# Model-Based Clustering, Discriminant Analysis, and Density Estimation [1]

Chris Fraley and Adrian E. Raftery
University of Washington

Working Paper no. 11
Center for Statistics and the Social Sciences
University of Washington
Box 354322
Seattle, WA 98195-4322, USA

October 2000

[1]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 98195-4322. Email: `fraley/raftery@stat.washington.edu`; Web: `www.stat.washington.edu/fraley` and `www.stat.washington.edu/raftery` This research was supported by the Office of Naval Research under 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 useful comments and discussions, and to Simon Byers for the `NNclean` denoising software.

**Abstract**

Cluster analysis is the automated search for groups of related observations in a data set. 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 outline 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.

# Contents



# List of Figures

# 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, 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. dendogram 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 are there?", "Which clustering method should be used?" and "How should outliers be handled?". Moreover, the statistical properties of these methods are generally unknown, precluding the possibility of formal inference.

It was realized early on cluster analysis can also be based on probablility models (see Bock 1996, 1998 for a survey). This realization has provided insight into when a particularly 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 particular probability models. For example, standard $k$ means clustering and Ward's method are equivalent to known procedures for approximately maximizing the multivariate normal classfication 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 paper 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 multivariate normal mixture models with covariances parametrized 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, while inititialization is critical in EM since the likelihood surface tends to have multiple modes, although EM typically produces improved partitions when started from reasonable ones. By initializing the EM iteration with partitions from model-based hierarchical agglomeration and using approximate Bayes factors with the 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 algorithm was extended by Fraley and Raftery (1998) to select the parametrization of the model as well as the number of clusters simultaneously using the BIC.

Figure 1*a* shows the two-group model-based classification of a data set used for breast cancer diagnosis (Magasarian et al. 1995). Although no information about the known malignant vs. 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 paper, which makes use of the known classifications of the UCI data. Nearly 96% of these new data points are correctly classified by this procedure. This data set is discussed in more detail in Section 8.1.

This paper reviews the model-based approach to clustering, and shows how it can also

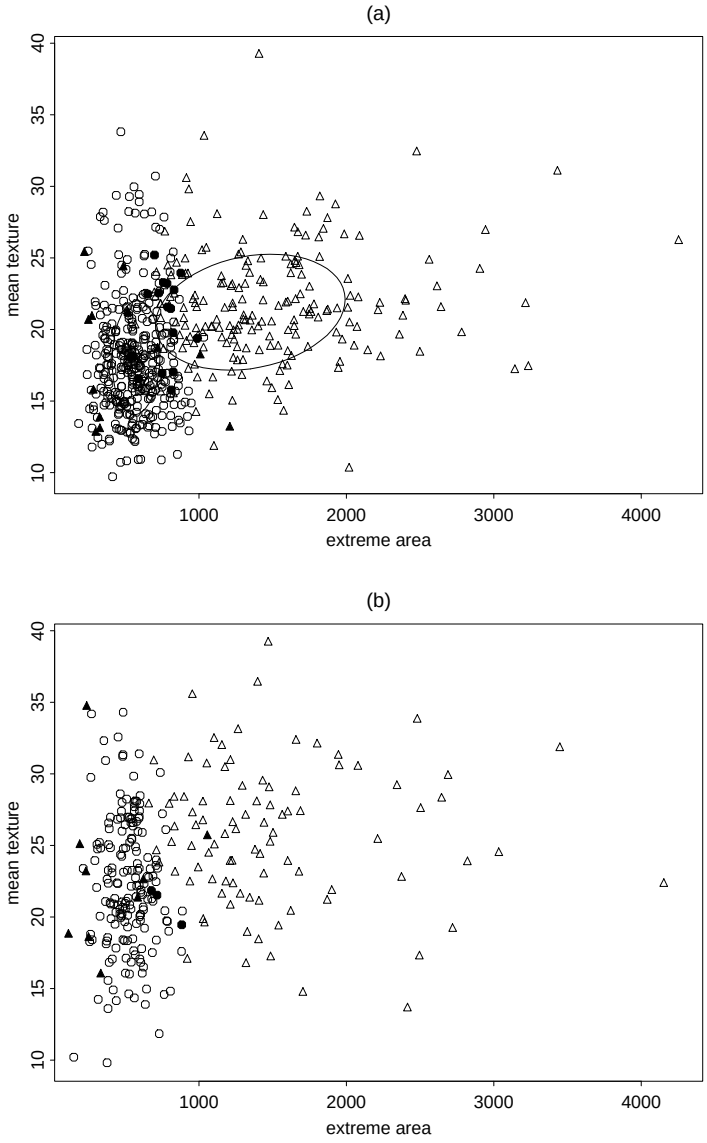


Figure 1: ($a$) A Projection of the UCI Wisconsin Diagnostic Breast Cancer Data Showing the Two-Group Model-Based Classification. The ellipses shown are projections of the ellipsoids defined by the covariances of the two multivariate normal components in the mixture model fitted to the data. There are 569 observations. Although no information about the known malignant vs. 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. ($b$) A Projection of 280 Additional Observations. This shows 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. Circles represent benign observations; triangles malignant ones. Filled symbols represent misclassified observations. The resulting out-of-sample classification is nearly 96% correct.

be applied in discriminant analysis and multivariate density estimation. The organization is as follows. Section 2 is a discussion of mixture models, including the multivariate normal model and the geometric interpretation of its parametrization by eigenvalue decomposition. The EM iteration for maximum likelihood estimation and its specialization to mixtures is the topic of Section 3. 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. Examples illustrating these methods are given in Section 8. Section 9 gives sources for model-based clustering software. Section 10 discusses some limitations of the method and suggests extensions to overcome them, including strategies for large data sets.

# 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

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

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

Most commonly, $f_k$ is the multivariate normal (Gaussian) density $\phi_k$, parametrized by its mean $\mu_k$ and covariance matrix $\Sigma_k$:

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

Data generated by mixtures of multivariate normal densities are characterized by groups or clusters centered at the means $\mu_k$, 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 $\Sigma_k$, which may also be parametrized 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 $\Sigma_k$, 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, while $d(d+1)/2$ and $G(d(d+1)/2)$ parameters are required for constant $\Sigma_k$ and unrestricted $\Sigma_k$, respectively, if the data are $d$-dimensional.

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

$$\Sigma_k = \lambda_k D_k A_k D_k^T, \tag{3}$$

where $D_k$ is the orthogonal matrix of eigenvectors, $A_k$ is a diagonal matrix whose elements are proportional to the eigenvalues, and $\lambda_k$ is an associated constant of proportionality. Their idea was to treat $\lambda_k$, $A_k$ 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 $k$th component of the mixture, $A_k$ its shape, and $\lambda_k$ its volume, which is proportional to $\lambda_k^d \det(A_k)$. For example, if the largest eigenvalue of $\Sigma_k$ is much larger than the other eigenvalues, the $k$-th cluster will be concentrated close to a line in $d$-space, which will be the first principal component of the distribution of the $k$-th group. Similarly, if the two largest eigenvalues are of the same magnitude and dominate the other eigenvalues, the $k$-th cluster will be concentrated close to a plane in $d$-space. The $k$-th cluster will be roughly spherical if the largest and smallest eigenvalues of $\Sigma_k$ 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 — $\lambda I$, equal variance, and unconstrained variance — mentioned above, as well as other useful models, such as $\Sigma_k = \lambda_k I$ where the clusters are spherical, but with 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 parsimonious parametrizations of covariance matrices have been proposed that could be applied in the context of cluster analysis. These include the intra-class correlation or one-factor model, in which all 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 parametrizations 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), either in a Euclidean or a deformed space (Sampson and Guttorp 1992).

# 3 The EM Iteration for Mixture Models

The EM (Expectation–Maximization) 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 $\theta$, 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,

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

The MLE for $\theta$ 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 loglikelihood given the observed data and the current parameter estimates is computed, and an 'M-step' in which parameters that maximize the expected loglikelihood 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 in order 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, Laird and Rubin 1977; Boyles 1983; Wu 1983; McLachlan and Krishnan 1997). Although these conditions do not always hold in practice, the EM iteration 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

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

Assuming that each $\mathbf{z}_i$ is independent and identically distributed 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 loglikelihood is

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

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

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

while the M-step involves maximizing (6) in terms of $\tau_k$ and $\theta_k$ 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 $\{m : z_{im}^* = \max_k z_{ik}^*\}$, so that $(1 - \max_k z_{ik}^*)$ is a measure of the uncertainty in the classification (Bensmail et al., 1997).

