Stoch Environ Res Risk Assess (2018) 32:1415–1434
https://doi.org/10.1007/s00477-017-1457-1

ORIGINAL PAPER

A hidden Markov model for earthquake prediction

Cheuk Fung Yip[1] · **Wai Leong Ng**[2] · **Chun Yip Yau**[2]

Published online: 21 September 2017

Abstract Earthquake occurrence is well-known to be associated with structural changes in underground dynamics, such as stress level and strength of electromagnetic signals. While the causation between earthquake occurrence and underground dynamics remains elusive, the modeling of changes in underground dynamics can provide insights on earthquake occurrence. However, underground dynamics are usually difficult to measure accurately or even unobservable. In order to model and examine the effect of the changes in unobservable underground dynamics on earthquake occurrence, we propose a novel model for earthquake prediction by introducing a latent Markov process to describe the underground dynamics. In particular, the model is capable of predicting the change-in-state of the hidden Markov chain, and thus can predict the time and magnitude of future earthquake occurrences simultaneously. Simulation studies and applications on a real earthquake dataset indicate that the proposed model successfully predicts future earthquake occurrences. Theoretical results, including the stationarity and ergodicity of the proposed model, as well as consistency and asymptotic normality of model parameter estimation, are provided.

Keywords Change-point · ETAS model · EM algorithm · Structural break

1 Introduction

Earthquakes constitute a natural disaster that threatens the lives of millions of people worldwide. It is crucial to analyze earthquake data in order to help elucidate earthquake mechanisms and protect lives through a reliable earthquake early warning system (Cameletti et al. 2016; Finazzi and Fassò 2016).

In seismology, earthquake data are commonly recorded in the form of an *earthquake catalogue*, which includes the exact date and time of the occurrence, as well as the latitude, longitude, depth, and magnitude of earthquakes in a particular region. A typical earthquake catalogue is shown in Fig. 1. In particular, earthquakes are only detectable by seismometer stations if their magnitudes are larger than a certain threshold, referred to as a *magnitude of completeness*. This threshold mainly depends on the accuracy and working mechanism of the seismometers, as well as the geology around the stations.

One important feature typically observed in an earthquake catalogue is that the earthquakes are clustered, as shown in Fig. 2. Sometimes, there is a sequence of frequent arrivals of earthquakes; whereas, at other times, it undergoes a period of fewer arrivals, called seismic quiescence. Although earthquake clustering is typically difficult to classify, it is possibly explained by structural changes in some underground dynamics. Reid (1910) proposed the elastic-rebound theory in an attempt to elucidate

Electronic supplementary material The online version of this article (doi:10.1007/s00477-017-1457-1) contains supplementary material, which is available to authorized users.

✉ Cheuk Fung Yip
1155015637@link.cuhk.edu.hk

Wai Leong Ng
1009613922@link.cuhk.edu.hk

Chun Yip Yau
cyyau@sta.cuhk.edu.hk

[1] Department of Medicine and Therapeutics, Chinese University of Hong Kong, Shatin, Hong Kong, China

[2] Department of Statistics, Chinese University of Hong Kong, Shatin, Hong Kong, China

Fig. 1 An example of an earthquake catalogue

"Date"	"Time"	"Magnitude"	"Latitude"	"Longitude"	"Depth"
"2010/01/01"	"02:27:45.96"	2.18	32.471	-115.214	6
"2010/01/01"	"02:33:42.82"	3.24	32.454	-115.203	6
"2010/01/01"	"02:55:04.37"	2.79	35.985	-117.298	0.7
"2010/01/01"	"03:25:30.18"	2.86	36.032	-117.783	0.7
"2010/01/01"	"06:52:51.40"	2.56	32.465	-115.215	6
"2010/01/01"	"07:36:17.14"	2.06	32.49	-115.193	6
"2010/01/01"	"07:40:08.08"	2.48	32.456	-115.198	6
"2010/01/01"	"07:52:00.46"	2.22	32.492	-115.2	6
"2010/01/01"	"09:30:55.04"	2.11	32.488	-115.273	6
"2010/01/01"	"14:06:44.28"	2.78	32.447	-115.208	6
"2010/01/01"	"15:43:02.44"	2.49	32.268	-115.335	6
"2010/01/01"	"17:51:53.43"	2.64	32.362	-115.251	6
"2010/01/02"	"03:15:45.20"	2.81	33.576	-118.889	11.2
"2010/01/02"	"04:08:49.18"	2.11	34.58	-119.846	4.7
"2010/01/02"	"14:04:21.08"	2.02	34.67	-116.336	4.8
"2010/01/02"	"22:37:14.14"	2.38	32.545	-115.23	6
"2010/01/02"	"23:11:22.50"	2.04	32.938	-115.554	15.1
"2010/01/03"	"07:04:44.23"	2.78	32.472	-115.209	6
"2010/01/03"	"17:25:52.94"	2.93	32.545	-115.239	6
"2010/01/03"	"17:56:19.70"	2.09	32.55	-115.234	6

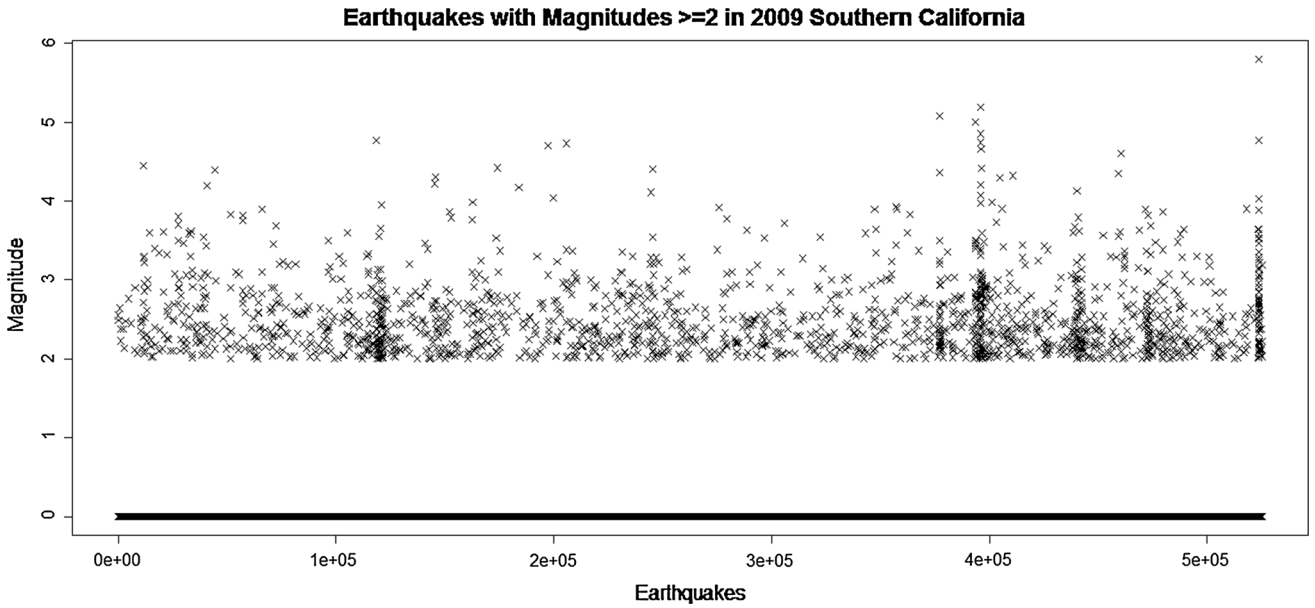

Fig. 2 Earthquakes with magnitudes $\geqslant M_{min} = 2$ in 2009 Southern California, where M_{min} is the magnitude of completeness of the catalogue

earthquake mechanisms. In particular, stress builds up slowly in underground faults due to tectonic plate movement. Earthquakes occur when the accumulated stress exceeds a certain threshold, and the energy is then released. Therefore, the threshold divides the underlying stress level into two states that correspond to high or low frequency of earthquakes, respectively. However, while earthquakes are observable, the underground stress level is difficult to measure. Hence, to model the earthquake mechanism, the stress level should be modeled as a hidden process that switches between the two states. As the Markov chain is commonly used to model state transitions, the hidden Markov model (HMM) is a natural candidate for modeling earthquake data.

The HMM involves a stochastic process $\{(Y_n, S_n), n \in \mathbb{N} = \{1, 2, 3, \ldots\}\}$, where $\{Y_n\}$ represents an observed process and $\{S_n\}$ is a hidden process that governs the distribution of $\{Y_n\}$. In particular, $\{S_n\}$ is a Markov chain with state space $\mathcal{S}$ and transition probability matrix $\mathcal{P}$. For $s, s' \in \mathcal{S}$, the (s, s') entry of $\mathcal{P}$, denoted by $\mathcal{P}(s, s')$, is the probability of going from state s to s'. Note that $\sum_{s' \in \mathcal{S}} \mathcal{P}(s, s') = 1$. Given the hidden Markov chain $\{S_n\}$, $\{Y_n\}$ is a sequence of conditional independent random variables with probability density $f(Y_n|S_n)$, which is known as state-dependent distribution. In the literature, different forms of state dependent distribution $f(Y_n|S_n)$ have been considered. For example, Ebel et al. (2007) modeled the inter-arrival time between two earthquakes

(Y_n) by an exponential distribution; Orfanogiannaki et al. (2010) employed a Poisson model for earthquake arrival frequency (Y_n). These HMMs are mainly designed for explaining earthquake clustering, and information about earthquake magnitude is not taken into account. Moreover, in these models, no structure on the hidden Markov process is assumed, i.e., the transition probability matrix $\mathcal{P}$ of the hidden Markov process $\{S_n\}$ is constant. As a result, they lack the power to predict the future hidden states, and thus the frequency and magnitude, of future earthquakes.

In this paper, we develop a novel HMM for earthquake modeling and forecasting. We use earthquake amplitude as the observation process, and thus the arrival time and magnitudes of earthquakes can be modeled simultaneously. More importantly, we propose a dynamic structure for the hidden Markov process. Specifically, the transition probabilities are modeled by possibly time-varying covariates and past observations. This enables prediction of future hidden states, and thus future earthquakes' frequency and magnitude.

The paper is organized as follows. Section 2 introduces the model and derives the likelihood function. Section 3 discusses computational issues involved in the maximum likelihood estimation. Section 4 introduces an algorithm for predicting future earthquake occurrences. Section 5 reports the results of simulation experiments. Section 6 presents an application to Southern California earthquake occurrences. The proof of strict stationarity, ergodicity of the model, and the consistency and asymptotic normality of the maximum likelihood estimation are given in the online Supplementary Material.

2 Model description

In an earthquake catalogue, we observe the magnitude $\{A_n\}$ where $n = 1, 2, \ldots$, is the time index. Earthquake occurrence is defined as an earthquake with A_n greater than the magnitude of completeness M_{min}. In other words, A_n equals zero when no earthquake occurs, or a value greater than M_{min} when an earthquake occurs. As the occurrence of an event is naturally modeled by a Bernoulli random variable, A_n can be modeled by the product of a left-truncated exponential random variable and a Bernoulli random random variable as

$$A_n = M_n \mathbb{I}_n, \tag{1}$$

where

$$M_n \sim \text{Left} - \text{truncated}\exp(\lambda_n, \text{M}_{\text{min}}), \quad \mathbb{I}_n \sim \text{Bernoulli}(\pi_n), \tag{2}$$

(Ogata 1988, 1998). The rate of the exponential distribution λ_n controls the magnitude, and the probability of earthquake occurrence π_n controls the frequency of earthquake occurrence.

Suggested by the elastic-rebound theory of Reid (1910) that the unobservable stress level can be divided into two states that correspond to high and low level of earthquake frequency, we introduce a hidden state process $\{S_n\}$ that governs π_n and λ_n. At time n, S_n equals 0 or 1 when the stress level is in the state of low or high level of earthquake frequency respectively. Moreover, the parameters λ_n and π_n are modeled by

$$\lambda_n = \begin{cases} \lambda_0, & S_n = 0, \\ \lambda_1, & S_n = 1, \end{cases} \quad \pi_n = \begin{cases} \pi_0, & S_n = 0, \\ \pi_1, & S_n = 1. \end{cases}$$

Given the current state S_n, the conditional distribution of A_n is called the state dependent distribution, and can be derived from (1) and (2) as

$$f_{A_n|S_n}(a_n|s_n) = \begin{cases} \pi_{s_n}\lambda_{s_n}e^{-\lambda_{s_n}(a_n - M_{min})}, & M_{min} \le a_n < \infty, \\ 1 - \pi_{s_n}, & a_n = 0. \end{cases} \tag{3}$$

Although we restrict S_n to be of two states, it can be easily extended to a general S-state model with $S \ge 2$. Next, since the frequency and magnitude of earthquake occurrence are highly dependent on the hidden state S_n, in order to predict future earthquakes, it is important to model the dynamic of the hidden state S_n. However, in the literature, HMMs for earthquake data usually assumes that the transition probabilities between hidden states are constant throughout the whole series (Ebel et al. 2007; Orfanogiannaki et al. 2010). Therefore, they are not capable of predicting future hidden states. To overcome this problem, we propose a dynamic model that incorporates covariates and allows us to predict the future dynamic of S_n, which in turn enables us to predict changes in the frequency and magnitude of earthquake occurrence.

