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, UK

ilya@stats.gla.ac.uk

S.A. Zuyev

Department of Statistics and Modelling Science,
University of Strathclyde, Glasgow G1 1XH, U.K.

sergei@stams.strath.ac.uk

Summary

In this paper, 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. Thus, we derive a Ward style 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 style 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. Discretising over a grid, at each iteration step the current intensity function is increased at the points where the gradient is minimal at the expense of regions with a large gradient value. The algorithm is applied to both synthetic data (a toy 1-dimensional example and a simulation from a popular spatial cluster model) as well as to a real life data set concerning the positions of redwood seedlings from Strauss (1975). Finally, the relative merits of our approach compared to classical hierarchical and agglomerative clustering techniques as well as modern model based clustering methods using (Markov) point processes and mixture distributions are discussed.

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 dividing data 'points' representing individuals or objects into groups. Such techniques are widely used in exploratory data analysis and implemented in all major commercial statistical packages.

Classical clustering techniques are often *hierarchical* in nature, building a tree or the so-called dendrogram based on some distance measure. Thus, starting from clusters consisting of a single point, at each step the pair of clusters that are closest to each other are merged until arriving at a single cluster containing all data points. The distance between two clusters may be defined in a various ways, e. g., as the minimum distance from a member of one group to a point of the other as in the single linkage algorithm (Sneath, 1957), the maximum such distance as in the complete linkage algorithm or some average between pairs of points chosen from the two groups. Alternatively, 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, e. g., Hartigan (1975) and Jardine & Sibson (1971).

In contrast, *partition techniques* are based on iteratively allocating points to clusters. Fixing in advance the number of clusters (say k), initially k points are chosen as cluster centres and all other points are assigned to the nearest centre. Re-allocation of a point is then based on 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, 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, although their efficiency depends on the shape and other characteristics of the clusters. 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. Thus, writing $(y_1, \ldots, y_m)$ for the vector of observations, let $\phi(j) \in \{1, \ldots, k\}$ denote the component label of y_j. Since $\phi(\cdot)$ is not observed, we are in a missing data situation, and the goal is to estimate the missing component indicators $\phi(\cdot)$, as well as any unknown model parameters θ. More specifically, let $f_i(\cdot;\theta)$ be the density for the i^{th} component. Then, assuming independence, the complete data likelihood is

$$L(\theta;\phi) = \prod_{j=1}^{m} f_{\phi(j)}(y_j;\theta).$$

If each component is normally distributed with mean μ_i and the same covariance matrix Σ, the log likelihood reduces to

$$\begin{aligned} l(\theta;\phi) &= \log L(\theta;\phi) && (1) \\ &= -\frac{1}{2}\sum_{i=1}^{k} n_i \log|\Sigma| - \frac{1}{2}\sum_{i=1}^{k}\sum_{y_j:\phi(j)=i} (y_j-\mu_i)^\top \Sigma^{-1}(y_j-\mu_i). \end{aligned}$$

If $\Sigma = \sigma^2 I$ and the means μ_i are estimated by the sample means of the components, optimisation of (1) leads to the Ward criterion. For general Σ, we re-obtain the Friedman and Rubin criterion. More details and variations on this theme can be found in Banfield & Raftery (1993), Diebolt & Robert (1994), Green & Richardson (1997), McLachlan & Basford (1988) and Scott & Simons (1971). 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.

A disadvantage of most approaches outlined above is that the number of clusters is decided 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. These difficulties are avoided by taking a point process approach. For instance, Baddeley & Van Lieshout (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. Currently, Baddeley and Van Lieshout, in collaboration with amongst others N. Fisher of CSIRO are generalising such an approach to spatial interpolation and extrapolation problems. 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, cf. Van Lieshout (2000). See also Lawson (1993) and Lund (1999).

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 assumptions are made. The total intensity of the point process of parents is prefixed, and its spatial distribution is chosen so as to minimise the Ward criterion.

