Model-Based Clustering, Discriminant Analysis, and Density Estimation

Chris FRALEY and Adrian E. RAFTERY

Cluster analysis is the automated search for groups of related observations in a dataset. Most clustering done in practice is based largely on heuristic but intuitively reasonable procedures, and most clustering methods available in commercial software are also of this type. However, there is little systematic guidance associated with these methods for solving important practical questions that arise in cluster analysis, such as how many clusters are there, which clustering method should be used, and how should outliers be handled. We review a general methodology for model-based clustering that provides a principled statistical approach to these issues. We also show that this can be useful for other problems in multivariate analysis, such as discriminant analysis and multivariate density estimation. We give examples from medical diagnosis, minefield detection, cluster recovery from noisy data, and spatial density estimation. Finally, we mention limitations of the methodology and discuss recent developments in model-based clustering for non-Gaussian data, high-dimensional datasets, large datasets, and Bayesian estimation.

KEY WORDS: Bayes factor; Breast cancer diagnosis; Cluster analysis; EM algorithm; Gene expression microarray data; Markov chain Monte Carlo; Mixture model; Outliers; Spatial point process.

1. INTRODUCTION

Cluster analysis is the identification of groups of observations that are cohesive and separated from other groups. Interest in clustering has increased recently due to the emergence of several new areas of application. These include datamining, which started from the search for groupings of customers and products in massive retail datasets; document clustering and the analysis of Web use data; gene expression data from microarrays, where one goal is to find of genes that act together; and image analysis, where clustering is used for image segmentation and quantization.

Most clustering done in practice is based largely on heuristic but intuitively reasonable procedures, and most clustering methods available in commercial statistical software are also of this type. One widely used class of methods involves hierarchical agglomerative clustering, in which two groups chosen to optimize some criterion are merged at each stage of the algorithm. Popular criteria include the sum of within-group sums of squares (Ward 1963) and the shortest distance between groups, which underlies the single-link method. Another common class of methods is based on iterative relocation (also called iterative partitioning), in which data points are moved from one group to another until there is no further improvement in some criterion. Iterative relocation with the sum of squares criterion is often called k-means clustering (MacQueen 1967). Although there has been considerable research in this area (e.g., dendrogram analysis for hierarchical clustering), there is little systematic guidance associated with these methods for solving basic practical questions that arise in cluster analysis, such as how many clusters there are, which clustering method should be used, and how outliers should be handled. Moreover, the statistical properties of these methods are generally unknown, precluding the possibility of formal inference.

It was realized early on that cluster analysis can also be based on probability models (see Bock 1996, 1998a, 1998b, for a survey). This realization has provided insight into when a particular clustering method can be expected to work well (i.e., when the data conform to the model), and has led to the development of new clustering methods. It has also been shown that some of the most popular heuristic clustering methods are approximate estimation methods for certain probability models. For example, standard k-means clustering and Ward's method are equivalent to known procedures for approximately maximizing the multivariate normal classification likelihood when the covariance matrix is the same for each component and proportional to the identity matrix.

Finite mixture models have often been proposed and studied in the context of clustering (Wolfe 1963, 1965, 1967, 1970; Edwards and Cavalli-Sforza 1965; Day 1969; Scott and Symons 1971; Duda and Hart 1973; Binder 1978). More recently, it has been recognized that these models can provide a principled statistical approach to the practical questions that arise in applying clustering methods (McLachlan and Basford 1988; Banfield and Raftery 1993; Cheeseman and Stutz 1995; Fraley and Raftery 1998). In finite mixture models, each component probability distribution corresponds to a cluster. The problems of determining the number of clusters and of choosing an appropriate clustering method can be recast as statistical model choice problems, and models that differ in numbers of components and/or in component distributions can be compared. Outliers are handled by adding one or more components representing a different distribution for outlying data.

In this article we describe and review a methodological framework that underlies a powerful approach not just to cluster analysis, but also to some other basic problems of multivariate statistics—discriminant analysis and multivariate density estimation. This strategy arose from the demonstrated promise in clustering applications of two methods based on

Chris Fraley is a research staff member and Adrian E. Raftery is Professor of Statistics and Sociology, Department of Statistics, University of Washington, Box 354322, Seattle WA 98195 (E-mail: *fraley/raftery@stat.washington.edu*). This research was supported by Office of Naval Research grants N00014-96-1-0192 and N00014-96-1-0330. The authors are grateful to William Wolberg for valuable correspondence about the Wisconsin Diagnostic Breast Cancer Data and for providing additional data; to John Castelloe, Gilles Celeux, Danny Walsh, and Naisyin Wang for comments and discussions; and to Simon Byers for the NNclean denoising software.

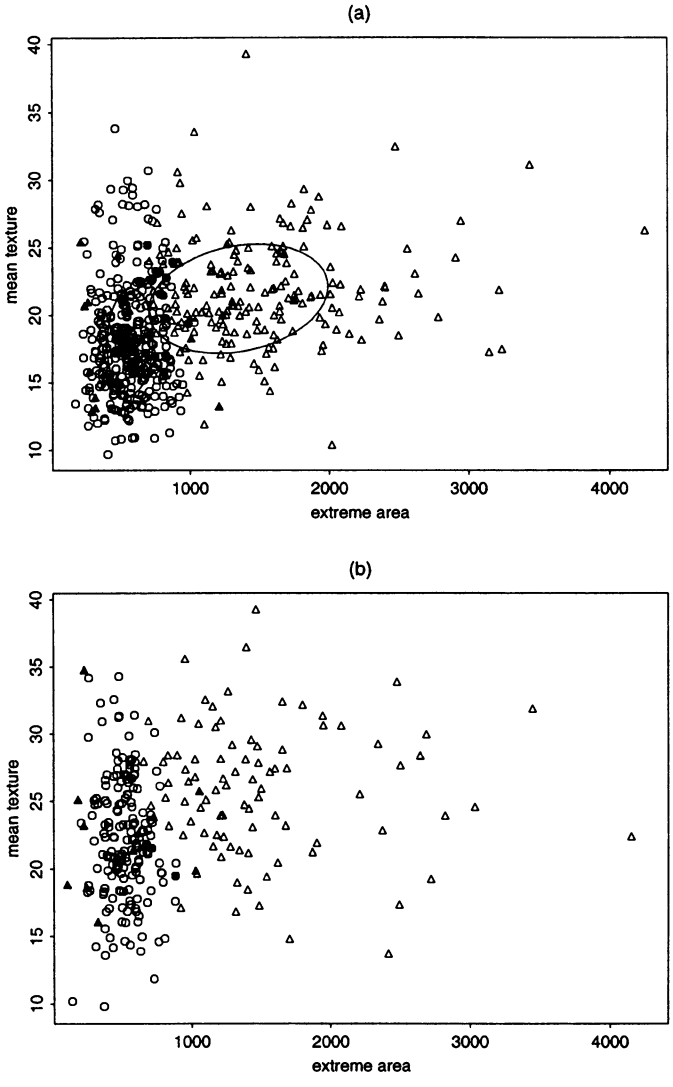

Figure 1. A Projection of the UCI Wisconsin Diagnostic Breast Cancer Data Showing the Two-Group Model-Based Classification (a) and a Projection of 280 Additional Observations (b). The ellipses shown are projections of the ellipsoids defined by the covariances of the two multivariate normal components in the mixture model fitted to the data. There are 569 observations. Although no information about the known malignant versus benign classifications is used by the clustering method, and there is considerable overlap between the two groups, model-based clustering produces a partition that is nearly 95% correct. In (b), the classification produced by the EM-based discriminant analysis technique of Section 6.2, using the UCI Wisconsin Diagnostic Breast Cancer Data as a training set is shown. Circles represent benign observations; triangles, malignant observations. Filled symbols represent misclassified observations. The resulting out-of-sample classification is nearly 96% correct.

multivariate normal mixture models with covariances parameterized by eigenvalue decomposition. These methods are hierarchical agglomeration based on the classification likelihood (Murtagh and Raftery 1984; Banfield and Raftery 1993) and the EM algorithm for maximum likelihood estimation of multivariate mixture models (McLachlan and Basford 1988; Celeux and Govaert 1995). The two approaches are complementary; model-based hierarchical agglomeration tends to produce reasonably good partitions even when started without any information about the groupings, whereas initialization is critical in expectation-maximization (EM) because the likelihood surface tends to have multiple modes, although EM typically produces improved partitions when started from reasonable ones. By initializing EM with partitions from model-based hierarchical agglomeration and using approximate Bayes factors with the Bayesian Information Criterion (BIC) approximation (Schwarz 1978) to determine the number of groups present in the data, Dasgupta and Raftery (1998) achieved good results for some difficult problems in minefield and seismic fault detection. Their method was extended by Fraley and Raftery (1998) to select the parameterization of the model as well as the number of clusters simultaneously using BIC.

Figure 1(a) shows the two-group model-based classification of a dataset used for breast cancer diagnosis (Mangasarian, Street, and Wolberg 1995). Although no information about the known malignant versus benign classifications was used by the clustering method, and there is considerable overlap between the two groups, model-based clustering produced a partition that is nearly 95% correct. Figure 1(b) shows 280 additional data points classified by discriminant analysis with a model-based method described in this article, which makes use of the known classifications. Nearly 96% of these new data points are correctly classified by this procedure. This dataset is discussed in more detail in Section 8.1.

This article reviews the model-based approach to clustering and shows how it can also be applied in discriminant analysis and multivariate density estimation. The organization is as follows. Sections 2–5 include a review of material covered in earlier work (Fraley and Raftery 1998) and elsewhere. Section 2 discusses mixture models, including the multivariate normal model and the geometric interpretation of its parameterization by eigenvalue decomposition. Section 3 covers the EM algorithm for maximum likelihood estimation and its specialization to mixtures. Section 4 gives background on Bayes factors, their approximation via BIC, and their use for selecting the number of clusters and the clustering model. Section 5 describes the overall clustering methodology that combines hierarchical agglomeration, EM, and BIC. Section 6 shows how these ideas can be applied to discriminant analysis, and Section 7 does the same for multivariate density estimation. Section 8 gives examples illustrating these methods. Section 9 gives sources for model-based clustering software. Finally, Section 10 discusses some limitations of the method and suggests extensions to overcome them, including strategies for large datasets.

2. MIXTURE MODELS

Given data $\mathbf{y}$ with independent multivariate observations $\mathbf{y}_1, \ldots, \mathbf{y}_n$, the likelihood for a mixture model with G components is

$$\mathcal{L}_{MIX}(\theta_1, \ldots, \theta_G; \tau_1, \ldots, \tau_G \mid \mathbf{y}) = \prod_{i=1}^{n} \sum_{k=1}^{G} \tau_k f_k(\mathbf{y}_i \mid \theta_k), \quad (1)$$

where f_k and θ_k are the density and parameters of the kth component in the mixture and τ_k is the probability that an observation belongs to the kth component ($\tau_k \geq 0$; $\sum_{k=1}^{G} \tau_k = 1$).

Most commonly, f_k is the multivariate normal (Gaussian) density ϕ_k, parameterized by its mean μ_k and covariance

matrix Σ_k,

$$\phi_k(\mathbf{y}_i \mid \mu_k, \Sigma_k) \equiv \frac{\exp\left\{-\frac{1}{2}(\mathbf{y}_i - \mu_k)^T \Sigma_k^{-1}(\mathbf{y}_i - \mu_k)\right\}}{\sqrt{\det(2\pi\Sigma_k)}}. \quad (2)$$

Data generated by mixtures of multivariate normal densities are characterized by groups or clusters centered at the means μ_k, with increased density for points nearer the mean. The corresponding surfaces of constant density are ellipsoidal. Geometric features (shape, volume, orientation) of the clusters are determined by the covariances Σ_k, which may also be parameterized to impose cross-cluster constraints. Common instances include $\Sigma_k = \lambda I$, where all clusters are spherical and of the same size; $\Sigma_k = \Sigma$ constant across clusters, where all clusters have the same geometry but need not be spherical (Friedman and Rubin 1967); and unrestricted Σ_k, where each cluster may have a different geometry (Scott and Symons 1971). For $\Sigma_k = \lambda I$, only one parameter is needed to characterize the covariance structure of the mixture, whereas $d(d+1)/2$ and $G(d(d+1)/2)$ parameters are required for constant Σ_k and unrestricted Σ_k if the data are d-dimensional.

Banfield and Raftery (1993) proposed a general framework for geometric cross-cluster constraints in multivariate normal mixtures by parameterizing covariance matrices through eigenvalue decomposition in the form

$$\Sigma_k = \lambda_k D_k A_k D_k^T, \quad (3)$$

where D_k is the orthogonal matrix of eigenvectors, A_k is a diagonal matrix whose elements are proportional to the eigenvalues, and λ_k is an associated constant of proportionality. Their idea was to treat λ_k, A_k, and D_k as independent sets of parameters and either constrain them to be the same for each cluster or allow them to vary among clusters. When parameters are fixed, clusters will share certain geometric properties; D_k governs the orientation of the kth component of the mixture, A_k its shape, and λ_k its volume, which is proportional to $\lambda_k^d \det(A_k)$. For example, if the largest eigenvalue of Σ_k is much larger than the other eigenvalues, then the kth cluster will be concentrated close to a line in d-space, which will be the first principal component of the distribution of the kth group. Similarly, if the two largest eigenvalues are of the same magnitude and dominate the other eigenvalues, then the kth cluster will be concentrated close to a plane in d-space. The kth cluster will be roughly spherical if the largest and smallest eigenvalues of Σ_k are of the same magnitude.

This approach generalizes the work of Murtagh and Raftery (1984), who used the equal shape/equal volume model ($\Sigma_k = \lambda D_k A D_k^T$) for clustering in character recognition and other situations involving thin, highly linear, and possibly overlapping clusters with different orientations. It also subsumes the three most common models—λI, equal variance, and unconstrained variance—mentioned earlier, as well as other useful models, such as $\Sigma_k = \lambda_k I$, where the clusters are spherical but have different volumes, and $\Sigma_k = \lambda_k A_k$, where all covariances are diagonal but otherwise their shapes, sizes, and orientations are allowed to vary. For an extensive enumeration of possible models resulting from (3), see Celeux and Govaert 1995.

Other proposed parsimonious parameterizations of covariance matrices could be applied in the context of cluster analysis. These include the intraclass correlation or one-factor model, in which all of the off-diagonal elements of the correlation matrix are equal, generalizations of this based on factor analysis and structural equations (e.g., Jöreskog 1973; Bollen 1989), autoregressive and other parameterizations common in time series (Box and Jenkins 1976), and models common in geostatistics in which covariances are functions of distance (e.g., Journel and Huijbrechts 1978) in either a Euclidean or a deformed space (Sampson and Guttorp 1992).

