12-2011

Quasi-hidden Markov model and its applications in cluster analysis of earthquake catalogs

Zhengxiao WU
Singapore Management University, zxwu@smu.edu.sg

Published in Journal of Geophysical Research, 2011 December, Volume 116, Issue B12, Pages 1-12
http://doi.org/10.1029/2011JB008447

Quasi-hidden Markov model and its applications in cluster analysis of earthquake catalogs

Zhengxiao Wu[1]

Received 13 April 2011; revised 6 October 2011; accepted 11 October 2011; published 24 December 2011.

[1] We identify a broad class of models, quasi-hidden Markov models (QHMMs), which include hidden Markov models (HMMs) as special cases. Applying the QHMM framework, this paper studies how an earthquake cluster propagates statistically. Two QHMMs are used to describe two different propagating patterns. The “mother-and-kids” model regards the first shock in an earthquake cluster as “mother” and the aftershocks as “kids,” which occur in a neighborhood centered by the mother. In the “domino” model, however, the next aftershock strikes in a neighborhood centered by the most recent previous earthquake in the cluster, and therefore aftershocks act like dominoes. As the likelihood of QHMMs can be efficiently computed via the forward algorithm, likelihood-based model selection criteria can be calculated to compare these two models. We demonstrate this procedure using data from the central New Zealand region. For this data set, the mother-and-kids model yields a higher likelihood as well as smaller AIC and BIC. In other words, in the aforementioned area the next aftershock is more likely to occur near the first shock than near the latest aftershock in the cluster. This provides an answer, though not entirely satisfactorily, to the question “where will the next aftershock be?”. The asymptotic consistency of the model selection procedure in the paper is duly established, namely that, when the number of the observations goes to infinity, with probability one the procedure picks out the model with the smaller deviation from the true model (in terms of relative entropy rate).

Citation: Wu, Z. (2011), Quasi-hidden Markov model and its applications in cluster analysis of earthquake catalogs, *J. Geophys. Res.*, *116*, B12316, doi:10.1029/2011JB008447.

1. Introduction

1.1. Aftershocks

[2] The meaning of aftershock varies in seismological literature. Conventionally, the earthquake with the largest magnitude in an earthquake cluster is called the main shock, those before it are called foreshocks, and those after it are called aftershocks. Thus the foreshocks, the main shock, and the aftershocks cannot be identified until after the cluster ends. This is inconvenient for real-time monitoring and prediction. To avoid this complexity, many earthquake models [*Kagan and Knopoff*, 1981; *Rathbun*, 1993; *Ogata*, 1998; *Zhuang et al.*, 2002] define aftershocks based solely on their occurrence time, not factoring in the magnitudes. We adhere to this practice. In this article the first earthquake in an earthquake cluster is named first-shock and the rest its aftershocks.

[3] While earthquake prediction is generally very challenging, aftershock forecast is relatively tractable on account of aftershocks being near the first shock in space and time. (As a referee pointed out, this might be a statement too optimistic. Our understanding of the aftershock sequence evolution is still very poor.) In fact, all the earthquakes in a cluster are close to each other as the name cluster implies. A famous example is the successful prediction of the 7.3-magnitude earthquake in Haicheng, China, in 1975. Warnings were sent out and evacuations ordered days before the earthquake hit the city. Thousands of lives were saved. The telltale sign had been a series of small tremors that occurred before the big one [*Wang et al.*, 2006].

1.2. Cluster Models

[4] Branching models are widely employed for the study of earthquake clusters. An example is the epidemic-type aftershock sequence (ETAS) model proposed by *Ogata* [1998]. In the last decade, intensive research on the ETAS model has been conducted [*Zhuang et al.*, 2002, 2004; *Zhuang*, 2006; *Ogata and Zhuang*, 2006; *Vere-Jones and Zhuang*, 2008; *Helmstetter and Sornette*, 2002, 2003; *Helmstetter*, 2003]. The ETAS model is a branching process with immigrants, with each existent earthquake (“ancestor”) producing offspring independently. The immigrants are earthquakes without ancestors. They are interpreted as the background seismic activity. The functional form of the intensity of the point

[1] Department of Statistics and Applied Probability, National University of Singapore, Singapore.

processis often carefully selected so that it accommodates the Gutenberg-Richter frequency-magnitude law (G-R law) and the empirical Omori law for aftershock sequences.

[5] Aside from earthquake clusters, there are a large number of earthquakes that strike without any foreshocks or aftershocks. These are called single earthquakes. An analysis of an earthquake catalog often begins with the separation of the earthquakes into earthquake clusters and single earthquakes. *Zhuang et al.* [2002] proposed a stochastic declustering algorithm based on the ETAS model. The algorithm is "stochastic" in that random numbers are generated in the procedure and the resulting partition depends on the realization of these random numbers. The algorithm therefore gives a different partition each time it is run. However, a unique declustering is often preferred in practice.

[6] An alternative to the branching models is the "mother-and-kids" model suggested by *Wu* [2009, 2010]. In the work of *Wu* [2009, 2010], an earthquake catalog is modeled as a superposition of two independent processes: single earthquakes and earthquake clusters. The single earthquake occurrences follow a time-homogeneous Poisson process, while clusters are defined by a time-inhomogeneous Poisson process so that they are randomly initiated and eventually die out. The model is given thename "mother-and-kids" as the first-shock is considered the "mother" and the aftershocks the "kids". When the cluster is active, aftershocks occur in the neighborhood of the mother where the neighboring area is defined by a bivariate Gaussian distribution centered at the mother's location. Each occurrence of an aftershock has a probability p of deactivating the cluster, or each kid's birth has a probability p of sterilizing the mother. It is assumed that there is no more than one active cluster at a time. The model does not include the magnitudes of the events.

[7] *Wu* [2010] carefully compares the mother-and-kids model with the ETAS model [*Zhuang et al.*, 2002, 2004] and demonstrates that the mother-and-kids model outdoes the ETAS model in several ways. In particular, it naturally gives an algorithm that produces a unique declustering for earthquake catalogs. Furthermore, quite interestingly, the fact that the model does not include the magnitudes enables the discovery of two significantly different G-R laws for the single earthquakes and earthquake clusters, respectively. Thus one can conclude that the mechanisms that control the evolution of single earthquakes and earthquake clusters are different [*Knopoff*, 2000].

[8] Several reasons motivated the mother-and-kids model. One physical reason is that the ETAS model attempts to establish a causal relationship between events by assigning each aftershock a single, unique ancestor/trigger, which *Helmstetter and Sornette* [2002, p. 10-2] noted as an "important defect" on grounds that events may well be triggered by the combined loading and action of several previous earthquakes (in other words, an aftershock could have several triggers). By contrast, the mother-and-kids model abstracts the cause of the aftershock occurrences as "the cluster is active." Hence the first-shock is only a nominal mother, a precursor, but not necessarily the physical trigger of the aftershocks. Therefore the mother-and-kids model models precedence which suffices for aftershock forecast and, in doing so, bypasses the tricky problem of causality.

[9] *Wu* [2010] formulates the mother-and-kids model in the framework of a hidden Markov model (HMM). However, unlike a conventional HMM, the number of hidden states in the mother-and-kids model increases with the number of observations. In fact, the mother-and-kids model belongs to a more general class of models, called quasi-hidden Markov models (QHMMs) [*Wu*, 2011].

[10] The goal of this paper is to introduce the QHMM framework to seismology and to apply the framework to investigate further details about how an earthquake cluster propagates statistically. To that end, an alternative QHMM (the "domino" model) is proposed. In the model, the next aftershock strikes in a neighborhood centered by the most recent event in the cluster; hence aftershocks behave like dominoes. It mimics the branching model's idea that each existent earthquake can trigger offspring in its neighborhood. And it also obeys the simple intuition behind the mother-and-kids model: the earthquakes within a cluster are near one another in space and time. So where will the next aftershock be? Or more specifically, which fits the data better, the mother-and-kids model or the domino model? This is a model selection problem.

[11] The rest of the paper is organized as follows, Section 2 gives the QHMM formulation of the mother-and-kids model and section 3 introduces the domino model. In section 4, we examine an earthquake catalog of the central New Zealand region and conduct data analysis by using the two QHMMs, and the likelihood-based model selection criteria indicate that the mother-and-kids model fits this data set better than the domino model. Section 5 discusses extensions and future work. Appendix A describes the QHMMs and the associated algorithms. Appendix B studies the asymptotics of the likelihood inferences for QHMMs which justifies the model selection procedure in section 4.

2. Mother-and-Kids Model

2.1. Notations

[12] We use O_t and q_t to denote the observation and the hidden state at time t in a QHMM (see Appendix A for a review of QHMM). The O_t and q_t in the mother-and-kids model are defined as follows. Suppose an earthquake catalog contains T earthquake records. Earthquake t ($1 \leq t \leq T$) has the record $O_t = (x_t, y_t, \tau_t)$, which are the longitude and latitude of the epicenter and the origin time, respectively. The hidden state is defined to be $q_t = (J_t, C_t, A_t)$, where J_t is the index of the most recent mother quake up to t, i.e., $J_t = \max\{k \leq t$: earthquake k is a mother quake$\}$. If there are no cluster quakes and hence no mother quakes up to t, we let $J_t = 0$. C_t and A_t are two indicator functions: $C_t = 1$ if earthquake t is a cluster earthquake and $C_t = 0$ otherwise; $A_t = 1$ if the mother earthquake J_t is fertile at time τ_t (the cluster is still active), $A_t = 0$ otherwise.

