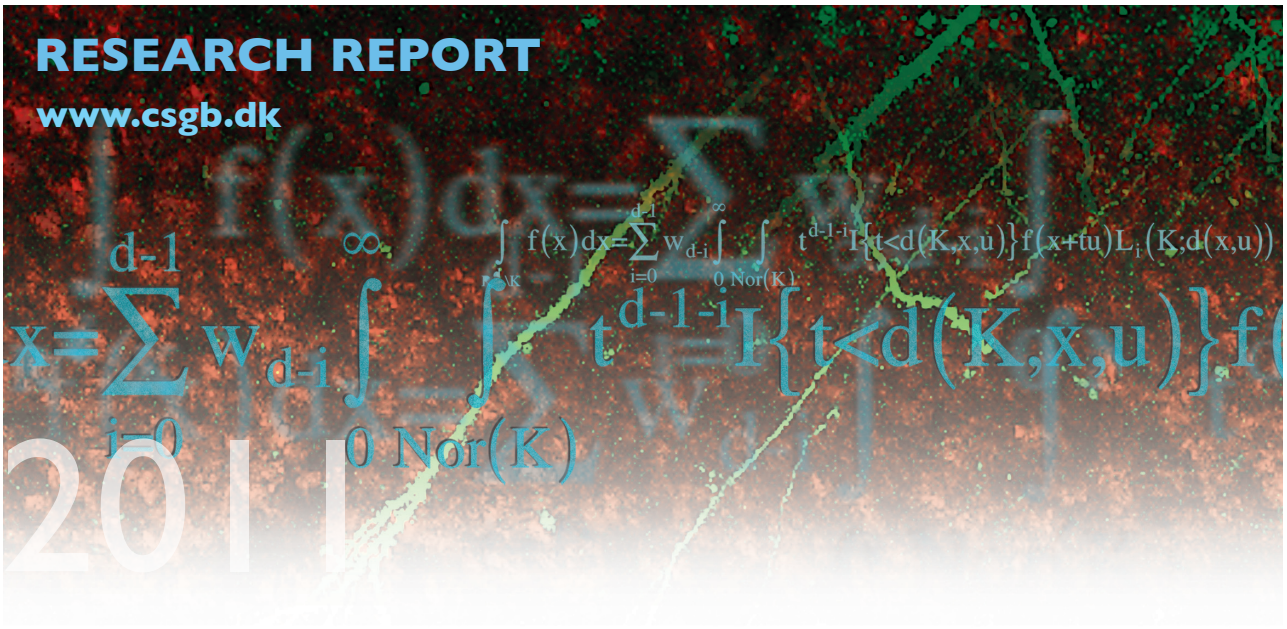
CENTRE FOR **STOCHASTIC GEOMETRY**
AND ADVANCED **BIOIMAGING**



Jakob Gulddahl Rasmussen

Bayesian Inference for Hawkes Processes

No. 01, February 2011

Bayesian Inference for Hawkes Processes

Jakob Gulddahl Rasmussen

Department of Mathematical Sciences
Aalborg University jgr@math.aau.dk

Abstract

The Hawkes process is a practically and theoretically important class of point processes, but parameter-estimation for such a process can pose various problems. In this paper we explore and compare two approaches to Bayesian inference. The first approach is based on the so-called conditional intensity function, while the second approach is based on an underlying clustering and branching structure in the Hawkes process. For practical use, MCMC (Markov chain Monte Carlo) methods are employed. The two approaches are compared numerically using three examples of the Hawkes process.

Keywords: Bayesian inference, cluster process, Hawkes process, Markov chain Monte Carlo, missing data, point process

1 Introduction

The (marked) Hawkes process (or self-exciting process) is an important class of marked point processes (Hawkes, 1971a,b, 1972; Hawkes and Oakes, 1974). It has seen many applications so far, primarily in seismology (e.g. Hawkes and Adamopoulos (1973); Ogata (1988, 1998)), but also in other areas such as neurophysiology (Chornoboy *et al.*, 2002), criminology (Mohler *et al.*, 2011) and biology (Balderama *et al.*, 2010). Thus it is of great importance to have efficient and precise estimation procedures for the fitting the parameters of the Hawkes process to real data.

As was pointed out in Hawkes and Oakes (1974), the Hawkes process can be defined in two equivalent ways: either using the so-called conditional intensity function or as a Poisson cluster process with a certain branching structure (see Sections 2.1 and 2.2). The definition using the conditional intensity function immediately leads to an expression for the likelihood function, and although it cannot be maximized analytically, it is possible to make numerical procedures for maximizing this function. However, it is observed in Veen and Schoenberg (2008) that such approximative maximum likelihood estimation can be numerically unstable. The definition using the Poisson cluster process formulation defines the Hawkes process using an (in practice unobserved) clustering and branching structure. In Veen and Schoenberg (2008) this is used to make an alternative maximum likelihood estimation procedure using the EM (expectation-maximization) algorithm, which they observe is more numerically stable.

In the present paper we will explore two approaches to Bayesian inference, each based on one of the two definitions. The first approach simply uses the conditional intensity function to define the likelihood function, and then approximates the posterior distribution of the parameters using an MCMC approach. In the second approach the clustering and branching structure is regarded as missing data, and the Hawkes process is separated into a number of Poisson processes. Again MCMC is used for estimation, but in this case the parameters and the missing data are estimated simultaneously.

The outline of the paper is as follows: Section 2 states the two definitions of the Hawkes process, and gives three examples. In Sections 3 and 4, Bayesian inference based on the conditional intensity function and based on the clustering and branching structure are described. In Section 5, the two approaches are compared using the three examples of Section 2, and Section 6 concludes the paper with possible extensions of the methods.

2 Two definitions of the Hawkes process

In this section the Hawkes process is defined in two equivalent ways and examplified.

2.1 Definition using conditional intensity function

Let $X = \{(t_i, \kappa_i)\}$ be a marked point process on the time line, where $t_i \in \mathbb{R}$ denotes the points (or events) of the point process, and $\kappa_i \in \mathbb{M}$ denotes the marks, where $\mathbb{M}$ is a measurable space called the mark space. Furthermore let N be its corresponding counting measure, i.e. $N(B)$ is the number of points falling in an arbitrary Borel set $B \subseteq \mathbb{R}$. See Daley and Vere-Jones (2003) for more details on point processes.

One way of defining a marked point process is by specifying its conditional intensity function and mark distribution. The conditional intensity function is defined as

$$\lambda^*(t) = \frac{E(N(\mathrm{d}t)|\{(t_i, \kappa_i)\}_{t_i<t})}{\mathrm{d}t}.$$

The intuitive interpretation of the conditional intensity function is that $\lambda^*(t)\mathrm{d}t$ is the mean number of points falling in an infinitesimal interval around t given the knowledge about all points in the past and their marks. Note that the dependence on the past is suppressed in the notation $\lambda^*(t)$; here the notation of Daley and Vere-Jones (2003) has been adopted, where $*$ is supposed to remind us of the fact that this function is allowed to depend on the past points and marks, i.e. $\{(t_i, \kappa_i)\}_{t_i<t}$. The mark distribution is most conveniently described by its density function (with respect to some reference measure on $\mathbb{M}$) given the past and the time of the point

$$\gamma^*(\kappa|t) = \gamma(\kappa|t, \{(t_i, \kappa_i)\}_{t_i<t})$$

again using the $*$ notation to represent the past.

