# Assessment and Pruning of Hierarchical Model Based Clustering

Jeremy Tantrum[*]
Department of Statistics
University of Washington
Seattle, WA 98195
tantrum@stat.washington.edu

Alejandro Murua
Department of Statistics
University of Washington
Seattle, WA 98195
murua@ieee.org

Werner Stuetzle[†]
Department of Statistics
University of Washington
Seattle, WA 98195
wxs@stat.washington.edu

## ABSTRACT

The goal of clustering is to identify distinct groups in a dataset. The basic idea of model-based clustering is to approximate the data density by a mixture model, typically a mixture of Gaussians, and to estimate the parameters of the component densities, the mixing fractions, and the number of components from the data. The number of distinct groups in the data is then taken to be the number of mixture components, and the observations are partitioned into clusters (estimates of the groups) using Bayes' rule. If the groups are well separated and look Gaussian, then the resulting clusters will indeed tend to be "distinct" in the most common sense of the word - contiguous, densely populated areas of feature space, separated by contiguous, relatively empty regions. If the groups are not Gaussian, however, this correspondence may break down; an isolated group with a non-elliptical distribution, for example, may be modeled by not one, but several mixture components, and the corresponding clusters will no longer be well separated. We present methods for assessing the degree of separation between the components of a mixture model and between the corresponding clusters. We also propose a new clustering method that can be regarded as a hybrid between model-based and nonparametric clustering. The hybrid clustering algorithm prunes the cluster tree generated by hierarchical model-based clustering. Starting with the tree corresponding to the mixture model chosen by the Bayesian Information Criterion, it progressively merges clusters that do not appear to correspond to different modes of the data density.

## Categories and Subject Descriptors

I.5.3 [**Pattern Recognition**]: Clustering; I.5.1 [**Pattern Recognition**]: Models—*Statistical*; G.3 [**Probability and Statistics**]: Multivariate Statistics

## General Terms

Model-based Clustering

## Keywords

Model-based Clustering, Nonparametric Clustering, Density Estimation, Unimodality

## 1. INTRODUCTION AND MOTIVATION

The goal of clustering is to identify distinct groups in a dataset $\mathcal{X} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\} \subset R^m$. For example, when presented with (a typically higher dimensional version of) a dataset like the one in Figure 1a we would like to detect that there appear to be two groups, and assign a group label to each observation. (Throughout this paper we distinguish between "groups" and "clusters", which are estimates for the groups.)

***Model-based clustering in a nutshell.*** To cast clustering as a statistical problem we regard the data $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ as a sample from some unknown probability density $p(\boldsymbol{x})$. *Model-based clustering* (see [9] and references therein) relies on the premise that each group $g$ is represented by a density $p_g(\boldsymbol{x})$ that is a member of some parametric family, typically the multivariate Gaussian distributions. In this case $p(\boldsymbol{x})$ is a Gaussian mixture:

$$p(\boldsymbol{x}) = \sum_{g=1}^{G} \pi_g \, p_g(\boldsymbol{x}; \boldsymbol{\mu}_g, \Sigma_g)\,, \qquad (1)$$

where $G$ is the number of groups, $\pi_g$ is the prior probability of group $g$, and $p(\boldsymbol{x}; \mu, \Sigma)$ denotes the Gaussian density with mean $\mu$ and covariance matrix $\Sigma$. For fixed $G$ we can estimate the parameters $\pi_g$, $\boldsymbol{\mu}_g$, and $\Sigma_g$ by maximum likelihood, using the EM-algorithm [9, Chapter 2.8]. There are many ways of estimating $G$ [9, Chapter 6], e.g. by maximizing the Bayesian Information Criterion (BIC) [11, 5]:

$$\hat{G} = \text{argmax}_G \left(2 \times L(G) - r \, \log(n)\right)\,. \qquad (2)$$

Here, $L(G)$ is the log-likelihood of the best $G$ component model, $r$ is the number of parameters of the model and $n$ is the number of observations.

[*] Supported by NSA grant 62-2948.

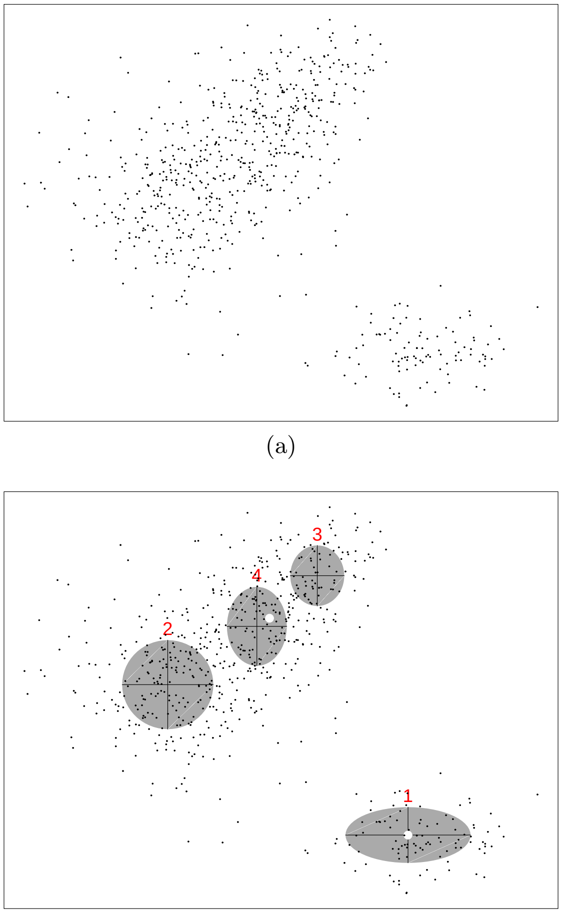
[†] Supported by NSF grant DMS-9803226 and NSA grant 62-2948.

*SIGKDD* '03 Washington, DC, USA
Copyright 2003 ACM 1-58113-737-0/03/0008...$5.00.

While attractive conceptually, the straight-forward approach to mixture modeling — fit models for many different values of $G$ using the EM algorithm, and then choose the model that maximizes the BIC — is slow. Following a suggestion by Fraley and Raftery [5], we address this problem by using a hierarchical approach: Find a model with $G-1$ components by merging the two groups of the $G$ component model for which the merge leads to the smallest decrease in log-likelihood. Among the sequence of models thus generated choose the one maximizing the BIC.

