Ruprecht-Karls-Universität Heidelberg
Fakultät für Mathematik und Informatik

Bayesian Point Process Modelling of Earthquake Occurrences

Diplomarbeit
von
Natalia Andrea Hernandez Vargas

Betreuer: Dr. Thordis Linda Thorarinsdottir
Prof. Dr. Tilmann Gneiting

November 2012

Hiermit versichere ich, dass ich meine Arbeit selbstständig unter Anleitung verfasst habe, dass ich keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe, und dass ich alle Stellen, die dem Wortlaut oder dem Sinne nach anderen Werken entlehnt sind, durch die Angabe der Quellen als Entlehnungen kenntlich gemacht habe.

Abgabedatum: November 12, 2012

Abstract

Branching process models are commonly applied in seismology to model earthquake occurrences. For the parameter estimation, maximum likelihood estimation (MLE) is applied where, usually, numerical maximization algorithms must be implemented, as no closedform solutions are available. However, realistic models for earthquake occurrences are highly complex and the log-likelihood functions are often very flat which renders this procedure difficult. We propose an alternative parameter estimation method based on Bayesian inference. The method involves the implementation of MCMC (Markov chain Monte Carlo) algorithms to perform posterior approximations. We estimate the epidemic-type aftershock sequence (ETAS) model with both the conventional MLE method and our proposed method, to demonstrate that the new alternative can be very accurate. As a case study, we model earthquake occurrences in California and compare predictive performance of the temporal ETAS model under maximum likelihood and Bayesian inference.

Zusammenfassung

Punktprozesse werden häufig in der Seismologie verwendet. Sie werden mit der Maximum-Likelihood-Methode geschätzt. Normalerweise werden numerische Maximierungsalgorithmen implementiert, da keine analytischen Lösungen verfügbar sind. Allerdings sind diese Modelle komplex und seine log-likelihood-Funktionen sind flach. Wir schlagen eine alternative Schätzsmethode vor, die auf Bayesscher Inferenz basiert ist. Die Methode beinhaltet MCMC-Verfahren, die a posteriori Approximationen durchführen. Wir schätzen das ETAS Modell mit beiden Methoden. Unsere Studienregion ist der Staat Kalifornien in den USA. Ausserdem vergleichen wir die Vorhersagefähigkeit der beiden Methoden.

Contents

1 Introduction

"All things have second birth;
The earthquake is not satisfied at once."

William Wordsworth (1770-1850)
major English Romantic poet

Point process models have long been used to describe earthquake occurrences, see e.g. Vere-Jones (1970, 1975). The models that are currently most common are branching process models, which are based on the assumption that all earthquakes can trigger aftershocks. Two models of this kind are investigated in this work: the epidemic-type aftershock sequence (ETAS) model (Ogata, 1988, 1998) and the short-term earthquake probabilites (STEP) model (Reasenberg and Jones, 1989, 1990). Both of them produce forecasts based on prior seismicity only. The ETAS model assumes that aftershock sequences have an epidemic behavior, i.e. large earthquakes induce more aftershocks than small ones in a given interval of time. It also assumes that the larger the mainshock event is, the longer the time period of the aftershock sequence is (Harte (2010)). The STEP model is based on foreshock/aftershock statistics. It combines a background (time-independent) model with aftershock rates. This model is a generic forecast model for earthquake occurrences and it returns a description of the probability and number of events that are likely to occur after a mainshock of a given magnitude (Gerstenberger et al. (2005)).

Branching process models of this type are commonly fitted with maximum likelihood estimation (MLE). Usually, numerical maximization algorithms must be implemented, as no closed-form solutions are available. However, these models are complex and the log-likelihood functions are extremely flat. These effects were observed by Veen and Schoenberg (2008), who investigated an alternative estimation method for these models based on the expectation-maximization algorithm.

We propose an alternative parameter estimation method based on Bayesian inference. The method involves the implementation of MCMC (Markov chain Monte Carlo) algorithms to perform posterior approximations. We estimate the ETAS model with both the conventional MLE method and our proposed method, to demonstrate that the new alternative can be very efficient.

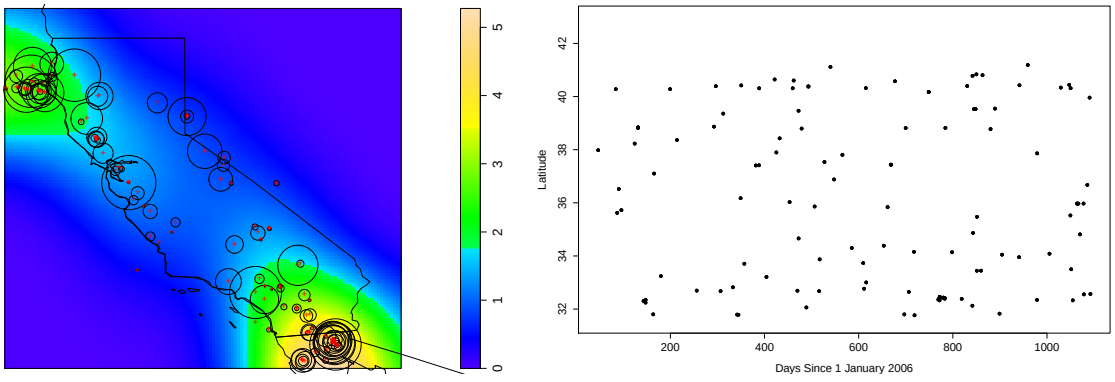
As a case study, we analyse data from the state of California in the United States, where earthquake occurrences are very common, mainly due to the San Andreas fault. The catastrophe caused by the 1906 earthquake in the San Francisco region marked the beginning of the study of California earthquakes and geology. See e.g. Stoffer (2006) for further details on the field geology.

The remainder of this thesis is organized as follows. In section 2, we give a description of the data set applied in our case study. In section 3, we state the definition of a marked Hawkes process. We describe and explain the current formulation of the models outlining the maximum likelihood estimation method and we propose a Bayesian version of the Epidemic-Type Aftershock Sequences (ETAS) model employing the fact that the ETAS model is a marked

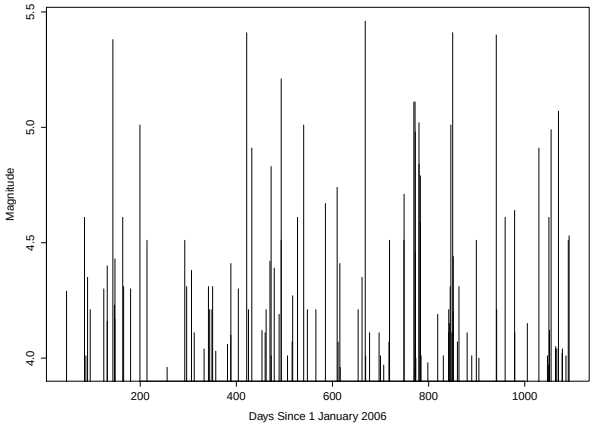
Hawkes process with dependent marks. In sections 4 and 5, we give a detailed description of the two estimation methods. We describe the implementation of a Metropolis-Hastings algorithm to estimate the parameters of the model in its two versions: the temporal and the spatio-temporal ETAS model. In section 6, we discuss the results obtained in sections 4 and 5 to compare the accuracy of the alternative estimation methods. In section 7, we describe the thinning method applied to simulate the point process with a desired conditional intensity function and present the procedure to produce earthquake forecasts under the temporal ETAS model. In section 8, we discuss forecast verification methods for such data, the L- and N-test and residual analysis. We also outline other evaluation methods proposed by Clements et al. (2011). We conclude with a discussion in section 9.

2 Data Description

Our data set is an earthquake catalog for the state of California, which contains the estimated earthquake hypocenter locations and the magnitudes. It was obtained from the Advanced National Seismic System (ANSS)[1]. The ranges of the observation window are $[-125.7, -113.15]$ degrees of longitude and $[31.55, 42.95]$ degrees of latitude. Our full data set consists of all events from January 1, 2006 to September 1, 2009 of magnitude greater than or equal to 3.95 on the Richter scale, a total of 142 events. We use the 121 events that occurred between January 1, 2006 and December 31, 2008 as a training set to estimate the model parameters while the models are evaluated on the 21 events that occurred between January 1, 2009 and September 1, 2009. The training set is shown in Figure 1.



(a) (b)



(c)

Figure 1: Seismicity in California during 2006-2008 with magnitude greater than or equal to 3.95 on the Richter scale. (a) The red crosses indicate the locations of the events and the circles represent their magnitudes. The color grid shows a kernel intensity estimate on the window. (b) Latitude of event versus occurrence time. (c) Magnitude of event versus occurrence time.

[1]It was found in the *Collaboratory for the Study of Earthquake Predictability (CSEP) Development website* (2012) **URL:** *http://northridge.usc.edu/trac/csep/wiki*

3 One-Day Models for Earthquake Occurrences

3.1 Marked Hawkes Process

A marked point process is a stochastic process model with a point process component, which contains the so-called marks. These marks are the information about the locations in time, space, or space-time of events that may themselves have a stochastic structure and stochastic dependency relations, see e.g. Daley and Vere-Jones (2003) for more details. The Hawkes process is a cluster process, which is commonly applied not only in seismology, but also e.g. in epidemiology and neurophysiology. A marked version of the temporal Hawkes process is described in Rasmussen (2011) as follows.

Let $X = \{(t_i, \kappa_i)\}$ be a marked point process on the time line, where

$t_i \in \mathbb{R}$ denotes an event of the point process,

$\kappa_i \in \mathbb{M}$ denotes the corresponding mark and $\mathbb{M}$ denotes a measurable space called mark space.

A marked point process can be defined by its conditional intensity function, where we condition on past events, and the mark distribution. The conditional intensity function is given by

$$\lambda^*(t) = \frac{\mathbb{E}(N(dt)|\mathcal{H}_t)}{dt}, \tag{1}$$

where N denotes the corresponding counting measure and dt denotes an infinitesimal interval around t. The notation of Daley and Vere-Jones (2003) is employed, where the star is used to indicate that the function is allowed to depend on past events and marks given by $\mathcal{H}_t := \{(t_i, \kappa_i)\}_{t_i<t}$. The mark distribution γ^* is described by its density function γ given the past $\mathcal{H}_t$ and the time of the point, i.e.

$$\gamma^*(\kappa|t) = \gamma(\kappa|t, \mathcal{H}_t), \tag{2}$$

where the star again indicates dependence on the past of the process.

The conditional intensity function of the Hawkes process is assumed to be of the form

$$\lambda^*(t) = \mu(t) + \sum_{t_i<t} \alpha(\kappa_i)\beta(t - t_i, \kappa_i), \tag{3}$$

where

$\mu(t)$ is a non-negative function on $\mathbb{R}$ called *immigrant intensity* with parameter vector $\mu = (\mu_1, \ldots, \mu_{n_\mu})$,

$\alpha(\kappa)$ is a non-negative function on $\mathbb{M}$ called *total offspring intensity* with parameter vector $\alpha = (\alpha_1, \ldots, \alpha_{n_\alpha})$,

$\beta(t, \kappa)$ is a density function on $[0, \infty)$ called *normalised offspring intensity* with parameter vector $\beta = (\beta_1, \ldots, \beta_{n_\beta})$, which is allowed to depend on the mark κ.

3.2 Spatio-temporal Epidemic-Type Aftershock Sequences (ETAS) Model

3.2.1 Current Formulation

The formulation of the spatio-temporal epidemic-type aftershock sequence (ETAS) model by Ogata (1998) is in the form of its conditional intensity function; that is, the process is controlled by an intensity conditional on the observation history $\mathcal{H}_t$. The analysis is based on the formulation given by

$$\lambda\left(t,x,y,m|\mathcal{H}_t\right)=\lambda\left(t,x,y|\mathcal{H}_t\right)\Gamma_1(m), \tag{4}$$

$$\lambda\left(t,x,y|\mathcal{H}_t\right)=\mu(x,y)+A\sum_{i:t_i<t}\alpha(m_i)g(t-t_i)f(x-x_i,y-y_i|m_i) \tag{5}$$

where

$\mu(x,y)$ is the background intensity independent of time which is assumed to be a constant over the area in our case;

A is a constant;

$\alpha(m)$ is the expected number of events triggered from an event of a magnitude m, given by

$$\alpha(m)=\exp\left[\alpha_2\left(m-m_c\right)\right], \tag{6}$$

where α_2 is a constant and m_c is the magnitude threshold of observed earthquakes;

$g(t)$ is the probability density function of the occurrence times of the triggered events, given by

$$g(t)=\frac{\beta_2-1}{\beta_1}\left(1+\frac{t}{\beta_1}\right)^{-\beta_2}, \tag{7}$$

which is the modified Omori law, where β_1 and β_2 are constants;

$f(x,y|m)$ is the location distribution of the triggered events, given by

$$f(x,y|m)=\frac{1}{2\pi\gamma_2 e^{\alpha_2(m-m_c)}}\exp\left[-\frac{x^2+y^2}{2\,\gamma_2\,e^{\alpha_2(m-m_c)}}\right], \tag{8}$$

which is a short-range Gaussian decay, where γ_2 is a constant;

$\Gamma_1(m)$ is the probability density of the magnitudes of all events, independent of the other components of the model. It is given by

$$\Gamma_1(m)=\gamma_1\exp\left[-\gamma_1(m-m_c)\right], \tag{9}$$

which is the Gutenberg-Richter law, where γ_1 is linked to Gutenberg-Richter's b value by $\gamma_1=b\log 10$ and m_c is again the magnitude threshold.

