

WILEY

Board of the Foundation of the Scandinavian Journal of Statistics

Simulation Procedures and Likelihood Inference for Spatial Point Processes

Author(s): Charles J. Geyer and Jesper Møller

Source: *Scandinavian Journal of Statistics*, Vol. 21, No. 4 (Dec., 1994), pp. 359-373

Published by: Wiley on behalf of Board of the Foundation of the Scandinavian Journal of Statistics

Stable URL: <http://www.jstor.org/stable/4616323>

Accessed: 27-02-2018 16:02 UTC

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at <http://about.jstor.org/terms>



JSTOR

Board of the Foundation of the Scandinavian Journal of Statistics, Wiley are collaborating with JSTOR to digitize, preserve and extend access to Scandinavian Journal of Statistics

Simulation Procedures and Likelihood Inference for Spatial Point Processes

CHARLES J. GEYER

University of Minnesota

JESPER MØLLER

University of Aarhus

ABSTRACT. An alternative algorithm to the usual birth-and-death procedure for simulating spatial point processes is introduced. The algorithm is used in a discussion of unconditional versus conditional likelihood inference for parametric models of spatial point processes.

Key words: conditional maximum likelihood, Gibbsian point processes, Markov chain Monte Carlo, Metropolis–Hastings algorithms, spatial birth-and-death processes, Strauss process, unconditional maximum likelihood

1. Introduction

Statistical inference for spatial point processes has usually been performed conditional on n , the number of points in the pattern, see e.g. Ripley (1977, 1981, 1988), Diggle (1983), and Baddeley & Møller (1989). This is mainly for computational convenience, though it is usually not clear whether n is (approximately) ancillary. For example, to perform likelihood inference for parametric models of Gibbsian point processes one may approximate the normalizing constant (which typically cannot be determined explicitly) by Monte Carlo methods (Ogata & Tanemura, 1981, 1989; Penttinen, 1984; Ripley, 1988; Geyer & Thompson, 1992). A different approach used the Robbins–Munro algorithm (Younes, 1988; Moyeed & Baddeley, 1991). Hence efficient simulation procedures are needed, but the only well-known algorithm for the unconditional case seems to be based on running a spatial birth-and-death process which converges towards the Gibbsian point process (Preston, 1977; Ripley, 1977; Lotwick & Silverman, 1981; Møller, 1989). In contrast it is much simpler to use other algorithms for the case where n is fixed (Metropolis *et al.*, 1953; Ripley, 1979).

In this paper we suggest an alternative simulation procedure for the unconditional case using the Metropolis–Hastings algorithm (Metropolis *et al.*, 1953; Hastings, 1970), which is simpler than the spatial birth-and-death technique and which can be used for Markov chain Monte Carlo methods. Section 2 contains a few preliminaries, while our algorithm is described in section 3. Convergence of the algorithm is established in section 4. Further, in section 5 we discuss the convergence of our algorithm in practice when simulating a Strauss process, and the algorithm is used in a comparative study of the unconditional and conditional maximum likelihood estimators for the interaction parameter of the Strauss process. Finally, section 6 contains some concluding remarks.

2. Preliminaries

This section contains a few preliminaries on general finite point processes. We adapt notation in Baddeley & Møller (1989).

Let $(S, \mathcal{B}, \lambda)$ be a measure space with $0 < \lambda(S) < \infty$ and where $\mathcal{B}$ contains all singletons. Let $(\Omega, \mathcal{F}, \mu)$ denote the corresponding exponential space (see e.g. Carter & Prenter, 1972).

Informally, $\Omega = \bigcup_{n=0}^{\infty} \Omega_n$ where

$$\Omega_n = \{\{x_1, \dots, x_n\} \subset S\}$$

is the set of all point configurations of n (not necessarily distinct) points included in S ; especially, $\Omega_0 = \{\emptyset\}$ where $\emptyset$ denotes the empty point configuration. Moreover, $\mathcal{F}$ is a natural σ -field and μ is the Poisson process with intensity measure λ on S : for $F \in \mathcal{F}$ we have that

$$\mu(F) = \exp(-\lambda(S)) \left[1(\emptyset \in F) + \sum_{n=1}^{\infty} \frac{1}{n!} \int \cdots \int 1(\{x_1, \dots, x_n\} \in F) \lambda(dx_1) \cdots \lambda(dx_n) \right], \quad (2.1)$$

where $1(\cdot)$ denotes indicator function. Now, a measurable mapping from a probability space into $(\Omega, \mathcal{F})$ is called a finite point process on S . We shall only consider point processes with a distribution π on $(\Omega, \mathcal{F})$ which is absolutely continuous with respect to μ .

For a usual finite point process, S is a bounded Borel set of $\mathbb{R}^d$ equipped with the Borel σ -field $\mathcal{B}$ and Lebesgue measure λ restricted to S , in which case μ is just the usual homogeneous Poisson process with intensity 1 restricted to S . However, the algorithm presented in section 3 applies for general finite point processes. Thus for instance, marked point processes (see e.g. Baddeley & Møller, 1989) can be simulated as well. In fact also lattice processes can be included as explained in example 3 below.

3. A Metropolis–Hastings algorithm for spatial point processes

Let the situation be as in section 2 and suppose we want to simulate realizations of a point process with distribution π . Its density $f = d\pi/d\mu$ with respect to the Poisson process is assumed to be hereditary, that is

$$f(x) > 0 \Rightarrow f(y) > 0 \quad \text{if } y \subseteq x. \quad (3.1)$$

Here $y \subseteq x$ means that y is a subconfiguration of x . Except in rare cases it is impossible to make direct simulations, so alternatively one may run a homogeneous Markov chain which converges towards π .