There have been several recent advances in extending the normal mixture model to large datasets [2, 13].

*A conceptual problem with model-based clustering.* Model-based clustering relies on the premise that mixture components in the model correspond to distinct groups in the data. If the groups are Gaussian, then the resulting clusters will indeed tend to be "distinct" in the most common sense of the word - contiguous, densely populated areas of feature space, separated by contiguous, relatively empty regions [3]. If the groups are not Gaussian, however, the correspondence between groups and mixture components may break down. An isolated group with a non-elliptical distribution, for example, may be modeled by not one, but several mixture components, and the corresponding clusters will no longer be distinct. This problem is illustrated in Figure 1a. Most observers would probably agree that the data in this figure fall into two separate groups. The BIC criterion, however, chooses a mixture model with four components; Figure 1b shows regions containing 60% of the mass of each component.



(a)

(b)

**Figure 1: Data set with fitted Gaussian mixture. The modes of the mixture are indicated by the two white dots. (This example is referred to as the running example in the remainder of the paper.)**

*Contributions of the paper.* We present diagnostic tools for assessing the degree of separation between the components of a mixture model and between the corresponding clusters. We also propose an algorithm for pruning the cluster tree generated by hierarchical model-based clustering. The algorithm starts with the tree corresponding to the mixture model chosen by the Bayesian Information Criterion. It then progressively combines mixture components that do not appear to correspond to different modes of the data density and merges the corresponding clusters. Each cluster in the final partition may therefore be modeled by more than one mixture component. The resulting procedure can be regarded as a hybrid between nonparametric and model-based clustering: we look for modes in the data distribution using the mixture model as a density estimate.

## 2. ASSESSING SEPARATION BETWEEN MIXTURE COMPONENTS

Roughly speaking, we would expect mixture components modeling different groups in the data to be well separated. On the other hand, mixture components modeling parts of the same group would be expected to exhibit significant overlap.

We now put this concept in probability terms. Assume that we have modeled the distribution of the observed data by a mixture density $p(\boldsymbol{x}) = \sum_g \pi_g p_g(\boldsymbol{x})$. We can generate observations from this density by first generating a component label $Y$ with $P(Y = g) = \pi_g$, and then generating $X$ from $p_Y$. According to Bayes' rule, the posterior probability $P(Y = g|X)$ is

$$P(Y = g|X) = \frac{\pi_g p_g(X)}{\sum_{j=1}^{G} \pi_j p_j(X)} .$$

Component $g$ is well separated from all the other components if $P(Y = g|X)$ only takes extreme values, either close to zero or close to one - one for observations actually generated from component $g$, and zero for all others.

Exactly evaluating the distributions of $P(Y = g|X)$ for the $G$ components is generally impossible when the dimension $m$ is larger than 1. To see why this is so, define the random variable $h(X) = P(Y = g|X)$. Its distribution function $F_h(u)$ is given by

$$F_h(u) = P(h(X) \leq u) = \int I(h(\boldsymbol{x}) \leq u)\, p(\boldsymbol{x}) d\boldsymbol{x} , \qquad (3)$$

where $I(\cdot)$ denotes the indicator function. Except in trivial cases, (such as $G = 2$, $\Sigma_1 = \Sigma_2$) the region of feature space defined by the indicator function has a complex shape described in terms of conic sections. Hence in general this integral can not be evaluated analytically and we resort to Monte Carlo simulation.

In the following we present three methods for assessing the separation between mixture components, based on the posterior probabilities, the margins, and the misclassification probabilities. All these were estimated by simulating from the model.

## 2.1 Assessing separation using posterior probabilities

Figure 2 shows rootograms of the posterior probabilities $P(Y = g|X)$ for the four components of the mixture model in our running example. (A rootogram is a variant of a histogram where the heights of the bars encode the square roots of the bin counts, instead of the bin counts themselves. This makes low counts more visible.) The rootograms are based on 20,000 data points generated from the estimated mixture model. We have omitted the bin containing $P(Y = g|X) = 0$, because it would have by far the largest bin count and would obscure the information in the remaining bins.

The rootogram for component one (top panel) has a large peak at $P(Y = 1|X) = 1$ and is essentially zero elsewhere, indicating clear separation of component one from all the other components. On the other extreme, the rootogram for component four has no peak at $P(Y = 4|X) = 1$. This is due to the fact that component four is completely overlapped by components two and three, and hence there is always a substantial posterior probability that an observation generated from $p_4$ might have come from $p_2$ or $p_3$. Furthermore, the significant mass away from $P(Y = g|X) = 1$ in the rootograms for components two, three, and four shows that these components are not well separated.

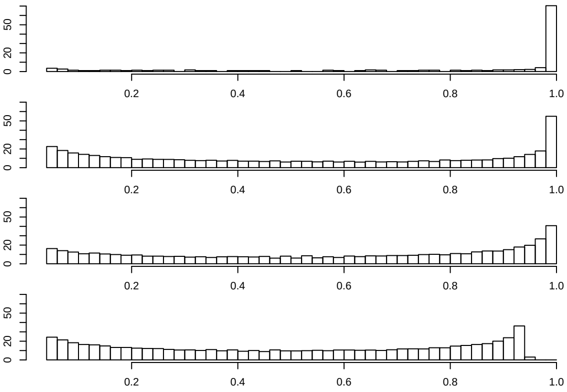


**Figure 2: Running example: Rootograms of the posterior probabilities $P(Y = g|X)$ for $X$ distributed according to the mixture model. The bin containing zero is not shown in order not to obscure the pattern in the other bins.**

## 2.2 Assessing separation using margins

An alternative to looking at the posterior probabilities is to consider the margins. Let $\hat{Y}(X)$ be the estimated component label assigned to $X$ by Bayes' rule:

$$\hat{Y}(X) = \arg\max_g P(Y = g|X)\,.$$

The margin of $X$ drawn from component $Y$ of the model is given by

$$\text{margin}(X, Y) = P(\hat{Y}(X) = Y|Y) - \max_{g \neq Y} P(\hat{Y}(X) = g|Y)\,.$$

| | 1 | 2 | 3 | 4 | $MC_g$ | $\pi_g$ |
|---|---|---|---|---|---|---|
| 1 | 0.998 | 0.002 | 0 | 0 | 0.002 | 0.162 |
| 2 | 0.001 | 0.879 | 0.001 | 0.119 | 0.121 | 0.388 |
| 3 | 0 | 0 | 0.908 | 0.092 | 0.092 | 0.191 |
| 4 | 0 | 0.060 | 0.105 | 0.835 | 0.165 | 0.259 |

**Table 1: Misclassification matrix for the running example.**

Note that a negative margin means that $X$ is assigned to the wrong component, and that a small margin means that $X$ lies in a region where components overlap significantly.

