, and MST, and is associated with density level $\lambda = 0$.

To determine the descendents of a node $N$ we find the lowest density level $\lambda$ or, equivalently, the longest edge $e$ in $T(N)$ with runt size larger than our chosen threshold. If there is no such edge then $N$ is a leaf of the tree.

Otherwise, we create daughter nodes $N_r$ and $N_l$ associated with density level

$$\lambda(N_l) = \lambda(N_r) = \frac{2^m}{nV\|e\|^m}\,.$$

Breaking all edges of $T(N)$ with length $\geq \|e\|$ results in a subgraph of $T(N)$; the sample cluster cores $\mathcal{X}(N_l)$ and $\mathcal{X}(N_r)$ consist of the observations in the fragments rooted at the ends of $e$. The high density clusters $D(N_l)$ and $D(N_r)$ are unions of spheres of radius $\|e\|/2$ centered at the observations in $\mathcal{X}(N_l)$ and $\mathcal{X}(N_r)$, respectively. The trees $T(N_l)$ and $T(N_r)$ are obtained by breaking the edge $e$ of $T(N)$.

We refer to the observations in $T(N)$ as the *sample cluster* or, if there is no ambiguity, simply as the cluster represented by $N$. If $N_1, \ldots, N_k$ are the leaves of the cluster tree, then the corresponding clusters form a partition of the sample. The cluster cores $\mathcal{X}(N_i)$ are subsets of the corresponding clusters located in the high density regions.

# 4 Runt pruning and single linkage clustering

The standard method for extracting clusters from a MST is single linkage clustering: to create $k$ clusters, break the $k-1$ longest edges in the MST. This approach can be successful if the groups are clearly separated, i.e., if the Hausdorff distance between groups is large compared to the typical nearest neighbor distance. For an illustration, see the "Bullseye"" example in Section 5.2. However, in situations where the grouping is not quite as obvious, single linkage clustering tends to fail, and it has acquired a (deservedly) bad reputation. There are two reasons for this failure.

First, single linkage clustering tends to generate many small clusters because the longest edges of the minimal spanning tree will be in low density regions, which typically are at the periphery of the data: long edges tend to connect stragglers to the bulk.

Second, choosing a single edge length threshold for defining clusters is equivalent to looking at a single level set of the nearest neighbor density estimate. However, as Figure 2 illustrates, there are densities where no single level set will reveal all the modes. Therefore single linkage clustering cannot be "repaired" by simply discarding all small clusters and considering only the large ones as significant — the problem is more fundamental.

# 5 Examples

We present four examples. The first, simulated data with highly nonlinear cluster shapes, demonstrates that runt analysis can indeed find such structure for which other algorithms, like average linkage, complete linkage, and model-based clustering fail.

The second example, simulated data with spherical Gaussian clusters, is designed to be most favorable for model-based clustering and suggests that the performance penalty of runt pruning in such cases is not disastrous.

The third example, data on the chemical compositions of 572 olive oil samples from nine different areas of Italy, is used to illustrate how we might set a runt size threshold, and how we can use diagnostic plots to assess whether clusters are real or spurious.

The fourth example, 256-dimensional data encoding the shapes of handwritten digits, shows that runt pruning can be reasonably applied to high-dimensional data, despite the fact that it is based on a poor density estimate.

## 5.1 Comparing clustering methods

To evaluate clustering methods empirically we have to apply them to labeled data. We can then compare the partitions found by the various methods with the true partition defined by the labels. In simple situations, as in the "bulls-eye" example of Section 5.2, the comparison can be informal, but in general we want a figure of merit that does not rely on subjective judgments. This goal raises two questions: (a) how do we measure the degree of agreement between two partitions, and (b) how do we choose the size of the partition to be generated by the clustering method that we want to evaluate?

**Measuring agreement between partitions.** Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be two partitions of a set of $n$ objects. The partitions define a contingency table: let $n_{ij}$ be the number of objects that belong to subset $i$ of partition $\mathcal{P}_1$ and to subset $j$ of partition $\mathcal{P}_2$. We measure the agreement between $\mathcal{P}_1$ and $\mathcal{P}_2$ by the adjusted Rand index (Hubert and Arabie 1985) defined as

$$R = \frac{\sum_{ij}\binom{n_{ij}}{2} - \sum_i \binom{n_{i\cdot}}{2}\sum_j \binom{n_{\cdot j}}{2}/\binom{n}{2}}{\frac{1}{2}\left(\sum_i \binom{n_{i\cdot}}{2} + \sum_j \binom{n_{\cdot j}}{2}\right) - \sum_i \binom{n_{i\cdot}}{2}\sum_j \binom{n_{\cdot j}}{2}/\binom{n}{2}} .$$

Here $n_{i\cdot} = \sum_j n_{ij}$, and $n_{\cdot j}$ is defined analogously.

