



ELSEVIER

Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

Computational Statistics & Data Analysis 46 (2004) 33–56

COMPUTATIONAL
STATISTICS
& DATA ANALYSIS

www.elsevier.com/locate/csda

A network flow approach in finding maximum likelihood estimate of high concentration regions

Xiaoming Huo^{*,1}, Jye-Chyi Lu

*School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta,
GA 30332-0205, USA*

Received 31 December 2002; received in revised form 21 May 2003

Abstract

A maximum likelihood estimation (MLE) method of high density regions in spatial point processes is introduced. The method is motivated from a network flow approach for flexibly incorporating geometric restrictions in computing the MLEs. An easy-to-implement computational algorithm having a low order of complexity is provided. Simulation studies show that it performs very well in many difficult situations, and reaches the global optimality. Two real data sets illustrate the applicability of the proposed method.

© 2003 Elsevier B.V. All rights reserved.

Keywords: Envelope; High concentration region; High density region; Point processes; High activity regions

1. Introduction

Finding high concentration (or density) regions (HCRs) in spatial point processes is of interest in many applications, including data mining. The following briefly show two examples: (1) detection of minefield based on aerial observations (e.g., Allard and Fraley, 1997), and (2) analysis of seismic data in North California (e.g., Dasgupta and Raftery, 1998). Several image segmentation problems are also connected to this topic. See Geiger et al. (1995) and Jermyn and Ishikawa (1999) for examples.

In the HCR literature, Silverman (1986) introduced a method of finding a high density region based on a combination of density estimation and thresholding. Dasgupta and Raftery (1998) proposed a model-based method with illustrations of several data

* Corresponding author. Tel.: +1-404-385-0354.

E-mail address: xiaoming.huo@isye.gatech.edu (X. Huo).

¹ Dr. Huo is partially supported by NSF grant 0140587.

sets. Some other methods include Voronoi-tessellation based method by Allard and Fraley (1997), and a K-nearest-neighbor (K-NN) approach by Byers and Raftery (1998). Picard and Bar-Hen (2000) reported a comparison study of several existing methods. Allard and Fraley (1997) connects the search of HCRs to the maximum likelihood estimation (MLE). The last work is insightful; because the MLE has many known optimal statistical properties.

This paper introduces an alternative approach based on designing a *digraph* for the observed data. This digraph will facilitate a search of cycles to optimize an objective function that is equivalent to the one in the MLE. This eventually renders a method to find the HCRs that satisfy given geometric constraints. Our method presents the search of the HCR, MLE, and contour of an envelope (see Section 2 for definitions) in a unified framework. Comparing with existing methods, our method has the following three advantages:

- (1) *Ability of flexibly incorporating geometric conditions on the regions.* This will be done through defining the connectivity of the digraph. See Section 2 for details.
- (2) *Fast algorithm.* The worst-case order-of-complexity is shown to be $O(n^3)$, which is at worst a polynomial time algorithm.
- (3) *Generalization to solving several related problems.* In this paper, we focused on solving the MLE problem. Our method can be generalized to solve other related problems. See Sections 2 and 4 for examples.

The main idea of our method comes from the network flow approach, which is used widely in the *operations research* (OR) community, e.g., Ahuja et al. (1993). By revisiting the data taken from the literature, our studies show a great potential of applying the network flow approach to solve a difficult data mining problem of locating the HCR. Section 2 motivates the study by formulating the equivalence of three related problems; all of them are solvable by the proposed method. Section 3 describes an easy-to-implement algorithm. Numerical studies are reported in Section 4. Section 5 concludes this article and discusses some future work.

2. Problem formulation

2.1. Motivation—three related problems

Our method is motivated from solving the following three related problems:

- (1) finding the *envelope*, which is the contour of a probability density function,
- (2) locating the high concentration region (*HCR*), and
- (3) computing the *MLE* of a high concentration region.

The notations *envelope*, *HCR* and *MLE* will be used to refer to these three problems, respectively.

Consider a spatial point process with data occurring as “points” in a Euclidean space. To simplify the description, this article will focus on the two-dimensional case, in which

data are described as points having Cartesian coordinates (x_i, y_i) , $i = 1, 2, \dots, N$, where N is the total number of observations. The probability density function (p.d.f.) of this point process is denoted by $f(x, y)$. The probability of having an observation in region A is $P\{A\} = \int_A f(x, y) dx dy$.

- (1) *Envelope*. Picard and Bar-Hen (2000) studied the following problem in their data analysis problem. For a given α ,

$$\begin{aligned} \max_c \quad & c \\ \text{s.t.} \quad & P\{A(c)\} \geq 1 - \alpha, \end{aligned} \quad (2.1)$$

where $A(c) = \{(x, y) : f(x, y) > c\}$, and $P\{A(c)\}$ is defined as the probability of having an observation in $A(c)$. If $f(x, y)$ is a continuous function, then the boundary of A (denoted by ∂A) is a closed curve, or a set of non-intersecting closed curves. The ∂A is called an *envelope*, which also is a *contour*.

- (2) *High concentration region (HCR)*. A high concentration region is a subspace that has “high probability” of containing these points. We study the following problem: for a given α ,

$$\begin{aligned} \min_A \quad & |A| \\ \text{s.t.} \quad & P\{A\} \geq 1 - \alpha, \end{aligned} \quad (2.2)$$

where $|A| = \int_A 1 dx dy$ is the area of the region A .

- (3) *Maximum likelihood estimate (MLE)*. Allard and Fraley (1997) studied the following problem in estimating features in a spatial point process:

$$\max_A \left(P\{A\} \ln \frac{P\{A\}}{|A|} + (1 - P\{A\}) \ln \frac{1 - P\{A\}}{|\Omega| - |A|} \right), \quad (2.3)$$

where Ω and $|\Omega|$ denote the entire domain of f and its size. When the underlying density function is a mixture of two uniform distributions, the solution to the above optimization problem gives the MLE of region A .

2.2. Equivalence between envelope and HCR

Among all regions whose probability mass is no less than $1 - \alpha$, the “high concentration” region has the smallest size. Based on the idea of Neyman-Pearson (e.g., Lehmann, 1986, p. 72), in the region where the objective function in (2.2) is minimized, the p.d.f. f must have large value; on the other hand, if f has large value, it must be in the solution of (2.2). This implies that in searching for the HCR, one can find the largest c such that $P\{A(c)\} > 1 - \alpha$. The envelope ∂A will then encircle a region that has the highest concentration of probability mass.

2.3. “Equivalence” between HCR and MLE

By establishing the connection between the HCR and MLE, the solution to HCR provides a way to progressively locate the MLE. Without loss of generality, we assume

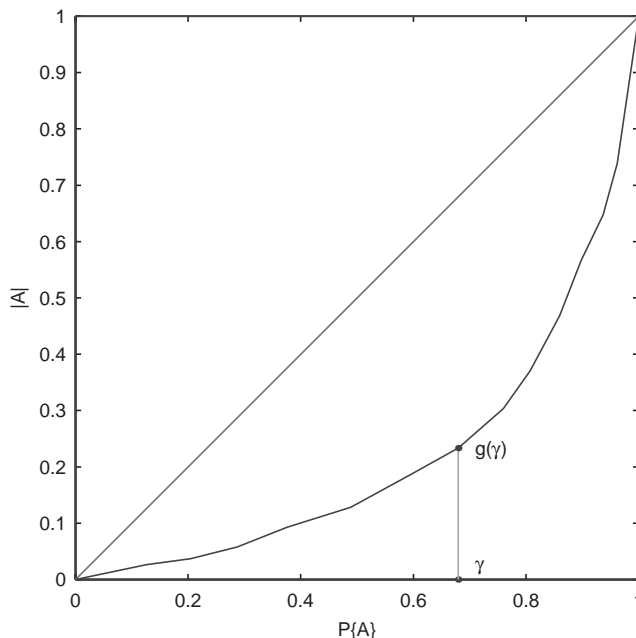


Fig. 1. An example of function g .

$|\Omega| = 1$. Let $g(\gamma)$ denote the optimal value of HCR in (2.2):

$$g(\gamma) = \min_A |A|$$

s.t. $P\{A\} \geq \gamma$.

The difference between this equation and Eq. (2.2) is that an equality condition is added to make sure that the function $g(\cdot)$ is well-defined. The following summarizes a few key properties of $g(\cdot)$:

- (1) $g(0) = 0$.
- (2) $g(1) \leq 1$, from the fact that $|\Omega| = 1$.
- (3) The function $g(\gamma)$ is non-decreasing.

