Detecting Features in Spatial Point Processes With Clutter via Model-Based Clustering

Abhijit DASGUPTA and Adrian E. RAFTERY

We consider the problem of detecting features, such as minefields or seismic faults, in spatial point processes when there is substantial clutter. We use model-based clustering based on a mixture model for the process, in which features are assumed to generate points according to highly linear multivariate normal densities, and the clutter arises according to a spatial Poisson process. Nonlinear features are represented by several densities, giving a piecewise linear representation. Hierarchical model-based clustering provides a first estimate of the features, and this is then refined using the EM algorithm. The number of features is estimated from an approximation to its posterior distribution. The method gives good results for the minefield and seismic fault problems. Software to implement it is available on the World Wide Web.

KEY WORDS: Bayes factor; BIC; EM algorithm; Minefield; Poisson process; Seismic fault.

1. INTRODUCTION

Consider the problem of detecting surface-laid minefields on the basis of an image from a reconnaissance aircraft. After processing, such an image is reduced to a list of objects, some of which may be mines and some of which may be "clutter," such as other metal objects or rocks. The objects are small and can be represented by points without losing much information. The analyst's task is to determine whether or not minefields are present, and where they are. A typical dataset is shown in Figure 1. [Although actual minefield data were not available to us, the data in Fig. 1 were simulated according to specifications developed at the Naval Coastal Systems Station, Panama City, Florida, to represent minefield data encountered in practice; see Muise and Smith (1992).]

A similar problem that we also consider is to determine the location of seismic faults on the basis of earthquake catalogs only. Earthquakes tend to cluster close to the faults, but many earthquakes also happen far from the main fault lines. The faults are analogous to the minefields in the first example, whereas the earthquakes that happen away from the main faults are analogous to the clutter.

Our solution is an extension of the model-based clustering methodology introduced by Banfield and Raftery (1993), extending the work of Murtagh and Raftery (1984). This is based on a mixture model in which features are represented by highly linear multivariate normal densities and clutter is represented by a spatial Poisson process. By saying that a multivariate normal density is "highly linear," we mean that it is highly concentrated about its first principal component. Hierarchical clustering then partitions the points between the features and the clutter.

The methodology of Banfield and Raftery (1993) has been used successfully in a variety of situations and is the basis for the `mclust` software in S-PLUS. (The `mclust` software is also available free of charge as a Fortran program at http://lib.stat.cmu.edu/general/mclust, or by sending the e-mail message "send mclust from general" to statlib@stat.cmu.edu. An updated version of the S-PLUS function is available at http://lib.stat.cmu.edu/S/mclust, or by sending the e-mail message "send mclust from S" to statlib@stat.cmu.edu.) For the present problem it works reasonably well even when the amount of clutter is very large, but there is room for improvement. It provides estimates of the number of features based on the approximate weight of evidence (AWE), a crude approximation to twice the logarithm of the Bayes factor for k clusters against one cluster derived from the hierarchical model-based clustering results and using the maximized classification likelihood. These estimates of the number of clusters are reasonably good when the amount of clutter is small, but less so when the amount of clutter is large.

Here we improve on the method of Banfield and Raftery (1993) by refining the final partition using the EM algorithm and by using approximate Bayes factors to select the number of clusters. Figure 2 shows the results of applying our method, which we call `mclust-em`, to the minefield data of Figure 1. `mclust` alone has a detection rate of 89% and a false-positive rate of 22%, and the improved version, `mclust-em`, has a detection rate of 97% and a false-positive rate of 4%. Figure 2 clearly shows that `mclust-em` reconstructs the true minefield extremely well.

In Section 2 we review the model-based clustering method of Banfield and Raftery (1993) (`mclust`) and point out some limitations that are important in the present context. We review the EM algorithm for classification and show how it can be used to estimate the shape parameter, which determines how linear the clusters are and which in `mclust` must be specified by the user. We also show how to choose the number of clusters using approximate Bayes factors. In Section 3 we apply the methods to several minefield configurations and to the estimation of seismic faults based on earthquake catalogs. Finally, in Section 4 we dis-

Abhijit Dasgupta is a graduate student, Department of Biostatistics, and Adrian E. Raftery is Professor of Statistics and Sociology, Department of Statistics, University of Washington, Box 354322, Seattle, WA 98195-4322; E-mail: raftery@stat.washington.edu; Web: www.stat.washington.edu/raftery. This research was supported by Office of Naval Research grants N00014-91-J-1074, N00014-96-1-0192, and N00014-96-1-0330. The authors are grateful to Peter Guttorp, Girardeau Henderson, and Robert Muise for helpful discussions, to the editor, the associate editor, and an anonymous referee for useful comments that improved the article, and to Simon Byers for improving the software to implement the methods.

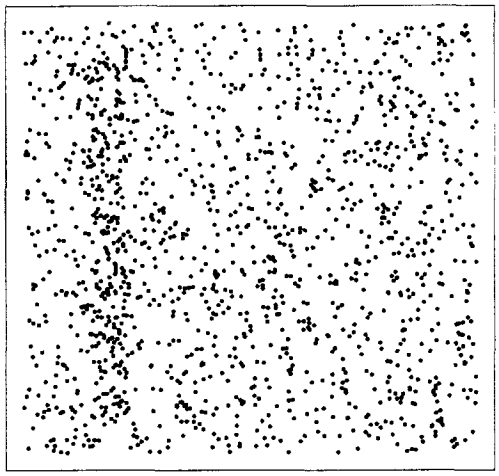

Figure 1. A Simulated Minefield With Clutter.

cuss remaining limitations of the methods proposed here, how they might be overcome, and connections to related literature.