The plan of this paper is as follows. In § 2, we propose considering 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), Molchanov & Zuyev (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 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 $\mathbf{y} = \{y_1, \ldots, y_m\}$ 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) $\mathbf{x} = \{x_1, \ldots, x_k\}$, $k = 0, 1, \ldots,$ explaining the data. This can be done by minimising the following Ward-style criterion

$$\left[\text{trace}\left\{\sum_{x_i \in \mathbf{x}} \sum_{y_j \in Z_{\mathbf{x}}(x_i)} (y_j - x_i)(y_j - x_i)^\top\right\}\right] = \sum_{x_i \in \mathbf{x}} \sum_{y_j \in Z_{\mathbf{x}}(x_i)} \rho^2(x_i, y_j)\,, \quad (2)$$

where $Z_{\mathbf{x}}(x_i)$ is the collection of points in the plane closer to x_i than to any other parent $x_j \in \mathbf{x}$, $j \neq i$. In other words, $Z_{\mathbf{x}}(x_i)$ are the Voronoi cells generated by the set $\mathbf{x}$, see Okabe et al. (2000). Minimisation problems for the functional (2) 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) of a configuration $\mathbf{y}$ in agglomerative clustering, or the mailbox problem discussed by Okabe et al. (2000, Chapter 9). In all these instances the number k has to be predetermined and steepest descent type minimisation algorithms are used to find a configuration $\mathbf{x}$ that minimises (2). This involves optimising in a space of moderate dimension of 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 $\mathbf{x}$ 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 non-homogeneous 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 number 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. As $\mu(D)$ is finite by assumption, the total number of points in Π is almost surely finite as well.

Now, replacing $\mathbf{x}$ with Π in (2) and taking the expectation of the random variable thus obtained yields our objective functional that can be written as

$$f(\mu) = E_\mu \left\{ \sum_{x_i \in \Pi} \sum_{y_j \in Z_{\mathbf{x}}(x_i)} \rho^2(x_i, y_j) \right\} . \tag{3}$$

The subscript μ under the expectation or probability signs 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 (3) (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 daughters represent subscribers of the network, the parents correspond to stations. Writing $\rho(y, \Pi)$ for the minimal distance between y and a point of Π, (3) can be reformulated as

$$f(\mu) = \sum_{j=1}^{m} E_\mu \left\{ \rho^2(y_j, \Pi) \right\} . \tag{4}$$

Note that with positive probability Π is empty, in which case the distance $\rho^2(y_j, \Pi)$ in (4) is ill-defined. Thus, we must assign some value u to $\rho(y_j, \emptyset)$. Since we are dealing with minimisation of $f(\mu)$, a natural choice for u is the diameter of D, i.e. 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 (4), yielding

$$f(\mu) = \sum_{j=1}^{m} \int_0^{u^2} \exp\left\{ -\mu(B_{t^{1/2}}(y_j) \cap D) \right\} dt \tag{5}$$

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 non-negative measures and can be extended using (5) to signed measures, although without immediate probabilistic interpretation readily available. An important implication of (5) is that the objective functional is convex in μ, that is for every pair of measures μ and η, and for each $c \in [0, 1]$,

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

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 $a > 0$. Doing so, the minimisation problem can be written as

$$f(\mu) \mapsto \min\,, \quad \mu(D) = a\,. \tag{6}$$

Further constrains on μ may be added to incorporate additional information about the parents, e.g. by weighing 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 of 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, the so-called high intensity solution $\bar{\mu}$.

In our context, suppose that the daughter points $\mathbf{y}$ have been sampled from a distribution with probability density $p_y(\cdot)$, perhaps obtained by kernel smoothing (Bowman & Azzalini, 1997) of $\mathbf{y}$. Then the objective function (4) transforms into

$$f(\mu) = \int_D E_\mu \left[\rho^2(z, \Pi)\right] \, p_y(z) \, dz \,. \tag{7}$$

The same functional (7) 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:

$$p_{\bar{\mu}}(z) \propto (p_y(z))^{d/(d+2)} \,. \tag{8}$$

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 e.g. the k-mean algorithm or the constrained optimisation problem (6), that can be further improved using descent methods.