Fig. 1 provides an illustration of a possible function g with respect to coordinates, $P\{\cdot\} = \gamma$ and $g(\gamma) = |\cdot|$. Define a *feasible region* as a set on Fig. 1 that contains all the possible points of $(P\{A\}, |A|)$ for all regions A 's. The curve given by $(x, g(x))$ forms the lower boundary of the feasible region. This curve will be called the *frontier* of the feasible region.

By using an idea similar to Neyman-Pearson (e.g., Lehmann, 1986, p. 72), it is possible to prove that $g(\gamma) \leq \gamma$. Since this property is not essential in our approach, we omit the details. This fact can be observed in Fig. 1 as well.

When $|\Omega| = 1$, Cover and Thomas (1991) stated that the objective function in (2.3) is the relative entropy, $RE(P\{A\}, |A|)$ (or the Kullback–Liebler distance), between two

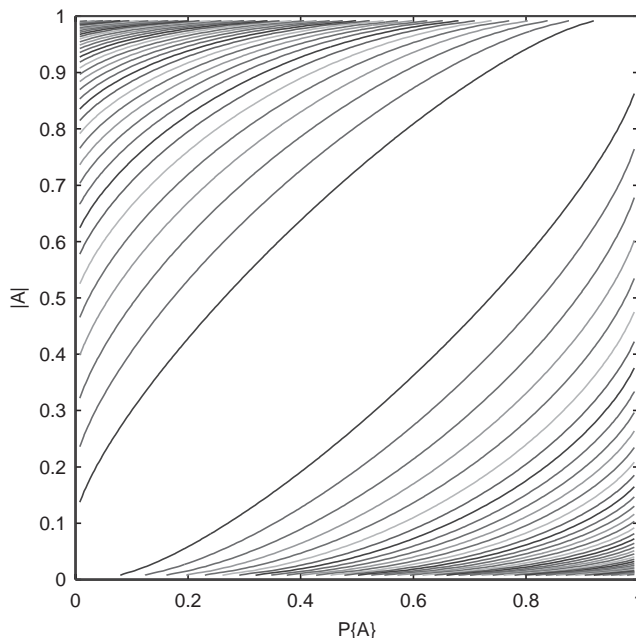


Fig. 2. Contour of the relative entropy function $RE(P\{A\}, |A|)$.

statistical distributions: Bernoulli($P\{A\}$) and Bernoulli($|A|$). The function $RE(\alpha, \beta)$ is a well-behaved (i.e., smooth) function. Fig. 2 provides a contour of $RE(\cdot, \cdot)$. The following states a few properties of the relative entropy.

- (1) $RE(\alpha, \beta) = RE(1 - \alpha, 1 - \beta)$. Thus, the function $RE(\cdot, \cdot)$ is symmetric about point $(1/2, 1/2)$.
- (2) $RE(P\{A\}, |A|)$ is non-negative.
- (3) For a fixed value p , $RE(p, y)$ is convex with respect to y , and the minimum point is at $y_0 = p$. When y “moves” away from p , $RE(p, y)$ increases.
- (4) Therefore, the minimum value of $RE(\cdot, \cdot)$ lies on the diagonal of the unit square, $\{(\gamma, \gamma) : 0 \leq \gamma \leq 1\}$, where the minimum function value is 0.
- (5) If the region A is an MLE, then the point $(P\{A\}, |A|)$ must be on the frontier determined by the function g . Hence, A is a solution to an HCR problem.

The first sentence of Item (5) is true. Because if the point $(P\{A\}, |A|)$ is *not* on the frontier, then there exists a region A' such that

$$P\{A\} = P\{A'\} \quad \text{and} \quad |A| > |A'|.$$

The pair $(P\{A'\}, |A'|)$ will give a larger likelihood, which indicates that region A cannot be the MLE. The second sentence of Item (5) will become evident after we reviewing the idea of Lagrange Multiplier in Section 2.5.

Note that Item (5) leads to a way of solving MLE. That is, by solving a sequence of HCRs, the curve $(x, g(x))$ can be progressively approximated. Therefore the MLE is the region A such that $x_0 = P\{A\}$, and x_0 maximizes $RE(x, g(x))$. The next section will provide more details.

2.4. Solving MLE via HCR

The solution to MLE can be obtained via the following algorithm, which is based on a comprehensive search of HCR.

Algorithm I

Initialize: $\tau = 1/2, cnt. = 1, opt. = 0$.

while continue condition is satisfied,

for $i = 1 : 2 : (2 * cnt. - 1)$, (i.e., i takes only odd numbers: 1, 3, ...)

 compute $g(i\tau)$;

if $\max RE(i\tau, g(i\tau)) > opt.$,

 store solution A that is associated with $g(i\tau)$;

 update $opt.$: $opt. \leftarrow \max RE(i\tau, g(i\tau))$;

end;

end;

 update τ and $cnt.$: $\tau \leftarrow \tau/2, cnt. \leftarrow 2 * cnt.$;

end.

Note that this algorithm is equivalent to sampling the frontier $(g(x))$, with a sampling rate of $\tau = 1/(2^k)$, where k is a natural number. Thus, the Algorithm I is computationally intensive, even though in theory, it is doable. In practice, it is not easy to apply this algorithm to solve the problem in (2.2). Hence, in the next subsection we study a variation based on Lagrange Multipliers.

2.5. Lagrange multiplier approach to solve HCR

Our method will solve the following sequence of problems: for a given $\mu > 0, \mu \in \mathcal{R}$,

$$\min_A |A| - \mu P\{A\}. \quad (2.4)$$

The parameter μ will be chosen in the algorithm. Because problem (2.4) is an unrestricted optimization problem, solving problem in (2.4) is much easier than solving the problem in (2.2). The following observations will ensure that by solving a sequence of the above problems, the solution to MLE can be found:

(1) For a fixed μ , when $A(\mu)$ is the solution to the problem in (2.4), the line

$$y - \mu x = |A(\mu)| - \mu P\{A(\mu)\} \quad (2.5)$$

on the plane (x, y) is a tangent line of the feasible region formed by all possible points $(P\{A\}, |A|)$. The point— $(P\{A(\mu)\}, |A(\mu)|)$ —is a tangent point.

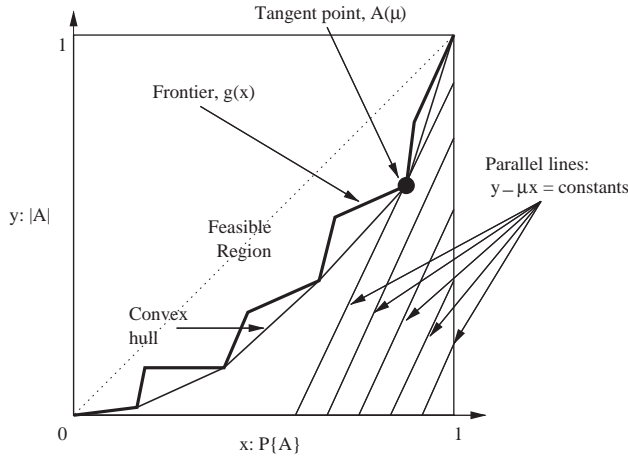


Fig. 3. An illustration of the idea of Lagrange Multiplier. When the μ varies, the slope of the parallel lines varies.

- (2) Consider the convex hull of the feasible region, i.e., $\{(x, y) : y \geq g(x), 0 \leq x \leq 1\}$. A line in (2.5) is also a tangent line of this convex hull.
- (3) It can be easily verified that the contour of $RE(x, y)$ is a convex function in the unit square. Hence, the MLE that maximizes the relative entropy must be one of the tangent point in (2.5) with an appropriate μ .
- (4) Based on all of the above observations, one can obtain the MLE by solving a sequence of problems as in (2.4).

The above observations follow closely the spirit of Lagrange multiplier. We have explained earlier that the MLE is achieved if and only if the pair $(P\{A\}, |A|)$ is on the frontier of the feasible region. Moreover, by following the same argument, one can conclude that the MLE must be on the convex hull of the feasible region. To determine the boundary of the convex hull of the feasible region, a Lagrange Multiplier approach is adopted. We consider a set of parallel lines given by $y - \mu x = C$, where C is a constant. If we minimize the value of the function $y - \mu x$ in the feasible region, the minimizer will render a tangent line from this set and a tangent point, as illustrated in Fig. 3. The value of μ is the slope of this set of parallel lines. By choosing different values of the slope μ , one can explore the boundary of the convex hull of the feasible region. Eventually, the MLE can be computed.