The Hawkes process can now be defined using a particular form of the conditional intensity function. Firstly we need to define the following functions, where the choice of names and the interpretations of the functions should be clear in Section 2.2:

- Immigrant intensity: $\mu(t)$ is a non-negative function on $\mathbb{R}$ with parameter vector $\mu = (\mu_1, \ldots, \mu_{n_\mu})$.
- Total offspring intensity: $\alpha(\kappa)$ is a non-negative function on $\mathbb{M}$ with parameter vector $\alpha = (\alpha_1, \ldots, \alpha_{n_\alpha})$.
- Normalised offspring intensity: $\beta(t, \kappa)$ is a density function on $[0, \infty)$ with parameter vector $\beta = (\beta_1, \ldots, \beta_{n_\beta})$ which is allowed to depend on the mark κ.
- Mark density: $\gamma^*(\kappa|t)$ is a density function on $\mathbb{M}$ depending on t and the past before time t (although we will restrict this dependence somewhat in Section 2.2) with a parameter vector $\gamma = (\gamma_1, \ldots, \gamma_{n_\gamma})$.

The product $\alpha(\kappa)\beta(t, \kappa)$ is called the offspring intensity. Using these functions we can define the Hawkes process as the point process with the conditional intensity function

$$\lambda^*(t) = \mu(t) + \sum_{t_i < t} \alpha(\kappa_i)\beta(t - t_i, \kappa_i), \tag{2.1}$$

where $t \in \mathbb{R}$.

2.2 Definition as a Poisson cluster process

An alternative way of defining the Hawkes process is to define it as a marked Poisson cluster process, where the clusters are generated by a certain branching structure. Here we distinguish between two types of points - immigrants and offspring - and have the following definition:

1. The immigrants I follow a Poisson process with intensity $\mu(t)$.
2. Each immigrant $t_i \in I$ has an associated mark κ_i with density function $\gamma(\kappa_i|t_i)$.
3. Each immigrant $t_i \in I$ generates a cluster C_i, and these clusters are independent.
4. A cluster C_i consists of points of generations of order $n = 0, 1, \ldots$ with the following branching structure: Generation 0 consists simply of the immigrant and its mark (t_i, κ_i). Recursively, given the $0, \ldots, n$ generations in C_i, each t_j of generation n generates a Poisson process O_j of offspring of generation $n + 1$ with intensity function $\alpha(\kappa_j)\beta(t - t_j, \kappa_j)$. Each offspring $t_k \in O_j$ has an associated mark κ_k with density function $\gamma(\kappa_k|t_k, (t_j, \kappa_j))$.
5. Finally, X consists of the union of all clusters.

If $t_j \in O_i$, we say that t_j is the child (or first order offspring) of t_i or that t_i is the parent (or first order ancestor) of t_j. We also denote the index of the parent t_i of t_j by $i = \mathrm{pa}(j)$. The collection of relations between all points and their parents (if any), we call the branching structure. The branching structure is conveniently represented as $Y = \{y_j\}$, where $y_j = i$ if $t_j \in O_i$ or $y_j = 0$ if t_j is an immigrant. The names used for $\mu(\cdot)$, $\alpha(\cdot)$ and $\beta(\cdot)$ in Section 2.1 should make sense when viewed in connection with the definition given in this section; note that $\beta(t, \kappa)$ is the density

function for the length of the time interval between a child and its parent, while $\alpha(\kappa)$ is the mean number of children of a point with mark κ.

One important thing to notice is that this definition is a restriction of the definition in Section 2.1, since the mark is not allowed to depend on the whole past, but only the time and mark of the parent (or nothing in the case of an immigrant). Since the definition in Section 2.1 does not distinguish between point types, the γ^* used there becomes a mixture distribution with mixture weights proportional to the intensities of being different types of points, i.e.

$$\gamma^*(\kappa|t) = \frac{1}{\lambda^*(t)}\left(\mu(t)\gamma(\kappa|t) + \sum_{t_i<t} \alpha(\kappa_i)\beta(t-t_i,\kappa_i)\gamma(\kappa|t,(t_i,\kappa_i))\right). \tag{2.2}$$

Note that in the case of i.i.d. (independent and identically distributed) marks this simplifies to $\gamma^*(\kappa|t) = \gamma(\kappa|t) = \gamma(\kappa|t,(t_i,\kappa_i))$.

2.3 Examples

We will consider three examples in this paper. These are chosen to test whether the methods in this paper work on examples of the following the cases: an unmarked Hawkes process, a marked Hawkes process with independent marks, and a marked Hawkes process with dependent marks.

Example 1 is one of the most simple cases of a Hawkes process: an unmarked process with exponentially decaying offspring intensity. More precisely,

$$\begin{aligned}\mu(t) &= \mu_1 \mathbf{1}(t \geq 0),\\ \alpha(\kappa) &= \alpha_1,\\ \beta(t,\kappa) &= \beta_1 e^{-\beta_1 t}.\end{aligned}$$

Here $\mathbf{1}(\cdot)$ denotes the indicator function. Note that the term $\mathbf{1}(t \geq 0)$ ensures that there are no points before time 0; this is done to avoid dealing with so-called edge-effects which is outside the scope of this paper (see Section 6 for more details). Also note that the unmarked Hawkes process is a special case of the marked Hawkes process, and all formulas in this paper apply simply by removing $\gamma^*(\cdot)$ or $\gamma(\cdot)$ if present.

Example 2 can be thought of as a simple model for a reproducing population with exponential survival times, where the individuals reproduce uniformly throughout their survival times. This example is defined by

$$\begin{aligned}\mu(t) &= \mu_1 \mathbf{1}(t \geq 0),\\ \alpha(\kappa) &= \alpha_1 \kappa,\\ \beta(t,\kappa) &= \mathbf{1}(t \in (0,\kappa))/\kappa,\\ \gamma^*(\kappa|t) &= \gamma_1 e^{-\gamma_1 \kappa}.\end{aligned}$$

Note that γ_1 is the inverse mean survival time and α_1/γ_1 is the mean number of children of any point.

Example 3 is an example of the ETAS (epidemic type aftershock sequences) model, commonly used for modeling the times, magnitudes and sometimes positions

of earthquakes. A lot of research has gone into modelling earthquakes, but since the main reason for including the model in this paper is illustration of the methods, it will be kept fairly simple here in order to focus on how well the methods handle dependent marks. The ETAS model used in this paper includes the mark $\kappa = (m, x, y)$, i.e. magnitude $m \in (0, \infty)$ and coordinates $(x, y) \in W$ of the epicenter of an earthquake, where W is some observation window of the positions. It is defined by the functions

$$\begin{aligned}\mu(t) &= \mu_1 \mathbf{1}(t \geq 0),\\ \alpha(\kappa) &= \alpha_1 e^{\alpha_2 m},\\ \beta(t, \kappa) &= \frac{\beta_2}{\beta_1}\left(1 + \frac{t}{\beta_1}\right)^{-\beta_2 - 1},\end{aligned}$$

the mark density for the immigrants

$$\gamma(\kappa|t) = \gamma_1 e^{-\gamma_1 m} \frac{\mathbf{1}((x, y) \in W)}{|W|},$$

and the mark density for the offspring

$$\gamma(\kappa|t, (t_{\text{pa}}, \kappa_{\text{pa}})) = \gamma_1 e^{-\gamma_1 m} \frac{1}{2\pi\gamma_2^2} \exp\left(-\frac{\|(x, y) - (x_{\text{pa}}, y_{\text{pa}})\|^2}{2\gamma_2^2}\right),$$

where the index pa denotes the parent. Note that this means that the magnitudes are i.i.d. exponential variables, and the positions follow a uniform distribution for the main earthquakes and a normal distribution centered on the parent for the aftershocks.

