How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis [1]

C. Fraley and A. E. Raftery

Technical Report No. 329

Department of Statistics
University of Washington
Box 354322
Seattle, WA 98195-4322 USA

[1] Funded by the Office of Naval Research under contracts N00014-96-1-0192 and N00014-96-1-0330. Thanks go to Simon Byers for providing the `NNclean` denoising procedure.

Abstract

We consider the problem of determining the structure of clustered data, without prior knowledge of the number of clusters or any other information about their composition. Data are represented by a mixture model in which each component corresponds to a different cluster. Models with varying geometric properties are obtained through Gaussian components with different parameterizations and cross-cluster constraints. Noise and outliers can be modeled by adding a Poisson process component. Partitions are determined by the EM (expectation-maximization) algorithm for maximum likelihood, with initial values from agglomerative hierarchical clustering.

Models are compared using an approximation to the Bayes factor based on the Bayesian Information Criterion (BIC); unlike significance tests, this allows comparison of more than two models at the same time, and removes the restriction that the models compared be nested. The problems of determining the number of clusters and the clustering method are solved simultaneously by choosing the best model. Moreover, the EM result provides a measure of uncertainty about the associated classification of each data point.

Examples are given, showing that this approach can give performance that is much better than standard procedures, which often fail to identify groups that are either overlapping or of varying sizes and shapes.

Contents

List of Tables

List of Figures

1 Introduction

We consider the problem of determining the intrinsic structure of clustered data when no information other than the observed values is available. This problem is known as *cluster analysis*, and should be distinguished from the related problem of *discriminant analysis*, in which known groupings of some observations are used to categorize others and infer the structure of the data as a whole.

Probability models have been proposed for quite some time as a basis for cluster analysis. In this approach, the data are viewed as coming from a mixture of probability distributions, each representing a different cluster. Recently, methods of this type have shown promise in a number of practical applications, including character recognition (Murtagh and Raftery [53]), tissue segmentation (Banfield and Raftery [7]), minefield and seismic fault detection (Dasgupta and Raftery [27]), identification of textile flaws from images (Campbell et al. [21]), and classification of astronomical data (Celeux and Govaert [24], Mukerjee et al. [51]).

Bayes factors, approximated by the Bayesian Information Criterion (BIC), have been applied successfully to the problem of determining the number of components in a model [27], [51] and for deciding which among two or more partitions most closely matches the data for a given model [21]. We describe a clustering methodology based on multivariate normal mixtures in which the BIC is used for direct comparison of models that may differ not only in the number of components in the mixture, but also in the underlying densities of the various components. Partitions are determined (as in [27]) by a combination of hierarchical clustering and the EM (expectation-maximization) algorithm (Dempster, Laird and Rubin [28]) for maximum likelihood. This approach can give much better performance than existing methods. Moreover, the EM result also provides a measure of uncertainty about the resulting classification. Figure 1 shows an example in which model-based classification is able to match the clinical classification of a biomedical data set much more closely than single-link (nearest-neighbor) or standard k-means, in the absence of any training data.

This paper is organized as follows. In Section 2, we give the necessary background in multivariate cluster analysis, including discussions of probability models, the EM algorithm for clustering and approximate Bayes factors. The basic model-based strategy and modifications for handling noise are described in Sections 2.5 and 2.6, respectively. A detailed analysis of the multivariate data set shown in Figure 1 is given in Section 3.1, followed by an example from minefield detection in the presence of noise in Section 3.2. Information on available software for the various procedures used in this approach is given in Section 4. A final section summarizes and indicates extensions to the method.

2 Model-Based Cluster Analysis

2.1 Cluster Analysis Background

By *cluster analysis* we mean the partitioning of data into meaningful subgroups, when the number of subgroups and other information about their composition may be unknown; good introductions include Hartigan [36], Gordon [35], Murtagh [52], McLachlan and Basford [46], and Kaufman and Rousseeuw [42]. Clustering methods range from those that are largely

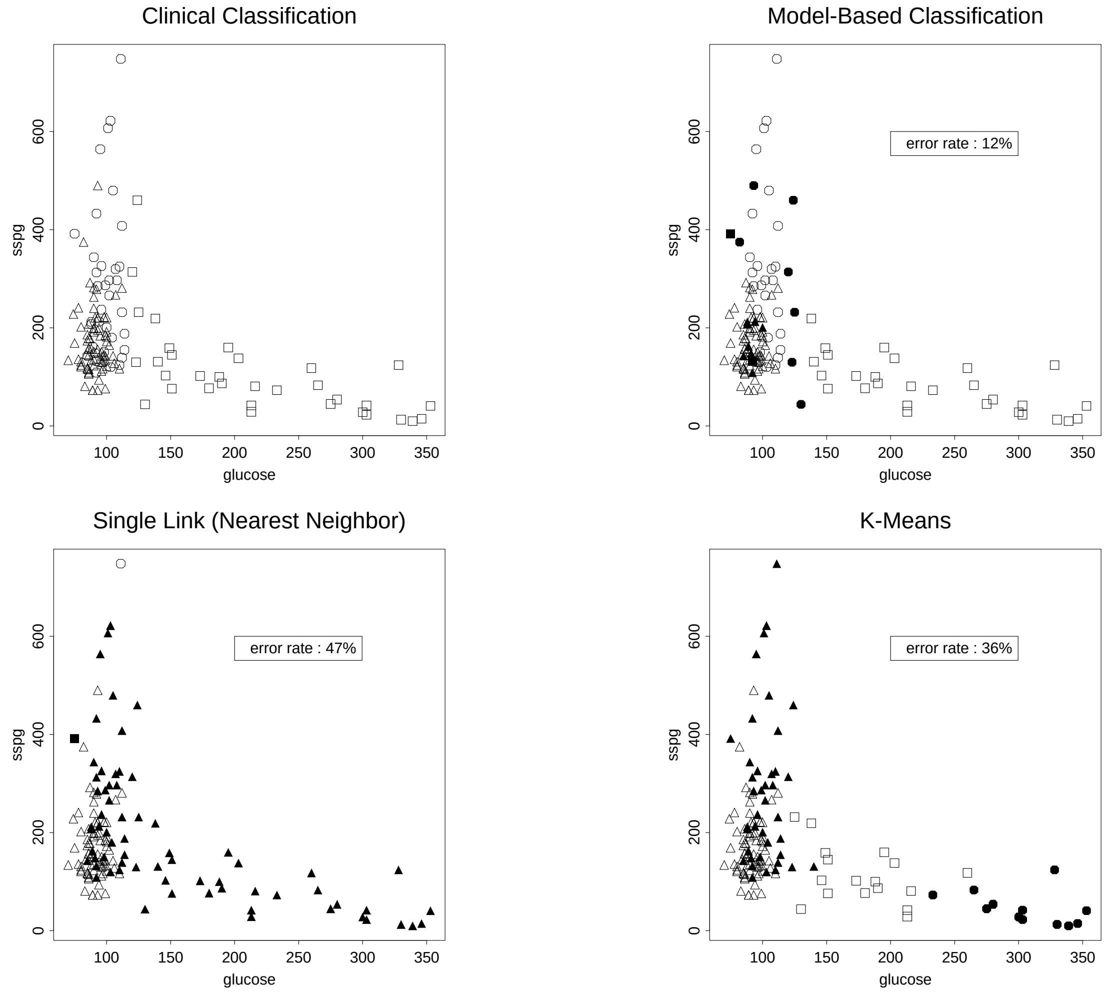

Figure 1: A projection of the three-group classification of the diabetes data from Reaven and Miller [56] using single link or nearest neighbor, standard k-means, and the unconstrained model-based approach. Filled symbols represent misclassified observations.

heuristic to more formal procedures based on statistical models. They usually follow either a hierarchical strategy, or one in which observations are relocated among tentative clusters.