Figure 3 shows the cumulative distribution function (cdf) of the margin for observations drawn from the four component mixture model of our running example. There is a large proportion of small margins indicating substantial overlap between the components.

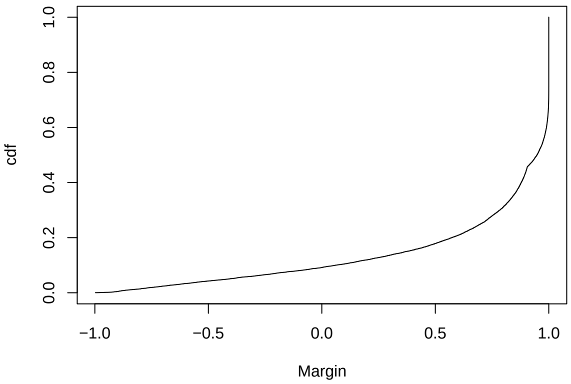


**Figure 3: Running example: Cumulative distribution function of the margin.**

## 2.3 Assessing separation using misclassification probabilities

When the number of clusters is moderate, we can look at the misclassification matrix to detect well separated as well as overlapping components of a mixture model. Table 1 shows the misclassification matrix for the mixture model in our running example. Let $m_{gg'}$ be the probability that the Bayes' rule assigns an observation from component $g$ to component $g'$.

From the misclassification matrix we can extract information at three different levels of detail. At the coarsest level we can look at the overall misclassification probability given by $\sum_g \pi_g(1 - m_{gg})$. The lower this probability is, the better the separation. At the next higher level of detail, we can look at the component-wise misclassification probabilities $MC_g$. In our example (Table 1) the misclassification probability for component one is very small ($MC_1 = 0.002$), indicating that component one is well separated. The misclassification probabilities for the other components are substantially larger. On the most detailed level, the values of $m_{gg'}$ and $m_{g'g}$ indicate which other components overlap component $g$. The pattern of entries in Table 1 shows that components two, three and four are mutually overlapping. We could not see this from the less detailed views.

# 3. ASSESSING SEPARATION BETWEEN CLUSTERS

A mixture model is only an estimate for the true underlying density of the data. Therefore the degree of separation between mixture components (or lack thereof) does not always accurately reflect the actual separation between the clusters.

We cannot compute the matrix of misclassification probabilities for the observed data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, nor the margins, (as we did in the previous section for observations simulated from the estimated model) because those require knowing the true labels. However we can compute the posterior probabilities $P(Y = g|\boldsymbol{x}_i)$, and therefore generate a plot analogous to Figure 2, shown in Figure 4. The rootogram for $P(Y = 4|\boldsymbol{x}_i)$ (bottom panel) looks basically flat, from which we can conclude that cluster 4 almost certainly does not correspond to a distinct group in the data.

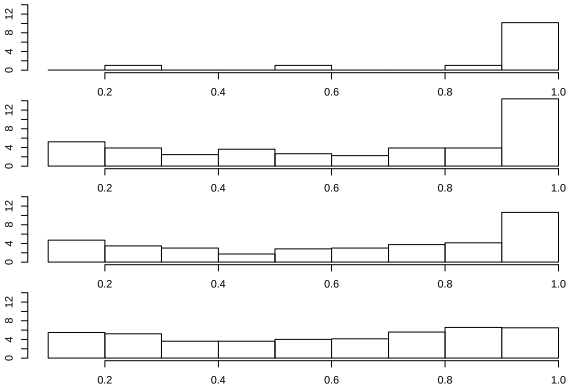


**Figure 4: Running example: Rootograms of the posterior probabilities $P(Y = g|\boldsymbol{x}_i)$ for the data.**

# 4. HYBRID CLUSTERING

Hierarchical model-based clustering generates a hierarchy of mixture models: The model with $m - 1$ mixture components is obtained by merging the two clusters of the $m$ component model for which the change leads to the smallest decrease in log-likelihood. The result of this merging process can be represented by a binary tree $T$. The leaves of the tree are the observations. Each interior node $N$ of the tree is assigned a *generation* between 1 and $n - 1$, indicating where in the sequence of merges it was generated. The interior node corresponding to the $i$-th merge in the sequence is assigned generation $n - i$; the root node therefore has generation 1. Each node $N$ is also associated with the cluster formed by its descendent leaves.

The merge sequence defines a sequence of trees: $T_m$ is obtained from $T$ by removing the offspring of all nodes with generation greater than or equal to $m$. By construction, $T_m$ has $m$ leaves and corresponds to a mixture model with $m$ mixture components. Let $G$ be the number of mixture components chosen by the BIC, and let $T_G$ be the corresponding tree.

If the distinct groups in the data all have Gaussian distributions, then we expect roughly a one-to-one correspondence between groups and mixture components associated with the leaves of $T_G$. Also, the clusters associated with the leaves of $T_G$ will be similar to the groups. ("Roughly" because $G$, after all, is only an estimate.) If the groups are not Gaussian, however, each group may be modeled by more than one mixture component, and consequently will be the union of several clusters.

The idea of hybrid clustering is to test, for each node of $T_G$ whose daughters are leaves, whether the corresponding clusters are well separated. If they are not, then the clusters probably correspond to the same group, and we merge them. The new cluster is then modeled by the sum of the mixtures modeling the daughters that were merged. This pruning process is repeated until no further clusters can be merged.