3 Bayesian inference based on the conditional intensity function

In this section we define one approach to Bayesian inference for the Hawkes process based on the definition using the conditional intensity function in Section 2.1. We will call this the conditional intensity based method.

3.1 Likelihood, prior and posterior

Assume we have observed a dataset of points given by a marked point pattern $x = \{(t_1, \kappa), \ldots, (t_n, \kappa_n)\}$ on $[0, T) \times \mathbb{M}$ for some fixed time $T > 0$, and no points have occurred before 0. Then by Proposition 7.3.III in Daley and Vere-Jones (2003), the likelihood function is given by

$$p(x|\mu, \alpha, \beta, \gamma) = \left(\prod_{i=1}^{n} \lambda^*(t_i)\gamma^*(\kappa_i|t_i)\right) \exp(-\Lambda^*(T)), \tag{3.1}$$

where

$$\Lambda^*(t) = \int_0^t \lambda^*(s)\mathrm{d}s = M(t) + \sum_{t_i < t} \alpha(\kappa_i) B(t - t_i, \kappa_i), \qquad M(t) = \int_0^t \mu(s)\mathrm{d}s$$

and B is the distribution function corresponding to the density function β.

If we denote the prior by $p(\mu, \alpha, \beta, \gamma)$, we get the posterior

$$p(\mu, \alpha, \beta, \gamma|x) \propto p(\mu, \alpha, \beta, \gamma)p(x|\mu, \alpha, \beta, \gamma). \tag{3.2}$$

3.2 Markov chain Monte Carlo

The posterior (3.2) is not on a form that allows us to find the maximum or mean of the posterior for the parameters analytically, so instead we turn to MCMC. More specifically, we use a Metropolis-within-Gibbs algorithm where each parameter is updated one at a time. As proposal distributions we use normal distributions.

For updating μ_k for $k = 1, \ldots, n_\mu$ we draw $\tilde{\mu}_k$ from a normal distribution with the current parameter value μ_k as mean and some fixed standard deviation σ_{μ_k}. Similar updates are used for the other parameters, but with σ_{α_k}, σ_{β_k}, and σ_{γ_k} as standard deviations. From (3.2) we immediately get the Hastings ratios for these updates; for example, the Hastings ratio for updating μ_k to the proposed value $\tilde{\mu}_k$ is given by

$$\begin{aligned} H_{\mu_k} &= \frac{p(\tilde{\mu}, \alpha, \beta, \gamma)}{p(\mu, \alpha, \beta, \gamma)} \frac{p(\tilde{\mu}, \alpha, \beta, \gamma|x)}{p(\mu, \alpha, \beta, \gamma|x)} \\ &= \frac{p(\tilde{\mu}, \alpha, \beta, \gamma)}{p(\mu, \alpha, \beta, \gamma)} \left(\prod_{i=1}^{n} \frac{\tilde{\mu}(t_i) + \sum_{j<i} \alpha(\kappa_j)\beta(t_i - t_j|\kappa_i)}{\mu(t_i) + \sum_{j<i} \alpha(\kappa_j)\beta(t_i - t_j|\kappa_i)} \right) \exp\big(M(t) - \tilde{M}(t)\big) \end{aligned}$$

where $\tilde{\mu}$, $\tilde{\mu}(\cdot)$ and $\tilde{M}(\cdot)$ denotes μ, $\mu(\cdot)$ and $M(\cdot)$ with the proposed value $\tilde{\mu}_k$ inserted instead of μ_k and all other parameters are left unchanged. Similar expressions can easily be obtained for the Hastings ratios for updating the other parameters.

4 Bayesian inference based on the clustering and branching structure

In this section we define another approach to Bayesian inference. This is based on the definition of the Hawkes process as a Poisson cluster process in Section 2.2. We call this the cluster based method.

4.1 Bayesian inference with missing data

As in Section 3.1 we assume that we have a dataset of marked points given by $x = \{(t_1, \kappa), \ldots, (t_n, \kappa_n)\}$ in $[0, T) \times \mathbb{M}$. If we want to base an estimation approach on the Poisson cluster process formulation, we run into the problem that the branching structure Y is unobserved. Treating this as missing data, we simultaneously have to estimate the missing data Y and the parameters $(\mu, \alpha, \beta, \gamma)$. For this we need the conditional distributions of both the parameters given the missing data and the missing data given the parameters. These are used for setting up a Gibbs sampler in Section 4.2.

Knowing the branching structure, we can separate the dataset into a number of independent marked Poisson processes: I is the process of marked immigrants and O_j is the process of marked children of t_j for $j = 1, \ldots, n$. It follows from Section 2.2 that these processes have intensity functions

$$\lambda_I(t) = \mu(t) \qquad \text{and} \qquad \lambda_{O_j}(t) = \alpha(\kappa_j)\beta(t - t_j, \kappa_j), \tag{4.1}$$

and mark densities

$$\gamma_I(\kappa|t) = \gamma(\kappa|t) \qquad \text{and} \qquad \gamma_{O_j}(t) = \gamma(\kappa|t, (t_j, \kappa_j)). \tag{4.2}$$

If we condition on the branching structure $Y = y$, the independence means that we get the following conditional likelihood

$$p(x|y, \mu, \alpha, \beta, \gamma) = p(I|y, \mu, \gamma) \prod_{j=1}^{n} p(O_j|y, \alpha, \beta, \gamma).$$

Working with the term for the immigrants first, we get from (3.1), (4.1) and (4.2) that

$$p(I|y, \mu, \gamma) = \left(\prod_{t_i \in I} \mu(t_i)\gamma(\kappa_i|t_i) \right) \exp(-M(T)). \tag{4.3}$$

For offspring process O_j, again using (3.1), (4.1) and (4.2), we get that

$$\begin{aligned} p(O_j|y, \alpha, \beta, \gamma) = &\left(\prod_{t_i \in O_j} \alpha(\kappa_j)\beta(t_i - t_j, \kappa_j)\gamma(\kappa_i|t_i, (t_j, \kappa_j)) \right) \\ &\times \exp\left(-\alpha(\kappa_j)B(T - t_j, \kappa_j)\right). \end{aligned}$$

Since the offspring processes are independent, we can multiply these to get the joint likelihood for the offspring processes

$$\begin{aligned} p(O|y, \alpha, \beta, \gamma) = &\left(\prod_{t_i \in O} (\alpha(\kappa_{\mathrm{pa}(i)})\beta(t_i - t_{\mathrm{pa}(i)}, \kappa_{\mathrm{pa}(i)})\gamma(\kappa_i|t_i, (t_{\mathrm{pa}(i)}, \kappa_{\mathrm{pa}(i)})) \right) \\ &\times \exp\left(-\sum_{t_j \in x} \alpha(\kappa_j)B(T - t_j) \right), \end{aligned} \tag{4.4}$$

where $O = (O_1, \ldots, O_n)$ denotes the collection of offspring processes.