A general construction considered by Hastings (1970) can be described as follows in the present situation: the transition probabilities are given by

$$Q(F | x) = \int_{y \in F} A(y | x) P(dy | x), \quad x \in \Omega, F \in \mathcal{F}, x \notin F, \quad (3.2)$$

and

$$Q(\{x\} | x) = P(\{x\} | x) + \int_{y \neq x} [1 - A(y | x)] P(dy | x),$$

where P is a Markov kernel and $0 \leq A(\cdot | \cdot) \leq 1$. We call P the proposal kernel and A the acceptance probability. We require that these are chosen so that the Markov chain is time reversible with respect to π , that is

$$\int_{G_m} Q(F_n | x) f(x) \mu(dx) = \int_{F_n} Q(G_m | x) f(x) \mu(dx)$$

or equivalently by (3.2),

$$\int_{G_m} \int_{F_n} A(y | x) f(x) P(dy | x) \mu(dx) = \int_{F_n} \int_{G_m} A(y | x) f(x) P(dy | x) \mu(dx) \quad (3.3)$$

for all $F_n \in \mathcal{F}_n$, $G_m \in \mathcal{F}_m$, $m \geq n \geq 0$, where $\mathcal{F}_n$ denotes $\mathcal{F}$ restricted to Ω_n . In turn this implies that π is invariant: for any $F \in \mathcal{F}$,

$$\int Q(F | x) f(x) \mu(dx) = \pi(F).$$

Observe that if Q_0 and Q_1 are two transition kernels which both satisfy (3.3), then

$$Q_p(F | x) = (1-p)Q_0(F | x) + pQ_1(F | x), \quad 0 \leq p \leq 1, \quad (3.4)$$

also defines a transition kernel which satisfies (3.3). The algorithm we propose is a mixture like (3.4) and generalizes the method of Norman & Filinov (1969), see also Allen & Tildesley (1987, sect. 4.6). There are three different types of transitions: either a point is displaced or removed or a new point is added as described below. Here Q_0 controls displacements and Q_1 controls deletion and addition of points.

Specifically, let $x = \{x_1, \dots, x_n\}$ with $f(x) > 0$. Then $Q_0(\cdot | x)$ is defined through the proposal kernel

$$P_0(F | x) = \begin{cases} \frac{1}{n} \sum_{i=1}^n \int 1[(x \setminus \{x_i\}) \cup \{y_i\} \in F] \lambda(dy_i) / \lambda(S) & \text{if } n \geq 1, \\ 1[\emptyset \in F] & \text{if } n = 0, \end{cases}$$

and the acceptance probability

$$A_0(y | x) = \min \{1, f(y)/f(x)\},$$

so $Q_0(\cdot | x)$ is concentrated on $H_n = \Omega_n \cap H$ with $H = \{f > 0\}$. This is nothing but the procedure in Metropolis *et al.* (1953), and Q_0 is easily seen to satisfy (3.3). In fact any other kernel which satisfies (3.3) and is concentrated on H_n could be used as well in the following; examples of such algorithms can be found in Ripley (1979) and Møller (1993).

Next, the proposal kernel of $Q_1(\cdot | x)$ is constructed so that it is concentrated on $H_{n-1} \cup H_n \cup H_{n+1}$ (or $H_0 \cup H_1$ if $n = 0$): with probability $q(x)$ we generate a new point ξ from some density $b(x, \xi)$ with respect to $\lambda(d\xi)$, and with probability $1 - q(x)$ we either delete a random point $\eta \in x$ which is selected with some probability $d(x \setminus \eta, \eta)$ or if $n = 0$ we do nothing. It is assumed that

$$d(x, \xi) > 0 \quad \text{whenever } x \cup \xi \in H, \quad (3.5)$$

and for all $x \in H_n$, $F_{n+1} \in \mathcal{F}_{n+1}$, $n \geq 0$ with

$$\int_{x \cup \xi \in F_{n+1}} f(x \cup \xi) \lambda(d\xi) > 0$$

we require that

$$\int_{x \cup \xi \in F_{n+1}} b(x, \xi) \lambda(d\xi) > 0. \quad (3.6)$$

Denoting the acceptance probability by A_1 , then for $F_m \in \mathcal{F}_m$ we have $Q_1(F_m | x) = 0$ if $|m - n| > 1$, while

$$\begin{aligned} Q_1(F_{n+1} | x) &= q(x) \int_{x \cup \xi \in F_{n+1}} b(x, \xi) A_1(x \cup \xi | x) \lambda(d\xi), \quad n \geq 0, \\ Q_1(F_n | x) &= 1[x \in F_n] \left\{ q(x) \int b(x, \xi) [1 - A_1(x \cup \xi | x)] \lambda(d\xi) \right. \\ &\quad \left. + (1 - q(x)) \sum_{\eta \in x} d(x \setminus \eta, \eta) [1 - A_1(x \setminus \eta | x)] \right\}, \quad n \geq 0, \end{aligned}$$

where $\sum_{\eta \in x} \dots$ is set to 1 when $n = 0$, and

$$Q_1(F_{n-1} | x) = (1 - q(x)) \sum_{\eta \in x} 1[x \setminus \eta \in F_{n-1}] d(x \setminus \eta, \eta) A_1(x \setminus \eta | x), \quad n \geq 1.$$

Combining this with (2.1) and (3.3) it is straightforwardly seen that time reversibility holds if and only if

$$\begin{aligned} & \int [1 - q(x \cup \xi)] d(x, \xi) A_1(x | x \cup \xi) f(x \cup \xi) 1[x \in F_n, x \cup \xi \in G_{n+1}] \lambda(dx_1) \dots \lambda(dx_n) \lambda(d\xi) \\ &= \int q(x) b(x, \xi) A_1(x \cup \xi | x) f(x) 1[x \in F_n, x \cup \xi \in G_{n+1}] \lambda(dx_1) \dots \lambda(dx_n) \lambda(d\xi) \end{aligned}$$

for all $F_n \in \mathcal{F}_n$, $G_{n+1} \in \mathcal{F}_{n+1}$, $n \geq 0$, where $\lambda(dx_1) \dots \lambda(dx_n)$ is substituted by $1[x = \emptyset]$ when $n = 0$. Consequently, using (3.1) the following detailed balance condition is seen to ensure time reversibility: for any $x \in \Omega$ and $\xi \in S$,

$$[1 - q(x \cup \xi)] d(x, \xi) A_1(x | x \cup \xi) f(x \cup \xi) = q(x) b(x, \xi) A_1(x \cup \xi | x) f(x) \quad \text{whenever } x \cup \xi \in H.$$

This is fulfilled if

$$q(x \cup \xi) < 1, \quad d(x, \xi) > 0, \quad q(x) > 0, \quad b(x, \xi) > 0 \quad \text{whenever } x \cup \xi \in H, \quad (3.7)$$

and

$$A_1(x | x \cup \xi) = \begin{cases} \min \{1, 1/r(x, \xi)\} & \text{if } x \cup \xi \in H, \\ 0 & \text{otherwise,} \end{cases} \quad (3.8)$$

and

$$A_1(x \cup \xi | x) = \begin{cases} \min \{1, r(x, \xi)\} & \text{if } x \cup \xi \in H, \\ 0 & \text{otherwise,} \end{cases} \quad (3.9)$$

where

$$r(x, \xi) = \frac{f(x \cup \xi)}{f(x)} \frac{1 - q(x \cup \xi)}{q(x)} \frac{d(x, \xi)}{b(x, \xi)} \quad \text{if } x \cup \xi \in H.$$

Then we need only to know the density f up to a constant of proportionality. Moreover, (3.7) clearly implies (3.5) and (3.6).

