Biometrika (2001), **88**, 4, *pp*. 1021–1033

Printed in Great Britain

Clustering methods based on variational analysis in the space of measures

BY M. N. M. VAN LIESHOUT
CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands
colette@cwi.nl

I. S. MOLCHANOV
Department of Statistics, University of Glasgow, Glasgow G12 8QW, U.K.
ilya@stats.gla.ac.uk

AND S. A. ZUYEV
Department of Statistics and Modelling Science, University of Strathclyde, Glasgow G1 1XH, U.K.
sergei@stams.strath.ac.uk

SUMMARY

We formulate clustering as a minimisation problem in the space of measures by modelling the cluster centres as a Poisson process with unknown intensity function. We derive a Ward-type clustering criterion which, under the Poisson assumption, can easily be evaluated explicitly in terms of the intensity function. We show that asymptotically, i.e. for increasing total intensity, the optimal intensity function is proportional to a dimension-dependent power of the density of the observations. For fixed finite total intensity, no explicit solution seems available. However, the Ward-type criterion to be minimised is convex in the intensity function, so that the steepest descent method of Molchanov & Zuyev (2001) can be used to approximate the global minimum. It turns out that the gradient is similar in form to the functional to be optimised. If we discretise over a grid, the steepest descent algorithm at each iteration step increases the current intensity function at those points where the gradient is minimal at the expense of regions with a large gradient value. The algorithm is applied to a toy one-dimensional example, a simulation from a popular spatial cluster model and a real-life dataset from Strauss (1975) concerning the positions of redwood seedlings. Finally, we discuss the relative merits of our approach compared to classical hierarchical and partition clustering techniques as well as to modern model based clustering methods using Markov point processes and mixture distributions.

Some key words: Cluster analysis; Optimisation on measures; Poisson point process; Steepest descent.

1. INTRODUCTION

The term 'cluster analysis' incorporates a wide class of techniques for partitioning data 'points' representing individuals or objects into groups. Classical clustering techniques are often hierarchical in nature, building a tree, the so-called dendrogram, based on some distance measure. The method of construction may be agglomerative or divisive, and the distance between two clusters may be defined in various ways. Ward (1963) argues that the loss of information caused by merging clusters may be measured by the increment of the pooled within groups sum of squared deviations, so that at each step one merges those groups whose fusion results in minimum increase in the sum of squares. Finally, the tree is thresholded in order to find the meaningful clusters; see for example Hartigan (1975) and Jardine & Sibson (1971).

In contrast, partition techniques are based on iteratively allocating points to clusters, using some optimality criterion, such as the trace or determinant of the pooled within groups sum of squares matrix. The former again is Ward's criterion, and the latter was proposed by Friedman & Rubin (1967). Similar techniques appear when finding the k-mean of a sample of points; see Hartigan (1975) and MacQueen (1967).

The techniques discussed above are essentially model-free, but recently there has been a surge of interest in mixture models. Here, the data is supposed to come from a mixture of k components representing the clusters. Suppose that $(y_1, \ldots, y_m)$ denotes the vector of observations, and let $\phi(j) \in \{1, \ldots, k\}$ denote the component label of y_j. Then one can define an optimal clustering by maximising a complete-data likelihood. It turns out that, if each component is normally distributed with mean m_i and the same covariance matrix $\Sigma = \sigma^2 I$, and if the means m_i are estimated by the sample means of the observations allocated to the component, then the criterion for clustering becomes the Ward criterion. For general Σ, we re-obtain the Friedman & Rubin criterion. More details and variations on this theme can be found in Banfield & Raftery (1993), Diebolt & Robert (1994), Richardson & Green (1997), McLachlan & Basford (1988), Scott & Simons (1971) and Stephens (2000). Further information on classical clustering methods can be found in Everitt (1974), Hartigan (1975), Johnson & Wichern (1982), Kaufman & Rousseeuw (1990), Mardia et al. (1979) and other textbooks on multivariate statistics.

Most approaches outlined above decide on the number of clusters in an ad hoc, subjective manner. Furthermore, the cluster centres only play an implicit role, approximated by the centre of gravity or other 'mean' of the detected clusters, if they appear at all. Such an approach may be natural in applications where the main aim is to detect groups in data, but less so in datasets of a biological or evolutionary nature such as those discussed in § 5. In such cases, a point process approach can be taken. For instance, Baddeley & van Lieshout (1993, 2001), Lawson (1993), van Lieshout (1995) and van Lieshout & Baddeley (1995) suggest an integrated model for the number of clusters, their centres and the data partition simultaneously. Coupling from the past ideas (Propp & Wilson, 1996) can be used to sample from the posterior distribution of cluster centres, facilitating the estimation of model parameters and other quantities of interest; see Baddeley & van Lieshout (2001) and van Lieshout (2000).

Here we propose an intermediate approach that is neither hierarchical nor strongly model-based. As above, we use a point process framework to allow a variable number of cluster centres. The parent process of cluster centres is assumed to be distributed as an inhomogeneous Poisson process, but no other model assumption is made. Instead of choosing the number of clusters or a threshold level in the dendrogram, we fix the total intensity of the point process of parents; its spatial distribution is chosen so as to minimise the Ward criterion. In contrast to the partition approach based on the same criterion, our optimisation problem is convex in the intensity function. This implies that a unique solution can be found by steepest descent techniques, independently of the initialisation of the algorithm.

