Nearest-Neighbor Clutter Removal for Estimating Features in Spatial Point Processes

Simon BYERS and Adrian E. RAFTERY

We consider the problem of detecting features in spatial point processes in the presence of substantial clutter. One example is the detection of minefields using reconnaissance aircraft images that identify many objects that are not mines. Our solution uses Kth nearest neighbor distances of points in the process to classify them as clutter or otherwise. The observed Kth nearest neighbor distances are modeled as a mixture distribution, the parameters of which are estimated by a simple EM algorithm. This method allows for detection of generally shaped features that need not be path connected. In the minefield example this method yields high detection and low false-positive rates. Another application, to outlining seismic faults, is considered with some success. The method works well in high dimensions. The method can also be used to produce very high-breakdown-point–robust estimators of a covariance matrix.

KEY WORDS: Breakdown point; Edge effects; EM algorithm; Image analysis; Minefield; Mixture model; Robust covariance estimation; Seismic fault.

1. INTRODUCTION

Consider the problem of detecting surface-laid minefields on the basis of an image from a reconnaissance aircraft. Some processing may be performed on the image to give a list of objects, some of which may be mines and some of which may be clutter. These objects are assumed to be small compared to the scale of the problem and may be represented as points without significant loss of information. An analyst may then be called on to determine whether or not minefields are present, and where they are. This can be thought of as searching for regions of higher density of the point process that is the result of the data collection.

Allard and Fraley (1997) developed a method to find the maximum likelihood solution using Voronoï polygons. They assumed that the feature consists of a single connected (but otherwise arbitrarily shaped) component and estimated it as the largest connected component of the (unconnected) nonparametric maximum likelihood solution. Dasgupta and Raftery (1998) used model-based clustering (Banfield and Raftery 1993), to search for minefields of linear or piecewise linear form in the presence of clutter. They used an EM algorithm to estimate the underlying mixture model.

These methods are both somewhat restricted in that Allard and Fraley (1997) assumed that there is a single connected feature, whereas Dasgupta and Raftery (1998) assumed the features to have a specific shape. Here we adopt a different approach, estimating and removing the *clutter* without making any assumptions about the shape or number of the features. This enables us to estimate features in quite general situations and also to remove clutter as an initial step before other analyses.

The method of detection proposed here is one based on the distance to the Kth nearest neighbor (NN) of a point in a process. Intuition suggests that for points in regions of higher density (i.e., those inside the features), the Kth nearest neighbor distance of a point will be less, on average, than that for a point in the clutter. Techniques are developed for separating the mines from the clutter on the basis of this measurement. There is an extensive literature on point processes and their properties, reviewed by Cressie (1991) and Ripley (1991). NN techniques have been used, but not in the context considered here.

Figure 1 shows a simulated minefield and the result of an application of this method. This method can search for minefields of any shape, and they need not be path connected. In the aforementioned example, the detection rate was 97.1%, with a false positive rate of 5.7%. Section 3.1 discusses this example further.

The methodology is described in Section 2, and in Section 3 it is applied to some examples, one of which was also considered by Dasgupta and Raftery (1998). In Section 4 the method is applied to the problem of robust covariance estimation. In Section 5 software available to implement the methods is described, and in Section 6 limitations and extensions of this method are discussed.

2. METHODOLOGY

2.1 Two-Dimensional Poisson Process Theory

We assume that the clutter is distributed as a homogeneous Poisson point process. The features are distributed also as a Poisson process restricted to a certain part of the image and overlaid on the clutter. This now defines a Poisson process of piecewise constant rate. We make no assumptions about the shape of the features or their number, so our method is quite general.

If we consider just one homogeneous Poisson process, we can find the distribution of the distance D_K from a randomly chosen point in the process to its Kth nearest

Simon Byers is Graduate Research Assistant and Adrian E. Raftery is Professor of Statistics and Sociology, Department of Statistics, University of Washington, Seattle, WA 98195; (E-mail: byers@stat.washington.edu and raftery@stat.washington.edu or www.stat.washington.edu/raftery). This research was supported by Office of Naval Research grant N00014-96-1-0192. The authors are grateful to Abhijit Dasgupta, Chris Fraley, and Werner Stuetzle for helpful discussions and to the editor, associate editor, and two anonymous referees for helpful comments on the first version of this article.