To model the dynamic of $\{S_n\}$, we model the transition probability $\mathbb{P}(S_n = s_n|S_{n-1}, \boldsymbol{Z}_{n-1})$ by a logistic function,

$$\mathbb{P}(S_n = 1|S_{n-1} = 0, \boldsymbol{z}_{n-1}) = p_{01_n} = \frac{\exp(\boldsymbol{z}'_{n-1}\boldsymbol{\alpha})}{1 + \exp(\boldsymbol{z}'_{n-1}\boldsymbol{\alpha})},$$
$$p_{00_n} = 1 - p_{01_n}, \tag{4}$$

and

$$\mathbb{P}(S_n = 0|S_{n-1} = 1, \boldsymbol{z}_{n-1}) = p_{10_n} = \frac{\exp(\boldsymbol{z}'_{n-1}\boldsymbol{\beta})}{1 + \exp(\boldsymbol{z}'_{n-1}\boldsymbol{\beta})},$$
$$p_{11_n} = 1 - p_{10_n}, \tag{5}$$

where $\boldsymbol{\alpha}$, $\boldsymbol{\beta} \in \mathbb{R}^d$ are parameters; and $\boldsymbol{z}_{n-1} \in \mathbb{R}^d$ is a d dimensional time-varying covariate. Here, the logistic function is used as a convenience choice to transform covariates

into probabilities between zero and one; other functions, such as the probit link, can be used (Hughes et al. 1999). The transition probabilities can be summarized by the matrix

$$\mathbf{P_n} = \begin{pmatrix} p_{00_n} & p_{01_n} \\ p_{10_n} & p_{11_n} \end{pmatrix}. \quad (6)$$

Figure 3 depicts a typical relationship between $\{A_n\}$ and $\{S_n\}$. Note that earthquakes occur more frequently in State 1 than in State 0. This is in line with the clustering feature observed empirically in earthquake catalogues.

Remark 1 There are two reasons for adopting a discrete time framework instead of using a continuous-time HMM. First, in practice, earthquake catalogues and covaraites are recorded in discrete time. Even continuous time models have to be discretized in order to incorporate the discrete time covariates. Moreover, as the proposed discrete time model takes observations in a 1 min fine grid, it should give sufficiently accurate approximation to the continuous time process of earthquake occurrence.

Second, and more importantly, the estimation theory of discrete-time HMM is well developed and is easy to implement. It can be readily extended to the case that the dynamics of the hidden state depends on time-varying covariates. However, the continuous-time analog, i.e., the continuous-time HMM with non-homogeneous transition probability depending on time-varying covariates, is more difficult to make statistical inference. For example, when the transition probability P_t depends on the time-varying covariates, the infinitesimal generator $Q = \lim_{t\to 0} \frac{P_t - I}{t}$ its corresponding forward and backward probabilities are more complicated for establishing computational feasible estimation algorithms and studying other probabilistic behaviors of the model.

2.1 The choice of covariates

One of the main features of the proposed model is the incorporation of the covariates $\boldsymbol{Z}_n$ in the dynamic of the hidden state S_n [(4) and (5)], which facilitates the prediction of future hidden states. In practice, the covariates may be chosen according to some phenomena in seismology that provide information about the state of the underground underlying dynamics. For example, it is observed that the aftershocks initiated from an earthquake decrease with the passing of time (Utsu 1961; Ogata 1988). This phenomenon suggests that the time that has passed since the previous earthquake, denoted by T_n for time n, provides important information for the clustering pattern of upcoming observations (Sornette and Knopoff 1997). Thus, we choose T_n as a covariate for state transition.

Several potential precursors for earthquakes, including electromagnetic signals, electrical resistivity changes, magnetization changes, local stress interactions, or other

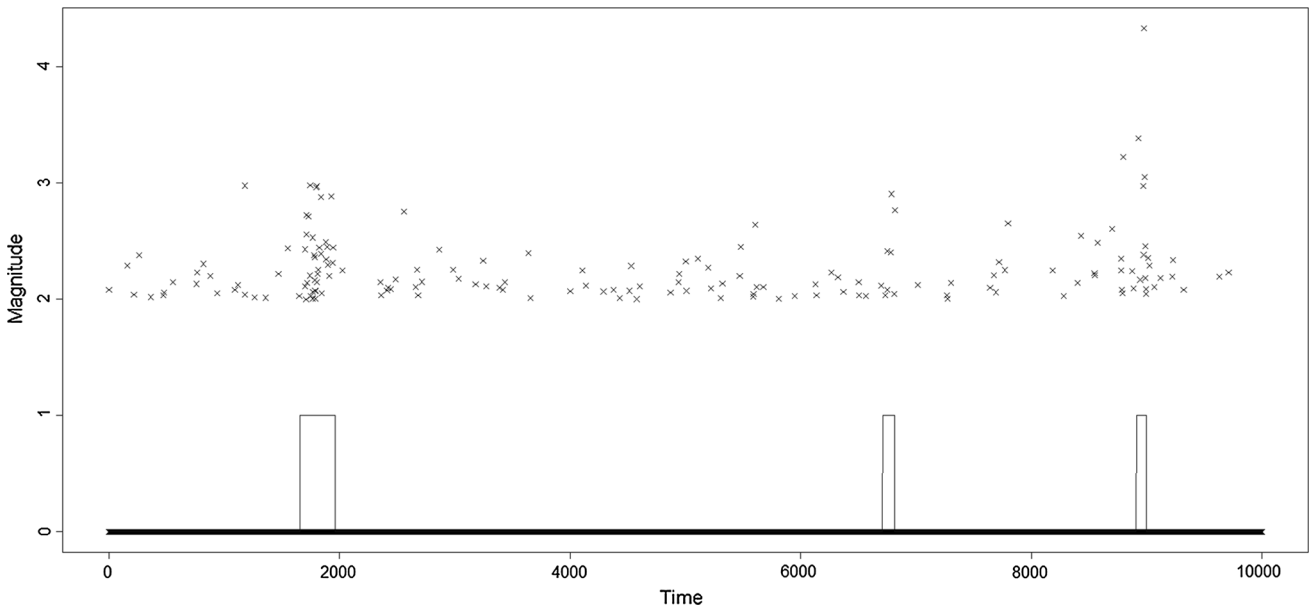

Fig. 3 Plot of simulated earthquake observations (crosses) with two hidden states (horizontal lines) and transitions (vertical lines) in between. $(\lambda_0, \lambda_1, \pi_0, \pi_1, \alpha_0, \alpha_1, \beta_0, \beta_1) = (5, 2, 0.01, 0.1, -6, -0.05, -4, -0.15)$

underground measurements have been proposed in literatures (e.g., Park et al. 1993; Seeber and Armbruster 2000). However, it is found that many actual observations of low-frequency electrical precursors deviate from the expected laboratory results. In some cases, precursors are recorded without corresponding the co-seismic signals or involve missing data. Moreover, the electrical signals are also susceptible to background noise (Huang and Ikeya 1998). As a result, these potential precursors cannot be readily included as covariates on earthquake prediction. For simplicity, we include T_n as the only time-varying covariate Z_n in the real data analysis.

Specifically, T_n is an observable discrete stochastic process with infinite state space $\mathrm{T} = \mathbb{Z}^+ = \{0, 1, 2, \ldots\}$. If an earthquake occurred at time point n, i.e., when $A_n \neq 0$, then we have $T_n = 0$. The value of T_n accumulates if no earthquake occurs at time point n, i.e.,

$$T_n|T_{n-1} = \begin{cases} T_{n-1}+1, & A_n = 0, \\ 0, & A_n \neq 0. \end{cases}$$

Note that T_n has an advantage of fully determined by A_n and T_{n-1}, hence does not involve prediction errors due to the modelling of covariates. Moreover, T_n introduces a self-exciting feature, in the sense that one earthquake can affect the dynamics of subsequent hidden states. The corresponding transition probabilities of the hidden states are particular cases of (4) and (5) with $\boldsymbol{z}_n = t_n$, where t_n is the observed value of T_n at time n. In summary, the earthquake prediction model for the earthquake catalogue $\{A_n\}$ is given as

$$A_n = M_n \mathbb{I}_n, \tag{7}$$

where $M_n \sim$ Left-truncated $\exp(\lambda_{S_n}, M_{min})$, $\mathbb{I}_n \sim \text{Bernoulli}(\pi_{S_n})$,

$$T_n|T_{n-1} = \begin{cases} T_{n-1}+1, & A_n = 0, \\ 0, & A_n \neq 0. \end{cases}$$

$$\mathbb{P}(S_n = 1|S_{n-1} = 0, T_{n-1} = t_{n-1}) = p_{01_n} = \frac{\exp(\alpha_0 + \alpha_1 t_{n-1})}{1 + \exp(\alpha_0 + \alpha_1 t_{n-1})},$$

$$\mathbb{P}(S_n = 0|S_{n-1} = 1, T_{n-1} = t_{n-1}) = p_{10_n} = \frac{\exp(\beta_0 + \beta_1 t_{n-1})}{1 + \exp(\beta_0 + \beta_1 t_{n-1})},$$

$p_{00_n} = 1 - p_{01_n}$, and $p_{11_n} = 1 - p_{10_n}$.

The parameters of interest are collected as

$$\boldsymbol{\theta} = \{\lambda_0, \lambda_1, \pi_0, \pi_1, \alpha_0, \alpha_1, \beta_0, \beta_1\} \in \boldsymbol{\Theta} = \mathbb{R}^+ \times \mathbb{R}^+ \times (0,1) \times (0,1) \times \mathbb{R} \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}. \tag{8}$$

Since the covariate T_n depends on the $\{A_k, k \leq n\}$, the state transition dynamic depends on the past observations. Thus, Model (7) cannot be covered by traditional HMM, where the state transition dynamic does not depend on past observations. Therefore, the probabilistic properties, such as stationarity and ergodicity, and asymptotic theories of inference in traditional HMMs cannot be directly applied to Model (7). In the Supplementary Material, we provide the proof of the following theorem.

Theorem 1 *The process* $\{X_n = (S_n, A_n, T_n)'\}_{n\in\mathbb{Z}}$ *following Model* (7) *is strictly stationary and ergodic.*

For a general model that incorporates an additional d-dimensional time-varying covariates $\{Z_n \in \mathbb{R}^d\}_{n\in\mathbb{Z}}$, Model (7) can be extended as

$$A_n = M_n \mathbb{I}_n, \tag{9}$$

where $M_n \sim$ Left-truncated $\exp(\lambda_{S_n}, M_{min})$, $\mathbb{I}_n \sim$ Bernoulli (π_{S_n}),

$$T_n|T_{n-1} = \begin{cases} T_{n-1}+1, & A_n = 0, \\ 0, & A_n \neq 0. \end{cases}$$

As, denoting $\mathbb{P}(S_n = 1|(s,t,z))$ as $\mathbb{P}(S_n = 1|S_{n-1} = s, T_{n-1} = t, Z_{n-1} = z)$, the transition probabilities of the hidden states are

$$\mathbb{P}(S_n = 1|(0, t_{n-1}, z_{n-1})) = p_{01_n} = \frac{\exp(\alpha_0 + \alpha_1 t_{n-1} + z'_{n-1}\boldsymbol{\alpha})}{1 + \exp(\alpha_0 + \alpha_1 t_{n-1} + z'_{n-1}\boldsymbol{\alpha})},$$

$$\mathbb{P}(S_n = 0|(1, t_{n-1}, z_{n-1})) = p_{10_n} = \frac{\exp(\beta_0 + \beta_1 t_{n-1} + z'_{n-1}\boldsymbol{\beta})}{1 + \exp(\beta_0 + \beta_1 t_{n-1} + z'_{n-1}\boldsymbol{\beta})},$$

$p_{00_n} = 1 - p_{01_n}$, and $p_{11_n} = 1 - p_{10_n}$.

The parameters of interest are collected as $\boldsymbol{\theta} = \{\lambda_0, \lambda_1, \pi_0, \pi_1, \alpha_0, \boldsymbol{\alpha}, \beta_0, \beta_1, \boldsymbol{\beta}\} \in \boldsymbol{\Theta} = \mathbb{R}^+ \times \mathbb{R}^+ \times (0,1) \times (0,1) \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}^d$. The following corollary extends Theorem 1 for Model (9).

Corollary 1 *Let* $\{Z_n \in \mathbb{R}^d\}_{n\in\mathbb{Z}}$ *be a d-dimensional time-varying covariate which is a bounded, strictly stationary and ergodic Markov Chain process. Assume that* $Z_n|Z_{n-1}$ *and* $A_n|S_n$ *are independent for all* $n \in \mathbb{N}$. *Then the process* $\{X_n = (S_n, A_n, T_n, Z_n)'\}_{n\in\mathbb{Z}}$ *following Model* (9) *is strictly stationary and ergodic.*

2.2 Likelihood

In this section, we derive the likelihood function for parameter estimation. Define $a_1^N = (a_1, a_2, \ldots, a_N)$ and $s_1^N = (s_1, s_2, \ldots, s_N)$ as the observations and states from time 1 to N. Since we have no observations to determine T_n and S_n for $n \leq 0$, we set $T_0 = 0$ and introduce an initial distribution for S_1 by

$$\mathbb{P}(S_1 = s_1) = \delta_{s_1}, \tag{10}$$

where $s_1 \in \mathcal{S} = \{0, 1\}$ and $\delta_0 + \delta_1 = 1$. The use of the initial distribution can be justified by the stationarity of $\{(S_n, A_n, T_n)'\}$ in Theorem 1. The likelihood function of the observations a_1^N is

$$\begin{aligned} L(\boldsymbol{\theta}, a_1^N) &= \mathbb{P}\left(A_1^N = a_1^N\right) \\ &= \sum_{s_1^N \in \mathcal{S}^N} \mathbb{P}\left(A_1^N = a_1^N, S_1^N = s_1^N\right) \\ &= \sum_{s_1^N \in \mathcal{S}^N} \left[\mathbb{P}(S_1 = s_1) \prod_{k=2}^{N} \mathbb{P}(S_k = s_k | s_{k-1}, t_{k-1}) \right. \\ &\quad \left. \prod_{k=1}^{N} \mathbb{P}(A_k = a_k | s_k) \right] \\ &= \sum_{s_1^N \in \mathcal{S}^N} \delta_{s_1} \prod_{k=2}^{N} p_{s_{k-1} s_k} \prod_{k=1}^{N} f_{A_k|S_k}(a_k | s_k), \end{aligned} \tag{11}$$

where $s_{k-1}, s_k \in \mathcal{S} = \{0, 1\}$, δ_{s_1}, $p_{s_{k-1} s_{kk}}$, and $f_{A_k|S_k}(a_k|s_k)$ are defined in (10), (7), and (3), respectively. For simplicity, $p_{s_{k-1} s_{kk}}$ is written as $p_{s_{k-1} s_k}$. Note that for any sequence of observations, the maximizer (δ_0, δ_1) of the likelihood will either be (1, 0) or (0, 1) (Levinson et al. 1983; Leroux and Puterman 1992; Zucchini et al. 2008).