We notice that the definition of the location distribitution $f(x,y|m)$ in (8) seems to indicate that the function $\alpha(m)$ in (6) cancels out against the normalizing constant in the function $f(x,y|m)$. However, we present this formulation as it is been given by Zhuang et al. (2004).

Parameter Estimation The current method to estimate the parameters of the ETAS model is to maximize the log-likelihood function given by

$$\log L(\theta) = \sum_k \log \lambda_\theta \left(t_k, x_k, y_k | \mathcal{H}_{t_k}\right) - \int_0^T \iint_S \lambda_\theta \left(t, x, y | \mathcal{H}_t\right) \, dxdydt, \tag{10}$$

where $\theta = (\mu, A, \alpha, c, p, d, \beta)$ are the parameters to be estimated and k runs over all events in the region S and in the time interval $[0, T]$. For this we minimize the `neglogLik` function (negative Log Likelihood of a Point Process Model) provided in the `R` package `PtProcess` of Harte (2010) using the `nlm` function (Non-Linear Minimization) of the `R` package `stat`. This procedure is discussed in more detail in section 4.1.

3.2.2 Bayesian Version

The new proposed formulation uses the notation of Rasmussen (2011) and it is based on the formulation implemented in the `R` package `PtProcess` in order to estimate the parameters in a Bayesian setting. The new formulation of the conditional intensity function implemented in this analysis is given by

$$\lambda_\theta \left(t, x, y, m | \mathcal{H}_t\right) = \lambda^*_\theta \left(t | \mathcal{H}_t\right) \gamma^*_\theta (x, y, m | \mathcal{H}_t), \tag{11}$$

$$\lambda^*_\theta \left(t | \mathcal{H}_t\right) = \mu + \alpha_1 \sum_{i:t_i<t} \alpha(m_i)\beta(t - t_i), \tag{12}$$

$$\gamma^*_\theta \left(x, y, m | \mathcal{H}_t\right) = \frac{1}{\lambda^*_\theta(t)} \left[\frac{\mu}{|W|} + \alpha_1 \sum_{i:t_i<t} \alpha(m_i)\beta(t - t_i)\Gamma_2(x - x_i, y - y_i | m_i)\right] \Gamma_1(m), \tag{13}$$

where

μ is again the background intensity independent of time which is assumed to be a constant;

α_1 is a constant;

$\alpha(m)$ is the expected number of events triggered from an event of a magnitude m, given by

$$\alpha(m) = \exp\left[\alpha_2 \left(m - m_c\right)\right], \tag{14}$$

where α_2 is a constant and m_c is the magnitude threshold;

$\beta(t)$ is the probability density function of the occurrence times of the triggered events, given by

$$\beta(t) = \left(1 + \frac{t}{\beta_1}\right)^{-\beta_2}, \tag{15}$$

which is the modified Omori law, where β_1 and β_2 are constants;

$\Gamma_2(x, y|m)$ describes the location distribution of the triggered events, given by

$$\Gamma_2(x, y|m) = \exp\left[-\frac{x^2 + y^2}{2\gamma_2\,\alpha(m)}\right], \tag{16}$$

which is a short-range Gaussian decay, where γ_2 is a constant and $\alpha(m)$ is as defined in (14);

$\Gamma_1(m)$ is the probability density of magnitudes of all the events, independent of the other components of the model, with an exponential distribution given by

$$\Gamma_1(m) = \gamma_1 \exp\left[-\gamma_1(m - m_c)\right], \tag{17}$$

i.e.

$$m - m_c \sim \text{Exp}(\gamma_1), \tag{18}$$

which is the Gutenberg-Richter law, where γ_1 is linked to the Gutenberg-Richter's b value by $\gamma_1 = b \log 10$ and m_c is again the magnitude threshold;

W is the observation window.

The main difference between both formulations presented in this section is that the proposed Bayesian version is a point process with dependent marks that treats the spatial location as a mark, whereas the formulation proposed by Ogata (1998) sees the spatial location as a component of the events of the point process. However, both formulations lead to the same model, as the mark distribution defined in (13) is similar to the conditional intensity function of the Ogata formulation in (5). We mentioned in Section 3.2.1 a possible discrepancy between the normalizing constant of $f(x, y|m)$ in (8) and $\alpha(m)$ in (6). This is not the case in the formulation proposed by Rasmussen (2011), as the definition of the location distribution $\Gamma_2(x, y|m)$ in (16) does not have a normalizing constant. This second formulation matches the formulation used in the `R` package `PtProcess` implemented in this work.

In order to accurately analyse the model, we investigate two versions, the temporal and the spatio-temporal ETAS model. The temporal ETAS model uses only the occurrence time and the magnitude of each event, whereas the spatio-temporal model has also a location component as described above.

Temporal ETAS model The formulation of the temporal ETAS model is given by (11) without the spatial component:

$$\lambda_\theta\,(t, m|\mathcal{H}_t) = \lambda^*_\theta\,(t|\mathcal{H}_t)\,\gamma^*_\theta(m|\mathcal{H}_t), \tag{19}$$

$$\lambda^*_\theta\,(t|\mathcal{H}_t) = \mu + \alpha_1 \sum_{i:t_i<t} \alpha(m_i)\beta(t - t_i), \tag{20}$$

$$\gamma^*_\theta\,(m|\mathcal{H}_t) = \Gamma_1(m), \tag{21}$$

where all components are as described in (11).

Parameter Estimation The ETAS model is an example of a Hawkes process with dependent marks. We consider the process a point process in time with marks that consist of the magnitude $m \in (0, \infty)$ and the spatial location $(x, y) \in S_x \times S_y = W$ of the hypocenter of the earthquake occurrence, where W is the observation window.

Let $x = \{(t_1, x_1, y_1, m_1), ...(t_n, x_n, y_n, m_n)\}$ on $[0, T) \times S_x \times S_y \times (0, \infty)$ be a marked point pattern for some fixed time T, and assume that no points have ocurred before time 0. Then Daley and Vere-Jones (2003) propose that the likelihood function is given by

$$p\left(x|\mu, \alpha_1, \alpha_2, \beta_1, \beta_2, \gamma_1, \gamma_2\right) = \left[\prod_{k=1}^{n} \lambda_\theta\left(t_k, x_k, y_k, m_k|\mathcal{H}_{t_k}\right)\right] \exp\left(-\Lambda^*(T)\right), \tag{22}$$

where $\lambda_\theta\left(t_k, x_k, y_k, m_k|\mathcal{H}_{t_k}\right)$ is given by (11), and

$$\begin{aligned} \Lambda^*(t) &= \int_0^t \lambda_\theta^*\left(s|\mathcal{H}_s\right)\, ds \\ &= M(t) + \alpha_1 \sum_{i:t_i<t} \alpha(m_i) B(t - t_i), \end{aligned} \tag{23}$$

where

$M(t)$ is the integral of μ given by

$$M(t) = \int_0^t \mu\, ds, \tag{24}$$

$B(t)$ is the integral of $\beta(t)$ given by

$$B(t) = \int_0^t \beta(s)\, ds, \tag{25}$$

$\alpha(m)$ is given by (14).

Thus, the log-likelihood function is given by

$$\log p\left(x|\mu, \alpha_1, \alpha_2, \beta_1, \beta_2, \gamma_1, \gamma_2\right) = \sum_{k=1}^{n} \log \lambda_\theta\left(t_k, x_k, y_k, m_k|\mathcal{H}_{t_k}\right) - \Lambda^*(T). \tag{26}$$

Denoting the parameters by $\theta = (\mu, \alpha_1, \alpha_2, \beta_1, \beta_2, \gamma_1, \gamma_2)$ and the prior by $p(\theta)$ the posterior is given by

$$p\left(\theta|x\right) \propto p(\theta) p\left(x|\theta\right), \tag{27}$$

where $p\left(x|\theta\right)$ is the likelihood function given by (22).

The prior of each parameter ω is a Gamma distribution given by

$$\omega \sim \Gamma\left(a_\omega, b_\omega\right), \tag{28}$$

where a_ω and b_ω are prior parameters for ω.

Metropolis-Hastings algorithm The posterior (27) has a complicated form that does not allow us to find the mean or the maximum of the posterior for the parameters analytically. Instead, we use a Markov chain Monte Carlo (MCMC) algorithm (Rasmussen (2011)). In this case a Metropolis-Hastings algorithm is implemented in order to update one parameter at a time. Truncated normal distributions are used as proposal distributions.

For updating each parameter ω_k for $k = 1, \ldots, n_\omega$, $\tilde{\omega}_k$ is drawn from a truncated normal distribution with the current parameter value ω_k as mean and some fixed standard deviation σ_{ω_k}. The Hastings ratios are calculated from (27) for the parameter updates, which are given by

$$H_\omega = \frac{p(\tilde{\theta}|x)}{p(\theta|x)} \frac{J(\theta|\tilde{\theta})}{J(\tilde{\theta}|\theta)} \propto \frac{p(\tilde{\theta})}{p(\theta)} \frac{p(x|\tilde{\theta})}{p(x|\theta)} \frac{J(\theta|\tilde{\theta})}{J(\tilde{\theta}|\theta)}, \tag{29}$$

where $\tilde{\theta}$ denotes the proposed parameter value and $J(\theta|\cdot)$ denotes the proposal distribution when θ is the current state of the chain.

Computing this ratio can be numerically unstable, so the logarithms of (29) are computed for each parameter. For the ETAS model, these are given by

$$\begin{aligned} \log H_\mu \;=\; & \log p\,(\tilde{\mu}) - \log p\,(\mu) + \log J(\mu|\tilde{\mu}) - \log J(\tilde{\mu}|\mu) \\ & + \sum_{i=1}^{n} \log \left(\tilde{\mu} + \alpha_1 \sum_{j<i} \alpha(m_j)\beta(t_i - t_j) \right) \\ & - \sum_{i=1}^{n} \log \left(\mu + \alpha_1 \sum_{j<i} \alpha(m_j)\beta(t_i - t_j) \right) \\ & + (\mu - \tilde{\mu})\, T, \end{aligned} \tag{30}$$

where $\tilde{\mu}$ denotes the proposed value;

$$\begin{aligned} \log H_{\alpha_1} \;=\; & \log p\,(\tilde{\alpha_1}) - \log p\,(\alpha_1) + \log J(\alpha_1|\tilde{\alpha_1}) - \log J(\tilde{\alpha_1}|\alpha_1) \\ & + \sum_{i=1}^{n} \log \left(\mu + \tilde{\alpha_1} \sum_{j<i} \alpha(m_j)\beta(t_i - t_j) \right) \\ & - \sum_{i=1}^{n} \log \left(\mu + \alpha_1 \sum_{j<i} \alpha(m_j)\beta(t_i - t_j) \right) \\ & + (\alpha_1 - \tilde{\alpha_1}) \sum_{i:t_i<T} \alpha(m_i) B(T - t_i), \end{aligned} \tag{31}$$

where $\tilde{\alpha_1}$ denotes the proposed value;

$$\begin{aligned}\log H_{\alpha_2} &= \log p\left(\tilde{\alpha_2}\right) - \log p\left(\alpha_2\right) + \log J(\alpha_2|\tilde{\alpha_2}) - \log J(\tilde{\alpha_2}|\alpha_2) \\ &+ \sum_{i=1}^{n} \log\left(\mu + \alpha_1 \sum_{j<i} \tilde{\alpha}(m_j)\beta(t_i - t_j)\right) \\ &- \sum_{i=1}^{n} \log\left(\mu + \alpha_1 \sum_{j<i} \alpha(m_j)\beta(t_i - t_j)\right) \\ &+ \alpha_1 \sum_{i:t_i<T} \left[\alpha(m_i) - \tilde{\alpha}(m_i)\right] B(T - t_i), \end{aligned} \tag{32}$$

where $\tilde{\alpha_2}$ and $\tilde{\alpha}(\cdot)$ denote the proposed value and $\alpha(\cdot)$ with $\tilde{\alpha_2}$ inserted, respectively;

$$\begin{aligned}\log H_{\beta_1} &= \log p\left(\tilde{\beta_1}\right) - \log p\left(\beta_1\right) + \log J(\beta_1|\tilde{\beta_1}) - \log J(\tilde{\beta_1}|\beta_1) \\ &+ \sum_{i=1}^{n} \log\left(\mu + \alpha_1 \sum_{j<i} \alpha(m_j)\tilde{\beta}(t_i - t_j)\right) \\ &- \sum_{i=1}^{n} \log\left(\mu + \alpha_1 \sum_{j<i} \alpha(m_j)\beta(t_i - t_j)\right) \\ &+ \alpha_1 \sum_{i:t_i<T} \alpha(m_i)\left[B(T - t_i) - \tilde{B}(T - t_i)\right], \end{aligned} \tag{33}$$

where $\tilde{\beta_1}$, $\tilde{\beta}(\cdot)$ and $\tilde{B}(\cdot)$ denote the proposed value and $\beta(\cdot)$ and $B(\cdot)$ with $\tilde{\beta_1}$ inserted, respectively;