The adjusted Rand index has a maximum value of 1 which is achieved when the two partitions are identical up to re-numbering of the subsets. It has expected value 0 under random assignment of the objects to the subsets of $\mathcal{P}_1$ and $\mathcal{P}_2$ that leave the marginals $n_{i\cdot}$ and $n_{\cdot j}$ fixed.

**Choosing a partition size.** Choosing a partition size is a difficult issue, especially for nonparametric clustering methods, for which there is as yet no automatic method, and subjective judgment is required. To eliminate the subjective element from the comparisons, we decompose the clustering problem into two subproblems: (a) determining the number of groups, and (b) finding the groups, given their number. We compare the performance on subproblem (b), using two different rules for setting the number of groups. First, we have each method produce the true number of groups. Second, we generate a range of partitions of different sizes, calculate the adjusted Rand index for each of them, and then report the maximum value of the index achieved by the method and the corresponding partition size.

## 5.2 Nonlinear clusters — Bullseye

The data used in this example are shown in Figure 4(a). There are 500 observations uniformly distributed over the center of the bullseye and the ring. Figure 4(b) shows the 2-partition generated by runt pruning of the MST. Figures 4(c),..., 4(e) show the 2-partitions generated by single, average, and complete linkage, respectively. Figure 4(f) shows the 2-partition generated by fitting Gaussian mixtures. We used the software described in Fraley and Raftery (1999). The Gaussians were constrained to have equal spherical covariance matrices.

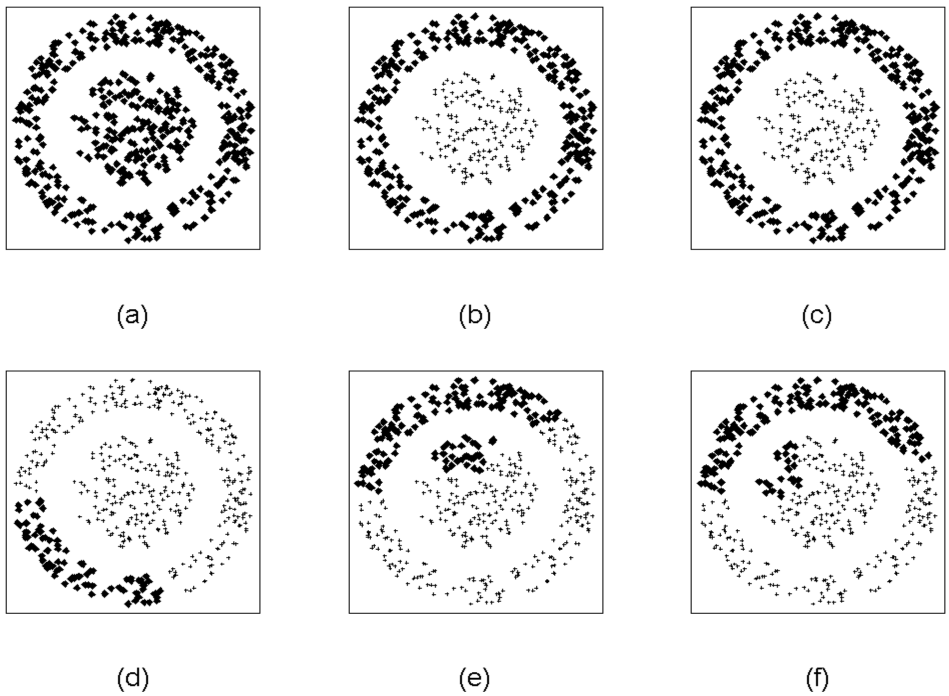


Figure 4: (a) Observations; (b)...(f) 2-partitions found by (b) runt analysis, (c) single linkage, (d) average linkage, (e) complete linkage, (f) model-based clustering.

Runt pruning correctly identifies the two clusters, as does single linkage clustering. Single linkage performs well in this example because the Hausdorff distance between the groups is large compared to the typical nearest neighbor distance. The other methods all fail in similar ways. This result is not surprising, because they are all designed to find roughly convex clusters.

## 5.3 Gaussian clusters — Simplex

The data in this example consist of spherical Gaussian clusters with common standard deviation $\sigma = 0.25$, centered at the vertices of the unit simplex in $p-1$ dimensions.

The first example is for $p = 3$, with cluster sizes 100, 200, and 300, respectively. The runt sizes of the MST are, in descending order, 194, 94, 29, 29, 20, 20, 20, 19, 15, 15, 12, 11, 11, 10, 10, .... There is a big gap after 94, suggesting the presence of three modes.

Figure 5(a) shows the cluster tree, with the root node selected. Panel (b) shows the descendents of the root node. In panel (c) we have selected the right daughter of the root node. Panel (d) shows its descendents.