From (11), the computation of likelihood function requires summing over all possible combinations of the hidden states s_1^N, which is a sum of 2^N terms, where each term is a product of $2N$ factors. This operation requires a computational order of $O(N2^N)$, and thus direct numerical maximization is infeasible when N is large. To simplify the computations, we can express the likelihood function as

$$L\left(\boldsymbol{\theta}, a_1^N\right) = \boldsymbol{\delta} \mathbf{F}(\mathbf{a_1}|\mathbf{s_1}) \mathbf{P_2} \mathbf{F}(\mathbf{a_2}|\mathbf{s_2}) \mathbf{P_3} \mathbf{F}(\mathbf{a_3}|\mathbf{s_3}) \cdots \mathbf{P_N} \mathbf{F}(\mathbf{a_N}|\mathbf{s_N}) \mathbf{1}', \tag{12}$$

where $\mathbf{1}' = (\mathbf{1}, \mathbf{1})'$, $\boldsymbol{\delta} = (\delta_0, \delta_1)$ is the initial state distribution, $\mathbf{P_n}$ is the transition probability matrix defined in (6), and

$$\mathbf{F}(a_n|s_n) = \begin{pmatrix} f_{A_n|S_n}(a_n|0) & 0 \\ 0 & f_{A_n|S_n}(a_n|1) \end{pmatrix},$$

(Zucchini and MacDonald 2009). Since (12) involves a product of $2N$ two-dimensional matrices $\mathbf{P_k}$ and $\mathbf{F}(\mathbf{a_k}|\mathbf{s_k})$, the number of operations required for computing $L(\boldsymbol{\theta}, a_1^N)$ is of order $O(N)$, which is much smaller than the order $O(N2^N)$ using (11).

We have the following results regarding the maximum likelihood estimate (MLE) of Model (7) and Model (9), for which the proof is provided in the Supplementary Material.

Theorem 2 *The MLE $\hat{\boldsymbol{\theta}}$ for the parameter $\boldsymbol{\theta} = \{\lambda_i, \pi_i, \alpha_0, \alpha_1, \beta_0, \beta_1,\ i \in \mathcal{S} = \{0, 1\}\} \in \boldsymbol{\Theta}$ of the process $\{(S_n, A_n, T_n)'\}_{n \in \mathbb{N}}$ following Model (7) is strongly consistent in the quotient topology, i.e., its Euclidean distance to the equivalent class convergence to 0, that is,*

$$\lim_{N \to \infty} d\left(\hat{\boldsymbol{\theta}}, \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \boldsymbol{\theta} \sim \boldsymbol{\theta}^{\star}\}\right) = 0, \quad \mathbb{P}^{\boldsymbol{\theta}^{\star}}\text{–a.s.}$$

where $\boldsymbol{\theta}^{\star}$ is the true parameter vector and the parameters $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \boldsymbol{\Theta}$ are equivalent, denote as $\boldsymbol{\theta}_1 \sim \boldsymbol{\theta}_2$, if $\mathbb{P}^{\boldsymbol{\theta}_1} = \mathbb{P}^{\boldsymbol{\theta}_2}$ almost surely. Furthermore, the MLE $\hat{\boldsymbol{\theta}}$ for the parameter $\boldsymbol{\theta}$ is asymptotically normally distributed.

Corollary 2 *Let $\hat{\boldsymbol{\theta}}$ be the MLE for the parameter $\boldsymbol{\theta} = \{\lambda_i, \pi_i, \alpha_0, \alpha_1, \boldsymbol{\alpha}, \beta_0, \beta_1, \boldsymbol{\beta},\ i \in \mathcal{S} = \{0, 1\}\} \in \boldsymbol{\Theta}$ of the process $\{(S_n, A_n, T_n, Z_n)'\}_{n \in \mathbb{N}}$ following Model (9). If the assumptions in Corollary1 is satisfied, then $\hat{\boldsymbol{\theta}}$ is strongly consistent in the quotient topology. Furthermore, the MLE $\hat{\boldsymbol{\theta}}$ for the parameter $\boldsymbol{\theta}$ is asymptotically normally distributed.*

3 Implementation

3.1 EM algorithm

Given the observations $\{a_1, \ldots, a_N\}$ from an earthquake catalogue, the parameters $\boldsymbol{\theta}$ of the Model (7) can be found by maximizing the likelihood (12). While direct maximization by standard numerical methods, such as the Broyden–Fletcher–Goldfarb–Shanno algorithm, is simple and fast, it could also be unstable. On the other hand, the expectation–maximization (EM) algorithm generally gives more stable performance even under poor initial values of the parameters (Bulla and Berzel 2008).

The EM algorithm is a popular method for computing MLE for problems involving missing data. Since the values of the hidden state in Model (7) are unobservable, they can be treated as missing data, and the EM algorithm is applicable (Baum et al. 1970; Baum 1972; Welch 2003). However, the Baum–Welch algorithm assumes that the transition probabilities in the hidden Markov chain are constant. As the likelihood function (11) involves a dynamic model for the hidden state process, the standard Baum-Welch algorithm can be modified to handle time-varying transition probabilities. The details of the EM algorithm and the derivation of the forward and backward probabilities used in the algorithm are illustrated in the Supplementary Material. Other possible estimation procedures using EM algorithm can be find in the literature, see Bartolucci and Farcomeni (2009, 2015) and Bartolucci et al. (2013).

Once the MLE $\hat{\boldsymbol{\theta}}$ is obtained, we can estimate the variance–covariance matrix of the MLE by inverting the observed information matrix, see Sect. 3.6 of Zucchini and MacDonald (2009). The observed information matrix can be computed by numerically approximating the Hessian of negative log-likelihood function evaluated at the estimated parameters, see Sect. 3.6.1 of Zucchini and MacDonald (2009). See also Bartolucci and Farcomeni (2015a) for alternative methods for computing the standard errors of the parameters estimates of the hidden Markov model. With the variance–covariance matrix and the asymptotically normality of the MLE proven in Sect. 3 of the Supplementary Material, we can construct confidence intervals for the parameters.

3.2 Further computational issues

In maximizing the likelihood (12), an issue in implementation is that some estimating parameters are in bounded domains. In particular, in (3) the probability of earthquake occurrence satisfies $\pi_{s_n} \in (0,1)$ and the rate of exponential distribution satisfies $\lambda_{s_n} > 0$. To avoid constrained maximization, these parameters can be reparameterized to have unbounded domains prior to maximization. Specifically, we set $\tilde{\pi}_{s_n} = \frac{1}{1+e^{\pi_{s_n}}}$ and $\tilde{\lambda}_{s_n} = \exp(\lambda_{s_n})$, so that the reparameterized parameters $\tilde{\pi}_{s_n}$ and $\tilde{\lambda}_{s_n}$ are constraint-free. Once the maximizers are obtained, they are transformed back to yield the maximizers in terms of the original parameters π_{s_n} and λ_{s_n}. The asymptotic variance–covariance matrix of the original parameters can also be transformed from the observed information matrix of the reparameterized parameters in the numerical maximization to enhance numerical stability, see Sect. 3.6 of Zucchini and MacDonald (2009).

Due to the complicated structure of the HMM, the likelihood is not a convex function of the parameters. Hence, both direct maximization and the EM algorithm may converge to a local maximum of the likelihood function. While both methods cannot guarantee convergence to the global maximum, the EM algorithm is less sensitive to poor initial values, and thus less likely to be trapped in the local maximum. Nevertheless, multiple random starting points can be specified to increase the probability of obtaining the global maximizer.

Finally, an issue exists regarding the identifiability of the parameters. In particular, the parameters of Model (7) are not identifiable in the sense that the indices of the states can be permuted without affecting the distribution of the observation process $\{A_n\}$. This issue is called the label-switching problem (Leroux 1992). For example, the parameter vectors $\{\lambda_0, \lambda_1, \pi_0, \pi_1, \alpha_0, \alpha_1, \beta_0, \beta_1\}$ and $\{\lambda_1, \lambda_0, \pi_1, \pi_0, \beta_0, \beta_1, \alpha_0, \alpha_1\}$, which correspond to permuting the indices of state 0 and state 1, give exactly the same value of likelihood. Therefore, we define State 0 as the state with a smaller probability of earthquake occurrence π. Similarly, we set State 1 as the state with a larger probability of earthquake occurrence.

4 Prediction

4.1 Estimation of hidden states

In this section we develop prediction procedures for the hidden states and future earthquake occurrences. First, to estimate the hidden states, the local decoding and the Viterbi algorithm are commonly employed (Viterbi 1967; Zucchini and MacDonald 2009). Local decoding estimates the hidden state at a particular time point $k \in [1, N]$ by the most probable state in terms of the posterior probability of states S_k given the observations a_1^N, i.e.,

$$\hat{s}_k = \underset{i=0,1}{\operatorname{argmax}} \mathbb{P}\left(S_k = i | a_1^N; \hat{\boldsymbol{\theta}}\right). \tag{13}$$

To estimate the whole hidden state process S_1^N, the local decoding can be performed for each of $n = \{1, \cdots, N\}$. Note that, in the implementation of local decoding, the posterior probability of states $\mathbb{P}(S_k = s_k | a_1^N; \hat{\boldsymbol{\theta}})$ can be computed efficiently by the forward and backward probabilities, see (32), (33), (34), and (36) in Sect. 4.1 of the Supplementary Material.

In contrast, the Viterbi algorithm works globally, in the sense that it estimates the entire hidden state process S_1^N simultaneously. Specifically, the estimated sequence of hidden states $\hat{s}_1^N$ is obtained by

$$\hat{s}_1^N = \underset{s_1^N \in \{0,1\}^N}{\operatorname{argmax}} \mathbb{P}\left(S_1^N = s_1^N | a_1^N; \hat{\boldsymbol{\theta}}\right),$$

(Viterbi 1967). Equivalently, $\hat{s}_1^N$ is the sequence that maximizes the complete data log-likelihood $J(\boldsymbol{\theta}, a_1^N, s_1^N)$, i.e.,

$$\begin{aligned} \log J(\boldsymbol{\theta}, a_1^N, s_1^N) &= \log \mathbb{P}\left(A_1^N = a_1^N, S_1^N = s_1^N\right) \\ &= \log \left[\delta_{s_1} \prod_{k=2}^{N} p_{s_{k-1} s_k} \prod_{k=1}^{N} f_{A_k|S_k}(a_k|s_k) \right] \\ &= \log \delta_{s_1} + \sum_{k=2}^{N} \log p_{s_{k-1} s_k} \\ &\quad + \sum_{k=1}^{N} \log f_{A_k|S_k}(a_k|s_k). \end{aligned} \tag{14}$$

with respect to the hidden states.

The Viterbi algorithm for computing $\hat{s}_1^N$ is described as follows. Define

$$V_{1i} = \mathbb{P}(S_1 = i, A_1 = a_1) = \delta_i f_{A_1|S_1}(a_1|s_1),$$

and

$$V_{ni} = \max_{s_1^{n-1} \in \mathcal{S}^{n-1}} \mathbb{P}\left(S_1^{n-1} = s_1^{n-1}, S_n = i, A_1^n = a_1^n\right), \tag{15}$$

for $n = 2, 3, \ldots, N$. For $i, j = 0, 1$, V_{nj} satisfies the recursive relation

$$\begin{aligned}
V_{nj} &= \max_{s_1^{n-1} \in \mathcal{S}^{n-1}} \mathbb{P}\left(A_n = a_n | S_n = j, s_1^{n-1}, a_1^{n-1}\right) \\
&\quad \mathbb{P}\left(S_n = j | s_1^{n-1}, a_1^{n-1}\right) \mathbb{P}\left(S_1^{n-1} = s_1^{n-1}, A_1^{n-1} = a_1^{n-1}\right) \\
&= \max_{s_1^{n-1} \in \mathcal{S}^{n-1}} \mathbb{P}(A_n = a_n | S_n = j) \mathbb{P}(S_n = j | s_{n-1}, t_{n-1}) \\
&\quad \mathbb{P}\left(S_1^{n-1} = s_1^{n-1}, A_1^{n-1} = a_1^{n-1}\right) \\
&= \max_{s_1^{n-2} \in \mathcal{S}^{n-2}} \max_{i \in \mathcal{S}} \mathbb{P}(A_n = a_n | S_n = j) \\
&\quad \mathbb{P}(S_n = j | S_{n-1} = i, t_{n-1}) \\
&\quad \mathbb{P}\left(S_1^{n-2} = s_1^{n-2}, S_{n-1} = i, A_1^{n-1} = a_1^{n-1}\right) \\
&= \max_{s_1^{n-2} \in \mathcal{S}^{n-2}} \max_{i \in \mathcal{S}} f_{A_n|S_n}(a_n|j) p_{ij_n} \\
&\quad \mathbb{P}\left(S_1^{n-2} = s_1^{n-2}, S_{n-1} = i, A_1^{n-1} = a_1^{n-1}\right) \\
&= \left[\max_{i \in \mathcal{S}} \left(V_{n-1i} \mathrm{p}_{\mathrm{ij_n}}\right)\right] f_{A_n|S_n}(a_n|j).
\end{aligned} \tag{16}$$

Using this recursive relationship, the value of V_{nj} from $n = 1$ to N can be computed efficiently. Using (16), the sequence $\hat{s}_N, \hat{s}_{N-1}, \ldots, \hat{s}_1$ can be computed recursively based on

Step 1 Compute $V_{1i}, \ldots, V_{Ni}$, i=0, 1, using (16).