[13] Five parameters are introduced in the model: γ is the space-homogeneous intensity of the point process for single earthquakes, λ is the extra intensity when a cluster is active, ϵ is the intensity of the initiation of a new cluster, d is the variance parameter of the bivariate Gaussian distribution, p is the probability that the mother earthquake becomes sterile after giving birth to one more kid. The mother earthquake is assumed to be born reproductive, which guarantees that each cluster contains at least two earthquakes. The two dimensional (longitude and latitude) region under study is

denoted R, and its area is denoted $|R|$. Therefore a uniformly distributed point in this region has the density $\mathbf{1}_{(x,y)\in R}/|R|$.

2.2. Simulate the Model

[14] To further illustrate the meaning the parameters, we outline how a synthetic catalogue can be simulated under the mother-and-kids model. One can carry out the simulation by generating the single earthquakes and the clustered earthquake separately and then merging the two sequences.

[15] Algorithm 1 is as follows:

[16] 1. Simulate the single earthquake sequence. Each single earthquake's location is independent and uniformly distributed in the specified region. The time interval between two consecutive single earthquakes is exponentially distributed with mean $1/\gamma$. (Recall that for a simple Poisson process with intensity γ, the waiting time is exponentially distributed with mean $1/\gamma$.)

[17] 2. Simulate the clustered earthquake sequence.

[18] (i) Simulate the mother earthquake: the location of the mother is uniformly distributed in the region, and the waiting time before a mother earthquake strikes is exponential with mean $1/\epsilon$. The cluster is activated right after the mother earthquake strikes.

[19] (ii) Simulate the next aftershock: the location of the next aftershock follows a bivariate Gaussian distribution, centered at its mother earthquake, with variance parameter d; the waiting time before the next aftershock strikes is exponential with mean $1/(\epsilon + \lambda)$. The occurrence of this aftershock has a probability p to deactivated the cluster. If the cluster is deactivated, go to i, otherwise, repeat ii.

[20] 3. Combine the two sequences and sort the sequence according to the occurrence time.

2.3. QHMM Formulation

[21] We give the precise QHMM formulation of the mother-and-kids model in this section. Recall that to fully determine the QHMM, it suffices to specify the conditional distribution of $P(O_{t+1}, q_{t+1}|\mathrm{O}_{1:t}, q_t)$, which satisfies

$$P(O_{t+1}, q_{t+1}|O_{1:t}, q_t) = P(O_{t+1}|O_{1:t}, q_{t+1}, q_t)P(q_{t+1}|O_{1:t}, q_t).$$

[22] The mother-and-kids model assumes that

$$P(O_{t+1}|O_{1:t}, q_{t+1}, q_t) = P(O_{t+1}|O_{1:t}, q_{t+1})$$

and

$$P(q_{t+1}|O_{1:t}, q_t) = P(q_{t+1}|q_t).$$

[23] These conditional distributions are discussed below in two cases and each case has several scenarios:

[24] 1. $A_t = 0$, i.e., no cluster is active at time τ_t. In this case, the next earthquake $t + 1$ is either a single earthquake (scenario a) or a mother earthquake which starts a new active cluster (scenario b). As the new cluster is initiated withintensity ϵ and the single earthquake occurs with intensity γ, it is straightforward that the next observation $t + 1$ is a single quake with probability $\frac{\gamma}{\epsilon+\gamma}$ and is a mother quake with probability $\frac{\epsilon}{\epsilon+\gamma}$. The waiting time $\tau_{t+1} - \tau_t$ follows an exponential distribution with parameter $\epsilon + \gamma$. The location (x_{t+1}, y_{t+1}) is uniformly distributed in R in both scenarios.

[25] (i) Earthquake $t + 1$ is a single earthquake, then there is still no active cluster at time τ_{t+1}, and J_{t+1} keeps the same value as J_t. Hence

$$P(q_{t+1}|q_t) = P(J_{t+1} = J_t, C_{t+1} = 0,\ A_{t+1} = 0|q_t) = \frac{\gamma}{\epsilon + \gamma},$$

and

$$\begin{aligned}P(O_{t+1}|O_{1:t}, q_{t+1}) = P(x_{t+1}, y_{t+1}, \tau_{t+1}|O_{1:t},\ J_{t+1} = J_t,\\ C_{t+1} = 0,\ A_{t+1} = 0)\\ = \frac{(\epsilon + \gamma)\exp\{-(\tau_{t+1} - \tau_t)(\epsilon + \gamma)\}\mathbf{1}_{(x_{t+1},y_{t+1})\in R}}{|R|}.\end{aligned}$$

[26] (ii) Earthquake $t + 1$ is a mother quake; it starts a new active cluster; J_{t+1} is equal to $t + 1$. Hence

$$P(q_{t+1}|q_t) = P(J_{t+1} = t + 1, C_{t+1} = 1,\ A_{t+1} = 1|q_t) = \frac{\epsilon}{\epsilon + \gamma},$$

and

$$\begin{aligned}P(O_{t+1}|O_{1:t}, q_{t+1}) = P(x_{t+1}, y_{t+1}, \tau_{t+1}|O_{1:t}, J_{t+1} = t + 1,\\ C_{t+1} = 1, A_{t+1} = 1)\\ = \frac{(\epsilon + \gamma)\exp\{-(\tau_{t+1} - \tau_t)(\epsilon + \gamma)\}\mathbf{1}_{(x_{t+1},y_{t+1})\in R}}{|R|}.\end{aligned}$$

[27] 2. $A_t = 1$, i.e., there is one active cluster at time τ_t. This case has three scenarios. An argument similar to the one in the previous case applies. The presence of an active cluster introduces an extra intensity λ, hence the waiting time $\tau_{t+1} - \tau_t$ follows an exponential distribution with parameter $\lambda + \epsilon + \gamma$.

[28] (i) Earthquake $t + 1$ is a single quake; the cluster is still active at time τ_{t+1}; J_{t+1} keeps the same value as J_t. Hence

$$P(q_{t+1}|q_t) = P(J_{t+1} = J_t, C_{t+1} = 0,\ A_{t+1} = 1|q_t) = \frac{\gamma}{\lambda + \epsilon + \gamma},$$

and

$$\begin{aligned}&P(O_{t+1}|O_{1:t}, q_{t+1})\\ &= P(x_{t+1}, y_{t+1}, \tau_{t+1}|O_{1:t}, J_{t+1} = J_i, C_{t+1} = 0, A_{t+1} = 1)\\ &= \frac{(\lambda + \epsilon + \gamma)\exp\{-(\tau_{t+1} - \tau_t)(\lambda + \epsilon + \gamma)\}\mathbf{1}_{(x_{t+1},y_{t+1})\in R}}{|R|}.\end{aligned}$$

[29] (ii) Earthquake $t + 1$ is a cluster quake (kid); it does not sterilize its mother (this occurs with probability $1\text{-}p$); J_{t+1} keeps the same value as J_t. In this and the next scenario, the location (x_{t+1}, y_{t+1}) follows a bivariate Gaussian distribution centered at its mother. Thus

$$P(q_{t+1}|q_t) = P(J_{t+1} = J_t, C_{t+1} = 1, A_{t+1} = 1|q_t) = \frac{(1 - p)(\lambda + \epsilon)}{\lambda + \epsilon + \gamma},$$

and

$$P(O_{t+1}|O_{1:t}, q_{t+1}) = P(x_{t+1}, y_{t+1}, \tau_{t+1}|O_{1:t}, J_{t+1} = J_t, C_{t+1} = 1, A_{t+1} = 1)$$
$$= \frac{(\lambda + \epsilon + \gamma)\exp\{-(\tau_{t+1} - \tau_t)(\lambda + \epsilon + \gamma)\}\exp\left\{-\frac{\left(x_{t+1}-x_{J_t}\right)^2+\left(y_{t+1}-y_{J_t}\right)^2}{2d}\right\}}{2\pi d}.$$

[30] (iii) Earthquake $t + 1$ is a cluster quake (kid); it sterilizes its mother (this occurs with probability p); J_{t+1} keeps the same value as J_t. Hence

$$P(q_{t+1}|q_t) = P(J_{t+1} = J_t, C_{t+1} = 1, A_{t+1} = 0|q_t) = \frac{p(\lambda + \epsilon)}{\lambda + \epsilon + \gamma},$$

and

$$P(O_{t+1}|O_{1:t}, q_{t+1}) = P(x_{t+1}, y_{t+1}, \tau_{t+1}|O_{1:t}, J_{t+1} = J_t, C_{t+1} = 1, A_{t+1} = 0)$$
$$= \frac{(\lambda + \epsilon + \gamma)\exp\{-(\tau_{t+1} - \tau_t)(\lambda + \epsilon + \gamma)\}\exp\left\{-\frac{\left(x_{t+1}-x_{J_t}\right)^2+\left(y_{t+1}-y_{J_t}\right)^2}{2d}\right\}}{2\pi d}.$$

[31] It is the last two scenarios in case 2 that describe how an earthquake cluster grows.