For multivariate normal mixtures, the E-step is given by (7) with $f_k$ replaced by $\phi_k$ as defined in (2), regardless of the parametrization. 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:

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

Computation of the covariance estimate $\hat{\Sigma}_k$ depends on its parametrization. For details of the M-step for $\Sigma_k$ parametrized by the eigenvalue decomposition (3), see 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 iteration 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 they contain are concentrated close to a linear subspace of lower dimension than the data.

A variant of EM called *classification EM* or 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) 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 questions

reduce 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 allows simultaneous selection of the number of groups and the clustering model. The problem then reduces to comparison among the members of a set of possible models.

There are tradeoffs between the choice of the number of clusters and that of the clustering model. If a simpler model is used, more clusters may be needed to provide a good representation of the data. If a more complex model is used, fewer clusters may suffice. As a simple example, consider the situation where there is 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 $k$ means) were used to fit this data, 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, then, by the law of total probability, $p(D|M_k)$ is obtained by integrating (not maximizing) over the parameters, i.e.

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

where $p(\theta_k|M_k)$ is the prior distribution of $\theta_k$, 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*, and if the prior model probabilities, $p(M_k)$, are the same, 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. In addition to

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 the use of 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 Bayesian Information Criterion or BIC:

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

where $\nu_k$ 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 is 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, while 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, 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 et al. 1997, 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) discuss 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 AWE, but in subsequent experiments it has consistently performed less well than BIC. Cheeseman and Stutz (1995) and Chickering and Heckerman (1997) use 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 et al. 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 are compared in Biernacki and Govaert (1999).

Bensmail et al. (1997) discuss an alternative approximation to the integrated likelihood for choosing both the number of groups and the clustering model based on Markov chain Monte Carlo 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. The parametrization (3) is used as the basis for a class of models that is sufficiently flexible to accomodate data with widely varying characteristics. The strategy consists of 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*

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

where the $\ell_i$ are labels indicating a unique classification of each observation: $\ell_i = k$ if $\mathbf{y}_i$ belongs to the $k$th 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 parametrized as in (3), with volume and shape ($\lambda_k$ 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 uniform spherical covariance $\lambda I$, the selection criterion is the well-known sum-of-squares criterion (Ward 1963).

Other common heuristic clustering criteria, such as the *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 generalizations to nonisotropic settings, or situations in which the groups differ markedly. The single link clustering method seems not to be 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 analogue 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 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 are discussed in 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 iteration. 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.
- Implement 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 parametrized 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 only for one of the models (e.g. unconstrained covariance), using the resulting partitions as starting values for EM with any other parametrization. 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

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

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 data set. Initial values for $z_{ik}$ are formed by augmenting the indicator variables from the hierarchical clustering step with a row of zeroes for each observation initially assessed as being noise, and a column of indicator variables giving the result of the denoising step (1 indicating noise; 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, $C$, is assumed to be known.

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

$$\Pr[\mathbf{y} \in \text{Class } c] = \frac{\tau_c f_c(\mathbf{y})}{\sum_{k=1}^{C} \tau_k f_k(\mathbf{y})}.$$

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 the $c$th class are multivariate normal, so that

$$f_c(\mathbf{y}) = \phi(\mathbf{y}|\mu_c, \Sigma_c). \tag{12}$$

If the covariance matrices for the different classes are the same, i.e. $\Sigma_c = \Sigma$ for $c = 1, \ldots, C$, and if maximum likelihood estimates of $\mu_c$ and $\Sigma$ 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 $\Sigma_c$ are allowed to differ without constraint, 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 subsections.

## 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 parametrization 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 (Section 4), which would typically be less demanding computationally. Biernacki and Govaert (1999) compare 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 for 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, 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 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

$$f_c(\mathbf{y} \mid \theta_k) = \sum_{k=1}^{G_c} \tau_{ck} \phi(\mathbf{y} \mid \mu_{ck}, \Sigma_{ck}). \tag{14}$$

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 or MDA (Hastie and Tibshirani 1996). In developing MDA, Hastie and Tibshirani made two assumptions: $(i)$ that all of the component covariance matrices are the same, i.e. $\Sigma_{ck} = \Sigma$ for each $c$, $k$; and $(ii)$ that the number of mixture components is known in advance for each class. When Learning Vector Quantization (Kohonen 1989) is used for initialization, however, 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 assumptions $(i)$ and $(ii)$ 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 parametrization of the covariance matrix and which number of mixture components is best suited to each class. We shall refer to this generalization of MDA as 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, LDA and QDA. Friedman (1989) had 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 data sets, 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, since 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), while 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 in 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 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

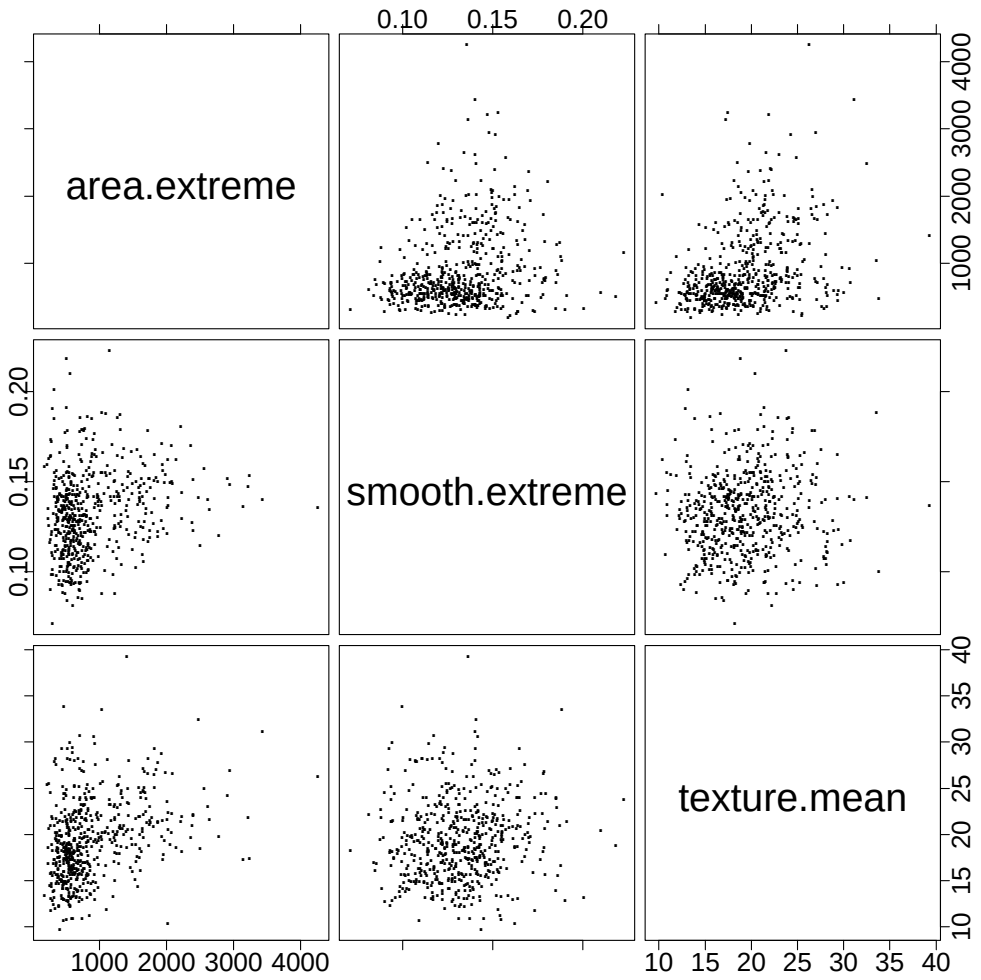


Figure 2: Pairs plots of the Wisconsin Diagnostic Breast Cancer Data from the UCI Machine Learning Repository, showing only the 3 explanatory variables used by Mangasarian et al. (1995). There are 569 observations.

to locate planes separating classes of data. Their 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 2, 3, and 4 features and 1 or 2 linear separating planes. Their 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 for these data the diagnoses are available, we first applied cluster analysis to the three attributes only, ignoring the "known" classifications. The model-based clustering methodology outlined in Section 5 yields the results shown in Figure 3. The maximum BIC value occurs for the 3-group unconstrained model; the difference in BIC values between the 2- and 3- group unconstrained models is close enough to conclude that there are either 2 or