Step 2 $\hat{s}_N = \underset{i=0,1}{\mathrm{argmax}} \mathrm{V_{Ni}}$.

Step 3 For $n = N-1, N-2, \ldots, 1$, $\hat{s}_n = \underset{i=0,1}{\mathrm{argmax}} \left(V_{ni} p_{i\hat{s}_{n+1}n}\right)$.

A simulation study is performed to compare the local decoding and the Viterbi algorithm. We generated a sample of size $N = 10,000$ by model (7) with the set of parameters $\{\lambda_0, \lambda_1, \pi_0, \pi_1, \alpha_0, \alpha_1, \beta_0, \beta_1\} = \{5, 2, 0.01, 0.1, -6, -0.05, -4, -0.15\}$. Figure 4 depicts the true states and the estimated states from local decoding and the Viterbi algorithm. Note that usually more change-points in the estimated hidden state occur by local decoding. One possible reason for this is that local decoding maximizes the posterior distribution of hidden states separately at each time point, and the information from the preceding and succeeding hidden states are not taken into account. Hence, the Viterbi algorithm appears to be a more reliable method.

4.2 Prediction of future earthquake occurrence

Based on the estimated model parameters and hidden state, we propose the following simulation-based procedure to predict future hidden state, occurrences, and magnitudes of future earthquakes.

Step 1 Apply local decoding or the Viterbi algorithm to find the estimated state $\hat{s}_N$ in time N.

Step 2 Observe the time that has passed since the last earthquake to obtain t_N.

Step 3 Based on the MLE $\hat{\boldsymbol{\theta}}$ and the estimated state $\hat{s}_N$, simulate M sample paths of hidden states $S_{N+1}^{N+m}(j)$ and observed earthquake magnitude $A_{N+1}^{N+m}(j)$, where m is a large integer and $j = 1, \ldots, M$.

Step 4 For each path of hidden state $S_{N+1}^{N+m}(j)$, identify the time of the k-th change-in-state, denote as $t_k(j)$, where $k \in \mathbb{Z}$. The 95% prediction interval for the time to the k-th change-in-state is the interval between the 2.5-th and 97.5-th percentiles of $\{t_k(j)\}_{j=1,\ldots,M}$.

Step 5 For each path of observations $A_{N+1}^{N+m}(j)$, identify the time and magnitude of the k-th earthquake, denote as $\tilde{t}_k(j)$ and $m_k(j)$, where $k \in \mathbb{Z}$. The 95% prediction interval for the time to the k-th earthquake is the interval between the 2.5-th and 97.5-th percentiles of $\{\tilde{t}_k(j)\}_{j=1,\ldots,M}$. Similarly, the 95% prediction interval for the magnitude of the k-th earthquake is the interval between the 2.5-th and 97.5-th percentiles of $\{m_k(j)\}_{j=1,\ldots,M}$.

Note that it suffices to choose m to be large enough such that each of the M paths contains not less than k changes in states and k earthquakes. Alternatively, one may simulate each of the paths separately until k changes in states and k earthquakes have occurred.

5 Simulation studies

In the simulation study, to mimic the scenario of real earthquake, the data is generated from a $\boldsymbol{\theta}$ similar to the parameter estimates of the dataset in Sect. 6. Specifically,

$$\begin{aligned}
\begin{pmatrix} \lambda_0 \\ \lambda_1 \end{pmatrix} &= \begin{pmatrix} 5 \\ 2 \end{pmatrix}, \begin{pmatrix} \pi_0 \\ \pi_1 \end{pmatrix} = \begin{pmatrix} 0.01 \\ 0.1 \end{pmatrix}, \begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix} \\
&= \begin{pmatrix} -6 \\ -0.05 \end{pmatrix}, \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} -4 \\ -0.15 \end{pmatrix},
\end{aligned}$$

with $M_{min} = 2$. As there is no information before time $n = 1$, set $t_0 = 0$. In addition, the time difference between consecutive observations is chosen to be 1 minute. As an earthquake catalogue typically includes records across several decades, the sample size N is huge. For example, in the real dataset in Sect. 6, the sample size is around $N = 14,000,000$. To reflect the situation, we explore the

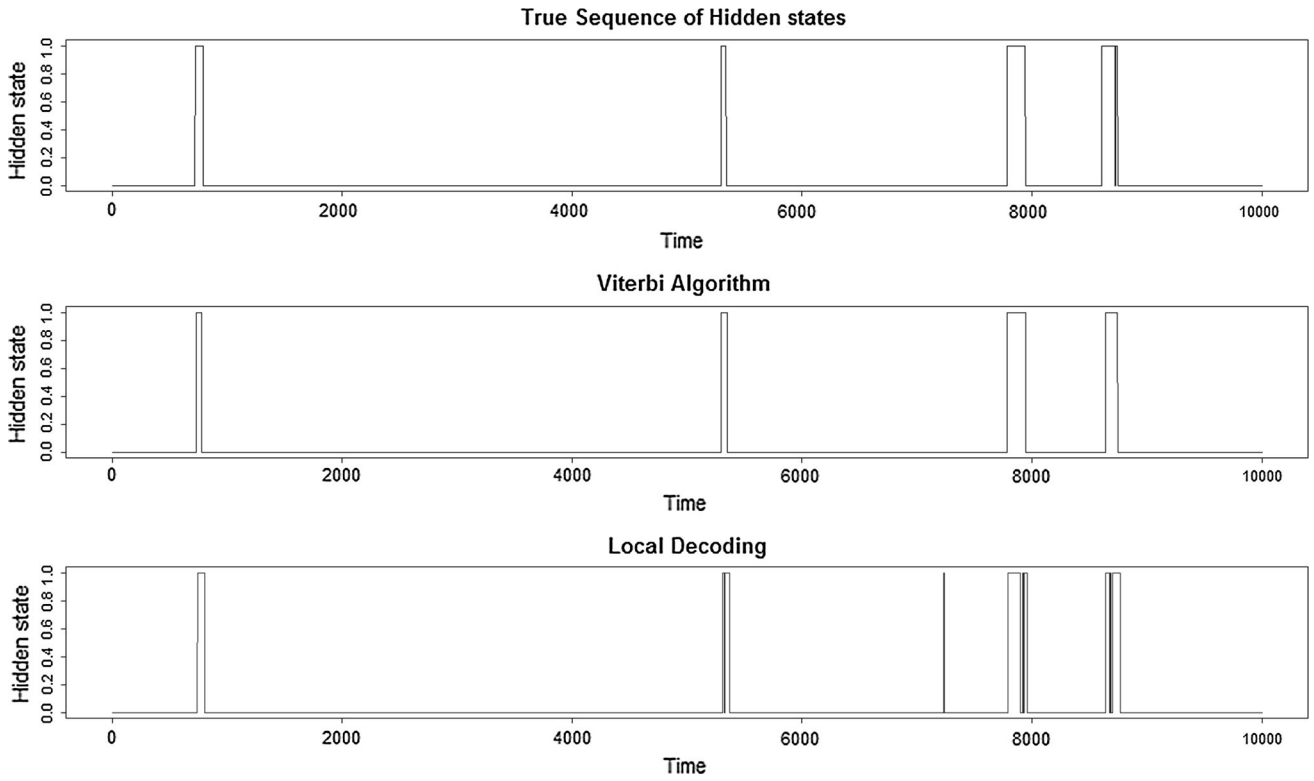

Fig. 4 Estimation of change-points by different algorithms. True change-points (top); change-points estimated by Viterbi algorithm (middle); change-points estimated by local decoding (bottom)

scenarios $N = 100,000$, 1, 000, 000 and 5, 000, 000. Both the EM algorithm and the direct numerical maximization are employed for parameter estimations. The mean, standard errors, median and root mean square errors of the parameter estimates for 200 replications are reported in Table 1. On the other hand, we include another two parameter sets with sample size $N = 100,000$ to simulate the parameter estimation of Model (9). Another time-varying covariate X_n is added on top of T_n to mimic the local stress interactions (Seeber and Armbruster 2000). In particular,

$$X_n|X_{n-1} = \begin{cases} X_{n-1}+1, & -150 \leqslant X_{n-1} < 150, \\ -150, & X_{n-1} = 150. \end{cases}$$

The results are summarized in Table 2. Note that both methods estimate the true parameters accurately. Nevertheless, the parameter estimates by EM algorithm have slightly lower standard errors than that by the direct numerical maximization. One possible reason for this is that direct numerical maximization is more susceptible to local maxima.

We next study the finite sample performance of hidden state estimations. Using the MLE obtained from the EM algorithm, a sequence of hidden states is estimated by local decoding and the Viterbi algorithm under the case $N = 5,000,000$. For illustration, a subsample of size $N = 200,000$ is depicted in Fig. 5. Note that the Viterbi algorithm recovers most of the change-points, and more false-positive change-points of states are found by local decoding. This agrees with the comparison of the two methods in Sect. 4.

Finally, we study the prediction of future changes in states and earthquakes occurrences. Based on one particular path under the case $N = 5,000,000$, we simulate $M = 1000$ future paths of length $m = 36,792,000$ to construct 95% prediction intervals for each of the 50 future changes in states, 20 future earthquakes with magnitudes $\geqslant 5$, and 10 future earthquakes with magnitudes $\geqslant 6$, respectively. The length $m = 36,792,000$ corresponds to the number of observations in 70 years, which is large enough to cover the desired number of future earthquakes over each magnitudes. The results of prediction are presented in Figs. 6, 7 and 8. Note that the prediction intervals cover well the future true change-points in hidden states, time and magnitude of future earthquakes, with only one earthquake of magnitudes $\geqslant 6$ missed. Thus, the prediction algorithm provides satisfactory prediction of the future changes in states, occurrence time, and magnitudes of large earthquakes.

Table 1 Mean, standard errors (in brackets), median (in italics), and root mean squared error (in underline) of parameter estimates based on 200 replications

N	EM algorithm			Direct maximization		
	100,000	1,000,000	5,000,000	100,000	1,000,000	5,000,000
$\hat{\lambda}_0$	5.0091	5.0051	4.9993	5.0166	4.9937	4.9994
	(0.1634)	(0.0524)	(0.0226)	(0.1744)	(0.0564)	(0.0227)
	5.0044	*5.0098*	*5.0001*	*5.0304*	*4.9928*	*5.0004*
	0.1632	0.0525	0.0226	0.1748	0.0566	0.0227
$\hat{\lambda}_1$	2.0047	1.9972	2.0001	1.9951	2.0006	2.0002
	(0.0855)	(0.0295)	(0.0129)	(0.0947)	(0.0259)	(0.0129)
	2.0081	*2.0005*	*2.0015*	*2.0026*	*2.0026*	*2.0011*
	0.0855	0.0295	0.0129	0.0946	0.0259	0.0129
$\hat{\pi}_0$	0.0100	0.0100	0.0100	0.0100	0.0100	0.0100
	(0.0003)	(0.0001)	(0.00005)	(0.0003)	(0.0001)	(0.00005)
	0.0100	*0.0100*	*0.0100*	*0.0100*	*0.0100*	*0.0100*
	0.0003	0.0001	0.00005	0.00003	0.0001	0.00005
$\hat{\pi}_1$	0.0990	0.0999	0.0999	0.0986	0.0998	0.0999
	(0.0055)	(0.0018)	(0.0007)	(0.0053)	(0.0024)	(0.0012)
	0.0988	*0.0998*	*0.1000*	*0.0989*	*0.1001*	*0.1000*
	0.0056	0.0018	0.0007	0.0055	0.0024	0.0012
$\hat{\alpha}_0$	−5.9520	−5.9817	−5.9936	−5.9770	−6.0007	−5.9977
	(0.5021)	(0.1299)	(0.0548)	(0.4783)	(0.1353)	(0.0608)
	−5.9998	*−5.9923*	*−5.9954*	*−5.9873*	*−5.9933*	*−6.0000*
	0.5031	0.1308	0.0550	0.4776	0.1349	0.0607
$\hat{\alpha}_1$	−0.0644	−0.0512	−0.0504	−0.0559	−0.0496	−0.0501
	(0.0536)	(0.0075)	(0.0030)	(0.0408)	(0.0074)	(0.0033)
	−0.0487	*−0.0508*	*−0.0503*	*−0.0495*	*−0.0490*	*−0.0497*
	0.0554	0.0076	0.0031	0.0412	0.0074	0.0033
$\hat{\beta}_0$	−3.9469	−3.9729	−3.9967	−4.0171	−3.9673	−3.9801
	(0.6955)	(0.2503)	(0.0568)	(0.8047)	(0.4518)	(0.2544)
	−3.9651	*−3.9710*	*−3.9923*	*−4.0332*	*−3.9690*	*−3.9846*
	0.6958	0.2512	0.0568	0.8028	0.4519	0.2546
$\hat{\beta}_1$	−0.3247	−0.1703	−0.1510	−0.4380	−0.2054	−0.1675
	(0.5142)	(0.1118)	(0.0116)	(1.1023)	(0.1818)	(0.0781)
	−0.1558	*−0.1533*	*−0.1521*	*−0.1318*	*−0.1495*	*−0.1560*
	0.5418	0.1134	0.0116	1.1367	0.1896	0.0798

6 Application to earthquake occurrence

In this section, we apply the proposed Model (7) to an earthquake catalogue in Southern California. Southern California is well-known to be an active tectonic region due to the San Andreas Fault beneath. Earthquake catalogues from 1981 to 2015 are obtained from the Southern California Earthquake Center (http://service.scedc.caltech.edu/ftp/catalogs/). The average magnitude of completeness is found to be approximately 1.8 by Hutton et al. (2010). The data from 1981 to 2007, which involve more than 90, 000 earthquakes, are used for model fitting and prediction. The data from 2008 and 2015, which involve 22 earthquakes of magnitude $\geqslant 5$ and an earthquake of magnitude $\geqslant 7$, are used as a testing dataset to study the prediction accuracy.