In § 2, we consider the cluster centres as a realisation of a Poisson process with unknown intensity surface. We formulate a clustering criterion in the spirit of Ward as the expected pooled within groups sum of squares. Section 3 considers an asymptotic solution by letting the expected number of clusters increase. If this number is instead set at a finite value, numerical optimisation is called for. We adapt the steepest descent algorithm of Molchanov & Zuyev (2000b, 2001) to the present context in § 4, and evaluate its performance on synthetic and real-life examples in § 5. The paper is concluded by a critical discussion and comparison with hierarchical, partition and model-based approaches.

2. Optimising the intensity of the Poisson parent process

Throughout this paper, the data pattern to be analysed consists of a set of points $y=\{y_1,\ldots,y_m\}$ in a bounded subset D of the d-dimensional Euclidean space $\mathbb{R}^d$. The Euclidean distance between two points $x, y \in D$ is denoted by $\rho(x, y)$. Our aim is to find a collection of cluster centres, or parents, $x=\{x_1,\ldots,x_k\}$, for $k=1,2,\ldots$, explaining the data. This can be done by minimising the Ward-type criterion

$$\left[\operatorname{tr}\left\{\sum_{x_i\in x}\sum_{y_j\in Z_x(x_i)}(y_j-x_i)(y_j-x_i)^{\mathrm{T}}\right\}\right]=\sum_{x_i\in x}\sum_{y_j\in Z_x(x_i)}\rho^2(x_i,y_j), \tag{1}$$

where $Z_x(x_i)$ is the collection of points in the plane closer to x_i than to any other parent $x_j \in x$ $(j \neq i)$. In other words, $Z_x(x_i)$ are the Voronoi cells generated by the set x; see Okabe et al. (2000). Minimisation problems for the functional (1), also with a general power $\beta>0$ instead of 2, can be traced to many other applications, including that of finding the k-mean (Hartigan, 1975, Ch. 4) of a configuration y in agglomerative clustering, or the mailbox problem discussed by Okabe et al. (2000, Ch. 9). In all these instances the number k has to be predetermined and steepest-descent-type minimisation algorithms are used to find a configuration x that minimises (1). This involves optimisation in a space of moderate dimension, dk, but the objective functional is not convex, so, as the initial configuration must be provided by the user, there is no guarantee that the descent algorithm ends up at a global rather than a local minimum.

The key innovation of the current paper is to interpret x as a realisation of a Poisson point process Π on D with finite intensity measure μ. For the homogeneous case, μ is proportional to Lebesgue measure, but we are mostly interested in the inhomogeneous case when μ becomes a general intensity measure. The total number of points of Π in a set B is a Poisson random variable with mean $\mu(B)$ and the numbers of points in disjoint sets are mutually independent. Therefore, constraints on the number of parent points can be rephrased as constraints on the total mass $\mu(D)$ which is also the mean number of Π-points in D. Since $\mu(D)$ is finite by assumption, the total number of points in Π is also almost surely finite.

Replacing x with Π in (1) and taking the expectation of the random variable thus obtained yields our objective functional that can be written as

$$f(\mu)=E_\mu\left\{\sum_{x_i\in\Pi}\sum_{y_j\in Z_\Pi(x_i)}\rho^2(x_i,y_j)\right\}. \tag{2}$$

The subscript μ under the expectation or probability sign is used to indicate that the expectation or probability is taken with respect to the distribution of a Poisson process with intensity measure μ. A functional of type (2), with an arbitrary power of $\rho(x_i, y_j)$, was considered by Molchanov & Zuyev (2000a) for optimising the locations of stations in telecommunication networks. In this context, the daughter points represent subscribers of the network, while the parent points correspond to stations. If we write $\rho(y, \Pi)$ for the minimal distance between y and a point of Π, (2) can be reformulated as

$$f(\mu) = \sum_{j=1}^{m} E_\mu\{\rho^2(y_j, \Pi)\}. \tag{3}$$

Note that with positive probability Π is empty, in which the case the distance $\rho^2(y_j, \Pi)$ in (3) is ill-defined. Thus, we must assign some value u to $\rho(y_j, \varnothing)$. Since we are dealing with minimisation of $f(\mu)$, a natural choice for u is the diameter of D, that is the maximal distance $\rho(x, y)$ between two points $x, y \in D$.

Since Π is a Poisson point process, it is relatively straightforward to compute the expectation in (3), yielding

$$f(\mu) = \sum_{j=1}^{m} \int_0^{u^2} \exp[-\mu\{B_{t^{1/2}}(y_j) \cap D\}]\, dt, \tag{4}$$

where $B_{t^{1/2}}(y_j)$ is the ball of radius $t^{1/2}$ centred at y_j. The interested reader is referred to the Appendix for a derivation of this formula.

The objective functional is defined on the set of all finite nonnegative measures and can be extended using (4) to signed measures, although without immediate probabilistic interpretation. An important implication of (4) is that the objective functional is convex in μ; that is, for every pair of measures μ and η and for each $c \in [0, 1]$,

$$f\{c\mu + (1-c)\eta\} \leqslant cf(\mu) + (1-c)f(\eta).$$

This is easily seen by using the fact that the function $\mu \mapsto e^{-\mu}$ is convex and observing that convexity is preserved by integration.