## 4.1 Illustration of hybrid clustering

Before describing its ingredients in more detail, let us see the pruning process in action. The upper panel of Figure 5 shows the tree $T_4$ whose leaves correspond to the mixture model fit to the data in our running example. The circled node is the one being tested. The lower panel of Figure 5 shows the projection of its associated cluster onto the *Fisher discriminant direction*, which is the direction that best separates the projections of the two daughter clusters [6][8, Chapter 11.5]. The grey curve is the kernel density estimate for the projected data with the smallest bandwidth that yields a unimodal density [12, Chapter 6.3 and 6.4]. The black curve is the kernel density estimate with the smallest bandwidth that yields a bimodal density. The dot plot of the projected data looks unimodal, and the unimodal and bimodal distributions are almost identical, which indicates that the daughter clusters are not well separated in feature space. A formal test for unimodality of the projected data (Section 4.2) would reject the null hypotheses of unimodality at level $\alpha = 0.49$, meaning that the evidence against unimodality is weak. We therefore prune the daughters. The new tree is the one shown in black in Figure 6. The diagnostic plot is qualitatively similar to the one in Figure 5; the daughter clusters of the node being tested do not seem to be well separated, with unimodality being rejected at level $\alpha = 0.12$. We therefore prune again and are left with the tree shown in Figure 7. Now the picture is different: The diagnostic plot reveals a clear separation between the clusters, and a formal test rejects the hypothesis of unimodality at level $\alpha = 0.002$. We conclude that there appear to be two distinct groups in the data, one modeled by three mixture components, and the other one modeled by one mixture component.

## 4.2 Testing for unimodality

In order to automate the pruning process described in Section 4.1 we need a way of measuring the amount of evidence against unimodality for a univariate data set (the projection of a cluster onto the Fisher discriminant direction best separating its daughters). Even if we carry out the pruning process interactively, by looking at diagnostic plots like the ones in Figures 5-7, such a measure of evidence still provides a useful guideline.

Let $x_1, \ldots, x_n$ be a set of (univariate) data sampled from some density $f(x)$, and let $F_n(x)$ be the empirical cdf of the sample. To test the null hypotheses that $f(x)$ is unimodal we use J.A. Hartigan and P.M. Hartigan's DIP test described in [7]. The test statistic is the DIP

$$D = \sup_x |F_n(x) - H(x)|,$$

where $H$ is the unimodal cdf closest to $F_n$. Bickel and Fan [1]

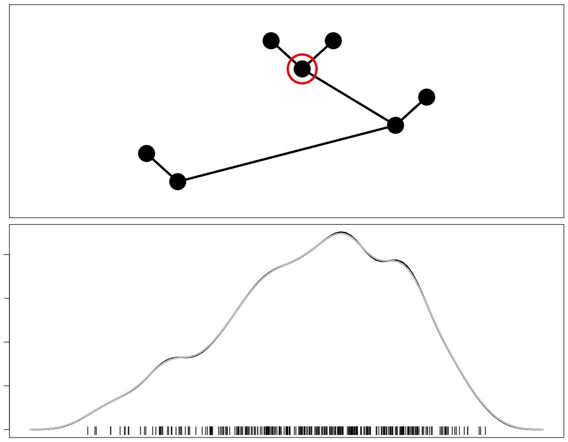


**Figure 5: Running example: Tree generated by hierarchical model-based clustering and diagnostic plot for the circled node.**

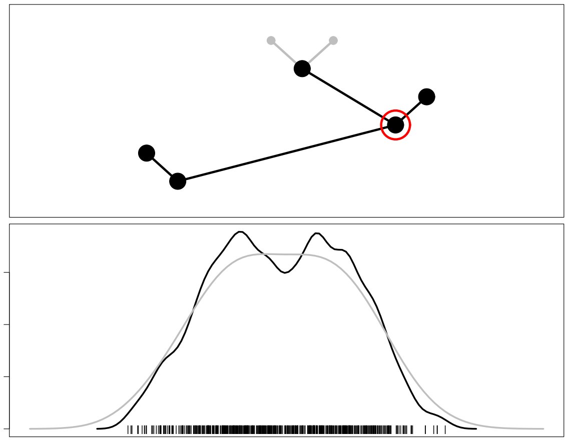


**Figure 6: Running example: Tree generated by hierarchical model-based clustering after first step of pruning, and diagnostic plot for the circled node.**

show that the nonparametric maximum likelihood estimate of the closest unimodal cdf, given the mode location $m_0$, is the greatest convex minorant of $F_n$ on $(-\infty, m_0]$ and the least concave majorant on $[m_0, \infty)$. (The greatest convex minorant of $F_n$ on $(-\infty, m_0]$ is the convex function $G$ not exceeding $F_n$ on $(-\infty, m_0]$ that minimizes $\sup_{x \leq m_0} |F_n(x) - G(x)|$. The least concave majorant is defined analogously.)

Bickel and Fan [1] also show that this estimate is robust against inaccuracy in the estimate of the mode. We could estimate the mode location by minimizing the DIP. However, this would be computationally expensive. Instead we estimate the mode using a kernel smoother, as suggested by Silverman [12, Chapter 6.3 and 6.4]. Figure 8 shows the empirical cdf of a sample (black curve), and the closest unimodal cdf (grey curve). The DIP is the maximum absolute difference between the two curves, indicated by the heavy vertical line. The estimated mode location is shown by the grey vertical line.

The distribution of the DIP under the null hypotheses of unimodality is not available in closed form but can be estimated by Monte Carlo. As before, let $H(x)$ be the unimodal cdf closest to $F_n(x)$. We generate $M$ samples of size $n$ from $H(x)$ ($M = 100$, say) and compute the DIPs $D_1, \ldots, D_M$. If the DIP $D_{orig}$ for the original sample is the $k$-th largest among $\{D_{orig}, D_1, \ldots, D_M\}$ then we reject the null hypotheses of unimodality at level $k/(M+1)$.