[32] Two characteristics set the mother-and-kids model apart from a conventional HMM. First, the hidden state space $\mathbf{S_t}$ is not fixed over time because J_t takes values in $\{0, 1, 2,\ldots,t\}$. In fact, there are roughly $4t$ possible combinations of (J_t, C_t, A_t) in $\mathbf{S_t}$.

[33] The second characteristic is more subtle: the generation distribution of $O_{t+1} = (x_{t+1}, y_{t+1}, \tau_{t+1})$ not only depends on $q_{t+1} = (J_{t+1}, C_{t+1}, A_{t+1})$ but also it depends on τ_t and possibly on (x_{J_t}, y_{J_t}). This violates the basic HMM assumption that the generation distribution of O_t is solely decided by q_t. Hence the mother-and-kids model is a genuine QHMM because τ_t and (x_{J_t}, y_{J_t}) are functions of q_{t+1} and $O_{1:t}$.

3. Domino Model

3.1. Background

[34] Though various physical models and seismological theories have been invented to try to explain the mechanism that generates aftershock sequences [see, e.g., *Harris*, 2001, and references therein], no consensus has been reached. This leaves room for statistically exploring the possibilities of how an earthquake cluster propagates.

[35] The domino model has the same setup as the mother-and-kids model. An earthquake is either a single earthquake or a member of an earthquake cluster. A time-homogeneous Poisson process governs the single earthquake's occurrence, while a time-inhomogeneous Poisson process models the earthquake clusters. When an earthquake cluster is alive, it grows with extra intensity. Each occurrence of an aftershock deactivates the cluster with a positive probability (a first-shock always activates the cluster though. This assumption ensures each cluster has at least two events), hence the cluster eventually dies out. The only part that differs from the mother-and-kids model is the next aftershock occurs in the neighborhood centered at the most recent earthquake in the cluster instead of centered at the first shock.

[36] It is interesting to note that although the domino model looks very similar to the mother-and-kids model, its state space at time t is only about half size of the state space of a mother-and-kids model. The detailed dynamics of the domino model is described below.

3.2. Dynamics of the Domino Model

[37] Hidden state in the domino model is defined as $q_t' = (J_t', C_t', A_t')$, where J_t' is the index of the most recent earthquake in the most recent cluster up to t. $C_t' = 1$ if earthquake t is a cluster earthquake and $C_t' = 0$ otherwise; $A_t' = 1$ if the most recent cluster is active, $A_t' = 0$ otherwise.

[38] There are five parameters in the model as well: γ' is the intensity of the point process for single earthquakes, λ' is the extra intensity when a cluster is active, ϵ' is the intensity of the initiation of a new cluster, d' is the variance parameter of the bivariate Gaussian distribution, p' is the probability that an aftershock kills (deactivates) the cluster to which it belongs.

[39] The conditional distribution of $P(O_{t+1}, q'_{t+1}|\mathrm{O}_{1:t}, q_t)$ satisfies

$$P(O_{t+1}, q'_{t+1}|O_{1:t}, q'_t) = P(O_{t+1}|O_{1:t}, q'_{t+1})P(q'_{t+1}|q'_t).$$

and is listed below in two cases:

[40] 1. $A_t' = 0$, i.e., no cluster is active at time τ_t. In this case, the next earthquake $t + 1$ is either a single quake (scenario a) or a first shock which starts a new active cluster (scenario b).

[41] (i) Earthquake $t + 1$ is a single quake, then there is still no active cluster at time τ_{t+1}, and J'_{t+1} keeps the same value as J_t'. Hence

$$P(q'_{t+1}|q'_t) = P(J'_{t+1} = J'_t, C'_{t+1} = 0, A'_{t+1} = 0|q'_t) = \frac{\gamma'}{\epsilon' + \gamma'},$$

and

$$P(O_{t+1}|O_{1:t}, q'_{t+1})$$
$$= P(x_{t+1}, y_{t+1}, \tau_{t+1}|O_{1:t}, J'_{t+1} = J_t, C'_{t+1} = 0, A'_{t+1} = 0)$$
$$= \frac{(\epsilon' + \gamma')\exp\{-(\tau_{t+1} - \tau_t)(\epsilon' + \gamma')\}\mathbf{1}_{(x_{t+1}, y_{t+1})\in R}}{|R|}.$$

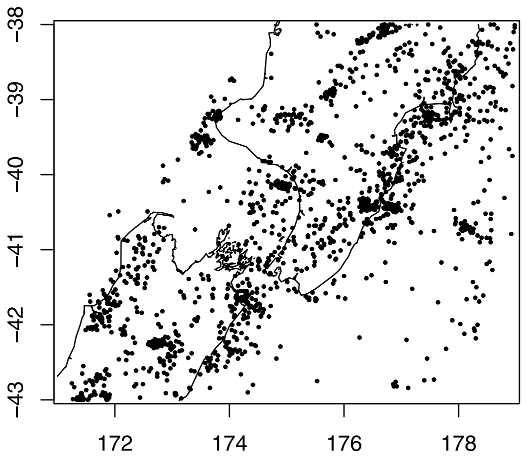

Figure 1. Epicentral locations in the central New Zealand region.

[42] (ii) Earthquake $t + 1$ is a first shock; it starts a new active cluster; J'_{t+1} is equal to $t + 1$. Hence

$$P(q'_{t+1}|q'_t) = P(J'_{t+1} = t + 1, C'_{t+1} = 1, A'_{t+1} = 1|q'_t) = \frac{\epsilon'}{\epsilon' + \gamma'},$$

and

$$\begin{aligned} &P(O_{t+1}|O_{1:t}, q'_{t+1}) \\ &\quad = P(x_{t+1}, y_{t+1}, \tau_{t+1}|O_{1:t}, J'_{t+1} = t + 1, C'_{t+1} = 1, A'_{t+1} = 1) \\ &\quad = \frac{(\epsilon' + \gamma') \exp\{-(\tau_{t+1} - \tau_t)(\epsilon' + \gamma')\}\mathbf{1}_{(x_{t+1}, y_{t+1}) \in R}}{|R|}. \end{aligned}$$

[43] 2. $A_t' = 1$, i.e., there is one active cluster at time τ_t. This case has three scenarios.

[44] (i) Earthquake $t + 1$ is a single quake; the cluster is still active at time τ_{t+1}; J'_{t+1} keeps the same value as J_t'. Hence

$$P(q'_{t+1}|q'_t) = P(J'_{t+1} = J_i', C'_{t+1} = 0, A'_{t+1} = 1|q'_t) = \frac{\gamma'}{\lambda' + \epsilon' + \gamma'},$$

and

$$\begin{aligned} &P(O_{t+1}|O_{1:t}, q'_{t+1}) \\ &\quad = P(x_{t+1}, y_{t+1}, \tau_{t+1}|O_{1:t}, J'_{t+1} = J_t', C'_{t+1} = 0, A'_{t+1} = 1) \\ &\quad = \frac{(\lambda' + \epsilon' + \gamma') \exp\{-(\tau_{t+1} - \tau_t)(\lambda' + \epsilon' + \gamma')\}\mathbf{1}_{(x_{t+1}, y_{t+1}) \in R}}{|R|}. \end{aligned}$$

[45] (ii) Earthquake $t + 1$ is a cluster quake (aftershock); it does not kill the cluster; J'_{t+1} takes value $t + 1$ as earthquake $t + 1$ becomes the most recent aftershock. Thus

$$\begin{aligned} P(q'_{t+1}|q'_t) &= P(J'_{t+1} = t + 1, C'_{t+1} = 1, A'_{t+1} = 1|q'_t) \\ &= \frac{(1 - p')(\lambda' + \epsilon')}{\lambda' + \epsilon' + \gamma'}, \end{aligned}$$

and

$$\begin{aligned} P(O_{t+1}|O_{1:t}, q'_{t+1}) &= P(x_{t+1}, y_{t+1}, \tau_{t+1}|O_{1:t}, J'_{t+1} = J_t, C'_{t+1} = 1, A'_{t+1} = 1) \\ &= \frac{(\lambda' + \epsilon' + \gamma') \exp\{-(\tau_{t+1} - \tau_t)(\lambda' + \epsilon' + \gamma')\} \exp\left\{-\frac{\left(x_{t+1} - x_{J_t'}\right)^2 + \left(y_{t+1} - y_{J_t'}\right)^2}{2d'}\right\}}{2\pi d'}. \end{aligned}$$

[46] (iii) Earthquake $t + 1$ is a cluster quake (aftershock); it kills the cluster to which it belongs; J'_{t+1} becomes $t + 1$. Hence

$$P(q'_{t+1}|q'_t) = P(J'_{t+1} = t + 1, C'_{t+1} = 1, A'_{t+1} = 0|q'_t) = \frac{p'(\lambda' + \epsilon')}{\lambda' + \epsilon' + \gamma'},$$

and

$$\begin{aligned} &P(O_{t+1}|O_{1:t}, q'_{t+1}) \\ &\quad = P(x_{t+1}, y_{t+1}, \tau_{t+1}|O_{1:t}, J'_{t+1} = J_t, C'_{t+1} = 1, A'_{t+1} = 0) \\ &\quad = \frac{(\lambda' + \epsilon' + \gamma') \exp\{-(\tau_{t+1} - \tau_t)(\lambda' + \epsilon' + \gamma')\} \exp\left\{-\frac{\left(x_{t+1} - x_{J_t'}\right)^2 + \left(y_{i+1} - y_{J_t'}\right)^2}{2d'}\right\}}{2\pi d'}. \end{aligned}$$

[47] It is the last two scenarios that set the domino model apart from the mother-and-kids model. Note that when $C_t' = 1$, earthquake t must be the most recent cluster earthquake up to t, i.e., $J_t' = t$. This fact implies that the size of the hidden state space is about $2t$ at time t.