If the function $g(x)$ is convex, then the solution to the problem in (2.4) is also the solution to the problem in (2.2) with $\alpha = 1 - P\{A(\mu)\}$. Hence with appropriate choice of μ , we can solve the problem HCR. On the other hand, if $g(x)$ is not convex, then there is no guarantee that the HCR can be solved by repeatedly solving the problems as in (2.4). Hence, to be more precise about the relationship between the HCR and MLE, we can consider the MLE a special case of the HCR.

In theory, by taking a large number of values of parameter μ , one can solve the problem of MLE within any given precision. The following subsection provides an

algorithm, based on solving a sequence of problems in (2.4) to find the MLE. Section 3 uses a network flow approach to implement this algorithm.

2.6. MLE solution

Algorithm II

Choose K , e.g., $K = 7$.

Let a_0 be a small fraction, e.g., $a_0 = 0.001$ and a_K be a large integer, e.g., $a_K = 1000$.

Let *continue* = “yes”.

Compute $A(a_0)$ and $A(a_K)$ as solutions to the problem in (2.4) when $\mu = a_0$ and $\mu = a_K$, respectively.

while *continue* = “yes”,

for $i = 1 : K - 1$,

 Assign $a_i = a_0 (\frac{a_K}{a_0})^{i/K}$.

 compute $A(a_i)$;

end

Let

$a^* = \operatorname{argmin}_{a_i, i=0,1,\dots,K} RE(P\{A(a_i)\}, |A(a_i)|)$.

Determine $k_1 < k_2$, such that for any $\mu < a_{k_1}$ or $\mu > a_{k_2}$, we have

$RE(P\{A(\mu)\}, |A(\mu)|) > RE(P\{A(a^*)\}, |A(a^*)|)$.

The above is due to the fact that $|A| - a_i P\{A\} = |A(a_i)| - a_i P\{A(a_i)\}$, $i = 0, 1, 2, \dots, K$ are tangent lines of the feasible region.

Update a_0 and $a_K : a_0 \leftarrow a_{k_1}$ and $a_K \leftarrow a_{k_2}$.

If

$|P\{A(a_K)\} - P\{A(a_0)\}| < \varepsilon$,

 then *continue* = “no”.

end.

Result: current $A(a^*)$ is an approximate to the MLE.

Let $A(a^{(0)})$ be the true MLE, then the above algorithm guarantees that

$$|P\{A(a^{(0)})\} - P\{A(a^*)\}| < \varepsilon.$$

Hence, one can approximate the MLE within any given precision. Note that in the above, we only guarantee that the region $A(a^*)$ generates a likelihood function that is large enough to be close to the maximum likelihood. We have *not* guaranteed that region $A(a^*)$ is close to $A(a^{(0)})$. To link $A(a^*)$ with $A(a^{(0)})$, we need to impose some distributional conditions on the probability density function, which is going to be heavily involved and is beyond the scope of the present paper. We will leave detailed discussion on this as future works.

3. Main approach: a network flow algorithm

This section proposes a network flow approach to solve the problem in (2.4). A *digraph* consists of several directional edges connecting a pair of vertices (e.g., data points) in a spatial point process (see Fig. 3 for an example). A *cycle* in a digraph is a chain of edges that returns to its origin. This section will establish an equivalence between a cycle on the digraph and a polygonal region in the domain of the point process. A propagating method is used to find the *minimum cost cycle*, which corresponds to finding a “minimum size” region A , and consequently solves the problem in (2.4).

This section is organized as follows. In Section 3.1, a link between the formulation in the continuous case and a point process is established. In Section 3.2, we provide the definition of a digraph. In Section 3.3, we establish the link between the objective function in Eq. (2.4) and the cost function that is associated with a cycle in a digraph. In Section 3.4, a “fast-algorithm” with an order of complexity no higher than $O(n^3)$ is presented. Finally, some complexity analysis is provided in Section 3.5.

3.1. Preliminary

This section discusses how to apply the network flow method for point processes. For a point process, $P\{A\}$ is replaced by the empirical probability, which by definition is the percentage of observed points that are in region A . The percentage of points is also called the empirical probability associated with region A , and is denoted by $\tilde{P}\{A\}$. The problem in (2.4) becomes

$$\min_A |A| - \mu \cdot \tilde{P}\{A\}. \quad (3.6)$$

The solution to the problem in (3.6) is not well-defined unless there are some restrictions on the region A . For example, if region A can have arbitrary shape, one can always put a tiny circle around each observed data point, so that the total area $|A|$ is nearly equal to zero and the empirical probability $\tilde{P}\{A\}$ is almost one. Then, the optimal solution in (3.6) is achieved. But such a solution is of little practical use. *In our approach, we limit our solution, A , to be a polygon. This polygon is single connected—the boundary does not intersect with itself—and possibly star shaped. The vertices of this polygon are a subset of the observed points.* More geometric restrictions can be added on the solution to include expert’s knowledge. For example, one can assume that there is an interior point (denoted by O) inside this polygon; the region is *star shaped*: when any vertex (denoted by V) on the polygon, the line segment OV is inside this polygon. To prevent the polygon from being too irregular, for an edge AB on this polygon, one can impose a lower bound on the angle: $\angle AOB \geq \pi/k$, where $k > 0$ is a given constant. Obviously, $\angle AOB \leq \pi$.

3.2. A digraph

The key objective of constructing a digraph is to create a cycle in the digraph, such that the region associated with the MLE corresponds to this cycle. Define *meaningful*

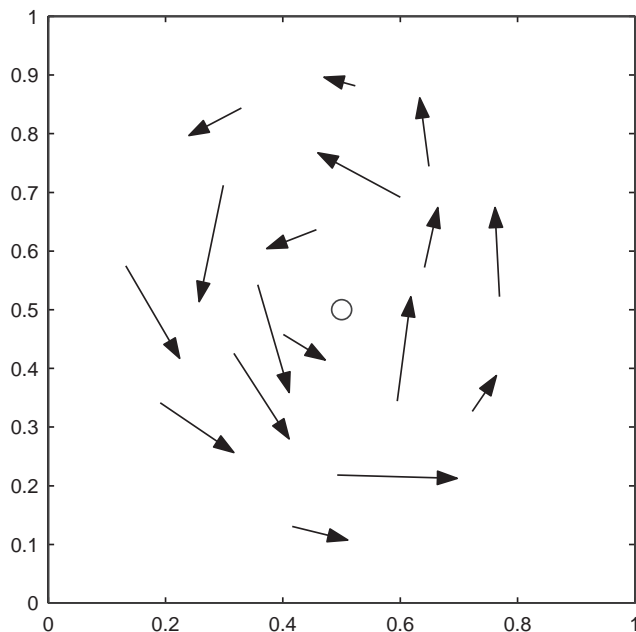


Fig. 4. An illustration of a flow around point $(1/2, 1/2)$.

data points as the vertices of edges formed from a subset of observed data points. The construction consists of two steps:

- (1) Specify a set of meaningful data points based on the “geometric restriction” of a digraph linking to the HCR;
- (2) Follow directions of these edges to determine which pairs of vertices are connected by directional edges, making sure that one of the cycles in this digraph corresponds to the HCR.

An illustration of such a digraph with a subset of directional edges is given in Fig. 4. Consider a point process in the unit square: $[0, 1] \times [0, 1]$. Assume that the only condition for the HCR is that it must include the center point $(1/2, 1/2)$. The meaningful data points can be all the observed data points that are different from $(1/2, 1/2)$. If and only if the vector \vec{ab} goes anticlockwise around the point $(1/2, 1/2)$, there is an edge linking data points a to b .

Note that under different geometric restrictions, the shape of the cycle can be different. For example, if the “anticlockwise” is the only restriction, any star shaped region with a common center $(1/2, 1/2)$ will be allowed. If the smoothness of the boundary of a region is needed, the curvature between two connecting edges (an in-edge and an out-edge at the same vertex) should be bounded. These conditions can make a big difference in the estimation results in practice. Section 4 will provide more discussion on this subject.

To facilitate the search for the minimum cost cycle, the meaningful data points are ordered. For example, in the above case, the meaningful points can be sorted by the angles of vectors starting from the central point and ending at the meaningful points. The purpose of this ordering becomes more evident when we describe a propagating algorithm in Section 3.4.