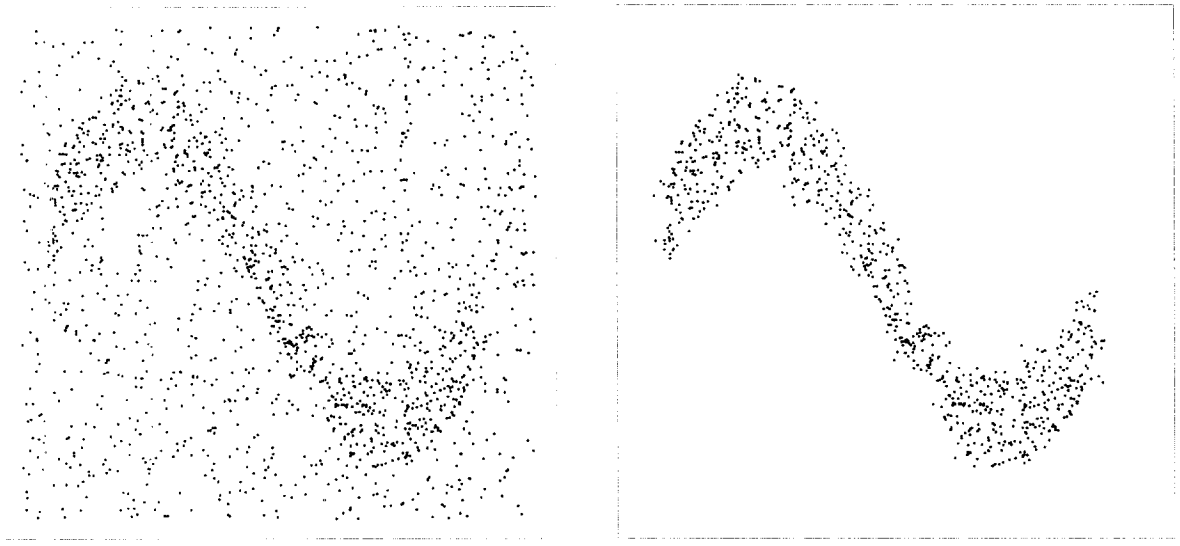


Figure 1. A Simulated Minefield that was Treated With 15th Nearest-Neighbor Clutter Removal to Give an Estimated Minefield.

neighbor. For $x \in [0, \infty)$,

$$P(D_K \geq x) = \sum_{k=0}^{K-1} \frac{e^{-\lambda\pi x^2}(\lambda\pi x^2)^k}{k!} = 1 - F_{D_K}(x).$$

This formula is obtained by imagining a circle of radius x around the point in consideration. If D_K is greater than x, then there must be one of $0, 1, \ldots, K-1$ points in this circle. Thus the density $f_{D_K}(x)$ can be found:

$$f_{D_K}(x) = \frac{dF_{D_K}(x)}{dx} = \frac{e^{-\lambda\pi x^2} 2(\lambda\pi)^K x^{2K-1}}{(K-1)!}$$

This is a transformed gamma random variable, $Y \sim \Gamma(K, \lambda\pi)$, where $Y = (D_K)^2$. We denote this situation by $D_K \sim \Gamma^{(1/2)}(K, \lambda\pi)$. This result was mentioned in Cressie (1991, p. 611) in connection with tests of complete spatial randomness. More generally, this is an instance of the generalized gamma distribution (Stacy 1967). Due to the convenient form of this distribution, maximum likelihood estimation of the rate, λ, given some observed values of D_K, is easy. Given the values of D_K from a homogeneous rate two-dimensional Poisson process, the maximum likelihood estimate (MLE) of the rate of the process from this method is given by

$$\hat{\lambda} = \frac{K}{\pi \sum_{i=1}^{n} d_i^2}$$

where the d_i are the observations of the Kth NN distances. In a realization of a point process, it is a simple matter to calculate actual Kth NN distances for all of the points.

2.2 Observed Kth Nearest-Neighbor Distances as a Mixture Distribution

Returning to the simple model of a feature with clutter as two superimposed Poisson processes, one may postulate that the distribution of D_K is (approximately) a mixture of two of the distributions seen in Section 2.1. The observed D_K can be displayed in a histogram, as in Figure 2. With a strong distinction between feature and clutter, this histogram becomes highly bimodal. Our postulated model for the D_K is

$$D_K \sim p\Gamma^{(1/2)}(K, \lambda_1\pi) + (1-p)\Gamma^{(1/2)}(K, \lambda_2\pi).$$

A simple application of the EM algorithm (Dempster, Laird, and Rubin 1977) can be used to estimate the parameters λ_1, λ_2, and p that characterize the postulated distribution of the D_K. The "missing data" in this problem are the classifications into the two components, with one $\delta_i \in \{0, 1\}$ for each data point, where $\delta_i = 1$ if the ith point is inside one of the features and $\delta_i = 0$ if not (in which case the ith point is clutter). Thus each data point has an observation d_i of D_K and an unknown δ_i.