Hierarchical methods proceed by stages producing a sequence of partitions, each corresponding to a different number of clusters. They can be either 'agglomerative', meaning that groups are merged, or 'divisive', in which one or more groups are split at each stage. Hierarchical procedures that use subdivision are not practical unless the number of possible splittings can somehow be restricted. In agglomerative hierarchical clustering, however, the number of stages is bounded by the number of groups in the initial partition. It is common practice to begin with each observation in a cluster by itself, although the procedure could be initialized from a coarser partition if some groupings are known. A drawback of agglomerative methods is that those that are practical in terms of time efficiency require memory usage proportional to the square of the number of groups in the initial partition.

At each stage of hierarchical clustering, the splitting or merging is chosen so as to optimize some criterion. Conventional agglomerative hierarchical methods use heuristic criteria, such as single link (nearest neighbor), complete link (farthest neighbor), or sum of squares [42]. In model-based methods, a maximum-likelihood criterion is used for merging groups [53, 7].

Relocation methods move observations iteratively from one group to another, starting from an initial partition. The number of groups has to be specified in advance and typically does not change during the course of the iteration. The most common relocation method — k-means (MacQueen [44], Hartigan and Wong [37]) — reduces the within-group sums of squares. For clustering via mixture models, relocation techniques are usually based on the EM algorithm [28] (see section 2.3).

Neither hierarchical nor relocation methods directly address the issue of determining the number of groups within the data. Various strategies for simultaneous determination of the number of clusters and cluster membership have been proposed (e. g. Engelman and Hartigan [31], Bock [12], Bozdogan [17] — for a survey see Bock [13]). An alternative is described in this paper.

2.2 Probability Models for Cluster Analysis

In model-based clustering, it is assumed that the data are generated by a mixture of underlying probability distributions in which each component represents a different group or cluster. Given observations $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_n)$, let $f_k(\mathbf{x}_i \mid \theta_k)$ be the density of an observation $\mathbf{x}_i$ from the kth component, where θ_k are the corresponding parameters, and let G be the number of components in the mixture. The model for the composite of the clusters is usually formulated in one of two ways. The *classification likelihood* approach maximizes

$$\mathcal{L}_C(\theta_1, \ldots, \theta_G; \gamma_1, \ldots, \gamma_n \mid \mathbf{x}) = \prod_{i=1}^{n} f_{\gamma_i}(\mathbf{x}_i \mid \theta_{\gamma_i}), \tag{1}$$

where the γ_i are discrete values labeling the classification : $\gamma_i = k$ if $\mathbf{x}_i$ belongs to the kth component. The *mixture likelihood* approach maximizes

$$\mathcal{L}_M(\theta_1, \ldots, \theta_G; \tau_1, \ldots, \tau_G \mid \mathbf{x}) = \prod_{i=1}^{n} \sum_{k=1}^{G} \tau_k f_k(\mathbf{x}_i \mid \theta_k), \tag{2}$$

where τ_k is the probability that an observation belongs to the kth component ($\tau_k \geq 0$; $\sum_{k=1}^{G} \tau_k = 1$).

We are mainly concerned with the case where $f_k(\mathbf{x}_i \mid \theta_k)$ is multivariate normal (Gaussian), a model that has been used with considerable success in a number of applications [53, 7, 24, 27, 21, 51]. In this instance, the parameters θ_k consist of a mean vector μ_k and a covariance matrix Σ_k, and the density has the form

$$f_k(\mathbf{x}_i \mid \mu_k, \Sigma_k) = \frac{\exp\left\{-\frac{1}{2}(\mathbf{x}_i - \mu_k)^T \Sigma_k^{-1}(\mathbf{x}_i - \mu_k)\right\}}{(2\pi)^{\frac{p}{2}} |\Sigma_k|^{\frac{1}{2}}}. \tag{3}$$

Clusters are ellipsoidal, centered at the means μ_k. The covariances Σ_k determine their other geometric characteristics.

Banfield and Raftery [7] developed a model-based framework for clustering by parameterizing the covariance matrix in terms of its eigenvalue decomposition in the form

$$\Sigma_k = \lambda_k D_k A_k D_k^T, \tag{4}$$

where D_k is the orthogonal matrix of eigenvectors, A_k is a diagonal matrix whose elements are proportional to the eigenvalues of Σ_k, and λ_k is a scalar. The orientation of the principal components of Σ_k is determined by D_k, while A_k determines the shape of the density contours; λ_k specifies the volume of the corresponding ellipsoid, which is proportional to $\lambda_k^p |A_k|$.[1] Characteristics (orientation, volume and shape) of distributions are usually estimated from the data, and can be allowed to vary between clusters, or constrained to be the same for all clusters.

This approach subsumes several earlier proposals based on Gaussian mixtures: $\Sigma_k = \lambda I$ gives the sum of squares criterion, long known as a heuristic (Ward [65]), in which clusters are spherical and have equal volumes; $\Sigma_k = \Sigma = \lambda DAD^T$, in which all clusters have the same shape, volume and orientation (Friedman and Rubin [33]); unconstrained $\Sigma_k = \lambda_k D_k A_k D_k^T$, which is the most general model (Scott and Symons [60]); and $\Sigma_k = \lambda D_k A D_k$ (Murtagh and Raftery [53]), in which only the orientations of the clusters may differ. Table 1 shows the geometric interpretation of the various parameterizations discussed in [7]. A more extensive set of models within the same framework is treated in [24].

Table 1: Parameterizations of the covariance matrix Σ_k in the Gaussian model and their geometric interpretation. The models shown here are those discussed in Banfield and Raftery [7].