3.3. Network flow

Suppose (x_i, y_i) and (x_j, y_j) are the coordinates of the starting and ending vertices of an edge, respectively. Without loss of generality, consider the edge neither vertical nor horizontal. Define a trapezoid T determined by the points

$$(x_i, y_i), (x_j, y_j), (0, y_j), \text{ and } (0, y_i).$$

Let $|T|$ be the area of T , and $P\{T\}$ the probability mass $P\{T\} = \int_T f$ in this trapezoid, where f is the p.d.f. Denoted by $f_{i \rightarrow j}$ the cost function assigned to this edge. The absolute value of $f_{i \rightarrow j}$ is equal to $|T| - \mu P\{T\}$, and its sign is negative if $y_i > y_j$ (edge going down), and positive if $y_i < y_j$ (edge going up). Note that the value of $f_{i \rightarrow j}$ depends on μ , which is the same parameter as in (2.4).

Let us discuss the one-to-one mapping between the polygons in the domain of the point process and cycles in the digraph. Suppose vertices v_1, v_2, \dots, v_n form a cycle in the digraph: $v_1 \rightarrow v_2 \rightarrow \dots \rightarrow v_n \rightarrow v_1$. Let R be the polygon associated with this cycle. Fig. 4 provides an illustration of such a polygon with $n = 7$ vertices. It can be easily verified that

$$\sum_{i=1}^{n-1} f_{i \rightarrow (i+1)} + f_{n \rightarrow 1} = |R| - \mu P\{R\}.$$

Note that the right-hand side of the above equation is exactly the objective function in (2.4). Hence to solve the problem in (2.4), one only needs to find the cycle in the digraph that minimizes the overall cost of all possible cycles. Many branches of science have used similar ideas in solving their problems. As an example, see Geiger et al. (1995) and Jermyn and Ishikawa (1999) in segmenting medical images (Fig. 5).

3.4. A propagation algorithm

Finding the minimum cost cycle is a well-studied problem in operations research, cf. Ahuja et al. (1993). Other references are in the *dynamic programming* literature, e.g., the “label correcting algorithms” as in Bertsekas (1995). In this section, we present an algorithm that is based on a propagating method. This method can be easily implemented in a digraph. Note that these algorithms all have the same order of complexity (in the worst case scenario). See Section 4.3 for details about the complexity.

Suppose v_1, v_2, \dots, v_N are the ordered vertices in the digraph, where N is the total number of vertices. We can construct an N by $2N$ table (see Table 1) such that when $i < j \leq N$, the entry (i, j) is the cost (denoted by $d(i, j)$) of the minimum cost path going from v_i to v_j ; when $i \leq N < j \leq 2N$, the entry (i, j) is the cost (denoted by $d(i, j - N)$) of the minimum cost path going from v_i to v_{j-N} .

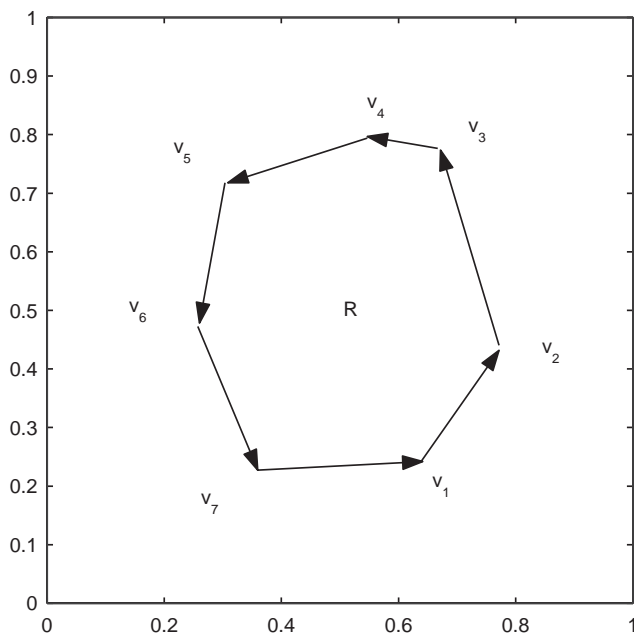
Fig. 5. Polygon R circumented by edges.

Table 1
The minimum “distance” table for meaningful data points

	v_1	v_2	v_3	\dots	v_N	v_1	v_2	v_3	\dots	v_N
v_1	—	$d(1,2)$	$d(1,3)$		$d(1,N)$	$d(1,1)$	—	—		—
v_2	—	—	$d(2,3)$		$d(2,N)$	$d(2,1)$	$d(2,2)$	—		—
v_3	—	—	—		$d(3,N)$	$d(3,1)$	$d(3,2)$	$d(3,3)$		—
\vdots				\ddots	\vdots	\vdots	\vdots		\ddots	\vdots
v_N	—	—	—		—	$d(N,1)$	$d(N,2)$	$d(N,3)$		$d(N,N)$

Apparently, the minimum value of $d(i, i)$ is associated with the minimum cost cycle. The entries in the above table can be computed recursively based on a formular that will be defined below: Eq. (3.8). For simplicity, let

$$d(i, j) = d(i, j - N), \quad \text{if } j > N.$$

Then, for $i < j \leq i + N$, $d(i, j)$ is the minimum of $f_{i \rightarrow j}$ and the quantity $\tilde{d}(i, j)$ defined below:

$$\tilde{d}(i, j) = \min_{i < k < j} d(i, k) + f_{k \rightarrow j}, \quad (3.7)$$

s.t. there exist edges from v_i to v_k and v_k to v_j ,

when $k > N$ (or $j > N$ only), $f_{k \rightarrow j} = f_{(k-N) \rightarrow (j-N)}$ (or $f_{k \rightarrow (j-N)}$); similarly, v_k or v_j is v_{k-N} or v_{j-N} , if k or j is larger than N . In summary, we have

$$d(i, j) = \min\{\tilde{d}(i, j), f_{i \rightarrow j}\}. \quad (3.8)$$

Following the relation in (3.8) and (3.7), all the entries in the above table can be computed recursively.

3.5. Complexity

Suppose there are N meaningful data points, and it takes $O(1)$ to compute the value of a $f_{i \rightarrow j}$. Since there are at most $\binom{N}{2} \asymp N^2/2$ edges in the digraph, it would take no more than $O(N^2)$ to compute the values of all possible functions $f_{i \rightarrow j}$. To generate the propagation table, from the relationships in (3.8) and (3.7), it would take no more than $O(N)$ steps to *update* each entry. Since there are N^2 entries, it should take no more than $O(N^3)$ operations to compute all the entries in the above table. Overall, the complexity of this method is no more than $O(N^3)$. In practice, the computational complexity could be much lower. The more restrictive the geometric restrictions are, the less edges there will be. Hence the overall computational complexity could be reduced dramatically. Under the terminology *circular shortest path*, researchers have proposed efficient exact and approximate algorithms in solving the above problem. We list the work by Sun and Pallottino (2003) as an example. Providing more computationally efficient approaches will be a future research activity.

4. Simulations and real-life examples

This section explores the properties of the proposed method. In Section 4.1, four examples with synthetic data are described. They demonstrate the effectiveness of this approach for different types of underlying regions. Section 4.2 uses two real-life datasets to address the applicability of the proposed method. In Section 4.3, effects of wrongly specified geometric information is illustrated in some numerical experiments. In Section 4.4, experiments with data having a range of degree of noisiness are presented. They empirically show the stability of the proposed method.

4.1. Synthetic datasets

Four cases illustrated in Fig. 6 are considered. In each case, the points are in the unit square, $[0, 1] \times [0, 1]$.

In the first example, a subsquare $[0.4, 0.6] \times [0.4, 0.6]$ contains 20% of the data points. The subsquare forms a HCR, as we can see in Fig. 7(a). In this case, the geometric constraint is that the solution must be a polygon, which includes the interior point (0.5,0.5) and its boundary going anticlockwise around the interior point.