3. THE EXPECTATION–MAXIMIZATION ALGORITHM FOR MIXTURE MODELS

The EM algorithm (Dempster, Laird, and Rubin 1977; McLachlan and Krishnan 1997) is a general approach to maximum likelihood estimation for problems in which the data can be viewed as consisting of n multivariate observations $\mathbf{x}_i$ recoverable from $(\mathbf{y}_i, \mathbf{z}_i)$, in which $\mathbf{y}_i$ is observed and $\mathbf{z}_i$ is unobserved. If the $\mathbf{x}_i$ are independent and identically distributed (iid) according to a probability distribution f with parameters θ, then the *complete-data likelihood* is

$$\mathcal{L}_C(\mathbf{x}_i \mid \theta) = \prod_{i=1}^{n} f(\mathbf{x}_i \mid \theta).$$

Further, if the probability that a particular variable is unobserved depends only on the observed data $\mathbf{y}$ and not on $\mathbf{z}$, then the *observed-data likelihood*, $\mathcal{L}_O(\mathbf{y} \mid \theta)$, can be obtained by integrating $\mathbf{z}$ out of the complete-data likelihood,

$$\mathcal{L}_O(\mathbf{y} \mid \theta) = \int \mathcal{L}_C(\mathbf{x} \mid \theta)\, d\mathbf{z}. \quad (4)$$

The maximum likelihood estimate (MLE) for θ based on the observed data maximizes $\mathcal{L}_O(\mathbf{y} \mid \theta)$.

The EM algorithm alternates between two steps, an "E step," in which the conditional expectation of the complete-data log-likelihood given the observed data and the current parameter estimates is computed, and an "M step," in which parameters that maximize the expected log-likelihood from the E step are determined. The unobserved portion of the data may involve values that are missing due to nonresponse and/or quantities that are introduced to reformulate the problem for EM. Under fairly mild regularity conditions, EM can be shown to converge to a local maximum of the observed-data likelihood (e.g., Dempster et al. 1977; Boyles 1983; Wu 1983; McLachlan and Krishnan 1997). Although these conditions do not always hold in practice, the EM algorithm has been widely used for maximum likelihood estimation for mixture models with good results.

In EM for mixture models, the "complete data" are considered to be $\mathbf{x}_i = (\mathbf{y}_i, \mathbf{z}_i)$, where $\mathbf{z}_i = (z_{i1}, \ldots, z_{iG})$ is the unobserved portion of the data, with

$$z_{ik} = \begin{cases} 1 & \text{if } \mathbf{x}_i \text{ belongs to group } k \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

Assuming that each $\mathbf{z}_i$ is iid according to a multinomial distribution of one draw from G categories with probabilities

$\tau_1, \ldots, \tau_G$, and that the density of an observation $\mathbf{y}_i$ given $\mathbf{z}_i$ is given by $\prod_{k=1}^{G} f_k(\mathbf{y}_i \mid \theta_k)^{z_{ik}}$, the resulting complete-data log-likelihood is

$$l(\theta_k, \tau_k, z_{ik} \mid \mathbf{x}) = \sum_{i=1}^{n} \sum_{k=1}^{G} z_{ik} \log[\tau_k f_k(\mathbf{y}_i \mid \theta_k)]. \quad (6)$$

The E step of the EM algorithm for mixture models is given by

$$\hat{z}_{ik} \leftarrow \frac{\hat{\tau}_k f_k(\mathbf{y}_i \mid \hat{\theta}_k)}{\sum_{j=1}^{G} \hat{\tau}_j f_j(\mathbf{y}_i \mid \hat{\theta}_j)}, \quad (7)$$

while the M step involves maximizing (6) in terms of τ_k and θ_k with z_{ik} fixed at the values computed in the E step, $\hat{z}_{ik}$. The value z_{ik}^* of $\hat{z}_{ik}$ at a maximum of (1) is the estimated conditional probability that observation i belongs to group k. The maximum likelihood classification of observation i is $\{j \mid z_{ij}^* = \max_k z_{ik}^*\}$, so that $(1 - \max_k z_{ik}^*)$ is a measure of the uncertainty in the classification (Bensmail, Celeux, Raftery, and Robert 1997).

For multivariate normal mixtures, the E step is given by (7) with f_k replaced by ϕ_k as defined in (2), regardless of the parameterization. For the M step, estimates of the means and probabilities have simple closed-form expressions involving the data and $\hat{z}_{ik}$ from the E step,

$$\hat{\tau}_k \leftarrow \frac{n_k}{n}; \qquad \hat{\mu}_k \leftarrow \frac{\sum_{i=1}^{n} \hat{z}_{ik} \mathbf{y}_i}{n_k}; \qquad n_k \equiv \sum_{i=1}^{n} \hat{z}_{ik}. \quad (8)$$

Computation of the covariance estimate $\widehat{\Sigma}_k$ depends on its parameterization. Details of the M step for Σ_k parameterized by the eigenvalue decomposition (3) have been given by Celeux and Govaert (1995).

EM estimation for mixture models has a number of limitations. First, the rate of convergence can be slow. However, EM typically gives good results if the data conform reasonably well to the model and the iteration is started at reasonable values. Second, the EM algorithm for multivariate normal mixtures breaks down when the covariance associated with one or more components is singular or nearly singular. It may either fail or give inaccurate results if one or more clusters contain only a few observations (which can happen if there are too many components in the mixture), or if the observations that they contain are concentrated close to a linear subspace of lower dimension than the data.

A variant of EM called *classification EM* (CEM) (Celeux and Govaert 1992), in which the $\hat{z}_{ik}$ are converted to a discrete classification before performing the M step, is equivalent to standard k-means clustering (MacQueen 1967; Hartigan and Wong 1978) when a uniform spherical Gaussian distribution is used as the probability model. It should be noted that CEM is a procedure for maximizing the classification likelihood (10) discussed in Section 5.1 rather than the mixture likelihood (Celeux and Govaert 1993).

4. MODEL SELECTION

Two basic issues arising in applied cluster analysis are selection of the clustering method and determination of the number of clusters. In the mixture modeling approach, these issues can be reduced to a single concern, that of model selection. Recognizing that each combination of a number of groups and a clustering model corresponds to a different statistical model for the data reduces the problem to comparison among the members of a set of possible models.

There are trade-offs between the choice of the number of clusters and that of the clustering model. If a simpler model is used, then more clusters may be needed to provide a good representation of the data. If a more complex model is used, then fewer clusters may suffice. As a simple example, consider the situation with a single Gaussian cluster whose covariance matrix corresponds to a long, thin ellipsoid. If a model with equal-volume spherical components (the model underlying Ward's method and the k-means method) were used to fit this data, then more than one hyperspherical cluster would be needed to approximate the single elongated ellipsoid.

Our approach to the problem of model section in clustering is based on Bayesian model selection via Bayes factors and posterior model probabilities (e.g., Kass and Raftery 1995). The basic idea is that if several models, $M_1, \ldots, M_K$, are considered, with prior probabilities $p(M_k), k = 1, \ldots, K$ (often taken to be equal), then, by Bayes's theorem, the posterior probability of model M_k given data D is proportional to the probability of the data given model M_k, times the model's prior probability, namely

$$p(M_k|D) \propto p(D|M_k)p(M_k).$$

When there are unknown parameters, by the law of total probability, $p(D|M_k)$ is obtained by integrating (not maximizing) over the parameters, that is,

$$p(D|M_k) = \int p(D|\theta_k, M_k)p(\theta_k|M_k)\, d\theta_k,$$

where $p(\theta_k|M_k)$ is the prior distribution of θ_k, the parameter vector for model M_k. The quantity $p(D|M_k)$ is known as the *integrated likelihood* of model M_k.

A natural Bayesian approach to model selection is then to choose the model that is most likely a posteriori. If the prior model probabilities, $p(M_k)$, are the same, then this amounts to choosing the model with the highest integrated likelihood. For comparing two models, M_1 and M_2, the Bayes factor is defined as the ratio of the two integrated likelihoods, $B_{12} = p(D|M_1)/p(D|M_2)$, with the comparison favoring M_1 if $B_{12} > 1$ and conventionally being viewed as providing very strong evidence for M_1 if $B_{12} > 100$ (Jeffreys 1961). Often, values of $2\log(B_{12})$ rather than B_{12} are reported, and on this scale, rounding, very strong evidence corresponds to a threshold of 10 (Kass and Raftery 1995).

This approach is appropriate in the present context, because it applies when there are more than two models and can be used for comparing nonnested models. Besides being a Bayesian solution to the problem, it has some desirable frequentist properties. For example, if one has just two models and they are nested, then basing model choice on the Bayes factor minimizes the total error rate, which is the sum of the type I and type II error rates (Jeffreys 1961).

The main difficulty in using Bayes factors is the evaluation of the integral that defines the integrated likelihood. For

regular models, the integrated likelihood can be approximated simply by the BIC,

$$2\log p(D|M_k) \approx 2\log p(D|\hat{\theta}_k, M_k) - \nu_k log(n) = BIC_k, \quad (9)$$

where ν_k is the number of independent parameters to be estimated in model M_k (Schwarz 1978; Haughton 1988). This approximation is particularly good when a unit information prior is used for the parameters, that is, a prior that contains the amount of information provided on average by one observation (Kass and Wasserman 1995; Raftery 1995). The reasonableness of this prior has been discussed by Raftery (1999).

Finite mixture models do not satisfy the regularity conditions that underly the published proofs of (9), but several results suggest its appropriateness and good performance in the model-based clustering context. Leroux (1992) showed that basing model selection on a comparison of BIC values will not underestimate the number of groups asymptotically, and Keribin (1998) showed that BIC is consistent for the number of groups. Roeder and Wasserman (1997) showed that if a mixture of (univariate) normals is used for one-dimensional nonparametric density estimation, then using BIC to choose the number of components yields a consistent estimator of the density. Finally, in a range of applications of model-based clustering, model choice based on BIC has given good results (Campbell, Fraley, Murtagh, and Raftery 1997; Campbell, Fraley, Stanford, Murtagh, and Raftery 1999; DasGupta and Raftery 1998; Fraley and Raftery 1998; Stanford and Raftery 2000).

Several other approaches to choosing the number of clusters in model-based clustering have been proposed. McLachlan and Basford (1988) discussed the use of resampling in this context. Banfield and Raftery (1993) derived an approximation to the integrated likelihood based on the classification likelihood, called the Approximate Weight of Evidence (AWE), but in subsequent experiments it has consistently performed less well than BIC. Cheeseman and Stutz (1995) and Chickering and Heckerman (1997) used a different approximation to the integrated likelihood; other approaches include an informational complexity criterion called ICOMP (Bozdogan 1994), an entropy criterion called NEC (Celeux and Soromenho 1996; Biernacki, Celeux, and Govaert 1999), the integrated classification likelihood (Biernacki et al. 2000), and cross-validated likelihood (Smyth 2000). These methods were developed for choosing the number of clusters, but presumably they could be either applied or extended to choose the clustering model as well. The performances of some of these criteria were compared by Biernacki and Govaert (1999). Bensmail et al. (1997) discussed an alternative approximation to the integrated likelihood for choosing both the number of groups and the clustering model based on Markov chain Monte Carlo (MCMC) estimation of the models.

5. CLUSTER ANALYSIS

The purpose of *cluster analysis* is to classify data of previously unknown structure into meaningful groupings. In this section we outline a strategy for cluster analysis based on mixture models. We use the parameterization (3) as the basis for a class of models that is sufficiently flexible to accommodate data with widely varying characteristics. The strategy comprises three core elements: initialization via model-based hierarchical agglomerative clustering, maximum likelihood estimation via the EM algorithm, and selection of the model and the number of clusters using approximate Bayes factors with the BIC approximation.

5.1 Model-Based Hierachical Clustering

Model-based hierarchical agglomerative clustering is an approach to computing an approximate maximum for the *classification likelihood*,

$$\mathcal{L}_{CL}(\theta_1, \ldots, \theta_G; \ell_1, \ldots, \ell_n \mid \mathbf{y}) = \prod_{i=1}^{n} f_{\ell_i}(\mathbf{y}_i \mid \theta_{\ell_i}), \quad (10)$$

where the ℓ_i are labels indicating a unique classification of each observation, $\ell_i = k$ if $\mathbf{y}_i$ belongs to the kth component. In the mixture likelihood (1), each component is weighted by the probability that an observation belongs to that component. The presence of the class labels in the classification likelihood (10) introduces a combinatorial aspect that makes exact maximization impractical.

Murtagh and Raftery (1984) successfully applied model-based agglomerative hierarchical clustering to problems in character recognition using a multivariate normal model parameterized as in (3), with volume and shape (λ_k and A_k) held constant across clusters. This approach was generalized by Banfield and Raftery (1993) to other models and applications, including tissue segmentation in medical images.

Model-based agglomerative hierarchical clustering proceeds by successively merging pairs of clusters corresponding to the greatest increase in the classification likelihood (10) among all possible pairs. In the absence of any information about groupings, the procedure starts by treating each observation as a singleton cluster. When the probability model in (10) is multivariate normal with the equal-volume spherical covariance λI, the selection criterion is the well-known sum-of-squares criterion (Ward 1963).

Other common heuristic clustering criteria, such as *single link* (nearest neighbor), *complete link* (farthest neighbor), and *average link*, have no known associated statistical model. However, there may be relationships that have yet to be uncovered. The criterion underlying complete link clustering is close to, but not the same as, the classification likelihood for a model in which each group is uniformly distributed on a hypersphere, with the same radius for each group. The criterion underlying average-link clustering has some similarities with the classification likelihood for a model in which each group has a multivariate isotropic Laplace distribution, with density $f(\mathbf{y}) \propto \exp\{-|\mathbf{y}-\mu|/\sigma\}$. Further investigation of such connections may provide insight into when complete-link and average-link clustering are most likely to work well. They may also point to more fully model-based methods along the same lines, as well as to generalizations to nonisotropic settings or situations in which the groups differ markedly. The single-link clustering method seems to be not related to a statistical model and does not perform well in instances where clusters are not well separated (e.g., Fraley and Raftery 1998).

However, nearest-neighbor classification, the supervised analog of single-link clustering, often works well for discriminant analysis.

In the heuristic methods, the computational cost of merging pairs of clusters remains fixed as long as the clusters remain unchanged, and computational methods that store and update these costs are much faster than the alternatives, provided that sufficient memory is available. Many model-based methods can also be implemented in this way, although evaluating the merge criterion can involve a relatively expensive computation, such as a determinant or an eigenvalue decomposition. Hierarchical agglomeration should be avoided with those multivariate normal models, such as constant variance, for which there is no advantage in storing the cost of merging pairs, unless an initial partition with a small number of groups is available. An alternative model, such as the one with unconstrained variance, can be used in these cases. Efficient numerical algorithms for agglomerative hierarchical clustering based on (10) with multivariate normal models have been discussed by Fraley (1998).

5.2 Combining Hierarchical Agglomeration, EM, and Bayes Factors

In hierarchical agglomeration, each stage of merging corresponds to a unique number of clusters and a unique partition of the data. A given partition can be transformed into indicator variables (5), which can then be used as conditional probabilities in an M step of EM for parameter estimation, initializing an EM algorithm. This, combined with Bayes factors as approximated by BIC for model selection, yields a comprehensive clustering strategy:

- Determine a maximum number of clusters, M, and a set of mixture models to consider.
- Perform hierarchical agglomeration to approximately maximize the classification likelihood for each model, and obtain the corresponding classifications for up to M groups.
- Apply the EM algorithm for each model and each number of clusters $2, \ldots, M$, starting with the classification from hierarchical agglomeration.
- Compute BIC for the one-cluster case for each model and for the mixture model with the optimal parameters from EM for $2, \ldots, M$ clusters.

Strong evidence for a model and an associated number of clusters is taken to correspond to a decisive maximum of the BIC.

