Bayesian nonparametric marked Hawkes processes for earthquake modeling

Hyotae Kim and Athanasios Kottas *

December 14, 2024

Abstract

We propose a Bayesian nonparametric model for marked Hawkes processes (MHPs). The conditional intensity function of the processes is decomposed into the ground process intensity and the mark density function. Our primary focus is on modeling the ground process intensity, while also providing a specifically defined model for the mark density function and its flexible alternative. The prior probability model for the intensity is carefully designed to balance flexibility with tractable posterior inference, achieved through a novel mixture modeling method. This model is motivated by seismology applications, where magnitude is treated as a mark associated with the time of earthquake occurrences. Accordingly, the mixture model basis is defined as a function of occurrence time and magnitude, with its functional form chosen to ensure model flexibility and alignment with earthquake data characteristics, such as the fact that larger magnitude earthquakes generate more subsequent shocks than smaller ones. The model is illustrated with three synthetic data examples and an earthquake occurrence data set.

Keywords: Marked Hawkes process; Nonparametric modeling; Bayesian Earthquake occurrence model; Erlang mixtures; Gamma process; Markov chain Monte Carlo

*Hyotae Kim (hyotae.kim@duke.edu) is Postdoctoral Researcher, Department of Biostatistics & Bioinformatics, Duke University, and Athanasios Kottas (thanos@soe.ucsc.edu) is Professor, Department of Statistics, University of California, Santa Cruz. This work is part of the Ph.D. dissertation of H. Kim, completed at University of California, Santa Cruz. The research was supported in part by the National Science Foundation under award SES 1950902.

1 Introduction

The Hawkes process (HP) is a stochastic process characterized by the property that each event increases the chance of subsequent events occurring within a certain period of time. This so-called *self-exciting* nature causes an event to generate descendant events, each of which can, in turn, trigger its own group of descendant events, yielding event clusters akin to a family tree. Consequently, the intensity at a given time is influenced by previous events, making the process inherently non-Markovian. A representative example is earthquake occurrence, where each earthquake acts as a catalyst for subsequent events, often forming clusters consisting of a main shock followed by multiple aftershocks. HPs have been widely applied, particularly in seismology (e.g., Ogata, 1988, 1998; Zhuang et al., 2002; Veen and Schoenberg, 2008), as well as in criminology (e.g., Mohler et al., 2011), finance (e.g., Da Fonseca and Zaatour, 2014; Hardiman et al., 2013; Rambaldi et al., 2015), and biology (e.g., Balderama et al., 2012).

The marked Hawkes process (MHP) is an extension of the HP that incorporates marks, which are observable variables associated with each time point (Hawkes, 1971a,b). Marks provide additional features beyond time in the point pattern. For instance, the magnitude of an earthquake can serve as a mark in seismology applications (e.g., Ogata, 1988, 1998). In criminology, the type of crime can be used as a mark to study the dynamics of specific crime types (Mohler, 2014). Similarly, in epidemiology, the finetype (strain) of a disease can be treated as a mark for forecasting strain-specific epidemics of infectious diseases (Meyer et al., 2012).

The theoretical background for (marked) Hawkes processes can be found in Daley et al. (2003). Comprehensive reviews of general Hawkes processes are available in Laub et al. (2015) and Reinhart (2018). Despite the widespread application of MHPs, their modeling

approaches have received limited attention and are largely confined to a single parametric model family. The Epidemic Type Aftershock Sequences (ETAS) model, originally developed for earthquake forecasting in Ogata (1988), remains the most widely used framework for MHPs, along with extensions for spatio-temporal applications (e.g., Kagan, 1991; Musmeci and Vere-Jones, 1992; Ogata, 1998). ETAS models have been further enhanced by allowing model parameters to vary over time (Kumazawa and Ogata, 2014) and space (Ogata et al., 2003; Nandan et al., 2017).

The ETAS model has a structural limitation: the distribution of aftershock occurrence times—specifically the waiting time from the parent earthquake—is assumed to be independent of the parent's magnitude (mark history). This assumption can pose a significant drawback. We demonstrate that our alternative model provides new insights into seismic patterns and improves the accuracy of future event predictions. The spatio-temporal extension of the ETAS model also retains this limitation, maintaining independence between time and magnitude. Therefore, our primary contribution is the development of a novel, nonparametric flexible modeling framework that relaxes the potentially unrealistic separability assumption between aftershock occurrence times and parent shock magnitudes. Some spatio-temporal models in Ogata (1998) introduce dependencies between aftershock occurrence location and magnitude using joint non-separable parametric functions. While the spatio-temporal extension is beyond the scope of this paper, our framework can readily be extended to capture dependencies not only between time and magnitude but also between location and magnitude. We briefly outline this extension as a direction for future research in the discussion section.

For inference about models for (marked) Hawkes processes, including the ETAS model, Rasmussen (2013) proposed Bayesian methods based on the likelihood, defined either directly via the HP conditional intensity or through the HP cluster representation. The

latter uses data augmentation, with the branching structure required for the representation and treated as missing data. Ross (2021) improved the cluster representation-based approach, enhancing its computational efficiency. We propose a Bayesian nonparametric (BNP) modeling and inference method for MHPs. To the best of our knowledge, nonparametric approaches have not been explored in MHP modeling.

Our proposal models the (temporal) MHP conditional intensity function, with particular emphasis on earthquake applications, though it is not limited to such cases as long as the data include a continuous mark, as in seismology. The intensity function is decomposed into the *ground process* intensity function and the mark density function. The ground process intensity determines the expected size and temporal distribution of events in an MHP, while the mark density specifies the distribution of the associated marks.

Our primary focus is to model the ground process intensity function efficiently, ensuring both model flexibility and tractability for posterior inference. We introduce a specifically designed mark density function that provides practical advantages. However, the choice of the functional form for the mark density function is relatively open, as it does not adversely affect the flexibility or computational efficiency of our model for the ground process intensity function. This is because the likelihoods for the ground process and the mark density are separated.

The rest of the paper is organized as follows. Section 2 provides background on MHPs, including the definition of the MHP intensity function, the likelihood, and stability conditions for the process. Section 3 details the modeling and inference methodology for the MHP intensity function and introduces a predictive performance measure for model assessment. The proposed modeling approach is demonstrated with synthetic and real data in Sections 4 and 5, respectively. Finally, Section 6 concludes with a summary and general discussion.

2 Background

Marked point processes, including the MHP, have intensities of the form $\lambda^*(t,\kappa) = \lambda_g^*(t) f^*(\kappa|t)$ with time $t > 0$ and mark $\kappa \in \mathcal{K}$, where $\mathcal{K}$ denotes a mark space. The term $\lambda_g^*(t)$ in the factorized intensity function is referred to as the *ground process* intensity function and is used to generate a sequence of time points for a marked point process. $f^*(\kappa|t)$ denotes a mark density function from which we can draw a mark sequence for the point process. The asterisk $(*)$ on the functions indicates dependence on the past history of the process, defined as $\mathcal{H}(t) = \{(t_i, \kappa_i) : t_i < t\}$. In principle, the mark density function is conditioned upon the time of the current event as well as the past history of times and marks. In practice, however, simplified mark density functions are widely used; for example, the ETAS model assumes i.i.d. marks with an exponential density function, $f^*(\kappa|t) = f(\kappa) = \text{Exp}(\kappa|\psi)$. The ETAS model will be discussed in more detail later in this section.

The MHP's ground process intensity resembles the typical Hawkes process intensity function in its functional form, except that it doesn't only depend on time history but also on mark history. That is,

$$\lambda_g^*(t) \equiv \lambda_g(t|\mathcal{H}(t)) = \mu(t) + \sum_{t_i<t} h(t - t_i, \kappa_i),$$

where $\mu(t) > 0$ for $t > 0$ and $h(x_i, \kappa_i) > 0$ for $x_i = t - t_i > 0$ and $\kappa_i \in \mathcal{K}$ are the background intensity (immigrant intensity) and excitation (offspring intensity) functions, respectively. Denote by $\alpha(\kappa) = \int_0^\infty h(u,\kappa)du$ the *total offspring intensity* at κ. This function determines how many offspring a predecessor with mark κ has and serves as the branching ratio of the typical HP. For any $\kappa \in \mathcal{K}$, the total intensity must be finite for a stable process, one that is not susceptible to exploding as time passes. The following are the MHP stability conditions under the assumption of $f^*(\kappa|t) = f(\kappa)$, (Daley et al., 2003, p. 203)

(i) $\alpha(\kappa) = \int_0^\infty h(u,\kappa)du < \infty$ for any $\kappa \in \mathcal{K}$; and

(ii) $\rho = \mathrm{E}(\alpha(\kappa)) = \int_{\mathcal{K}} \alpha(\kappa) f(\kappa) d\kappa < 1$.

The expected total intensity ρ is the average number of offspring; an immigrant has the potential to produce ρ number of offspring, each of whom has the same potential. Thus, the expected size of a cluster arising from an immigrant can be expressed as an infinite sum in terms of the expected total intensity, that is, $1 + \rho + \rho^2 + \ldots = \sum_{r=0}^{\infty} \rho^r$. Accordingly, the restriction of $\rho < 1$ ensures that the geometric series is finite, and so is the size of a cluster. Consequently, the MHP becomes stable given a finite number of immigrants $M = \int_0^T \mu(t) dt$ within an observation window $(0, T)$.

Using the finite total intensity function, we can factorize the excitation function into $\alpha(\kappa)$ and $g_\kappa(x) = h(x, \kappa)/\alpha(\kappa)$. Particularly, $g_\kappa(x)$ is a density function such that $g_\kappa(x) > 0$ for $x > 0$ and $\kappa \in \mathcal{K}$ and $\int_0^\infty g_\kappa(u) du = 1$. Throughout this paper, we will call it the *offspring density function* of the MHP.

Denote by $\{(t_i, \kappa_i) : i = 1, \ldots, n\}$ a realization from a MHP with intensity $\lambda^*(t, \kappa) = \lambda_g^*(t) f(\kappa)$ on $(0, T) \times \mathcal{K}$. The likelihood of the MHP realization is defined as (Daley et al., 2003, Proposition 7.3.III)

$$\Big[\prod_{i=1}^{n} f(\kappa_i)\Big] \exp\Big\{ - \int_0^T \lambda_g^*(u) du\Big\} \Big[\prod_{i=1}^{n} \lambda_g^*(t_i)\Big].$$

As in the typical HP, the ground process intensity $\lambda_g^*(t_i)$ can be expressed using the HP cluster representation. So, given the branching structure $\{y_i : i = 1, \ldots, n\}$, the MHP likelihood can be denoted as

$$\begin{aligned} &\Big[\prod_{i=1}^{n} f(\kappa_i)\Big] \Big[\exp\Big\{ - \int_0^T \mu(u) du\Big\} \prod_{\{t_i \in I\}} \mu(t_i)\Big] \\ &\times \Big[\exp\Big\{ - \sum_{i=1}^{n} \int_0^{T - t_i} h(u, \kappa_i) du\Big\} \prod_{\{i : t_i \in O\}} h(t - t_{y_i}, \kappa_{y_i})\Big], \end{aligned} \tag{1}$$

where I is the immigrant process such that $I = \{t_i : y_i = 0, i = 1, \ldots, n\}$, and O the superposition of the offspring processes such that $O = \{t_i : y_i \neq 0, i = 1, \ldots, n\}$.

