UNIVERSITY OF CALIFORNIA
SANTA CRUZ

FLEXIBLE BAYESIAN MODELING AND INFERENCE METHODS FOR HAWKES PROCESSES

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

STATISTICAL SCIENCE

by

Hyotae Kim

December 2021

The Dissertation of Hyotae Kim
is approved:

Professor Athanasios Kottas, Chair

Professor Juhee Lee

Professor Bruno Sansó

Peter F. Biehl
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Flexible Bayesian modeling and inference methods for Hawkes processes

by

Hyotae Kim

We propose a Bayesian nonparametric modeling and inference framework for Hawkes processes. The objective is to increase the inferential scope for this practically important class of point processes by exploring flexible models for its conditional intensity function.

As a building block for conditional intensity models, we develop a prior probability model for temporal Poisson process intensities through structured mixtures of Erlang densities with common a scale parameter, mixing on the integer shape parameters. The mixture weights are constructed through increments of a cumulative intensity function modeled nonparametrically with a gamma process prior. This model specification provides a novel extension of Erlang mixtures for density estimation to the intensity estimation setting.

Turning to the main dissertation component, we develop different types of nonparametric prior models for the Hawkes process immigrant intensity and for the excitation function (or its normalized version, the offspring density), the two functions that define the point 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 methodology is further elaborated to construct a flexible and computationally efficient model for marked Hawkes processes. The motivating application involves earthquake data modeling, where the mark is given by the earthquake magnitude. The proposed model builds from a prior for the excitation function that allows flexible shapes for mark-dependent offspring densities. In the context of our motivating application, the modeling approach enables estimation of aftershock densities that can vary with the magnitude of the main shock, unlike existing marked Hawkes process models for earthquake occurrences.

For all proposed models, we develop approaches to prior specification, and design posterior simulation algorithms to obtain inference for different point process functionals. The modeling approaches are studied empirically using several synthetic and real data examples, including data on earthquake occurrences from Japan and from the southwestern US.

To all my family, who have provided me with endless love, support, and advice.

Acknowledgments

The material of Chapter 2 includes an adapted reprint of the previously published article:

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

The co-author of this publication supervised the research underlying the chapter. I am also grateful to the editor of Statistics and Computing, particularly an associate editor, and an anonymous referee for their helpful comments. The research for Chapter 2, as well as for Chapter 3 and 4, was supported in part by the National Science Foundation under award SES 1950902.

I would like to express my deepest gratitude to my advisor, Thanasis Kottas. His excellent work ethic and rigorous approach to research helped me establish a solid foundation as a researcher. In addition to his acclaimed expertise, Thanasis is dedicated to the education and training of the next generation of scientists. From recognizing motivations to developing models to solve problems, he taught me the methods of thinking. Furthermore, he encouraged me to maintain a positive outlook on life, which lifted my spirits whenever I failed and felt depressed. He has always stood up for me both inside and outside the department. Without his guidance, I would not be able to complete this thesis.

I am indebted to many individuals in the department. My sincere thanks go to Juhee Lee, who is a member of my dissertation committee, for her insightful comments

that have contributed to the enhancement of my dissertation. It is an honor to have Bruno Sansó on my committee, and I appreciate his insights and suggestions, which have enriched my work. I feel like Raquel Prado has been my second advisor. The support and encouragement she has provided me over the years have been invaluable. I also wish to thank Laura Baracaldo Lancheros, Rui Meng, and Wenjie Zhao for their support and blessing. These people have brought me so much joy during my Ph.D. journey. I would like to extend my appreciation for the assistance provided by my academic siblings Chunyi Zhao and Xiaotian Zheng.

The gratitude of my family, Hyeongyeong Mo, Jonggyu Kim, and Suyeon Kim, cannot be overstated. Having their support throughout my endeavors is such a great gift.

Chapter 1

Introduction

A (temporal) point process results in a sequence of random arrival/event/occurrence times. The point process can be interpreted as a counting process, which involves cumulative counts of the number of arrivals/events/occurrences over time. The point process and the counting process terminologies are interchangeable; we can consider a point process a sequence of times at which the counting process has jumped by 1. Point processes have been applied to research for seismology, finance, sociology, and numerous other disciplines (e.g., Ogata, 1988; Filimonov and Sornette, 2015; Fox et al., 2016; Mohler et al., 2011; Balderama et al., 2012; Schoenberg et al., 2019). Point processes can also be defined over space or both time and space, referred to as spatial point processes and spatio-temporal point processes, respectively.

This dissertation focuses on temporal point processes with prior probability modeling for their conditional intensity functions. The conditional intensity function is the conditional expected rate of arrivals at t, given the history of process up to time

t, denoted by $\mathcal{H}(t)$. Denote by $N(\cdot)$ the counting process. The conditional intensity is expressed as

$$\lambda^*(t) = \lambda(t \mid \mathcal{H}(t)) = \lim_{h \to 0} \frac{\mathrm{E}(N(t+h) - N(t) | \mathcal{H}(t))}{h}.$$

As such, the cumulative conditional intensity function, defined as $\Lambda(t) = \int_0^t \lambda^*(u) du$, indicates the expected number of arrivals in time window $(0, t)$. In the following two sections, we will review two point processes – the non-homogeneous Poisson process and the Hawkes process – which will be mainly studied in the thesis.

1.1 Non-homogeneous Poisson processes

Poisson processes play a key role in both theory and applications of point processes. They form a widely used class of stochastic models for point patterns that arise in biology, ecology, engineering, and finance among many other disciplines. The relatively tractable form of the non-homogeneous Poisson process (NHPP) likelihood is one of the reasons for the popularity of NHPPs in applications involving point process data.

A NHPP on $\mathbb{R}^+$ can be defined through its intensity function, $\lambda(t)$, for $t \in \mathbb{R}^+$, a non-negative and locally integrable function such that: (a) for any bounded $B \subset \mathbb{R}^+$, the number of events in B, $N(B)$, is Poisson distributed with mean $\Lambda(B) = \int_B \lambda(u)\, \mathrm{d}u$; and (b) given $N(B) = n$, the times t_i, for $i = 1, ..., n$, that form the point pattern in B arise independently and identically distributed (i.i.d.) according to density $\lambda(t)/\Lambda(B)$. Consequently, the likelihood for the NHPP intensity function, based on the

point pattern $\{0 < t_1 < ... < t_n < T\}$ observed in time window $(0, T)$, is proportional to $\exp(-\int_0^T \lambda(u)\,\mathrm{d}u) \prod_{i=1}^n \lambda(t_i)$.

Theoretical background for the Poisson process can be found, for example, in Kingman (1993) and Daley and Vere-Jones (2003). Regarding Bayesian nonparametric modeling and inference, prior probability models have been developed for the NHPP mean measure (e.g., Lo, 1982, 1992), and mainly for the intensity function of NHPPs over time and/or space. Modeling methods for NHPP intensities include: mixtures of non-negative kernels with weighted gamma process priors for the mixing measure (e.g., Lo and Weng, 1989; Wolpert and Ickstadt, 1998; Ishwaran and James, 2004; Kang et al., 2014); piecewise constant functions driven by Voronoi tessellations with Markov random field priors (Heikkinen and Arjas, 1998, 1999); Gaussian process priors for logarithmic or logit transformations of the intensity (e.g., Møller et al., 1998; Brix and Diggle, 2001; Adams et al., 2009; Rodrigues and Diggle, 2012); and Dirichlet process mixtures for the NHPP density, i.e., the intensity function normalized in the observation window (e.g., Kottas, 2006; Kottas and Sansó, 2007; Taddy and Kottas, 2012).

1.2 Hawkes processes

The Hawkes process (HP), originally developed in Hawkes (1971a), 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; Ogata, 1998; Zhuang et al., 2002; Veen and Schoenberg, 2008), as well as in crime data modeling (e.g., Mohler et al., 2011, Mohler, 2014), finance (e.g., Fonseca and Zaatour, 2014; Hardiman et al., 2013; Rambaldi et al., 2015), biology (e.g., Balderama et al., 2012), and epidemiology (e.g., Meyer et al., 2012; Schoenberg et al., 2019).

The standard form of the HP conditional intensity is expressed as

$$\lambda^*(t) = \lambda(t \mid \mathcal{H}(t)) \; = \; \mu + \sum_{t_j<t} h(t - t_j), \quad t > 0 \tag{1.1}$$

where, as before, $\mathcal{H}(t)$ denotes the point process history up to time t, $\mu > 0$ is the background (immigrant) intensity, and $h : \mathbb{R}^+ \to \mathbb{R}^+$ is the excitation function, which accounts for the effect of previous events on the current intensity. The excitation function (offspring intensity) must be integrable, and it can thus be equivalently represented in terms of the branching ratio, $\gamma = \int_0^\infty h(u)\mathrm{d}u$, and the offspring density $f(t) = h(t)/\gamma$. The original definition of the HP, and several of its applications, focus on constant immigrant rate μ, in which case the HP is stationary provided $\gamma \in (0, 1)$. Our methodology of Chapter 3 handles also the more general case of a non-constant immigrant intensity function, $\mu(t) : \mathbb{R}^+ \to \mathbb{R}^+$.

The *immigrant* and *offspring* terminology originates from the HP cluster representation (Hawkes and Oakes, 1974), which is key for our modeling and inference

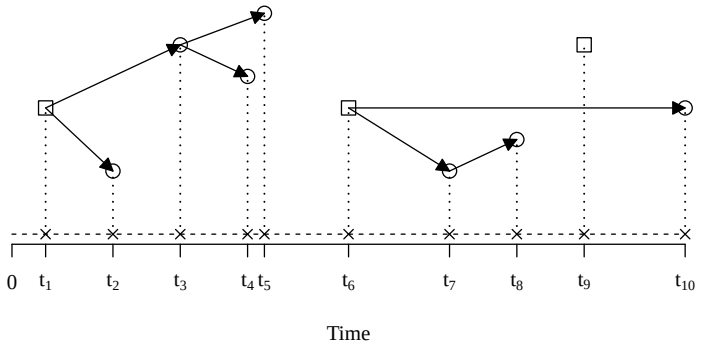

Figure 1.1: Illustration of the HP branching structure. Squares and circles indicate immigrant and offspring points. For example, t_1 is an immigrant and has two offspring points t_2 and t_3. The offspring t_3 also gives birth to two offspring points t_4 and t_5. Crosses at the bottom denote point observations, composing the HP point pattern.

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, $I = \{t_j : y_j = 0\}$, and sets of offspring, $O_j = \{t_i : y_i = j\}$, where O_j collects all offspring of t_j. For example, in Figure 1.1, the latent variables for time points t_2 and t_3 are $y_2 = 1$ and $y_3 = 1$ because t_1 begets the two points. Point t_1 is an immigrant point with $y_1 = 0$. The branching structure splits the point pattern into $I = \{t_1, t_6, t_9\}$, $O_1 = \{t_2, t_3\}$, $O_3 = \{t_4, t_5\}$, $O_6 = \{t_7, t_{10}\}$, and $O_7 = \{t_8\}$. 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 μ (or

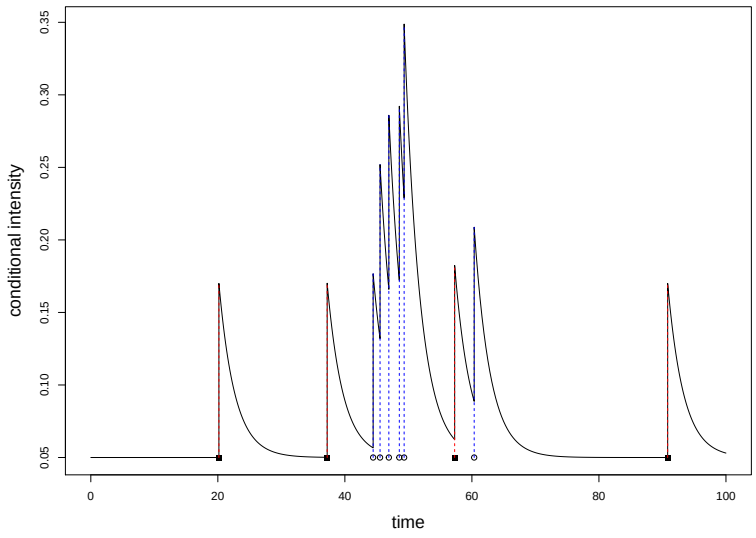

Figure 1.2: An example of a HP conditional intensity function. Squares and circles refer to immigrant and offspring time points. Red and blue dashed lines indicate the increase of conditional intensity at the immigrant or offspring points.

$\mu(t)$) and $h(t-t_j) = \gamma f(t-t_j)$ for I and O_j, respectively. Figure 1.2 shows an example of the conditional intensity function with the exponential density with rate 0.4 for $f(t)$, constant immigrant intensity $\mu = 0.05$, and branching ratio $\gamma = 0.3$.

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 constant immigrant intensity and simple parametric offspring intensities. The HP cluster representation provides a practically useful alternative, using the branching structure latent variables. Consider the general case with a time-varying immigrant intensity. Then, given $\boldsymbol{y}$, the

augmented likelihood can be expressed as

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

where $O=\cup_{j=1}^{n}O_j$ is the set of all offspring points. The second exponential term incorporates the probability that point t_j has no offspring in $(0,T)$, for all j with $O_j=\emptyset$.

Regarding classical inference for HPs, which is based mainly on maximum likelihood estimation, we refer to the reviews by Laub et al. (2015) and Reinhart (2018), the latter focusing on space-time HPs. Bayesian modeling and inference for HPs has received relatively less attention in the literature. Working with constant immigrant intensity and parametric forms for the excitation function, Rasmussen (2013) compared posterior simulation methods based on either the likelihood defined directly through the conditional intensity or the augmented likelihood that utilizes the branching structure. Donnet et al. (2020) studied Bayesian nonparametric priors for multivariate HPs, but their main contributions are theoretical, without particular emphasis on practical properties of the prior models. Zhang et al. (2018) introduced another Bayesian nonparametric model for HPs, focusing on the excitation function. They restrict the immigrant intensity to be constant to retain stationarity, required for their inference methods.

1.3 Motivation and objectives

The main objective of the dissertation is to provide a Bayesian nonparametric modeling and inference framework for Hawkes processes. We will introduce flexible

models for the immigrant intensity and the excitation function accompanied by efficient posterior inference methods for the HP conditional intensity function and other point process functionals. We will also describe quantitative tools for comprehensive model comparison. The models will be extended by adding marks for marked HPs, motivated by earthquake modeling applications.

As a building block for the priors for HP intensity functions, in Chapter 2 we develop a flexible and computationally efficient model for NHPPs. We focus on temporal intensities to motivate the modeling approach and to detail the methodological development, and then extend the model for spatial NHPPs. The NHPP intensity over time is represented as a weighted combination of Erlang densities indexed by their integer shape parameters and with a common scale parameter. Thus, different from existing mixture representations, the proposed mixture model is more structured with each Erlang density identified by the corresponding mixture weight. The non-negative mixture weights are defined through increments of a cumulative intensity on $\mathbb{R}^+$. Under certain conditions, the Erlang mixture intensity model can approximate in a pointwise sense general intensities on $\mathbb{R}^+$ (see Section 2.1.1). A gamma process prior is assigned to the primary model component, that is, the cumulative intensity that defines the mixture weights. Mixture weights driven by the gamma process prior result in flexible intensity function shapes, and, at the same time, ready prior-to-posterior updating given the observed point pattern. Indeed, a key feature of the model is that it can be implemented with an efficient Markov chain Monte Carlo (MCMC) algorithm that does not require approximations, complex computational methods, or restrictive prior modeling

assumptions in order to handle the NHPP likelihood normalizing term. The intensity model is extended to the two-dimensional setting through products of Erlang densities for the mixture components, with the weights built from a measure modeled again with a gamma process prior. The extension to spatial NHPPs retains the appealing aspect of computationally efficient MCMC posterior simulation.

Turning to HPs, in Chapter 3 we study models for temporal HPs without marks. According to the HP cluster representation, the process can be viewed in terms of independent NHPPs for immigrants and offspring. This representation allows us to use the Erlang mixture model for the immigrant intensity or, with modifications, for the excitation function, which yields an immigrant or an offspring semiparametric model. Unlike the former, the Erlang mixture for the excitation function must incorporate a stability condition. Therefore, we adjust the Erlang mixture with gamma priors having special-form shape parameters for mixture weights, which ensures the stability condition. The prior choice still provides the conjugacy with ready expressions for hyperparameters in posterior sampling. In addition to model flexibility and efficiency in model implementation, the single common scale parameter of the Erlang mixture facilitates parsimonious modeling.

The other semiparametric modeling framework is based on nonparametric uniform mixtures for the offspring density, based on the Dirichlet process (DP) or geometric weights prior. Such a model specializes in non-increasing offspring density estimation. This modeling approach is motivated by standard HP models, whose parametric offspring densities have decreasing patterns with either exponential or polynomial tail

behavior.

We also provide a fully nonparametric modeling framework for the HP conditional intensity function, constructed by combining the immigrant semiparametric model and either of the offspring semiparametric models. Such a model enables general inference for HPs, providing flexibility to both the immigrant intensity and the excitation function.

The HP model is further elaborated in Chapter 4 to include marks. The motivating application includes earthquake data modeling, where the mark is given by earthquake magnitude. The excitation function of the marked HP involves marks as well as time. To model the more general excitation function, we consider weighted combinations of basis functions that have a multiplicative form, consisting of an Erlang density for time and a polynomial function for the mark. We define the mixture weights through increments of a measure defined on $\mathbb{R}^+ \times \mathcal{K}$, where $\mathcal{K}$ denotes the mark space. A gamma process prior is assigned to the measure for model flexibility. The prior choice also enables tractable inference, producing (gamma) conjugate priors for mixture weights. Efficient handling of the normalizing constant is another key benefit, as in the Erlang mixture model for NHPPs. To our knowledge, the model is the first nonparametric method for marked HPs applied to earthquake data. In the context of our motivating application, the modeling approach enables estimation of aftershock densities that can vary with the magnitude of the main shock, unlike existing marked HP models for earthquake occurrences.

Chapter 2

Erlang Mixture Modeling for Poisson Process Intensities

2.1 Methodology for temporal Poisson processes

The mixture model for NHPP intensities is developed in Section 2.1.1, including a discussion of model properties and theoretical justification. Sections 2.1.2 and 2.1.3 present a prior specification approach and the posterior simulation method, respectively.

2.1.1 The mixture modeling approach

We develop a model for the NHPP intensity function, $\lambda(t)$. Denote by $\text{Ga}(\cdot \mid \alpha, \beta)$ the gamma density (or distribution, depending on the context) with mean α/β. The proposed intensity model involves a structured mixture of Erlang densities, $\text{Ga}(t \mid j, \theta^{-1})$, mixing on the integer shape parameters, j, with a common scale parameter

θ. The non-negative mixture weights are defined through increments of a cumulative intensity function, H, on $\mathbb{R}^+$, which is assigned a gamma process prior. More specifically,

$$\begin{gathered}\lambda(t) \equiv \lambda(t \mid H, \theta) \;=\; \sum_{j=1}^{J} \omega_j \,\mathrm{Ga}(t \mid j, \theta^{-1}), \;\; t \in \mathbb{R}^+ \\ \omega_j = H(j\theta) - H((j-1)\theta), \quad H \sim \mathcal{G}(H_0, c_0),\end{gathered} \tag{2.1}$$

where $\mathcal{G}(H_0, c_0)$ is a gamma process specified through H_0, a (parametric) cumulative intensity function, and c_0, a positive scalar parameter (Kalbfleisch, 1978). A (real-valued) stochastic process $H = (H(t); t \geq 0)$ follows the *(time-inhomogeneous) gamma process* if $H(0) = 0$, and if it has independent, gamma distributed increments, i.e., $H(t) - H(s)$ is gamma distributed with shape $D_0(t) - D_0(s)$ for $s < t$ and rate $c_0 > 0$, where D_0 is a general increasing and everywhere right continuous function and has left limits everywhere. We adopt the following parameterization $c_0 H_0 = D_0$. Accordingly, for any $t \in \mathbb{R}^+$, the gamma process H has $\mathrm{E}(H(t)) = H_0(t)$ and $\mathrm{Var}(H(t)) = H_0(t)/c_0$, and thus H_0 plays the role of the centering cumulative intensity, whereas c_0 is a precision parameter. As an independent increments process, the $\mathcal{G}(H_0, c_0)$ prior for H implies that, given θ, the mixture weights are independent $\mathrm{Ga}(\omega_j \mid c_0\, \omega_{0j}(\theta), c_0)$ distributed, where $\omega_{0j}(\theta) = H_0(j\theta) - H_0((j-1)\theta)$. As shown in Section 2.1.3, this is a key property of the prior model with respect to implementation of posterior inference.

The model in (2.1) is motivated by Erlang mixtures for density estimation, under which a density g on $\mathbb{R}^+$ is represented as $g(t) \equiv g_{J,\theta}(t) = \sum_{j=1}^{J} p_j \,\mathrm{Ga}(t \mid j, \theta^{-1})$, for $t \in \mathbb{R}^+$. Here, $p_j = G(j\theta) - G((j-1)\theta)$, where G is a distribution function on $\mathbb{R}^+$; the last weight can be defined as $p_J = 1 - G((J-1)\theta)$ to ensure that $(p_1, ..., p_J)$ is a probability vector. Erlang mixtures can approximate general densities on the positive

real line, in particular, as $\theta \to 0$ and $J \to \infty$, $g_{J,\theta}$ converges pointwise to the density of distribution function G that defines the mixture weights. This convergence property can be obtained from more general results from the probability literature that studies Erlang mixtures as extensions of Bernstein polynomials to the positive real line (e.g., Butzer, 1954); a convergence proof specifically for the distribution function of $g_{J,\theta}$ can be found in Lee and Lin (2010). Density estimation on compact sets via Bernstein polynomials has been explored in the Bayesian nonparametrics literature following the work of Petrone (1999a,b). Regarding Bayesian nonparametric modeling with Erlang mixtures, we are only aware of Xiao et al. (2021) where renewal process inter-arrival distributions are modeled with mixtures of Erlang distributions, using a Dirichlet process prior (Ferguson, 1973) for distribution function G. Venturini et al. (2008) study a parametric Erlang mixture model for density estimation on $\mathbb{R}^+$, working with a Dirichlet prior distribution for the mixture weights.

Therefore, the modeling approach in (2.1) exploits the structure of the Erlang mixture density model to develop a prior for NHPP intensities, using the density/distribution function and intensity/cumulative intensity function connection to define the prior model for the mixture weights. In this context, the gamma process prior for cumulative intensity H is the natural analogue to the Dirichlet process prior for distribution function G; recall that the Dirichlet process can be defined through normalization of a gamma process (e.g., Ghosal and van der Vaart, 2017). To our knowledge, this is a novel construction for NHPP intensities that has not been explored for intensity estimation in either the classical or Bayesian nonparametrics literature. The following lemma,

which can be obtained applying Theorem 2 from Butzer (1954), provides theoretical motivation and support for the mixture model.

Lemma. Let h be the intensity function of a NHPP on $\mathbb{R}^+$, with cumulative intensity function $H(t) = \int_0^t h(u)\,\mathrm{d}u$, such that $H(t) = O(t^m)$, as $t \to \infty$, for some $m > 0$. Consider the mixture intensity model $\lambda_{J,\theta}(t) = \sum_{j=1}^{J}\{H(j\theta) - H((j-1)\theta)\}\,\mathrm{Ga}(t \mid j, \theta^{-1})$, for $t \in \mathbb{R}^+$. Then, as $\theta \to 0$ and $J \to \infty$, $\lambda_{J,\theta}(t)$ converges to $h(t)$ at every point t where $h(t) = \mathrm{d}H(t)/\mathrm{d}t$.

The form of the prior model for the intensity in (2.1) allows ready expressions for other NHPP functionals. For instance, the total intensity over the observation time window $(0, T)$ is given by $\int_0^T \lambda(u)\,\mathrm{d}u = \sum_{j=1}^{J} \omega_j K_{j,\theta}(T)$, where $K_{j,\theta}(T) = \int_0^T \mathrm{Ga}(u \mid j, \theta^{-1})\,\mathrm{d}u$ is the j-th Erlang distribution function at T. In the context of the MCMC posterior simulation method, this form enables efficient handling of the NHPP likelihood normalizing constant. Moreover, the NHPP density on interval $(0, T)$ can be expressed as a mixture of truncated Erlang densities. More specifically,

$$f(t) = \frac{\lambda(t)}{\int_0^T \lambda(u)\,\mathrm{d}u} = \sum_{j=1}^{J} \omega_j^*\, k(t \mid j, \theta), \;\; t \in (0, T), \tag{2.2}$$

where $\omega_j^* = \omega_j K_{j,\theta}(T)/\{\sum_{r=1}^{J} \omega_r K_{r,\theta}(T)\}$, and $k(t \mid j, \theta)$ is the j-th Erlang density truncated on $(0, T)$.

Regarding the role of the different model parameters, we reiterate that (2.1) corresponds to a structured mixture. The Erlang densities, $\mathrm{Ga}(t \mid j, \theta^{-1})$, play the role of basis functions in the representation for the intensity. In this respect, of primary importance is the flexibility of the nonparametric prior for the cumulative intensity

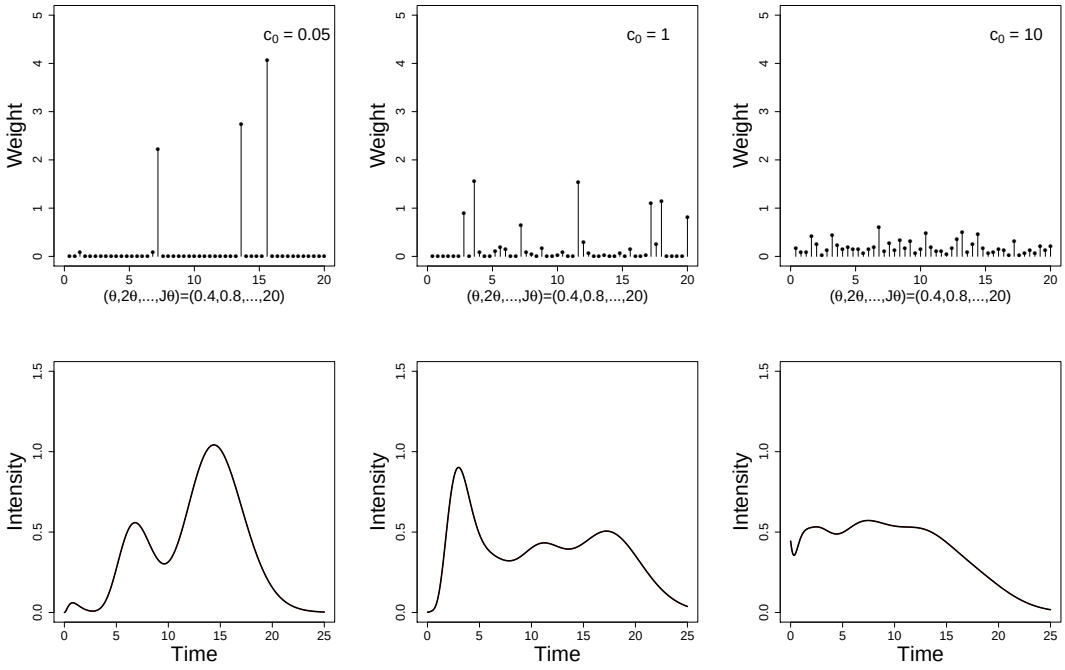

Figure 2.1: Prior realizations for the mixture weights (top panels) and the corresponding intensity function (bottom panels) for three different values of the gamma process precision parameter, $c_0 = 0.05, 1, 10$. In all cases, $J = 50$, $\theta = 0.4$, and $H_0(t) = t/2$.

function H that defines the mixture weights. In particular, the gamma process prior provides realizations for H with general shapes that can concentrate on different time intervals, thus favoring different subsets of the Erlang basis densities through the corresponding ω_j. Here, the key parameter is the precision parameter c_0, which controls the variability of the gamma process prior around H_0, and thus the effective mixture weights. As an illustration, Figure 2.1 shows prior realizations for the weights ω_j (and the resulting intensity function) for different values of c_0, keeping all other model parameters the same. Note that as c_0 decreases, so does the number of practically non-zero weights.

The prior mean for H is taken to be $H_0(t) = t/b$, i.e., the cumulative intensity (hazard) of an exponential distribution with scale parameter $b > 0$. Although it is possible to use more general centering functions, such as the Weibull $H_0(t) = (t/b)^a$, the exponential form is sufficiently flexible in practice, as demonstrated with the synthetic data examples of Section 2.2. Based on the role of H in the intensity mixture model, we typically anticipate realizations for H that are different from the centering function H_0, and thus, as discussed above, the more important gamma process parameter is c_0. Moreover, the exponential form for H_0 allows for an analytical result for the prior expectation of the Erlang mixture intensity model. Under $H_0(t) = t/b$, the prior expectation for the weights is given by $\mathrm{E}(\omega_j \mid \theta, b) = \theta/b$. Therefore, conditional on all model hyperparameters, the expectation of $\lambda(t)$ over the gamma process prior can be written as

$$\mathrm{E}(\lambda(t) \mid b, \theta) = \frac{\theta}{b} \sum_{j=1}^{J} \mathrm{Ga}(t \mid j, \theta^{-1}) = \frac{\exp(-(t/\theta))}{b} \sum_{m=0}^{J-1} \frac{(t/\theta)^m}{m!}, \quad t \in \mathbb{R}^+,$$

which converges to b^{-1}, as $J \to \infty$, for any $t \in \mathbb{R}^+$ (and regardless of the value of θ and c_0). In practice, the prior mean for the intensity function is essentially constant at b^{-1} for $t \in (0, J\theta)$, which, as discussed below, is roughly the effective support of the NHPP intensity. This result is useful for prior specification as it distinguishes the role of b from that of parameters θ and c_0.