Multivariate normal mixtures parameterized through eigenvalue decomposition as in (3) represent a good set of models for clustering in many situations arising in practice. With these models, computation can be saved by doing hierarchical agglomeration for only one of the models (e.g., unconstrained covariance), using the resulting partitions as starting values for EM with any other parameterization. This method for model-based clustering is illustrated in the examples of Sections 8.1 and 8.2.

5.3 Modeling Noise and Outliers

Noise and outliers can often be handled in this framework by adding a term or terms to the mixture to represent "nonconforming" data. A mixture in which one component models noise as a homogeneous Poisson process has been used successfully in a number of applications (Banfield and Raftery 1993; Dasgupta and Raftery 1998; Campbell et al. 1997, 1999). The corresponding model is

$$\widetilde{\mathcal{L}}_{MIX}(\theta_1, \ldots, \theta_G; \tau_0, \tau_1, \ldots, \tau_G \mid \mathbf{y}) = \prod_{i=1}^{n}\left[\frac{\tau_0}{V} + \sum_{k=1}^{K}\tau_k \phi_k(\mathbf{x}_i \mid \theta_k)\right], \quad (11)$$

in which V is the hypervolume of the data region, $\tau_k \geq 0$, and $\sum_{k=0}^{G}\tau_k = 1$. Isolated outliers can sometimes be treated by *iterated sampling* (e.g., Fayyad and Smyth 1996), in which points of low probability are removed from clusters and the clustering/removal process is repeated until all remaining observations have relatively high density. Alternatively, noise can be modeled in mixtures via the t distribution (Peel and McLachlan 2000).

When the data contain a great deal of noise, the basic model-based clustering method of Section 5.2 needs to be modified as follows:

- Obtain an initial categorization of each observation as being "data" or "noise." Some possible methods for denoising include a Voronoï method (Allard and Fraley 1997) and a nearest-neighbor method (Byers and Raftery 1998).
- Apply hierarchical clustering to the denoised data.
- Apply EM based on the Gaussian model with the added noise term(s) to the entire dataset. Initial values for z_{ik} are formed by augmenting the indicator variables from the hierarchical clustering step with a row of 0s for each observation initially assessed as being noise and a column of indicator variables giving the result of the denoising step (1 indicating noise and 0 otherwise).

An example of model-based clustering with very noisy data is given in Section 8.3.

6. DISCRIMINANT ANALYSIS

6.1 Discriminant Analysis Background

In discriminant analysis, also known as *supervised classification*, known classifications of some observations (the "training set") are used to classify others (e.g., McLachlan 1992; Ripley 1996). The number of classes, G, is assumed to be known.

Many discriminant analysis methods are probabilistic, based on the assumption that the observations in the kth class are generated by a probability distribution specific to that class, $f_k(\cdot)$. Then, if τ_k is the proportion of members of the population that are in class k, Bayes's theorem says that the posterior probability that an observation $\mathbf{y}$ belongs to class k is

$$\Pr[\mathbf{y} \in \text{class } j] = \frac{\tau_j f_j(\mathbf{y})}{\sum_{k=1}^{G}\tau_k f_k(\mathbf{y})}.$$

Assigning $\mathbf{y}$ to the class to which it has the highest posterior probability of belonging minimizes the expected misclassification rate; this is called the Bayes classifier.

Most commonly used discriminant analysis methods are based on the assumption that the observations in each class are multivariate normal, so that

$$f_k(\mathbf{y}) = \phi(\mathbf{y}|\mu_k, \Sigma_k). \tag{12}$$

If the covariance matrices for the different classes are the same (i.e., $\Sigma_k = \Sigma$ for $k = 1, \ldots, G$), and if maximum likelihood estimates of μ_k and Σ from training data are used, then the (conditional) Bayes classifier is Fisher's linear discriminant analysis (LDA) rule. In that case, the classification rule is defined by whether or not a linear combination of the components of $\mathbf{y}$ exceeds a threshold. This reduces the discrimination to a one-dimensional problem and produces a classification rule that is a simple thresholding. If the covariance matrices Σ_k are allowed to differ without constraint, then the resulting method is standard quadratic discriminant analysis (QDA), in which the classification function is a quadratic form in the components of $\mathbf{y}$. The ideas discussed in this review allow the standard LDA and QDA to be extended in several ways, described in more detail in the next two sections.

6.2 Eigenvalue Decomposition Discriminant Analysis

Bensmail and Celeux (1996) imposed cross-group constraints on the class covariance matrices in (12) for discriminant analysis, based on the parameterization by eigenvalue decomposition (3) originally proposed for model-based clustering. This approach, called eigenvalue decomposition discriminant analysis (EDDA), has the advantage of permitting more flexibility than LDA while at the same time allowing more structure than the unconstrained model underlying QDA, which may have too many parameters to perform optimally. They considered 14 possible models for the covariances based on (3), allowing the data to choose between them using cross-validation. The best model could alternatively be chosen using approximate Bayes factors, as we have proposed for clustering (Sec. 4), which would typically be less demanding computationally. Biernacki and Govaert (1999) compared a number of different criteria, including BIC, in simulation studies of model-based clustering and discriminant analysis. In a related but different context, Stanford and Raftery (2000) found that BIC and cross-validation tended to choose similar models, with BIC requiring far less computation.

A single EM iteration provides a simple way of assigning new observations to known classes, so that the framework described earlier for model-based clustering can easily be adapted to discriminant analysis. First, an M step is carried out for the appropriate model, with indicator variables corresponding to the known discrete labels of the training set as starting values (5). This yields approximate parameters $\tilde{\theta}$ and mixing proportions $\tilde{\tau}$ for the model. (The mixing proportions can be treated separately if they are known in advance.) Then an E step is computed for the new observations using the parameters from the "discrete" M step to obtain the conditional probability that each new object belongs to each of the possible groups in the mixture. An observation $\mathbf{y}_i$ is assigned to the group for which it has the highest conditional probability,

$$\max_j \frac{\tilde{\tau}_j f_j(\mathbf{y}_i \mid \tilde{\theta}_j)}{\sum_{k=1}^{G} \tilde{\tau}_k f_k(\mathbf{y}_i \mid \tilde{\theta}_k)}. \tag{13}$$

If the parameter estimates were replaced with the true parameters for the population, then this discriminant rule would correspond to the optimal Bayes rule.

A simple extension allows all of the data (training and new) to be taken into account when estimating the parameters, even when the size of the training set is too small to provide a basis for standard discriminant analysis techniques. The EM algorithm is applied as before to all of the data, except that the $\hat{z}_{ik}$ for the training data are constrained to be 0 or 1 throughout the algorithm, reflecting the known group memberships.

6.3 Mixture Discriminant Analysis

An alternative model-based approach to generalizing LDA and QDA is to allow the density for each class itself to be a mixture of normals, namely

$$f_j(\mathbf{y} \mid \theta_k) = \sum_{k=1}^{G_j} \tau_{jk} \phi(\mathbf{y} \mid \mu_{jk}, \Sigma_{jk}). \tag{14}$$

This idea has been suggested a number of times in the literature (e.g., Scott 1992; McLachlan 1992) and is the basis of mixture discriminant analysis (MDA) (Hastie and Tibshirani 1996). In developing MDA, Hastie and Tibshirani made two assumptions: that all of the component covariance matrices are the same (i.e., $\Sigma_{jk} = \Sigma$ for each j, k) and that the number of mixture components is known in advance for each class. But when learning vector quantization (Kohonen 1989) is used for initialization, only the total combined number of mixture components for all classes needs to be specified at the outset. Hastie and Tibshirani also proposed several extensions of the method under these assumptions. In a similar approach, Ormoneit and Tresp (1998) use unconstrained mixtures with a fixed number of components, averaged over parameters estimated via EM with a number of different random starting values.

MDA can also be extended by relaxing the aforementioned two assumptions and applying model-based clustering to the members of each class in the training set. This would allow the component covariance matrices to vary, both within and between classes, perhaps with some cross-component constraints. The data would then determine which parameterization of the covariance matrix and which number of mixture components is best suited to each class. We call this generalization of MDA MclustDA.

The basic idea of the model-based discriminant analysis methods described here is to allow more flexibility than is possible with the traditional methods of LDA and QDA. Friedman (1989) earlier proposed an approach to this problem called regularized discriminant analysis (RDA), which chooses a linear combination of the LDA and QDA models that best fits the data. EDDA (Bensmail and Celeux 1996) provides a class of models that are intermediate between LDA and QDA, while remaining geometrically or substantively interpretable.

Mixture-based MDA and MclustDA further improve on EDDA by expanding the discriminant model from a single Gaussian component to a mixture. In particular, this approach allows close approximation of nonlinear and nonmonotonic classification boundaries. Under fairly weak conditions, a mixture model can approximate a given density arbitrarily

closely given enough components, allowing great flexibility. In MclustDA, the data choose both the number of components in each class and the form of the covariance matrices, so that the method could revert to LDA or QDA for some datasets and use a large number of components (and thus be almost "nonparametric") for others.

7. DENSITY ESTIMATION

In density estimation, it is the value of the mixture likelihood at individual points that is of interest, rather than the membership of the components, which is important in clustering or discriminant analysis. Roeder and Wasserman (1997) used normal mixtures for univariate density estimation, with BIC to determine the number of components. The model-based clustering method of Section 5 can be viewed as leading to a multivariate extension of their method, because the parameter estimates for the best model define a multivariate mixture density for the data. However, the issue of choosing a probability model for the individual components is less critical in one dimension and was not discussed by Roeder and Wasserman (1997). In one dimension there are only two possible models (equal and unequal variance), whereas many more models are possible in the multivariate case, so that the available set of models and model selection procedures play a critical role in density estimation by multivariate normal mixtures. Results of simulations for two-dimensional analogs of the univariate mixtures from Marron and Wand (1992) that were studied by Roeder and Wasserman (1997) are presented in Section 8.5, and some applications are illustrated in Sections 8.1.3 and 8.4.

An alternative approach to density estimation using normal mixtures models the normal parameters as coming from a Dirichlet process. This approach was proposed for one-dimensional density estimation by Escobar and West (1995) and MacEachern and Müller (1998), and extended to the multivariate case by Müller, Erkanli, and West (1996). Roeder and Wasserman (1997) argued for directly selecting the number of components rather than modeling it using a Dirichlet process, on the grounds that the former allows direct control over the number of components.

8. EXAMPLES

8.1 UCI Wisconsin Diagnostic Breast Cancer Data

8.1.1 Cluster Analysis. In widely publicized work (e.g., Mangasarian et al. 1995), 176 consecutive future cases were successfully diagnosed from 569 instances through the use of linear programming techniques to locate planes separating classes of data. These results were based on 3 out of 30 attributes: extreme area, extreme smoothness, and mean texture. The three explanatory variables were chosen via cross-validation comparing methods using all subsets of two, three, and four features and one or two linear separating planes. The training data is available from the UCI Machine Learning Repository at *http://www.ics.uci.edu/AI/ML/MLDBRepository.html*. The three variables of interest are shown in Figure 2.

Although the diagnoses are available for these data, we first applied cluster analysis to the three attributes only, ignoring the "known" classifications. The model-based clustering

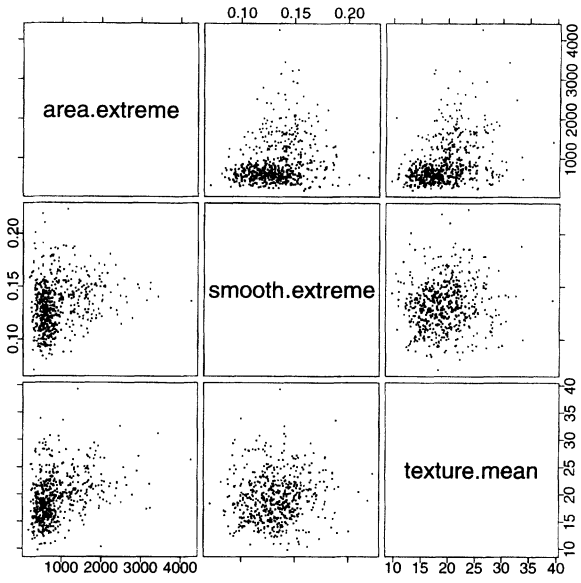

Figure 2. Pairs Plots of the Wisconsin Diagnostic Breast Cancer Data From the UCI Machine Learning Repository. Only the three explanatory variables used by Mangasarian et al. (1995) are shown. There are 569 observations.

methodology outlined in Section 5 yields the results shown in Figure 3. The maximum BIC value occurs for the three-group unconstrained model; the difference in BIC values between the two- and three-group unconstrained models is small enough to conclude that there are either two or three groups in the data [Fig. 3(a)]. The two-group classification matches the clinical diagnosis for all but 29 of the 569 observations (see Fig. 1). Note that the most uncertain points tend to fall in the same region between the two clusters as the misclassified data [Fig. 3(c)], whereas the location of uncertainty of the misclassified observations relative to the uncertainty of all of the observations [Fig. 3(d)] confirms that the more uncertain observations are also the ones most likely to be misclassified.

The finding that there may be three groups in the data is clinically important, because it is necessary to have some idea of the chance of malignancy to determine an appropriate course of action. Tumors of the intermediate class would be followed up by biopsy under local anesthesia, whereas those likely to be malignant would be followed up by a more invasive biopsy under general anesthesia.