4. Earthquakes in Central New Zealand Region

4.1. Data and Declustering

[48] The data set contains the earthquakes that occurred in the period January 1970 to August 1999, in the rectangular area 38°–43°S and 171°–179°E (the central New Zealand region), with magnitudes greater than 4.0 and depths less than 40 km. They are available from the New Zealand local catalog recorded by the Institute of Geology and Nuclear Sciences (Figure 1), which can be obtained at http://www.geonet.org.nz/. This data set is used by *Zhuang et al.* [2002] to demonstrate their stochastic declustering procedure. Figure 2 plots the occurrence time and the locations of these earthquakes. Visually, it is hard to tell which of the two models is better.

[49] Applying the forward algorithm and the BFGS procedure (see Appendix A), we find the MLEs of the two QHMMs as follows: in the mother-and-kids model $\hat{\gamma}$ = 0.1086, $\hat{\lambda}$ = 2.0787, $\hat{\epsilon}$ = 0.0104, $\hat{d}$ = 0.0031, $\hat{p}$ = 0.3181 and in the domino model $\hat{\gamma}'$ = 0.1084, $\hat{\lambda}'$ = 2.2824, $\hat{\epsilon}'$ = 0.0103, $\hat{d}'$ = 0.0037, $\hat{p}'$ = 0.3214.

[50] We set these MLE as the model parameters and run the forward-backward algorithm. The conditional probabilities of $P(C_t|O_{1:T})$ and $P(C_t'|O_{1:T})$ are calculated for all earthquakes t, $1 \leq t \leq T$. Figure 3a and Figure 3b depict the histograms of $P(C_t|O_{1:T})$ and $P(C_t'|O_{1:T})$, respectively. It can be seen that most counts are concentrated near probability 0 or 1, which indicates that both models can classify events with high statistical confidence. About 3.6% of the events have probability of being in clusters ranging from 0.1 to 0.9 under the domino mother (Figure 2b) and the percentage reduces to 3.4% for the mother-and-kids model (Figure 2a).

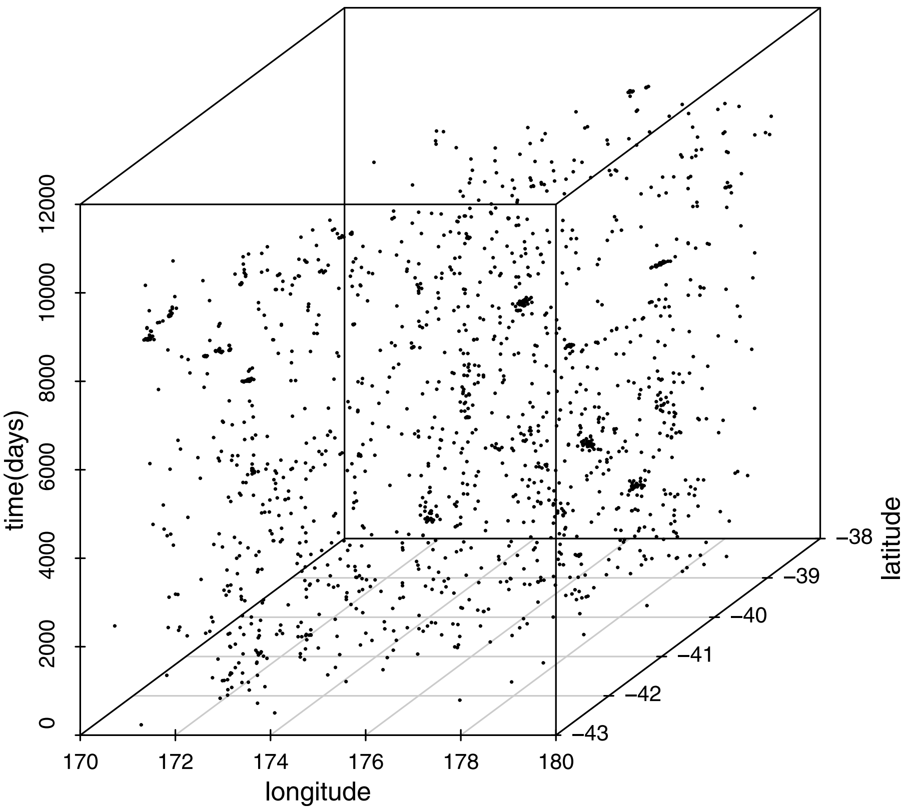

Figure 2. Occurrence time (day 1 is 1 January 1970) and epicentral locations in the central New Zealand region.

[51] For comparison, note that the stochastic declustering procedure proposed by *Zhuang et al.* [2002] is based on the aftershock probabilities calculated under the ETAS model. Figure 5a in the work of *Zhuang et al.* [2002] shows a histogram of those aftershock probabilities. Although that histogram also has most counts near 0 and 1, there are 31.7% of the events having aftershock probability ranging from 0.1 to 0.9. In this sense, the QHMM models give more decisive declustering than ETAS. The possible reasons are discussed by *Wu* [2010].

[52] Although a stochastic declustering is available in the QHMM models, in practice a unique and deterministic declustering is often preferred. We employ the Viterbi algorithm and find the most likely cluster sequence for both models. The results are plotted in Figure 4 and Figure 5 for the mother-and-kids model and the domino model, respectively. The declustering performances of the two models are comparable. In the mother-and-kids model, there are 105 earthquake clusters consisting of 438 events and the remaining 1132 events are identified as single earthquakes,

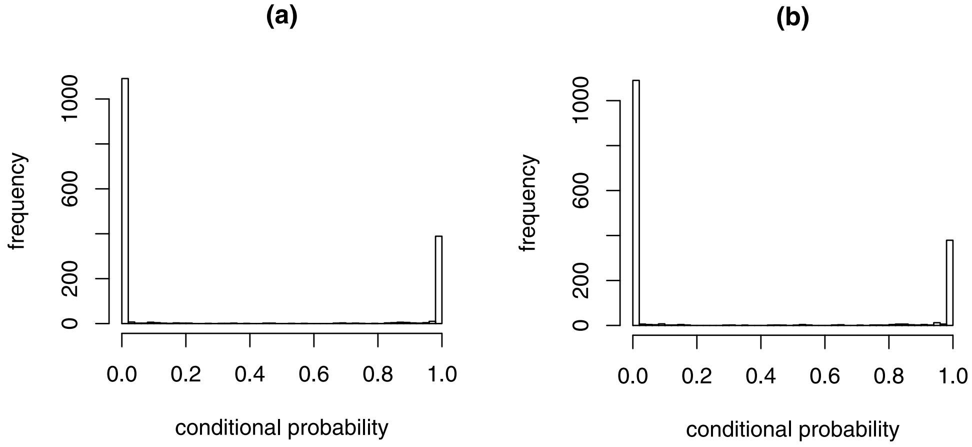

Figure 3. Histograms of conditional probabilities to be in a cluster under (a) the mother-and-kids model and (b) the domino model.

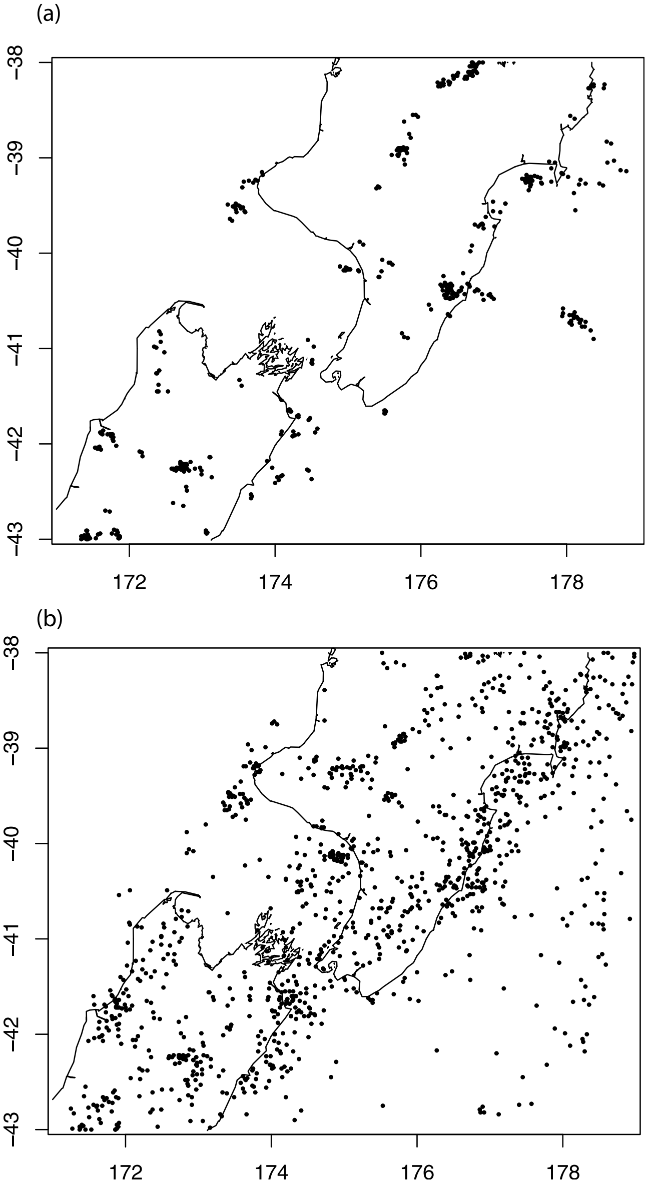

Figure 4. Results from applying Viterbi algorithm to the data set under the mother-and-kids model. (a) Epicentral locations of the most likely earthquake clusters. (b) Epicentral locations of the most likely single earthquakes.

while the domino model picks out 108 clusters that contain 441 cluster earthquakes and 1129 events are single earthquakes. Only 23 earthquakes (about 1.5% of the data) are in different categories under these two declusterings. This is not surprising since both models start from the intuition that the earthquakes within a cluster are near one another.