Finally, we review the ETAS model (Ogata, 1988) to clarify its limitation and the motivation for our model. Denote by $\text{Powlaw}(x|p, c)$ the power-law density function, defined as $pc^p/(c + x)^{p+1}$ for $p, c > 0$. The ETAS model is given by

$$\begin{aligned} \lambda_g^*(t) &= \mu + \sum_{t_i<t} a_\alpha \exp\{\eta(\kappa_i - \kappa_0)\}\text{Powlaw}(t - t_i|p, c) \\ f^*(\kappa|t) &= f(\kappa) = \text{Exp}(\kappa|\psi) = \psi \exp\{-\psi\kappa\}, \;\; \kappa \in [\kappa_0, \infty) = \mathcal{K}, \end{aligned} \tag{2}$$

with parameters $a_\alpha, \eta, \psi > 0$ to be estimated and a given constant $\kappa_0 > 0$ for the mark space $\mathcal{K}$. In the modeling framework, the power-law density $\text{Powlaw}(t - t_i|p, c)$ plays a role as the offspring density function. Since the function is invariant to κ_i, it cannot account for variations in mark history. The demand for a mark-dependent offspring density function is a driving force behind our modeling method; our case study with earthquake data in Section 5 illustrates how the temporal dynamics of offspring (aftershock) vary according to marks.

In addition, the total intensity function in the simple parametric form of $a_\alpha \exp\{\eta(\kappa_i - \kappa_0)\}$ in the ETAS model may not adequately explain the observed aftershock counts; several alternative (but still parametric) functions have been proposed rather than the simple exponential function (Ogata and Zhuang, 2006). It is noted that the total intensity functions maintain the increasing property over κ_i, which reflects the fact that a shock of greater magnitude will generate more subsequent shocks. As such, a more flexible $\alpha(\kappa_i)$ that retains the increasing pattern in κ_i may yield more relevant estimates for the evolution of offspring (aftershock) counts with respect to the mark (magnitude).

3 Methodology

3.1 Model formulation

In this section, we propose a model that addresses the limitations of the ETAS model mentioned above. As such, the modeling framework features a mark-dependent offspring density function and a flexible (nonparametric) total intensity function. In addition, it is capable of tractable posterior inference using a general MCMC method, outlined in Section 3.4.

We develop a Bayesian prior probability model for the MHP intensity function, namely the ground process intensity function, where the immigrant intensity function is assumed to be constant, without loss of generality, for simplicity in the model description. We are primarily concerned with modeling the excitation function of the ground process in order to accomplish the aforementioned characteristics of the offspring density and total intensity functions. Consequently, the section will mainly concentrate on $h(x, \kappa)$. As an alternative to the constant immigrant, Section 3.2 reviews a nonparametric modeling method, originally developed for immigrant intensity of the general Hawkes process in **Kim and Kottas (2024)**. By substituting the nonparametric model for the constant immigrant function, we can obtain a fully nonparametric model for the ground process intensity function, and Section 5 presents the results of fitting this model to an actual earthquake data set.

Denote by $\text{Ga}(x|a, b)$ for $x \in \mathbb{R}^+$, the gamma density function with shape a, rate b (or scale b^{-1}), and mean ab, which is also called the Erlang density with positive integer shape parameter, $a \in \mathbb{N}$. We model $h(x, \kappa)$ as a weighted combination of basis functions, each

formed by the product of an Erlang density and a polynomial function $b_m(\kappa; d)$, as follows:

$$\begin{aligned} h(x,\kappa) &= \sum_{l=1}^{L}\sum_{m=1}^{M} \nu_{lm}\mathrm{Ga}(x|l,\theta^{-1})b_m(\kappa;d), \quad x\in\mathbb{R}^+ \text{ and } \kappa\in(\kappa_0,\kappa_{max}) = \mathcal{K}\subset\mathbb{R} \\ \nu_{lm} &= H(A_{lm}), \quad H|H_0,c_0 \sim \mathcal{G}(H_0,c_0), \end{aligned} \tag{3}$$

where $A_{lm} = [(l-1)\theta, l\theta) \times [(m-1)/M, m/M)$, $l = 1,\ldots,L$ and $m = 1,\ldots,M$. The mixture weights are defined through increments of a random measure H on $\mathbb{R}^+ \times (0,1)$. Let $u_\kappa \equiv u(\kappa;\kappa_0,\kappa_{max}) = (\kappa-\kappa_0)/(\kappa_{max}-\kappa_0)$ for $\kappa\in(\kappa_0,\kappa_{max})$. Then, the polynomial function in the mixture basis is

$$b_m(\kappa;d) \equiv b(\kappa,d,m,M) = M u_\kappa^{(m-1)^d},$$

with parameter $d\in\mathbb{R}$ to be estimated. Under the modeling framework, the total offspring intensity is derived as $\alpha(\kappa) = \sum_{m=1}^{M} V_m b_m(\kappa;d)$, where $V_m = \sum_{l=1}^{L}\nu_{lm}$. Each component $b_m(\kappa;d)$ presents a non-decreasing function over κ with increasing rates varying in m (see Figure 1), which results in an increasing function of κ for the total offspring intensity for $M > 1$. Again, this is a favorable characteristic in modeling aftershock frequency, since a stronger earthquake would be expected to produce more subsequent quakes. A larger M enhances the expressiveness of the derived total offspring intensity model by incorporating more mixture components with different slopes. Model parameter d allows for the adjustment of the slopes (of the component functions) in such a way that a small value close to 0 results in a nearly linear increasing function for the total intensity, whereas a large value supports the intensity function with a more radical increase over κ. We can further improve the model's flexibility by adding an additional parameter $e > 0$, such that $b_m(\kappa;d,e) = M u_\kappa^{[e(m-1)]^d}$. The model with $0 < e < 1$ can represent even functions whose increasing rates diminish over κ (i.e., the second derivative of the intensity function is

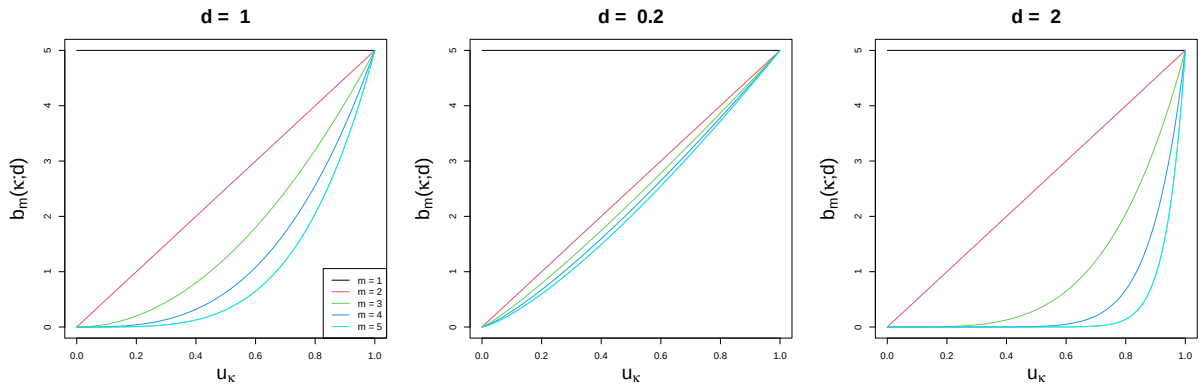

Figure 1: Polynomial functions of $b_m(\kappa; d)$, $m = 1, \ldots, M$, for $d = 1$ (left), 0.2 (middle), and 2 (right) with $M = 5$.

negative). In this regard, the simple extension may be useful in applications where a more diverse shape of increasing total intensity is expected.

The gamma process, $\mathcal{G}(H_0, c_0)$, in (3) has a centering measure H_0 on $\mathbb{R}^+ \times (0, 1)$ and a precision parameter $c_0 > 0$. Placing $\mathcal{G}(H_0, c_0)$ prior on H implies that $H(A_{lm})$, $l = 1, \ldots, L$, $m = 1, \ldots, M$, follow independent gamma distributions with mean $H_0(A_{lm})$ and variance $H_0(A_{lm})/c_0$ for any partition A_{lm} of $(0, L\theta) \times (0, 1)$. Let $A_{lm} = A_l \times A_m = [(l-1)\theta, l\theta) \times [(m-1)/M, m/M)$. We consider a productive-formed centering measure, $H_0(A_{lm}) = H_{0x}(A_l) \times H_{0\kappa}(A_m)$, where the two measures are based on a Weibull cumulative hazard function with shape b_1 (and scale 1) and an exponential cumulative hazard function with rate b_2, respectively, such that

$$H_{0x}(A_l) = H_{0x}\big(l\theta\big) - H_{0x}\big((l-1)\theta\big)\big) = \big(l\theta\big)^{b_1} - \big((l-1)\theta\big)^{b_1};$$
$$H_{0\kappa}(A_m) = H_{0\kappa}\big(m/M\big) - H_{0\kappa}\big((m-1)/M\big) = b_2/M.$$

The precision parameter c_0 governs not only the variability of H around H_0 but also the flexibility of the prior models for the offspring density function and the total intensity

function (this will be discussed in more detail later). The centering functions $H_{0x}(A_l)$ and $H_{0\kappa}(A_m)$ are relatively flexible in choice compared to the key parameter c_0; the posterior estimation results under the model are robust to the choice of functions, unless c_0 is fixed at a large constant, thereby preventing shrinkage to a small value close to zero. Empirically, the shape of the prior model for the offspring density function is determined by $H_{0x}(\cdot)$, specifically its density function, that is, the derivative of $H_{0x}(\cdot)$. For example, the Weibull cumulative hazard function with $b_1 < 1$ has a decreasing derivative, $H'_{0x}(x) = b_1 x^{b_1-1}$, and, thus, a decreasing (prior) offspring density function, as shown later in Figure 2 for prior simulations. Note that, in earthquake applications, $g_\kappa(x)$ should decrease with x (the distance between an aftershock and its predecessor) as above so that the chance of a subsequent quake occurring diminishes as time passes from the current shock. In general, no information is available regarding the shape of the offspring density; a simple exponential cumulative hazard function can be used as the default, which provides a constant (prior) offspring density function.

To complete the modeling approach for the MHP intensity function, we also need to define the mark density along with the ground process intensity model described above. Instead of the exponential mark density of the ETAS model, we consider the beta density function of $\text{Be}(u_\kappa|a_\beta, b_\beta)$ for u_κ, which provides an analytical expression for the expected total intensity of ρ, given by (6). Accordingly, the MHP stability condition of $\rho < 1$ can be achieved by controlling parameters in ρ, as described in Section 3.3. For greater flexibility, a mixture of beta densities can serve as an alternative, such as $\pi\text{Be}(u_\kappa|a_{\beta,1}, b_{\beta,1}) + (1 - \pi)\text{Be}(u_\kappa|a_{\beta,2}, b_{\beta,2})$, which can be generalized as $\sum_{r=1}^{R} \pi_r \text{Be}(u_\kappa|a_{\beta,r}, b_{\beta,r})$. This approach addresses potential distortion issues in the estimation results caused by misspecification of the mark density function model.

3.1.1 Model property

Under our modeling method, we can obtain ready expressions for the following key functions. The total offspring intensity function is given by

$$\alpha(\kappa) = \int_0^\infty h(x,\kappa)dx = \sum_{l=1}^{L}\sum_{m=1}^{M} \nu_{lm} b_m(\kappa; d) = \sum_{m=1}^{M} V_m \Big\{ M u_\kappa^{(m-1)^d} \Big\}, \tag{4}$$

where $V_m = \sum_{l=1}^{L} \nu_{lm} \overset{ind.}{\sim} \text{Ga}\big(c_0 \sum_{l=1}^{L} H_0(A_{lm}), c_0\big)$. The prior mean $\text{E}^\nu(\alpha(\kappa))$ and variance $\text{Var}^\nu(\alpha(\kappa))$ of the total intensity over the weights of V_m, $m = 1, \ldots, M$, are available in closed-form,

$$\begin{aligned} \text{E}^\nu(\alpha(\kappa)) &= \sum_{m=1}^{M} \text{E}(V_m) M u_\kappa^{(m-1)^d} = (L\theta)^{b_1} b_2 \sum_{m=1}^{M} u_\kappa^{(m-1)^d} \\ \text{Var}^\nu(\alpha(\kappa)) &= \sum_{m=1}^{M} \text{Var}(V_m) \Big\{ M u_\kappa^{(m-1)^d} \Big\}^2 = \frac{(L\theta)^{b_1} b_2 M}{c_0} \sum_{m=1}^{M} u_\kappa^{2(m-1)^d}. \end{aligned} \tag{5}$$

The precision parameter, c_0, of the gamma process prior is a major contributor to the prior uncertainty of $\alpha(\kappa)$ and negatively associated with it. In contrast, the number of polynomial component functions, M, has a positive impact on the prior uncertainty.