2. METHODOLOGY

2.1 Model-Based Clustering

Banfield and Raftery (1993) proposed a method for model-based clustering of d-dimensional data based on a mixture of Gaussian distributions, with an additional (optional) component consisting of a homogeneous spatial Poisson process to represent "noise" or "clutter." They developed a hierarchical clustering method aimed at maximizing the classification likelihood

$$L(\theta, \gamma) = \prod_{i=1}^{n} f_{\gamma_i}(\mathbf{x}_i; \theta) = \prod_{k=0}^{G} \prod_{\mathbf{x}_i \in E_k} f_k(\mathbf{x}_i; \theta), \qquad (1)$$

where $E_k = \{\mathbf{x}_i: \gamma_i = k\}$ is the set of observations generated by the kth component of the mixture. Here $f_k(\mathbf{x}; \theta)$ is a multivariate normal $(\mu_k, \mathbf{\Sigma}_k)$ density for $k = 1, \ldots, G$ and a uniform density over the region of interest for $k = 0$ (for the Poisson noise). The maximization is over both the parameters θ and the partition γ; the method tends to give a good suboptimal partition rather than a global maximum.

Banfield and Raftery (1993) parameterized their model using the following modification of the standard spectral decomposition of $\mathbf{\Sigma}_k$:

$$\mathbf{\Sigma}_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T, \qquad (2)$$

where λ_k is the largest eigenvalue of $\mathbf{\Sigma}_k$, $\mathbf{D}_k$ is the matrix of eigenvectors, and $\mathbf{A}_k = \text{diag}\{1, \alpha_{2k}, \ldots, \alpha_{dk}\}$ is a $d \times d$ diagonal matrix. These factors have nice geometric interpretations: λ_k controls the volume of the kth cluster, $\mathbf{D}_k$ its orientation, and $\mathbf{A}_k$ its shape. Banfield and Raftery (1993) developed a range of clustering methods appropriate for different situations by constraining some or all of λ_k, $\mathbf{D}_k$, and $\mathbf{A}_k$ to be equal across clusters. Bensmail and Celeux (1996) applied these models to discriminant analysis.

Of particular interest in the present context is the model in which the shapes of the clusters are the same but their volumes and orientations are different; that is, $\mathbf{A}_k = \text{diag}\{1, \alpha\}$ $(k = 1, \ldots, G)$, where $\alpha < 1$ (because for spatial point processes we deal only with two dimensions, $d = 2$). Here α is the ratio of the second to the first eigenvalue; when α is much smaller than 1, the resulting clusters will tend to be long and narrow (or highly linear), whereas when α is close to 1, the clusters will tend to be almost circular in shape. For this reason, α is called the "shape parameter." Minefields, seismic faults, and similar features tend to be long and narrow with fairly similar shapes; this is represented by α being small. Banfield and Raftery (1993) denoted the resulting criterion for merging clusters by S^*. This methodology is implemented in the Fortran and S-PLUS software `mclust`.

To select the number of clusters, Banfield and Raftery (1993) developed the AWE criterion, which is a crude approximation to twice the log posterior probability that there are G clusters, plus a constant. The AWE approximation works reasonably well when the amount of clutter is small (as in most clustering applications), but turns out not to work well when the clutter is a high proportion of the data.

As can be seen from Figure 2, `mclust` alone works reasonably well for detecting features in a great deal of clutter, such as in Figure 1, but there is substantial room for improvement. It yields only a suboptimal solution to the estimation problem and does not give the probability that each point belongs to the feature, only a single partition. It also requires the user to specify α in advance, and does not provide a way of estimating it. Its method for finding the number of clusters works much less well when there is a lot of clutter.

In the following subsections we outline a series of improvements to `mclust`, based on the EM algorithm, that overcome these limitations and lead to improved performance. The actual (mixture) likelihood for the parameters

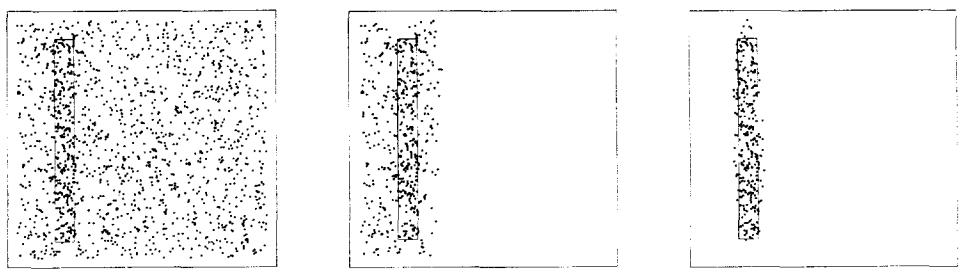

Figure 2. Simulated Minefield With Box Representing the Mined Region (a), Estimated Minefield Provided by `mclust` *(b), and Estimated Minefield by Our Method,* `mclust-em` *(c).*

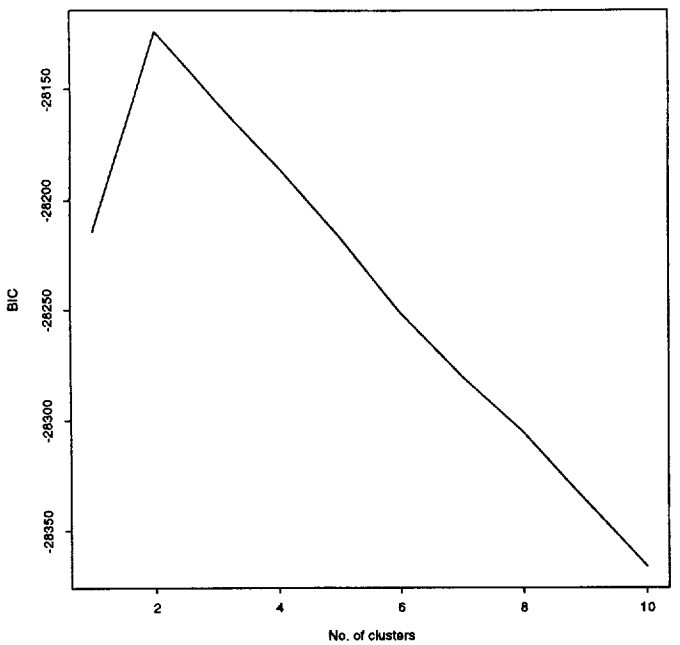

Figure 3. BIC Values for Different Numbers of Clusters in the Simulated Minefield.

is used, rather than the classification likelihood. This yields a way of estimating the shape parameter α, posterior probabilities of belonging to a feature or to the clutter for each point, and an approximation to the posterior probabilities of the number of clusters that works well in examples.