Since the value of $f(\mu)$ can be made arbitrarily small as the total mass of μ increases unboundedly, we have to constrain $\mu(D)$ to some fixed $a > 0$. The minimisation problem can then be written as

$$f(\mu) \mapsto \min, \quad \mu(D) = a. \tag{5}$$

Further constraints on μ may be added to incorporate additional information about the parents, for example by weighting their possible positions with a 'cost' function and considering only those μ that do not exceed the total cost; see Molchanov & Zuyev (2000a) for a general framework for optimising functionals of Poisson point processes.

3. An asymptotic solution

Molchanov & Zuyev (2000a) suggested a framework for asymptotic analysis of minimisation problems for functionals on measures with growing total mass. Referring to Molchanov & Zuyev (2000a) for details, consider a sequence of measures μ_a $(a > 0)$ such that μ_a minimises $f(\mu)$ over all measures with total mass a. Then under certain technical conditions the normalised intensities $a^{-1}\mu_a$ converge to a limit as $a \to \infty$, the so-called high intensity solution $\bar{\mu}$.

In our context, suppose that the daughter points y have been sampled from a distribution with probability density $p_y(.)$, perhaps given a priori or estimated by kernel smoothing (Bowman & Azzalini, 1997, Ch. 1) of y. Then the objective function (3) transforms into

$$f(\mu) = \int_D E_\mu\{\rho^2(z, \Pi)\} p_y(z)\, dz. \tag{6}$$

The same functional (6) was considered by Molchanov & Zuyev (2000a) in a telecommunication application, where it was shown that the density of a high intensity solution $\bar{\mu}$ is proportional to a power of the daughter density:

$$p_{\bar{\mu}}(z) \propto \{p_y(z)\}^{d/(d+2)}. \tag{7}$$

The interpretation of this result is that, if a large number of parent points are taken into account, they can be sampled from a density proportional to $\{p_y(z)\}^{d/(d+2)}$. Such a sample provides a natural initial configuration for, for example, the k-means algorithm or the constrained optimisation problem (5), that can be further improved using descent methods.

4. Steepest descent algorithm

Functionals of measures can be minimised efficiently using steepest descent algorithms, as described in Molchanov & Zuyev (2001). At every step, the idea is to move from μ to $\mu+\eta$ for some signed measure η chosen in such a way that the value of the objective function decreases as fast as possible and the constraints are not violated. In our case, this means that the total mass of $\mu+\eta$ must be the same as that of μ.

The steepness of a particular update from μ to $\mu+\eta$ is characterised by the directional derivative of $f(\mu)$ evaluated with respect to η, defined by

$$\lim_{t\downarrow 0} t^{-1}\{f(\mu+t\eta)-f(\mu)\} = \int g_\mu(z)\eta(dz). \tag{8}$$

The function $g_\mu(.)$ is called the gradient of $f(\mu)$. For the objective function $f(\mu)$ given by (2), the gradient equals

$$g_\mu(z) = -\sum_{j=1}^{m} \int_{\rho^2(y_j,z)}^{u^2} \exp[-\mu\{B_{t^{1/2}}(y_j)\cap D\}]\, dt. \tag{9}$$

A derivation of this expression can be found in the Appendix. Note that the gradient (9) resembles $f(\mu)$ as in (4), except for the integration interval.

The steepest descent algorithm iteratively redistributes mass of μ in the direction determined by this gradient. Clearly, to keep the total mass of $\mu+\eta$ constant, the added term η must have zero total mass, and hence η is necessarily a signed measure. The size ε of a step is controlled by the mass of the positive, or negative, part of η. To minimise the right-hand side of (8) one should place an atom of mass ε at the minimum of $g_\mu(.)$, or distribute it between several global minima if they exist. Similarly, the negative mass $-\varepsilon$ should ideally be placed at the maximum of $g_\mu(.)$, which amounts to taking away a mass ε from μ at this point. This can seldom be done, however, since the current μ may not have enough mass at this point, if any. Thus, we should remove mass from regions where $g_\mu(z)$ is large until an amount ε has been taken. More precisely, Molchanov & Zuyev (2001) proved that the steepest descent direction η is obtained when the mass of μ is redistributed in such a way that all mass of μ is taken from $D_t=\{x\in D: g_\mu(z)\geqslant t\}$ for a suitable $t\geqslant 0$, and placed at the point where g_μ is minimal. The threshold value t can be found from the condition $\mu(D_t)=\varepsilon$. If the equality has no solution, then we choose the smallest t satisfying $\mu(D_t)\leqslant\varepsilon$ and remove mass $\varepsilon-\mu(D_t)$ by reducing the μ-content of points $z\in D$ with $g_\mu(z)$ as close as possible to, but smaller than, t.

At the beginning of the algorithm, the step size ε is set at some arbitrary value. Iteratively, in the direction specified by the steepest gradient, mass of amount ε is redistributed in the manner described above. If this step does not lead to a decrease of the objective function, the step size is reduced and the procedure repeated. Note that, since (4) is convex in μ, the steepest descent algorithm converges to the global minimum from every initial state.