With the beta mark density function of $\text{Be}(u_\kappa | a_\beta, b_\beta)$, the expected total intensity is derived as,

$$\begin{aligned} \rho &= \int_{\mathcal{K}} \alpha(\kappa) f^*(\kappa|t) d\kappa = \int_0^1 \alpha(\kappa) \text{Be}(u_\kappa | a_\beta, b_\beta) du_\kappa \\ &= \sum_{m=1}^{M} V_m M \frac{\text{Beta}\big(a_\beta + (m-1)^d, b_\beta\big)}{\text{Beta}(a_\beta, b_\beta)}, \end{aligned} \tag{6}$$

where $\text{Beta}(a, b)$ is a beta function. The stability condition, $\rho < 1$, cannot be strictly enforced with analytical solutions for some parameters due to the presence of random variables V_m and d. Instead, we specify priors for model parameters in a way that the

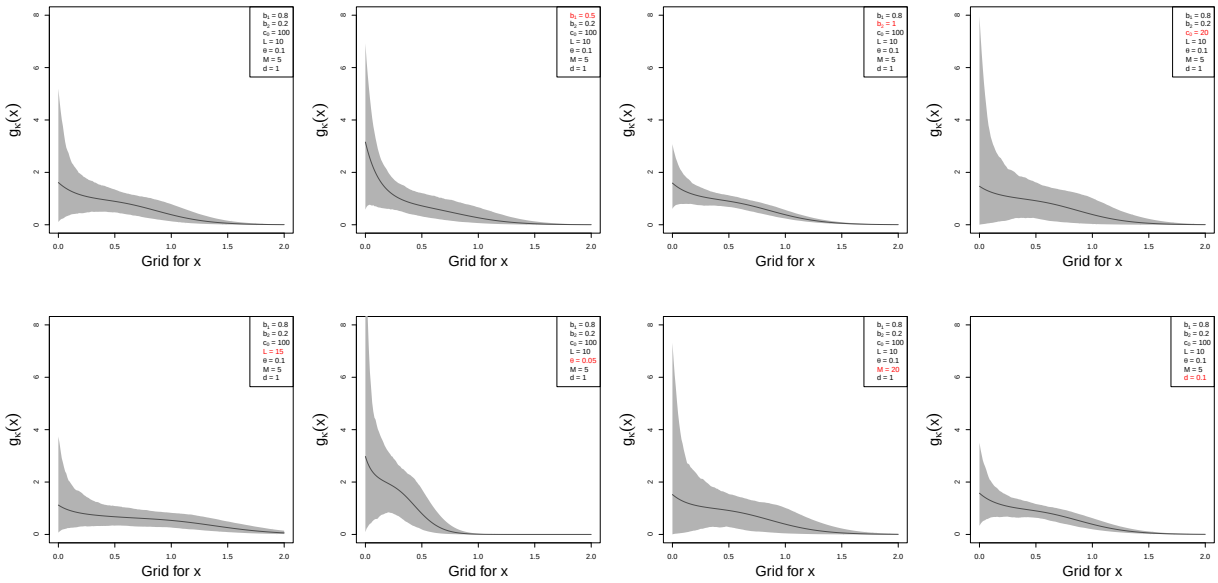

Figure 2: Empirical results of $1,000$ realizations for the offspring density function with different choices of model parameters: Prior mean (solid line) and 95% prior uncertainty bands (shaded area) for $g_\kappa(x)$ with $\kappa = 5.5$.

random total intensity has most of its probability on $\rho < 1$, detailed in the following section.

The offspring density function can be expressed as a weighted combination of Erlang densities. That is,

$$g_\kappa(x) = \frac{h(x,\kappa)}{\alpha(\kappa)} = \sum_{l=1}^{L} \Big(\frac{\sum_{m=1}^{M} \nu_{lm} b_m(\kappa; d)}{\sum_{m=1}^{M} V_m b_m(\kappa; d)} \Big) \mathrm{Ga}(x|l, \theta^{-1}) = \sum_{l=1}^{L} W_l(\kappa) \mathrm{Ga}(x|l, \theta^{-1}). \quad (7)$$

The mixture weights of $W_l(\kappa)$, $l = 1, \ldots, L$, are functions of κ, which enables the offspring density function to change with the mark. We illustrate the implications of model parameters $(b_1, b_2, c_0, L, \theta, M, d)$ on the prior model for $g_\kappa(x)$ using $1,000$ realizations of the function, for which ν_{lm}'s are drawn $1,000$ times from gamma distributions of $\mathrm{Ga}\big(c_0 H_0(A_{lm}), c_0\big)$.

Figure 2 shows how the prior mean and variance of $g_\kappa(x)$ vary with changes in parameter values. The first two panels demonstrate that b_1 of $H_{0x}(A_l)$ affects the prior mean of $g_\kappa(x)$, leading to an alteration in the shape of the offspring density function. L and θ also influence the mean; the product of $L\theta$ provides a rough indication of the effective support of $g_\kappa(x)$. As an example, the first two bottom left panels with $L\theta = 1.5$ and 0.5 contain the majority of probabilities in $(0, 1.5)$ and $(0, 0.5)$. Other parameters, (b_2, c_0, M, d), are associated with the prior model uncertainty (also called model flexibility): it has negative relationships with (b_2, c_0) and positive relationships with (M, d). Furthermore, we found that, for a fixed $L\theta$, a smaller θ (and thus a larger L) yields greater model flexibility (not shown here).

3.2 Immigrant intensity modeling

As a more generic approach, we relax the assumption of constant immigrant intensity and take a nonparametric modeling method present in **Kim and Kottas (2024)**. The following is a brief description of the model for immigrant intensity.

Let $\mu(t)$ denote the immigrant intensity function, represented as an Erlang mixture model given by

$$\begin{aligned} \mu(t) &= \sum_{j=1}^{J} \omega_j \mathrm{Ga}(t \mid j, \phi^{-1}) \\ \omega_j &= G\big(j\phi\big) - G\big((j-1)\phi\big), \;\; j = 1, \ldots, J, \quad G \mid G_0, e_0 \sim \mathcal{G}(G_0, e_0). \end{aligned} \tag{8}$$

Again, $\mathcal{G}(G_0, e_0)$ is the gamma process with precision parameter $e_0 > 0$, and centering cumulative intensity function G_0, where $\mathrm{E}(G(t)) = G_0(t)$ and $\mathrm{Var}(G(t)) = G_0(t)/e_0$. The centering function is defined as $G_0(t) = t/b_{G_0}$. As $\phi \to 0$ and $J \to \infty$, the model converges pointwise to the intensity function of G. Thus, with the stochastic process prior on G, the

immigrant intensity function acquires flexibility, allowing it to represent any function $f(t) : \mathbb{R}^+ \to \mathbb{R}^+$. For practical model implementation, a constant J is used, with larger values required to capture the dynamics of a non-standard and more complex underlying function. Posterior inferences about the model parameters can be obtained using a general MCMC method; a majority of the parameters, including the mixture weights ω_j, $j = 1, \ldots, J$, have prior conjugacy, resulting in posterior full conditional distributions in closed-form.

3.3 Prior specification

This section outlines a strategy for specifying the fixed parameters L and M, as well as the prior for $(\theta, c_0, d, b_1, b_2, a_\beta, b_\beta)$ in the proposed model. First, θ is assigned a Lomax distribution $\text{Lo}(2, d_\theta)$ prior, as in the Erlang mixture model for NHPPs (Kim and Kottas, 2022). The shape parameter of 2 results in infinite variance for θ, which makes the model robust to the choice of d_θ. Recall that $L\theta$ serves as the upper bound of the effective support for the offspring density function. The scale parameter d_θ is determined using the probability approximation, $\Pr(0 < \theta < T_O) \approx 0.99$, which gives a conservative choice for θ that, in the following simulation studies and real data analysis, presents moderate prior-to-posterior learning for finding an adequately small θ value for any choice of L, including $L = 1$. T_O is the upper bound of the time difference $x_i = t_i - t_{y_i}$. Unless the exact branching structures of y_i are provided, the precise range of x_i cannot be determined. In real-world applications, limited information about the branching structure, which may still be useful for selecting T_O, is often available. For example, the Japan earthquake data in Section 5 includes an index for differentiating between main shocks (immigrant) and aftershocks (offspring). Then, subtracting the offspring time points from their nearest preceding immigrant time points provides a rough estimate of the range of x_i values. Previous findings or known facts may also be used. As an example, you can find the results of a study on the waiting time

between an aftershock and a main shock in Reasenberg et al. (1999).

The precision parameter, c_0, of the gamma process prior takes an exponential prior with rate a_{c_0}. The hyperparameter is set to provide a conservative upper bound (empirically chosen) for c_0; we used Exp(0.005) prior (with mean of 200 and 90 percentile of 461) for the following examples of Sections 4 and 5 and observed substantial prior-to-posterior learning toward 0.

We place an exponential prior on d and set the prior mean to 1, an empirical choice that leads to moderate prior-to-posterior learning in all the following examples.

We assign Exp(a_{a_β}) and Exp(a_{b_β}) priors to parameters a_β and b_β of the beta mark density function. By setting the prior means $1/a_{a_\beta}$ and $1/a_{b_\beta}$ to 1 and plugging the means into the parameters, the beta mark density function turns into a uniform density function. This conservative choice that does not assume a particular shape for mark density would be suitable for use as a default. We can also take advantage of observed marks data, namely their mean κ_{avg}, for the hyperparameter specification. The beta mark density function has the mean of $a_\beta/(a_\beta + b_\beta)$ for u_κ, and we match it to the average $(\kappa_{avg} - \kappa_0)/(\kappa_{max} - \kappa_0)$. By replacing a_β with $\mathrm{E}(a_\beta) = 1$ (or b_β with $\mathrm{E}(b_\beta) = 1$) and solving the equation for the other parameter, an appropriate value for b_β (or a_β) can be acquired. We use this value for specifying the prior as $\mathrm{E}(b_\beta)$ (or $\mathrm{E}(a_\beta)$).

Parameter b_1 of H_{0x} determines the shape of $g_\kappa(x)$, as shown in Figure 2. The decreasing characteristic of $g_\kappa(x)$ for earthquake applications can be attained by setting b_1 to less than 1. We assign the Exp(1) prior to b_1, which places the majority of probabilities on the unit interval.

Parameter b_2 takes Exp(a_{b_2}) prior, with a_{b_2} specified using the stability condition, $\rho < 1$. The prior density function of ρ and its cumulative distribution function play a key role in the specification. Since the functions are not available in analytical form, a kernel density

estimate and an empirical distribution are employed in practice, for which prior samples of ρ are produced as follows. We first draw prior samples of V_m and then feed them into equation (6) for sampling ρ. This procedure requires model parameters $(\theta, d, c_0, b_1, b_2, a_\beta, b_\beta)$, which we replace with their prior means. The equation also involves L and M, and we will discuss a way of finding plausible values for these variables later in this section. Lastly, the prior mean of $\mathrm{E}(b_2) = 1/a_{b_2}$ is demanded to substitute b_2. The hyperparameter is adaptively chosen to ensure ρ can take most values within the unit interval with non-negligible prior densities, and more importantly, $\Pr(\rho < 1) \approx 1$ under the empirical distribution.