Σ_k	Distribution	Volume	Shape	Orientation	Reference
λI	Spherical	equal	equal	NA	[65, 53, 7, 24]
$\lambda_k I$	Spherical	variable	equal	NA	[7, 24]
λDAD	Ellipsoidal	equal	equal	equal	[33, 60, 7, 24]
$\lambda_k D_k A_k D_k$	Ellipsoidal	variable	variable	variable	[60, 7, 24]
$\lambda D_k A D_k$	Ellipsoidal	equal	equal	variable	[53, 7, 24]
$\lambda_k D_k A D_k$	Ellipsoidal	variable	equal	variable	[7, 24]

The classification likelihood can be used as the basis for agglomerative hierarchical clustering [53], [7]. At each stage, a pair of clusters is merged so as to maximize the resulting likelihood. Fraley [32] developed efficient algorithms for hierarchical clustering with the various parameterizations (4) of Gaussian mixture models.

2.3 EM Algorithms for Clustering

Iterative relocation methods for clustering via mixture models are possible through EM and related techniques [46]. The EM algorithm [28, 47] is a general approach to maximum

[1] Conventions for normalizing λ_k and A_k include requiring $|A_k| = 1$ [24], so that $\lambda_k = |\Sigma_k|^{1/p}$, or else requiring $\max(A_k) = 1$ [7], so that λ_k is the largest eigenvalue of Σ_k.

likelihood in the presence of incomplete data. In EM for clustering, the “complete” data are considered to be $y_i = (\mathbf{x}_i, \mathbf{z}_i)$, where $\mathbf{z}_i = (z_{i1}, \ldots, z_{iG})$ with

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

constitutes the “missing” data. The relevant assumptions are that the density of an observation $\mathbf{x}_i$ given $\mathbf{z}_i$ is given by $\prod_{k=1}^{G} f_k(\mathbf{x}_i \mid \theta_k)^{z_{ik}}$ and that each $\mathbf{z}_i$ is independent and identically distributed according to a multinomial distribution of one draw on G categories with probabilities $\tau_1, \ldots, \tau_G$. The resulting *complete-data loglikelihood* is

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

The quantity $\hat{z}_{ik} = E[z_{ik} | \mathbf{x}_i, \theta_1, \ldots, \theta_G]$ for the model (6) is the conditional expectation of z_{ik} given the observation $\mathbf{x}_i$ and parameter values. The value z^*_{ik} of $\hat{z}_{ik}$ at a maximum of (2) is the conditional probability that observation i belongs to group k; the classification of an observation $\mathbf{x}_i$ is taken to be $\{j \mid z^*_{ij} = \max_k z^*_{ik}\}$.

The EM algorithm iterates between an E-step in which values of $\hat{z}_{ik}$ are computed from the data with the current parameter estimates, and an M-step in which the complete-data loglikelihood (6), with each z_{ik} replaced by its current conditional expectation $\hat{z}_{ik}$, is maximized with respect to the parameters (see Figure 2). Celeux and Govaert [24] detail both the E and M steps for the case of multivariate normal mixture models parameterized via the eigenvalue decomposition in (4). Under certain conditions (Boyles [16], Wu [66], McLachlan and Krishnan [47]), the method can be shown to converge to a local maximum of the mixture likelihood (2). Although the conditions under which convergence has been proven do not always hold in practice, the method is widely used in the mixture modeling context with good results. Moreover, for each observation i, $(1 - \max_k z^*_{ik})$ is a measure of uncertainty in the associated classification (Bensmail et al. [9]).

The EM algorithm for clustering has a number of limitations. First, the rate of convergence can be very slow. This does not appear to be a problem in practice for well-separated mixtures when started with reasonable values. Second, the number of conditional probabilities associated with each observation is equal to the number of components in the mixture, so that the EM algorithm for clustering may not be practical for models with very large numbers of components. Finally, EM breaks down when the covariance matrix corresponding to one or more components becomes ill-conditioned (singular or nearly singular). In general it cannot proceed if clusters contain only a few observations or if the observations they contain are very nearly colinear. If EM for a model having a certain number of components is applied to a mixture in which there are actually fewer groups, then it may fail due to ill-conditioning.

A number of variants of the EM algorithm for clustering presented above have been studied. These include the *stochastic EM* or SEM algorithm (Broniatowski, Celeux and Diebolt [18], Celeux and Diebolt [22]), in which the $\hat{z}_{ik}$ are simulated rather than estimated in the E-step, and the *classification EM* or CEM algorithm (Celeux and Govaert [23]), which converts the $\hat{z}_{ik}$ from the E-step to a discrete classification before performing the M-step. The standard k-means algorithm can be shown to be a version of the CEM algorithm corresponding to the uniform spherical Gaussian model $\Sigma_k = \lambda I$ [23].

initialize $\hat{z}_{ik}$ (this can be from a discrete classification (5))

repeat

M-step: maximize (6) given $\hat{z}_{ik}$ (f_k as in (3)

$$\begin{aligned} n_k &\leftarrow \textstyle\sum_{i=1}^{n} \hat{z}_{ik} \\ \hat{\tau}_k &\leftarrow \frac{n_k}{n} \\ \hat{\mu}_k &\leftarrow \frac{\sum_{i=1}^{n} \hat{z}_{ik}\mathbf{x}_i}{n_k} \\ \hat{\Sigma}_k &: \text{depends on the model — see Celeux and Govaert [24]} \end{aligned}$$

E-step: compute $\hat{z}_{ik}$ given the parameter estimates from the M-step

$$\hat{z}_{ik} \leftarrow \frac{\hat{\tau}_k f_k(\mathbf{x}_i \mid \hat{\mu}_k, \hat{\Sigma}_k)}{\sum_{j=1}^{G} \hat{\tau}_j f_j(\mathbf{x}_i \mid \hat{\mu}_j, \hat{\Sigma}_j)}, \quad \text{where } f_k \text{ has the form (3).}$$

until convergence criteria are satisfied

Figure 2: EM algorithm for clustering via Gaussian mixture models. The strategy described in this paper initializes the iteration with indicator variables (5) corresponding to partitions from hierarchical clustering, and terminates when the relative difference between successive values of the mixture loglikelihood falls below a small threshold.

2.4 Bayesian Model Selection in Clustering

One advantage of the mixture-model approach to clustering is that it allows the use of approximate Bayes factors to compare models. This gives a systematic means of selecting not only the parameterization of the model (and hence the clustering method), but also the number of clusters. For a recent review of Bayes factors emphasizing the underlying concepts and scientific applications, see Kass and Raftery [41].