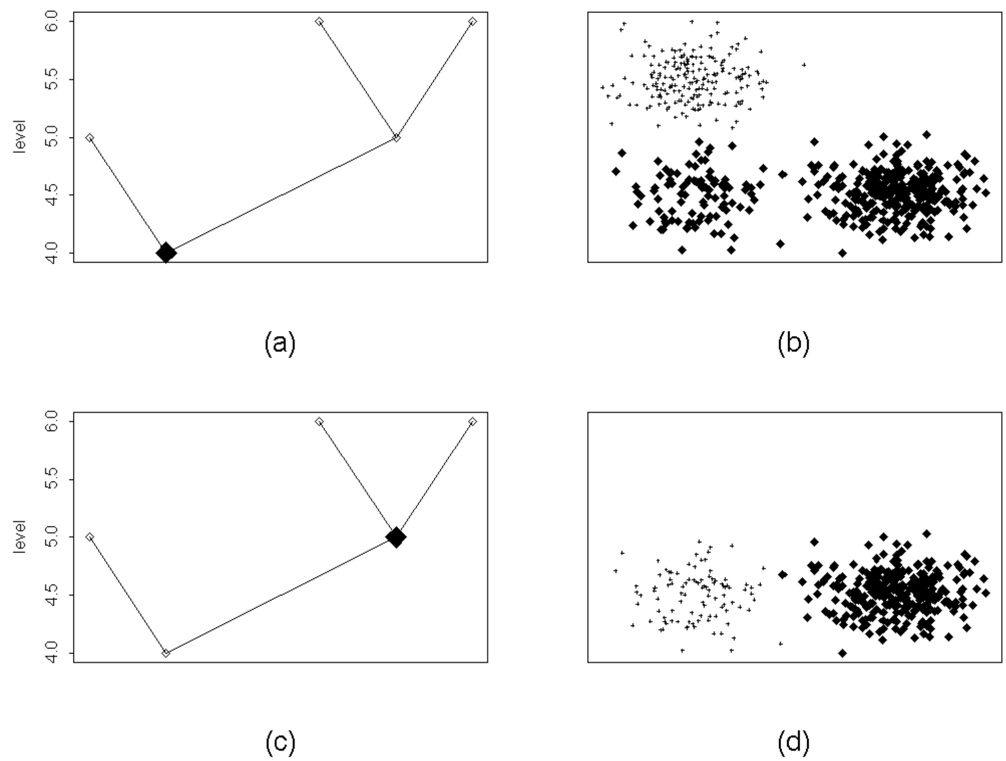


Figure 5: (a) Cluster tree for tri-modal Gaussian data with root node selected; (b) left and right descendents of root node; (c) cluster tree with right daughter of root node selected; (d) left and right descendents.

In this simple example, average and complete linkage, runt analysis, and model-based clustering all do an excellent job of finding the groups when asked to produce a 3-partition. The exception is single linkage clustering. Breaking the two longest edges of the MST results in two clusters of size 1 and one cluster containing the rest of the data.

| SL | AL | CL | RP | MC-EI | MC-VI |
|---|---|---|---|---|---|
| 0.0 (0.03) | 0.92 (0.01) | 0.92 (0.04) | 0.82 (0.05) | 0.93 (0.02) | 0.92 (0.01) |
| 0.0 (0.08) | 0.92 (0.01) | 0.92 (0.02) | 0.90 (0.03) | 0.93 (0.02) | 0.92 (0.01) |
| 16 | 7 | 7 | 8 | 7 | 7 |

Table 1: Comparison of single, average, and complete linkage, runt pruning, and two versions of model-based clustering for seven-dimensional simplex data. First row: adjusted Rand index if methods are made to generate seven clusters; second row: adjusted Rand index for optimal partition size; third row: optimal partition size. Numbers in parentheses are standard errors.

We next consider dimensionality $p = 7$, with cluster sizes 50, 60, …, 110. The runt sizes are 80, 73, 60, 38, 35, 26, 14, 10, 9, 9, 8, 7, 6, 6, 6.

Table 1 summarizes the performance of single, average, and complete linkage, runt analysis, and two versions of model-based clustering, fitting spherical Gaussians with equal variance and fitting spherical Gaussians with unequal variances. The first row of the table contains the values of the adjusted Rand index when the methods are asked to construct a 7-partition. The second row contains the optimal values of the index (optimized over partition size). Numbers in parentheses are standard errors obtained by half-sampling (Shao and Tu 1995, Section 5.2.2). All methods except single linkage clustering perform well, although runt pruning appears to fall off a little.

## 5.4 Olive oil data

The data for this example consist of measurements of eight chemical concentrations on 572 samples of olive oil. The samples come from three different regions of Italy. The regions are further partitioned into nine areas: areas A1…A4 belong to region R1, areas A5 and A6 belong to region R2, and areas A7…A9 belong to region R3. We did not scale or sphere the data, because the variables are already on the same scale. The largest runt sizes were 168, 97, 59, 51, 42, 42, 33, 13, 13, 12, 11, 11, 11, 10, 10,…. The gap after 33 suggests the presence of eight modes. We thus chose runt size threshold 33 in the construction of the cluster tree. (The picture often is not as clear.)