$$\begin{aligned}\log H_{\beta_2} &= \log p\left(\tilde{\beta_2}\right) - \log p\left(\beta_2\right) + \log J(\beta_2|\tilde{\beta_2}) - \log J(\tilde{\beta_2}|\beta_2) \\ &+ \sum_{i=1}^{n} \log\left(\mu + \alpha_1 \sum_{j<i} \alpha(m_j)\tilde{\beta}(t_i - t_j)\right) \\ &- \sum_{i=1}^{n} \log\left(\mu + \alpha_1 \sum_{j<i} \alpha(m_j)\beta(t_i - t_j)\right) \\ &+ \alpha_1 \sum_{i:t_i<T} \alpha(m_i)\left[B(T - t_i) - \tilde{B}(T - t_i)\right], \end{aligned} \tag{34}$$

where $\tilde{\beta_2}$, $\tilde{\beta}(\cdot)$ and $\tilde{B}(\cdot)$ denote the proposed value and $\beta(\cdot)$ and $B(\cdot)$ with $\tilde{\beta_2}$ inserted, respec-

tively;

$$\begin{aligned}\log H_{\gamma_2} &= \log p\left(\tilde{\gamma_2}\right) - \log p\left(\gamma_2\right) + \log J(\gamma_2|\tilde{\gamma_2}) - \log J(\tilde{\gamma_2}|\gamma_2) \\ &\quad + \sum_{i=1}^{n} \log\left(\frac{\mu}{|W|} + \alpha_1 \sum_{j<i} \alpha(m_j)\beta(t_i - t_j)\tilde{\Gamma}_2(x_i - x_j, y_i - y_j|m_j)\right) \\ &\quad - \sum_{i=1}^{n} \log\left(\frac{\mu}{|W|} + \alpha_1 \sum_{j<i} \alpha(m_j)\beta(t_i - t_j)\Gamma_2(x_i - x_j, y_i - y_j|m_j)\right), \end{aligned} \tag{35}$$

where $\tilde{\gamma_2}$ and $\tilde{\Gamma}_2(\cdot)$ denote the proposed value and $\Gamma_2(\cdot)$ with $\tilde{\gamma_2}$ inserted, respectively.

The parameter γ_1 has a known posterior, which is a Gamma distribution with parameters given by

$$\gamma_1 \sim \Gamma\left(n + a_{\gamma_1}, b_{\gamma_1} + \sum_{i=1}^{n}(m_i - m_c)\right), \tag{36}$$

where a_{γ_1} and b_{γ_1} are prior parameters for γ_1.

3.3 Short-Term Earthquake Probabilities (STEP) Model

The Short-Term Earthquake Probabilities (STEP) model by Reasenberg and Jones (1989, 1990) is an aftershock process, which is defined as a nonhomogeneous Poisson process in time with intensity $N(t)$ given by

$$N(t) = K\,(t+c)^{-p}\,, \tag{37}$$

which is the modified Omori law, where K, c and p are constants.

The magnitude distribution is given by

$$N(M) = A\,10^{-bM}\,, \tag{38}$$

which is the Gutenberg-Richter relation, where M is the aftershock magnitude, and A and b are constants.

Following these assumptions the rate of aftershocks λ with magnitude M or larger at time t after a mainshock of magnitude M_m is given by

$$\lambda(t, M) = 10^{a+b\,(M_m - M)}(t+c)^{-p}\,, \tag{39}$$

where a, b, p and c are constants. The probability P of one or more earthquakes occurring in a time interval $[S, T]$ is

$$P = 1 - \exp\left[\int_S^T \lambda(t, M)\,dt\right]\,. \tag{40}$$

Parameter Estimation The model parameters $\theta = (a, b, p, c)$ are estimated separately for each earthquake sequence with maximum likelihood method using earthquake data.

The rate of aftershocks of the STEP model given by (39) is similar to the terms of the sum in the conditional intensity of the ETAS model given by (12). Therefore, we focus only in the ETAS model in the following.

4 Estimation of the Temporal ETAS Model

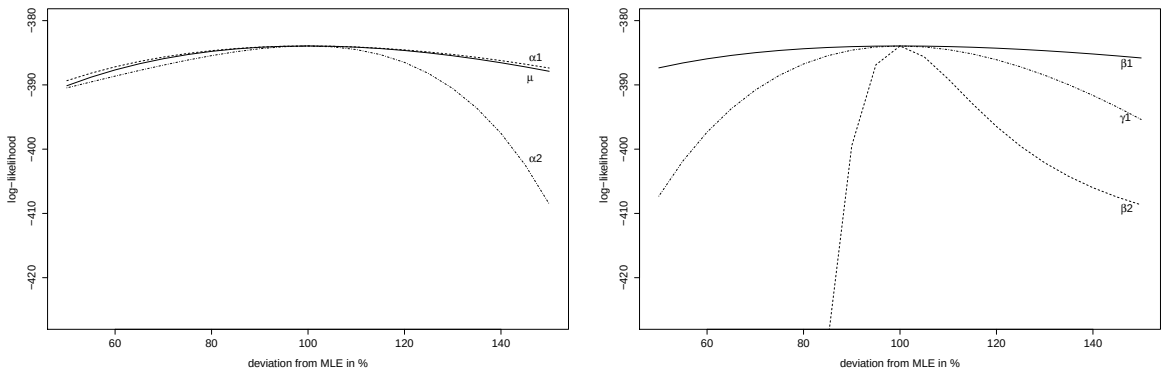
4.1 Maximum Likelihood Estimation of the Temporal ETAS Model

In order to compute the maximum likelihood estimator $\hat{\theta}$, where $\theta = (\mu, \alpha_1, \alpha_2, \beta_1, \beta_2, \gamma_1)$, the `mpp` function (Marked Point Process Object) of the `R` package `PtProcess` is used to create the marked point process object. The `neglogLik` function (Negative Log-Likelihood) of the same `R` package is then used as an input function for the `nlm` function (Non-Linear Minimization) of the `R` package `stat` in order to maximize the log-likelihood function. The estimated $\hat{\theta}$ under MLE for the California data set is listed in Table 1 (ahead on page 16).

As described in section 3 the MLE maximizes the log-likelihood function (10) of the model. Numerical maximization algorithms must be employed, e.g. the method proposed in the `R` package `PtProcess` described above, as no closed form solutions are available. However, the log-likelihood function of the ETAS model tends to be flat in the vicinity of its maximum, which leads to convergence problems of the optimization algorithms, and the results can also be influenced by the choice of initial values (Veen and Schoenberg (2008)).

Figure 2 shows the log-likelihood function where one component of θ is varied at a time by up to 50% around the MLE estimate $\hat{\theta}$. The function is quite flat around $\hat{\theta}$, especially when μ, α_1 and β_1 are varied, which means that these parameters are particularly difficult to estimate. Nevertheless, the parameters α_2, β_2 and γ_1 show clear maximums with peaked log-likelihood functions and can therefore be estimated more precisely. Similar effects were observed by Veen and Schoenberg (2008), who also investigated estimation procedures for this model.

Veen and Schoenberg (2008) further observed that if the log-likelihood function is extremely flat, the choice of initial values can influence the results. Table 2 (ahead on page 20) lists the results of performing the MLE method with three different initial values. These values are chosen randomly in the vicinity of the MLE $\hat{\theta}$ listed in Table 1 (ahead on page 16).



(a) (b)

Figure 2: Flatness of the log-likelihood function for the temporal ETAS model when one parameter is varied at a time. Two plots are shown to improve the legibility.

4.2 Temporal ETAS Estimation Using Bayesian Inference

The Metropolis-Hastings ratios defined in (30)-(34) and (36) are used in order to estimate the parameters. The training data described in Section 2 is used as the observation history to calculate the posterior approximations. The algorithm generates $15,000$ values $(\theta^{(1)}, \ldots, \theta^{(15000)})$, where $\theta = (\mu, \alpha_1, \alpha_2, \beta_1, \beta_2, \gamma_1)$. The first $5,000$ iterations are considered burn-in and are discarded. The fitted model has a conditional intensity function $\lambda_{\hat{\theta}}(t, m|\mathcal{H}_t)$, where $\hat{\theta}$ are the means of the chains obtained with the Metropolis-Hastings algorithm. The resulting chains are shown in figures 4, 5 and 6 in section 6.1, where we analyse them further. The posterior means $\hat{\theta}$ are listed in Table 1 (ahead on page 16).

In order to investigate if the performance of this method is also influenced by the choice of its initial values, we have run the algorithm multiple times using the same random initial values as we applied to the MLE method in section 4.1. The results are shown in Table 2 (ahead on page 20).

5 Estimation of the Spatio-temporal ETAS Model

5.1 Maximum Likelihood Estimation of the Spatio-temporal ETAS Model

The maximum likelihood estimator $\hat{\theta}$ is computed as explained in section 4.1 but using the intensity function (11), where $\theta = (\mu, \alpha_1, \alpha_2, \beta_1, \beta_2, \gamma_1, \gamma_2)$. The `R` package `PtProcess` provides a trial version for the estimation of the spatio-temporal ETAS model, which has a similar formulation as described in Section 3.2.1. The results obtained are listed in Table 3 (ahead on page 22).

For the spatio-temporal version of the ETAS model, similar numerical maximization algorithms are employed to maximize the log-likelihood function as for the temporal version even though it has the same suboptimal features as discussed above. In Figure 3 the log-likelihood is shown, where the parameter γ_2 is varied around the estimated parameter $\hat{\theta}$. The function is again flat in the vicinity of $\hat{\theta}$, so that the parameter value γ_2 is difficult to estimate.

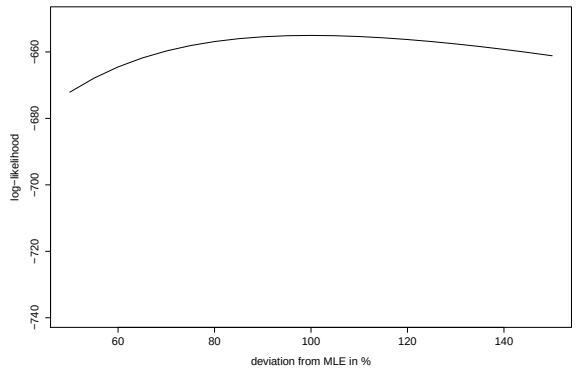


Figure 3: Flatness of the log-likelihood function for the spatio-temporal ETAS model when the parameter γ_2 is varied around its MLE.

5.2 Spatio-temporal ETAS Estimation Using Bayesian Inference

The Metropolis-Hastings ratios defined in (30)-(34) and (36) are implemented in order to estimate the parameters for the spatio-temporal version as explained in section 4.2, as the posterior distributions of the parameter chains obtained in the temporal case are independent of the additional spatial parameter γ_2. The ratio defined in (35) is employed to estimate separately the parameter chain of γ_2 using the values obtained in the temporal case described in Section 4.2. Therefore, the parameter value $\hat{\theta}_{-\gamma_2}$ is the one obtained in the temporal case, where $\theta_{-\gamma_2} = (\mu, \alpha_1, \alpha_2, \beta_1, \beta_2, \gamma_1)$. In Table 3 (ahead on page 22) the estimated $\hat{\theta}$ are listed, where $\theta = \left(\frac{\mu}{|W|}, \alpha_1, \alpha_2, \beta_1, \beta_2, \gamma_1, \gamma_2\right)$. The new μ is so defined, as the positions of the background events are assumed to have an uniform distribution.

In Section 6.2 we show the results obtained with both methods. Knowing the disadvantages of the maximum likelihood method presented in the temporal case, different initial values are used to see if there is again an influence of the choice of initial values. However, the results obtained with the maximum likelihood method are not as expected. We can not obtain satisfying results no matter which initial values we choose. Table 3 (ahead on page 22) lists three results using different initial values. This is probably because of the functions for spatio-temporal models implemented from the `R` package `PtProcess`, as they are trial versions. This does not allow us to compare the performance of the proposed method with the previous one.

