Ann Inst Stat Math (2012) 64:521–544
DOI 10.1007/s10463-010-0320-7

Markov-modulated Hawkes process with stepwise decay

Ting Wang · Mark Bebbington · David Harte

Received: 15 April 2010 / Revised: 30 August 2010 / Published online: 30 December 2010

Abstract This paper proposes a new model—the Markov-modulated Hawkes process with stepwise decay (MMHPSD)—to investigate the variation in seismicity rate during a series of earthquake sequences including multiple main shocks. The MMHPSD is a self-exciting process which switches among different states, in each of which the process has distinguishable background seismicity and decay rates. Parameter estimation is developed via the expectation maximization algorithm. The model is applied to data from the Landers–Hector Mine earthquake sequence, demonstrating that it is useful for modelling changes in the temporal patterns of seismicity. The states in the model can capture the behavior of main shocks, large aftershocks, secondary aftershocks, and a period of quiescence with different background rates and decay rates.

Keywords Markov-modulated Hawkes process with stepwise decay · EM algorithm · ETAS model · Simulation · Landers

1 Introduction

Fedotov (1968) and Mogi (1968) described the concept of a seismic cycle including foreshocks and relative quiescence. Statistically, there have been few attempts to identify, or exploit this. Zhuang (2000) analyzed the earthquake sequence off Cape Palliser in New Zealand from 1978 to 1996 and divided the seismicity into

T. Wang (✉) · M. Bebbington
Volcanic Risk Solutions, Massey University, Private Bag 11222,
Palmerston North 4442, New Zealand
e-mail: t.wang@massey.ac.nz

D. Harte
Statistics Research Associates, PO Box 12 649, Wellington 6144, New Zealand

four periods: background seismicity, a relatively quiescent period, main shock and aftershock sequence, and a postseismic period. According to variations in occurrence rate and probability distribution of inter-event time, Pievatolo and Rotondi (2008) decomposed the observed seismicity from the Kresna region of Bulgaria into different phases and found two seismic cycles which have similar patterns as background activity—foreshocks, main shocks, and aftershock sequences.

Bebbington et al. (2010) examined a cyclical model for the San Francisco Bay region, showing that it reproduces observed accelerated moment release (AMR) behavior (Bufe and Varnes 1993). This study is notable in that the state variable was the accumulation of tectonic strain, not directly reflected in the observed data. Hidden Markov models (HMMs) form a remarkably general statistical framework for modelling such partially observed systems, by assuming that the unobserved (or hidden) process is a Markov chain, and the observations are conditionally independent given the hidden states. However, earthquakes considered as point processes often occur in a self-exciting way, i.e., previous events often trigger new ones. This combination of a self-exciting process and a continuous-time hidden process is not considered in any of the existing HMMs and their extensions, discrete-time HMMs (see Rabiner 1989 for a detailed review), Markov-modulated Poisson processes (MMPPs; Fischer and Meier-Hellstern 1993; Rydén 1994, 1996), nonhomogeneous hidden Markov models (NHMMs; Hughes and Guttorp 1994), Markov-modulated generalized linear models (MMGLMs; Harte 2005), or Markov-modulated renewal processes (Bebbington 2007). Hence, to model earthquakes in the HMM framework, further extension is required.

An MMPP is a doubly stochastic Poisson process, the intensity of which is controlled by a finite, non-observable, continuous-time Markov chain. It is parameterized by an r-state continuous-time Markov chain with infinitesimal generator $Q = (q_{ij})_{r\times r}$ and r Poisson arrival rates $\lambda_1, \lambda_2, \ldots, \lambda_r$. When an MMPP is in one specific state, say s, the event occurrences follow a Poisson process with rate λ_s. This process has been applied to model bursty point processes, especially in telecommunications (see e.g., Heffes and Lucantoni 1986), but as pointed out above, it is not suitable for earthquakes, due to the occurrence rate being constant in any given state, hence not self-exciting.

Hence, it is necessary to extend the MMPP to the case when the observed process follows a self-exciting point process sojourning in each state. The self-exciting Hawkes process with exponential decay rate (Hawkes 1971) will be used to formulate the event occurrence rate in each hidden state. This is because this Hawkes process has the Markovian property and is thus mathematically tractable. We will examine the effect of this choice later.

The next section introduces the Markov-modulated Hawkes process with stepwise decay (MMHPSD), a model in which the process switches among a finite number of states, each characterized by a self-exciting occurrence rate of events from a Hawkes process. Section 3 presents a method of estimating the parameters involved in the model using the EM algorithm, including a detailed implementation algorithm for parameter estimation. The goodness-of-fit problem is discussed in Sect. 4. A simulation algorithm for the process is provided as part of a study of the consistency of the parameter estimation procedure in Sect. 5. Section 6 discusses a case study of the model using the earthquake data for the Landers–Hector Mine series.

2 Markov-modulated Hawkes process with stepwise decay

The Hawkes process has an extensive application history in seismology (see e.g., Hawkes and Adamopoulos 1973), epidemiology, neurophysiology (see e.g., Brémaud and Massoulié 1996), and econometrics (see e.g., Bowsher 2007). It is a point-process analogue of the autoregressive model in time series. Let N be a simple point process on $\mathbb{R}$ with successive occurrence times $T_0 = t_0 = 0, T_1 = t_1, \ldots, T_n = t_n$. The history $\mathcal{H}_t$ of the process at time t is defined by $\mathcal{H}_t = \{t_l : t_l < t\}$. The conditional intensity function of the Hawkes process is $\lambda(t) \stackrel{\triangle}{=} \lambda(t \mid \mathcal{H}_t) = \lambda + \int_{-\infty}^{t} g(t-u)\mathrm{d}N(u)$, where $\lambda \geq 0$, $g(u) \geq 0$ and $\int_0^\infty g(u)\mathrm{d}u < 1$ (Hawkes 1971). The immigrants arrive in a Poisson process of rate λ, and each immigrant generates offspring in a non-stationary Poisson stream of rate $g(t-u)$. If we let $g(t-u) = \nu\eta e^{-\eta(t-u)}$, the process will then have the Markovian property (Daley and Vere-Jones 2003) and become mathematically tractable. This model can be imbedded into the hidden Markov model framework with restriction of the intensity function to stepwise decay for reasons of computational feasibility (cf. Wang 2010). Thus for a point process N with occurrence times $t_0 = 0, t_1, \ldots, t_n$, the intensity of the process changes only at the times $\lim_{\epsilon \downarrow 0}(t_k + \epsilon)$, i.e., for $t_k < t < t_{k+1}$, $\lambda^*(t) = \lambda^*(t_{k+1})$, and $\lambda^*(t_{k+1}) = \lambda + \nu\eta \sum_{t_j < t_k} e^{-\eta(t_k - t_j)}$. The general conditional intensity function for the process can be written as

$$\lambda^*(t) = \lambda + \nu\eta \sum_{t_j < \max\{t_l : t_l < t\}} e^{-\eta(\max\{t_l : t_l < t\} - t_j)}. \tag{1}$$

We will call this restricted process as a Hawkes process with stepwise decay.

Consider a Hawkes process with stepwise decay the parameters of which vary according to an r-state irreducible Markov process. Denote the infinitesimal generator of the underlying Markov process $\{Y(t)\}$ by $Q = (q_{ij})_{r\times r}$ and $q_i = -q_{ii}$, $i = 1, \ldots, r$. The observed Hawkes process with stepwise decay is characterized by $\Lambda^*(t)$, an $r \times r$ diagonal matrix with diagonal elements $\lambda_1^*(t), \ldots, \lambda_r^*(t)$, where

$$\lambda_i^*(t) = \lambda_i + \nu_i\eta_i \sum_{t_j < \max\{t_l : t_l < t\}} e^{-\eta_i(\max\{t_l : t_l < t\} - t_j)}, \tag{2}$$

t_k is the occurrence time of the kth event, and λ_i, ν_i and η_i, $i = 1, \ldots, r$, are parameters. This process is said to be in state i, $1 \leq i \leq r$, when the underlying Markov process is in state i. When this process is in state i at time t, events occur according to a Hawkes process with stepwise decay rate $\lambda_i^*(t)$ as defined in (2). This process we will call a Markov-modulated Hawkes process with stepwise decay (MMHPSD) of order r. Denote the inter-event times of this process by $X_1 = T_1 - T_0, \ldots, X_n = T_n - T_{n-1}$. Let Y_0 be the state of the Markov process having generator Q at time $t = 0$, and let $X_0 = 0$. The kth event of the MMHPSD is associated with the corresponding state Y_k of the underlying Markov process as well as the time X_k, $k \geq 1$, between the $(k-1)$st and the kth event.

The state of the underlying Markov process at time t is $Y(t)$. Let $N_n(u)$ denote the number of events in the MMHPSD in $(t_{n-1}, t_{n-1} + u]$, where t_n denotes the nth

event time. In order to obtain the transition probability matrix, we need the following Lemma. See the Appendix for the proof.

Lemma 1 *Given the history* $\mathcal{H}_{t_{n-1}}$*, the transition probability without arrival,* $H_{ij}^{(n)}(u) = P\{Y(t_{n-1}+u) = j, N_n(u) = 0 \mid Y(t_{n-1}) = i, \mathcal{H}_{t_{n-1}}\}$*, can be expressed as* $H^{(n)}(u) = \exp\{(Q - \Lambda^*(t_n))u\}$ *for* $u \geq 0$*, where* $H^{(n)}(u) = \{H_{ij}^{(n)}(u)\}$.

The transition probability matrix of the MMHPSD is then given by $F^{(n)}(x) = \int_0^x H^{(n)}(u)\Lambda^*(t_{n-1}+u)\mathrm{d}u$, where the elements are $F_{ij}^{(n)}(x) = P(Y_n = j, X_n \leq x \mid Y_{n-1} = i, \mathcal{H}_{t_{n-1}})$. Thus, the transition density matrix is $f^{(n)}(x) = \exp\{(Q - \Lambda^*(t_n))x\}\Lambda^*(t_n)$. Let $\lambda = (\lambda_1, \ldots, \lambda_r)$, $\nu = (\nu_1, \ldots, \nu_r)$, and $\eta = (\eta_1, \ldots, \eta_r)$. Denote $\Theta = \{Q; \lambda, \nu, \eta\}$. Let $\pi = (\pi_1, \ldots, \pi_r)$ denote the initial distribution vector of the Markov process. If $x_1, \ldots, x_n$ are the observed inter-event times, then the likelihood of the parameter Θ is $\mathcal{L}(\Theta; x_1, \ldots, x_n) = \pi \left\{\prod_{i=1}^n f^{(i)}(x_i; \Theta)\right\} \mathbf{1}$, where $\mathbf{1}$ is an $r \times 1$ vector of ones.