4.2 Markov chain Monte Carlo

Again we have to use a Metropolis-within-Gibbs algorithm, where we update each of the parameters and some of the y_i in the branching structure. Actually in some simple cases the conditional distributions of some of the parameters are well-known distributions, so we can update these parameters without employing a Metropolis update; e.g. in Example 1 the conditional distribution of μ_1 is a Gamma distribution. However, this is the exception rather than the rule, so we will ignore this simplification and instead focus on the general case.

For the parameter updates we again use normally distributed proposals, and from (4.3) and (4.4) we get the Hastings ratios for each type of updates

$$
\begin{aligned}
H_{\mu_k} &= \frac{p(\tilde{\mu},\alpha,\beta,\gamma)}{p(\mu,\alpha,\beta,\gamma)} \prod_{t_i\in I}\left(\frac{\tilde{\mu}(t_i)}{\mu(t_i)}\right) \exp\Big(M(T)-\tilde{M}(T)\Big) \\
H_{\alpha_k} &= \frac{p(\mu,\tilde{\alpha},\beta,\gamma)}{p(\mu,\alpha,\beta,\gamma)} \prod_{t_i\in O}\left(\frac{\tilde{\alpha}(\kappa_{\mathrm{pa}(i)})}{\alpha(\kappa_{\mathrm{pa}(i)})}\right) \exp\bigg(\sum_{t_j\in x}(\alpha(\kappa_j)-\tilde{\alpha}(\kappa_j))B(T-t_j,\kappa_j)\bigg) \\
H_{\beta_k} &= \frac{p(\mu,\alpha,\tilde{\beta},\gamma)}{p(\mu,\alpha,\beta,\gamma)} \prod_{t_i\in O}\left(\frac{\tilde{\beta}(t_i-t_{\mathrm{pa}(i)},\kappa_{\mathrm{pa}(i)})}{\beta(t_i-t_{\mathrm{pa}(i)},\kappa_{\mathrm{pa}(i)})}\right) \\
&\quad \times \exp\bigg(\sum_{t_j\in x}\alpha(\kappa_j)\Big(B(T-t_j,\kappa_j)-\tilde{B}(T-t_j,\kappa_j)\Big)\bigg) \\
H_{\gamma_k} &= \frac{p(\mu,\alpha,\beta,\tilde{\gamma})}{p(\mu,\alpha,\beta,\gamma)} \prod_{t_i\in I}\left(\frac{\tilde{\gamma}(\kappa_i|t_i)}{\gamma(\kappa_i|t_i)}\right) \prod_{t_i\in O}\left(\frac{\tilde{\gamma}(\kappa_i|t_i,(t_{\mathrm{pa}(i)},\kappa_{\mathrm{pa}(i)}))}{\gamma(\kappa_i|t_i,(t_{\mathrm{pa}(i)},\kappa_{\mathrm{pa}(i)}))}\right),
\end{aligned}
$$

where $\tilde{\mu}(\cdot)$, etc. denotes $\mu(\cdot)$, etc. with the proposed values inserted.

For the missing data, we use three types of updates: changing an immigrant to an offspring, changing an offspring to an immigrant, and changing the parent of an offspring. The first two updates are handled together: we change an immigrant to an offspring with probability p_{IO} and an offspring to an immigrant with probability $1-p_{IO}$. If we choose to change an immigrant into an offspring, we draw the immigrant t_i from a uniform distribution on all the immigrants, and choose its new parent t_j from a uniform distribution on all points before t_i. If we choose to change an offspring into an immigrant, we draw the offspring t_i from a uniform distribution on all the offspring, and we denote its current parent by t_j. In either case we denote the new immigrant and offspring process by $\tilde{I}$ and $\tilde{O}$. From (4.3) and (4.4) we get the Hastings ratio for changing an immigrant into an offspring given by

$$
\begin{aligned}
H_{I\to O} &= \frac{p(\tilde{I}|y,\mu,\gamma)p(\tilde{O}|y,\alpha,\beta,\gamma)}{p(I|y,\mu,\gamma)p(O|y,\alpha,\beta,\gamma)} \times \frac{(1-p_{IO})\frac{1}{n_O+1}}{p_{IO}\frac{1}{n_I}\frac{1}{i-1}} \\
&= \frac{\alpha(\kappa_j)\beta(t_i-t_j,\kappa_j)\gamma(\kappa_i|t_i,(t_j,\kappa_j))(1-p_{IO})n_I(i-1)}{\mu(t_i)\gamma(\kappa_i|t_i)p_{IO}(n_O+1)},
\end{aligned}
$$

where n_I denotes the number of immigrants and n_O denotes the number of offsprings. Similarly, the Hastings ratio for changing an offspring into an immigrant is given by

$$
H_{O\to I} = \frac{\mu(t_i)\gamma(\kappa_i|t_i)p_{IO}n_O}{\alpha(\kappa_j)\beta(t_i-t_j,\kappa_j)\gamma(\kappa_i|t_i,(t_k,\kappa_j))(1-p_{IO})(n_I+1)(i-1)}.
$$

This combined update is referred to as a type $I\leftrightarrow O$ update.

When we update the parent of an offspring, we pick a random offspring t_i uniformly from all the offspring, and afterwards we pick a random point $t_{\tilde{j}}<t_i$ as its new parent, denoting the current parent by t_j. Letting $\tilde{O}$ denote the collection of offspring processes when $\mathrm{pa}(i)=\tilde{j}$ instead of j, we get the Hastings ratio

$$
H_O = \frac{p(\tilde{O}|y,\alpha,\beta,\gamma)}{p(O|y,\alpha,\beta,\gamma)}\frac{\frac{1}{n_O}\frac{1}{i-1}}{\frac{1}{n_O}\frac{1}{i-1}} = \frac{\alpha(\kappa_{\tilde{j}})\beta(t_i-t_{\tilde{j}},\kappa_{\tilde{j}})\gamma(\kappa_i|t_i,(t_{\tilde{j}},\kappa_{\tilde{j}}))}{\alpha(\kappa_j)\beta(t_i-t_j,\kappa_j)\gamma(\kappa_i|t_i,(t_j,\kappa_j))}
$$

This update is refered to as a type O update.

We use a Gibbs sample to combine all of these updates, where in each step we go through each parameter one at a time, and then we repeat the type $I \leftrightarrow O$ and type O a number of times.

5 Comparison

In this section we make simulations of the examples defined in Section 2.3 and compare how well the two methods work in each case. The reason for using simulated data rather than real data is that with simulated data we know what the algorithms are supposed to produce, and this gives a better foundation for comparing the algorithms. The simulations can be made using either one of the definitions, but it is easiest to simulate it as a combination of independent Poisson processes as given in the definition in Section 2.2; see e.g. Møller and Rasmussen (2005, 2006) for details on how to simulate a Hawkes process, or Ogata (1981) for a general, but slower, algorithm for simulating any point process specified by a conditional intensity.

