Nonparametric Bayesian Modeling and Inference for Hawkes Processes

Hyotae Kim* Athanasios Kottas†

Abstract

We propose a nonparametric Bayesian modeling and inference framework for temporal Hawkes processes. The objective is to broaden the inferential scope for this practically important class of point processes by exploring flexible models for its conditional intensity function. In particular, we develop different types of nonparametric prior models for the offspring density and for the immigrant intensity, the two functions that define the Hawkes process conditional intensity. The prior models are carefully constructed such that, along with the Hawkes process branching structure, they enable efficient handling of the complex likelihood normalizing terms in implementation of inference. The modeling framework includes a prior for general decreasing offspring densities, which is relevant to applications of Hawkes processes in earthquake modeling. We discuss prior specification for model hyperparameters, design posterior simulation algorithms, and develop inference techniques for model assessment and for prediction. The modeling approach is studied empirically using several synthetic data examples, and is illustrated through reexamination of a data set involving earthquake occurrences.

Keywords: Bayesian nonparametrics; Dirichlet process; Erlang mixtures; Gamma process; Markov chain Monte Carlo; Non-homogeneous Poisson process.

1 Introduction

We develop nonparametric Bayesian (NPB) modeling and inference methods for temporal Hawkes processes. The Hawkes process (HP) (Hawkes, 1971a,b) is a versatile stochastic model for point processes built from structured conditional intensity functions that model *self-excitation*, i.e., the property that the occurrence of an event increases the rate of occurrence for some period of time in the future. Such a structure yields point patterns with events that are naturally clustered in time. For example, earthquake occurrences are grouped into clusters consisting of a main shock and subsequent shocks (aftershocks). Indeed, including extensions to incorporate marks and information on spatial location, HPs have found applications in seismology (e.g., Ogata, 1988, 1998; Zhuang et al., 2002; Veen and Schoenberg, 2008; Fox et al., 2016; Molkenthin et al., 2022), as well as biology and epidemiology (e.g., Balderama et al., 2012; Holbrook et al., 2022), finance (e.g., Hawkes, 2018), and crime data modeling (e.g., Mohler et al., 2011). Details on the theory of HPs are given in Daley and Vere-Jones (2003, 2007). Reviews of HP models and their applications can be found in, e.g., Reinhart (2018), Laub et al. (2021), and Bernabeu et al. (2025).

*Department of Biostatistics & Bioinformatics, Duke University. Email: hyotae.kim@duke.edu
†Department of Statistics, University of California, Santa Cruz. Email: thanos@soe.ucsc.edu

The general form of the HP conditional intensity is given by

$$\lambda^*(t) \equiv \lambda(t \mid \mathcal{H}(t)) = \mu(t) + \sum_{t_j<t} h(t-t_j), \quad t \in \mathbb{R}^+, \tag{1}$$

where $\mathcal{H}(t)$ is the point process history up to time t (with t_j denoting a generic time event), $\mu : \mathbb{R}^+ \to \mathbb{R}^+$ is the background (immigrant) intensity function, and $h : \mathbb{R}^+ \to \mathbb{R}^+$ is the excitation function (offspring intensity), which accounts for the effect of previous events on the current intensity. The excitation function takes time lags as its arguments (denoted generically by x), and it must be integrable over $\mathbb{R}^+$. It can thus be equivalently represented in terms of the branching ratio, $\gamma = \int_0^\infty h(x)\,\mathrm{d}x$, and the offspring density $f(x) = h(x)/\gamma$, for $x \in \mathbb{R}^+$. The original definition of the HP, and several of its applications, focus on immigrant rate which is constant in time ($\mu(t) \equiv \mu > 0$), in which case the HP is stationary provided $\gamma \in (0, 1)$. Our methodology handles the general case of a time-dependent immigrant intensity function.

The *immigrant/offspring* terminology originates from the HP cluster representation (Hawkes and Oakes, 1974), which is key for inference methods. Consider a HP realization, $\{0 < t_1 < ... < t_n < T\}$, observed in time window $(0, T]$. The branching structure for the point pattern can be described by latent variables $\boldsymbol{y} = \{y_i : i = 1, ..., n\}$, such that $y_i = 0$ if t_i is an immigrant point, and $y_i = j$ if point t_i is the offspring of t_j. Hence, given $\boldsymbol{y}$, the HP point pattern is partitioned into the set of immigrants, $I = \{t_i : y_i = 0\}$, and sets of offspring, $O_j = \{t_i : y_i = j\}$, where O_j collects all offspring of t_j. Moreover, conditioning on the branching structure $\boldsymbol{y}$, the HP can be constructed as the superposition of independent Poisson processes corresponding to I and the O_j, with intensity $\mu(t)$ and $h(t - t_j) = \gamma\, f(t - t_j)$ for I and O_j, respectively.

The HP likelihood based on observed point pattern $\{0 < t_1 < ... < t_n < T\}$ is given by $\exp\left(-\int_0^T \lambda^*(u)\,\mathrm{d}u\right) \prod_{i=1}^n \lambda^*(t_i)$. Owing to the additive form of the conditional intensity, it is challenging to work with the likelihood, even under simple parametric offspring intensities. The HP cluster representation provides an effective alternative. Augmented with the branching structure latent variables $\boldsymbol{y}$, the HP likelihood can be expressed as:

$$\begin{aligned} &\exp\left(-\int_0^T \mu(u)\,\mathrm{d}u\right) \left\{\prod_{\{i:\, t_i \in I\}} \mu(t_i)\right\} \\ &\times \exp\left(-\sum_{j=1}^n \int_0^{T-t_j} h(u)\,\mathrm{d}u\right) \left\{\prod_{\{i:\, t_i \in O\}} h(t_i - t_{y_i})\right\}, \end{aligned} \tag{2}$$

where $O = \cup_{j=1}^n O_j$ is the set of all offspring points. For all j with $O_j = \emptyset$, the second exponential term incorporates the probability that point t_j has no offspring in $(0, T]$. The distribution for the latent variables, which yields the HP likelihood upon marginalization over $\boldsymbol{y}$, is $\delta_0(y_1) \prod_{i=2}^n \mathrm{discUnif}(y_i \mid 0, 1, \ldots, i-1)$, where $\mathrm{discUnif}(\mid 0, 1, \ldots, i-1)$ denotes the discrete uniform distribution supported on $\{0, 1, \ldots, i-1\}$, and $\delta_s(\cdot)$ denotes the point mass at s.

We refer to Laub et al. (2021) for discussion of classical inference for HPs, which is based mainly on maximum likelihood estimation. Bayesian inference for parametric HP models is discussed in, e.g., Rasmussen (2013) and Ross (2021); the former compares Markov chain Monte Carlo (MCMC) posterior simulation based on either the likelihood defined directly through the conditional intensity or the augmented likelihood that utilizes the branching structure. The more recent literature also

includes semiparametric Bayesian modeling approaches, discussed in Section 6 relative to our NPB methods.

In this article, we develop a general NPB modeling and inference framework for temporal Hawkes processes. Our key motivation is to expand the scope of modeling methods for inference and prediction under this practically important class of self-exciting point processes. We develop different NPB models, based on mixture structures/basis representations, for the components of the HP conditional intensity, including prior models for the immigrant intensity, excitation function, and offspring density. Posterior simulation is facilitated by the HP cluster representation, and, importantly, by the structure of the proposed prior models. The model construction seeks to achieve a balance between flexible shapes for the HP intensity functions and computationally tractable implementation. Indeed, the MCMC simulation methods yield full posterior inference and prediction, without approximations to the HP likelihood or to the posterior distribution. As part of the modeling framework, we study a nonparametric prior that supports general decreasing offspring densities. This option is relevant for HP modeling for analysis of earthquake occurrences. It incorporates the constraint of decreasing aftershock densities, supported by seismology theory (e.g., Ogata, 1988), while allowing for general tail behavior, including either exponential or polynomial offspring density tails (corresponding to the two most commonly used parametric choices of exponential or Lomax densities). Another component of the proposed framework involves fully NPB models that combine the prior for the immigrant intensity with either of the priors for the excitation function or offspring density. As demonstrated with synthetic data examples, the fully NPB models can uncover non-standard shapes for the different components of the HP conditional intensity. In addition to MCMC posterior simulation algorithms, we develop inference techniques for model assessment and for prediction.

The rest of the paper is organized as follows. In Section 2, we present the different NPB models for the HP conditional intensity function. Technical details on prior specification and posterior simulation are deferred to the Supplementary Material. Section 3 develops methods for inference beyond model estimation, including model assessment techniques, and predictive inference. In Section 4, we present a subset of the results from a comprehensive simulation (with additional results given in the Supplementary Material). An application involving earthquake data analysis follows in Section 5. Finally, Section 6 concludes with a summary and discussion.

2 Modeling methods for the HP intensity function

This section develops the prior models for the two components of the HP conditional intensity, the immigrant intensity function (Section 2.1) and the excitation function (Section 2.2), with fully NPB models emerging from their combination (Section 2.3).

The model for the immigrant intensity and one of the models for the excitation function are built from weighted combinations of Erlang densities. Relevant to those modeling approaches is the Erlang mixture model for density estimation, which we review here. Denote by $\text{ga}(a, b)$ the gamma density (or distribution, depending on the context) with mean a/b, and consider the following representation for density f on $\mathbb{R}^+$:

$$f(t) = \sum_{l=1}^{L} \psi_l \, \text{ga}(t \mid l, \theta^{-1}), \quad t \in \mathbb{R}^+ \tag{3}$$

where $\psi_l = F(l\theta) - F((l-1)\theta)$, for $l = 1, \ldots, L-1$, and $\psi_L = 1 - F((L-1)\theta)$, for a distribution function F on $\mathbb{R}^+$. This is a structured mixture for which the Erlang densities play the role of basis densities with increasing means and a single parameter θ controlling their dispersion. The relative magnitude of the weights assigned to the basis densities is driven by the shape of distribution F. Under this formulation for the weights, as $\theta \to 0$ and $L \to \infty$, the Erlang mixture density converges pointwise to the density of F (e.g., Butzer, 1954; Lee and Lin, 2010). The convergence result for the Erlang mixture structure provides the foundation for NPB modeling, treating F as a random distribution. There has been some relevant work on NPB modeling for renewal processes (Xiao et al., 2021) and for survival analysis (Li et al., 2024), using in both cases a Dirichlet process (DP) prior (Ferguson, 1973) for F.

2.1 Modeling the immigrant intensity function

Temporal HPs are typically applied with immigrant rate μ which is constant in time. To achieve general inference and prediction, we seek a flexible prior model for immigrant intensity functions $\mu(t)$. In particular, in earthquake modeling, the immigrant intensity explains the dynamics of main shock occurrences, and it is thus natural to expect improved prediction if we can effectively estimate time-dependent immigrant intensities. Indeed, this is illustrated with the earthquake data analysis of Section 5.

Given the branching structure, the immigrant intensity plays the role of the non-homogeneous Poisson process (NHPP) intensity for the immigrants point pattern. Existing methods for temporal NHPPs typically build from Gaussian process (GP) priors for (transformations of) the NHPP intensity (e.g., Adams et al., 2009), or DP mixture priors for the NHPP density over the observation window (e.g., Taddy and Kottas, 2012). As the prior for the immigrant intensity $\mu(t)$, we employ the NPB model developed in Kim and Kottas (2022), which extends the Erlang mixture model in (3) for intensity estimation. Relative to GP-based methods, this model is more computationally tractable. Compared with DP mixture models, it offers a more parsimonious representation for the intensity over $\mathbb{R}^+$, rather than only over $(0, T]$, which is important for predictive inference based on the estimated HP conditional intensity (Section 3.2).