Also key are the two remaining model parameters, the number of Erlang basis densities J, and their common scale parameter θ. Parameters θ and J interact to control both the effective support and shape of NHPP intensities arising under (2.1). Regarding intensity shapes, as the lemma suggests, smaller values of θ and larger values

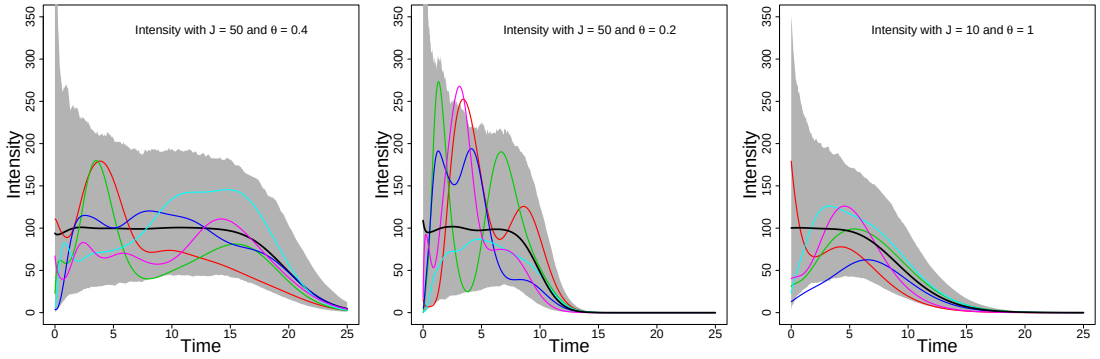

Figure 2.2: Prior mean (black line), prior 95% interval bands (shaded area), and five individual prior realizations for the intensity under the Erlang mixture model in (2.1) with $(\theta, J) = (0.4, 50)$ (left panel), $(\theta, J) = (0.2, 50)$ (middle panel), and $(\theta, J) = (1, 10)$ (right panel). In all cases, the gamma process prior is specified with $c_0 = 0.01$ and $H_0(t) = t/0.01$.

of J generally result in more variable, typically multimodal intensities. Moreover, the representation for $\lambda(t)$ in (2.1) utilizes Erlang basis densities with increasing means $j\theta$, and thus $(0, J\theta)$ can be used as a proxy for the effective support of the NHPP intensity. Of course, the mean underestimates the effective support, a more accurate guess can be obtained using, say, the 95% percentile of the last Erlang density component. For an illustration, Figure 2.2 plots five prior intensity realizations under three combinations of (θ, J) values, with $c_0 = 0.01$ and $b = 0.01$ in all cases. Also plotted are the prior mean and 95% interval bands for the intensity, based on 1000 realizations from the prior model. The left panel corresponds to the largest value for $J\theta$ and, consequently, to the widest effective support interval. The value of $J\theta$ is the same for the middle and right panels, resulting in similar effective support. However, the intensities in the middle panel show larger variability in their shapes, as expected since the value of J is increased and the value of θ decreased relative to the ones in the right panel.

2.1.2 Prior specification

To complete the full Bayesian model, we place prior distributions on the parameters c_0 and b of the gamma process prior for H, and on the scale parameter θ of the Erlang basis densities. A generic approach to specify these hyperpriors can be obtained using the observation time window $(0, T)$ as the effective support of the NHPP intensity.

We work with exponential prior distributions for parameters c_0 and b. Using the prior mean for the intensity function, which as discussed in Section 2.1.1 is roughly constant at b^{-1} within the time interval of interest, the total intensity in $(0, T)$ can be approximated by T/b. Therefore, taking the size n of the observed point pattern, as a proxy for the total intensity in $(0, T)$, we can use T/n to specify the mean of the exponential prior distribution for b. Given its role in the gamma process prior, we anticipate that small values of c_0 will be important to allow prior variability around H_0, as well as sparsity in the mixture weights. Experience from prior simulations, such as the ones shown in Figure 2.1, is useful to guide the range of "small" values. Note that the pattern observed in Figure 2.1 is not affected by the length of the observation window. In general, a value around 10 can be viewed as a conservative guess at a high percentile for c_0. For the data examples of Section 2.2, we assigned an exponential prior with mean 10 to c_0, observing substantial learning for this key model hyperparameter with its posterior distribution supported by values (much) smaller than 1.

Also given the key role of parameter θ in controlling the intensity shapes, we recommend favoring sufficiently small values in the prior for θ, especially if prior infor-

mation suggests a non-standard intensity shape. Recall that θ, along with J, control the effective support of the intensity, and thus "small" values for θ should be assessed relative to the length of the observation window. Again, prior simulation, as in Figure 2.2, is a useful tool. A practical approach to specify the prior range of θ values involves reducing the Erlang mixture model to the first component. The corresponding (exponential) density has mean θ, and we thus use $(0, T)$ as the effective prior range for θ. Because T is a fairly large upper bound, and since we wish to favor smaller θ values, rather than an exponential prior, we use a Lomax prior, $p(\theta) \propto (1+d_\theta^{-1}\theta)^{-3}$, with shape parameter equal to 2 (thus implying infinite variance), and median $d_\theta(\sqrt{2}-1)$. The value of the scale parameter, d_θ, is specified such that $\Pr(0 < \theta < T) \approx 0.999$. This simple strategy is effective in practice in identifying a plausible range of θ values. For the synthetic data examples of Section 2.2, for which $T = 20$, we assigned a Lomax prior with scale parameter $d_\theta = 1$ to θ, obtaining overall moderate prior-to-posterior learning for θ.

Finally, we work with fixed J, the value of which can be specified exploiting the role of θ and J in controlling the support of the NHPP intensity. In particular, J can be set equal to the integer part of T/θ^*, where θ^* is the prior median for θ. More conservatively, this value can be used as a lower bound for values of J to be studied in a sensitivity analysis, especially for applications where one expects non-standard shapes for the intensity function. In practice, we recommend conducting prior sensitivity analysis for all model parameters, as well as plotting prior realizations and prior uncertainty bands for the intensity function to graphically explore the implications

of different prior choices.

The number of Erlang basis densities is the only model parameter which is not assigned a hyperprior. Placing a prior on J complicates significantly the posterior simulation method, as it necessitates use of variable-dimension MCMC techniques, while offering relatively little from a practical point of view. The key observation is again that the Erlang densities play the role of basis functions rather than of kernel densities in traditional (less structured) finite mixture models. Also key is the nonparametric nature of the prior for function H that defines the mixture weights which *select* the Erlang densities to be used in the representation of the intensity. This model feature effectively guards against over-fitting if one conservatively chooses a larger value for J than may be necessary. In this respect, the flexibility afforded by random parameters c_0 and θ is particularly useful. Overall, we have found that fixing J strikes a good balance between computational tractability and model flexibility in terms of the resulting inferences.

2.1.3 Posterior simulation

Denote as before by $\{0 < t_1 < ... < t_n < T\}$ the point pattern observed in time window $(0, T)$. Under the Erlang mixture model of Section 2.1.1, the NHPP likelihood is proportional to $\exp\left(-\int_0^T \lambda(u)\,\mathrm{d}u\right) \prod_{i=1}^n \lambda(t_i)$

$$
\begin{aligned}
&= \exp\left(-\sum\nolimits_{j=1}^{J} \omega_j K_{j,\theta}(T)\right) \prod_{i=1}^{n} \left\{ \sum_{j=1}^{J} \omega_j \,\mathrm{Ga}(t_i \mid j, \theta^{-1}) \right\} \\
&= \prod_{j=1}^{J} \exp(-\omega_j K_{j,\theta}(T)) \prod_{i=1}^{n} \left\{ \left(\sum\nolimits_{r=1}^{J} \omega_r\right) \sum_{j=1}^{J} \left(\frac{\omega_j}{\sum_{r=1}^{J} \omega_r} \right) \mathrm{Ga}(t_i \mid j, \theta^{-1}) \right\},
\end{aligned}
$$

where $K_{j,\theta}(T) = \int_0^T \mathrm{Ga}(u \mid j, \theta^{-1})\,\mathrm{d}u$ is the j-th Erlang distribution function at T.

For the posterior simulation approach, we augment the likelihood with auxiliary variables $\boldsymbol{\gamma} = \{\gamma_i : i = 1, \ldots, n\}$, where γ_i identifies the Erlang basis density to which time event t_i is assigned. Then, the augmented, hierarchical model for the data can be expressed as follows:

$$\begin{aligned}
\{t_1, ..., t_n\} \mid \boldsymbol{\gamma}, \boldsymbol{\omega}, \theta &\sim \prod_{j=1}^{J} \exp(-\omega_j K_{j,\theta}(T)) \prod_{i=1}^{n} \left\{ \left(\textstyle\sum_{r=1}^{J} \omega_r \right) \mathrm{Ga}(t_i \mid \gamma_i, \theta^{-1}) \right\} \\
\gamma_i \mid \boldsymbol{\omega} &\overset{i.i.d.}{\sim} \sum_{j=1}^{J} \left(\frac{\omega_j}{\sum_{r=1}^{J} \omega_r} \right) \delta_j(\gamma_i), \quad i = 1, ..., n \\
\theta, c_0, b, \boldsymbol{\omega} &\sim p(\theta)\, p(c_0)\, p(b) \prod_{j=1}^{J} \mathrm{Ga}(\omega_j \mid c_0\, \omega_{0j}(\theta), c_0), \qquad (2.3)
\end{aligned}$$

where $\boldsymbol{\omega} = \{\omega_j : j = 1, ..., J\}$, and $p(\theta)$, $p(c_0)$, and $p(b)$ denote the priors for θ, c_0, and b. Recall that, under the exponential distribution form for $H_0 = t/b$, we have $\omega_{0j}(\theta) = \theta/b$.

We utilize Gibbs sampling to explore the posterior distribution. The sampler involves ready updates for the auxiliary variables γ_i, and, importantly, also for the mixture weights ω_j. More specifically, the posterior full conditional for each γ_i is a discrete distribution on $\{1, ..., J\}$ such that $\Pr(\gamma_i = j \mid \theta, \boldsymbol{\omega}, \text{data}) \propto \omega_j\, \mathrm{Ga}(t_i \mid j, \theta^{-1})$, for $j = 1, ..., J$.

Denote by $N_j = |\{t_i : \gamma_i = j\}|$, for $j = 1, ..., J$, that is, N_j is the number of time points assigned to the j-th Erlang basis density. The posterior full conditional

distribution for ω is derived as follows:

$$\begin{aligned} p(\boldsymbol{\omega} \mid \theta, c_0, b, \boldsymbol{\gamma}, \text{data}) &\propto \left\{ \prod_{j=1}^{J} \exp(-\omega_j K_{j,\theta}(T)) \right\} \left(\sum\nolimits_{r=1}^{J} \omega_r \right)^n \\ &\times \left\{ \prod_{j=1}^{J} \omega_j^{N_j} \left(\sum\nolimits_{r=1}^{J} \omega_r \right)^{-N_j} \right\} \left\{ \prod_{j=1}^{J} \text{Ga}(\omega_j \mid c_0\, \omega_{0j}(\theta), c_0) \right\} \\ &\propto \prod_{j=1}^{J} \exp(-\omega_j K_{j,\theta}(T))\, \omega_j^{N_j}\, \text{Ga}(\omega_j \mid c_0\, \omega_{0j}(\theta), c_0) \\ &= \prod_{j=1}^{J} \text{Ga}(\omega_j \mid N_j + c_0\, \omega_{0j}(\theta), K_{j,\theta}(T) + c_0), \end{aligned}$$

where we have used the fact that $\sum_{j=1}^{J} N_j = n$. Therefore, given the other parameters and the data, the mixture weights are independent, and each ω_j follows a gamma posterior full conditional distribution. This is a practically important feature of the model in terms of convenient updates for the mixture weights, and with respect to efficiency of the posterior simulation algorithm as it pertains to this key component of the model parameter vector.

Finally, each of the remaining parameters, c_0, b, and θ, is updated with a Metropolis-Hastings (M-H) step, using a log-normal proposal distribution in each case.

2.1.4 Model extensions to incorporate marks

Here, we discuss how the Erlang mixture prior for NHPP intensities can be embedded in semiparametric models for point patterns that include additional information on marks.

Consider the setting where, associated with each observed time event t_i, marks $\boldsymbol{y}_i \equiv \boldsymbol{y}_{t_i}$ are recorded (marks are only observed when an event is observed). Without loss

of generality, we assume that marks are continuous variables taking values in mark space $\mathcal{M} \subseteq \mathbb{R}^d$, for $d \geq 1$. As discussed in Taddy and Kottas (2012), a nonparametric prior for the intensity of the temporal process, $\mathcal{T}$, can be combined with a mark distribution to construct a semiparametric model for marked NHPPs. In particular, consider a generic marked NHPP $\{(t, \boldsymbol{y}_t) : t \in \mathcal{T}, \boldsymbol{y}_t \in \mathcal{M}\}$, that is: the temporal process $\mathcal{T}$ is a NHPP on $\mathbb{R}^+$ with intensity function λ; and, conditional on $\mathcal{T}$, the marks $\{\boldsymbol{y}_t : t \in \mathcal{T}\}$ are mutually independent. Now, assume that, conditional on $\mathcal{T}$, the marks have density m_t that depends only on t (i.e., it does not depend on any earlier time $t' < t$). Then, by the "marking" theorem (e.g., Kingman, 1993), we have that the marked NHPP is a NHPP on the extended space $\mathbb{R}^+ \times \mathcal{M}$ with intensity $\lambda^*(t, \boldsymbol{y}_t) = \lambda(t)\, m_t(\boldsymbol{y}_t)$. Therefore, the likelihood for the observed marked point pattern $\{(t_i, \boldsymbol{y}_i) : i = 1, ..., n\}$ can be written as $\exp\left(-\int_0^T \lambda(u)\,\mathrm{d}u\right) \prod_{i=1}^n \lambda(t_i) \prod_{i=1}^n m_{t_i}(\boldsymbol{y}_i)$ (the integral $\int_0^T \int_{\mathcal{M}} \lambda^*(u, \boldsymbol{z})\,\mathrm{d}u\mathrm{d}\boldsymbol{z}$ in the normalizing term reduces to $\int_0^T \lambda(u)\,\mathrm{d}u$, since m_t is a density). Hence, the MCMC method of Section 2.1.3 can be extended for marked NHPP models built from the Erlang mixture prior for intensity λ, and any time-dependent model for the mark density m_t.

2.2 Data examples

To empirically investigate inference under the proposed model, we present three synthetic data examples corresponding to decreasing, increasing, and bimodal intensities. We also consider the coal-mining disasters data set, which is commonly used to illustrate NHPP intensity estimation.

We used the approach of Section 2.1.2 to specify the priors for c_0, b and θ, and the value for J. In particular, we used the exponential prior for c_0 with mean 10 for all data examples. For the three synthetic data sets (for which $T = 20$), we used the Lomax prior for θ with shape parameter equal to 2 and scale parameter equal to 1. Prior sensitivity analysis results are provided in Section 2.2.3. Overall, results from prior sensitivity analysis (also conducted for all other data examples) suggest that the prior specification approach of Section 2.1.2 is effective as a general strategy. Moreover, more dispersed priors for parameters c_0, b and θ have little to no effect on the posterior distribution for these parameters and essentially no effect on posterior estimates for the NHPP intensity function, even for point patterns with relatively small size, such as the one ($n = 112$) for the data example of Section 2.2.3.

Section 2.2.5 compares our model with a Bayesian nonparametric model based on a Gaussian process (GP) prior for the logit transformation of the NHPP intensity, called the sigmoidal Gaussian Cox process (SGCP) model (Adams et al., 2009). Such a model is popular in both the statistics and the machine learning literature. Since we were not able to find publicly available software, results are based on our implementation of the SGCP model, which allows for a detailed comparison involving full inference results and computational efficiency.

Appendix A provides also computational details about the MCMC posterior simulation algorithm, including study of the effect of the number of basis densities (J) and the size of the point pattern (n) on effective sample size and computing time.

2.2.1 Decreasing intensity synthetic point pattern

The first synthetic data set involves 491 time points generated in time window $(0, 20)$ from a NHPP with intensity function $\beta^{-1}\alpha(\beta^{-1}t)^{\alpha-1}$, where $(\alpha, \beta) = (0.5, 8 \times 10^{-5})$. This form corresponds to the hazard function of a Weibull distribution with shape parameter less than 1, thus resulting in a decreasing intensity function.

The Erlang mixture model was applied with $J = 50$, and an exponential prior for b with mean 0.04. The model captures the decreasing pattern of the data generating intensity function; see Figure 2.3. We note that there is significant prior-to-posterior learning in the intensity function estimation; the prior intensity mean is roughly constant at value about 25 with prior uncertainty bands that cover almost the entire top left panel in Figure 2.3. Prior uncertainty bands were similarly wide for all other data examples.

2.2.2 Increasing intensity synthetic point pattern

We consider again the form $\beta^{-1}\alpha(\beta^{-1}t)^{\alpha-1}$ for the NHPP intensity function, but here with $(\alpha, \beta) = (6, 7)$ such that the intensity is increasing. A point pattern comprising 565 points was generated in time window $(0, 20)$. The Erlang mixture model was applied with $J = 50$, and an exponential prior for b with mean 0.035. Figure 2.4 reports inference results. This example demonstrates the model's capacity to effectively recover increasing intensity shapes over the bounded observation window, even though the Erlang basis densities are ultimately decreasing.

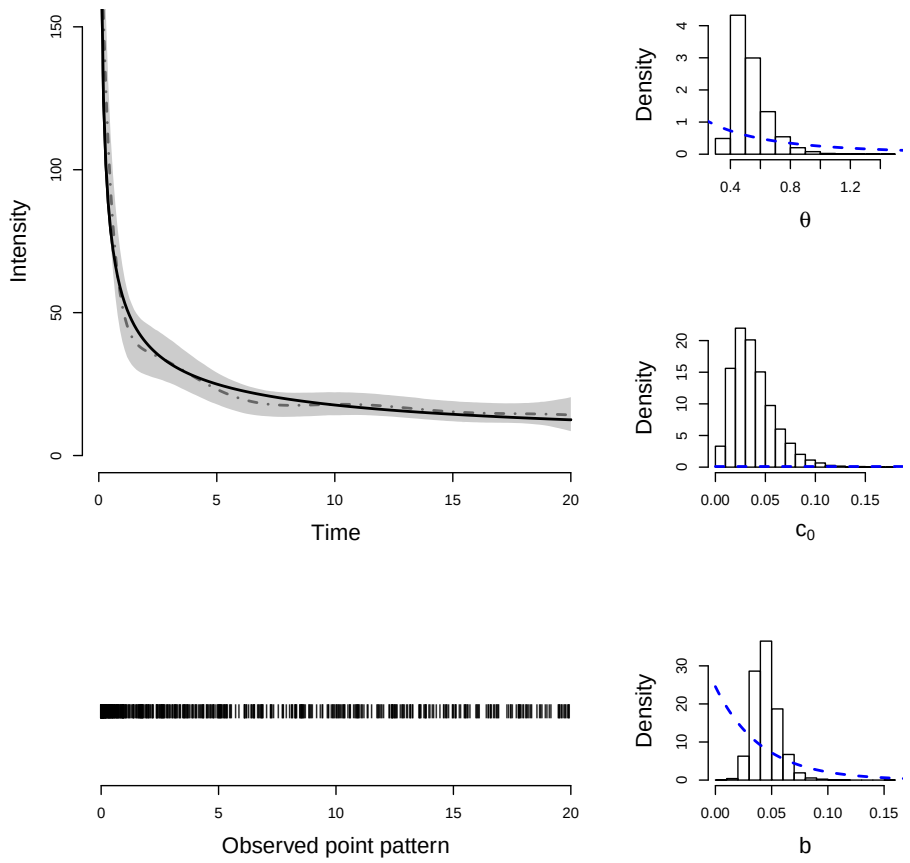

Figure 2.3: Synthetic data from temporal NHPP with decreasing intensity. The top left panel shows the posterior mean estimate (dashed-dotted line) and posterior 95% interval bands (shaded area) for the intensity function. The true intensity is denoted by the solid line. The point pattern is plotted in the bottom left panel. The three plots on the right panels display histograms of the posterior samples for the model hyperparameters, along with the corresponding prior densities (dashed lines).

2.2.3 Bimodal intensity synthetic point pattern

The data examples in Sections 2.2.1 and 2.2.2 illustrate the model's capacity to uncover monotonic intensity shapes, associated with a parametric distribution different from the Erlang distribution that forms the basis of the mixture intensity model. Here,

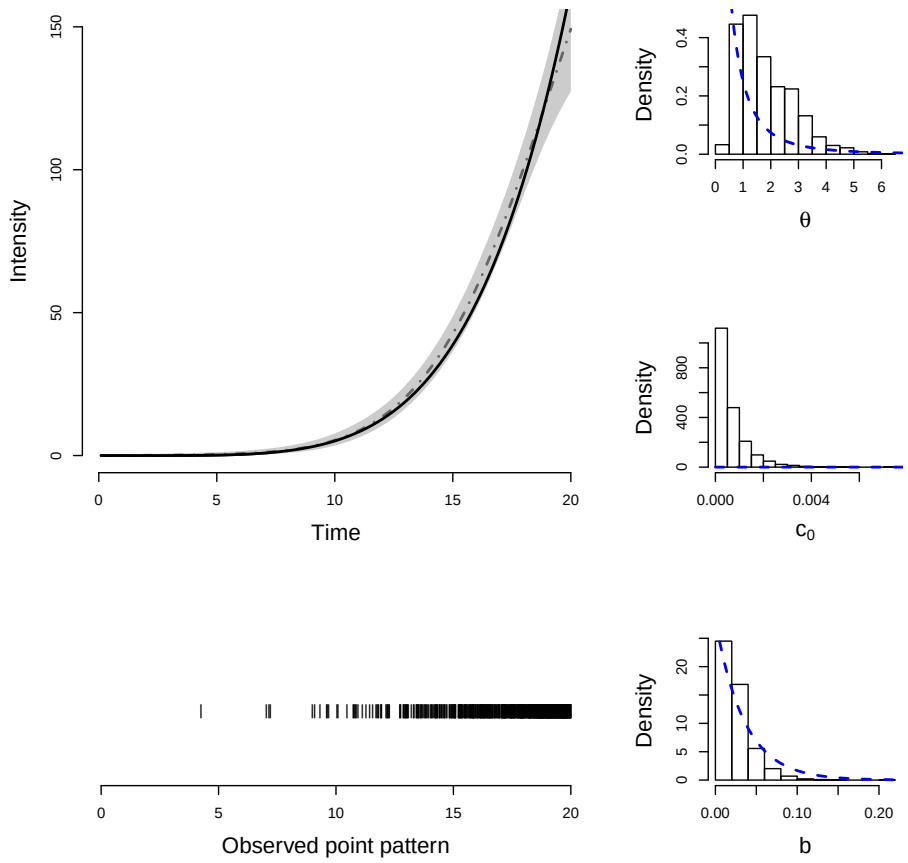

Figure 2.4: Synthetic data from temporal NHPP with increasing intensity. The top left panel shows the posterior mean estimate (dashed-dotted line) and posterior 95% interval bands (shaded area) for the intensity function. The true intensity is denoted by the solid line. The point pattern is plotted in the bottom left panel. The three plots on the right panels display histograms of the posterior samples for the model hyperparameters, along with the corresponding prior densities (dashed lines).

we consider a point pattern generated from a NHPP with a more complex intensity function, $\lambda(t) = 50\,\mathrm{We}(t \mid 3.5, 5) + 60\,\mathrm{We}(t \mid 6.5, 15)$, where $\mathrm{We}(t \mid \alpha, \beta)$ denotes the Weibull density with shape parameter α and mean $\beta\,\Gamma(1 + 1/\alpha)$. This specification results in a bimodal intensity within the observation window $(0, 20)$ where a synthetic

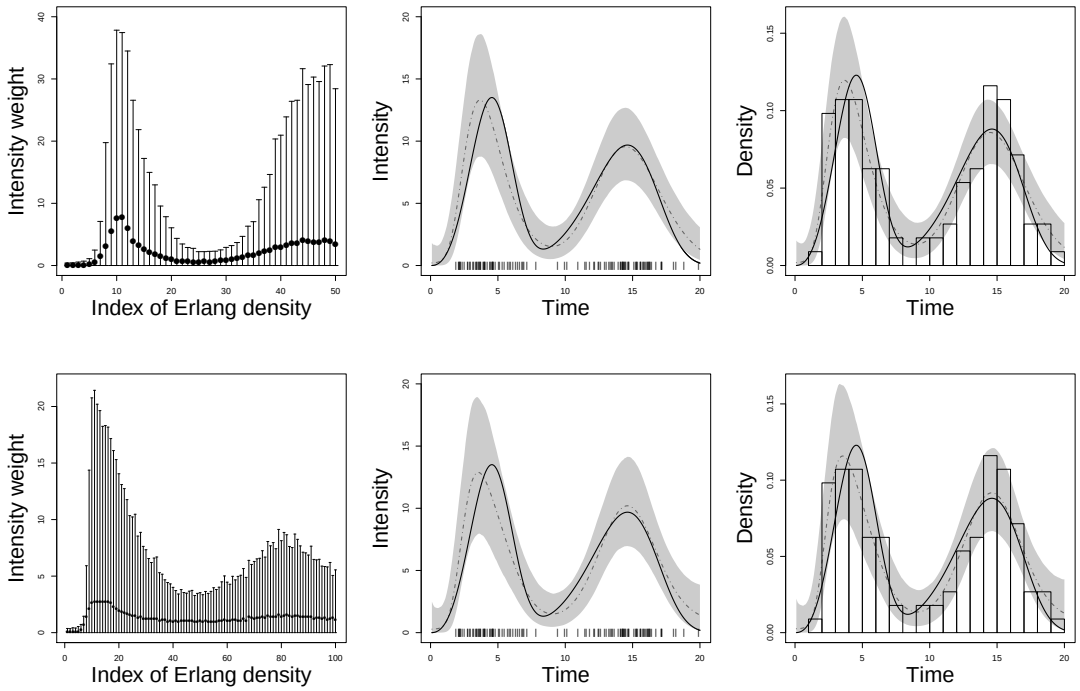

Figure 2.5: Synthetic data from temporal NHPP with bimodal intensity. Inference results are reported under $J = 50$ (top row) and $J = 100$ (bottom row). The left column plots the posterior means (circles) and 90% interval estimates (bars) of the weights for the Erlang basis densities. The middle column displays the posterior mean estimate (dashed-dotted line) and posterior 95% interval bands (shaded area) for the NHPP intensity function. The true intensity is denoted by the solid line. The bars on the horizontal axis indicate the point pattern. The right column plots the posterior mean estimate (dashed-dotted line) and posterior 95% interval bands (shaded area) for the NHPP density function on the observation window. The histogram corresponds to the simulated times that comprise the point pattern.

point pattern of 112 time points is generated; see Figure 2.5.

We used an exponential prior for b with mean 0.179. Anticipating an underlying intensity with less standard shape than in the earlier examples, we compare inference results under $J = 50$ and $J = 100$; see Figure 2.5. The posterior point and interval estimates capture effectively the bimodal intensity shape, especially if one takes into ac-

count the relatively small size of the point pattern. (In particular, the histogram of the simulated random time points indicates that they do not provide an entirely accurate depiction of the underlying NHPP density shape.) The estimates are somewhat more accurate under $J = 100$. The estimates for the mixture weights (left column of Figure 2.5) indicate the subsets of the Erlang basis densities that are utilized under the two different values for J. The posterior mean of θ was 0.366 under $J = 50$, and 0.258 under $J = 100$, that is, as expected, inference for θ adjusts to different values of J such that $(0, J\theta)$ provides roughly the effective support of the intensity.

We present results from sensitivity analysis to the prior choices for θ, c_0, and b. We focus on this data set, mainly because it involves the smallest sample size ($n = 112$) among all data examples, but also due to the non-standard, bimodal shape of the underlying non-homogeneous Poisson process (NHPP) intensity function. We have conducted prior sensitivity analysis for all other data examples observing levels of robustness to the prior choice that are either higher or the same with the ones reported here.

Here, we study the effect of the priors for θ, c_0, and b, under $J = 50$. Inference results for the intensity function and for the model parameters are reported in Figure 2.6, under four different prior choices. The top row corresponds to the prior specification approach described in Section 2.1.2, and thus to the earlier results (under $J = 50$). For each of the other three cases, we change one of the priors for θ, c_0, and b relative to the "default" prior specification. In particular, results in the second row are based on a more dispersed Lomax prior for θ, with scale parameter $d_\theta = 9$ (instead of $d_\theta = 1$), such that $\Pr(0 < \theta < T) \approx 0.9$ (instead of $\Pr(0 < \theta < T) \approx 0.999$), where $T = 20$.

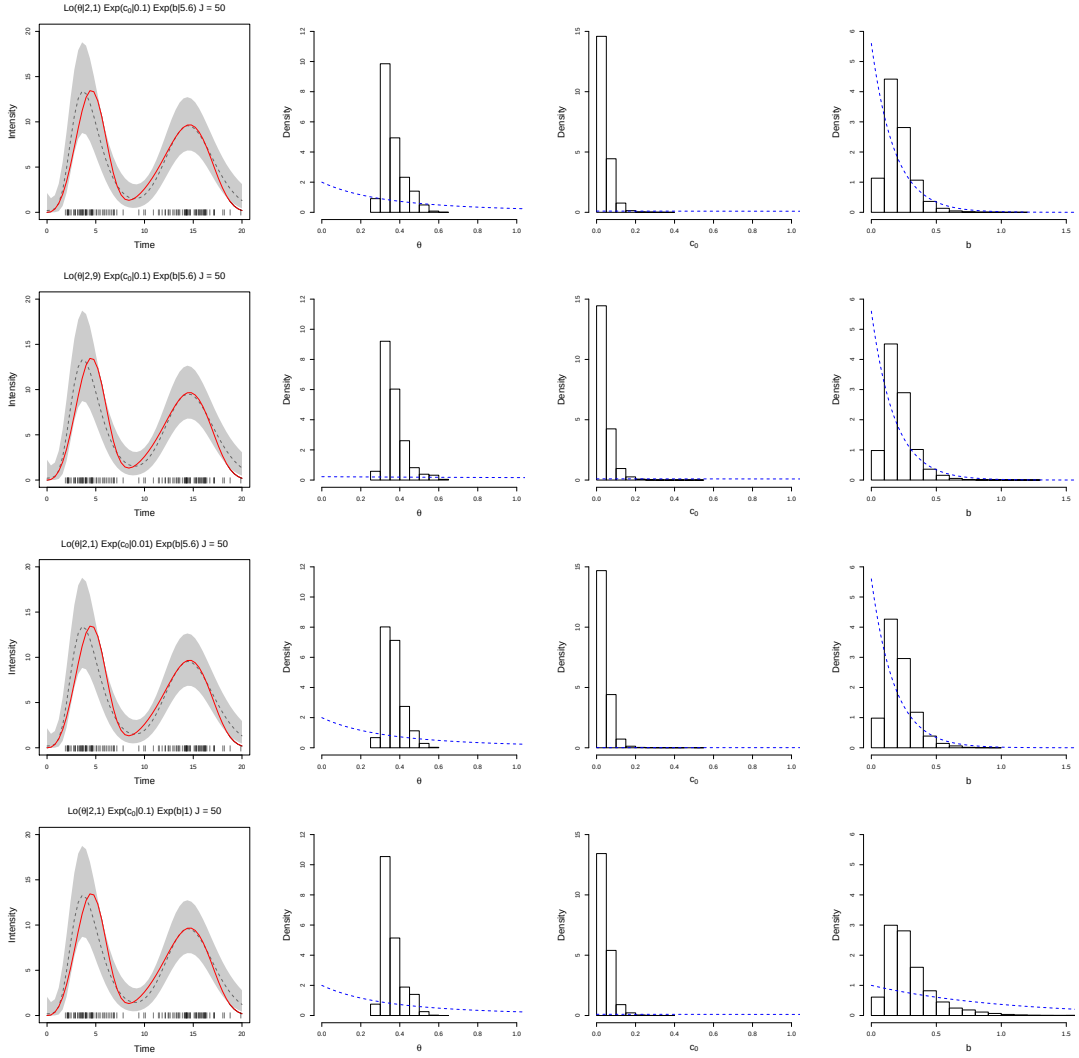

Figure 2.6: Synthetic data from temporal NHPP with bimodal intensity. Erlang mixture model inference results under four different prior choices for θ, c_0, and b. The left column shows the posterior mean estimate (dashed line) and posterior 95% interval bands (shaded area) for the intensity function, with the true intensity denoted by the red solid line. The other three columns plot histograms of the posterior samples for θ, c_0, and b, along with the corresponding prior densities (blue dashed lines).