Figure 6 shows the cluster tree. We have labeled each leaf with the predominant area for the olive oil samples falling into this leaf. Table 2 shows

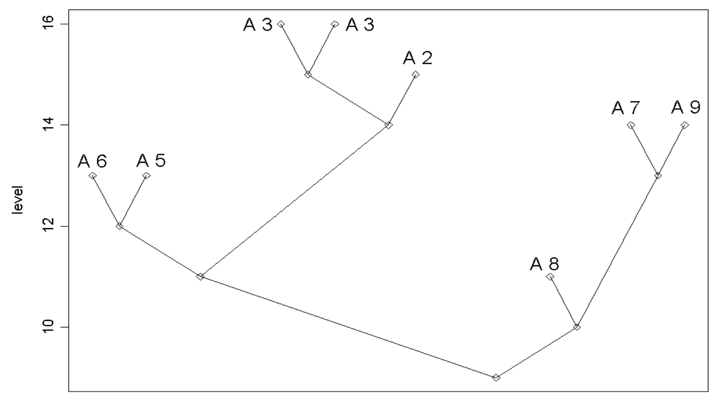


Figure 6: Cluster tree for olive oil data. Nodes have been labeled with predominant area.

the distribution of areas over clusters. Table 3 summarizes the performance of clustering methods for the olive oil date. Runt pruning with threshold 33 (eight clusters) gives an adjusted Rand index of 0.57. Again, single linkage is the lone outlier.

The split represented by the root node separates region R3 from regions R1 and R2. The left daughter separates region R1 from region R2. The method erroneously splits area A3 into two clusters and was not successful in identifying areas A1 and A4. This raises two questions: (a) do the olive oils from area A3 really have a bimodal density, and (b) does the density really have modes corresponding to areas A1 and A4?

The first question is at least partly answered by Figure 7. Panel (a) shows the cluster tree. We have selected the node $N$ that has partitioned area A3. Panel (b) shows a rootogram (histogram with the roots of the cell counts plotted on the vertical axis) of the cluster represented by $N$, projected onto the Fisher discriminant direction calculated to separate the daughter clusters. (The idea for this diagnostic plot comes from Gnanadesikan, Kettenring, and Landwehr (1982). The Fisher discriminant direction maximizes the ratio of between-cluster variance to within-cluster variance of the projected data (Mardia, Kent, and Bibby 1979, Section 11.5). In that sense it is the direction

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| A1 | 0 | 1 | 0 | 0 | 0 | 17 | 0 | 7 |
| A2 | 0 | 51 | 1 | 0 | 0 | 4 | 0 | 0 |
| A3 | 90 | 11 | 103 | 1 | 0 | 0 | 1 | 0 |
| A4 | 5 | 13 | 4 | 0 | 0 | 14 | 0 | 0 |
| A5 | 0 | 0 | 0 | 64 | 1 | 0 | 0 | 0 |
| A6 | 0 | 0 | 0 | 0 | 33 | 0 | 0 | 0 |
| A7 | 0 | 3 | 0 | 0 | 0 | 43 | 0 | 4 |
| A8 | 0 | 2 | 0 | 0 | 0 | 2 | 45 | 1 |
| A9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 51 |

Table 2: Olive oil data: cluster number (horizontal axis) tabulated against area (vertical axis).

that best reveals separation between clusters.) The rootogram does not look bimodal. While this does not conclusively show that there is only one mode — there might be two modal regions with nonlinear shapes, so that the separation does not manifest itself in the projection — it is an indication, and we might want to prune the tree by removing the daughters. In contrast, the rootogram in panel (d) where we have selected the node separating area A2 from area A3, shows clear bimodality. Note that this diagnostic can be used in practical problems because it does not require knowing the true labels of the observations.

To answer the question whether areas A1 and A4 really are separated from areas A2 and A3 and correspond to modes of the density, we project the observations from the four areas onto the first two discriminant coordinates (Mardia, Kent, and Bibby 1979, Section 12.5). Figure 8 shows that while areas A2 (open circle) and A3 (filled circle) form fairly obvious groups, this is not true for areas A1 (triangle) and A4 (cross). Again, finding this is not strictly conclusive because we are seeing only a two-dimensional projection of the eight-dimensional data but it is a good indication. Note that this approach is not an "operational" diagnostic, because in practice the true labels of the observations would be unknown. We use it here merely to help evaluate the performance of our method.