2.2 The EM Algorithm for Mixture Models

The EM algorithm (Dempster, Laird, and Rubin 1977) was originally proposed as a general method for obtaining maximum likelihood estimates in the presence of missing data. Its use for the estimation of mixture models has been studied in detail by McLachlan and Basford (1988) and Redner and Walker (1984).

Consider a population whose distribution follows a mixture density of the form

$$f(\mathbf{x};\theta) = \sum_{k=0}^{G} \pi_k f_k(\mathbf{x};\theta),$$

where the f_ks are distinct densities and $\sum_{k=0}^{G} \pi_k = 1$. Then for n observations from this mixture distribution, the "complete" data would be $y_i = (\mathbf{x}_i, \mathbf{z}_i)$, where $\mathbf{z}_i = (z_{i0}, z_{i1}, \ldots, z_{iG})$, with

$$z_{ik} = \begin{cases} 1 & \text{if the } i\text{th observation is in the } k\text{th cluster} \\ 0 & \text{otherwise.} \end{cases}$$

The vector $\mathbf{z}_i$ has a multinomial distribution with parameters $(1;\pi)$, where $\pi = (\pi_0, \ldots, \pi_G)$. This leads to the "complete-data log-likelihood," namely

$$l(y;\theta,\pi) = \sum_{i=1}^{n}\sum_{k=0}^{G} z_{ik}\{\log \pi_k + \log f_k(\mathbf{x}_i;\theta)\}.$$

The E step requires the computation of $\hat{z}_{ik} = E(z_{ik}|x_1, \ldots, x_n, \theta, \pi)$, which is the posterior probability that $\mathbf{x}_i$ is in the kth cluster. Maximum likelihood estimates of θ and π are obtained in the M step, which consists of maximizing the expected complete-data log-likelihood, namely

$$l^*(y;\theta,\pi) = \sum_{i=1}^{n}\sum_{k=0}^{G} \hat{z}_{ik}\{\log \pi_k + \log f_k(\mathbf{x}_i;\theta)\}.$$

This process is iterated to convergence.

2.3 Estimating the Shape Parameter

The hierarchical clustering methods of Banfield and Raftery (1993) (and the `mclust` software) require the shape parameter, α, to be specified by the user. Here we show how maximum likelihood estimates of α may be obtained, by including α in the M step of the EM algorithm.

Suppose that we have d-dimensional data generated by a mixture of G Gaussian distributions satisfying (2), with $\mathbf{A}_k = \mathbf{A} = \text{diag}\{1, \alpha, \ldots, \alpha\}$ $(k = 1, \ldots, G)$ and of clutter distributed as a homogeneous spatial Poisson process. Let the estimated covariance matrix $\hat{\mathbf{\Sigma}}_k$ (estimated from all the data, weighted by $\hat{z}_{ik}$) for the kth cluster have the spectral decomposition $\hat{\mathbf{\Sigma}}_k = L_k \Omega_k L_k^T$, where $\Omega_k = \text{diag}\{\omega_{k1}, \ldots, \omega_{kd}\}$. Then the expected complete-data log-likelihood, given α and maximized over the other parameters, is

$$\begin{aligned} &2 \text{ log-likelihood} \\ &= 2\sum_{k=0}^{G} \log(\hat{n}_k/n) \\ &\quad - (n - \hat{n}_0)(d\log(2\pi) + d(1 - \log(d)) + \log(|\mathbf{A}|)) \\ &\quad - d\sum_{k=1}^{G} \hat{n}_k \log(\text{trace}(\Omega_k \mathbf{A}^{-1})/n_k) - 2\hat{n}_0 \log(V), \end{aligned} \quad (3)$$

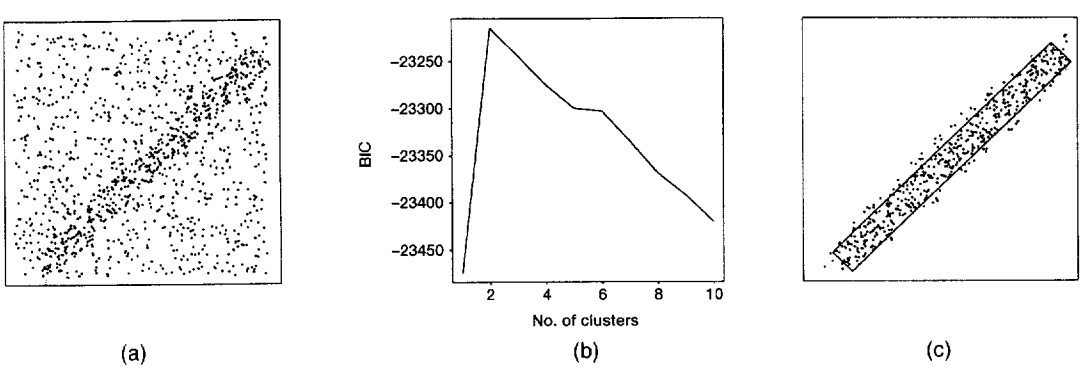

Figure 4. The Diagonal Minefield (a), the BIC Graph for the Diagonal Minefield (b), and the `mclust-em` *Solution (c).*