We work with fixed L and M, which are specified using sensitivity analysis as outlined below. As L and M significantly impact the model's computational efficiency in posterior inference, we recommend starting with small values, such as 5 / 10. You may also set L's initial value based on the relationship between $L\theta$ and T_O as $L^{\text{init}} = \lceil T_O/\hat{\theta} \rceil$, where $\lceil x \rceil$ is the ceiling function of x and $\hat{\theta}$ the prior mean of θ. Those values are gradually increased until no noticeable differences are observed in the two key functions of $\alpha(\kappa)$ and $g_\kappa(x)$. In principle, having a larger number of mixture components should yield better estimates of the underlying functions. However, in practice, we seek values of L and M that are relatively small yet provide accurate estimation and prediction comparable to those achieved with higher values.

Prior(s) for the immigrant intensity function depend on its functional form. For constant immigrant intensity, an exponential distribution $\mathrm{Exp}(a_\mu)$ prior is used, with the hyperparameter derived by matching the observed number of immigrant points n_I to the expected number of immigrant points within the observation window $(0, T)$, such that $n_I = T \times \mathrm{E}(\mu) = T/a_\mu$. In the case of unknown n_I, the total number of observed points n can be an alternative, which is expected to have a higher degree of prior-to-posterior learning. Prior specification for the Erlang mixture-based immigrant intensity function

follows the strategy of **Kim and Kottas (2024)**; we employ the same approach (detailed in Section A.1 of the reference) to choose priors for our model parameters (ϕ, e_0, b_{G_0}) and a fixed value for J.

3.4 Posterior inference

Denote by $\{0 < t_1 < t_2 < \cdots < t_n < T\}$ an observed point pattern in the observation window $(0, T)$ and by $\{\kappa_i \in \mathcal{K}\colon\ i = 1, \ldots, n\}$ the observed marks in the mark space $\mathcal{K} = (\kappa_0, \kappa_{max})$. Given the branching structure $\boldsymbol{y} = (y_1, \ldots, y_n)$, the MHP likelihood under our modeling framework is derived as

$$\begin{aligned}&\Big[\prod_{i=1}^{n} \mathrm{Be}(u_{\kappa_i}|a_\beta, b_\beta)\Big] \exp\Big\{-\int_0^T \mu(u)du\Big\} \exp\Big\{-\sum_{l=1}^{L}\sum_{m=1}^{M} \nu_{lm} K_{lm}(\theta, d)\Big\} \\ &\times \prod_{t_i \in I} \mu(t_i) \prod_{t_i \in O} \Big\{\sum_{l=1}^{L}\sum_{m=1}^{M} \nu_{lm} \mathrm{Ga}(t_i - t_{y_i}|l, \theta^{-1}) b_m(\kappa_{y_i}; d)\Big\},\end{aligned} \tag{9}$$

where $K_{lm}(\theta, d) \equiv K(\theta, d, l, m) = \sum_{i=1}^{n} b_m(\kappa_i; d) \int_0^{T-t_i} \mathrm{Ga}(s|l, \theta^{-1})ds$. Here, $\mu(t)$ represents a generic immigrant intensity function, which we model as either a constant function or an Erlang mixture in this paper. I and O are an immigrant process and the superposition of offspring processes, defined as $I = \{t_i : y_i = 0, i = 1, \ldots, n\}$ and $O = \{t_i : y_i \neq 0, i = 1, \ldots, n\}$.

We augment the likelihood with auxiliary variables $\boldsymbol{\xi} = \{\boldsymbol{\xi_i} : i = 1, \ldots, n_O\}$, where $\boldsymbol{\xi_i} = (\xi_{i1}, \xi_{i2})$ identifies the basis function to which an event $(t_i, \kappa_i) \in O$ is assigned. n_O is the number of offspring points categorized by the branching structure, calculated by $|O| = |\{t_i : y_i \neq 0, i = 1, \ldots, n\}|$, where $|A|$ is the cardinality of set A. Similarly, we can define $n_I = |I| = |\{t_i : y_i = 0, i = 1, \ldots, n\}|$. Let $\boldsymbol{\Theta} = (\boldsymbol{y}, \boldsymbol{\tau}_\mu, \theta, d, \boldsymbol{\xi}, \boldsymbol{\nu}, a_\beta, b_\beta)$, where $\boldsymbol{\tau}_\mu$ is a vector of parameter(s) in $\mu(t)$. Then, the hierarchical model for the augmented data can

be represented as

$$
\begin{aligned}
p\big((\boldsymbol{t},\boldsymbol{\kappa})|\boldsymbol{\Theta}\big) &\propto \Big[\prod_{i=1}^{n}\mathrm{Be}(u_{\kappa_i}|a_\beta,b_\beta)\Big] \\
&\quad\times \exp\Big\{-\int_0^T \mu(u)du\Big\}\exp\Big\{-\sum_{l=1}^{L}\sum_{m=1}^{M}\nu_{lm}K_{lm}(\theta,d)\Big\} \\
&\quad\times \prod_{t_i\in I}\mu(t_i)\prod_{t_i\in O}\Big\{\Big(\sum_{r_1=1}^{L}\sum_{r_2=1}^{M}\nu_{r_1r_2}\Big)\mathrm{Ga}(t_i-t_{y_i}|\xi_{i1},\theta^{-1})b_{\xi_{i2}}(\kappa_{y_i};d)\Big\}; \\
p(\boldsymbol{\xi_i}|\boldsymbol{\nu}) &\propto \sum_{l=1}^{L}\sum_{m=1}^{M}\frac{\nu_{lm}}{\Big(\sum_{r_1=1}^{L}\sum_{r_2=1}^{M}\nu_{r_1r_2}\Big)}\delta_{(l,m)}(\xi_{i1},\xi_{i2}), \quad i=1,\ldots,n_O; \\
p(\boldsymbol{y}) &\propto \delta_0(y_1)\prod_{i=2}^{n}\mathrm{Unif}(y_i|0,1,\ldots,i-1); \\
\nu_{lm}|c_0,b_1,b_2,\theta &\overset{ind.}{\sim} \mathrm{Ga}(c_0H_0(A_{lm}),c_0), \quad l=1,\ldots,L, \quad m=1,\ldots,M; \\
(\boldsymbol{\tau}_\mu,\theta,d) &\sim p(\boldsymbol{\tau}_\mu)\mathrm{Lo}(2,d_\theta)\mathrm{Exp}(a_d); \\
(c_0,b_1,b_2) &\sim \mathrm{Exp}(a_{c_0})\mathrm{Exp}(a_{b_1})\mathrm{Exp}(a_{b_2}); \\
(a_\beta,b_\beta) &\sim \mathrm{Exp}(a_{a_\beta})\mathrm{Exp}(a_{b_\beta}),
\end{aligned}
\tag{10}
$$

where $\delta_{(a,b)}(x,y)$ is the Dirac delta function such that $\delta_{(a,b)}(x,y)=1$ if $x=a$ & $y=b$ and $\delta_{(a,b)}(x,y)=0$ otherwise. $\mathrm{Unif}(x|0,1,\ldots,i-1)$ is the discrete uniform probability mass function. $p(\boldsymbol{\tau}_\mu)$ is the prior distribution for $\boldsymbol{\tau}_\mu$, with being either an exponential distribution for constant immigrant intensity (μ) or a product of a Lomax distribution and two exponential distributions for parameters (ϕ, e_0, b_{G_0}) of the immigrant intensity Erlang mixture model of (8). Note that the other parameters – mixture weights (ω_j) – in the immigrant intensity mixture model take gamma distributions due to the gamma process prior placed on G.

Posterior inference was carried out using the Gibbs sampler, in which simple and efficient

updates of $\boldsymbol{\xi}$, $\boldsymbol{y}$, and $\boldsymbol{\nu}$ were made through their closed-form posterior full conditional distributions. We denote the observed data set as $\boldsymbol{D} = \{\boldsymbol{t}, \boldsymbol{k}\}$, which is used in the following derivations of posterior full conditional distributions.

As the posterior full conditional for auxiliary variable $\boldsymbol{\xi_i}$, we achieve an independent discrete distribution with support of $\{(l, m) : l = 1, \ldots, L, \ m = 1, \ldots, M\}$, such that $\Pr\big((\xi_{i1} = l, \xi_{i2} = m)|\boldsymbol{y}, \theta, d, \boldsymbol{\nu}, \boldsymbol{D}\big) \propto \nu_{lm}\mathrm{Ga}(t_i - t_{y_i}|l, \theta^{-1})b_m(\kappa_{y_i}; d)$.

Given the discrete uniform priors, the branching structures y_i for $i = 2, \ldots, n$ also have discrete distributions as their posterior full conditionals with probabilities

$$\Pr(y_i = k|\mu, \theta, \boldsymbol{\nu}, d, \boldsymbol{D}) = \begin{cases} \dfrac{\mu}{\mu + \sum_{r=1}^{i-1}\sum_{l=1}^{L}\sum_{m=1}^{M}\nu_{lm}\mathrm{Ga}(t_i - t_r|l, \theta^{-1})b_m(\kappa_r; d)}, & k = 0; \\ \dfrac{\sum_{l=1}^{L}\sum_{m=1}^{M}\nu_{lm}\mathrm{Ga}(t_i - t_k|l, \theta^{-1})b_m(\kappa_k; d)}{\mu + \sum_{r=1}^{i-1}\sum_{l=1}^{L}\sum_{m=1}^{M}\nu_{lm}\mathrm{Ga}(t_i - t_r|l, \theta^{-1})b_m(\kappa_r; d)}, & k = 1, \ldots, i-1. \end{cases}$$

Let $n_{lm} = |\{t_i : \xi_{i1} = l, \xi_{i2} = m, y_i \neq 0\}|$ for $l = 1, \ldots, L$ and $m = 1, \ldots, M$. The posterior full conditional distribution for $\boldsymbol{\nu}$ is derived as $p(\boldsymbol{\nu}|\boldsymbol{\xi}, \theta, c_0, b_1, b_2, \boldsymbol{\kappa}) \propto \prod_{l=1}^{L}\prod_{m=1}^{M}\Big[\exp\Big\{-\nu_{lm}K_{lm}(\theta, d)\Big\}\nu_{lm}^{n_{lm}}\mathrm{Ga}(\nu_{lm}|c_0H_0(A_{lm}), c_0)\Big]$. Hence, the mixture weights are independently gamma distributed, such that $\nu_{lm} \overset{ind.}{\sim} \mathrm{Ga}(c_0H_0(A_{lm}) + n_{lm}, c_0 + K_{lm}(\theta, d))$ for $l = 1, \ldots, L$ and $m = 1, \ldots, M$.

Other parameters θ, d, c_0, b_1, b_2, a_β, and b_β are updated with the Metropolis-Hastings (M-H) algorithm, in which log-normal distributions are used as proposal distributions.