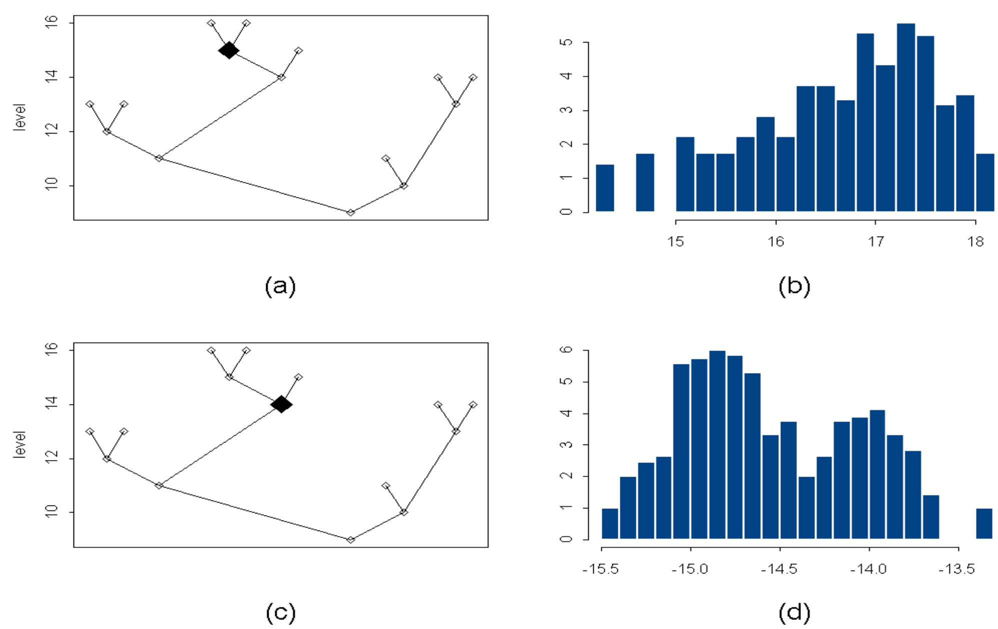


Figure 7: Diagnostic plots. Panel (a): cluster tree with node splitting area A3 selected; (b) projection of data in node on the Fisher discriminant direction separating daughters; (c) cluster tree with node separating area A3 from area A2 selected; (d) projection of data on the Fisher discriminant direction.

## 5.5 Handwritten digit data

The data for this example are 2,000 $16 \times 16$ grey level images of handwritten digits; the data therefore are of dimensionality 256. (The data were previously used to evaluate machine learning algorithms). The runt sizes are 288, 283, 90, 84, 74, 47, 37, 35, 22, 21, 21, 19, 19, 18, 13, 12, 12,.... The gap after 35 (vaguely) suggests presence of nine groups.

Table 4 summarizes the performance of various clustering methods on the handwritten digit data. The clear winners are model-based clustering with identical, spherical covariance matrices, and runt pruning. Runt pruning with threshold 35 (nine clusters) gives an adjusted Rand index of 0.64. The poor performance of average linkage is surprising.

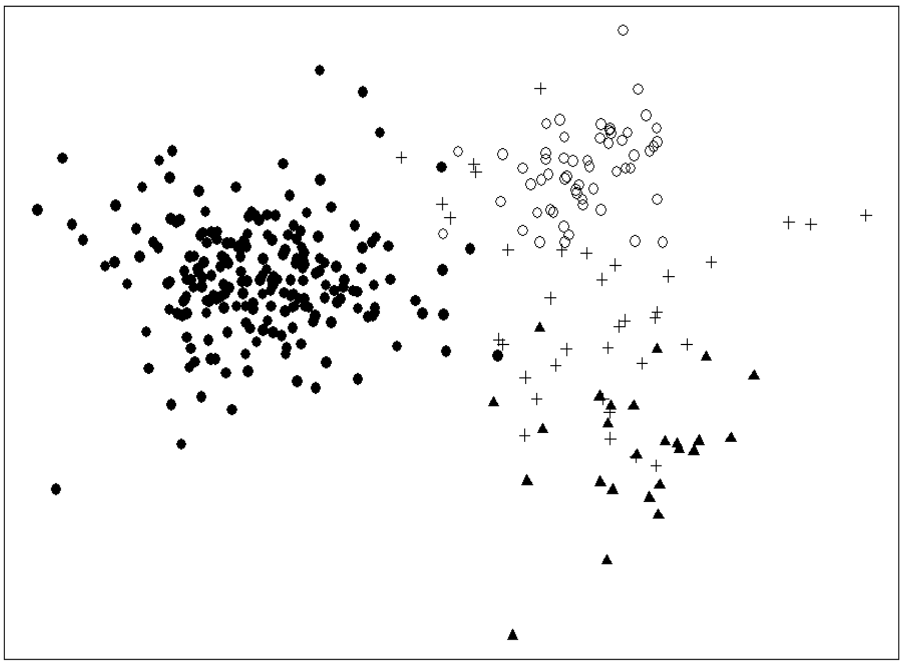

Figure 8: Projection of areas A1 (triangle), A2 (open circle), A3 (filled circle), and A4 (cross) on the plane spanned by first two discriminant coordinates.

# 6 Remarks

We address three issues: (a) relationship between parametric and nonparametric clustering; (b) distinction between clustering and compact partitioning; (c) non-modal groups.