5.1 Example 1

We start by generating a simulation of the model given by Example 1 in Section 2.3 and try to estimate the posterior distribution of the parameters using both methods for MCMC-based Bayesian inference in order to compare the two approaches. We generate a point pattern on the time interval $[0, 10)$ using the parameters $(\mu_1, \alpha_1, \beta_1) = (0.5, 0.9, 10)$. This point pattern is shown in Figure 1. Here we can clearly see that this process is indeed a model for clustered point patterns (but note that the clusters visible in the point pattern may actually contain multiple clusters from the definition of clusters in Section 2.2). Three points, t_{14}, t_{19}, and t_{49}, have been marked with arrows for later use.

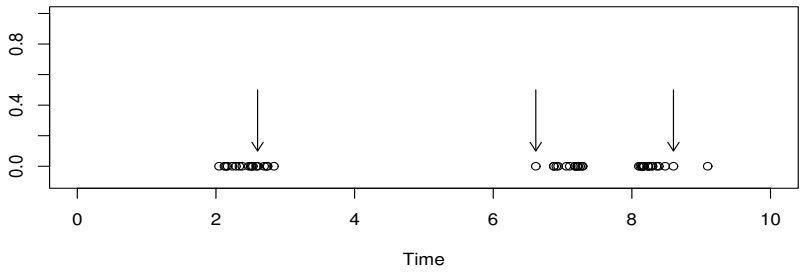


Figure 1: A simulated point pattern with 50 points. Points number 14, 19 and 49 have been marked with arrows.

In order to estimate the posterior distribution of the parameters, we need to equip the model with a prior distribution, and since there is no actual data, we have no information to put into this prior. It is tempting to use an improper uniform prior $p(\mu_1, \alpha_1, \beta_1) \propto \mathbf{1}[\mu_1, \alpha_1, \beta_1 > 0]$, but this does not yield a proper posterior. To see this, consider the likelihood function given by (3.1), fix μ_1 and α_1, and let β_1

tend to infinity. Inserting the expression for the model of this example, we get that

$$\lim_{\beta_1 \to \infty} p(x|\mu_1, \alpha_1, \beta_1) = \mu_1^n \exp(-\mu_1 T - \alpha_1 n) > 0.$$

Since the posterior does not even tend to 0 when β_1 tends to infinity, it cannot be a proper posterior, so instead we need some proper priors. Here we use independent exponential priors for the three parameters $(\mu_1, \alpha_1, \beta_1)$ with hyperparameters $(0.01, 0.01, 0.01)$. These priors are very flat and influence the results little.

To estimate the posterior distributions we use both the conditional intensity based method and the cluster based method. For both cases, the proposal parameters $(\sigma_{\mu_1}, \sigma_{\alpha_1}, \sigma_{\beta_1}) = (1, 0.5, 5)$ have been adjusted to give average acceptance probabilities roughly around 0.25 (Roberts *et al.*, 1997). For the cluster based method, we have adopted 10 updates of Y (both type $I \leftrightarrow O$ and type O updates) for each step in the algorithm. For both methods 20000 steps have been used, where the first 500 steps have been discarded as burnin. Trace plots (not shown here) for the three parameters $(\mu_1, \alpha_1, \beta_1)$, and also n_I for the cluster based method, reveal good mixing with no discernible patterns, and furthermore shows that 500 steps seem to be an appropriate burnin. The marginal posterior distributions approximated by both methods are shown in Figure 2. From these histograms we can see that the marginal posteriors are almost equal for the two methods, which suggests that both of the methods work equally well in estimating the parameters. The approximated posterior means for the conditional intensity based method is given by $(\hat{\mu}_1, \hat{\alpha}_1, \hat{\beta}_1) = (0.794, 0.883, 13.08)$ and for the cluster based method by $(\hat{\mu}_1, \hat{\alpha}_1, \hat{\beta}_1) = (0.792, 0.881, 12.24)$. Comparing with the original parameters used in the simulation, $(\mu_1, \alpha_1, \beta_1) = (0.5, 0.9, 10)$, these agree somewhat well; however, we should not put too much value into a comparison with the original parameters, since the particular simulation used may well be atypical (indeed the Hawkes process produces point patterns that are highly varying due to the high variation in the number of points in a cluster), and furthermore the prior also incorporates some information into the example.

So far in the cluster based method, we have considered the missing data Y as a set of nuisance parameters, which only have been estimated in order to estimate the model parameters. However, estimating Y may be a relevant problem all by itself. For example, if the Hawkes process was used to model the spread of a disease, estimating Y may tell likely paths that the disease has taken, thus providing valuable information on how to stop such a disease. In other words, we might as well enjoy the benefits of having used the time to estimate these. From the MCMC runs with the cluster based method, we can estimate the posterior distribution of the missing data by considering how often a particular point t_i has been an immigrant and how often it has been an offspring of point t_j for all $j = 1, \ldots, i-1$. We call the distribution of $\mathrm{pa}(i)$ on $\{0, \ldots, i-1\}$ the (marginal) posterior parent distribution of t_i, where the value 0 represents the point being an immigrant, while the values $j = 1, \ldots, i-1$ represents the point being an offspring of t_j. Note that occasionally it is more convenient to use the value i for t_i being an immigrant (this is done in Figures 4, 7 and 10).

Figure 3 shows histograms approximating the marginal posterior parent distributions of points t_{14}, t_{19}, and t_{49}. The three points have been marked with arrows

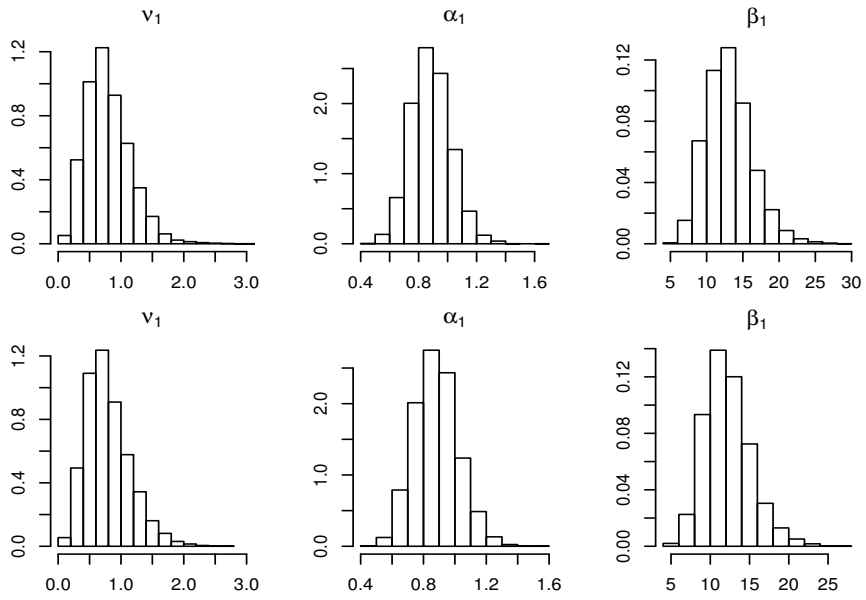


Figure 2: Histograms of marginal posterior distributions of $(\mu_1, \alpha_1, \beta_1)$. The upper row shows the result from the conditional intensity based method and the lower row the cluster based method.