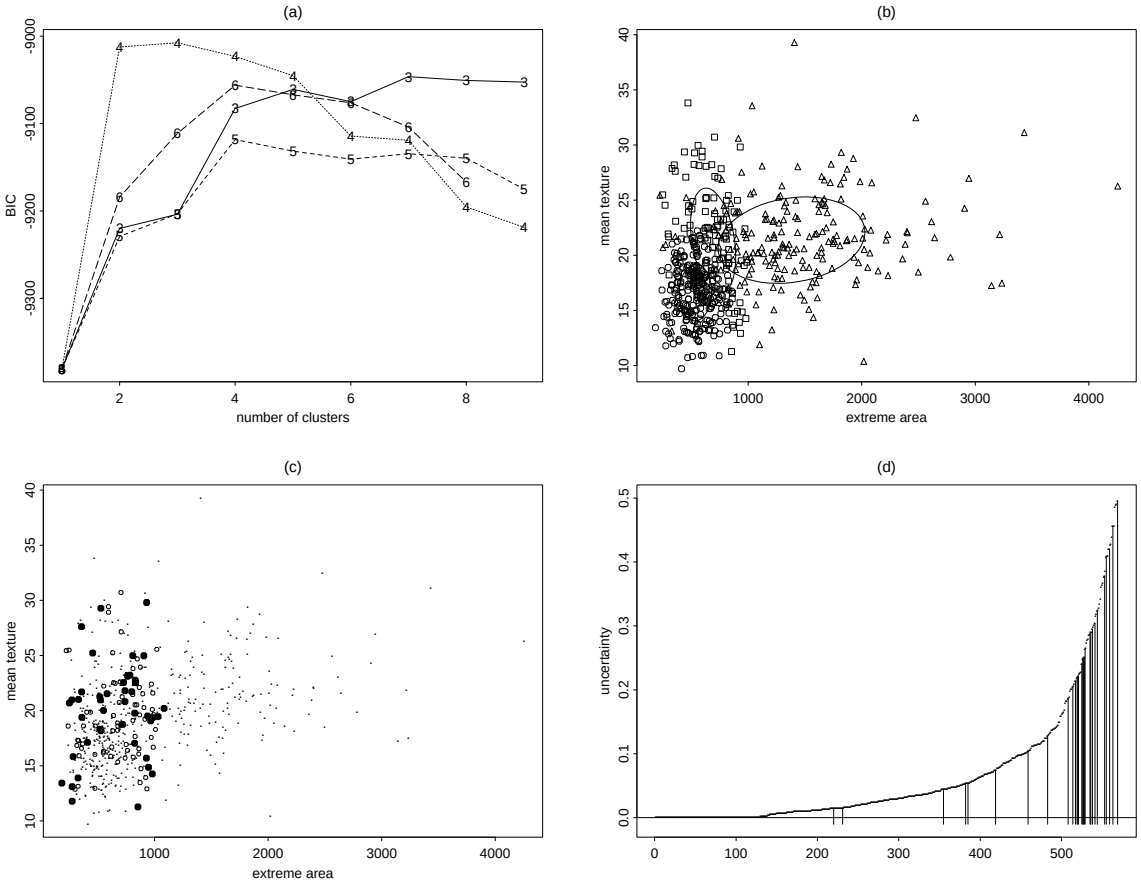


Figure 3: Cluster analysis of the Wisconsin Diagnostic Data reduced to the three explanatory variables. ($a$) BIC values, excluding those for the two spherical models since they fall well below the others. Models 3–6 correspond to $\Sigma$ (equal variance), $\Sigma_k$ (unconstrained), $\lambda D_k A D_k^T$ (common shape and volume), and $\lambda_k D_k A D_k^T$ (common shape), respectively. Model 4 is the best model. ($b$) The 3-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 2-group model-based classification (shown in Figure 1). Small dots correspond to observations with uncertainty less than .1; open circles to those with uncertainty in the interval [.1, .25); 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 2-group model-based classification.

3 groups in the data (Figure 3*a*). The 2-group classification matches the clinical diagnosis for all but 29 of the 569 observations (see Figure 1). Note that the most uncertain points tend to fall in the same region between the two clusters as the misclassified data (Figure 3*c*), while the location of uncertainty of the misclassified observations relative to the uncertainty of all of the observations (Figure 3*d*) confirms that the more uncertain observations are also the ones most likely to be misclassified.

Three groups are clinically important because it is necessary to have some idea of the chance of malignancy in order to determine an appropriate course of action. Tumors of the intermediate class would be followed up by biopsy under local anesthesia, while 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 in Mangasarian et al. (1995) correctly diagnosed 176 consecutive new patients as of November 1995. Since only the training set is available from the UCI repository, we obtained additional data for discriminant analysis from Dr. William Wolberg, M.D., 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 (Figure 1*b*). The model-based approach has the advantage over the linear programming method of Mangasarian et al. (1995) that it generalizes easily to data in which more than two groups are present, and that 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, chapter 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 vs. nondiseased observations from plasma lipid data are shown in Scott (1992), p. 250–251. The parametric density estimate was obtained by fitting a single normal to each of two sets of observations, while the nonparametric estimate is an average shifted histogram. Scott considered only two possibilities: a completely parametric (multivariate normal) den-

sity, and a fully nonparametric approach via kernel density estimation. MclustDA includes a single normal density as a special case, and will collapse down 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 Wisconsion 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 for the malignant ones. 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 if 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 actually are mines, but that are identified as nonmines). Because of the considerable linear dependence among 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 4-group nonconstant spherical model. 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 over 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 in Murtagh et al. (2000) that is based on the problem of locating galaxies in a noisy astronomical image. 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 respectively $(10, 20)$ and $(18, 10)$. There are