Similar to the standard discrete time HMM problem, if we define the forward probability as $\alpha_t(i) = P\{T_1 = t_1, \ldots, T_{N(t)} = t_{N(t)}, 0 < t_{N(t)} \leq t < t_{N(t)+1}, Y(t) = i\}$, and the backward probability as $\beta_t(j) = P\{T_{N(t)+1} = t_{N(t)+1}, \ldots, T_n = t_n, t_{N(t)} < t \leq t_{N(t)+1} < t_n, Y(t) = j\}$, then the likelihood can be expressed as $\mathcal{L} = \sum_{i=1}^r \alpha_t(i)\beta_t(i)$.

2.1 Comparison with the ETAS model

The Epidemic Type Aftershock Sequence (ETAS; Ogata 1988) model and its extensions (Ogata 1998 and references therein) are nowadays almost invariably used to investigate the characteristics of aftershock sequences. Hence, we need to compare our formulation with that of the ETAS model. The conditional intensity function for the ETAS model is given by $\lambda(t \mid \mathcal{H}_t) = \mu + \sum_{i:t_i<t} e^{\alpha(M_i - M_0)} K(t - t_i + c)^{-p}$, where μ is the occurrence rate for the background seismic activity; the parameter c is a time delay constant; the parameter p characterizes the aftershock decay rate; the parameter α provides a measure of the power of a shock in generating its aftershocks; and the parameter K can be explained as the productivity of events dependent on the magnitude threshold M_0.

The main differences between the ETAS model and the Hawkes process are in the term involving magnitude and in the decay rate term. The ETAS model assumes that for each event, both the time and magnitude of the event influence the intensity. However, the conditional intensity function of the Hawkes process is time-dependent only. Although the decay rate of the ETAS model follows a power law, whereas that of the Hawkes process is an exponential decay, these are per ancestor event, and the actual decay rates, as we see later, appear rather different. For very small and very large delays $t - t_i$, the decay rate term in the ETAS model is always larger, implying a higher seismicity rate immediately following an event. Moreover, in the ETAS model, this is multiplied by the exponential of the magnitude (which is always larger than or equal to 1). The relationship between ν in the Hawkes process and K in the ETAS model is not as obvious. Assuming that the distribution of earthquake

magnitudes M follows the Gutenberg–Richter distribution $b\log(10)10^{-b(M-M_0)}$, the average number of offspring (triggered per immigrant) predicted by the ETAS model is $a_E = n_0\int_0^{+\infty}(t+1)^{-p}\mathrm{d}t$, where $\alpha^* = \alpha/\log(10)$ and $n_0 = Kb/c^{p-1}(b-\alpha^*)$ (cf. Helmstetter and Sornette 2002). For the Hawkes process, it is $\int_0^{+\infty}\nu\eta e^{-\eta t}\mathrm{d}t = \nu$. Given the same number of offspring, we thus have $\nu = Kc^{-(p-1)}b(b-\alpha^*)^{-1}\int_0^{+\infty}(t+1)^{-p}\mathrm{d}t$. For example, for $b = 1$, $\alpha = 1.468$, i.e., $\alpha^* = 0.638$, restraining $0 < \nu = a_E < 1$ ensures that each immigrant has a finite total number of offspring with probability one. However, in the ETAS model, there is no such constraint on the parameter K, i.e., there are cases when $a_E \geq 1$ due to the self-exciting nature of the process, which compounds with the magnitude term.

3 An EM algorithm for parameter estimation

In order to estimate the parameters, we will construct a version of the EM algorithm.

3.1 The complete likelihood

First, we formulate the complete likelihood of an MMHPSD process. Suppose that the occurrence times $t_0 = 0, t_1, \ldots, t_n = T$ from an MMHPSD of order r are observed. Let $x_1 = t_1 - t_0, \ldots, x_n = t_n - t_{n-1}$, which are the inter-event times. Assume that the hidden Markov process $\{Y(t)\}$ has transitions at the time-points $0 < u_1 < u_2 < \cdots < u_m < T$. Define $u_0 = 0$ and $u_{m+1} = T$, and write $I_k = [u_{k-1}, u_k)$, and $\Delta u_k = u_k - u_{k-1}$, for $1 \leq k \leq m+1$. Moreover, denote the state of $\{Y(t)\}$ during I_k by s_k, let z_k be the number of events in I_k (do not count the event at $t = 0$), and let $N(t) = \#\{k : 0 < k \leq n, t_k \leq t\}$, i.e., the number of events up to time t. Let $\rho_0 = 0$ and $\rho_k = z_1 + \cdots + z_k$, for $1 \leq k \leq m+1$. Then the complete likelihood of Θ can be written as

$$\begin{aligned}\mathcal{L}^c = \pi_{s_1}&\left\{\prod_{k=1}^{m} q_{s_k}e^{-q_{s_k}\Delta u_k}\times\frac{q_{s_k,s_{k+1}}}{q_{s_k}}\right\}e^{-q_{s_{m+1}}\Delta u_{m+1}}\\ &\times\left\{\prod_{k=1}^{m+1}\prod_{i=1}^{z_k}\lambda^*_{s_k}(t_{\rho_{k-1}+i})\exp\left\{-\int_{t_{\rho_{k-1}+i-1}}^{t_{\rho_{k-1}+i}}\lambda^*_{s_k}(t)\mathrm{d}t\right\}\right\},\end{aligned}\tag{3}$$

where $q_k = -q_{kk}$. Taking the logarithm of the complete likelihood (3) and simplifying, it follows that $\log\mathcal{L}^c = L_1 + L_2$, where

$$L_1 = \sum_{i=1}^{r} I\{Y(0) = i\}\log\pi_i - \sum_{i=1}^{r} D_i q_i + \sum_{i=1}^{r}\sum_{\substack{j=1\\ j\neq i}}^{r} w_{ij}\log q_{ij}\tag{4}$$

and

$$L_2 = \sum_{i=1}^{r}\sum_{k=1}^{n}\log\lambda^*_i(t_k)I\{Y(t_k) = i\} - \sum_{i=1}^{r}\int_0^T\lambda^*_i(t)I\{Y(t) = i\}\mathrm{d}t,\tag{5}$$

for $i \neq j$, $w_{ij} = \#\{k : 1 \leq k \leq m, s_k = i, s_{k+1} = j\} = \#\{t : 0 < t \leq T, Y(t-) = i, Y(t) = j\}$ is the number of jumps of $Y(t)$ from State i to State j in $[0, T]$, $D_i = \sum_{\{k:1\leq k\leq m+1, s_k=i\}} \Delta u_k = \int_0^T I\{Y(t) = i\}\mathrm{d}t$ is the time $\{Y(t)\}$ spends in state i during $[0, T]$.

3.2 The EM algorithm

The EM algorithm can now be used to estimate the parameters. If Θ_0 is a given parameter estimate, then the E-step of the EM algorithm is to calculate the expectation $\mathcal{Q}(\Theta; \Theta_0) = \mathbb{E}_{Y,\Theta_0}[\log \mathcal{L}^c(\Theta; N(t), Y) \mid N(t)]$ with respect to Y and Θ_0. The M-step then maximizes the $\mathcal{Q}$ function to obtain the new estimate $\widehat{\Theta} = \arg\max_\Theta \mathcal{Q}(\Theta; \Theta_0)$. The expected complete log-likelihood $\mathcal{Q}(\Theta; \Theta_0)$ is obtained by taking the expectation of the complete likelihood over Y at the current parameter estimate Θ_0 conditional on the observed data $\{N(s), 0 \leq s \leq T\}$.

The parameters involved in Eq. (4), the initial distribution π of the Markov process and the infinitesimal generator Q, can be estimated using a similar approach as outlined for the MMPP in Rydén (1996). The conditional expectation of the statistics D_i and w_{ij} are

$$\widehat{D}_i = \mathbb{E}\left\{ \int_0^T I\{Y(t) = i\}\mathrm{d}t \;\middle|\; N(s), 0 \leq s \leq T \right\} = \int_0^T \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^r \alpha_s(j)\beta_s(j)}\mathrm{d}t \tag{6}$$

and

$$\begin{aligned} \widehat{w}_{ij} &= \mathbb{E}\left\{\#\{t : 0 < t \leq T, Y(t-) = i, Y(t) = j\} \mid N(s), 0 \leq s \leq T\right\} \\ &= \int_0^T \frac{\alpha_t(i)q_{ij}\beta_t(j)}{\sum_{k=1}^r \alpha_s(k)\beta_s(k)}\mathrm{d}t. \end{aligned} \tag{7}$$

The remaining problem is to estimate the parameters in the intensity function of the Hawkes process with stepwise decay rate. This can be done by maximizing the conditional expectation of L_2 in Eq. (5) given the observations $\{N(s), 0 \leq s \leq T\}$. This conditional expectation can be expressed as

$$\begin{aligned} \mathcal{Q}_2(\Theta; \Theta_0) &= \mathbb{E}\{L_2 \mid N(s), 0 \leq s \leq T\} \\ &= \sum_{i=1}^r \sum_{k=1}^n \left\{ \frac{\alpha_{t_k}(i)\beta_{t_k}(i)}{\mathcal{L}} \log\left(\lambda_i^*(t_k)\right) - \int_{t_{k-1}}^{t_k} \frac{\alpha_t(i)\beta_t(i)}{\mathcal{L}} \lambda_i^*(t_k)\mathrm{d}t \right\}. \end{aligned} \tag{8}$$

In order to get the estimates of the parameters involved in the intensity part, we will calculate the forward and backward probabilities in the E-step, and then carry out numerical optimization to maximize $\mathcal{Q}_2(\Theta; \Theta_0)$ in Eq. (8) in the M-step.