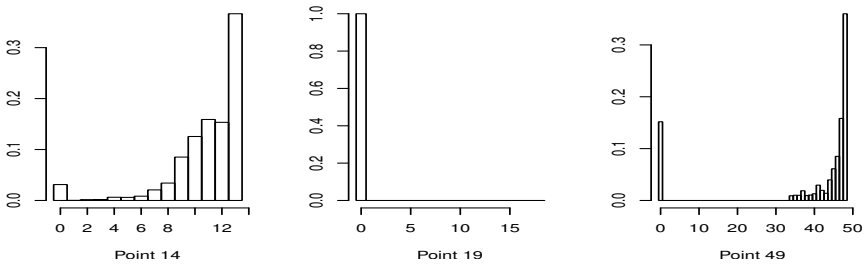


Figure 3: Histograms showing the marginal posterior parent distributions for t_{14}, t_{19} and t_{49}.

in Figure 1.

Comparing the histograms with the placement of the point, we see that these posterior distributions fit nicely with intuition:

- Point 14 is located in the middle of a cluster with many points close before it. This is reflected in the posterior distribution, since there are high probabilities that t_{14} is an offspring of any of the points just before it (in particular t_{13}), and only a small probability that t_{14} is an immigrant.

- Point 19 is located fairly isolated with no points located close before it. The posterior distribution shows that it is an immigrant with a probability close to 1.

- Point 49 is located at the end of a cluster just slightly separated from the points before it. The posterior distribution shows a probability around 0.15

that this point is an immigrant, but a larger probability that it is an offspring of one of the 15 closest points before it.

Note that the marginal posterior parent distributions are monotonously increasing (disregarding the probability of $\mathrm{pa}(i) = 0$). The reason for this is that the offspring intensity is monotonously decreasing in this example, so the most likely parent of point t_i is always t_{i-1}.

Next we try to visualise all the marginal posterior parent distributions together. Figure 4 shows these distributions in a grey scale image: the columns corresponds to the posterior distributions where bright colours show high probabilities. The probability of being an immigrant has now been placed at the diagonal rather than 0 since this is more convenient in this figure.

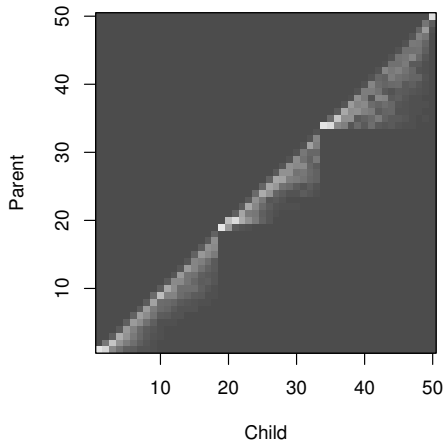


Figure 4: Marginal posterior parent distributions; see the text for a detailed description.

The figure shows the separation into the three large clusters also seen in the data in Figure 1, as well as points which are clearly immigrant (bright dots on the diagonal). In other words, the missing data seem to have been well estimated. The dotted look in the plot is an artifact; if we had run the algorithm of the cluster based method for more steps, or had used more than 10 updates for Y for each step, this would have been smoothed out.

Finally, we should note that the above conclusions is not specific to the one simulation considered here. Other simulations with the same parameters show similar conclusions. The conclusions are also the same for other parameter settings, unless the clustering becomes washed-out: this for example often happens for the parameter setting $(\mu_1, \alpha_1, \beta_1) = (0.5, 0.9, 1)$ where the points of each cluster are more spread out making it hard to discern any clustering, both visually and for the algorithms. Note that this applies both to both methods (since the conditional intensity based method relies on the clustering indirectly through the conditional intensity function), which both fluctuates wildly rather than converging. Note that in practice an informative prior (if available) might remedy this, but for the simulated examples in this paper with uninformative priors both algorithms fails.

5.2 Example 2

For the second example a simulation of the model with parameters $(\mu_1, \alpha_1, \gamma_1) = (0.5, 0.9, 1)$ is generated on $[0, 10] \times [0, \infty)$. The simulated data, which consists of 22 times of events and associated survival times, is shown in Figure 5. Note that three of the survival times extend beyond the observation time interval $[0, 10]$; we will assume here that these survival times are known, although in practice we might have to deal with some sort of right censoring here.

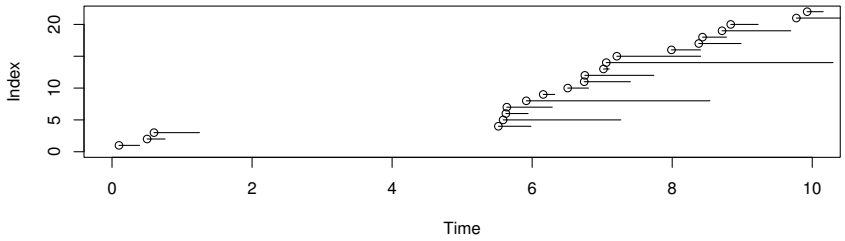


Figure 5: Times and marks (survival times) of a simulated dataset; x and y coordinates shows the time and index number of each point, while the line segment shows the survival times.

We now estimate the three parameters using both methods again equipping the model with independent exponential priors with inverse mean 0.01 for each parameter. The estimated marginal posterior distributions of the parameters are shown in Figure 6 using the cluster structure based method (the estimated marginal posteriors are very similar for the conditional intensity based method). The posterior means are estimated to be $(\hat{\mu}_1, \hat{\alpha}_1, \hat{\gamma}_1) = (1.04, 0.827, 1.34)$. The first parameter $\hat{\mu}_1$ is much higher than the one used in the simulation, while $\hat{\alpha}_1$ is a bit lower; note that these two parameters are negatively correlated since μ_1 controls how many points are immigrants, while α_1 controls how many points are offspring, which may partly explain this. The slightly high $\hat{\gamma}_1$ is easily explained by the fact that this parameter is the inverse mean survival times, and the survival times happens to be low in the simulation.

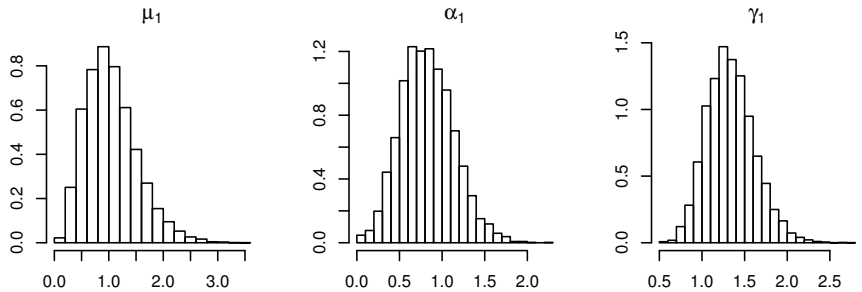


Figure 6: Histograms of the marginal posterior distributions of $(\mu_1, \alpha_1, \gamma_1)$.

Figure 7 shows the marginal posterior parent distribution in the same manner as in Figure 4. Comparing Figure 7 with Figure 4, we see some structural differences.

Figure 7 is characterised by long lines of bright squares; the reason for this is that some events have long survival times, thus being plausible parents for many later events, while other events are have died out earlier. Furthermore, comparing the colors vertically (disregarding the diagonal and the dark grey elements), we see that they are almost the same for all of the possible parents; this is explained by the fact that any living event produces offspring processes with the same intensity in the model. In short, the estimated branching structure corresponds closely to what we might expect.