4.2. G-R Laws

[53] Recall the G-R law for earthquake occurrence:

$$\log_{10} N_M = a - bM$$

where N_M is the number of earthquakes which are greater than M in magnitude. The variations of the b values in different kind of events are of central interest in statistical seismology [*Utsu*, 1966; *Smith*, 1981; *Knopoff et al.*, 1982; *Wyss and Wiemer*, 2000]. For example, *Smith* [1981] observes the statistically significant differences in b values between foreshock and aftershock sequences; hence he proposes b value as an earthquake precursor. However, *Knopoff et al.* [1982] argues that the differences may be due to the windowing algorithm used by *Smith* [1981].

[54] On the basis of the declusterings given by the Viterbi algorithm, we examine the G-R laws for the single earthquakes and cluster earthquakes under these two models.

[55] Figure 6 and Figure 7 plot $\log_{10} N_M$ against M for single earthquakes and cluster earthquakes under the two models, respectively. The maximum likelihood method is used to estimate the b values of the G-R laws [*Shi and Bolt*, 1981]. The results are summarized in Table 1. It shows that under both models the G-R law for the single earthquakes is

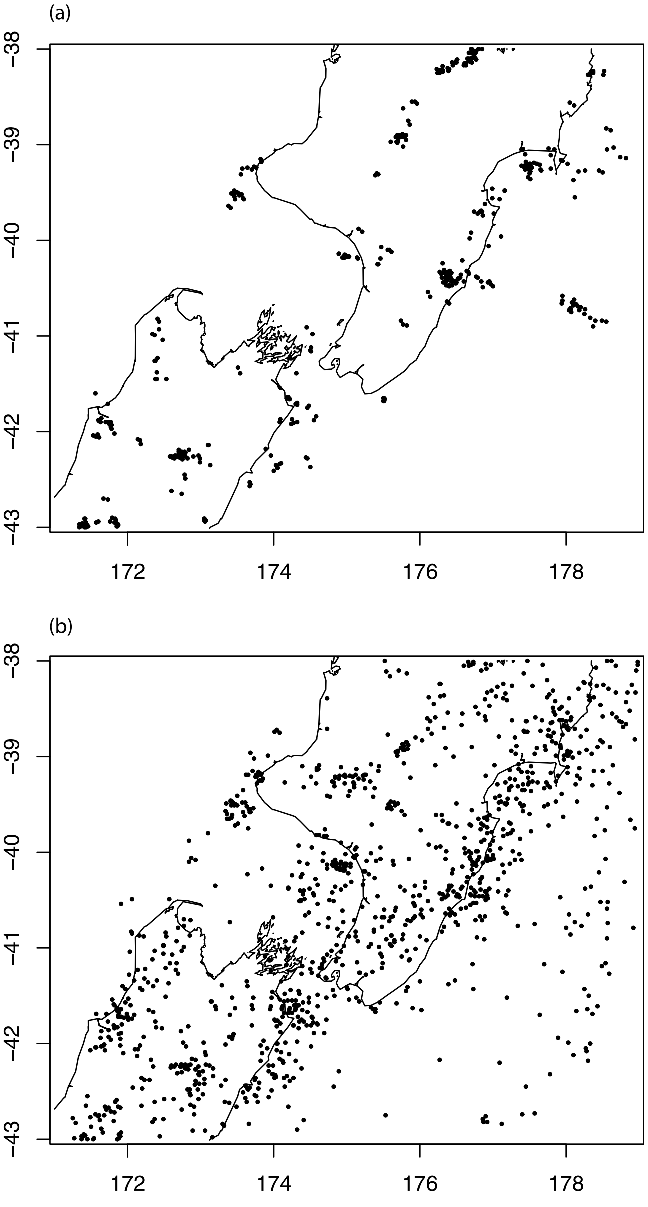

Figure 5. Results from applying Viterbi algorithm to the data set under the domino model. (a) Epicentral locations of the most likely earthquake clusters. (b) Epicentral locations of the most likely single earthquakes.

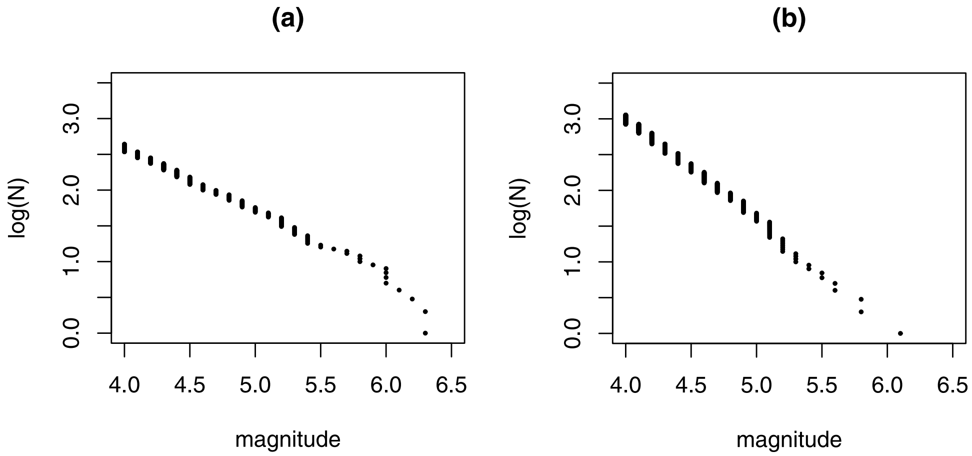

Figure 6. G-R law under the mother-and-kids model: (a) $\log_{10}N_M$ versus M for cluster earthquakes; (b) $\log_{10}N_M$ versus M for single earthquakes.

statistically different from that for the cluster earthquakes. In particular, the *b* values of the single earthquakes are significantly bigger than the *b* values of the cluster earthquakes. This phenomenon is also observed for earthquakes in the central and western Honshu area of Japan [*Wu*, 2010]. Such an observation brings into question the undifferentiated treatment of background events with and without offspring in ETAS models [*Zhuang et al.*, 2002, 2004].

[56] As both models do not incorporate any information of the earthquake magnitude, the resulting different G-R laws for single earthquakes and earthquake clusters can be viewed as evidence of reasonable declustering. In fact, if the models do not make sense at all, then one would expect that the partitioned events resemble two random samples, in which case no statistically significant difference would be found for the two categories' magnitudes.

[57] A full earthquake occurrence model should certainly take into account the magnitudes. However, at the starting stage of modeling when the distributions of magnitudes are not well understood, not including the magnitudes can be an advantage. Here the QHMM models discover two different G-R laws which cannot be seen by the ETAS model because ETAS model a priori assumes that there is one magnitude distribution for all earthquakes.

4.3. Model Selection

[58] Which model fits the data better, the mother-and-kids model or the domino model? The answer helps predict the location of the next aftershock in an earthquake cluster.

[59] The forward algorithm gives log likelihoods $\log(L) = -8087$ for the mother-and-kids model and $\log(L') = -8186$ for the domino model, respectively. The mother-and-kids model has a much greater likelihood since $L'/L = \exp(-99)$ which is essentially 0. Hence the mother-and-kids model fits the data set better according to the maximum likelihood principle. In this sense, the next aftershock is "statistically closer" to the first shock than to the most recent event in the cluster. This provides a partial answer to the question "where will the next aftershock be?".

[60] As both models have five parameters, the same conclusion can be drawn for other model selection criteria such as Akaike information criterion (AIC) [*Akaike*, 1974] and Bayesian information criterion (BIC) [*Schwarz*, 1978]. For example, according to a rule of thumb in the work of

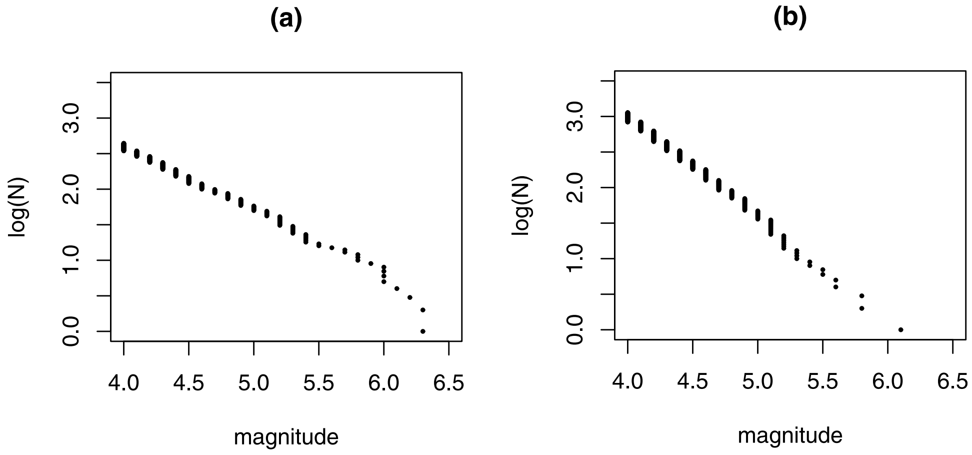

Figure 7. G-R law under the domino model: (a) $\log_{10}N_M$ versus M for cluster earthquakes; (b) $\log_{10}N_M$ versus M for single earthquakes.