The Erlang mixture model for the HP immigrant intensity is given by:

$$\begin{aligned} \mu(t) &= \sum_{l=1}^{L} \nu_l \, \text{ga}(t \mid l, \theta^{-1}), \quad t \in \mathbb{R}^+ \\ \nu_l &= G(l\theta) - G((l-1)\theta), \;\; l = 1, \ldots, L; \quad G \mid c_0, G_0 \sim \mathcal{G}(G_0, c_0). \end{aligned} \tag{4}$$

Here, $\mathcal{G}(G_0, c_0)$ denotes the gamma process with precision parameter $c_0 > 0$, and centering cumulative intensity G_0, such that $\text{E}(G(t) \mid G_0) = G_0(t)$ and $\text{Var}(G(t) \mid c_0, G_0) = G_0(t)/c_0$ (Kalbfleisch, 1978). We set $G_0(t) = t/b_{G_0}$, i.e., the cumulative hazard of the exponential distribution with scale parameter b_{G_0}. Hyperpriors are placed on c_0, b_{G_0}, and θ. Owing to the independent increments of the gamma process, the weights ν_l are independently gamma distributed. In particular, $\nu_l \mid \theta, c_0, b_{G_0} \overset{ind.}{\sim} \text{ga}(c_0\, \theta\, b_{G_0}^{-1}, c_0)$, under the exponential cumulative hazard for G_0.

The model for the immigrant intensity builds from a weighted combination of Erlang densities with common scale parameter θ and (specified) integer shape parameters. The weights are constructed through increments of random cumulative intensity function G. In the context of NHPPs, Kim and Kottas (2022) present a convergence result that extends the one for Erlang mixture densities: as

$\theta \to 0$ and $L \to \infty$, $\mu(t) = \sum_{l=1}^{L} \{G(l\theta) - G((l-1)\theta)\}\, \mathrm{ga}(t \mid l, \theta^{-1})$ converges pointwise to the NHPP intensity that corresponds to cumulative intensity G (provided $G(t) = O(t^q)$, as $t \to \infty$, for some $q > 0$). This result provides theoretical support for the model structure in (4) and motivates the nonparametric prior for G. The gamma process prior allows flexible cumulative intensity realizations G with their effective support concentrated on different time intervals, resulting in weights ν_l that concentrate on different subsets of the Erlang basis densities. The precision parameter c_0 controls the variability of realizations G around G_0, and thus the effective mixture weights. In general, a smaller value for c_0 implies a smaller number of practically non-zero weights. Moreover, smaller values of θ yield more variable, typically multimodal shapes for the intensity.

Note that we are using a parsimonious mixture formulation with the Erlang densities defined in terms of a single parameter, and playing the role of basis functions in the model representation for the immigrant intensity. Working with a specified, sufficiently large number of basis densities, the nonparametric prior for G is key in the selection of the subset of them to be used in the representation of the intensity. An approach to specify L and the priors for c_0, b_{G_0}, and θ is discussed in the Supplementary Material.

Let $\{0 < t_1 < ... < t_n < T\}$ be the point pattern observed in time window $(0, T]$. The model is implemented using the HP cluster representation, based on the branching structure variables $\boldsymbol{y}$, along with additional latent variables $\boldsymbol{\xi} = \{\xi_i : i \in I\}$, where ξ_i indicates the Erlang basis density to which immigrant time point $t_i \in I$ is assigned. Consider a generic excitation function h. Then, building from the HP augmented likelihood in (2), the first stage of the hierarchical model for the observed point pattern, given $\boldsymbol{y}$ and $\boldsymbol{\xi}$, can be written as

$$\begin{aligned} \exp\left(-\sum_{l=1}^{L} \nu_l \int_0^T \mathrm{ga}(u \mid l, \theta^{-1})\, \mathrm{d}u\right) \left\{\prod_{\{i:\, t_i \in I\}} \left(\textstyle\sum_{k=1}^{L} \nu_k\right) \mathrm{ga}(t_i \mid \xi_i, \theta^{-1})\right\} \\ \times \exp\left(-\sum_{j=1}^{n} \int_0^{T-t_j} h(u)\, \mathrm{d}u\right) \left\{\prod_{\{i:\, t_i \in O\}} h(t_i - t_{y_i})\right\} \end{aligned} \tag{5}$$

where the ξ_i are i.i.d., given $\boldsymbol{\nu} = \{\nu_l : l = 1, ..., L\}$, with ξ_i taking value l with probability $\nu_l / (\sum_{k=1}^{L} \nu_k)$, for $l = 1, ..., L$. Note that, marginalizing over the ξ_i, the second term in (5) becomes $\prod_{\{i:\, t_i \in I\}} \left\{\sum_{l=1}^{L} \nu_l\, \mathrm{ga}(t_i \mid l, \theta^{-1})\right\} = \prod_{\{i:\, t_i \in I\}} \mu(t_i)$.

The Supplementary Material includes MCMC posterior simulation details for the immigrant intensity model parameters, $(\boldsymbol{\nu}, \theta, c_0, b_{G_0})$, and for latent variables $\boldsymbol{y}$ and $\boldsymbol{\xi}$. A key feature of the Erlang mixture model structure is that the weights ν_l can be sampled independently from gamma posterior full conditional distributions.

2.2 Prior probability models for the excitation function

Here, we turn our attention to modeling the excitation function. Regarding estimation, this is a challenging task, even under the augmented likelihood in (2). The challenge arises from likelihood component $\exp\left(-\sum_{j=1}^{n} \int_0^{T-t_j} h(u)\, \mathrm{d}u\right)$, which generally includes in a complicated fashion all model parameters for the excitation function.

We develop two models inspired by the direct connection between the excitation function h and the offspring density f, i.e., $f(x) = h(x)/\gamma$, where $\gamma = \int_0^{\infty} h(x)\, \mathrm{d}x$ is the branching ratio. The

first model uses a weighted combination of Erlang densities to represent the excitation function, with weights defined such that the implied prior for the offspring density is the Erlang mixture model for density estimation in (3). Since the excitation function and offspring density have the same functional shape, this modeling approach allows for general excitation functions. Moreover, it results in efficient posterior simulation, since, as for the model of Section 2.1, the mixture weights can be updated through gamma distributions.

Under the second approach, we model directly the offspring density with a nonparametric uniform mixture structure that represents all decreasing densities on $\mathbb{R}^+$. The constraint of a decreasing offspring density is motivated by the two parametric forms that are most commonly used, i.e., the exponential or Lomax (power-law) offspring density. We use $\text{Lo}(p, c)$ to denote the Lomax distribution with shape parameter $p > 0$, scale parameter $c > 0$, and density $f(x \mid p, c) = p\, c^p/(c + x)^{p+1}$, for $x > 0$.

The exponential density is typically chosen for convenience as it provides a simple parametric form for estimation, as well as analytical results for HP functionals, dating back to the study by Hawkes (1971a). The Lomax density is, to our knowledge, the only parametric form that can be potentially justified in the context of a specific scientific area, in particular, by *Omori's law* for the decay over time of the frequency of earthquake aftershocks (e.g., Ogata, 1988). From a modeling perspective, the assumption of a decreasing offspring density may be viewed as natural in other applications, given the role of the excitation function in the generation of HP point patterns. Hence, this modeling approach offers flexibility with respect to the shape and tail behavior of the offspring density; note that the two most commonly used parametric models have different tails (exponential vs polynomial).

A consideration for any excitation function model involves the branching ratio γ, which controls whether the process explodes or not. In the context of the HP branching structure, if $\gamma \geq 1$, an immigrant generates infinitely many offspring on average, whereas, when $\gamma \in (0, 1)$, each cluster size is (almost surely) finite, with expected size $1/(1 - \gamma)$ (e.g., Hawkes and Oakes, 1974; Daley and Vere-Jones, 2003). The stability condition, $\gamma \in (0, 1)$, can be incorporated under both prior models presented next.

2.2.1 Erlang mixture prior model for the excitation function

We model the excitation function as a weighted combination of Erlang densities, similar in spirit with the prior model in (4) for the immigrant intensity. However, we adjust the weights construction to account for the property that the excitation function is integrable over $\mathbb{R}^+$. The key idea is to define the weights for the excitation function representation such that the corresponding offspring density admits the prior model in (3). This can be accomplished using the DP definition, and the constructive definition of the Dirichlet distribution through normalization of independent gamma variables.

More specifically, the Erlang mixture model for the HP excitation function:

$$\begin{aligned} h(x) &= \sum_{k=1}^{K} \tilde{\nu}_k \, \text{ga}(x \mid k, \phi^{-1}), \quad x \in \mathbb{R}^+ \\ \tilde{\nu}_k \mid \phi, b_{F_0} &\overset{ind.}{\sim} \text{ga}(\alpha_0 A_k, \eta), \;\; k = 1, \ldots, K, \end{aligned} \tag{6}$$

where $A_k = F_0(k\phi) - F_0((k-1)\phi)$, for $k = 1, \ldots, K-1$, $A_K = 1 - F_0((K-1)\phi)$, $\alpha_0 > 0$, and F_0 is

a distribution function on $\mathbb{R}^+$. We take F_0 to be the exponential distribution with scale parameter b_{F_0}.

Under the prior model in (6), the branching ratio is given by $\gamma = \sum_{k=1}^{K} \tilde{\nu}_k$, and thus the offspring density can be written as

$$f(x) \;=\; \gamma^{-1} h(x) \;=\; \sum_{k=1}^{K} \psi_k \, \mathrm{ga}(x \mid k, \phi^{-1}), \quad x \in \mathbb{R}^+ \tag{7}$$

where $\psi_k = \tilde{\nu}_k / \sum_{k'=1}^{K} \tilde{\nu}_{k'}$, for $k = 1, ..., K$. The vector of weights, $(\psi_1, \ldots, \psi_K)$, follows a Dirichlet prior distribution with parameters $(\alpha_0 A_1, \ldots, \alpha_0 A_K)$. Hence, based on the DP definition, we can equivalently write $\psi_k = F(k\phi) - F((k-1)\phi)$, for $k = 1, \ldots, K-1$, and $\psi_K = 1 - F((K-1)\phi)$, where F is a random distribution function on $\mathbb{R}^+$ that arises from a $\mathrm{DP}(\alpha_0, F_0)$, that is, a DP prior with total mass parameter α_0 and centering distribution F_0. Since the aforementioned Dirichlet distribution defines the DP finite dimensional distributions, this connection between the two formulations for the weights ψ_k holds true for any K. Consequently, the implied prior model for the offspring density is the DP-Erlang mixture model in (3).

Note that the model in (6) implies the $\mathrm{ga}(\alpha_0, \eta)$ prior distribution for the branching ratio. Although the stability condition, $\gamma \in (0, 1)$, cannot be enforced with probability 1, parameters α_0 and η can be specified such that the condition is supported with high prior probability. Note also the practical utility of allowing a general rate parameter η in the prior for the weights $\tilde{\nu}_k$. Even though $\eta = 1$ suffices for the connection with the DP prior discussed above, by allowing for general η values we can control both the DP total mass parameter α_0, and the prior for the branching ratio.

A useful model property involves the prior expectation for the offspring density. As shown in the Supplementary Material, the limit, as $K \to \infty$, of $\mathrm{E}(f(x) \mid \phi, b_{F_0})$ is the exponential density with rate parameter $\phi^{-1}\{1 - \exp(-\phi/b_{F_0})\}$. The accuracy of the approximation depends on ϕ; e.g., $\mathrm{E}(f(x) \mid \phi, b_{F_0})$ becomes practically identical to the exponential density for $K = 100$ ($K = 10$) when $\phi = 0.1$ ($\phi = 1$). This result provides a connection with a common choice for the offspring density in parametric HP models, and it also facilitates prior specification. The Supplementary Material discusses an approach to specify K, α_0 and η, and the hyperpriors for ϕ and b_{F_0}.

