EM algorithm for the estimation of the RETAS model

Tom Stindl and Feng Chen†
Department of Statistics, UNSW Sydney
August 17, 2023

Abstract

The Renewal Epidemic-Type Aftershock Sequence (RETAS) model is a recently proposed point process model that can fit event sequences such as earthquakes better than pre-existing models. Evaluating the log-likelihood function and directly maximizing it has been shown to be a viable approach to obtain the maximum likelihood estimator (MLE) of the RETAS model. However, the direct likelihood maximization suffers from numerical issues such as premature termination of parameter searching and sensitivity to the initial value. In this work, we propose to use the Expectation-Maximization (EM) algorithm as a numerically more stable alternative to obtain the MLE of the RETAS model. We propose two choices of the latent variables, leading to two variants of the EM algorithm. As well as deriving the conditional distribution of the latent variables given the observed data required in the E-step of each EM-cycle, we propose an approximation approach to speed up the E-step. The resulting approximate EM algorithms can obtain the MLE much faster without compromising on the accuracy of the solution. These newly developed EM algorithms are shown to perform well in simulation studies and are applied to an Italian earthquake catalog.

Keywords: MLE, point process, renewal process, seismology, self-exciting.

---

*This research includes computations using the Linux computational cluster Katana supported by the Faculty of Science, UNSW Sydney, and the National Computational Infrastructure (NCI) supported by the Australian Government.
†Chen was partly supported by a UNSW SFRGP grant.
1 Introduction

The self-exciting point process of Hawkes (1971) has an intensity process which consists of a long-term constant background rate plus short-term self-excitement effects due to the points (events) in its history. By the cluster process representation of the Hawkes process (Hawkes and Oakes, 1974), each event of the process can be interpreted as either an immigrant or an offspring, where the immigrants arrive according to a homogeneous Poisson process and each immigrant potentially generates offspring according to a common inhomogeneous Poisson process, and each offspring can then trigger additional offspring of their own according to the same inhomogeneous Poisson process.

Since the introduction of the Hawkes (1971) process, several model innovations have been proposed in the literature with widespread use in many application domains. The RHawkes process (Wheatley et al., 2016) is one such innovation that replaces the constant background rate in the intensity of the Hawkes process with a time-dependent function that resets on the arrival of each immigrant and then evolves according to a fixed function of time. This is equivalent to replacing the homogeneous Poisson immigrant arrival process with a renewal process where the waiting times between successive events have the fixed function of time as their common hazard rate function. Similar to the Poisson process, the renewal process has independent and identically distributed (i.i.d.) inter-event waiting times, but it differs in that a general waiting time distribution can be used instead of the exponential distribution. With this generalization of the immigrant arrival process it becomes possible to model diverse temporal clustering characteristics such as those with roughly evenly distributed clusters, and those with occasional large gaps between events.

The ETAS model (Ogata, 1988) is an extended version of the Hawkes self-exciting point process where the excitation effect of an event is allowed to depend on the mark associated with the event (e.g. the magnitude of an earthquake). The ETAS model also has a spatiotemporal extension (Ogata, 1998) which has been influential in various fields, such as seismology (Ogata and Zhuang, 2006; Guo et al., 2015), criminology (Mohler et al., 2011; Lewis and Mohler, 2011; Zhuang and Mateu, 2019) and epidemiology (Meyer et al., 2012). Specifically, in seismology, the immigrant events can be interpreted as main-shocks
and the offspring events as aftershocks. In the same spirit of the RHawkes process being a modification to the Hawkes process, the RETAS model (Stindl and Chen, 2022, 2023) is a modification to the ETAS model, where the main-shock arrival times follow a renewal process rather than a homogeneous Poisson process. The RETAS model might be considered a special case of the general modulated renewal process proposed by Lin and Fine (2009) with the backward recurrence time $V_i$ defined as the time lapsed from the previous immigrant (mainshock), although an important difference to note is that the backward recurrence time is unobserved in the RETAS model, while it is assumed part of the data in Lin and Fine (2009).

Likelihood based inferences such as maximum likelihood estimation are typically straightforward to implement for self-exciting point processes. However, this is not the case for the RETAS model since the conditional intensity process depends on the unobserved index of the most recent immigrant, and the standard formula for point processes likelihood cannot be used directly to calculate its likelihood. Instead, Stindl and Chen (2022) proposed a recursive algorithm to calculate the likelihood and demonstrated that the MLE can be obtained by numerically maximizing the likelihood. However, direct optimization of the likelihood might be numerically unstable or computationally costly, as noted by Veen and Schoenberg (2008) for the ETAS model. These issues can be mitigated using the Expectation-Maximization (EM) algorithm of Dempster et al. (1977), as shown in Marsan and Lengliné (2008); Veen and Schoenberg (2008); Lewis and Mohler (2011).

Rather than maximizing the log-likelihood directly, the EM algorithm iteratively maximizes the conditional expectation of the complete data log-likelihood given the observed data and the current estimate of the parameter before setting the maximizer as the updated parameter. This achieves the same objective of maximizing the observed data log-likelihood while mitigating some of the numerical instabilities of direct maximization caused by multimodality and flat surfaces. Although the EM algorithm does not completely resolve the issues associated with converging to a local maximum, its numerical performance tends to be more stable. The estimation of the ETAS and RETAS model can be posed in the EM framework where the observed data consists of the occurrence times of the earthquakes and
the associated marks such as the epicenters and magnitudes and the branching structure consisting of the indicators for main-shocks and triggered aftershocks serves as the set of unobserved latent variables.

Wheatley et al. (2016) proposed an EM algorithm in the spirit of Dempster et al. (1977) for the RHawkes process. This is a promising endeavor due to its success for the Hawkes and ETAS models. The approach by Wheatley et al. (2016) modified the EM algorithms of Veen and Schoenberg (2008); Marsan and Lengliné (2008) and Lewis and Mohler (2011) to allow for a renewal immigrant arrival process. They also introduced a version of the EM algorithm with a reduced set of latent variables that only included a binary variable for immigrants rather than the complete branching structure. However, in their implementation of the EM algorithm, the expectation in the E-step is calculated with respect to a conditional distribution that conditions on a varying subset of the observed data rather than the entire set. For the classical ETAS model this did not cause concern because of the conditional independence of past immigrant/offspring indicators and future events times given past event times. However, this is not the case for the RETAS model and the objective of this article is to amend the aforementioned EM algorithms of Wheatley et al. (2016) by conditioning on the whole set of observed data in implementing the E-step.

Hence, we propose a recursive algorithm to calculate the conditional probabilities needed in the E-step. We first calculate the filtered most recent immigrant probabilities which only condition on a subset of the observed data. Then, we smooth the filtered probabilities to obtain the smoothed most recent immigrant probabilities which condition on the entire observed data. These probabilities facilitate the calculation of the smoothed branching structure probabilities such as the immigrant and parent probabilities. The M-step can then be implemented as in Wheatley et al. (2016). We also provide a strategy to reduce the computational time of the EM algorithm that follows the procedure discussed in Stindl and Chen (2021) to accelerate the calculation of the log-likelihood for the RHawkes process. This approach truncates the distribution of the most recent immigrant and restricts the longevity of the self-excitation effects.

The remainder of this article is devoted to the following. Section 2 describes the RETAS
model and the log-likelihood evaluation algorithm. Section 3 presents the whole data EM algorithm for the RETAS model. Section 4 introduces an approximation in the E-step to accelerate the EM algorithms. Section 5 provides numerical evidence that the EM algorithms converge to the MLE using simulated data, and a comparison to direct log-likelihood maximization to obtain the MLE is also presented. The article concludes with a case study of an Italian earthquake catalog in Section 6.