3.3 Implementation

We will now provide the implementation steps to carry out the parameter estimation. For the scaling of forward–backward recursions and the evaluation of matrix exponential integrals, to avoid a lengthy explanation, the readers are referred to Rydén (1996), Roberts et al. (2006) and Van Loan (1978) for the version corresponding to the MMPP. To estimate the parameters via the EM algorithm, we provide the following implementation algorithm.

Algorithm 1 Given the initial values $\lambda_0 = (\lambda_{01}, \ldots, \lambda_{0r})$, $\nu_0 = (\nu_{01}, \ldots, \nu_{0r})$, and $\eta_0 = (\eta_{01}, \ldots, \eta_{0r})$, $Q_0 = (q_{0ij})_{r\times r}$ and $\pi_0 = (\pi_{01}, \ldots, \pi_{0r})$, and given the observed occurrence times $t_0 = 0, t_1, \ldots, t_n = T$ from an MMHPSD of order r, with inter-event times $x_1 = t_1 - t_0, \ldots, x_n = t_n - t_{n-1}$, the EM algorithm for the MMHPSD can be carried out as follows.

(1) Let $L(0) = \pi_0$, and for $k = 1, \ldots, n$, let $L(k) = L(k-1) f^{(k)}(x_k)/L(k-1) f^{(k)}(x_k)\mathbf{1}$.
(2) Let $R(n+1) = \mathbf{1}$, and for $k = n, \ldots, 1$, let $R(k) = f^{(k)}(x_k)R(k+1)/L(k-1) f^{(k)}(x_k)\mathbf{1}$.
(3) For $k = 1, \ldots, n$, let

$$\mathcal{C}_k = \begin{pmatrix} Q - \Lambda(t_k) & \Lambda^*(t_k)R(k+1)L(k-1) \\ 0 & Q - \Lambda(t_k) \end{pmatrix}$$

and calculate the matrix $e^{\mathcal{C}_k x_k}$. Then set $\mathcal{I}_k$ as the $r \times r$ upper-right block of this matrix.
(4) Calculate the conditional expectations (6) and (7), $w = Q \odot \sum_{k=1}^{n} \mathcal{I}_k^T / c_k$ and $D_i = w_{ii}/q_{0ii}$, where $\odot$ denotes element-by-element multiplication of the two matrices.
(5) We have the new estimates $\widehat{q}_{ij} = w_{ij}/D_i$, for $i, j = 1, \ldots, r, i \neq j$, $\widehat{q}_i = \sum_{j=1, j\neq i}^{r} \widehat{q}_{ij}$, for $i = 1, \ldots, r$, and $\widehat{\pi}_i = \pi_{0i} e_i^T R(1)$, for $i = 1, \ldots, r$, where e_i is an $r \times 1$ vector, the elements of which are zeros except the ith entry which is 1.
(6) For $i = 1, \ldots, r$ and $k = 1, \ldots, n$, set $B_{ik} = L(k)e_i e_i^T R(k+1) = L_i(k)R_i(k+1)$.
(7) Numerically optimize $\mathcal{Q}_2(\Theta; \Theta_0) = \sum_{i=1}^{r}\sum_{k=1}^{n} B_{ik} \log(\lambda_i^*(t_k)) - \sum_{i=1}^{r}\sum_{k=1}^{n} \mathcal{I}_{ik}\lambda_i^*(t_k)/c_k$, and obtain estimates for the parameters λ, ν, and η.
(8) Calculate the log likelihood $LL_{old} = \sum_{k=1}^{n} \log(L(k-1) f^{(k)}(x_k)\mathbf{1})$. Then substitute $(\widehat{Q}, \widehat{\mathbf{\Lambda}}, \widehat{\nu}, \widehat{\eta})$ in place of $(Q_0, \mathbf{\Lambda}_0, \nu_0, \eta_0)$ in step (1) with $L(0) = \widehat{\pi}$, and calculate the 'new' log likelihood $LL_{new} = \sum_{k=1}^{n} \log(L(k-1) f^{(k)}(x_k)\mathbf{1})$.
(9) Substitute $(\widehat{Q}, \widehat{\mathbf{\Lambda}}, \widehat{\nu}, \widehat{\eta})$ in place of $(Q_0, \Lambda_0, \nu_0, \eta_0)$ and repeat step (1) to step (8) until the difference in the log likelihoods becomes less than the terminative condition.

Note that the parameter estimation of $\mathcal{Q}_2$ in Step (7) requires some numerical optimization technique, such as the standard nonlinear optimization technique suggested by Fletcher and Powell (1963). To perform this, we need the gradients and Hessian of $\mathcal{Q}_2$, which are the partial first- and second-order derivatives of the function $\mathcal{Q}_2$ with

respect to the parameters. See the Appendix for the detailed gradients and Hessian calculations.

4 Goodness-of-fit

4.1 Estimated intensity function of the observed process

After fitting an MMHPSD to a set of observed data, the probability of the hidden state occupying a specific state at time t given the entire observed process can provide a clear illustration of the underlying process. As discussed in Zucchini and Guttorp (1991) and MacDonald and Zucchini (1997), this probability can be directly calculated using the forward and backward probabilities with the estimated parameters,

$$P(S_t = i \mid \mathcal{H}_T) = \frac{P(T_1 = t_1, \ldots, T_n = t_n, Y(t) = i)}{P(T_1 = t_1, \ldots, T_n = t_n)} = \frac{\alpha_t(i)\beta_t(i)}{\mathcal{L}}. \tag{9}$$

The estimated intensity function can be obtained by

$$\widehat{\lambda}^*(t) = \sum_{i=1}^{r} \left(\widehat{\lambda}_i + \widehat{\nu}_i \widehat{\eta}_i \sum_{t_j < \max\{t_l : t_l < t\}} e^{-\widehat{\eta}_i (\max\{t_l : t_l < t\} - t_j)} \right) \frac{\alpha_t(i)\beta_t(i)}{\mathcal{L}}. \tag{10}$$

We will use this as the basis of a goodness-of-fit test.

Algorithm 2 After fitting an MMHPSD to the data, the estimates of the parameters $\widehat{\lambda} = (\widehat{\lambda}_1, \ldots, \widehat{\lambda}_r)$, $\widehat{\nu} = (\widehat{\nu}_1, \ldots, \widehat{\nu}_r)$, and $\widehat{\eta} = (\widehat{\eta}_1, \ldots, \widehat{\eta}_r)$, $\widehat{Q} = (\widehat{q}_{ij})_{r \times r}$ and $\widehat{\pi} = (\widehat{\pi}_1, \ldots, \widehat{\pi}_r)$ are obtained. Given the observed occurrence times $t_0 = 0, t_1, \ldots, t_n = T$ with inter-event times $x_1 = t_1 - t_0, \ldots, x_n = t_n - t_{n-1}$, the procedure of estimating the probability of the hidden state occupying a specific state at time t and the intensity function is as follows.

(1) Let $L(0) = \widehat{\pi}$, and for $k = 1, \ldots, n$, let $L(k) = L(k-1) f^{(k)}(x_k)/L(k-1) f^{(k)}(x_k)\mathbf{1}$.
(2) Let $R(n+1) = \mathbf{1}$, and for $k = n, \ldots, 1$, let $R(k) = f^{(k)}(x_k)R(k+1)/L(k-1) f^{(k)}(x_k)\mathbf{1}$.
(3) For $i = 1, \ldots, r$ and $t_{k-1} \le t < t_k$, set $\widehat{p}_i(t) = L(k-1)\exp\{(\widehat{Q} - \widehat{\Lambda}^*(t_k))(t - t_{k-1})\} e_i e_i^T \exp\{(\widehat{Q} - \widehat{\Lambda}^*(t_k))(t_k - t)\}\widehat{\Lambda}^*(t_k)R(k+1)/c_k$. Thus $\widehat{p}_i(t)$ is the estimated probability of the hidden state occupying state i at time t for (9).
(4) For $i = 1, \ldots, r$ and $t_{k-1} \le t < t_k$, set the estimated intensity function (10) as $\widehat{\lambda}^*(t) = \sum_{i=1}^{r} (\widehat{\lambda}_i + \widehat{\nu}_i \widehat{\eta}_i \sum_{t_j < \max\{t_l : t_l < t\}} e^{-\widehat{\eta}_i (\max\{t_l : t_l < t\} - t_j)}) \widehat{p}_i(t)$.

4.2 Residual analysis

Residual analysis (see e.g., Ogata 1988; Bebbington and Harte 2001) can be used to assess the goodness-of-fit of the MMHPSD. According to Theorem 7.4.I in Daley and Vere-Jones (2003), if a point process in time with intensity function $\lambda(t)$ is

rescaled using the random time change $\tau = \Lambda(t) = \int_0^t \lambda(s)\mathrm{d}s$, then the rescaled time τ is a stationary Poisson process with unit rate. For an observed point process with occurrence times $t_1, t_2, \ldots, t_n$, the true model for the data is usually unknown. If we fit a model to the data and rescale the points as given above, then the fitted model is a good approximation to the true model if the residual point process

$$\tau_i = \Lambda(t_i), \tag{11}$$

is indistinguishable from a stationary Poisson process with unit rate. Residual analysis differs in intent from other model selection criteria such as the log likelihood, Akaike Information Criterion (AIC; Akaike 1974) and Bayesian information criterion (BIC; Schwarz 1978) in that it is used to identify systematic deviation of the data from the fitted model. Therefore, it provides evidence of a good fit. The latter three criteria, however, only suggest which model is relatively better and do not guarantee that the model is a good approximation of the true model. Standard tests can be used to investigate whether the residual point process is a stationary Poisson process with unit rate.

5 Simulation

5.1 Simulation algorithm

To see how well the parameter estimation method works, an MMHPSD process can be simulated, and then the model can be fitted to the simulated data set to get the parameter estimates. By comparing the estimated parameters with the true parameters, one can see how the parameter estimation method performs.

Algorithm 3 Given the initial state $y_1 = j$, the parameters Q, λ_l, ν_l and η_l, $l = 1, \ldots, r$, and the history data set $\mathcal{H} \subset \{t \leq 0\}$, the following steps can be carried out to generate n events from an MMHPSD process.