Of course (3.7)–(3.9) leave us with many possibilities for choosing the density $b(x, \cdot)$ and the probabilities $q(x)$ and $d(x, \cdot)$. In the examples below we consider one of the simplest situations, namely when

$$q(\cdot) \equiv 1/2, \quad b(\cdot, \cdot) \equiv \frac{1}{\lambda(S)}, \quad d(x, \cdot) \equiv \frac{1}{n+1} \quad \text{if } x \in \Omega_n, \quad (3.10)$$

and hence

$$r(x, \xi) = \frac{f(x \cup \xi)}{f(x)} \frac{\lambda(S)}{n+1} \quad \text{if } x \cup \xi \in H_{n+1}. \quad (3.11)$$

Choice of the mixing probability p is discussed in section 5.

Example 1. For most applications f will be the density of a usual point process defined on a bounded Borel set $S \subset \mathbb{R}^d$ with Lebesgue measure $\lambda(S) > 0$, cf. section 2. Then given that the current state of the Markov chain is $x = \{x_1, \dots, x_n\}$ we generate the next proposal y

as follows. If $n = 0$ then the proposal either

- (a) remains in the empty point configuration, or
- (b) becomes a single uniformly distributed point in S .

If $n \geq 1$ then the proposal is either obtained by

- (c) replacing a randomly picked point in x by a uniformly distributed point in S , or
- (d) deleting a randomly picked point in x , or
- (e) adding a new uniformly distributed point in S to x .

Here (a) and (b) occur with probabilities $1 - p/2$ and $p/2$, respectively, while (c), (d), (e) have probabilities $1 - p$, $p/2$, $p/2$, respectively. Finally, in any of the cases (a)–(e) the proposal is accepted with probability $A(y | x) = \min \{1, \lambda(S)^m - n f(y) n! / (f(x) m!)\}$ if $y \in \Omega_m$, cf. (3.8)–(3.11).

A more rapid algorithm may be obtained by generating the uniform point y_i in (c) within some “neighbourhood” $D(x_i) \subset S$ of the deleted point x_i ; e.g. $D(x_i)$ could be a square centred at x_i as in Metropolis *et al.* (1953). Allen & Tildesley (1987, sect. 4.4) discuss how to adjust the side of the square during the simulation. $\square$

Example 2. A concrete case of the preceding example is the unconditional Strauss process (Strauss, 1975; Kelly & Ripley, 1976) on the unit square $S = [0, 1]^2$. This is an exponential family having densities

$$f_\theta(x) = \frac{1}{c(\theta)} \exp(\langle t(x), \theta \rangle) \quad (3.12)$$

with respect to the Poisson process with unit intensity on S , where $\theta = (\theta_1, \theta_2)$ is the canonical parameter and $t(x) = (n(x), s(x))$ is the canonical statistic, $n(x)$ being the number of points in x and $s(x)$ being the number of unordered pairs of points $\{x_i, x_j\} \subseteq x$ such that $\|x_i - x_j\| < r$, where $r > 0$ is a fixed constant not treated here as an unknown parameter, and where $\langle t(x), \theta \rangle$ denotes the inner product $n(x)\theta_1 + s(x)\theta_2$. In order that the process exist, it is necessary that $\theta_2 \leq 0$. If $\theta_2 = 0$ we have a Poisson process μ_{θ_1} with intensity $\exp(\theta_1)$. If $\theta_2 < 0$, the process models repulsion between points. The limit $\theta_2 = -\infty$ gives a hard-core process, the same process that is obtained by conditioning μ_{θ_1} on the event that $\|x_i - x_j\| \geq r$ for all distinct i and j . Simulated point patterns for different values of θ_2 can be seen in Baddeley & Møller (1989). Simulated distributions of the canonical statistics and maximum likelihood estimates will be discussed in section 5. $\square$

Example 3. Point processes as in example 1 can often be approximated by a lattice process, cf. Besag *et al.* (1982) and Jensen & Møller (1991). Moreover, most Metropolis–Hastings algorithms have been developed for lattice processes with a finite state space. It may therefore be illuminating to describe our algorithm for such cases. For ease of presentation we consider a binary state space $\mathcal{Z} = \{0, 1\}^N$ with $0 < N < \infty$ a probability mass function $p(z) > 0$ for all $z \in \mathcal{Z}$, though more general state spaces can easily be included by extending the following to marked point processes.

Clearly, any $z = (z_1, \dots, z_N) \in \mathcal{Z}$ is in one-to-one correspondence with an $x = \{x_1, \dots, x_n\}$ of n distinct points in $S = \{1, \dots, N\}$ so that $z_k = 1$ if and only if $x_i = k$ for some i . Letting λ be the counting measure on S , then the density with respect to μ is seen to be

$$f(x) = n! \exp\{\lambda(S)\} p(z)$$

if $x = \{x_1, \dots, x_n\}$ with $x_i \neq x_j$ for all $i \neq j$, and $f(x) = 0$ otherwise. Hence the acceptance probability becomes the usual Metropolis criterion, i.e. $A(y | x) = \min \{1, p(y)/p(x)\}$ whenever the proposal y corresponds to some $w \in \mathcal{Z}$; if y contains multiple points then y is always rejected. The generation of the proposal can also be described by (a)–(e) as in example 1.

For practical purposes this seems only to be a convenient algorithm when $p(z)$ is naturally expressed in terms of the non-zero z_j s. One example of this is provided by the autologistic model introduced in Besag (1972). $\square$

4. Convergence

In this section we prove convergence of the Markov chain $\{X(k): k = 1, 2, \dots\}$ generated by the Metropolis–Hastings algorithm introduced in section 3.