2 RETAS model

Let the occurrence time be denoted by $\tau_i$ with $\tau_{i-1} < \tau_i$ and the associated mark be denoted by $m_i$. The cumulative number of events up to and including time $t$ is given by the counting process $\{N(t), t \geq 0\}$ with $N(t) = \sum_{i=1}^n 1\{\tau_i \leq t\}$, where $1\{\cdot\}$ is the indicator function. Suppose that the marked point process is realized over the interval $[0, T]$ with $T$ denoting the censoring time, $n = N(T)$ being the total number of events, $\tau_{1:n} = (\tau_1, \ldots, \tau_n)$ containing all the observed event times and $m_{1:n} = (m_1, \ldots, m_n)$ containing all the corresponding event marks. The conditional intensity process of the RETAS model describes the instantaneous arrival rate of events at time $t$ which depends on all past event occurrence times and marks. We define $H_{t-} = \{N(t-), (\tau_i, m_i) : i = 1, 2, \ldots, N(t-)\}$ to be the internal history of the process just prior to time $t$. We also extend the definition of internal history to include the index of the most recent immigrant $I(t) = \max \{i : \tau_i < t, B_i = 0\}$ where $B_i = 0$ indicates that the $i$-th event is an immigrant, and define $\tilde{H}_{t-} = H_{t-} \cup \{I(s), 0 \leq s \leq t\}$ to be the extended history.

The conditional intensity process for the RETAS model with respect to the extended history $\tilde{H}_{t-}$ is a superposition of a time-varying background rate that renews at immigrant arrivals and a self-exciting term consisting of cumulative excitation effects due to past events

$$\lambda(t|\tilde{H}_{t-}) = \mu(t - \tau_{I(t)}) + \sum_{j=1}^{N(t-)} \psi(t - \tau_j, m_j),$$

see Eq. (1) in Stindl and Chen (2019). The function $\mu : \mathbb{R}^+ \to \mathbb{R}^+$ is a deterministic hazard rate function of the renewal distribution for the waiting time between successive immigrants.
The excitation function $\psi(t - \tau_j, m_j)$ describes the intensity of an inhomogenous Poisson process for an offspring originating at the occurrence time $\tau_j$ with mark $m_j$. Herein, we assume that the excitation function is separable, $\psi(t, m) = g(t)k(m)$ and comprises of a temporal response function $g : \mathbb{R}^+ \to \mathbb{R}^+$ and a boost function $k : \mathcal{M} \to \mathbb{R}^+$ where $\mathcal{M}$ represents the mark space. If $g$ is a density function such that $\int_0^\infty g(t)dt = 1$ then the boost function $k$ determines the expected number of offspring due to a single event (with mark $m$) while the temporal response function describes the arrival time of the offspring event relative to its parent.

This model generalization introduces intricate dependencies between the clusters where a cluster consists of an immigrant and its multi-generational sequence of offspring. This additional flexibility comes at a cost as the classical formula of the observed data log-likelihood (see e.g. Daley and Vere-Jones, 2003) given by

$$\ell(\theta) = \sum_{i=1}^{N(T)} \log \lambda(\tau_i|H_{\tau_i-}) - \int_0^T \lambda(s|H_{s-})ds,$$

(2)

is not convenient for likelihood evaluation, since for the RETAS model, knowledge of the most recent immigrant event is required to define the renewal times. However, Chen and Stindl (2018) bypassed the formula in (2) and developed an algorithm to evaluate the likelihood recursively for the RHawkes process by conditioning on the most recent immigrant. The algorithm provides an alternate but equivalent expression to calculate the likelihood in (2) without resorting to first calculating the intensity process with respect to the internal history $H_t$, whose form is rather complex.

For ease of reference, we reproduce the log-likelihood evaluation algorithm for the RETAS model (adapted from Stindl and Chen, 2021). Let the filtered probabilities for the index of the most recent immigrant be denoted by $p_{ij} = \mathbb{P}(I(\tau_i) = j|\tau_1:i-1, m_1:i-1)$ for $j = 1, \ldots, i-1; i = 2, \ldots, n+1$. Define

$$\phi(t) = \sum_{j: \tau_j < t} g(t - \tau_j)k(m_j), \quad \Phi(T) = \sum_{j: \tau_j < T} k(m_j) \int_{\tau_j}^T g(t - \tau_j)dt,$$

and for $j = 1, \ldots, i-1; i = 2, \ldots, n+1$

$$d_{ij} = (\mu(\tau_i - \tau_j) + \phi(\tau_i))S_{ij}, \quad S_{ij} = e^{-\int_{\tau_{i-1}}^{\tau_i} \mu(t-\tau_i)dt},$$
where \( \tau_{n+1} := T \). Further define \( d_1 = \mu(\tau_1) e^{-\int_0^{\tau_1} \mu(t)dt} \). The log-likelihood of the RETAS model is given by
\[
\ell(\theta) = \log d_1 + \sum_{i=2}^{n} \log \left( \sum_{j=1}^{i-1} p_{ij} d_{ij} \right) + \log \left( \sum_{j=1}^{n} p_{n+1,j} S_{n+1,j} \right) + \Phi(T),
\]
and the \( p_{ij} \)'s are calculated recursively by
\[
p_{ij} = \begin{cases} 
p_{i-1,j} \Phi(\tau_{i-1}) S_{i-1,j}, & j = 1, \ldots, i - 2 \\
\sum_{k=1}^{i-2} p_{i-1,k} d_{i-1,k}, & i \geq 3,
\end{cases}
\]
with the initial condition \( p_{21} = 1 \) which implies that the first event is an immigrant.

### 3. EM algorithm for the RETAS model

The EM algorithm, originally proposed by Dempster et al. (1977) as a general method for obtaining the MLE of a statistical model, requires a choice of the set of unobserved latent variables, whose joint distribution with the observed data generally takes a tractable form. The EM algorithm is an iterative procedure that alternates between an Expectation step (E-step) which computes a criterion function, and a Maximization step (M-step) which maximizes this criterion function, until the maximizer in the M-step satisfies a pre-specified convergence criterion.

Let \( Y \) denote the observed data and \( B \) denote the unobserved latent variables. Then the criterion function in the E-step is the conditional expectation of the complete data log-likelihood \( \log L(\theta|Y, B) \) given the observed data \( Y \) and the current estimate of the model parameters \( \hat{\theta}^{[m]} \) (i.e. the maximizer from the previous M-step or the initial estimate), that is,
\[
Q(\theta) = \mathbb{E}_{B|Y,\hat{\theta}^{[m]}} \left[ \log L(\theta|Y, B) \right].
\]
Under rather general conditions, the sequence of maximizers in the M-step
\[
\hat{\theta}^{[m+1]} = \arg \max_\theta Q(\theta),
\]
will converge to a local maximizer of the observed data log-likelihood \( \log L(\theta|Y) \). It has been shown in Dempster et al. (1977) that each consecutive estimate of the EM cycle will not decrease the value of the observed data log-likelihood.

An EM-type algorithm has been used by Wheatley et al. (2016) to estimate the related but simpler (unmarked) RHawkes process. Their algorithm uses the branching structure as the unobserved latent variables \( B_{1:n} = (B_1, \ldots, B_n) \) in which \( B_i = 0 \) if \( \tau_i \) is the occurrence time of an immigrant, or \( B_i = j \) if \( \tau_i \) is the occurrence time of an offspring with parent event occurring at time \( \tau_j \). They also introduced a second EM algorithm in which the latent variable was reduced to a binary variable in which the label \( j \) is replaced with one to simply indicate that it is an offspring event. The observed data contains all the event times up to the censoring time \( T \). However, in their algorithms, while calculating the conditional distribution of the latent variable \( B_i \), instead of conditioning on the entire observed data \( \{\tau_{1:n}, \tau_{n+1} > T\} \), they only conditioned on the subset \( \{\tau_{1:i-1}\} \). Unfortunately, by reducing the conditioning information set, their algorithms do not lead to the MLE as expected of the EM algorithm (cf. the simulation studies in Chen and Stindl, 2018). Since the direct likelihood maximization can suffer from convergence issues and the EM algorithms of Wheatley et al. (2016) do not converge to the MLE, we propose bona fide EM algorithms that condition on the entire observed data in the E-step to overcome these issues. To distinguish from the EM algorithms of Wheatley et al. (2016), we refer to our EM algorithms as the whole data EM algorithms.