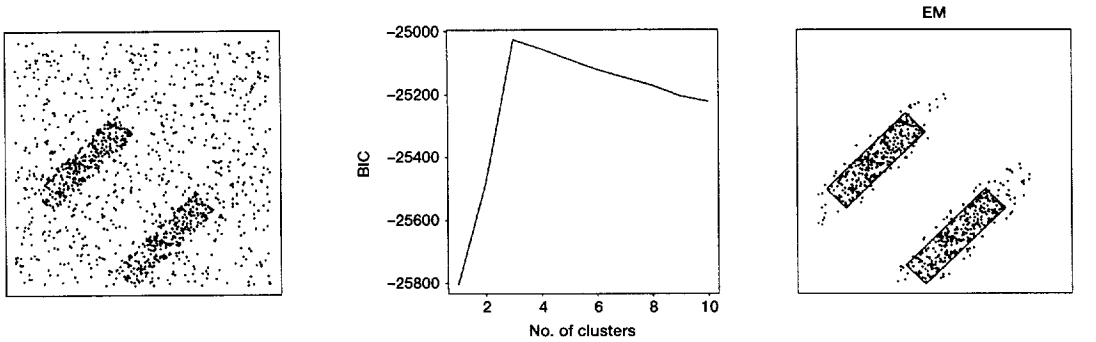

Figure 5. Two Parallel Minefields (a), the BIC Graph for the Parallel Minefield (b), and the `mclust-em` *Solution (c).*

where $\hat{n}_k = \sum_{i=1}^{n} \hat{z}_{ij}$, $(k = 0, 1, \ldots, G)$, and V is the volume occupied by the data in R^d. Operationally, we have used the definition $V = \prod_{j=1}^{d}(\max_{i=1,\ldots,n}\{x_{ij}\} - \min_{i=1,\ldots,n}\{x_{ij}\})$, which is the volume of the smallest hyperrectangle with sides parallel to the coordinate axes containing the data. But other, perhaps more satisfactory definitions could also be used, such as the volume of the smallest hyperrectangle with sides parallel to the principal components of the data or the volume enclosed by the convex hull of the data. For the examples considered in this article, these three definitions give similar results.

We then maximize this "profile likelihood" with respect to α. This results in the following likelihood equation:

$$\sum_{k=1}^{G} \left(\frac{\hat{n}_k}{n - \hat{n}_0} \right) \frac{\omega_{k2} + \cdots + \omega_{kd}}{\alpha\omega_{k1} + \omega_{k2} + \cdots + \omega_{kd}} = 1 - \frac{1}{d}. \quad (4)$$

This is a polynomial equation in α of order G, which will have at most G distinct roots. The roots of interest are those in [0, 1].

When there is one (nonclutter) group (i.e., $G = 1$) we have

$$\hat{\alpha} = \frac{\omega_2 + \cdots + \omega_d}{(d-1)\omega_1},$$

which is the maximum likelihood estimate of α (Anderson 1984). If we have two normal clusters (i.e., $G = 2$), the estimate of α is the solution of the following quadratic equation:

$$\mathrm{N}\left(1 - \frac{1}{d}\right)\omega_{11}\omega_{21}\alpha^2 + \left\{\mathrm{N}\left(1 - \frac{1}{d}\right)(\omega_{11}\omega_{22} + \omega_{12}\omega_{21}) - n_1\omega_{12}\omega_{21} - n_2\omega_{11}\omega_{22}\right\}\alpha + \mathrm{N}\left(\frac{1}{d}\right)\omega_{12}\omega_{22} = 0,$$

which can be solved by the usual algebraic methods, subject to $\hat{\alpha} \in [0, 1]$. Here $N = \hat{n}_1 + \hat{n}_2 = n - \hat{n}_0$. Note that if we have two balanced clusters (i.e., $\hat{n}_1 = \hat{n}_2$) in two dimensions, then $\hat{\alpha}$ is the geometric mean of the estimates of α from each of the two clusters individually.

For more than two clusters, it is possible to obtain exact solutions for $\hat{\alpha}$ by solving the appropriate polynomial equation. Computationally, however, it is easier to find the estimate of α by performing a grid search over the interval [0, 1] based on (4).

In general, mixture models can have multiple maximum likelihood solutions, including degenerate maximum likelihood estimates when all of a mixture component gets absorbed in d or fewer observations. We did not encounter this problem in our examples, and it seems unlikely to be as large a problem here as in other mixture models, for several reasons. The starting points given by `mclust` were always far enough from such degeneracies to prevent the EM algorithm from converging to any of them. Also, the imposition of a common shape on the covariance matrices implies that this can arise only if one of the λ_k is 0, or if $\alpha = 0$. In the latter case, all the Gaussian components will be degenerate, which will be a relatively rare occurrence.

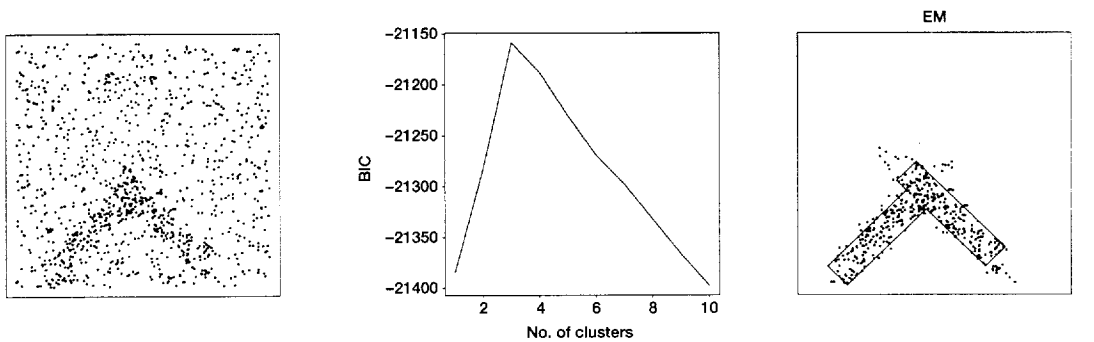

Figure 6. The Arrow-Shaped Minefield (a), the BIC Graph for the Arrow-Shaped Minefield (b), the `mclust-em` *Solution (c).*