For the constant immigrant intensity μ, the $\mathrm{Exp}(a_\mu)$ prior results in a gamma posterior full conditional distribution with shape $n_I + 1$ and rate $T + a_\mu$. For the Erlang mixture-

based immigrant intensity, we first represents the likelihood of the immigrant process, $\exp\Big\{-\sum_{j=1}^{J}\omega_j\int_0^T \text{Ga}(u \mid j,\phi^{-1})du\Big\}\prod_{t_i\in I}\sum_{j=1}^{J}\omega_j\text{Ga}(t_i \mid j,\phi^{-1})$, as a hierarchical model with auxiliary variables $\boldsymbol{\zeta}=\{\zeta_i : i=1,\ldots,n_I\}$, that is,

$$p(\boldsymbol{t}|\boldsymbol{y},\phi,\boldsymbol{\zeta},\boldsymbol{\omega}) \propto \exp\Big\{-\sum_{j=1}^{J}\omega_j\int_0^T \text{Ga}(u \mid j,\phi^{-1})du\Big\}\prod_{t_i\in I}\Big\{\Big(\sum_{j=1}^{J}\omega_j\Big)\text{Ga}(t_i \mid \zeta_i,\phi^{-1})\Big\};$$

$$p(\zeta_i|\boldsymbol{\omega}) \propto \sum_{j=1}^{J}\frac{\omega_j}{\Big(\sum_{r=1}^{J}\omega_r\Big)}\delta_j(\zeta_i), \quad i=1,\ldots,n_I,$$

where $\delta_a(x)$ is the Dirac delta function such that $\delta_a(x)=1$ if $x=a$ and $\delta_a(x)=0$ otherwise. Then, the mixture weights (ω_j), with the $\text{Ga}(e_0\phi/b_{G_0},e_0)$ prior, are updated using the posterior full conditional in closed-form, i.e., $\omega_j \overset{ind.}{\sim} \text{Ga}(e_0\phi/b_{G_0}+n_j, e_0+\omega_j\int_0^T \text{Ga}(u \mid j,\phi^{-1})du)$, with $n_j=|\{t_i:\zeta_i=j,y_i=0\}|$. Auxiliary variables (ζ_i) are given independent discrete posterior full conditional distributions with probabilities $\Pr(\zeta_i=j|\phi,\boldsymbol{\omega},\boldsymbol{D}) \propto \omega_j\text{Ga}(t_i|j,\phi^{-1})$. Updates of the other parameters (ϕ,e_0,b_{G_0}) are accomplished via the M-H algorithm with log-normal proposal distributions.

3.5 Prediction

As a criterion for evaluating and comparing models, we employ predictive performance, which involves inferring a posterior predictive distribution for the count over a given time period and a mark interval, then comparing it with the actual observed count. To do so, we first define a prediction window of $B \equiv [T_1^*,T_2^*]\times[k_1^*,k_2^*] \subset (0,T)\times(k_0,k_{max})$, where $(0,T)$ and (k_0,k_{max}) are the observation window and the mark space for a given entire data set. The observed count $N_{obs}(B)$ is the number of observed points that fall within the prediction window, B. The predictive count can be computed by conducting the posterior simulation described below.

1. By fitting a model to an observed point pattern that falls within $(0, T_1^*) \times (k_0, k_{max})$, that is, all data prior to the time interval of the prediction window, draw posterior samples $\{\boldsymbol{\theta}_s : s = 1, \ldots, S\}$ of the model parameters $\boldsymbol{\theta}$ using the MCMC method outlined in the previous section.

2. With each $\boldsymbol{\theta}_s$, construct an MHP intensity function and draw a realization $\boldsymbol{z}^s = \{(t_i^s, k_i^s) : i = 1, \ldots, n^s\}$ from the process with the intensity function. We employ the MHP simulation method described in Rasmussen (2013) to generate point patterns.

3. For each s, count the points within the prediction window, that is, $N_{pred}^s(B) \equiv |\{(t_i^s, k_i^s) : t_i^s \in [T_1^*, T_2^*],\ \ k_i^s \in [k_1^*, k_2^*],\ \ i = 1, \ldots, n^s\}|$, where $|A|$ is the cardinality of set A.

Then, the histogram of $\{N_{pred}^s(B) : s = 1, \ldots, S\}$ presents the posterior predictive distribution of the count over the prediction window and is plotted alongside a vertical line corresponding to $N_{obs}(B)$ for graphical comparison in Section 5.

4 Simulation study

In the following sections, we illustrate the proposed model using three synthetic data sets and highlight its characteristics by comparing it to two alternative models: the (parametric) ETAS model (2) and its extension to a semiparametric model obtained by replacing the

power-law offspring density function with a nonparametric approach, defined as

$$\begin{aligned} g(x) &= \int_{\mathbb{R}^+} \theta^{-1}\mathbf{1}_{[0,\theta)}(x)dG(\theta) \quad \text{with} \quad G \sim \text{DP}(F_0, \alpha_0) \quad \text{and} \quad F_0(\theta) = \text{IG}(a,b) \\ &\approx \sum_{l=1}^{L} \omega_l Z_l^{-1}\mathbf{1}_{[0,Z_l)}(x) \quad \text{with} \quad \omega_l \overset{i.i.d.}{\sim} \text{Be}(1,\alpha_0) \quad \text{and} \quad Z_l \overset{i.i.d.}{\sim} \text{IG}(a,b), \end{aligned} \tag{11}$$

where $\text{DP}(F_0, \alpha_0)$ is a Dirichlet process with precision parameter α_0 and centering cumulative distribution function F_0. In addition, $\text{IG}(a,b)$ and $\text{Be}(c,d)$ indicate an inverse gamma distribution with shape a and mean $b/(a-1)$, and a beta distribution with mean $c/(c+d)$ and variance $cd/\Big((c+d)^2(c+d+1)\Big)$. With the DP prior, the uniform mixture model for $g(x)$ can represent any non-increasing density function on $\mathbb{R}^+$. **Kim and Kottas (2024)** applied the model to the offspring density function of the general Hawkes process, and similarly, we adopt it for the offspring density function of the ground process of MHPs.

The ETAS model is assigned the following set of priors: Exp(2.5) with $a_\alpha < 1 - (\eta/\psi)$ for a_α, Exp(0.2) with $\eta < \psi$ for η, Exp(0.1) for p, Exp(0.1) for c, and Exp(0.1) for ψ. To ensure the model complies with the process stability condition, truncations are applied to the exponential priors for a_α and η. Under the ETAS model, the expected total intensity is derived as $\rho = a_\alpha\psi/(\psi - \eta)$, and we restrict it to the unit interval to maintain stability. Following the literature, priors for the semiparametric model are set to $\text{Ga}(5, 0.25)$ for α_0, Exp(1) for a, Exp(1.5) for b, and $L = 200$, along with the analogous priors for a_α, η, and ψ. Priors for our nonparametric model are chosen based on the strategy of Section 3.3, specified individually for each example below. Along with the prior distribution above for the excitation function, it is also imperative to establish a prior distribution for the constant immigrant intensity of μ for the models. Our prior specification strategy advocates exponential priors of Exp(11) for the first example, Exp(20.3) for the second example, and Exp(22) for the third example. Note that all the examples below presume constant immigrant intensities for the models as their underlying functions for simulation. Finally,

we designate the mark space as $\mathcal{K} = [4, \infty)$ for the ETAS and semiparametric models, and $\mathcal{K} = (4, 10)$ for the nonparametric model, since the support for the mark density function is confined to $(4, 10)$, reflecting the range from "light" to "great" earthquake magnitudes.

4.1 Power-law density example

The first data set comprises 906 time points, consisting of 102 immigrants and 804 offspring, recorded within the time interval $(0, T) = (0, 5000)$. They originated from a marked Hawkes process (MHP) characterized by a ground intensity function of $\lambda_g^*(t) = 0.02 + \sum_{t_i<t} 0.47 \exp\{0.5(\kappa_i - \kappa_0)\}\text{Powlaw}(t - t_i|20, 2)$ for $t > 0$, and a mark density function of $\text{Exp}(\kappa|1, 4, 10)$, which is a truncated exponential density function with rate of 1 and support $\kappa \in [4, 10]$.

According to the prior specification strategy in Section 3.3, we chose the following priors for the nonparametric model: $\text{Lo}(2, 0.1)$ for θ with $L = 20$; $\text{Exp}(1)$ and $\text{Exp}(8)$ for b_1 and b_2; $\text{Exp}(1)$ for d with $M = 15$; $\text{Exp}(0.005)$ for c_0; $\text{Exp}(1)$ and $\text{Exp}(0.204)$ for a_β and b_β.

This example uses samples derived from the MHP, with immigrant intensity and excitation functions matching those of the ETAS model. Consequently, it is unsurprising that the parametric model exhibits an impeccable fit to the data in Figure 3. The semiparametric model, which leverages the ETAS model's parametric function for $\alpha(\kappa)$, also demonstrates an adequate alignment of estimated values with the true total offspring intensity function, albeit with a minor discrepancy in estimating $g_\kappa(x)$ near zero. The nonparametric model shows commendable efficacy in the function estimations but performs slightly worse in estimating $\alpha(\kappa)$, with a minor deviation of the posterior means from the true function values. The offspring density estimates accurately reflect the mark-invariant underlying function, despite the model's feature of mark-dependent (estimated) offspring density function.

Figure 4 presents the sensitivity analysis results for varying L and M, highlighting the

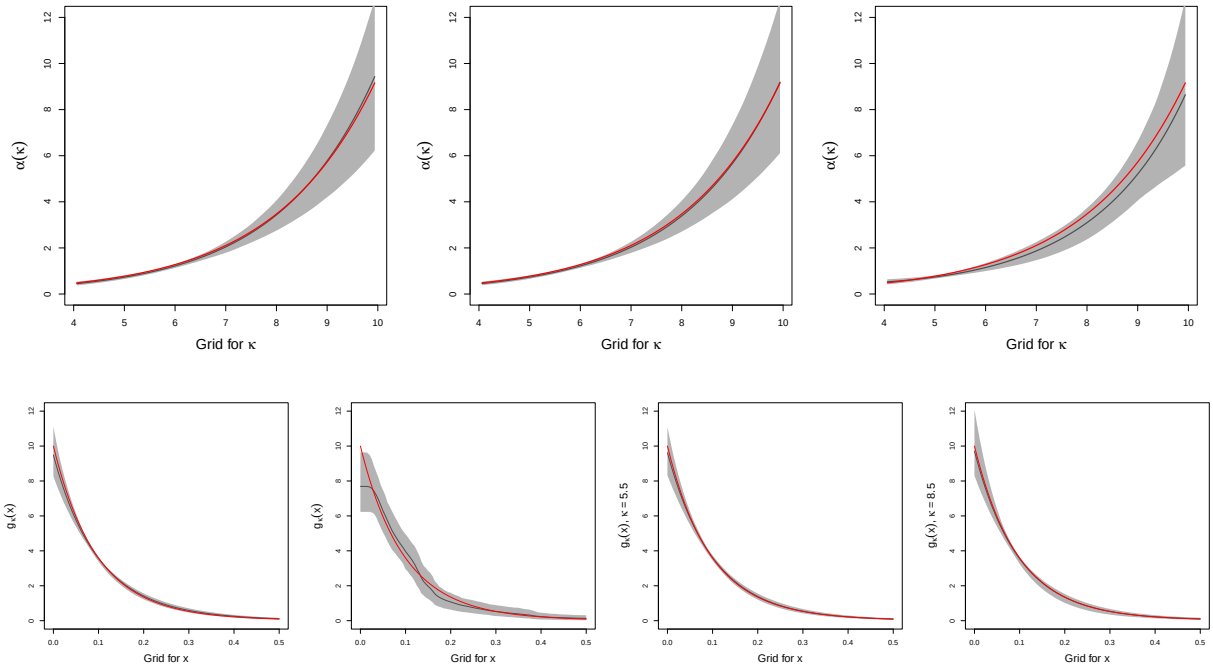

Figure 3: Power-law density example for model comparison: ETAS model in the first column, semiparametric model in the second column, and nonparametric model in the last (two) column(s). Posterior means (black line) and 95% uncertainty bands (shaded area) for $\alpha(\kappa)$ with the true intensity (red line) in the top row and for $g_\kappa(x)$ with the true density (red line) in the bottom row.

implications of boosting the number of mixture components. The top two rows juxtapose different values of M, indicating that $M = 5$ is insufficient to retrieve the true increasing trend of $\alpha(\kappa)$. In contrast, $M = 15$ produces significant improvements, with point estimates aligning much more closely with the true function. The second and third rows compare different values of L. When $L = 10$, the interval estimates for $g_\kappa(x)$ barely encompass the true underlying function, whereas those for $L = 20$ adequately capture it.