The E step of the algorithm consists of

$$E(\hat{\delta}_i^{(t+1)}) = \frac{\hat{p}^{(t)} f_{D_K}(d_i; \hat{\lambda}_1^{(t)})}{\hat{p}^{(t)} f_{D_K}(d_i; \hat{\lambda}_1^{(t)}) + (1-\hat{p}^{(t)}) f_{D_K}(d_i; \hat{\lambda}_2^{(t)})},$$

and the M step consists of

$$\hat{\lambda}_1^{(t-1)} = \frac{K \sum_{i=1}^{n} \hat{\delta}_i^{(t+1)}}{\pi \sum_{i=1}^{n} d_i^2 \hat{\delta}_i^{(t+1)}},$$

$$\hat{\lambda}_2^{(t+1)} = \frac{K \sum_{i=1}^{n} (1 - \hat{\delta}_i^{(t+1)})}{\pi \sum_{i=1}^{n} d_i^2 (1 - \hat{\delta}_i^{(t-1)})},$$

and

$$p^{(t+1)} = \sum_{i=1}^{n} \hat{\delta}_i^{(t-1)} / n.$$

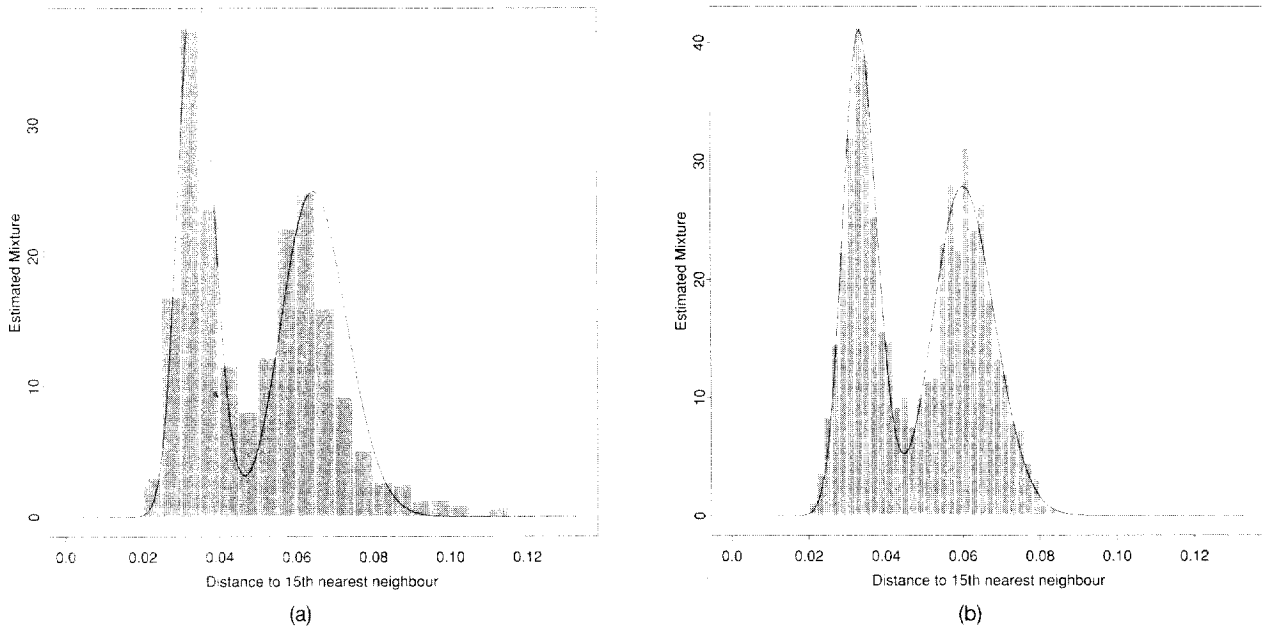


Figure 2. Two Estimates of a Mixture Distribution, (a) Ignoring and (b) Correcting for Edge Effects, From the Example Seen in Figure 1. Note that the right tail of the density is much better fit with the correction for edge effects.

With this information, the data points can be classified as mines or clutter based on some criterion. The simplest criterion is to classify according to the component of the mixture under which the observed D_K has higher density.

2.3 Edge Effects

Consideration of the points in the corner of the region under consideration reveals that some points may have different Kth NN characteristics due to the effects of the boundary. For example, a point right in the corner of a square region will have only one-quarter of the surrounding points that points in the center will have. This means that the Kth NN distance will be larger on average than if the region of consideration covered the plane. Methods for correction of effects such as these have been discussed by Cressie (1991, p. 607).

One method of correcting for this effect is to wrap the region of interest onto a toroid. This means that points on the edges have points from opposite edges as effective neighbors. This is easy to implement. Figure 2; shows two histograms of observed D_{15} from the data shown in Figure 1; 2b shows the D_{15} found from the toroidal edge–corrected data, and 2a shows the uncorrected version. One can see that a slight right tail of the distribution is present in the uncorrected version and that this is not well fit by the estimated density. The toroidal edge correction remedies this quite effectively.