4 Steepest descent algorithm

The minimisation of functionals of measures can be done 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 suitably chosen (signed) measure η such 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, which is defined by

$$\lim_{t \downarrow 0} t^{-1}\{f(\mu + t\eta) - f(\mu)\} = \int g_\mu(z)\eta(dz) \,. \tag{9}$$

The function $g_\mu(\cdot)$ is called the gradient of $f(\mu)$. For the objective function $f(\mu)$ given by (3), the gradient equals

$$g_\mu(z) = -\sum_{j=1}^{m} \int_{\rho^2(y_j,z)}^{u^2} \exp\left\{-\mu(B_{t^{1/2}}(y_j) \cap D)\right\} dt\,. \tag{10}$$

A derivation of this expression can be found in the Appendix. Note that the gradient (10) resembles $f(\mu)$ as in (5), 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, 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 (9) one should place an atom of mass ε at the minimum of $g_\mu(\cdot)$ (or distribute it between several global minima if they exist). Similarly, the negative mass $-\varepsilon$ should ideally be placed at the maximum of $g_\mu(\cdot)$, which amounts to taking away an amount 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 at all. Thus, we should remove mass from regions where $g_\mu(z)$ is large, until an amount ε has been taken. More precisely, Molchanov & Zuyev (2000b), 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 $D_t = \{x \in D : \; g_\mu(z) \geq t\}$ for a suitable $t \geq 0$, 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 $\mu(D_t) \leq \varepsilon$ 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, ε mass 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 (5) 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 Problem (6) to have a solution can be formulated as

$$\begin{cases} g_{\mu^*}(z) = c & \mu^* - \text{a.e.}\,, \\ g_{\mu^*}(z) \geq c & \text{for all } z\,, \end{cases} \tag{11}$$

for some constant c, given that measure μ^* minimises $f(\mu)$ over all non-negative measures with the given total mass. The constant c is, in fact, the Lagrange multiplier for the corresponding constrained optimisation problem. The necessary condition (11) 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 at all other points z in the support of μ, $g_\mu(z)$ is at least c. The described algorithm is implemented in Splus and R-languages, see, e. g., Venables & Ripley (1994) about statistical analysis using Splus/R. The code is available on the web at

```
www.stats.gla.ac.uk/~ilya
www.stams.strath.ac.uk/~sergei
```

and distributed as an R-language bundle `mesop`. Data sets used in the following examples can be obtained from the same source.

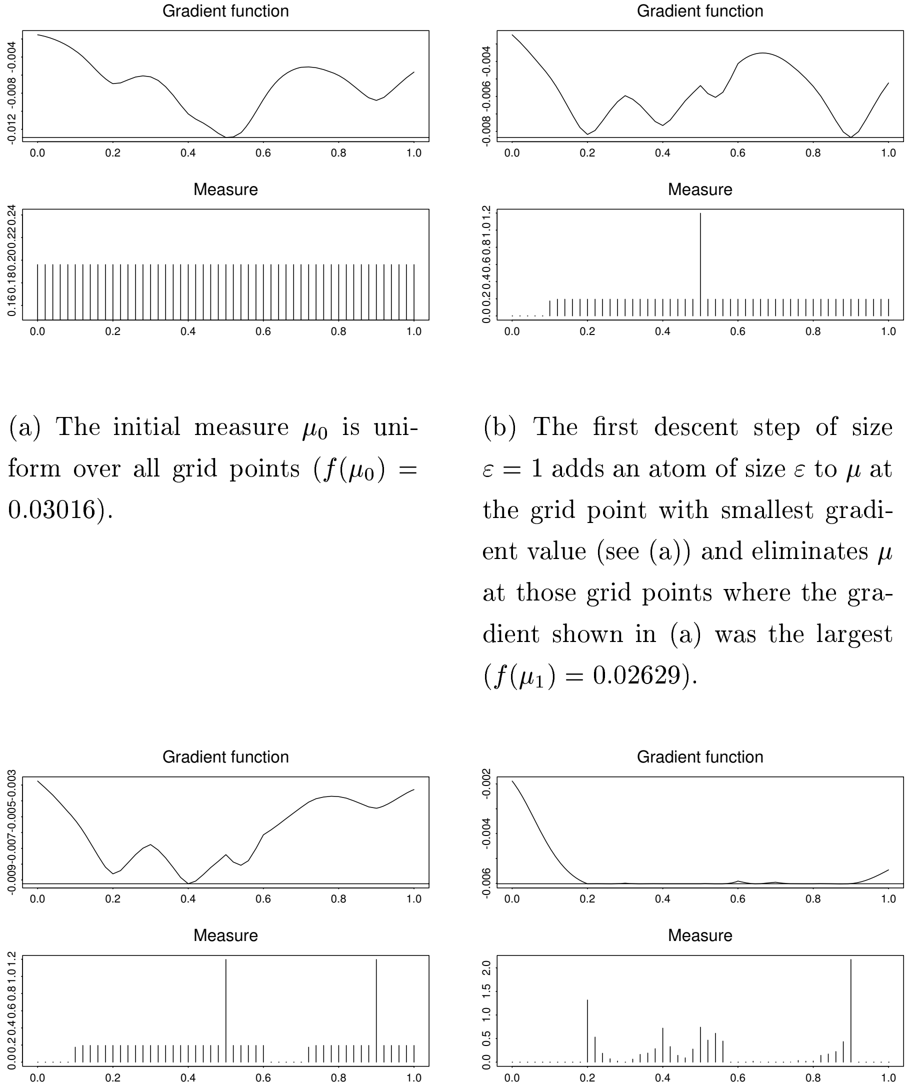
As an illustration, Figure 4.1 shows several steps of the steepest descent algorithm applied to a one-dimensional problem on $D = [0, 1]$ with $\mathbf{y} = \{0.2, 0.4, 0.5, 0.55, 0.9\}$ and the measure's total mass is fixed at $a = 10$. The parent space is discretised into a grid with mesh size s (in our example $s = 0.02$), and the intensity measure μ is atomic and supported on the grid. Note, however, that the data points $\mathbf{y}$ do not necessarily lie on the grid (e.g. the point 0.55 here). Consider a daughter point $y_j \in \mathbf{y}$. Then, the inner integrand in the objective functional (5) is a step function in t, with break points at the squared distances from y_j to grid points. Thus, if necessary rearranging the indices of the grid points in such a way that $\rho(x_1, y_1) \leq \rho(x_2, y_1) \leq \cdots \leq \rho(x_n, y_1)$, the integral $\int_0^{u^2} \exp\{-\mu(B_{t^{1/2}}(y_1) \cap D)\}\, dt$ can be written as