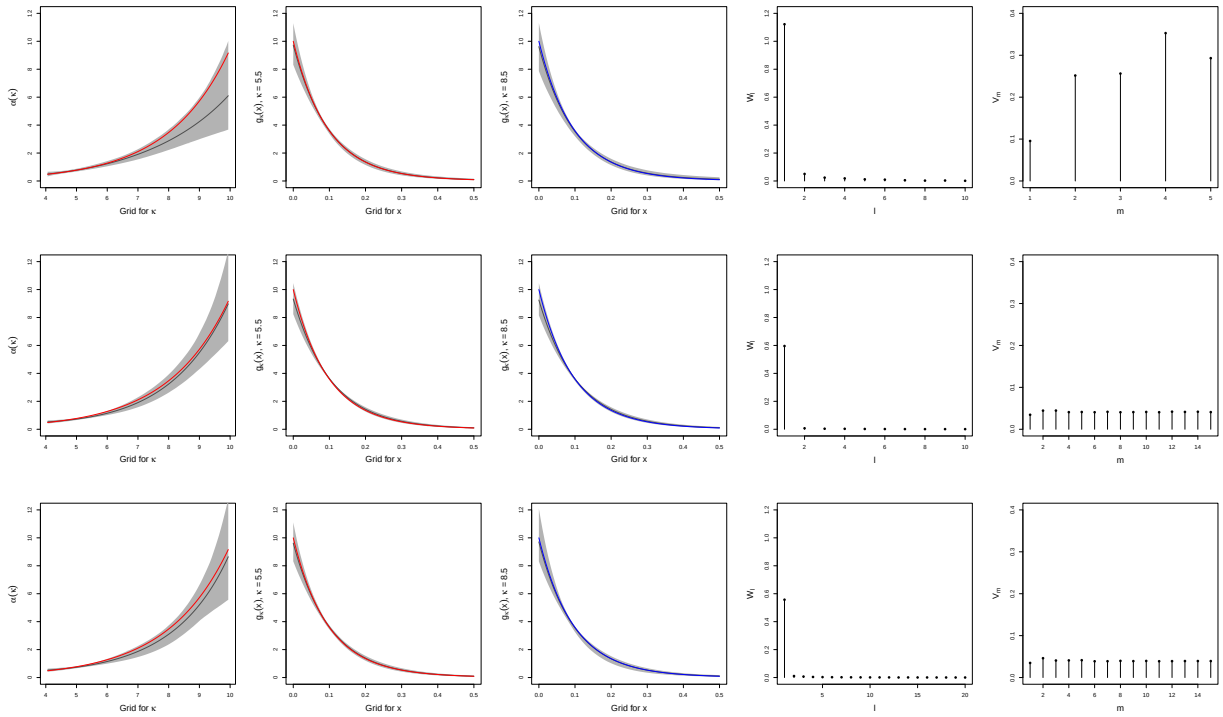

Figure 4: Sensitivity analysis for the power-law density example with varying L and M. The posterior means and 95% uncertainty bands of the total offspring intensity, the offspring density at two different marks, and the posterior means of mixture weights $W_l = \sum_{m=1}^{M} \nu_{lm}$ and $V_m = \sum_{l=1}^{L} \nu_{lm}$ are presented in the columns from left to right. The red and blue lines denote the true function values. Each row displays the results for a specific combination of L and M: from top to bottom, $(L, M) = (10, 5)$, $(10, 15)$, and $(20, 15)$.

4.2 Mark-dependent power-law density example

We produced a point pattern encompassing 493 points (48 immigrants and 445 offspring) from a MHP with ground intensity defined as $\lambda_g^*(t) = 0.01 + \sum_{t_i<t} 0.37 \exp\{0.45(\kappa_i - \kappa_0)\}\text{Powlaw}(t - t_i|5 + \kappa_i, 1)$ for $t \in (0, 5000)$ and mark density $\text{Exp}(\kappa|0.6, 4, 10)$. The

power-law offspring density function in this instance is influenced by the mark history κ_i through its shape parameter. This function reflects a potential seismic pattern: aftershocks from parent earthquakes of higher magnitude emerge sooner, with the concomitant density function more concentrated around 0.

The nonparametric model took the following priors: $\text{Lo}(2, 0.1)$ for θ with $L = 20$; $\text{Exp}(1)$ and $\text{Exp}(8)$ for b_1 and b_2; $\text{Exp}(1)$ for d with $M = 15$; $\text{Exp}(0.005)$ for c_0; $\text{Exp}(1)$ and $\text{Exp}(0.329)$ for a_β and b_β.

Due to the dependence of the underlying offspring density function on the mark, the ETAS and semiparametric models struggle to represent the true density function's behavior as κ varies, as shown in Figure 5. Conversely, the mark-dependent offspring density function of the nonparametric model accurately captures the slope variations in the true function with respect to κ. Regarding inferences about the total offspring intensity function, all models perform effectively, yielding posterior means that closely align with the authentic growing trend and uncertainty bands that fully encompass the true function.

4.3 Mixture of mark-dependent power-law densities example

This example explores a data set involving 454 points (50 immigrants and 404 offspring), generated through a MHP with a more intricate ground intensity function, $\lambda_g^*(t) = 0.01 + \sum_{t_i<t} 0.32 \exp\{0.5(\kappa_i - \kappa_0)\}[0.6\text{Powlaw}(t - t_i|10 + \kappa_i, 1) + 0.4\text{Powlaw}(t - t_i|10, 1 + \kappa_i)]$ for $t \in (0, 5000)$. The truncated exponential density $\text{Exp}(\kappa|0.6, 4, 10)$ continues to serve as the mark density function. Each power-law density delineated above exhibits mark-dependence in either the shape or the scale parameter. With a larger mark κ_i, the power-law mixture function has more probabilities near 0 and a longer tail. Consequently, as another conceivable behavior of aftershocks, a violent earthquake triggers numerous immediate aftershocks and also further shocks for weeks or even months thereafter.

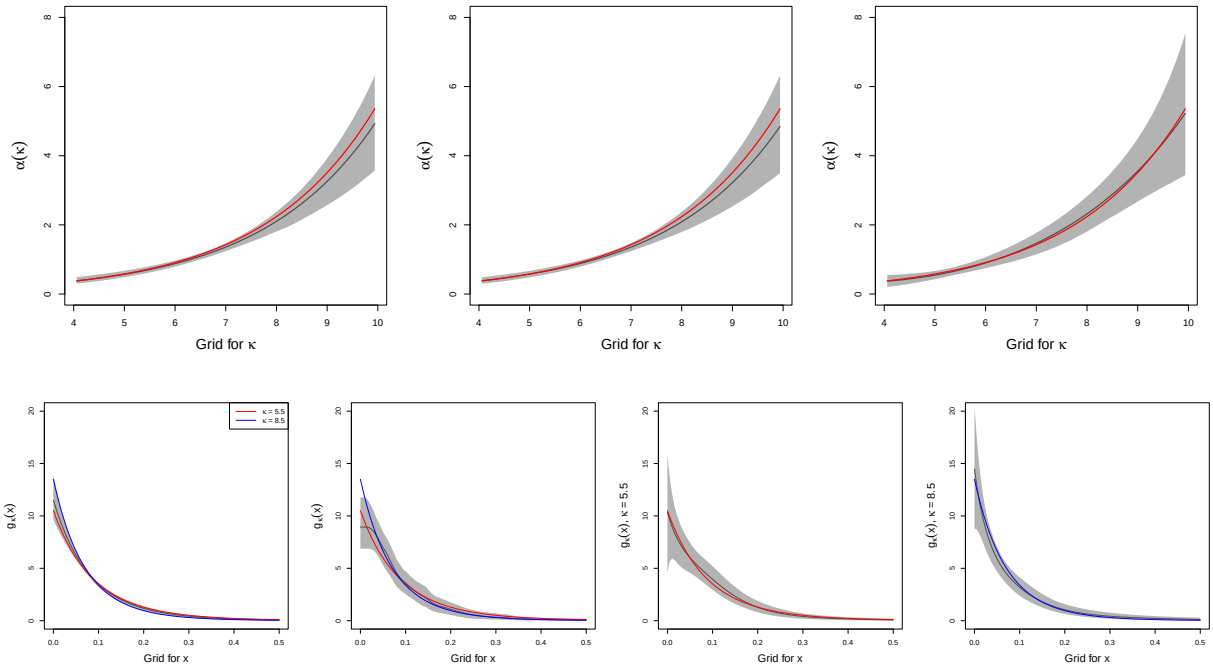

Figure 5: Mark-dependent power-law density example for model comparison: ETAS model in the first column, semiparametric model in the second column, and nonparametric model in the last (two) column(s). Posterior means (black line) and 95% uncertainty bands (shaded area) for $\alpha(\kappa)$ with the true intensity (red line) in the top row and for $g_\kappa(x)$ with the true densities (red and blue lines) in the bottom row.

For the priors of the nonparametric model, we assigned $\text{Lo}(2, 0.5)$ to θ with $L = 60$; $\text{Exp}(1)$ and $\text{Exp}(130)$ to b_1 and b_2; $\text{Exp}(1)$ to d with $M = 15$; $\text{Exp}(0.005)$ to c_0; $\text{Exp}(1)$ and $\text{Exp}(0.334)$ to a_β and b_β.

The mark-dependent underlying offspring density function poses a challenge for the ETAS and semiparametric models in capturing functional variations associated with the mark. The nonparametric model, however, excels at catching these variations: in Figure

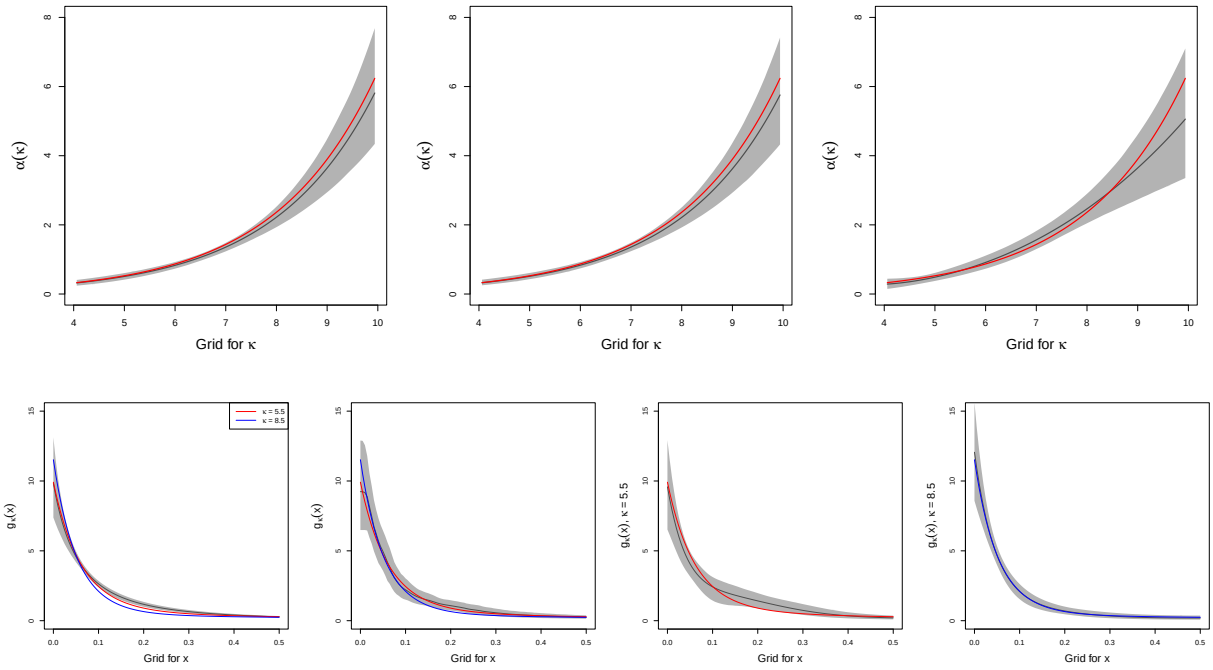

Figure 6: Mixture of mark-dependent power-law density example for model comparison: ETAS model in the first column, semiparametric model in the second column, and nonparametric model in the last (two) column(s). Posterior means (black line) and 95% uncertainty bands (shaded area) for $\alpha(\kappa)$ with the true intensity (red line) in the top row and for $g_\kappa(x)$ with the true densities (red and blue lines) in the bottom row.