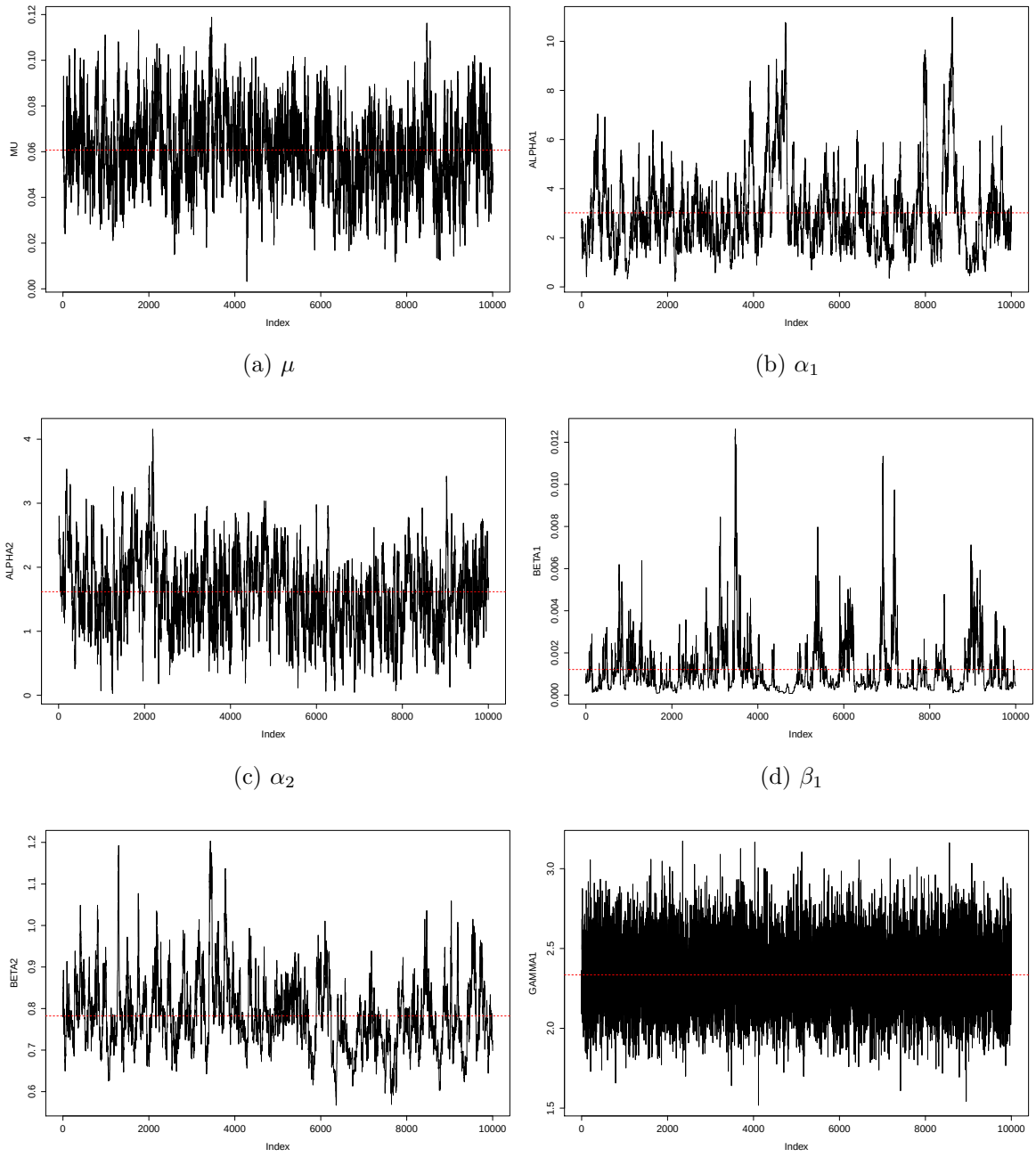
6 Estimation Results

6.1 Estimation under the Temporal ETAS Model

The estimation results for the two methods described in section 4 are listed in Table 1. Figure 4 and Figure 5 show the traceplots of the parameter chains obtained with the Bayesian method and their histograms, respectively. The estimated parameter values are similar to the ones obtained with the maximum likelihood method. The log-likelihood values are also very similar, i.e. our estimation method is indeed maximizing the log-likelihood function. The main difference is that the estimated value of $\hat{\beta}_1$ is here significantly greater than the MLE. The dependence of β_1 on the other parameter of the modified Omori law, β_2, can influence its estimation. We can see in the traceplots of both parameters that they have a positive correlation giving unexpectedly high values for β_1 that affect the posterior mean value of the $\hat{\beta}_1$ chain. Rasmussen (2011) sees this difficulty, too.

Table 1: Estimated parameter values obtained from fitting the temporal ETAS model to the California data set. The values shown for the Bayesian inference are the posterior means.

Method	μ	α_1	α_2	β_1	β_2	γ_1	$\log L$
MLE	**0.058**	**3.07**	**1.94**	**0.0005**	**0.77**	**2.34**	**−383.94**
Bayes	**0.061**	**3.02**	**1.62**	**0.0012**	**0.78**	**2.33**	**−385.59**

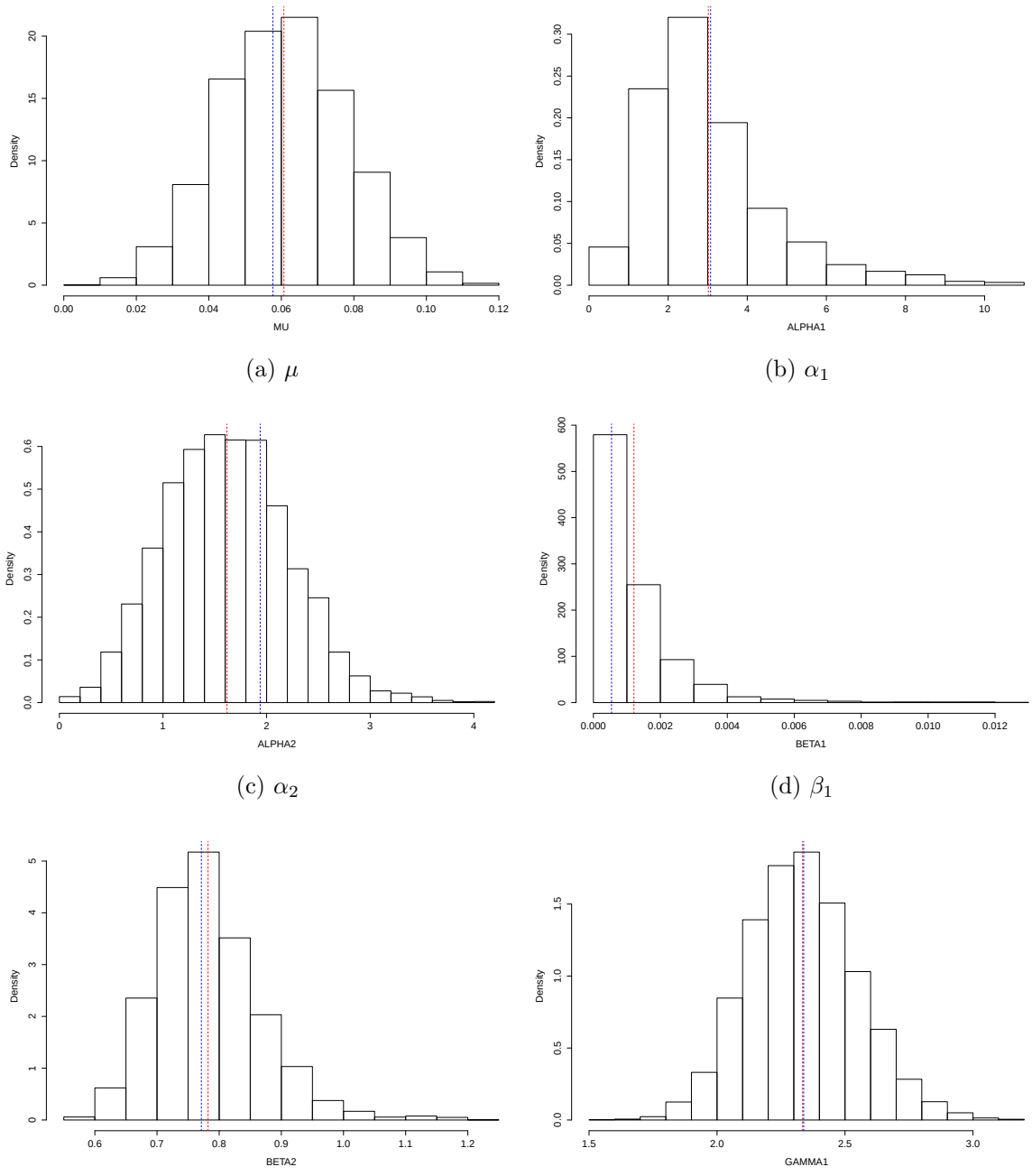


(a) μ (b) α_1

(c) α_2 (d) β_1

(e) β_2 (f) γ_1

Figure 4: Traceplots for each parameter chain for the temporal ETAS model. The red dotted line represents the posterior mean value.



(a) μ (b) α_1

(c) α_2 (d) β_1

(e) β_2 (f) γ_1

Figure 5: Histograms of each parameter chain for the temporal ETAS model. The red dotted line represents the posterior mean value and the blue dotted line represents the MLE listed in Table 1.

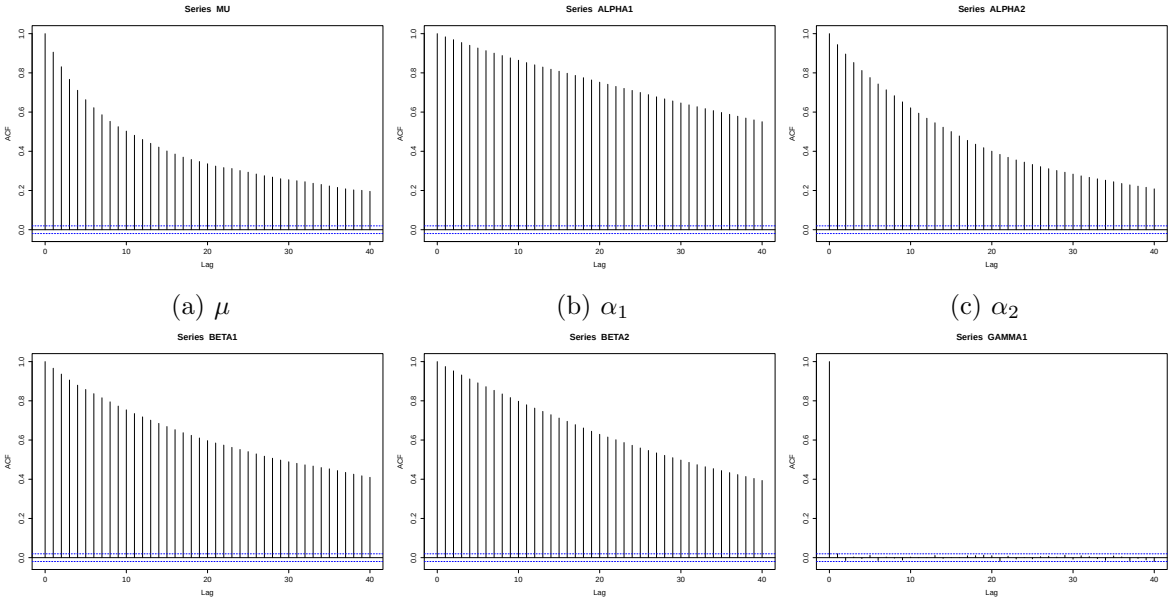
6.1.1 Convergence of Parameter Chains

The sample autocorrelation function, described in Hoff (2009), is computed for each parameter chain in order to see how good the approximation is. For a generic sequence of numbers $\Theta = \{\theta_1, \ldots, \theta_S\}$ the lag-t autocorrelation function estimates the correlation between the elements of the sequence that are t steps apart. The autocorrelation function is given by

$$\text{acf}_t(\Theta) = \frac{\frac{1}{S-t}\sum_{s=1}^{S-t}(\theta_s - \overline{\theta})(\theta_{s+t} - \overline{\theta})}{\frac{1}{S-1}\sum_{s=1}^{S}(\theta_s - \overline{\theta})^2}, \tag{41}$$

where S is the number of iterations. This function can be computed with the `acf` function of the `R` package `stats`.

Figure 6 shows the autocorrelation function of the parameter chains. The autocorrelation function of the parameter chain of γ_1 shown in Figure 6f has an optimal form. This result is expected for the parameter γ_1, because its posterior is known so that we can obtain a direct sample. In the other cases, the autocorrelation function shows that the chains have a high degree of correlation, specially the chain of α_1. Markov chains with such a high autocorrelation move around the parameter space slowly, taking a long time to achieve the correct balance among the different regions of the parameter space (Hoff (2009)). This was also expected because of the flatness of the log-likelihood function discussed in section 4.1. Figure 2 shows clear maximums of the parameters μ and α_2 with peaked log-likelihood functions, whereas the log-likelihood function was quite flat for the parameters α_1, β_1 and β_2.



(a) μ (b) α_1 (c) α_2

(d) β_1 (e) β_2 (f) γ_1

Figure 6: Autocorrelation functions of each parameter chain for the temporal ETAS model.

6.1.2 Influence of the Choice of Initial Values

Table 2 lists the results of the maximum likelihood method and the Bayesian method using three different initial values. The Bayesian method seems to be significantly more robust against changes in the initial values than the maximum likelihood estimation. In all cases, the Bayesian inference method converges to an estimated $\hat{\theta}$, which is always very close to the estimated maximum value listed in Table 1, whereas the estimation values obtained with the maximum likelihood method clearly show an influence of the choice of the initial values.

Table 2: Estimated parameter values under the temporal ETAS model for the California data set with three different initial values, where the parameter values shown as the fitted by Bayesian inference are the mean of the posterior approximations. In the second trial the MLE fails to converge.