8.1.2 Discriminant Analysis With One Gaussian Component per Group. According to the documentation for the Wisconsin Diagnostic Breast Cancer Data in the UCI Machine Learning Repository, the classifier proposed by Mangasarian et al. (1995) correctly diagnosed 176 consecutive new patients as of November 1995. Because only the training set is available from the UCI repository, we obtained additional data for discriminant analysis from Dr. William Wolberg of the University of Wisconsin, the oncologist involved in the original analysis of these data. Using parameter estimates generated via an M step of EM started from the known discrete classification of the UCI data (with two groups) model-based discriminant analysis via (13) classified 280 new observations with 95.7% accuracy [see Fig. 1(b)]. The model-based approach has some advantages over the linear programming method of

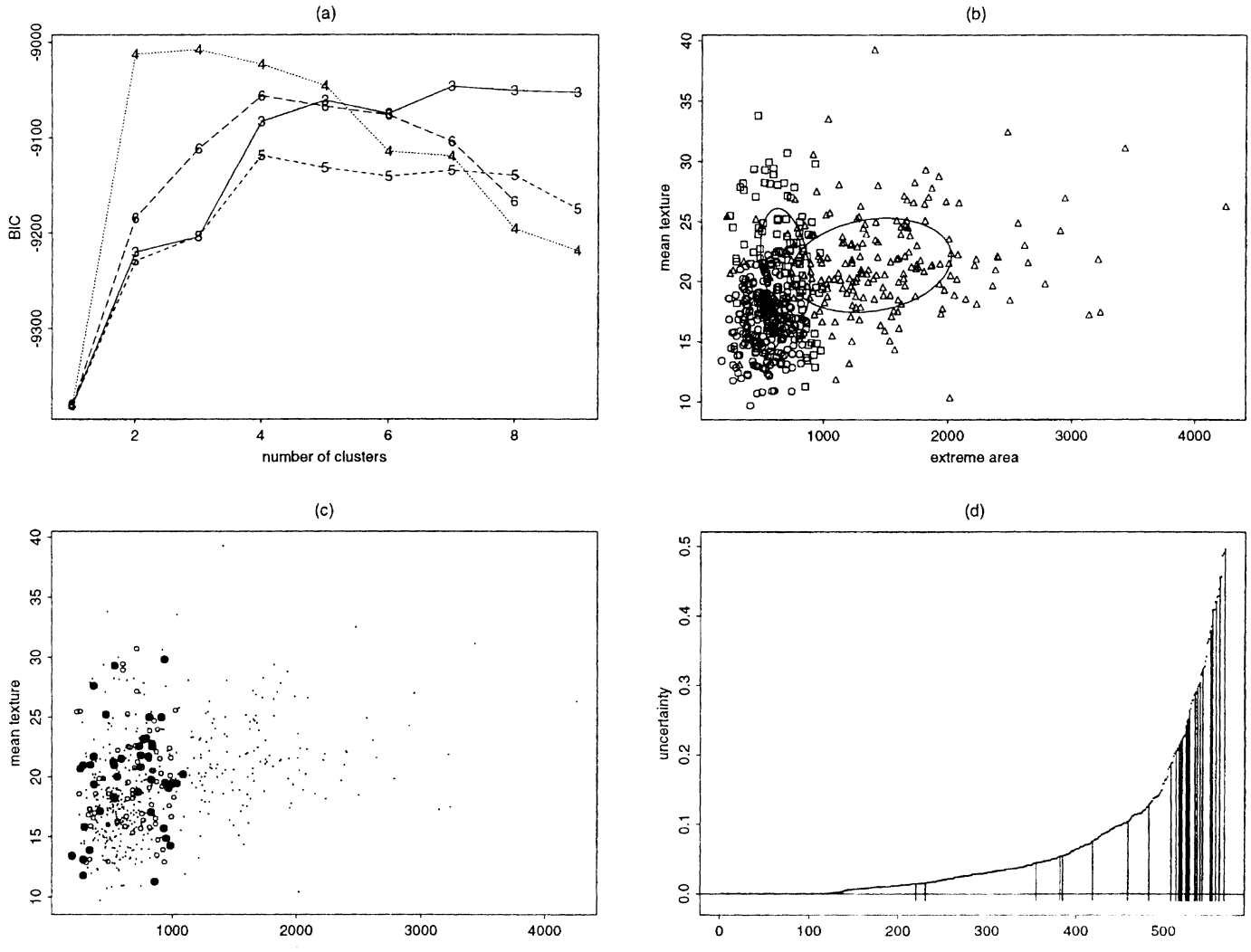

Figure 3. Cluster Analysis of the Wisconsin Diagnostic Data Reduced to the Three Explanatory Variables. (a) BIC values, excluding those for the two spherical models, because they fall well below the others. Models 3–6 correspond to Σ (equal variance), Σ_k (unconstrained), $\lambda D_k A D_k^T$ (common shape and volume), and $\lambda_k D_k A D_k^T$ (common shape). Model 4 is the best model. (b) The three-group unconstrained model-based classification of the data, showing the projections of the ellipses defined by the covariance of each of the three groups. (c) Uncertainty in the two-group model-based classification (shown in Fig. 1). Small dots correspond to observations with uncertainty less than .1; open circles, to those with uncertainty in the interval [.1, .25); and filled circles, to those with uncertainty greater than or equal to .25. (d) Location of the misclassified observations (vertical lines) relative to the uncertainties of all observations in the two-group model-based classification.

Mangasarian et al. (1995)—it generalizes easily to data in which more than two groups are present, and the groups need not be linearly separable.

8.1.3 Discriminant Analysis With a Mixture for Each Group. One application of density estimation is the computation of likelihood ratios for discriminant analysis (e.g., Scott 1992, chap. 9). A model is fitted to each of two sets of data known to have different values of a particular characteristic, and the ratio of their densities is computed over a range of values. When the model-based clustering methodology described here is used for each class, this is an application of MclustDA, the generalization of mixture discriminant analysis (Hastie and Tibshirani 1996) described in Section 6.3.

Contour and perspective plots of parametric and nonparametric likelihood ratio surfaces for diseased versus nondiseased observations from plasma lipid data are shown in Scott (1992, pp. 250–251). The parametric density estimate was obtained by fitting a single normal to each of two sets of observations, whereas the nonparametric estimate is an average shifted histogram. Scott considered only two possibilities: a completely parametric (multivariate normal) density and a fully nonparametric approach via kernel density estimation. MclustDA includes a single normal density as a special case and will reduce to that if the data do not warrant additional complexity. MclustDA can also be viewed as nonparametric, however, in the sense that it can approximate complex densities arbitrarily closely by adding components.

In a similar calculation, we applied MclustDA to the UCI Wisconsin Diagnostic Breast Cancer Data reduced to the two explanatory variables shown in the projections of Figure 1: extreme area and mean texture, treating the malignant and benign observations separately. A single ellipsoidal normal was obtained for the benign observations, and a mixture of two unconstrained normals was obtained for the malignant obser-

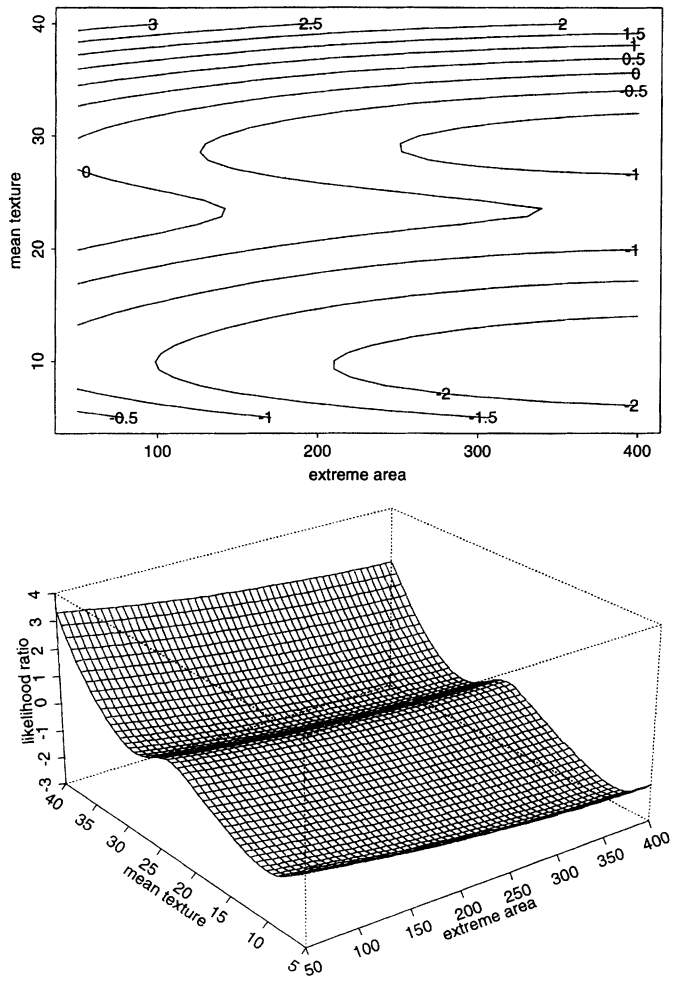

Figure 4. Contour and Perspective Plots of a Portion of the Log-Likelihood Ratio Surface for Two Covariates of the UCI Wisconsin Breast Cancer Data Obtained From Density Estimation via Model-Based Clustering.

vations. Contour and perspective plots of the resulting parametric likelihood ratio surface are shown in Figure 4. This ratio of density estimates captures the nonmonotonic nature of the likelihood ratio surface, while remaining satisfactorily smooth.

8.2 Minefield Detection

The Coastal Battlefield Reconnaissance and Analysis (COBRA) program (Witherspoon et al. 1995), developed by the U.S. Marine Corps, is intended to detect minefields in coastal areas via aerial reconnaissance. Figure 5 is a pairs plot of the measured intensity for all six bands of a COBRA reconnaissance image for each of 173 locations identified as possible mines on the basis of acquired images. Only 35 of the locations corresponded to actual mines; the other 138 were false positives. The goal here was to see whether model-based clustering could separate out the mines from the false positives based on the intensities, or at least identify a group containing the mines, so as to reduce the number of false positives. In this application, it is important to avoid false negatives (i.e., locations that are actually mines but are identified as non-mines). Because of the considerable linear dependence among

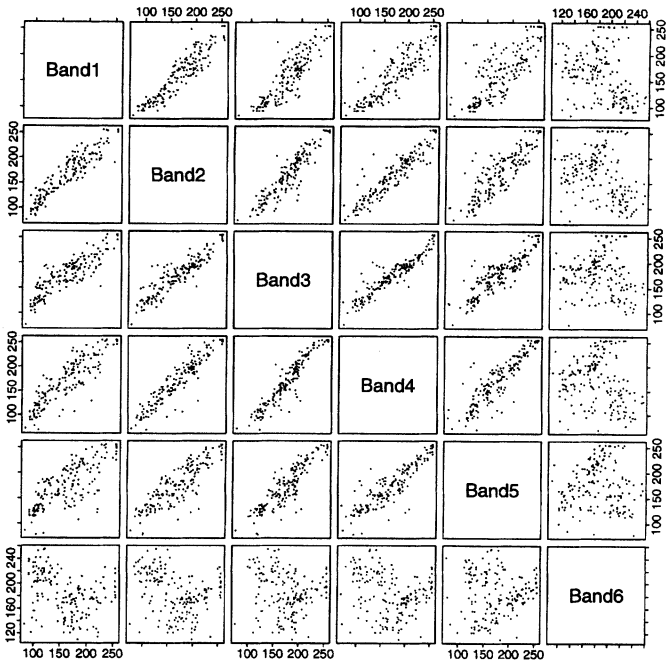

Figure 5. The Six Bands of a COBRA Reconnaissance Image.

the bands, we applied model-based clustering to the intensity measured in bands 1 and 6 only.

According to BIC, the best model is the four-group non-constant spherical model [see Figure 6]. In this grouping, all 35 mines are confined to one group containing a total of 89 points. By considering only the 89 points in that group as possible mines, the number of false positives is thus reduced by more than 60%, from 138 to 54, without introducing any false negatives.

8.3 Cluster Recovery From Noisy Data

We consider a problem in cluster recovery posed by Murtagh, Starck, and Berry (2000) that is based on the problem of locating galaxies in a noisy astronomical image.

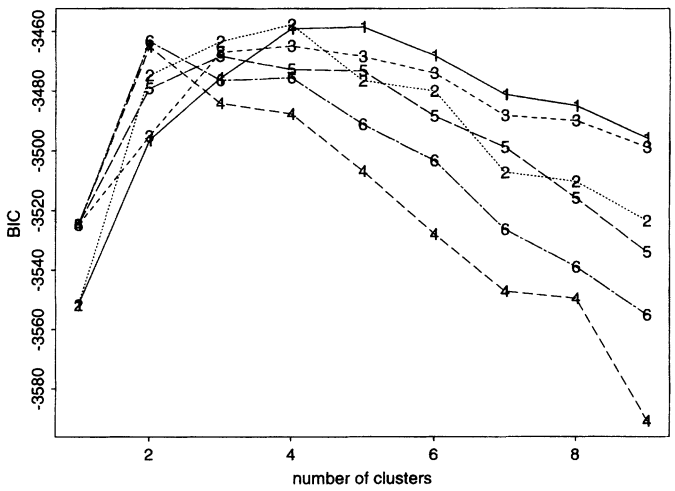

Figure 6. BIC for the COBRA Minefield Detection Problem, Using Bands 1 and 6. Models 1 and 2 are $\Sigma_k = \lambda I$ (constant spherical), and $\lambda_k I$ (nonconstant spherical), while models 3–6 are as given in Figure 3. The resulting classification reduced the number of false positives by more than 60% without introducing any false negatives.

The data consist of two simulated two-dimensional Gaussian clusters with centers (64, 64) and (190, 190) and with standard deviations in the x and y directions of (10, 20) and (18, 10). There are 300 data points in the first of these clusters and 250 in the second. Background noise is provided by adding 10,000 points from a Poisson distribution.

The results for this cluster recovery problem are shown in Figure 7. The model-based clustering strategy accurately determines the cluster means, although the clusters found are smaller than the true clusters (and they contain some noise points located within the cluster boundaries). A different threshold for determining the classification from the conditional probabilities could be used, as illustrated in Figure 7(d).

It should be noted that the method is sensitive to the value of V, the assumed volume of the data region, in (11). Here it is clear that V is the area of the image; Banfield and Raftery (1993) and Dasgupta and Raftery (1998) similarly used the volume of the smallest hyperrectangle with sides parallel to the axes that contains all of the data points. Other possibilities include taking V to be the smallest hyperrectangle with sides parallel to the principal components of the data that contains all the data points, or using the volume of the convex hull of the data (e.g., Bentley, Clarkson, and Levine 1993).

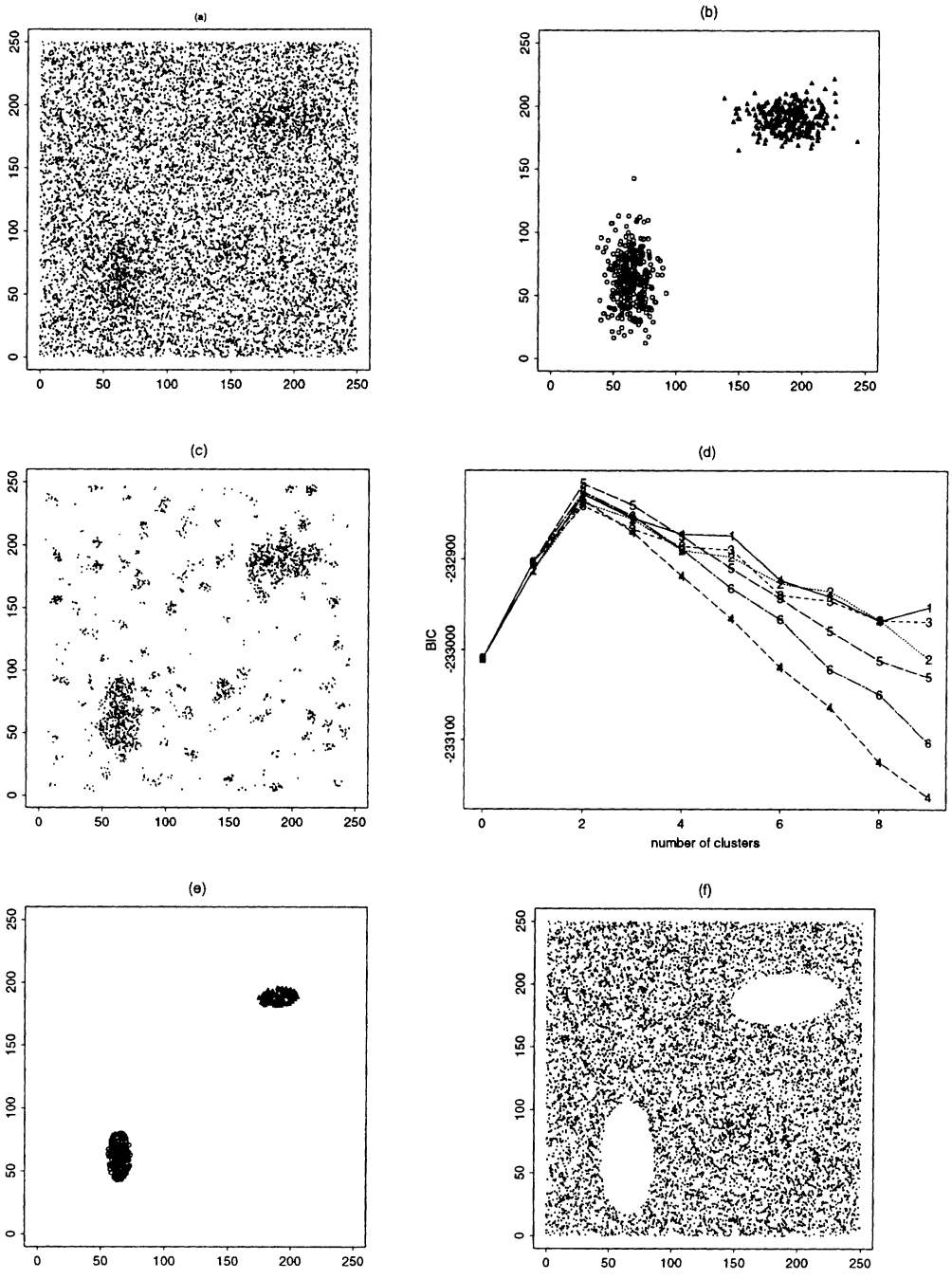

Figure 7. (a) An instance of the cluster recovery data, consisting of two Gaussian clusters with a total of 550 points, and 10,000 noise points. (b) The Gaussian clusters. (c) The data after 20 nearest-neighbor denoising with NNclean. (d) BIC from model-based clustering. In model 5, groups have equal shape and volumes. (e) Model-based classification. (f) Points with classification uncertainty less than .1.