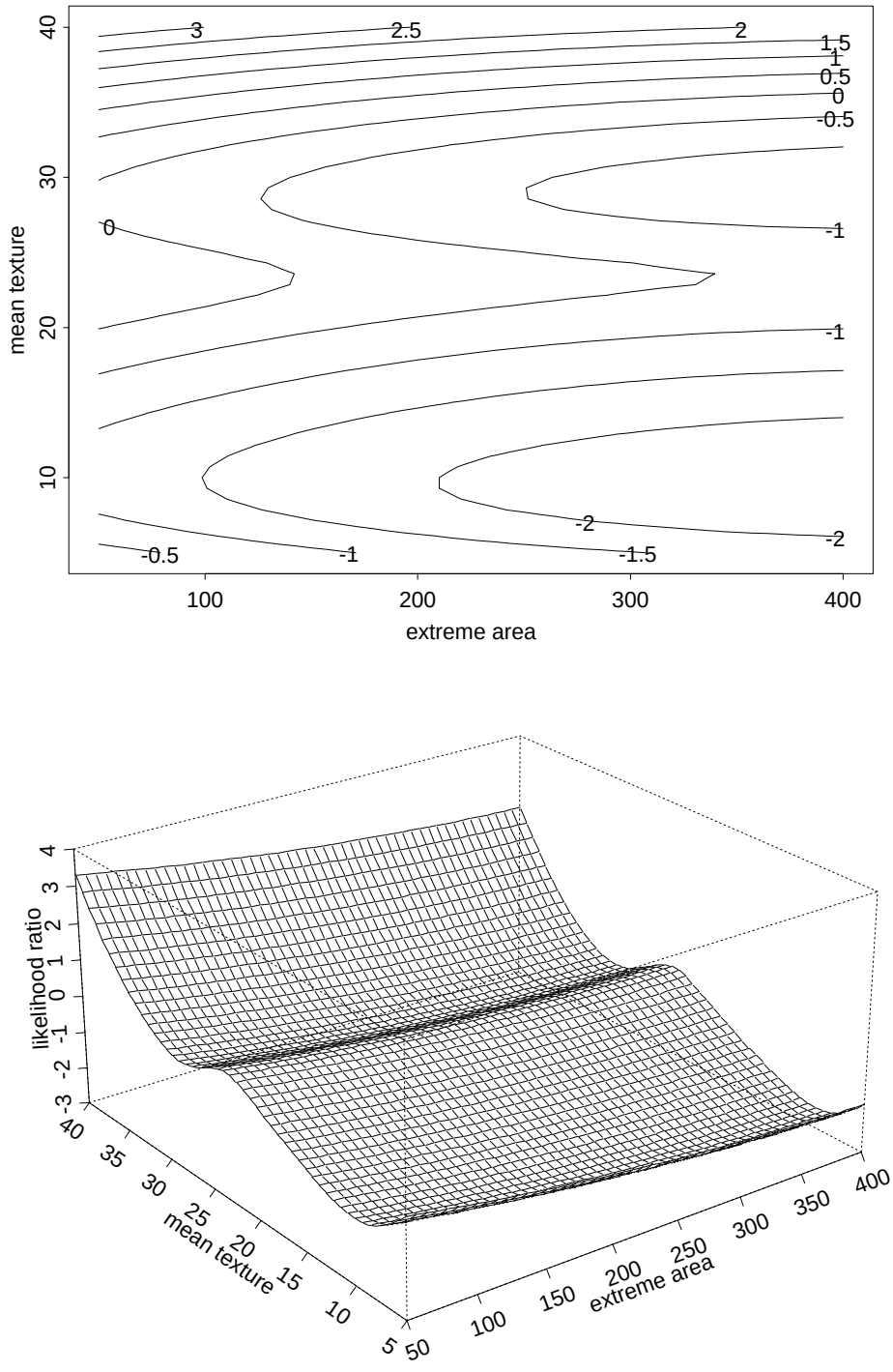


Figure 4: Using MclustDA: Contour and perspective plots of a portion of the loglikelihood ratio surface for two covariates of the UCI Wisconsin Breast Cancer Data obtained from density estimation via model-based clustering.

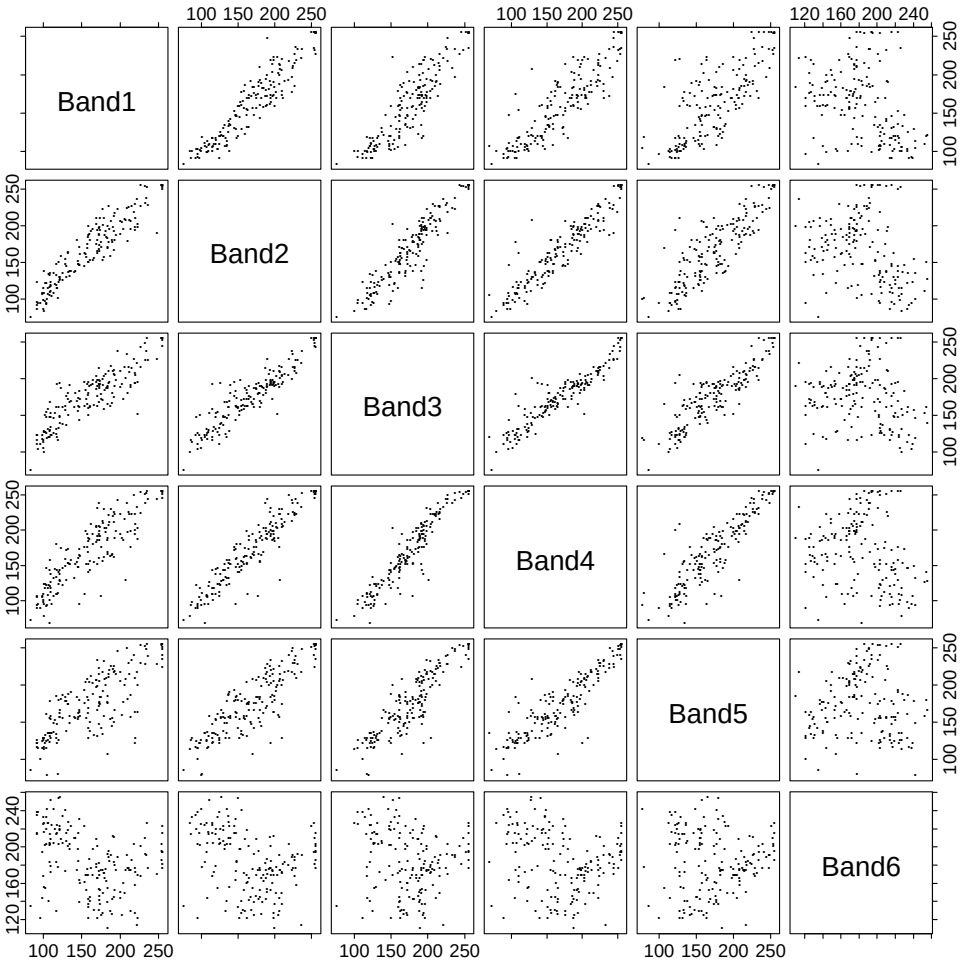


Figure 5: The six bands of a COBRA reconnaissance image.

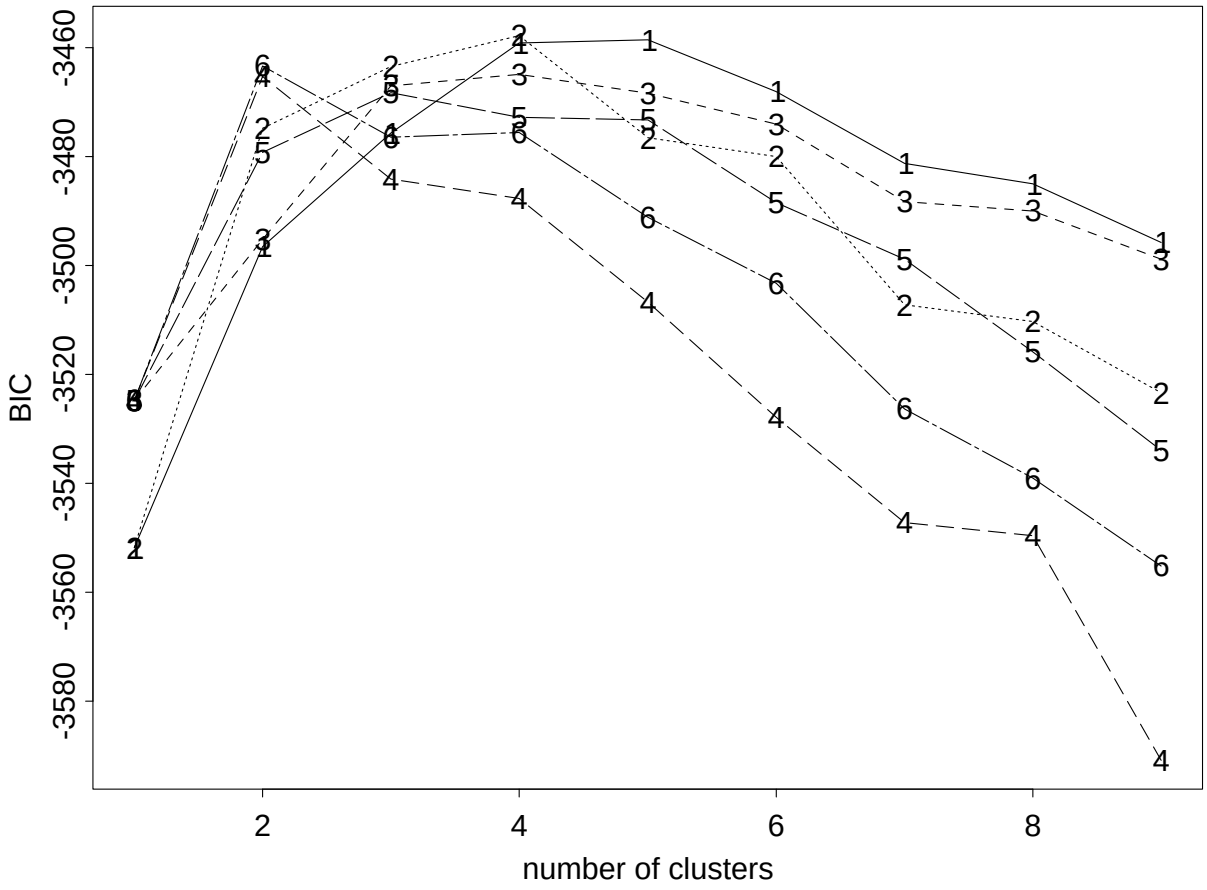


Figure 6: BIC for the COBRA minefield detection problem, using bands 1 and 6. Models 1 and 2 are $\Sigma_k = \lambda I$ (constant spherical), and $\lambda_k I$ (nonconstant spherical), while models 3–6 are as given in Figure 3. The resulting classification reduced the number of false positives by more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

300 data points in the first of these clusters, and there are 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 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 et al. 1993).