Trial	Method	μ	α_1	α_2	β_1	β_2	γ_1	$\log L$
I	Initial	0.857	0.74	0.05	0.1629	0.79	0.48	–
	MLE	**0.106**	**2.82**	**1.78**	$\mathbf{5.4 \times 10^5}$	$\mathbf{1.1 \times 10^8}$	**2.34**	**−391.94**
	Bayes	**0.061**	**3.04**	**1.64**	**0.0011**	**0.78**	**2.34**	**−385.42**
II	Initial	0.181	2.22	2.07	0.0254	2.10	1.33	–
	MLE	–	–	–	–	–	–	–
	Bayes	**0.062**	**2.79**	**1.56**	**0.0014**	**0.79**	**2.34**	**−384.83**
III	Initial	2.412	0.56	0.56	0.0598	1.45	1.68	–
	MLE	**0.058**	**3.07**	**1.94**	**0.0005**	**0.77**	**2.34**	**−383.94**
	Bayes	**0.061**	**2.89**	**1.68**	**0.0012**	**0.79**	**2.34**	**−384.91**

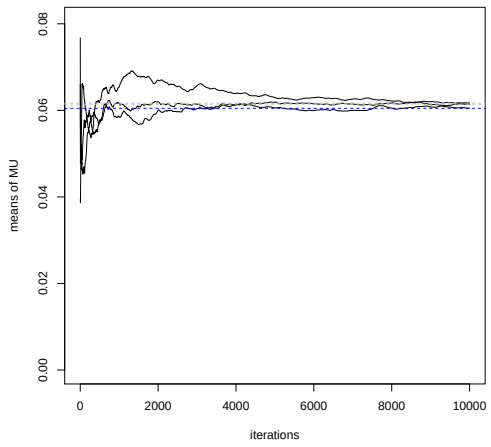
In the first case, the $\log L$ value obtained with the maximum likelihood method is close to the expected maximum value, but the estimated values of $\hat{\beta}_1$ and $\hat{\beta}_2$ are illogical. The optimization algorithm probably ignored the effect of the modified Omori law described in 15, i.e.

$$\beta(t - t_i) = \left(1 + \frac{t - t_i}{\beta_1}\right)^{-\beta_2} \xrightarrow{\beta_1 \to \infty} 1^{-\beta_2} = 1. \tag{42}$$

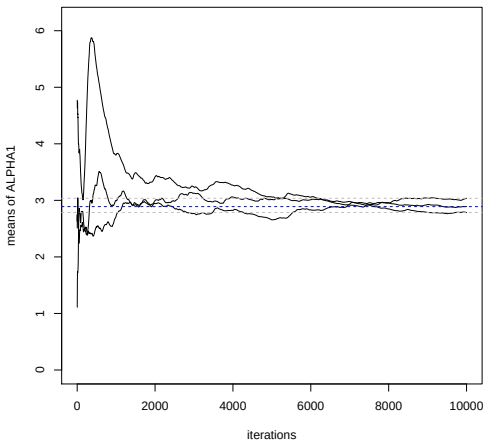
In the second case, the optimization algorithm for the maximum likelihood method fails to converge, whereas the Bayesian method gives a good result.

The third case leads us to the previous result, where the maximum likelihood method estimates the presumed maximum value as proposed with the Bayesian method yielding very similar values.

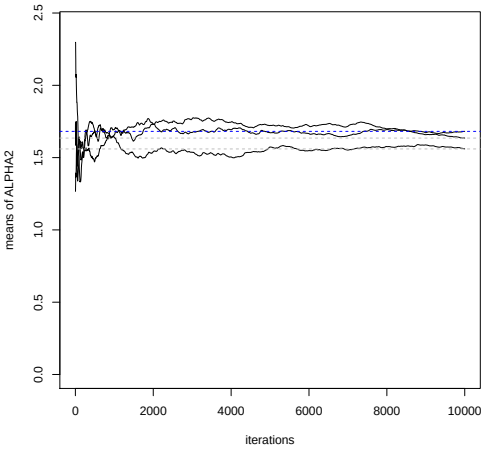
The Bayesian method shows apparently no dependency on the choice of initial values, as it seems to return very similar values every time, even though the initial values differ significantly. The cumulated posterior means, given by $\left(\frac{1}{i}\sum_{j=1}^{i}\theta^{(j)}\right)_{1\le i\le S}$, where S the number of iterations and $\theta^{(j)}$ is the jth value of the paremeter chain $\left(\theta^{(1)},\ldots,\theta^{(S)}\right)$, are calculated for all the parameter chains to observe the convergence of the chains. Figure 7 shows the results for the three different cases.



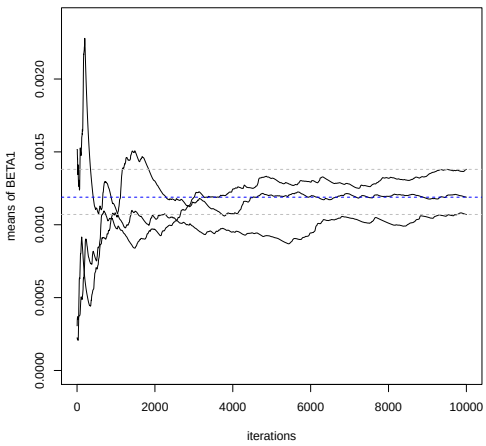
(a) μ



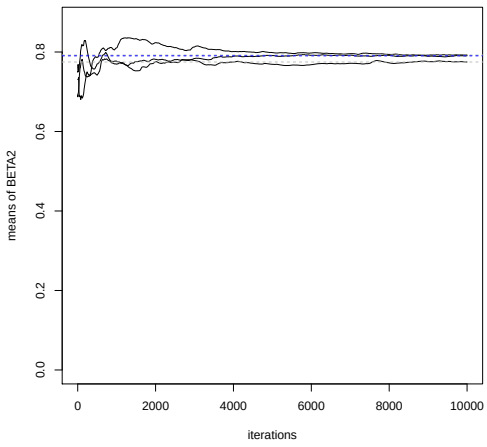
(b) α_1



(c) α_2



(d) β_1



(e) β_2

Figure 7: Convergence of means of the parameter chains under the temporal ETAS model choosing different initial values. The dotted lines represent the mean value of the 10,000 iterations.

6.2 Estimation of the Spatio-temporal ETAS Model

The estimation results of both methods described in section 5 are listed in Table 3 using three different initial values. Figure 8a and Figure 8b show the traceplot of the parameter chain of γ_2 obtained with the Bayesian method and its histogram in one of the cases, respectively.

Table 3: Estimated parameter values under the spatio-temporal ETAS model for the California data set with three different initial values, where the parameter values shown as the fitted by Bayesian inference are the mean of the posterior approximations.

Trial	Method	μ	α_1	α_2	β_1	β_2	γ_1	γ_2	$\log L$
I	Initial	0.8566	0.74	0.05	0.1629	0.79	0.47	0.001	–
	MLE	**0.0004**	$\mathbf{2.1 \times 10^{21}}$	$\mathbf{2.6 \times 10^{-11}}$	$\mathbf{3.4 \times 10^{-29}}$	**0.73**	**2.34**	**0.001**	**−702.04**
	Bayes	**0.0004**	**3.04**	**1.64**	**0.0011**	**0.77**	**2.34**	**0.035**	**−780.29**
II	Initial	0.1809	2.22	2.07	0.0254	2.10	1.33	0.199	–
	MLE	**0.0004**	**0.10**	**0.48**	**0.3309**	$\mathbf{1.8 \times 10^{-7}}$	**2.34**	**0.001**	**−773.67**
	Bayes	**0.0004**	**2.79**	**1.56**	**0.0014**	**0.79**	**2.34**	**0.042**	**−783.64**
III	Initial	2.4121	0.56	0.56	0.0598	1.45	1.68	1.220	–
	MLE	**0.0004**	$\mathbf{2.3 \times 10^{84}}$	$\mathbf{4.5 \times 10^{-237}}$	$\mathbf{1.4 \times 10^{-115}}$	**0.73**	**2.34**	**0.001**	**−702.04**
	Bayes	**0.0004**	**2.89**	**1.68**	**0.0012**	**0.79**	**2.34**	**0.034**	**−784.85**

In the first and third cases, the parameter values α_2 and β_1 under maximum likelihood extimation tend to zero and the parameter value α_1 is much too large. The implication of these values is significant for the model fitting. The estimated values of α_2 indicate that the optimization algorithm ignores the effects of $\alpha(m)$ in (14), i.e.

$$\alpha(m) = \exp\left[\alpha_2 \left(m - m_c\right)\right] \xrightarrow{\alpha_2 \to 0} 1. \tag{43}$$

The value α_1 is a normalizing parameter that affects considerably the offspring intensity (the summation in (12)) ignoring as well the effects of the expected number of triggered events $\alpha(m)$ and the modified Omori law $\beta(t)$ in (15). The effect of the modified Omori law is also ignored because of the estimated values of β_1, similar as explained in Section 6.1.2, but tending to zero, i.e.

$$\beta(t - t_i) = \left(1 + \frac{t - t_i}{\beta_1}\right)^{-\beta_2} \xrightarrow{\beta_1 \to 0} \infty. \tag{44}$$

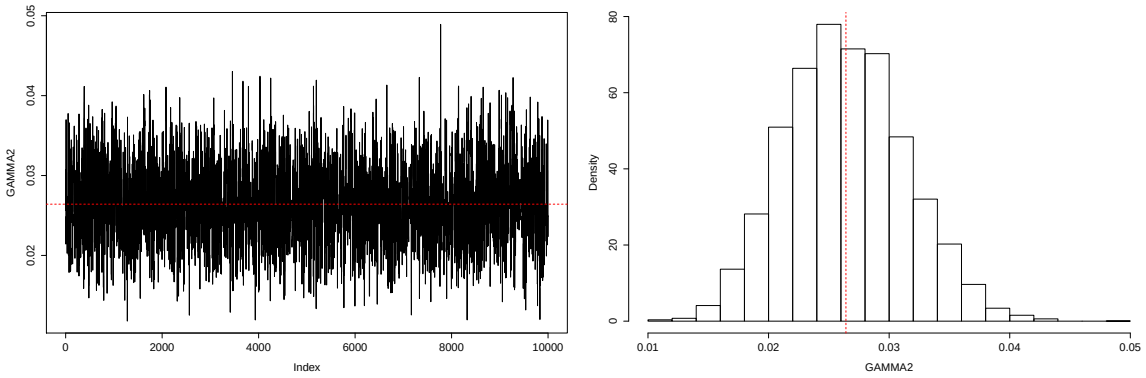
A similar situation occurs in the second case, where the estimated value of β_2 under maximum likelihood estimation tends also to zero, i.e.

$$\beta(t - t_i) = \left(1 + \frac{t - t_i}{\beta_1}\right)^{-\beta_2} \xrightarrow{\beta_2 \to 0} 1. \tag{45}$$

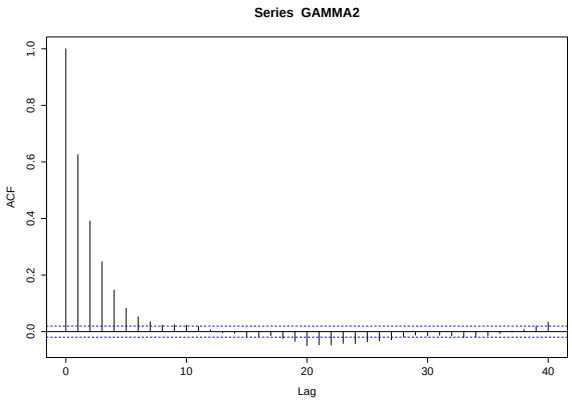
This result is probably an error in the `R` package `PtProcess`, as the functions implemented in the spatio-temporal case are trial versions.

The Bayesian method shows again apparently no dependency on the choice of initial values, as it seems to return similar values every time, even though the initial values differ significantly. The log-likelihood value obtained with our proposed method is close to the probable maximum value, i.e. the proposed estimation method is also maximizing the log-likelihood function of the spatio-temporal ETAS model.

Figure 8c shows the autocorrelation function of the parameter chain of γ_2. This chain has a low degree of correlation, which means that the approximation is good. This results was expected, as the other parameter values were estimated previously.



(a) Traceplot (b) Histogram



(c) acf

Figure 8: Traceplot, histogram and autocorrelation function of the parameter chain of γ_2, respectively, for the California dataset under the spatio-temporal ETAS model.

7 Prediction under the Temporal ETAS model

Prediction of earthquake occurrences is focused on determine the probability distribution of the time to the next event with a given magnitude. This distribution can be determined empirically by simulation (Harte (2010)).

To simulate a point process with the fitted conditional intensity function $\lambda_{\hat{\theta}}(t, m|\mathcal{H}_t)$ we apply the thinning method of Ogata (1981) as implemented in the `simulate` function (Simulate a Point Process) of the `R` package `PtProcess`. The method calculates an upper bound for the intensity function, simulating a value for the time to the next possible event using a rate equal to this upper bound. It then calculates the intensity at this simulated point. The ratio of this rate to the upper bound is compared with a uniform random number to randomly determine whether the simulated time is accepted or not (Harte (2010)). The result of one simulation is shown in Figure 9.