Regarding posterior simulation given point pattern $\{0 < t_1 < ... < t_n < T\}$ observed in time window $(0, T]$, we use again the branching structure variables $\boldsymbol{y}$, along with basis configuration latent variables. We again denote the latter variables by $\boldsymbol{\xi} = \{\xi_i : i \in O\}$, where now ξ_i indicates the Erlang basis density of the excitation function model in (6) to which offspring time point $t_i \in O$ is assigned. The ξ_i are i.i.d., given $\{\tilde{\nu}_k : k = 1, ..., K\}$, with $\xi_i = k$ with probability $\tilde{\nu}_k / (\sum_{k'=1}^{K} \tilde{\nu}_{k'})$, for $k = 1, ..., K$. Using the HP augmented likelihood in (2) with a generic immigrant intensity function μ, the first stage of the hierarchical model for the observed point pattern, given $\boldsymbol{y}$ and $\boldsymbol{\xi}$, can be expressed as:

$$\begin{aligned} &\exp\left(-\int_0^T \mu(u)\,\mathrm{d}u\right) \left\{ \prod_{\{i:\, t_i \in I\}} \mu(t_i) \right\} \\ &\times \exp\left(-\sum_{k=1}^{K} \tilde{\nu}_k \, C_k(\phi)\right) \left\{ \prod_{\{i:\, t_i \in O\}} \left(\sum\nolimits_{k=1}^{K} \tilde{\nu}_k\right) \mathrm{ga}(t_i - t_{y_i} \mid \xi_i, \phi^{-1}) \right\}, \end{aligned} \tag{8}$$

where $C_k(\phi) = \sum_{j=1}^{n} \int_0^{T-t_j} \text{ga}(u \mid k, \phi^{-1})\, \text{d}u$. Marginalizing over the ξ_i, the last component of (8) becomes $\prod_{\{i:\, t_i \in O\}} \left\{ \sum_{k=1}^{K} \tilde{\nu}_k \, \text{ga}(t_i - t_{y_i} \mid k, \phi^{-1}) \right\} = \prod_{\{i:\, t_i \in O\}} h(t_i - t_{y_i})$.

Recall that the main challenge in model estimation pertains to the HP augmented likelihood term $\exp\left(- \sum_{j=1}^{n} \int_0^{T-t_j} h(u)\, \text{d}u \right)$. Under our model, this term simplifies to $\prod_{k=1}^{K} \exp(-\tilde{\nu}_k \, C_k(\phi))$, where the coefficients $C_k(\phi)$ depend only on model parameter ϕ. Combined with the gamma prior distributions for the weights $\tilde{\nu}_k$ in (6) (equivalently, the DP-based prior for the weights ψ_k in (7)), the MCMC algorithm involves remarkably simple updates for the main model parameters. In particular, given the branching and basis configuration latent variables (and the other model parameters), the weights $\tilde{\nu}_k$ are independent, and each $\tilde{\nu}_k$ follows a gamma posterior full conditional distribution. Details of the MCMC algorithm are available in the Supplementary Material.

2.2.2 Uniform mixture modeling for decreasing offspring densities

For the second modeling approach, we combine a prior for the branching ratio, γ, with a prior model that supports general decreasing offspring densities. Such a model can be formulated using the probabilistic representation of non-increasing densities on $\mathbb{R}^+$ as scale uniform mixtures. More specifically, a density f on $\mathbb{R}^+$ is non-increasing if and only if there exists a distribution function F on $\mathbb{R}^+$ such that $f(x) = \int \text{Unif}(x \mid 0, \theta)\, \text{d}F(\theta)$. Hence, the uniform mixture model for decreasing HP offspring densities is given by:

$$f(x) \;=\; \int \theta^{-1} \mathbf{1}_{[0,\theta)}(x)\, \text{d}F(\theta) \;=\; \sum_{m=1}^{\infty} \omega_m \, Z_m^{-1} \mathbf{1}_{[0,Z_m)}(x), \quad x \in \mathbb{R}^+. \tag{9}$$

This representation has been used for NPB modeling of unimodal densities (e.g., Brunner and Lo, 1989; Lavine and Mockus, 1995; Kottas and Krnjajić, 2009). In our context, it yields a flexible model that, subject to the monotonicity constraint, can capture non-standard behavior close to 0 as well as in the tail of the offspring density.

The second expression in (9) signifies that the mixing distribution F is assigned a nonparametric prior that defines it as a (random) countable discrete distribution, $F = \sum_{m=1}^{\infty} \omega_m \, \delta_{Z_m}$. The atoms, Z_m, and corresponding weights, ω_m, are a priori independent. We have $Z_m \mid a_0, \beta \overset{i.i.d.}{\sim} F_0$, where F_0 is the centering distribution, taken to be the inverse-gamma distribution, $\text{IG}(a_0, \beta)$, with mean $\beta/(a_0 - 1)$ (provided $a_0 > 1$). A general approach to define the weights involves the stick-breaking construction:

$$\omega_1 = \varphi_1, \quad \text{and} \quad \omega_m = \varphi_m \prod_{r=1}^{m-1} (1 - \varphi_r), \quad \text{for} \quad m \geq 2. \tag{10}$$

We take $\varphi_m \mid \alpha \overset{i.i.d.}{\sim} \text{Be}(1, \alpha)$, where $\text{Be}(a, b)$ denotes the beta distribution with mean $a/(a+b)$, such that F is assigned the $\text{DP}(\alpha, F_0)$ prior. Hyperpriors are placed on the total mass parameter, α, and the centering distribution parameters, (a_0, β).

Note that, under the model in (9), the prior expectation for the offspring density is given by $\text{E}(f(x) \mid F_0) = \int \theta^{-1} \mathbf{1}_{[0,\theta)}(x) f_0(\theta)\, \text{d}\theta$, where f_0 is the density of F_0. The prior expectation can be set to either the exponential or the Lomax density through an appropriate choice of F_0. In particular, if F_0 is taken to be the $\text{ga}(2, \beta)$ distribution, then $\text{E}(f(x) \mid \beta)$ is the exponential density with rate parameter β. Moreover, if $f_0(\theta \mid p, c) \propto (\theta/c)\{1 + (\theta/c)\}^{-(p+2)}$, then $\text{E}(f(x) \mid p, c)$ is given by the $\text{Lo}(p, c)$ density. In this case, distribution F_0 corresponds to a scale transformation of a Beta

prime distribution, in particular, $f_0(\theta \mid p, c)$ is the density of $\theta = c\,z$, where z follows the Beta prime distribution with shape parameters 2 and p.

We work with the inverse-gamma distribution for F_0, since it facilitates MCMC posterior simulation. Under $F_0 = \text{IG}(a_0, \beta)$, the prior expectation can be expressed as $\text{E}(f(x) \mid a_0, \beta) = \{\beta\,\Gamma(a_0)\}^{-1}\,\gamma(a_0 + 1, \beta/x)$, where $\gamma(s,t) = \int_0^t u^{s-1}\exp(-u)\,\text{d}u$ is the lower incomplete gamma function. To study the tail of the prior mean offspring density, the expression can be derived analytically for integer values of shape parameter a_0. In particular, for $a_0 = 1$, $\text{E}(f(x) \mid \beta) = \beta^{-1}\{1 - (1 + \beta x^{-1})\exp(-\beta x^{-1})\}$, which indicates tail behavior that falls between that of the exponential and Lomax densities.

For implementation of MCMC posterior inference, we work with a truncated version of the mixing distribution, $F_M = \sum_{m=1}^{M} \omega_m\,\delta_{Z_m}$, the only adjustment pertaining to the definition of the last weight, $\omega_M = 1 - \sum_{m=1}^{M-1}\omega_m = \prod_{m=1}^{M-1}(1 - \varphi_m)$. The truncation level M can be specified to any desired level of accuracy, using standard properties of the DP prior (see the Supplementary Material). This approach is in the spirit of blocked Gibbs sampling for mixture models built from stick-breaking priors for the mixture weights (e.g., Ishwaran and James, 2001). However, given the HP likelihood form, MCMC posterior simulation is more involved than for density estimation.

The MCMC method details are given in the Supplementary Material, but we note here the key aspects. Using the distribution function for the uniform mixture kernel, the HP augmented likelihood term associated with the excitation function can be written as: $\exp\left(-\sum_{j=1}^{n}\int_0^{T-t_j} h(u)\,\text{d}u\right) = \prod_{m=1}^{M}\exp\left(-\gamma\,\omega_m\,D(Z_m)\right)$, where $D(Z_m) = \sum_{j=1}^{n} Z_m^{-1}\min\{Z_m, T - t_j\}$. The important feature of this factorization is that the m-th product component, $\exp\left(-\gamma\,\omega_m\,D(Z_m)\right)$, includes only the m-th atom parameter Z_m. The observed point pattern, $\{0 < t_1 < ... < t_n < T\}$, defines a partition of $\mathbb{R}^+$ that comprises $n+1$ intervals with endpoints $0 < T - t_n < ... < T - t_1$. This results in $n+1$ scenarios for the expression of function $D(Z_m)$, which correspond to Z_m falling in each of the $n+1$ intervals. Except for the case $0 < Z_m < T - t_n$ (when $D(Z_m) = n$), parameter Z_m appears in the expression for $D(Z_m)$ through Z_m^{-1}. Combined with the inverse-gamma distribution for F_0, this structure results in a posterior full conditional for each Z_m that can be expressed as a piecewise density with truncated inverse-gamma components. The updates for the mixture weights involve the latent parameters φ_m, for $m = 1, ..., M-1$. The posterior full conditional for each φ_m can be written proportional to $\exp(-c_1\varphi_m)\exp\{-c_2(1 - \varphi_m)\}\,\varphi_m^{c_3}\,(1 - \varphi_m)^{c_4}$, where c_s, for $s = 1, ..., 4$, depends on other model parameters and the data, but not on φ_m. Using slice sampling auxiliary variables to handle the exponential terms, we can update each φ_m through draws from uniform and truncated beta distributions.

Overall, MCMC posterior simulation is more demanding for this modeling approach compared with the Erlang mixture model for the excitation function. It is worth noting however that we can achieve effective posterior sampling for a HP stochastic model that may be useful in applications, given the structured offspring density.

For the earthquake data analysis presented in Section 5, we apply the uniform DP mixture model in (9) and (10). As an alternative to the DP prior, we have studied the geometric weights prior (e.g., Mena et al., 2011). This prior retains the same specification for the atoms (and the independence between weights and atoms), but involves a more parsimonious construction for the weights: $\omega_m = (1 - \zeta)\,\zeta^{m-1}$, for $m = 1, 2, \ldots$, where $\zeta \sim \text{Be}(a_\zeta, b_\zeta)$. This is therefore a more structured prior than the DP, resulting in ordered weights as opposed to the stochastically ordered weights of the DP prior. In the Supplementary Material, we discuss prior specification and posterior simulation for the

geometric weights model, and compare the two uniform mixture models in the context of simulation examples. We find that they produce similar estimation results for the branching structure and the offspring density, although the MCMC algorithm for the geometric weights model requires more care to effectively explore the posterior distribution for parameter ζ. The geometric weights formulation may provide a useful option for possible extensions of the uniform mixture model, e.g., for multivariate HPs, since it substantially reduces the number of model parameters for the weights.