8.4 Spatial Density Estimation

As an illustration of density estimation with multivariate mixtures (Sec. 7), we consider the density of the Lansing Woods maples (Gerrard 1969). Figure 8 shows the location of the maples, the model-based classification, the corresponding density, and a standard Gaussian kernel density estimate. The BIC [Fig. 8(a)] indicates that a varying-volume spherical model with six groups is the best model among those available. The Gaussian kernel density estimate [Fig. 8(d)] was computed with the S+SpatialStats software (Kaluzny, Vega, Cardoso, and Shelly 1998), using a bandwidth estimated by cross-validation using the sm software of Bowman and Azzalini (1997). Some advantages of the model-based approach are that there are no bandwidth parameters involved, and that it is easy to compute the density at points other than the data points.

8.5 Simulation Study for Two-Dimensional Density Estimation

In this section we give the results of simulations using two-dimensional analogs of the univariate normal densities from Marron and Wand (1992) that were studied by Roeder and Wasserman (1997). Figure 9 shows contour plots of the 10 densities used in the simultations.

Table 1 compares the average integrated mean squared error (MISE) for density estimation via model-based clustering, with those for Gaussian kernel density estimation using both the normal optimal bandwidth and cross-validated bandwidth, over 50 simulations for each of the 10 models (250 data points). The results for Gaussian kernel density estimation were obtained using the sm software of Bowman and Azzalini (1997). The numbers shown are the MISE for kernel density estimation divided by the MISE for model-based clustering for each of the two kernel methods. Only in one of the ten simulated situations does kernel estimation outperform model-based clustering: the Claw (Bart Simpson) density, the most complicated of the ten densities studied.

Figure 10 shows the density used to generate the data, as well as each of the estimated densities from model-based clustering, Gaussian kernel with optimal normal and cross-validated bandwidths for one dataset simulated from the trimodal density.

8.6 Gene Expression Microarray Data

New techniques in biotechnology, such as cDNA microarrays and high-density oligonucleotide chips, allow simultaneous monitoring of the expression of thousands of genes

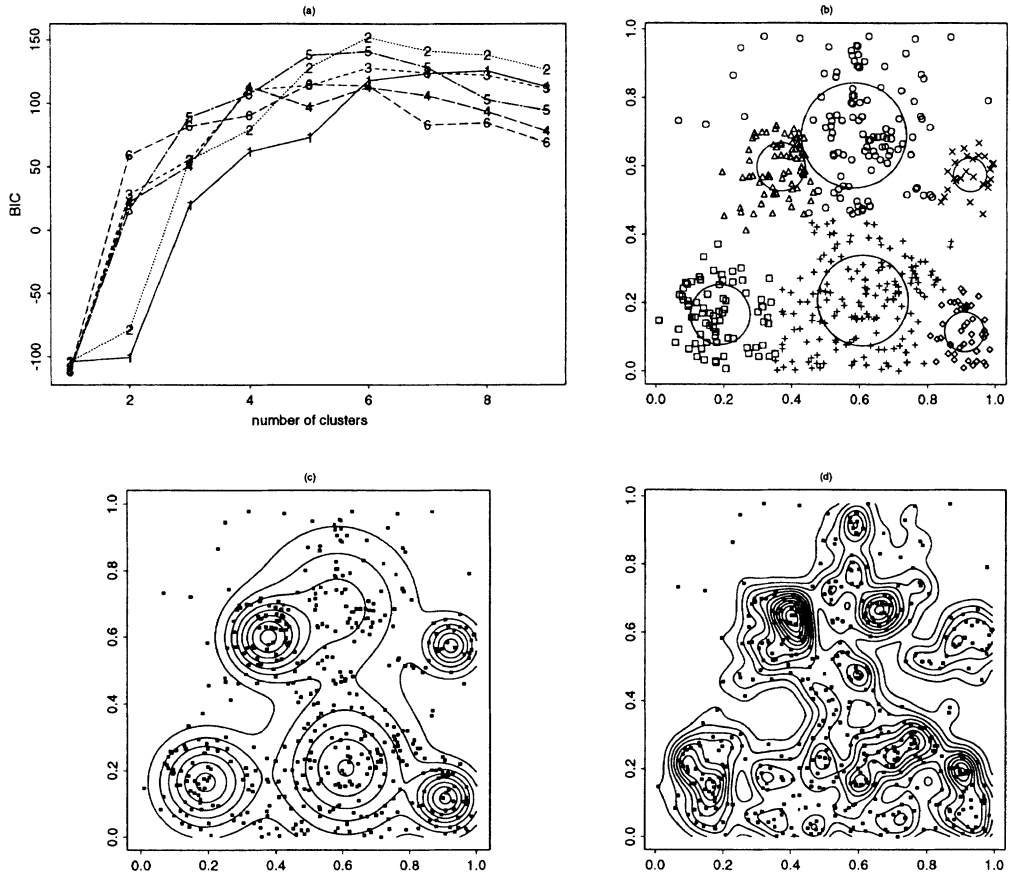

Figure 8. Density Estimation for the Lansing Woods Maples. (a) BIC from model-based clustering. The maximum BIC model is a six-component varying-volume spherical mixture. (b) Model-based classification, with circles indicating the circles defined by the estimated covariance of each of the six groups. (c) Contours of the density as determined by model-based clustering, with the location of the maples superimposed. (d) Contours of a standard Gaussian kernel density estimate with bandwidth selected by cross-validation.

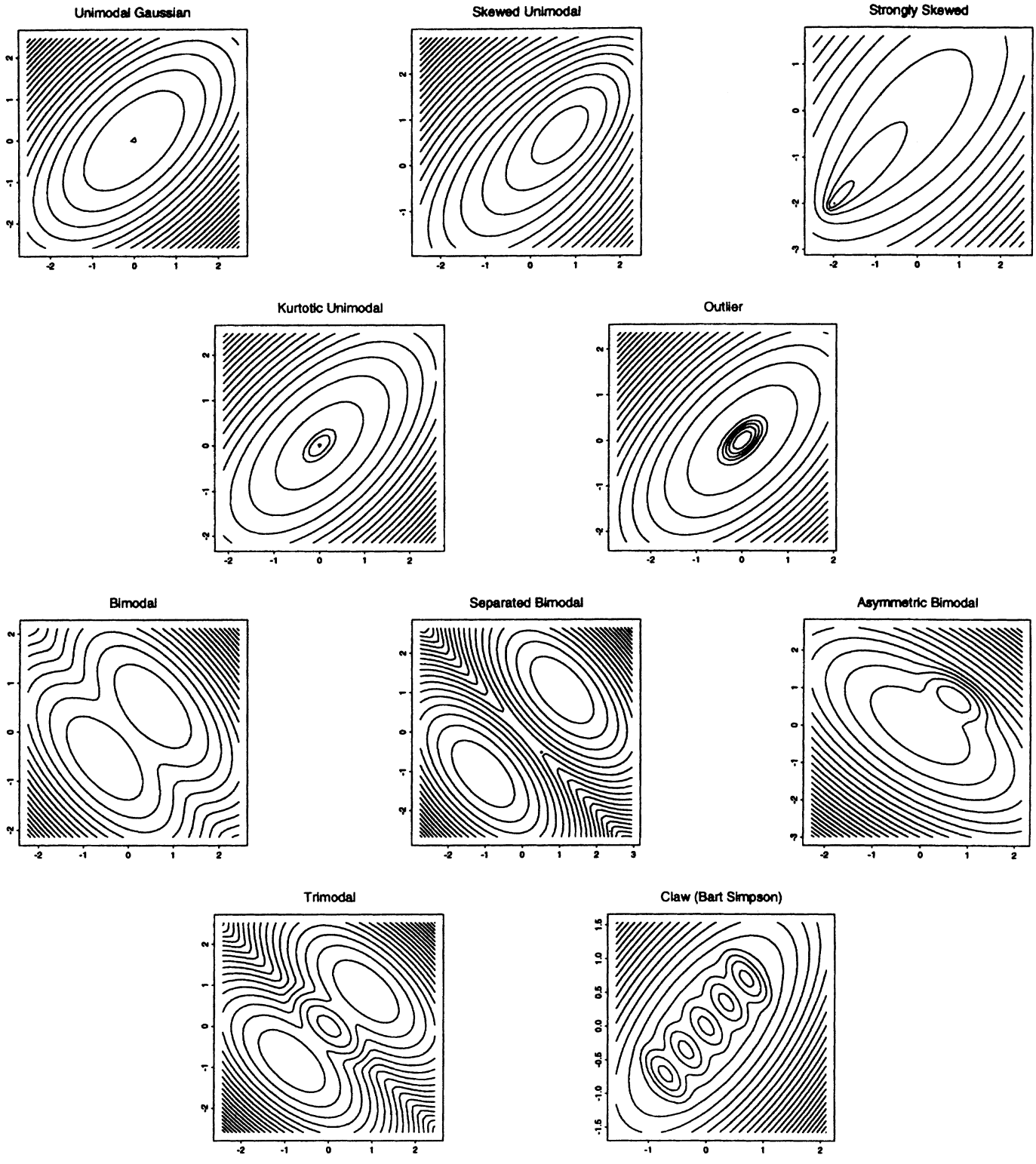

Figure 9. Contours of the 10 Two-Dimensional Simulation Densities.

under any number of desired conditions. Experiments for which these data are collected include time series, for example, phases of a cell cycle (Cho et al. 1998; Chu et al. 1998), group comparisons, such as benign versus malignant tumors (Perou et al. 1999; Ross et al. 2000), and population descriptions, such as response to drug treatments (Weinstein et al. 1997; Scherf et al. 2000). Ultimately the data are represented in the form of a matrix of measurements, with one dimension (usually the rows) for the genes or clones (pieces of the gene) and the other dimension for the experiments or samples. The amount and complexity of the data generated by these new technologies requires specialized bioinformatic and statistical tools to extract useful information. Clustering and classification are particularly important in this context because

Table 1. MISE Ratios for Density Estimation via Model-Based Clustering (MBC), Gaussian Kernel Density Estimation With Normal Optimal Bandwidth (NOB), and Gaussian Kernel Density Estimation With Cross-Validated Bandwidth (CVB)

Model	NOB/MBC	CVB/MBC
Unimodal Gaussian	4.5	4.4
Skewed unimodal	2.2	1.9
Strongly skewed	6.7	1.5
Kurtotic unimodal	12.5	3.5
Outlier	14.9	4.6
Bimodal	4.2	3.6
Separated bimodal	11.5	4.2
Asymmetric bimodal	4.1	2.6
Trimodal	4.1	2.1
Claw (Bart Simpson)	1.8	.7
Average	6.6	2.9

NOTE: The numbers shown are the ratios of the MISEs for NOB and CVB to that for MBC. For the strongly skewed model, the CVB result is averaged over 42 of the 50 replicates, because in the remaining 8 instances, the cross-validated bandwidth could not be computed with default parameters.

of the desire to identify genes whose activities are related in circumstances of interest, as well as the desire to group samples or experimental conditions on the basis of observed gene expression.

Clustering Gene Expression Data. A wide range of clustering methods have been applied to gene expression data, including hierarchical clustering (e.g., Weinstein et al. 1997; Eisen, Spellman, Brown, and Botstein 1998), self-organizing maps (e.g., Golub et al. 1999; Tamayo et al. 1999), graph-theoretic methods (e.g., Ben-Dor, Shamir, and Yakhini 1999), techniques related to principal components (e.g., Alter, Brown, and Botstein 2000; Hastie et al. 2000b; Yeung and Ruzzo 2001), k-means (e.g., Herwig et al. 1999; Tavazoie, Hughes, Campbell, Cho, and Church 1999), network models (e.g., Michaels et al. 1998), resampling approaches (e.g., Holmes and Bruno 2000; van der Laan and Bryan 2000), deterministic annealing (Alon et al. 1999), support vector machines (Brown et al. 2000), and fuzzy logic (Woolf and Wang 2000).

In many studies, both genes and samples are clustered and the results displayed simultaneously in colored block diagrams (e.g., Perou et al. 1999; Ross et al. 2000). There are also explicit methods for two-way or block clustering of gene expression data (e.g., Lazzeroni and Owen 2000; Tibshirani, Hastie, Ross, Botstein, and Brown 1999). Although gene expression clustering software that has been made available tends to be applied in other studies (e.g., Chu et al. 1998; Iyer et al. 1999; Alizadeh et al. 2000; Ross et al. 2000; Scherf et al. 2000 all use the method of Eisen et al. 1998), no single method has emerged as a method of choice even when restricted to a certain type of gene expression data, and new approaches are continuing to be proposed at a considerable pace.

Determining the Number of Clusters. Many techniques for cluster analysis of gene expression data rely on graphical display and visual inspection to determine the number of clusters (e.g., Eisen et al. 1998; Wen et al. 1998; Tamayo et al. 1999). Some studies develop statistics from which to compare the clustering obtained when different numbers of clusters are assumed. For example, Tavazoie et al. (1999) chose the number of clusters by quantifying a compromise between overclassification to avoid missing classes and good separation between the classes. Golub et al. (1999) used a *neighborhood analysis* technique for deciding whether or not a particular gene should belong to a tentative class. In the CAST algorithm of Ben-Dor et al. (1999), the number of clusters is determined automatically, but it depends on a user-defined cluster-affinity parameter whose relationship to the number of clusters is monotonic but not transparent.

Tibshirani, Walther, and Hastie (2000) proposed a gap statistic as a general method for determining the number of clusters. In *tree harvesting* for hierarchical clustering (Hastie, Tibshirani, Botstein, and Brown 2000a), the model size and hence the number of clusters is chosen by k-fold cross-validation. In plaid models (Lazzeroni and Owen 2000), layers are added only if they are determined to make a significant contribution to the model in terms of minimizing a sum of squared errors; the number of clusters is effectively determined by the number of layers added to the model. Van der Laan and Bryan (2000) suggested determining the number of clusters using the average silhouette width measure in the *partitioning about medoids* method of Kaufman and Rouseeuw (1990).