6, the estimates of $g_\kappa(x)$ reveal that the density function with a larger mark has a more dramatic declining pattern, consistent with the true function. Figure 7 depicts the tail probability function of κ for a given x, derived from the offspring density function. The true tail probability increases with the mark, as expected from the underlying density function, where a larger mark indicates a heavier tail. The nonparametric model in the plot illustrates such an increasing pattern, with uncertainty bands enveloping the true functions.

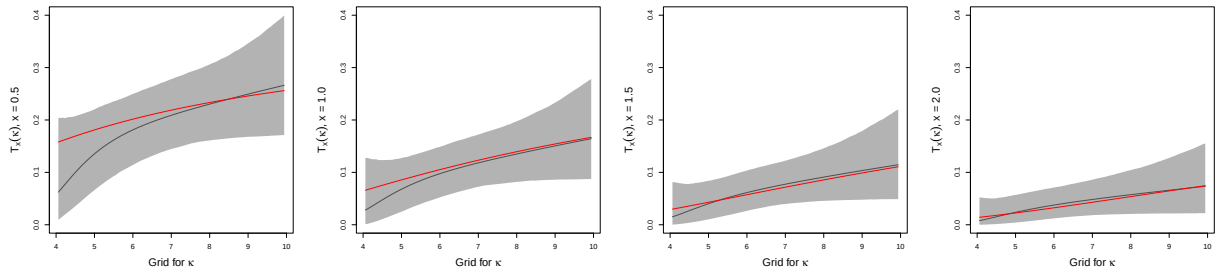

Figure 7: Posterior means and 95% uncertainty bands for the tail probability function, defined as $T_x(\kappa) = 1 - \int_0^x g_\kappa(u)du$ with four different values of x, under the nonparametric model.

We evaluate the tail probability function, $T_x(\kappa)$, at four different x values, with increasing x values expected to scale down the function. In all four cases, the model performs well, especially at elevated κ values. The relatively lower performance at small κ may relate to the reduced accuracy at $\kappa = 5.5$ compared to $\kappa = 8.5$ in the estimation of $g_\kappa(x)$ in Figure 6. It should be noted that this disparity is not an artifact of the nonparametric model; the ETAS model also fails to accurately capture the underlying function in the interval $(0.1, 0.3)$, possibly due to unrepresentative data. All three models produce acceptable estimates for the total offspring intensity function; however, the nonparametric model's posterior means underestimate the true function around $\kappa = 10$ slightly more than those of the other models.

5 Real data analysis

The section compares the ETAS (2), semiparametric (11), and nonparametric (3) models using real-world earthquake data. The data set is a catalog of earthquakes with magnitudes

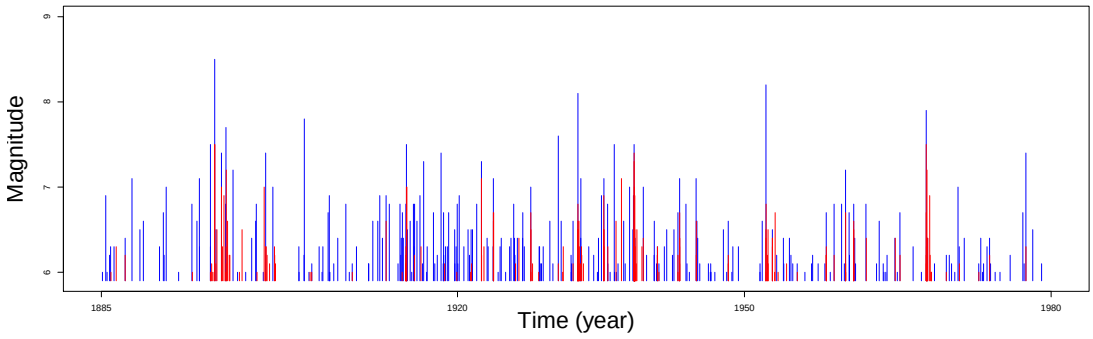

Figure 8: Japan earthquake point pattern, split into main shocks (blue) and aftershocks (red).

of six or greater that occurred in Japan and its vicinity from 1885 through 1980 (spanning approximately $34,711$ days), as given in Ogata (1988). We use 458 point observations (258 main shocks and 200 aftershocks) after removing 25 foreshocks from the original 483 observations. Figure 8 shows the earthquakes' occurrence times and magnitudes, color-coded by earthquake type: blue for main shocks and red for aftershocks. Model evaluation is accomplished based on the predictive performance (see Section 3.5 for a description of the prediction method). To do so, we first divided the data into training and test sets, using the last 30 years of data, from 1950/01/01 onward, as the test set. This division resulted in 340 and 118 observations for training and test sets, respectively. Next, we fit the models to the training data set to get posterior samples of the models' parameters, which are later used to construct the MHP intensity function for the process' realizations. Predictive performance is then assessed by comparing the posterior predictive distributions derived from the realizations under each model with the observed count (118).

Following is a list of the priors we used for each model: Exp(139) for μ, Exp(2.5) with

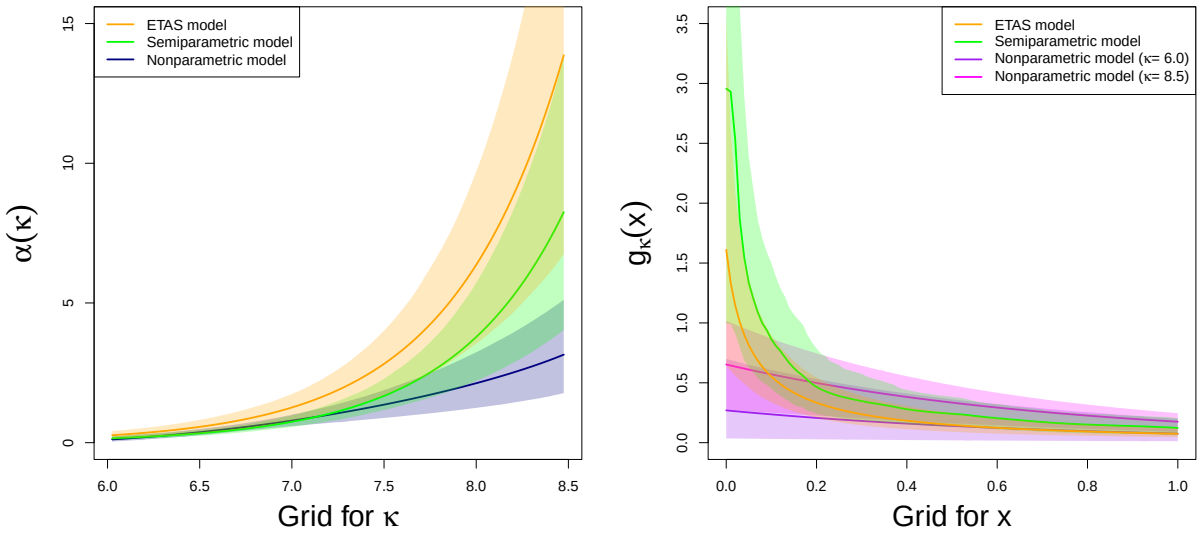

Figure 9: Posterior means (lines) and 95% uncertainty bands (shaded areas) for the total offspring intensity function (left), the offspring density function (middle), and its cumulative distribution function (right) for each model.

$a_\alpha < 1 - (\eta/\psi)$ for a_α, Exp(0.2) with $\eta < \psi$ for η, Exp(0.1) for p, Exp(0.1) for c, and Exp(0.1) for ψ for the ETAS model; Ga$(5, 0.25)$ for α_0, Exp(1) for a, Exp(1.5) for b, and $L = 200$, along with the analogous priors for μ, a_α, η, and ψ, for the semiparametric model. The nonparametric model is assigned Lo$(2, 9)$ for θ with $L = 80$, Exp(1) and Exp(3000) for b_1 and b_2, Exp(1) for d with $M = 10$, Exp(0.005) for c_0, Exp(1) and Exp(0.241) for a_β and b_β. The values of L and M were determined based on sensitivity analysis results, which are discussed later in this section. We tried not only a constant function but also an Erlang mixture (8) to model immigrant intensity. For the constant immigrant, Exp(139) for μ was used, while for the Erlang mixture, Lo$(2, 900)$ for ϕ with $J = 60$, Exp(0.1) for

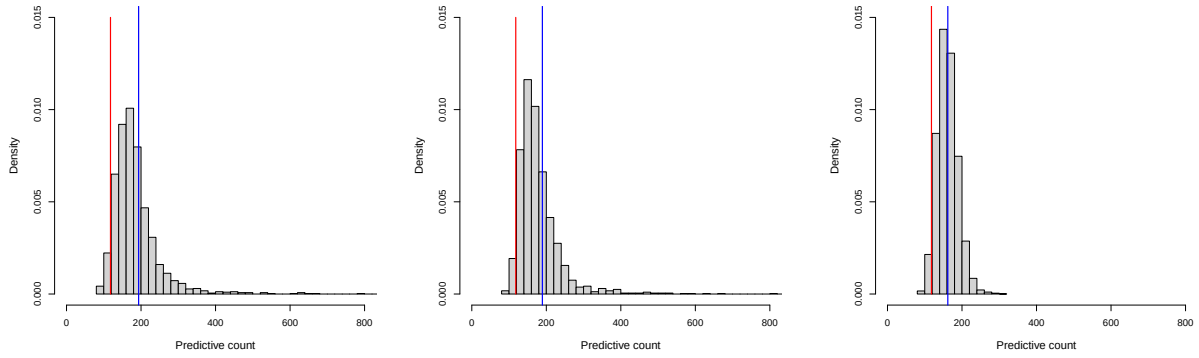

Figure 10: Histograms of posterior distributions for predictive counts under the ETAS (left), semiparametric (middle), and nonparametric (right) models, accompanied with their posterior means (blue lines) and the observed count (i.e., the test set size), $N_{\text{obs}}(B) = 118$ (red line).

e_0, and Exp(0.007) for b_{G_0} were chosen. We set the mark space to $\mathcal{K} = (5.9, 8.6)$ for the nonparametric model and $\mathcal{K} = [6, \infty)$ for the ETAS and semiparametric models, where the observed earthquake magnitude ranges from 6 to 8.5.

Figure 9 displays the posterior estimates of $\alpha(\kappa)$ and $g_\kappa(x)$. The estimated functions from the models exhibit notable discrepancies in shape, with the nonparametric model showing distinctly smaller slopes than the other models in both panels. In the $g_\kappa(x)$ plot for a day (24 hours) for x, the ETAS and semiparametric models present dramatic decreases in offspring density over time compared to the nonparametric model. Under the nonparametric model, larger magnitudes result in higher probabilities near zero, leading to shorter waiting times for subsequent shocks. Furthermore, the nonparametric model with $\kappa = 8.5$ exhibits a longer tail than the model with $\kappa = 6.0$ (not shown here); their 99.9th percentiles are 107 days and 95 days, respectively.

To evaluate the models, we apply an estimation reliability criterion, the misclassification rate, as described in **Kim and Kottas (2024)**, and the predictive performance measure introduced in Section 3.5. The misclassification rate (R) assesses the accuracy of the model in estimating branching structures and classifying points in the training set into main shocks and aftershocks. R is calculated as $R = (M_I + M_O)/n$, where $M_I = |\{i : y_i = 0,\ y_i^{true} \neq 0,\ i = 1, \ldots, n\}|$ is the immigrant misclassification count, and $M_O = |\{i : y_i \neq 0,\ y_i^{true} = 0,\ i = 1, \ldots, n\}|$ is the offspring misclassification count. Here, y_i and y_i^{true} for $i = 1, \ldots, n$ denote the estimated branching structures and observed earthquake type indices, respectively.