2.4 Choosing the Value of *K*

So far we have assumed that the user has chosen the value of K. In many problems some degree of user specification will be possible. For example, one can choose K to be about the size of the smallest feature that one wishes to detect or the largest one wishes to consider as clutter. Using a value $K = 8$ might miss features composed of only seven points, but using $K = 5$ might extract the feature well.

One might think of finding the joint distribution of D_K for several values of K and using the information in all of them. But the sum of the D_K^2 for the largest K present is sufficient for λ, so no extra information is gained beyond using the largest K present.

In the absence of user input, we recommend that the user perform the analysis for several increasing K, stop when the improvements in separation get small enough, and then use the results from the last K. We used an entropy-type measure of separation, $S = \sum_{i=1}^{n} \delta_i \log(\delta_i)$, where the δ_i are the probabilities of being in the first component of the mixture. Plotting these entropies sequentially and looking for a leveling-off changepoint in the graph is easy to do.

Figure 3 shows the entropies plotted for the method applied with increasing K to the first sine wave dataset of Section 3.1. Several values of K seem reasonable from this plot, and in our experience to date all of the values within a reasonable range tend to perform similarly. One simple and automatic ad hoc method is to estimate the changepoint in a piecewise linear saturation model of the entropies as a function of K, using the posterior mode from the approximate Bayesian method of Raftery (1994 sec. 2.3). Here this yields $K = 8$. In fact, the results reported in Section 3.1 are based on $K = 15$, which also seems reasonable given Figure 3, and these two values of K yield similar results in this example.

2.5 Extension to Higher Dimensions

The D_K can be found for a particular realization of a process as easily in d dimensions as in two dimensions.

To extend the distributional result, one looks at a d-dimensional hypersphere instead of a circle. The volume

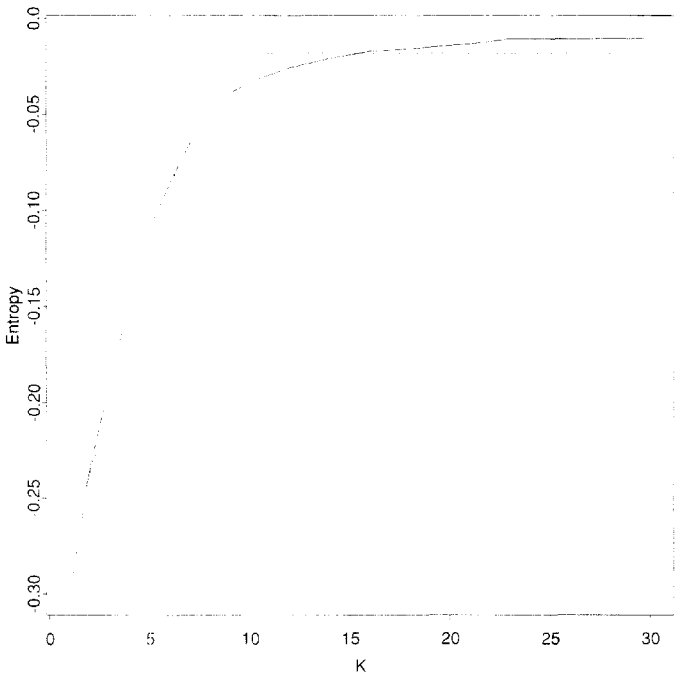


Figure 3. Plot of Classification Entropies for Various Values of K. ——, entropies, · · ·, changepoint model.

of the d-dimensional unit hypersphere is

$$\alpha_d = \frac{2\pi^{d/2}}{d\Gamma(d/2)}.$$

Then the procedure in Section 2.1 can be followed using $\alpha_d x^d$ instead of πx^2, to give the result that

$$Y \sim \Gamma(K, \lambda\alpha_d), \quad \text{where} \quad Y = (D_K)^d.$$

The extension to a mixture distribution also follows. The bias correction becomes more difficult, but is still possible in higher dimensions.

2.6 Summary of the Procedure

Our method for removing the clutter comprises the following steps:

- Choose a value for K.
- Find the Kth NN distance for each point in the data.
- Apply the EM algorithm to these distances to estimate λ_1, λ_2, and p.
- Use these estimates to classify the data points according to whether they have higher density under the clutter or feature component of the mixture.

This method can yield an estimate of the features, but it may also include residual elements of clutter in this estimate, because there may be a few points in the clutter that have sufficiently low Kth NN distance to be mistakenly included in the feature. One easy remedy for this is to treat the estimated feature as feature plus sparse clutter and apply the entire procedure again.