Analogously, the third and fourth rows correspond to more dispersed exponential priors for c_0 and b, respectively. Note that the posterior distribution for θ and c_0 is largely unaffected by the change in the dispersion of the prior distribution, whereas there is

some effect on the tail of the posterior density for b. Importantly, for a point pattern with a relatively small size, posterior estimates for the intensity function are essentially the same under the different prior choices.

2.2.4 Coal-mining disasters data

Our real data example involves the "coal-mining disasters" data (e.g., Andrews and Herzberg, 1985, p. 53-56), a standard dataset used in the literature to test NHPP intenstiy estimation methods. The point pattern comprises the times (in days) of $n = 191$ explosions of fire-damp or coal-dust in mines resulting in 10 or more casualties from the accident. The observation window consists of 40,550 days, from March 15, 1851 to March 22, 1962.

We fit the Erlang mixture model with $J = 50$, using a Lomax prior for θ with shape parameter 2 and scale parameter $2,000$, such that $\Pr(0 < \theta < 40,550) \approx 0.998$, and an exponential prior for b with mean 213. We also implemented the model with $J = 130$, obtaining essentially the same inference results for the NHPP functionals with the ones reported in Figure 2.7.

The estimates for the point process intensity and density functions (Figure 2.7, top row) suggest that the model successfully captures the multimodal intensity shape suggested by the data. The estimates for the mixture weights (Figure 2.7, bottom left panel) indicate the Erlang basis densities that are more influential to the model fit.

The bottom right panel of Figure 2.7 reports results from graphical model checking, using the "time-rescaling" theorem (e.g., Daley and Vere-Jones, 2003). If the

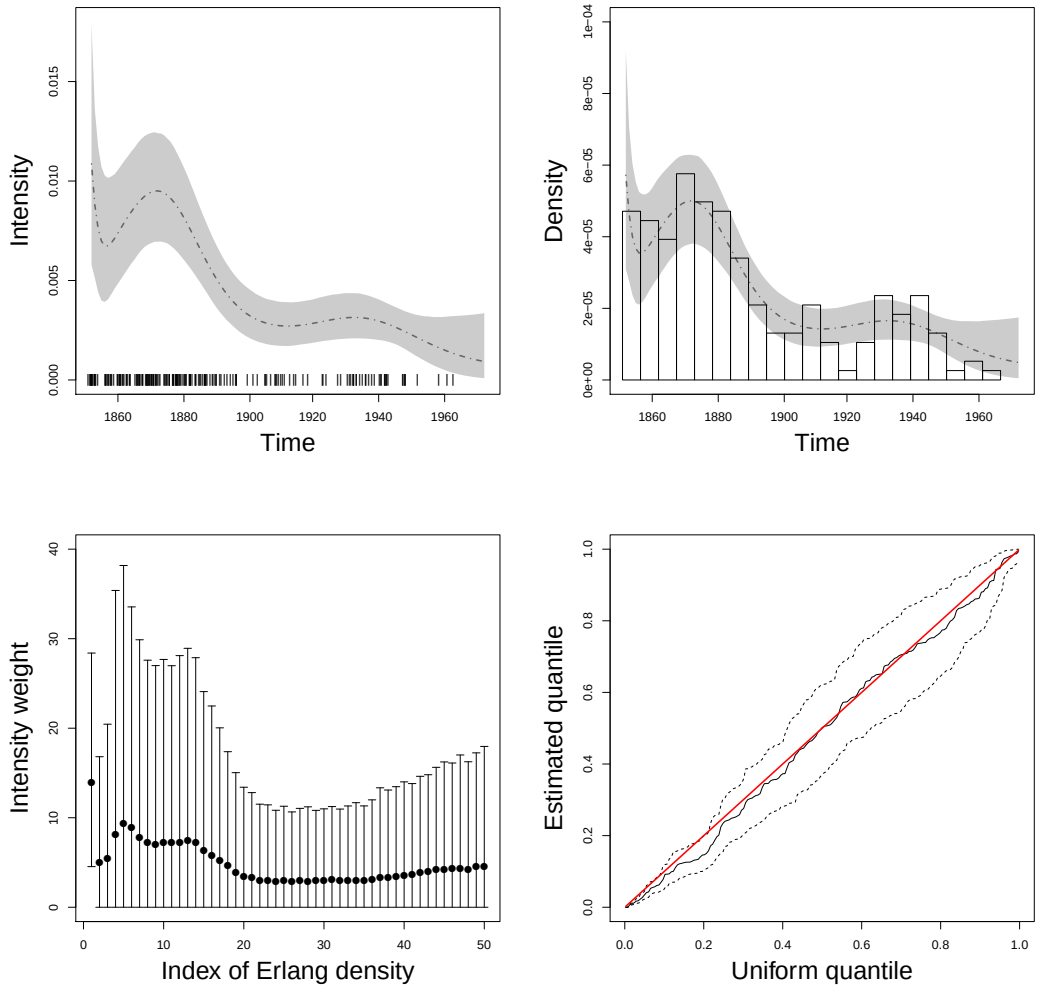

Figure 2.7: Coal-mining disasters data. The top left panel shows the posterior mean estimate (dashed-dotted line) and 95% interval bands (shaded area) for the intensity function. The bars at the bottom indicate the observed point pattern. The top right panel plots the posterior mean (dashed-dotted line) and 95% interval bands (shaded area) for the NHPP density, overlaid on the histogram of the accident times. The bottom left panel presents the posterior means (circles) and 90% interval estimates (bars) of the mixture weights. The bottom right panel plots the posterior mean and 95% interval bands for the time-rescaling model checking Q-Q plot.

point pattern $\{0 = t_0 < t_1 < ... < t_n < T\}$ is a realization from a NHPP with cumulative intensity function $\Lambda(t) = \int_0^t \lambda(u)\mathrm{d}u$, then the transformed point pattern $\{\Lambda(t_i) : i =$

$1, ..., n\}$ is a realization from a unit rate homogeneous Poisson process. Therefore, if we further transform to $U_i = 1 - \exp\{-(\Lambda(t_i) - \Lambda(t_{i-1}))\}$, where $\Lambda(0) \equiv 0$, then the $\{U_i : i = 1, ..., n\}$ are independent uniform$(0, 1)$ random variables. Hence, graphical model checking can be based on quantile–quantile (Q-Q) plots to assess agreement of the estimated U_i with the uniform distribution on the unit interval. Under the Bayesian inference framework, we can obtain a posterior sample for the U_i for each posterior realization for the NHPP intensity, and we can thus plot posterior point and interval estimates for the Q-Q graph. These estimates suggest that the NHPP model with the Erlang mixture intensity provides a good fit for the coal-mining disasters data.

2.2.5 Model comparison

Under the SGCP model, the temporal NHPP intensity is represented as $\lambda(t) = \lambda^* \sigma(g(t))$, where λ^* is an upper bound on the intensity function, $\sigma(z) = (1 + e^{-z})^{-1}$ is the logistic function, and g is a real-valued random function assigned a GP prior. Adams et al. (2009) develop MCMC posterior simulation using latent variables (associated with "thinned" events) to handle the intractable NHPP likelihood normalizing term.

To compare the Erlang mixture and SGCP models, we work with the synthetic data set of Section 2.2.3. This is a choice arising from practical considerations; as discussed below, the SGCP model is particularly computationally intensive, and we thus consider the data example with the smallest point pattern size.

We applied the SGCP model using a GP prior for function g with constant mean, μ, and squared-exponential covariance function, $\text{cov}(g(t_1), g(t_2)) = \theta_1 e^{-\theta_2(t_1 - t_2)^2}$.

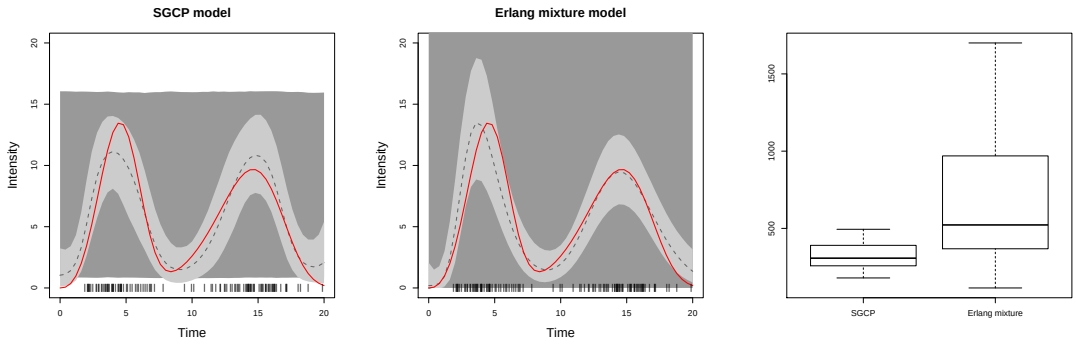

Figure 2.8: Synthetic data from temporal NHPP with bimodal intensity (Section 2.2.3). The left and middle panels provide estimates for the underlying intensity (red solid line) under the SGCP and Erlang mixture model, respectively: prior 95% interval bands (dark-gray shaded area), the posterior mean (dashed line), and posterior 95% interval bands (light-gray shaded area). The right panel shows for each model boxplots of 51 ESS values based on posterior samples for the intensity at 51 equally-spaced time points.

Therefore, the SGCP model parameters comprise λ^* and $(\mu, \theta_1, \theta_2)$, all four of which are assigned priors. We note that a prior specification strategy is not provided in Adams et al. (2009). We observed that, at least for this data set, inference for all SGCP model parameters is sensitive to the prior choice. Moreover, there is a conflict regarding the role of parameter λ^*: it controls the extent of prior uncertainty, with larger λ^* values resulting in wider prior interval bands for the intensity, but at the same time, increasing λ^* increases the number of latent variables and thus also the MCMC algorithm computing time.

We tuned the priors for the SGCP model parameters to obtain the best possible estimates for the intensity function. The resulting estimates are reported in the left panel of Figure 2.8. Even though we intentionally favored the SGCP model through the prior selection, the Erlang mixture model posterior mean estimate captures the peaks of the

underlying intensity more accurately than the SGCP model, indeed, under larger levels of prior uncertainty (middle panel of Figure 2.8).

The right panel of Figure 2.8 shows boxplots of 51 ESS values computed from the posterior samples for $\lambda(t)$ at 51 equally-spaced time points on a grid over $(0, T) = (0, 20)$. The ESS values are based on 15,000 posterior samples, obtained after discarding 5,000 burn-in samples. Evidently, the Erlang mixture model outperforms the SGCP model in terms of mixing of the MCMC algorithm as measured by the ESS. The benefit is more emphatic considering computing times: completing the 20,000 MCMC iterations took around 310 minutes under the SGCP model, while the corresponding computing time for the Erlang mixture model was about 5 minutes.

2.3 Modeling for spatial Poisson process intensities

In Section 2.3.1, we extend the modeling framework to spatial NHPPs with intensities defined on $\mathbb{R}^+ \times \mathbb{R}^+$. The resulting inference method is illustrated with synthetic and real data examples in Section 2.3.3 and 2.3.4, respectively.

2.3.1 The Erlang mixture model for spatial NHPPs

A spatial NHPP is again characterized by its intensity function, $\lambda(\boldsymbol{s})$, for $\boldsymbol{s} = (s_1, s_2) \in \mathbb{R}^+ \times \mathbb{R}^+$. The NHPP intensity is a non-negative and locally integrable function such that: (a) for any bounded $B \subset \mathbb{R}^+ \times \mathbb{R}^+$, the number of points in B, $N(B)$, follows a Poisson distribution with mean $\int_B \lambda(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u}$; and (b) given $N(B) = n$, the random locations $\boldsymbol{s}_i = (s_{i1}, s_{i2})$, for $i = 1, ..., n$, that form the spatial point pattern

in B are i.i.d. with density $\lambda(\boldsymbol{s})/\{\int_B \lambda(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}\}$. Therefore, the structure of the likelihood for the intensity function is similar to the temporal NHPP case. In particular, for spatial point pattern, $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$, observed in bounded region $D \subset \mathbb{R}^+ \times \mathbb{R}^+$, the likelihood is proportional to $\exp\{-\int_D \lambda(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}\} \prod_{i=1}^n \lambda(\boldsymbol{s}_i)$. As is typically the case in standard applications involving spatial NHPPs, we consider a regular, rectangular domain for the observation region D, which can therefore be taken without loss of generality to be the unit square.

Extending the Erlang mixture model in (2.1), we build the basis representation for the spatial NHPP intensity from products of Erlang densities, $\{\mathrm{Ga}(s_1 \mid j_1, \theta_1^{-1})\,\mathrm{Ga}(s_2 \mid j_2, \theta_2^{-1}) : j_1, j_2 = 1, ..., J\}$. Mixing is again with respect to the shape parameters (j_1, j_2), and the basis densities share a pair of scale parameters (θ_1, θ_2). Therefore, the model can be expressed as

$$\lambda(s_1, s_2) = \sum_{j_1=1}^{J} \sum_{j_2=1}^{J} \omega_{j_1 j_2}\,\mathrm{Ga}(s_1 \mid j_1, \theta_1^{-1})\,\mathrm{Ga}(s_2 \mid j_2, \theta_2^{-1}), \quad (s_1, s_2) \in \mathbb{R}^+ \times \mathbb{R}^+.$$

Again, a key model feature is the prior for the mixture weights. Here, the basis density indexed by (j_1, j_2) is associated with rectangle $A_{j_1 j_2} = [(j_1-1)\theta_1, j_1\theta_1) \times [(j_2-1)\theta_2, j_2\theta_2)$. The corresponding weight is defined through a random measure H supported on $\mathbb{R}^+ \times \mathbb{R}^+$, such that $\omega_{j_1 j_2} = H(A_{j_1 j_2})$. This construction extends the one for the weights of the temporal NHPP model. We again place a gamma process prior, $\mathcal{G}(H_0, c_0)$, on H, where c_0 is the precision parameter and H_0 is the centering measure on $\mathbb{R}^+ \times \mathbb{R}^+$. As a natural extension of the exponential cumulative hazard used in Section 2.1.1 for the gamma process prior mean, we specify H_0 to be proportional to area. In particular, $H_0(A_{j_1 j_2}) =$

$|A_{j_1 j_2}|/b = \theta_1\theta_2/b$, where $b > 0$. Using the independent increments property of the gamma process, and under the specific choice of H_0, the prior for the mixture weights is given by

$$\omega_{j_1 j_2} \mid \theta_1, \theta_2, c_0, b \overset{i.i.d.}{\sim} \text{Ga}(\omega_{j_1 j_2} \mid c_0\,\theta_1\,\theta_2\,b^{-1}, c_0), \quad j_1, j_2 = 1, ..., J,$$

which, as before, is a practically important feature of the model construction as it pertains to MCMC posterior simulation.

To complete the full Bayesian model, we place priors on the common scale parameters for the basis densities, (θ_1, θ_2), and on the gamma process prior hyperparameters c_0 and b. The role played by these model parameters is directly analogous to the one of the corresponding parameters for the temporal NHPP model, as detailed in Section 2.1.1. Therefore, we apply similar arguments to the ones in Section 2.1.2 to specify the model hyperpriors. More specifically, we work with (independent) Lomax prior distributions for scale parameters θ_1 and θ_2, where the shape parameter of the Lomax prior is set equal to 2 and the scale parameter is specified such that $\Pr(0 < \theta_1 < 1)\Pr(0 < \theta_2 < 1) \approx 0.999$. Recall that the observation region is taken to be the unit square; in general, for a square observation region, this approach implies the same Lomax prior for θ_1 and θ_2. The gamma process precision parameter c_0 is assigned an exponential prior with mean 10. The result of Section 2.1.1 for the prior mean of the NHPP intensity can be extended to show that $\text{E}(\lambda(s_1, s_2) \mid b, \theta_1, \theta_2)$ converges to b^{-1}, as $J \to \infty$, for any $(s_1, s_2) \in \mathbb{R}^+ \times \mathbb{R}^+$, and for any (θ_1, θ_2) (and c_0). The prior mean for the spatial NHPP intensity is practically constant at b^{-1} within its effective

support given roughly by $(0, J\theta_1) \times (0, J\theta_2)$. Hence, taking the size of the observed spatial point pattern as a proxy for the total intensity, b is assigned an exponential prior distribution with mean $1/n$. Finally, the choice of the value for J can be guided from the approximate effective support for the intensity, which is controlled by J along with θ_1 and θ_2. Analogously to the approach discussed in Section 2.1.2, the value of J (or perhaps a lower bound for J) can be specified through the integer part of $1/\theta^*$, where θ^* is the median of the common Lomax prior for θ_1 and θ_2.

2.3.2 Posterior simulation for spatial NHPPs

The posterior simulation method for the spatial NHPP model is developed through a straightforward extension of the approach detailed in Section 2.1.3. We work again with the augmented model that involves latent variables $\{\boldsymbol{\gamma}_i : i = 1, \ldots, n\}$, where $\boldsymbol{\gamma}_i = (\gamma_{i1}, \gamma_{i2})$ identifies the basis density to which observed point location (s_{i1}, s_{i2}) is assigned. The spatial NHPP model retains the practically relevant feature of efficient updates for the mixture weights, which, given the other model parameters and the data, have independent gamma posterior full conditional distributions.

Denote by $\{\boldsymbol{s}_i : i = 1, ..., n\}$, where $\boldsymbol{s}_i = (s_{i1}, s_{i2})$, the spatial point pattern observed in the unit square. Under the Erlang mixture model for spatial NHPPs,

developed in Section 2.3.1, the likelihood is proportional to

$$
\begin{aligned}
&\exp\Big(-\int_0^1\int_0^1 \lambda(u_1,u_2)\,\mathrm{d}u_1\mathrm{d}u_2\Big)\prod_{i=1}^{n}\lambda(s_{i1},s_{i2})\\
&=\exp\Big(-\sum_{j_1=1}^{J}\sum_{j_2=1}^{J}\omega_{j_1j_2}\,K_{j_1,\theta_1}(1)K_{j_2,\theta_2}(1)\Big)\\
&\times\prod_{i=1}^{n}\Big\{\sum_{j_1=1}^{J}\sum_{j_2=1}^{J}\omega_{j_1j_2}\,\mathrm{Ga}(s_{i1}\mid j_1,\theta_1^{-1})\mathrm{Ga}(s_{i2}\mid j_2,\theta_2^{-1})\Big\}\\
&=\prod_{j_1=1}^{J}\prod_{j_2=1}^{J}\exp\Big(-\omega_{j_1j_2}K_{j_1,\theta_1}(1)K_{j_2,\theta_2}(1)\Big)\\
&\times\prod_{i=1}^{n}\Big\{\Big(\sum_{r_1=1}^{J}\sum_{r_2=1}^{J}\omega_{r_1r_2}\Big)\sum_{j_1=1}^{J}\sum_{j_2=1}^{J}\Big(\frac{\omega_{j_1j_2}}{\sum_{r_1=1}^{J}\sum_{r_2=1}^{J}\omega_{r_1r_2}}\Big)\\
&\times\mathrm{Ga}(s_{i1}\mid j_1,\theta_1^{-1})\mathrm{Ga}(s_{i2}\mid j_2,\theta_2^{-1})\Big\}.
\end{aligned}
$$

where $K_{j,\theta}(1)=\int_0^1\mathrm{Ga}(u\mid j,\theta^{-1})\,\mathrm{d}u$.

Next, we introduce auxiliary variables $\Gamma=\{\boldsymbol{\gamma}_i : i=1,\ldots,n\}$, where $\boldsymbol{\gamma}_i=(\gamma_{i1},\gamma_{i2})$, to obtain the following hierarchical model representation:

$$
\begin{aligned}
\{\boldsymbol{s}_1,...,\boldsymbol{s}_n\}\mid\Gamma,\boldsymbol{\omega},\boldsymbol{\theta}\sim&\prod_{j_1=1}^{J}\prod_{j_2=1}^{J}\exp\Big(-\omega_{j_1j_2}K_{j_1,\theta_1}(1)K_{j_2,\theta_2}(1)\Big)\\
&\times\prod_{i=1}^{n}\Big\{\Big(\sum_{r_1=1}^{J}\sum_{r_2=1}^{J}\omega_{r_1r_2}\Big)\,\mathrm{Ga}(s_{i1}\mid\gamma_{i1},\theta_1^{-1})\mathrm{Ga}(s_{i2}\mid\gamma_{i2},\theta_2^{-1})\Big\}\\
\boldsymbol{\gamma}_i\mid\boldsymbol{\omega}\overset{i.i.d.}{\sim}&\sum_{j_1=1}^{J}\sum_{j_2=1}^{J}\Big(\frac{\omega_{j_1j_2}}{\sum_{r_1=1}^{J}\sum_{r_2=1}^{J}\omega_{r_1r_2}}\Big)\,\delta_{j_1}(\gamma_{i1})\delta_{j_2}(\gamma_{i2}),\quad i=1,\ldots,n\\
\boldsymbol{\theta},c_0,b,\boldsymbol{\omega}\sim&\,p(\theta_1)p(\theta_2)p(c_0)p(b)\prod_{j_1=1}^{J}\prod_{j_2=1}^{J}\mathrm{Ga}(\omega_{j_1j_2}\mid c_0\theta_1\theta_2b^{-1},c_0)
\end{aligned}
$$

where $\boldsymbol{\omega}=\{\omega_{j_1j_2}:j_1,j_2=1,...,J\}$, $\boldsymbol{\theta}=(\theta_1,\theta_2)$, and $p(\theta_1)$, $p(\theta_2)$, $p(c_0)$, and $p(b)$ denote the priors for θ_1, θ_2, c_0, and b.

As in the posterior inference method for the temporal NHPP model (Section 2.1.3), we explore the posterior distribution using Gibbs sampling. The posterior full

conditional for each $\boldsymbol{\gamma}_i$ is a discrete distribution on $\{1,...,J\} \times \{1,...,J\}$ such that $\Pr(\gamma_{i1} = j_1, \gamma_{i2} = j_2 \mid \boldsymbol{\theta}, \boldsymbol{\omega}, \text{data}) \propto \omega_{j_1 j_2} \, \text{Ga}(s_{i1} \mid j_1, \theta_1^{-1}) \text{Ga}(s_{i2} \mid j_2, \theta_2^{-1})$, for $j_1, j_2 = 1, ..., J$.

Let $N_{j_1 j_2} = |\{\boldsymbol{s}_i : \gamma_{i1} = j_1, \ \gamma_{i2} = j_2\}|$, for $j_1, j_2 = 1, ..., J$. With the conditionally conjugate priors for $\omega_{j_1 j_2}$, implied by the gamma process prior, the posterior full conditional distribution for the mixture weights is derived as follows:

$$
\begin{aligned}
p(\boldsymbol{\omega} \mid \Gamma, \boldsymbol{\theta}, c_0, b, \text{data}) &\propto \Big\{ \prod_{j_1=1}^{J} \prod_{j_2=1}^{J} \exp\Big(- \omega_{j_1 j_2} K_{j_1,\theta_1}(1) K_{j_2,\theta_2}(1) \Big) \Big\} \Big(\sum_{r_1=1}^{J} \sum_{r_2=1}^{J} \omega_{r_1 r_2} \Big)^n \\
&\times \Big\{ \prod_{j_1=1}^{J} \prod_{j_2=1}^{J} \omega_{j_1 j_2}^{N_{j_1 j_2}} \big(\sum_{r_1=1}^{J} \sum_{r_2=1}^{J} \omega_{r_1 r_2} \big)^{-N_{j_1 j_2}} \Big\} \\
&\times \Big\{ \prod_{j_1=1}^{J} \prod_{j_2=1}^{J} \text{Ga}(\omega_{j_1 j_2} \mid c_0 \theta_1 \theta_2 b^{-1}, c_0) \Big\} \\
&\propto \prod_{j_1=1}^{J} \prod_{j_2=1}^{J} \exp\Big(- \omega_{j_1 j_2} K_{j_1,\theta_1}(1) K_{j_2,\theta_2}(1) \Big) \omega_{j_1 j_2}^{N_{j_1 j_2}} \\
&\times \text{Ga}(\omega_{j_1 j_2} \mid c_0 \theta_1 \theta_2 b^{-1}, c_0) \\
&= \prod_{j_1=1}^{J} \prod_{j_2=1}^{J} \text{Ga}(\omega_{j_1 j_2} \mid N_{j_1 j_2} + c_0 \theta_1 \theta_2 b^{-1}, K_{j_1,\theta_1}(1) K_{j_2,\theta_2}(1) + c_0)
\end{aligned}
$$

with $\sum_{j_1=1}^{J} \sum_{j_2=1}^{J} N_{j_1 j_2} = n$. As in the temporal case, the mixture weights can be updated independently, given the other parameters and the data, from gamma posterior full conditional distributions. Therefore, the practical benefits of the Erlang mixture model structure – convenient updates for the mixture weights and computational efficiency of the MCMC algorithm – carry over to inference for spatial NHPPs.

Finally, parameters θ_1 and θ_2 and the hyperparameters, c_0 and b, of the gamma process prior for H are updated with Metropolis-Hastings steps, using log-normal pro-

posal distributions.

2.3.3 Synthetic data example

Here, we illustrate the spatial NHPP model using synthetic data based on a bimodal intensity function built from a two-component mixture of bivariate logit-normal densities. Denote by $\mathrm{BLN}(\boldsymbol{\mu}, \Sigma)$ the bivariate logit-normal density arising from the logistic transformation of a bivariate normal with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. A spatial point pattern of size 528 was generated over the unit square from a NHPP with intensity $\lambda(s_1, s_2) = 150\,\mathrm{BLN}((s_1, s_2) \mid \boldsymbol{\mu}_1, \Sigma) + 350\,\mathrm{BLN}((s_1, s_2) \mid \boldsymbol{\mu}_2, \Sigma)$, where $\boldsymbol{\mu}_1 = (-1, 1)$, $\boldsymbol{\mu}_2 = (1, -1)$, and $\Sigma = (\sigma_{11}, \sigma_{12}, \sigma_{21}, \sigma_{22}) = (0.3, 0.1, 0.1, 0.3)$. The intensity function and the generated spatial point pattern are shown in the top left panel of Figure 2.9.

The Erlang mixture model was applied setting $J = 70$ and using the hyperpriors for θ_1, θ_2, c_0 and b discussed in Section 2.3.1. Figure 2.9 reports inference results. The posterior mean intensity estimate successfully captures the shape of the underlying intensity function. The structure of the Erlang mixture model enables ready inference for the marginal NHPP intensities associated with the two-dimensional NHPP. Although such inference is generally not of direct interest for spatial NHPPs, in the context of a synthetic data example it provides an additional means to check the model fit. The marginal intensity estimates effectively retrieve the bimodality of the true marginal intensity functions; the slight discrepancy at the second mode can be explained by inspection of the generated data for which the second mode clusters are located slightly

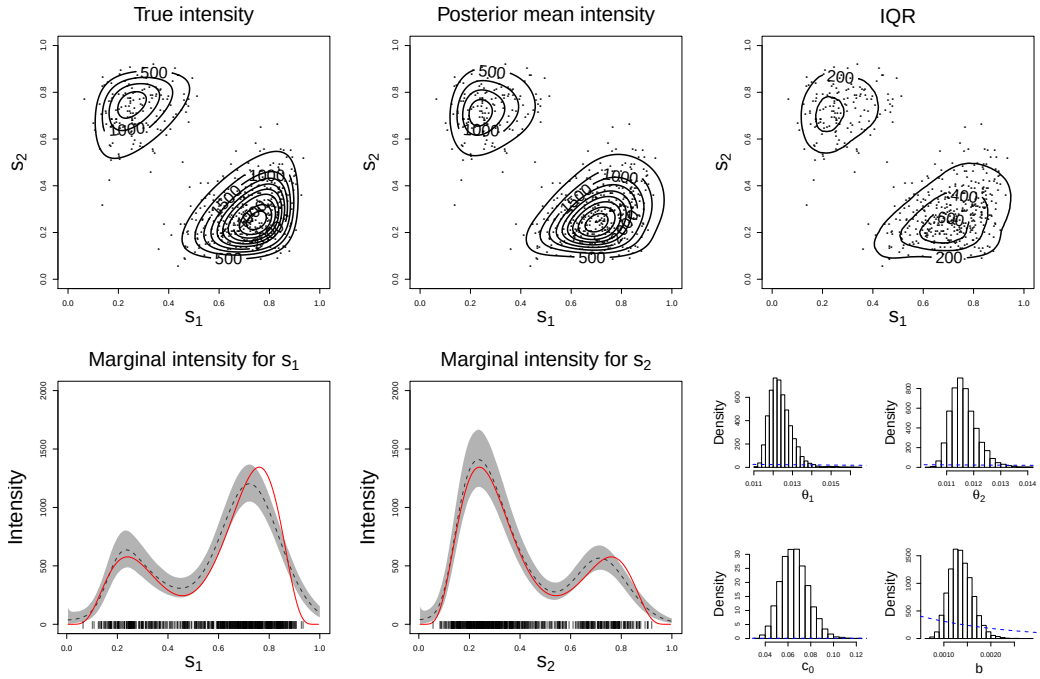

Figure 2.9: Synthetic data example from spatial NHPP. The top row panels show contour plots of the true intensity, and of the posterior mean and interquartile range estimates. The points in each panel indicate the observed point pattern. The first two panels at the bottom row show the marginal intensity estimates – the posterior mean (dashed line) and 95% uncertainty bands (shaded area) – along with the true function (red solid line) and corresponding point pattern (bars at the bottom of each panel). The bottom right panel displays histograms of posterior samples for the model hyperparameters along with the corresponding prior densities (dashed lines).

to the left of the theoretical mode. Finally, we note the substantial prior-to-posterior learning for all model hyperparameters.