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

A similar formula holds for the other summands in (5), 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, updates of the gradient and objective functional are easy to perform.

5 Examples

In all examples below we used the steepest descent algorithm described in Section 4 on the unit square $[0, 1] \times [0, 1]$ in the plane. The measures were defined on a uniform grid with mesh size $s = 0.02$ 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 fast enough (about one second per step on a SUN ULTRA 10 Workstation 360 MHz) for $\mathbf{y}$ consisting of 123 points as in the case study described below. Plausible results are obtained already for the tolerance level $\delta = 0.01$ in about 100 steps, while $\delta = 0.0001$ requires considerably more steps to be done (in the range of several thousands depending on the total mass of μ).



(a) The initial measure μ_0 is uniform over all grid points ($f(\mu_0) = 0.03016$).

(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 ($f(\mu_1) = 0.02629$).

(c) The second descent step of size $\varepsilon = 1$ ($f(\mu_2) = 0.02243$).

(d) The final solution $\bar{\mu}$ after 477 steps ($f(\bar{\mu}) = 0.01831$).

Figure 4.1: Several steps of the steepest descent algorithm applied to a one-dimensional problem on a grid of mesh size 0.02.

5·1 *Synthetic examples*

We analyse a synthetic data set 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 Figure 5.1 was obtained.