## 6.1 Parametric versus nonparametric clustering

The most appealing feature of parametric or model-based clustering is that it seems to offer a way of estimating the number of groups: Fit a mixture model

| SL | AL | CL | RP | MC-EI | MC-VI |
|---|---|---|---|---|---|
| 0.0 (0.03) | 0.61 (0.07) | 0.43 (0.04) | 0.50 (0.06) | 0.45 (0.04) | 0.37 (0.04) |
| 0.0 (0.1) | 0.64 (0.05) | 0.50 (0.04) | 0.70 (0.06) | 0.58 (0.05) | 0.61 (0.05) |
| 15 | 12 | 7 | 5 | 8 | 5 |

Table 3: Comparison of single, average, and complete linkage, runt pruning, and two versions of model-based clustering for the olive oil data. First row: adjusted Rand index if methods are made to generate nine clusters; second row: adjusted Rand index for optimal partition size; third row: optimal partition size. Numbers in parentheses are standard errors.

| SL | AL | CL | RP | MC-EI | MC-VI |
|---|---|---|---|---|---|
| 0.0 (0.0) | 0.07 (0.05) | 0.28 (0.05) | 0.58 (0.04) | 0.62 (0.03) | 0.33 (0.04) |
| 0.0 (0.0) | 0.29 (0.04) | 0.36 (0.03) | 0.69 (0.04) | 0.63 (0.03) | 0.36 (0.03) |
| 20 | 20 | 19 | 7 | 15 | 8 |

Table 4: Comparison of single, average, and complete linkage, runt analysis, and two versions of model-based clustering for the handwritten digit data. First row: adjusted Rand index if methods are made to generate ten clusters; second row: adjusted Rand index for optimal partition size; third row: optimal partition size. Numbers in parentheses are standard errors.

to the sample and use a criterion such as BIC (Schwartz 1978) to estimate the number of mixture components, which is then taken as an estimate for the number of groups (Fraley and Raftery 1998). The conceptual problem with this approach is that the optimal number of mixture components is not an intrinsic property of the data density but instead depends on the family of distributions that are being mixed. Fitting a mixture of uniforms, or a mixture of spherical Gaussians, or a mixture of general Gaussians will result in different estimates for the number of groups. Thus, the estimated number of groups depends on assumptions about the group distributions that are somewhat arbitrary and unverifiable, but can sometimes be gleaned from plots of the data.

Nonparametric clustering is based on the premise that groups correspond to modes of the density. While this assumption can also be questioned — see Section 6.3 — at least the number of modes is an intrinsic property of the density. Model-based clustering can be used in an attempt to find modes and estimate their number. A weakness of this approach is that mixture modeling estimates the number of components in a mixture, which in general will be different from the number of modes.

## 6.2 Clustering versus compact partitioning.

It is important to distinguish between clustering and *compact partitioning* or *dissection*. The goal of compact partitioning is to split a set of objects into groups that are spatially compact. The degree of compactness can for example be measured by the total squared distance between the observations and their closest group mean vectors, the figure of merit that is optimized by the k-means or Lloyd algorithm (MacQueen 1967). There are applications, such as vector quantization, where compact partitioning of the data makes perfect sense, even when there are no distinct groups at all. Some clustering algorithms, like complete linkage, are really better regarded as tools for compact partitioning. Compact partitioning algorithms might identify groups if they happen to be roughly spherical and well separated, but they will fail if this is not the case, as illustrated in Section 5.2.

## 6.3 Non-modal groups

We have assumed thus far that groups correspond to modes of the density, but this assumption may not hold. Most observers would agree that the data sets shown in Figure 9 each consist of two groups. However, these groups do not manifest themselves as modes. We are not aware of a general, automatic solution to this problem. There appear to be two approaches.

The first is to abandon the requirement that the method be automatic. Instead we graphically present the data to an observer and rely on the human cognitive ability for detecting groups. A large collection of tools has been developed for this purpose, ranging from mapping techniques like projection pursuit (Friedman and Tukey 1974; Friedman, Stuetzle, and Schroeder 1984; Friedman 1987) to systems like XGobi (Swayne, Cook, and Buja 1998) that are based on high interaction motion graphics.

The second approach is to transform the data such that in the transformed data grouping manifests itself as multimodality. The prototypical example for this approach is the Hough transform in computer vision (Ballard and Brown 1982, Section 4.3).

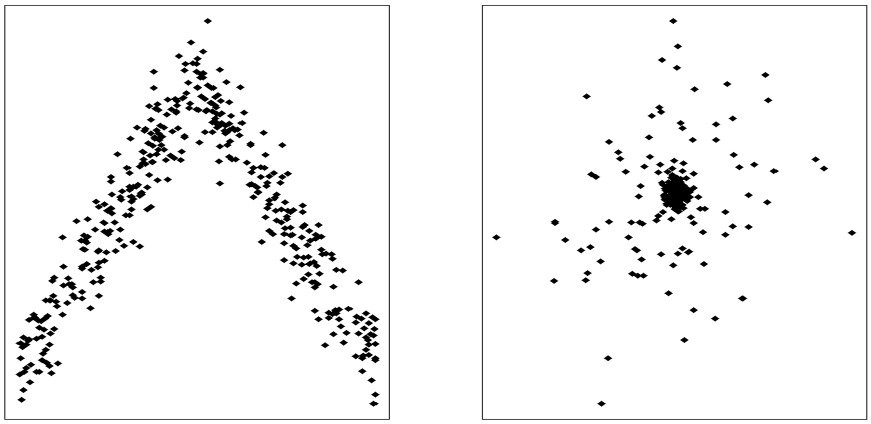

Figure 9: Groups that do not correspond to modes.