S-PLUS functions were written to perform this procedure in a user-friendly manner, as described in Section 5. Extensions to this procedure show promise. These include the following

- Application of the procedure multiple times to get a "better" end result. This would treat the estimated feature as a new dataset and apply the same method on this. This was used in the example in Section 3.2.
- Alteration of the EM algorithm to search for r groups, each with a different rate. We have implemented this on examples not shown here, and the performance is in line with the performance in the two-rate case. Each group might correspond to a set of features with a different density (e.g., seismic faults with different earthquake frequencies).

3. EXAMPLES

3.1 A Curvilinear Minefield

The minefield shown in Figure 1 is actually a region between two sine waves, but numerous other (and more bizarre) shapes have been explored. This one could be seen as an example of a section of a curvilinear minefield. One can see that the problem is very respectably cleaned up in Figure 1. This was performed using 15th NN methods. Table 1 gives the error rates from several minefield examples of this form, but with varying rates in the minefield. The method performs less well as the minefield rate decreases below twice the clutter rate. The effect of the edge correction apparently is to impede detection, but in reality it has a large corrective effect on the mixture fitting. The uncorrected method finds more mines and more false positives by a biasing artifact introduced by the edge effects. The influence of one point on estimation of the rate of a process based on a D_K is of the order r^2, hence the observed large changes in the mixture.

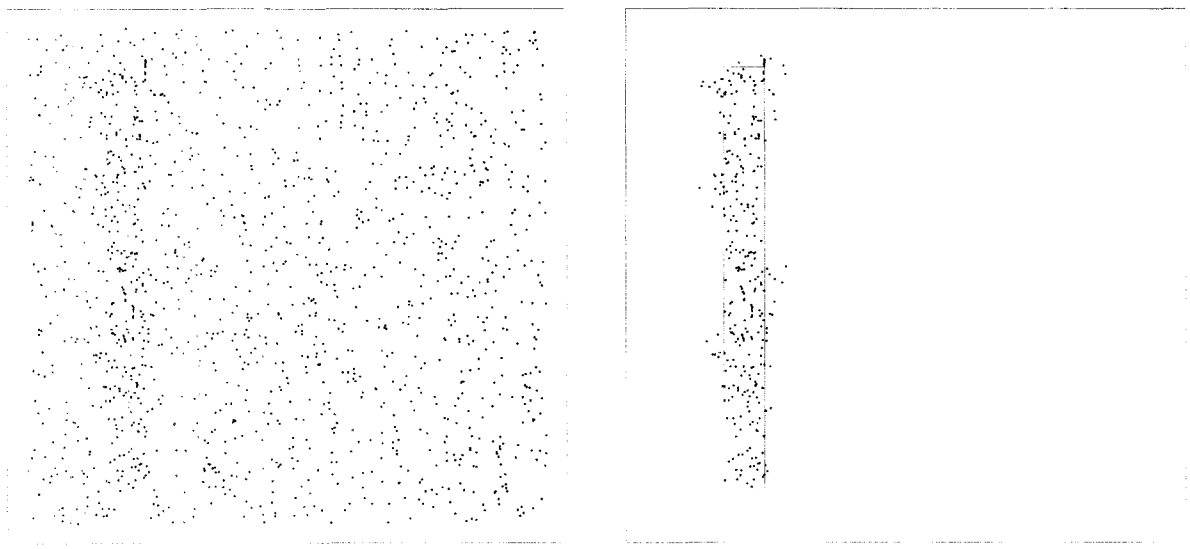
3.2 A Linear Minefield

The linear simulated minefield of Dasgupta and Raftery (1998), using the same (simulated) data, was subjected to the method developed here. The simulation was designed as described by Muise and Smith (1992) to capture the essential features of actual processed images of minefields. Figure 4 shows the initial data along with the results of minefield detection. This example was processed with 15th NN analysis with a further application to clean up. Table 2 shows the classification rates from the application of the new method to this example.

The NN method does almost as well here as the model-based `mclust-em` method of Dasgupta and Raftery (1998),

Table 1. Detection and False-Positive Rates (in Percent) for the Example in Figure 1

Rates λ_1, λ_2	Edge corrected		Not edge corrected	
	Detection	False positive	Detection	False positive
4,729, 1,200	98	4	99	6
3,553, 1,200	91	6	98	11
2,405, 1,200	82	12	96	46



(a) (b)

Figure 4. The Dasgupta and Raftery Minefield Example (a) Before and (b) After a Double Application of 15th NN-based Minefield Detection. The inner box shows the boundary of the minefield.

even though the minefield is fairly similar in shape (although not identical) to one for which the latter method would be optimal.