Recall that (3.1) and (3.7)–(3.9) are assumed to hold. For technical reasons we assume also that

$$P_i(H | x) = 1 \quad \text{and} \quad A_i(\cdot | x) = 1 \quad \text{whenever} \quad x \notin H, \quad i = 0, 1. \quad (4.1)$$

Assume for the moment that $p > 0$, which is the case of main interest in this paper. Then (3.1) and (3.5), (3.6) ensure that the Markov chain is π -irreducible, i.e. for any $x \in \Omega$ and $F \in \mathcal{F}$ with $\pi(F) > 0$ the k -step transition probability

$$P^k(F | x) = P(X(k) \in F | X(1) = x)$$

is strictly positive for some $k \geq 1$. This follows by noting that (4.1) and (3.5) give that $p^{k_1}(\emptyset | x) > 0$ for some k_1 , and then combining (3.6) and (2.1) with Fubini's theorem to obtain that $p^{k_2}(F | \emptyset) > 0$ for some k_2 . Further, modifying the proof of corol. 2 in Tierney (1994, sect. 3.1), it follows that the chain is positive Harris recurrent, that is

$$P(X(k) \in F \text{ for infinitely many } k | X(1) = x) = 1$$

for all $x \in \Omega$ and $F \in \mathcal{F}$ with $\pi(F) > 0$. Hence, if the chain is aperiodic, we have that

$$\lim_{k \rightarrow \infty} \sup_{F \in \mathcal{F}} \left| \pi(F) - \int P^k(F | x) \nu(dx) \right| = 0 \quad (4.2)$$

for any initial distribution ν by Orey's theorem (Nummelin, 1984, corol. 6.7), and for any measurable function $g: \Omega \rightarrow R$ such that $|g|$ is π -integrable,

$$\bar{g}_K \equiv \frac{1}{K} \sum_{k=1}^K g(X(k)) \rightarrow \pi g \equiv \int g(x) \pi(dx) \quad \text{as } K \rightarrow \infty \quad (4.3)$$

almost surely by Birkhoff's ergodic theorem (Tierney, 1994, th. 3). For example, the chain becomes aperiodic when $p < 1$ or $q(\emptyset) < 1$ as this implies that $P^k(\emptyset | \emptyset) > 0$ for all $k > 0$ and hence that $P^k(\emptyset | x) > 0$ for all $x \in \Omega$ and all sufficiently large k .

Similar results hold for the special case $p = 0$, which might be relevant if we want to simulate the conditional distribution π_n of π given that there are n points. Assuming that $\pi(\Omega_n) > 0$ it is easily seen as before that the Markov chain restricted to Ω_n becomes aperiodic, π_n -irreducible, and positive Harris recurrent, so that convergence results as (4.2), (4.3) hold (with π replaced by π_n and with ν concentrated on Ω_n).

It is also important to establish a central limit theorem when it exists, that is

$$\sqrt{K}(\bar{g}_K - \pi g) \xrightarrow{\mathcal{L}} N(0, \sigma^2), \quad \text{as } K \rightarrow \infty. \quad (4.4)$$

We establish it here only for the unconditional sampler described in example 1 of section 3 with $p > 0$ and for the Strauss process (3.12) using prop. 5.21 in Nummelin (1984) to show that the sampler is geometrically ergodic and an observation of Chan & Geyer (1994) that geometric ergodicity implies exponentially fast α -mixing hence a central limit theorem for all functionals in $L^{2+\varepsilon}(\pi)$ for any $\varepsilon > 0$. Below we shall use the following bounds for the acceptance probability: $A(y | x) \geq \min \{1, n(x)/[\beta\lambda(S)]\}$ if $x = y \cup \xi$, while $A(y | x) \leq \beta\lambda(S)/[n(x) + 1]$ if $y = x \cup \xi$.

In order to apply the proposition we need to characterize the "small sets" (Nummelin, 1984, sect. 2.3). A set A is "small" if there is an $\alpha > 0$, a probability measure ν , and an integer k such that the k -step transition probability satisfies $P^k(\cdot | x) \geq \alpha\nu(\cdot)$, for all $x \in A$. The probability of losing a point in one iteration starting at x is bounded below by $pn(x)/[2\beta\lambda(S)]$, where $\beta = \exp(\theta_1)$. Hence if $A = \{x: n(x) \leq a\}$, the probability of the chain being in the empty realization at step $m > a$ is positive and bounded below independently of the starting point x for any $x \in A$. Taking ν to be the probability measure concentrated at the empty realization, we see that any set A of this form is small.

Suppose that $a > \beta\lambda(S)$ (but otherwise unspecified, to be chosen later) so that for $x \notin A$ the probability of accepting a step down is 1 and the probability $Q(x)$ of accepting a step up is less than $\beta\lambda(S)/a$. Let $h(x) = 2^{n(x)}$. Then for $x \notin A$

$$\begin{aligned} E[h(X_{k+1}) | X_k = x] &= 2^{n(x)} \left[(1-p) + \frac{p}{2} (1-Q(x)) \right] + 2^{n(x)+1} \frac{p}{2} Q(x) + 2^{n(x)-1} \frac{p}{2} \\ &= 2^{n(x)} \left\{ 1 - \frac{p}{4} + \frac{p}{2} Q(x) \right\} \leq 2^{n(x)} \left\{ 1 - \frac{p}{4} + \frac{p}{2} \frac{\beta\lambda(S)}{a} \right\}. \end{aligned}$$

Prop. 5.21 in Nummelin requires that for some $r > 1$ the supremum over all $x \notin A$ of

$$\begin{aligned} E[rh(X_{k+1}) - h(X_k) | X_k = x] &\leq 2^{n(x)} \left\{ 1 - \frac{p}{4} + \frac{p}{2} \frac{\beta\lambda(S)}{a} \right\} - 2^{n(x)} \\ &= 2^{n(x)} \left\{ r - 1 - r \frac{p}{4} + r \frac{p}{2} \frac{\beta\lambda(S)}{a} \right\} \end{aligned}$$

be strictly negative, which it will be if r is chosen so that $r - 1 - r(p/4) < 0$ (i.e. $1 < r < 1/(1 - (p/4))$) and a is chosen large enough. The second condition in the proposition is obvious, because $E[h(X_{k+1}) | X_k = x] \leq 2^{n(x)+1}$ is bounded on A .

As noted by Geyer & Chan (1994) this implies that the stationary chain is α -mixing exponentially fast, hence by standard results for stationary processes the central limit theorem (4.4) holds for any $g \in L^{2+\varepsilon}(\pi)$ with $\varepsilon > 0$ and σ^2 given by

$$\sigma^2 = \sum_{k=-\infty}^{\infty} \gamma_k \quad (4.5)$$

where

$$\gamma_k = \gamma_{-k} = \text{cov}[g(X_0), g(X_k)] \quad (4.6)$$

when X_0 has the invariant distribution. For a Harris chain, the asymptotics do not depend on the initial distribution (Geyer, 1992). Central limit theorems for vector functionals with $2 + \varepsilon$ moments follow immediately from the scalar case and the Cramér–Wold device.

Similar results hold for samplers for the conditional Strauss process (take $p = 0$ so the sampler keeps the number of points constant). This case is much simpler. Since the density is bounded and bounded away from zero, so are the odds ratios in the Metropolis steps, and the Markov chain is uniformly ergodic and has a central limit theorem (Tierney, 1994, corol. 3).