## 8.4 Spatial Density Estimation

As an illustration of density estimation with multivariate mixtures (Section 7), we consider the density of the Lansing Woods maples (Gerrard 1969). Figure 8 shows the location of the maples, the model-based classsification, the corresponding density, and a standard Gaussian kernel density estimate. The BIC (Figure 8*a*) indicates that a nonuniform spherical model with six groups is the best model among those available. The Gaussian kernel density estimate (Figure 8*d*) was computed with the `S+SpatialStats` software (Kaluzny et al. 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 gives the average integrated mean squared error (MISE) for density estimation via model-based clustering, as well as 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

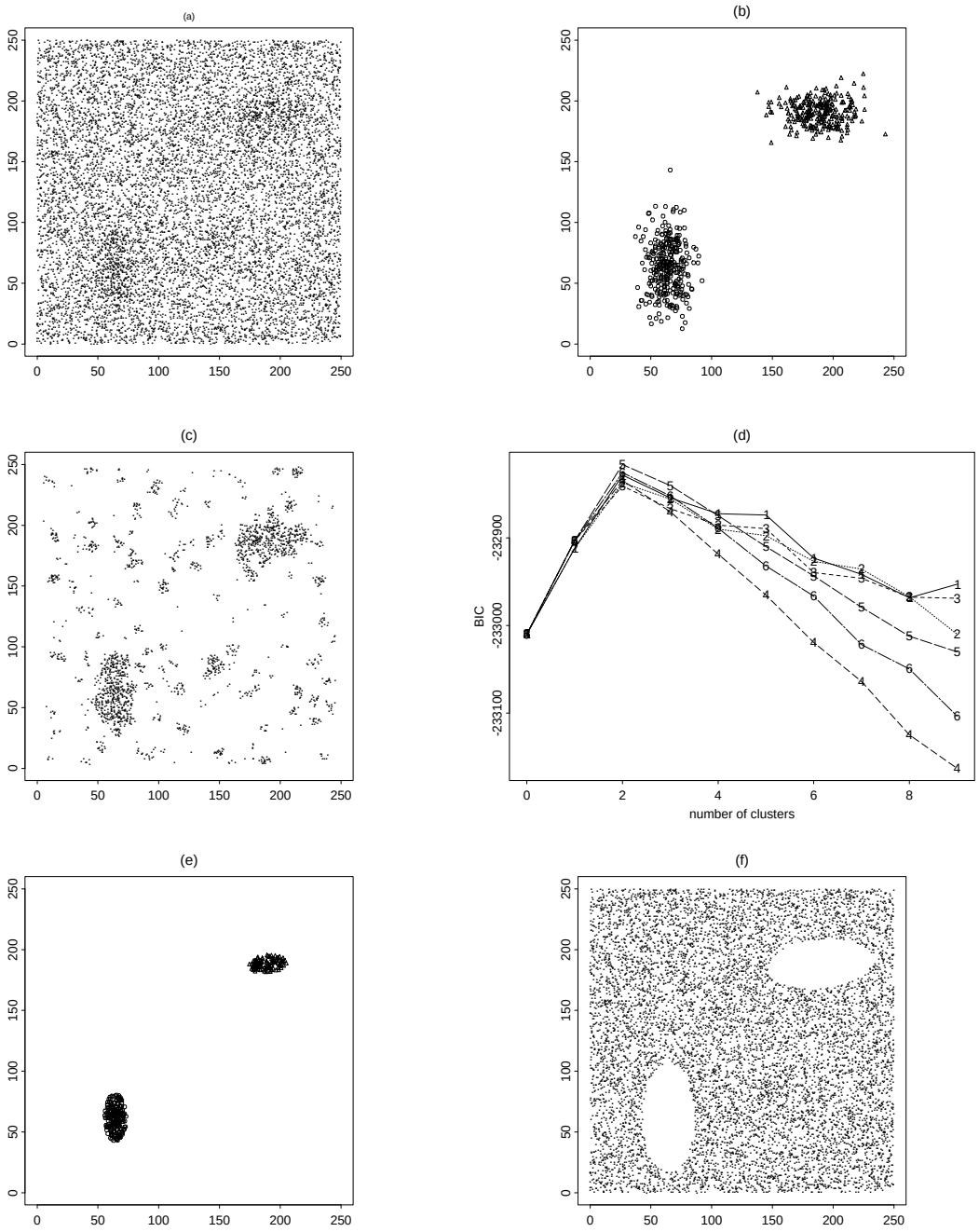


Figure 7: ($a$) An instance of the cluster recovery data, consisting of two Gaussian clusters with a total of 550 points, and 10,000 noise points. ($b$) The Gaussian clusters. ($c$) The data after 20 nearest-neighbor denoising with `NNclean`. ($d$) BIC from model-based clustering. In model 5, groups have equal shape and volumes. ($e$) Model-based classification. ($f$) Points with classification uncertainty less than 0.1.

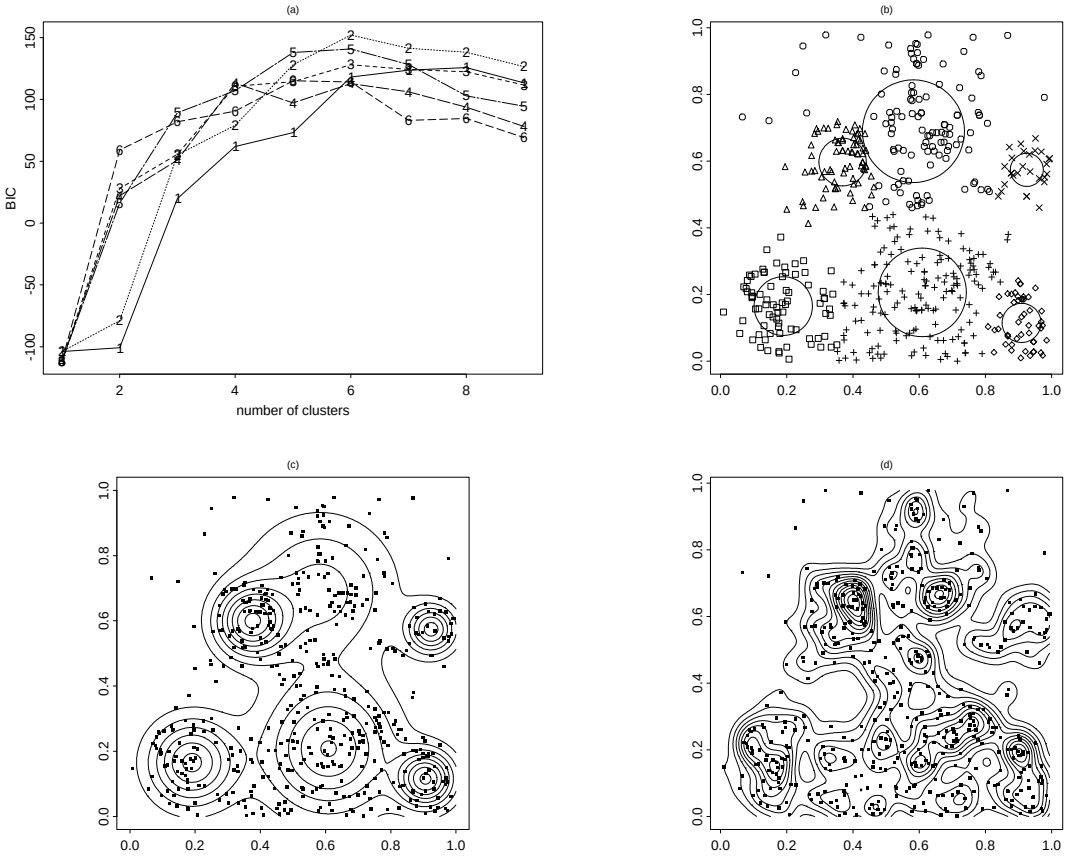


Figure 8: Density estimation for the Lansing Woods maples. ($a$) BIC from model-based clustering. The maximum-BIC model is a six-component nonuniform spherical mixture. ($b$) Model-based classification, with circles indicating the circles defined by the estimated covariance of each of the six groups. ($c$) 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.

are the MISE for kernel density estimation divided by the MISE for model-based clustering, for each of the two kernel methods. This provides a direct comparison between model-based clustering and kernel estimation in each of the simulated situations. Only in one of the ten simulated situations does kernel estimation outperform model-based clustering: the `Claw (Bart Simpson)` density, which is the most complicated of the 10 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.