Table 1. Maximum Likelihood Estimates of b Values for Cluster Earthquakes and Single Earthquakes Under Mother-and-Kids Model and Under the Domino Model

	$\hat{b}$ (95% C.I.)	
	Mother-and-Kids Model	Domino Model
Earthquakes clusters	1.023 ([0.930, 1.121])	1.023 ([0.930, 1.121])
Single earthquakes	1.589 ([1.498, 1.683])	1.591 ([1.500, 1.686])

Burnham and Anderson [2003], an AIC difference of 20 or above between two models clearly indicates that one model is superior to the other. We have an AIC difference of 198 for these two models.

[61] Another clue to the location of the next aftershock is the MLE $\hat{d} = 0.0031$ and $\hat{d}' = 0.0037$. Recall that d and d' are the variance parameters of bivariate Gaussian distributions, which can be considered a measure of distance. Since $\hat{d} < \hat{d}'$, it too corroborates the conclusion that the next aftershock will be closer to the first-shock than to the most recent earthquake in the cluster.

5. Discussion

[62] HMMs have been applied successfully in areas as diverse as speech recognition, gene finding, and financial data analysis, but only recently have a few HMM applications appeared in the seismological literature [*Granat and Donnellan*, 2002; *Ebel et al.*, 2007]. The QHMM introduced in this paper is a more flexible framework and it shares the computational feasibility of the HMM. Large amounts of data are available in seismology, and often the mechanism behind the data needs to be understood. This provides a perfect setting for QHMMs, as one can incorporate various geophysical hypotheses into QHMMs and have them tested. This paper illustrates this idea.

[63] The mother-and-kids and domino models have rather different structural assumptions to the ETAS model. The background events or the immigrants in the ETAS model are all potential "ancestors." The models discussed in this paper have two classes of immigrants. The first are infertile and do not have offspring, whereas the second class are fertile and have offspring. In the ETAS model each offspring then becomes a mother, and it can also have offspring in exactly the same way as its mother. This does not happen in the two models discussed in this paper. Only the original immigrant fertile "ancestor" events have children. Her offspring are infertile.

[64] In the ETAS model, a mother theoretically never becomes infertile. She just produces fewer offspring over time according to a power law decay (Omori law). Eventually the likelihood is so low (but nonzero) that she is effectively infertile. In the two models of this paper, the mother can become abruptly infertile at the birth of each child. Once infertile, she stays infertile, and so the aftershock sequence stops.

[65] In the ETAS model, the location spatial density of an offspring event is bound in some way (could be centered but not necessarily) to the location of its mother event (not the original ancestor). In the mother-and-kids model it is bound to the location of the mother, who is like an ancestor in the ETAS model, in that she is an immigrant into the system. The mother in the ETAS case could be a number of generations from the original immigrant event. In the domino model, the birth location of a child is bound to the location of its sibling born immediately before it. It is as though the mother is moving around, so the best predicted location of the next child is somewhere close to where she had the last child.

[66] The two models used in this paper can be generalized in many ways. For instance, by letting the hidden state include more indicator functions, the assumption of no more than one active cluster at a time can be relaxed. In the two models we assume for simplicity that while the mother is fertile, her fertility rate is constant. This assumption can be generalized so that the fertility rate decays over time as it does under the ETAS model. Also although exponential distribution are commonly used to model the interval time between earthquakes, the usage of other alternative distributions is straightforward in the QHMM framework. Another extension is to build the magnitude of earthquakes into the models.

[67] Actually, the ETAS model can be considered a submodel of a QHMM. The reason is that a QHMM framework can introduce multiple point processes into the model, whereas ETAS has only one. Take the ETAS model in the work of *Zhuang et al.* [2002], for example. It is a branching process with immigrants, and the subprocess composed of immigrants without offspring is independent of the rest of the events. Now, if this subprocess is allowed to have a different intensity from that of other immigrants, then we have two independent processes for the observations and a QHMM can be built. In the event that the parameters in the corresponding intensity functions coincide, it is reduced to the original ETAS model. However, we need to note that the generality of the QHMM does not come for free. It is often more difficult to set up and program a QHMM, and the computational complexity of the associated algorithms can be high.

[68] Cluster models (or seismicity-based models) such as ETAS are often used for seismic hazard assessments in regions (such as Japan) where earthquakes are relatively deep and the fault structure is not clearly identified. But for regions like Southern California where the fault structure is well known, fault-based models are more suitable [see *Holliday et al.*, 2008, and references therein]. The fault information can also be incorporated into a QHMM. Future work is to be done in this direction.

[69] The scheme of employing QHMMs to make statistical inferences on cluster growing patterns can also be applied in other areas. For instance, an important question in epidemiology is how an infectious disease spreads. With a related dataset, different infection hypotheses can be compared statistically.

Appendix A: Quasi-Hidden Markov Model

A1. Hidden Markov Model

[70] In a basic HMM [*Rabiner*, 1989; *Granat and Donnellan*, 2002], at time t, $t = 1, 2, \ldots, T$, one observes O_t while the actual state q_t is hidden. The hidden state q_t takes values in a finite discrete state space $\{S_1, S_2, \ldots, S_N\}$ and the

sequence q_1, q_2, … forms a first order homogeneous Markov chain with transition probability $a_{ij} = P(q_{t+1} = S_j|\mathrm{q}_t = S_i)$. The observation O_t is independent of everything else conditional on the hidden state q_t, and O_t follows a distribution $b_j(\cdot)$ when the actual state q_t is at S_j.

[71] Several algorithms facilitate the statistical inferences on a HMM. The forward algorithm [*Baum et al.*, 1970] computes the likelihood function $P(O_{1:T})$ ($O_{1:T}$ denotes $(O_1, O_2, \ldots, O_T)$ for short) of the model. Combining with a backward procedure, it leads to the forward-backward algorithm, which calculates the probability of being in state S_i at time t given the observation sequence, i.e., $P(q_t = S_i|\mathrm{O}_{1:T})$. The Viterbi algorithm [*Viterbi*, 1967] finds the most likely hidden state sequence (path). And the parameters of the model can be estimated by the maximum likelihood principle via the EM algorithm [*Dempster et al.*, 1977].

[72] Allowing the Markov chain formed by the hidden state sequence q_1, q_2, … to be nonhomogeneous, *Hughes and Guttorp* [1994] proposes an extension of the basic HMM, namely the Nonhomogeneous Hidden Markov Model (NHMM). The forward-backward algorithm, the Viterbi algorithm and the EM algorithm can be similarly applied to NHMM [*Diebolt et al.*, 1994; *Hughes et al.*, 1999].

A2. Definition of QHMM

[73] Likewise, we let O_t and q_t denote the observation and the hidden state at time t, respectively. However, the hidden state q_t is allowed to take values in a time-varying finite discrete space $\mathbf{S_t} = \{S_1, S_2, \ldots, S_{N_t}\}$. The observations $\{O_t\}$ and the hidden states $\{q_t\}$ are said to form a QHMM if the conditional probability distributions satisfy $P(O_{t+1}, q_{t+1}|\mathrm{O}_{1:t}, q_{1:t}) = P(O_{t+1}, q_{t+1}|\mathrm{O}_{1:t}, q_t)$. Note that the assumptions of HMM and NHMM imply $P(O_{t+1}, q_{t+1}|\mathrm{O}_{1:t}, q_{1:t}) = P(O_{t+1}, q_{t+1}|\mathrm{q}_t)$, hence HMM and NHMM are special cases of the QHMMs.

[74] Next we describe the foward-backward algorithm and the Viterbi algorithm associated with a QHMM, adopting the notations in the work of *Rabiner* [1989]. The detailed derivation of the algorithms can be found in the work of *Wu* [2011].

A3. Forward-Backward Algorithm

[75] We define the forward variables $\alpha_t(i)$ as

$$\alpha_t(i) = P(O_{1:t}, q_t = S_i), \; 1 \leq i \leq N_t$$

Then it is not hard to see that

$$\alpha_{t+1}(i) = \sum_{j=1}^{N_t} P(O_{t+1}, q_{t+1} = S_i|O_{1:t}, q_t = S_j)\alpha_t(j), \quad 1 \leq i \leq N_{t+1}. \tag{A1}$$

Hence one can compute $\alpha_t(i)$ inductively after setting $\alpha_1(i) = \pi(O_1, q_1 = S_i)$, $1 \leq i \leq N_1$, where $\pi(\cdot,\cdot)$ is the initial distribution of (O_1, q_1). The forward variables can be used in the computation of the likelihood of the model: $P(O_{1:T}) = \sum_{i=1}^{N_T} \alpha_T(i)$.