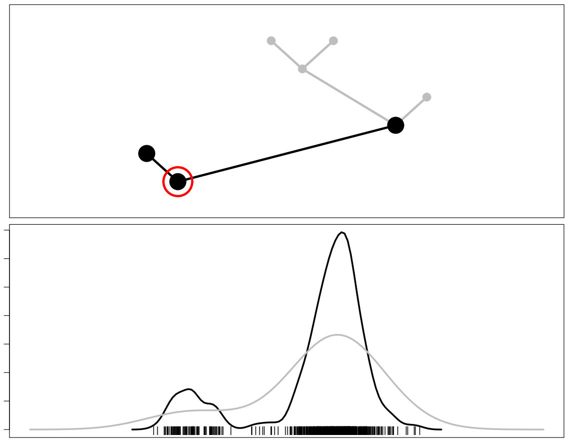


**Figure 7: Running example: Tree generated by hierarchical model-based clustering after second step of pruning, and diagnostic plot for the circled node.**

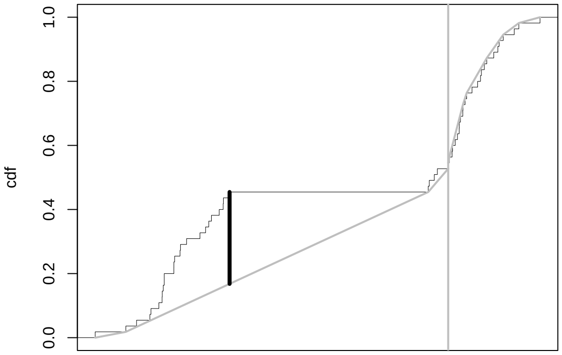


**Figure 8: Illustration of the DIP statistic.**

## 4.3 Remarks

Hybrid clustering is based on the premise that groups can correspond to collections of mixture components, not just individual components. The purpose of our method is to identify those collections, not to find a better fitting mixture model. This is in contrast to the work by Sand and Moore [10] on repairing faulty mixture models.

Automatic pruning requires specification of a significance level for the DIP tests; the larger the level, the larger the pruned tree. The significance level should not be taken too literally: the total pruning procedure does not constitute a level $\alpha$ test for unimodality of the multivariate feature distribution.

First, there is the problem of multiplicity: If we are carrying out many tests at a given level $\alpha$, then the probability of erroneously rejecting one or more of the null hypotheses is greater than $\alpha$.

Second, we are choosing the projection directions to maximize the separation between the clusters. This becomes an issue if the dimensionality of the feature space is large relative to the total number of observations in the two clusters which are under consideration. For example, if we have a total of $p + 1$ observations in a $p$ dimensional feature space then there will always be a direction for which the observations in the two clusters project onto exactly two points, one for each cluster. We deal with this problem by first projecting the combined observations from the two clusters onto their $k$ largest principal components and then finding the Fisher discriminant direction in this lower dimensional subspace. We chose $k$ to be one third of the total number of observations in the two clusters.

## 5. EXAMPLES

We show two examples of Hybrid clustering, one with real data and one with synthetic data simulated from a known Gaussian mixture.

### 5.1 Example 1

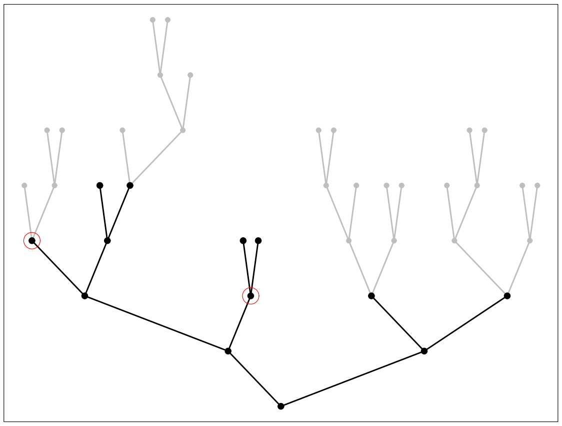


**Figure 9: Olive oil data: Original tree (all nodes) and pruned tree (black nodes).**

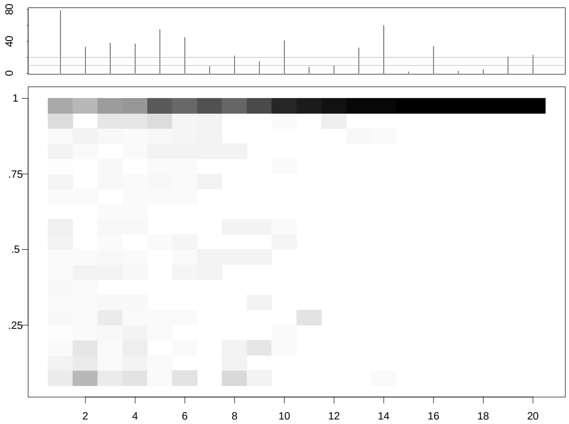


**Figure 10: Olive oil data: Histograms of posterior probabilities P($Y = g|\boldsymbol{x}_i$) for the data, before pruning**

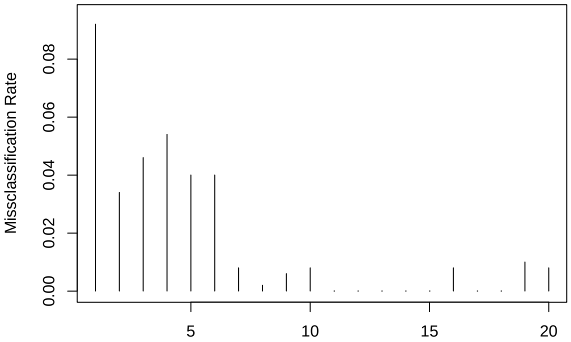


**Figure 11: Olive oil data: Misclassification probabilities $MC_g$ for the 20 components of the mixture model.**

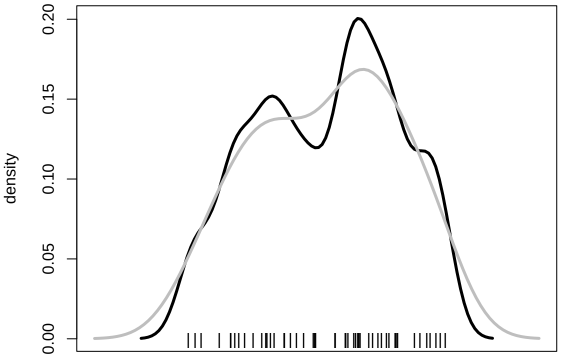


**Figure 12: Pruned node of olive oil tree**

The data for our first example consist of measurements of eight chemical concentrations on 572 samples of olive oils from nine different areas of Italy. Applying hierarchical model-based clustering with diagonal covariance matrices and using the BIC to estimate the number of mixture components results in a mixture model with 20 components, corresponding to the 20 leaves of the tree shown in Figure 9. The 20 columns of Figure 10 are histograms of $P(Y = g|\boldsymbol{x}_i)$ for $g = 1, \ldots, 20$, with the counts encoded as grey levels; the columns thus are a different graphical representation of the rootograms making up the rows of Figure 4. The bars in the upper panel of Figure 10 encode the observation counts in the clusters. If the clusters were all well separated, then each observation would have posterior probability one for one of the mixture components and zero for all the others, and the plot would have a solid black stripe at the top and be white elsewhere. We are obviously quite far removed from this ideal situation. This impression is confirmed by Figure 11. Some of the mixture components are not very isolated; observations generated from mixture component 1, for example, have roughly an 9% probability of being assigned to some other component.