If this problem does arise, however, an ad hoc solution would be to recognize that in practice data are measured discretely, and to impose corresponding lower bounds on the covariance matrices. For example, if the data in Figure 1 were accurate to three decimal places where the data region is the unit square, then we could reasonably impose the restriction that four standard deviations along the second principal component be at least .001; that is, $4(\alpha\lambda_1)^{1/2} > .001$, or $\alpha\lambda_1 > 2.5 \times 10^{-7}$. The likelihood becomes large only for extremely small values of $(\alpha\lambda_1)$, so a restriction such as this should be enough to avoid the degenerate solutions (see also Titterington, Smith, and Makov 1985, sec. 4.3). If an iteration of the EM algorithm leads to parameter values that violate the restriction, then the size of the EM step can be reduced, or the algorithm can be restarted with different starting values. This solution has been successfully applied in a different context by Stanford and Raftery (1997).

2.4 Choosing the Number of Clusters

Choosing the number of clusters is a vital issue in the applications that we consider here. In cluster analysis more generally it is a critical part of many applications, and no fully general and satisfactory solution seems available.

Here we attempt to develop a solution to the problem, building on the representation of cluster analysis in terms of mixture models. We determine the number of clusters by representing each choice considered as a statistical model and then comparing the resulting rival models using (approximate) *Bayes factors*. The Bayes factor for one model against another is the posterior odds for that model against the other when neither model is favored over the other a priori. (For a recent review of Bayes factors emphasizing the concepts underlying them and their use in scientific applications, see Kass and Raftery 1995.)

Reasons for using Bayes factors rather than frequentist significance tests in the present context include the fact that we are considering several possible models rather than just the two that can be compared by frequentist methods and general arguments for using Bayes factors rather than p values and the associated tests, reviewed by Kass and Raftery (1995). The asymptotic distribution of the likelihood ratio test statistic for finite-mixture models is nonstandard (Titterington et al. 1985; Wolfe 1971) and does not seem to be known in general.

The Bayes factors that we need for the present problem are ratios of high-dimensional integrals and usually cannot easily be evaluated analytically. Banfield and Raftery (1993) approximated them using a heuristically derived penalized version of the maximized classification log-likelihood, called the AWE. Our experience is that this works reasonably well when there is not much clutter, as in the examples considered by Banfield and Raftery (1993), but it often performs poorly when the amount of clutter is large, as in the examples we consider here.

Now, thanks to the EM algorithm, we can find the maximized *mixture* likelihood. Here we base our approximations to Bayes factors on the Bayesian information criterion (BIC) approximation (Schwarz 1978), namely

$$2 \log p(x|G) \approx 2l(x;\hat{\theta},G) - m_G \log n = \text{BIC},$$

where $p(x|G)$ is the (integrated) likelihood of the data given that there are G clusters, $l(x;\hat{\theta},G)$ is the maximized mixture log-likelihood with G clusters, and m_G is the number of independent parameters to be estimated in the G-cluster model. If each model (i.e., each number of clusters) is equally likely a priori, then $p(x|G)$ is proportional to the posterior probability that there are G clusters. The larger the value of BIC, the better the model according to this criterion; differences exceeding 10 can be viewed as representing very strong evidence. As we show later, this approximation works well even when the amount of clutter is large.

The available general theoretical justifications of the BIC approximation (Kass and Raftery 1995; Kass and Wasserman 1995; Schwarz 1978) rely on regularity conditions that do not hold for finite-mixture models (Titterington et al. 1985). However, several results are available that do support its use in this context. Leroux (1992) has shown that the estimator of the number of components that maximizes BIC does not underestimate the true number of components, asymptotically. Roeder and Wasserman (1997) have shown that when a finite mixture of normal densities is used to estimate a density "nonparametrically," the density estimator that uses BIC to select the number of components in the mixture is consistent. They also reported a simulation study in which the BIC estimator of the number of components performed very well. Our own results here are also encouraging for the BIC estimator. Kass and Wasserman (1995) have shown that for a wide class of models, BIC yields a close approximation to the Bayes factor with a proper, "unit-information" prior. If this result were extended to finite-mixture models, then it would provide further support for the use of BIC as the basis for approximate Bayes factors.

Figure 3 shows the values of BIC for each number of clusters up to 10 for the simulated minefield data of Figure 1. This decisively chooses the correct number of clusters—namely two, corresponding to the minefield and the clutter. (The one-cluster model is defined here by all the data being clutter-generated by a homogeneous spatial Poisson process; the one cluster is thus clutter rather than "feature." This makes sense in the minefield context, because it corresponds to the "null" situation in which there is no minefield, only clutter. However, in other applications it might well make more sense to consider the other possible one-cluster model—that in which all of the data are generated by a single multivariate normal distribution and there is no clutter.)

2.5 Implementation of `mclust-em`

We now describe our modification of model-based clustering, which we call `mclust-em`. We first do an agglomerative hierarchical model-based clustering of the data using `mclust` and an initial guess of the shape parameter α. This provides an initial partition of the data for each