After taking different values of μ , solutions to the problem in (2.4) are drawn in Fig. 7(b), with the values of μ in the parentheses. In Fig. 7(c), the tangent lines given

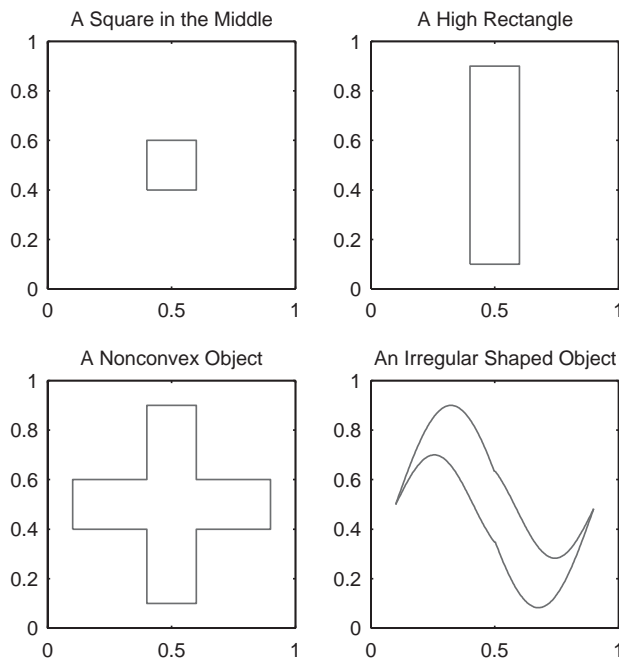


Fig. 6. Four types of regions.

by (2.5) are drawn, together with the contour of the relative entropy function $RE(\cdot, \cdot)$ in dotted lines.

In Fig. 7(c), one can see that the fourth and fifth solutions *approximately* provide the maximum value of the likelihood function. In Fig. 7(b), one can verify that the fourth and fifth polygons are the closest to the underlying subsquare, which is bounded by the relatively straight red lines.

In Fig. 7(d), the MLEs in nine repetitions with independently generated point processes are provided. The MLEs are computed according to Algorithm II. It shows that in this case, the MLE is a fairly good method of estimating the underlying HCR.

In the first set of examples, a regular shaped region—a square—is studied.

In the next set of examples, to explore *whether the shape of the region will influence the estimate*, we consider an elongated rectangle. In Fig. 8(a), the rectangle $[0.4, 0.6] \times [0.1, 0.9]$ contains 20% of the data points. The geometric restrictions for the underlying HCR remain the same as in the first example. In Fig. 8(b), solutions associated with a range of μ 's are drawn. Again, the fourth and fifth solutions are the closest to the underlying rectangle. From Fig. 8(c), one can see how the MLE could be found. Fig. 8(b) is generated by the same method as Fig. 7(b). The Fig. 8(d) provides several MLEs from repeatedly generating independent point processes. In summary, the second set of experiments shows that deformation of the underlying region does not destroy the consistency of the MLE.

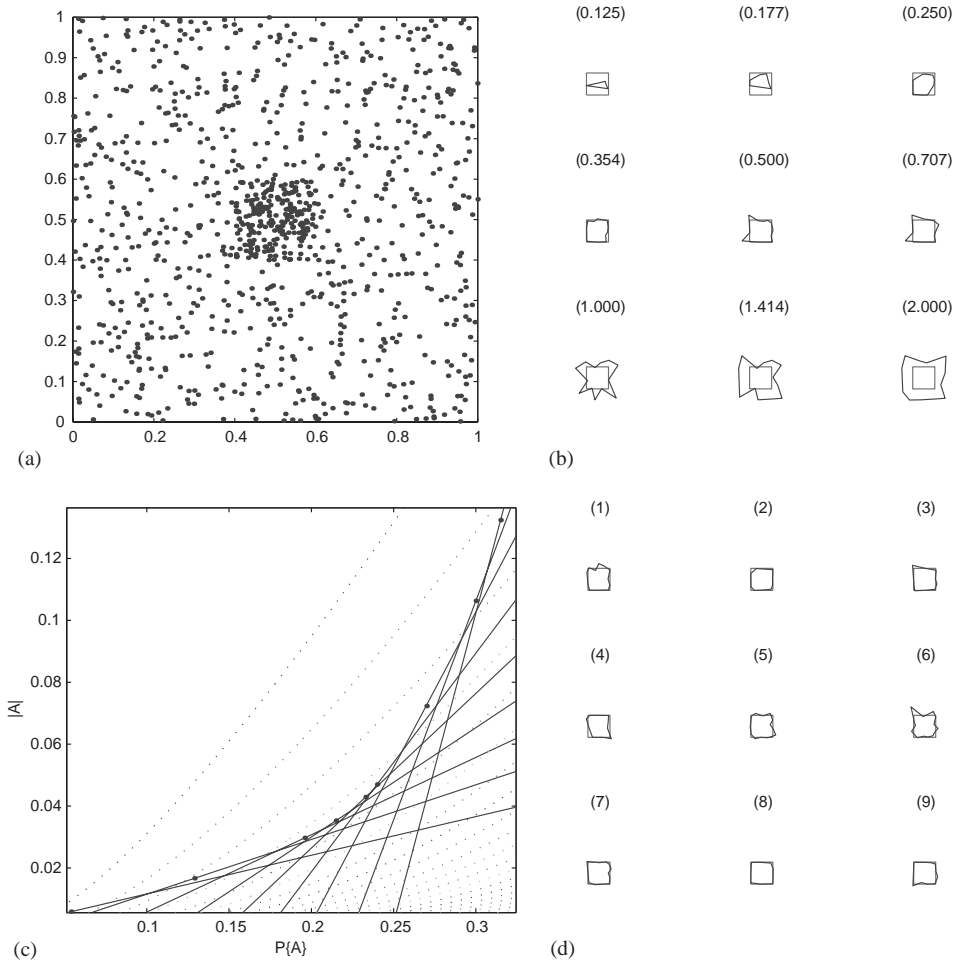


Fig. 7. Simulation on a square in the middle.

In the above two experiments, the underlying high concentration regions are convex. Our method can deal with *non-convex regions* as well. For example, the HCR in the point process in Fig. 9(a) is a cross with the vertical rectangle given by $0.4 \leq x \leq 0.6$ and $0.1 \leq y \leq 0.9$, and the horizontal rectangle is a 90 degree rotation with respect to the point $(0.5, 0.5)$, which is also the center of the vertical rectangle. This region contains 50% of the data points. Although it is not convex, it is still star shaped. The same geometric restrictions as in the previous two examples are imposed. Hence, the numerical algorithm is exactly the same as used in the previous two cases. Via a visual inspection of Fig. 9(c), the fifth solution nearly gives the MLE. Note that in Fig. 9(b), both the fifth and the sixth solutions are very close to the underlying region. Similar to what has been done earlier, in Fig. 9(d), we provide some examples of MLEs for

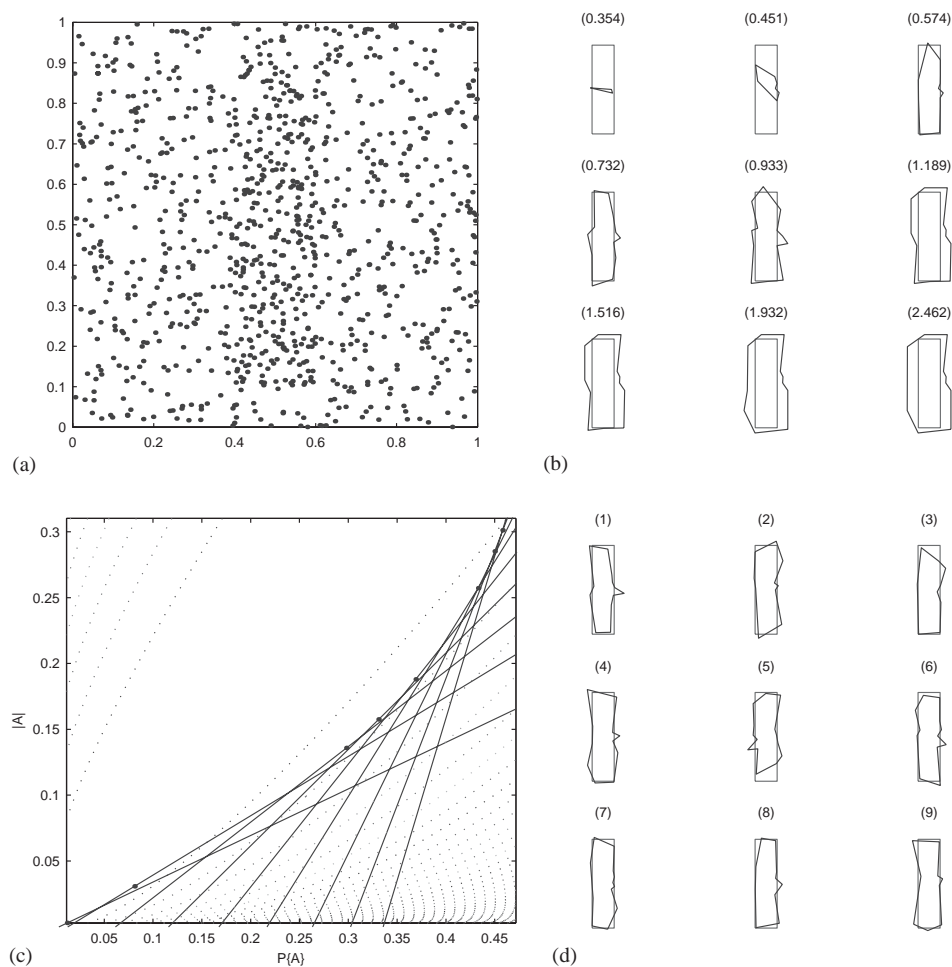


Fig. 8. A high rectangle.

independently generated point processes. It again shows the effectiveness of our method in locating the MLE.