Applying our pruning algorithm with significance level $\alpha = 0.01$ prunes the nodes shown in grey in Figure 9 and results in 7 clusters, four of which are modeled by more than one mixture component. Figure 12 shows a typical diagnostic plot for a node whose daughters are pruned ($\alpha = 0.88$), and Figure 13 shows a typical plot for a node whose daughters are retained ($\alpha = 0.01$). These two nodes are circled in

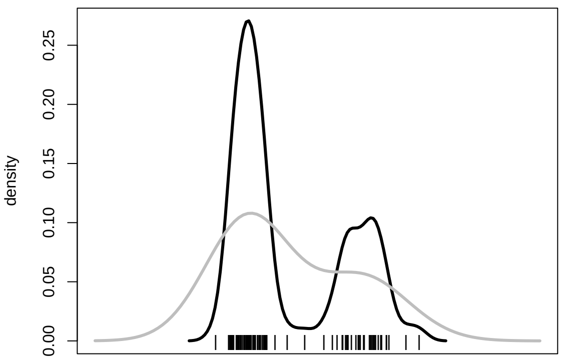


**Figure 13: Non pruned node of olive oil tree**

Figure 9.

Figure 14 is the post-pruning analog to Figure 10. It is much closer to the ideal of "black stripe, white elsewhere". The misclassification probabilities shown in Figure 15 also have decreased significantly; the largest one is now 1.5% instead of 9%.

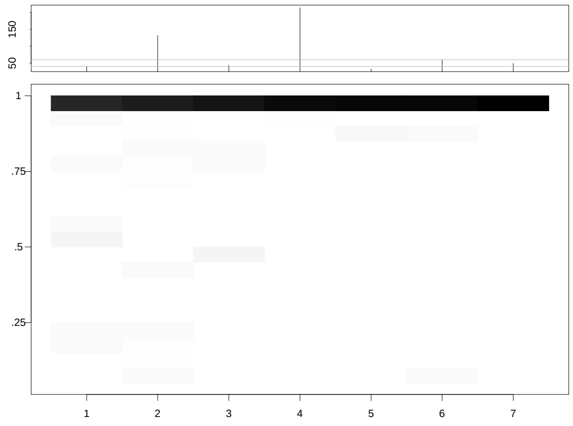


**Figure 14: Olive oil data: Histograms of posterior probabilities $\mathbf{P}(Y = g|\boldsymbol{x}_i)$ for the data, after pruning.**

Figure 16 shows the cdf's of the margins for the two clusterings, pre-pruning in black, post-pruning in grey. If the mixture components were perfectly separated then the cdf of the margin would be a step function with a single step at margin = 1. Pruning brings us closer to this ideal.

In our example we know the group labels of the observations - we know the area of origin for each olive oil and it seems reasonable to assume that any groups in the data reflect the areas of origin. We therefore assess how closely the clusters match the areas. Figure 17 shows a two way contingency table of areas on the vertical axis versus clusters on the horizontal axis, before pruning. Notice that areas 3, 8 and 9 are each broken up into several clusters and the clustering procedure has not been able to separate out areas 1, 2 and 4. Figure 18 shows the corresponding contingency table after pruning. Note that areas 3, 8, and 9 now correspond to single clusters and 1, 2, and 4 have been combined into one cluster. This raises the question how well areas 1, 2, and 4 are in fact separated in the 8 dimensional feature space. Figure 19 shows a projection of the observations from those areas onto the two linear discriminant coordinates [8, Chapter 11.5]. There is no obvious separation of the points into groups.

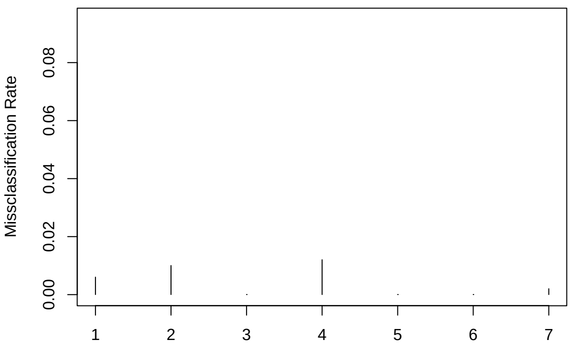


**Figure 15: Olive oil data:Misclassification probabilities for the model, after pruning.**

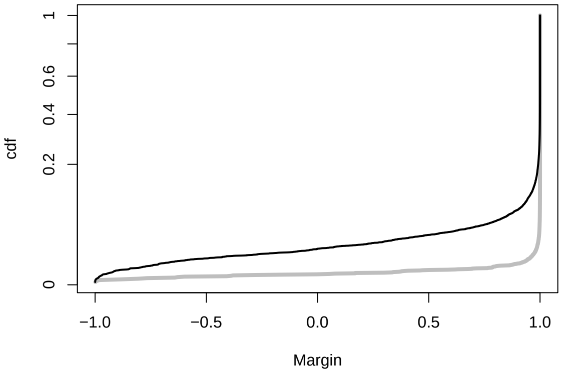
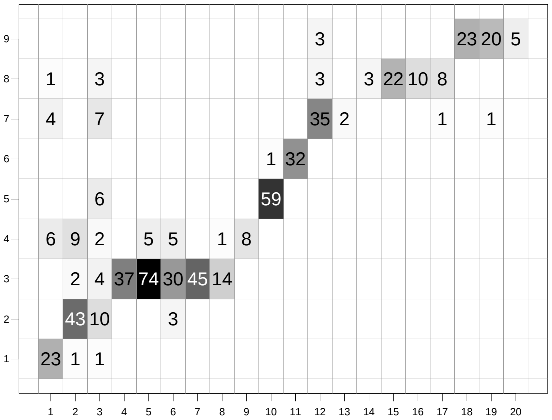


**Figure 16: Olive oil data: Cumulative distribution function of the margins before pruning (black line) and after pruning (grey line).**

It is convenient to have a numerical measure summarizing the degree of agreement between groups (areas) and clusters. We use the Fowlkes-Mallows index [4] for this purpose. The index is the geometric mean of two probabilities: the probability that two randomly chosen observations are in the same cluster given that they are in the same group, and the probability that two randomly chosen observations are in the same group given that they are in the same cluster. Hence a Fowlkes-Mallows index near 1 means that the clusters are a good estimate of the groups. For our example, the Fowlkes-Mallows index before pruning is 0.52, compared to an index of 0.81 after pruning. This shows that pruning substantially improved the agreement between groups and clusters.