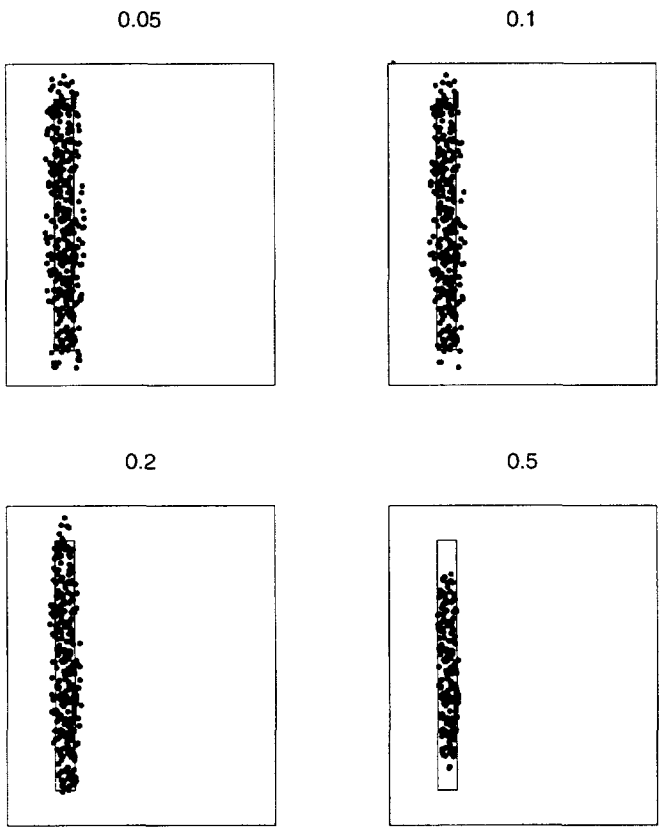

Figure 7. The Points Whose Probability of Being a Mine is at Least Equal to the Threshold Value Shown, for the Dataset in Figure 1, Based on `mclust-em`.

number of clusters that we wish to consider. We then use the EM algorithm described in Section 2.3 to improve the clustering for each of the models (i.e., for each number of clusters considered). Once the EM iterations converge, we use the maximized mixture likelihood to compute the BIC for each model. This provides information about the appropriate choice of model, as discussed in Section 2.4.

This procedure has some advantages over `mclust`. First, it provides an estimate of the shape parameter α based on the data. Even though `mclust-em` requires input of an initial value of α, the estimate of α that it provides is fairly robust to the choice of initial value. Second, it provides values of $\hat{z}_{ik}$ for each observation. This gives a measure of the uncertainty associated with the classification.

3. EXAMPLES

3.1 Detecting Minefields

We now consider the minefield detection problem in more detail. Given a processed image such as that of Figure 1, possibly containing both minefields and a substantial number of false identifications or "clutter," the task is (a) to determine whether or not a minefield is present; (b) to find out how many minefields there are; and (c) to determine the boundaries of the minefields. We have seen how `mclust-em` successfully carries out these tasks for the data of Figure 1.

We now consider three additional simulated minefield scenarios in which the minefield configuration and the ratio of the density of mines to that of clutter are varied. In the first scenario the minefield is diagonal rather than parallel to the axes in the image, in the second there are two disjoint minefields, and in the third the minefield is highly nonlinear.

We presented the `mclust-em` solution to the first scenario in Figure 1. The `mclust-em` solutions for the other three scenarios are shown in Figures 4, 5, and 6. In each case the BIC indicates decisively and correctly which model should be selected. The solutions displayed are of points that have at least a 20% chance of being in the minefield. Other choices may be made depending on the user's preference. Figure 7 shows that the results do not change greatly as this threshold is varied.

We now consider the performance of `mclust-em`, both in absolute terms and compared to `mclust`. Table 1 gives the detection and false-positive rates for the `mclust` and `mclust-em` solutions in the four scenarios. The detection rate is defined as the proportion of observations in the minefield region that are classified as such, and the false-positive rate is defined as the number of false positives classified as being in the minefield region divided by the total number of false positives in the data.

`mclust-em` performs extremely well, with almost perfect detection rates and low false-positive rates. Figures 4, 5, and 6 show that `mclust-em` does indeed give an excellent visual impression of where the minefields are located and of their boundaries. The original `mclust` of Banfield and Raftery (1993) also does fairly well and correctly locates the bulk of the minefield in each case, at least in a broad qualitative sense. But it is substantially less accurate than the improved version presented here (`mclust-em`), with appreciably lower detection rates *and* higher false-positive rates. Usually there is a trade-off between these two aspects of performance of a statistical method, but here we have been able to improve both simultaneously.

Note that the true minefields in these examples are rectangular, whereas the boundaries reconstructed by our methods are ellipses, so the data did not come from the model underlying our method. Nevertheless, the quality of the reconstructions is good.

3.2 Detecting Seismic Faults

We consider the problem of detecting seismic faults based on an earthquake catalog. The idea is that earthquake epicenters occur along seismically active faults and are measured with some error. So over time, observed earth-

Table 1. Detection Rates (DR) and False Positives (FP) for the Four Minefield Scenarios, all in Percentages

	DR	FP	DR	FP	
Type of minefield	`mclust-em`		`mclust-em`		$\hat{\alpha}$
Vertical	88	22	97	4	.015
Diagonal	90	12	100	9	.013
Arrow	84	25	100	9	.05
Parallel	81	38	100	16	.07

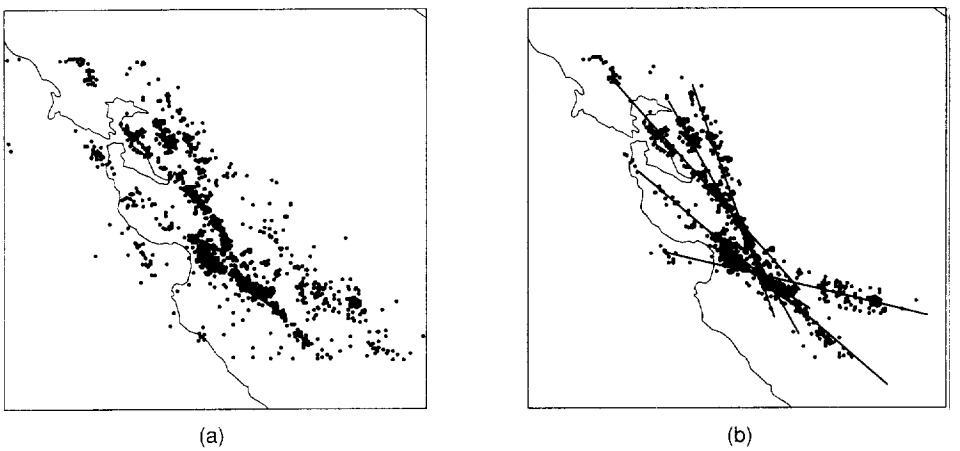

Figure 8. Earthquake Catalog (a) and the Solution Given by `mclust-em` *(b). The lines are the first principal axis of the clusters.*

quake epicenters should be clustered along such faults. We considered an earthquake catalog recorded over a 40,000 km^2 region of the central coast ranges in California from 1962–1981 (McKenzie, Miller, and Uhrhammer 1982). An advantage of looking at this region is that the known fault structure is well documented.