The Bayes factor is the posterior odds for one model against the other assuming neither is favored a priori. Banfield and Raftery [7] used a heuristically derived approximation to twice the log Bayes factor called the 'AWE' to determine the number of clusters in hierarchical clustering based on the classification likelihood. When EM is used to find the maximum mixture likelihood, a more reliable approximation to twice the log Bayes factor called the Bayesian Information Criterion or 'BIC' (Schwarz [59]) is applicable:

$$2 \log p(x|\mathcal{M}) + \text{const.} \approx 2l_{\mathcal{M}}(x, \hat{\theta}) - m_{\mathcal{M}} log(n) \equiv \text{BIC},$$

where $p(x|\mathcal{M})$ is the (integrated) likelihood of the data for the model $\mathcal{M}$, $l_{\mathcal{M}}(x, \hat{\theta})$ is the maximized mixture loglikelihood for the model, and $m_{\mathcal{M}}$ is the number of independent parameters to be estimated in the model. The number of clusters is not considered an independent parameter for the purposes of computing the BIC. If each model is equally likely a priori, then $p(x|\mathcal{M})$ is proportional to the posterior probability that the data conform to the model $\mathcal{M}$. Accordingly, the larger the value of the BIC, the stronger the evidence for the model.[2]

The fit of a mixture model to a given data set can only improve (and the likelihood can only increase) as more terms are added to the model. Hence likelihood cannot be used directly in assessment of models for cluster analysis. In the BIC, a term is added to the loglikelihood penalizing the complexity of the model, so that it may be maximized for more parsimonious parameterizations and smaller numbers of groups than the loglikelihood. The BIC can be used to compare models with differing parameterizations, differing numbers of components, or both. Bayesian criteria other than the BIC have been used in cluster analysis (e. g. Bock [12], Binder [11]). Although regularity conditions for the BIC do not hold for mixture models, there is considerable theoretical and practical support for its use in this context [43, 58, 27, 21, 51].

A standard convention for calibrating BIC differences is that differences of less than 2 correspond to weak evidence, differences between 2 and 6 to positive evidence, differences between 6 and 10 to strong evidence, and differences greater than 10 to very strong evidence (Jeffreys [40], Kass and Raftery [41]).

2.5 Model-Based Strategy for Clustering

In practice, agglomerative hierarchical clustering based on the classification likelihood (1) with Gaussian terms often gives good, but suboptimal partitions. The EM algorithm can refine partitions when started sufficiently close to the optimal value. Dasgupta and Raftery

[2]Kass and Raftery [41] and other authors define the BIC to have the opposite sign as that given here, in which case the smaller (more negative) the BIC, the stronger the evidence for the model. We have chosen to reverse this convention in order to make it easier to interpret the plots of BIC values that we present later.

[27] were able to obtain good results in a number of examples by using the partitions produced by model-based hierarchical agglomeration as starting values for an EM algorithm for constant-shape Gaussian models, together with the BIC to determine the number of clusters. Their approach forms the basis for a more general model-based strategy for clustering:

- Determine a maximum number of clusters to consider (M), and a set of candidate parameterizations of the Gaussian model to consider. In general M should be as small as possible.
- Do agglomerative hierarchical clustering for the unconstrained Gaussian model,[3] and obtain the corresponding classifications for up to M groups.
- Do EM for each parameterization and each number of clusters $2, \ldots, M$, starting with the classification from hierarchical clustering.
- Compute the BIC for the one-cluster model for each parameterization, and for the mixture likelihood with the optimal parameters from EM for $2, \ldots, M$ clusters. This gives a matrix of BIC values corresponding to each possible combination of parameterization and number of clusters.
- Plot the BIC values for each model. A decisive first local maximum indicates strong evidence for a model (parameterization + number of clusters).

It is important to avoid applying this procedure to a larger number of components than necessary. One reason for this is to minimize computational effort; other reasons have been discussed in Section 2.3. A heuristic that works well in practice is to select the number of clusters corresponding to the first decisive local maximum (if any) over all the parameterizations considered. There may in some cases be local maxima giving larger values of BIC due to ill-conditioning rather than a genuine indication of a better model (for further discussion, see section 3.1).

2.6 Modeling Noise and Outliers

Although the model-based strategy for cluster analysis as described in Section 2.5 is not directly applicable to noisy data, the model can be modified so that EM works well with a reasonably good initial identification of the noise and clusters. Noise is modeled as a constant-rate Poisson process, resulting in the mixture likelihood

$$\tilde{\mathcal{L}}_M(\theta_1, \ldots, \theta_G; \tau_0, \tau_1, \ldots, \tau_G \mid \mathbf{x}) = \prod_{i=1}^n \left[\frac{\tau_0}{V} + \sum_{k=1}^{G} \tau_k f_k(\mathbf{x}_i \mid \theta_k) \right], \tag{7}$$

where V is the hypervolume of the data region, $\sum_{k=0}^{G} \tau_k = 1$, and each $f_k(\mathbf{x}_i \mid \theta_k)$ is multivariate normal. An observation contributes $1/V$ if it belongs to the noise; otherwise it contributes a Gaussian term.

[3] While there is a hierarchical clustering method corresponding to each parameterization of the Gaussian model, it appears to be sufficient in practice to use only the unconstrained model for initialization.

The basic model-based procedure for noisy data is as follows. First, it is necessary to obtain an initial estimate of the noise. Possible approaches to denoising include the nearest-neighbor method of Byers and Raftery [20] and the method of Allard and Fraley [1], which uses Voronoï tessellations. Next, hierarchical clustering is applied to the denoised data. In a final step, EM based on the augmented model (7) is applied to the entire data set with the Gaussian components initialized with the hierarchical clustering partitions, and the noise component initialized with the result of the denoising procedure. The BIC is then used to select the best model representing the data.

,

3 Examples

3.1 Diabetes Diagnosis

In this section we illustrate the model-based approach to clustering using a three-dimensional data set involving 145 observations used for diabetes diagnosis (Reaven and Miller [56]). Figure 3 is a pairs plot showing the clinical classification, which partitions the data into three groups. The variables have the following meanings:

`glucose` - plasma glucose response to oral glucose,
`insulin` - plasma insulin response to oral glucose,
`sspg` - degree of insulin resistance.

The clusters are overlapping and are far from spherical in shape. As a result, many clustering procedures would not work well for this application. For example, Figure 1 shows the $(1, 3)$ projection of three-cluster classifications obtained by the single-link (nearest-neighbor) method, standard k-means, and the model-based method for an unconstrained Gaussian mixture. Of the possible group assignments, those shown were chosen so as to minimize the error rate in each case. The assumption of three classes is artificial for single link and k-means, while for the model-based method the BIC was used to determine the number of groups (see below).

Neither standard k-means nor single link perform well in this example. Two of the clusters identified by single link are singletons, so that nearly all of the data are assigned to one class. While all three classes resulting from standard k-means are nontrivial, two of the classes are confined to one of the long thin extensions while the third class subsumes the other extension as well as their conjunction. In the clinical classification, each of the two long extensions roughly represents a cluster, while the third cluster is concentrated closer to the origin. Most clustering methods that are currently in common use work well when clusters are well separated, but many break down when clusters overlap or intersect.

It is important, however, to distinguish between single-link clustering and nearest-neighbor *discrimination*. In discrimination, there is a 'training set' of data whose group memberships are known in advance, while in clustering, all group memberships are unknown. Nearest-neighbor discrimination assigns a data point to the same group as the point in the training set nearest to it. It often works very well (e.g. Ripley [57]), but its success depends entirely on the available training set.