It is shown in Molchanov & Zuyev (2000a) that a necessary condition for a measure μ^* to solve problem (5) can be formulated as

$$g_{\mu^*}(z)\begin{cases} = c, & \mu^*\text{-almost everywhere,} \\ \geqslant c, & \text{for all } z, \end{cases} \tag{10}$$

for some constant c. The constant c is the Lagrange multiplier for the corresponding constrained optimisation problem. The necessary condition (10) can be used as a stopping rule for the steepest descent algorithm described above: stop if over all points in the support of the current μ the variation of g_μ is a constant c within a predetermined small number δ, and if, at all other points z in the support of μ, $g_\mu(z)$ is at least c. The algorithm is implemented in the S-Plus and R languages. The code is available at www.stats.gla.ac.uk/~ilya and www.stams.strath.ac.uk/~sergei and is distributed as an R-language bundle mesop. The datasets used in the following examples can be obtained from the same sources.

Figure 1 shows several steps of the steepest descent algorithm applied to a one-dimensional problem on $D = [0, 1]$ with $y = \{0{\cdot}2, 0{\cdot}4, 0{\cdot}5, 0{\cdot}55, 0{\cdot}9\}$ and the measure's total mass fixed at $a = 10$. The parent space is discretised into a grid with mesh size s, with $s = 0{\cdot}02$ in our example, and the intensity measure μ is atomic and supported on the grid. Note, however, that the data points y do not necessarily lie on the grid; see for example the point 0·55 here. The inner integrand in the objective functional (4) is a step function in t, with break points at the squared distances from y_j to grid points. To see this, consider y_1. If necessary we rearrange the indices of the grid points in such a way that $\rho(x_1, y_1) \leqslant \rho(x_2, y_1) \leqslant \ldots \leqslant \rho(x_n, y_1)$, and then the integral $\int_0^{u^2} \exp[-\mu\{B_{t^{1/2}}(y_1) \cap D\}]\, dt$ can be written as

$$\rho^2(x_1, y_1) + \{\rho^2(x_2, y_1) - \rho^2(x_1, y_1)\}e^{-\mu(\{x_1\})} + \ldots + \{u^2 - \rho^2(x_n, y_1)\}e^{-\mu(\{x_1\}) - \ldots - \mu(\{x_n\})}.$$

A similar formula holds for the other summands in (4) and for the gradient. Therefore, if for each y_j a record is kept of the grid points sorted according to their distance to y_j as well as the increments in squared distance, it is easy to perform updates of the gradient and objective functional.

5. Examples

5·1. *Preamble*

In all the examples below we used the steepest descent algorithm described in § 4 on the unit square $[0, 1] \times [0, 1]$ in the plane. The measures were defined on a uniform grid with mesh size $s = 0{\cdot}02$ in both directions. The stopping rule was such that the descent is terminated if the variation of the gradient over all atoms of μ with mass greater than δa is less than δ multiplied by the total range of the gradient, i.e. the difference between its maximum and minimum. The descent works acceptably fast, taking about one second per step on a SUN ULTRA 10 Workstation, 360 MHz, for y consisting of 123 points as in the case study described below. Plausible results are obtained for the tolerance level $\delta = 0{\cdot}01$ in about 100 steps, while $\delta = 0{\cdot}0001$ requires considerably more steps to be done, of the order of several thousands depending on the total mass of μ.

We have opted to present the results in the form of contour plots of the optimal measure