1. Set $i = s = 1$ and $t_i = t_s = 0$.
2. Use $r_i = q_j + \lambda_j + \nu_j \eta_j \sum_{t_l < t_i} e^{-\eta(t_i - t_l)}$ as the rate and generate an inter-event time τ_i from the exponential distribution.
3. Set $t_{s+1} = t_s + \tau_i$. Generate a uniform random variable $U \in (0, 1)$.
   - If $U > q_j / r_i$, then add $t_{i+1} = t_{s+1}$ into the history. Set $i = i + 1$, $s = s + 1$ and go to Step 2.
   - If $U \leq q_j / r_i$, then this point is a state transition point. Use $(q_{jk}/q_j)_{1 \leq k \leq r}$ to generate the next state y_2. Set $j = y_2$, $s = s + 1$ and go to Step 2.

   If $i = n$, then stop.
4. The sequence of times simulated is the generated MMHPSD process.

As we will see in the example of the Landers earthquake sequences, the magnitude is state dependent. If one wants to simulate earthquake magnitudes in the MMHPSD, we refer to Jaumé and Bebbington (2004) for discussion of ways in which a state-dependent magnitude distribution might be parameterized.

5.2 Consistency of the parameter estimation procedure

We now investigate the performance of the parameter estimation algorithm proposed in Sect. 3. The ETAS model has been well studied and has often been used to simulate earthquake catalogues. In order to get some reasonable parameters of the MMHPSD from an analogue earthquake sequence, we therefore use the ETAS model to simulate a long homogeneous earthquake catalogue and then fit the MMHPSD model with two states to the simulated sequence. We can then use these estimated parameters to simulate MMHPSD sequences of varying lengths and examine the parameter estimation method for consistency before we use this model to analyze a real earthquake catalogue.

Zhuang (2000) used the ETAS model to study the seismicity of the earthquake sequence which occurred off Cape Palliser at the southeastern tip of North Island, New Zealand from January 1, 1978 to May 31, 1996. The whole period was divided visually according to the magnitude–time plot into four stages: early background period, relatively quiescent period, main shock and aftershock sequence, and an active period of post-aftershocks.

We use the same study area as in Zhuang (2000), selecting 920 events with minimum magnitude 2.0 from the SSLib package (http://homepages.paradise.net.nz/david.harte/SSLib/, last accessed on September 22, 2009) within the cylinder centered at latitude 41.686S and longitude 175.508E with a radius of 36 km and a depth of 40 km from January 1, 1978 to May 31, 1996. Note that the catalogue has changed slightly since Zhuang's work appeared. An ETAS model fitted to this sequence produced maximum likelihood estimates of the parameters, $\hat{\mu} = 0.025$, $\hat{K} = 11.002$, $\hat{\alpha} = 1.468$, $\hat{c} = 0.004$ and $\hat{p} = 1.127$. These estimated parameters were then used to simulate a sequence of 3,000 events from the ETAS model.

Next, the MMHPSD with two states is fitted to the simulated series of events. The estimated parameters are $\widehat{\lambda}_1 = 0.003$, $\widehat{\lambda}_2 = 6.324$, $\widehat{\nu}_1 = 0.192$, $\widehat{\nu}_2 = 0.583$, $\widehat{\eta}_1 = 0.044$, $\widehat{\eta}_2 = 19.502$, $\widehat{q}_1 = 0.083$, and $\widehat{q}_2 = 8.036$. These estimated parameters are then used to simulate four groups of MMHPSD events. Each group consists of 100 sequences. Each sequence has 500 events in Group 1, 1,000 events in Group 2, 2,000 events in Group 3, and 5,000 events in Group 4. Then the MMHPSD is refitted to each of the simulated sequences, and the parameters are estimated.

The boxplot of the estimated parameters for each group is shown in Fig. 1. We see that the mean of the parameter estimates of each of the parameters is very close to the true value, and the variation in the parameter estimates decreases with the number of events, indicating consistency.

6 Example: the Landers–Hector Mine series of earthquakes

6.1 Earthquake data around Landers

The magnitude 7.3 Landers earthquake on June 28, 1992 is one of the most significant, and the most studied, earthquakes in Southern California. It was preceded by the Joshua Tree earthquake of magnitude 6.1 on April 22, 1992, and followed by the

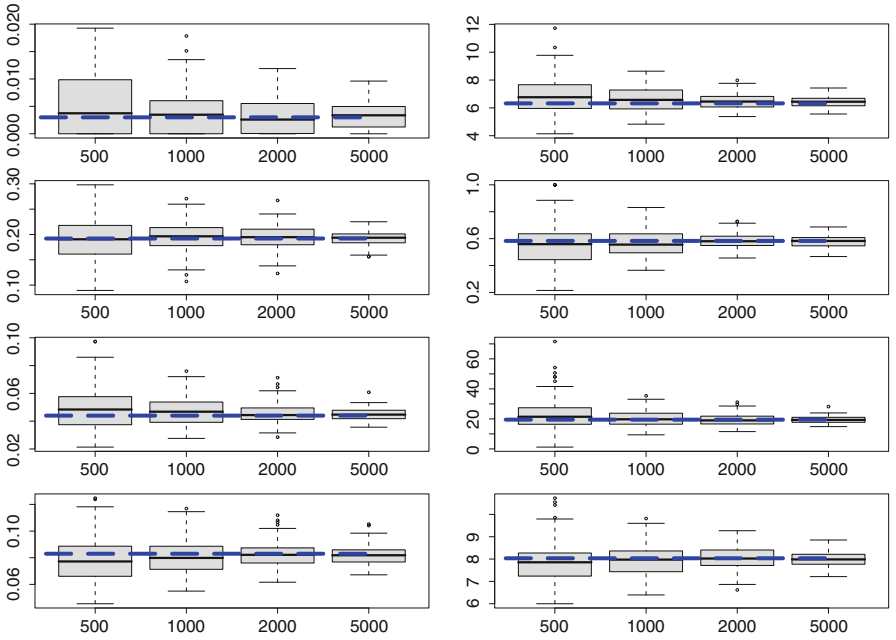

Fig. 1 Boxplot of the estimated parameters $\hat{\lambda}_1$, $\hat{\lambda}_2$, $\hat{\nu}_1$, $\hat{\nu}_2$, $\hat{\eta}_1$, $\hat{\eta}_2$, $\hat{q}_1$ and $\hat{q}_2$ for each group. The true parameters are $\lambda_1 = 0.003$, $\lambda_2 = 6.324$, $\nu_1 = 0.192$, $\nu_2 = 0.583$, $\eta_1 = 0.044$, $\eta_2 = 19.502$, $q_1 = 0.083$ and $q_2 = 8.036$ which are indicated using *dashed lines*

magnitude 6.4 Big Bear earthquake 3 h later, about 35 km west of the Landers epicenter. The Landers aftershock sequence contains more than ten large aftershocks of magnitude 5.0 or above. The 1999 magnitude 7.1 Hector Mine earthquake occurred about 30 km northeast of the Landers source region 7 years later.

The seismicity in the region around the above events will be studied. In order to choose the magnitude threshold, the cutoff should ensure that the earthquakes above this magnitude follow the Gutenberg–Richter (G–R) law because this is assumed in the ETAS formulation. While the series contains multiple seismic cycles, the number of events should not be too large, otherwise the parameter estimation will be very time-consuming. Thus, we select earthquakes between latitude 33.8N and 34.8N, and between longitude 117.1W and 116.1W, with minimum magnitude 3 from January 1, 1976 to December 31, 2008, which are complete according to the G–R law.

Some 2,431 events are thus selected, including the magnitude 7.3 Landers earthquake which occurred on June 28, 1992 and the magnitude 7.1 Hector Mine earthquake on October 16, 1999. The data are from the SCSN catalogue which is available from the southern California earthquake data center web site (http://www.data.scec.org/index.html, last accessed on September 19, 2009). The locations of the earthquakes are plotted in Fig. 2. Figure 3 shows the cumulative number of earthquakes in the study area during the 33 years. We can clearly see several jumps which indicate seismicity rate changes. We will fit the MMHPSD to this set of data and discuss the

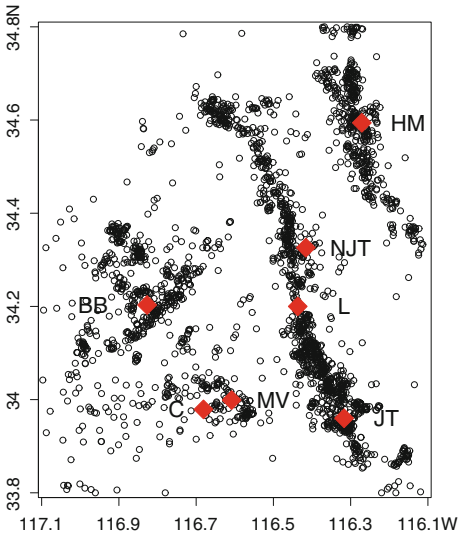

Fig. 2 Location map of the earthquakes with minimum magnitude 3.0 in the area between latitude 33.8N to 34.8N, and longitude 117.1W to 116.1W from January 1, 1976 to December 31, 2008. The *large symbols* (*filled diamonds*) indicate the above-mentioned earthquakes, the March 15, 1979 $M = 5.2$ earthquake (*NJT*), the July 8, 1986 $M = 5.7$ earthquake (*MV*), the December 15, 1988 $M = 5.0$ earthquake (*C*), the April 22, 1992 $M = 6.1$ Joshua Tree earthquake (*JT*), the June 28, 1992 $M = 7.3$ Landers earthquake (*L*), the June 28, 1992 $M = 6.4$ Big Bear (*BB*) and the October 16, 1999 $M = 7.1$ Hector Mine earthquake (*HM*)

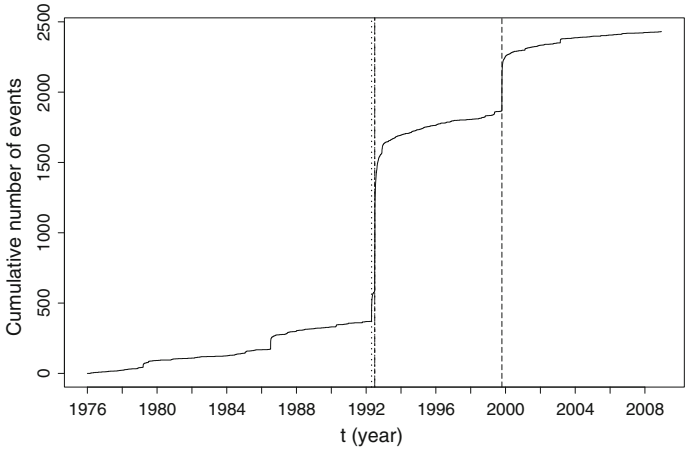

Fig. 3 Cumulative number of earthquakes with minimum magnitude 3.0 in the area between latitude 33.8N to 34.8N, and longitude 117.1W to 116.1W from January 1, 1976 to December 31, 2008. The *vertical lines* from *left* to *right*, respectively, indicate the occurrence times of the magnitude 6.1 Joshua Tree earthquake on April 22, 1992, the magnitude 7.3 Landers earthquake on June 28, 1992, the magnitude 6.4 Big Bear earthquake on June 28, 1992, and the magnitude 7.1 Hector Mine earthquake on October 16, 1999

features this model captures. For comparison, we also examine the data using the ETAS model with magnitude threshold $M_0 = 3.0$. The occurrence times are expressed in days.