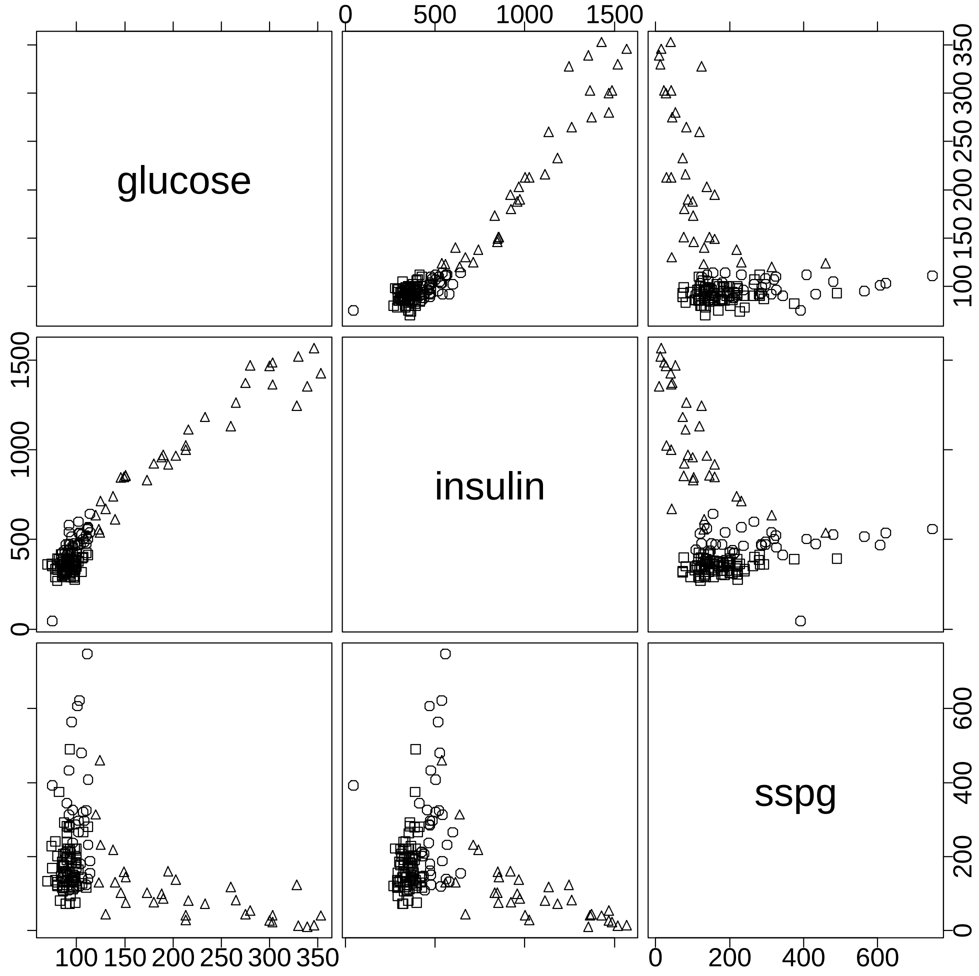

Figure 3: Pairs plot showing the clinical classification of the diabetes data. The symbols have the following interpretation: squares – normal; circles – chemical diabetes; triangles – overt diabetes.

Figure 4 gives a plot of the BIC for six model-based methods (spherical models with equal and varying volumes, constant variance, unconstrained variance, and constant shape models with equal and varying volumes). The first local maximum (in this case also the global maximum) occurs for the unconstrained model with three clusters, for which the classification assignment is shown in Figure 1. For initial values in EM, we used the z_{ik} given by equation (5) for the discrete classification from agglomerative hierarchical clustering for the unconstrained model $(\lambda_k D_k A_k D_k^T)$ in all cases, leaving the model selection to the EM phase.

Of note is that no values of the BIC are given in Figure 4 for the spherical, varying-volume model for 9 clusters and for the unconstrained model for 8 and 9 clusters. In these cases, the covariance matrix associated with one or more of the mixture components is ill-conditioned, so that the loglikelihood and hence the BIC cannot be computed. Hierarchical clustering for the spherical, varying-volume model produces a 9-cluster solution in which one cluster is a singleton, and for the unconstrained model it produces 8- and 9-cluster solutions in which one cluster contains three points. Because the data are three-dimensional, a minimum of four

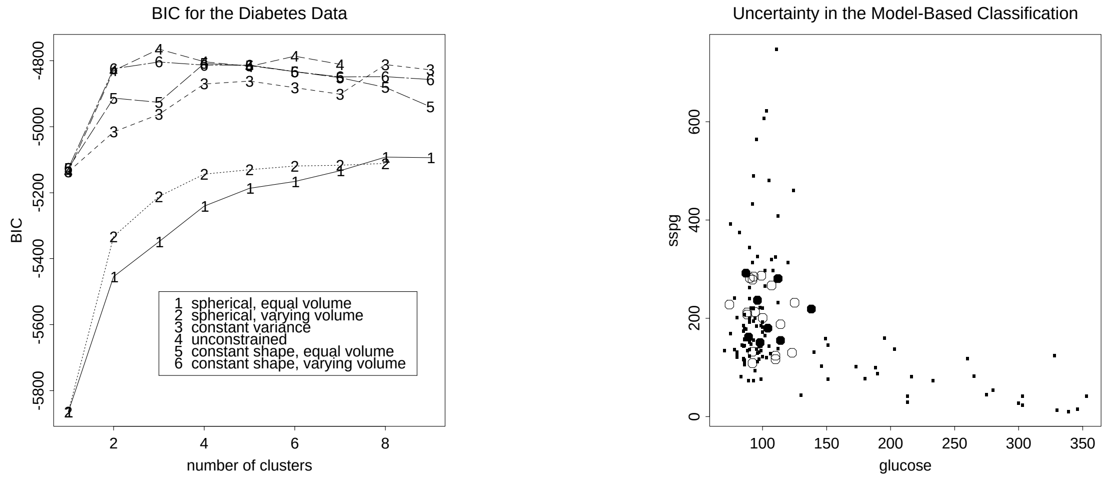

Figure 4: The plot on the left shows the Bayesian Information Criterion (BIC) for model-based methods applied to the diabetes data. The first local (also global) maximum occurs for the unconstrained model with three clusters. The plot on the right depicts the uncertainty of the classification produced by the best model (unconstrained, 3 clusters) indicated by the BIC. The symbols have the following interpretation: dots < 0.1; open circles ≥ 0.1 and < 0.2; filled circles ≥ 0.2.

points is required for the estimate of the covariance matrix to be nonsingular. The algorithms used for EM and for computing the BIC monitor an estimate of the reciprocal condition number (smallest to largest eigenvalue ratio) of the covariances. This latter quantity falls in the range $[0, 1]$, and values near zero imply ill-conditioning [34]. Computations are less reliable for ill-conditioned problems, and as a result ill-conditioning may cause anomalies before reaching the point of actual failure. In our implementation, EM terminates with a warning if one or more estimated covariance matrices are judged to be too close to singularity, and the BIC calculation produces a missing value under the same circumstances. Table 2 shows reciprocal condition estimates for six different Gaussian mixture models for up to 9 clusters. It should also be clear that EM started from partitions obtained by hierarchical clustering should not be continued for higher numbers of clusters once ill-conditioning is encountered.