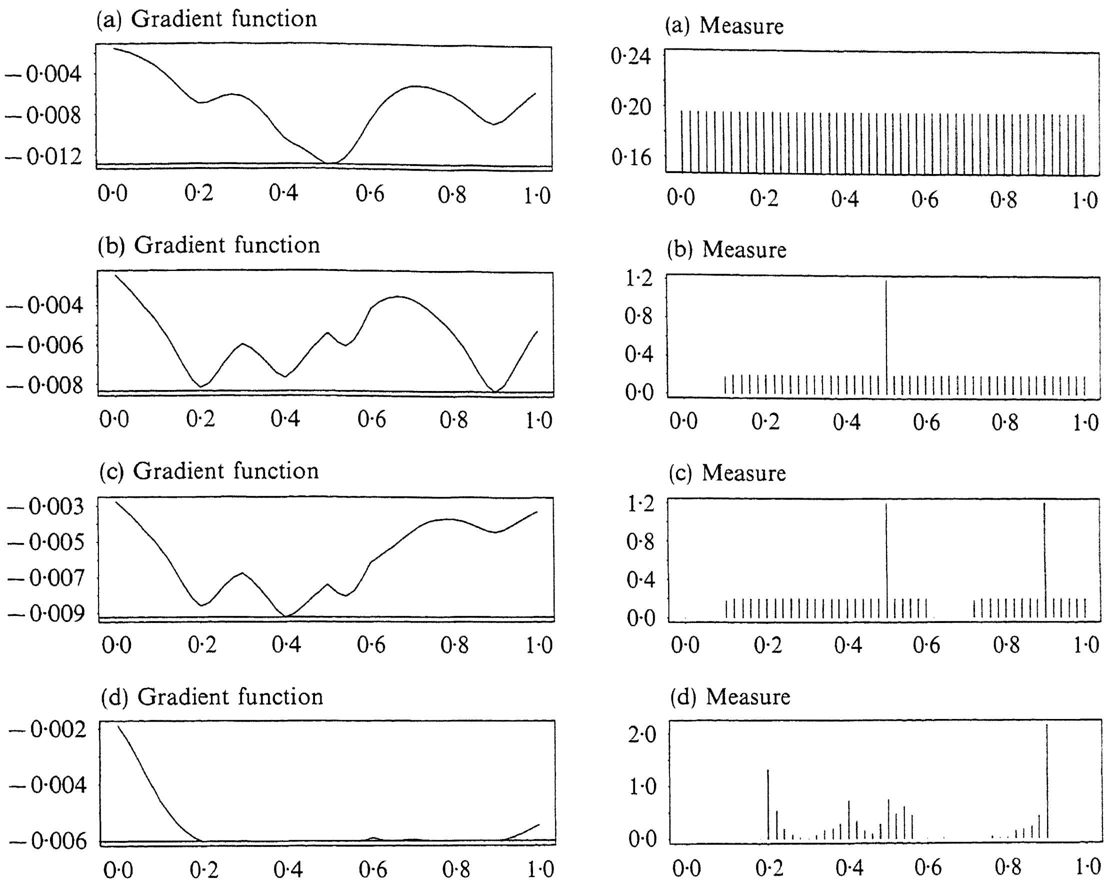

Fig. 1. Plots of the gradient function and measure at several steps of the steepest descent algorithm applied to a one-dimensional problem on a grid of mesh size 0·02. (a) The initial measure μ_0 is uniform over all grid points, $f(\mu_0) = 0{\cdot}03016$. (b) The first descent step of size $\varepsilon = 1$ adds an atom of size ε to μ at the grid point with smallest gradient value, see (a), and eliminates μ at those grid points where the gradient shown in (a) was the largest, $f(\mu_1) = 0{\cdot}02629$. (c) The second descent step of size $\varepsilon = 1$, $f(\mu_2) = 0{\cdot}02243$. (d) The final solution μ^* after 477 steps, $f(\mu^*) = 0{\cdot}01831$.

μ^* for a range of total mass values a; these give a good visual indication of clustering. Although the optimal measures are not hierarchical with respect to the total mass, if a dendrogram is required it can easily be obtained as follows. For each leaf y_j, a 'parent' node in the tree is found by optimising $\{\mu^*(y_k) - \mu^*(y_j)\}/\rho(y_k, y_j)$ subject to the constraint that $\mu^*(y_k) > \mu^*(y_j)$ and possibly a threshold constraint on $\rho(y_k, y_j)$, where μ^* is the density of the continuous measure obtained by spreading each atom of the measure μ^* over the associated pixel centred at that atom. The resulting family trees give the required hierarchy; see Koontz et al. (1976) and Silverman (1986, p. 131). Roughly speaking, cluster boundaries will tend to follow the valleys in the intensity surface.

5·2. *Synthetic examples*

We analyse a synthetic dataset sampled from a stochastic cluster process. The parents follow a Poisson point process with intensity 10; each parent has a Poisson number of daughters with mean 10, scattered independently and uniformly in a disc of radius 0·1 around the parent. After truncation to the unit square, the pattern of 73 points shown in Fig. 2 was obtained.

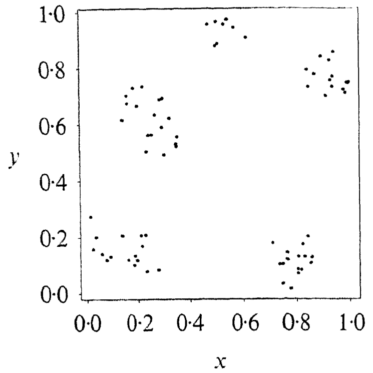

Fig. 2. A synthetic two-dimensional dataset.

Figure 3 shows the results of applying the numerical procedure of the previous section. The optimal measure is shown for a range of total mass levels. If the total mass is small in comparison to the number of data points, the contours of the optimal intensity surface suggest a few large components. If we increase the total mass, these groups split themselves into smaller clusters. Asymptotically, since the high-intensity solution (7) is a power of