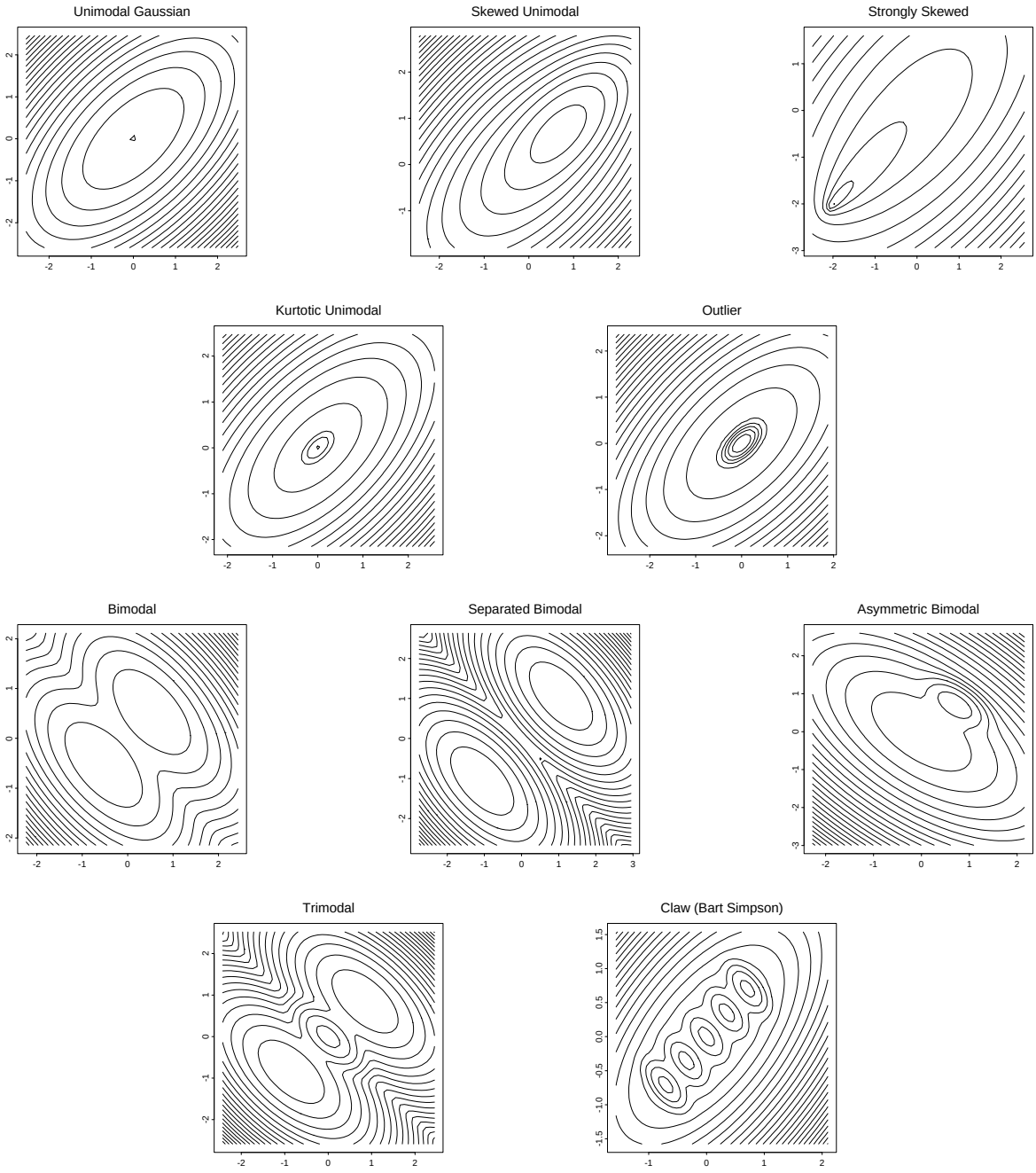


Figure 9: Contours of the 10 2-dimensional simulation densities.

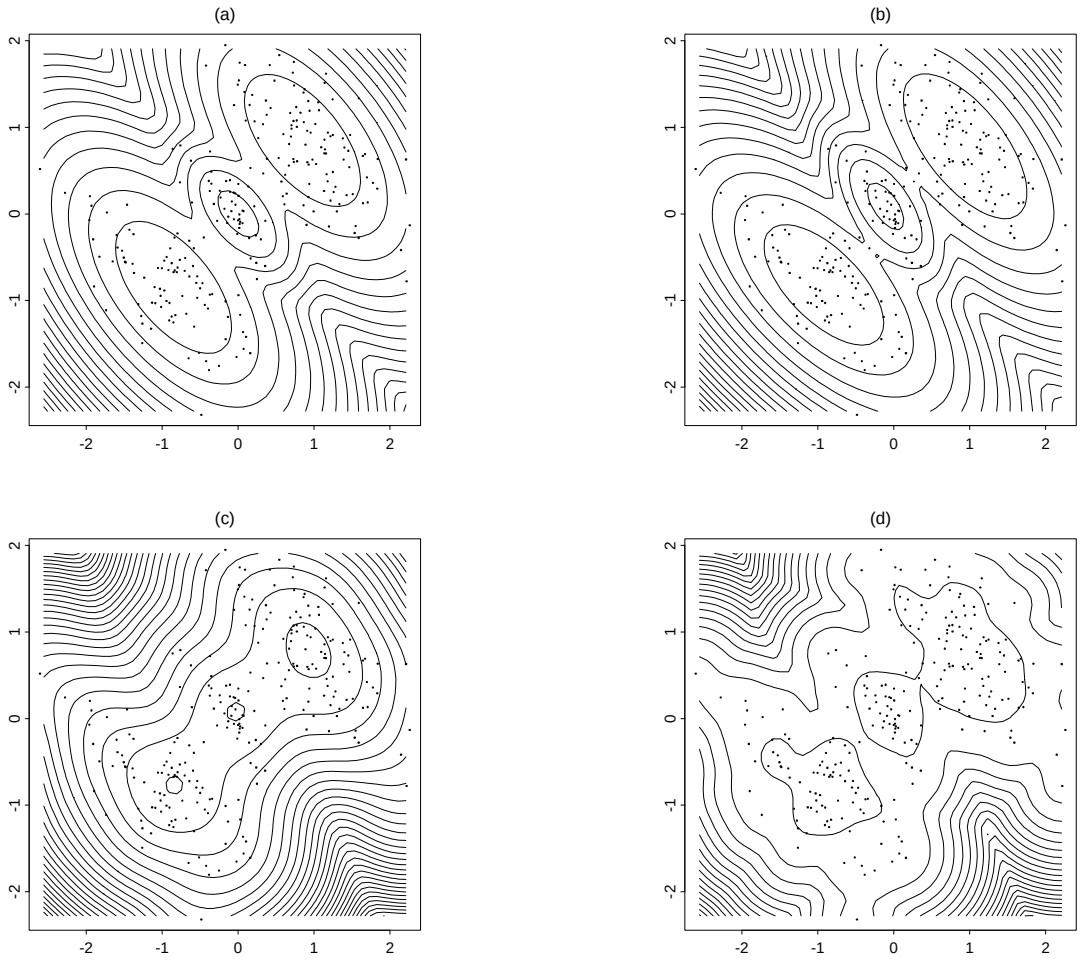


Figure 10: Comparative density estimates for trimodal MVN data. ($a$) Density used to generate the data. ($b$) Density estimate via model-based clustering. ($c$) Gaussian kernel density estimate with normal optimal bandwidth. ($d$) Gaussian kernel density estimate with cross-validated bandwidth. There are 250 data points, superimposed on the contours to show their location.

Table 1: MISE for Density Estimation via Model-Based Clustering (MBC), Gaussian Kernel Density Estimation with Normal Optimal Bandwdith (NOB), and Gaussian Kernel Density Estimation with Cross-validated Bandwidth (CVB). The numbers shown are the ratios of the MISEs for NOB and CVB respectively to that for MBC.

| 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 | 0.7 |
| Average | 6.6 | 2.9 |

NOTE: For the `Strongly Skewed` model, the CVB result is averaged over 42 of the 50 replicates, since in the remaining 8 instances, the cross-validated bandwidth could not be computed with default parameters in the `sm` software.