In seismology, the Gutenberg–Richter law asserts that if the magnitude m and the total number of earthquakes having magnitudes greater than m (denoted by N_m) have a log-linear relationship, then the earthquake magnitudes are exponentially distributed (Cosentino et al. 1977). In particular, if for some constants a and b,

$$\log_{10} N_m = a - bm,$$

then the probability of an earthquake with magnitude M larger or equal to m is

$$\mathbb{P}(M \geqslant m) = e^{-(b \ln 10)m}.$$

For the earthquake catalogue of Southern California, a plot of $\log N_m$ against m is provided in Fig. 9. This justifies the use of exponential distribution in Model (7) for modeling the earthquake magnitudes.

Using the proposed estimation and prediction procedures, we fit the Model (7) to the data, obtaining parameter estimates and prediction intervals for future changes in hidden states, time, and magnitude of future earthquakes.

Table 2 Mean, standard errors (in brackets), median (in italics), and root mean squared error (in underline) of two sets of feasible parameter sets based on 100 replications

	Set 1			Set 2		
	True value	EM algorithm	Direct maximization	True value	EM algorithm	Direct maximization
$\hat{\lambda}_0$	10	9.9824 (0.1332) *9.9688* 0.1337	9.9830 (0.1397) *9.9682* 0.1400	12	12.0497 (0.2417) *12.0490* 0.2456	12.0384 (0.2514) *12.0239* 0.2531
$\hat{\lambda}_1$	1	1.0007 (0.0082) *1.0004* 0.0081	1.0006 (0.0087) *1.0005* 0.0087	2	1.9993 (0.0203) *1.9980* 0.0202	2.0021 (0.0182) *2.0013* 0.0182
$\hat{\pi}_0$	0.1	0.1000 (0.0013) *0.1000* 0.0013	0.1001 (0.0015) *0.1002* 0.0015	0.05	0.0498 (0.0011) *0.0498* 0.0011	0.0500 (0.0010) *0.0501* 0.0010
$\hat{\pi}_1$	0.3	0.3000 (0.0022) *0.3002* 0.0022	0.3001 (0.0019) *0.3001* 0.0019	0.25	0.2501 (0.0021) *0.2502* 0.0021	0.2500 (0.0020) *0.2500* 0.0020
$\hat{\alpha}_0$	−0.5	−0.4840 (0.0885) *−0.4752* 0.0895	−0.4871 (0.0817) *−0.4907* 0.0824	−1	−1.0200 (0.2190) *−1.0333* 0.2189	−0.9847 (0.2489) *−1.0009* 0.2482
$\hat{\alpha}_1$	−0.2	−0.1998 (0.0138) *−0.1997* 0.0138	−0.2016 (0.0129) *−0.1999* 0.0130	−0.3	−0.3009 (0.0320) *−0.2986* 0.0319	−0.3093 (0.0379) *−0.3029* 0.0388
$\hat{\alpha}_2$	0.1	0.1000 (0.0022) *0.1002* 0.0022	0.1000 (0.0022) *0.0999* 0.0022	0.3	0.3037 (0.0343) *0.3002* 0.0343	0.3137 (0.0447) *0.3068* 0.0465
$\hat{\beta}_0$	−2	−1.9968 (0.1498) *−2.0111* 0.1491	−2.0125 (0.1668) *−2.0080* 0.1664	−2.5	−2.5041 (0.1127) *−2.5096* 0.1122	−2.5120 (0.1174) *−2.5194* 0.1174
$\hat{\beta}_1$	−0.1	−0.0973 (0.0360) *−0.0963* 0.0359	−0.1063 (0.0378) *−0.1110* 0.0382	−0.1	−0.1010 (0.0201) *−0.1014* 0.0200	−0.0995 (0.0275) *−0.1016* 0.0274
$\hat{\beta}_2$	0.2	0.2023 (0.0119) *0.2011* 0.0121	0.2034 (0.0117) *0.2023* 0.0121	0.1	0.1002 (0.0021) *0.1001* 0.0021	0.1001 (0.0023) *0.1002* 0.0023

In particular, we set $M_{min} = 2$, the time between each time point equals 1 min, and $t_0 = 0$. The results are summarized in the following section.

6.1 Results

We applied the EM algorithm and direct numerical maximization to obtain the MLE of the model parameters. Different initial values have been used so that the resulting MLE is likely to achieve the global maximum. The standard errors of the MLE are estimated by inverting the observed information matrix, as discussed in Sect. 3. The results are summarized in Table 3.

Compared with direct numerical maximization, the EM algorithm gives a larger value of likelihood function and smaller values of standard errors of parameter estimates. Thus, the parameter estimates found by the EM algorithm are used for the prediction of hidden states and future earthquakes. The interpretation of the fitted model is as follows. With $(\hat{\pi}_0, \hat{\pi}_1) = (0.0042, 0.098)$, it is 25 times more likely to experience an earthquake at state 1 (active state) than state 0 (inactive state). Moreover, from

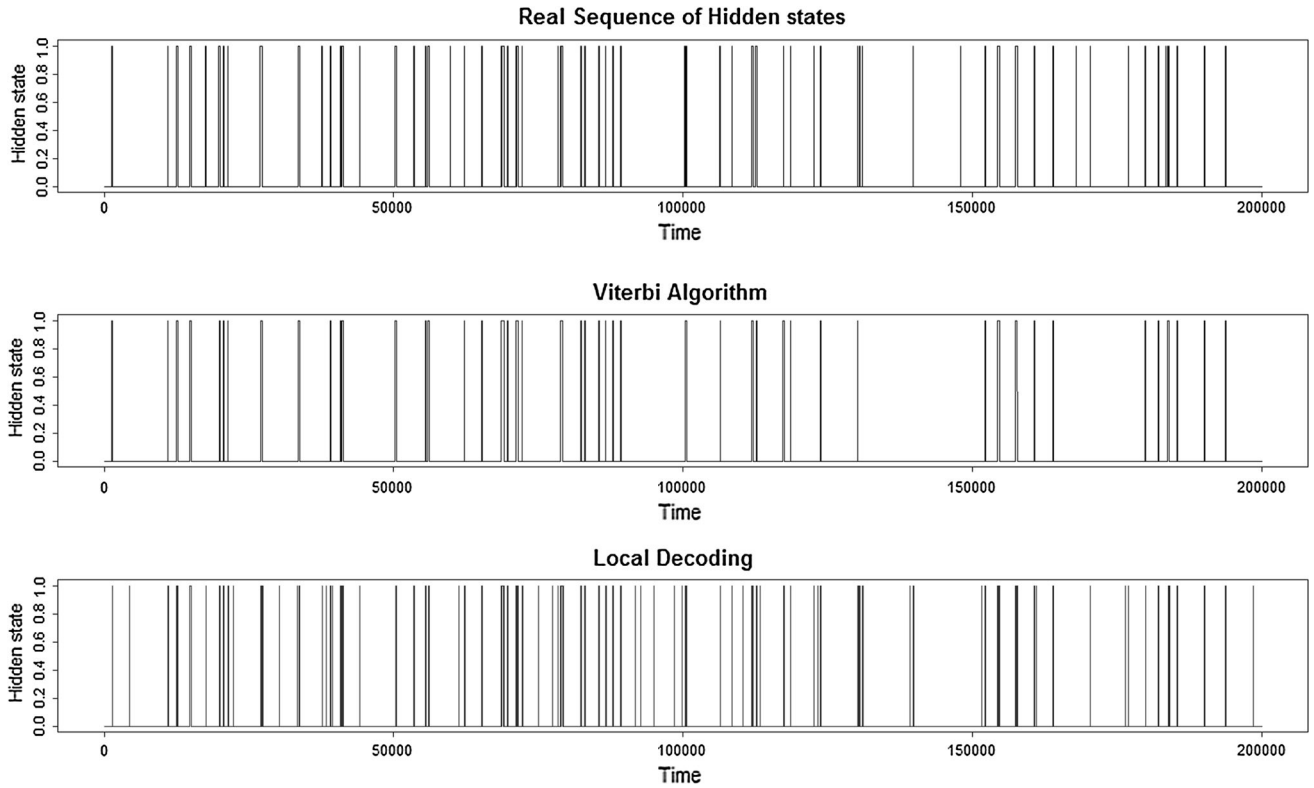

Fig. 5 Estimates of states for a subsample of size $N = 200,000$ under the case $N = 5,000,000$. True states (top); estimated states by the Viterbi algorithm (middle); estimated states by local decoding (bottom)

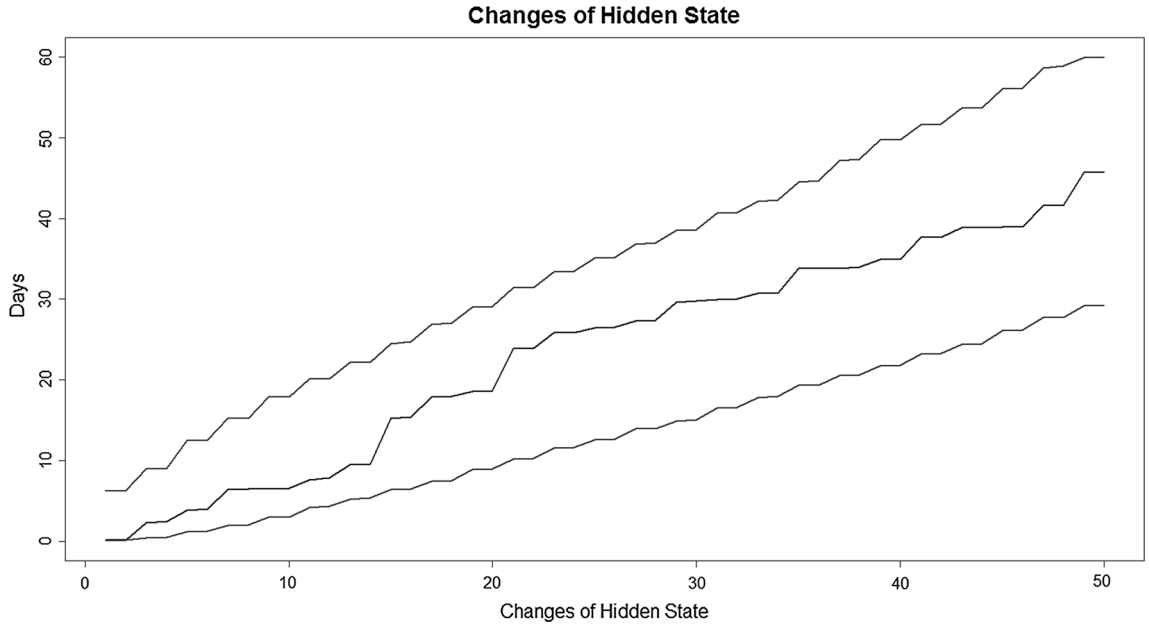

Fig. 6 Prediction of 50 future changes in hidden state with prediction intervals; true change-points (middle); the prediction intervals (top and bottom)

$(\hat{\lambda}_0, \hat{\lambda}_1) = (2.5402, 1.9564)$, state 0 has smaller earthquake magnitudes than state 1 on average. The parameters α_0 and β_0 represent the baseline effect on the transition probabilities from state 0 to 1 and from state 1 to 0, respectively. In particular, with covariate $T_n = 0$, the baseline transition probability matrix is