3.2 Minefield Detection in the Presence of Noise

Figure 5 shows the results of the model-based strategy for noise (section 2.6) on simulated minefield data (Muise and Smith [50] — see also [27]). The data arise from the processing of a series of images taken by a reconnaissance aircraft in which a large number of points are identified as representing possible mines, but many of these are in fact false positives (noise). The assumption is that the imaged area does not lie completely within a minefield, and that if there is a minefield it will occur in an area where there is a higher density of identified points. The goals are to determine whether the image contains one or more minefields, and to give the location of any minefields that may be present.

The initial denoising for Figure 5 was carried out using the `NNclean` procedure for nearest-neighbor denoising [20]. The BIC is clearly maximized at a value of 3 (2 clusters plus

Table 2: Minimum reciprocal condition estimates for covariances in model-based methods applied to the diabetes data. Rows correspond to models and columns to numbers of components. Values near zero are cases in which there is either a very small cluster, or one whose points are very nearly colinear.

	1	2	3	4	5	6	7	8	9
λI	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$\lambda_k I$	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0
Σ	0.33	0.062	0.10	0.047	0.053	0.053	0.053	0.027	0.031
Σ_k	0.33	0.020	0.0060	0.0064	0.0067	10^{-7}	10^{-7}	10^{-32}	10^{-32}
$\lambda D_k A D_k^T$	0.0025	0.0048	0.0044	0.0049	0.0072	0.0070	0.0070	0.0017	0.0024
$\lambda_k D_k A D_k^T$	0.0025	0.0035	0.0070	0.0065	0.0065	0.0063	0.0046	0.0039	0.0027

noise), and favors the uniform-shape, equal-volume model. The two clusters together give an accurate reconstruction of the actual minefield.

It should be noted that the method is sensitive to the value of V, the assumed volume of the data region. Here it is clear that V is the area of the image; Banfield and Raftery [7] and Dasgupta and Raftery [27] similarly used the volume of the smallest hyperrectangle with sides parallel to the axes that contains all the data points. However, this value could overestimate V in many cases. Another possibility is to take V to be the smallest hyperrectangle with sides parallel to the principal components of the data that contains all the data points. Our implementation uses the smaller of these two alternatives as a default, but also allows specification of V by the user. A better solution might be to use the volume of the convex hull of the data, although this may not be practical to compute in higher dimensions.

4 Software

Software implementing state-of-the-art algorithms for hierarchical clustering [32] and EM based on the various parameterizations of Gaussian clustering models is available through the internet — for details see

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

Included are functions that incorporate hierarchical clustering, EM, and BIC in the model-based cluster analysis strategy described in this paper. This software is designed to interface with the commercial interactive software S-PLUS[4]. An earlier version of the model-based hierarchical clustering software is included in the S-PLUS package as the function `mclust`. Subscription information for a mailing list for occasional announcements such as software updates can also be found on the same web page.

An S-PLUS function `NNclean` implementing the nearest neighbor denoising method [20] is available at `http://lib.stat.cmu.edu/S/nnclean`.

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

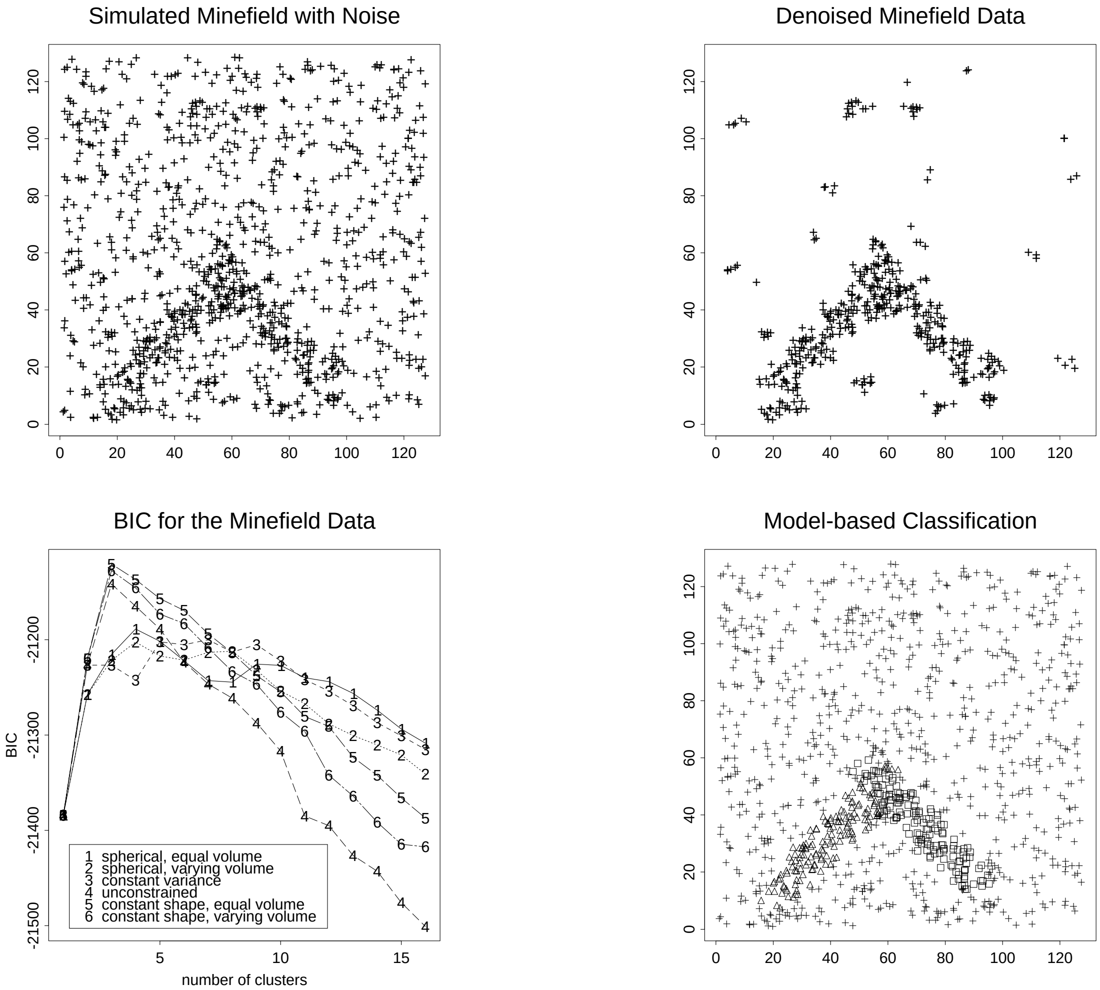

Figure 5: Model-based classification of a simulated minefield with noise. Hierarchical clustering was first applied to data after 5 nearest neighbor denoising. EM was then applied to the full data set with the noise term included in the model.