### 3.1 Whole data EM algorithm

Let \( Y = \{\tau_{1:n}, m_{1:n}, \tau_{n+1} > T\} \) denote the entire set of observed data. Similar to Wheatley et al. (2016), we use the branching structure \( B_{1:n} = (B_1, \ldots, B_n) \) as the latent variables. The complete data likelihood or the joint density of the observed data \( Y \) and the latent
The criterion function to be maximized in the M-step is the conditional expectation of $\log L(\theta)$ given the observed data $Y$ and the current parameter estimate $\hat{\theta}^{[m]}$ and is given by

$$Q_1(\theta) = \mathbb{E}_{B_1:n|Y,\hat{\theta}^{[m]}}[\log L(\theta)]$$

$$= \log \mu(\tau_1) - \int_0^{\tau_1} \mu(s)ds + \sum_{i=2}^{n-1} \sum_{j=1}^{i-1} \mathbb{P}(B_i = 0, I(\tau_i) = j|Y, \hat{\theta}^{[m]}) \log \mu(\tau_i - \tau_j)$$
$$- \sum_{i=2}^{n-1} \sum_{j=1}^{i-1} \mathbb{P}(I(\tau_i) = j|Y, \hat{\theta}^{[m]}) \int_{\tau_i}^{\tau_1} \mu(s - \tau_j)ds$$
$$- \sum_{j=1}^{n} \mathbb{P}(I(T) = j|Y, \hat{\theta}^{[m]}) \int_{\tau_n}^{T} \mu(s - \tau_j)ds$$
$$+ \sum_{i=2}^{n} \sum_{j=1}^{i-1} \mathbb{P}(B_i = 0|Y, \hat{\theta}^{[m]}) \log \psi(\tau_i - \tau_j, m_j) - \int_0^{T} \phi(s)ds. \quad (8)$$

Before optimization can be implemented in the M-step, a procedure to compute the conditional probabilities in (8) is needed. This can be obtained by drawing upon a recursive algorithm similar to that used for stochastic declustering of earthquakes using the RETAS model (Stindl and Chen, 2023). Let $\pi^{[m]}_{ij}$ denote the conditional probability that the $j$-th event is the parent to the $i$-th event

$$\pi^{[m]}_{ij} = \mathbb{P}(B_i = j|Y, \hat{\theta}^{[m]}). \quad (9)$$

Define $\omega^{[m]}_{ij}$ to be the probability that the $i$-th event is an immigrant event and the most recent immigrant prior to this event was the $j$-th event

$$\omega^{[m]}_{ij} = \mathbb{P}(B_i = 0, I(\tau_i) = j|Y, \hat{\theta}^{[m]}). \quad (10)$$
Define \( q_{ij}^{[m]} \) to be the most recent immigrant probability conditional on the entire observed data
\[
q_{ij}^{[m]} = \mathbb{P}(I(\tau_i) = j | Y, \hat{\theta}^{[m]}),
\]
with the convention that \( q_{n+1,j}^{[m]} = \mathbb{P}(I(T) = j | Y, \hat{\theta}^{[m]}) \) is the most recent immigrant probabilities at the censoring time.

The most recent immigrant probabilities conditional on the whole observed data are calculated using
\[
q_{ij} = \frac{f_{ij}}{\sum_{k=1}^{i-1} f_{ik} p_{ik}},
\]
where \( f_{ij} = p(\tau_{i:n+1} | Y_{1:i-1}, I(\tau_i) = j) \) and is given by
\[
f_{ij} = S_{ij} e^{-\{\Phi(\tau_i) - \Phi(\tau_{i-1})\}} \left[ f_{i+1,j} \cdot \phi(\tau_i) + f_{i+1,i} \cdot \mu(\tau_i - \tau_j) \right],
\]
which depends on \( f_{i+1,j} \) and hence a backward recursion is implemented with initial conditions for \( j = 1, \ldots, n, \)
\[
f_{n+1,j} = P(\tau_{n+1} > T | \tau_{1:n}, m_{1:n}, I(\tau_{n+1}) = j) = e^{-\{U(T-\tau_j) - U(\tau_n-\tau_j)\} - \Phi(T) + \Phi(\tau_n)},
\]
where \( U(t) = \int_0^t \mu(s) ds \) is the cumulative hazard rate function. The derivation of (12) and (13) are provided in Appendix A.

### 3.1.1 Expectation Step (E-step)

The recursive procedure to compute the most recent immigrant probabilities \( q_{ij} \) facilitates the calculation of the smoothed branching structure probabilities. The probabilities \( \omega_{ij} \) are calculated using
\[
\omega_{ij} = \frac{f_{i+1,i} \cdot \mu(\tau_i - \tau_j)}{f_{i+1,j} \cdot \phi(\tau_i) + f_{i+1,i} \cdot \mu(\tau_i - \tau_j)} \cdot q_{ij},
\]
for \( i = n, \ldots, 3 \) and \( j = 1, \ldots, i - 1. \) The smoothed parent probabilities are calculated using
\[
\pi_{ij} = \sum_{k=1}^{i-1} \frac{f_{i+1,k} \cdot \psi(\tau_i - \tau_j, m_j)}{f_{i+1,j} \cdot \phi(\tau_i) + f_{i+1,i} \cdot \mu(\tau_i - \tau_k)} \cdot q_{ik},
\]
for \( i = n, \ldots, 3 \) and \( j = 1, \ldots, i - 1. \) The derivation of (14) and (15) are provided in Appendix B.
3.1.2 Maximization Step (M-step)

Suppose that the immigrant hazard rate function $\mu(\cdot)$ and the excitation function $\psi(\cdot, \cdot)$ are parameterized by separate parameters $\theta_\mu$ and $\theta_\psi$ respectively. Then the optimization of the objective function in the M-step can simplify to two lower-order optimization problems. The M-step updates the estimated parameters for $\theta_\mu$ by maximizing the function

$$Q_\mu(\theta_\mu) = \log \mu(\tau_1) - \sum_{i=1}^{n-1} \sum_{j=1}^{\tau_i - \tau_j} \left[ \omega_{ij} \log \mu(\tau_i - \tau_j) - q_{ij} U(\tau_i - \tau_j) \right] - \sum_{j=1}^{n} q_{n+1,j} U(T - \tau_j),$$

(16)

where $U(t) = \int_0^t \mu(s) \, ds$ and the excitation function parameters $\theta_\psi$ are updated by maximizing the function

$$Q_\psi(\theta_\psi) = \sum_{i=2}^{n} \sum_{j=1}^{\pi_{i,j}^{\text{m}}} \log \psi(\tau_i - \tau_j, m_j) - \int_0^T \phi(s) \, ds.$$  

(17)

Here, the functions $\mu$ and $U$ depend on $\theta_\mu$ while $\psi$ and $\phi$ depend on $\theta_\psi$. The probabilities $q_{ij}^{\text{m}}$, $\omega_{ij}^{\text{m}}$ and $\pi_{ij}^{\text{m}}$ are calculated based on the parameter estimate $\hat{\theta}^{\text{m}}$. Solving these two optimization problems provides a new estimate $\hat{\theta}^{\text{m+1}}$. The E-step and M-step are then continually repeated until some convergence criterion is satisfied.

3.2 Whole data EM algorithm with a reduced set of latent variables

Wheatley et al. (2016) proposed an additional EM algorithm for the RHawkes process with a reduced set of latent variables. Similar to their EM2 algorithm, we reduce the latent variables by only labelling immigrants with $Z_i = 0$ and offspring with $Z_i = 1$ so that $Z$ is a dichotomized variable of $B$. For the reduced latent variables, the E-step replaces the calculation of the parent probabilities $\pi_{ij}$ with the probability that an event is an offspring $\pi_i := \mathbb{P}(Z_i = 1) = \mathbb{P}(B_i \neq 0 | Y)$ and this can be calculated by $\pi_i = 1 - \sum_{j=1}^{i-1} \omega_{ij}$. Hence, we only calculate $\omega_{ij}$ at each iteration while avoiding the calculation of $\pi_{ij}$ completely. The M-step maximizes the same objective function $Q_\mu$ in (16) to update the parameters $\theta_\mu$ but
instead of maximizing $Q_\psi$ in (17) to obtain updated parameter estimates of $\theta_\psi$, we now maximize

$$\tilde{Q}_\psi(\theta_\psi) = \sum_{i=2}^{n} \pi_i^{[m]} \log \phi(\tau_i) - \int_0^T \phi(s) ds.$$  

(18)

4 Accelerating the whole data EM algorithms

The EM algorithms discussed in Section 3 demand quadratic computational time and can be impractical for obtaining estimates of model parameters for sequences with a large number of events. This issue is also encountered in the likelihood evaluation for the RHawkes process. Stindl and Chen (2021) proposed a linear time algorithm to approximate the likelihood function of the RHawkes process. Their approximation to the likelihood was shown to be 20-40 times faster on sequences with around 10,000 events and the resultant estimates had similar finite sample performances compared to the MLE obtained based on the exact likelihood function.