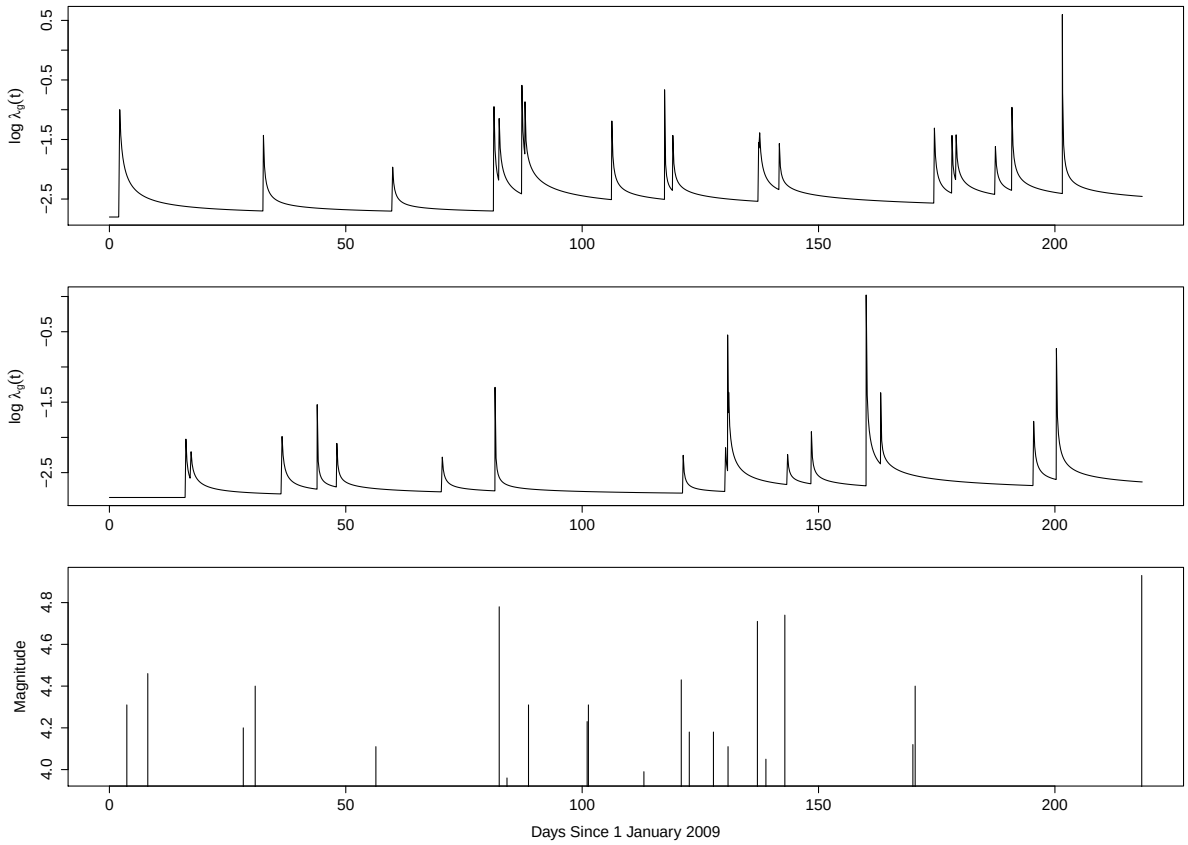


Figure 9: Example of one simulation. The last plot shows the magnitudes of events from the control data set versus occurrence times. The first plot shows $\log \lambda_{\hat{\theta}}(t, m|\mathcal{H}_t)$ fitted with the proposed method to the simulated dataset versus the occurrence time. The second plot shows $\log \lambda_{\hat{\theta}}(t, m|\mathcal{H}_t)$ fitted with the maximum likelihood method to the simulated dataset versus the occurrence time.

The California training data set has an observation period that finishes at midnight on December 31, 2008 (day 1096 since January 1, 2006), and we want to determine the probability distribution of the time to the next event with magnitude greater than or equal to 4.75. We

choose this magnitude, since it is the first strong earthquake occurrence that it is registered in the control data set (at day 82.45). The simulation stops only by meeting the given stopping condition. 2000 simulations are performed and from each the time to the first magnitude greater than or equal to 4.75 is recorded in order to plot a histogram of these times (in days from January 1, 2009). The prediction results using both methods are shown in Figure 10 and 11. The 0.5, 0.8, 0.9, 0.95 and 0.99 quantiles of the empirical distributions are listed in Table 4.

Table 4: The 0.5, 0.8, 0.9, 0.95 and 0.99 quantiles of the empirical distribution of the times to the given event.

Method	0.50	0.80	0.90	0.95	0.99
Bayes	**27.27**	**63.49**	**96.89**	**127.46**	**186.49**
MLE	**36.70**	**90.40**	**132.42**	**173.10**	**235.53**

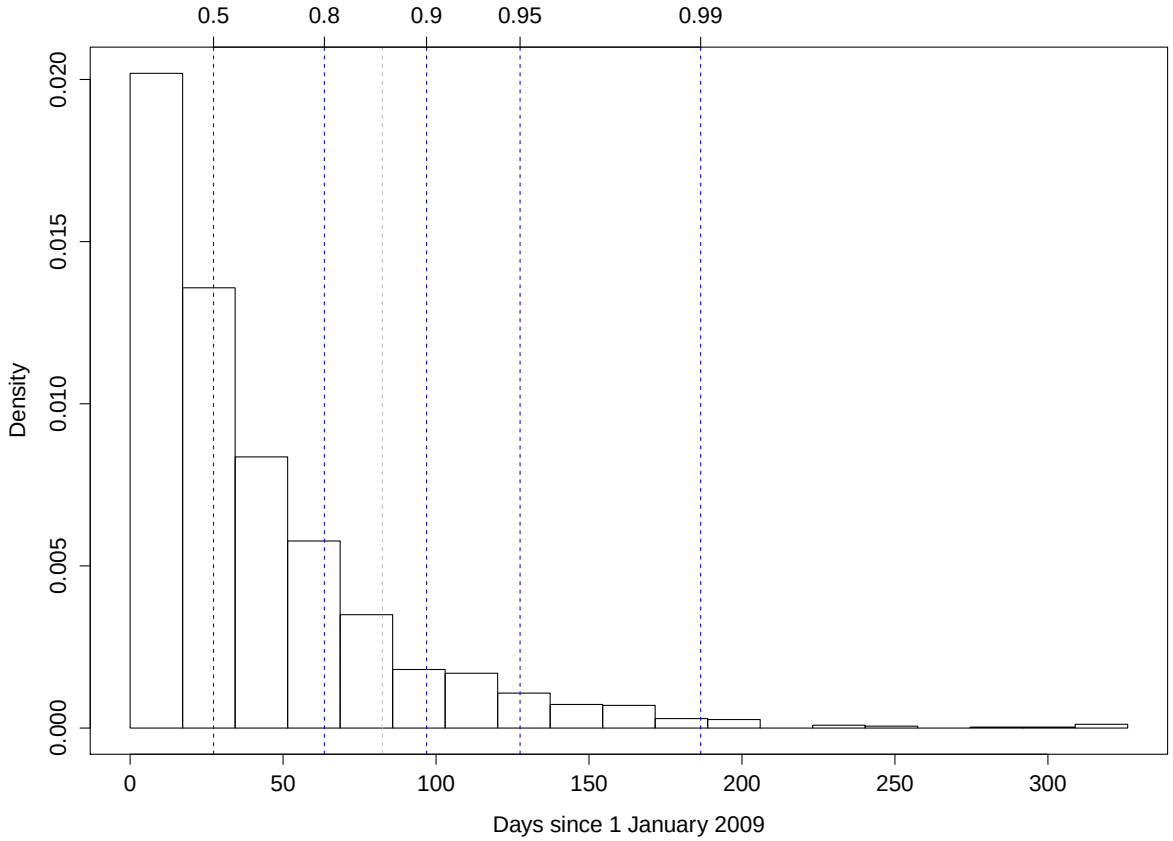


Figure 10: The temporal ETAS model fitted with the proposed method to the California data set. 2000 simulations were performed from January 1, 2009 until the first magnitude ≥ 4.75 event in each occured. The histogram represents the empirical distribution of the times to this event. The dotted blue lines represent the 0.5, 0.8, 0.9, 0.95 and 0.99 quantiles. The gray line represents the first event registered with the given magnitude in the control data set (day 82.45).

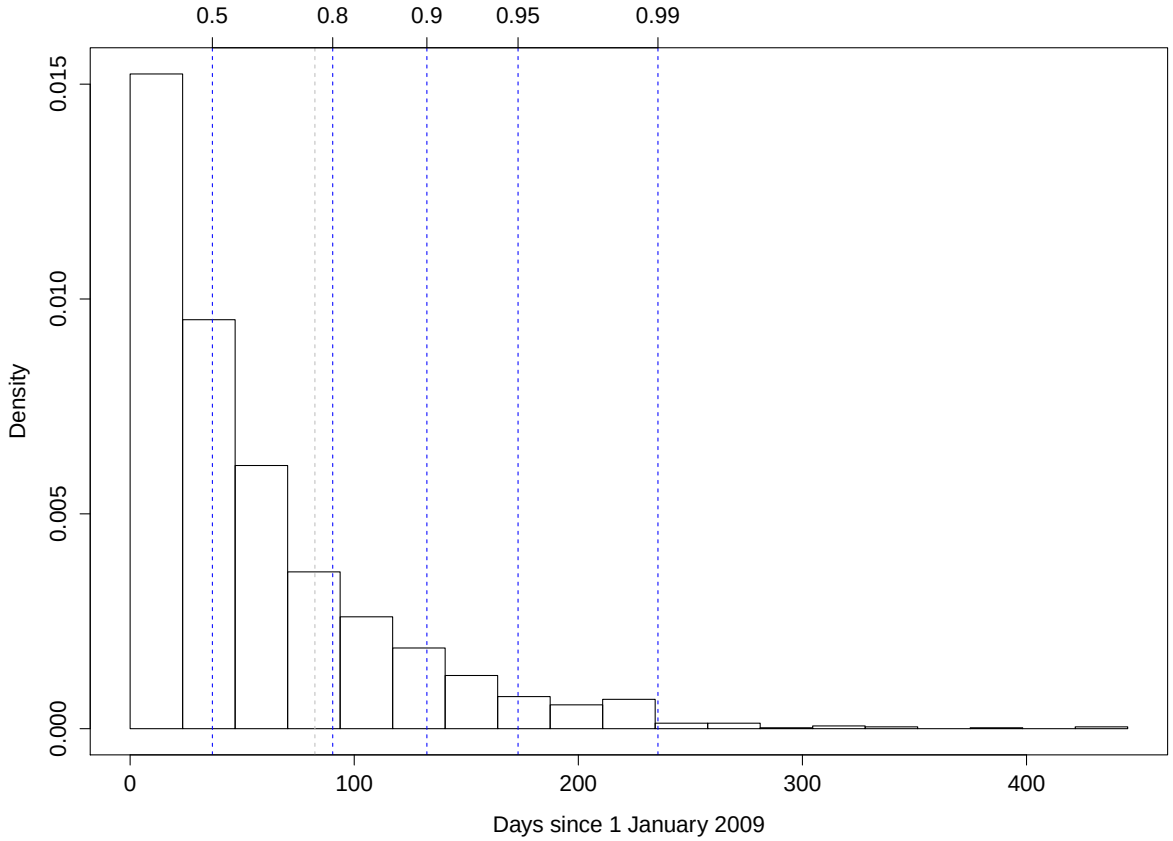


Figure 11: The temporal ETAS model fitted with the MLE to the California data set. 2000 simulations were performed from January 1, 2009 until the first magnitude ≥ 4.75 event in each occured. The histogram represents the empirical distribution of the times to this event. The dotted blue lines represent the 0.5, 0.8, 0.9, 0.95 and 0.99 quantiles. The gray line represents the first event registered with the given magnitude in the control data set (day 82.45).

Predictive Performance We want to compare the predictive perfomance using our different estimation methods. We produce 218 predictions as the one shown above taking as starting point every day from January 1, 2009 until August 6, 2009.

A useful forecast measure for comparing different methods on the same data set is the mean absolute error (MAE) (Hyndman and Koehler (2006)). The forecast error is defined as

$$e_t = |F_t - x_t|\,, \tag{46}$$

where F_t is the median of the predictive distribution and x_t is the observation at time t. Thus the mean absolute error is given by

$$\mathbf{MAE}(F, x) = \frac{1}{n}\sum_{i=1}^{n} |F_t - x_i| = \frac{1}{n}\sum_{i=1}^{n} e_i, \tag{47}$$

where F is the prediction vector and x is the observation vector. The MAE values calculated are listed in Table 5.

Scoring rules provide summary measures for the evaluation of probabilistic forecasts, by assigning a numerical score based on the predictive distribution and on the event or value that should predict. They measure the quality of the probabilistic forecasts in order to rank competing prediction procedures (Gneiting and Raftery (2007)).

Gneiting and Raftery (2007) state that the restriction to predictive densities is often unpractical, like in our case, since our predictive distributions are expressed in terms of samples. Therefore, it is better to define scoring rules directly in terms of predictive cumulative distribution functions. The continuous ranked probability score (CRPS) is defined as

$$\mathbf{CRPS}(F, x) = \int_{-\infty}^{\infty} (F(y) - \mathbf{1}_{\{y \geq x\}})^2\, dy, \tag{48}$$

where F is the cumulative distribution function of a probabilistc forecast.

Applications of the CRPS have been restricted by a lack of readily computable solutions to the integral in (48). Therefore, the integral often can be evaluated in a closed form given by

$$\mathbf{CRPS}(F, x) = \frac{1}{M}\sum_{m=1}^{M} |x_m - x| - \frac{1}{2M^2}\sum_{m=1}^{M}\sum_{n=1}^{M} |x_m - x_n|\,, \tag{49}$$

where F is again the cumulative distribution function from a forecast ensemble of size M, x is the verifying observation and $|\cdot|$ denotes the Euclidean norm. We calculate this generalization of the CRPS presented by Grimit et al. (2006) in our simulations, the results are listed in Table 5.