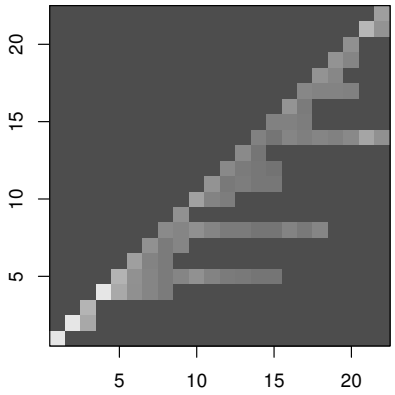


Figure 7: Marginal posterior parent distributions.

5.3 Example 3

For the last example, we generate a simulation of data from the model in Example 3 using the parameters $\mu_1 = 0.5$, $(\alpha_1, \alpha_2) = (0.4, 0.5)$, $(\beta_1, \beta_2) = (0.1, 2)$ and $(\gamma_1, \gamma_2) = (1, 1)$, and with $W = [0, 10] \times [0, 10]$; this is shown in Figure 8. Note that these parameters are not chosen to be realistic, but merely to generate a dataset to test the algorithms; the primary reason for choosing this example is to test how well the methods work in the case of dependent marks (the position of an aftershock depends on the position of the parent) and not to provide a statistical analysis of an earthquake dataset - this has been studied in many works for more realistic models than the one given here. To focus on this we will simplify the inference a bit further: We assume that $\beta_2 = 2$ is a known parameter not to be estimated (estimating β_1 and β_2 simultaneously in the present model gives certain approximate non-identifiability issues, which we will not go into here), and we ignore edge effects both in time and space (earthquakes may appear as aftershocks from earthquakes before time 0 or outside W).

Again using MCMC we estimate the marginal posterior distributions of the parameters using both methods. These are shown in Figure 9 for the cluster based method (they are not visibly different for the other method). The posterior means are given by $\hat{\mu}_1 = 1.027$, $(\hat{\alpha}_1, \hat{\alpha}_2) = (0.729, 0.0116)$, $\hat{\beta}_1 = 0.0760$ and $(\hat{\gamma}_1, \hat{\gamma}_2) = (0.934, 0.937)$. Comparing to the simulation parameters, we can see that $\hat{\beta}_1$, $\hat{\gamma}_1$, and $\hat{\gamma}_2$ are fairly close. The immigrant intensity μ_1 is about twice that of the

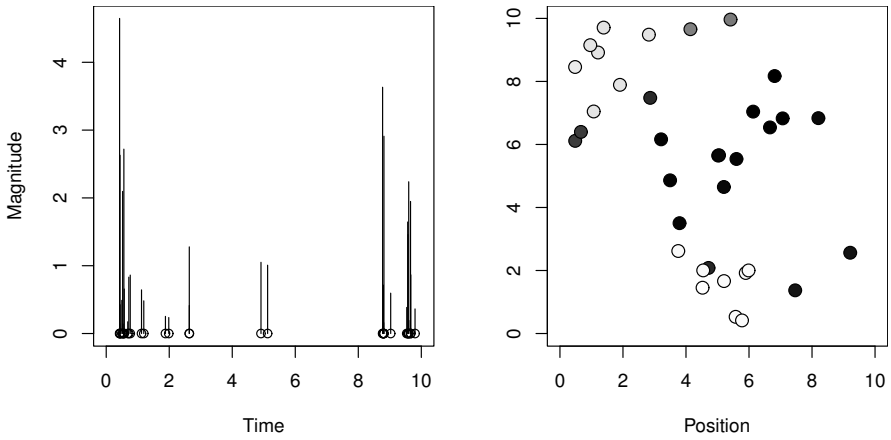


Figure 8: Left: Times and magnitudes of earthquakes. Right: Positions of earthquakes, where the dark colors represent early earthquakes and light colors represent later ones.

parameter used in the simulation, but as we will see later, this could be explained by the fact that there are more than the expected number of clusters in the data. The α parameters are somewhat off, though it is unclear why this happened for this particular simulation.

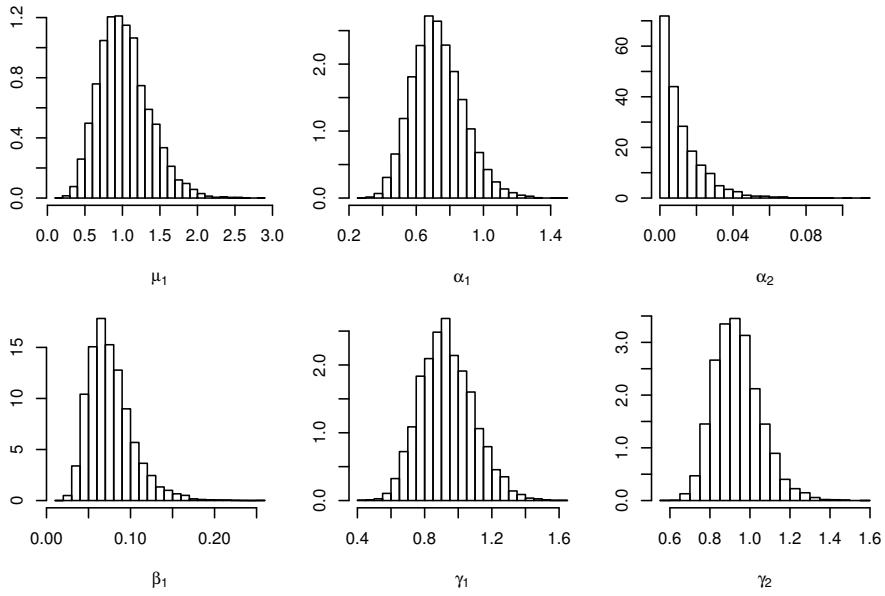


Figure 9: Histograms of marginal posterior distributions of $\mu_1, \alpha_1, \alpha_2, \beta_1, \gamma_1$, and γ_2.

The estimated posterior distribution of the missing data is shown in Figure 10. The left plot shows the marginal posterior parent distributions. Compared to Figure 3, the clusters are less well-defined in this plot. The explanation for this is that the spatial proximity is also relevant for estimating which events are the parents of which events. The right plot shows an estimate of the branching structure viewed spatially. Here for each point i, the mode of the marginal distribution of $\text{pa}(i)$ is shown by adding an arrow going from $(x_{\tilde{\text{pa}}(i)}, y_{\tilde{\text{pa}}(i)})$ to (x_i, y_i) where $\tilde{\text{pa}}(i)$ is the estimated mode of the marginal posterior parent distribution (if the mode is 0, there

is no arrow). Visually the clusters are clearly marked in his plot, and they seem to fit well with intuition. Also note that there are more than 5 clusters (the expected number of clusters using the parameters in this example is $\mu_1 T = 5$), which may explain the high estimate of μ_1 as mentioned earlier.