2.3 Fully NPB models for the HP intensity function

NPB prior models for the general HP conditional intensity in (1) arise by combining the Erlang mixture model in (4) for the immigrant intensity with either of the two priors of Section 2.2, that is, the Erlang mixture model in (6) for the excitation function or the uniform DP mixture model in (9) and (10) for decreasing offspring densities.

Approaches to study model properties and for prior specification follow directly from the methods in Sections 2.1 and Section 2.2. Moreover, the form of the HP augmented likelihood in (2) facilitates MCMC posterior simulation for the fully NPB models. Given the branching structure latent variables, we can separately update the parameters of the immigrant intensity and of the excitation function, using the MCMC algorithms detailed in the Supplementary Material.

A practical question pertains to the capacity to effectively estimate non-standard shapes for the immigrant intensity function and the excitation function under a model that uses nonparametric priors for both functions. It is also relevant to study the effect of using a general NPB model when standard (parametric) forms for the immigrant intensity and/or the excitation function may suffice. We address these questions through a simulation study. The main results for the fully NPB models are presented in Section 4, with additional details included in the Supplementary Material.

3 Model assessment and prediction

The MCMC posterior simulation methods for the models developed in Section 2 provide directly full inference for the immigrant intensity function and for the excitation function (or the offspring density). Since all models are implemented using the HP augmented likelihood, we also obtain posterior samples for the branching structure latent variables. Here, we discuss additional inference methods, including techniques for model assessment (Section 3.1) and prediction (Section 3.2).

3.1 Model assessment

A general purpose technique for (graphical) assessment of a point process model builds from the time-rescaling theorem (e.g., Daley and Vere-Jones, 2003). Consider a realization, $\{0 = t_0 < t_1 < ... < t_n < T\}$, from a point process with conditional intensity function $\lambda^*(t)$, and let $\Lambda^*(t) = \int_0^t \lambda^*(u)\,\mathrm{d}u$ (with $\Lambda^*(0) \equiv 0$). Then, the transformed point pattern $\{\Lambda^*(t_1), \ldots, \Lambda^*(t_n)\}$ is a realization from a unit rate homogeneous Poisson process, and thus the $U_i = 1 - \exp\{-(\Lambda^*(t_i) - \Lambda^*(t_{i-1}))\}$, for $i = 1, ..., n$, are independent Unif$(0, 1)$ random variables. Therefore, model assessment can be based on a quantile-quantile (Q-Q) plot to evaluate agreement of the estimated U_i with the uniform distribution on $(0, 1)$. Under a Bayesian modeling approach, we can apply this technique for every posterior sample of $\lambda^*(t)$, and thus report posterior mean and interval estimates for the Q-Q graph.

We next discuss two model assessment tools that require different levels of information regarding

the underlying branching structure. As such, these tools are useful mainly for settings where models are assessed/compared using synthetic point patterns, generated from the HP using its cluster representation. Indeed, we apply both tools in the context of a simulation study designed to evaluate different aspects of the proposed models; see Section 4.

Suppose that for the point pattern $\{0 < t_1 < ... < t_n < T\}$ we have access to the classification of each time point as either an immigrant or an offspring. Therefore, we know the "true" branching variables to the extent of $y_i^{\text{true}} = 0$ (t_i is an immigrant) or $y_i^{\text{true}} \neq 0$ (t_i is an offspring point). We thus also have the actual number of immigrants, $n_I^{\text{true}} = |\{i : y_i^{\text{true}} = 0\}|$, and number of offspring, $n_O^{\text{true}} = |\{i : y_i^{\text{true}} \neq 0\}|$. Evidently, these counts can be used to assess the posterior distribution for the immigrant size, $n_I = |\{i : y_i = 0\}|$, and the offspring size, $n_O = |\{i : y_i \neq 0\}|$, as estimated by a particular model. For additional evaluation of the branching structure estimation, we consider two types of misclassification rates. Note that, for a given MCMC posterior sample, $|\{i : y_i \neq 0,\ y_i^{\text{true}} = 0\}|$ is the number of immigrants incorrectly classified as offspring, and, analogously, $|\{i : y_i = 0,\ y_i^{\text{true}} \neq 0\}|$ is the number of offspring incorrectly classified as immigrants. We therefore define the immigrant and offspring misclassification rates (M_I and M_O, respectively) as follows:

$$M_I = \frac{|\{i : y_i \neq 0,\ y_i^{\text{true}} = 0\}|}{n_I^{\text{true}}} \quad \text{and} \quad M_O = \frac{|\{i : y_i = 0,\ y_i^{\text{true}} \neq 0\}|}{n_O^{\text{true}}}. \tag{11}$$

For the combined misclassification rate, we use the weighted average of the two rates, $R = (n_I^{\text{true}} M_I + n_O^{\text{true}} M_O)/n$.

The immigrant/offspring classification is most often not available in applications. However, the earthquake data set considered in Section 5 includes the classification of each shock as a main shock or aftershock, and we can thus use the technique discussed above as an extra means to compare across different models.

To motivate the second technique, suppose that we seek an empirical estimate of the offspring density. With the branching structure information available, such an estimate for the immigrant intensity is readily available, but the problem requires more consideration for the offspring density. For any parent point t_j (i.e., any point that has offspring within the observation window $(0, T]$), its offspring are generated in the interval $(t_j, T]$ according to the NHPP with intensity defined by the excitation function. The corresponding density is given by $q_{t_j}(x) = h(x)\,\mathbf{1}_{(0,T-t_j]}(x)/\{H(T-t_j)\}$, where $H(t) = \int_0^t h(s)\,\mathrm{d}s$. To account for the fact that the support depends on the time of the parent event, we use a weighted combination of these densities, referred to as the *aggregate offspring density function*:

$$\sum_{k=1}^{\tilde{n}} \left\{ \frac{H(T - t_{\tilde{y}_k})}{\sum_{r=1}^{\tilde{n}} H(T - t_{\tilde{y}_r})} \right\} q_{t_{\tilde{y}_k}}(x) \;=\; \frac{\sum_{k=1}^{\tilde{n}} h(x)\,\mathbf{1}_{(0,T-t_{\tilde{y}_k})}(x)}{\sum_{k=1}^{\tilde{n}} H(T - t_{\tilde{y}_k})}. \tag{12}$$

Here, $\{\tilde{y}_k : k = 1, ..., \tilde{n}\}$ are the distinct values in the branching structure $\{y_1, \ldots, y_n\}$, which therefore identify the index for all time events in the point pattern that have offsrping within the observation window. Under the HP branching structure representation, the k-th weight (left-hand-side of (12)) is proportional to the average number of offspring generated by the parent time point identified by $\tilde{y}_k$.

The estimates for the aggregate offspring density can be compared with the data histogram that uses the distances between the offspring times and their parent time (for all parent time

events). This tool is practically relevant for synthetic point patterns, for which we have access to the required branching structure information. In the context of our Bayesian simulation study (based on a single synthetic point pattern for every simulation setting), the aggregate offspring density estimates provide a useful method to assess estimation accuracy for the offspring density, additional to comparing with the true functions used to generate the synthetic point patterns.

3.2 Prediction

The Bayesian modeling approach provides a natural setting for posterior predictive inference for essentially any point process stochastic model for which we can simulate realizations from the point process. The posterior predictive distribution for a point process functional can be sampled by computing the functional based on point patterns simulated from the point process, with a different point pattern simulated for each posterior sample for the conditional intensity (e.g., Leininger and Gelfand, 2017).

Given the pattern $\{0 < t_1 < ... < t_n < T\}$ observed in time window $(0, T]$, we seek prediction within the future time interval $(T, T^*]$. For instance, in the context of the earthquake data analysis in Section 5, we obtain posterior forecasts for the number of earthquakes over a period of 20 years after the end of the observation time window.

We next present a computational approach for conditional prediction, that is, predictive inference within the time interval $(T, T^*]$ given the observed point pattern, which thus defines the history for the HP conditional intensity up to time T. For every MCMC posterior sample, we simulate a HP realization in the prediction time interval, using the immigrant intensity and excitation functions defined by the model parameters in the posterior sample. To simulate the point process realization, we use the HP cluster representation, taking into account the observed time points that fix the realization in $(0, T]$. Since we utilize the HP branching definition, the HP realization in $(T, T^*]$ is formed by the superposition of multiple generations of NHPP patterns. This is key in that, owing to the Poisson process definition, each NHPP simulation can be performed in $(T, T^*]$ without the need to condition on the history up to time T.

More specifically, we explore HP simulation in the prediction window $(T, T^*]$, conditional on the observed point pattern, $0 < t_1 < ... < t_n < T$. For each MCMC posterior sample, we have the posterior realization for the immigrant intensity μ (or $\mu(t)$), and for the excitation function h (equivalently, for the offspring density f and the branching ratio γ). The HP realization in $(T, T^*]$ is formed by combining: (a) the new immigrants in $(T, T^*]$, and all points from subsequent generations that fall in $(T, T^*]$; and, (b) the new time events generated within $(T, T^*]$ as offspring of the observed time points, and all points in $(T, T^*]$ from subsequent generations.

Regarding item (a), for models with immigrant intensity constant in time, the number of new immigrants in $(T, T^*]$ is sampled from the Poisson distribution with mean $\mu\,(T^* - T)$, and their times are drawn independently from the uniform distribution on $(T, T^*]$. For fully NPB models with time-dependent immigrant intensity, the number of new immigrants is drawn from the Poisson distribution with mean $\int_T^{T^*} \mu(t)\,\mathrm{d}t$, and the new immigrant times are independent draws from the density $\mu(t)/\{\int_T^{T^*} \mu(u)\,\mathrm{d}u\}$. Under the prior model of Section 2.1, this density is a mixture of the Erlang basis densities in (4), truncated over $(T, T^*]$, with weights proportional to $\nu_l \int_T^{T^*} \mathrm{ga}(u \mid l, \theta^{-1})\,\mathrm{d}u$. Next, we simulate all subsequent generations, where, for any point that has offspring within $(T, T^*]$, its offspring are generated as a point pattern from a NHPP with intensity defined through the excitation function.

The first step for item (b) involves generating within the prediction interval any possible offspring from the time points in the observed point pattern. For any observed time event t_i, we draw the number of its offspring, $\tilde{m}_i$, in $(T, T^*]$ from the Poisson distribution with mean $\int_{T-t_i}^{T^*-t_i} h(u)\,\mathrm{d}u$. If $\tilde{m}_i = 0$, there is no contribution to the predictive count from observed time point t_i. If $\tilde{m}_i \geq 1$, then we generate the $\tilde{m}_i$ offspring of t_i in $(T, T^*]$ by drawing them independently from the density $h(x-t_i)/\{\int_T^{T^*} h(s-t_i)\mathrm{d}s\}$, for $x \in (T, T^*]$, or, equivalently, by drawing them independently from the density $f(x)/\{\int_{T-t_i}^{T^*-t_i} f(u)\mathrm{d}u\}$, for $x \in (T-t_i, T^*-t_i]$. Next, for every offspring of t_i in $(T, T^*]$, we generate its complete family tree within $(T, T^*]$; again, this involves simulation from (independent) NHPPs with intensities driven by the excitation function.

Let N_1 be the count corresponding to item (a), i.e., the total number of time points in the prediction interval that originate from the new immigrants (including the new immigrant points). Similarly, let N_2 be the count associated with item (b), i.e., the total number of all offspring of $\{t_1, ..., t_n\}$, and all time points from all subsequent generations, that fall in the prediction interval. Then, the predictive draw for the number of events in $(T, T^*]$ is $N_1 + N_2$. Repeating over all posterior samples, we obtain the corresponding posterior forecast distribution.