Figure 8a shows the locations of the earthquakes recorded in the catalog. Some linear structures are clearly visible in the plot. Figure 8b shows the `mclust-em` solution. Figure 9 shows the BIC values for the different numbers of clusters and their successive differences. We selected a classification with seven clusters (six nonnoise clusters and one noise cluster), because the BIC attains a local maximum there and the successive differences in the BIC values are small thereafter.

We find that the classification obtained using six (nonnoise) clusters corresponds well with the available documentation of faults in the region of interest (Fig. 10). In particular we find that the activity along the San Andreas, Calveras, and Hayward faults is well captured by `mclust-em`. One or two clusters do not correspond to any of the documented faults. One possible reason for this is that there are small pockets of intense activity (spherical clusters), which `mclust-em` classifies with other clusters to form the long, narrow clusters for which we are searching. Another possibility is that there are undocumented faults that are fairly active and need to be investigated.

4. DISCUSSION

We have introduced a new method, called `mclust-em`, for detecting features in spatial point processes in the presence of large amounts of clutter. This method uses model-based clustering based on a mixture model in which the features of interest are represented by multivariate normal densities with high linearity, specified by the shape being the same across features, with a low value of the shape parameter α, the ratio of the second to the first eigenvalue of the feature's covariance matrix. The clutter is represented by a homogeneous spatial Poisson process.

The model is estimated via the EM algorithm, with a good starting point provided by agglomerative hierarchical clustering based on the same model, as implemented in `mclust`. The number of features is estimated using approximate Bayes factors. The overall method works well for detecting and finding the boundaries of minefields in the presence of large amounts of clutter and for detecting

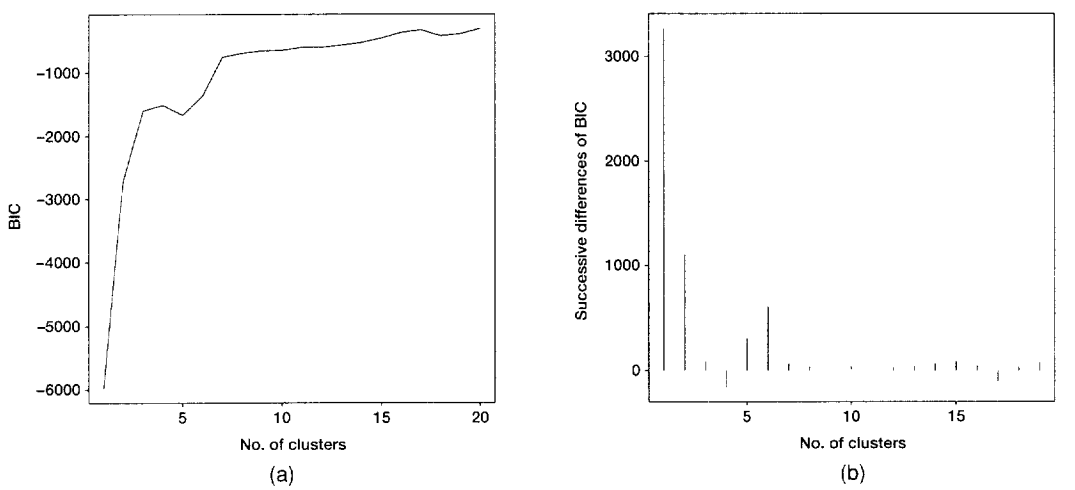

Figure 9. BIC Values for Different Cluster Sizes for the Earthquake Data (a), and Successive Differences of the BIC Values (b).

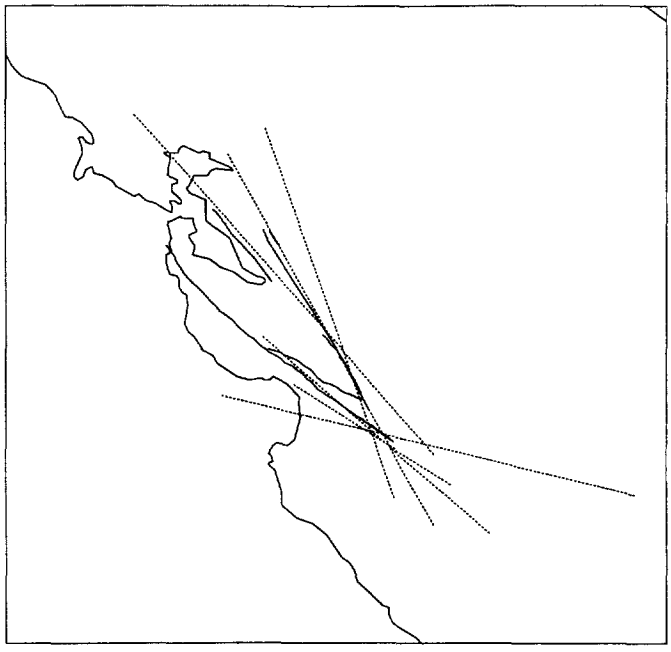

Figure 10. Correspondence of the `mclust-em` *Solution With the Documentation. The solid lines are the documented faults; the dotted lines are the first principal axes of the* `mclust-em` *clusters.*

seismic faults. Nonlinear features can be represented in a piecewise linear fashion, by several clusters rather than just one cluster.