5. Empirical results and Monte Carlo maximum likelihood

As a concrete example we use the unconditional Strauss process (3.12) with the following choices of parameters: $\theta_1 = \log(100) \approx 4.6$, $\theta_2 = \log(0.75) \approx -0.288$, $r = 0.1$ (the interaction radius), $p = 0.1$ (the fraction of steps that attempt to change the number of points). It turns out that our choice of $p = 0.1$ was not optimal. According to Allen & Tildesley (1987, p. 128) the value $p = 2/3$ is commonly employed in simulations of molecular systems. However, our later investigations showed that $p = 1$ is optimal and mixed about five times faster than if $p = 0.1$. For example, the integrated autocorrelation time (i.e. the sample size needed to get the same accuracy as with independent sampling) for $n(x)$ reduces from 10.11 to 1.41 and for $s(x)$ from 7.55 to 1.37 when $p = 0.1$ and $p = 1$, respectively.

The Markov chain appears to reach equilibrium starting from the empty state in less than 2500 basic update steps, but for safety we used a "burn-in" of 40,000 basic update steps before collecting samples. The spacing between samples was 200 basic update steps, and 1000 samples were collected. Altogether this makes 240,000 basic update steps, which took 95 seconds on a workstation that does about 1.5 million floating point operations per second.

The time series of the samples are shown in Fig. 1. Some autocorrelation is evident in the time series, but the autocorrelation appears to be significant only out to lag 7 or 8. Geyer (1992) explains why this degree of autocorrelation can give more efficient estimation than using very wide sample spacing in an attempt to get almost uncorrelated samples. We did not, however, attempt to find the optimal sample spacing as described by Geyer (1992). Figure 2 shows the scatter plot of the canonical statistics $n(x)$ and $s(x)$.

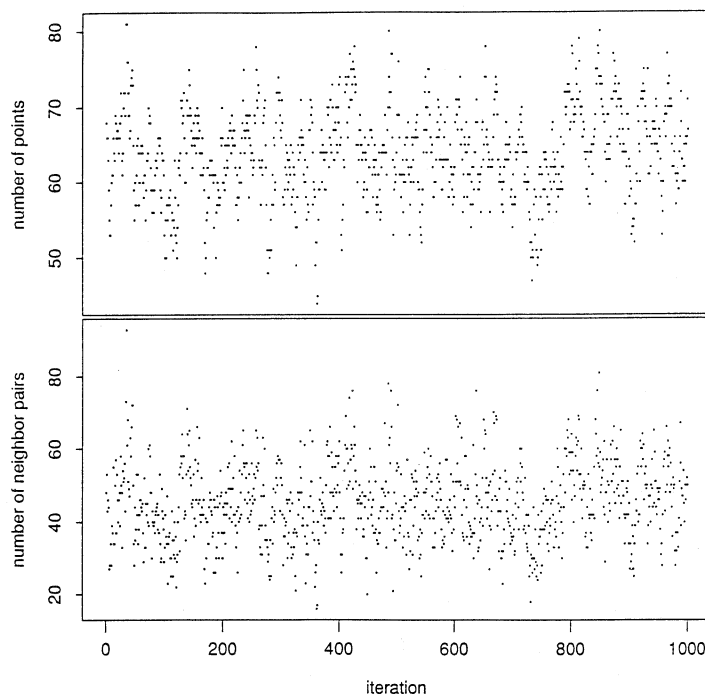


Fig. 1. Time series for the canonical statistics $n(x)$ and $s(x)$ for an unconditional Strauss process on the unit square with canonical parameters $\theta_1 = \log(100)$ and $\theta_2 = \log(0.75)$ and interaction radius $r = 0.1$. The Metropolis algorithm had a warm-up of 200 iterations followed by a run of 1000 iterations. Each iteration comprised 200 basic update steps.

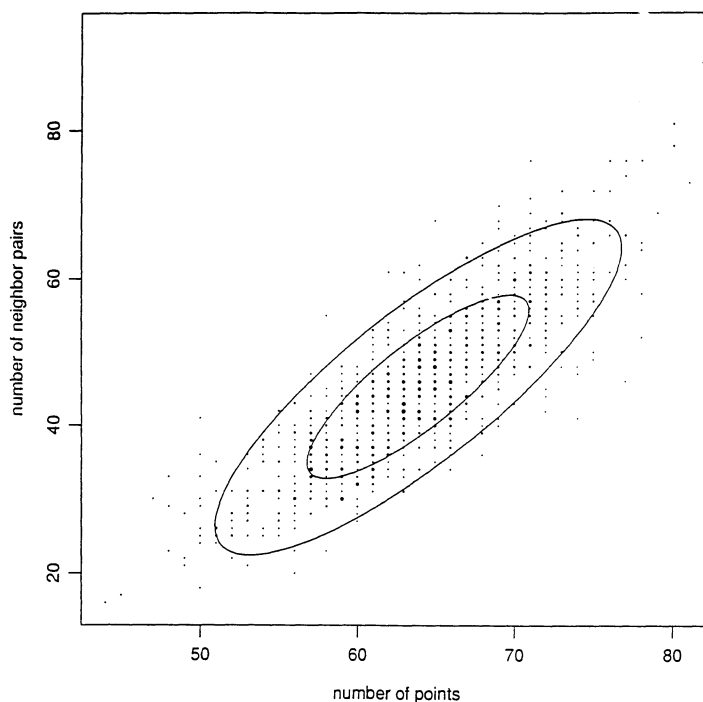


Fig. 2. Scatter plot for the canonical statistics $n(x)$ and $s(x)$ for an unconditional Strauss process of Fig. 1. The ellipses are the 50% and 90% contours of the normal approximation. Multiple points in the plot are indicated by the variable size dots.

Conditional and unconditional Monte Carlo maximum likelihood were done following Geyer & Thompson (1992) and Geyer (1993). Convergence of the algorithm and asymptotic normality of the Monte Carlo error are discussed by Geyer (1994). Details of the calculations will be explained after the results are discussed. Every tenth point in the sample shown in Figs. 1 and 2 was treated as an observation of the process and the MLE of θ and CMLE of θ_2 (maximizing the conditional likelihood given the number of points) were calculated giving the 100 maximum likelihood estimates shown in Fig. 3. The second component θ_2 of the MLE is plotted against the conditional MLE of θ_2 in Fig. 4. The differences between the two estimators in Fig. 4 appear to be no more than Monte Carlo error in the likelihood calculations. Whatever the exact differences are, these calculations do not show them.