We note that an alternative method for conditional HP simulation arises by extending the standard thinning algorithm for NHPP simulation; see Ogata (1981), and Chapter 4 in Laub et al. (2021). This sampling method generates the time points (arrival times) for the HP realization sequentially by stochastically accepting candidate points produced given the process history. More specifically, if t_i is the current time point in the HP realization, the candidate for the next time point (t_{i+1}) is $t_i + E$, where E is a draw from an exponential distribution with rate parameter $B = \lambda^*(t_i^+)$. Here, t_i^+ denotes the time right after arrival t_i (in practice, B can be computed as $\lambda^*(t_i + \epsilon)$, for small $\epsilon > 0$). Hence, provided the HP conditional intensity is non-increasing during periods of no arrivals, B yields an upper bound for the conditional intensity $\lambda^*(t)$, for $t \in (t_i, t_{i+1}]$. Then, $t_i + E$ is accepted as the next time point of the HP sample path if $U \leq \lambda^*(t_i + E)$, where U is a draw from the uniform distribution on $(0, B)$. The sequential sampling procedure stops once a time point is generated outside the prediction interval. Again, the approach can be applied for all conditional intensity posterior samples to produce a forecast distribution for, say, the count in interval $(T, T^*]$.

Although the simulation method based on the thinning algorithm appears to be the more commonly utilized for temporal HPs, it is limited in scope compared to the approach that uses the HP branching representation. In particular, as aforementioned, the thinning algorithm requires that the HP conditional intensity is non-increasing during periods between successive time events. This condition holds true for standard parametric models, based on immigrant intensity constant in time, and an exponential or Lomax offspring density. It is also satisfied for the uniform mixture model for decreasing offspring densities in (9), provided the immigrant intensity is constant in time. However, the condition is not necessarily satisfied for HP models with time-dependent immigrant intensity, or with excitation function that is not restricted to be decreasing.

All prediction results for the earthquake data (Section 5 and Supplementary Material) are based on the simulation method that uses the HP cluster representation. Although results are not shown, we have tested the thinning algorithm for the two models considered in Section 5 for which it applies: a parametric model, and the semiparametric model with immigrant intensity constant in time, and the uniform DP mixture model in (9) and (10) for the offspring density. The prediction results for these two models were practically the same under the two different methods for conditional HP simulation.

4 Simulation study

We study the performance of the different NPB models, presented in Section 2, using several synthetic point patterns generated from HPs with varying structure for the immigrant intensity and the excitation function. The simulation experiment follows our Bayesian inference setting, such that results for each simulation example are conditional on a specific point pattern. As the point patterns are simulated using the HP cluster representation, we can apply the model assessment tools discussed in Section 3.1.

In this section, we focus on results for the fully NPB models, based on the Erlang mixture prior for the immigrant intensity function, and either the Erlang mixture model for the excitation function (Section 4.1) or the uniform DP mixture model for the offspring density (Section 4.2). We also discuss some of the additional results and examples, which can be found in the Supplementary Material.

4.1 Erlang mixture model for the excitation function

In this section, we consider the fully NPB model based on the Erlang mixture priors in (4) and (6) for the immigrant intensity and the excitation function, respectively.

The data generating HP conditional intensity is of the general form in (1) with $h(x) = \gamma f(x)$. We evaluate performance of the model under three scenarios. The first example uses a simple HP form, based on immigrant intensity constant in time and an exponential offspring density:

$$\mu(t) \equiv \mu = 0.02, \qquad f(x) = \text{Exp}(x \mid 1), \tag{13}$$

with $\gamma = 0.8$, where $\text{Exp}(x \mid \theta)$ denotes the density of the exponential distribution with rate parameter θ. The second example introduces a non-standard (bimodal) offspring density combined with a unimodal immigrant intensity:

$$\mu(t) = 200\,\text{We}(t \mid 3, 5000), \qquad f(x) = 0.6\,\text{We}(x \mid 2, 2) + 0.4\,\text{We}(x \mid 5, 10), \tag{14}$$

with $\gamma = 0.6$, where $\text{We}(t \mid \theta_1, \theta_2)$ is the density of the Weibull distribution with shape parameter θ_1 and scale parameter θ_2. The third example involves non-standard shapes for both the immigrant intensity and the offspring density, in particular,

$$\mu(t) = 120\,\text{We}(t \mid 1.5, 2000) + 80\,\text{We}(t \mid 7, 7500) \tag{15}$$

with $\gamma = 0.6$, and the offspring density $f(x)$ in (14). The data generating immigrant intensities and offspring densities are depicted by the red lines in Figure 1.

The synthetic point pattern is generated within time interval $(0, 5000]$ for the first example, and within time interval $(0, 10000]$ for the second and third examples. The HP realization includes 534 points (105 immigrants, 429 offspring) for the first example, 500 points (202 immigrants, 298 offspring) for the second example, and 542 points (206 immigrants, 336 offspring) for the third example.

The priors used for the model hyperparameters in each example are given in the Supplementary Material. Posterior mean and interval estimates for the immigrant intensity, offspring density, and aggregate offspring density are shown in Figure 1. Recall from Section 3.1 that the aggregate

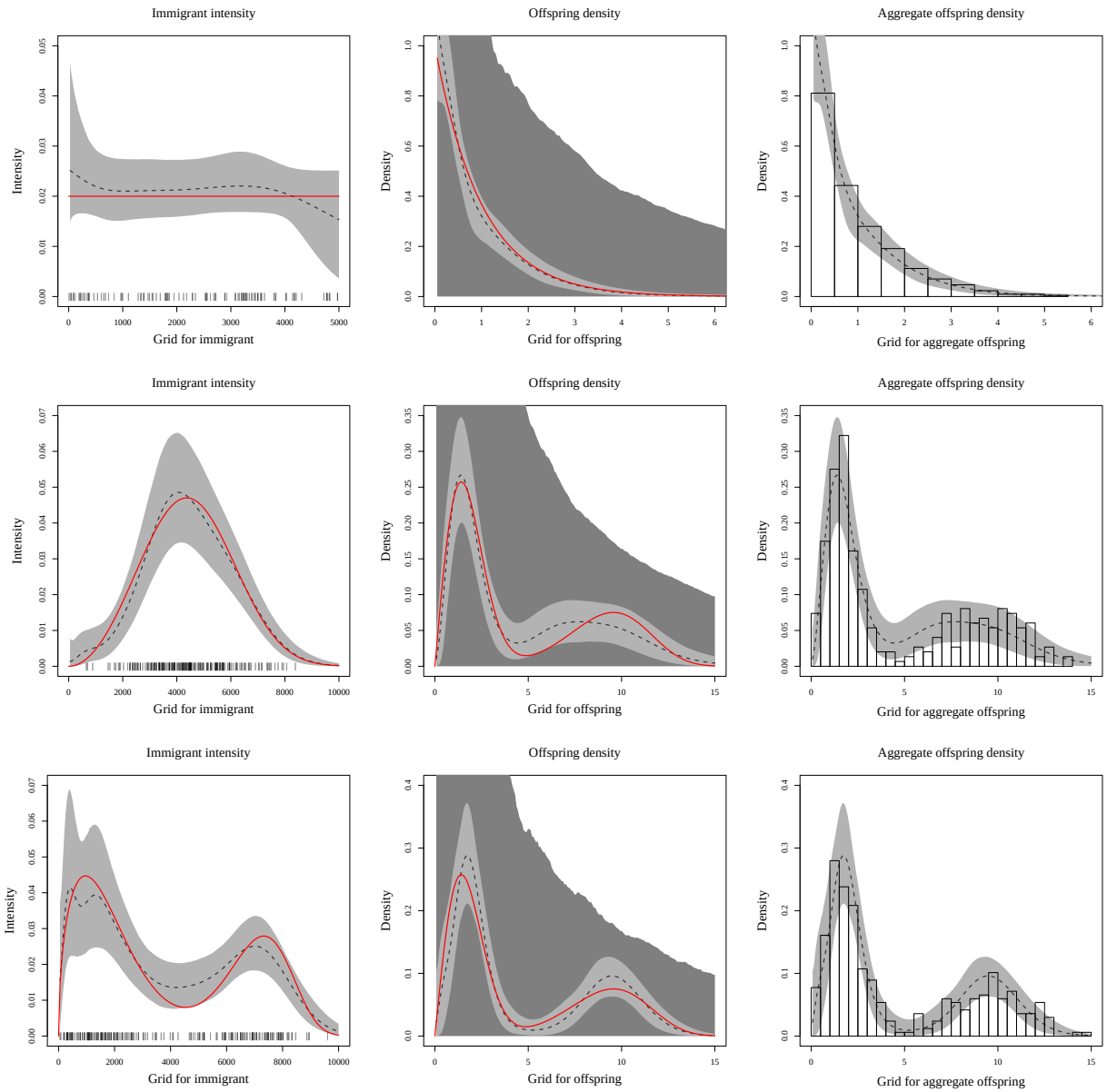


Figure 1: Simulation example for the fully NPB model with the Erlang mixture excitation function (Section 4.1). The top, middle, and bottom row correspond to the first, second, and third example, respectively. Left column: Posterior mean (dashed line) and 95% interval estimates (light gray bands) for the immigrant intensity. The red lines depict the data generating HP immigrant intensities, and the bars on the horizontal axes indicate the immigrant occurrence times. Middle column: Posterior mean estimate (dashed line), 95% posterior interval estimates (light gray bands), and 95% prior interval estimates (dark gray bands) for the offspring density. The red lines depict the data generating HP offspring densities. Right column: Posterior mean (dashed line) and 95% interval estimates (light gray bands) for the aggregate offspring density. The histograms are based on the distances between offspring times and their corresponding parent times (for all parent points).

offspring density provides a data-based means to evaluate estimation accuracy for the offspring density. It is encouraging that the model captures the non-standard shapes for the immigrant intensity and/or the offspring density, but also reassuring that it has the capacity to recover simpler parametric forms. Figure 1 includes the 95% prior interval bands for the offspring densities (we

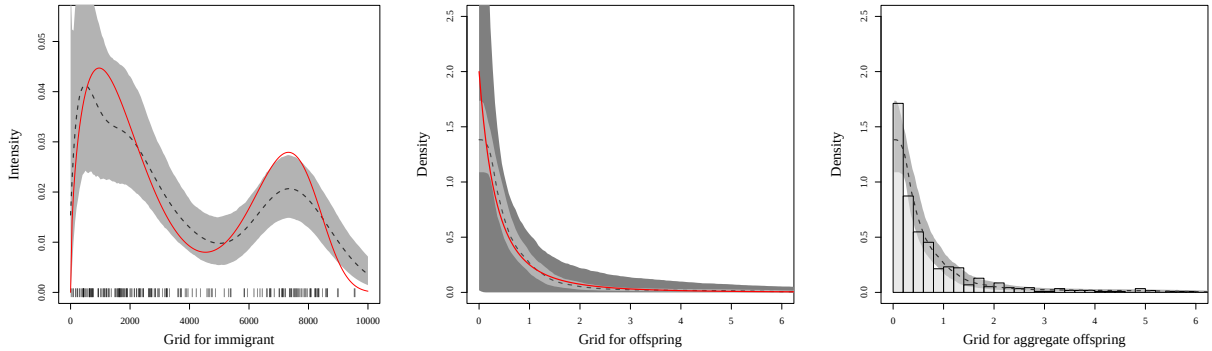


Figure 2: Simulation example for the fully NPB model with the uniform DP mixture offspring density (Section 4.2). Posterior mean (dashed line) and 95% interval estimates (light gray bands) for the immigrant intensity (left panel), the offspring density (middle panel), and the aggregate offspring density (right panel). The bars on the horizontal axis of the left panel indicate the immigrant occurrence times. The histogram in the right panel is based on the distances between the offspring times and their parent time (for all parent points). The middle panel includes the 95% prior interval estimates (dark gray bands). The data generating HP immigrant intensity and offspring density are shown by the red line in the left and middle panel, respectively.

omit the corresponding bands for the immigrant intensity, since they cover essentially the entire panel). The prior interval bands reflect the uncertainty implied by the NPB prior model, and they also demonstrate the extent of prior-to-posterior learning.