Motivated by these promising reductions in computational time, we propose a similar approximation to our whole data EM algorithms. The approximation truncates the distribution of the most recent immigrant and the longevity of the self-excitation effects. For the former truncation, the potential candidates for the most recent immigrant are restricted to a relatively small number of recent events. For the latter truncation, the excitation effect of an event is restricted to a small number of future events rather than lasting indefinitely. We use a dynamic approach to determine the range of truncation at different event times based on the data, parameters, and preset approximation tolerance.

It is also possible to speed up the convergence of the exact EM1 algorithm by employing the pure quasi-Newton accelerator as proposed in Jamshidian and Jennrich (1997). This method only requires the EM step and can considerably reduce the number of iterations required until convergence. However, for our numerical experiments in Section 5 we did not observe a decrease in the number of iterations because the true $p$ parameter was close to the boundary and multiple adjustments to the step length are required to get the estimate back in the interior of the parameter space.
4.1 Truncation of the most recent immigrant distribution

The truncation of the most recent immigrant distribution assumes that only the most recent $R_i$ events are potential candidates to be the most recent immigrant at time $\tau_i$ rather than all previous $i-1$ events. This approximation assumes that events beyond the most recent $R_i$ events have zero probability of being the most recent immigrant. The different methods used to determine $R_i$ will result in different approximations to the conditional distribution of $I(\tau_i)$ given $Y$ defined by the conditional probabilities $q_{ij}$ in (12).

The strategy to determine $R_i$ is based on the data, parameters, and preset approximation tolerance. The value of each $R_i$ is determined such that the sum of the most recent $R_i$ immigrant probabilities exceeds $1 - \epsilon$ for a small positive $\epsilon \in (0, 1)$. If we let $c_{ij}$ denote the cumulative sum of the most recent $j$ events at time $\tau_i$ then $c_{ij} = \sum_{k=1}^{j} p_{i,i-k}$, and $R_i = \min \{ j \geq 1 ; c_{ij} \geq 1 - \epsilon \}$. Here we use $p_{ij}$ rather than $q_{ij}$ since $p_{ij}$ are calculated prior to the calculation of $q_{ij}$ and rather than calculate all $p_{ij}$ we can truncate here first to avoid unnecessary calculations of $p_{ij}$ for $j = 1, \ldots, i - R_i - 1$. The probabilities $p_{ij}$ and $q_{ij}$ must sum to one, and therefore, the probabilities $p_{ij}$ and $q_{ij}$ for $j = i - R_i, \ldots, i - 1$ must be renormalized with the remaining $p_{ij}$ and $q_{ij}$ for $j = 1, \ldots, i - R_i - 1$ all set equal to zero.

4.2 Truncation of the longevity of self-excitation effects

The evaluation of $\omega_{ij}$ in (14) and $\pi_{ij}$ in (15) also depend on the calculation of the self-excitation effects $\phi(\tau_i) = \sum_{j=1}^{i-1} \psi(\tau_i - \tau_j, m_j)$. Furthermore, in the calculation of $Q_\psi$ in (17), for each event, we must evaluate the excitation function $\psi(\cdot, \cdot)$ at the time lags for all past events. Therefore, an approximation that restricts the longevity of the self-excitation effects will be employed to guarantee that the computation time stays bounded for each $i$. Similar truncations have also been used by Halpin (2013) and Stindl and Chen (2021).

For this purpose, we assume that the $i$-th event only contributes to the self-excitation effects for the next $F_i$ events. Hence, we assume that $\psi(\tau_i - \tau_j, m_j) = 0$ for all $j < i - F_i - 1$. This implies the following approximation for the self-excitation effects

$$\phi(\tau_i) \approx \sum_{j=i-F_i}^{i-1} \psi(\tau_i - \tau_j, m_j).$$  (19)
We choose \( F_i \) as the smallest integer \( j \leq i - 1 \) such that \( G(\tau_i - \tau_{i-j}) \geq (1 - \delta)G(\infty) \) for a small positive \( \delta > 0 \) where \( G(t) = \int_0^t g(s)ds \). That is, assuming that the temporal response function \( g \) is a density

\[
F_i = \min \{ j \leq i - 1; G(\tau_i - \tau_{i-j}) \geq 1 - \delta \} = \min \{ j \leq i - 1; \tau_i - \tau_{i-j} \geq G^{-1}(1 - \delta) \},
\]

where \( G^{-1}(\cdot) \) denotes the quantile of the temporal response function \( g(\cdot) \), and when the set is empty we define \( F_i = i - 1 \).

### 4.3 E-step of the accelerated EM algorithm

We define \( \tilde{q}_{ij} \), \( \tilde{\omega}_{ij} \), \( \tilde{\pi}_{ij} \) to be the approximate probabilities using the truncations discussed in Subsection 4.1 and 4.2 for a given level of approximation tolerance \( \epsilon \) and \( \delta \). With these choices of approximation, the objective function \( Q_\mu \) in (16) is approximate by

\[
Q_\mu(\theta) \approx \log \mu(\tau_1) - U(\tau_1)
\]

\[
+ \sum_{i=2}^n \sum_{j=i-R_i}^{i-1} \tilde{\omega}_{ij}^{[m]} \log \mu(\tau_i - \tau_j) - \tilde{q}_{ij}^{[m]} U(\tau_i - \tau_j) - \sum_{j=n-R_{n+1}+1}^n \tilde{q}_{n+1,j}^{[m]} U(T - \tau_j),
\]

(20)

and the objective function \( Q_\psi \) in (17) is approximated by

\[
Q_\psi \approx \sum_{i=2}^n \sum_{j=i-F_i}^{i-1} \tilde{\pi}_{ij}^{[m]} \log \psi(\tau_i - \tau_j, m_j) - \int_0^T \phi(s)ds.
\]

(21)

A similar approximation holds for \( \tilde{Q}_\psi \) in (18) in which \( \phi(\tau_i) \) is replaced with (19).

### 5 Numerical experiments

This section performs numerical experiments to show that the whole data EM algorithms from Section 3 and the accelerated versions from Section 4 are suitable strategies for the estimation of the RETAS model. A comparison with the direct likelihood maximization method (Chen and Stindl, 2018) is also provided. We show that the direct maximization method can fail in some situations, such as when using unintelligent starting values for the optimizer, but the EM estimation method has better numerical stability. This section
concludes with a simulation study to show that the accelerated EM algorithms can obtain the MLE much faster without compromising on the accuracy of the solution.

The simulation model is motivated by the seismological context, and we consider a gamma main-shock renewal process with hazard function

$$
\mu(t) = \frac{1}{\Gamma(t/\beta, \kappa)} t^{\kappa-1} e^{-t/\beta}, \quad t \geq 0,
$$

where $\kappa > 0$ and $\beta > 0$ are the shape and scale parameters respectively, and $\Gamma(x, k) = \int_x^\infty s^{k-1} e^{-s} ds$ is the upper incomplete gamma function. The parameters are set to $\kappa = 0.5$ and $\beta = 200$ which implies that the average waiting time between immigrants (or main-shocks in the seismology context) is 100 days. The form of the temporal response function is motivated by Omori’s Law (Omori, 1894; Utsu, 1961) with density function

$$
g(t) = \frac{p-1}{c} \left(1 + \frac{t}{c}\right)^{-p}, \quad t \geq 0,
$$

where $p > 1$ is a shape parameter indicating the decay rate of aftershocks and $c > 0$ is a scale parameter. The parameters are set to $p = 1.1$ and $c = 0.01$ which are typical values for an earthquake catalog. The boost function is given by $k(m) = Ae^{\alpha(m-m_0)}$ where $A > 0$ and $\alpha > 0$ are parameters and $m_0$ is the threshold magnitude of the catalog. The earthquake magnitudes are modelled using a shifted-exponential distribution as motivated by the Gutenberg-Richer law (Gutenberg and Richter, 1944) with density $J(m) = \gamma e^{-\gamma(m-m_0)}$ where $\gamma > 0$ is a rate parameter. With these forms for the boost function and magnitude density, the productivity, which is the average number of aftershocks for an earthquake is calculated as $E[k(M)] = \frac{A\gamma}{\gamma-\alpha}$. For the simulation model we set $A = 0.25$, $\alpha = 1$ and $\gamma = 2$ which implies that the productivity is 0.5 and hence on average half the earthquake are main-shocks and half aftershocks.