# 9 Model-based Clustering Software

The `MCLUST` software (Fraley and Raftery 1999), implementing model-based clustering and discriminant analysis as described this paper, is available through the Internet at

`http://www.stat.washington.edu/fraley/mclust`.

It is designed to interface with the commercial interactive software package S-PLUS[1].

Other software packages for model-based clustering include `EMMIX` (McLachlan et al. 1999) and `AutoClass` (Cheesman and Stutz 1995). Software for mixture discriminant analysis (MDA) and some of its generalizations is also available (see Hastie and Tibshirania 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

To date, the clustering methods based on multivariate normal mixture models we have described in this paper have been used with success in applications including minefield and

[1]MathSoft, Inc., Seattle, WA USA — `http://www.mathsoft.com/splus`

seismic fault detection (Dasgupta and Raftery, 1998), identifying flaws in textiles from images (Campbell et al.,1997, 1999), and the classification of astronomical data (Mukherjee et al., 1998). However, their practical use without modification can be limited for non-Gaussian, high-dimensional, or large data sets.

## 10.1 Non-Gaussian Data

Multivariate normal mixtures can accomodate data of varying structure. 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. Dasguspta and Raftery 1998; Fraley and Raftery 1998). For example if one component is concentrated about a nonlinear curve, 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 (Section 8.2), observations identified as not being mines were located in several groups in the model-based classification, while the true mines were confined to a single mixture component. An explicit approach to the problem of clusters that are 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 (Section 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 et al. 2000), and mixtures of distributions for angular data (Peel et al. 2000).

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 that are

generated, for example, by tracking visits to Web sites. 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, the covariance estimates in the ellipsoidal models will often be singular, causing the EM algorithm to break down, although the spherical and possibly 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 for example in the COBRA minefield detection problem of Section 8.2 or in the gamma ray bursts analyzed in 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 et al. 1991). A satisfactory solution remains a major research challenge, although new model-based multidimensional scaling techniques (e.g. Oh and Raftery 2000) may help bring the benefits of model-based clustering to this setting.

## 10.3 Large Data Sets

One reason for the current explosion of interest in clustering is the desire to use it for finding patterns in very large data sets, sometimes called "datamining". Model-based clustering as described in this paper does not scale to large data sets 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 there may not be adequate memory available for processing large data sets, memory requirements can be reduced if some of the observations can be grouped together in advance. Posse (2000) proposed the use of the minimum spanning tree to obtain initial partitions for hierarchical agglomeration for large data sets.

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 data set 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 MRI 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 one is seeking small groups in very large data sets. Small groups may not be represented at all in a sample, or else 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 two billion objects, which they solved by *iterated sampling* (see Section 5.3). The problem could also be approached via a modification of the simple sampling method. One version of this would be as follows. 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 data set. Select out the observations $i$ such that $f_i$ is below some threshold, i.e. those that are not well represented by any of the clusters identified so far. Form a second sample, including all 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). 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 $\tau_k$. 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 easily be held in memory, increases greatly when this is not the case. In this context, there has been considerable work on computational techniques for making the EM algorithm more efficient when applied

to large data sets (Bradley et al. 1998; Moore and Lee 1998; Moore 1999; Thiesson et al. 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 data sets as a whole, numerical error due to finite precision arithmetic would remain an obstacle. This limitation favors the traditional approach we have mentioned, that of 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 that 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 data sets (Meila 1999; Posse 2000; Smyth 2000; Cadez et al. 2000).

## 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 may wish to use Bayesian methods for estimation too, 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 proba-

bilities of belonging to groups. From the EM algorithm it is easy to calculate approximate posterior probabilities conditional on the maximum likelihood estimators of the model parameters, and the error in doing this typically declines to zero quickly, at rate $O(n^{-1/2})$. However, because this ignores the uncertainty in the parameter estimates, it is likely to underestimate the overall uncertainty, and so to bias estimated posterior probabilities towards greater certainty, i.e. towards 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 maximum likelihood estimator, as suggested by Dempster et al. (1977). This is likely to go a long way towards alleviating the first and most important of the three problems mentioned, although it will not solve the first two.

The problem of specifying the prior remains. If informative prior information is available, this should be used. If not, it would be desirable to have an easy way of specifying a prior, and 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, may be useful for estimation also, 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 there has been a great deal of work on Bayesian estimation of mixture models using Markov chain Monte Carlo. 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 iteration. 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 it has proved difficult to apply it to multivariate mixtures such as those that arise in clustering. 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, for example, by Richardson and Green (1997). This arises because one can switch the labeling of the mixture components without changing the likelihood. Since there are $G!$ labellings, 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 consisted of ordering the components *a priori* in some way (e.g. Celeux et al. 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 seem much more promising (Celeux 1998; Celeux, Hurn and Robert 1999; Stephens 1997, 2000). These consist basically of clustering the MCMC output itself according to the apparent labeling in operation, and 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 the application of model-based clustering itself to this "meta-problem" might be useful.

# References


Allard, D. and C. Fraley (1997). Nonparametric maximum likelihood estimation of features in spatial point processes using Voronoï tessellation. *Journal of the American Statistical Association 92*, 1485–1493.

Banfield, J. D. and A. E. Raftery (1992). Ice floe identification in satellite images using mathematical morphology and clustering about principle curves. *Journal of the American Statistical Association 87*, 7–16.

Banfield, J. D. and A. E. Raftery (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics 49*, 803–821.

Becker, M. P. and I. Yang (1998). Latent class marginal models for cross-classifications of counts. *Sociological Methodology 28*, 293–326.

Bensmail, H. and G. Celeux (1996). Regularized Gaussian discriminant analysis through eigenvalue decomposition. *Journal of the American Statistical Association 91*, 1743–1748.

Bensmail, H., G. Celeux, A. E. Raftery, and C. P. Robert (1997). Inference in model-based cluster analysis. *Statistics and Computing 7*, 1–10.

Bentley, J. L., K. L. Clarkson, and D. B. Levine (1993). Fast linear expected-time algorithms for computing maxima and convex hulls. *Algorithmica 9*, 168–183.

Biernacki, C., G. Celeux, and G. Govaert (1999). An improvement of the NEC criterion for assessing the number of clusters in a mixture model. *Pattern Recognition Letters 20*, 267–272.

Biernacki, C., G. Celeux, and G. Govaert (2000). Assessing a mixture model for clustering with the integrated complete likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. (to appear).

Biernacki, C. and G. Govaert (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.

Bock, H. H. (1998a). Probabilistic approaches in cluster analysis. *Bulletin of the International Statistical Institute 57*, 603–606.

Bock, H. H. (1998b). Probabilistic aspects in classification. In C. Hayashi, K. Yajima, H. H. Bock, N. Oshumi, Y. Tanaka, and Y. Baba (Eds.), *Data science, classification and related methods*, pp. 3–21. Springer-Verlag.

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

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

Boyles, R. A. (1983). On the convergence of the EM algorithm. *Journal of the Royal Statistical Society, Series 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 E. Diday,

Y. Lechevallier, M. Schader, P. Bertrand, and B. Burtschy (Eds.), *New Approcahes in Classification and Data Analysis*, pp. 169–177. Springer-Verlag.

Bradley, P. S., U. Fayyad, and C. Reina (1998). Scaling EM (expectation-maximization) clustering to large databases. Technical Report MSR-TR-98-35, Microsoft Research.

Byers, S. D. and A. E. Raftery (1998). Nearest neighbor clutter removal for estimating features in spatial point processes. *Journal of the American Statistical Association 93*, 577–584.

Cadez, I., D. Heckerman, C. Meek, P. Smyth, and S.White (2000). Visualization of navigation patterns on a web site using model-based clustering. Technical Report MSR-TR-2000-18, Microsoft Research.

Campbell, J. G., C. Fraley, F. Murtagh, and A. E. Raftery (1997). Linear flaw detection in woven textiles using model-based clustering. *Pattern Recognition Letters 18*, 1539–1548.

Campbell, J. G., C. Fraley, D. Stanford, F. Murtagh, and A. E. Raftery (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 R. Payne and P. Green (Eds.), *COMPSTAT 1998*, pp. 227–232. Physica-Verlag.

Celeux, G., D. Chaveau, and J. Diebolt (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 G. Govaert (1992). A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis 14*, 315–332.