2.3.4 Real data illustration

Our final data example involves a spatial point pattern that has been previously used to illustrate NHPP intensity estimation methods (e.g., Diggle, 2014; Kottas and Sansó, 2007). The data set involves the locations of 514 maple trees in a 19.6 acre

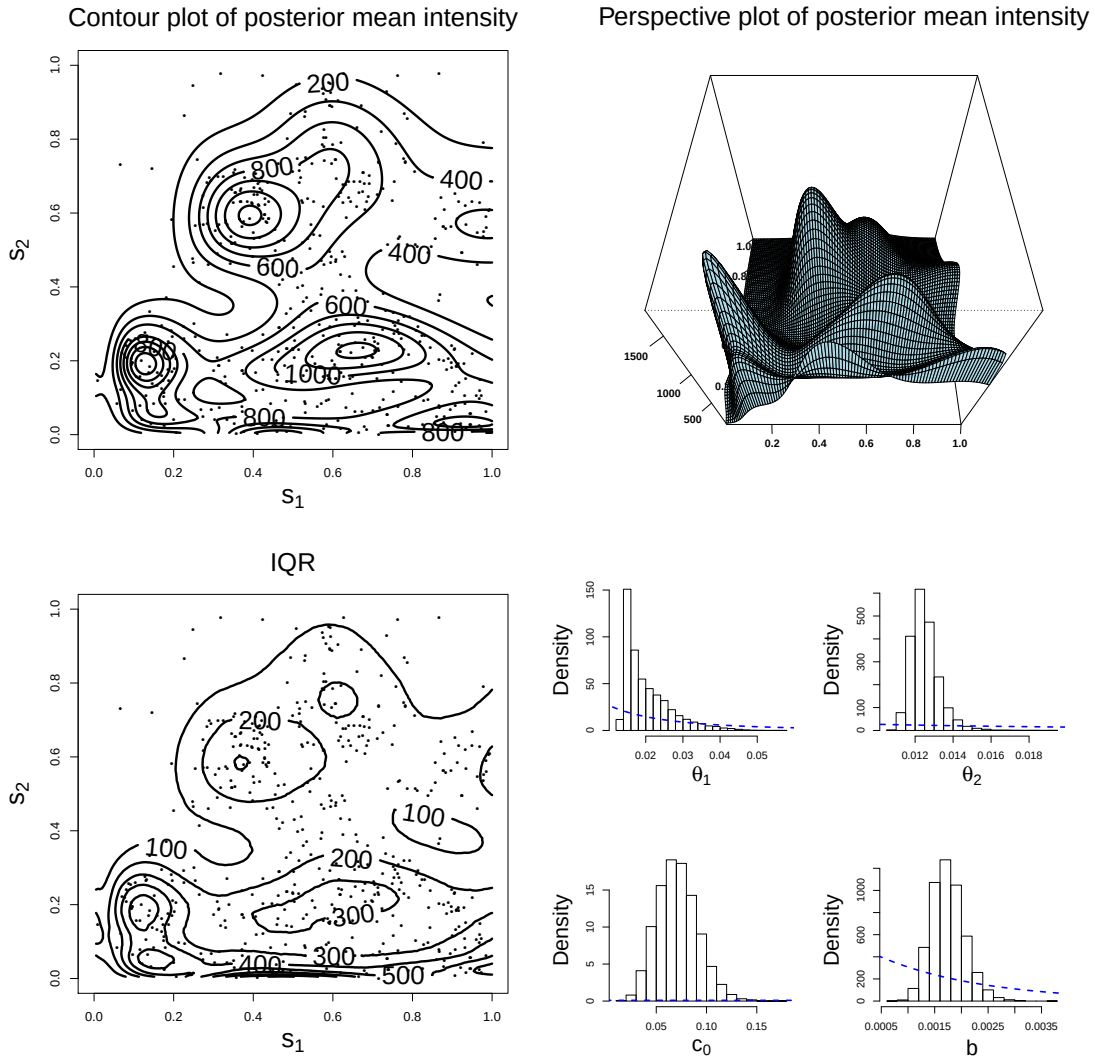

Figure 2.10: Maple trees data. The top row panels show the posterior mean estimate for the intensity function in the form of contour and perspective plots. The bottom left panel displays the corresponding posterior interquartile range contour plot. The bottom right panel plots histograms of posterior samples for the model hyperparameters along with the corresponding prior densities (dashed lines). The points in the left column plots indicate the locations of the 514 maple trees.

square plot in Lansing Woods, Clinton County, Michigan, USA; the maple trees point pattern is included in the left column panels of Figure 2.10.

To apply the spatial Erlang mixture model, we specified the hyperpriors for θ_1, θ_2, c_0 and b following the approach discussed in Section 2.3.1, and set $J = 70$. As with the synthetic data example, the posterior distributions for model hyperparameters are substantially concentrated relative to their priors; see the bottom right panel of Figure 2.10. The estimates for the spatial intensity of maple tree locations reported in Figure 2.10 demonstrate the model's capacity to uncover non-standard, multimodal intensity surfaces.

2.4 Discussion

We have proposed a Bayesian nonparametric modeling approach for Poisson processes over time or space. The approach is based on a mixture representation of the point process intensity through Erlang basis densities, which are fully specified save for a scale parameter shared by all of them. The weights assigned to the Erlang densities are defined through increments of a random measure (a random cumulative intensity function in the temporal NHPP case) which is modeled with a gamma process prior. A key feature of the methodology is that it offers a good balance between model flexibility and computational efficiency in implementation of posterior inference. Such inference has been illustrated with synthetic and real data for both temporal and spatial Poisson process intensities.

To discuss our contribution in the context of Bayesian nonparametric modeling methods for NHPPs (briefly reviewed in the Introduction), note that the main

approaches can be grouped into two broad categories: placing the prior model on the NHPP intensity function; or, assigning separate priors to the total intensity and the NHPP density (both defined over the observation window).

In terms of applications, especially for spatial point patterns, the most commonly explored class of models falling in the former category involves Gaussian process (GP) priors for logarithmic (or logit) transformations of the NHPP intensity (e.g., Møller et al., 1998; Adams et al., 2009). The NHPP likelihood normalizing term renders full posterior inference under GP-based models particularly challenging. This challenge has been bypassed using approximations of the stochastic integral that defines the likelihood normalizing term (Brix and Diggle, 2001, Brix and Møller, 2001), data augmentation techniques (Adams et al., 2009), and different types of approximations of the NHPP likelihood along with integrated nested Laplace approximation for approximate Bayesian inference (Illian et al., 2012, Simpson et al., 2016). In contrast, the Erlang mixture model can be readily implemented with MCMC algorithms that do not involve approximations to the NHPP likelihood or complex computational techniques. Section 2.2.5 and Appendix B include comparison of the proposed model with two GP-based models: the sigmoidal Gaussian Cox process (SGCP) model (Adams et al., 2009) for temporal NHPPs; and the log-Gaussian Cox process (LGCP) model for spatial NHPPs, as implemented in the `R` package `lgcp` (Taylor et al., 2013). The results, based on the synthetic data considered in Sections 2.2.3 and 2.3.3, suggest that the Erlang mixture model is substantially more computationally efficient than the SGCP model, as well as less sensitive to model/prior specification than LGCP models for which the choice of

the GP covariance function can have a large effect on the intensity surface estimates.

Since it involves a mixture formulation for the NHPP intensity, the proposed modeling approach is closer in spirit to methods based on Dirichlet process mixture priors for the NHPP density (e.g., Kottas and Sansó, 2007; Taddy and Kottas, 2012). Both types of approaches build posterior simulation from standard MCMC techniques for mixture models, using latent variables that configure the observed points to the mixture components. Models that build from density estimation with Dirichlet process mixtures benefit from the wide availability of related posterior simulation methods (e.g., the number of mixture components in the NHPP density representation does not need to be specified), and from the various extensions of the Dirichlet process for dependent distributions that can be explored to develop flexible models for hierarchically related point processes (e.g., Taddy, 2010; Kottas et al., 2012; Xiao et al., 2015; Rodriguez et al., 2017). However, by construction, this approach is restricted to modeling the NHPP intensity only on the observation window, in fact, with a separate prior for the NHPP density and for the total intensity over the observation window. The Erlang mixture model overcomes this limitation. For instance, in the temporal case, the prior model supports the intensity on $\mathbb{R}^+$, and the priors for the total intensity and the NHPP density over $(0, T)$ (given in Equation (2.2)) are compatible with the prior for the NHPP intensity.

Chapter 3

Bayesian Nonparametric Modeling and Inference for Hawkes Processes

3.1 Prior models for the HP intensity function

The Erlang mixture prior model is useful as a building block towards Bayesian nonparametric inference for Hawkes processes that can be represented as hierarchically structured, clustered NHPPs. Beyond its flexibility and computational efficiency, the Erlang mixture model facilitates prediction of future events. Not all existing models offer this benefit. DP-based mixture models, for example, consider intensity functions defined on a bounded interval, the observation window. Therefore, intensity estimates cannot be used to make predictions for intervals beyond the observation window. This is not a major concern when modeling NHPPs, since their conditional intensities do not involve the process history. The HP, on the other hand, utilizes the conditional intensity

estimates for predictive inference. Restricting the model for the HP immigrant intensity on the observation window would be a practical limitation. In the following section, we describe a semiparametric model using the Erlang mixture prior for the Hawkes process immigrant (background) intensity function.

3.1.1 Mixture model for the immigrant intensity function

According to the cluster representation, the HP can be viewed as independent NHPPs for immigrants and offspring. So, we can treat the HP immigrant intensity function as the intensity of the NHPP for immigrants. Similarly, the excitation function becomes the intensity of the NHPP for offspring. The connection allows us to model the immigrant intensity and excitation functions using the Erlang mixture, introduced in Section 2.1.1. This section focuses on the Erlang mixture model for the immigrant intensity. Substituting the Erlang mixture for the constant immigrant intensity μ of (1.1) enables flexible immigrant intensity modeling, which is desirable for real data analyses (e.g., Ogata, 1998; Helmstetter et al., 2006; Ogata and Zhuang, 2006, for earthquake applications). Following are the details of the model formulation and the prior specification.

The Erlang mixture model is a structured mixture of Erlang densities with a common scale parameter, mixing on integer shape parameters. The mixture weight is constructed through increments of a cumulative hazard function. This model converges (pointwise) to the hazard function of the cumulative hazard. Incorporating a nonparametric prior on the cumulative hazard gives flexibility to the (cumulative) haz-

ard function and eventually to the model due to convergence. We employ the gamma process prior for the cumulative hazard function. The prior choice allows for efficient handling of the likelihood normalizing constant with ready prior-to-posterior updating for the mixture weight.

Denote by $\mathrm{Ga}(t|a,b)$ for $t \in \mathbb{R}^+$, a gamma density function with shape a and mean a/b. We also use the term for an Erlang density with an integer shape parameter. With the Erlang mixture for the immigrant intensity, our semiparametric model is defined as

$$\begin{aligned} \lambda^*(t) &= \sum_{l=1}^{L} \nu_l \mathrm{Ga}(t|l,\theta^{-1}) + \gamma \sum_{t_i<t} \mathrm{Exp}(t-t_i|\alpha), \quad t \in \mathbb{R}^+ \\ \nu_l &= G(l\theta) - G((l-1)\theta), \;\; l = 1,\ldots,L, \quad G \sim \mathcal{G}(G_0, c_0), \end{aligned} \tag{3.1}$$

where $\mathcal{G}(G_0, c_0)$ denotes the gamma process with a non-decreasing function $G_0(\cdot)$ on $\mathbb{R}^+$ and a positive constant c_0, such that $\mathrm{E}(G(t)) = G_0(t)$ and $\mathrm{Var}(G(t)) = G_0(t)/c_0$. So, G_0 and c_0 can be viewed as the centering function and the precision parameter of G. We consider the exponential cumulative hazard for the centering function as follows: $G_0(t) = t/b_{G_0}$ with $b_{G_0} > 0$. This function, which has only one parameter b_{G_0} to be estimated, is chosen for its brevity. The mixture weight, ν_l, has the gamma prior distribution $\mathrm{Ga}(c_0\theta/b_{G_0}, c_0)$, driven by the gamma process prior. $\mathrm{Exp}(t-t_i|\alpha)$ of (3.1) denotes the exponential density function with mean $1/\alpha$, which is the common choice for the offspring density alongside the power-law density function, defined as $ab^a/(b+t)^{a+1}$, $a, b > 0$.

We follow the strategy given in Section 2.1.2 to specify the Erlang mixture of

the semiparametric model, with different notations. The prior mean of the immigrant intensity function is roughly constant at $1/b_{G_0}$. So, T/b_{G_0}, integrated immigrant intensity over $(0, T)$, would be a reasonable approximation to the total immigrant intensity on $(0, T)$. To specify b_{G_0}, we match T/b_{G_0} with its proxy: the number of observed immigrant points; or half of the point pattern size. For the precision parameter c_0, we take the Exp(0.1) prior with mean 10. Based on empirical evidence, 10 was regarded as a conservative value, and substantial prior-to-posterior learning was observed in the simulation study presented in Section 3.3.2.1. The parameter θ, the common scale of Erlang density basis functions, is assigned the Lomax prior with shape 2 and scale d_θ. The hyperparameter is determined by considering the least flexible model of a single component (i.e., $L = 1$). As a rough guess about the effective support for the immigrant intensity, the model returns $(0, L\theta) = (0, \theta)$. Using the observation window as a proxy for the effective support, d_θ is chosen, such that $\Pr(0 < \theta < T) \approx 0.999$. Such a choice is conservative because more flexible models with $L > 1$ would result in a smaller θ by the support's upper bound $L\theta$. Using $(0, L\theta^*)$ for the effective support, the number of mixture components, L, is chosen and set to the integer part of T/θ^*, where θ^* denotes the prior median for θ. In order to achieve a more conservative value for L, we recommend conducting sensitivity analysis based on the value as a lower bound. A larger L may improve the estimation of immigrant intensity in cases where underlying immigrant intensities are assumed to be non-standard. The branching ratio, γ, is assigned the $\text{Ga}(2, a_\gamma)$ prior with shape 2 and mean $2/a_\gamma$. We determine the hyperparameter using the HP stability condition $\gamma \in (0, 1)$, such that $\Pr(0 < \gamma < 1) \approx 0.9$. Lastly, α of

the exponential offspring density has the Exp(a_α) prior. Hyperparameter a_α is chosen based on a prior guess at (L_O, U_O), a bound for the distance between an offspring and its parent. We take a_α, such that $\Pr(L_O < (t - t_i) < U_O) \approx 0.9$. In Section 3.2.1.1, we present the hierarchical model representation and outline the posterior inference procedure.

3.1.2 Mixture models for the excitation function

In this section, we turn our attention to the estimation of the excitation function. For the offspring density function, exponential and power-law densities are typically chosen. The resulting decreasing (in time) excitation functions have dominated parametric modeling (e.g., Hawkes, 1971a; Ogata, 1988; Balderama et al., 2012; Mohler, 2014). It may be natural to assume a decreasing excitation function in certain applications such as earthquake forecasting. Indeed, the odds of aftershocks occurring will decrease as time passes from the main shock. But, recent classical nonparametric approaches cast doubt on the decreasing-shape offspring functions and show non-decreasing estimation results from real data analyses (e.g., Mohler et al., 2011; Lewis and Mohler, 2011; Zhou et al., 2013). Even when decreasing offspring densities are justified, the two prevalent options for the offspring density function have the issue of potentially incompatible estimation results caused by their different tail behavior. Consequently, model misspecification may result in distorted estimates. These considerations have motivated our nonparametric approach to modeling the excitation function, for which we propose two different options: an Erlang mixture-based model in Section 3.1.2.1;

and uniform-mixture-based models for decreasing offspring densities in Section 3.1.2.2.

3.1.2.1 Erlang mixture for the excitation function

As a necessary condition of the excitation function, it must be integrable so that it can be factorized into the branching ratio and the offspring density. Furthermore, the branching ratio should be within the unit interval for HP stability. The process will explode without such a scheme due to the divergence in the expected cluster size, which is represented by a geometric series (Daley and Vere-Jones, 2003). The exclusive conditions of the excitation function preclude the use of the Erlang mixture model we applied to the immigrant intensity. Under the modeling framework, the branching ratio can be expressed as $\gamma = \sum_{j=1}^{J} \omega_j = H(J\theta)$. Because the gamma process prior that is assigned to H regards $\mathbb{R}^+$ as the state space, limiting the values of γ makes the prior choice irrelevant.

Alternatively, we propose a novel mixture of Erlang densities with weights (directly) assigned independent gamma priors, as opposed to defined by increments of a cumulative hazard function. The model satisfies the excitation function integrability condition needed for the HP definition. The modeling method also allows for effective handling of the stability condition via hyperparameters of the gamma priors. With the common choice, a positive constant μ, for the immigrant intensity, we define a semiparametric model as

$$\lambda^*(t) = \mu + \sum_{t_i<t} \Big[\sum_{l=1}^{L} \nu_l \mathrm{Ga}(t - t_i | l, \theta^{-1}) \Big], \quad t \in \mathbb{R}^+$$
$$\nu_l \overset{ind.}{\sim} \mathrm{Ga}(A_l, \eta), \tag{3.2}$$

where $A_l \equiv \alpha_0[F_0(l\theta) - F_0((l-1)\theta)]$, $l = 1, \ldots, L-1$ and $A_L \equiv \alpha_0[1 - F_0((L-1)\theta)]$. As a result of these hyperparameter choices for the gamma priors, the branching ratio, $\gamma = \sum_{l=1}^{L} \nu_l$, becomes a gamma random variable with shape $\alpha_0 = \sum_{l=1}^{L} A_l$ and rate η. So, the necessary condition of integrability for the excitation function is satisfied, and the stability condition can be managed by adjusting just two hyperparameters. F_0 is a cumulative distribution function and is assigned the exponential cumulative distribution for simplicity.

The model in (3.2) can be expressed as $\lambda^*(t) = \mu + \sum_{t_i<t} \gamma \Big[\sum_{l=1}^{L} \omega_l \mathrm{Ga}(t - t_i | l, \theta^{-1}) \Big]$ with the normalized weight $\omega_l = \nu_l / \sum_{k=1}^{L} \nu_k$. Due to the characteristic property that a Dirichlet random vector can be represented by independent gamma random variables, we can regard the weights $(\omega_1, \omega_2, \ldots, \omega_L)$ as a Dirichlet random vector with concentration parameters A_l, $l = 1, \ldots, L$. This model expression establishes a connection between the Erlang mixture and the Erlang-density-based DP mixture. Suppose the normalized mixture weights are defined as $\omega_l = F(l\theta) - F((l-1)\theta)$ for $l = 1, \ldots, L-1$ and $\omega_L = 1 - F((L-1)\theta)$ with a cumulative distribution function F. Placing the DP prior on F with precision parameter α_0 and centering cumulative distribution function F_0 recovers the Erlang DP mixture $\sum_{l=1}^{L} \omega_l \mathrm{Ga}(t - t_i | l, \theta^{-1})$, where the mixture weight has the Dirichlet distribution $\mathrm{Dir}(A_1, A_2, \ldots, A_L)$. Consequently,

the Erlang mixture in (3.2) for the excitation function is equivalent to the Erlang DP mixture for the offspring density function multiplied by the special-form branching ratio, $\gamma = \sum_{l=1}^{L} \nu_l \sim \text{Ga}(\alpha_0, \eta)$. According to the pointwise convergence of the Erlang mixture in distribution in Lee and Lin (2010) and the convergence theorem for the density function in Butzer (1954), the Erlang DP mixture converges to the density function of F. Therefore, the DP prior for F, prompted by the gamma prior for ν_l, provides flexibility for the prior model.

To complete the probability model, we place the following set of priors on the model parameters. The constant immigrant μ is assigned the $\text{Exp}(a_\mu)$ prior, under which the posterior full condition for μ is available in closed-form. The rate parameter is chosen by linking T/a_μ, expected cumulative immigrant intensity over $(0, T)$, to the size of observed immigrant points (or simply $n/2$, half of the observed point pattern size). The branching ratio is gamma distributed with shape α_0 and rate η. The hyperparameters are specified by the stability condition, such that $\Pr(0 < \gamma < 1) \approx 0.9$.

As in the immigrant Erlang mixture model, we take the Lomax prior for θ with shape 2 and scale d_θ, such that $\Pr(0 < \theta < T_O) \approx 0.999$, where the interval $(0, T_O)$ is the effective support for the excitation function of distance $t - t_i$. The upper bound T_O can be chosen by experts or general theories for the range of $t - t_i$, which may differ depending on the application to which the model is applied. We use $(0, L\theta^*)$ as a rough guess about the effective support and take the integer part of T_O/θ^* to set L, where θ^* indicates the prior median of θ. The rate parameter b_{F_0} of the exponential centering distribution, $F_0(t) = 1 - \exp\{-b_{F_0} t\}$, is assigned the $\text{Exp}(a_{b_{F_0}})$ prior. We

can determine the hyperparameter using the offspring density function expected over the mixture weight. Appendix F provides the asymptotic expected offspring density function, that is, the exponential density function with rate $(1 - \exp\{-\theta/b_{F_0}\})/\theta$. According to the appendix, the mean distance (between an offspring and its parent) under the asymptotic density function is given by $\theta/(1 - \exp\{-\theta/b_{F_0}\})$. We regard the upper bound T_O of the effective support as a conservative choice for the mean distance and solve the equation $\theta^*/(1 - \exp\{-\theta^*/b_{F_0}\}) = T_O$ in terms of b_{F_0}. Then, we have $b_{F_0} = \theta^*/ - \log(1 - \theta^*/T_O)$, which will be used for the expectation of b_{F_0}, such that $\mathrm{E}(b_{F_0}) = 1/a_{b_{F_0}} = \theta^*/ - \log(1 - \theta^*/T_O)$.

We conducted sensitivity analysis to the choice of T_O and even tried T as an extremely conservative selection. The posterior inference for the excitation function was found to be robust in the simulation study of Section 3.3.2. Section 3.2.1.2 addresses the hierarchical model representation and posterior inference methods.

3.1.2.2 Uniform mixtures for the non-increasing offspring density function

In HP parametric modeling, both exponential and power-law densities are commonly used as the offspring density function. But, these two densities exhibit different tail behavior, owing to their exponential tail or polynomial tail. In the real data analysis of Section 3.4, we applied each density function to HP parametric modeling for earthquake occurrences. The choice of the offspring density function brought about remarkably different estimation results. Our new modeling approach was inspired by the dramatic change in inferences from the density function. The section describes a

semiparametric modeling framework, whose offspring density specializes in estimating non-increasing densities.

Our modeling method is based on the fact that for any non-increasing density f on the positive real line there exists a distribution function F on $[0, \infty)$ such that $f(t|F) = \int \theta^{-1} \mathbf{1}_{[0,\theta)}(t) dF(\theta)$. By placing a nonparametric prior on the mixing distribution, F, we can construct a uniform mixture model which can represent any non-increasing density function on $\mathbb{R}^+$.

For F, we consider two stochastic process priors: the Dirichlet process (DP) and geometric weights (GW). Both priors can be defined via stick-breaking constructions, $F(\cdot) = \sum_{l=1}^{\infty} \omega_l \delta_{Z_l}(\cdot)$. From the infinite-sum representation of the priors, the target function (i.e., the offspring density function such that $f = g$) can be derived as $g(t|F) = \int \theta^{-1} \mathbf{1}_{[0,\theta)}(t) dF(\theta) = \sum_{l=1}^{\infty} \omega_l Z_l^{-1} \mathbf{1}_{[0,Z_l)}(t)$, where the weights ω_l and locations Z_l are random, with $\sum_{l=1}^{\infty} \omega_l = 1$ almost surely. The weight is independent of $Z_l \overset{i.i.d.}{\sim} F_0$, where F_0 is the centering distribution function. By substituting the uniform mixture for the offspring density function (the target function), we can define a semiparametric model as follows:

$$\lambda^*(t) = \mu + \sum_{t_j < t} \gamma g(t - t_j) = \mu + \sum_{t_j < t} \gamma \Big[\sum_{l=1}^{\infty} \omega_l Z_l^{-1} \mathbf{1}_{[0,Z_l)}(t) \Big]. \quad t > 0 \qquad (3.3)$$

Following is a way to specify mixture weight for the priors, which clarifies the difference between their modeling frameworks. Denote by $\text{Be}(a, b)$ the beta distribution with mean $a/(a + b)$. The weight of the Dirichlet process prior is defined through the

following stick-breaking mechanism,

$$\omega_1 = \nu_1 \qquad \omega_l = \nu_l \prod_{r=1}^{l-1}(1-\nu_r), \quad l \geq 2, \tag{3.4}$$

with $\nu_l \overset{i.i.d.}{\sim} \text{Be}(1,\alpha)$ and $\alpha \sim \text{Ga}(a_\alpha, b_\alpha)$. On the other hand, the weight of the GW prior is given in the form

$$\omega_l = (1-\zeta)\zeta^{l-1}, \quad l = 1, 2, \ldots, \tag{3.5}$$

with $\zeta \sim \text{Be}(a_\zeta, b_\zeta)$. The method of generating ω_l in (3.5) can be envisioned as a simplified stick-breaking technique in which the stick is always broken with the same size of $1-\zeta$. By the use of a single parameter ζ, we can achieve ordered weights, $\omega_1 > \omega_2 > \ldots$, which resolves the identifiability issue that arises in inference about weights and locations when using the DP prior. But, we focus on the estimation of the offspring density, neither the weight nor location parameters; the feature of the GW prior brings no remarkable difference in estimating the offspring density (see, e.g., Section 3.3.3). Practically, posterior sampling ζ is easier to implement and has a lower time complexity than sampling ν_l, $l = 1, \ldots, L$ in the DP model. Empirically, however, the GW model requires more mixing components L and, therefore, more computing time for a similar uncertainty level in the prior model. Note that we use the blocked Gibbs sampler for posterior inference with truncation approximation to F, such that $F_L(\cdot) = \sum_{l=1}^{L} \omega_l \delta_{Z_l}(\cdot)$. The number of components, L, is a key factor for computing time.

The following strategy is used to specify the prior for the models. For the common parameter μ, we take the $\text{Exp}(a_\mu)$ prior with rate $a_\mu = 2T/n$, chosen as in

Section 3.1.2.1. The branching ratio, independent of the offspring density function under the models, is assigned the $\mathrm{Ga}(a_\gamma, b_\gamma)$ prior. We set the hyperparameters, considering the stability condition, such that $\Pr(0 < \gamma < 1) \approx 1$. We choose the inverse gamma distribution for the centering function F_0, which affords a recognizable posterior full conditional distribution – the piecewise truncated inverse gamma distribution – for Z_l.

The scale parameter β of the inverse gamma prior for Z_l is chosen using the offspring density expected over the random mixing distribution, $\mathrm{E}(g(x))$, $x \equiv t - t_y$, where t_y is the parent of t. Then, the expectation $\mathrm{E}(X)$ of the density function indicates the mean distance from the parent, which can be derived as $\beta/4$ with the inverse gamma centering distribution with shape 3 and scale β (Appendix G). The mean distance should be within the effective support, $(0, T_O)$, for the offspring density function. We choose T_O, considering expert advice or existing theories for the distance x in each application to which the model is applied. The hyperparameter β is selected such that $\Pr(0 < \beta/4 < T_O) \approx 0.999$.

The other key parameters of the models are α of the DP prior and (a_ζ, b_ζ) of the GW prior. Generally, α is well known for modulating the discreteness and the variability of F around F_0. In the uniform DP mixture, α also contributes to the smoothness of $f(t|F)$. Since the kernel is the uniform density function, the finite DP mixture yields a stepwise function. Larger values of α produce less discrete mixing distributions with more distinct locations and, thus, more steps in $f(t|F)$. Empirically, we select the gamma prior for α whose effective support reaches 20, which offers a good balance of the model flexibility and smoothness.

Finally, the truncation level L of the uniform mixtures is determined, taking into account the prior expectation of the partial sum of weights: $\mathrm{E}(\sum_{l=1}^{L}\omega_l|\alpha) = 1 - \{\alpha/(\alpha+1)\}^L$ for DP and $\mathrm{E}(\sum_{l=1}^{L}\omega_l|\zeta) = 1-\zeta^L$ for GW, where the expectation of the partial sum marginalized over α or ζ is approximately 1 under the selected L. Section 3.2.1.2 will present the hierarchical representation of the above uniform-mixture-based models as well as an overview of posterior inference methods.

3.1.3 Fully NPB model for the full HP intensity function

To achieve fully nonparametric model extensions, we propose models obtained by combining a semiparametric immigrant model with a semiparametric offspring model. Denote by I the immigrant process and by O the superposition of offspring processes. With the branching structure, the fully nonparametric Bayesian (NPB) model based on two Erlang mixtures is defined as

$$\begin{aligned}
\lambda^*(t) &= \sum_{l=1}^{L_I}\nu_{l,I}\mathrm{Ga}(t|l,\theta_I^{-1}) + \sum_{t_i<t}\Big[\sum_{l=1}^{L_O}\nu_{l,O}\mathrm{Ga}(t-t_i|l,\theta_O^{-1})\Big], \quad t\in\mathbb{R}^+ \\
\nu_{l,I} &\overset{ind.}{\sim} \mathrm{Ga}(A_{l,I},c_0), \;\; A_{l,I}\equiv c_0[G_0(l\theta_I) - G_0((l-1)\theta_I)], \\
\nu_{l,O} &\overset{ind.}{\sim} \mathrm{Ga}(A_{l,O},\eta), \;\; A_{l,O}\equiv \alpha_0[F_0(l\theta_O) - F_0((l-1)\theta_O)],
\end{aligned} \tag{3.6}$$

where $G_0(t) = t/b_{G_0}$ and $F_0(t) = 1-\exp\{-t/b_{F_0}\}$. The prior is specified, following the strategy given for each Erlang-mixture-based semiparametric model in Sections 3.1.1 and 3.1.2.1.