The Supplementary Material elaborates this part of the simulation study. In particular, for the first example, we compare with the parametric model that corresponds to the data generating mechanism, and with the NPB model that uses the Erlang mixture excitation function with immigrant intensity constant in time. Evidently, neither of the NPB models is needed for this example. It is noteworthy that the two NPB models yield similar estimation for the branching structure, and that the fully NPB model performs only slightly worse than the correct parametric model. For instance, under the parametric model, the posterior mean (standard deviation) for the branching ratio, γ, and the combined misclassification rate, R, are 0.798 (0.039), and 0.036 (0.006), respectively. The corresponding estimates under the fully NPB model are 0.806 (0.039), and 0.041 (0.008). Along with the estimates in Figure 1, these results provide clear empirical evidence that the fully NPB model avoids (significant) overfitting, while it can identify different functional forms for the immigrant intensity and the excitation function.

4.2 Uniform DP mixture model for the offspring density

Here, we test the model that combines the Erlang mixture prior in (4) for the immigrant intensity with the uniform DP mixture prior in (9) and (10) for the offspring density.

The synthetic point pattern comprises 777 points (193 immigrants and 584 offspring) generated within time interval $(0, 10000]$. The data generating HP conditional intensity is defined by the immigrant intensity in (15), $\gamma = 0.8$, and a Lomax offspring density, $f(x) = \text{Lo}(x \mid 2, 1)$. The Weibull mixture structure yields the bimodal immigrant intensity also used for the third example of Section 4.1. As discussed in Section 2.2, the Lomax and the exponential densities are the two most commonly used parametric forms for the offspring density.

Figure 2 reports posterior mean and interval estimates for the immigrant intensity, offspring

γ	n_I	n_O	M_I	M_O	R
0.75 (0.03)	193 (10)	584 (10)	0.149 (0.035)	0.049 (0.009)	0.074 (0.009)

Table 1: Simulation example for the fully NPB model with the uniform DP mixture offspring density (Section 4.2). Posterior mean (standard deviation) for the branching ratio (γ), immigrant and offspring group sizes (n_I and n_O), and misclassification rates (M_I, M_O, and R).

density, and aggregate offspring density. Again, the results are encouraging with respect to the model's capacity to uncover the shapes of the underlying functions. Effective model performance is also supported by inference results for the underlying branching structure (Table 1).

The Supplementary Material includes additional simulation examples for the uniform mixture offspring density model, under the simpler setting where the data generating HP has immigrant intensity constant in time. Accordingly, we implement the semiparametric model with $\mu(t) \equiv \mu$, and the uniform mixture model in (9) for the offspring density. As previously discussed, we explore both the DP prior and the geometric weights prior for the mixing distribution. We consider the same branching ratio ($\gamma = 0.8$), and two scenarios for the data generating offspring density: the same Lomax density used here ($f(x) = \text{Lo}(x \mid 2, 1)$); and, a decreasing Weibull density. The NPB models yield effective estimation for the offspring density and the branching structure. As expected, the semiparametric uniform DP mixture offspring density model results in smaller misclassification rates than in Table 1, and it also provides some improvement on the estimation of the underlying Lomax density, for x values close to 0.

Overall, the simulation study provides empirical evidence that the synthesis of the NPB prior models for the immigrant intensity and the excitation function (offspring density) offers a versatile inference framework for temporal HPs. We elaborate the empirical investigation with the earthquake data analysis of the next section, where prediction is added to the inference objectives.

5 Earthquake data analysis

We consider the catalog of earthquakes reported in Table 1 of Ogata (1988), which includes the occurrence times for earthquakes (of magnitude ≥ 6, and depth < 100 kilometers) recorded in Japan and its vicinity from 1885 through 1980. The geographic region that corresponds to this earthquake catalog is part of the northwestern Pacific seismic belt; see Figure 1 of Ogata (1988).

An important feature of this earthquake catalog is that each shock is classified as a main shock (258 points), aftershock (200 points) or foreshock (25 points). We work with the point pattern that comprises the 458 occurrence times which exclude the foreshocks. The frequency histogram of these times is shown in Figure 3. The figure extends to the end of year 2000, using data we have obtained from the United States Geological Survey (USGS) earthquake archive (USGS, 2026); see Section 5.2 for details.

Earthquake modeling builds mainly from the Epidemic Type Aftershock Sequences (ETAS) model, including extensions to handle space-time patterns and to incorporate earthquake magnitude as a mark variable; see, e.g., the review in Bernabeu et al. (2025). The ETAS model for temporal earthquake point patterns is defined through a HP with immigrant intensity constant in time, $\mu(t) \equiv \mu$, and Lomax offspring density, $f(x) = \text{Lo}(x \mid p, c)$. Bayesian inference for the ETAS model is implemented for the purposes of comparison with the proposed NPB models.

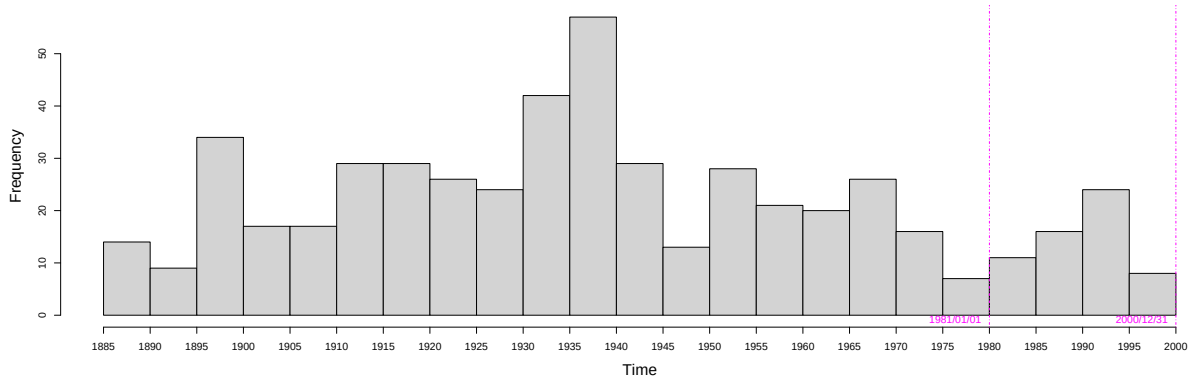


Figure 3: Frequency histogram of earthquake occurrence times (earthquakes of magnitude ≥ 6, and depth < 100 kilometers) recorded in Japan and its vicinity. The magenta dashed vertical lines indicate the prediction time interval (January 1, 1981 to December 31, 2000).

As discussed in Section 2.2, parametric models for earthquake occurrences (including the ETAS model) are based on a decreasing offspring density, the decreasing shape suggested by seismology theory. We thus focus on the uniform DP mixture prior in (9) and (10) to explore a general NPB model for decreasing offspring densities. In fact, we consider two versions of the NPB model relating to the prior choice for the immigrant intensity: the fully NPB model with the Erlang mixture prior in (4) for the immigrant intensity; and, the simpler version with immigrant intensity constant in time. Hence, there is a progression in the complexity of the models – from the ETAS, to the NPB, to the fully NPB model – which facilitates examining the effect of the two conditional intensity components in inference and prediction. Although results are not shown, we note that the NPB model with the Erlang mixture prior for the excitation function yields decreasing offspring density realizations, given the specific earthquake point pattern.

Regarding the hyperparameter priors, we used a ga$(2, 8)$ prior for the branching ratio (for all models), and an Exp(153) prior for μ (for the models with immigrant intensity constant in time). An Exp(1) prior was assigned to both parameters of the ETAS model Lomax offspring density. For the uniform DP mixture offspring density prior model, we placed an Exp(0.0769) prior to a_0, Exp(0.0083) prior to β, and ga$(4, 0.5)$ prior to α (the DP truncation level was set to $M = 85$). Finally, for the Erlang mixture immigrant intensity prior model, we used a Lo$(2, 1800)$ prior for θ, Exp(0.1) prior for c_0, and Exp(0.0065) prior for b_{G_0} (the number of basis functions was set to $L = 120$).

All models are implemented given the point pattern in $(1885, 1980]$, with the time interval $(1981, 2000]$ used for prediction. The corresponding results are discussed in Sections 5.1 and 5.2. The main shock/aftershock classification available in this catalog is not taken into account in fitting any of the models. However, we use this information to evaluate the different models in terms of their estimation of the branching structure.

The main shock/aftershock classification for each time point is also useful in providing simple, data-based estimates for γ and μ, which can be used as reference points in comparing model-based inference results. If the HP is stationary, i.e., it has immigrant intensity constant in time and $\gamma \in (0, 1)$, its first-order intensity is given by $\lambda = \mu/(1 - \gamma)$. In our analysis, the time unit is days, with the total time period for the observation window comprising about $T = 35000$ days. Hence,

Model	γ	Cluster size		Misclassification		
		n_I	n_O	M_I	M_O	R
ETAS	0.697 (0.175)	234 (20)	224 (20)	0.293 (0.056)	0.257 (0.036)	0.278 (0.024)
NPB	0.459 (0.066)	250 (25)	208 (25)	0.238 (0.070)	0.268 (0.041)	0.251 (0.029)
Fully NPB	0.436 (0.064)	261 (24)	197 (24)	0.208 (0.067)	0.285 (0.042)	0.242 (0.027)

Table 2: Earthquake data analysis. Posterior mean (standard deviation) for the branching ratio, immigrant/offspring cluster sizes, and misclassification rates. NPB: uniform DP mixture offspring density model with immigrant intensity constant in time; Fully NPB: the extension of the NPB model with the Erlang mixture immigrant intensity.

using n $(= 458)$ as a proxy for λT, the average intensity in the observation window, the estimate for λ is n/T. Under immigrant intensity constant in time, and given the branching structure, the main shocks form a realization from a Poisson process with intensity μ. Now, using the number of earthquakes classified as main shocks ($n_I^{\text{true}} = 258$) as a proxy for μT, the total immigrant intensity over the observation window, the estimate for μ is $\tilde{\mu} = n_I^{\text{true}}/T$. Finally, the estimate for γ can be obtained as $\tilde{\gamma} = 1 - (n_I^{\text{true}}/n)$. The estimates for the specific data set are $\tilde{\mu} = 0.0074$, and $\tilde{\gamma} = 0.4367$.

5.1 Inference and model comparison results

The estimation results for the branching structure (Table 2) suggest that the ETAS model overestimates the branching ratio (note the data-based estimate, $\tilde{\gamma} = 0.4367$). Both NPB models improve on the estimation of the branching ratio, and the immigrant and offspring cluster sizes. This is also reflected in the combined misclassification rate estimates, according to which the fully NPB model provides the best performance.

The ga$(2, 8)$ prior distribution for the branching ratio is chosen such that the stability condition, $\gamma \in (0, 1)$, receives essentially all the prior probability. In contrast to the NPB models, the ETAS model posterior distribution for γ is sensitive to the prior choice, supporting larger values under more dispersed priors. For instance, under a ga$(2, 4)$ prior for γ, the posterior mean (standard deviation) for the branching ratio and the combined misclassification rate increase to 0.897 (0.291) and 0.292 (0.025), respectively.