Model-Based Clustering. Recently, Yeung, Fraley, Murua, Raftery, and Ruzzo (2001) applied the model-based method of Section 5.2 to the gene clustering problem in both real and simulated gene expression datasets for which the clone groupings were known in advance. They found that unspecialized model-based clustering methods showed performance comparable to that of a leading heuristic alternative specifically designed for gene expression data, CAST (Ben-Dor et al. 1999). Moreover, model-based clustering provided a way of selecting the clustering model and the number of clusters.

To apply model-based clustering to the experiment clustering problem, the set of covariance models would have to be restricted, because the number of cases is typically smaller than the number of variables. Extension to a more general context would be possible, however, via Bayesian estimation with a proper prior (see Sec. 10.4).

Discriminant Analysis. There is also considerable interest in discriminant analysis, or supervised classification, of gene expression data (e.g., Chu et al. 1998; Wasserman and Fickett 1998; Golub et al. 1999; Brown et al. 2000; Ben-Dor et al. 2000; Hastie et al. 2000a, 2000b; West et al. 2000). Dudoit, Fridlyand, and Speed (2000a) applied various approaches to cancer classification via gene expression data analysis in a comparative study. The methods compared included nearest-neighbor classifiers, LDA, and classification trees, as well as machine learning approaches such as bagging and boosting. Discriminant analysis techniques related to model-based clustering could also be applied here.

9. MODEL–BASED CLUSTERING SOFTWARE

The MCLUST software (Fraley and Raftery 1999), implementing model-based clustering and discriminant analysis as described this article, is available at *http://www.stat.washington.edu/mclust*. It is designed to interface with the commercial interactive software package S-PLUS.

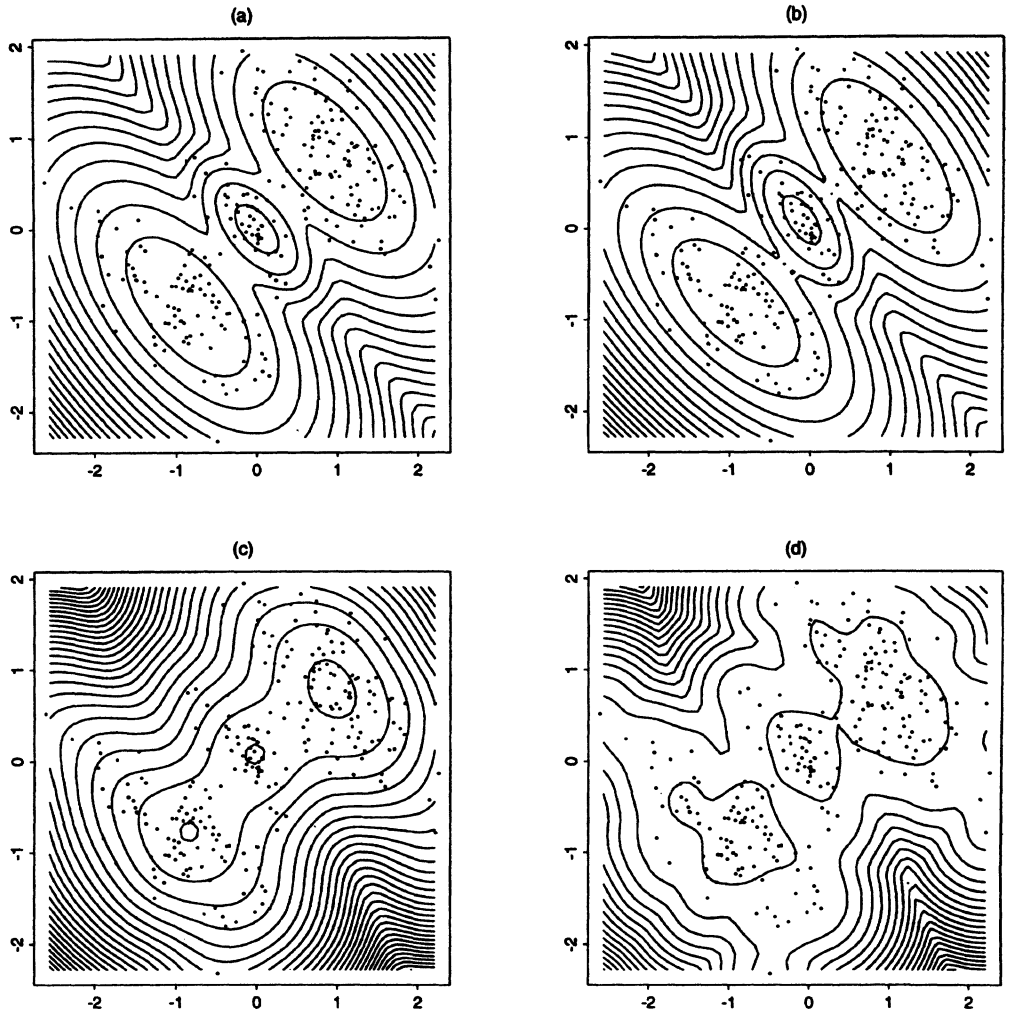

Figure 10. Comparative Density Estimates for Trimodal Multivariate Normal Data. (a) Density used to generate the data; (b) density estimate via model-based clustering; (c) Gaussian kernel density estimate with normal optimal bandwidth; (d) Gaussian kernel density estimate with cross-validated bandwidth. There are 250 data points, superimposed on the contours to show their location.

Other software packages for model-based clustering include EMMIX (McLachlan, Peel, Basford, and Adams 1999) and AutoClass (Cheesman and Stutz 1995). Software for MDA and some of its generalizations is also available (see Hastie and Tibshirani 1996). An S-PLUS function implementing the nearest-neighbor denoising method (Byers and Raftery 1998) used in the example of Section 8.3 is available through StatLib at *http://lib.stat.cmu.edu/S/nnclean.*

10. LIMITATIONS AND EXTENSIONS

The clustering methods based on multivariate normal mixture models that we have described in this article have been used with success in such applications as detection of minefields and seismic faults (Dasgupta and Raftery 1998), identification of flaws in textiles from images (Campbell et al. 1997, 1999), and classification of astronomical data (Mukherjee et al. 1998). However, their practical use without modification can be limited for non-Gaussian, high-dimensional, or large datasets.

10.1 Non-Gaussian Data

Multivariate normal mixtures can accommodate data of varying structures. The component distributions are concentrated around surfaces of lower dimension; for example, a highly linear distribution is concentrated around a line, which is the first principal component. Sometimes clusters are concentrated around lower-dimensional manifolds that are not linear. A non-Gaussian component can often be approximated by several Gaussian ones (e.g., Dasgupta and Raftery 1998; Fraley and Raftery 1998). For example, if one component is concentrated about a nonlinear curve, then it may be possible to provide a piecewise linear approximation, which could be represented by several Gaussian clusters, each one concentrated about a linear subspace. In the COBRA minefield example (Sec. 8.2), observations identified as not being mines were located in several groups in the model-based classification, whereas the true mines were confined to a single mixture component. An explicit approach to the problem of clusters concentrated around nonlinear curves rather than lines is to model the curves nonparametrically but smoothly using the concept of *principal curves* (Hastie and Stuetzle 1989). This

idea of clustering about principal curves was proposed and developed by Banfield and Raftery (1992) and Stanford and Raftery (2000).

The model-based framework is flexible and need not be restricted to multivariate normal mixtures. In the example of cluster recovery from noisy data (Sec. 8.3), the cluster structure was recaptured by preprocessing the data to remove some of the noise in the hierarchical clustering phase and adding a Poisson term to the mixture to model the noise in the EM phase. Other mixture models that have been applied in clustering and related contexts include mixtures of t distributions (Peel and McLachlan 2000), mixtures of trees (Meila 1999), mixtures of first-order Markov chains (Cadez, Heckerman, Meek, Smyth, and White 2000), and mixtures of distributions for angular data (Peel, Whitten, and McLachlan 2001).

Mixture models for multivariate discrete data, often called latent class models, have been developed over a long period (Lazarsfeld 1950; Lazarsfeld and Henry 1968; Clogg and Goodman 1984; Becker and Yang 1998), and could be used for clustering within the framework described here. More recently, Chickering and Heckerman (1997) pointed out that a finite mixture model is a graphical Markov model with a single hidden node. This has opened up the possibility of applying the technology of graphical models and Bayes nets to the clustering problem, particularly for high-dimensional discrete data of the kind generated by, for example, tracking visits to websites. Handling data in which attributes or dimensions are of different kinds (e.g., discrete, ordinal, continuous and censored) is currently a major challenge for model-based clustering.

10.2 High-Dimensional Data

A limitation of model-based clustering with high-dimensional data is that the number of parameters per component in multivariate normal mixtures that allow orientation to vary between clusters grows as the square of the dimension of the data. Moreover, if the dimension of the data is high relative to the number of observations, then the covariance estimates in the ellipsoidal models will often be singular, causing the EM algorithm to break down, although the more parsimonious models such as the spherical and diagonal models may still be applicable.

When the data are of high dimension, some sort of dimension reduction strategy is inevitable. Sometimes correlations or other relationships among variables are evident, so that selecting a subset of the variables with which to work is relatively easy, as in, for example, the COBRA minefield detection problem of Section 8.2 or the gamma ray bursts analyzed by Mukherjee et al. (1998). Principal components are often used for dimension reduction (e.g., Smyth 2000), but in some instances transforming the data into principal components may obscure rather than reveal groupings of interest (Chang 1983). Recent research has found that the wavelet transform is effective for dimension reduction in some clustering applications (Murtagh et al. 2000).

Another approach to high-dimensional data is to replace the data by distances or dissimilarities between data points. This is prevalent in applications such as document clustering or information retrieval, where each dimension corresponds to a word or term that may or may not appear in the document. Clustering methods that are not model-based have been developed for this situation, and many hierarchical agglomerative methods can be adapted to this problem. Model-based clustering can also be combined with multidimensional scaling (e.g., DeSarbo, Howard, and Jedidi 1991). A satisfactory solution remains a major research challenge, although new model-based multidimensional scaling techniques (e.g., Oh and Raftery 2001) may help bring the benefits of model-based clustering to this setting.

10.3 Large Datasets

One reason for the current explosion of interest in clustering is the desire to use it for finding patterns in very large datasets, sometimes called "datamining." Model-based clustering as described in this article does not scale to large datasets without modification. A major limiting factor is that time-efficient methods for model-based hierarchical agglomeration have initial memory requirements proportional to the square of the number of groups in the initial partition, which by default assigns each observation to a group with a single element. Although in the default procedure adequate memory may not be available for processing large datasets, memory requirements can be reduced if some of the observations can be grouped together in advance. Posse (2001) proposed using the minimum spanning tree to obtain initial partitions for hierarchical agglomeration for large datasets.

When the sample size is moderately large, a general and simple approach is to take a random sample of the data and then apply model-based clustering to the sample. The results are then extended to the full dataset using discriminant analysis, with the sample viewed as the training set, essentially basing inference on the sample rather than on the full population. Banfield and Raftery (1993) applied this idea in segmenting an magnetic resonance image, which they cast as a problem of clustering the 26,000 or so nonbackground pixels in the image. They took an initial sample of only 500 pixels, clustered them, and then classified the remaining 25,500 pixels on the basis of the results. With the methodology described here, the discriminant analysis is straightforward; a final E step is applied to the remaining data to obtain conditional probabilities, using the parameter estimates derived from the sample.

The simple sampling strategy just described may break down when seeking small groups in very large datasets. Small groups may not be represented at all in a sample, or they may have too few representatives to be distinguished as a cluster. Fayyad and Smyth (1996) considered one such instance, finding a group of about 40 quasars in a catalog of about 2 billion objects, which they solved by *iterated sampling* (see Sec. 5.3). The problem could also be approached a modification of the simple sampling method. One version of this is as follows:

1. In the final E step from the simple sampling method, compute $f_i = \max_k f_k(\mathbf{y}_i|\hat{\theta}_k)$ for each observation i in the full dataset.
2. Select out the observations i such that f_i is below some threshold, that is, those that are not well represented by any of the clusters identified so far.

3. Form a second sample, including all of the poorly represented data points identified, together with a stratified sample from the clusters that have been identified (e.g., roughly equal numbers from each cluster).
4. Apply model-based clustering to the new sample, and apply the E step to the full sample as before. A final application of the M step to the full sample might also be needed, especially to estimate the proportions τ_k. These steps could be iterated until a stable solution is found.

So far we have discussed difficulties with moderately large datasets—large enough that a set of interpoint distances cannot be held in memory, although the data can. Datamining is often concerned with even larger datasets. The computation time for an EM iteration, which depends only on the data dimension when the all of the data can be easily held in memory, increases greatly when this is not the case. In this context, considerable work has been done on computational techniques for making the EM algorithm more efficient when applied to large datasets (Bradley, Fayad, and Reina 1998; Moore and Lee 1998; Moore 1999; Thiesson, Meek, and Heckerman 1999). One focus is the development of "one-pass" methods, in which each part of the data needs to be loaded into memory only once. However, even with memory resources and processor speeds large enough for handling massive datasets as a whole, numerical error due to finite precision arithmetic would remain an obstacle. This limitation favors the traditional approach that we have mentioned, clustering a subset of the data for use as a training set, and then applying a discriminant rule for classification.

A number of assumptions in the mixture modeling approach may be at odds with the realities of massive data entities, so straightforward application of the simple or iterated sampling approach may not work well. First, it is assumed that the data come from a mixture model and are present in the data collection in the appropriate proportions. Second, it is assumed that somehow a training set can be selected from the data in the correct proportions, which may be unrealistic for large out-of-core databases that cannot be sampled randomly. Despite these apparent obstacles, model-based clustering seems to be emerging as an important component within schemes for the classification of large datasets (Meila 1999; Smyth 2000; Cadez et al. 2000; Posse 2001).

10.4 Bayesian Estimation

In this review we have focused on frequentist estimation, mostly via maximum likelihood, for the mixture models underlying model-based clustering. We have found approximate Bayesian methods more useful for model selection, however. Some statisticians also may wish to use Bayesian methods for estimation, for reasons of statistical principle, or because informative prior information is available.

For other statisticians, we can think of three reasons why they might be interested in adopting a Bayesian approach to estimation. The first, and probably most important from a practical viewpoint, is that the EM algorithm for maximizing the likelihood can converge to degenerate solutions with infinite likelihood, corresponding to small and/or highly linear clusters. This also makes it difficult to identify small clusters, especially with the more complex models. A Bayesian approach can alleviate this problem by effectively smoothing the likelihood so that its many uninteresting infinite spikes are removed.

The second reason has to do with interval estimation. There are many ways of calculating approximate standard errors from the EM algorithm (e.g., McLachlan and Krishnan 1997, chap. 4), and they can be combined with an assumption of approximate normality to obtain approximate confidence intervals. However, one may want more precise estimation intervals, and these can be obtained from a Bayesian approach.

The third reason has to do with the assessment of uncertainty in the posterior probabilities of belonging to groups. From the EM algorithm, it is easy to calculate approximate posterior probabilities conditional on the MLEs of the model parameters, and the error in doing this typically declines to zero quickly, at rate $O(n^{-1/2})$. But because this ignores the uncertainty in the parameter estimates, it is likely to underestimate the overall uncertainty and so to bias estimated posterior probabilities toward greater certainty (i.e., toward 0 or 1), albeit to an extent that declines to zero as sample size increases.