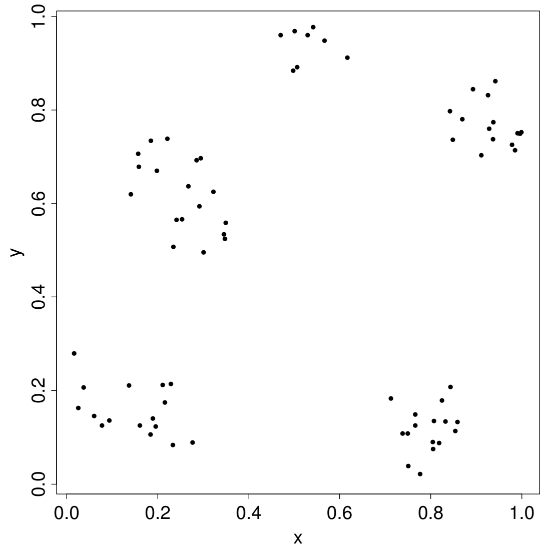
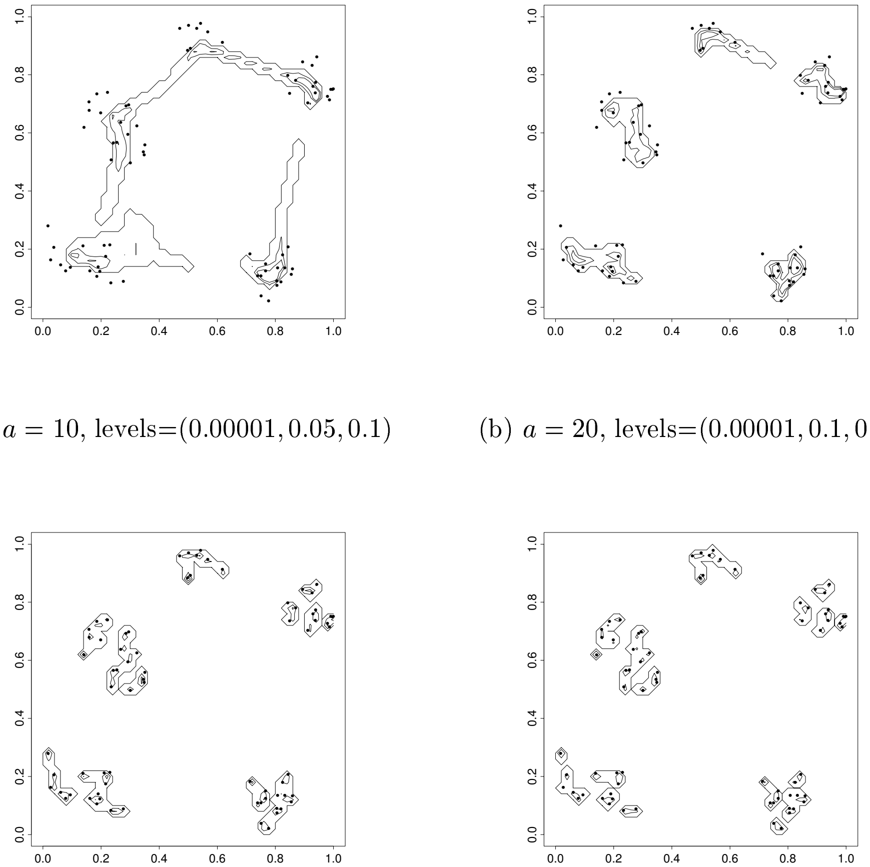


Figure 5.1: A synthetic two-dimensional data set.

Figure 5.2 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. Increasing the total mass, these groups split themselves in smaller clusters. A more detailed Bayesian analysis based on the cluster process described above and a repulsive Markov prior can be found in Van Lieshout (2000).



(a) $a = 10$, levels=$(0.00001, 0.05, 0.1)$

(b) $a = 20$, levels=$(0.00001, 0.1, 0.2)$

(c) $a = 70$, levels=$(0.00001, 0.5, 1.0)$

(d) $a = 100$, levels=$(0.00001, 0.8, 2.0)$

Figure 5.2: Contour plots of the optimal measures (with varying total mass a) for the synthetic data set. The contours are taken at the specified levels.

5·2 *Redwood data*

Figure 5.3 shows the locations of redwood seedlings extracted from a larger data set in Strauss (1975). The plot suggests aggregation of the seedlings, which Strauss attributes to the presence of stumps of older redwoods, whose position has not been recorded. The tree positions shown in Figure 5.3 contains those seedlings in region II of Strauss (1975, Figure 1, p. 474), a roughly triangular area containing almost all of the redwood stumps.

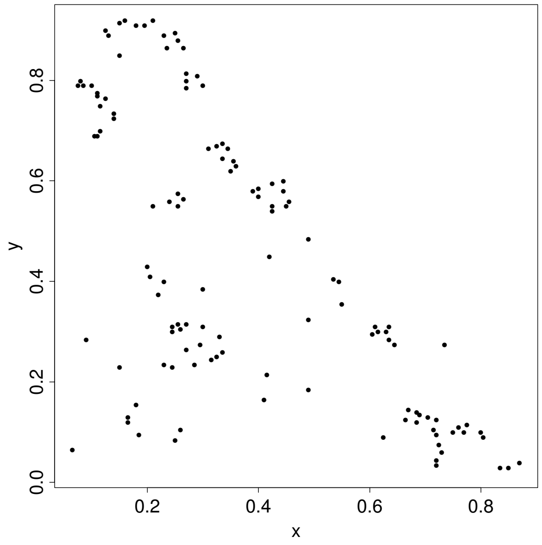


Figure 5.3: Locations of redwood seedlings.