5 Discussion

We have described a clustering methodology based on multivariate normal mixture models and shown that it can give much better performance than existing methods. This approach uses model-based agglomerative hierarchical clustering to initialize the EM algorithm for a variety of models, and applies Bayesian model selection methods to determine the best clustering method along with the number of clusters. The uncertainty associated with the final classification can be assessed through the conditional probabilities from EM.

This approach has some limitations, however. The first is that computational methods for hierarchical clustering have storage and time requirements that grow at a faster than linear rate relative to the size of the initial partition, so that they cannot be directly applied to large data sets. One way around this is to determine the structure of some subset of the data according to the strategy given here, and either use the resulting parameters as initial values for EM with all of the data, or else classify the remaining observations via supervised classification or discriminant analysis [7]. Bensmail and Celeux [8] have developed a method for regularized discriminant analysis based on the full range of parameterizations of Gaussian mixtures (4). Alternatively, fast methods for determining an initial rough partition can be used to reduce computational requirements. Posse [55] suggested a method based on the minimum spanning tree for this purpose, and has shown that it works well in practice.

Second, although experience to date suggests that models based on the multivariate normal distribution are sufficiently flexible to accommodate many practical situations, the underlying assumption is that groups are concentrated locally about linear subspaces, so that other models or methods may be more suitable in some instances. In Section 3.2, we obtained good results on noisy data by combining the model-based methodology with a separate denoising procedure. This example also suggests that nonlinear features can in some instances be well represented in the present framework as piecewise linear features, using several groups. There are alterative models in which classes are characterized by different geometries such as linear manifolds (e. g. Bock [12], Diday [29], Späth [61]). When features are strongly curvilinear, curves about which groups are centered can be modeled by using *principal curves* (Hastie and Stuetzle [38]). Clustering about principal curves has been successfully applied to automatic identification of ice-floe contours [5, 6], tracking of ice floes [3], and modeling ice-floe leads [4]. Initial estimation of ice-floe outlines is accomplished by means of mathematical morphology (e.g. [39]). Principal curve clustering in the presence of noise using BIC is discussed in Stanford and Raftery [62].

In situations where the BIC is not definitive, more computationally intensive Bayesian analysis may provide a solution. Bensmail et al. [9] showed that exact Bayesian inference via Gibbs sampling, with calculations of Bayes factors using the Laplace-Metropolis estimator, works well in several real and simulated examples.

Approaches to clustering based on the classification likelihood (1) are also known as *classification maximum likelihood* methods (e. g. McLachlan [45], Bryant and Williamson [19]) or *fixed-classification* methods (e. g. Bock [14, 13, 15]). There are alternatives to the classification and mixture likelihoods given in section 2.2, such as the classification likelihood of Symons [64]

$$\mathcal{L}_C(\theta_1, \ldots, \theta_G; \tau_1, \ldots, \tau_G; \gamma_1, \ldots, \gamma_n \mid \mathbf{x}) \\ = \prod_{i=1}^{n} \tau_{\gamma_i} f_{\gamma_i}(\mathbf{x}_i \mid \theta_{\gamma_i}),$$

and the posterior likelihood of Anderson [2]

$$\mathcal{L}_P(\theta_1, \ldots, \theta_G; \tau_1, \ldots, \tau_G; z_{11}, z_{12}, \ldots, z_{nn} \mid \mathbf{x}) \\ = \prod_{k=1}^{G} \prod_{i=1}^{n} \tau_k^{z_{ik}} f(\mathbf{x}_i \mid \theta_k)^{z_{ik}}.$$

The former is the complete data likelihood for the EM algorithm when the z_{ik} are restricted to be indicator variables (5), while the later has the same form as the complete data likelihood for the EM algorithm, but includes the z_{ik} as parameters to be estimated. Fuzzy clustering methods (Bezdek [10]), which are not model-based, also provide degrees of membership for observations.

The k-means algorithm has been applied not only to the classical sum-of-squares criterion but also to other model-based clustering criterion (e. g. Bock [12, 13, 15], Diday and

Govaert [30], Diday [29], Späth [61], Celeux and Govaert [24]). Other model-based clustering methodologies include Cheeseman and Stutz [25, 63], implemented in the `AutoClass` software, and McLachlan et al. [46, 48, 54], implemented in the `EMMIX` (formerly `MIXFIT`) software. `AutoClass` handles both discrete data and continuous data, as well as data that has both discrete and continuous variables. Both `AutoClass` for continuous data and `EMMIX` rely on the EM algorithm for the multivariate normal distribution; `EMMIX` allows the choice of either equal, unconstrained, or diagonal covariance matrices, while in `Autoclass` the covariances are assumed to be diagonal. As in our approach, `AutoClass` uses approximate Bayes factors to choose the number of clusters (see also Chickering and Heckerman [26]), although their approximation differs from the BIC. `EMMIX` determines the number of clusters by resampling, and has the option of modeling outliers by fitting mixtures of multivariate t-distributions (McLachlan and Peel [49]). In `Autoclass`, EM is initialized using random starting values, the number of trials being determined through specification of a limit on the running time. Options for initializing EM in `EMMIX` include the most common heuristic hierarchical clustering methods, as well as k-means, whereas we use the model-based hierarchical clustering solution as an initial value.

References

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

[2] J. J. Anderson. Normal mixtures and the number of clusters problem. *Computational Statistics Quarterly*, 2:3–14, 1985.

[3] J. D. Banfield. Automated tracking of ice floes : A statistical approach. *IEEE Transactions on Geoscience and Remote Sensing*, 29(6):905–911, November 1991.

[4] J. D. Banfield. Skeletal modeling of ice leads. *IEEE Transactions on Geoscience and Remote Sensing*, 30(5):918–923, September 1992.

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

[6] J. D. Banfield and A. E. Raftery. Identifying ice floes in satellite images. *Naval Research Reviews*, 43:2–18, 1992.

[7] J. D. Banfield and A. E. Raftery. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49:803–821, 1993.

[8] H. Bensmail and G. Celeux. Regularized Gaussian discriminant analysis through eigenvalue decomposition. *Journal of the American Statistical Association*, 91:1743–1748, December 1996.

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

[10] J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum, 1981.

[11] D. A. Binder. Bayesian cluster analysis. *Biometrika*, 65:31–38, 1978.

[12] H. H. Bock. *Automatische Klassifikation (Clusteranalyse)*. Vandenhoek & Ruprecht, 1974.

[13] H. H. Bock. Probability models and hypotheis testing in partitioning cluster analysis. In P. Arabie, L. Hubert, and G. DeSorte, editors, *Clustering and Classification*, pages 377–453. World Science Publishers, 1996.