Table 5: Predictive performance for predicting number of days until the next earthquake event of magnitude 4.75 or greater in our study region for daily forecasts from January 1, 2009 until August 6, 2009. The performance is measured by the mean absolute error (MAE) and the mean continuous ranked probability score (CRPS). The other two values are the MAE and CRPS of the predictions between January 1, 2009 and March 26, 2009 and the ones between March 27, 2009 and August 6, 2009, respectively.

Method	MAE		CRPS	
Bayes	**34.24**		**23.17**	
	21.67	41.81	16.88	26.96
MLE	**33.41**		**23.02**	
	22.02	40.28	17.79	26.18

The two methods show similar predictive performance, with the MLE method performing slightly better on average under both the MAE and the CRPS. Both methods seem extremely insensitive to changes in time and the daily predictions are very similar throughout the entire test period. For the MLE method, the predictive median is always between 35 and 46 days while the predictive median for the Bayes methods lies between 33 and 40 days. That is, the model doesn't seem to adapt to there being a higher or a lower chance of a large earthquake even if such an event just happened. This might partly be due to the fact that we have used constant parameter estimates over the test set. Furthermore, our entire data set only contains 22 events of magnitude 4.75 or greater. A larger study is thus needed to assess the robustness of the methods. However, this small example demonstrates that forecast verification methods for point process models may be extended beyond the currently used residual methods.

8 Verification and Comparison

Verification methods are useful to evaluate the goodness of fit of models. Clements et al. (2011) presents residual analysis methods for temporal and spatio-temporal point processes, which are applied to earthquake occurrences models in California. We calculate simple, low-power means (L-test and N-test) and residuals as proposed in Harte (2010) to evaluate the fit of both models in the temporal case. Further methods are outlined but not implemented, as the tools are not available in the `PtProcess` package.

8.1 L-test and N-test

The Likelihood-test (L-test) and the Number-test (N-test) are goodness of fit tests. The L-test evaluates the quality of a model in the likelihood space, and the N-test compares the total predicted rate with the observated rate. A detailed description of these methods can be found in Schorlemmer et al. (2007).

Likelihood-test The L-test works by simulating s realizations from the fitted model. The log-likelihood is calculated for the observed data set (l_{obs}) and each simulation (l_j, for $j = 1, \ldots, s$). The quantile score η is given by

$$\eta = \frac{1}{s} \sum_{j=1}^{s} \mathbf{1}_{\{l_j < l_{\text{obs}}\}}, \tag{50}$$

where $\mathbf{1}_{\{.\}}$ denotes the indicator function. If η is close to zero, then the model is considered to be inconsistent with the data set, and can be rejected. Otherwise, the model is not rejected and further tests are necessary (Clements et al. (2011)).

Number-test The N-test is similar to the L-test, except that the quantile score shows instead the fraction of simulations that contain fewer points than the observed data set. This quantile score δ is given by

$$\delta = \frac{1}{s} \sum_{j=1}^{s} \mathbf{1}_{\{N_j < N_{\text{obs}}\}}, \tag{51}$$

where N_j is the number of points contained in the jth simulation of the model and N_{obs} is the number of points contained in the observed data set. If δ is close to 0 or 1, the model is rejected.

We apply two different methods to obtain predictions for earthquakes in California from January 1, 2009 to September 1, 2009 under the Bayesian model. The Metropolis-Hastings algorithm was implemented once in order to get parameter chains based on the training data as described in section 7. We then proceeded to simulate 250 random point patterns using the posterior means of the model parameters. In a second alternative method, we draw an independent sample from the posterior distributions of the parameters for each simulated point pattern. For comparison, we also simulate 250 patterns from the temporal ETAS model where the parameter estimates are obtain with maximum likelihood estimation, see section 7.

Table 6 shows the results implementing the two methods and the MLE. The L-test indicates that for both methods the model can not be rejected, because the η scores are not close to zero. According to the N-test the model evaluated with the method I is not significantly over-predicting the total number of earthquakes, whereas the other two results have good δ scores.

Table 6: Results of the L- and N-test for earthquake predictions in California from January 1, 2009 to September 1, 2009 under the temporal ETAS model. Method I denotes the Bayesian method under the posterior mean and method II denotes the Bayesian method with a random sample from the posterior distribution. Listed are the log-likelihood of the observed data set l_{obs}, the quantile score η, the number of observed events N_{obs} and the quantile score δ.

Method	l_{obs}	η	N_{obs}	δ
I	-72.921	**0.572**	21	**0.32**
II	-72.921	**0.308**	21	**0.56**
MLE	-72.533	**0.296**	21	**0.52**

According to the L-test the model evaluated with method I has a better performance as the other two. Nevertheless, the model evaluated with method II and the one fitted with MLE show better results according to the N-test. This discrepancy between both tests leads us to believe that they are not accurate. Clements et al. (2011) consider that these tests have very low power. Therefore, they propose instead residual methods, such as the residuals presented in Section 8.2 and the methods outlined in Section 8.3.

8.2 Residuals of Point Process Models

The `R` package `PtProcess` provides a function named `residuals` that calculates the residuals of a point process with the fitted conditional intensity function $\lambda_{\hat{\theta}}(t, m|\mathcal{H}_t)$. This method evaluates the goodness of fit of a model calculating a so- called residual process (Harte (2010)).

Let t_i be the times of the observed events. The transformed times are defined as

$$\tau_i = \int_0^{t_i} \lambda_{\hat{\theta}}(t, m|\mathcal{H}_t)\, dt, \tag{52}$$

where $\lambda_{\hat{\theta}}(t, m|\mathcal{H}_t)$ is the fitted intensity function given by (19). If the dataset is sampled from a process with intensity $\lambda_{\hat{\theta}}(t, m|\mathcal{H}_t)$, then the transformed time points form a stationary Poisson process with rate parameter one, which is the residual process (Aalen and Hoem (1978)).

A simple graphical diagnostic test of the goodness of fit of the model is to plot the event number i versus the trasformed time τ_i. The points should roughly follow a straight line $y = x$ (Harte (2010)). Significant departures from the straight line show a weakness in the model. If the line has a slope less than one in a given interval, then the transformed times τ_i are too small, which means that the fitted intensity function $\lambda_{\hat{\theta}}(t, m|\mathcal{H}_t)$ is too small in the given interval. In the same way, if the slope is greater than one, the fitted intensity function $\lambda_{\hat{\theta}}(t, m|\mathcal{H}_t)$ is too large. Figure 12 is the resulting plot implementing the fitted models to the

California training data set.

An alternative representation of this test is to plot τ_i on the vertical scale. This plot is called cusum and was devised by Page (1974). It represents the cumulative sum of the inter-event residual times, and the substraction of i removes the mean (Harte (2010)). Therefore, when the process is consistent, the cusum has a zero slope. A positive or negative slope means that the fitted intensity function is either too large or too small, respectively. Figure 13 is the cusum plot for both fitted models for the California training data set.

The results of the residual analysis show that the model fitted with the MLE method has a better goodness of fit as the model fitted with the proposed Bayesian method. Figures 12 and 13 indicate that the model fitted with our proposed method is overestimating the data set. This is probably because of the estimation of the modified Omori law parameter values. The model fitted with MLE shows also subintervals, where the departures from the straight line are quite significant. This could be a consecuence of the few points of the training data set.

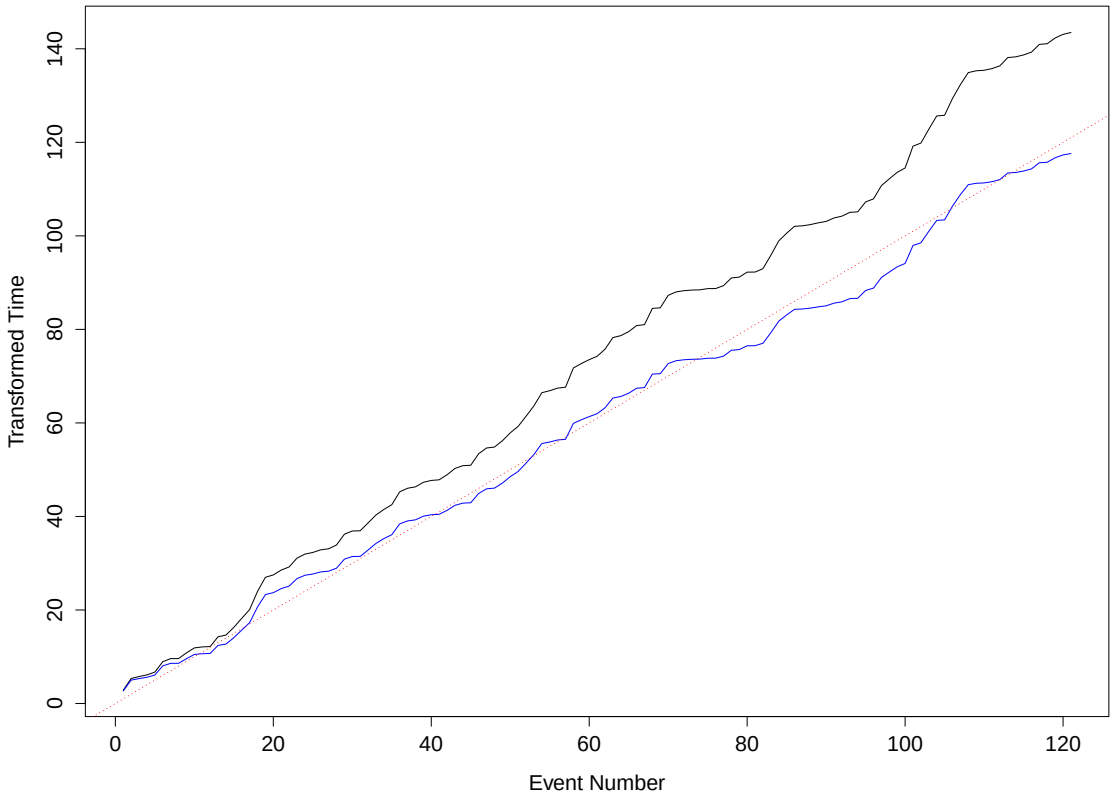


Figure 12: Residual process times for the temporal ETAS model fitted to the California data set with both models. The black line represents the model using the proposed Bayesian method. The blue line represents the model using the MLE method. The red dotted line is $y = x$.

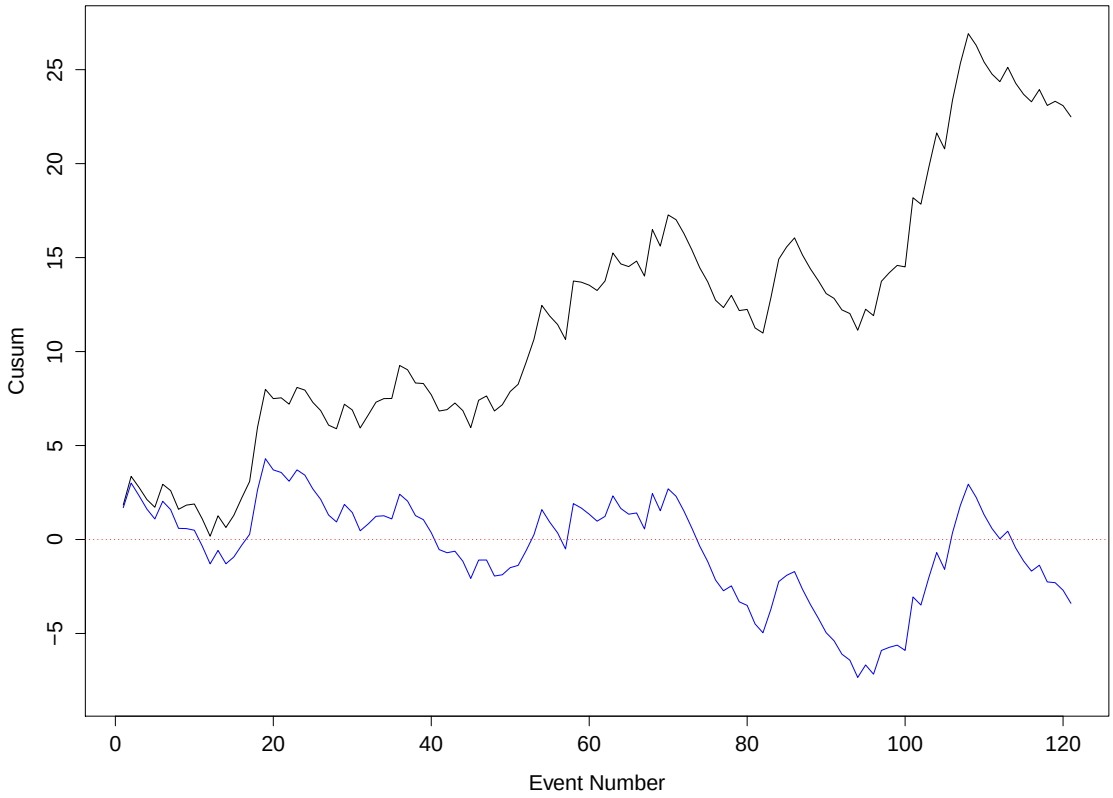


Figure 13: Cusum of the residual process times for the temporal ETAS model fitted to the California data set with both models. The black line represents the model using the proposed Bayesian method. The blue line represents the model using the MLE method. The red dotted line is $y = 0$.