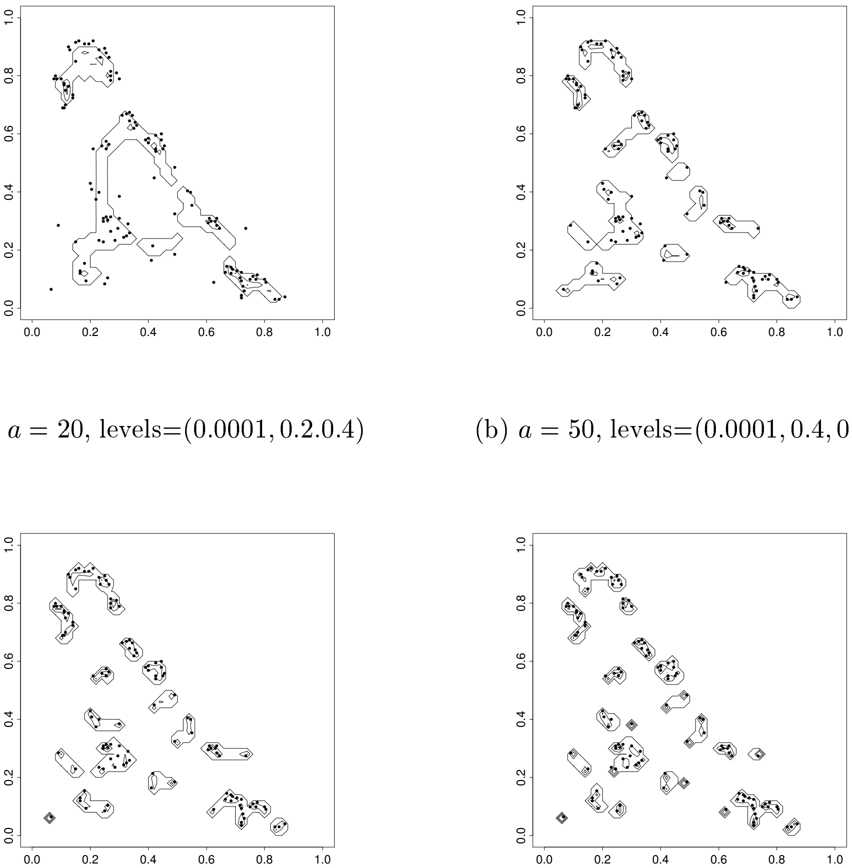
In Strauss (1975) a point process model was fitted to the redwood data, later shown in Kelly & Ripley (1976) to be ill-defined. Surprisingly, although the yet smaller square extracted by Ripley (1977) appears frequently in the spatial statistics literature, the full data set seems to have been reanalysed 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 data set (corresponding to the top left corner of Figure 5.3) previous analyses include Diggle (1983), where a Gaussian scatter model with a Poisson parent process was fitted 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. Using a uniform distribution for the daughters instead of a Gaussian one yielded similar results, see Diggle (1978). Finally, Lawson (1993) fitted a similar Gaussian scatter point process, but failing to include a repulsive parent model led to the implausibly large number of 16 parents.

Below we report the results of using the optimisation algorithm for the Problem (6). Figure 5.4 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 to consider a range of values. As it can be seen from Figure 5.4, for small values of a, a few large components explain most of the mass in the optimal measure; increasing the value of a, the support of the optimal measure splits into more and more groups.

6 Discussion

In this paper, we treated partitioning 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, the contour lines of which provide an indication of the plausible clusters.

We defined the parent and daughter processes on the same space D, but our approach is equally valid if the parent process were defined on some



(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=$(0.0001, 1.0, 2.0)$

Figure 5.4: Contour plots of measures solving (6) for the Redwood data with varying total mass a. The levels of contours are specified.

bounded $E \supseteq D$, a modification that is especially useful whenever edge effect are a concern. Also, the Ward style pooled within groups sum of squares criterion (3) may be replaced by other objective functionals. Additional analysis is necessary in this case to verify the validity of the conditions for the asymptotic results outlined in Section 3 to hold, see Molchanov & Zuyev (2000a) for details.