We simulate $N = 1,000$ catalogs from the simulation model with a censoring time $T = 25,000$. The mean total number of earthquakes for the simulated catalogs is 472.1. For each catalog, we estimate the RETAS model parameters by directly maximizing the log-likelihood (MLE), the whole data EM algorithm with the full branching structure $B_{1:n}$ used as latent variables (EM1), and with the reduced set $Z_{1:n}$ used as latent variables (EM2). Each method begins its search from three different starting values from different parts of the
parameter space given by $\theta_1 = (0.5, 100, 1.2, 0.1, 0.25, 1)$, $\theta_2 = (0.25, 50, 1.1, 0.05, 0.125, 0.5)$ and $\theta_3 = (1, 400, 1.4, 0.2, 0.5, 2)$ where $\theta_1$ contains the true parameters. For the MLE, the BFGS method was implemented using the `optim` function in R to compute the minimizer of the negative log-likelihood function with `reltol` set to $1 \times 10^{-8}$. The EM algorithms consist of cycling between the E-step and M-step until the distance between each component of the parameter vector from consecutive iterations are all less than $1 \times 10^{-3}$. The M-step also calls the `optim` function in R using the Nelder-Mead method with `reltol` set to $1 \times 10^{-8}$. The Nelder-Mead method was used in the M-step because it was numerically more stable.

Table 1 and Table 2 present the results for the direct ML and whole data EM estimation methods respectively. Each table contains the bias (Bias), the standard deviation of the parameter estimates (SE), the average of the standard error estimates obtained by inverting the observed information matrix ($\widehat{SE}$), the root mean squared error (RMSE), the average running time (RT) on Intel Xeon X5675 processors (12M cache, 3.06 GHz, 6.4GT/S QPI) and the number of iterations required until convergence (Iter). An iteration is defined as the number of log-likelihood evaluations for the direct MLE or the number of E- and M-step cycles needed for the whole data EM algorithm. The standard errors for the estimates are computed based on an approximated Hessian obtained using numerical differentiation on the exact log-likelihood function. We use the `hessian` function in the R package `numDeriv` with the default Richardson extrapolation method to obtain the approximate Hessian which can then be inverted to obtain the standard errors.

Veen and Schoenberg (2008) illustrated the difficulties that can arise when directly maximizing the log-likelihood in (2) using numerical optimization routines such as the Nelder-Mead method (Nelder and Mead, 1965) for the ETAS model. The authors confirm that the optimization routine might perform poorly when the ETAS log-likelihood surface is extremely flat or has multiple local maxima, unless intelligent starting values are chosen, or clever supervision is imposed during the optimization routine. The same issues are also evident for the direct maximization of the log-likelihood of the RETAS model. The direct MLE results show poor performance when the optimization routine does not have intelligent starting values as seen in Table 1. For instance, both the bias and the root mean
Table 1: Direct MLE for the RETAS model based on 1,000 simulated catalogs with three different starting values used for the BFGS optimization routine given by $\theta_1 = (0.5, 100, 1.2, 0.1, 0.25, 1)$, $\theta_2 = (0.25, 50, 1.1, 0.05, 0.125, 0.5)$ and $\theta_3 = (1, 400, 1.4, 0.2, 0.5, 2)$ where $\theta_1$ contains the true parameters.

<table>
<thead>
<tr>
<th></th>
<th>$\kappa$</th>
<th>$\beta$</th>
<th>$p$</th>
<th>$c$</th>
<th>$A$</th>
<th>$\alpha$</th>
<th>RT</th>
<th>Iter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Truth</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bias</td>
<td>0.0143</td>
<td>0.18</td>
<td>0.0223</td>
<td>0.00232</td>
<td>0.0087</td>
<td>0.0022</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SE</td>
<td>0.0769</td>
<td>26.89</td>
<td>0.0853</td>
<td>0.00553</td>
<td>0.0762</td>
<td>0.1373</td>
<td>148</td>
<td>29</td>
</tr>
<tr>
<td>$\bar{SE}$</td>
<td>0.0814</td>
<td>33.52</td>
<td>0.0868</td>
<td>0.00553</td>
<td>0.0719</td>
<td>0.1327</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMSE</td>
<td>0.0782</td>
<td>26.88</td>
<td>0.0881</td>
<td>0.00600</td>
<td>0.0766</td>
<td>0.1373</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

|       | $\theta_1$ |          |      |      |       |           |     |      |
| Bias  | 0.0459     | -6.58    | 0.0268 | 0.00253 | 0.0163 | -0.0013  |     |      |
| SE    | 0.2442     | 31.29    | 0.0842 | 0.00545 | 0.1683 | 0.1390   | 285 | 60   |
| $\bar{SE}$ | 0.0835 | 30.94    | 0.0858 | 0.00551 | 0.0699 | 0.1318   |     |      |
| RMSE  | 0.2483     | 31.96    | 0.0883 | 0.00601 | 0.1690 | 0.1389   |     |      |

|       | $\theta_3$ |          |      |      |       |           |     |      |
| Bias  | -0.0690    | 157.07   | -0.1038 | 0.00182 | -0.2246 | -0.0774  |     |      |
| SE    | 0.0781     | 77.15    | 0.1010 | 0.09807 | 1.5369 | 0.5880   | 297 | 59   |
| $\bar{SE}$ | 0.0709 | 39.56    | 0.0959 | 0.00655 | 0.0631 | 0.1310   |     |      |
| RMSE  | 0.1041     | 174.97   | 0.1448 | 0.09803 | 1.5523 | 0.5928   |     |      |

Squared error can be incredibly large when starting from a value away from the global optimizer. Indeed, the empirical standard errors consistently over-estimate the standard errors and this holds even when the optimization routine is initialised at the true value. Furthermore, the direct maximization of the log-likelihood often fails to converge, with failure to converge observed on 200 of the simulated datsets when starting the iterations from an initial value away from the truth.

However, the EM algorithms are numerically more stable compared to the direct MLE method as seen in Table 2. The EM estimation results perform consistently well no matter the starting value. There are small biases compared to the standard errors and the empirical standard errors are more consistent with the standard errors compared to the direct
Table 2: EM estimation for the RETAS model based on 1,000 simulated catalogs with three different starting values $\theta_1 = (0.5, 100, 1.2, 0.1, 0.25, 1)$, $\theta_2 = (0.25, 50, 1.1, 0.05, 0.125, 0.5)$ and $\theta_3 = (1, 400, 1.4, 0.2, 0.5, 2)$ where $\theta_1$ contains the true parameters.