Similarly, for the offspring density function supposed to be non-increasing, the uniform mixtures given in Section 3.1.2.2 can be substituted for the offspring Erlang

mixture of (3.6). With the branching structure, the fully NPB model using the uniform mixtures is defined as

$$\begin{aligned}
\lambda^*(t) &= \sum_{l=1}^{L_I} \nu_{l,I} \mathrm{Ga}(t|l, \theta_I^{-1}) + \sum_{t_i<t} \Big[\int \theta^{-1} \mathbf{1}_{[0,\theta)}(t-t_i) dF(\theta) \Big], \quad t \in \mathbb{R}^+ \\
\nu_{l,I} &\overset{ind.}{\sim} \mathrm{Ga}(A_{l,I}, c_0), \;\; A_{l,I} \equiv c_0[G_0(l\theta_I) - G_0((l-1)\theta_I)] \\
F &\sim \mathrm{DP}(F_0, \alpha_0) \;\; / \;\; \mathrm{GW}(F_0, \zeta),
\end{aligned} \tag{3.7}$$

where the centering distribution F_0 in both DP and GW priors is set to an inverse gamma distribution with shape 3 and mean $\beta/2$. Again, the prior for the model parameters is specified in the same manner as in Sections 3.1.1 and 3.1.2.2.

3.2 Posterior inference

3.2.1 Hierarchical model representation

3.2.1.1 Mixture model for the immigrant intensity function

Here is the hierarchical representation of the semiparametric model incorporating the Erlang mixture for the immigrant intensity function, as presented in Section 3.1.1. We will refer to this as: the semiparametric model with the immigrant Erlang mixture; or simply the immigrant Erlang mixture model.

$$
\begin{aligned}
p(\boldsymbol{t}|\boldsymbol{\nu},\theta,\boldsymbol{\xi},\boldsymbol{y}) &= \Big[\prod_{t_i\in I}(\sum_{k=1}^{L}\nu_k)\mathrm{Ga}(t_i|\xi_i,\theta^{-1})\Big]\exp\Big\{-\sum_{l=1}^{L}\nu_l\int_0^T \mathrm{Ga}(u|l,\theta^{-1})du\Big\} \\
&\quad\times\Big[\prod_{t_i\in O}\gamma\mathrm{Exp}(t_i-t_{y_i}|\alpha)\Big]\exp\Big\{-\sum_{t_i\in\boldsymbol{t}}\gamma\int_0^{T-t_i}\mathrm{Exp}(s|\alpha)ds\Big\} \\
p(\xi_i|\boldsymbol{\nu}) &\propto \sum_{l=1}^{L}\frac{\nu_l}{\sum_k\nu_k}\delta_l(\xi_i),\quad i:t_i\in I \\
p(\boldsymbol{y}) &= \delta_0(y_1)\prod_{i=2}^{n}\mathrm{Unif}(y_i|0,1,\ldots,i-1),
\end{aligned}
\tag{3.8}
$$

where $\mathrm{Unif}(y_i|0,1,\ldots,i-1)$ denotes a discrete uniform probability mass function suppored on a set of non-negative integers $\{0,1,\ldots,i-1\}$. We adopted the discrete uniform prior choice for the branching structure, which was introduced in Ross (2021) for HP parametric models. $\delta_l(t)$ is the Dirac delta function, such that $\delta_l(t)=1$ for $t=l$. This representation omits priors for ν_l, θ, c_0, and b_{G_0}, specified in Section 3.1.1.

The HP cluster representation factorizes the likelihood so that we can estimate separately the immigrant intensity and the excitation function. So, given the branching structure, we can directly apply the inference method for the NHPP model in Section 2.1.3 to estimating the immigrant intensity. For better mixing in the posterior distribution of θ, we replace the Metropolis-Hastings (M-H) algorithm with the Hamiltonian Monte Carlo (HMC) algorithm. Appendix C details the MCMC posterior simulation for the model.

3.2.1.2 Mixture models for the excitation function

The hierarchical representation of the semiparametric model with the Erlang mixture for the excitation function is as follows:

$$p(\boldsymbol{t}|\boldsymbol{\nu},\theta,\boldsymbol{\xi},\boldsymbol{y}) = \mu^{n_I}\Big[\prod_{i:t_i\in O}(\sum_{k=1}^{L}\nu_k)\mathrm{Ga}(t_i - t_{y_i}|\xi_i,\theta^{-1})\Big]$$
$$\times \exp\{-\mu T\}\exp\Big\{-\sum_{t_i\in\boldsymbol{t}}\sum_{l=1}^{L}\nu_l\int_0^{T-t_i}\mathrm{Ga}(u|l,\theta^{-1})du\Big\} \qquad (3.9)$$
$$p(\xi_i|\boldsymbol{\nu}) \propto \sum_{l=1}^{L}\frac{\nu_l}{\sum_k\nu_k}\delta_l(\xi_i),\ \ i: t_i\in O,$$

where n_I denotes the number of observations belonging to the immigrant process, $|\{t_i : y_i = 0,\ \ i = 1,\ldots,n\}|$. Latent variables $\{y_1, y_2, \ldots, y_n\}$ are assigned the discrete uniform prior, as in (3.8), with the Dirac delta function for y_1. You can find the priors for model parameters ν_l, θ, and b_{F_0} in Section 3.1.2.1. The model will be called: the semiparametric model with the offspring Erlang mixture; or the offspring Erlang mixture model.

We can take posterior samples for model parameters using the general Gibbs sampling method. Although the Erlang mixture has been adjusted for the excitation function, the weights ν_l retain conjugate (gamma) priors. So, we can draw posterior samples for ν_l using ready prior-to-posterior updating. The constant immigrant intensity μ also has a well-known distribution for posterior sampling under its exponential prior. Details of the MCMC posterior simulation are available in Appendix D.

The other semiparametric approach, based on uniform mixtures, has the fol-

lowing model representation

$$p(\boldsymbol{t}|\boldsymbol{y},\mu,\gamma,\boldsymbol{\xi},\boldsymbol{Z}) \propto \mu^{n_I}\gamma^{n_O}\Big[\prod_{i:t_i\in O}\frac{1}{Z_{\xi_i}}1_{(0,Z_{\xi_i})}(t_i-t_{y_i})\Big]\exp\{-\mu T\}\exp\Big\{-\sum_{l=1}^{L}\gamma\omega_l K(Z_l)\Big\}$$
$$p(\xi_i|\boldsymbol{\omega}) = \sum_{l=1}^{L}\omega_l\delta_l(\xi_i),\ \ i:t_i\in O, \tag{3.10}$$

$$\boldsymbol{\omega} = \begin{cases} \omega_1=\nu_1,\ \ \omega_l=\nu_l\prod_{r=1}^{l-1}(1-\nu_r),\ \ l=2,\ldots,L-1,\ \ \omega_L=\prod_{r=1}^{L-1}(1-\nu_r), \\ \qquad\qquad \text{with } \nu_l\sim \text{Be}(1,\alpha),\ \ \alpha\sim\text{Ga}(a_\alpha,b_\alpha),\ \ l=1,2,\ldots,L-1 \quad \text{for DP;} \\ \omega_l=(1-\zeta)\zeta^{l-1},\ \ l=1,2,\ldots,L-1,\ \ \omega_L=\zeta^{L-1},\ \ \zeta\sim\text{Be}(a_\zeta,b_\zeta) \quad \text{for GW,} \end{cases}$$

$$K(Z_l) = \begin{cases} \frac{1}{Z_l}\Big(\sum_{i=1}^{n}(T-t_i)\Big) \ \text{ for } \ Z_l > T-t_1; \\ r+\frac{1}{Z_l}\Big(\sum_{i=r+1}^{n}(T-t_i)\Big) \ \text{ for } \ T-t_{r+1}<Z_l\le T-t_r,\ \ r=1,\ldots,n-1; \\ n \ \text{ for } \ 0<Z_l\le T-t_n. \end{cases}$$

The hierarchical representation excludes the priors for the branching structure, but they are the same as in (3.8), and for model parameters Z_l, μ, γ, and β, given in Section 3.1.2.2. Throughout the thesis, the models are also referred to as uniform-mixture-based models.

The posterior inference method for the models is based on the blocked Gibbs sampler. Specifically, we use the algorithm for estimating the offspring density function, modeled by the uniform mixture with the DP or GW prior. With an exponential prior for the constant immigrant intensity μ, we can simply draw the posterior sample for the parameter from a well-known distribution. The key parameter Z_l also has a conjugate

(inverse gamma) prior. An exponential term in (3.10) involves the function $K(Z_l)$, which is defined differently based on the postion of Z_l in a partition of $\mathbb{R}^+$. The parameter Z_l, therefore, has a piecewise truncated (inverse gamma) distribution for its posterior sampling. A complete description of the MCMC inference method are given in Appendix E.

3.2.1.3 Fully NPB model

Our fully nonparametric models are constructed by combining two semiparametric models. Correspondingly, the hierarchical model representation is created by incorporating semiparametric model representations. For example, we can derive the NPB model representation based on the immigrant and offspring Erlang mixtures by replacing μ^{n_I} and $\exp\{-\mu T\}$ of (3.9) with $\left[\prod_{t_i \in I}(\sum_{k=1}^{L} \nu_k)\mathrm{Ga}(t_i|\xi_i, \theta^{-1})\right]$ and $\exp\left\{-\sum_{l=1}^{L} \nu_l \int_0^T \mathrm{Ga}(u|l, \theta^{-1})du\right\}$ of (3.8). The latent variables of the branching structure remain the discrete uniform priors and the Dirac delta function.

Given the branching structure, we can separately estimate the immigrant intensity and the excitation function. Therefore, the inference methods for each mixture (appendices C, D, E) remain applicable to the nonparametric model for each function estimation.

3.2.2 Inference for functional

3.2.2.1 First- and second-order intensities

Presented here are the first- and second-order intensity functions, which can be used to characterize HPs in conjunction with the conditional intensity function. We begin by defining the HP conditional intensity function using the counting process, which helps clarify the notion of the two functions. The conditional intensity function can be interpreted as the conditional expected rate of arrivals at t, given the history $\mathcal{H}(t)$, times observed in $[0, t)$. So, we can write the function as $\lambda^*(t) = \mathrm{E}(dN(t)/dt|\mathcal{H}(t))$, where $N(t)$ is a counting process at t. The first-order intensity represents an averaged intensity function over the history. In other words, the first-order intensity $\lambda(t)$ is the expectation of the HP conditional intensity function, defined as $\lambda(t) = \mathrm{E}(\lambda^*(t))$. The second-order intensity $r_t(\tau)$ is the HP covariance structure, given by $r_t(\tau) \equiv \mathrm{Cov}(dN(t)/dt, dN(t+\tau)/d\tau)$, $\tau > 0$ and $r_t(\tau) \equiv \lambda(t)$, $\tau = 0$. Hawkes (1971a) presented derivations of the two intensities under a simple parametric model consisting of a constant immigrant intensity and an exponential offspring density function.

To derive the first-order intensity under our models, we start with a general expression of the first-order intensity, $\lambda(t) = \mu + \int_0^t h(u)\lambda(t-u)du$ (Laub et al., 2015). Taking the Laplace transform of the expression yields

$$\begin{aligned}
\mathcal{L}[\lambda(t)](s) &= \mathcal{L}[\mu](s) + \mathcal{L}\Big[\int_0^t h(u)\lambda(t-u)du\Big](s), \quad \text{where} \\
\mathcal{L}[\mu](s) &= \int_0^\infty \exp\{-st\}\mu dt = \frac{\mu}{s} \qquad (3.11) \\
\mathcal{L}\Big[\int_0^t h(u)\lambda(t-u)du\Big](s) &= \mathcal{L}[h(t)](s)\mathcal{L}[\lambda(t)](s).
\end{aligned}$$

Under the offspring Erlang mixture model, for which the offspring density is defined as $h(t) = \sum_{l=1}^{L} \nu_l(\theta^{-l}/\Gamma(l))t^{l-1}\exp\{-t\theta^{-1}\}$, $\nu_l = \gamma\omega_l$, the Laplace transform of the excitation function is derived as

$$\begin{aligned}\mathcal{L}[h(t)](s) &= \int_0^\infty \exp\{-st\}\Big(\sum_{l=1}^{L}\nu_l\frac{\theta^{-l}}{\Gamma(l)}t^{l-1}\exp\{-t\theta^{-1}\}\Big)dt \\ &= \sum_{l=1}^{L}\nu_l\int_0^\infty \frac{\theta^{-l}}{\Gamma(l)}t^{l-1}\exp\{-t(s+\theta^{-1})\}dt \\ &= \sum_{l=1}^{L}\nu_l\Big(\frac{1}{s\theta+1}\Big)^l.\end{aligned} \tag{3.12}$$

Substituting (3.12) into (3.11) provides an analytical form of the Laplace transform of the first-order intensity, $\mathcal{L}[\lambda(t)](s) = \mu/\big[s\big(1-\sum_{l=1}^{L}\nu_l(s\theta+1)^{-l}\big)\big]$.

Similarly, under the uniform-mixture-based models with the excitation function $h(t) = \sum_{l=1}^{L}(\nu_l/\theta_l)1_{(0,\theta_l)}(t)$, $\nu_l = \gamma\omega_l$, we can derive the Laplace transform of the first-order intensity as follows

$$\mathcal{L}[h(t)](s) = \int_0^\infty \exp\{-st\}\Big(\sum_{l=1}^{L}\frac{\nu_l}{\theta_l}1_{(0,\theta_l)}(t)\Big)dt = \frac{1}{s}\sum_{l=1}^{L}\frac{\nu_l}{\theta_l}(1-\exp\{-s\theta_l\}). \tag{3.13}$$

Substituting (3.13) into (3.11) yields an analytical form of the Laplace transformed first-order intensity, $\mathcal{L}[\lambda(t)](s) = \mu/\big[s\big(1-(\sum_{l=1}^{L}(\nu_l/\theta_l)(1-\exp\{-s\theta_l\}))/s\big)\big]$. By numerical inverse transforms, we can readily obtain the first-order intensities under the two semiparametric models. Further development of the intensity can be accomplished by replacing μ with $\mu(t)$ for our nonparametric models. Since the immigrant intensity of the nonparametric models is the Erlang mixture, we can obtain an analogous result to (3.12) for the Laplace transform of $\mu(t)$. Therefore, Laplace transforms of the intensity of the models remain analytically available.

In terms of second-order intensity, no analytical solution to the Laplace transform of the function is available. As an alternative, we propose a simpler, intuitive, but computationally intensive method.

The Monte Carlo integration underlies the method and is used to approximate the expectations in $r_t(\tau) = \mathrm{E}((dN(t)/dt)(dN(t+\tau)/d\tau)) - \mathrm{E}(dN(t)/dt)\mathrm{E}(dN(t+\tau)/d\tau)$. Each expectation of the second term denotes the first-order intensities at t and $t+\tau$, for example, $\mathrm{E}(dN(t)/dt) = \mathrm{E}(\mathrm{E}(dN(t)/dt|\mathcal{H}(t))) = \mathrm{E}(\lambda^*(t)) = \lambda(t)$. With the Monte Carlo integration, we can find the expectation by averaging realizations of $\lambda^*(t)$. We can obtain the expectation $\mathrm{E}(dN(t+\tau)/d\tau)$ in the same manner. For posterior inference, we draw multiple realizations of $\lambda^*(t)$ and $\lambda^*(t+\tau)$ at each MCMC iteration and average each sequence to find $\lambda(t)$ and $\lambda(t+\tau)$. Collections of $\lambda(t)$ and $\lambda(t+\tau)$ from each iteration provide the posterior distributions of the two first-order intensities at t and $t+\tau$.

By the double expectation theorem, the first expectation of the second-order intensity can be rewritten as $\mathrm{E}((dN(t)/dt)(dN(t+\tau)/d\tau)) = \mathrm{E}(\mathrm{E}((dN(t)/dt)(dN(t+\tau)/d\tau)|\mathcal{H}(t+\tau))) = \mathrm{E}(dN(t)/dt\mathrm{E}(dN(t+\tau)/d\tau|\mathcal{H}(t+\tau))) = \mathrm{E}(dN(t)/dt\lambda^*(t+\tau))$. For posterior simulation, we sample realizations of $dN(t)/dt\lambda^*(t+\tau)$, and then taking the average gives us an approximation to $\mathrm{E}(dN(t)/dt\lambda^*(t+\tau))$ at each MCMC iteration. Unlike the simple sampling of $\lambda^*(t+\tau)$, computing $dN(t)/dt = \lim_{m\to 0}(N(t+m) - N(t))/m$ may be challenging. We bypass the calculation by replacing the limit with a small value of m. We empirically choose m such that, for any $m_1, m_2 < m$ and fixed $t > 0$, $(N(t+m_1) - N(t))/m_1 \approx (N(t+m_2) - N(t))/m_2$. Some results on the first- and

second-order intensities are provided in Section 3.3.2.3.

3.2.2.2 Predicted counts of future events

In this section, we describe a simple inference method for prediction. Denote by $(0, T)$ the observation window and by $A = [T, T^*)$ the prediction interval. The predictive count $N(A)$ indicates the number of points within the prediction interval, A. For posterior prediction, we draw a realization, on $(0, T) \cup A = (0, T^*)$, from a HP with intensity specified by model parameters sampled at each MCMC iteration. Counting points lying in A produces the predictive count $N(A)$, and the collection of $N(A)$ from each iteration gives the posterior distribution for $N(A)$. As alternatives, we consider $N_I(A)$ and $N_O(A)$ for the immigrant predictive count and the offspring predictive count. The alternative counts are obtained by splitting the predictive count, $N(A)$, using the branching structure sampled at each MCMC iteration. So, the number of points in A whose latent variables are equal to 0 becomes $N_I(A)$. Similarly, $N_O(A)$ can be defined under the condition that latent variables are not 0.

3.3 Simulation study

Using synthetic data examples, we investigate here the performance of our models in estimating intensity/density. Section 3.3.2.1 and Section 3.3.2.2 focus on the semiparametric models with the Erlang mixture for the immigrant intensity or the excitation function. Section 3.3.2.3 mainly illustrates the nonparametric model and highlights the model performance by comparison with other models. Section 3.3.3 deals

with decreasing offspring density examples for the semiparametric models specializing in non-increasing function estimation.

In the following sections, we will evaluate models primarily by graphical comparison between estimated functions and underlying intensity/density functions. For more objective comparison, Section 3.3.1 introduces some quantitative measures.

3.3.1 Criteria for model assessment and comparison

The graphical comparison between estimated and underlying functions is an intuitive method. However, a discrepancy in the graphical comparison may occur when data lacks representativeness. To clarify the source of discrepancies, we should also conduct the following comparison:

- the histogram of immigrant points and the normalized immigrant intensity; and
- the histogram of distances between offspring and their parents and the normalized offspring intensity.

Unlike the simple-form normalized immigrant intensity $\mu(t)/\int_0^T \mu(u)du$, the normalized offspring intensity corresponding to the histogram of the distances is a mixture of offspring density functions supported on $(0, T - t_i)$, where $\{t_i\}$ is a set of all parents. The shape of the offspring function arises from a sampling method for the offspring points. In a HP sampling method based on the cluster representation (e.g., Møller and Rasmussen, 2005, 2006), we draw the offspring distances from a parent t_i using the offspring density function defined on $(0, T - t_i)$. So, the collection of all the offspring

distances has a mixture of the offspring density functions. We come up with a function, named the *aggregate offspring density function*, that is a mixture of offspring densities comparable with the histogram of the offspring distances. Denoted by $\boldsymbol{t}$ an observed point pattern, $\{t_1, t_2, \ldots, t_n\}$, the aggregate offspring density function is

$$q(x|\boldsymbol{t}, \tilde{\boldsymbol{y}}, h) = \frac{\sum_{k=1}^{\tilde{n}} H(T - t_{\tilde{y}_k}) \times \left(\frac{h(x)}{H(T-t_{\tilde{y}_k})} \mathbf{1}_{(0,T-t_{\tilde{y}_k})}(x) \right)}{\sum_{k=1}^{\tilde{n}} \int_0^{T-t_{\tilde{y}_k}} h(s)ds} = \frac{\sum_{k=1}^{\tilde{n}} h(x) \mathbf{1}_{(0,T-t_{\tilde{y}_k})}(x)}{\sum_{k=1}^{\tilde{n}} H(T - t_{\tilde{y}_k})}. \tag{3.14}$$

where $H(t)$ denotes the cumulative excitation function, defined as $H(t) = \int_0^t h(s)ds$, and $\tilde{\boldsymbol{y}} = \{\tilde{y}_k\}$, $k = 1, 2, \ldots, \tilde{n}$, the set of unique values of the branching structure $\{y_1, \ldots, y_n\}$ with $\tilde{n} < n$. In the numerator, $\left(\frac{h(x)}{H(T-t_{\tilde{y}_k})} \mathbf{1}_{(0,T-t_{\tilde{y}_k})}(x) \right)$ represents the offspring density function of the offspring distance from parent $t_{\tilde{y}_k}$, as in the cluster representation-based sampling algorithm. The unnormalized weight $H(T - t_{\tilde{y}_k})$ is also taken from the algorithm, indicating the expected number of points that will be drawn from the offspring density function. To cover all offspring distances, we add up the unnormalized mixtures. Hence, normalizing the mixture yields a distribution to which the offspring histogram can conform.

During the remainder of this section, we will discuss some numerical criteria that enable quantitative model comparisons. The first is the total variation distance (TVD). Based on the target functions to which our estimated cumulative distributions are matched, we define three TVDs:

- TVD^a for the immigrant cumulative distribution;
- TVD^b for the offspring cumulative distribution; and

- TVD^c for the empirical offspring distribution.

TVD^c has a target function based on data, while the other TVDs compute the targets using the underlying intensity functions, such that $\int_0^t \mu(u)du / \int_0^T \mu(s)ds$, $t \in (0, T)$ for TVD^a and $\int_0^x g(u)du$, $x > 0$ for TVD^b. So, we can use TVD^c as a goodness-of-fit measure as well as the aggregate offspring density function. We calculate TVD^c by matching the cumulative aggregate offspring density function to the empirical offspring distribution.

As another quantitative tool for model assessment, we consider two criteria based on the branching structure: the immigrant/offspring cluster sizes and the immigrant/offspring misclassification with its standardized summary, the misclassification rate.

The cluster sizes are defined as $n_I = |\{i : y_i = 0, \ \ i = 1, \ldots, n\}|$ for the immigrant size and $n_O = |\{i : y_i \neq 0, \ \ i = 1, \ldots, n\}|$ for the offspring size. We can compute the sizes by plugging the branching structure, sampled at each MCMC iteration, into the equations. Then, we can compare their posterior means with observed counts of immigrant/offspring points to evaluate the model.

Next, we introduce the three different misclassification evaluation tools as follows:

- $M_I = |\{i : y_i = 0, \ \ y_i^{true} \neq 0, \ \ i = 1, \ldots, n\}|$ for the immigrant misclassification;
- $M_O = |\{i : y_i \neq 0, \ \ y_i^{true} = 0, \ \ i = 1, \ldots, n\}|$ for the offspring misclassification;
- $R = (M_I + M_O)/n$ for the misclassification rate,

where y_i^{true}, $i = 1, \ldots, n$ denotes the observed branching structure. The immigrant misclassification identifies how many points classified as immigrants by the branching structure are actually not immigrants. In a similar way, the offspring misclassification is also explicable. The misclassification rate gives aggregate and standardized information about the misclassification criteria.

3.3.2 Synthetic data examples for Erlang mixture models

3.3.2.1 Illustration of immigrant Erlang mixture model

We generated 504 points (397 for immigrants and 107 for offspring) from a HP with intensity $\lambda^*(t) = 400[0.6\text{We}(t|1.5, 2000) + 0.4\text{We}(t|7, 8000)] + 0.2\sum_{t_i<t}\text{Exp}(t-t_i|2)$, $t \in (0, 10000)$. The immigrant intensity function has a non-standard shape, based on a mixture of two Weibull densities. The simulation study is designed to examine the flexibility of the immigrant Erlang mixture of the semiparametric model given by (3.1).

Following the prior specification in Section 3.1.1, we took the set of priors: the $\text{Lo}(2, 500)$ prior for θ with $L = 80$; $\text{Exp}(0.1)$ for c_0; $\text{Exp}(0.0252)$ for b_{G_0}, $\text{Ga}(2, 4)$ for γ, and $\text{Exp}(1)$ for a.

Prior specification provides enough prior uncertainty to cover the true function and even the entire panel (left of Figure 3.1). This is the case for all models with the immigrant Erlang mixture. Therefore, in the following examples, we omit the prior uncertainty bands from the panels for immigrant intensity estimates.

The posterior mean for the immigrant intensity function reveals the bimodality of the underlying intensity, the interval estimates encompassing the true values (left

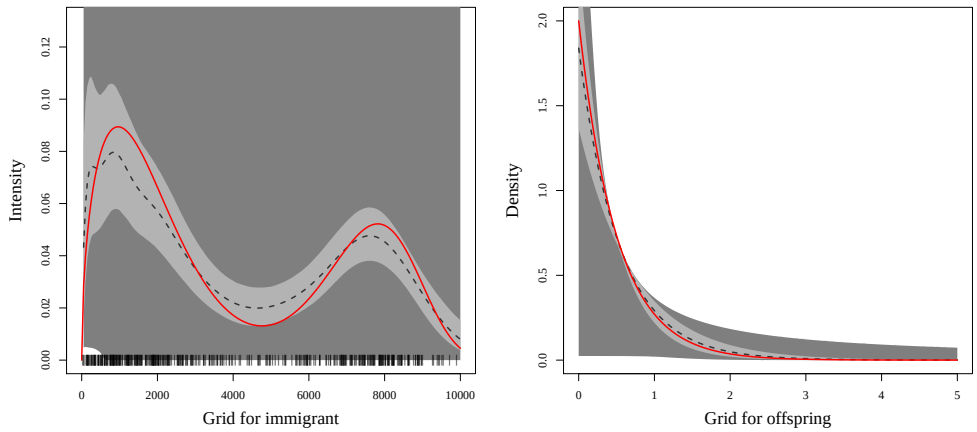

Figure 3.1: Weibull mixture immigrant and exponential offspring example. Semiparametric model with Erlang mixture immigrant intensity. The left panel displays the posterior mean (dashed) and the posterior 95% interval estimates (light gray) for the immigrant intensity function with: the prior 95% uncertainty bands (dark gray); the underlying immigrant intensity (red solid); and the point pattern (bar) at the bottom. Similarly, the right panel demonstrates the prior uncertainty bands (dark gray), the posterior mean (dashed), and the interval estimates (light gray) with the underlying offspring density function (red solid).

of Figure 3.1). As in the underlying function, the model has an exponential density function for the offspring density. As a result of choosing such a model, the posterior mean offspring density function is almost identical to the actual value (right of Figure 3.1). For both cluster sizes, the posterior means for n_I and n_O are 392 and 112 with a standard deviation of 6. The estimates are comparable with the true immigrant and offspring sizes, 397 and 107. The posterior means of the misclassification are $M_I = 22$, $M_O = 27$, and $R = 0.098$, respectively. The branching ratio, γ, has the posterior mean of 0.225 and the posterior 95% interval estimates of (0.180,0.275), which contains true $\gamma = 0.2$. The inferences are made on the basis of $20,000$ posterior samples taken after $20,000$ burn-in steps.

3.3.2.2 Illustration of offspring Erlang mixture model

We explore the semiparametric model appearing in (3.2) with two examples, in which the underlying offspring density functions are either the unimodal Weibull or the two-component Weibull mixture. The offspring functions simulate two possibilities. Secondary events may be activated after a period of idleness or secondary events may be intensified twice at specific times.

The HP intensity function of the first example is defined as

$$\lambda^*(t) = 0.01 + 0.8 \sum_{t_i<t} \text{We}(t - t_i|2, 2), \quad t \in (0, 10000).$$

We sampled 534 time points (105 for immigrants and 429 for offspring) from the intensity function. The priors we used for model fitting are : Exp(37.5) for μ, Lo$(2, 0.3)$ for θ with $L = 50$, Exp(0.168) for b_{F_0}, and fixed hyperparameters $(2, 4)$ for (α_0, η).

The second example intensity function is defined as

$$\lambda^*(t) = 0.01 + 0.8 \sum_{t_i<t} [0.6\text{We}(t - t_i|2, 2) + 0.4\text{We}(t - t_i|5, 10)], \quad t \in (0, 10000),$$

from which we generated 500 points (101 for immigrants 399 for offspring). We placed priors Exp(38.6) on μ, Lo$(2, 0.7)$ on θ with $L = 50$, Exp(0.0673) on b_{F_0}, and fixed hyperparameters $(2, 4)$ on (α_0, η) for model fitting.

Table 3.1 provides posterior estimates of the immigrant intensity, the branching ratio, the cluster sizes, and the misclassification. The posterior means of μ and γ are close to the true values of $(0.01, 0.8)$. The estimated cluster sizes are comparable to the true sizes: immigrant/offspring sizes (105/429) for the Weibull example and

	μ	γ	Cluster size		Misclassification		
			n_I	n_O	M_I	M_O	R
Ex1	0.011(0.001)	0.798(0.039)	107(3)	427(3)	8(3)	6(1)	0.026(0.006)
Ex2	0.011(0.001)	0.786(0.041)	106(5)	394(5)	20(5)	15(1)	0.071(0.010)

Table 3.1: Weibull (top) and Weibull mixture (bottom) examples. Semiparametric model with Erlang mixture offspring intensity. The posterior means and standard deviations for parameters (μ, γ) and quantitative measures.

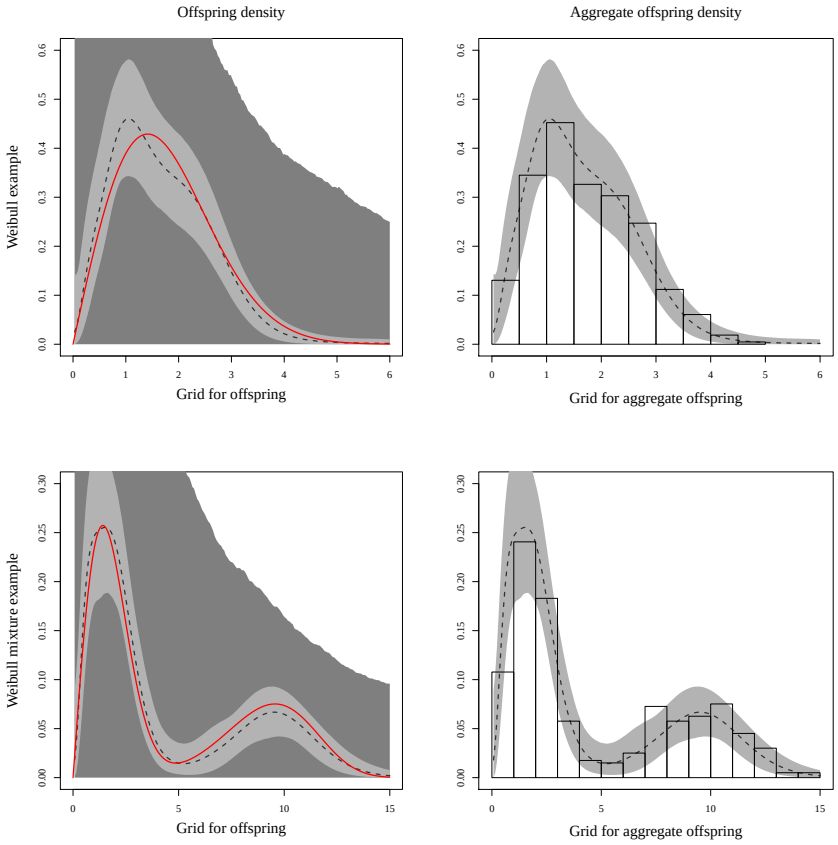

Figure 3.2: Weibull (top) and Weibull mixture (bottom) examples. Semiparametric model with Erlang mixture offspring intensity. The left panels present the posterior means (dashed) and posterior 95% interval estimates (light gray) for offspring density functions with: prior uncertainty bands (dark gray) and underlying density functions (red solid). The right panels display the posterior means and interval estimates for aggregate offspring density functions as well as data histograms.