In all the above three examples, the underlying objects are star shaped. Namely, there exists an interior point, from which all points on the boundary are visible. In Fig. 10(a), we have an irregular shaped HCR, which is a region enclosed by two trigonometric curves. The region is *not star shaped*. To utilize the network flow approach, one first needs to specify two points, e.g., $(0.3, 0.8)$ and $(0.7, 0.2)$. Then, specify a flow going anticlockwise around the line segment connecting these two points. Note that there is no longer a center point. Following the same framework, the solutions associated with a range of values of μ is given in Fig. 10(b). In Fig. 10(c), one can observe that the fifth solution is close to the MLE. As been done earlier, several MLEs in this setting are shown in Fig. 10(d). It is again shown that the MLE performs very well.

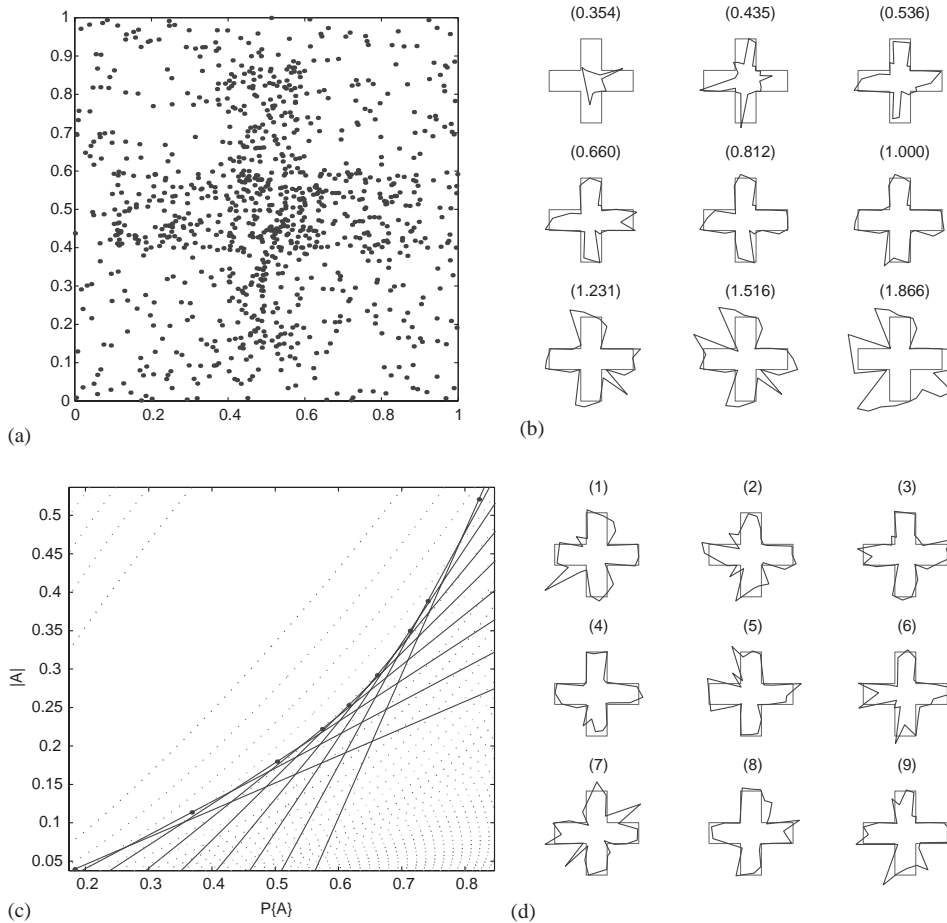


Fig. 9. A non-convex object: a cross.

4.2. “Real-life” examples

This section examines how well our approach works for two data sets that have been used in the literature. We first study a tree data that is taken at Paracou site and is studied in Picard and Bar-Hen (2000). A normalized version (inside the unit square $[0, 1] \times [0, 1]$) is shown in Fig. 11(a), where the horizontal axis is the tree diameter, and the vertical axis is the annual diameter increment. There are 7357 data points as of that many trees. The second data set contains the earthquake locations in California. This data set has been studied in a sequence of papers, see Allard and Fraley (1997), Byers and Raftery (1998), and Dasgupta and Raftery (1998). The pattern in this data should be closely related to the geological structures. A normalized version of this data is shown in Fig. 12(a). Our intension is to use these two examples to show the applicability of our algorithm. It is *not* the authors’ intention to carry out this part

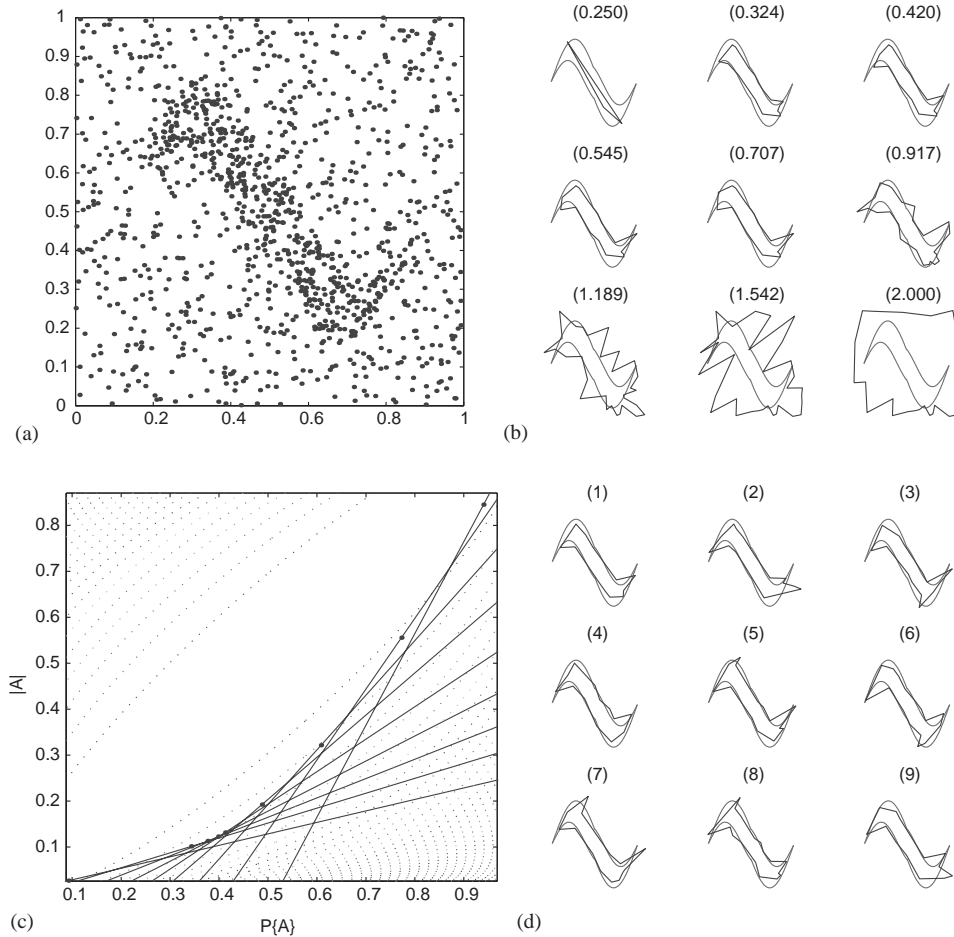


Fig. 10. An irregular shape.

as a thorough *data analysis* project. Due to this, it is not argued that our method is the “best” method to be used. We will not intensively engage in model interpretation either.