<table>
<thead>
<tr>
<th></th>
<th>$\kappa$</th>
<th>$\beta$</th>
<th>$p$</th>
<th>$c$</th>
<th>$A$</th>
<th>$\alpha$</th>
<th>RT</th>
<th>Iter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Truth</td>
<td></td>
<td>0.5</td>
<td>200</td>
<td>1.1</td>
<td>0.01</td>
<td>0.25</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Bias</td>
<td>0.0138</td>
<td>1.05</td>
<td>0.0168</td>
<td>0.00173</td>
<td>0.0090</td>
<td>0.0046</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SE</td>
<td>0.0825</td>
<td>33.41</td>
<td>0.0799</td>
<td>0.00529</td>
<td>0.0808</td>
<td>0.1324</td>
<td>440</td>
<td>35</td>
</tr>
<tr>
<td>$\hat{SE}$</td>
<td>0.0813</td>
<td>33.04</td>
<td>0.0823</td>
<td>0.00516</td>
<td>0.0696</td>
<td>0.1290</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMSE</td>
<td>0.0836</td>
<td>33.41</td>
<td>0.0816</td>
<td>0.00557</td>
<td>0.0813</td>
<td>0.1324</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EM1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>0.0111</td>
<td>0.51</td>
<td>0.0232</td>
<td>0.00199</td>
<td>0.0051</td>
<td>0.0090</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SE</td>
<td>0.0828</td>
<td>33.38</td>
<td>0.0851</td>
<td>0.00555</td>
<td>0.0823</td>
<td>0.1338</td>
<td>714</td>
<td>55</td>
</tr>
<tr>
<td>$\hat{SE}$</td>
<td>0.0801</td>
<td>32.59</td>
<td>0.0849</td>
<td>0.00534</td>
<td>0.0678</td>
<td>0.1291</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMSE</td>
<td>0.0835</td>
<td>33.37</td>
<td>0.0882</td>
<td>0.00589</td>
<td>0.0824</td>
<td>0.1340</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EM2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>0.0164</td>
<td>1.91</td>
<td>0.0138</td>
<td>0.00163</td>
<td>0.0153</td>
<td>0.0020</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SE</td>
<td>0.0849</td>
<td>34.14</td>
<td>0.0817</td>
<td>0.00529</td>
<td>0.1070</td>
<td>0.1331</td>
<td>845</td>
<td>67</td>
</tr>
<tr>
<td>$\hat{SE}$</td>
<td>0.0825</td>
<td>33.45</td>
<td>0.0822</td>
<td>0.00514</td>
<td>0.0712</td>
<td>0.1289</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMSE</td>
<td>0.0865</td>
<td>34.18</td>
<td>0.0828</td>
<td>0.00553</td>
<td>0.1080</td>
<td>0.1331</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EM2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\theta_3$</td>
<td>0.0138</td>
<td>1.07</td>
<td>0.0169</td>
<td>0.00174</td>
<td>0.0092</td>
<td>0.0047</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SE</td>
<td>0.0825</td>
<td>33.45</td>
<td>0.0802</td>
<td>0.00532</td>
<td>0.0835</td>
<td>0.1325</td>
<td>450</td>
<td>34</td>
</tr>
<tr>
<td>$\hat{SE}$</td>
<td>0.0813</td>
<td>33.05</td>
<td>0.0824</td>
<td>0.00517</td>
<td>0.0695</td>
<td>0.1290</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMSE</td>
<td>0.0836</td>
<td>33.45</td>
<td>0.0820</td>
<td>0.00559</td>
<td>0.0839</td>
<td>0.1325</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EM2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\theta_3$</td>
<td>0.0113</td>
<td>0.57</td>
<td>0.0229</td>
<td>0.00198</td>
<td>0.0056</td>
<td>0.0087</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SE</td>
<td>0.0828</td>
<td>33.41</td>
<td>0.0852</td>
<td>0.00555</td>
<td>0.0850</td>
<td>0.1338</td>
<td>752</td>
<td>55</td>
</tr>
<tr>
<td>$\hat{SE}$</td>
<td>0.0802</td>
<td>32.62</td>
<td>0.0848</td>
<td>0.00533</td>
<td>0.0680</td>
<td>0.1291</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMSE</td>
<td>0.0836</td>
<td>33.40</td>
<td>0.0882</td>
<td>0.00589</td>
<td>0.0851</td>
<td>0.1340</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EM2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\theta_3$</td>
<td>0.0163</td>
<td>1.84</td>
<td>0.0142</td>
<td>0.00164</td>
<td>0.0193</td>
<td>0.0023</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SE</td>
<td>0.0849</td>
<td>34.12</td>
<td>0.0817</td>
<td>0.00530</td>
<td>0.2239</td>
<td>0.1332</td>
<td>885</td>
<td>67</td>
</tr>
<tr>
<td>$\hat{SE}$</td>
<td>0.0825</td>
<td>33.43</td>
<td>0.0823</td>
<td>0.00515</td>
<td>0.0713</td>
<td>0.1289</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMSE</td>
<td>0.0864</td>
<td>34.15</td>
<td>0.0829</td>
<td>0.00555</td>
<td>0.2246</td>
<td>0.1332</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
MLE method. We noticed that the EM2 algorithm is on average slower than the EM1 algorithm. We suggest that even though the EM2 algorithm speeds up the calculation of the probabilities in the E-step, it often requires more calculations of the objective function $Q_\psi$ in (17) in the M-step when performing the optimization using Nelder-Mead. Since the criterion function in the M-step requires quadratic computational time it can lead to longer running times for estimation compared to the EM1 algorithm. The running times for the direct MLE are on average faster than that of both the EM algorithms. However, by applying the accelerated EM algorithms discussed in Section 4 we can reduce the running times but preserve the accurate estimates of the EM algorithms.

Since the running time of the EM1 algorithm is faster than the EM2 algorithm, we only apply the accelerated EM algorithm with the entire branching structure $B_{1:n}$ used as latent variables. The approximation parameters $\epsilon$ and $\delta$ are set equal and take values $0.01, 0.001, 1 \times 10^{-4}, 1 \times 10^{-10}$. Table 3 presents the estimation results for the accelerated EM algorithm with the same convergence criterion as before. For small values of $\epsilon$ and $\delta$, the estimation results perform similarly in terms of bias and RMSE but with significantly reduced running times. The average running time for the exact EM1 algorithm was 440 seconds while for the accelerated EM1 algorithm with $\epsilon = \delta = 1 \times 10^{-4}$ the average running time was only 139 seconds which is slightly faster than the direct MLE at 148 seconds.

The EM algorithms can avoid some of the issues that direct likelihood maximization suffer, and the accelerated EM algorithms are able to provide accurate estimates while significantly reducing the running time. Hence, we suggest the following estimation procedure for real data. First, fit the RETAS model using the accelerated EM algorithm with a sequence of decreasing values of $\delta$ and $\epsilon$ in which the previous iterations estimate is used as the starting value for the next iteration. Then implement the exact EM1 algorithm to obtain a final estimate of the parameters. Use numerical differentiation to obtain an approximate Hessian based on the exact log-likelihood function which can then be inverted to obtain standard errors. We opt for numerical differentiation of the log-likelihood function rather than methods that differentiate the EM operator due to its ease of application and good performance (see e.g. Jamshidian and Jennrich (2000)).
Table 3: Accelerated EM1 estimation for the RETAS model based on 1,000 simulated catalogs with tuning parameters $\epsilon$ and $\delta$ taking the same value from the set \{0.01, 0.001, $1 \times 10^{-4}$, $1 \times 10^{-10}$\} with the starting value for the EM algorithm set at the true parameter $\theta_1 = (0.5, 100, 1.2, 0.1, 0.25, 1)$.