We found that the nonparametric model achieved the smallest posterior mean and standard deviation for R: 0.217 and 0.016, respectively, compared to 0.228 and 0.023 for the semiparametric model, and 0.266 and 0.024 for the ETAS model. Figure 10 illustrates the posterior predictive distributions under each model, with the true value of $N_{\text{obs}}(B)$ for $B \equiv [1950, 1980] \times [6, 8.5]$. The nonparametric model outperforms the others, exhibiting the smallest difference between the true count and the posterior predictive mean, as well as the lowest posterior predictive variance.

Figures 11 and 12 show the sensitivities of the key functions $\alpha(\kappa)$ and $g_\kappa(x)$ to M and L. Here, M and L represent the numbers of mixture basis components: polynomial functions $b_m(\kappa; d)$ and Erlang density functions, respectively, with each parameter primarily influencing the estimation results for $\alpha(\kappa)$ and $g_\kappa(x)$, as demonstrated in Figure 4 for the power-law density example.

In Figure 11 (or Figure 12), from left to right, M (or L) is doubled from 5 (or 40), while L (or M) is held constant. The left and middle panels exhibit notable differences in the posterior mean estimates for each function, whereas the right panels closely resemble the middle panels. Based on these observations, we ceased increasing the parameter values

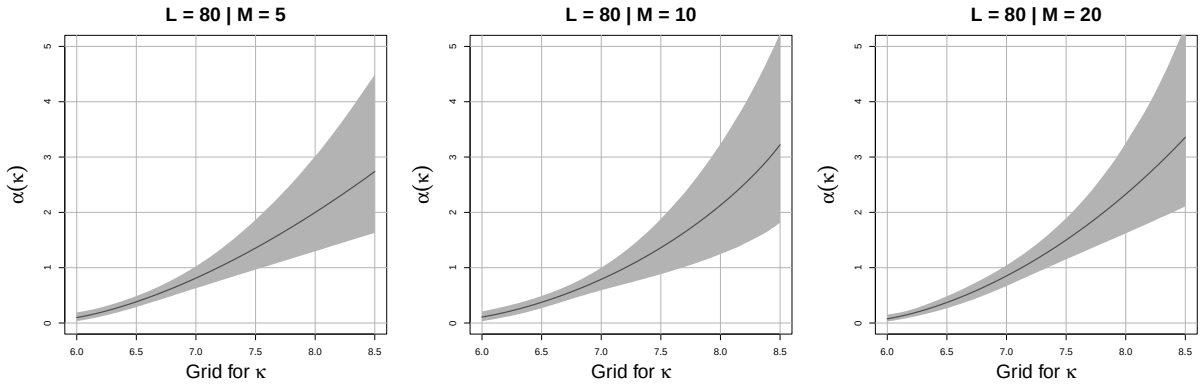

Figure 11: Sensitivity analysis of $\alpha(\kappa)$ with several choices of $M \in \{5, 10, 20\}$ for $L = 80$.

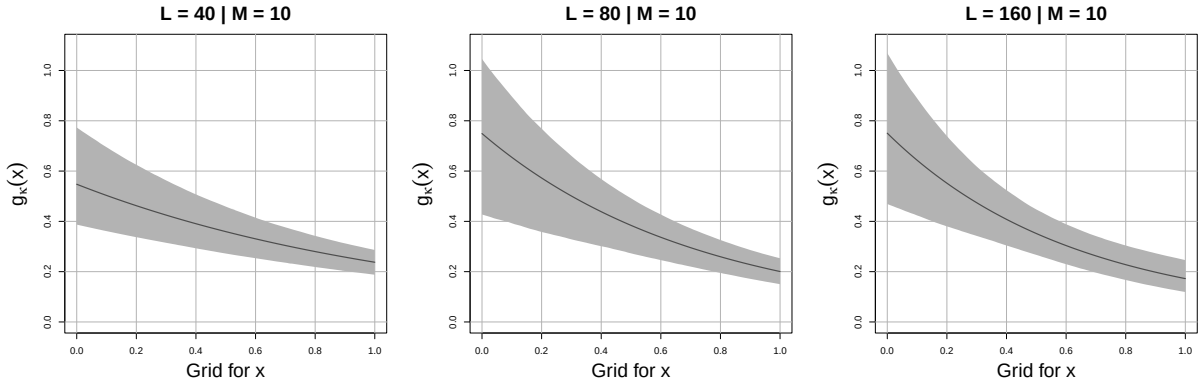

Figure 12: Sensitivity analysis of $g_\kappa(x)$, $\kappa = 7.25$ with several choices of $L \in \{40, 80, 160\}$ for $M = 10$.

and selected M and L as 10 and 80, respectively.

6 Discussion

We propose a flexible Bayesian modeling method for MHPs, specifically for the ground process conditional intensity function. The excitation function is modeled as a mixture

of basis functions, where each basis function comprises an Erlang density for time and a polynomial function for the mark. The mixture weights are represented as increments of a measure, to which a gamma process prior is assigned. For simplicity, we assume a constant function for the immigrant intensity of the ground process conditional intensity function, but it can be extended to an Erlang mixture, enabling a fully nonparametric approach to modeling the conditional intensity.

Posterior inference is conducted using a general MCMC method with the Gibbs sampler. The proposed model ensures computational efficiency, as most model parameters have closed-form posterior full conditionals. In particular, the prior conjugacy of the mixture weights is a key benefit of the model's efficient handling of the normalizing constant. Recall that the HP cluster representation expresses the HP likelihood using Poisson process likelihoods, which are notoriously difficult to handle because of the normalizing constant (e.g., see Adams et al., 2009, for the doubly intractable posterior distribution in the context of nonparametric intensity estimation for inhomogeneous Poisson processes). Our model simplifies this complexity by representing the integral part of the process likelihood for the excitation function via cumulative Erlang distribution functions multiplied by polynomial functions.

More importantly, the proposed model accommodates mark dependence in the offspring density function. The ETAS model in Ogata (1988) assumes that the excitation function factors into two components, $\alpha(\kappa)$ and $g(x)$. In contrast, our model introduces a mark-dependent offspring density function, $g_\kappa(x)$, enabling the exploration of aftershock occurrence patterns, namely the density of waiting times between an aftershock and its predecessor, as a function of the predecessor's magnitude. Indeed, we applied the model to a Japan earthquake data set and identified two distinct patterns in the density function for aftershocks. Earthquakes with larger magnitudes tend to have aftershocks that occur

closer in time, resulting in a more concentrated density around zero. Conversely, larger-magnitude earthquakes are also associated with a few aftershocks occurring relatively long after their antecedents, resulting in a longer-tailed density for aftershocks.

The models were evaluated using a predictive performance measure and a branching structure classification criterion, referred to as the misclassification rate. In a case study using the Japan earthquake dataset, our model outperforms the ETAS model and a semi-parametric alternative in both prediction and classification, while also offering the new findings discussed above.

While the proposed model has many positive attributes, it also has a couple of downsides. Selecting the number of mixture components, L and M, relies on sensitivity analysis, necessitating multiple MCMC runs with increasing L and M values. Additionally, unlike other models, the proposed model requires an upper bound for the mark space. Thus, the range of observed marks is needed unless a theoretical basis exists for a specific bounded mark space.

Our model is based on the unpredictable mark assumption, where the mark density function is independent of the current time and the MHP (time and mark) history. A potential direction for future research involves extending the model to incorporate a more general mark density function that depends on the current time and the process history. Another promising direction is to extend the model to spatio-temporal Hawkes processes, which may involve replacing the basis function with a product of a function for the mark history and several Erlang density functions for time and space. While this paper focuses on continuous marks, adapting the model to accommodate various types of marks would further enhance its practical applicability.

References

Adams, R. P., Murray, I., and MacKay, D. J. (2009). Tractable nonparametric bayesian inference in poisson processes with gaussian process intensities. In *Proceedings of the 26th annual international conference on machine learning*, pages 9–16.

Balderama, E., Schoenberg, F. P., Murray, E., and Rundel, P. W. (2012). Application of branching models in the study of invasive species. *Journal of the American Statistical Association*, 107(498):467–476.

Da Fonseca, J. and Zaatour, R. (2014). Hawkes process: Fast calibration, application to trade clustering, and diffusive limit. *Journal of Futures Markets*, 34(6):548–579.

Daley, D. J., Vere-Jones, D., et al. (2003). *An introduction to the theory of point processes: volume I: elementary theory and methods*. Springer.

Hardiman, S. J., Bercot, N., and Bouchaud, J.-P. (2013). Critical reflexivity in financial markets: a hawkes process analysis. *The European Physical Journal B*, 86(10):1–9.

Hawkes, A. G. (1971a). spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90.

Hawkes, A. G. (1971b). point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 33(3):438–443.

Kagan, Y. Y. (1991). Likelihood analysis of earthquake catalogues. *Geophysical journal international*, 106(1):135–148.

Kim, H. and Kottas, A. (2022). Erlang mixture modeling for poisson process intensities. *Statistics and Computing*, 32:1–15.

Kumazawa, T. and Ogata, Y. (2014). Nonstationary etas models for nonstandard earthquakes. *The Annals of Applied Statistics*, 8(3):1825–1852.

Laub, P. J., Taimre, T., and Pollett, P. K. (2015). Hawkes processes. *arXiv preprint arXiv:1507.02822*.

Meyer, S., Elias, J., and Höhle, M. (2012). A space–time conditional intensity model for invasive meningococcal disease occurrence. *Biometrics*, 68(2):607–616.

Mohler, G. (2014). Marked point process hotspot maps for homicide and gun crime prediction in chicago. *International Journal of Forecasting*, 30(3):491–497.

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., and Tita, G. E. (2011). Self-exciting point process modeling of crime. *Journal of the american statistical association*, 106(493):100–108.

Musmeci, F. and Vere-Jones, D. (1992). A space-time clustering model for historical earthquakes. *Annals of the Institute of Statistical Mathematics*, 44(1):1–11.

Nandan, S., Ouillon, G., Wiemer, S., and Sornette, D. (2017). Objective estimation of spatially variable parameters of epidemic type aftershock sequence model: Application to california. *Journal of Geophysical Research: Solid Earth*, 122(7):5118–5143.

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical association*, 83(401):9–27.

Ogata, Y. (1998). Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics*, 50(2):379–402.

Ogata, Y., Katsura, K., and Tanemura, M. (2003). Modelling heterogeneous space–time occurrences of earthquakes and its residual analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52(4):499–509.

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

Rambaldi, M., Pennesi, P., and Lillo, F. (2015). Modeling foreign exchange market activity around macroeconomic news: Hawkes-process approach. *Physical Review E*, 91(1):012819.

Rasmussen, J. G. (2013). Bayesian inference for hawkes processes. *Methodology and Computing in Applied Probability*, 15(3):623–642.

Reasenberg, P. A., Poland, C. D., and Engineers, D. (1999). Earthquake aftershocks—entering damaged buildings.

Reinhart, A. (2018). A review of self-exciting spatio-temporal point processes and their applications. *Statistical Science*, 33(3):299–318.

Ross, G. J. (2021). Bayesian estimation of the etas model for earthquake occurrences. *Bulletin of the Seismological Society of America*, 111(3):1473–1480.

Veen, A. and Schoenberg, F. P. (2008). Estimation of space–time branching process models in seismology using an em–type algorithm. *Journal of the American Statistical Association*, 103(482):614–624.

Zhuang, J., Ogata, Y., and Vere-Jones, D. (2002). Stochastic declustering of space-time earthquake occurrences. *Journal of the American Statistical Association*, 97(458):369–380.