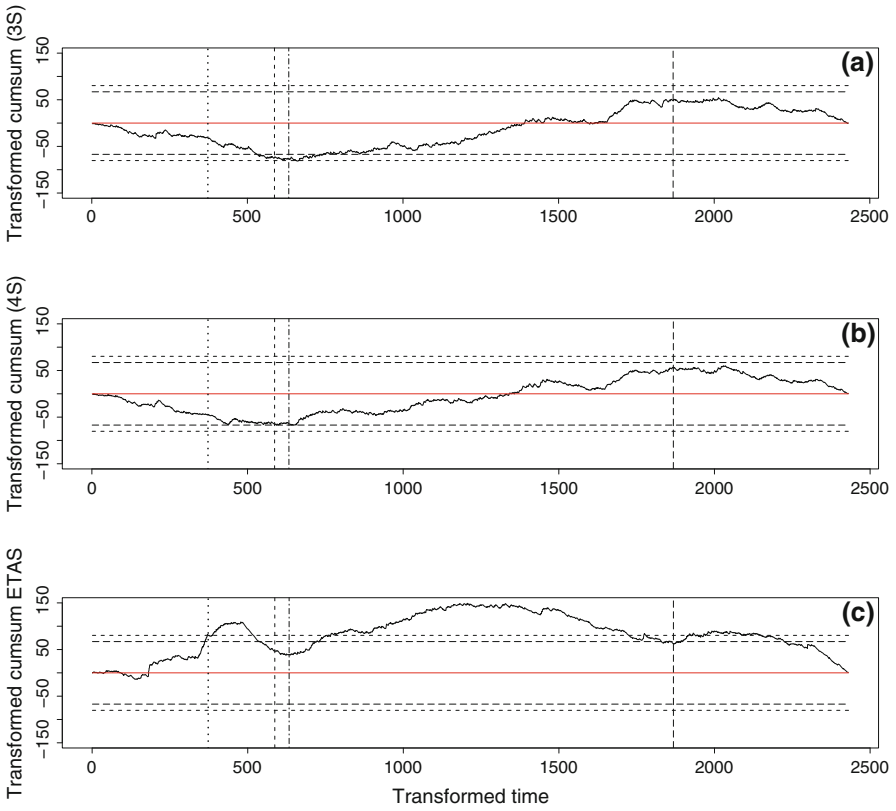

Fig. 4 The deviation of the cumulative number of events in the residual process from the null hypothesis of stationarity for **a** the MMHPSD with three states, **b** the MMHPSD with four states, and **c** the ETAS model, fitted to the earthquakes around Landers. The *solid straight line* is the theoretical behavior under the null hypothesis. The *dashed lines* indicate the two-sided 95% (*long dash*) and 99% (*dashed*) confidence limits of the Kolmogorov–Smirnov statistic. The *vertical lines* from *left* to *right* are the same as in Fig. 3

6.2 Exploratory data analysis using the MMHPSDs

Starting from a two-state MMHPSD, we add one state at a time until the residual point process of the current model is a stationary Poisson process with unit rate. The result will thus be a good approximation of the true model. For each model, the transformed time τ_i in (11) is calculated, and the cumulative number of the residual process versus the transformed time $\{\tau_i\}$ is plotted. For the fitted MMHPSDs with two and three hidden states, the 95% confidence limits are exceeded not long before the Landers earthquake, and the 99% limits are further exceeded after the Big Bear earthquake (see Fig. 4a for the plot of the three-state MMHPSD). Note that, in order to better see the deviation of the calculated curves from the theoretical one, we subtracted the latter from the former. The residual process of the MMHPSD with four hidden states, however, is well approximated by the standard stationary Poisson process, as shown in Fig. 4b.

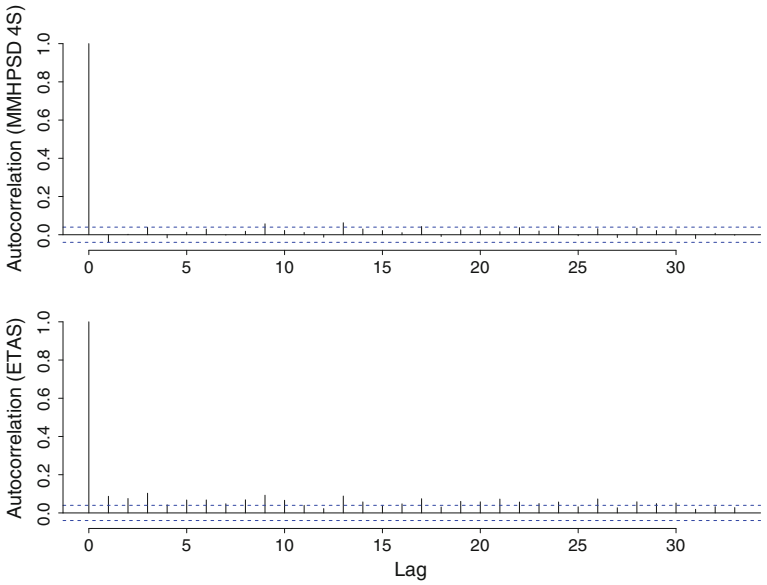

Fig. 5 The autocorrelation plot for the U_k series (*top* MMHPSD with four states; *bottom* ETAS), with *dashed lines* indicating the 95% confidence intervals

Let $E_k = \tau_k - \tau_{k-1} = \Lambda(t_k) - \Lambda(t_{k-1})$, $k = 1, \ldots, n$, and $U_k = 1 - \exp(-E_k)$. For the MMHPSD with four hidden states, we then test whether the U_k have a uniform distribution on $[0, 1)$. Testing the U_k versus a null hypothesis of uniformity produces a Kolmogorov–Smirnov statistic of 0.0206 ($P = 0.2535$), from which we see that we cannot reject the null hypothesis of uniformity. Therefore, the inter-arrival times appear exponentially distributed.

The t-statistic of the hypothesis test with the null hypothesis that the correlation coefficient between $\log(E_k)$ and $\log(E_{k+1})$ equals zero is -1.4514 with a P value of 0.1468. A scatter plot of U_{k+1} against U_k suggests no particular pattern. The autocorrelation of U_k is shown in Fig. 5, which indicates that except at lag 9 and 13 the autocorrelation between U_k lies within the 95% confidence band. Therefore, there is no evidence that the inter-arrival times are not independent, and hence, the fitted MMHPSD with four hidden states is a good approximation of the true intensity of the data.

The parameters estimated via the EM algorithm for the four state MMHPSD are shown in Table 1. Note that the states are ordered according to the decay rate η. The estimated Q-matrix is

$$\hat{Q} = \begin{pmatrix} -0.0066 & 0.0000 & 0.0064 & 0.0002 \\ 0.0001 & -0.4677 & 0.0002 & 0.4674 \\ 0.0001 & 0.0643 & -6.1022 & 6.0378 \\ 58.0805 & 132.6127 & 121.2158 & -311.9090 \end{pmatrix}, \tag{12}$$

Table 1 Estimated parameters of the four-state MMHPSD fitted to the data around Landers

State	1	2	3	4
$\hat{\lambda}$	0.022	0.000	0.783	154.098
$\hat{\nu}$	0.254	0.545	0.914	0.999
$\hat{\eta}$	0.026	0.521	19.286	188.787

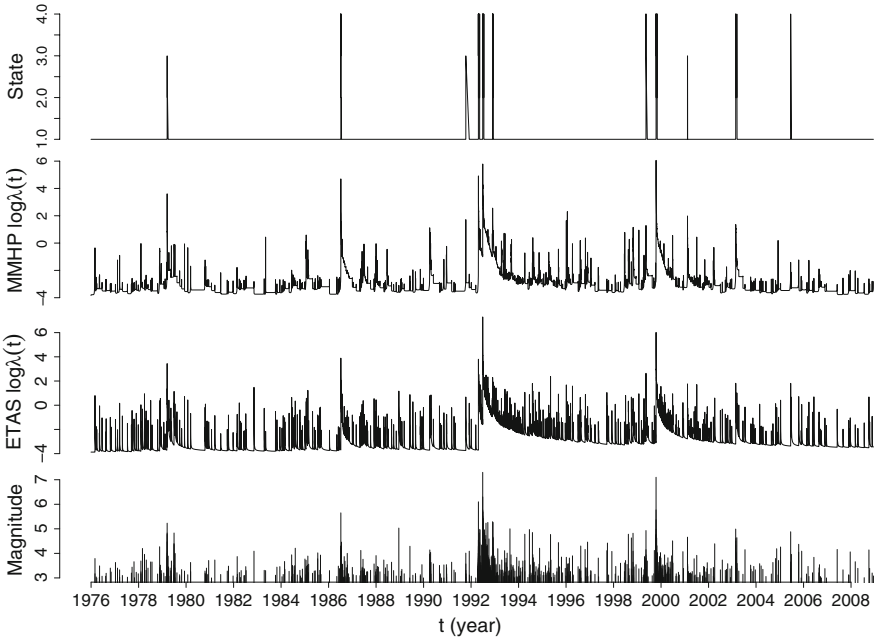

Fig. 6 Illustration of the fitted four-state MMHPSD and the fitted ETAS model. *Top plot* the tracked Viterbi path for the hidden states given the observed sequence; *next two plots* the estimated intensity functions for the fitted four-state MMHPSD and the fitted ETAS model; *bottom* the magnitude–time plot for the Landers earthquakes

with stationary distribution $(0.9652, 0.0315, 0.0032, 0.0001)$. The tracked Viterbi path (see Bebbington 2007) and the estimated intensity function of the four-state MMHPSD are shown in Fig. 6. Note that the residual analysis shows that the MMHPSD with four hidden states is the simplest MMHPSD model which captures the main features of the data. Hence, according to Occam's Razor, it can be concluded that the four-state model is the 'best' in this sense. The log likelihood, AIC (Akaike 1974, 1978; Shibata 1980, 1981) and BIC (Schwarz 1978) values for the fitted MMHPSDs with two, three, and four hidden states are listed in Table 2. Among the three models, the AIC and BIC suggest that the fitted MMHPSD with four hidden states is the best fit. Although the AIC and BIC may improve with additional states, these parameters cannot be justified by the residual analysis, which shows that the data are already explained by the four-state model.