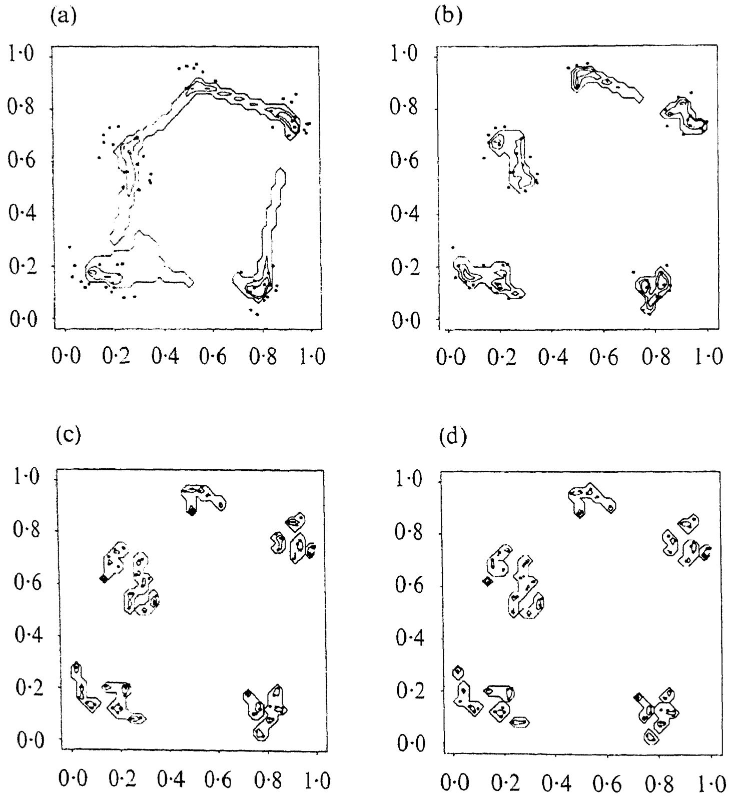

Fig. 3. Contour plots of the optimal measures, with varying total mass a, for the synthetic dataset. The contours are taken at the specified levels:
(a) $a = 10$, levels $= (0{\cdot}00001, 0{\cdot}05, 0{\cdot}1)$;
(b) $a = 20$, levels $= (0{\cdot}00001, 0{\cdot}1, 0{\cdot}2)$;
(c) $a = 70$, levels $= (0{\cdot}00001, 0{\cdot}5, 1{\cdot}0)$;
(d) $a = 100$, levels $= (0{\cdot}00001, 0{\cdot}8, 2{\cdot}0)$.

the daughter density, its contour lines are those of $p_y(.)$, albeit at different levels. It has been observed in numerous simulation experiments that, if the total mass a is approximately half of the number of daughter points, the method leads to optimal measures that describe the cluster structure well.

A more detailed Bayesian analysis based on the cluster process described above and a repulsive Markov prior can be found in van Lieshout (2000).

5·3. *Redwood data*

Figure 4 shows the locations of redwood seedlings extracted from a larger dataset in Strauss (1975). The plot suggests aggregation of the seedlings, which Strauss attributes to the presence of stumps of older redwoods, the positions of which have not been recorded. The tree positions shown in Fig. 4 represent those seedlings falling in region II of Straus (1975, Fig. 1), a roughly triangular area containing almost all of the redwood stumps.

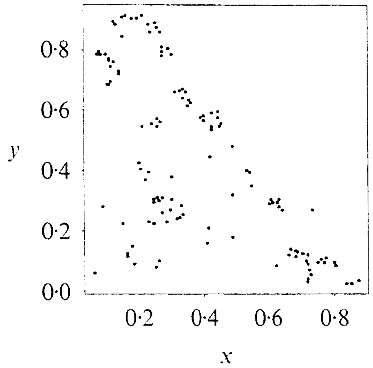

Fig. 4. Locations of redwood seedlings.

In Strauss (1975) a point process model was fitted to the redwood data, later shown in Kelly & Ripley (1976) to be ill defined. Surprisingly, although the even smaller square extracted by Ripley (1977) appears frequently in the spatial statistics literature, the full dataset seems to have been reanalysed only in van Lieshout (1995), where a cluster process was fitted with points scattered according to a Gaussian distribution around parents that are distributed according to a repulsive point process model and the posterior intensity surface of cluster locations was computed. For the smaller dataset, corresponding to the top left corner of Fig. 4, Diggle (1983, pp. 78–81) fitted a Gaussian scatter model with a Poisson parent process using a least squares approach. That yielded an estimated number of 26 stumps, which is implausible from a biological point of view. The least squares approach does not allow for estimation of cluster positions as such. Use of a uniform distribution for the daughters instead of a Gaussian yielded similar results (Diggle, 1978). Finally, Lawson (1993) fitted a similar Gaussian scatter point process, but failure to include a repulsive parent model led to the implausibly large number of 16 parents.

We applied the optimisation algorithm for the problem defined by (5). Figure 5 shows contour plots of several optimal measures with varying total mass a. The choice of a is obviously subjective, and, as in hierarchical clustering algorithms, we recommend consideration of a range of values. As can be seen from Fig. 5, for small values of a a few large components explain most of the mass in the optimal measure; if the value of a is increased, the support of the optimal measure splits into more and more groups.

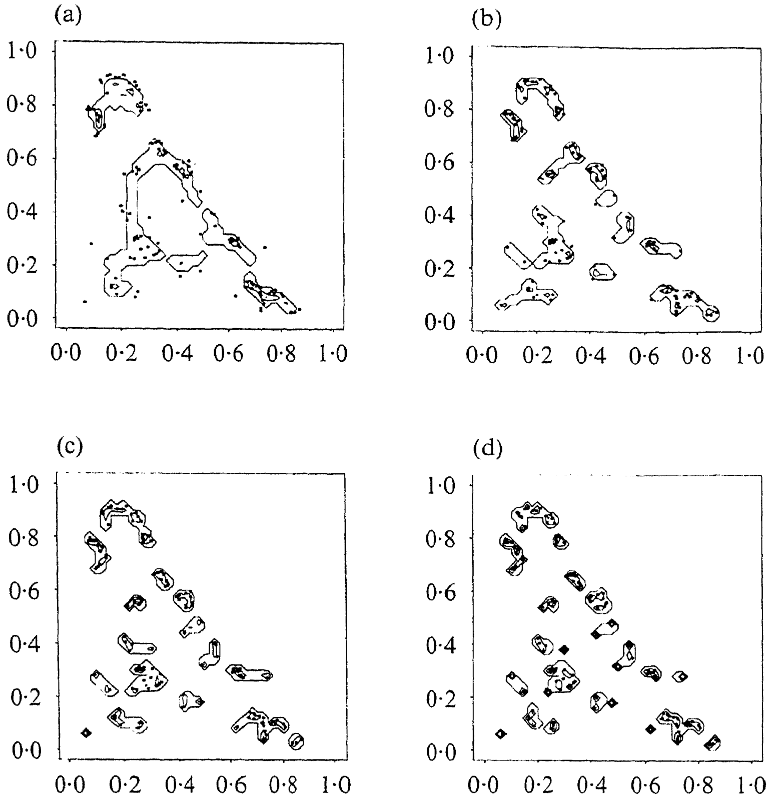

Fig. 5. Contour plots of measures solving (5) for the Redwood data with varying total mass a. The levels of contours are specified: (a) $a = 20$, levels = (0·0001, 0·2, 0·4); (b) $a = 50$, levels = (0·0001, 0·4, 0·8); (c) $a = 100$, levels = (0·0001, 0·6, 1·2); (d) $a = 200$, levels = (0·0001, 1·0, 2·0).