(101/399) for the Weibull mixture example. The estimated functions corresponding to the non-standard-shape underlying offspring densities corroborate the model flexibility for the offspring density function (left of Figure 3.2). In the right column are shown the differences between the aggregate offspring density function and the data histogram. The top-right panel indicates that the local discrepancy around $x = 1.8$ in offspring density estimation on the left is due to the (non-representative) data set. For inference, we used $20,000$ posterior samples, obtained by $20,000$ burn-in.

3.3.2.3 Illustration of fully NPB model

The section involves three examples with underlying HP intensity functions:

- $\lambda^*(t) = 0.02 + 0.8 \sum_{t_i<t} \text{Exp}(t - t_i|1),\ t \in (0, 5000)$
- $\lambda^*(t) = 200\text{We}(t|3, 5000) + 0.6 \sum_{t_i<t} [0.6\text{We}(t - t_i|2, 2) + 0.4\text{We}(t - t_i|5, 10)],\ t \in (0, 10000)$
- $\lambda^*(t) = 200[0.6\text{We}(t|1.5, 2000) + 0.4\text{We}(t|7, 7500)] + 0.6 \sum_{t_i<t} [0.6\text{We}(t - t_i|2, 2) + 0.4\text{We}(t - t_i|5, 10)],$ $t \in (0, 10000)$

Each of the HPs generates 534 (105 for immigrants/429 for offspring), 500 (202/298), and 542 (206/336) point observations, respectively. We made the following inferences based on: $30,000$ posterior samples (after 10,000 burn-in) for the first two examples; and $10,000$ posterior samples (after 30,000 burn-in) for the last example.

In the first example, the data-generating HP uses a conventional choice for conditional intensity. Hawkes (1971a) introduced the HP with the intensity function, and Adamopoulos (1976) utilized the HP intensity function in earthquake modeling. The exponential density function remains a common choice to account for the time

behavior in the excitation function (e.g., Mohler, 2014). In this section, we will compare our nonparametric model with a parametric model with the conventional intensity. Since the data set is derived from the parametric model, the parametric model serves as the gold standard. We also fit the semiparametric model (3.2) (a simpler alternative to the nonparametric model with constant immigrant) to the data for the comparison.

The prior for each model is as follows: (parametric model) Exp(18.7) for μ, Ga$(1, 5)$ for γ, and Exp(1) for rate parameter of the exponential offspring density; (semiparametric model) Exp(18.7) for μ, Lo$(2, 0.5)$ for θ with $L = 50$, Exp(0.101) for b_{F_0}, and fixed hyperparameters $(2, 4)$ for (α_0, η); (nonparametric model) Lo$(2, 250)$ for θ_I with $L_I = 50$, Exp(0.1) for c_0, Exp(0.0534) for b_{G_0}, Lo$(2, 0.5)$ for θ_O with $L_O = 50$, Exp(0.101) for b_{F_0}, and fixed hyperparameters $(2, 4)$ for (α_0, η).

Figure 3.3 illustrates that our models perform well when it comes to intensity/density estimation. The Erlang mixtures of our models, however, have relatively large posterior interval estimates compared to the parametric model. Despite the fact that the immigrant Erlang mixture of the nonparametric model consists of decreasing or unimodal components, the posterior estimated intensity function captures well the underlying constant function. Furthermore, the Erlang mixture-based immigrant function retrieves the constant function without over-fitting.

Table 3.2 presents quantitative evaluations of the models. The results of our semiparametric and nonparametric models are comparable to those of the parametric model, with the exception of relatively large TVDs.

Figure 3.4 compares estimated first- and second-order intensity functions with

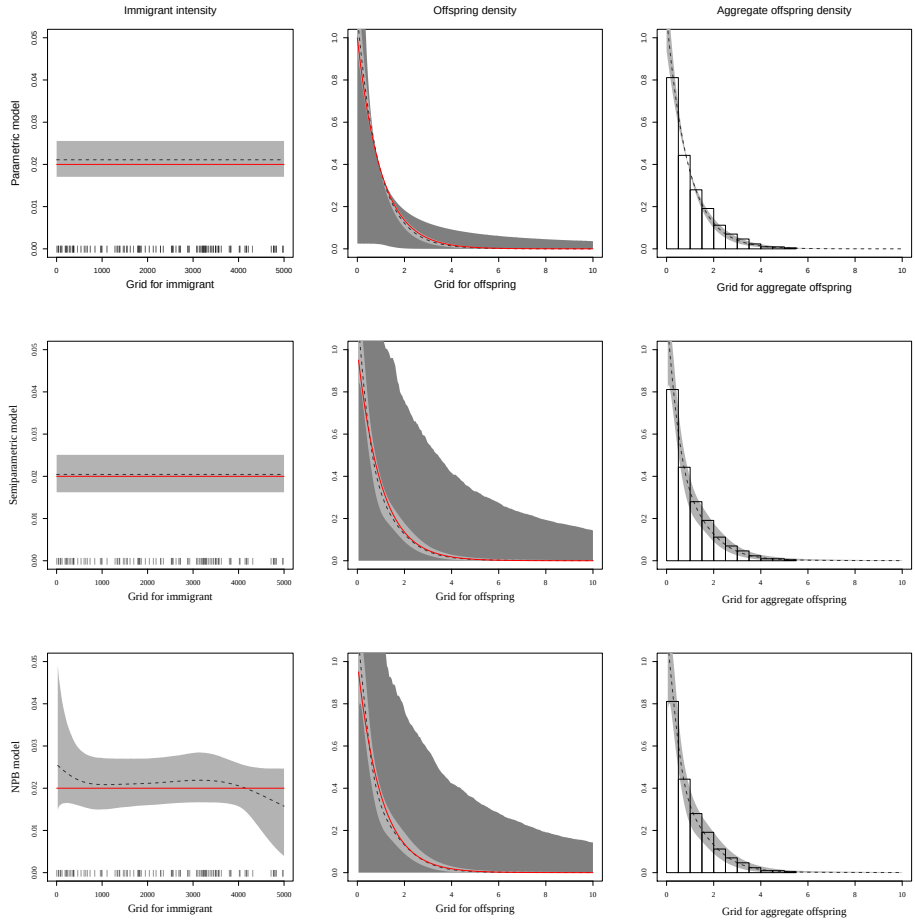

Figure 3.3: Constant immigrant and exponential offspring example. Parametric (top row), semiparametric (middle row), and nonparametric (bottom row) models. The left column represents immigrant intensity estimates. The middle and right columns display offspring density and aggregate offspring density estimates.

the true functions, which are analytically available in Hawkes (1971a). All models overestimate the true functions. Nevertheless, we found that their posterior interval estimates covered the true intensities (not shown). As the parametric model is considered to be the gold standard, our proposed models are justified in light of comparable results.

The second example is an extension of the second example in Section 3.3.2.2. In

Ex1	γ	Cluster size		Misclassification		
		n_I	n_O	M_I	M_O	R
Para	0.798(0.039)	105(4)	429(4)	10(3)	10(2)	0.036(0.006)
Semi	0.808(0.040)	101(5)	433(5)	9(3)	13(4)	0.041(0.008)
NPB	0.805(0.039)	103(5)	431(5)	10(3)	12(3)	0.040(0.008)
	TVD^a	TVD^b	TVD^c			
Para	0	0.038(0.025)	0.052(0.021)			
Semi	0	0.053(0.024)	0.055(0.020)			
NPB	0.037(0.029)	0.056(0.024)	0.056(0.021)			
Ex2	γ	Cluster size		Misclassification		
		n_I	n_O	M_I	M_O	R
Semi	0.604(0.042)	198(12)	302(12)	56(9)	60(6)	0.231(0.017)
NPB	0.603(0.042)	198(12)	302(12)	56(9)	60(6)	0.231(0.017)
	TVD^a	TVD^b	TVD^c			
Semi	0.035(0.020)	0.082(0.029)	0.093(0.029)			
NPB	0.047(0.020)	0.080(0.028)	0.090(0.028)			
Ex3	γ	Cluster size		Misclassification		
		n_I	n_O	M_I	M_O	R
NPB	0.622(0.039)	205(10)	337(10)	53(8)	55(4)	0.199(0.014)
	TVD^a	TVD^b	TVD^c			
NPB	0.064(0.023)	0.072(0.025)	0.080(0.026)			

Table 3.2: The posterior means and standard deviations for γ and quantitative measures (Ex1: constant immigrant and exponential offspring, Ex2: Weibull immigrant and Weibull mixture offspring, and Ex3: Weibull mixtures for both immigrant and offspring).

this example, we have replaced the constant immigrant intensity function with a Weibull density-based intensity function. Comparison is made between semiparametric and nonparametric models, where the constant immigrant of the semiparametric model is replaced by Weibull density-based intensity. Due to the analogy between the immigrant intensity function and the underlying intensity, the semiparametric model will perform better in immigrant intensity estimation, serving as a benchmark for evaluating the nonparametric model.

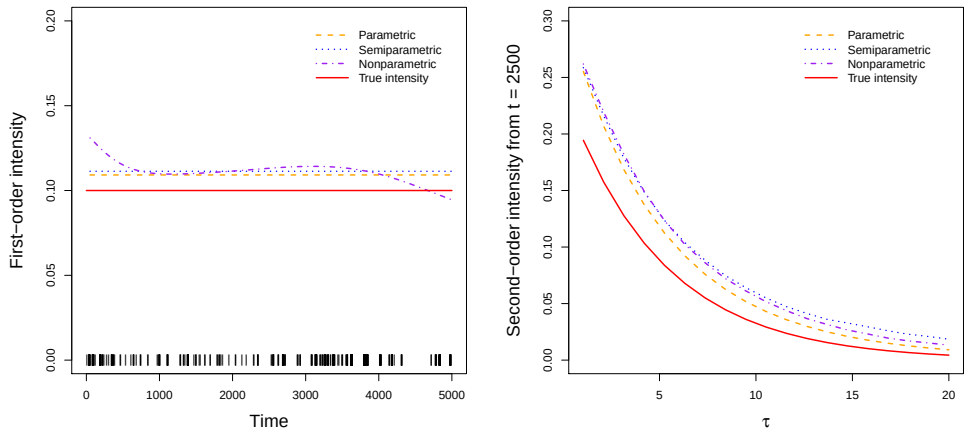

Figure 3.4: Constant immigrant and exponential offspring example. First-order (left) and second-order (right) intensity estimates from each model. Bars at the bottom of the left panel indicate the observed point pattern.

We set the model parameters to the following priors and values: (semiparametric model) Exp(0.0018), Exp(0.1), and Exp(0.0001) for (μ, a_ψ, b_ψ), Lo$(2, 0.7)$ for θ with $L = 50$, Exp(0.0673) for b_{F_0}, and fixed hyperparameters $(2, 4)$ for (α_0, η); (nonparametric model) Lo$(2, 500)$ for θ_I with $L_I = 50$, Exp(0.1) for c_0, and Exp(0.025) for b_{G_0}, Lo$(2, 0.7)$ for θ with $L = 50$, Exp(0.0673) for b_{F_0}, and fixed hyperparameters $(2, 4)$ for (α_0, η).

For the immigrant and offspring functions in Figure 3.5, the posterior estimates retrieve the Weibull and the Weibull mixture global patterns and contain the true functions within their 95% posterior interval estimates. In Table 3.2, the relatively lower TVD^a of the semiparametric model is explained by the fact that it has the same immigrant intensity as the underlying function. Each model struggles to find the second mode of the offspring density function (second column of Figure 3.5). The estimated

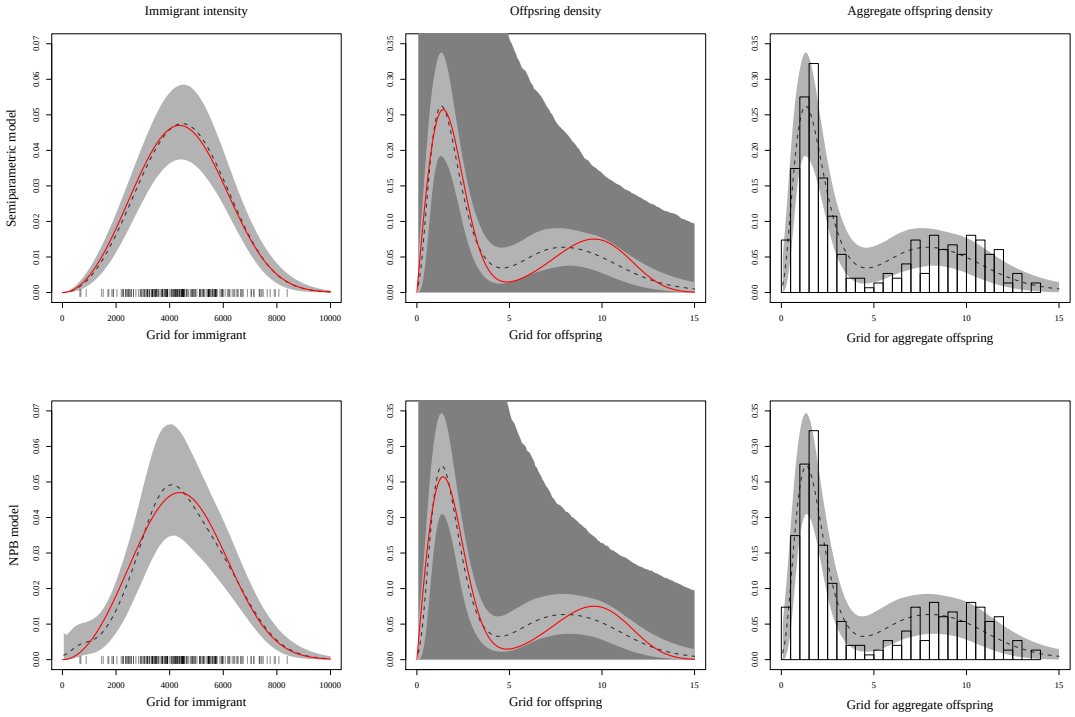

Figure 3.5: Weibull immigrant and Weibull mixture offspring example. Semiparametric (top) and nonparametric (bottom) models. Estimates for the immigrant intensity (left), offspring density (middle), and aggregate offspring density (right) functions.

aggregate offspring density functions of the figure suggest that the difficulty is unrelated to the data set. Large TVD^c also supports this claim with a large discrepancy between the estimated function and the data. The reason for this could be found from the relatively large misclassification rate in Table 3.2, compared to the other examples. From the fact that $M_O = 60$ of $n_O = 302$ points classified as offspring were actually immigrants, it can be conjectured that the misclassified offspring points contributed to the shift in the second mode.

The intensity function of the last example is more challenging for standard parametric models to cover. We fit only the nonparametric model to the data to see

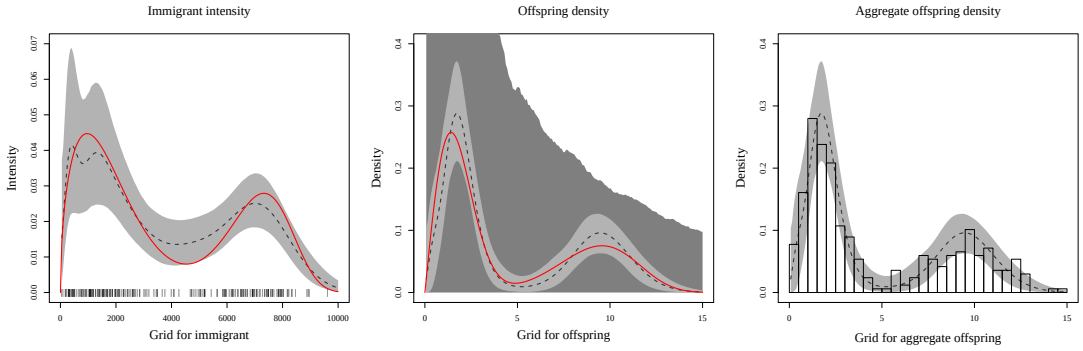

Figure 3.6: Example of Weibull mixtures for both immigrant and offspring. Nonparametric model. Estimates of the immigrant intensity (left), offspring density (middle), and aggregate offspring density (right) functions.

the model flexibility.

We specified the model with prior choices: $\text{Lo}(2, 500)$ for θ_I with $L_I = 50$, $\text{Exp}(0.1)$ for c_0, $\text{Exp}(0.0271)$ for b_{G_0}, $\text{Lo}(2, 0.7)$ for θ_O with $L_O = 50$, $\text{Exp}(0.0673)$ for b_{F_0}, and fixed hyperparameters $(2, 4)$ for (α_0, η).

The estimated functions shown in Figure 3.6, reproduce the global pattern of two bimodalities with interval estimates covering the underlying functions. But, there are some discrepancies in immigrant intensity estimation, in particular around $t = 5,000$, as well as the relatively large TVD^a in Table 3.2, compared to other examples. As illustrated in the last panel of Figure 3.6, it is not surprising that the estimated density function in the middle panel has larger scales than the true values at the modes. In the middle panel, one can also observe a small rightward shift of the first mode. In the right panel, we are also able to see a slight shift to the right from the data at the mode, indicating that the shift is not connected with the data. The large TVD^c in Table 3.2 supports the claim, as it implies substantial discrepancies between the estimated

function and the observed data. Possibly, the large misclassification $M_O = 55$ in the table contributed to the movement.

3.3.3 Synthetic data examples for uniform-mixture-based models

To illustrate the other type of semiparametric model given in Section 3.1.2.2, we present two examples where the offspring density functions are decreasing. The intensity functions are defined as

- $\lambda^*(t) = 0.02 + 0.8 \sum_{t_i<t} 2/(1 + t - t_i)^3$ for $t \in (0, 5000)$; and
- $\lambda^*(t) = 0.01 + 0.8 \sum_{t_i<t} \text{We}(t - t_i|0.5, 2)$ for $t \in (0, 10000)$.

The power-law and the Weibull (with shape < 1) densities in the excitation functions have heavy-tailed distributions with polynomial tails. The power-law offspring density has been widely used for HP modeling in many applications, such as the ETAS model for earthquake occurrences in Ogata (1988). We generated 534 points (105 for immigrants/429 for offspring) from each HP intensity function.

Following the prior specification in Section 3.1.2.2, we assigned, in the first example, priors Exp(18.7) to μ, Ga$(2, 4)$ to γ, Exp(0.0667) to β, Ga$(5, 0.25)$ to α, and $L = 200$ for the DP-based model; and the same priors to the common parameters, (μ,γ,β), and Be$(3, 3)$ to ζ with $L = 50$ for the GW-based model. The second example took priors: Exp(37.5) for μ, Ga$(2, 4)$ for γ, Exp(0.0143) for β, Ga$(5, 0.25)$ for α, and $L = 200$ for the DP-based model; and, with the same priors for the common parameters, Be$(3, 3)$ for ζ and $L = 50$ for the GW-based model.

Example	Model	μ	γ	Cluster size		
				n_I	n_O	M_I
Power-law	DP	0.022(0.002)	0.797(0.040)	107(4)	427(4)	12(3)
	GW	0.022(0.002)	0.797(0.039)	107(4)	427(4)	11(3)
Weibull	DP	0.010(0.001)	0.805(0.040)	103(5)	431(5)	18(4)
	GW	0.010(0.001)	0.806(0.041)	103(6)	431(6)	18(4)
		Misclassification		TVD[b]	TVD[c]	
		M_O	R			
Power-law	DP	9(2)	0.039(0.007)	0.062(0.023)	0.065(0.020)	
	GW	9(1)	0.039(0.006)	0.059(0.021)	0.057(0.017)	
Weibull	DP	20(2)	0.072(0.009)	0.070(0.018)	0.071(0.018)	
	GW	20(3)	0.072(0.008)	0.074(0.016)	0.076(0.018)	

Table 3.3: Decreasing offspring density examples. Semiparametric models based on the Dirichlet process (DP) or geometric weights (GW) prior. The posterior means and standard deviations of (μ,γ) and quantitative measures.

In Figure 3.7, the left two columns illustrate the flexibility of semiparametric models in covering the underlying density functions with their interval estimates. According to the estimated aggregate offspring density functions in the right two columns, the shapes of the posterior mean density functions are justified. For instance, the L-shaped decreasing pattern around $x = 1$ in the second example can be explained by the relatively small sample size at the point, relative to the underlying Weibull density.

Both semiparametric models produce similar inferences. In addition to the comparable results in offspring density estimation, the quantitative measures in Table 3.3 provide analogous values as well. For instance, the posterior estimates of (μ, γ, n_I, n_O) from each model are almost identical to each other, which corresponds to the true values: $(0.02, 0.8, 105, 429)$ for the power-law example and $(0.01, 0.8, 105, 429)$ for the Weibull example. Accordingly, we will use the DP-based model as the semipara-

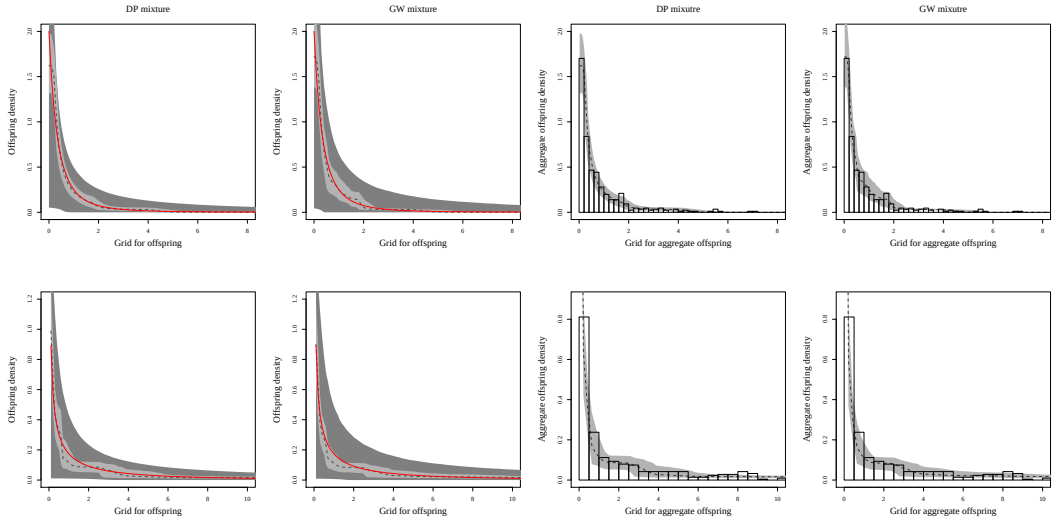

Figure 3.7: Decreasing offspring density examples with: power-law density (top) and Weibull density (bottom) functions. Semiparametric models based on the Dirichlet process (DP) or geometric weights (GW) prior. The first two columns present posterior point (dashed) and posterior 95% interval (light-gray) estimates for the offspring density function with prior 95% uncertainty bands (dark-gray) and underlying functions (red solid). The last two columns display posterior point (dashed) and interval (light-gray) estimates for the aggregate offspring density function with the histogram of all distances between offspring points and their parents.

metric model for the following real data analysis and simply extend it to the nonparametric model by replacing the constant immigrant with the immigrant Erlang mixture.

All the inferences are based on $20,000$ posterior samples, obtained by discarding $20,000$ burn-in samples.

3.4 Real data analysis

Ogata (1988) provided a catalog of earthquakes with a magnitude of six or greater that occurred in Japan and its vicinity from 1885 through 1980 (over approximately $34,711$ days). We analyzed 458 point observations (258 main shocks and 200

aftershocks), which resulted from removing 25 foreshocks from a total of 483 observations.

In order to evaluate the proposed models, we compared our semiparametric (Semipara) and nonparametric (Nonpara) models with the ETAS model (Para-P). We have added the classical parametric HP model (Para-E), which has the exponential density for the offspring density function, to the comparison. The semiparametric model was derived by substituting the uniform DP mixture for the power-law formed offspring density in the ETAS model. The nonparametric model replaced the constant immigrant intensity of the semiparametric model with the immigrant Erlang mixture. The nested relationships among the models allow us to examine the effect of immigrant intensity and offspring density substitution. Since the seismic statistical literature suggests decreasing functions in time for the excitation function (e.g., Ogata, 1988; Kagan, 1991; Musmeci and Vere-Jones, 1992; Ogata, 1998), we modeled the offspring density function using the uniform DP mixture, which focuses on non-increasing densities.

We chose the following set of priors for models' parameters: Exp(1) for the rate parameter of the exponential density function for Para-E; and Exp(1) for the two parameters p and c of the power-law density function for Para-P; Exp(0.15) for β and Ga$(5, 0.25)$ for α with $L = 200$ for Semipara; Lo$(2, 1800)$ for θ_I, $L_I = 80$, Exp(0.1) for c_0, and Exp(0.0065) for b_{G_0}, along with the priors for the uniform DP mixture in the semiparametric model for Nonpara. We assigned Exp(153) and Ga$(2, 4)$ priors to the common parameters μ and γ.

We can categorize the quantitative results in Table 3.4 into the ETAS model

	γ	Cluster size		Misclassification		
		n_I	n_O	M_I	M_O	R
Para-P	0.9(0.359)	223(21)	235(21)	48(7)	83(16)	0.288(0.026)
Para-E	0.29(0.029)	326(6)	132(6)	83(5)	15(3)	0.213(0.009)
Semipara	0.299(0.028)	322(5)	136(5)	81(4)	17(3)	0.215(0.009)
Nonpara	0.295(0.028)	324(6)	134(6)	82(4)	16(3)	0.216(0.009)

Table 3.4: The posterior means and standard deviations of the branching ratio, cluster sizes, and misclassification under each model.

(Para-P) and the other models. Para-P has a large branching ratio estimate compared to the alternatives, resulting in more points being classified as descendants. It is also noteworthy that the estimate of offspring misclassification under Para-P is greater than that of the other group; more importantly, the misclassification rate (R) of the Para-P group is relatively high. In spite of the fact that the estimated cluster sizes from Para-P are better (closer to the observed values of 258/200), its large R makes one doubt the results of classification. Additionally, Para-P has larger posterior variances for all quantitative measures in the table.

The first row of Figure 3.8 shows Q-Q plots for the time-rescaling theorem, a graphical measure of model validity (e.g., Daley and Vere-Jones, 2003). Para-E, for example, performs worse than the other models in the criterion by missing the standard line around 0.5. In this respect, the nonparametric model would be preferred because of its wide posterior uncertainty bands, which help contain the standard line within the interval estimates.

On the second row of the figure are the estimated immigrant intensities from each model and an estimated function based on the data (purple solid line). The pos-

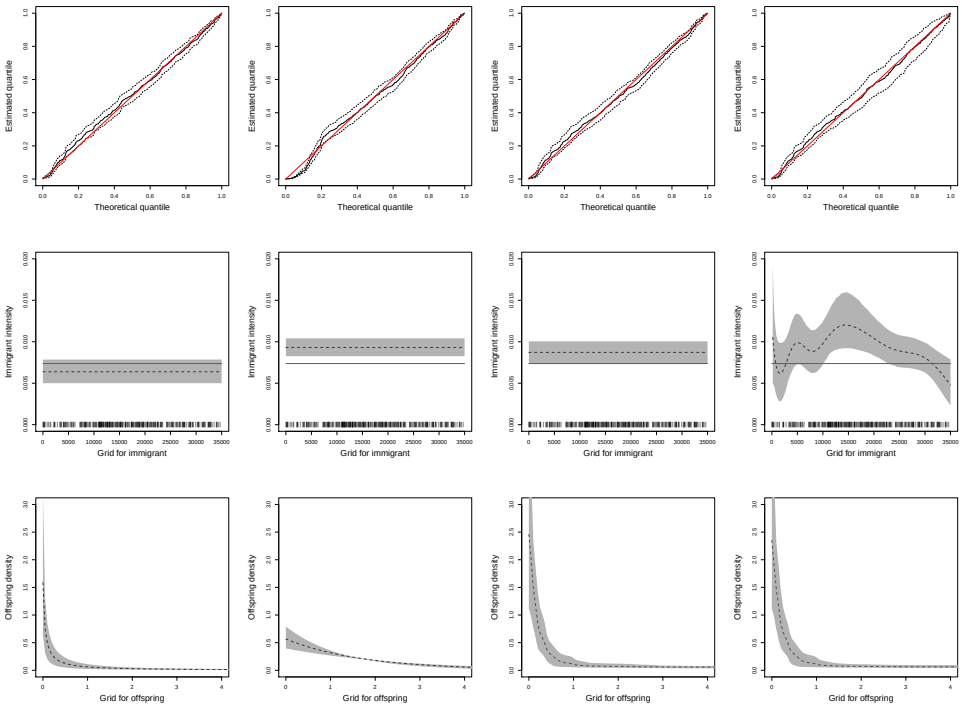

Figure 3.8: Earthquake example. Parametric models with the power-law density function (first column) and the exponential density function (second column), the semiparametric model (third column), and the nonparametric model (fourth column). Q-Q plots in the first row display results from applying the time-rescaling theorem. The next two rows demonstrate estimated functions for the immigrant intensity and the offspring density. The solid line (purple) in the second row indicates a data-based estimate $\tilde{\mu}$ of μ, such that $\int_0^T \tilde{\mu} dt = 258$, the number of immigrant points. Bars at the bottom of the panel indicate a point pattern of main shocks.

terior interval estimates from Para-E do not cover the purple line, which indicates that Para-E has a larger immigrant intensity estimated function than the data set suggests. Unlike the other models, the nonparametric model, due to the immigrant Erlang mixture, yields estimates of the immigrant intensity that are nonstandard in shape. We should note that the shape of the immigrant intensity estimates is associated with the shape of estimated first-order intensities, as seen in Figure 3.9. Further discussion will

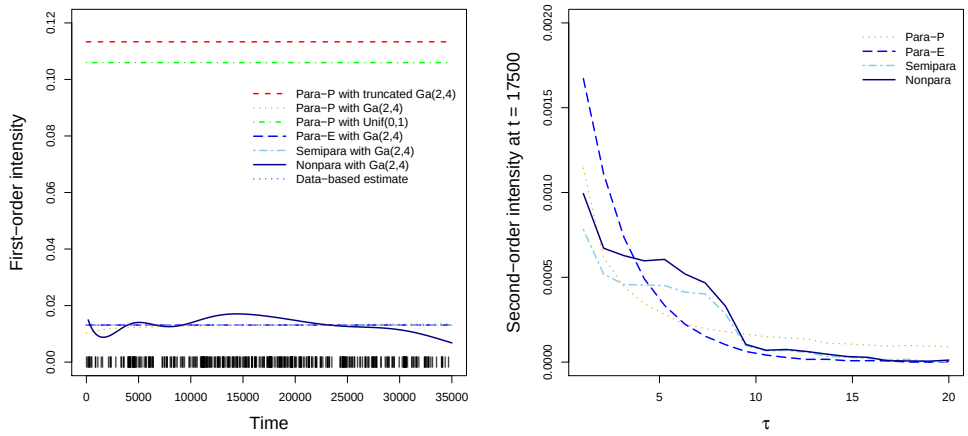

Figure 3.9: Earthquake example. The posterior means for the first-order (left) and the second-order (right) intensities. The left panel includes all point observations (bar) and a data-based estimate (purple-dotted), defined as $\tilde{\lambda}(u) = c$ with a constant c such that $\int_0^T \tilde{\lambda}(u)du = n = 458$.

follow.