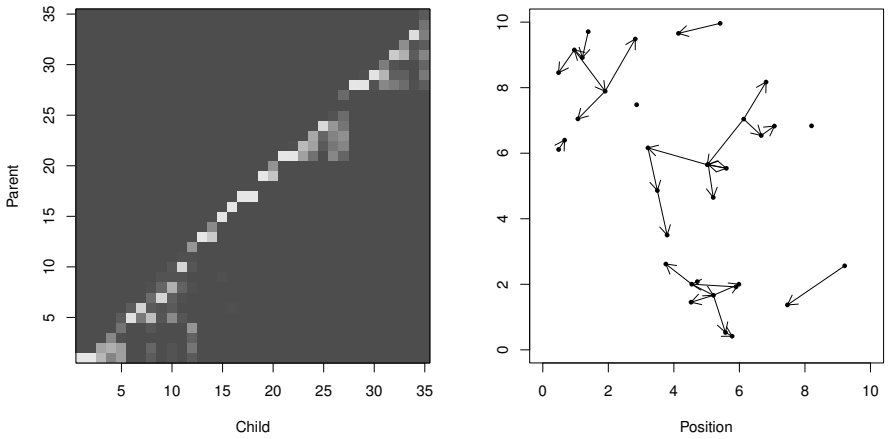


Figure 10: Left: Marginal posterior parent distributions. Right: modes of these distribution represented spatially by arrows.

5.4 Summary of comparison

The question is now, which of the methods performed the best. We will compare them with respect to various measures.

- *Estimates*: In all three examples the estimated posterior distributions of the parameters have been indistinguishable.

- *Running time*: The running time measured in seconds for the particular implementations used in this paper various much depending on the model. For Example 1 and 2, the MCMC runs are faster for the cluster based method roughly by a factor of 2. For Example 3, the difference is much more pronounced; here there is a factor of 100, and part of the reason is that the dependent marks in this model means that (2.2) has to be used in conditional intensity based method. Although (2.2) seems harmless, it appears many times each time a Hastings ratio needs to be evaluated. Of course these conclusions depend on the actual implementations of the algorithms, but this indicates that the cluster based method is faster.

- *Complexity*: From a theoretical point of view, the number of terms in the Hastings ratios for the parameter updates in the cluster based method grows linearly with the number of points, while the number of terms for the missing data updates are independent of the number of points (but here it seems reasonable to let the number of missing data updates grow linearly with the number of points to get good mixing). As a comparison, the number of terms in the Hastings ratios grows quadratically for the conditional intensity based

approach, so for sufficiently large datasets, we would expect that the cluster based method should be the fastest method for large datasets as was observed in the running times.

- *Missing data*: One distinct advantage of using the cluster based method is that this provides an estimate of the branching structure, which the other method does not. Obviously this is only an advantage if the branching structure has any interest in the particular data modelled by the Hawkes process. Alternatively viewing the estimates of the branching structure could be used to check the mixing of the MCMC algorithms or as a model check to see if the Hawkes process produce a reasonable fit to the branching structure.

6 Extensions

This section discusses some extensions and modifications the methods presented in this paper.

In Møller and Torrisi (2007) the spatial Hawkes process is defined using a similar definition as in Section 2.2. The term spatial here refers to the fact that the points are defined in a region of space rather than the time line, and no reference to the time of a point is given. A consequence of this is that there is no natural order of the points. The idea of considering the branching structure as missing data that can be estimated using MCMC together with the parameters can be immediately transfered to this setup; however, the actual implementation of the updates for the missing data will need some modifications since the data is not ordered anymore and thus any point can potentially be a child of any other point. Not being careful, we may thus encounter such absurd cases as several points being the parents of each other in a chain (e.g. point 1 is the parent of point 2, point 2 is the parent of point 3, and point 3 is the parent of point 1), something we never encounter when the points are ordered in time. Thus it may take some work to obtain efficient MCMC-based Bayesian inference procedures for the spatial Hawkes processes using ideas similar to the cluster based method.

Another issue that was completely ignored in the present paper is the problem of edge-effects. Here we assumed that the immigrant intensity was zero before time zero, but in practice we rarely observe data for the Hawkes process from its beginning. Thus there might be unobserved points before time zero causing offspring inside the dataset. In such cases the cluster based method will misclassify such points as immigrants or offsprings of points in the observed data, thus leading to biased estimates of the parameters; typically $\mu(t)$ will be estimated too high, and $\alpha(\kappa)$ will be too low, while it depends on the choice of model how $\beta(t, \kappa)$ is influenced. For large datasets such effects are negligible, but for small datasets this may influence the estimates, in particular if $\beta(t, \kappa)$ is heavy tailed. The conditional intensity based method also suffers from this, since it implicitely depends on the branching structure. It would be an interesting and practically relevant extension of the algorithms to include this.

In this paper Gibbs samplers have been applied in all cases of MCMC. While it is implementationally easy and computationally fast to use updates for one parameter

at a time, it may not be optimal. For example $\mu(t)$ and $\alpha(\kappa)$ is typically negatively correlated, and if the correlation is strong, the Gibbs sampler may have problems exploring the parameter space. Other MCMC approaches may well perform better.

Acknowledgements

The research was supported the Danish Natural Science Research Council (grant `09-072331`, *Point process modelling and statistical inference*) and by the Centre for Stochastic Geometry and Advanced Bioimaging, funded by a grant from the Villum Foundation.

References

Balderama, E., Schoenberg, F., Murray, E., and Rundel, P. (2010). Application of branching point process models to the study of invasive red banana plants in Costa Rica. *Submitted*.

Chornoboy, E. S., Schramm, L. P., and Karr, A. F. (2002). Maximum likelihood identification of neural point process systems. *Adv. Appl. Probab.*, **34**, 267–280.

Daley, D. J. and Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes, Volume I: Elementary Theory and Methods*. Springer, New York, 2nd edition.

Hawkes, A. G. (1971a). Point spectra of some mutually exciting point processes. *J. Roy. Statist. Soc. Ser. B*, **33**, 438–443.

Hawkes, A. G. (1971b). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, **58**(1), 83–90.

Hawkes, A. G. (1972). Spectra of some mutually exciting point processes with associated variables. In P. A. W. Lewis, editor, *Stochastic Point Processes*, pages 261–271. Wiley, New York.

Hawkes, A. G. and Adamopoulos, L. (1973). Cluster models for earthquakes – regional comparisons. *Bull. Int. Statist. Inst.*, **45**, 454–461.

Hawkes, A. G. and Oakes, D. (1974). A cluster representation of a self-exciting process. *J. Appl. Probab.*, **11**, 493–503.

Mohler, G., Short, M., Brantingham, P., Schoenberg, F., and Tita, G. (2011). Self-exciting point process modeling of crime. *J. Amer. Statist. Assoc., to appear*.

Møller, J. and Rasmussen, J. G. (2005). Perfect simulation of Hawkes processes. *Adv. in Appl. Probab.*, **37**(3), 629–646.

Møller, J. and Rasmussen, J. G. (2006). Approximate simulation of Hawkes processes. *Methodol. Comput. Appl. Probab.*, **8**, 53–65.

Møller, J. and Torrisi, G. L. (2007). The pair correlation function of spatial Hawkes processes. *Statistics & Probability Letters*, **77**, 995–1003.

Ogata, Y. (1981). On Lewis' simulation method for point processes. *IEEE Transactions on Information Theory*, **IT-27**(1), 23–31.

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *J. Amer. Statist. Assoc.*, **83**(401), 9–27.

Ogata, Y. (1998). Space-time point-process models for earthquake occurrences. *Ann. Inst. Statist. Math.*, **50**(2), 379–402.

Roberts, G. O., Gelman, A., and Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. App. Probab.*, **7**, 110–120.

Veen, A. and Schoenberg, F. P. (2008). Estimation of space-time branching process models in seismology using an EM-type algorithm. *J. Amer. Statist. Assoc.*, **103**(482), 614–624.