Both the MLE and CMLE were constrained to satisfy $\theta_2 \leq 0$ by the simple expedient of setting $\theta_2 = 0$ whenever the unconstrained MLE or CMLE was positive and (for the MLE) also setting $\theta_1 = \log(n(x))$, the correct MLE when the constraint is enforced (and the estimated distribution is a Poisson process). This gives the correct MLE because the log likelihood is concave and hence there is either a unique maximum in the interior of the parameter space or a unique maximum on the boundary.

It might be surmised that for stronger repulsion (larger negative θ_2) there would be some difference between the MLE and CMLE, but further calculations did not reveal any. It appears that for the Strauss process the MLE and CMLE are approximately equivalent for the same reason that these estimators are equivalent in linear regression: both the unconditional and the conditional models are exponential families and the conditional expectation of $s(x)$ given $n(x)$ is approximately a linear function of $n(x)$ as indicated by the elliptical contours for $t(x) = (n(x), s(x))$, see Fig. 2. Hence, if $\hat{\theta}^{(u)}$ and $\hat{\theta}_2^{(c)}$ denote the MLE and

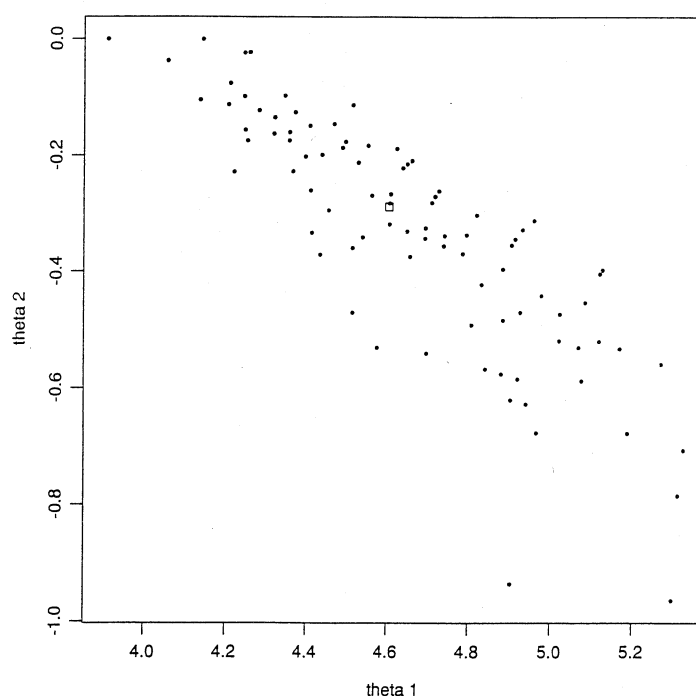


Fig. 3. Scatter plot for the unconditional MLE for the unconditional Strauss process of Fig. 1. The square marks the true parameter value.

CMLE, respectively, the likelihood equations become $t(x) = E_{\hat{\theta}^{(u)}} t(x)$ and $s(x) = E_{\hat{\theta}^{(c)}}(s(x) | n(x))$, and so assuming that $E_{\theta_2}(s(x) | n(x)) \doteq a(\theta_2)n(x) + b(\theta_2)$ is approximately linear in $n(x)$ we find that $a_2(\hat{\theta}_2^{(u)})n(x) + b(\hat{\theta}_2^{(u)}) \doteq a_2(\hat{\theta}^{(c)})n(x) + b(\hat{\theta}_2^{(c)})$ which is easily seen to imply that $\hat{\theta}_2^{(u)} \doteq \hat{\theta}_2^{(c)}$.

The maximum likelihood estimates were calculated using the mixture method of Geyer (1993). Samples were collected from seven different Strauss processes with the sample sizes and parameter values given in Table 1. The samples were then collectively considered a single sample from the mixture distribution with density $f_{\text{mix}} = \sum_j p_j f_{\theta_j}$, where the p_j are non-negative and sum to one. This density is not known exactly since the normalizing constants $c(\theta_j)$ which appear in the formula (3.12) for the density are unknown. They can, however, be accurately estimated from the samples by reverse logistic regression (Geyer, 1993). The log

Table 1. *Parameter values for the mixture sample used in calculating the Monte Carlo log likelihood*

Sample size	θ_1	θ_2
1000	4.61	-0.288
300	4.58	-0.578
300	5.03	-0.289
300	4.20	0.000
300	5.15	-0.479
300	5.15	-0.859
300	4.04	-0.052

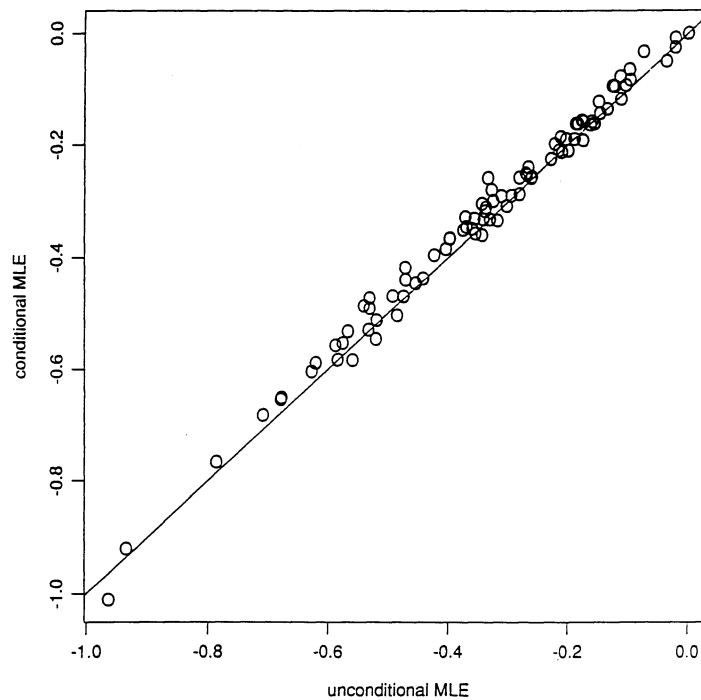


Fig. 4. Scatter plot for the conditional and unconditional MLEs of θ_2 for the unconditional Strauss process of Fig. 1. The line is the locus of points where the two estimators agree.

likelihood for an observation x (discarding some terms not containing θ) is then approximated by

$$l_m(\theta) = \langle t(x), \theta \rangle - \log \left(\sum_{i=1}^m \frac{\exp(\langle t(y_i), \theta \rangle)}{f_{\text{mix}}(y_i)} \right) \quad (5.1)$$

where $y_1, \dots, y_m$ are the point configurations of the mixture sample. Maximizing l_m gives the unconditional Monte Carlo MLE. The conditional likelihood is given by

$$l_m(\theta_2 | n(x)) = s(x)\theta_2 - \log \left(\sum_{i=1}^m \frac{\exp(s(y_i)\theta_2)}{f_{\text{mix}}(y_i)} \right) \quad (5.2)$$

the summation now being over the point configurations y_i in the sample satisfying the conditioning $n(y_i) = n(x)$. Maximizing $l_m(\cdot | n(x))$ gives the Monte Carlo CMLE.