In contrast to partition or mixture methods, the advantage of modelling the cluster centres by a point process is that the number of cluster centres need not be set in advance (nor be decided by ad hoc thresholding as in hierarchical clustering). Furthermore, since the objective functional (5) is convex, a global optimum is reached, rather than the locally optimal partitions produced by hierarchical or partition-based techniques. 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 Monte Carlo or coupling from the past methods, cf. Van Lieshout (1995), Van Lieshout (2000) and Van Lieshout & Baddeley (1995) or Lund (1999) for the special case where clusters consist of at most a single point. A similar remark can be made for Bayesian mixture models with a random number of components such as those in Green & Richardson (1997).

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 the intensity measure than the usual non-informative Lebesgue measure.

Acknowledgements

Van Lieshout's work was carried out under project PNA4.3 'Stochastic geometry'. The research of the first two authors was funded by a British Council/NWO Research grant JRP544/BR62-477.

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 $\mu(\cdot)$, and $\nu(\cdot)$ denotes a finite measure on D. For (4), $\nu(\cdot)$ assigns equal mass 1 to each data point y_j, $j = 1, \ldots, m$. Recall that $\rho(z, \Pi)$ is set to the diameter u of D if Π is empty. Then

$$\begin{aligned}
E_\mu F(\Pi) &= E_\mu \left(u^2\, \mathbb{1}_{\Pi=\emptyset}\right) \nu(D) + E_\mu \left\{ \int_D \rho^2(y, \Pi)\ \mathbb{1}_{X\neq\emptyset}\ \nu(dy) \right\} \\
&= u^2\, e^{-\mu(D)} \nu(D) + \int_D \int_0^\infty \mathrm{pr}_\mu\{\rho^2(y, \Pi) > t; \Pi \neq \emptyset\}\, dt\, \nu(dy) \\
&= u^2\, e^{-\mu(D)} \nu(D) \\
&\qquad + \int_D \int_0^\infty \mathrm{pr}_\mu[\Pi \cap B_{t^{1/2}}(y) = \emptyset; \Pi \cap \{D \setminus B_{t^{1/2}}(y)\} \neq \emptyset]\, dt\, \nu(dy) \\
&= u^2\, e^{-\mu(D)} \nu(D) + \int_D \int_0^\infty e^{-\mu\{B_{t^{1/2}}(y) \cap D\}} \left[1 - e^{-\mu\{D \setminus B_{t^{1/2}}(y)\}}\right]\, dt\, \nu(dy)\,.
\end{aligned}$$

Note that when t exceeds u^2, the inner integrand vanishes. Thus,

$$\begin{aligned}
f(\mu) = E_\mu F(\Pi) &= u^2\, e^{-\mu(D)}\, \nu(D) + \int_D \int_0^{u^2} \left[e^{-\mu\{B_{t^{1/2}}(y) \cap D\}} - e^{-\mu(D)}\right]\, dt\, \nu(dy) \\
&= \int_D \int_0^{u^2} e^{-\mu\{B_{t^{1/2}}(y) \cap D\}} dt\, \nu(dy)\,.
\end{aligned}$$

The gradient. Here we calculate the gradient. Firstly, the directional derivative of $f(\mu)$ can be written as

$$\lim_{s\downarrow 0} s^{-1}\left(\int_D\int_0^{u^2}\left[e^{-\mu\{B_{t^{1/2}}(y)\cap D\}-s\eta\{B_{t^{1/2}}(y)\cap D\}}-e^{-\mu\{B_{t^{1/2}}(y)\cap D\}}\right]dt\,\nu(dy)\right)$$

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

To express $f(\mu)$ as an integral with respect to $\eta(\cdot)$, note that

$$\int_0^{u^2}e^{-\mu\{B_{t^{1/2}}(y)\cap D\}}\eta\{B_{t^{1/2}}(y)\cap D\}\,dt=\int_0^{u^2}e^{-\mu\{B_{t^{1/2}}(y)\cap D\}}\int_{z\in D:\,\rho(z,y)\le t^{1/2}}\eta(dz)\,dt$$

$$=\int_D\eta(dz)\int_{\rho^2(z,y)}^{u^2}e^{-\mu\{B_{t^{1/2}}(y)\}}\,dt\,.$$

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

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

References

Baddeley, A. J. & Van Lieshout, M. N. M. (1993). Stochastic geometry in high-level vision. In Mardia, K. V. & Kanji, G. K., editors, *Statistics and images*, volume 1 of *AdVances in Applied Statistics*, pages 231–256, Carfax. Abingdon.