8.3 Further Verification Methods

Clements et al. (2011) propose several methods to evaluate the fit of a point process model. These evaluation tools include residual point process methods such as rescaling, thinning and superposition, comparative quadrat methods such as Pearson residuals and deviance residuals, and weighted second-order statistics to evaluate particular features of a model such as its background rate or the degree of spatial clustering.

Rescaling, thinning and superposition Rescaled residuals are useful to assess the overall fit of a model, as well as thinned and superimposed residuals.

Rescaled residuals are the result of rescaling the temporal coordinates of a multivariate point process according to the integrated conditional intensity in order to form a sequence of stationary Poisson processes.

Thinned residuals are useful to evaluate the spatial fit of a spatio-temporal point process model and to reveal locations where the model is fitting poorly (Clements et al. (2011)). These ones have the advantage that the coordinates of the points are not transformed as in the case of rescaled residuals. Therefore, the resulting residuals may be easier to interpret. To calculate thinned residuals, each point $(x_i, y_i, t_i) \in S$ is kept independently with probability

$$\frac{b}{\hat{\lambda}\left(t_i, x_i, y_i\right)}, \tag{53}$$

where $b = \inf\{\hat{\lambda}(t, x, y) \ : \ (t, x, y) \in S\}$ is the infimum of the estimated intensity over the entire observed space-time window, S. The remaining points are the so-called thinned residual points, which should be homogeneous Poisson with rate b if and only if the fitted model for λ is consistent (Schoenberg (2003)).

Superposition is a residual analysis method similar to thinned residuals, but instead of substracting points, new points are simulated to be added to the data (Clements et al. (2011)). Points are simulated at each location (t, x, y) according to a Cox process with intensity $c - \hat{\lambda}\left(x_i, y_i, t_i, m_i\right)$, where $c = \sup_S\{\hat{\lambda}(x, y, t)\}$. If the model is consistent, the union of the superimposed residuals and observed points are homogeneous Poisson.

However, these methods are generally unpractical when λ is spatially volatile, as these methods have limited power when the modeled conditional intensity assumes extremely low or high values in the observation regions, which is commonly the case for earthquake occurrences models (Clements et al. (2011)).

Comparative quadrat methods Pixel-based approaches introduced by Baddeley et al. (2005), such as Pearson residuals and deviance residuals, are used to compare models. They are based on comparing the total number of point of an observation region to the number predicted by the model.

Raw residuals are defined by Baddeley et al. (2005) as the number of observed points minus

the number of expected points in each pixel, i.e.

$$R(B_i) = N(B_i) - \int_{B_i} \hat{\lambda}(t, x, y)\, dtdxdy, \tag{54}$$

where $N(B_i)$ is the number of points in bin i. Pearson residuals are rescaled raw residuals with mean 0 and variance approximately equal to 1. They are given by

$$R_P(B_i) = \sum_{(x_i,y_i)\in B_i} \frac{1}{\sqrt{\hat{\lambda}(x_i, y_i)}} - \int_{B_i} \sqrt{\hat{\lambda}(x, y)}\, dxdy, \tag{55}$$

for all $\hat{\lambda}(x, y) > 0$.

These residuals are a good technique to identify individual bins containing earthquake occurrences that require an adjustment in their forecast rates, however they generally fail to identify other locations where the models may fit relatively well or poorly (Clements et al. (2011)).

Weighted second-order statistics The Ripley's K-function (Ripley (1981)) is a common tool to detect clustering or inhibition in a point process. It is defined as the average number of points within r of any given point divided by the overall rate λ, i.e.

$$K(r) = A\, N^{-2} \sum_{i<j,\, \|\mathbf{x}_i - \mathbf{x}_j\| < r} s(\mathbf{x}_i, \mathbf{x}_j), \tag{56}$$

where A is the area of the observation region, N is the total number of observed points, and $s(\mathbf{x}_i, \mathbf{x}_j)^{-1}$ is the proportion of area of the ball centered at $\mathbf{x}_i$ and passing through $\mathbf{x}_j$ that falls within the observation region.

The weighted K-function is good to test the degree of clustering in the model. The standard estimate of the weighted K-function is defined as

$$K_W(r) = \frac{b}{\int_S \hat{\lambda}_0(\mathbf{x})\, d\mathbf{x}} \sum_i \hat{\lambda}_0(\mathbf{x}_i)^{-1} \sum_{i\neq j} \hat{\lambda}_0(\mathbf{x}_j)^{-1} \mathbf{1}_{\{\|\mathbf{x}_i - \mathbf{x}_j\| \leq r\}}, \tag{57}$$

where $b = \min(\hat{\lambda})$, $\mathbf{1}_{\{.\}}$ is the indicator function, and $\hat{\lambda}_0(\mathbf{x}_i)$ is the conditional intensity at point $\mathbf{x}_i$ under a null hypothesis.

The weighted second-order statistics are especially useful for comparisons of competing models (Clements et al. (2011)).

9 Conclusions

We present a Bayesian method to estimate the Epidemic-Type Aftershock Sequences (ETAS) model. The advantages of this method compared to the conventional maximum likelihood estimation are substantial in terms of convergence and accuracy independently of the choice of initial values.

In Section 6.1 we show the results using as a case study a California earthquake occurrence data set, which demonstrate the good performance of the proposed method. The difficulties in estimating the parameter values of the modified Omori law could be fixed generating more samples with the algorithm described in Section 4.2. The advantage of convergence of our method may be a useful way to obtain initial values for the optimization algorithms of a maximum likelihood estimation, in case numerical maximization is preferred (Veen and Schoenberg (2008)).

In Section 6.2 we show the limited results, which can be obtained with the current tools in the `R` package implemented. Although we can not compare the results with the ones obtained with the conventional maximum likelihood method, we expect that the proposed method is also applicable to the spatio-temporal ETAS model.

In Section 7 we show the predictive performance with both methods in the temporal case. The two methods show similar predictive performance, with the MLE method performing slightly better on average under both the MAE and the CRPS. Both methods seem extremely insensitive to changes in time and the daily predictions are very similar throughout the entire test period.

On the other hand, according to the L- and N-test in Section 8 we can not reject the Bayesian method, but the residual analysis shows that its fit is not that accurate as the fit of the other method. However, according to the residual process the model fitted with MLE shows subintervals, where the departures from the straight line are quite significant. This could be a consecuence of the number of points of the training data set. We only have 121 events in a interval of time of 1096 days, which could lead us to unexpected results in the estimation procedures as well as in the prediction performances.

The model proposed by Ogata (1998) ignores possible spatial covariate effects. In the case of the region of California it is known that the San Andreas fault may have an influence in the earthquake occurrences. The San Andreas fault is a geological feature, which can be termed as lineament. It would be of interest to predict earthquake occurences from the lineament pattern (Baddeley et al. (2005)). The lineament pattern of this region would be a possible spatial covariate, therefore, it should be included in the analysis. As described in Baddeley et al. (2005), the null model would propose that the earthquake occurrences are a homogeneous Poisson process, which means that it is assumed that there is no dependence on the lineaments. An alternative model would propose, for example, that the density of earthquake occurrences depends on the distance from the nearest lineament.

References

Aalen, O. O. and Hoem, J. M. (1978), 'Random time changes for multivariate counting processes', *Scandinavian Journal of Statistics* **5**, 81–101.

Baddeley, A. and Turner, R. (2005), 'Spatstat: an R package for analyzing spatial point patterns', *Journal of Statistical Software* **12**(6), 1–42. URL: `www.jstatsoft.org`, ISSN: 1548-7660.

Baddeley, A., Turner, R., Møller, J. and Hazelton, M. (2005), 'Residual analysis for spatial point processes', *Journal of the Royal Statistical Society: Series B* **67**, 617–666.

Berthelsen, K. K. and Møller, J. (2003), 'Likelihood and non-parametric Bayesian MCMC inference for spatial point processes based on perfect simulation and path sampling', *Scandinavian Journal of Statistics* **30**, 549–564.

Clements, R. A., Schoenberg, F. P. and Schorlemmer, D. (2011), 'Residual analysis methods for space-time point processes with applications to earthquake forecast models in California', *Annals of Applied Statistics* **4**, 2549–2571.

Collaboratory for the Study of Earthquake Predictability (CSEP) Development website (2012). **URL:** *http://northridge.usc.edu/trac/csep/wiki*

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

Gerstenberger, M. C., Wiemer, S., Jones, L. M. and Reasenberg, P. A. (2005), 'Real-time forecasts of tomorrow's earthquakes in California', *Nature* **435**, 328–331.

Gneiting, T. and Raftery, A. E. (2007), 'Strictly proper scoring rules, prediction, and estimation', *Journal of the American Statistical Association* **102**, 359–378.

Grimit, E. P., Gneiting, T., Berrocal, V. and Johnson, N. A. (2006), 'The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification', *Quarterly Journal of the Royal Meteorological Society* **132**, 2925–2942.

Harte, D. (2010), 'PtProcess: An R package for modelling marked point processes indexed by time', *Journal of Statistical Software* **35**(8), 1–32.
URL: *http://www.jstatsoft.org/v35/i08/*

Hoff, P. D. (2009), *A First Course in Bayesian Statistical Methods*, Springer.

Hyndman, R. J. and Koehler, A. B. (2006), 'Another look at measures of forecast accuracy', *International Journal of Forecasting* **22**, 679–688.

Möller, A., Lenkoski, A. and Thorarinsdottir, T. L. (2012), 'Multivariate probabilistic forecasting using ensemble Bayesian model averaging and copulas', *Quarterly Journal of the Royal Meteorological Society* **98**, 789–795.

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

Ogata, Y. (1988), 'Statistical models for earthquake occurrences and residual analysis for point processes', *Journal of the American Statistical Association* **83**, 9–27.

Ogata, Y. (1989), 'Statistical model for standard seismicity and detection of anomalies by residual analysis', *Tectonophysics* **169**, 159–174.

Ogata, Y. (1998), 'Space-time point-process models for earthquake occurrences', *Annals of the Institute of Statistical Mathematics* **50**, 379–402.

Ogata, Y., Katsura, K. and Tanemura, M. (2003), 'Modelling heterogeneous space-time occurences of earthquakes and its residual analysis', *Applied Statistics* **52**, 499–509.

Ogata, Y. and Zhuang, J. (2006), 'Space-time ETAS models and an improved extension', *Tectonophysics* **413**, 13–23.

Page, E. S. (1974), 'Continuous inspection schemes', *Biometrika* **41**, 100–115.

Rasmussen, J. G. (2011), 'Bayesian inference for Hawkes processes', *Methodology and Computing in Applied Probability* pp. 1–20.
URL: *http://dx.doi.org/10.1007/s11009-011-9272-5*

Reasenberg, P. A. and Jones, L. M. (1989), 'Earthquake hazard after a mainshock in California', *Science* **243**, 1173–1176.

Reasenberg, P. A. and Jones, L. M. (1990), 'California aftershock hazard forecast', *Science* **247**, 345–346.

Reasenberg, P. A. and Jones, L. M. (1994), 'Earthquake aftershocks: Update', *Science* **265**, 1251–1252.

Ripley, B. (1981), *Spatial Statistics*, Wiley.

Schoenberg, F. P. (2003), 'Multidimensional residual analysis of point process models for earthquake occurrences', *Journal of the American Statistical Association* **98**, 789–795.

Schorlemmer, D., Gerstenberger, M. C., Wiemer, S., Jackson, D. D. and Rhoades, D. A. (2007), 'Earthquake likelihood model testing', *Seismological Research Letters* **78**, 17–27.

Stoffer, P. W. (2006), *Where's the San Andreas Fault? A Guidebook to Tracing the Fault on Public Lands in the San Francisco Bay Region*, U.S. Geological Survey, General Information Product 16.

Turcotte, D. L., Holliday, J. R. and Rundle, J. B. (2007), 'BASS: an alternative to ETAS', *Geophysical Research Letters* **34**.

Veen, A. and Schoenberg, F. P. (2008), 'Estimation of space-time branching process models in seismology using an EM-type algorithm', *Journal of the American Statistical Association* **103**, 614–624.

Vere-Jones, D. (1970), 'Stochastic models for earthquake occurrence', *Journal of the Royal Statistical Society, Series B* **32**, 1–62.

Vere-Jones, D. (1975), 'Stochastic models for earthquake sequences', *Geophysical Journal of the Royal Astronomical Society* **42**, 811–826.

Woessner, J., Christophersen, A., Zechar, J. D. and Monelli, D. (2010), 'Building self-consistent, short-term earthquake probability (STEP) models: improved strategies and calibration procedures', *Annals of Geophysics* **53**, 141–154.

Wordsworth, W. (1888), *The Complete Poetical Works*, Macmillan and Co., London. Bartleby.com, 1999.

Zhuang, J. (2006), 'Second-order residual analysis of spatiotemporal point processes and applications in model evaluation', *Journal of the Royal Statistical Society: Series B* **68**, 635–653.

Zhuang, J., Ogata, Y. and Vere-Jones, D. (2004), 'Analyzing earthquake clustering features by using stochastic reconstruction', *Journal of Geophysical Research* **109**.