3.3 Higher Dimensions

We constructed a 10-dimensional example to examine the performance in higher dimensions. It was constructed to be analogous to the previous example in that it is a 10-dimension hyperrectangle with one long dimension, but was larger in a few of the dimensions than the very slim example discussed previously. Due to the problems associated with displaying 10-dimensional data, only the first two dimensions are shown in Figure 5.

The "minefield" was composed of only 18 "mines" in clutter of 1,782 points, giving a total of 1,800 data points. Note that these 18 mines were not spottable by eye from *any* two-dimensional view of the data. The S-PLUS functions `brush`, `spin`, and `pairs` give no hint of even the presence of the "minefield."

The result of applying NN methods to these data was 100% detection and no false positives. Numerous repetitions of the exercise occasionally give a false positive (.0005% as a rate).

The reason that performance in this example is so stunning is that despite the fairly large apparent volume of the minefield, the entire region was a factor of 6×10^8 bigger in terms of volume. Thus the process rate is a factor of 6×10^6 higher in the minefield, due to the effects of increasing dimensionality.

Table 2. Detection and False-Positive Rates from Application to the Dasgupta and Raftery Example

Method	*Detection rate*	*False-positive rate*
First application	96	8
Second application	96	5
`mclust` and EM	97	4

3.4 Earthquake Data

We now illustrate the algorithm by applying it to a geological example. Data were collected as to the position of earthquakes in California. Supposing that earthquakes occur mainly near faults, it may be possible to look for faults by removing earthquakes that are away from the main body of occurrences. Figure 6a shows the locations of all earthquakes with a Richter intensity above 2.5 in central California from 1962–1981. An application of 5th NN clutter removal produced the results on the right.

In comparison with the results of the treatment of this data by Allard and Fraley (1997), the two methods perform very similarly. One key difference is the isolated cluster in the bottom right that NN methods pick up but that the connected component part of Allard and Fraley's method leaves out. This cluster is treated as one end of a linear cluster of earthquakes in the analysis of Dasgupta and Raftery (1998). They end up filling in the sparse part between it and other clusters with clutter to produce the linear form that they search for. It would seem that the `mclust-em` method is more suited to finding features such as faults that are supposed to be roughly linear, but the differences exposed here show that less-structured methods do have contributions to make in structured situations.

4. ROBUST COVARIANCE ESTIMATION

The presence of outliers in multivariate data can damage estimates of the moments and covariance. Methods have been developed to yield estimates that are robust to certain amounts of contamination. An extensive literature exists on this and related problems. (Surveys can be found in Barnett and Lewis 1978 and Rousseeuw and Leroy 1987.) The most commonly used robust estimator of covariance appears to be the minimum volume ellipsoid (MVE) method of Rousseeuw and van Zomeron (1990). The link to the methods here is that one might think of outliers as light clutter surrounding a cluster of points.

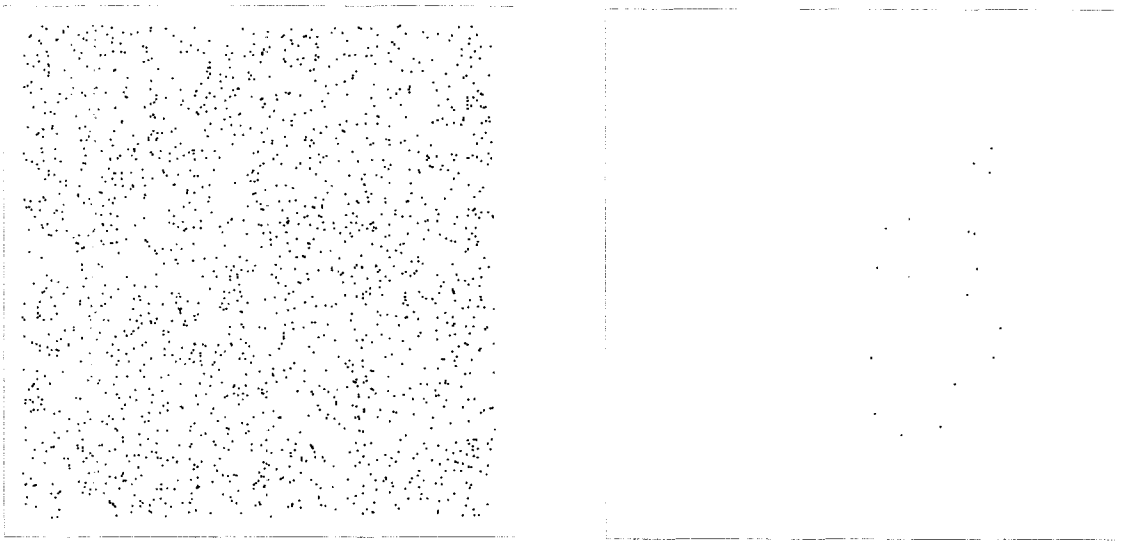


Figure 5. A 10-Dimensional Example Seen in the First 2 Dimensions: (a) Actual Data; (b) the Reconstructed "Minefield." The method achieved 100% detection and 0% false positive in this example, which had only 18 mines in 1,792 clutter.