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

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

Diebolt, J. & Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *J. Roy. Statist. Soc. Ser. B* **56**, 363–375.

DIGGLE, P. J. (1978). On parameter estimation for spatial point processes. *J. Roy. Statist. Soc. Ser. B* **40**, 178–181.

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

EVERITT, B. (1974). *Cluster Analysis.* Heinemann Educational Books, London.

FRIEDMAN, H. P. & RUBIN, J. (1967). On some invariant criteria for grouping data. *J. Amer. Statist. Assoc.* **62**, 1159–1178.

GREEN, P. & RICHARDSON, S. (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. Roy. Statist. Soc. Ser. B* **59**, 731–792.

HARTIGAN, J. A. (1975). *Clustering Algorithms.* Wiley.

JARDINE, N. & SIBSON, R. (1971). *Mathematical Taxonomy.* Wiley, London.

JOHNSON, R. A. & WICHERN, D. W. (1982). *Applied Multivariate Statistical Analysis.* Prentice Hall, New Jersey.

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

KELLY, F. P. & RIPLEY, B. D. (1976). On Strauss's model for clustering. *Biometrika* **63**, 357–360.

LAWSON, A. (1993). Discussion contribution. *J. Roy. Statist. Soc. Ser. B* **55**, 61–62.

LUND, J. (1999). *Statistical Inference and Perfect Simulation for Point Processes with Noise.* PhD thesis, The Royal Veterinary and Agricultural University, Copenhagen.

MACQUEEN, J. (1967). Some methods for classification and analysis of multivariate observations. In LeCam, L. & Neyman, J., editors, *Proceedings of the fifth Berkeley Symposium on mathematical statistics and probability,* volume 1, pages 281–297.

MARDIA, K. V., KENT, J. T. & BIBBY, J. M. (1979). *Multivariate Analysis.* Academic Press, London.

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

MOLCHANOV, I. & ZUYEV, S. (2000a). Variational analysis of functionals of a Poisson process. *Math. Oper. Res.* To appear.

MOLCHANOV, I. & ZUYEV, S. (2000b). Variational calculus in space of measures and optimal design. In Atkinson, A. C., Bogacka, B. & Zhigljavsky, A., editors, *Optimum Experimental Design: Prospects for the New Millennium.* Kluwer, Dordrecht. To appear.

MOLCHANOV, I. & ZUYEV, S. (2001). Steepest descent algorithms in space of measures. *Statist. Comput.* **???**, ??? Submitted.

OKABE, A., BOOTS, B., SUGIHARA, K. & CHIU, S. N. (2000). *Spatial Tessellations — Concepts and Applications of Voronoi Diagrams.* Wiley, Chichester, second edition.

PROPP, J. G. & WILSON, D. B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures Algorithms* **9**, 223–252.

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

SCOTT, A. J. & SIMONS, M. J. (1971). Clustering methods based on likelihood ratio criteria. *Biometrics* **27**, 387–397.

SNEATH, P. H. A. (1957). The application of computers to taxonomy. *Journal of General Microbiology* **17**, 201–226.

STRAUSS, D. J. (1975). A model for clustering. *Biometrika* **63**, 467–475.

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

VAN LIESHOUT, M. N. M. (2000). *Markov Point Processes and their Ap-*

plications. Imperial College Press/World Scientific Publishing, Where published.

Van Lieshout, M. N. M. & Baddeley, A. J. (1995). Markov chain Monte Carlo methods for clustering of image features. In *Proceedings of the fifth international conference on image processing and its applications*, volume 410 of *IEE Conference Publication*, pages 241–245, London. IEE.

Venables, W. N. & Ripley, B. D. (1994). *Modern Applied Statistics with S-PLUS.* Springer, New York.

Ward, J. H. (1963). Hierarchical groupings to optimize an objective function. *J. Amer. Statist. Assoc.* **58**, 236–244.