4.2.1. Paracou forest data

In Fig. 11(b), polygons associated with different μ 's are drawn. In determining the network flow, the point (0.1,0.1) is chosen as the interior point. In fact, in determining the “meaningful data points,” the points that are too close to point (0.1,0.1) has little chance to be on the boundary of a polygon. Thus, they are excluded. By doing so, the problem size is much reduced. In Fig. 11(c), the straight lines outline the shape of the function $g(\gamma)$. Hence, the approximate solution to problem envelope can be “read-off” from this figure. The MLE in this case is provided in Fig. 11(d).

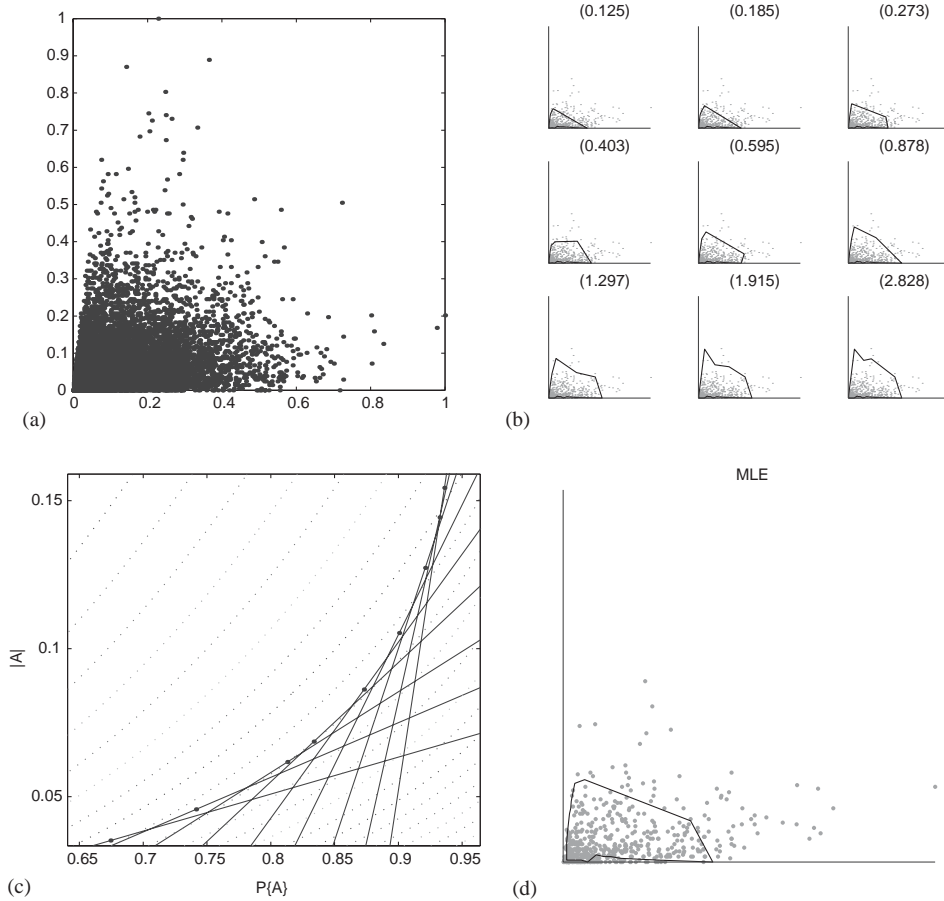


Fig. 11. The Cirad-foreêt, Paracou data set.

4.2.2. Earthquake positions

Because there are apparently multiple disconnected and irregular-shaped clusters, this data set is very difficult to analyze. To simplify the study, this section finds the HCR by imposing the following assumption: the estimate of the HCR is a single-connected polygon that includes the interior point (0.7631, 0.4), which is picked by visual inspection of the Fig. 12(a).

The tangent lines (in solid straight lines) and the contour of the likelihood function (in dotted lines) are drawn in Fig. 12(c). This set of estimates provide information on the concentration of the positions of the earthquake. The MLE is shown in Fig. 12(d).

Since the underlying features should be highly linear due to the existing geological structures, and our assumption that the estimates are polygons, our estimates do not show the underlying features that well. But, it helps to identify the main concentrated region in the point process. This is helpful in choosing which methods should be used

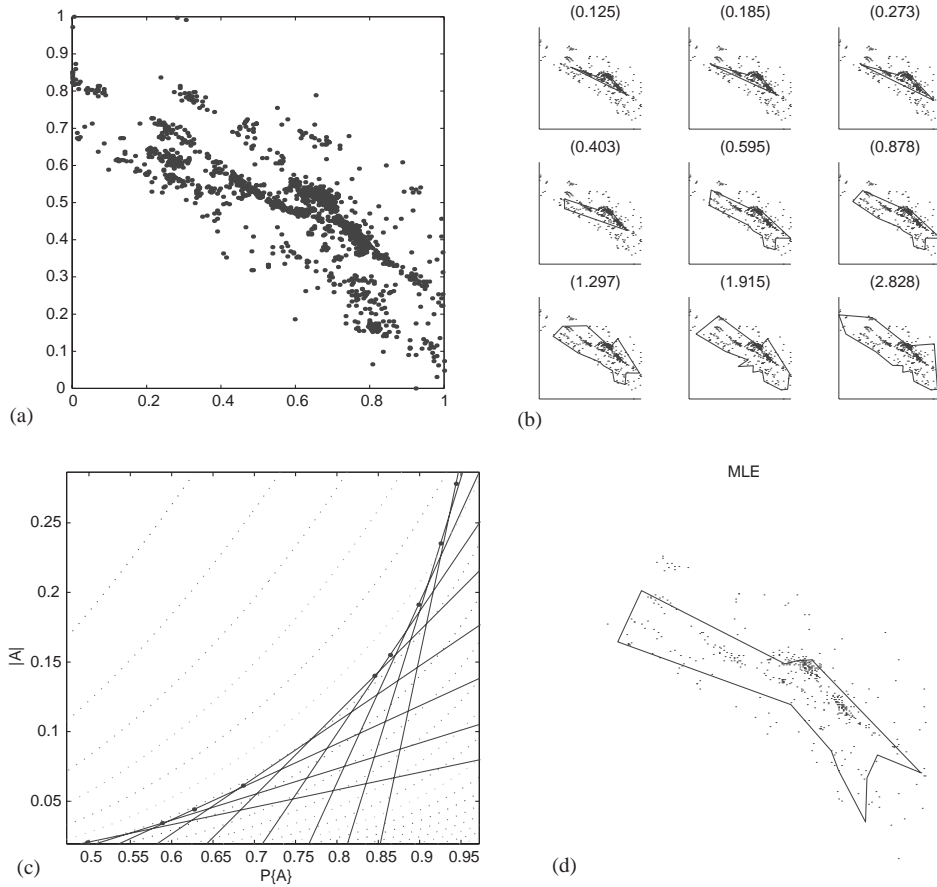


Fig. 12. Earthquake position data.

in the next steps. It would be very interesting to design a multiple-flow method to extract multiple clusters in a point process. We leave this to a future research project.

As pointed out by an anonymous referee, the directional linear features are the most significant feature in this data set. Due to this, a method that is designed for line features could be more ideal to analyze this data. As stated earlier, our intention is to show the applicability of the network flow approach. There is no implication that our MLE gives the most desirable estimate. Which estimates to choose always depends on the context of the applications.

4.3. Effects of wrongly specified conditions

Recall that in the network flow approach, the center(s) and the boundary condition of the underlying region need to be specified in advance. How wrongly specified

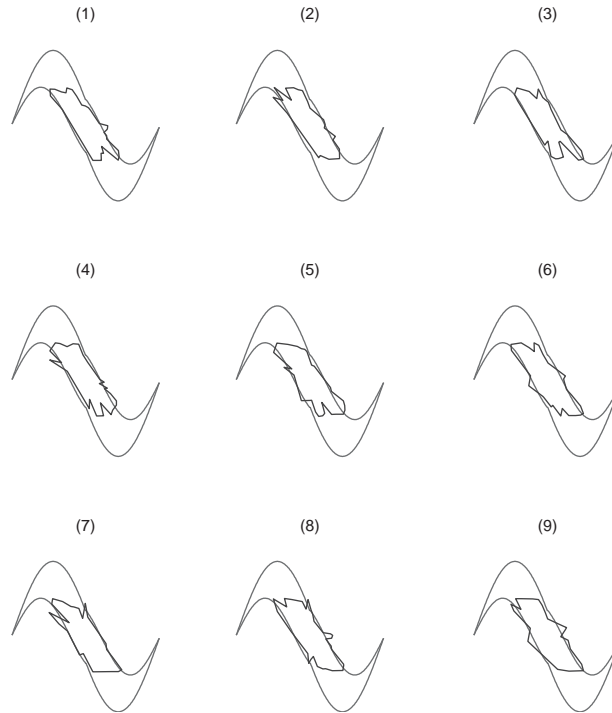


Fig. 13. Nine computational studies when the geometric information of the underlying region is wrongly specified. The blue curves is the MLE. The red (double sine) curves are the boundary of the true underlying region.

information will affect the output of the proposed algorithm? A thorough study can be done by utilizing the tools that have been developed in mathematical statistics. In this paper, some simulation studies are reported to illustrate this effect. We postpone the theoretical analysis for future publications.