Table 2 The log likelihood (LL), AIC, and BIC for the fitted MMHPSDs, respectively, with two states (MMHP2S), three states (MMHP3S) and four states (MMHP4S), and the ETAS model

	MMHP2S	MMHP3S	MMHP4S	ETAS
LL	1,760.669	1,942.465	2,019.240	2,020.587
AIC	−3,505.338	−3,854.930	−3,990.48	−4,031.174
BIC	−3,458.966	−3,767.983	−3,851.365	−4,002.192

We see from Table 1 and Fig. 6 that State 4 most likely corresponds to the major main shocks and its immediate aftermath. In this state, there is a very high immigration rate λ, indicating that the events are not triggered by other events. The decay rate η is also very high, so these events do not themselves trigger other events, and $\nu \sim 1$, indicating a period of almost constant activity rate. State 3 appears most likely to account for the major aftershocks, while State 2, with a negligible immigration rate and a small decay rate, is the principal aftershock state (cf. Fig. 8). Finally, State 1 has a low immigration rate, a very low decay rate, and a small value of ν. Coupled with the long sojourn time implied by (12), we see that this corresponds to a ground, or quiescent, state.

Further examining the transition matrix in (12), we see that State 4 has a very short sojourn time (approximately 5 min on average), and exits to one of the aftershock states. State 3 appears to possibly have some precursory properties for State 4, commensurate with the Joshua Tree–Landers triggering (Hill et al. 1993, 1995). This is further supported by the preferred transition from State 1 to State 3. In short, our seismic cycle is identified: State 1 (quiescent) – State 3 (precursory/main shock) – [State 4 (main shock) – State 3 (primary aftershocks)] – State 2 (aftershocks) – State 1, with the steps in brackets being optional, so to speak.

Aftershocks are generally considered to exhibit a power-law decay rate (see e.g., Utsu et al. 1995). Figure 7 enlarges the time period of about a year just after the Landers main shock with the corresponding Viterbi path of the hidden states (shown in the top frame) allowing us to clearly see the state transitions and the estimated intensity functions of both the MMHPSD and the ETAS models. Even though the 'per event' decay rate in the former model is exponential, the effect is closer to a power law (a straight line on the log-log axis) than is that of the ETAS model. The properties of mixtures often differ significantly from those of their components (cf. Vere-Jones et al. 2001). In order to see the pattern of state changes better, Fig. 8 presents the data from Fig. 7 in transformed, or residual, time. We can see from Fig. 7 that after approximately 20 days of very high seismicity rate following the Landers main shock, the Viterbi sequence entered State 1 although the seismicity rate remained relatively high due to the large number of contributing ancestor events immediately upstream, and then switched back to the high seismicity regime after approximately another 130 days (about 150 days after Landers main shock) on the appearance of a cluster of large aftershocks. It reentered State 1 after a further 20 days of high seismicity rate (about 170 days after Landers main shock). In comparison to the temporal study of the Landers aftershock sequence by Ogata et al. (2003), which detected a first change

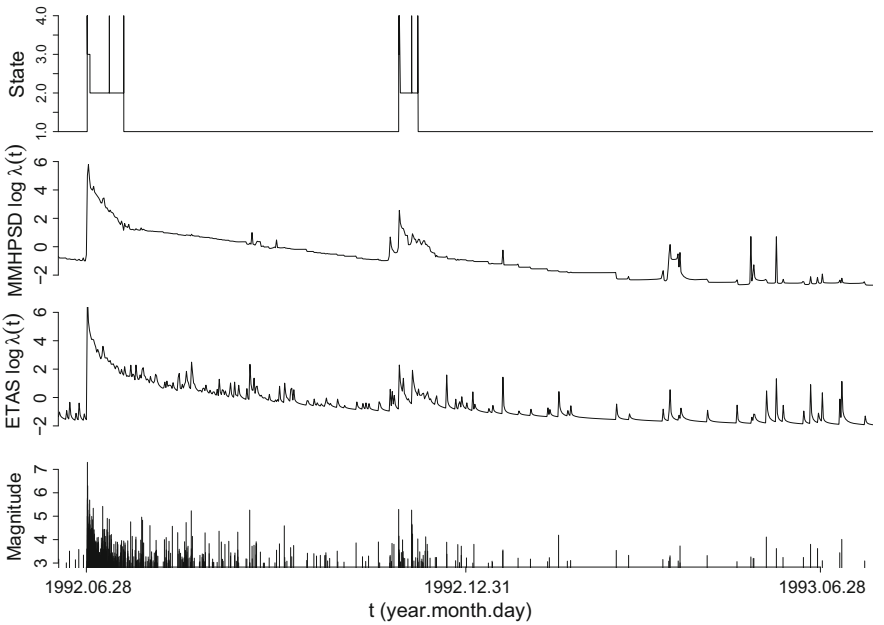

Fig. 7 Enlarged plots for the time period of about a year just after the Landers main shock with the x-axis in log scale. *Top* the tracked Viterbi path for the hidden states given the observed sequence; *next two plots* the estimated intensity functions for the fitted four-state MMHPSD and the fitted ETAS model; *bottom* the magnitude–time plot for the Landers earthquakes

point at about 190 days after the Landers main shock, the Viterbi path of the MMHPSD suggests the possibility of two further change points at about 20 and 150 days.

The conditional intensity function of a point process in time can be used to interpret the inter-event times. Larger intensities correspond to shorter inter-event times. From the above MMHPSD analysis of the data, it is not difficult to notice that the inter-event times are highly correlated with magnitude. Take the fitted four-state model for example. State 4 with the highest conditional intensity function captures the feature of the largest earthquake; State 3, which has medium-large intensities, is related to the large-magnitude aftershocks; States 1 and 2 with low intensities correspond to smaller earthquakes. The modelled inter-event times indirectly reflect the magnitude influence. Although we do not include the magnitude effect in our formulation of the conditional intensity function, the fitted model still accounts for the magnitude effect.

6.3 Comparison with the ETAS model

Now let us examine the same data using the ETAS model with magnitude threshold $M_0 = 3.0$. The maximum likelihood estimates of the ETAS parameters are, $\mu = 0.0208$, $K = 1.4217$, $\alpha = 1.6265$, $c = 0.0381$, and $p = 1.2230$. The estimated intensity function as plotted in Fig. 6 shows large spikes for almost every event, whereas the intensity function of the MMHPSD tends to have fewer, and on average lower,

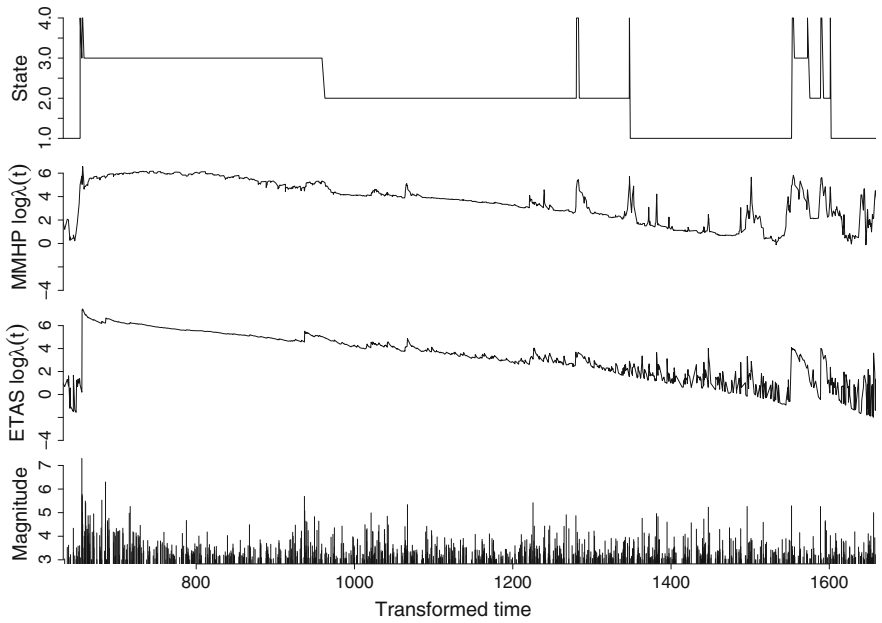

Fig. 8 Enlarged plots for the time period of about a year just after the Landers main shock with the x-axis in transformed time scale. *Top* the tracked Viterbi path for the hidden states given the observed sequence; *next two plots* the estimated intensity functions for the fitted four-state MMHPSD and the fitted ETAS model; *bottom* the magnitude–time plot for the Landers earthquakes

spikes. Recall from the discussion in Sect. 2.1 that for very small $t - t_i$, the ETAS model always has larger intensity than the Hawkes process and is very sensitive to event magnitude.

The log likelihood, AIC, and BIC values for the fitted ETAS model are given in Table 2. Both the AIC and BIC indicate that the ETAS model is a 'better' fit than the MMHPSD with four states, but this does not guarantee that this 'best' model captures the main features of the data. In order to examine whether the intensity of the fitted ETAS model is a good approximation of the true intensity of the observed data, we will examine the corresponding residual point process.