Although the time-rescaling Q-Q plots are similar across models (Figure 4), there is suggestion of lack of fit for the ETAS model (at smaller quantiles). Figure 4 reports also posterior mean and interval estimates for the immigrant intensity and the offspring density. Using the data-based estimate, $\tilde{\mu} = 0.0074$, as a reference point, the NPB model outperforms the ETAS model in immigrant intensity estimation. The ETAS model underestimates the immigrant intensity, which is consistent with the inference results for the branching structure (Table 2). The fully NPB model estimates time-dependent immigrant intensity with local fluctuations in the earlier part of the time period.

To facilitate visualization, the estimates for the offspring density are plotted over the (first) one-day interval for the time lag between aftershocks and their parent shock. There is a noticeable difference between the ETAS model and the NPB models in terms of the estimated offspring density at small x values. This is reflected in the estimates for the offspring probability within the first day from the time of the parent shock. The two NPB models yield essentially identical offspring density estimates, which provides further empirical evidence that the fully NPB model has the capacity to

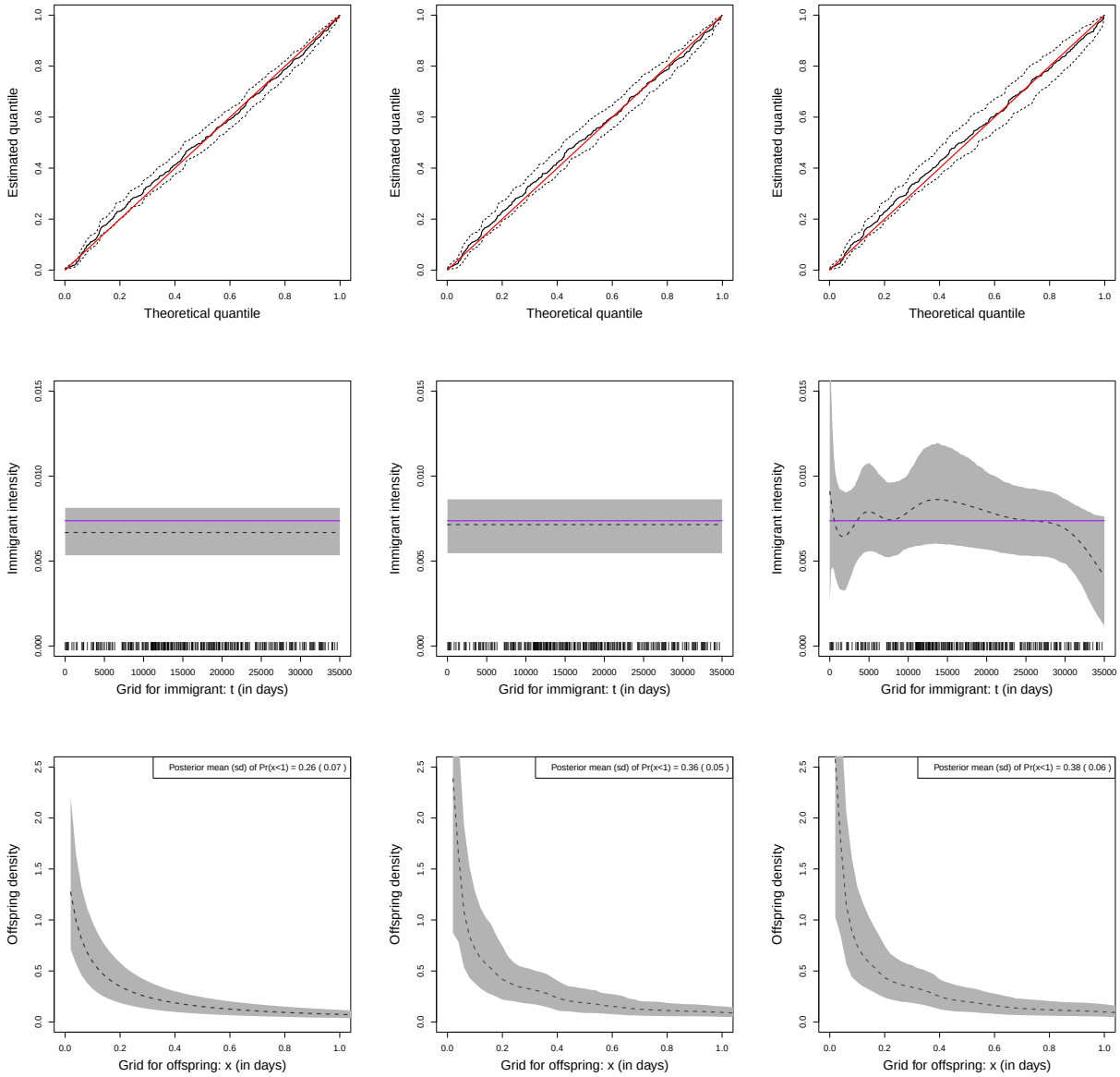


Figure 4: Earthquake data analysis. Inference results from the ETAS model (left column), NPB model (middle column), and fully NPB model (right column). Top row: posterior mean and 95% interval estimates for the time-rescaling model assessment Q-Q plots. Middle row: posterior mean estimate and 95% interval bands for the immigrant intensity (the bars on the horizontal axis of each panel indicate the main shock occurrence times). The solid purple line indicates the data-based estimate for immigrant intensity constant in time, $\tilde{\mu} = 0.0074$. Bottom row: posterior mean estimate and 95% interval bands for the offspring density. Each panel includes the estimates (posterior mean and standard deviation) of the probability under the offspring density over the first day from the time of the parent shock.

identify the immigrant intensity and offspring density functions.

5.2 Prediction

To study the predictive performance of the NPB models, we consider the forecast for the earthquake count over the 20 years after the end of the observation time window. The objective is to provide an

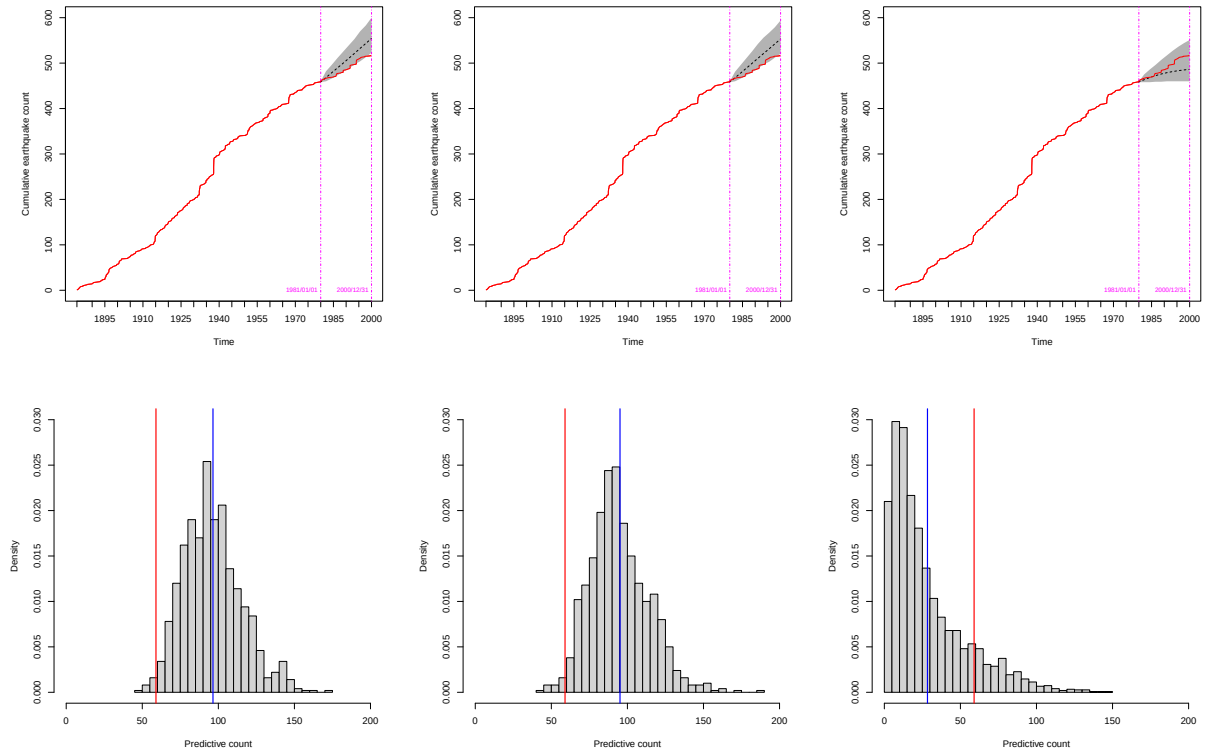


Figure 5: Earthquake data analysis. Prediction results based on the ETAS model (left column), NPB model (middle column), and fully NPB model (right column). The models are fitted to the earthquake point pattern recorded in $(1885, 1980]$, with predictions evaluated in time interval $(1981, 2000]$ using earthquake occurrences from the USGS earthquake archive (see Section 5.2). Top row: posterior predictive means (dashed lines) and 95% interval bands (shaded areas) for the cumulative earthquake count in the prediction time interval. The red line denotes the observed cumulative count of earthquakes over the entire time window from 1885 to 2000. Bottom row: posterior predictive distribution for the earthquake count over $(1981, 2000]$. The blue line indicates the posterior predictive mean, and the red line the earthquake count (59) in $(1981, 2000]$ obtained from the USGS database.

illustration with a prediction time interval of reasonable length relative to the observation window. To evaluate the predictions, we obtained the point pattern of earthquakes that occurred between January 1, 1981 and December 31, 2000 from the USGS earthquake archive. The USGS database includes filters for both earthquake magnitude and depth, and we can thus select the specification that matches the observed point pattern with respect to these marks. However, since the USGS archive allows for specification of rectangular (or circular) geographical regions, we work with earthquake occurrences within the smallest rectangle that contains the polygonal region shown in Figure 1 of Ogata (1988). The corresponding number of earthquakes of magnitude ≥ 6, depth < 100 kilometers, recorded in $(1981, 2000]$ is 59.

Figure 5 illustrates prediction results, using, for all models, the conditional simulation approach based on the HP cluster representation (Section 3.2). Overall, the ETAS and NPB models result in predictions that overestimate the cumulative earthquake count, as well as the total number of earthquakes in the prediction interval $(1981, 2000]$. The fully NPB model yields a noticeable improvement in the predictive results.

To quantify the differences in predictive performance between the three models, we compute the interval score (IS) (Gneiting and Raftery, 2007). Let $\tilde{y}$ $(= 59)$ be the count to be predicted, and (l, u)

the $(1-\alpha)100\%$ predictive interval. Then, the interval score is given by $\text{IS} = (u-l) + (2/\alpha)(l-\tilde{y})\mathbf{1}(\tilde{y} < l) + (2/\alpha)(\tilde{y}-u)\mathbf{1}(\tilde{y} > u)$, thus combining a reward for a narrow predictive interval with a penalty when $\tilde{y}$ is not included within the interval. The 95% posterior predictive interval is: $(65, 142)$ (IS = 317) under the ETAS model; $(63, 136)$ (IS = 233) under the semiparametric model; and, $(2, 93)$ (IS = 91) under the nonparametric model.