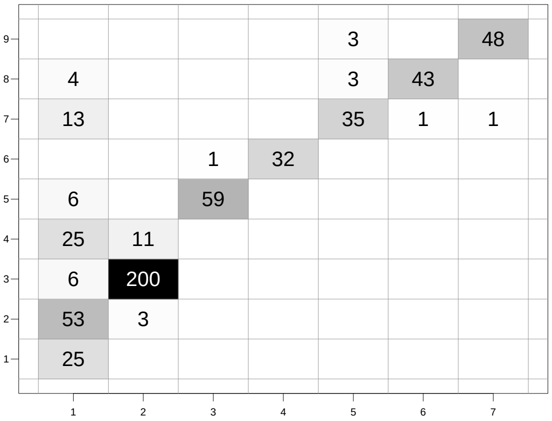
## 5.2 Example 2

In Example 1, pruning was successful in that it significantly improved the agreement between clusters and areas. The purpose of the second example is to illustrate how hybrid clustering performs on data which were in fact generated from a Gaussian mixture model. We choose a mixture model that mimics the olive oil data: we estimate mean



|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 9 |  |  |  |  |  |  |  |  |  |  |  | 3 |  |  |  |  |  | 23 | 20 | 5 |
| 8 | 1 |  | 3 |  |  |  |  |  |  |  |  | 3 |  | 3 | 22 | 10 | 8 |  |  |  |
| 7 | 4 |  | 7 |  |  |  |  |  |  |  |  | 35 | 2 |  |  |  | 1 |  | 1 |  |
| 6 |  |  |  |  |  |  |  |  |  | 1 | 32 |  |  |  |  |  |  |  |  |  |
| 5 |  |  | 6 |  |  |  |  |  |  | 59 |  |  |  |  |  |  |  |  |  |  |
| 4 | 6 | 9 | 2 |  | 5 | 5 |  | 1 | 8 |  |  |  |  |  |  |  |  |  |  |  |
| 3 |  | 2 | 4 | 37 | 74 | 30 | 45 | 14 |  |  |  |  |  |  |  |  |  |  |  |  |
| 2 |  | 43 | 10 |  |  | 3 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1 | 23 | 1 | 1 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

**Figure 17: Olive oil data: Two way contingency table of areas on the vertical axis versus clusters on the horizontal axis**



|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 9 |  |  |  |  | 3 |  | 48 |
| 8 | 4 |  |  |  | 3 | 43 |  |
| 7 | 13 |  |  |  | 35 | 1 | 1 |
| 6 |  |  | 1 | 32 |  |  |  |
| 5 | 6 |  | 59 |  |  |  |  |
| 4 | 25 | 11 |  |  |  |  |  |
| 3 | 6 | 200 |  |  |  |  |  |
| 2 | 53 | 3 |  |  |  |  |  |
| 1 | 25 |  |  |  |  |  |  |

**Figure 18: Olive oil data: Two way contingency table of areas on the vertical axis versus clusters on the horizontal axis**

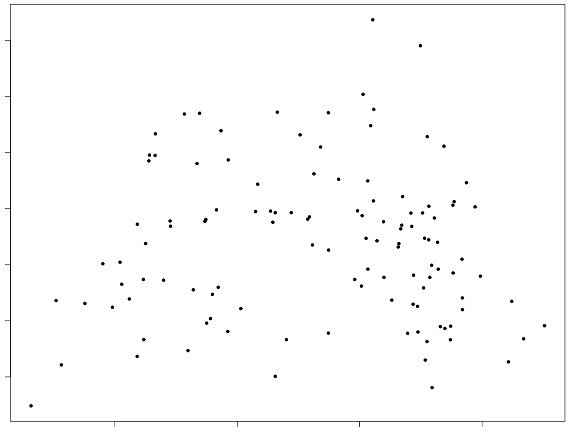


**Figure 19: Olive oil data: Projection of areas 1, 2, and 4 on linear discriminant directions.**

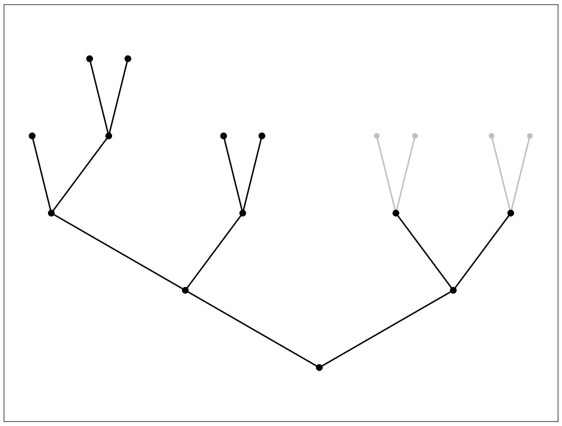


**Figure 20: Simulated olive oil data: Original tree (all nodes) and pruned tree (black nodes).**

and covariance for each area and then generate a sample of the same size as the olive oil data from the corresponding mixture model.

Applying hierarchical model-based clustering with diagonal covariance matrices and using the BIC to estimate the number of mixture components results in a mixture model with 9 components, corresponding to the 9 leaves of the tree shown in Figure 20. Pruning leads to a partition into 7 clusters corresponding to the leaves of the subtree drawn in black, and increases the Fowlkes-Mallows index from 0.71 to 0.86. Figures 21 and 22 show the contingency tables of areas versus clusters before and after pruning, respectively. Pruning removes the split of area 3 and merges the two impure clusters 1 and 2.