[76] In a similar manner, the backward variables $\beta_t(i)$ are defined as

$$\beta_t(i) = P(O_{t+1:T}|O_{1:t}, q_t = S_i), \quad 1 \leq i \leq N_t.$$

Note that the recursive relationship

$$\beta_t(i) = \sum_{j=1}^{N_{t+1}} P\big(O_{t+1}, q_{t+1} = S_j|O_{1:t}, q_t = S_i\big)\beta_{t+1}(j), \quad 1 \leq i \leq S_t \tag{A2}$$

holds. Hence setting $\beta_T(i) = 1$, $1 \leq i \leq S_T$, the remaining $\beta_t(i)$ can be computed backward in time inductively.

[77] Combining the forward procedure and the backward procedure, the probability of the hidden state is S_i at time t conditional on the observations is given by

$$P(q_t = S_i|O_{1:T}) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N_t} \alpha_t(j)\beta_t(j)}.$$

A4. Viterbi Algorithm

[78] The Viterbi algorithm uses dynamic programming to find the most likely hidden state sequence, namely, $\mathrm{argmax}_{q_1,q_2,\ldots,q_T} P(q_1,q_2,\ldots,q_T|\mathrm{O}_{1:T})$. We define the quantity $\delta_t(i) = \max_{q_1,\ldots,q_{t-1}} P(O_{1:t}, q_1,\ldots,q_{t-1}, q_t = S_i)$ and observe that

$$\delta_{t+1}(i) = \max_j \big[P\big(O_{t+1}, q_{t+1} = S_i|O_{1:t}, q_t = S_j\big)\delta_t(j)\big]. \tag{A3}$$

We also define $\psi_{t+1}(i) = \arg\max_j [P(O_{t+1}, q_{t+1} = S_i|\mathrm{O}_{1:t}, q_t = S_j)\, \delta_t(j)]$ to keep track the index that maximizes (A3). Let $q_T^* = S_{\mathrm{argmax}_j\, \delta_T(j)}$ and $q_t^* = S_{\psi_{t+1}\,(q_{t+1}^*)}$ for $t = T - 1$, $T - 2$, …, 1, then one can show that $\{q_1^*, q_2^*, \ldots, q_T^*\}$ is the most likely hidden state sequence, i.e., $P(q_1^*,q_2^*,\ldots,q_T^*|\mathrm{O}_{1:T}) = \max_{q_1,q_2,\ldots,q_T} P(q_1,q_2,\ldots,q_T|\mathrm{O}_{1:T})$.

A5. Search for MLE

[79] The EM algorithm typically does not apply to a QHMM. However, since the forward algorithm computes the likelihood function of a QHMM efficiently, most standard optimization procedures are applicable to find the maximum likelihood estimators (MLE) of the parameters. For instance, we use the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method [*Nocedal and Wright*, 2006] in the numerical study in section 4.

Appendix B: Asymptotics of the Model Selection

[80] Likelihood-based model selection criteria such as AIC and BIC are known to be effective for independent, identically distributed (i.i.d.) observations. They are often justified by their asymptotic properties. It is natural to ask what if the observations are from a HMM or a QHMM. A series of papers by *Baum and Petrie* [1966], *Leroux* [1992], *Bickel et al.* [1998, 2002], *Jensen and Petersen*, [1999], and *Douc and Matias* [2001] study the asymptotics of the likelihood inferences for HMMs. The general conclusion is: “HMM likelihoods and their derivatives behave like i.i.d ones” (title of *Bickel et al.* [2002]). Here we show that it is also true for the two QHMM introduced in this paper: mother-and-kids model and the domino model.

[81] Let $(\Omega, \mathcal{B}, P)$ be a probability space and suppose that $\{X_1, X_2, \ldots\}$ are i.i.d. with each X_i taking values in a standard Borel space. Clearly, the marginal distribution P_1 for

X_1 determines the joint distribution P_n for $(X_1, X_2, \ldots, X_n)$, i.e. $P_n = P_1 \times P_1 \times \ldots \times P_1$. Assume that P_1 is absolutely continuous with respect to a sigma-finite measure M_1 and $f = \frac{dP_1}{dM_1}$. Let $L_T(X_1, \ldots, X_n)$ denote the likelihood function under the true model, then by the strong law of large numbers,

$$\frac{1}{n}\log L_T(X_1, X_2, \ldots, X_n) = \frac{1}{n}\log\prod_{i=1}^{n} f(X_i) = \frac{1}{n}\sum_{i=1}^{n}\log f(X_i)$$
$$\rightarrow E\log f(X_1), P - a.s..$$

Suppose that one specifies a family of parametric density functions $\{g_\theta : \theta \in \Theta\}$ to approximate the marginal distribution density f. Let $L_\theta(X_1, \ldots, X_n)$ denote the likelihood function under such a model, then

$$\frac{1}{n}\log L_\theta(X_1, X_2, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n}\log g_\theta(X_i) \rightarrow E\log g_\theta(X_1), P - a.s..$$

Jensen's inequality implies that E log $f(X_1) - E$ log $g_\theta(X_1)$ is nonnegative for any θ. In fact, let $dQ_1^\theta = g_\theta dM_1$, E log $f(X_1)-$ E log $g_\theta(X_1)$ can be recognized as the Kullback-Leibler divergence between P_1 and Q_1^θ [*Kullback and Leibler*, 1951]. Therefore maximum likelihood estimation is asymptotically equivalent to finding the model closest to the true model based on their Kullback-Leibler divergence. This is the starting point of the classical arguments on the consistency and asymptotic normality of MLE.

[82] The observations O_1, O_2, … in section 3 are certainly not i.i.d., but an analogous result holds if O_1, O_2, … are assumed to be stationary and ergodic. Recall that $O_i = (x_i, y_i, t_i)$ takes values in $\mathcal{O} = R \times [0,\infty)$, where R is the two-dimensional region under study. A convenient reference measure is the Lesbesgue measure, which is used to describe the mother-and-kids model and the domino model in section 2. Assume that the true model admitsdensities $p_0(o_1, \ldots, o_n)$ relative to the Lebesgue measure on $\mathcal{O} \times \mathcal{O} \times \ldots \times \mathcal{O}$. Then the generalized Shannon-McMillan-Breiman theorem [*Barron*, 1985; *Algoet and Cover*, 1988] asserts that

$$\frac{1}{n}\log p_0(O_1, O_2, \ldots, O_n) = \frac{1}{n}\log\prod_{i=1}^{n} p_0(O_i|O_{i-1}, \ldots, O_1)$$
$$= \frac{1}{n}\sum_{i=1}^{n}\log p_0(O_i|O_{i-1}, \ldots, O_1)$$
$$\rightarrow H_0, P - a.s.$$

where $H_0 = \lim_{n\to\infty}E\{\log\ p_0(O_n|\mathrm{O}_{n-1}, \ldots, O_1)\}$ is the relative entropy rate of the process $\{O_i\}$.

[83] While the true model will never be known, the mother-and-kids model and the domino model serve as two possible approximations. Let p_{m_θ} $(o_1,\ldots,o_n)$ denote the density (i.e., the likelihood) of the mother-and-kids model, and $p_{d_{\theta'}}$ $(o_1,\ldots,o_n)$ denote the density of the domino model, where θ and θ' are the parameter vectors. Then one can correspondingly have

$$\frac{1}{n}\log p_{m_\theta}(O_1, O_2, \ldots, O_n) = \frac{1}{n}\sum_{i=1}^{n}\log p_{m_\theta}(O_i|O_{i-1}, \ldots, O_1)$$
$$\rightarrow H_{m_\theta}, P - a.s., \qquad \text{(B1)}$$

where $H_{m_\theta} = \lim_{n\to\infty} E\{\log\ p_{m_\theta}\ (O_n|\mathrm{O}_{n-1},\ldots,O_1)\}$, and

$$\frac{1}{n}\log p_{d_{\theta'}}(O_1, O_2, \ldots, O_n) = \frac{1}{n}\sum_{i=1}^{n}\log p_{d_{\theta'}}(O_i|O_{i-1}, \ldots, O_1)$$
$$\rightarrow H_{d_{\theta'}}, P - a.s., \qquad \text{(B2)}$$

where $H_{d_{\theta'}} = \lim_{n\to\infty} E\{\log\ p_{d_{\theta'}}\ (O_n|\mathrm{O}_{n-1},\ldots,O_1)\}$.

[84] Likewise, Jensen's inequality implies that $H_0 - H_{m_\theta}$ (or $H_0 - H_{d_{\theta'}}$) are nonnegative for any θ (or θ'). This difference between the relative entropy rates quantifies how far the proposed model deviates from the truemodel.

[85] It is reasonable to assume the parameter space Θ is compact, as one can select a very small $\delta > 0$ (for example, δ can be the smallest positive number a computer can store) and let the parameters γ, λ, ϵ, d be in the range of $[\delta, 1/\delta]$ and $p \in [\delta, 1 - \delta]$.