In the last row of the figure, the first two panels demonstrate how drastically different the estimated offspring density functions can be. The difference in offspring density modeling between the two parametric competitors causes huge discrepancies in all inferences: the quantitative measures, the Q-Q plot for model validity, the immigrant intensity function, and the offspring density function. The results highlight a distortion that may result from the parametric model misspecification, supporting a nonparametric model for the offspring density function. As indicated by the last two panels of the row, the change in immigrant intensity modeling has a negligible effect on offspring density estimation. Furthermore, the change does not produce notable differences in inference about the quantitative measures. The immigrant Erlang mixture of the nonparametric model, however, impacts immigrant intensity estimates and associated functions, such

as the first- and second-order intensities.

The left panel of Figure 3.9 demonstrates the posterior means for the first-order intensity under each model. It indicates an association between the shapes of the estimated immigrant intensity function and the estimated first-order intensity. The point pattern at the bottom of the panel can help explain the non-standard shape of the estimated first-order intensity under the nonparametric model. For example, the points congregated around $t = 0$ and $t = 5000$ exhibit large intensities near those positions, which correspond to the first two modes.

Unlike the models, our common prior choice $\text{Ga}(2, 4)$ for γ in the ETAS model leads to some posterior samples of $\gamma > 1$, despite the prior placing almost all probabilities on $(0, 1)$. Under the model, a branching ratio greater than 1 results in a negative first-order intensity, defined as $\mu/(1 - \gamma)$. Attempts were made to avoid the negative intensity sampling by using priors supported by $(0, 1)$. However, the posterior samples of $\gamma \approx 1$, induced by the priors, brought about outliers and highly right-skewed posterior distributions. Consequently, the outliers contributed to relatively large posterior means, as shown in Figure 3.9.

The right panel of Figure 3.9 presents estimates of the second-order intensity at $t = 17,000$ for τ ranging from 0 to 20. To empirically examine the estimates from each model, we extracted the data points lying in (17400,17600). For the collected 12 points, we computed the distance $d_{ij} = t_i - t_j$, $i = 2, \ldots, 12$, $j = 1, \ldots, 11$, $i > j$. It was followed by counting $D(k) = |\{d_{ij} : k - 0.5 < d_{ij} < k + 0.5\}|$, $k = 0, 1, \ldots, 20$. As a result, $D(0) = 16$, $D(1) = 20$, $D(5) = 8$ were found to be the three largest values. The

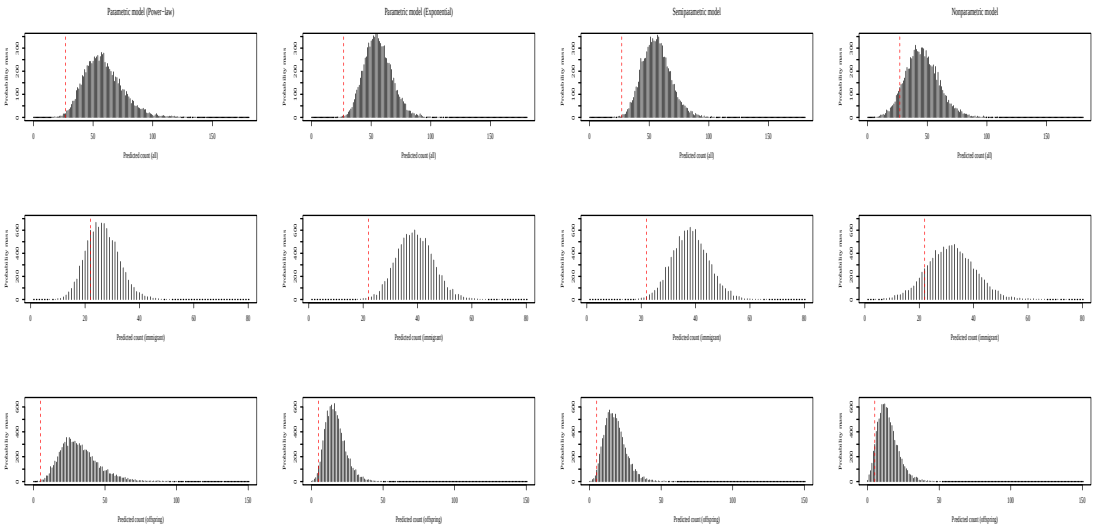

Figure 3.10: Earthquake example. Posterior predictive distributions, under each model, for the total count (first row) and associated immigrant (second row) and offspring (third row) counts of earthquakes that occurred in (1970,1980). The red dashed lines in each panel indicate the observed counts (27/22/5).

less sharp decreasing rate around $\tau = 5$ under the semiparametric and nonparametric models can be explained by the large $D(5)$.

Figure 3.10 shows the results of the predicted count, introduced in Section 3.2.2.2. We split the observation window into two sub-intervals: $(1885, 1969)$ for the training set and $(1970, 1980)$ for the test set. A comparison was made between the predictive counts and the observed numbers, $n_{\text{test}} = 27$ (22 main shocks and 5 aftershocks), of earthquakes that occurred during the last 11 years (about 4000 days). The nonparametric model (last column of the figure) has superior predictive performance, providing posterior predictive means that are closer to the observed counts. Considering the prediction of immigrants, the ETAS model achieves a competitive result, but its severe overprediction of the offspring count (aftershock) renders the model no longer competi-

tive. The nonparametric model has the best performance in predicting offspring counts, as the mean is closer to the actual count and the variance is smaller, as well as the second best prediction result for immigrant counts. Lastly, even though the nonparametric model performs better than the alternatives, it still overpredicts the test size. It may be the result of the relatively small test set size averaged over time, $n_{\text{test}}/11 = 2.45$ earthquakes/year, compared to the averaged training set size, $n_{\text{training}}/85 = 5.07$, where n_{training} is the number of earthquakes that occurred from 1885 through 1969.

3.5 Discussion

We have proposed Bayesian nonparametric modeling approaches for the HP conditional intensity function. Using the cluster representation, the HP can be modeled through independent NHPPs for immigrants and offspring. Therefore, we can utilize the Erlang mixture introduced for NHPPs for the immigrant intensity and the excitation function (with modifications). In this study, we have developed semiparametric models for either the immigrant intensity or the excitation function, which employ the Erlang mixture modeling framework. Additionally, we have proposed a semiparametric modeling approach for decreasing offspring densities based on uniform mixtures with the DP/GW prior. Finally, we have developed nonparametric modelling methods by combining the immigrant intensity model with one of the offspring models.

We have presented two types of mixture models pertaining to offspring functions, as well as an Erlang mixture for the immigrant intensity. Therefore, the question

arises as to which model should be utilized when. The data can assist in determining the immigrant intensity to be used. This intensity is modeled via either a constant function or an Erlang mixture. Experimental results have revealed a connection between the shape of immigrant intensity and that of first-order intensity. The first-order intensity is a non-conditional (averaged over the history) intensity function for the complete data set. As such, the data histogram provides insight into the shape of first-order intensity and, by extension, the shape of immigrant intensity. The excitation function is chosen either based on expert advice or relevant theories in the field to which the model is applied. A uniform-mixture-based model can, for example, be a flexible option, provided the offspring density is expected to decrease over time. The Erlang mixture excitation function, on the other hand, provides a less structured choice because it can depict functions of more general shapes.

Bayesian nonparametric models for HPs have rarely been explored (some of the methods were briefly discussed in the Introduction). Donnet et al. (2020) studied nonparametric models for multivariate HPs with an emphasis on theoretical properties, such as posterior concentration rates, but without practical discussion on the models. Inference methods based on Gaussian process (GP) priors have been investigated in the machine learning literature (e.g., Zhang et al., 2018; Zhou et al., 2019, 2020). Zhang et al. (2018) transformed the excitation function such that $h(x) = g(f(x))$, where $g(y) = y^2$ and $f(\cdot)$ is assigned a GP prior. They utilized a specific covariance function in the form of the Mercer expansion, which facilitated the establishment of closed-form HP likelihood along with the squared link function, $g(y)$. Zhou et al. (2019) further developed the GP

model by applying the GP structure to the immigrant intensity for a more flexible model. Furthermore, they used a variational inference method to achieve efficient inference. With the squared link function, the variational method allows one to bypass intractable integrals of exponential terms in the HP likelihood. Note that, given the branching structure (which is used by all GP-based models), the HP likelihood is decomposed into NHPP intensities, which contain intractable integrals in the likelihood normalizing term. Zhou et al. (2020) employed the sigmoid link function rather than the squared link for the immigrant intensity and excitation functions, that is, $h(x) = \lambda^* g(f(x))$, where $g(y) = (1 + e^{-y})^{-1}$, and $\lambda^* > 0$, is an upper bound on $h(x)$. The use of a Gaussian mixture representation of the link function, involving the Pólya-Gamma distribution, (Polson et al., 2013) allows efficient handling of the intractable normalizing term in conjunction with the Campbell's theorem.

The GP-based approaches have been developed in such a way as to improve posterior inferences. However, the computational complexity associated with inference remains the greatest obstacle to their application. For example, the most recent model in Zhou et al. (2020) has a time complexity greater than $\mathcal{O}(N_I^3 + N_O^3)$ for sampling multivariate normal function values, driven by GPs, in its Gibbs sampling approach, where N_I or N_O are the number of points classified as immigrants or offspring based upon the branching structure. To reduce the time complexity, they applied a sparse GP approximation by inducing points. But, such an approximation in Zhou et al. (2019) and Zhou et al. (2020) has a limitation that the inference results vary depending on the method of locating the inducing points as well as the number of inducing points. In

contrast, our modeling approach offers significant computational advantages. We have a parameter for the number of mixture components that may increase time complexity, but it has no bearing on the size of the point pattern. Moreover, each of the mixture components in our model has just a single common parameter, and the mixture weights have ready prior-to-posterior updating by independent gamma conjugate priors. Turning back to the GP-based methods, a lack of prior sensitivity analysis in the literature raises questions regarding the prior for model parameters and the choice of covariance function. More importantly, such modeling approaches do not provide information regarding the branching ratio, which is crucial to checking the HP stability condition, whereas our model allows us to control the prior for this key parameter. In addition to the conditional intensity function estimation, one may be interested in estimating the first- or second-order intensity, which also characterizes HPs. Our model framework enables tractable inference for these intensities as well.

Chapter 4

Marked Hawkes processes for earthquake occurrences: A Bayesian semiparametric modeling approach

4.1 Introduction

The marked Hawkes process (MHP) is an extension of the HP obtained by adding marks, observable variables associated with each time point (Hawkes, 1971a,b). The marks provide additional features beyond time to the point pattern. For instance, the magnitude of an earthquake is an important mark for the MHP in seismology applications. Parametric models, e.g., the Epidemic Type Aftershock-Sequences (ETAS) model, have been widely used to represent and estimate the MHP intensity function for forecasting earthquakes (e.g., Ogata, 1988, 1998). As another example, the type of crime serves as a mark for the MHP to detect the dynamics of crime in criminology (Mohler,

2014). Each type of crime is assigned a parameter, which gives a constant multiplicative effect on the intensity function. Meyer et al. (2012) introduced a spatio-temporal MHP model for predicting the incidence of an infectious disease, called invasive meningococcal disease. They defined the mark as finetype, a unique combination of serogroup; each finetype had a different level of an exponential multiplicative effect on the intensity function.

This chapter develops models for temporal MHPs with a single continuous mark. The focus is on applications to estimation of the intensity of earthquake occurrences, with earthquake magnitude providing the mark.

Regarding inference about the MHP intensity function in seismology applications, Ogata (1988) introduced a temporal MHP model referred to as the (ordinary) ETAS model, which was later extended to spatio-temporal MHP models (e.g., Kagan, 1991; Musmeci and Vere-Jones, 1992; Ogata, 1998). The temporal model and the spatio-temporal extension have been further extended by making some model parameters to vary in time (Kumazawa and Ogata, 2014) or space (Ogata et al., 2003; Nandan et al., 2017). When it comes to Bayesian approaches, Rasmussen (2013) illustrated Bayesian inference for the ETAS model with methods based on: the MHP intensity function; the cluster representation and NHPP intensity functions. Ebrahimian et al. (2014) and Ebrahimian and Jalayer (2017) proposed seismicity forecasting approaches based on Bayesian inference under the ETAS model. Ross (2021) applied the cluster representation-based method for a more computationally efficient algorithm.

Here, we propose a flexible and computationally efficient model for temporal

MHP intensity functions, motivated by earthquake modeling applications. Our model represents a key part of the intensity function, the ground process excitation function, as a weighted combination of basis functions. Each basis has a multiplicative form of an Erlang density for time and a polynomial function for the mark. The non-negative mixture weights are defined through increments of a random measure H, to which we assign a gamma process prior. Mixture weights driven by the prior specification on H result in flexible shapes for the excitation function, and in tractable posterior sampling for the primary model parameters (i.e, the mixture weights). To the best of our knowledge, the proposed model is the first nonparametric method for marked HPs applied to earthquake data. Moreover, unlike all existing ETAS models, our model does not assume factorization of the excitation function into two separate functions for time and the mark. As an important practical consequence, we can estimate magnitude-dependent aftershock densities.

The outline of the chapter is as follows. Section 4.2 provides background for the MHP, including the definition of the MHP intensity function, the likelihood, and stability conditions MHPs. Section 4.3 presents the modeling and inference methodology for MHP intensities. The modeling approach is illustrated with synthetic and real data in Sections 4.4 and 4.5. Finally, Section 4.6 concludes with a summary and a general discussion.

4.2 Background

Denote by κ the mark for the process, assumed to be a continuous and positive-valued variable (the earthquake magnitude in our motivating application). Then, the conditional intensity function $\lambda^*(t, \kappa)$ of the MHP is defined as $\lambda^*(t, \kappa) = \lambda_g^*(t) f^*(\kappa|t)$, where $\lambda_g^*(t)$ denotes the intensity function of the ground process. The difference is in the history $\mathcal{H}(t)$ of the intensity function: in the ground process intensity, the history consists of previous times t_i and marks κ_i such that $\mathcal{H}(t) = \{(t_i, \kappa_i) : t_i < t\}$ not just a sequence of times as in the HP intensity function. More specifically, the ground process intensity function takes the form

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

with $\mu(t) > 0$ for $t > 0$ and $h(x_i, \kappa_i) > 0$ for $x_i = t - t_i > 0$.

The other factor of the MHP intensity is the mark density function $f^*(\kappa|t)$; generally, the density function is conditioned on both time and the history. But, the mark density function independent of the history, such that $f^*(\kappa|t) = f(\kappa|t)$, has been widely used. Such a mark is named the *unpredictable mark*, and, as a representative example, the ETAS model has a special-type unpredictable mark whose density function is defined as $f^*(\kappa|t) = f(\kappa)$. Separately, if a process has the mark density independent of the history, but also the ground process is independent of the mark history (still, dependent on the time history), the mark is called the *independent mark* (Daley and Vere-Jones, 2003, Proposition 7.3.V). We will focus on the special-type unpredictable mark as in the ETAS model.

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 and Vere-Jones, 2003, Proposition 7.3.III)

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

Since the ground process intensity is in the form of the HP intensity function, the cluster representation is applicable to the intensity. With the branching structure $\{y_i : i = 1, \ldots, n\}$, the MHP likelihood can be represented as

$$\begin{aligned} &\Big[\prod_{i=1}^{n} f^*(\kappa_i|t_i)\Big]\Big[\exp\Big\{ - \int_0^T \mu(u) du\Big\} \prod_{\{t_i \in I\}} \mu(t_i)\Big] \\ &\times \Big[\exp\Big\{ - \sum_{i=1}^{n} \int_0^{T-t_i} h(u, \kappa_i) du\Big\} \prod_{\{i: t_i \in O\}} h(t - t_{y_i}, \kappa_{y_i})\Big], \end{aligned} \tag{4.1}$$

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 $O = \{t_i : y_i \neq 0, i = 1, \ldots, n\}$.

As in the HP intensity function modeling in Chapter 3, the excitation function $h(x_i, \kappa_i)$ of the ground process requires conditions to prevent the MHP from exploding. Under the assumption of $f^*(\kappa|t) = f(\kappa)$, the conditions are (Daley and Vere-Jones, 2003, Proposition 6.4.VII)

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

(ii) $\rho = \mathrm{E}(\alpha(\kappa)) = \int_{\mathcal{K}} \alpha(\kappa) f(\kappa) d\kappa < \infty$ is a necessary condition for MHP. Restricting the value of ρ to the unit interval ensures MHP stability.

The first condition enables the excitation function to be factorized into the total offspring intensity function $\alpha(\kappa)$, and the mark-dependent offspring density function $g_\kappa(x) =$

$h(x,\kappa)/\alpha(\kappa)$. The condition is required for defining MHPs. $\alpha(\kappa)$ determines the expected number of points of the offspring Poisson process with intensity $h(x,\kappa)$ for any mark $\kappa \in \mathcal{K}$. The quantity, $\rho = \mathrm{E}(\alpha(\kappa))$ of the second condition provides the expected offspring count that is averaged over the mark. Therefore, ρ plays the same role as the branching ratio for unmarked HPs; using ρ, we can define the expected size of a cluster arising from an immigrant as an infinite sum $\sum_{r=0}^{\infty} \rho^r$. The stability condition $\rho \in (0,1)$ guarantees the geometric series is finite.

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

$$\begin{aligned} \lambda_g^*(t) &= \mu + \sum_{t_i<t} a_\alpha \exp\{\eta(\kappa_i - \kappa_0)\}\mathrm{Powlaw}(t-t_i|p,c) \\ f^*(\kappa|t) &= f(\kappa) = \psi \exp\{-\psi\kappa\}, \;\; \kappa \in [\kappa_0, \infty) = \mathcal{K}, \end{aligned} \tag{4.2}$$

with positive parameters a_α, η, and ψ to be estimated and fixed $\kappa_0 > 0$.

It makes sense for the occurrence of aftershocks to be affected by the magnitude of the main shock (or of other precedent aftershocks). For example, the main shock with a larger magnitude may have more aftershocks occur sooner than later. It may be possible for subsequent shocks after such a main earthquake to take place in a long time, over a week or even a month. Therefore, we plan to develop a model for mark-dependent offspring densities, which also allows flexible increasing shape for $\alpha(\kappa)$. The exponential function of $\alpha(\kappa)$ in the ETAS model may be too restrictive, and indeed, different formed $\alpha(\kappa)$ has been introduced for seismic MHP modeling (Ogata and Zhuang, 2006). In that

sense, our model benefits by providing a more flexible framework for $\alpha(\kappa)$.

4.3 Methodology

4.3.1 Model formulation

In this section, we propose a Bayesian nonparametric model for $h(x, \kappa)$ of the MHP for earthquake applications. Unlike other disciplines, $\alpha(\kappa)$ is supposed to increase over κ, earthquake magnitude. In view of the fact that an earthquake of greater magnitude generates more aftershocks, the increasing total offspring intensity function makes sense. Notice that $\alpha(\kappa)$ determines the number of offspring for $\kappa \in \mathcal{K}$. We propose a model that takes into consideration the trait of $\alpha(\kappa)$ for aftershock occurrences.

The chapter focuses on modeling the excitation function with a constant immigrant intensity function for simplicity. By substituting the immigrant Erlang mixture in Section 3.1.1 for the constant intensity, one obtains a fully nonparametric model for the ground process intensity function, with additional flexibility for the immigrant intensity function.

We consider a parametric mark density function for the special type of unpredictable marks that are independent of both the current time and the MHP history. Such a parametric mark density provides a ready expression for ρ, the key quantity related to the HP stability condition as described in Section 4.2. This expression illustrates clearly how the integrability modeling condition is satisfied, which will be shown in Section 4.3.2. Moreover, it allows for efficient handling of the MHP stability condition

by tuning hyperparameters. Section 4.3.3 provides a strategy for the prior specification based on the expression and the condition.

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. $\text{Ga}(x|a,b)$ is also called the Erlang density with integer shape a. Our model for $h(x,\kappa)$ is a weighted combination of basis functions, each of which is constructed by a product of two functions: an Erlang density $\text{Ga}(x|l,\theta^{-1})$ and a polynomial function $b_m(\kappa;d)$. The mixture weights are defined through increments of a measure H on $\mathbb{R}^+ \times \mathcal{K}$, assigned a gamma process prior. That is,

$$\begin{aligned} h(x,\kappa) &= \sum_{l=1}^{L}\sum_{m=1}^{M} \nu_{lm}\text{Ga}(x|l,\theta^{-1})b_m(\kappa;d), \quad x \in \mathbb{R}^+ \text{ and } \kappa \in (\kappa_0,\kappa_{max}) = \mathcal{K} \\ \nu_{lm} &= H(A_{lm}), \quad H \sim \mathcal{G}(H_0,c_0), \end{aligned} \tag{4.3}$$

where $A_{lm} = [(l-1)\theta, l\theta) \times [(m-1)/M, m/M)$, $l = 1,\ldots,L$ and $m = 1,\ldots,M$. 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, $b_m(\kappa;d)$ is defined as

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

with a real-valued scalar d. Under the modeling framework, the total offspring intensity is derived as $\alpha(\kappa) = \sum_{m=1}^{M} V_m b_m(\kappa;d)$, where $V_m = \sum_{l=1}^{L}\nu_{lm}$. Since the total offspring intensity function is a mixture of increasing functions $b_m(\kappa;d)$, $m = 1,\ldots,M$, it satisfies the increasing characteristic of $\alpha(\kappa)$ regarding aftershock occurrences. The polynomial function involves coefficients m^d and M, and we will next discuss their role in the modeling framework.

We desired more flexibility on $\alpha(\kappa)$ for larger M, so introduced a multiplicative

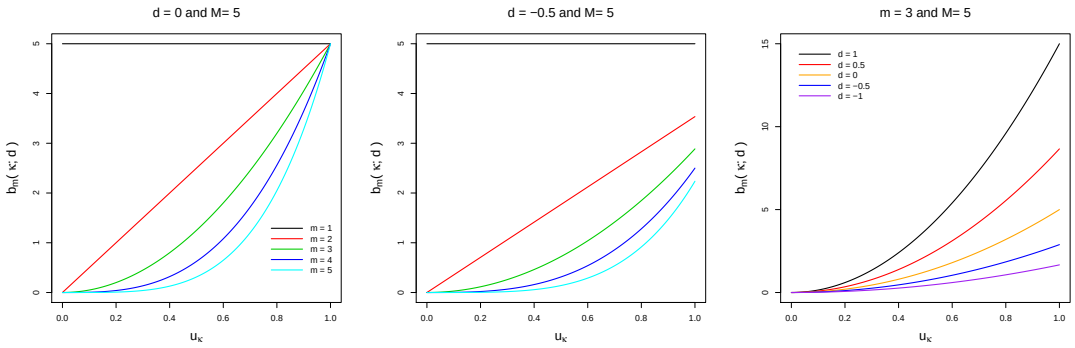

Figure 4.1: Polynomial functions $b_m(\kappa; d)$: for different values of m under fixed $d = 0$ (left) and -0.5 (middle); for different values of d under fixed $m = 3$ (right). M is constant at 5 in all three panels.

effect M into the function $b_m(\kappa; d)$. The effect ensures that the flexibility of $\alpha(\kappa)$ is enhanced with increasing M (shown in Section 4.3.2). The use of m^d is associated with the increasing rate of $b_m(\kappa; d)$. The first two panels of Figure 4.1 show the effect of m^d on $b_m(\kappa; d)$ by comparison, where the functions in the middle panel are built by adding the m^d effect to the functions in the left panel. Contrary to the second panel, a positive value of d under $M = 5$ will result in $b_m(\kappa; d)$ that has the relatively large increasing rate, as compared to the first panel. The third panel of the figure describes the m^d effect on $b_m(\kappa; d)$ for different values of d under fixed $m = 3$ and $M = 5$. We will make d random to have the best set of $b_m(\kappa; d)$, $m = 1, \ldots, M$ for fixed M, in terms of the increasing rate.

The gamma process, $\mathcal{G}(H_0, c_0)$, in (4.3) has a centering measure H_0 on $\mathbb{R}^+ \times (0, 1)$ and a precision parameter $c_0 > 0$. Placing the 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 A_{lm} of a partition 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-form centering measure, $H_0(A_{lm}) = H_{0x}(A_l) \times H_{0\kappa}(A_m)$, such that

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

The Weibull cumulative hazard with shape b_1 (and scale 1) and the exponential cumulative hazard with rate b_2 underlie the two separate measures, $H_{0x}(A_l)$ and $H_{0\kappa}(A_m)$. Unlike the Erlang mixture modeling for the NHPP or HP intensities, the MHP modeling approach does not hold any convergence of the mixture $h(x, \kappa)$ (cf. Lemma in Section 2.1.1). Yet, the centering measure is linked to the prior mean of $g_\kappa(x)$, for example the shape of the hazard function of $H_{0x}(A_l)$ determines the shape of the prior mean for $g_\kappa(x)$. In earthquake applications, because $g_\kappa(x)$ is generally expected to decrease over x (the distance between an aftershock and its parent), we apply the Weibull hazard function, which decreases over x under a shape parameter $b_1 < 1$. Unlike $H_{0x}(A_l)$, the exponential cumulative hazard function for $H_{0\kappa}(A_m)$ was chosen for the sake of brevity (providing a simple-form $H_{0\kappa}(A_m) = b_2/M$ with a single parameter b_2). Following is a discussion of the effect of model parameters $(b_1, b_2, c_0, L, \theta, M, d)$ on $g_\kappa(x)$ and $\alpha(\kappa)$.

Figure 4.2 illustrates the results of $1,000$ realizations of the offspring density function for different values of the model parameters. In comparison to the first panel, which represents the baseline model, we can observe the impact of each parameter on the prior mean or the prior model uncertainty. For example, the first two panels of the top row indicate that the shape parameter, b_1, of $H_{0x}(A_l)$ plays a key role in the

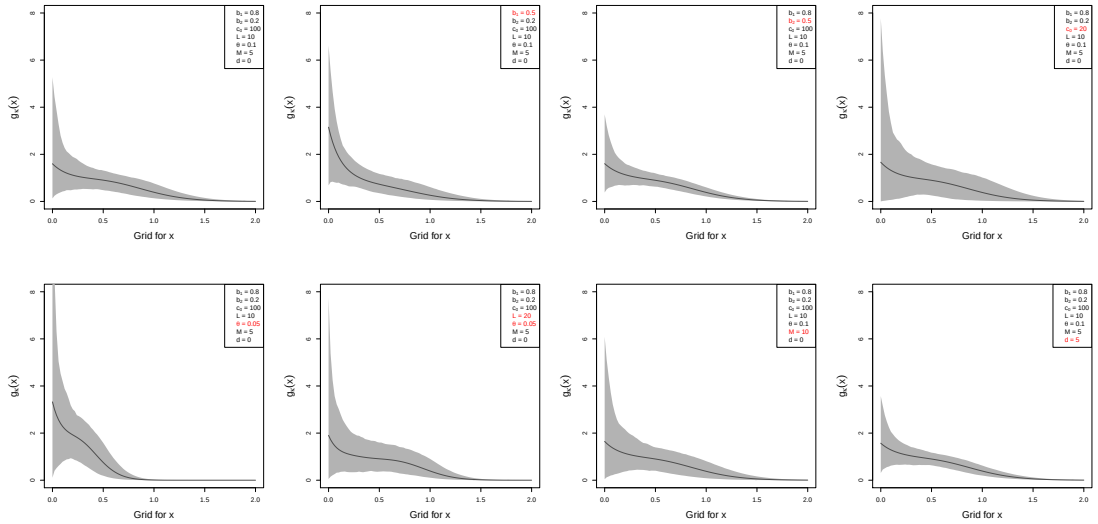

Figure 4.2: Prior mean (solid line) and prior 95% uncertainty bands (shaded area) for the offspring density function, $g_\kappa(x)$, $\kappa = 5.5$, under different choices of model parameters.

shape of $g_\kappa(x)$. The first two panels at the bottom confirm that L and θ still provide a rough guess about effective support for the offspring density function, as in the NHPP modeling.

As a result, our model provides the prior mean $\mathrm{E}^\nu(\alpha(\kappa))$ and the prior variance $\mathrm{Var}^\nu(\alpha(\kappa))$ for $\alpha(\kappa)$ over the weights ν_{lm} in closed-form so that we can detect the effect of model parameters on $\alpha(\kappa)$ more explicitly. Under the gamma process prior for H, the mean and variance are derived as

$$\begin{aligned} \mathrm{E}^\nu(\alpha(\kappa)) &= \sum_{l=1}^{L}\sum_{m=1}^{M} \mathrm{E}(\nu_{lm}) b_m(\kappa; d) = (L\theta)^{b_1} b_2 \sum_{m=1}^{M} m^d u_\kappa^{m-1} \\ \mathrm{Var}^\nu(\alpha(\kappa)) &= \sum_{l=1}^{L}\sum_{m=1}^{M} \mathrm{Var}(\nu_{lm}) (b_m(\kappa; d))^2 = \frac{(L\theta)^{b_1} b_2 M}{c_0} \sum_{m=1}^{M} (m^d u_\kappa^{m-1})^2. \end{aligned} \tag{4.4}$$

The precision parameter, c_0, of the gamma process is a key parameter for the prior uncertainty of $\alpha(\kappa)$. Also, as intended, the prior uncertainty increases as M, the number

of basis functions that compose $\alpha(\kappa)$, increases.