In Fig. 13, the point clouds have the same distribution as the ones that are used in Fig. 10. However, in the algorithm, the underlying region is assumed to be centered at $(0.5, 0.5)$ and is star-shaped.

In the nine simulations, the MLE's are significantly different from the underlying region. This demonstrates that the conditions that are made for the underlying region do play a critical role, which is expected.

4.4. Effect of the relative density in the high concentration region

It is essential to study how the relative density of a high concentration region will affect the result of our network flow algorithm. Intuitively, if the density within the high concentration region is *not* significantly higher than the density outside, the MLEs should appear relatively random. To study this problem, the following experiment is

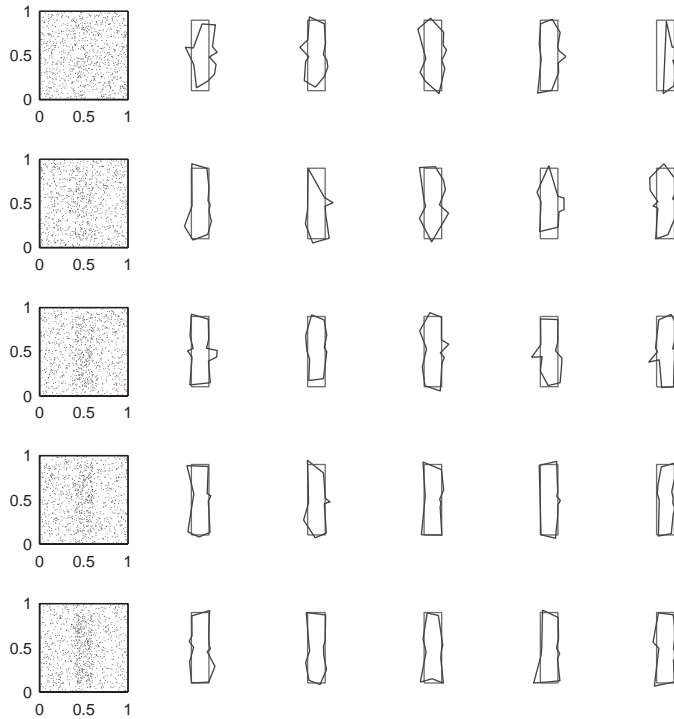


Fig. 14. MLEs of point processes with different relative densities in the high concentration region. The density of these point processes are given in Eq. (4.9). The values of α are 0.05, 0.10, 0.15, 0.20, and 0.25, which correspond to the first to the last rows. The first figure on each row is a realization of the point process. The rest five figures are the MLEs, for independently generated point processes, by running the network flow algorithm.

conducted. In Fig. 14, the point processes have the following density function:

$$f(x) = \begin{cases} 1 - \alpha + \frac{25}{4}\alpha & \text{if } x \in [0.4, 0.6] \times [0.1, 0.9], \\ 1 - \alpha & \text{otherwise.} \end{cases} \quad (4.9)$$

In Fig. 14, each row corresponds to one value of α . The values of α are (from top to bottom) $\alpha = 0.05, 0.10, 0.15, 0.20$, and 0.25 . The first figure of each row is a realization of this point process. The rest five figures are the results of the network flow algorithm, for five independently generated point processes. It is observed that when the value of α is small, the relative density of the HCR is small, hence the HCR is less observable in the generated point process, and the MLEs may deviate significantly from the HCR. On the other hand, when the relative density is high, the HCR becomes more visible, and the MLEs are closer to the underlying HCR. Analytical analysis on the relation between the relative density and the MLE will be carried out in the future.

4.5. Miscellaneous topics

Miscellaneous topics are presented in this section. In Section 4.5.1, some future experiments and analysis are listed. In Section 4.5.2, we emphasize the flexibility of our approach. In Section 4.5.3, the computer running time is reported.

4.5.1. Discussion

In the current approach, prior knowledge on the underlying region is required. We have studied some techniques, such as using density estimation as a pre-processing step, to estimate the center of a high concentration region. It will be interesting to design an iterative approach—between the center estimation and the likelihood maximization—to integrate our proposed approach and the acquisition of the prior knowledge. In numerical studies, it will also be interesting to study the scenarios, under which the underlying distribution is not mixed Uniform distribution.

4.5.2. Flexibility of our approach

The framework in our method is very flexible. For different geometric restrictions, different types of flow can be designed. Then, the rest of the steps will be the same. In principle, we can solve problems with arbitrary shapes. In this article, only the assumption of single-connectedness is imposed.

4.5.3. Computing time

The above algorithms are implemented in MATLAB. Because MATLAB is an interpreting computer language, its execution is expected to be slower than a precompiled function, such as a function in C or Fortran. In most of the above cases, it took about 30 s to solve each optimization problem on a SGI Octane workstation with a 300 MHz processor. This demonstrates that since the algorithm has low complexity, it does not require “low-level” software implementation, which can be very time consuming. Thus, the proposed method is computationally efficient, and has the potential of solving problem with large number of observations. We have dealt with thousands of data in several examples that were illustrated above. Considering the MATLAB based implementation, this is definitely not the limit of the proposed method.

5. Conclusion and discussion

This article proposes a new method that can solve three types of problems that are associated with the high-density region. Our approach is the first of this kind of searching for the polygons that precisely optimize a global objective function. Numerical studies with examples that are taken in the literature demonstrate the effectiveness of our approach. The method is flexible to accommodate various restrictions (e.g., non-convexity) on the shapes of the high density regions, and it finds the global optimal estimate with fast and easily implementable algorithms.

It is easy to incorporate other information of the point process into our method. One only needs to change the definition of the cost function associated with each edge

in the digraph. This is useful in the example with California earthquake positions, due to the fact that the magnitude of earthquakes may be correlated to the geological structures. Future work will present more theoretical analysis of the MLE by giving a better description of its limiting behavior. Studying the possibility of using multiple flows to deal with non-single-connected region is very interesting.

Acknowledgements

The comments from two anonymous referees and the associate editor helped to improve the presentation of this paper.

References

- Ahuja, R.K., Magnanti, T.L., Orlin, J.B., 1993. *Network Flows: Theory, Algorithms, and Applications*. Prentice-Hall, New York.
- Allard, D., Fraley, C., 1997. Nonparametric maximum likelihood estimation of features in spatial point processes using voronoi tessellation. *J. Amer. Statist. Assoc.* 92 (440), 1485–1493.
- Bertsekas, D., 1995. *Dynamic Programming and Optimal Control*, Vol. 1. Athena Scientific, Belmont, MA.
- Byers, S., Raftery, A.E., 1998. Nearest-neighbor clutter removal for estimating features in spatial point processes. *J. Amer. Statist. Assoc.* 93 (442), 577–583.
- Cover, T.M., Thomas, J.A., 1991. *Elements of Information Theory*. Wiley, New York.
- Dasgupta, A., Raftery, A., 1998. Detecting features in spatial point processes with clutter via model-based clustering. *J. Amer. Statist. Assoc.* 93 (441), 294–302.
- Geiger, D., Gupta, A., Costa, L.A., Vlontzos, J., 1995. Dynamic programming for detecting, tracking and matching deformable contours. *IEEE Trans. Pattern Anal. Mach. Intell.* 17 (3), 294–302.
- Jermyn, I., Ishikawa, H., 1999. Globally optimal regions and boundaries. In: *Proceedings of the Seventh ICCV, Kerkyra, Greece, September*.
- Lehmann, E.L., 1986. *Testing Statistical Hypothesis*. Springer, New York.
- Picard, N., Bar-Hen, A., 2000. Estimation of the envelop of a point set with loose boundaries. *Appl. Math. Lett.* 13, 13–18.
- Silverman, B., 1986. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London.
- Sun, C., Pallottino, S., 2003. Circular shortest path in images. *Pattern Recognition* 36 (3), 709–719.