The cumulative number versus the transformed time plot for the fitted ETAS model as shown in Fig. 4c lies well outside of the 99% confidence limits for much of the time between Joshua Tree and Landers, and between Big Bear and Hector Mine. The Kolmogorov–Smirnov test under the null hypothesis of stationarity produces a P value of 0.0002. In comparison with the results for the MMHPSD with four states, for the ETAS model, the t-statistic of the hypothesis test with the null hypothesis that the correlation coefficient between $\log(E_k)$ and $\log(E_{k+1})$ equals zero is 5.0352 with a P value less than 0.001. The autocorrelation of U_k for the ETAS model is shown in Fig. 5, which indicates autocorrelation between U_k for majority of the lags. Moreover, this autocorrelation is consistently positive. Therefore, the inter-arrival times for the ETAS model are not independent.

The interpretation of these results is that the fitted ETAS model is not a good approximation of the true intensity of the data. The reason for this involves the log likelihood of a point process,

$$\log L = \sum_{i=1}^{n} \log \lambda(t_i) - \int_0^T \lambda(t)\, \mathrm{d}t. \tag{13}$$

It is clear that the form of the ETAS intensity (see Sect. 2.1) allows the first term of the log likelihood to be enlarged greatly at each event occurrence. The second term, however, is much less influenced by such spikes at each event occurrence. The Hawkes process contributes much less to the log likelihood on the occurrence of each event. As $\lambda(t)$ is left continuous, provided that events are occurring closely enough in time, $\lambda(t)$ for the ETAS model will be higher, especially as the effect is added to by each subsequent event. Using maximum likelihood estimation, the integral term in (13) is the number of events in the data set, and hence, the likelihood of the ETAS model is greater than that of the MMHPSD. However, the residual process (11) is analogous to a CUMSUM, which will detect model discrepancies that continue over a period of time, in contrast to the log likelihood which isolates discrepancies at individual event locations t_i. In particular, the sum of the contributions from ancestor events in the ETAS model decays too slowly.

7 Discussion

The parameter estimation for an HMM type model incorporating a time-varying conditional intensity function is nowhere near trivial. The proposed MMHPSD switches among a finite number of states according to a Markov transition rate matrix with a self-exciting occurrence rate of the events from a Hawkes process in which the intensity of this process changes after each event occurrence but remains a constant between each two consecutive events. The conditional intensity function of this process is (1), not the original Hawkes process. This simplification is for the sake of parameter estimation. Note that when calculating the transition probability without arrival, $H_{ij}^{(n)}(u)$, an ordinary differential equation (14) has to be solved. If the conditional intensity function takes the form of the original Hawkes process, then we have to calculate an integral, the integrand of which is a matrix exponential with each element of the matrix being a function of t (cf. Wang 2010), a much more formidable computational challenge.

We see from Fig. 4c that the most significant deviations of the ETAS model from the observed data are for the period after the Joshua Tree earthquake but before the Landers earthquake, and the period not long after the Big Bear earthquake. These two periods were shown to be relatively quiescent (Ogata et al. 2003; Marsan and Nalbant 2005). The ETAS model fits well for the period immediately after the Landers earthquake. This suggests that the ETAS model captures well the features of an aftershock sequence, but may not be as good a fit for the entire seismic cycle, including the quiescence period of a sequence. This problem may be inherited from the assumption of the ETAS model. In the ETAS model, the frequency of the aftershocks

triggered by an event with $\{t_i, M_i\}$ is assumed to be $K_i/(t - t_i + c)^p$. For different magnitudes M_i, K_i are different with $K_i = Ke^{\alpha(M_i - M_0)}$. This suggests that for an event with magnitude M_0, the frequency of aftershocks triggered by the event is $K/(t - t_i + c)^p$, and for a larger event with magnitude $M_i > M_0$, the frequency becomes $Ke^{\alpha(M_i - M_0)}/(t - t_i + c)^p$, which is $e^{\alpha(M_i - M_0)}$ times of that of an event with magnitude M_0. Note that the empirical formula $e^{\alpha(M_i - M_0)}$ was found for a sequence of aftershocks triggered by a main shock with magnitude M_i, which describes the total number of aftershocks in the sequence. But in the ETAS model, this is assigned to all the events, including aftershocks and secondary aftershocks. This may result in a higher intensity than is consistent with the actual process. In the MMHPSD, however, there is an indirect coupling of the state and magnitude through the aftershock sequence characteristic of each class.

The advantage of the MMHPSD over the ETAS model is that the former incorporates seismic cycles rather than only an immigration-birth framework, and when the features of the event occurrences evolve towards a different attractor, the model automatically switches into a different regime (or state) in a seismic cycle. This is especially useful for a long sequence with several state changes. While one might use change-point analysis for the ETAS model to account for the seismicity changes, estimating the change point is not an easy procedure in this case, as the number of change points is indeterminate. Our current MMHPSD model is purely temporal, and further development might occur along spatial lines, as has been done for the ETAS model (cf. Ogata 1998).

In order to formulate main shocks and aftershocks in one model, a two-node (main shock/aftershock) stress release/transfer model was presented by Borovkov and Bebbington (2003) and Bebbington (2008), which provided an alternative to the ETAS model for aftershock sequences with a physically attractive mechanism for main shock interaction. Borovkov and Bebbington (2003) also pointed out that, to account for 'secondary' aftershocks, a third node might be added to the two-node model. However, the weakness is that earthquakes have to be identified as main shocks, aftershocks, etc., with main shocks assumed to be from the largest fault or faults, aftershocks assumed to be from the secondary faults, and so on (Borovkov and Bebbington 2003). We outlined how an MMHPSD might be used to perform this task 'automatically' producing, as a by-product, a model of the seismic cycle. Note that this is in some sense a generalization of the two-node or three-node stress transfer model, and the elastic rebound controlling the main shock events is represented by the first passage time between the main shock state.

8 Conclusion

Assuming the existence of an earthquake cycle (for example, mainshock–aftershock–quiescence–precursory seismicity), a new HMM type model is proposed in this paper. Unlike the traditional HMMs and MMPPs, the distinctive feature of this new HMM type model, MMHPSD, is the incorporation of a self-exciting point process into a continuous-time hidden Markov chain. The existing self-exciting models can only capture one or several fixed or pre-identified phases in one (seismic) cycle. For example,

the ETAS model formulates aftershock sequences, while the two-node stress release/transfer model can capture main shocks and aftershocks. The new model, however, while characterizing the self-exciting feature of each phase, switches into a new regime automatically whenever the feature of the event occurrences evolves towards a different attractor. A method for estimating the parameters via EM algorithm was developed, which involves a numerical optimization in the M-step for estimating the parameters in the Hawkes intensity function. Residual analysis techniques for point processes are used to demonstrate the goodness-of-fit for this model. Exploratory data analysis using the MMHPSD on the Landers–Hector Mine earthquake sequences shows that this model is useful in modelling changes of the temporal patterns of seismicity.

Appendix

Proof of Lemma 1

Proof Given the history, the transition probability without arrival is

$$H_{ij}^{(n)}(u) = P\{Y(t_{n-1}+u) = j, \quad N_n(u) = 0 \mid Y(t_{n-1}) = i, \mathcal{H}_{t_{n-1}}\}.$$

For $1 \le i, j \le r$ and for $\Delta u > 0$ we have

$$H_{ij}^{(n)}(u+\Delta u) = \sum_{k=1}^{r} H_{kj}^{(n)}(\Delta u) H_{ik}^{(n)}(u),$$

and the probability of the process remaining in state j without arrival is

$$H_{jj}^{(n)}(\Delta u) = \exp\{-q_j \Delta u\} \exp\{-\lambda_j^*(t_n)\Delta u\},$$

and

$$\lim_{\Delta u \to 0} \frac{H_{jj}^{(n)}(\Delta u) - 1}{\Delta u} = -q_j - \lambda_j^*(t_n).$$

Given that

$$\lim_{\Delta u \to 0} \frac{H_{ij}^{(n)}(\Delta u)}{\Delta u} = q_{ij},$$

we thus have

$$\begin{cases} H_{ij}^{(n)'}(u) = -H_{ij}^{(n)}(u)(q_j + \lambda_j^*(t_n)) + \sum_{k=1, k\neq j}^{r} H_{ik}^{(n)}(u) q_{kj}, & u > 0 \\ H_{ij}^{(n)'}(0) = \delta_{ij}, \end{cases}$$

where $\delta_{ij} = 1$ for $i = j$, and 0 otherwise, or in matrix format,

$$\begin{cases} H^{(n)'}(u) = H^{(n)}(u)(Q - \Lambda^*(t_n)) \\ H^{(n)'}(0) = I, \end{cases} \tag{14}$$

where $H^{(n)}(u) = \{H_{ij}^{(n)}(u)\}$, which yields $H^{(n)}(u) = \exp\{(Q - \Lambda^*(t_n))u\}$ for $u \geq 0$. □

The gradients and Hessian of the function $\mathcal{Q}_2$

To simplify the notations in the derivatives, denote

$$\begin{aligned}
\xi_{ik} &= \eta_i \sum_{t_j < t_{k-1}} e^{-\eta_i (t_{k-1} - t_j)}, \\
\zeta_{ik} &= \frac{1}{\lambda_i + \nu_i \eta_i \sum_{t_j < t_{k-1}} e^{-\eta_i (t_{k-1} - t_j)}}, \\
\varphi_{ik} &= \sum_{t_j < t_{k-1}} e^{-\eta_i (t_{k-1} - t_j)} - \eta_i \sum_{t_j < t_{k-1}} e^{-\eta_i (t_{k-1} - t_j)} (t_{k-1} - t_j), \\
\psi_{ik} &= -2\nu_i \sum_{t_j < t_{k-1}} e^{-\eta_i (t_{k-1} - t_j)} (t_{k-1} - t_j) + \nu_i \eta_i \sum_{t_j < t_{k-1}} e^{-\eta_i (t_{k-1} - t_j)} (t_{k-1} - t_j)^2.
\end{aligned}$$