<table>
<thead>
<tr>
<th>$\epsilon = \delta$</th>
<th>Truth</th>
<th>$\kappa$</th>
<th>$\beta$</th>
<th>$p$</th>
<th>$c$</th>
<th>$A$</th>
<th>$\alpha$</th>
<th>RT</th>
<th>Iter</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>Bias</td>
<td>0.0126</td>
<td>1.41</td>
<td>0.0360</td>
<td>0.002883</td>
<td>0.0071</td>
<td>0.0073</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.0843</td>
<td>33.67</td>
<td>0.1406</td>
<td>0.008945</td>
<td>0.0829</td>
<td>0.1351</td>
<td>126</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>$\widehat{SE}$</td>
<td>0.0824</td>
<td>33.16</td>
<td>0.1042</td>
<td>0.006742</td>
<td>0.0688</td>
<td>0.1293</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.0852</td>
<td>33.68</td>
<td>0.1450</td>
<td>0.009394</td>
<td>0.0831</td>
<td>0.1352</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.001</td>
<td>Bias</td>
<td>0.0145</td>
<td>1.60</td>
<td>0.0157</td>
<td>0.001680</td>
<td>0.0107</td>
<td>0.0031</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.0836</td>
<td>33.71</td>
<td>0.0834</td>
<td>0.005379</td>
<td>0.0814</td>
<td>0.1326</td>
<td>132</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>$\widehat{SE}$</td>
<td>0.0820</td>
<td>33.23</td>
<td>0.0817</td>
<td>0.005147</td>
<td>0.0697</td>
<td>0.1286</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.0848</td>
<td>33.73</td>
<td>0.0848</td>
<td>0.005633</td>
<td>0.0821</td>
<td>0.1326</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$1 \times 10^{-4}$</td>
<td>Bias</td>
<td>0.0145</td>
<td>1.61</td>
<td>0.0153</td>
<td>0.001664</td>
<td>0.0109</td>
<td>0.0029</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.0837</td>
<td>33.70</td>
<td>0.0820</td>
<td>0.005338</td>
<td>0.0817</td>
<td>0.1325</td>
<td>139</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>$\widehat{SE}$</td>
<td>0.0820</td>
<td>33.24</td>
<td>0.0813</td>
<td>0.005124</td>
<td>0.0699</td>
<td>0.1287</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.0848</td>
<td>33.73</td>
<td>0.0834</td>
<td>0.005589</td>
<td>0.0824</td>
<td>0.1325</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$1 \times 10^{-10}$</td>
<td>Bias</td>
<td>0.0145</td>
<td>1.61</td>
<td>0.0152</td>
<td>0.001659</td>
<td>0.0109</td>
<td>0.0029</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.0836</td>
<td>33.72</td>
<td>0.0818</td>
<td>0.005321</td>
<td>0.0813</td>
<td>0.1325</td>
<td>178</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>$\widehat{SE}$</td>
<td>0.0820</td>
<td>33.24</td>
<td>0.0813</td>
<td>0.005121</td>
<td>0.0699</td>
<td>0.1287</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.0848</td>
<td>33.74</td>
<td>0.0831</td>
<td>0.005572</td>
<td>0.0820</td>
<td>0.1324</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

6 Analysis of an Italian earthquake catalog

Italy is a seismically active region with the L’Auila region in central Italy known for its persistent and severe earthquakes (Molkenthin et al., 2022). This section analyses an Italian earthquake catalog that contains earthquakes from 2001-02-04 to 2020-03-28 in a rectangular region from 35°-49° latitude and 5°-20° longitude. The catalog was obtained
from the National Institute of Geophysics and Volcanology (http://terremoti.ingv.it/). The threshold magnitude of the catalog is $m_0 = 3$ resulting in 4,100 earthquakes in the catalog. Figure 1 plots the earthquakes with the area of the circle indicating the relative size of the magnitude. The right plots contain the magnitude against occurrence time and cumulative count of earthquakes over the period.

![Figure 1: Left plot: Location of earthquakes in and around Italy from 2001-02-04 to 2020-03-28 with the area of the circle representing the relative size of the magnitude. Right plot: Cumulative number of earthquakes and dot-plot of magnitude against time.](image)

Our proposed EM algorithm is used to estimate the RETAS model parameters for the Italian earthquake catalog. We fit RETAS models with Poisson, gamma renewal and Weibull renewal main-shock arrival processes where the gamma hazard rate function is given in (22), and the Weibull hazard rate function is given by

$$
\mu(t) = \kappa \frac{t^{\kappa-1}}{\beta^{\kappa}}, \quad t \geq 0.
$$

(24)
We use a power-law temporal response function $g(t) = \frac{p-1}{c} (1 + \frac{t}{c})^{-p}$ and exponential boost function $k(m) = Ae^{\alpha(m-m_0)}$. With a Poisson arrival process, the model reduces to the classical ETAS model. We use the Akaike Information Criterion (AIC) to compare the three models. To fit the models, we sequentially apply the accelerated EM1 algorithm with progressively smaller tolerance parameters $\epsilon$ and $\delta$ which are set equal and take values $1 \times 10^{-2}$, $1 \times 10^{-3}$, $1 \times 10^{-4}$. In each iteration, the converged parameter value from the previous iteration is used as the starting value for the current iteration. In the final iteration, the exact EM1 algorithm is used to obtain the final estimate of the RETAS model parameters. The parameter estimates, their standard errors, and the AIC value are presented in Table 4. The standard errors were obtained by inverting the Hessian matrix computed at the EM estimate using the exact observed data log-likelihood in an additional step once the final estimate was obtained.

<table>
<thead>
<tr>
<th></th>
<th>$\kappa$</th>
<th>$\beta$</th>
<th>$p$</th>
<th>$c$</th>
<th>$A$</th>
<th>$\alpha$</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poisson</td>
<td>1.00</td>
<td>20.31</td>
<td>1.152</td>
<td>0.0189</td>
<td>0.625</td>
<td>1.143</td>
<td>267.01</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>(4.133)</td>
<td>(0.0151)</td>
<td>(0.00266)</td>
<td>(0.0409)</td>
<td>(0.0507)</td>
<td></td>
</tr>
<tr>
<td>Gamma</td>
<td>0.498</td>
<td>20.58</td>
<td>1.181</td>
<td>0.0214</td>
<td>0.486</td>
<td>1.276</td>
<td>178.71</td>
</tr>
<tr>
<td></td>
<td>(0.0345)</td>
<td>(2.823)</td>
<td>(0.0159)</td>
<td>(0.00306)</td>
<td>(0.0329)</td>
<td>(0.0551)</td>
<td></td>
</tr>
<tr>
<td>Weibull</td>
<td>0.646</td>
<td>3.96</td>
<td>1.229</td>
<td>0.0233</td>
<td>0.351</td>
<td>1.424</td>
<td>37.54</td>
</tr>
<tr>
<td></td>
<td>(0.0180)</td>
<td>(0.330)</td>
<td>(0.0189)</td>
<td>(0.00330)</td>
<td>(0.0270)</td>
<td>(0.0600)</td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Whole data EM algorithm estimation results for the Italian earthquake catalog with Poisson, gamma renewal, and Weibull renewal main-shock arrival processes.

The parameter pairs $\{\kappa, \beta\}$, $\{p, c\}$ and $\{A, \alpha\}$ each describe a different feature of the earthquake sequence. The parameters $\kappa$ and $\beta$ describe the waiting time distribution between main-shocks. Since the gamma renewal and Weibull renewal processes are Poisson processes when their respective shape parameter $\kappa$ is set equal to one it suggests that the RETAS models show a significant departure from the Poisson assumption of the ETAS model with an approximate 95% confidence interval for $\kappa$ given by $(0.430, 0.565)$ and $(0.611, 0.681)$ for the gamma and Weibull models respectively. The mean waiting time be-
between main-shocks for the three models are 20.31, 10.24, and 5.45 for the Poisson, gamma, and Weibull models, respectively. Figure 2 plots the estimated hazard rate function $\hat{\mu}(t)$ for the three models. The Poisson and the gamma hazard functions only differ in their short-term behaviour with the longer-term hazard rate approaching similar levels. However, the estimated Weibull hazard function continues to decrease over the longer term and there is a noticeable difference in hazard rate for shorter waiting times. Both the gamma and Weibull hazard functions are monotonically decreasing from infinity with the Weibull hazard decreasing at a slower rate for smaller values of $t$.

Figure 2: Estimated hazard function $\hat{\mu}(t)$ for the RETAS models with Poisson, gamma renewal, and Weibull renewal main-shock processes for the Italian earthquake catalog.

The parameters $p$ and $c$ of the temporal response function $g$ describe the temporal
distribution of aftershocks relative to their trigger earthquakes. The estimated parameters \( \hat{\rho} \) and \( \hat{\alpha} \) for the three models are similar, although the Weibull model has aftershock effects that occur more rapidly after the triggering earthquake since it has the largest estimated \( \hat{\rho} \).

The estimate of the magnitude density parameter is \( \hat{\gamma} = 2.85 \) and hence the productivity measure \( \hat{\lambda}(\hat{\gamma}/(\hat{\gamma} - \hat{\alpha})) \) is 1.043, 0.880, and 0.701 for the Poisson, gamma renewal and Weibull renewal models respectively. This is consistent with the mean waiting time between main-shocks for the three models since for example the Weibull model has the smallest waiting time between main-shocks and hence has the smallest level of self-excitation effects.

The RETAS model with a Weibull renewal main-shock process is the preferred choice with a markedly smaller AIC value compared to the other two models. Indeed the Poisson main-shock process is not even a stationary model with a productivity that exceeds one. Hence, we conclude that the RETAS model outperforms the classical ETAS model and the whole data EM estimation method is able to avoid some of the convergence issues associated with direct maximization of the log-likelihood.