The Monte Carlo approximation of the score in the unconditional case is

$$\frac{\sum_{i=1}^m [t(x) - t(y_i)] \exp(\langle t(y_i), \theta \rangle) / f_{\text{mix}}(y_i)}{\sum_{j=1}^m \exp(\langle t(y_j), \theta \rangle) / f_{\text{mix}}(y_j)}. \quad (5.3)$$

Since the denominator of (5.3) converges to a constant almost surely, there will be a central limit theorem for the Monte Carlo approximation of the score if the functional

$$g_\theta(y) = [t(x) - t(y)] \frac{\exp(\langle t(y), \theta \rangle)}{f_{\text{mix}}(y)} \quad (5.4)$$

has $2 + \varepsilon$ moments with respect to (3.12), cf. section 4. Here we verify this if one of the components of the mixture is Poisson ($\theta_2 = 0$), which was true for the mixture we used (see Table 1). Having $2 + \varepsilon$ moments is ensured if

$$\left[\frac{[t_k(x) - t_k(y)] \exp(\langle t(y), \theta \rangle)}{\sum_i c_i \exp(\langle t(y), \psi_i \rangle)} \right]^{2+\varepsilon} \sum_j c_j \exp(\langle t(y), \psi_j \rangle)$$

is integrable with respect to the Poisson process for $k = 1, 2$, where $t(\cdot) = (t_1(\cdot), t_2(\cdot)) = (n(\cdot), s(\cdot))$, the ψ_j index the components of the mixture, and the c_j are normalizing constants. Let J be the index such that $\psi_{J,2} = 0$. Then

$$\begin{aligned} & \left[\frac{[t_k(x) - t_k(y)] \exp(\langle t(y), \theta \rangle)}{\sum_i c_i \exp(\langle t(y), \psi_i \rangle)} \right]^{2+\varepsilon} \sum_j c_j \exp(\langle t(y), \psi_j \rangle) \\ & \leq \left[\frac{[t_k(x) - t_k(y)] \exp(\langle t(y), \theta \rangle)}{c_J \exp(\langle t(y), \psi_J \rangle)} \right]^{2+\varepsilon} \sum_j c_j \exp(\langle t(y), \psi_j \rangle) \\ & = c_J^{-(2+\varepsilon)} \sum_j c_j [t_k(x) - t_k(y)]^{2+\varepsilon} \exp\{\langle t(y), (2+\varepsilon)(\theta - \psi_J) + \psi_J \rangle\}. \end{aligned}$$

Hence (5.4) always has $2 + \varepsilon$ moments, because $(2 + \varepsilon)(\theta_2 - \psi_{J,2}) + \psi_{J,2}$ is non-positive. In practice the c_j are not constants but estimated. However, the asymptotics are the same when the consistent estimator of the c_j given by reverse logistic regression is plugged in (Geyer, 1993, ths. 3 and 4).

Geyer (1994, th. 7) gives a central limit theorem for $\sqrt{K}(\hat{\theta}_K - \hat{\theta})$, the deviation of the Monte Carlo MLE $\hat{\theta}_K$ from the exact MLE $\hat{\theta}$. Here the assumptions are easily verified because of the exponential family structure and since there is a central limit theorem for (5.3). At least this gives a central limit theorem when f_{mix} is considered to be known. But then Geyer (1993, ths. 3 and 4) shows that the central limit theorem holds as the normalizing constants in f_{mix} are consistently estimated by reverse logistic regression.

Similarly, in the conditional case, we obtain central limit theorems for the Monte Carlo approximation of the conditional score as well as the deviation of the Monte Carlo CMLE from the exact CMLE.

The point of the mixture estimation is to get information about the tails of the distribution of the canonical statistic, which is best done by sampling not only from the true distribution but also from other distributions in the family that put mass at various points in the tails of the true distribution. This makes Monte Carlo maximum likelihood a stable estimation procedure with approximately the same Monte Carlo error over a wide range of observed values of the canonical statistic. Thus, from one approximate log likelihood (5.1) we can get estimates for 100 different observations that span the probable values of the canonical statistic. The scatter plot of these MLEs (Fig. 2) approximates the sampling distribution of the MLE.

6. Discussion

As mentioned in section 1, the well-known algorithm for point processes is a spatial birth-and-death process. This is a time-homogeneous jump process, and simple conditions for existence, convergence, and geometrically fast convergence can be found in Preston (1977) and Møller (1989). However, central limit theorems as well as Monte Carlo methods

and their practical aspects are yet not much explored for spatial birth-and-death processes. In contrast, our Metropolis–Hastings algorithm is much easier to analyse as demonstrated in sections 4 and 5. It involves fewer calculations and is also easier to program, but whether it runs faster depends of course on the implementation.

In section 5 we established a central limit theorem for the deviation of the Monte Carlo approximation from the exact MLE (or CMLE). This central limit theorem, call it the CLT for the Monte Carlo error, is determined by the properties of the Metropolis–Hastings algorithm and is unrelated to the central limit theorem for the exact MLE as the size of the window containing the point process is increased, call that the CLT for the exact MLE. For most point processes, it is not known whether there is a central limit theorem for the MLE in the latter sense; results seem to be available only for processes which are very close to the Poisson process (Jensen, 1991, 1993). For example, in order to apply th. 1 in Jensen (1991) the Strauss process (3.12) should satisfy the condition that

$$r \exp(1 + \theta_1)(1 - \exp(\theta_2)) < 1,$$

which for $\theta_1 = \log(100)$ and $r = 0.1$ implies that $\exp(\theta_2) > 0.96$, which is practically the Poisson case with no repulsion.

The simulation reported in section 5 does not depend on the existence of a CLT for the exact MLE. Given enough simulated observations to get an accurate enough approximation to the log likelihood, the sampling distribution of the MLE for a fixed window can be determined by arbitrary accuracy. The situation changes, however, if our simulation becomes a parametric bootstrap. If given one observation x of a point process, we assume a Strauss model and estimate the Monte Carlo MLE $\hat{\theta}(x)$ and then simulate realizations $x_1^*, \dots, x_m^*$ from the process with parameters $\hat{\theta}(x)$ and determine the MLE $\hat{\theta}(x_i^*)$ for each (using perhaps the same Monte Carlo approximation to the log likelihood), the $\hat{\theta}(x_i^*)$ are a parametric bootstrap approximation to the sampling distribution of the MLE. But lacking a CLT for the exact MLE, we cannot give a justification for the bootstrap of the usual sort, i.e. that it gives the same asymptotics. Though the parametric bootstrap seems reasonable, we have no theory to support it.