The simplest Bayesian estimation approach is to use the EM algorithm to find the posterior mode rather than the MLE, as suggested by Dempster et al. (1977). This is likely to go a long way toward alleviating the first and most important of the three problems mentioned, although it will not solve the other two.

The problem of specifying the prior remains. If informative prior information is available, then this should be used. If not, then it would be desirable to have an easy way of specifying a prior. Standard reference priors do not seem to be directly applicable to the models considered here. A unit information prior, either in the form proposed for testing by Kass and Wasserman (1995), in the slightly different form given by Raftery (1995), or in a diagonal form with the off-diagonal elements set to zero, also may be useful for estimation as a kind of reference prior. Raftery (1999) argued that such priors can provide a reasonable approximation to the elicited prior of someone who knows something, but not much, about the problem at hand. They also have the desirable property of being fairly flat over the part of parameter space where the likelihood is substantial, without being much greater elsewhere. These priors are proper, albeit mildly data dependent, and have the desirable smoothing properties mentioned earlier.

Recently, much work has been done on Bayesian estimation of mixture models using MCMC. The basic idea is to compute the joint posterior distribution of the model parameters and the "missing data," $\mathbf{z}$, defined in the same way as in the EM algorithm. This is typically done by Gibbs sampling or random walk Metropolis–Hastings, updating the components of the posterior distribution one at a time. Lavine and West (1992) were the first to do this, using Gibbs sampling and applying the results to clustering in the context of a mixture of multivariate normal distributions. They considered only the model with unconstrained covariance matrices. Working independently, Diebolt and Robert (1994) applied Gibbs sampling to Bayesian estimation of a one-dimensional normal mixture model. Bensmail et al. (1997) extended these results to the full range of clustering models considered here, and showed

how the Bayesian method can be effective when there are very small clusters, which would stump the frequentist approach.

Reversible jump MCMC (Green 1995) was an important development and was applied to one-dimensional normal mixtures by Richardson and Green (1997). This allows the MCMC sampler to move between different models as well as between different parameter values, and hence to yield estimates of Bayes factors and posterior model probabilities directly. Implementing this method seems somewhat challenging, however, and so far applying it to multivariate mixtures, such as those that arise in clustering, has proven difficult. Castelloe (1999) has succeeded in applying this approach to a two-dimensional model-based clustering problem with particular constraints.

A major difficulty with Bayesian estimation of mixtures in general, and MCMC implementations of it in particular, is the label-switching problem discussed by, for example, Richardson and Green (1997). This arises because one can switch the labeling of the mixture components without changing the likelihood. Because there are $G!$ labelings, it follows that there are $G!$ components of the posterior distribution, which are identical except for the labeling if the prior is symmetric with respect to labelings. This has various perverse consequences; for example, the posterior means of the means of the mixture components will all be the same.

Various solutions to the label-switching problem have been proposed. Early proposals involved ordering the components a priori in some way (e.g., Celeux, Chaveau, and Diebolt 1996; Mengersen and Robert 1996; Richardson and Green 1997), but this does not solve the problem in general. Recent proposals to postprocess the MCMC output (Celeux 1998; Celeux, Hurn, and Robert 1999; Stephens 1997, 2000) seem much more promising. These consist basically of clustering the MCMC output itself according to the apparent labeling in operation, then relabeling the sampled parameters so that they all correspond to the same labeling. Proposed methods for doing this include a k-means clustering algorithm and a transportation algorithm for optimization. One could imagine that applying model-based clustering itself to this "meta-problem" might be useful.

[Received October 2000. Revised October 2001.]

REFERENCES

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J. J., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O., and Staudt, L. M. (2000), "Distinct Types Diffuse Large B-Cell Lymphoma Identified by Gene Expression Profiling," *Nature*, 403, 503–511.

Allard, D., and Fraley, C. (1997), "Nonparametric Maximum Likelihood Estimation of Features in Spatial Point Processes Using Voronoï Tessellation," *Journal of the American Statistical Association*, 92, 1485–1493.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. (1999), "Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays," *Cell Biology*, 99, 6745–6750.

Alter, O., Brown, P. O., and Botstein, D. (2000), "Singular Value Decomposition for Genome-Wide Expression Data and Modeling," *Proceedings of the National Academy of Sciences*, 97, 10101–10106.

Banfield, J. D., and Raftery, A. E. (1992), "Ice Floe Identification in Satellite Images Using Mathematical Morphology and Clustering About Principle Curves," *Journal of the American Statistical Association*, 87, 7–16.

——— (1993), "Model-Based Gaussian and Non-Gaussian Clustering," *Biometrics*, 49, 803–821.

Becker, M. P., and Yang, I. (1998), "Latent Class Marginal Models for Cross-Classifications of Counts," *Sociological Methodology*, 28, 293–326.

Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M., and Yakhini, Z. (2000), "Tissue Classification With Gene Expression Profiles," in *RECOMB 2000: Proceedings of the 4th Annual International Conference on Computational Molecular Biology*, pp. 54–64.

Ben-Dor, A., Shamir, R., and Yakhini, Z. (1999), "Clustering Gene Expression Patterns," *Journal of Computational Biology*, 6, 281–297.

Bensmail, H., and Celeux, G. (1996), "Regularized Gaussian Discriminant Analysis Through Eigenvalue Decomposition," *Journal of the American Statistical Association*, 91, 1743–1748.

Bensmail, H., Celeux, G., Raftery, A. E., and Robert, C. P. (1997), "Inference in Model-Based Cluster Analysis," *Statistics and Computing*, 7, 1–10.

Bentley, J. L., Clarkson, K. L., and Levine, D. B. (1993), "Fast Linear Expected-Time Algorithms for Computing Maxima and Convex Hulls," *Algorithmica*, 9, 168–183.

Biernacki, C., Celeux, G., and Govaert, G. (1999), "An Improvement of the NEC Criterion for Assessing the Number of Clusters in a Mixture Model," *Pattern Recognition Letters*, 20, 267–272.

——— (2000), "Assessing a Mixture Model for Clustering With the Integrated Completed Likelihood," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22, 719–725.

Biernacki, C., and Govaert, G. (1999), "Choosing Models in Model-Based Clustering and Discriminant Analysis," *Journal of Statistical Computation and Simulation*, 64, 49–71.

Binder, D. A. (1978), "Bayesian Cluster Analysis," *Biometrika*, 65, 31–38.

Bock, H. H. (1996), "Probabilistic Models in Cluster Analysis," *Computational Statistics and Data Analysis*, 23, 5–28.

——— (1998a), "Probabilistic Approaches in Cluster Analysis," *Bulletin of the International Statistical Institute*, 57, 603–606.

——— (1998b), "Probabilistic Aspects in Classification," in *Data Science, Classification and Related Methods*, eds. C. Hayashi, K. Yajima, H. H. Bock, N. Oshumi, Y. Tanaka, and Y. Baba, New York: Springer-Verlag, pp. 3–21.

Bollen, K. A. (1989), *Structural Equations With Latent Variables*, New York: Wiley.

Bowman, A. W., and Azzalini, A. (1997), *Applied Smoothing Techniques for Data Analysis*, Oxford, U.K.: Clarendon Press.

Box, G. E. P., and G. M. Jenkins (1976). *Time-series Analysis: Forcasting and Control*, Holden-Day.

Boyles, R. A. (1983), "On the Convergence of the EM Algorithm," *Journal of the Royal Statistical Society*, Ser. B, 45, 47–50.

Bozdogan, H. (1994), "Choosing the Number of Clusters, Subset Selection of Variables, and Outlier Detection on the Standard Mixture-Model Cluster Analysis," in *New Approaches in Classification and Data Analysis*, eds. E. Diday, Y. Lechevallier, M. Schader, P. Bertrand, and B. Burtschy, New York: Springer-Verlag, pp. 169–177.

Bradley, P. S., Fayyad, U., and Reina, C. (1998), "Scaling EM (Expectation-Maximization) Clustering to Large Databases," Technical Report MSR-TR-98-35, Microsoft Research.

Brown, M. P. S., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, J. M., and Haussler, D. (2000), "Knowledge-Based Analysis of Microarray Gene Expression Data by Using Support Vector Machines," *Proceedings of the National Academy of Sciences*, 97, 262–267.

Byers, S. D., and Raftery, A. E. (1998), "Nearest-Neighbor Clutter Removal for Estimating Features in Spatial Point Processes," *Journal of the American Statistical Association*, 93, 577–584.

Cadez, I., Heckerman, D., Meek, C., Smyth, P., and White, S. (2000), "Visualization of Navigation Patterns on a Web Site Using Model-Based Clustering," Technical Report MSR-TR-2000-18, Microsoft Research.

Campbell, J. G., Fraley, C., Murtagh, F., and Raftery, A. E. (1997), "Linear Flaw Detection in Woven Textiles Using Model-Based Clustering," *Pattern Recognition Letters*, 18, 1539–1548.

Campbell, J. G., Fraley, C., Stanford, D., Murtagh, F., and Raftery, A. E. (1999), "Model-Based Methods for Real-Time Textile Fault Detection," *International Journal of Imaging Systems and Technology*, 10, 339–346.

Castelloe, J. (1999), "Reversible Jump Markov Chain Monte Carlo Analysis of Spatial Point Poisson Cluster Processes With Bivariate Normal Displacement," in *Computing Science and Statistics: Proceedings of the 31st Symposium on the Interface*, pp. 306–315.

Celeux, G. (1998), "Bayesian Inference for Mixtures: The Label-Switching Problem," in *COMPSTAT 1998*, eds. R. Payne and P. Green, Heidelberg and Vienna: Physica-Verlag, pp. 227–232.

Celeux, G., Chaveau, D., and Diebolt, J. (1996), "Stochastic Versions of the EM Algorithm: An Experimental Study in the Mixture Case," *Journal of Statistical Computation and Simulation*, 55, 287–314.

Celeux, G., and Govaert, G. (1992), "A Classification EM Algorithm for Clustering and Two Stochastic Versions," *Computational Statistics and Data Analysis*, 14, 315–332.

——— (1993), "Comparison of the Mixture and the Classification Maximum Likelihood in Cluster Analysis," *Journal of Statistical Computation and Simulation*, 47, 127–146.

——— (1995), "Gaussian Parsimonious Clustering Models," *Pattern Recognition*, 28, 781–793.

Celeux, G., Hurn, M., and Robert, C. (2000), "Computational and Inferential Difficulties With Mixture Posterior Distributions," *Journal of the American Statistical Association*, 95, 957–970.

Celeux, G., and Soromenho, G. (1996), "An Entropy Criterion for Assessing the Number of Clusters in a Mixture," *Journal of Classification*, 13, 195–212.

Chang, W. C. (1983), "On Using Principal Components Before Separating a Mixture of Two Multivariate Normal Distributions," *Applied Statistics*, 32, 267–275.

Cheeseman, P., and Stutz, J. (1995), "Bayesian Classification (AutoClass): Theory and Results," in *Advances in Knowledge Discovery and Data Mining*, eds. U. Fayyad, G. Piatesky-Shapiro, P. Smyth, and R. Uthurusamy, AAAI Press, pp. 153–180.

Chickering, D. M., and Heckerman, D. (1997), "Efficient Approximations for the Marginal Likelihood of Bayesian Networks With Hidden Variables," *Machine Learning*, 29, 181–212.

Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J., and Davis, R. W. (1998), "A Genone-Wide Transcriptional Analysis of the Mitotic Cell Cycle," *Molecular Cell*, 2, 65–73.

Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. O., and Herskowitz, I. (1998), "The Transcriptional Program of Sporulation in Budding Yeast," *Science*, 282, 699–705.

Clogg, C. C., and Goodman, L. A. (1984), "Latent Structure Analysis of a Set of Multidimensional Contingency Tables," *Journal of the American Statistical Association*, 79, 762–771.

Dasgupta, A., and Raftery, A. E. (1998), "Detecting Features in Spatial Point Processes With Clutter via Model-Based Clustering," *Journal of the American Statistical Association*, 93, 294–302.

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

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood for Incomplete Data via the EM Algorithm (with discussion)," *Journal of the Royal Statistical Society*, Ser. B, 39, 1–38.

DeSarbo, W. S., Howard, D. J., and Jedidi, K. (1991), "MULTICLUS: A New Method for Simultaneously Performing Multidimensional Scaling and Cluster Analysis," *Psychometrika*, 56, 121–136.

Diebolt, J., and Robert, C. (1994), "Estimation of Finite Mixtures Through Bayesian Sampling," *Journal of the Royal Statistical Society*, Ser. B, 56, 363–375.

Duda, R. O., and Hart, P. E. (1973), *Pattern Classification and Scene Analysis*, New York: Wiley.

Dudoit, S., Fridlyand, J., and Speed, T. P. (2000), "Comparison of Discrimination Methods for the Classification of Tumors in Gene Expression Data," Technical Report 576, University of California Berkeley, Dept. of Statistics.

Edwards, A. W. F., and Cavalli-Sforza, L. L. (1965), "A Method for Cluster Analysis," *Biometrics*, 21, 362–375.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998), "Cluster Analysis and Display of Genome-Wide Expression Patterns," *Proceedings of the National Academy of Sciences*, 95, 14863–14868.

Escobar, M. D., and West, M. (1995), "Bayesian Density Estimation and Inference Using Mixtures," *Journal of the American Statistical Association*, 90, 1301–1312.

Fayyad, U., and Smyth, P. (1996), "From Massive Data Sets to Science Catalogs: Applications and Challenges," in *Statistics and Massive Data Sets: Report to the Committee on Applied and Theoretical Statistics*, eds. J. Kettenring and D. Pregibon, National Research Council.

Fraley, C. (1998), "Algorithms for Model-Based Gaussian Hierarchical Clustering," *SIAM Journal on Scientific Computing*, 20, 270–281.

Fraley, C., and Raftery, A. E. (1998), "How Many clusters? Which Clustering Method? Answers via Model-Based Cluster Analysis," *The Computer Journal*, 41, 578–588.

——— (1999), "MCLUST: Software for Model-Based Cluster Analysis," *Journal of Classification*, 16, 297–306.

Friedman, H. P., and Rubin, J. (1967), "On Some Invariant Criteria for Grouping Data," *Journal of the American Statistical Association*, 62, 1159–1178.

Friedman, J. H. (1989), "Regularized Discriminant Analysis," *Journal of the American Statistical Association*, 84, 165–175.

Gerrard, D. J. (1969), "Competition Quotient: A New Measure of the Competition Affecting Individual Forest Trees," Research Bulletin No. 20, Agricultural Experimental Station, Michigan State University.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999), "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, 286, 531–537.

Green, P. J. (1995), "Reversible Jump MCMC Computation and Bayesian Model Determination," *Biometrika*, 82, 711–732.

Hartigan, J. A., and Wong, M. A. (1978), "Algorithm AS 136: A k-Means Clustering Algorithm," *Applied Statistics*, 28, 100–108.

Hastie, T., and Stuetzle, W. (1989), "Principal Curves," *Journal of the American Statistical Association*, 84, 502–516.

Hastie, T., and Tibshirani, R. (1996), "Discriminant Analysis by Gaussian Mixtures," *Journal of the Royal Statistical Society*, Ser. B, 58, 155–176.