6. Discussion

We have treated the partitioning of a pattern of points into clusters as an optimisation problem in the space of measures by assuming the parent process of cluster centres to be an inhomogeneous Poisson process. Thus, the output of the steepest descent algorithm is the optimal parent intensity measure.

We defined the parent and daughter processes on the same space D, but our approach is equally valid if the parent process is defined on some bounded $E \supseteq D$, a modification that is especially useful whenever edge effects are a concern. Also, the criterion (2) may be replaced by other objective functionals. Additional analysis is necessary in this case to verify that the conditions for the asymptotic results outlined in § 3 hold; see Molchanov & Zuyev (2000a) for details.

In contrast to partition or mixture methods, when we model the cluster centres by a point process the number of cluster centres need not be set in advance nor be decided by ad hoc thresholding as in hierarchical clustering. Instead, the total mass of the intensity function has to be constrained. Since the objective functional (4) is convex, a global optimum is reached, rather than the locally optimal partitions produced by hierarchical or partition-based techniques. Asymptotically, the optimal intensity is a power of the daughter density, so that its peaks and valleys coincide with those of the density. As usual, if one were to estimate the density by kernel estimation (Silverman, 1986) the choice of bandwidth would affect the result.

It should be noted that our model assumptions are very mild indeed. Alternatively, a parametric Markov point process model could be employed, allowing estimation of the model parameters, the posterior parent intensity measure and cluster labels. However, the computational cost is higher than for our steepest descent algorithm, relying on methods based on Monte Carlo or coupling from the past methods; see Baddeley & van Lieshout (2001), van Lieshout (1995, 2000) and van Lieshout & Baddeley (1995) or J. Lund's 1999 Ph.D. thesis from the Royal Veterinary and Agricultural University in Copenhagen for the special case where clusters consist of at most a single point. A similar remark can be made about Bayesian mixture models with a random number of components such as those dealt with in Richardson & Green (1997) and Stephens (2000).

Finally, the optimal measure μ^* can be used as input to a subsequent more detailed analysis. For instance, the spatial Markov model approach requires a reference Poisson point process, and μ^* would be a more natural candidate for its intensity measure than the usual noninformative choice of Lebesgue measure.

Acknowledgement

The research of M. N. M. van Lieshout and I. S. Molchanov was funded by a British Council/Nederlandse Organisatie voor Wetenschappelijk Onderzoek research grant. The authors are grateful to the referee and the associate editor for thoughtful comments.

Appendix

The objective function and the gradient

The objective function. Here we compute the expectation of

$$F(\Pi) = \int_D \rho^2(y, \Pi)\nu(dy)$$

if Π is an inhomogeneous Poisson process on D with intensity measure $\mu(.)$, and $\nu(.)$ denotes a finite measure on D. For (3), $\nu(.)$ assigns equal mass 1 to each data point y_j, for $j = 1, \ldots, m$. Recall that $\rho(y, \Pi)$ is set to the diameter u of D if Π is empty. Then

$$\begin{aligned} E_\mu\{\rho^2(y, \Pi)\} &= \int_0^{u^2} \mathrm{pr}_\mu\{\rho^2(y, \Pi) > t\}\, dt = \int_0^{u^2} \mathrm{pr}_\mu\{\Pi \cap B_{t^{1/2}}(y) = \varnothing\}\, dt \\ &= \int_0^{u^2} \exp[-\mu\{B_{t^{1/2}}(y) \cap D\}]\, dt \end{aligned}$$

and so

$$f(\mu) = E_\mu F(\Pi) = \int_D \int_0^{u^2} \exp[-\mu\{B_{t^{1/2}}(y) \cap D\}]\, dt\, \nu(dy).$$

The gradient. The directional derivative (8) of $f(\mu)$ can be written as

$$-\int_D \int_0^{u^2} \exp[-\mu\{B_{t^{1/2}}(y) \cap D\}]\eta\{B_{t^{1/2}}(y) \cap D\}\, dt\, \nu(dy).$$

To express it as an integral with respect to $\eta(.)$, note that, for $h(t)=\exp[-\mu\{B_{t^{1/2}}(y)\cap D\}]$,

$$\int_0^{u^2} h(t)\eta\{B_{t^{1/2}}(y)\cap D\}\,dt = \int_0^{u^2} h(t)\int_{z\in D:\rho(z,y)\leqslant t^{1/2}} \eta(dz)\,dt = \int_D \eta(dz)\int_{\rho^2(z,y)}^{u^2} h(t)\,dt.$$

Therefore, the gradient of $f(\mu)$ is given by

$$g_\mu(z) = -\int_D \int_{\rho^2(z,y)}^{u^2} \exp[-\mu\{B_{t^{1/2}}(y)\cap D\}]\,dt\,\nu(dy).$$

References

Baddeley, A. J. & Lieshout, M. N. M. van (1993). Stochastic geometry in high-level vision. In *Statistics and Images*, **1** of *Advances in Applied Statistics*, Ed. K. V. Mardia and G. K. Kanji, pp. 231–56, Abingdon: Carfax.