Celeux, G. and G. Govaert (1993). Comparison of the mixture and the classification maximum likelihood in cluster analysis. *Journal of Statistical Computation and Simulation 47*, 127–146.

Celeux, G. and G. Govaert (1995). Gaussian parsimonious clustering models. *Pattern Recognition 28*, 781–793.

Celeux, G., M. Hurn, and C. Robert (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association 95*,

957–970.

Celeux, G. and G. Soromenho (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 J. Stutz (1995). Bayesian classification (AutoClass): Theory and results. In U. Fayyad, G. Piatesky-Shapiro, P. Smyth, and R. Uthurusamy (Eds.), *Advances in Knowledge Discovery and Data Mining*, pp. 153–180. AAAI Press.

Chickering, D. M. and D. Heckerman (1997). Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables. *Machine Learning 29*, 181–212.

Clogg, C. C. and L. A. Goodman (1984). Latent structure analysis of a set of multidimensional contingency tables. *Journal of the American Statistical Association 79*, 762–771.

Dasgupta, A. and A. E. Raftery (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., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood for incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B 39*, 1–38.

DeSarbo, W. S., D. J. Howard, and K. Jedidi (1991). MULTICLUS: A new method for simultaneously performing multidimensional scaling and cluster analysis. *Psychometrika 56*, 121–136.

Diebolt, J. and C. Robert (1994). Estimation of finite mixtures through Bayesian sampling. *Journal of the Royal Statistical Society, Series B 56*, 363–375.

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

Edwards, A. W. F. and L. L. Cavalli-Sforza (1965). A method for cluster analysis. *Biometrics 21*, 362–375.

Escobar, M. D. and M. West (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association 90*, 1301–1312.

Fayyad, U. and P. Smyth (1996). From massive data sets to science catalogs: applications and challenges. In J. Kettenring and D. Pregibon (Eds.), *Statistics and Massive Data*

*Sets: Report to the Committee on Applied and Theoretical Statistics*. National Research Council.

Fraley, C. (1998). Algorithms for model-based Gaussian hierarchical clustering. *SIAM Journal on Scientific Computing 20*, 270–281.

Fraley, C. and A. E. Raftery (1998). How many clusters? Which clustering method? - Answers via model-based cluster analysis. *The Computer Journal 41*, 578–588.

Fraley, C. and A. E. Raftery (1999). MCLUST: Software for model-based cluster analysis. *Journal of Classification 16*, 297–306.

Friedman, H. P. and J. Rubin (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.

Green, P. J. (1995). Reversible jump MCMC computation and Bayesian model determination. *Biometrika 82*, 711–732.

Hartigan, J. A. and M. A. Wong (1978). Algorithm AS 136 : A $k$-means clustering algorithm. *Applied Statistics 28*, 100–108.

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

Hastie, T. and R. Tibshirani (1996). Discriminant analysis by Gaussian mixtures. *Journal of the Royal Statistical Society, Series B 58*, 155–176.

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.

Jeffreys, H. (1961). *Theory of Probability* (3rd ed.). Clarendon.

Jöreskog, K. G. (1973). A general method for estimating a linear structural equation system. In A. S. Goldberger and O. D. Duncan (Eds.), *Structural Equation Models in the Social Sciences*, pp. 85–112. Seminar Press.

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

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

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

Kass, R. E. and L. Wasserman (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association 90*, 928–934.

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.). Springer.

Lavine, M. and M. West (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 E. A. Schulman, P. F. Lazarsfeld, S. A. Starr, and J. A. Clausen (Eds.), *Studies in Social Psychology in World War II. Vol. 4: Measurement and Prediction*, pp. 362–412. Princeton University Press.

Lazarsfeld, P. F. and N. W. Henry (1986). *Latent Structure Analysis*. Houghton Mifflin.

Leroux, M. (1992). Consistent estimation of a mixing distribution. *The Annals of Statistics 20*, 1350–1360.

MacEachern, S. N. and P. Müller (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 L. M. L. Cam and J. Neyman (Eds.), *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, pp. 281–297. University of California Press.

Mangasarian, O. L., W. N. Street, and W. H. Wolberg (1995). Breast cancer diagnosis and prognosis via linear programming. *Operations Research 43*, 570–577.

Marron, J. S. and M. P. Wand (1992). Exact mean integrated squared error. *The Annals of Statistics 20*, 712–536.

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

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

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

McLachlan, G. J., D. Peel, K. E. Basford, and P. Adams (1999). The EMMIX software for the fitting of mixtures of normal $t$-components. *Journal of Statistical Software 4*. (on-line publication).

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

Mengersen, K. and C. P. Robert (1996). Testing for mixtures: A Bayesian entropic approach. In J. M. Bernardo, J. O. Berger, A. P. David, and A. F. M. Smith (Eds.), *Bayesian Statistics 5*, pp. 255–276. Oxford University Press.

Moore, A. (1999). Very fast mixture-model-based clustering using multiresolution kd-trees. In M. Kearns and D. Cohn (Eds.), *Advances Neural Information Processing Systems*, Volume 10.

Moore, A. and M. S. Lee (1998). Cached sufficient statistics for efficient machine learning with large datasets. *Journal of Artificial Intelligence Research 8*, 67–91.

Mukherjee, S., E. D. Feigelson, G. J. Babu, F. Murtagh, C. Fraley, and A. E. Raftery (1998). Three types of gamma ray bursts. *The Astrophysical Journal 508*, 314–327.

Müller, P., A. Erkanli, and M. West (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika 83*, 67–80.

Murtagh, F. and A. E. Raftery (1984). Fitting straight lines to point patterns. *Pattern Recognition 17*, 479–483.

Murtagh, F., J.-L. Starck, and M. W. Berry (2000). Overcoming the curse of dimensionality by means of the wavelet transform. *The Computer Journal 43*, 107–120.

Oh, M.-S. and A. E. Raftery (2000, September). Bayesian multidimensional scaling and choice of dimension. Technical Report 379, University of Washington, Department of Statistics.

Ormoneit, D. and V. Tresp (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 G. J. McLachlan (2000). Robust mixture modeling using the $t$-distribution. *Statistics and Computing*. (to appear).

Peel, D., W. J. Whitten, and G. J. McLachlan (2000). Fitting mixtures of Kent distributions to aid in joint set identification. *Journal of the American Statistical Association*. (to appear).

Posse, C. (2000). Hierarchical model-based clustering for large data sets. *Journal of Computational and Graphical Statistics*. (to appear).

Raftery, A. E. (1995). Bayesian model selection in social research (with discussion). *Sociological Methodology 25*, 111–193.

Raftery, A. E. (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 P. J. Green (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, Series B 59*, 731–758.

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

Roeder, K. and L. Wasserman (1997). Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association 92*, 894–902.

Sampson, P. D. and P. Guttorp (1992). Nonparametric estimation of nonstationary spatial covaraince structure. *Journal of the American Statistical Association 87*, 108–119.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*, 461–464.

Scott, A. J. and M. J. Symons (1971). Clustering methods based on likelihood ratio criteria. *Biometrics 27*, 387–397.

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

Smyth, P. (2000). Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and Computing 10*, 63–72.

Stanford, D. and A. E. Raftery (2000). Principal curve clustering with noise. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*, 601–609.

Stephens, M. (1997). Contribution to the discussion of Richardson and Green, 1997: on the Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, Series B 59*, 768–769.

Stephens, M. (2000). Dealing with label-switching in mixture models. *Journal of the Royal Statistical Society, Series B 62*, 795–809.

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

Titterington, D. M., A. F. Smith, and U. E. Makov (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley.

Ward, J. H. (1963). Hierarchical groupings to optimize an objective function. *Journal of the American Statistical Association 58*, 234–244.

Wolfe, J. H. (1963). Object cluster analysis of social areas. Master's thesis, University of California, Berkeley.

Wolfe, J. H. (1965). A computer program for the maximum-likelihood analysis of types. USNPRA Technical Bulletin 65-15, US Naval Personnel Research Activity, San Diego, CA.

Wolfe, J. H. (1967). NORMIX: Computational methods for estimating the parameters of multivariate normal mixture distributions. Technical Bulletin USNPRA SRM 68-2, US Naval Personnel Research Activity, San Diego, CA.

Wolfe, J. H. (1970). Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research 5*, 329–350.

Wu, C. F. J. (1983). On convergence properties of the EM algorithm. *The Annals of Statistics 11*, 95–103.