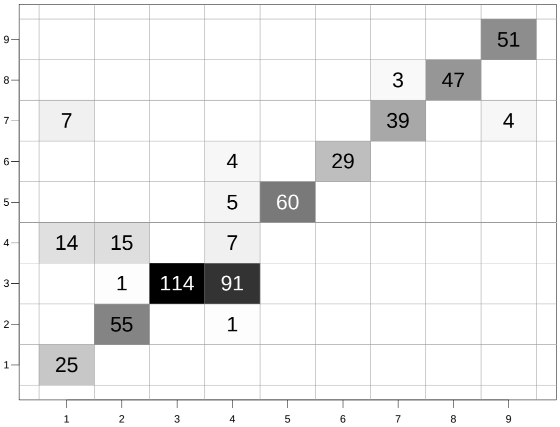
## 6. SUMMARY

The basic premise of model-based clustering is that each group in the data corresponds to a single component of the estimated mixture density. If this premise holds, then the ability to estimate the number of mixture components (equal to the number of groups) is a major strength of model-based clustering compared to nonparametric clustering methods.

On the other hand, if the premise does not hold, the result of model-based clustering can be misleading, because several mixture components may model the same group. Consequently the number of mixture components will overestimate the number of groups, and the clusters corresponding to individual mixture components will no longer be well separated. It is therefore important to be able to decide whether or not the premise holds and, in case the premise does not hold, to determine which mixture components correspond to the same group.

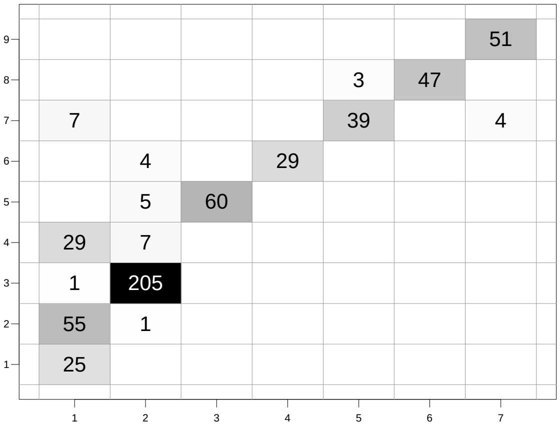
We have introduced methods for assessing the degree of separation between the components of a mixture model, and between the corresponding clusters. We have also presented an algorithm for pruning the cluster tree generated by hierarchical model-based clustering. The algorithm starts with the tree corresponding to the mixture model chosen by the BIC. It then progressively merges clusters that do not appear to correspond to different modes of the data density.

We have applied model-based clustering to a simple synthetic example in which the premise was violated. In this case the method indeed exhibited the deficiencies that we had anticipated. We have also shown that our proposed



|   | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 9 |  |  |  |  |  |  |  |  | 51 |
| 8 |  |  |  |  |  |  | 3 | 47 |  |
| 7 | 7 |  |  |  |  |  | 39 |  | 4 |
| 6 |  |  |  | 4 |  | 29 |  |  |  |
| 5 |  |  |  | 5 | 60 |  |  |  |  |
| 4 | 14 | 15 |  | 7 |  |  |  |  |  |
| 3 |  | 1 | 114 | 91 |  |  |  |  |  |
| 2 |  | 55 |  | 1 |  |  |  |  |  |
| 1 | 25 |  |  |  |  |  |  |  |  |

**Figure 21: Simulated olive oil data: Two way contingency table of areas versus clusters before pruning.**



|   | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 9 |  |  |  |  |  |  | 51 |
| 8 |  |  |  |  | 3 | 47 |  |
| 7 | 7 |  |  |  | 39 |  | 4 |
| 6 |  | 4 |  | 29 |  |  |  |
| 5 |  | 5 | 60 |  |  |  |  |
| 4 | 29 | 7 |  |  |  |  |  |
| 3 | 1 | 205 |  |  |  |  |  |
| 2 | 55 | 1 |  |  |  |  |  |
| 1 | 25 |  |  |  |  |  |  |

**Figure 22: Simulated olive oil data: Two way contingency table of areas versus clusters after pruning.**

diagnostic tools reveal the true structure of the data and lead to more accurate clustering. Application of our new techniques to a real-world example has also been encouraging. Our diagnostics have shown that most probably the premise of model-based clustering was violated in this case as well, and our Hybrid clustering method has significantly improved the quality of the clustering.

## 7. REFERENCES

[1] P. Bickel and J. Fan. Some problems of the estimation of unimodal densities. *Statistica Sinica*, 6:23–45, 1996.

[2] P. Bradley, U. Fayyad, and C. Reina. Scaling EM (expectation-maximization) clustering to large databases. Technical Report MSR-TR-98-35, Microsoft Research, 1999.

[3] J.W Carmichael, G.A. George, and R.S. Julius. Finding natural clusters. *Systematic Zoology*, 17:144–150, 1968.

[4] E. B. Fowlkes and C. L. Mallows. A method for comparing two hierarchical clusterings. *J. American Statistical Association*, 78:553–569, 1983.

[5] C. Fraley and A. Raftery. How many clusters? which clustering method? answers via model-based cluster analysis. *The Computer Journal*, 41(8):578–588, 1998.

[6] R. Gnanadesikan, J.R. Kettenring, and J.M. Landwehr. Projection plots for displaying clusters. In *Statistics and Probability: Essays in Honor of C. R. Rao*, pages 269–280. Elsevier/N.Holland, 1982.

[7] J.A. Hartigan and P.M. Hartigan. The dip test of unimodality. *Annals of Statistics*, 13:70–84, 1985.

[8] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis.* Academic Press, London, 1979.

[9] G.J. McLachlan and D. Peel. *Finite Mixture Models.* John Wiley & Sons, 2000.

[10] P. Sand and A. Moore. Repairing faulty mixture models using density estimation. In *Machine Learning: Proceedings of the eighteenth International Conference*, pages 457–464, 2001.

[11] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:497–511, 1978.

[12] B W Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman & Hall, 1986.

[13] Jeremy Tantrum, Alejandro Murua, and Werner Stuetzle. Hierarchical model-based clustering of large datasets through fractionation and refractionation. In *Proc. 8th Int. Conf. on Knowledge Discovery and Data Mining (KDD02)*, pages 183–190, 2002.