Hastie, T., Tibshirani, R., Botstein, D., and Brown, P. (2000a), "Supervised Harvesting of Expression Trees," preprint, Stanford University, Dept. of Statistics.

Hastie, T., Tibshirani, R., Eisen, M., Brown, P., Roos, D., Scherf, U., Weinstein, J., Alizadeh, A., Staudt, L., and Botstein, D. (2000b), "Gene Shaving: A New Class of Clustering Methods for Expression Arrays," preprint, Stanford University, Dept. of Statistics.

Haughton, D. M. A. (1988), "On the Choice of a Model to Fit Data From an Exponential Family," *The Annals of Statistics*, 16, 342–355.

Herwig, R., Poustka, A. J., Müller, C., Bull, C., Lehrach, H., and O'Brien, J. (1999), "Large-Scale Clustering of cDNA Fingerprinting Data," *Genome Research*, 9, 1093–1105.

Holmes, I., and Bruno, W. J. (2000), "Finding Regulatory Elements Using Joint Likelihoods for Sequence and Expression Profile Data," in *ISMB 2000—Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology*, pp. 202–210.

Iyer, V. R., Eisen, M. B., Ross, D. T., Schuler, G., Moore, T., Lee, J. C. F., Trent, J. M., Staudt, L. M., Hudson, J. J., Boguski, M. S., Lashkari, D., Shalon, D., Botstein, D., and Brown, P. O. (1999), "The Transcriptional Program in the Response of Human Fibroblasts to Serum," *Science*, 283, 83–87.

Jeffreys, H. (1961), *Theory of Probability* (3rd ed.), Oxford, U.K.: Clarendon Press.

Jöreskog, K. G. (1973), "A General Method for Estimating a Linear Structural Equation System," in *Structural Equation Models in the Social Sciences*, eds. A. S. Goldberger and O. D. Duncan, New York: Seminar Press, pp. 85–112.

Journel, A. G., and Huibregts, C. J. (1978), *Mining Geostatistics*, New York: Academic Press.

Kaluzny, S. P., Vega, S. C., Cardoso, T. P., and Shelly, A. A. (1998), *S + SpatialStats: User's Manual for Windows and UNIX*, New York: Springer.

Kass, R. E., and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795.

Kass, R. E., and Wasserman, L. (1995), "A Reference Bayesian Test for Nested Hypotheses and Its Relationship to the Schwarz Criterion," *Journal of the American Statistical Association*, 90, 928–934.

Kaufman, L., and Rousseeuw, P. J. (1990), *Finding Groups in Data*, New York: Wiley.

Keribin, C. (1998), "Consistent Estimate of the Order of Mixture Models," *Comptes Rendues de l'Academie des Sciences, Série I—Mathématiques*, 326, 243–248.

Kohonen, T. (1989), *Self-Organization and Associative Memory* (3rd ed.), New York: Springer.

Lavine, M., and West, M. (1992), "A Bayesian Method for Classification and Discrimination," *Canadian Journal of Statistics*, 20, 451–461.

Lazarsfeld, P. F. (1950), "The Logical and Mathematical Foundation of Latent Structure Analysis," in *Studies in Social Psychology in World War II, Vol. 4: Measurement and Prediction*, eds. E. A. Schulman, P. F. Lazarsfeld, S. A. Starr, and J. A. Clausen, Princeton, NJ: Princeton University Press, pp. 362–412.

Lazarsfeld, P. F., and Henry, N. W. (1968), *Latent Structure Analysis*, New York: Houghton Mifflin.

Lazzeroni, L., and Owen, A. (2000, March), "Plaid Models for Gene Expression Data," Technical Report, Stanford University, Dept. of Statistics.

Leroux, M. (1992), "Consistent Estimation of a Mixing Distribution," *The Annals of Statistics*, 20, 1350–1360.

MacEachern, S. N., and Müller, P. (1998), "Estimating a Mixture of Dirichlet Process Models," *Journal of Computational and Graphical Statistics*, 7, 223–238.

MacQueen, J. (1967), "Some Methods for Classification and Analysis of Multivariate Observations," in *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, eds. L. M. L. Cam and J. Neyman, Berkeley, CA: University of California Press, pp. 281–297.

Mangasarian, O. L., Street, W. N., and Wolberg, W. H. (1995), "Breast Cancer Diagnosis and Prognosis via Linear Programming," *Operations Research*, 43, 570–577.

Marron, J. S., and Wand, M. P. (1992), "Exact Mean Integrated Squared Error," *The Annals of Statistics*, 20, 712–536.

McLachlan, G. J. (1992), *Discriminant Analysis and Statistical Pattern Recognition*, New York: Wiley.

McLachlan, G. J., and Basford, K. E. (1988). *Mixture Models: Inference and Applications to Clustering*, New York: Marcel Dekker.

McLachlan, G. J., and Krishnan, T. (1997), *The EM Algorithm and Extensions*, New York: Wiley.

McLachlan, G. J., Peel, D., Basford, K. E., and Adams, P. (1999), "The EMMIX Software for the Fitting of Mixtures of Normal t-Components," *Journal of Statistical Software*, 4, (on-line publication) www.jstatsoft.org.

Meila, M. (1999), "Learning Mixtures of Trees," Ph.D. thesis, Massachusetts Institute of Technology.

Mengersen, K., and Robert, C. P. (1996), "Testing for Mixtures: A Bayesian Entropic Approach," in *Bayesian Statistics 5*, eds. J. M. Bernardo, J. O. Berger, A. P. David, and A. F. M. Smith, Oxford, U.K.: Oxford University Press, pp. 255–276.

Michaels, G. S., Carr, D. B., Askenazi, M., Fuhrman, S., Wen, X., and Somogyi, R. (1998), "Cluster Analysis and Data Visualization of Large-Scale Gene Expression Data," *Pacific Symposium on Biocomputing*, 3, 42–53.

Moore, A. (1999), "Very Fast EM-based Mixture Model Clustering Using Multiresolution kd-Trees," in *Advances Neural Information Processing Systems*, Vol. 11, eds. M. Kearns, S. Solla, and D. Cohn. MIT Press, pp. 543–549.

Moore, A., and Lee, M. S. (1998), "Cached Sufficient Statistics for Efficient Machine Learning With Large Datasets," *Journal of Artificial Intelligence Research*, 8, 67–91.

Mukherjee, S., Feigelson, E. D., Babu, G. J., Murtagh, F., Fraley, C., and Raftery, A. E. (1998), "Three Types of Gamma Ray Bursts," *The Astrophysical Journal*, 508, 314–327.

Müller, P., Erkanli, A., and West, M. (1996), "Bayesian Curve Fitting Using Multivariate Normal Mixtures," *Biometrika*, 83, 67–80.

Murtagh, F., and Raftery, A. E. (1984), "Fitting Straight Lines to Point Patterns," *Pattern Recognition*, 17, 479–483.

Murtagh, F., Starck, J.-L., and Berry, M. W. (2000), "Overcoming the Curse of Dimensionality by Means of the Wavelet Transform," *The Computer Journal*, 43, 107–120.

Oh, M.-S., and Raftery, A. E. (2001), "Bayesian Multidimensional Scaling and Choice of Dimension," *Journal of the American Statistical Association*, 96, 1031–1044.

Ormoneit, D., and Tresp, V. (1998), "Averaging, Maximum Penalized Likelihood and Bayesian Estimation for Improving Gaussian Mixture Probability Density Estimates," *IEEE Transactions on Neural Networks*, 9, 639–649.

Peel, D., and McLachlan, G. J. (2000), "Robust Mixture Modeling Using the t-Distribution." *Statistics and Computing*, 10, 335–344.

Peel, D., Whitten, W. J., and McLachlan, G. J. (2001), "Fitting Mixtures of Kent Distributions to Aid in Joint Set Identification," *Journal of the American Statistical Association*, 96, 56–63.

Perou, C. M., Jeffrey, S. S., van de Rijn, M., Rees, C. A., Eisen, M. B., Ross, D. T., Pergamenschikov, A., Williams, C. F., Zhu, S. X., Lee, J. C. F., Lashkari, D., Brown, P. O., and Botstein, D. (1999), "Genome-Wide Analysis of DNA Copy-Number Changes Using cDNA microarrays," *Nature Genetics*, 23, 41–46.

Posse, C. (2001), "Hierarchical Model-Based Clustering for Large Datasets," *Journal of Computational and Graphical Statistics*, 10, 464–486.

Raftery, A. E. (1995), "Bayesian Model Selection in Social Research (with discussion)," *Sociological Methodology*, 25, 111–193.

——— (1999), "Bayes Factors and BIC: Comment on 'A Critique of the Bayesian Information Criterion for Model Selection'," *Sociological Methods and Research*, 27, 411–427.

Richardson, S., and Green, P. J. (1997), "On Bayesian Analysis of Mixtures With an Unknown Number of Components," *Journal of the Royal Statistical Society*, Ser. B, 59, 731–758.

Ripley, B. D. (1996), *Pattern Recognition and Neural Networks*, Cambridge, U.K.: Cambridge University Press.

Roeder, K., and Wasserman, L. (1997), "Practical Bayesian Density Estimation Using Mixtures of Normals," *Journal of the American Statistical Association*, 92, 894–902.

Ross, D. T., Scherf, U., Eisen, M. B., Perou, C. M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S. S., Van de Rijn, M., Waltham, M., Pergamenschikov, A., Lee, J. C. F., Lashkari, D., Shalon, D., Myers, T. G., Weinstein, J. N., Botstein, D., and Brown, P. O. (2000), "Systematic Variation in Gene Expression Patterns in Human Cancer Cell Lines," *Nature Genetics*, 24, 227–235.

Sampson, P. D., and Guttorp, P. (1992), "Nonparametric Estimation of Nonstationary Spatial Covariance Structure," *Journal of the American Statistical Association*, 87, 108–119.

Scherf, U., Ross, D. T., Waltham, M., Smith, L. H., Lee, J. K., Tanabe, L., Kohn, K. W., Reinhold, W. C., Myers, T. G., Andrews, D. T., Scudiero, D. A., Eisen, M. B., Sausville, E. A., Pommier, Y., Botstein, D., and Brown, P. O. (2000), "A gene expression database for the Molecular Pharmacology of Cancer," *Nature Genetics*, 24, 236–244.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Scott, A. J., and Symons, M. J. (1971), "Clustering Methods Based on Likelihood Ratio Criteria," *Biometrics*, 27, 387–397.

Scott, D. W. (1992), *Multivariate Density Estimation*, New York: Wiley.

Smyth, P. (2000), "Model Selection for Probabilistic Clustering Using Cross-Validated Likelihood," *Statistics and Computing*, 10, 63–72.

Stanford, D., and Raftery, A. E. (2000), "Principal Curve Clustering With Noise," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22, 601–609.

Stephens, M. (1997), Discussion of "On the Bayesian Analysis of Mixtures With an Unknown Number of Components," by S. Richardson and P. J. Green, *Journal of the Royal Statistical Society*, Ser. B, 59, 768–769.

——— (2000), "Dealing With Label-Switching in Mixture Models," *Journal of the Royal Statistical Society*, Ser. B, 62, 795–809.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S., and Golub, T. R. (1999), "Interpreting Patterns of Gene Expression With Self-Organizing Maps: Methods and Application to Hematopoietic Differentiation," *Proceedings of the National Academy of Sciences*, 96, 2907–2912.

Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., and Church, G. M. (1999), "Systematic Determination of Genetic Network Architecture," *Nature Genetics*, 22, 281–285.

Thiesson, B., Meek, C., and Heckerman, D. (1999), "Accelerating EM for Large Databases," Technical Report MSR-TR-99-31, Microsoft Research.

Tibshirani, R., Hastie, T., Ross, D., Botstein, D., and Brown, P. (1999), "Clustering Methods for the Analysis of DNA Microarray Data," preprint, Stanford University, Dept. of Health Research and Policy.

Tibshirani, R., Walther, G., and Hastie, T. (2000), "Estimating the Number of Clusters in a Dataset via the Gap Statistic," preprint, Stanford University, Dept. of Statistics.

van der Laan, M. J., and Bryan, J. F. (2000), "Gene Expression Analysis With the Parametric Bootstrap," preprint, University of California, Berkeley, Division of Biostatistics.

Ward, J. H. (1963), "Hierarchical Groupings to Optimize an Objective Function," *Journal of the American Statistical Association*, 58, 234–244.

Wasserman, W. W., and Fickett, J. W. (1998), "Identification of Regulatory Regions Which Confer Muscle-Specific Gene Expression," *Journal of Molecular Biology*, 278, 167–181.

Weinstein, J. N., Myers, T. G., O'Connor, P. M., Friend, S. H., Fornace, A. J. Jr., Kohn, K. W., Fojo, T., Bates, S. E., Rubenstein, L. V., Anderson, N. L., Buolamwini, J. K., van Osdol, W. W., Monks, A. P., Scudiero, D. A., Sausville, E. A., Zaharevitz, D. A., Bunow, B., Viswanadhan, V. N., Johnson, G. S., Wittes, R. E., and Paull, K. D. (1997), "An Information Intensive Approach to the Molecular Pharmacology of Cancer," *Science*, 275, 343–349.

Wen, X., Fuhrman, S., Michaels, G. S., Carr, D. B., Smith, S., Barker, J. L., and Somogyi., R. (1998), "Large-Scale Temporal Gene Expression Mapping of Central Nervous System Development," *Proceedings of the National Academy of Sciences*, 95, 334–339.

West, M., Nevins, J. R., Marks, J. R., Blanchette, C., Spang, R., and Zuzan, H. (2000), "Bayesian Regression Analysis in the 'Large p, Small n' Paradigm With Application in DNA Microarray Studies," preprint, Duke University, Institute of Statistics and Decision Sciences.

Witherspoon, N. H., J. H. Holloway Jr., K. S. Davis, R. W. Miller, and A. C. Dubey (1995). "The Coastal Battlefield Reconnaissance and Analysis (COBRA) Program for Minefield Detection," in *Detection and remediation technologies for mines and minelike targets*, Vol. 2496, ed. A. C. Dubey, *Proceedings of SPIE*, pp. 500–508. Orlando, FL.

Wolfe, J. H. (1963), "Object Cluster Analysis of Social Areas," Master's thesis, University of California, Berkeley.

——— (1965), "A Computer Program for the Maximum-Likelihood Analysis of Types," USNPRA Technical Bulletin 65-15, U.S. Naval Personnel Research Activity, San Diego.

——— (1967), "NORMIX: Computational Methods for Estimating the Parameters of Multivariate Normal Mixture Distributions," Technical Bulletin USNPRA SRM 68-2, U.S. Naval Personnel Research Activity, San Diego.

——— (1970), "Pattern Clustering by Multivariate Mixture Analysis," *Multivariate Behavioral Research*, 5, 329–350.

Woolf, P. J., and Wang, Y. (2000), "A Fuzzy Logic Approach to Analyzing Gene Expression Data," *Physiological Genomics*, 3, 9–15.

Wu, C. F. J. (1983), "On Convergence Properties of the EM Algorithm," *The Annals of Statistics*, 11, 95–103.

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E., and Ruzzo, W. L. (2001), "Model-Based Clustering and Data Transformation for Gene Expression Data," *Bioinformatics*, 17, 977–987.

Yeung, K. Y., and Ruzzo, W. L. (2001), "Principal Component Analysis for Clustering Gene Expression Data," *Bioinformatics*, 17, 763–774.