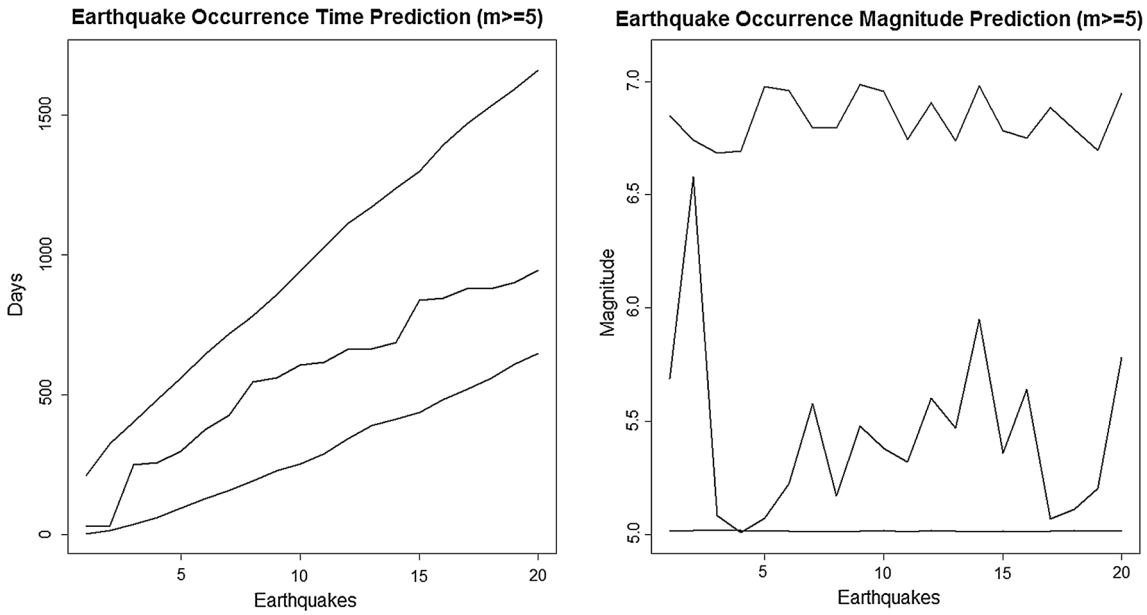

Fig. 7 Prediction of 20 future earthquakes of magnitudes $\geqslant 5$ with prediction intervals on occurrence time (left) and magnitudes (right); true change-points (middle); prediction intervals (top and bottom)

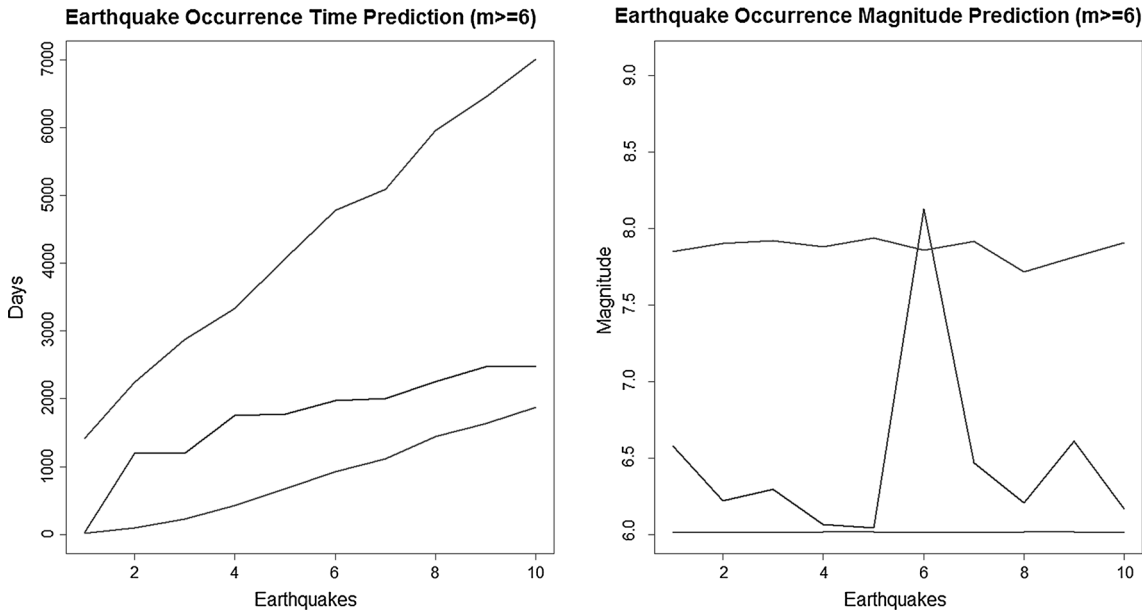

Fig. 8 Prediction of 10 future earthquakes of magnitudes $\geqslant 6$ with prediction intervals on occurrence time (left) and magnitudes (right); true change-points (middle); prediction intervals (top and bottom)

$$\mathbf{P} = \begin{pmatrix} \dfrac{1}{1+\exp(\hat{\alpha}_0)} & \dfrac{\exp(\hat{\alpha}_0)}{1+\exp(\hat{\alpha}_0)} \\ \dfrac{\exp(\hat{\beta}_0)}{1+\exp(\hat{\beta}_0)} & \dfrac{1}{1+\exp(\hat{\beta}_0)} \end{pmatrix} = \begin{pmatrix} 0.9995 & 0.0005 \\ 0.0172 & 0.9828 \end{pmatrix}, \tag{17}$$

indicating that it is more likely for the hidden process to transit from state 1 to 0 than from state 0 to 1. Moreover, in the long run, the approximate distribution of hidden states is

$$\hat{\mathbf{P}}^n = \begin{pmatrix} 0.9995 & 0.0005 \\ 0.0172 & 0.9828 \end{pmatrix}^n \rightarrow \begin{pmatrix} 0.9718 & 0.0282 \\ 0.9718 & 0.0282 \end{pmatrix} \text{as } n \rightarrow \infty, \tag{18}$$

and thus the system is more likely to remain in state 0.

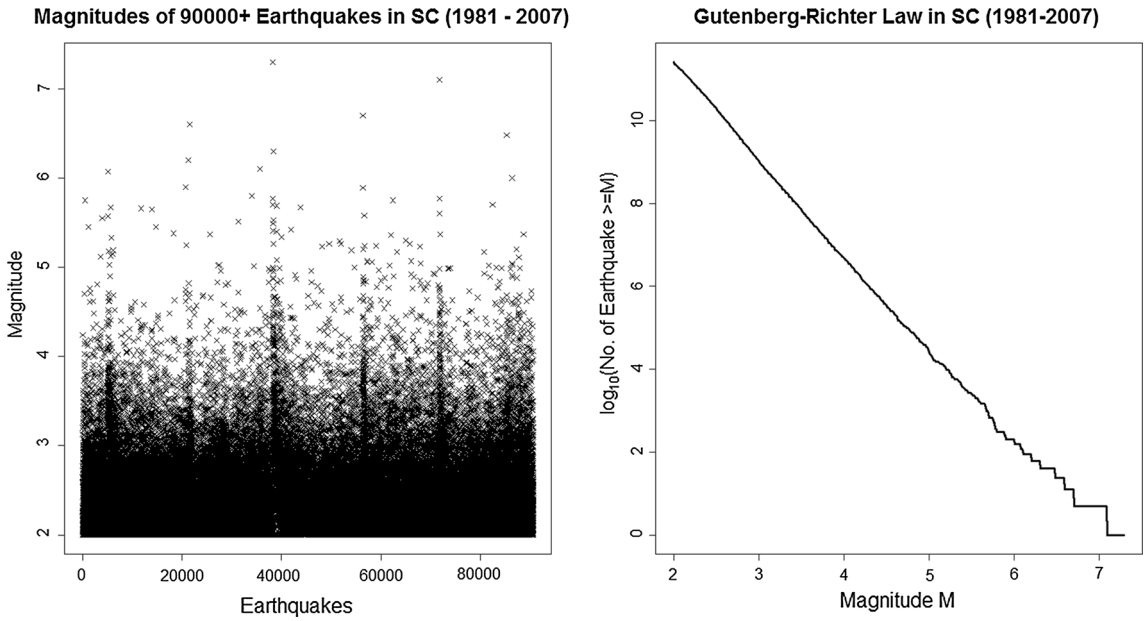

Fig. 9 Magnitudes of earthquakes with $M_{min} = 2$ between 1981 and 2007 in Southern California (left) and logN_m against magnitude m, where N_m is the number of earthquakes with magnitude greater or equal to m (right)

Table 3 MLE and standard errors (in brackets) by the EM algorithm and direct numerical maximization of likelihood function of Models (7) and (19)

Model (7)			Model (19)		
MLE	EM algorithm	Direct numerical maximization	MLE	EM algorithm	Direct numerical maximization
$\begin{pmatrix}\hat{\lambda}_0\\ \hat{\lambda}_1\end{pmatrix}$	2.5402 (0.0106) 1.9564 (0.0114)	2.5399 (0.0110) 1.9565 (0.0119)	$\begin{pmatrix}\hat{\lambda}_0\\ \hat{\lambda}_1\end{pmatrix}$	2.5354 (0.0105) 1.9504 (0.0114)	2.5354 (0.0108) 1.9504 (0.0115)
$\begin{pmatrix}\hat{\pi}_0\\ \hat{\pi}_1\end{pmatrix}$	0.0042 (0.00002) 0.0980 (0.0005)	0.0042 (0.00001) 0.0980 (0.0010)	$\begin{pmatrix}\hat{\pi}_0\\ \hat{\pi}_1\end{pmatrix}$	0.0042 (0.00002) 0.0904 (0.0005)	0.0042 (0.00002) 0.0904 (0.0005)
$\begin{pmatrix}\hat{\alpha}_0\\ \hat{\alpha}_1\end{pmatrix}$	−7.6489 (0.0296) −0.007902 (0.0003)	−7.6534 (0.0538) −0.007927 (0.0005)	$\hat{p}_{01}$	0.0001 (0.000004)	0.0001 (0.000004)
$\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix}$	−4.0452 (0.0281) −0.137088 (0.0049)	−4.0569 (0.0641) −0.136614 (0.0140)	$\hat{p}_{10}$	0.0044 (0.0002)	0.0044 (0.0002)
Maximum likelihood	−498408.8	−498410.9	Maximum likelihood	−498968.7	−498968.7

Next, α_1 and β_1 measure the effect of the time that has passed since the last earthquake T_n on transition probabilities. The negative estimates $\hat{\alpha}_1 = -0.008$ and $\hat{\beta}_1 = -0.137$ indicate that a large T_n favors the hidden process remaining in the same state. Indeed, whether a large T_n results in a longer expected waiting time until the next earthquake has been considered by Davis et al. (1989) and Sornette and Knopoff (1997). It is found that when T_n follows heavy-tailed distributions, the expected waiting time for the next earthquake increases when T_n increases,

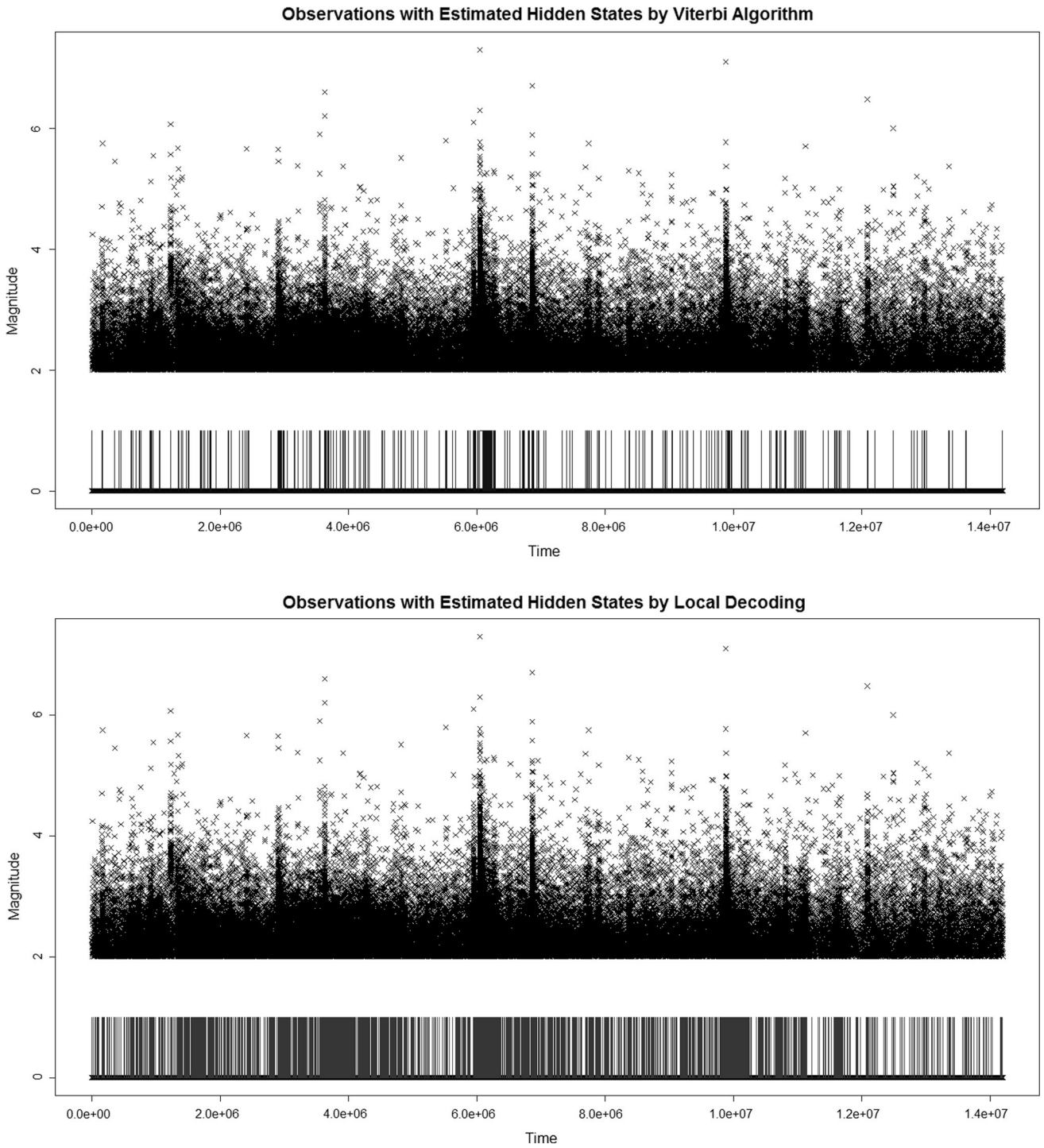

Fig. 10 Estimate of hidden states by the Viterbi algorithm (top) and local decoding (bottom)

and vice versa for T_n following thin-tailed distributions. As a heavy-tailed distributed T_n corresponds to less frequent earthquake occurrence (state 0), a longer waiting time for the next earthquake agrees with our quantitative finding that the hidden process tends to stay in state 0. Analogously, as a thin-tailed distributed T_n corresponds to frequent earthquake occurrence (state 1), a shorter waiting time for the next earthquake agrees with our quantitative finding that the hidden process tend to remain in state 1.