[14] H. H. Bock. Probability models in partitional cluster analysis. *Computational Statistics and Data Analysis*, 23:5–28, 1996.

[15] H. H. Bock. Probability models in partitional cluster analysis. In A. Ferligoj and A. Kramberger, editors, *Developments in Data Analysis*, pages 3–25. FDV, Metodoloski zvezki 12, Ljubljana, Slovenia, 1996.

[16] R. A. Boyles. On the convergence of the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 45:47–50, 1983.

[17] H. Bozdogan. Choosing the number of component clusters in the mixture model using a new informational complexity criterion of the inverse Fisher information matrix. In O. Opitz, B. Lausen, and R. Klar, editors, *Information and Classification*, pages 40–54. Springer-Verlag, 1993.

[18] M. Broniatowski, G. Celeux, and J. Diebolt. Reconnaissance de mélanges de densités par un algorithme d'apprentissage probabiliste. In E. Diday, M. Jambu, L. Lebart, J.-P. Pagès, and R. Tomassone, editors, *Data Analysis and Informatics, III*, pages 359–373. Elsevier Science, 1984.

[19] P. Bryant and A. J. Williamson. Maximum likelihood and classification : a comparison of three approaches. In W. Gaul and R. Schader, editors, *Classification as a Tool of Research*, pages 33–45. Elsevier Science, 1986.

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

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

[22] G. Celeux and J. Diebolt. The SEM algorithm : A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly*, 2:73–82, 1985.

[23] G. Celeux and G. Govaert. A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis*, 14:315–332, 1992.

[24] G. Celeux and G. Govaert. Gaussian parsimonious clustering models. *Pattern Recognition*, 28:781–793, 1995.

[25] P. Cheeseman and J. Stutz. Bayesian classification (AutoClass): Theory and results. In U. Fayyad, G. Piatesky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 153–180. AAAI Press, 1995.

[26] D. M. Chickering and D. Heckerman. Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables. *Machine Learning*, 29:181–244, 1997.

[27] A. Dasgupta and A. E. Raftery. Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association*, 93(441):294–302, March 1998.

[28] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood for incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

[29] E. Diday. *Optimisation en classification automatique*. INRIA (France), 1979.

[30] E. Diday and G. Govaert. Classification avec distances adaptives. *Comptes Rendues Acad. Sci. Paris, série A*, 278:993, 1974.

[31] L. Engelman and J. A. Hartigan. Percentage points of a test for clusters. *Journal of the American Statistical Association*, 64:1647, 1969.

[32] C. Fraley. Algorithms for model-based Gaussian hierarchical clustering. *SIAM Journal on Scientific Computing*, 20(1):270–281, 1998.

[33] H. P. Friedman and J. Rubin. On some invariant criteria for grouping data. *Journal of the American Statistical Association*, 62:1159–1178, 1967.

[34] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins, 3rd edition, 1996.

[35] A. D. Gordon. *Classification: Methods for the Exploratory Analysis of Multivariate Data*. Chapman and Hall, 1981.

[36] J. A. Hartigan. *Clustering Algorithms*. Wiley, 1975.

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

[38] T. Hastie and W. Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84:502–516, 1989.

[39] H. J. A. M. Heijmans. Mathematical morphology: A modern approach in image processing based on algebra and geometry. *SIAM Review*, 37(1):1–36, March 1995.

[40] H. Jeffreys. *Theory of Probability.* Clarendon, 3rd edition, 1961.

[41] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:773–795, 1995.

[42] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data.* Wiley, 1990.

[43] M. Leroux. Consistent estimation of a mixing distribution. *The Annals of Statistics*, 20:1350–1360, 1992.

[44] J. MacQueen. Some methods for classification and analysis of multivariate observations. In L. M. Le Cam and J. Neyman, editors, *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967.

[45] G. McLachlan. The classification and mixture maximum likelihood approaches to cluster analysis. In P. R. Krishnaiah and L. N. Kanal, editors, *Handbook of Statistics*, volume 2, pages 199–208. North-Holland, 1982.

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

[47] G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley, 1997.

[48] G. J. McLachlan and D. Peel. Mixfit: An algorithm for automatic fitting and testing of normal mixtures. In *Proceedings of the 14th International Conference on Pattern Recognition*, volume 1, pages 553–557. IEEE Computer Society, 1998.

[49] G. J. McLachlan and D. Peel. Robust cluster analysis via mixtures of multivariate t-distributions. In A. Amin, D. Dori, P. Pudil, and H. Freeman, editors, *Lecture Notes in Computer Science*, volume 1451, pages 658–666. Springer, 1998.

[50] R. Muise and C. Smith. Nonparametric minefield detection and localization. Technical Report CSS-TM-591-91, Coastal Systems Station, Panama City, Florida, 1991.

[51] S. Mukerjee, E. D. Feigelson, G. J. Babu, F. Murtagh, C. Fraley, and A. E. Raftery. Three types of gamma ray bursts. *The Astrophysical Journal*, 508:314–327, November 1998.

[52] F. Murtagh. *Multidimensional Clustering Algorithms*, volume 4 of *CompStat Lectures.* Physica-Verlag, 1985.

[53] F. Murtagh and A. E. Raftery. Fitting straight lines to point patterns. *Pattern Recognition*, 17:479–483, 1984.

[54] D. Peel and G. J. McLachlan. User's guide to EMMIX - version 1.0. University of Queensland, Australia, 1998.

[55] C. Posse. Hierarchical model-based clustering for large data sets. Technical report, University of Minnesota, School of Statistics, 1998.

[56] G. M. Reaven and R. G. Miller. An attempt to define the nature of chemical diabetes using a multidimensional analysis. *Diabetologia*, 16:17–24, 1979.

[57] B. D. Ripley. Neural networks and related methods for classification. *Journal of the Royal Statistical Society, Series B*, 56:409–456, 1994.

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

[59] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

[60] A. J. Scott and M. J. Symons. Clustering methods based on likelihood ratio criteria. *Biometrics*, 27:387–397, 1971.

[61] H. Späth. *Cluster Dissection and Analysis: Theory, Fortran Programs, Examples*. Ellis Horwood, 1985.

[62] D. Stanford and A. E. Raftery. Principal curve clustering with noise. Technical Report 317, University of Washington, Department of Statistics, February 1997.

[63] J. Stutz and P. Cheeseman. AutoClass - a Bayesian approach to classification. In J. Skilling and S. Sibisi, editors, *Maximum Entropy and Bayesian Methods, Cambridge 1994*. Kluwer, 1995.

[64] M. J. Symons. Clustering criteria and multivariate normal mixtures. *Biometrics*, 37:35–43, 1981.

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

[66] C. F. J. Wu. On convergence properties of the EM algorithm. *The Annals of Statistics*, 11:95–103, 1983.