The gradients of $\mathcal{Q}_2$ are then given by

$$\begin{aligned}
\frac{\partial \mathcal{Q}_2}{\partial \lambda_i} &= \sum_{k=1}^{n} B_{ik} \zeta_{ik} - \sum_{k=1}^{n} \frac{\mathcal{I}_{ik}}{c_k}, \\
\frac{\partial \mathcal{Q}_2}{\partial \nu_i} &= \sum_{k=1}^{n} B_{ik} \zeta_{ik} \xi_{ik} - \sum_{k=1}^{n} \frac{\mathcal{I}_{ik}}{c_k} \xi_{ik}, \\
\frac{\partial \mathcal{Q}_2}{\partial \eta_i} &= \sum_{k=1}^{n} B_{ik} \zeta_{ik} \nu_i \varphi_{ik} - \sum_{k=1}^{n} \frac{\mathcal{I}_{ik}}{c_k} \nu_i \varphi_{ik}.
\end{aligned}$$

The Hessian of the function $\mathcal{Q}_2$ is given by

$$\begin{aligned}
\frac{\partial^2 \mathcal{Q}_2}{\partial \lambda_i^2} &= -\sum_{k=1}^{n} B_{ik} \zeta_{ik}^2, \\
\frac{\partial^2 \mathcal{Q}_2}{\partial \lambda_i \partial \nu_i} &= -\sum_{k=1}^{n} B_{ik} \zeta_{ik}^2 \xi_{ik}, \\
\frac{\partial^2 \mathcal{Q}_2}{\partial \lambda_i \partial \eta_i} &= -\sum_{k=1}^{n} B_{ik} \zeta_{ik}^2 \nu_i \varphi_{ik},
\end{aligned}$$

$$\frac{\partial^2 \mathcal{Q}_2}{\partial \nu_i^2} = -\sum_{k=1}^{n} B_{ik}\zeta_{ik}^2\xi_{ik}^2,$$

$$\frac{\partial^2 \mathcal{Q}_2}{\partial \nu_i \partial \eta_i} = -\sum_{k=1}^{n} B_{ik}\zeta_{ik}^2\nu_i\varphi_{ik}\xi_{ik} + \sum_{k=1}^{n} B_{ik}\zeta_{ik}\varphi_{ik} - \sum_{k=1}^{n} \frac{\mathcal{I}_{ik}}{c_k}\varphi_{ik},$$

$$\frac{\partial^2 \mathcal{Q}_2}{\partial \eta_i^2} = -\sum_{k=1}^{n} B_{ik}\zeta_{ik}^2\nu_i^2\varphi_{ik}^2 + \sum_{k=1}^{n} B_{ik}\zeta_{ik}\psi_{ik} - \sum_{k=1}^{n} \frac{\mathcal{I}_{ik}}{c_k}\psi_{ik}.$$

Acknowledgments This work was supported by the Marsden Fund, administered by the Royal Society of New Zealand. We are grateful to Professor David Vere-Jones for insightful suggestions on using the self-exciting Hawkes process as the conditional intensity function of the observed point process. We would like to thank three anonymous reviewers for constructive suggestions which have greatly improved an earlier manuscript and have made the paper more concise. We also express our thanks to Marco Brenna, Takaki Iwata, Chin-Diew Lai and Roger Littlejohn for helpful comments.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723.

Akaike, H. (1978). A Bayesian analysis of the minimum AIC procedure. *Annals of the Institute of Statistical Mathematics, 30*(1), 9–14; also included in E. Parzen et al. (Eds.) (1998), *Selected papers of Hirotugu Akaike* (pp. 275–280). Berlin: Springer.

Bebbington, M. S. (2007). Identifying volcanic regimes using hidden Markov models. *Geophysical Journal International, 171*, 921–942.

Bebbington, M. S. (2008). Estimating rate- and state-fraction parameters using a two-node stochastic model for aftershocks. *Tectonophysics, 457*, 71–85.

Bebbington, M. S., Harte, D. S. (2001). On the statistics of the linked stress release model. *Journal of Applied Probability, 38A*, 176–187.

Bebbington, M. S., Harte, D. S., Jaumé, S. C. (2010). Repeated intermittent earthquake cycles in the San Francisco Bay Region. *Pure and Applied Geophysics, 167*, 801–818.

Borovkov, K., Bebbington, M. S. (2003). A stochastic two-node stress transfer model reproducing Omori's law. *Pure and Applied Geophysics, 160*, 1429–1445.

Bowsher, C. G. (2007). Modelling security market events in continuous time: intensity based, multivariate point process models. *Journal of Econometrics, 141*, 876–912.

Brémaud, P., Massoulié, L. (1996). Stability of nonlinear Hawkes processes. *Annals of Probability, 24*, 1563–1588.

Bufe, C. G., Varnes, D. J. (1993). Predictive modeling of the seismic cycle of the greater San Francisco Bay region. *Journal of Geophysical Research, 98*, 9871–9883.

Daley, D. J., Vere-Jones, D. (2003). *Introduction to the theory of point processes* (2nd ed.). New York: Springer.

Fedotov, S. A. (1968). The seismic cycle, quantitative seismic zoning, and long-term seismic forecasting. In S. V. Medvedev (Eds.), *Seismic zoning in the USSR* (pp. 133–166). Moscow: Izdatel'stvo Nauka.

Fischer, W., Meier-Hellstern, K. S. (1993). The Markov-modulated Poisson process (MMPP) cookbook. *Performance Evaluation, 18*(2), 149–171.

Fletcher, R., Powell, M. J. D. (1963). A rapidly convergent method for minimization. *The Computer Journal, 6*, 163–168.

Harte, D. S. (2005). *Package "HiddenMarkov": discrete time hidden Markov models.* R statistical program routines. Wellington: Statistics Research Associates. http://cran.at.r-project.org/web/packages/HiddenMarkov.

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

Hawkes, A. G., Adamopoulos, L. (1973). Cluster models for earthquakes-regional comparisons. *Bulletin of the International Statistical Institute, 45*, 454–461.

Heffes, H., Lucantoni, D. (1986). A Markov modulated characterization of packetized voice and data traffic related statistical performance. *IEEE Journal on Selected Areas in Communications, 4*, 856–868.
Helmstetter, A., Sornette, D. (2002). Subcritical and supercritical regimes in epidemic models of earthquake aftershocks. *Journal of Geophysical Research, 107*. doi:10.1029/2001JB001580
Hill, D. P., Reasenberg, P. A., Michael, A., Arabaz, W., Beroza, G. C., Brune, J. N., Brumbaugh, D., Davis, S., DePolo, D., Ellsworth, W. L., Gomberg, J., Harmsen, S., House, L., Jackson, S. M., Johnston, M., Jones, L., Keller, R., Malone, S., Nava, S., Pechmann, J. C., Sanford, A., Simpson, R. W., Smith, R. S., Stark, M., Stickney, M., Walter, S., Zollweg, J. (1993). Seismicity in the western United States remotely triggered by the $M7.4$ Landers, California, earthquake of June 28, 1992. *Science, 260*, 1617–1623.
Hill, D. P., Johnston, M. J. S., Langbein, J. O., Bilham, R. (1995). Response of Long Valley caldera to the $M_w = 7.3$ Landers, California, earthquake. *Journal of Geophysical Research, 100*, 12985–13005.
Hughes, J. P., Guttorp, P. (1994). A class of stochastic models for relating synoptic atmospheric patterns to regional hydrologic phenomena. *Water Resources Research, 30*, 1535–1546.
Jaumé, S. C., Bebbington, M. S. (2004). Accelerating seismic release from a self-correcting stochastic model. *Journal of Geophysical Research, 109*, B12301. doi:10.1029/2003JB002867
MacDonald, I., Zucchini, W. (1997). *Hidden-Markov and other models for discrete-valued time series*. New York: Chapman and Hall.
Marsan, D. (2003). Triggering of seismicity at short timescales following Californian earthquakes. *Journal of Geophysical Research, 108*, 2266. doi:10.1029/2002JB001946
Marsan, D., Nalbant, S. S. (2005). Methods for measuring seismicity rate changes: a review and a study of how the M_w 7.3 Landers earthquake affected the aftershock sequence of the M_w 6.1 Joshua Tree earthquake. *Pure and Applied Geophysics, 162*, 1151–1185.
Mogi, K. (1968). *Source locations of elastic shocks in the fracturing process in rocks* (1). *Bulletin of Earthquake Research Institute, 46*, 1103–1125.
Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association, 83*(401), 9–27.
Ogata, Y. (1998). Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics, 50*, 379–402.
Ogata, Y., Jones, L. M., Toda, S. (2003). When and where the aftershock activity was depressed: contrasting decay patterns of the proximate large earthquakes in southern California. *Journal of Geophysical Research, 108*(B6), 2318. doi:10.1029/2002JB002009 (ESE1-12).
Pievatolo, A., Rotondi, R. (2008). Statistical identification of seismic phases. *Geophysical Journal International, 173*, 942–957.
Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*, 257–286
Roberts, W. J. J., Ephraim, Y., Dieguez, E. (2006). On Rydén's EM algorithm for estimating MMPPs. *IEEE Signal Processing Letters, 13*(6), 373–376.
Rydén, T. (1994). Parameter estimation for Markov modulated Poisson processes. *Communications in Statistics–Stochastic Models, 10*(4), 795–829.
Rydén, T. (1996). An EM algorithm for estimation in Markov-modulated Poisson processes. *Computational Statistics & Data Analysis, 21*, 431–447.
Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461–464.
Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *The Annals of Statistics, 8*, 147–164.
Shibata, R. (1981). An optimal selection of regression variables. *Biometrika, 68*(1), 45–54.
Utsu, T., Ogata, Y., Matsu'ura, R. S. (1995). The centenary of the Omori formula for a decay law of aftershock activity. *Journal of Physics of the Earth, 43*, 1–33.
Van Loan, C. F. (1978). Computing integrals involving the matrix exponential. *IEEE Transactions on Automatic Control, AC-23*(3), 395–404.
Vere-Jones, D., Robinson, R., Yang, W. (2001). Remarks on the accelerated moment release model: problems of model formulation, simulation and estimation. *Geophysical Journal International, 144*, 517–531.
Wang, T. (2010). *Statistical models for earthquakes incorporating ancillary data*. PhD thesis, New Zealand: Massey University.
Zhuang, J. (2000). Statistical modeling of seismicity patterns before and after the 1990 Oct 5 Cape Palliser earthquake, New Zealand. *New Zealand Journal of Geology and Geophysics, 43*, 447–460.
Zucchini, W., Guttorp, P. (1991). A hidden Markov model for space-time precipitation. *Water Resources Research, 27*(8), 1917–1923.