6.2 Prediction

Using the parameters estimated by the EM algorithm, the hidden states are estimated by the Viterbi algorithm and

local decoding, see Fig. 10. Compared to the Viterbi algorithm, local decoding seems to overestimate the frequency of changes in the hidden state, which is similar to the findings in the simulation. Thus, the estimated current state $\hat{s}_N$ from the Viterbi algorithm is used for prediction. The prediction algorithm introduced in Sect. 4 is carried out to predict the future hidden state and earthquakes with magnitudes $\geqslant 5$ and $\geqslant 6$ between 2008 and 2015. Specifically, 1000 sample paths are simulated to construct the 95% prediction intervals (Figs. 11, 12 and 13).

Among the first 20 future earthquakes of magnitudes $\geqslant 5$, 19, and 18 earthquakes are within the 95% prediction intervals of the occurrence time and magnitude, respectively. For the one missed earthquake in the occurrence time prediction, the difference of the lower bound of the prediction interval and the earthquake is only 1 day. For the two missed earthquakes in the magnitude prediction, one has magnitude 5.00 while the lower bound is 5.01. Another missed earthquake with magnitude 7.2 is captured when we consider the prediction of earthquakes with magnitudes $\geqslant 6$ in Fig. 13. In addition, note that while several aftershocks with magnitudes $\geqslant 5$ occurred after the large earthquake of magnitude 7.2, which is shown as a long horizontal line in Fig. 12, the prediction intervals still cover the aftershocks.

6.3 Comparison with other models

In this section we compare the proposed Model (7) with a model without time-varying covariate, and a popular model for earthquake dynamics, the Epidemic Type Aftershock Sequence (ETAS) model by Ogata (1988, 1998).

6.3.1 Model without time-varying covariate

This model simplifies Model (7) by removing the time-varying covariate to model $\{S_n\}$, which reduces to a traditional homogeneous HMM. The modeling of earthquake catalogue $\{A_n\}$ follows (7) except that $\{T_n\}$ is not included in the modeling of $\{S_n\}$. In summary, the reduced earthquake prediction model for the earthquake catalogue $\{A_n\}$ is given as

$$A_n = M_n \mathbb{I}_n, \tag{19}$$

where $M_n \sim$ Left-truncated $\exp(\lambda_{S_n}, M_{min})$, $\mathbb{I}_n \sim$ Bernoulli (π_{S_n}),

$$\begin{aligned}
&\mathbb{P}(S_n = 1 | S_{n-1} = 0) = p_{01_n} = p_{01}, \\
&\mathbb{P}(S_n = 0 | S_{n-1} = 1) = p_{10_n} = p_{10}, \\
&\quad p_{00_n} = 1 - p_{01}, \text{and } \&\mathrm{p}_{11_n} = 1 - \mathrm{p}_{10}.
\end{aligned}$$

The parameters of interest are collected as

$$\begin{aligned}
&\boldsymbol{\theta} = \{\lambda_0, \lambda_1, \pi_0, \pi_1, p_{01}, p_{10}\} \in \\
&\boldsymbol{\Theta} = \mathbb{R}^+ \times \mathbb{R}^+ \times (0,1) \times (0,1) \times (0,1) \times (0,1).
\end{aligned}$$

The estimation results are summarized in Table 3. Note that the estimates of λ_s and π_s in Models (7) and (19) are similar because the earthquake occurrence and magnitude depend on the state itself but not on the state dynamics. On the other hand, Model (19) does not have α_1 and β_1 as in Model (7), and thus does not capture the effect of covariates.

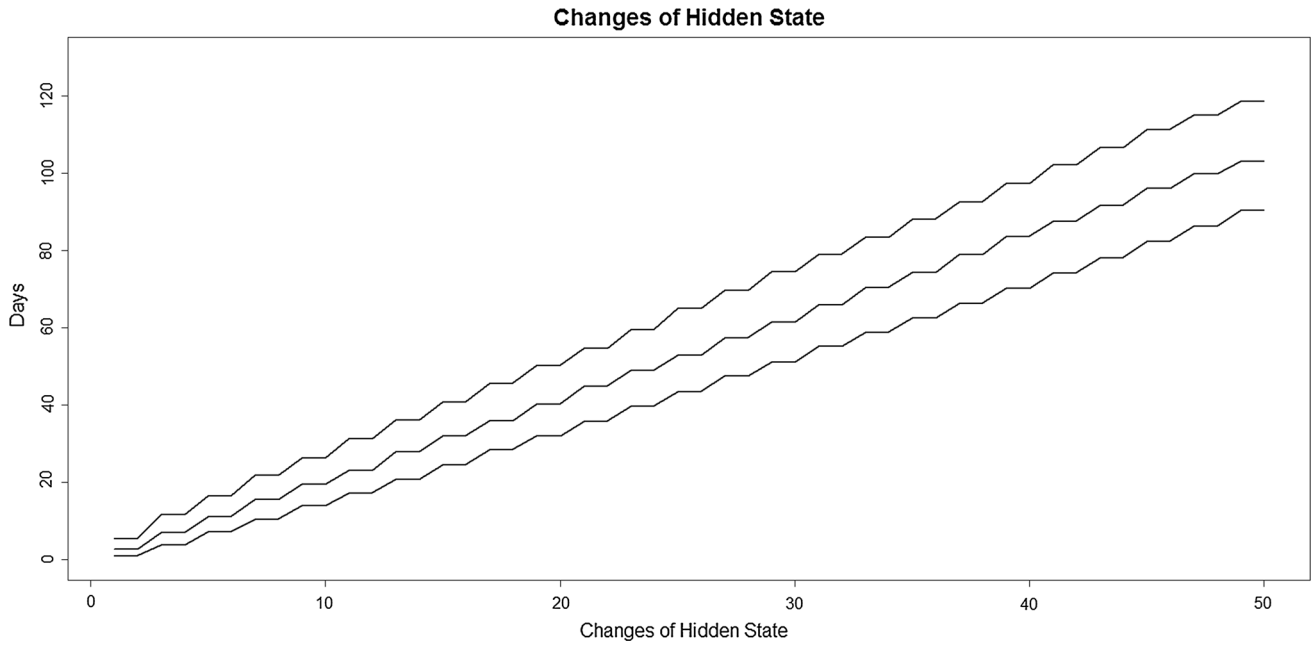

Fig. 11 Prediction of 50 future changes in hidden state with line plots; median (middle); interquartile range (top and bottom)

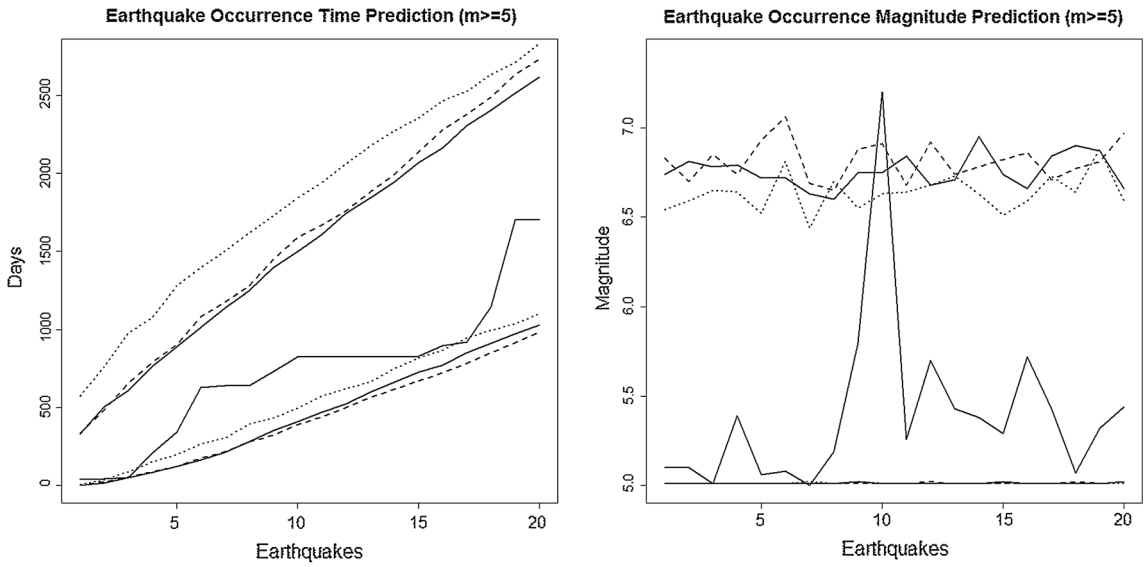

Fig. 12 Prediction of occurrence time (left) and magnitudes (right) of 20 future earthquakes of magnitudes greater or equal to 5; true value (middle; solid); 95% prediction intervals by Model (7) (top and bottom; solid); 95% prediction intervals by model without time covariate (top and bottom; dashed); 95% prediction intervals by the ETAS model (top and bottom; dotted)

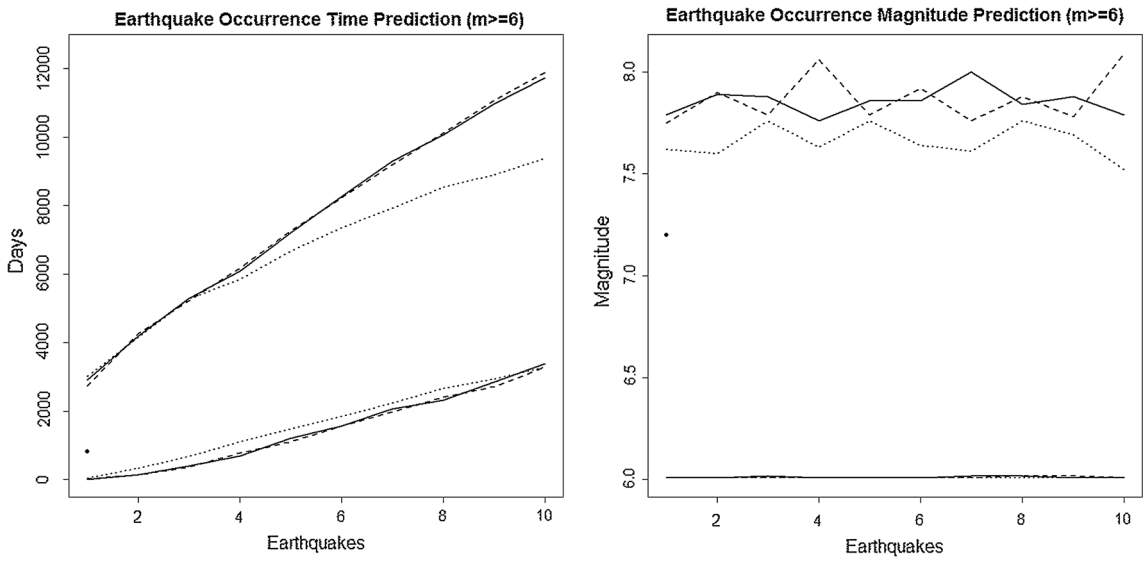

Fig. 13 Prediction of occurrence time (left) and magnitudes (right) of 10 future earthquakes of magnitudes greater or equal to 6; true value (middle); 95% prediction intervals by Model (7) (top and bottom; solid); 95% prediction intervals by model without time covariate (top and bottom; dashed); 95% prediction intervals by the ETAS model (top and bottom; dotted)

6.3.2 ETAS model

The ETAS model consists of a marked Poisson process $N(t)$ at a particular time point $t \in [0, T]$ with magnitude M as the mark, where $[0, T]$ is the period of study. The process has a self-exciting conditional intensity $\lambda(t, M|\mathcal{D}_t)$ depending on the history $\mathcal{D}_t = \{(t_n, M_n); t_n < t\}$ of the earthquake occurrence time $\{t_n\}$ and magnitude $\{M_n\}$ up to time t, where $n = 1, 2, \ldots, N$ and N is the total number of earthquakes in the catalogue. In particular, the conditional intensity is defined by

$$\lambda(t, M|\mathcal{D}_t) = j(M)\lambda(t|\mathcal{D}_t), \tag{20}$$

where $j(M)$ is the distribution of the magnitude, assumed to be conditionally independent of the occurrence time and locations, and