Some limitations remain. The method yields point estimates of the model parameters θ (namely the μ_k, λ_k, D_k, π_k, and α), but not standard errors or confidence intervals. Here interest focuses on the classification (for which uncertainty statements *are* available) rather than on the model parameters, so this does not seem too serious. The posterior probabilities of the classification are conditional on the parameter estimates and hence understate the true uncertainty, although to an extent that vanishes asymptotically.

Standard errors for the parameters could be found using the supplemented EM algorithm (Meng and Rubin 1991), and these could then be used to correct the posterior probabilities of the classification via the Laplace method (Tierney and Kadane 1986). A more direct approach might be via the weighted likelihood bootstrap (WLB; Newton and Raftery 1994), which can use the EM code for maximizing the likelihood directly to compute a full Bayesian posterior distribution. The WLB would have the advantage of involving only a relatively small addition to the methods developed here. A more ambitious approach would be to do a fully Bayesian analysis of the mixture model directly using Markov chain Monte Carlo (Diebolt and Robert 1994; Lavine and West 1992). This has been implemented for the `mclust` models of Banfield and Raftery (1993) without clutter (Bensmail, Celeux, Raftery, and Robert 1997), and could be extended to the case where there is clutter without great difficulty, at least in principle.

One way of improving the results would be to use more information. In our examples we have used only the spatial positions of points, but other information is also available, such as estimated physical characteristics of identified mines or the intensity of earthquakes. It would be possible to incorporate this auxiliary information directly in the mixture model and clustering criterion used; one way of doing this was suggested by Banfield (1988), with good results.

The substantial literature on the statistical analysis of spatial point patterns has been reviewed by Cressie (1991), Diggle (1984), and Ripley (1991). Surprisingly, this literature focuses on highly *local* dependencies and regularities in point processes and does not seem to contain methods directly applicable to the present problem of detecting features, such as minefields or seismic faults, that are more *global* (at least on the scale of the images to be analyzed). Our methods assume that the observed points are generated independently, and although we expect our results to be robust to this assumption, it would be interesting to refine them by allowing for local interactions and regularities in addition to the global features.

One notable exception is the work of Ogata and Katsura (1988) and Ogata and Tanemura (1985), who have analyzed the spatial point process of earthquake occurrences off Japan. But their aims were somewhat different from ours. They used smoothing methods to produce a smooth estimate of the intensity of earthquake occurrences at each point, and did not provide estimates of seismic faults as such. Our methods could be used to estimate earthquake intensities as well.

One interesting alternative to parametric model-based methods such as those proposed here is due to Allard and Fraley (1997), who developed a nonparametric maximum likelihood approach to the same problem that uses Voronoi tesselations and mathematical morphology.

A collection of S-PLUS functions to implement `mclust-em` is available at http://www.stat.washington.edu/raftery/Research/Mclust/mclust_software.html.

[Received October 1995. Revised February 1997.]

REFERENCES

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

Anderson, T. W. (1984), *An Introduction to Multivariate Statistical Analysis* (2nd ed.), New York: Wiley.

Banfield, J. D. (1988), "Constrained Cluster Analysis and Image Understanding," Ph.D. dissertation, University of Washington, Dept. of Statistics.

Banfield, J. D., and Raftery, A. E. (1993), "Model-Based Gaussian and Non-Gaussian Clustering," *Biometrics*, 49, 803–821.

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

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

Cressie, N. A. (1991), *Statistics for Spatial Data*, New York: Wiley.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 39, 1–37.

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

Diggle, P. J. (1984), *Statistical Analysis of Spatial Point Patterns*, New York: Academic Press.

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

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

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

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

McKenzie, M., Miller, R., and Uhrhammer, R. (1982), *Bulletin of the Seismographic Stations*, 53, Nos. 1–2, Berkeley, CA: University of California Press.

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

Meng, X.-L., and Rubin, D. B. (1991), "Using EM to Obtain Asymptotic Variance–Covariance Matrices: The SEM Algorithm," *Journal of the American Statistical Association*, 86, 899–909.

Muise, R., and Smith, C. (1992), "Nonparametric Minefield Detection and Localization," Technical Report CSS-TM-591-91, Coastal Systems Station, Panama City, FL.

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

Newton, M. A., and Raftery, A. E. (1994), "Approximate Bayesian Inference With the Weighted Likelihood Bootstrap" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 56, 3–48.

Ogata, Y., and Katsura, K. (1988), "Likelihood Analysis of Spatial Inhomogeneity of Marked Point Patterns," *Annals of the Institute of Statistical Mathematics*, 40, 29–39.

Ogata, Y., and Tanemura, M. (1985), "Estimation of Interaction Potentials of Marked Spatial Point Patterns Through the Maximum Likelihood Method," *Biometrics*, 41, 421–433.

Redner, R. A., and Walker, H. F. (1984), "Mixture Densities, Maximum Likelihood and the EM Algorithm," *SIAM Review*, 26, 195–239.

Ripley, B. D. (1991), *Statistical Inference for Spatial Processes*, Cambridge, U.K.: Cambridge University Press.

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

Stanford, D., and Raftery, A. E. (1997), "Principal Curve Clustering with Noise," Technical Report 317, University of Washington, Dept. of Statistics. (http://www.stat.washington.edu/tech.reports/tr317.ps).

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

Tierney, L., and Kadane, J. B. (1986), "Accurate Approximations for Posterior Moments and Marginal Densities," *Journal of the American Statistical Association*, 81, 82–86.

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

Wolfe, J. (1971), "A Monte Carlo Study of the Sampling Distribution of the Likelihood Ratio for Mixtures of Multinormal Distributions," Technical Report STB 72-2, U.S. Naval Personnel and Training Research Laboratory, San Diego.