# 7 Summary and future work

We presented *runt pruning*, a new clustering method that attempts to find modes of a density by analyzing the MST of a sample. The method exploits the connection between the MST and nearest neighbor density estimation. It does not rely on assumptions about the specific form of the data density or the geometric shapes of the clusters. Our (admittedly very limited) experiments suggest that the price in performance paid for this generality is small.

There are a number of areas for future work. Probably the most important and most difficult problem in nonparametric clustering is determining the correct size of the cluster tree, i.e., estimating the number of modes. Even the simpler problem of testing for unimodality has not found a fully satisfactory solution in the multivariate case. Existing methods, such as the DIP test (Hartigan and Hartigan 1985), the RUNT test (Hartigan and Mohanty 1992), and the MAP test (Rozál and Hartigan 1994), strictly speaking, test the hypotheses that the data are multivariate Gaussian or uniform, using a test statistic that is sensitive against multimodal alternatives.

Absent an automatic solution to the problem of estimating the size of a cluster tree, we have to rely on diagnostic tools that require human interaction. The diagnostic plots shown in Section 5 only work if the clusters can be linearly separated. If the clusters are highly nonlinear, as in the example of Section 5.2, projection of the data on the Fisher discriminant direction might not show a clear bimodality, and more powerful methods are needed.

A weakness of runt pruning is its reliance on the nearest neighbor density estimate. One would like to use better density estimates such as $k$-th nearest neighbor or (adaptive) kernel estimates (see, for example, Silverman 1986, Section 2), or a Projection Pursuit estimate (Friedman *et al.* 1984; Friedman 1987). We have implemented a recursive partitioning version of Wishart's (1969) hierarchical mode analysis which can be used with any density estimate. However, exactly computing the connected components of level sets for kernel or projection pursuit estimates appears to be difficult. We have had to resort to heuristics (just as Wishart did), which is an esthetic if not practical drawback.

A simpler alternative is to combine runt pruning with Wong's $k$-th nearest neighbor clustering (Wong 1979; Wong and Lane 1983), which is a generalization of single linkage clustering using the $k$-th nearest neighbor density estimate. Wong's method has three steps: (a) Construct a graph connecting each observation to its $k$ nearest neighbors; (b) Assign edge weight $w(e) = 1/\hat{p}_k(\mathbf{x}_i) + 1/\hat{p}_k(\mathbf{x}_j)$ to an edge with endpoints $\mathbf{x}_i, \mathbf{x}_j$; (c) Calculate the MST $T_w$ of the edge weighted graph. The discussion of Section 4 suggests forming clusters by applying runt pruning to $T_w$ instead of simply breaking the longest edges. We have conducted preliminary experiments with this method, and the results appear to be somewhat but not greatly better than those obtained by runt pruning of the MST.

Finally there is the problem of estimating cluster trees from large samples. Runt pruning consists of two steps, finding the MST and computing the cluster tree from the MST. For the MST we currently use a simple $O(n^2)$ algorithm (Algorithm 2 of Prim 1957) that does not require storing the interpoint distance matrix, but instead evaluates distances as needed. On a 1.5 Ghz Pentium PC, computing the MST for a data set with 10,000 observations in 10 dimensions takes roughly a minute. There are algorithms which are faster, at least in theory, for example the $O(n \log n)$ algorithm of Bentley and Friedman (1978). However, the actual performance of these algorithms degrades rapidly with increasing dimensionality; see Nene and Nayar (1997) for a discussion of this effect.

The time required for computing the cluster tree from the MST depends on the data and the runt size threshold. In our experiments it tended to be much less than the time required for finding the MST. So it appears that runt pruning is certainly practical for data sets of size $n = 10^4$, and is applicable for $n = 10^5$, given some patience. (Of course there is no experience regarding its performance for such large problems.)

The articles by Cutting, Karger, Pedersen, and Tukey (1992) and Cutting, Karger, and Pedersen (1993) contain some interesting ideas for clustering very large data sets. The question is how such approaches can be made to fit into the framework of estimating level sets.

ANKERST, M., BREUNING, M.M., KRIEGEL, H.P., and SANDER, J. (1999), "Optics: Ordering Points to Identify the Clustering Structure", *Proceedings, ACM SIGMOD International Conference on Management of Data (SIGMOD'99)*, 49-60.