Overall, the analysis of this earthquake point pattern reveals the possible benefits of the proposed NPB modeling framework. Practical utility emerges from the NPB prior model that supports more general shapes and tail behavior for the offspring density than a specific parametric family, such as the Lomax of the ETAS model. Such utility is observed even when we relax the parametric assumption only for the offspring density, and despite the fact that the underlying space is fairly structured comprising decreasing densities on $\mathbb{R}^+$. Additional improvement, especially in forecasting, is offered from the general NPB model that allows for local temporal structure in the immigrant intensity.

6 Discussion

We have developed a Bayesian nonparametric modeling framework for temporal Hawkes processes. The primary motivation is to enhance the scope for inference and prediction under this popular class of self-exciting point processes. We have proposed models for the immigrant intensity and for the excitation function (or offspring density), the two components of the HP conditional intensity. The modeling framework includes an option that supports general decreasing offspring densities, which may be appealing for HP applications to earthquake modeling. We have designed MCMC posterior simulation algorithms to implement the different models, and developed inference methods for prediction and for model assessment. Different aspects of the methodology have been explored through a comprehensive simulation study, and with the analysis of high-magnitude earthquake occurrences in Japan and its vicinity.

Here, we explore modeling for the standard version of the HP over time. HP stochastic models admit more general forms to accommodate space-time and multivariate point patterns, including incorporation of marks. Temporal marked HPs are defined through an excitation function that includes the mark variable(s) in its arguments, with the function typically factorized in applications. Current HP models for space-time point patterns (with or without marks) expand the excitation function with a multiplicative component to account for the spatial distance between offspring points and their parent points. Multivariate (mutually exciting) HPs model a collection of K point patterns such that the conditional intensity for each point process is driven by both self-arrivals and by arrivals at the other processes. In particular, the conditional intensity for dimension k extends the form in (1) such that past events from all dimensions $\ell \in \{1, ..., K\}$ may affect the occurrence of events in dimension k, the influence controlled by excitation functions $h_{k,\ell}$. Even though multivariate HPs are typically applied with immigrant intensities constant in time, they involve a significant estimation challenge as K increases.

HP stochastic models are extensively used in applications, and thus the extent of the literature on estimation for HPs is not surprising. Focusing on model-based inference, most methods are based on parametric specifications for the HP intensity. The more recent literature includes a relatively small collection of semiparametric Bayesian methods. Taking advantage of the HP cluster representation, the easier to address model component is the immigrant intensity, using NPB models for NHPP

intensities. For example, Molkenthin et al. (2022) use a (transformed) GP prior for the spatially dependent background intensity of a space-time HP model (with parametric excitation function). Regarding NPB models for the excitation function, Donnet et al. (2020) studied multivariate HPs based on piecewise constant functions or Beta mixtures for the excitation functions $h_{k,\ell}$ (with immigrant rates constant in time). This work is effective with respect to theoretical properties (posterior concentration rates). However, the assumption of bounded support for the excitation functions is a modeling restriction, at least as it pertains to predictive inference; recall from Section 3.2 the role of the excitation function in HP simulation. Donnet et al. (2020) implement the prior constructed from piecewise constant functions, although details are not provided on handling the different HP likelihood components. We note that some of the literature employs approximations to the HP likelihood. In particular, a drastic approximation to the second exponential term in (2) amounts to setting $\int_0^{T-t_j} f(u)\,\mathrm{d}u \approx 1$, for all time points t_j; see, e.g., Jiang and Rodriguez (2024) and further references therein. Inference methods based on (transformed) GP priors for the excitation function have also been investigated in the machine learning literature (e.g., Zhang et al., 2019; Zhou et al., 2020). This work is focused almost exclusively on computational estimation methods. Such methods avoid approximations to the HP likelihood, but involve augmentation with two additional sets of latent variables (in addition to the branching structure variables). More critically, the GP is arguably not very well suited as a prior for the excitation function, which requires more structure than a NHPP intensity. In particular, this literature does not appear to include study of prior model properties that pertain to the HP definition, such as the integrability of the excitation function, and the implied prior for the branching ratio.

While acknowledging the contributions of existing NPB work, we argue that there is scope to enhance the impact of NPB methodology to modeling/applications of HPs. To this end, this paper laid the groundwork with modeling for temporal HPs. Key considerations are to: connect properties of the NPB priors with HP properties; accommodate prior structure relevant to specific applications (decreasing aftershock densities in earthquake modeling); and, balance flexible shapes for the HP functions with computational efficiency in model implementation. In particular, the two models for the HP excitation function/offspring density (Section 2.2) allow for different prior structure, and provide flexible inference without approximations of the HP likelihood.

The modeling framework proposed in this work can be extended to handle the more general HP models discussed above. A prior model for the excitation function of temporal HPs is constructed in Kim and Kottas (2025), building from the Erlang basis representation in (6) to include a continuous mark variable. The approach is developed for earthquake modeling where the mark is the earthquake magnitude. A practical feature relative to existing methods involves the capacity to estimate aftershock densities that change with the magnitude of the parent shock. The Erlang mixture model for the excitation function can be further elaborated to allow for interaction between time and space, as well as interaction with marks. Finally, we note that both the Erlang mixture and the uniform mixture prior model for the excitation function are built from a random distribution F (on $\mathbb{R}^+$). This offers potential to design prior models for the excitation functions $h_{k,\ell}$ of a multivariate HP through NPB models for the corresponding distributions $F_{k,\ell}$. In particular, allowing for structured dependence in the prior for the $F_{k,\ell}$ will enable incorporation of prior beliefs, and may lead to practical benefits in estimation. We will report on these extensions in future work.

Supplementary Material

A. Prior specification. Strategies for specifying the priors of model parameters.

B. MCMC details. Details of the MCMC posterior simulation algorithms.

C. Simulation study. Complete collection of simulation study results.

D. Sensitivity analysis. Prior sensitivity analysis results for the NPB models of Section 2.2 for the excitation function.

References

Ryan Prescott Adams, Iain Murray, and David JC MacKay. 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, 2009.

Earvin Balderama, Frederic Paik Schoenberg, Erin Murray, and Philip W Rundel. Application of branching models in the study of invasive species. *Journal of the American Statistical Association*, 107:467–476, 2012.

A. Bernabeu, J. Zhuang, and J. Mateu. Spatio-temporal Hawkes point processes: A review. *Journal of Agricultural, Biological, and Environmental Statistics*, 30:89–119, 2025.

Lawrence J. Brunner and Albert Y. Lo. Bayes methods for a symmetric unimodal density and its mode. *The Annals of Statistics*, 17:1550–1566, 1989.

PL Butzer. On the extensions of bernstein polynomials to the infinite interval. *Proceedings of the American Mathematical Society*, 5:547–553, 1954.

Daryl J Daley and David Vere-Jones. *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*. Springer, 2003.

Daryl J Daley and David Vere-Jones. *An Introduction to the Theory of Point Processes: Volume II: General Theory and Structure*. Springer, 2007.

Sophie Donnet, Vincent Rivoirard, and Judith Rousseau. Nonparametric bayesian estimation for multivariate Hawkes processes. *The Annals of Statistics*, 48:2698–2727, 2020.

T. S. Ferguson. A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1: 209–230, 1973.

E. W. Fox, F. P. Schoenberg, and J. S. Gordon. Spatially inhomogeneous background rate estimators and uncertainty quantification for nonparametric hawkes point process models of earthquake occurrences. *The Annals of Applied Statistics*, 10:1725–1756, 2016.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102:359–378, 2007.

Alan G Hawkes. Hawkes processes and their applications to finance: a review. *Quantitative Finance*, 18:193–198, 2018.

Alan G Hawkes and David Oakes. A cluster process representation of a self-exciting process. *Journal of Applied Probability*, 11:493–503, 1974.

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

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

Andrew J Holbrook, Xiang Ji, and Marc A Suchard. From viral evolution to spatial contagion: a biologically modulated hawkes model. *Bioinformatics*, 38:1846–1856, 2022.

H. Ishwaran and L. F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96:161–173, 2001.

A. Z. Jiang and A. Rodriguez. Improvements on scalable stochastic Bayesian inference methods for multivariate Hawkes process. *Statistics and Computing*, 34, 85, 2024.

J. D. Kalbfleisch. Non-parametric bayesian analysis of survival time data. *Journal of the Royal Statistical Society: Series B*, 40:214–221, 1978.

Hyotae Kim and Athanasios Kottas. Erlang mixture modeling for poisson process intensities. *Statistics and Computing*, 32, 3, 2022.

Hyotae Kim and Athanasios Kottas. Bayesian nonparametric marked Hawkes processes for earthquake modeling. arXiv:2511.22538, 2025.

A. Kottas and M. Krnjajić. Bayesian semiparametric modelling in quantile regression. *Scandinavian Journal of Statistics*, 36:297–319, 2009.

Patrick J Laub, Young Lee, and Thomas Taimre. *The Elements of Hawkes Processes*. Springer, 2021.

M. Lavine and A. Mockus. A nonparametric Bayes method for isotonic regression. *Journal of Statistical Planning and Inference*, 46:235–248, 1995.

Simon CK Lee and X Sheldon Lin. Modeling and evaluating insurance losses via mixtures of erlang distributions. *North American Actuarial Journal*, 14:107–130, 2010.

T. J. Leininger and A. E. Gelfand. Bayesian inference and model assessment for spatial point patterns using posterior predictive samples. *Bayesian Analysis*, 12:1–30, 2017.

Y. Li, J. Lee, and A. Kottas. Bayesian nonparametric Erlang mixture modeling for survival analysis. *Computational Statistics and Data Analysis*, 191:107874, 2024.

Ramsés H Mena, Matteo Ruggiero, and Stephen G Walker. Geometric stick-breaking processes for continuous-time Bayesian nonparametric modeling. *Journal of Statistical Planning and Inference*, 141:3217–3230, 2011.

George O Mohler, Martin B Short, P Jeffrey Brantingham, Frederic Paik Schoenberg, and George E Tita. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106:100–108, 2011.

C. Molkenthin, C. Donner, S. Reich, G. Zöller, S. Hainzl, M. Holschneider, and M. Opper. GP-ETAS: semiparametric Bayesian inference for the spatio-temporal epidemic type aftershock sequence model. *Statistics and Computing*, 32:29, 2022.

Yosihiko Ogata. On Lewis' simulation method for point processes. *IEEE Transactions on Information Theory*, 27:23–31, 1981.

Yosihiko Ogata. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association*, 83:9–27, 1988.

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

Jakob Gulddahl Rasmussen. Bayesian inference for hawkes processes. *Methodology and Computing in Applied Probability*, 15:623–642, 2013.

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

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

Matthew A Taddy and Athanasios Kottas. Mixture modeling for marked poisson processes. *Bayesian Analysis*, 7:335–362, 2012.

USGS. Earthquake Hazards Program, Search Earthquake Catalog, 2026. URL https://earthquake.usgs.gov/earthquakes/search/.

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

Sai Xiao, Athanasios Kottas, Bruno Sansó, and Hyotae Kim. Nonparametric bayesian modeling and estimation for renewal processes. *Technometrics*, 63:100–115, 2021.

Rui Zhang, Christian Walder, Marian-Andrei Rizoiu, and Lexing Xie. Efficient non-parametric bayesian hawkes processes. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 4299–4305, 2019.

Feng Zhou, Zhidong Li, Xuhui Fan, Yang Wang, Arcot Sowmya, and Fang Chen. Efficient inference for nonparametric hawkes processes using auxiliary latent variables. *Journal of Machine Learning Research*, 21:1–31, 2020.

Jiancang Zhuang, Yosihiko Ogata, and David Vere-Jones. Stochastic declustering of space-time earthquake occurrences. *Journal of the American Statistical Association*, 97:369–380, 2002.

Acknowledgments

The research was supported in part by the National Science Foundation under award SES 1950902.