In the case of a bivariate normal cluster of points with extensive, very outlying clutter, the problem could be considered as a feature in clutter. Removal of the clutter followed by moment estimation could prove to be a viable procedure. The established methods for this type of problem, such as MVE, often have a breakdown point of about 50% outliers. Removal of clutter by NN methods, followed by estimation can give results that have a higher breakdown point than MVE in this case. In fact, rather than removing clutter, we weighted the observations by the probabilities that they are not clutter as estimated by the EM algorithm.

Table 3 shows some results from simulated data. The data are bivariate normal with mean zero and covariance matrix diag$(1, 25)$. The outliers have the same distribution as the data but are multiplied by 10. NN cleaning, MVE and the standard covariance estimate were each applied to 100 simulated datasets under each of the outlier levels. The mean estimated covariance was then found under each case, as displayed in Table 3.

It can be seen from the table that the standard estimator breaks down immediately, whereas MVE breaks down somewhere before 50% contamination in this case. NN cleaning remains steady above 50% contamination, but it begins to break down by 67%. Both the MVE estimator and the NN-cleaned estimator underestimate the variances where there are no outliers, the NN method more severely. Results not shown in the table show that in the presence of only one moderately large outlier, NN cleaning often performs extremely well. The D_K^2 influence of the one large NN distance can be enough to attract an entire component of the mixture, hence leaving the rest of the data to determine the estimated covariance.

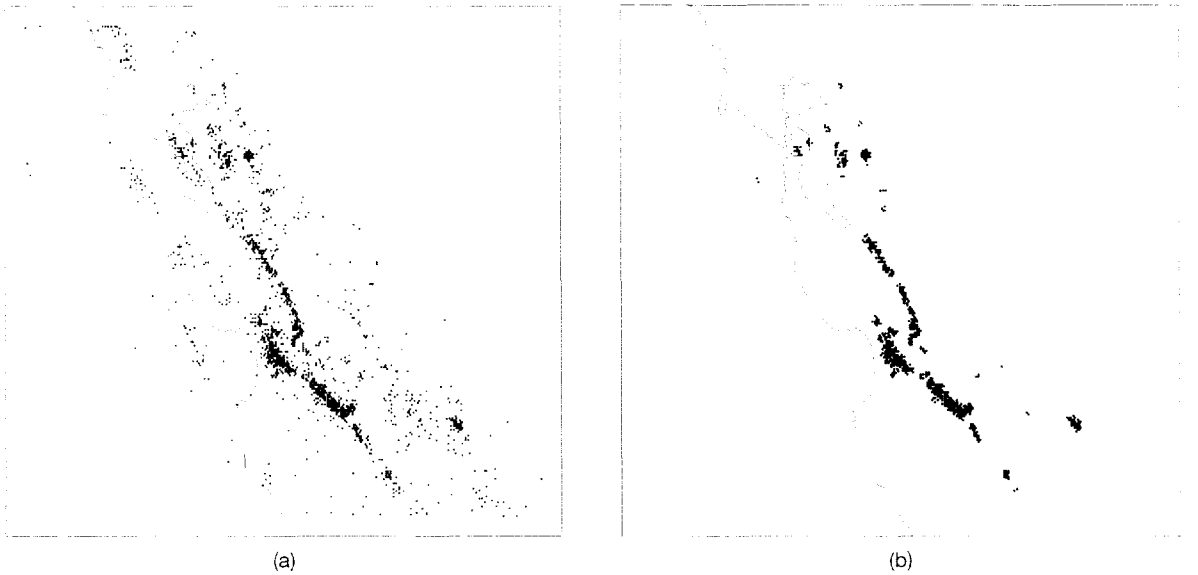


Figure 6. Positions of Earthquakes in California (a) and the Dense Clusters in the Data With the Noisiness Removed by 5th NN Methods (b).