Baddeley, A. J. & Lieshout, M. N. M. van (2001). Extrapolating and interpolating spatial patterns. In *Spatial Cluster Modelling*, Ed. D. Denison and A. B. Lawson. To appear. London: Chapman and Hall.

Banfield, J. D. & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49**, 803–21.

Bowman, A. W. & Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis: the Kernel Approach with S-Plus Illustrations.* Oxford: Oxford University Press.

Diebolt, J. & Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *J. R. Statist. Soc.* B **56**, 363–75.

Diggle, P. J. (1978). On parameter estimation for spatial point processes. *J. R. Statist. Soc.* B **40**, 178–81.

Diggle, P. J. (1983). *Statistical Analysis of Spatial Point Patterns.* London: Academic Press.

Everitt, B. (1974). *Cluster Analysis.* London: Heinemann Educational.

Friedman, H. P. & Rubin, J. (1967). On some invariant criteria for grouping data. *J. Am. Statist. Assoc.* **62**, 1159–78.

Hartigan, J. A. (1975). *Clustering Algorithms.* New York: Wiley.

Jardine, N. & Sibson, R. (1971). *Mathematical Taxonomy.* London: Wiley.

Johnson, R. A. & Wichern, D. W. (1982). *Applied Multivariate Statistical Analysis.* Englewood Cliffs, NJ: Prentice Hall.

Kaufman, L. & Rousseeuw, P. J. (1990). *Finding Groups in Data. An Introduction to Cluster Analysis.* New York: Wiley.

Kelly, F. P. & Ripley, B. D. (1976). On Strauss's model for clustering. *Biometrika* **63**, 357–60.

Koontz, W. L. G., Narendra, P. M. & Fukunaga, K. (1976). A graph-theoretic approach to nonparametric cluster analysis. *IEEE Trans. Comp.* **25**, 936–43.

Lawson, A. (1993). Discussion of a paper by A. F. M. Smith and G. O. Roberts. *J. R. Statist. Soc.* B **55**, 61–2.

Lieshout, M. N. M. van (1995). *Stochastic Geometry Models in Image Analysis and Spatial Statistics*, *CWI Tract*, **108**. Amsterdam: Stichting Mathematisch Centrum, Centrum voor Wiskunde en Informatica.

Lieshout, M. N. M. van (2000). *Markov Point Processes and their Applications.* Singapore: Imperial College Press/World Scientific Publishing.

Lieshout, M. N. M. van & Baddeley, A. J. (1995). Markov chain Monte Carlo methods for clustering of image features. In *Proc. 5th Int. Conf. on Image Process. & Appl., IEE Conference Publication,* **410**. Ed. M. C. Fairhurst et al., pp. 241–5. London: IEE.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proc. 5th Berkeley Symp. Math. Statist. Prob.*, **1**, Ed. L. LeCam and J. Neyman, pp. 281–97. Berkeley, CA: Univ. Calif. Press.

Mardia, K. V., Kent, J. T. & Bibby, J. M. (1979). *Multivariate Analysis.* London: Academic Press.

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

Molchanov, I. S. & Zuyev, S. A. (2000a). Variational analysis of functionals of a Poisson process. *Math. Oper. Res.* **25**, 485–508.

Molchanov, I. S. & Zuyev, S. A. (2000b). Variational calculus in space of measures and optimal design. In *Optimum Design* 2000: *Prospects for the New Millenium*, Ed. A. Atkinson, B. Bogacka and A. Zhigljavsky, pp. 79–90. Dordrecht: Kluwer.

Molchanov, I. S. & Zuyev, S. A. (2001). Steepest descent algorithms in space of measures. *Statist. Comp.* To appear.

Okabe, A., Boots, B., Sugihara, K. & Chiu, S. N. (2000). *Spatial Tessellations — Concepts and Applications of Voronoi Diagrams*, 2nd ed. Chichester: Wiley.

PROPP, J. G. & WILSON, D. B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Struct. Algor.* **9**, 223–52.

RICHARDSON, S. & GREEN, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with Discussion). *J. R. Statist. Soc.* B **59**, 731–92.

RIPLEY, B. D. (1977). Modelling spatial patterns (with Discussion). *J. R. Statist. Soc.* B **39**, 172–212.

SCOTT, A. J. & SIMONS, M. J. (1971). Clustering methods based on likelihood ratio criteria. *Biometrics* **27**, 387–97.

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* London: Chapman and Hall.

STEPHENS, M. (2000). Bayesian analysis of mixture models with an unknown number of components — an alternative to reversible jump methods. *Ann. Statist.* **28**, 40–74.

STRAUSS, D. J. (1975). A model for clustering. *Biometrika* **63**, 467–75.

WARD, J. H. (1963). Hierarchical groupings to optimize an objective function. *J. Am. Statist. Assoc.* **58**, 236–44.

[*Received June* 2000. *Revised May* 2001]