As mentioned previously, we have made no attempt to find optimal parameters for the simulations, being content with what past experience with other problems has shown to be reasonable. Finding an optimal design for the simulations would be a very complicated problem. Each MLE calculation has a different Monte Carlo error, since it involves the Hessian of the Monte Carlo log likelihood at $\hat{\theta}$ and the central limit theorem for the score at $\hat{\theta}$, and both of these vary with $\hat{\theta}$. Moreover, the mixture reweighting method involves a choice of sampling distributions making up the mixture. The accuracy of the Monte Carlo approximation to the log likelihood depends on the choice of the mixture in a complicated manner for which there is as yet no theory.

Despite the lack of optimality criteria, the Metropolis–Hastings algorithm is a useful method for simulating spatial point processes, and Monte Carlo maximum likelihood is a useful method for estimating the parameters of the processes and investigating the statistical properties of the estimators.

Acknowledgements

This research was started during the summer of 1992 at the Center for Spatial Statistics and the Department of Statistics, University of Washington. We thank Julian Besag and Elizabeth Thompson for inviting us. The research was also partially supported by the

National Science Foundation grant DMS-9007833. Masaharu Tanemura called our attention on Allen & Tildesley (1987). We appreciate also the referees' comments.

References

- Allen, M. P. & Tildesley, D. J. (1987). *Computer simulation of liquids*. Oxford University Press, Oxford.
- Baddeley, A. & Møller, J. (1989). Nearest-neighbour Markov point processes and random sets. *Int. Statist. Rev.* **57**, 89–121.
- Besag, J. (1972). Nearest-neighbour systems and the auto-logistic model for binary data. *J. Roy. Statist. Soc. Ser. B* **34**, 75–83.
- Besag, J., Milne, R. & Zachary, S. (1982). Point process limits of lattice processes. *J. Appl. Probabil.* **19**, 210–216.
- Carter, D. S. & Prenter, P. M. (1972). Exponential spaces and counting processes. *Z. Wahr. verw. Geb.* **21**, 1–19.
- Chan, K. S. & Geyer, C. J. (1994). Discussion of Tierney (1994). *Ann. Statist.* To appear.
- Diggle, P. J. (1983). *Statistical analysis of spatial point patterns*. Academic Press, London.
- Geyer, C. J. (1993). Estimating normalizing constants and reweighting mixtures in Markov chain Monte Carlo. Technical Report No. 568, School of Statistics, University of Minnesota. (Revised version).
- Geyer, C. J. (1992). Practical Markov chain Monte Carlo. *Statist. Sci.* **7**, 473–511.
- Geyer, C. J. (1994). On the convergence of Monte Carlo maximum likelihood calculations. *J. Roy. Soc. Ser. B* **56**, 261–274.
- Geyer, C. J. & Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *J. Roy. Statist. Soc. Ser. B* **54**, 657–699.
- Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.
- Jensen, J. L. (1991). A note on asymptotic normality in the thermodynamic limit at low densities. *Adv. Appl. Math.* **12**, 387–399.
- Jensen, J. L. (1993). Asymptotic normality of estimates in spatial point processes. *Scand. J. Statist.* **20**, 97–107.
- Jensen, J. L. & Møller, J. (1991). Pseudolikelihood for exponential family models of spatial point processes. *Ann. Appl. Probabil.* **1**, 445–461.
- Kelly, F. P. & Ripley, B. D. (1976). A note on Strauss' model for clustering. *Biometrika* **63**, 357–360.
- Lotwick, H. W. & Silverman, B. W. (1981). Convergence of spatial birth-and-death processes. *Math. Proc. Camb. Phil. Soc.* **90**, 155–165.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1092.
- Møller, J. (1989). On the rate of convergence of spatial birth-and-death processes. *Ann. Inst. Statist. Math.* **41**, 565–581.
- Møller, J. (1994). Markov chain Monte Carlo and spatial point processes. In preparation.
- Moyeed, R. A. & Baddeley, A. J. (1991). Stochastic approximation of the MLE for a spatial point pattern. *Scand. J. Statist.* **18**, 39–50.
- Norman, G. E. & Filinov, V. S. (1969). Investigations of phase transitions by a Monte-Carlo method. *High Temp. (USSR)* **7**, 216–222.
- Nummelin, E. (1984). *General irreducible Markov chains and non-negative operators*. Cambridge University Press, Cambridge.
- Ogata, Y. & Tanemura, M. (1981). Estimation of interaction potentials of spatial point patterns through the maximum likelihood procedure. *Ann. Inst. Statist. Math.* **33B**, 315–338.
- Ogata, Y. & Tanemura, M. (1989). Likelihood estimation of soft-core interaction potentials for Gibbsian point patterns. *Ann. Inst. Statist. Math.* **41**, 583–600.
- Penttinen, A. (1984). Modelling interaction in spatial point patterns: parameter estimation by the maximum likelihood method. *Jyväskylä Stud. Comput. Sci. Econ. Statist.* **7**.
- Preston, C. J. (1977). Spatial birth-and-death processes. *Bull. Int. Statist. Inst.* (2) **46**, 371–391.
- Ripley, B. D. (1977). Modelling spatial patterns (with discussion). *J. Roy. Statist. Soc. Ser. B* **39**, 172–212.
- Ripley, B. D. (1979). Simulating spatial patterns: dependent samples from a multivariate density. Algorithm AS 137. *Appl. Statist.* **28**, 109–112.
- Ripley, B. D. (1981). *Spatial statistics*. Wiley, New York.

- Ripley, B. D. (1988). *Statistical inference for spatial processes*. Cambridge University Press, Cambridge.
- Strauss, D. J. (1975). A model for clustering. *Biometrika* **62**, 467–475.
- Tierney, L. (1994). Markov chains for exploring posterior distributions. To appear in *Ann. Statist.*
- Younes, L. (1988). Estimation and annealing for Gibbsian fields. *Ann. Inst. H. Poincaré* **24**, 269–294.

Received March 1993, in final form September 1993

Charles J. Geyer, School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA.
Jesper Møller, Department of Theoretical Statistics, Institute of Mathematics, University of Aarhus, DK-8000, Aarhus C, Denmark.