Table 3. Estimates of Covariance for Various Levels of Contamination in Bivariate Normal Data with 500 Observations, the Mean of the Covariances from 100 Simulations being shown in each case

Method	Percent outliers							
	0%		33%		50%		67%	
NN cleaned	2.6	0	3.5	.1	3.9	.1	9.4	.2
	0	15.8	.1	21.9	.1	23.0	.2	58.8
MVE estimate	3.5	0	4.2	.1	8.5	0	126.1	9.1
	.1	21.9	.1	26.6	0	53.6	9.1	798.7
Standard	3.9	0	134.5	−3.3	202.2	2.7	268.0	1.6
	0	25.1	−3.3	863.1	2.7	1,265.9	1.6	1,674.5
True covariance of data	4.0	0						
	0	25.0						

5. SOFTWARE

The foregoing examples were implemented in S-PLUS 3.3. A specially written function, `nnclean` implemented the whole procedure. It is available at the Statlib S archive, accessible at http://lib.stat.cmu.edu/S/nnclean or by sending the e-mail message "`send nnclean from S`" to statlib@stat.cmu.edu.

The function takes as input an $n \times p$ matrix of the data points with a value for K and returns the estimated parameters of the mixture distribution along with a classification of the data points. The CPU time for the 1,800-point sine wave example was 1.5 seconds on a Silicon Graphics Impact running IRIX64. The use of quad tree NN finding enables application to datasets of more than 50,000 points without much trouble.

6. DISCUSSION

We have introduced a simple and intuitive method for estimating regions of differing point densities in a point process. It can be applied without user input about the shapes of the regions. This is a strength when the shape of the feature is not known. The application to the data from Dasgupta and Raftery (1998) shows that its performance is close to that of `mclust-em`, though the time it requires and its complexity are significantly less.

In comparison to the method of Allard and Fraley (1997), this method shows some conceptual similarities. A Voronoï polygon is closely related to the concept of NN. The NN method has the advantage that it searches for any feature, of whatever shape and whether path-connected or not. The postprocessing that can be used with the method of Allard and Fraley leads to a single path-connected component. The method of Allard and Fraley has the property that an actual region is yielded as an estimate, by virtue of the Voronoï polygons. It is also easy to overestimate the feature with this method, something that it may be desirable to do in some applications (such as the minefield application). The NN method does not immediately give a region, but such an estimate could be formed. Expansion of the NN-estimated feature could also be performed by some morphological postprocessing or by biasing the splitting criterion.

One very useful feature of the method described here is the potential for easy use in higher dimensions. A characteristic of most other methods available for this type of problem is the possibility of extension to problems in higher dimensions, but also the tediousness of doing so. Our method is as simple to implement in higher dimensions as in two; the 10-dimensional example illustrates this.

Our method can search for r regions of different point densities, and it has been modified to do so in examples not discussed here. For example, consider a minefield on a beach. The rate of clutter may be different on the beach than in the water; thus this becomes three regions, two regions of (different) low densities and one region of high density.

The method can be applied to robust covariance estimation. It outperforms existing methods when the proportion of outliers is very high, at least under the conditions explored.

Our method is based on the assumption that the feature is a Poisson process overlaid with a clutter process. We investigated the effects of several departures from this assumption in work not reported in detail here. When we applied the method to a minefield in clutter with inhibition, its performance improved. Also, application to a regular minefield in Poisson clutter yields an improvement.

The value of K to be used must be specified by the user. We have given some guidelines for the user of this method, but this area could benefit from further investigation.

In Section 2.3 we discussed one method for correcting for edge effects, based on wrapping the region onto a toroid. Another method is to explicitly bias correct in the estimation of the mixture of densities. Instead of the probability that a Poisson process has each of $0, \ldots, K-1$ points in a circle radius x, take the intersection of the circle radius x, centered at the point and the plotting region and use this area instead.

[Received July 1996. Revised May 1997.]

REFERENCES

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

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

Barnett, V., and Lewis, T. (1978), *Outliers in Statistical Data*, New York: Wiley.

Cressie, N. (1991), *Spatial Statistics*, Oxford, U.K.: Wiley.

Dasgupta, A., and Raftery, A. E. (1998), "Detecting Features in Spatial

Point Processes With Clutter via Model-Based Clustering," *Journal of the American Statistical Association*, 93, 294–302.

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

Muise, R., and Smith, C. (1992), "Nonparametric Minefield Detection and Localization," Technical Report CSS-TM-591-91, Naval Surface Warfare Center, Coastal Systems Station.

Raftery, A. E. (1994), "Change Point and Change Curve Modelling in Stochastic Processes and Spatial Statistics," *Journal of Applied Statistical Science*, 1, 403–424.

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

Rousseeuw, P. J., and Leroy, A. M. (1987), *Robust Regression and Outlier Detection*, New York: Wiley.

Rousseeuw, P. J., and van Zomeron, B. C. (1990), "Unmasking Multivariate Outliers and Leverage Points," *Journal of the American Statistical Association*, 85, 633–639.

Stacy, E. W. (1962), "A Generalization of the Gamma Distribution," *Annals of Mathematical Statistics*, 33, 1187–1192.