[86] The theorem states

[87] If O_1, O_2, … is stationary and ergodic, and the parameter space Θ is compact as above, then $\sup_\theta H_{m_\theta} >$ $\sup_{\theta'} H_{d_{\theta'}}$ implies that

$$\mathbf{1}_{\left\{\sup_\theta p_{m_\theta}(O_1,O_2,\ldots,O_n) > \sup_{\theta'} p_{d_{\theta'}}(O_1,O_2,\ldots,O_n)\right\}} \rightarrow 1, P - a.s.$$

and similarly, $\sup_\theta H_{m_\theta} < \sup_{\theta'} H_{d_{\theta'}}$ implies that

$$\mathbf{1}_{\left\{\sup_\theta p_{m_\theta}(O_1,O_2,\ldots,O_n) < \sup_{\theta'} p_{d_{\theta'}}(O_1,O_2,\ldots,O_n)\right\}} \rightarrow 1, P - a.s.$$

The proof of this theorem follows the standard argument in the work of *Wald* [1949], with (4) and (5) taking the place of the strong law of large numbers. The conditions in the work of *Wald* [1949, p. 596] can be verified in our setting easily. This proposition formulates the strong consistency of the model selection procedure used in this paper: as the number of observations goes to infinity, the procedure picks out the model that is closer to the true model with probability one.

[88] **Acknowledgments.** This work is jointly supported by NUS grants R-155-000-070-133 and R-155-000-105-112. The author benefits from many helpful discussions with Thomas G. Kurtz. The author also thanks Jiancang Zhuang who kindly provided the data set and Ying Luo and David Harte for the helpful suggestions.

References

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

Algoet, P. H., and T. M. Cover (1988), A sandwich proof of the Shannon-McMillan-Breiman theorem, *Ann. Prob.*, *16*(2), 899–909.

Barron, A. R. (1985), The strong ergodic theorem for densities: Generalized Shannon-McMillan-Breiman theorem, *Ann. Prob.*, *13*(4), 1292–1303.

Baum, L. E., and T. Petrie (1966), Statistical inference for probabilistic functions of finite state Markov chains, *Ann. Math. Stat.*, *37*, 1554–1563.

Baum, L. E., T. Petrie, G. Soules, and N. Weiss (1970), A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, *Ann. Math. Stat.*, *41*, 164–171.

Bickel, P. J., Y. Ritov, and T. Rydén (1998), Asymptotic normality of the maximum likelihood estimator for general hidden Markov models, *Ann. Stat.*, *26*, 1614–1635.

Bickel, P. J., Y. Ritov, and T. Rydén (2002), Hidden Markov model likelihoods and their derivatives behave like I.I.D ones, *Ann. Inst. Henri Poincaré*, *38*(6), 825–846.

Burnham, K. P., and D. R. Anderson (2003), *Model Selection and Multimodel Inference*, 2nd ed., Springer, New York.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977), Maximum likelihood from incomplete data via the EM algorithm, *J. R. Stat. Soc., Ser. B*, *39*(1), 1–38.

Diebolt, F. X., J. H. Lee, and G. C. Weinbach (1994), Regime switching with time varying transition probabilities, in *Nonstationary Time Series*

Analysis and Cointegration, edited by C.P. Hargreaves, pp. 283–302, Oxford Univ. Press, Oxford, U.K.

Douc, R., and C. Matias (2001), Asymptotics of the maximum likelihood estimator for general hidden Markov models, *Bernoulli*, *7*(3), 381–420.

Ebel, J., D. Chambers, A. Kafka, and J. Baglivo (2007), Non-poissonian earthquake clustering and the hidden Markov model as bases for earthquake forecasting in California, *Seismol. Res. Lett.*, *78*(1), 57–65.

Granat, R., and A. Donnellan (2002), A hidden markov model based tool for geophysical data exploration, *Pure Appl. Geophys.*, *159*(10), 2271–2283.

Harris, R. A. (2001), Stress triggers, stress shadows, and seismic hazard, in *International Handbook of Earthquake and Engineering Seismology*, edited by W. H. K. Lee et al., chap. 73, pp. 1217–1232, Chapman and Hall, New York.

Helmstetter, A. (2003), Is earthquake triggering driven by small earthquakes?, *Phys. Rev. Lett.*, *91*, 058501.

Helmstetter, A., and D. Sornette (2002), Subcritical and supercritical regimes in epidemic models of earthquake aftershocks, *J. Geophys. Res.*, *107*(B10), 2237, doi:10.1029/2001JB001580.

Helmstetter, A., and D. Sornette (2003), Predictability in the ETAS model of interacting triggered seismicity, *J. Geophys. Res.*, *108*(B10), 2482, doi:10.1029/2003JB002485.

Holliday, J. R., D. L. Turcotte, and J. B. Rundle (2008), A review of earthquake statistics: Fault and seismicity-based models, ETAS and BASS, *Pure Appl. Geophys.*, *165*, 1003–1024.

Hughes, J. P., and P. Guttorp (1994), A class of stochastic models for relating synoptic atmospheric patterns to regional hydrologic phenomena, *Water Resour. Res.*, *30*, 1535–1546.

Hughes, J. P., P. Guttorp, and S. P. Charles (1999), A non-homogeneous hidden Markov model for precipitation occurrence, *J. R. Stat. Soc., Ser. C*, *48*, 15–30.

Jensen, J., and N. Petersen (1999), Asymptotic normality of the maximum-likelihood estimator in state space models, *Ann. Stat.* , *27*, 514–535.

Kagan, Y. Y., and L. Knopoff (1981), Stochastic synthesis of earthquake catalogs, *J. Geophys. Res.*, *86*, 2853–2862.

Knopoff, L. (2000), The magnitude distribution of declustered earthquakes in southern California, *Proc. Natl. Acad. Sci. U.S.A.*, *97*(22), 11,880–11,884.

Knopoff, L., Y. Y. Kagan, and R. Knopoff (1982), b Values for foreshocks and aftershocks in real and simulated earthquake sequences, *Bull. Seismol. Soc. Am.*, *72*(5), 1663–1676.

Kullback, S., and R. A. Leibler (1951), On information and sufficiency, *Ann. Math. Stat.*, *22*(1), 79–86.

Leroux, B. G. (1992), Maximum-likelihood estimation for hidden Markov models, *Stochastic Proc. Appl.*, *40*, 127–143.

Nocedal, J., and S. J. Wright (2006), *Numerical Optimization*, 2nd ed., Springer, New York.

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

Ogata, Y., and J. Zhuang (2006), Space-time ETAS models and an improved extension, *Tectonophysics*, *413*(1–2), 13–23.

Rabiner, L. R. (1989), A tutorial on hidden Markov models and selected applications in speech recognition, *Proc. IEEE*, *77*(2), 257–286.

Rathbun, S. L. (1993), Modeling marked spatio-temporal point patterns, *Bull. Int. Stat. Inst.*, *55*(2), 379–396.

Schwarz, G. E. (1978), Estimating the dimension of a model, *Ann. Stat.*, *6*(2), 461–464.

Shi, Y., and B. A. Bolt (1981), The standard error or the magnitude-frequency *b* value, *Bull. Seismol. Soc. Am.*, *72*(5), 1677–1687.

Smith, W. D. (1981), The *b*-values as an earthquake precursor, *Nature*, *289*, 136–139.

Utsu, T. (1966), A statistical significance test of the difference in *b*-value between two earthquake groups, *J. Phys. Earth.*, *14*, 37–40.

Vere-Jones, D., and J. Zhuang (2008), On the distribution of the largest event in the critical ETAS model, *Phys. Rev. E*, *78*, 047102, doi:10.1103/PhysRevE.78.047102.

Viterbi, A. J. (1967), Error bounds for convolutional codes and an asymptotically optimum decoding algorithm, *IEEE Trans. Inf. Theory*, *13*, 260–269.

Wald, X. (1949), Note on the consistency of the maximum likelihood estimate, *Ann. Math. Stat.*, *20*(4), 595–601.

Wang, K., Q. Chen, S. Sun, and A. Wang (2006), Predicting the 1975 Haicheng earthquake, *Bull. Seismol. Soc. Am.*, *96*, 757–795, doi:10.1785/0120050191.

Wu, Z. (2009), A cluster identification framework illustrated by a filtering model for earthquake occurrences, *Bernoulli*, *15*(2), 357–379.

Wu, Z. (2010), A hidden Markov model for earthquake declustering, *J. Geophys. Res.*, *115*, B03306, doi:10.1029/2008JB005997.

Wu, Z. (2011), Quasi hidden Markov model and its applications in multiple-change-point problems, technical report, Natl. Univ. of Singapore, Singapore.

Wyss, M., and S. Wiemer (2000), Change in the probability for earthquakes in southern California due to the Landers magnitude 7.3 earthquake, *Science*, *290*, 1334–1338.

Zhuang, J. (2006), Second-order residual analysis of spatiotemporal point processes and applications in model evaluation, *J. R. Stat. Soc., Ser. B*, *68*(4), 635–653, doi:10.1111/j.1467-9868.2006.00559.

Zhuang, J., Y. Ogata, and D. Vere-Jones (2002), Stochastic declustering of space-time earthquake occurrences, *J. Am. Stat. Assoc.*, *97*, 369–380.

Zhuang, J., Y. Ogata, and D. Vere-Jones (2004), Analyzing earthquake clustering features by using stochastic reconstruction, *J. Geophys. Res.*, *109*, B05301, doi:10.1029/2003JB002879.

Z. Wu, Department of Statistics and Applied Probability, National University of Singapore, 6 Science Drive 2, 117546, Singapore. (stawz@nus.edu.sg)