BALLARD. D.H., and BROWN, C.M. (1982), "Computer Vision", Prentice Hall.

BENTLEY, J.L., and FRIEDMAN, J.H. (1978), "Fast Algorithms for Constructing Minimal Spanning Trees in Coordinate Spaces", *IEEE Transactions on Computers, C-27(2)*, 97-105.

CUTTING, D.R., KARGER, D.R., and PEDERSEN, J.O. (1993), "Constant Interaction Time Scatter/Gather Browsing of Very Large Document Collections", *Proceedings, Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 126-134.

CUTTING, D.R., KARGER, D.R., PEDERSEN, J.O., and TUKEY, J.W. (1992), "Scatter/Gather: A Cluster-Based Approach to Browsing Large Document Collections", *Proceedings, Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 318-329.

ESTER, M., KRIEGEL, H.P., SANDER, J., and Xu, X. (1996), "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", in *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 226–231, 1996.

FRALEY, C., and RAFTERY, A. (1998), "How Many Clusters? Which Clustering Method? - Answers Via Model-Based Cluster Analysis", *The Computer Journal, 41*, 578-588.

FRALEY, C. and RAFTERY, A. (1999), "Mclust: Software for Model-Based Cluster Analysis", *Journal of Classification, 16*, 297-306.

FRIEDMAN, J.H. (1987), "Exploratory Projection Pursuit", *Journal of the American Statistical Association, 82*, 249-266.

FRIEDMAN, J.H., STUETZLE, W., and SCHROEDER, A. (1984), "Projection Pursuit Density Estimation", *Journal of the American Statistical Association, 79*, 599-608.

FRIEDMAN, J.H., and TUKEY, J.W. (1974), "A Projection Pursuit Algorithm for Exploratory Data Analysis", *IEEE Transactions on Computers, C-23*, 881-890.

GNANADESIKAN, R., KETTENRING, J.R., and LANDWEHR, J.M. (1982), "Projection Plots for Displaying Clusters", in *Statistics and Probability: Essays in Honor of C. R. Rao*, Elsevier/N.Holland, 269-280.

GOWER, J.C., and ROSS, G.J.S. (1969), "Minimal Spanning Trees and Single Linkage Cluster Analysis", *Applied Statistics, 18*, 54-64.

HARTIGAN, J.A. (1975), "Clustering Algorithms", Wiley.

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

HARTIGAN, J.A. (1985), "Statistical Theory in Clustering", *Journal of Classification, 2*, 63-76.

HARTIGAN, J.A., and HARTIGAN, P.M. (1985), "The DIP Test of Unimodality", *Annals of Statistics, 13*, 70-84.

HARTIGAN, J.A., and MOHANTY, S. (1992), "The RUNT Test for Multimodality", *Journal of Classification, 9*, 63-70.

HUBERT, L., and ARABIE, P. (1985), "Comparing Partitions", *Journal of Classification, 2*, 193-218.

MACQUEEN, J. (1967), "Some Methods for Classification and Analysis of Multivariate Observations", in *Proceedings of the 5th Berkeley Symposium on Mathematics, Statistics and Probability, 1*, 281-296.

MARDIA, K., KENT, J., and BIBBY, J. (1979), "Multivariate Analysis", Academic Press.

MCLACHLAN, G.J., and Peel, D. (2000), "Finite Mixture Models", Wiley.

NENE, S.A., and NAYAR, S.K. (1997), "A Simple Algorithm for Nearest Neighbor Search in High Dimensions", *IEEE Transactions on Pattern Analysis and Machine Intelligence, 19*, 989-1003.

PRIM, R.C. (1957), "Shortest Connection Matrix Network and Some Generalizations", *Bell Systems Technical Journal, 36*, 1389-1401.

ROZÁL, G.P.M., and HARTIGAN, J.A. (1994), "The MAP test for Multimodality", *Journal of Classification, 11*, 5-36.

SCHWARTZ, G. (1978), "Estimating the Dimension of a Model", *Annals of Statistics, 6*, 461-464.

SHAO, J., and TU, D. (1995), "The Jackknife and Bootstrap", Springer.

SILVERMAN, B.W. (1986), "Density Estimation for Statistics and Data Analysis", Chapman & Hall.

SWAYNE, D.F., COOK, D., and BUJA, A. (1998), "Xgobi: Interactive Dynamic Data Visualization in the X Window System", *Journal of Computational and Graphical Statistics, 7*, 113-130.

TUKEY, P.A., and TUKEY, J.W. (1989), "Data Driven View Selection, Agglomeration, and Sharpening", in *Interpreting Multivariate Data*, Ed., V. Barnett, Wiley, 215-243.

WISHART, D. (1969), "Mode Analysis: A Generalization of Nearest Neighbor which Reduces Chaining Effects", in *Numerical Taxonomy*, Ed., A.J. Cole, Academic Press, 282-311.

WONG, M.A. (1979), "Hybrid Clustering", PhD thesis, New Haven CT: Department of Statistics, Yale University.

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