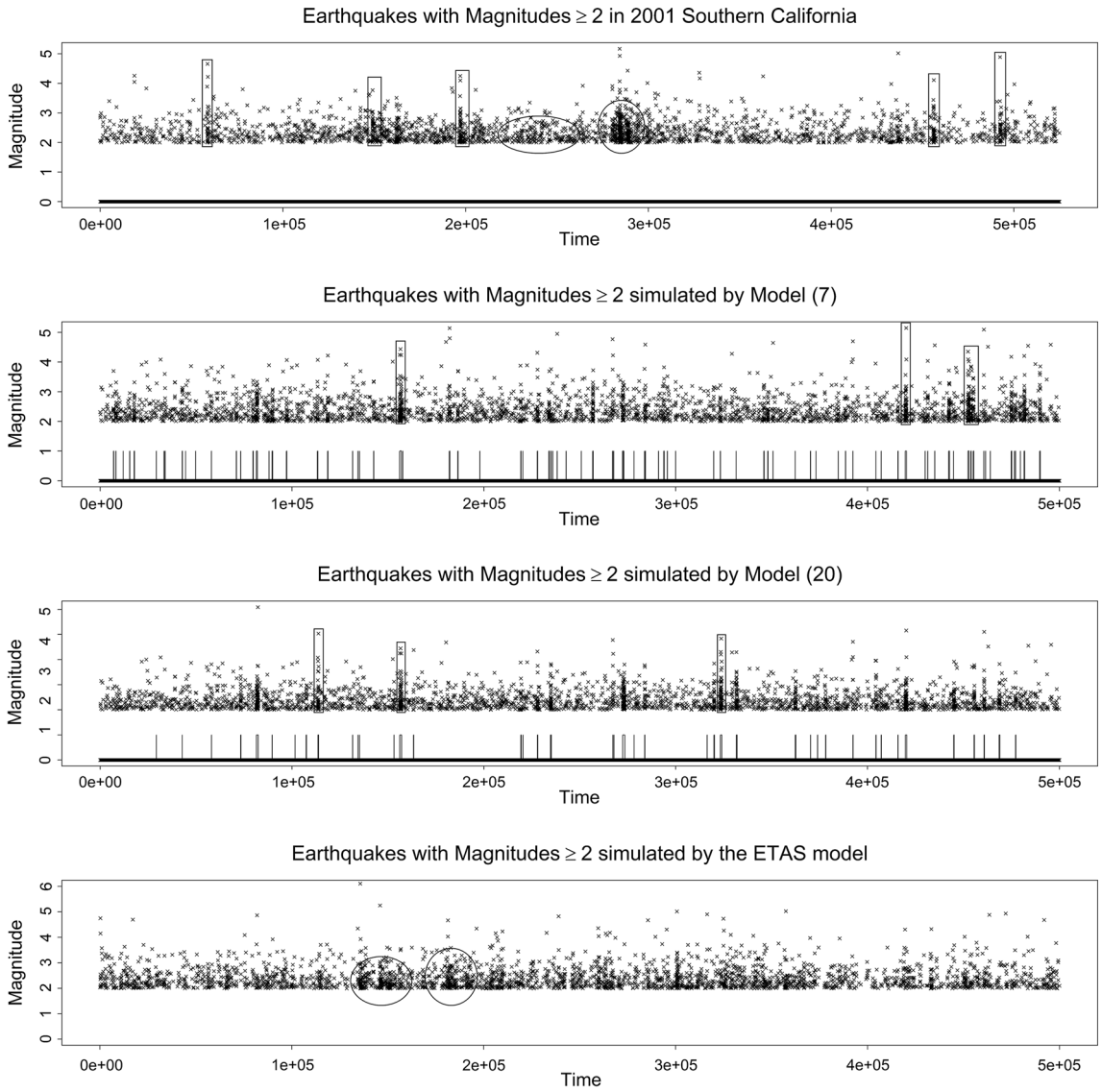

Fig. 14 Earthquakes with magnitudes $\geqslant M_{min} = 2$, where M_{min} is the magnitude of completeness of the catalogue. In 2001 Southern California (top); simulated by Model (7) (row 2); simulated by Model (19) (row 3); simulated by the ETAS model (bottom)

$$\lambda(t|\mathcal{D}_t) = \mu + \sum_{\{n:t_n<t\}} Ae^{\alpha(M_n - M_{min})}\left(1 + \frac{t - t_n}{c}\right)^{-p}. \qquad (21)$$

The parameters of interest are $\boldsymbol{\theta} = (\mu, A, \alpha, c, p)$. Note that the second term in (21) describes the *aftershocks* effect from each previous earthquake, which helps to explain the earthquake clustering behavior.

6.3.3 Prediction comparison

In this section, we compare the performance of the proposed Model (7) with Model (19) and the ETAS model using the earthquake catalogue in Southern California. Analogous to Sect. 6.2, we fit the Model (19) and ETAS model using the data from 1981 to 2007, and follow Steps 3 and 5 of the prediction algorithm in Sect. 4 to predict future earthquakes from 2008 to 2015. The R package **PtProcess** is used for the fitting of the ETAS model, and the thinning method by Ogata (1981) is used for simulating the marked point processes (Team 2016). The results are shown in Figs. 12, 13 and 14.

First, we compare the fitting of Model (7), (19) and the ETAS model. Figure 14 shows a typical year of earthquake observations in Southern California, and simulated earthquake observations by Model (7), (19) and the ETAS model using the corresponding estimated parameters. Observe that clusters of earthquakes exist in both figures of simulated observations. However, the clusters from the ETAS model are usually triangular-shaped and appear after an earthquake with a large magnitude (indicated by circles

in Fig. 14), while the clusters from Model (7) and (19) are usually finger-shaped (indicated by rectangles in Fig. 14). Model (7) gives finger-shaped clusters with varying width while Model (19) gives thin finger-shaped clusters. The triangular-shaped features of the ETAS model is a consequence of (21) that aftershocks depend on the magnitude of the original earthquake and decays with the passing of time. On the other hand, the finger-shaped feature in the proposed model results from the dynamic of the hidden Markov process in which it tends to remain in the active state for a relatively short period. Interestingly, it appears that both triangular-shaped and finger-shaped features appear in real earthquake observations. Thus, the theory of aftershock and state-changes in underground dynamics are both adequate in describing real earthquake data.

Finally, we compare the prediction results of the three models. From Fig. 12, the 95% prediction intervals by Models (7) and (19) both cover 19 earthquake occurrences and 18 earthquake magnitudes, while the corresponding 95% prediction intervals by the ETAS model only covers 18 earthquake occurrences and 17 earthquake magnitudes. More importantly, the proposed Model (7) provides a more accurate prediction than Model (19) and the ETAS model, with a narrower 95% prediction interval for predicting the first 20 earthquakes of magnitude $\geqslant 5$. This suggests that Model (7), which incoporates covariates in the dynamic of the hidden state, gives better performance than the models without covariates. For the prediction of first 10 earthquakes of magnitude $\geqslant 6$, Fig. 13 shows that the ETAS model gives a narrower 95% prediction interval comparing to the Models (7) and (19). However, the accuracy of the prediction is confined by the small number of earthquakes with magnitude $\geqslant 6$ in the dataset, which is only 10 among more than 90,000 earthquakes in the dataset for model fitting. Also, there is only 1 earthquake with magnitude $\geqslant 6$ among 20,000 earthquakes in the testing dataset for prediction. Thus, more data are required to compare the accuracy of the prediction of all the three models in earthquakes of magnitude $\geqslant 6$.

7 Conclusion and future research directions

In this paper, we develop a novel Hidden Markov Model for earthquake modeling and forecasting by introducing a latent Markov process to model the unobservable state of the underground dynamics. The proposed model is capable of predicting the change-in-state of the hidden Markov chain, and thus can predict the arrival time and magnitude of future earthquake occurrences simultaneously. Specifically, the transition probabilities are modeled by possibly time-varying covariates and past observations using a link function. This enables better understandings of the underground dynamics of earthquakes and prediction of future hidden states, and thus future earthquakes' frequencies and magnitudes. In particular, the time that has passed since the previous earthquakes and logistic function are chosen as the covariate and the link function respectively in our study. Simulation studies and applications with comparison with the ETAS model on a real earthquake dataset indicate the validity of the proposed model on modeling earthquakes occurrence. Theoretical results, including the stationarity and ergodicity of the proposed model, as well as consistency and asymptotic normality of model parameter estimation, are established in the Supplementary Material.

Some directions of further investigations related to extending the scope of the model or improving the modeling of earthquake mechanism are summarized as follows. For example, the periodicities and long-term trend of earthquake time series (Telesca et al. 2015), and the occurrence of large earthquakes when a long time has passed from the last large earthquakes (Fierro and Leiva 2017), may be included in the model. Moreover, the generalization to a space-time model can incorporate spatial dependence and possible location-specific covariates on modeling earthquake occurrences (Adelfio and Chiodi 2015).

References

Adelfio G, Chiodi M (2015) Alternated estimation in semi-parametric space-time branching-type point processes with application to seismic catalogs. Stoch Environ Res Risk Assess 29:443450

Bartolucci F, Farcomeni A (2009) A multivariate extension of the dynamic logit model for longitudinal data based on a latent Markov heterogeneity structure. J Am Stat Assoc 104:816–831

Bartolucci F, Farcomeni A (2015) A discrete time event-history approach to informative drop-out in mixed latent Markov models with covariates. Biometrics 71:80–89

Bartolucci F, Farcomeni A (2015a) Information matrix for hidden Markov models with covariates. Stat Comput 25:515–526

Bartolucci F, Farcomeni A, Pennoni F (2013) Latent Markov models for longitudinal data. Chapman & Hall, CRC Press, Boca Raton

Baum LE, Petrie T, Soules G, Weiss N (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Ann Math Stat 41(1):164–171

Baum LE (1972) An equality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. Inequalities 3:1–8

Bulla J, Berzel A (2008) Computational issues in parameter estimation for stationary hidden Markov models. Comput Stat 23(1):1–18

Cameletti M, De Rubies V, Ferrari C, Sbarra P, Tosi P (2016) An ordered probit model for seismic intensity data. Stoch Environ Res Risk Assess 1–10. doi:10.1007/s00477-016-1260-4

Cosentino P, Ficarra V, Luzio D (1977) Truncated exponential frequency-magnitude relationship in earthquake statistics. Bull Seismol Soc Am 67(6):1615–1623

Davis PM, Jackson DD, Kagan YY (1989) The longer it has been since the last earthquake, the longer the expected time till the next? Bull Seismol Soc Am 79(5):1439–1456

Ebel JE, Chambers DW, Kafka AL, Baglivo JA (2007) Non-Poissonian earthquake clustering and the hidden Markov model as bases for earthquake forecasting in California. Seismol Res Lett 78(1):57–65

Fierro R, Leiva V (2017) A stochastic methodology for risk assessment of a large earthquake when a long time has elapsed. Stoch Environ Res Risk Assess 1–10. doi:10.1007/s00477-016-1288-5

Finazzi F, Fassò A (2016) A statistical approach to crowdsourced smartphone-based earthquake early. Stoch Environ Res Risk Assess 1–10. doi:10.1007/s00477-016-1240-8

Huang Q, Ikeya M (1998) Seismic electromagnetic signals (SEMS) explained by a simulation experiment using electromagnetic waves. Phys Earth Planet Inter 109(3):107–114

Hughes JP, Guttorp P, Charles SP (1999) A non-homogeneous hidden Markov model for precipitation occurrence. J R Stat Soc Ser C (Appl Stat) 48(1):15–30

Hutton K, Woessner J, Hauksson E (2010) Earthquake monitoring in southern California for seventy-seven years (1932–2008). Bull Seismol Soc Am 100(2):423–446

Leroux BG (1992) Maximum-likelihood estimation for hidden Markov models. Stoch Process Appl 40(1):127–143

Leroux BG, Puterman ML (1992) Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models. Biometrics 48(2):545–558

Levinson SE, Rabiner LR, Sondhi MM (1983) An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. Bell Syst Tech J 62(4):1035–1074

Ogata Y (1981) On Lewis' simulation method for point processes. IEEE Trans Inf Theory 27(1):23–31

Ogata Y (1988) Statistical models for earthquake occurrences and residual analysis for point processes. J Am Stat Assoc 83(401):9–27

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

Orfanogiannaki K, Karlis D, Papadopoulos GA (2010) Identifying seismicity levels via Poisson hidden Markov models. Pure Appl Geophys 167(8–9):919–931

Park SK, Johnston MJ, Madden TR, Morgan FD, Morrison HF (1993) Electromagnetic precursors to earthquakes in the ULF band: a review of observations and mechanisms. Rev Geophys 31(2):117–132

Reid HF (1910) The mechanics of the earthquake, vol. II of Lawson, A.C., chairman, The California earthquake of April 18, 1906: report of the State Earthquake Investigation Commission: Carnegie Institution of Washington Publication, p 192 (reprinted in 1969)

Seeber L, Armbruster JG (2000) Earthquakes as beacons of stress change. Nature 407(6800):69

Sornette D, Knopoff L (1997) The paradox of the expected time until the next earthquake. Bull Seismol Soc Am 87(4):789–798

Team RC (2016) R: language and environment for statistical computing. R Foundation for Statistical Computing, 2005; Vienna, Austria

Telesca L, Giocoli A, Lapenna V, Stabile TA (2015) Robust identification of periodic behavior in the time dynamics of short seismic series: the case of seismicity induced by Pertusillo Lake, southern Italy. Stoch Environ Res Risk Assess 29:1437. doi:10.1007/s00477-014-0980-6

Utsu T (1961) A statistical study on the occurrence of aftershocks. Geophys Mag 30(4):521–605

Viterbi AJ (1967) Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Trans Inf Theory 13(2):260–269

Welch LR (2003) Hidden Markov models and the Baum–Welch algorithm. IEEE Inf Theory Soc Newsl 53(4):10–13

Zucchini W, MacDonald IL (2009) Hidden Markov models for time series: an introduction using R. CRC Press, Boca Raton

Zucchini W, Raubenheimer D, MacDonald IL (2008) Modeling time series of animal behavior by means of a latent-state model with feedback. Biometrics 64(3):807–815