References


**A Most recent immigrant probabilities conditional on the whole observed data**

Let $Y_{1:i-1} = \{\tau_{1:i-1}, m_{1:i-1}\}$ be a subset of the observed data $Y = \{\tau_{1:n}, m_{1:n}, \tau_{n+1} > T\}$ and define $\tau_{i:n+1}$ to be $\{\tau_i, \ldots, \tau_n, \tau_{n+1} > T\}$. To calculate $q_{ij}$, we smooth the probabilities $p_{ij}$ that facilitated likelihood evaluation in (4), by applying a suitable backward recursion. By applying Bayes’ theorem

$$q_{ij} = \frac{p(\tau_{i:n+1}|Y_{1:i-1}, I(\tau_i) = j)}{\sum_{k=1}^{i-1} p(\tau_{k:n+1}|Y_{1:i-1}, I(\tau_k) = k) \cdot p_{ik}} \cdot p_{ij}. \tag{25}$$

Define $f_{ij}$ to be the joint density of $\tau_{i:n+1}$ conditional on $Y_{1:i-1}$ and $I(\tau_i) = j$. Then it holds that

$$f_{ij} = p(\tau_{i:n+1}|Y_{1:i-1}, I(\tau_i) = j) = p(\tau_i|Y_{1:i-1}, I(\tau_i) = j) \cdot p(\tau_{i+1:n+1}|Y_{1:i}, I(\tau_i) = j). \tag{26}$$
The conditional density of the event time $\tau_i$ conditional on the previous times, marks and index of the most recent immigrant $I(\tau_i) = j$ is given by

$$p(\tau_i|Y_{1:i-1}, I(\tau_i) = j) = (\mu(\tau_i - \tau_j) + \phi(\tau_i)) e^{-\{U(\tau_i - \tau_j) - U(\tau_{i-1} - \tau_j)\} - \{\Phi(\tau_i) - \Phi(\tau_{i-1})\}}.$$  

(27)

To calculate $p(\tau_{i+1:n+1}|Y_{1:i}, I(\tau_i) = j)$ in (26) we condition on the most recent immigrant at time $\tau_{i+1}$ denoted by $I(\tau_{i+1}) = k$ to obtain the following

$$\sum_{k=1}^{n} p(\tau_{i+1:n+1}|Y_{1:i}, I(\tau_i) = j, I(\tau_{i+1}) = k) \cdot \mathbb{P}(I(\tau_{i+1}) = k|Y_{1:i}, I(\tau_i) = j).$$  

(28)

Since the conditioning information set contains both $I(\tau_i) = j$ and $I(\tau_{i+1}) = k$, it implies that the only indexes contributing to the summation in (28) are when $k = j$ or $k = i$. Also, since we condition on the most recent immigrant at time $\tau_{i+1}$ we can remove $I(\tau_i) = j$ from the conditioning information set since only the most recent renewal time provides information about the future evolution of the process given the history. Hence by the definition of $f_{ij}$, the summation in (28) reduces to

$$\sum_{k \in \{j, i\}} f_{i+1,k} \cdot \mathbb{P}(I(\tau_{i+1}) = k|Y_{1:i}, I(\tau_i) = j).$$  

(29)

The most recent immigrant index indicator function at time $\tau_{i+1}$, that is $I(\tau_{i+1})$, is equal to $I(\tau_i)$ if $B_i \neq 0$ or $i$ if $B_i = 0$. Therefore

$$\sum_{k \in \{j, i\}} f_{i+1,k} \cdot \mathbb{P}(I(\tau_{i+1}) = k|Y_{1:i}, I(\tau_i) = j) = [f_{i+1,j} \cdot \mathbb{P}(I(\tau_{i+1}) = j|Y_{1:i}, I(\tau_i) = j) + f_{i+1,i} \cdot \mathbb{P}(I(\tau_{i+1}) = i|Y_{1:i}, I(\tau_i) = j)]$$

$$= [f_{i+1,j} \cdot \mathbb{P}(B_i \neq 0|Y_{1:i}, I(\tau_i) = j) + f_{i+1,i} \cdot \mathbb{P}(B_i = 0|Y_{1:i}, I(\tau_i) = j)]$$

$$= \left[ f_{i+1,j} \cdot \frac{\phi(\tau_i)}{\mu(\tau_i - \tau_j) + \phi(\tau_i)} + f_{i+1,i} \cdot \frac{\mu(\tau_i - \tau_j)}{\mu(\tau_i - \tau_j) + \phi(\tau_i)} \right].$$  

(30)

Hence, by substituting (27) and (30) into (26) we obtain the following backward recursion to calculate $f_{ij}$

$$f_{ij} = S_{ij} e^{-\{\Phi(\tau_i) - \Phi(\tau_{i-1})\}} \left[ f_{i+1,j} \cdot \phi(\tau_i) + f_{i+1,i} \cdot \mu(\tau_i - \tau_j) \right],$$  

(31)

with initial conditions

$$f_{n+1,j} = P(\tau_{n+1} > T|\tau_{1:n}, m_{1:n}, I(\tau_{n+1}) = j) = e^{-\{U(T - \tau_j) - U(\tau_n - \tau_j)\} - \{\Phi(T) - \Phi(\tau_n)\}},$$

for $j = 1, \ldots, n$. 

28
B Branching structure probabilities conditioned on the whole observed data

The recursive procedure to compute the most recent immigrant probabilities $q_{ij}$ in Appendix A facilitates the calculation of the branching structure probabilities needed in the E-step. For the probabilities $\omega_{ij}$ we have

$$\omega_{ij} := \Pr(B_i = 0, I(\tau_i) = j | Y)$$

$$= \Pr(B_i = 0 | Y, I(\tau_i) = j) \cdot \Pr(I(\tau_i) = j | Y)$$

$$= \frac{p(\tau_{i+1,n+1}, \tau_{n+1} > T | Y_{1:i}, B_i = 0, I(\tau_i) = j) \cdot \Pr(B_i = 0 | Y_{1:i}, I(\tau_i) = j)}{p(\tau_{i+1:n+1} | Y_{1:i}, I(\tau_i) = j)} \cdot q_{ij}$$

$$= \frac{p(\tau_{i+1,n+1} | Y_{1:i}, I(\tau_i+1) = i) \cdot \Pr(B_i = 0 | Y_{1:i}, I(\tau_i) = j)}{p(\tau_{i+1:n+1} | Y_{1:i}, I(\tau_i) = j)} \cdot q_{ij}.$$  \hspace{1cm} (32)

Hence, by the definition of $f_{i+1,j}$ and the expression provided in Appendix A, to compute the denominator in (30), we obtain the following expression

$$\omega_{ij} = \frac{f_{i+1,i} \cdot \mu(\tau_i - \tau_j)}{f_{i+1,j} \cdot \phi(\tau_i) + f_{i+1,i} \cdot \mu(\tau_i - \tau_j)} \cdot q_{ij}. \hspace{1cm} (33)$$

The parent probabilities conditional on the whole data $Y$ required in the E-step can be calculated as follows

$$\pi_{ij} = \Pr(B_i = j | Y)$$

$$= \sum_{k=1}^{i-1} \Pr(B_i = j | Y, I(\tau_i) = k) \cdot q_{ik}$$

$$= \sum_{k=1}^{i-1} \frac{p(\tau_{i+1,n+1} | Y_{1:i}, I(\tau_i) = k, B_i = j) \cdot \Pr(B_i = j | Y_{1:i}, I(\tau_i) = k)}{p(\tau_{i+1:n+1} | Y_{1:i}, I(\tau_i) = k)} \cdot q_{ik}.$$ 

Again, the denominator was computed in (30), and using the definition of $f_{ij}$ we obtain the following

$$\pi_{ij} = \sum_{k=1}^{i-1} \frac{f_{i+1,k} \cdot \psi(\tau_i - \tau_j, m_j)}{f_{i+1,k} \cdot \phi(\tau_i) + f_{i+1,i} \cdot \mu(\tau_i - \tau_k)} \cdot q_{ik}. \hspace{1cm} (34)$$