We utilize a beta density for transformed marks such that $f(\kappa) = \text{Be}(u_\kappa|a_\beta, b_\beta)$, which facilitates an analytical form of ρ, as shown in the following section.

4.3.2 Model property

The proposed model provides a parsimonious representation of the excitation function with a common scale parameter, θ, for $\text{Ga}(x|l, \theta^{-1})$ and a common single parameter, d, for $b_m(\kappa; d)$.

The model also exhibits a good balance between model flexibility and computational efficiency in its implementation of posterior inference: for example, the gamma prior for the mixture weight, driven by the gamma process prior for H, is conjugate, which allows us to derive the posterior sample for ν_{lm} from independent gamma distributions. Our model can be implemented with an efficient Markov chain Monte Carlo (MCMC) algorithm that does not require complex computational methods or approximations to handle the normalizing constant term (provided in Section 4.3.4).

Under the model, we can obtain ready expressions for key functions. For instance, the total offspring intensity function is derived as

$$\alpha(\kappa) = \int_0^\infty h(x, \kappa)dx = \sum_{l=1}^{L}\sum_{m=1}^{M} \nu_{lm} b_m(\kappa; d) = \sum_{m=1}^{M} V_m(m^d M u_\kappa^{m-1}), \tag{4.5}$$

where $V_m = \sum_{l=1}^{L} \nu_{lm}$. It is a weighted combination of $b_m(\kappa; d)$ with the mixture weights of gamma random variables $V_m \sim \text{Ga}(c_0 \sum_{l=1}^{L} H_0(A_{lm}), c_0)$. As expected in earthquake applications, we can observe an increasing pattern over κ. Furthermore, the

representation establishes that $\alpha(\kappa)$ satisfies the first MHP stability condition in Section 4.2.

The expectation of $\alpha(\kappa)$ over κ in terms of the beta mark density is available in analytical form. That is

$$\rho = \int_{\kappa_0}^{\kappa_{max}} \alpha(\kappa) f^*(\kappa) d\kappa = \sum_{m=1}^{M} V_m m^d M \text{Beta}(a_\beta + m - 1, b_\beta)/\text{Beta}(a_\beta, b_\beta), \tag{4.6}$$

where $\text{Beta}(a, b)$ is a beta function. The finite mixture fulfills the necessary MHP condition, the finite expectation of the excitation function. As the equation contains random variables, such as V_m, the second stability condition, $\rho \in (0, 1)$, cannot be strictly enforced. Instead, our model allows us to control random ρ by tuning hyperparameters, placing most probabilities on (0,1).

Most importantly, the offspring density function can be expressed as a weighted combination of Erlang densities, as follows:

$$g_\kappa(x) = \frac{h(x, \kappa)}{\alpha(\kappa)} = \sum_{l=1}^{L} \Big(\frac{\sum_{m=1}^{M} \nu_{lm} b_m(\kappa; d)}{\sum_{m=1}^{M} V_m b_m(\kappa; d)} \Big) \text{Ga}(x|l, \theta^{-1}) = \sum_{l=1}^{L} W_l(\kappa) \text{Ga}(x|l, \theta^{-1}). \tag{4.7}$$

The mixture weight becomes a function of κ, which permits a mark-dependent offspring density function.

4.3.3 Prior specification

This section outlines a strategy for specifying two fixed parameters L and M and the prior for $(\theta, d, b_1, b_2, c_0, a_\beta, b_\beta)$ of the proposed model. We adopt the prior specification given in Section 2.1.2 for the parameters θ, L (equivalent to J of the NHPP

model), and c_0. For example, we place, on θ, a Lomax prior with shape 2 (implying infinite variance), scale d_θ, and median $d_\theta(\sqrt{2}-1)$, for which we specify d_θ such that $\Pr(0 < \theta < T_O) \approx 0.9$, where T_O is the upper bound of effective support for $g_\kappa(x)$.

By matching a rough guess $L\theta$ about the effective support to its proxy $(0, T_O)$, we set L to the integer part of T_O/θ^*, where θ^* is the prior median of θ. A selected L may serve as a lower bound for sensitivity analysis in applications that require a more conservative choice of L, i.e., where an intensity function will have non-standard shapes.

The precision parameter, c_0, of the gamma process prior takes an exponential prior with rate a_{c_0}. The hyperparameter is set to a value that gives a conservative upper bound (empirically chosen) for c_0. The following examples of Sections 4.4 and 4.5 have been performed with $a_{c_0} = 0.005$, and we observed significant prior-posterior learning toward 0.

Two independent exponential priors are assigned to (a_β,b_β), taking into consideration their brevity with only two single parameters needed to define them. To specify the hyperparameters, we compute the mean, κ_{avg}, of the observed marks and match it to the expectation of the beta mark density such that $a_\beta/(a_\beta + b_\beta) = (\kappa_{avg} - \kappa_0)/(\kappa_{max} - \kappa_0)$. We replace a_β and b_β with their respective expectations $\mathrm{E}(a_\beta)$ and $\mathrm{E}(b_\beta)$. With a simple choice $\mathrm{E}(a_\beta) = 1$, $E(b_\beta)$ can be derived as $E(b_\beta) = \mathrm{E}(a_\beta)(\kappa_{max} - \kappa_0)/(\kappa_{avg} - \kappa_0) - \mathrm{E}(a_\beta) = (\kappa_{max} - \kappa_0)/(\kappa_{avg} - \kappa_0) - 1$. With regard to the equation for $\mathrm{E}(b_\beta)$, we investigated other values of $\mathrm{E}(a_\beta)$ and observed robustness to the choice in the examples of Sections 4.4 and 4.5.

The parameter d is assigned the normal distribution $\mathrm{N}(0, 100)$ prior with mean

0 and variance 100. We set the prior mean to 0, which implies $b_m(\kappa; d)$ has no m^d effect. We observe that the variance of 100 provides effective support of $(-30, 30)$ for d, which is empirically large enough for inferences regarding d: in the examples of this chapter, we saw that the posterior distribution of d was more condensed, with means ranging from -5 to 5.

Unlike the other fixed parameter L, choosing M is challenging for the model. The strategy for selecting relevant M for the prior model is unavailable. d and M do not have any proxies to which their functions can be compared, whereas θ and L have $(0, T_O)$ for effective support, and comparing it to $L\theta$ justifies the choice of L. So, we determine the value for M by conducting sensitivity analysis, which requires multiple runs of MCMC. For example, we begin the posterior simulation with a small value of M, such as 3 or 4, iterate with increasing M values until the impact of increase in M on estimating the excitation function becomes negligible. Additionally, we may use the Q-Q plot that is derived from the time-rescaling theorem, terminating the iteration based on the graphical evaluation of the Q-Q plot.

The shape parameter b_1 of H_{0x} determines the shape of $g_\kappa(x)$, as seen Section 4.3.1. One can attain the decreasing characteristic of $g_\kappa(x)$ by setting b_1 to less than 1 for earthquake data sets. We assign the Exp(1) prior to b_1, which places large probabilities on the unit interval.

We employ an exponential prior for b_2 and specify the rate parameter using the stability condition $\rho \in (0, 1)$. We can achieve random ρ by substituting selected values of L and M, and prior means for all parameters into the equation for ρ given

in (4.6) along with gamma priors for V_m. Then, the hyperparameter for b_2 is chosen using the distribution of ρ, which should cover as much of the unit interval as possible together with $\Pr(0 < \rho < 1) \approx 1$.

4.3.4 Posterior inference

Denote by $\{0 < t_1 < t_2 < \cdots < t_n < T\}$ the point pattern observed in time window $(0, T)$ and by $\{\kappa_i \in \mathcal{K}\colon\ i = 1, \ldots, n\}$ the observed marks. Note that, unlike the ETAS model, our model defines the mark space as a bounded interval $\mathcal{K} = (\kappa_0, \kappa_{max})$, and the additional information about the upper bound κ_{max} is used to transform a mark κ_i to u_{κ_i}. Under our modeling framework with the constant immigrant intensity and the beta mark density, the MHP likelihood is derived as

$$\begin{aligned}&\Big[\prod_{i=1}^{n} \mathrm{Be}(u_{\kappa_i}|a_\beta, b_\beta)\Big] \exp\Big\{-\int_0^T \mu du\Big\} \exp\Big\{-\sum_{l=1}^{L}\sum_{m=1}^{M} \nu_{lm} K_{lm}(\theta, d)\Big\} \\ &\times \prod_{t_i \in I} \mu \prod_{t_i \in O} \Big\{\sum_{l=1}^{L}\sum_{m=1}^{M} \nu_{lm} \mathrm{Ga}(t_i - t_{y_i}|l, \theta^{-1}) b_m(\kappa_{y_i}; d)\Big\},\end{aligned} \tag{4.8}$$

where $K_{lm}(\theta, d) \equiv K(\theta, d, l, m) = \sum_{i=1}^{n} b_m(\kappa_i; d) \int_0^{T-t_i} \mathrm{Ga}(s|l, \theta^{-1}) ds$. I and O are an immigrant process and the superposition of offspring processes, respectively.

We augment the likelihood with auxiliary variables $\boldsymbol{\xi} = \{\boldsymbol{\xi_i} : i = 1, \ldots, n\}$, where $\boldsymbol{\xi_i} = (\xi_{i1}, \xi_{i2})$ identifies the basis function to which an event $(t_i, \kappa_i) \in O$ is assigned. Then, the hierarchical model, with the branching structure $\boldsymbol{y} = (y_1, \ldots, y_n)$,

for the augmented data can be represented as

$$
\begin{aligned}
(\boldsymbol{t},\boldsymbol{\kappa})|\boldsymbol{y},\mu,\boldsymbol{\xi},\theta,\boldsymbol{\nu},d,a_\beta,b_\beta \sim & \Big[\prod_{i=1}^{n}\mathrm{Be}(u_{\kappa_i}|a_\beta,b_\beta)\Big]\exp\Big\{-T\mu\Big\}\mu^{n_I} \\
& \times\prod_{l=1}^{L}\prod_{m=1}^{M}\exp\Big\{-\nu_{lm}K_{lm}(\theta,d)\Big\} \\
& \times\prod_{t_i\in O}\Big\{\big(\sum_{r_1=1}^{L}\sum_{r_2=1}^{M}\nu_{r_1r_2}\big)\mathrm{Ga}(t_i-t_{y_i}|\xi_{i1},\theta^{-1})b_{\xi_{i2}}(\kappa_{y_i};d)\Big\} \\
\boldsymbol{\xi_i}|\boldsymbol{\nu} \overset{i.i.d.}{\sim} & \sum_{l=1}^{L}\sum_{m=1}^{M}\frac{\nu_{lm}}{\Big(\sum_{r_1=1}^{L}\sum_{r_2=1}^{M}\nu_{r_1r_2}\Big)}\delta_{(l,m)}(\xi_{i1},\xi_{i2}),i=1,\ldots,n_O \\
\boldsymbol{y} \sim & \ \delta_0(y_1)\prod_{i=2}^{n}\mathrm{Unif}(y_i|0,1,\ldots,i-1) \\
\nu_{lm}|c_0,b_1,b_2,\theta \overset{ind.}{\sim} & \ \mathrm{Ga}(\nu_{lm}|c_0H_0(A_{lm}),c_0),\ \ l=1,\ldots,L,\ \ m=1,\ldots,M,
\end{aligned}
\tag{4.9}
$$

where n_I/n_O individually indicates the number of immigrant/offspring points categorized by the branching structure. $\delta_{(a,b)}(x,y)$ is the Dirac delta function such that $\delta_{(a,b)}(x,y)=1$ if $x=a$ and $y=b$. $\mathrm{Unif}(x|0,1,\ldots,i-1)$ is the discrete uniform probability mass function. We assign an exponential distribution $\mathrm{Exp}(a_\mu)$ prior to μ, and the other model parameters $(\theta,d,a_\beta,b_\beta,c_0,b_1,b_2)$ have the priors enunciated in Section 4.3.3.

Gibbs sampling method is used to undertake posterior inference, resulting in ready updates for $\boldsymbol{y}$, $\boldsymbol{\xi}$, μ, and $\boldsymbol{\nu}$ through standard-form posterior full conditional distributions.

Denoted by $\boldsymbol{D}=\{\boldsymbol{t},\boldsymbol{k}\}$ the data set, the posterior full conditional for the

branching structure $\boldsymbol{y}$ is a discrete distribution such that

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

The posterior full conditional for each $\boldsymbol{\xi_i}$ is an independent discrete distribution on $\{(l, m) : l = 1, \ldots, L, \ \ m = 1, \ldots, M\}$ such that $\Pr((\xi_{i1} = l, \xi_{i2} = m)|\theta, \boldsymbol{\nu}, d, \boldsymbol{y}, \boldsymbol{D}) \propto \nu_{lm}\mathrm{Ga}(t_i - t_{y_i}|l, \theta^{-1})b_m(\kappa_{y_i}; d)$.

With the $\mathrm{Exp}(a_\mu)$ prior, the immigrant intensity μ is given a gamma posterior full conditional distribution with shape $n_I + 1$ and rate $T + a_\mu$.

Denote by $n_{lm} = |\{t_i \in O : \xi_{i1} = l, \xi_{i2} = m\}|$ for $l = 1, \ldots, L$ and $m = 1, \ldots, M$, the number of offspring points assigned to (l, m)-th basis. The posterior full conditional distribution for $\boldsymbol{\nu}$ is derived as $p(\boldsymbol{\nu}|\boldsymbol{\xi}, \theta, c_0, b_1, b_2, \boldsymbol{\kappa}) \propto \prod_{l=1}^{L}\prod_{m=1}^{M}\Big[\exp\Big\{-\nu_{lm}K_{lm}(\theta, d)\Big\}\nu_{lm}^{n_{lm}}\mathrm{Ga}(\nu_{lm}|c_0H_0(A_{lm}), c_0)\Big]$. Therefore, the mixture weights are independently gamma distributed, $\mathrm{Ga}(\nu_{lm}|c_0H_0(A_{lm}) + n_{lm}, c_0 + K_{lm}(\theta, d))$ for $l = 1, \ldots, L$, $m = 1, \ldots, M$.

Finally, each of the remaining parameters, θ, d, a_β, b_β, c_0, b_1, and b_2 is updated with a Metropolis-Hastings (M-H) step, using a log-normal proposal distribution in each case except for d, for which we use a normal proposal distribution.

4.4 Simulation study

Using three synthetic data sets, we illustrate the proposed model and present its characteristics by comparison with two alternative models: the (parametric) ETAS model and a semiparametric model defined as follows:

$$h(x,\kappa) = \alpha(\kappa)g(x) = a_\alpha \exp\{\eta(\kappa-\kappa_0)\}\int_{\mathbb{R}^+} \theta^{-1}\mathbf{1}_{[0,\theta)}(x)dG(\theta), \quad G \sim \mathrm{DP}(G_0,\alpha_0).$$

The semiparametric model shares $\alpha(\kappa)$ with the ETAS model, and we adopt the uniform DP mixture in Section 3.1.2.2 to model $g(x)$. The mark density function for the alternative models assumes an exponential density for $\kappa \geq \kappa_0$, whereas our proposed model has a beta density function for the transformed mark $u_\kappa = (\kappa-\kappa_0)/(\kappa_{max}-\kappa_0)$.

Since our nonparametric modeling framework targets the excitation function, $h(x,\kappa)$, the section focuses on inference about $\alpha(\kappa)$ and $g_\kappa(x)$.

Priors for the alternative models are chosen, considering their prior uncertainty. In light of the fact that our nonparametric model features a high degree of flexibility, it is not feasible to make all models' uncertainty consistent. For the alternatives, we select priors that may not have the same level of uncertainty as the proposed model, but that cover at least true $\alpha(\kappa)$ and $g_\kappa(x)$. The ETAS model has a set of priors as following: Exp(4) for a_α, Exp(1) for η, Exp(0.1) for p and c, and Exp$(0.5,\eta)$ with support $\psi > \eta$ for ψ. The semiparametric model takes the priors, Ga$(5, 0.25)$ for α_0, Exp(1) for a_Z, Exp(1.5) for b_Z, and $L = 200$, along with the analogous priors for a_α, η, and ψ. Priors for the nonparametric model will be given later in each section. The constant immigrant intensity μ, common for all models, takes exponential priors: Exp(12.4) for the first

example, Exp(9.8) for the second example, and Exp(11.34) for the third example. We set the mark space to $\mathcal{K} = (4, 10)$ for the nonparametric model and $\mathcal{K} = [4, \infty)$ for the ETAS and semiparametric models.

For inference, we collect $2,000$ posterior samples by taking every 9th element after $2,000$ burn-in from the total $20,000$ MCMC iterations.

4.4.1 Power-law density example

The first data set comprises 808 time points (92 immigrants/716 offspring), observed over $(0, T) = (0, 5000)$. They arose from a MHP with ground intensity $\lambda_g^*(t) = 0.02 + \sum_{t_i<t} 0.8 \exp\{0.1(\kappa_i - \kappa_0)\}\text{Powlaw}(t - t_i|20, 2)$ for $t > 0$ and mark density $\text{Exp}(\kappa|1, 4, 10)$, a truncated exponential density function with rate 1 and support $\kappa \in [4, 10]$. The offspring density function in the ground process intensity is independent of κ, so we can detect if the mark-dependent structure of $g_\kappa(x)$ in the nonparametric model distorts the density estimation.

Following the strategy for the prior specification in Section 4.3.3, we chose the following priors for the nonparametric model parameters: $\text{Lo}(2, 0.2)$ for θ with $L = 5$; $\text{Exp}(1)$ and $\text{Exp}(2.5)$ for b_1 and b_2; $\text{N}(0, 100)$ for d with $M = 6$; $\text{Exp}(0.005)$ for c_0; $\text{Exp}(1)$ and $\text{Exp}(0.2)$ for a_β and b_β.

In Figure 4.3, all models contain true $\alpha(\kappa)$ within their interval estimates, while the nonparametric model has a relatively large posterior variance. Note that the alternative models have the same form of $\alpha(\kappa)$ as the true exponential function. Yet, the nonparametric model yields a competitive posterior estimated function, relatively close

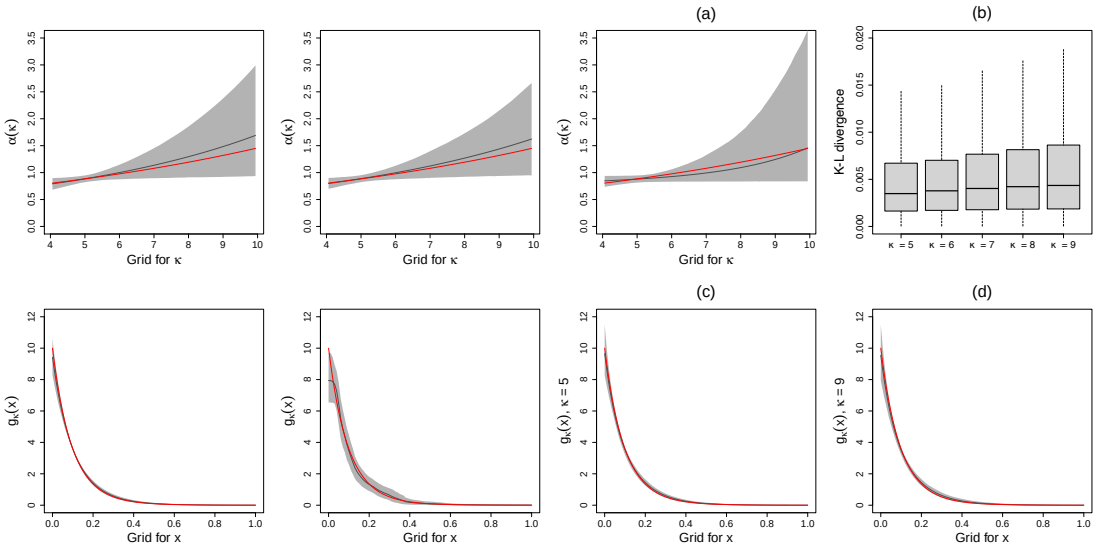

Figure 4.3: Power-law density example. ETAS (first column), semiparametric (second column), and nonparametric (third and fourth columns) models. The posterior means (black line) and 95% interval estimates (shaded area) for: $\alpha(\kappa)$ with the true intensity (red line) in the first row; $g_\kappa(x)$ with the true densities (red and blue lines) in the second row. (b) denotes the Kullback-Leibler divergence under different values of κ.

to the true at some κ. When it comes to $g_\kappa(x)$ estimation, the nonparametric model encompasses posterior mean estimated functions that correspond to true $g_\kappa(x)$, covered by the interval estimates, regardless of the κ values (see, (c) and (d) of Figure 4.3). They are comparable with the parametric estimated function, whose prior model for $g_\kappa(x)$ has the analogous framework to the true function. Additionally, we calculated the K-L divergence to quantify the difference between the true and nonparametric estimated density functions (panel (b) of the figure). There is very little divergence across all models: they are all roughly 0.004 in median.

Unlike the other models, the absence of a decreasing function in the semiparametric modeling framework for $g_\kappa(x)$ results in posterior estimates missing true densities

at some initial points of the grid.

4.4.2 Mark-dependent power-law density example

We generated a point pattern of size 1020 (207 immigrants and 813 offspring) from a MHP with ground intensity $\lambda_g^*(t) = 0.04 + \sum_{t_i<t} 0.35 \exp\{0.6(\kappa_i - \kappa_0)\}\text{Powlaw}(t - t_i|10 + \kappa_i, 1)$ for $t \in (0, 5000)$ and mark density $\text{Exp}(\kappa|1, 4, 10)$. The power-law offspring density function of the example has dependence on the mark history κ_i in the shape parameter.

The intensity function expresses a plausible seismic pattern that aftershocks associated with greater magnitudes of their parents will probably occur sooner rather than later, which results in an offspring density function that is more concentrated around 0.

The nonparametric model took the following priors: $\text{Lo}(2, 0.03)$ for θ with $L = 30$; Exp(1) and Exp(2.5) for b_1 and b_2; $\text{N}(0, 100)$ for d with $M = 6$; Exp(0.005) for c_0; Exp(1) and Exp(0.2) for a_β and b_β.

In Figure 4.4, estimated $\alpha(\kappa)$ of the nonparametric model is comparable with those of the alternatives, whose prior models for $\alpha(\kappa)$ are analogous to the true function. The ETAS and semiparametric models struggle to recapture the underlying $g_\kappa(x)$ for $\kappa = 9$, but the mark-dependent offspring density estimates under the nonparametric model cover well both the true functions for $\kappa = 5$ and 9 (see, (c) and (d) of Figure 4.4). Panel (b) of the figure presents the posterior distributions of the probability mass in $(0, 0.1)$ for different values of κ. Their increasing pattern over κ corresponds to the

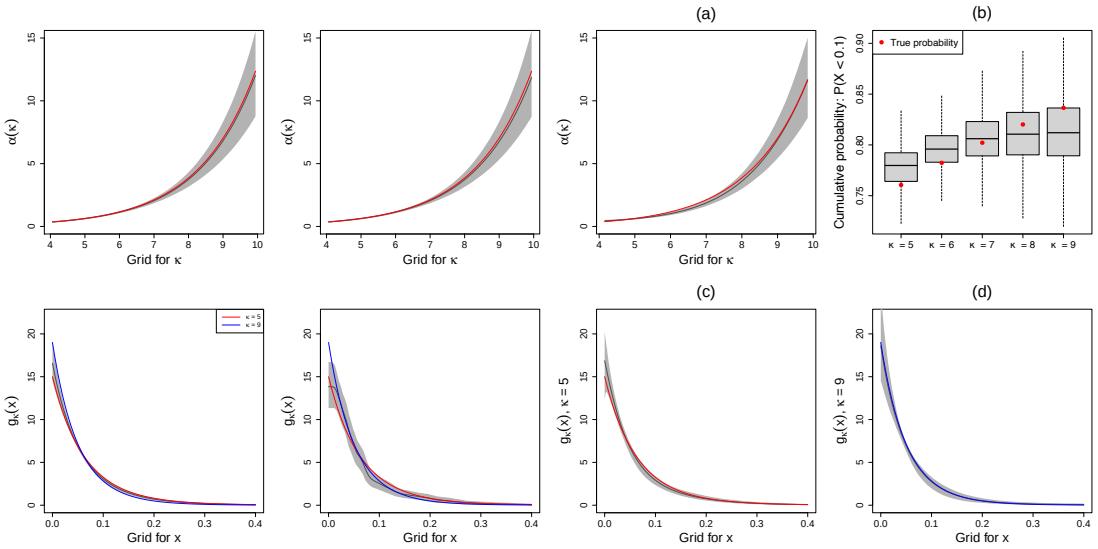

Figure 4.4: Mark-dependent power-law density example. ETAS (first column), semi-parametric (second column), and nonparametric (third and fourth columns) models. The posterior means (black line) and 95% interval estimates (shaded area) for: $\alpha(\kappa)$ with the true intensity (red line) in the first row; $g_\kappa(x)$ with the true densities (red and blue lines) in the second row. (b) denotes cumulative probabilities for $X < 0.1$ under different values of κ.

mark-dependent behavior of the underlying density function.

4.4.3 Mixture of mark-dependent power-law densities example

We consider a data set consisting of 882 points (109 immigrants and 773 offspring), generated from a MHP with more complex ground intensity, $\lambda_g^*(t) = 0.02 + \sum_{t_i<t} 0.25 \exp\{0.8(\kappa_i - \kappa_0)\}[0.6\text{Powlaw}(t - t_i|10 + \kappa_i, 1) + 0.4\text{Powlaw}(t - t_i|10, 1 + \kappa_i)]$ for $t \in (0, 5000)$. Each power-law density of $g_\kappa(x)$ has a mark-dependence either on the shape or the scale parameter. With a larger mark κ_i, the density function has more probabilities near 0 and a longer tail. Such a pattern is another possible behavior of aftershocks. In particular, the longer tail allows aftershocks to occur weeks or even

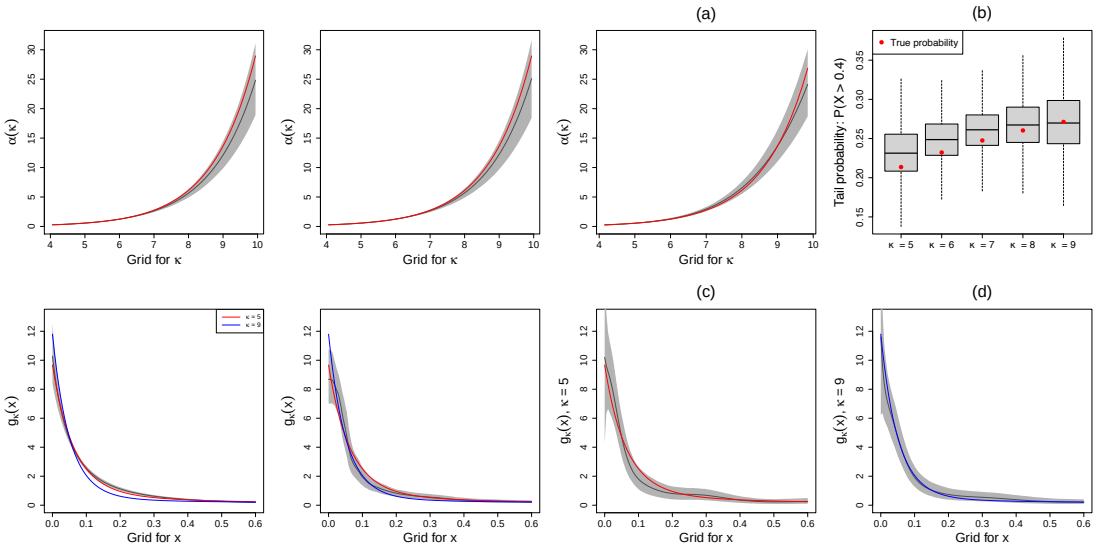

Figure 4.5: Mixture of mark-dependent power-law densities example. ETAS (first column), semiparametric (second column), and nonparametric (third and fourth columns) models. The posterior means (black line) and 95% interval estimates (shaded area) for: $\alpha(\kappa)$ with the true intensity (red line) in the first row; $g_\kappa(x)$ with the true densities (red and blue lines) in the second row. (b) denotes tail probabilities for $X > 4$ under different values of κ.

months after the parent earthquake with a large magnitude has occurred. The truncated exponential density $\text{Exp}(\kappa|1, 4, 10)$ is still used as the mark density function.

As for the prior for the nonparametric model, we assigned $\text{Lo}(2, 0.005)$ to θ with $L = 200$; $\text{Exp}(1)$ and $\text{Exp}(2.5)$ to b_1 and b_2; $\text{N}(0, 100)$ to d with $M = 6$; $\text{Exp}(0.005)$ to c_0; $\text{Exp}(1)$ and $\text{Exp}(0.2)$ to a_β and b_β.

In Figure 4.5, the nonparametric model outperforms the alternatives for $\alpha(\kappa)$ estimation, providing the posterior mean intensity closest to the true function. Limitation of the ETAS model becomes manifest in $g_\kappa(x)$ estimation: the estimated density function struggles to retrieve true $g_\kappa(x)$ for the two κ values, especially for $\kappa = 9$. The semiparametric model works better with larger posterior interval estimates to cover

most of the true values except for the densities near 0 for $\kappa = 9$. The nonparametric model exhibits the most accurate performance in $g_\kappa(x)$ estimation, with its posterior intervals covering the true densities for both $\kappa = 5$ and 9. With various κ values, we observed that each estimated function had a true $g_\kappa(x)$ within the 95% posterior interval estimates. Panel (b) of Figure 4.5 indicates the posterior distributions of the tail probability for $X > 0.4$ in terms of $g_\kappa(x)$ for different κ values. The nonparametric model provides tail probabilities that increase over κ, as intended by the true density function.

4.5 Real data analysis

In order to illustrate the proposed model, we present two earthquake examples. In addition to the analysis of the Japan earthquake catalog discussed in Section 3.4, we also analyze Southwestern USA earthquakes that have occurred in California and Nevada over the course of the past century in Section 4.5.2.

In the synthetic data analysis, we compare the ETAS, semiparametric, and proposed nonparametric models. In order to assess the models, we primarily use the predictive count criterion, described in Section 3.2.2.2.

We took the following priors for the ETAS and semiparametric models: Exp(4) for a_α, Exp(1) for η, Exp(0.1) for p and c, and Exp$(0.5, \eta)$ with support $\psi > \eta$ for ψ for the ETAS model; Ga$(5, 0.25)$ for α_0, Exp(1) for a_Z, Exp(1.5) for b_Z, and $L = 200$, along with analogous priors for a_α, η, and ψ, for the semiparametric model. The constant

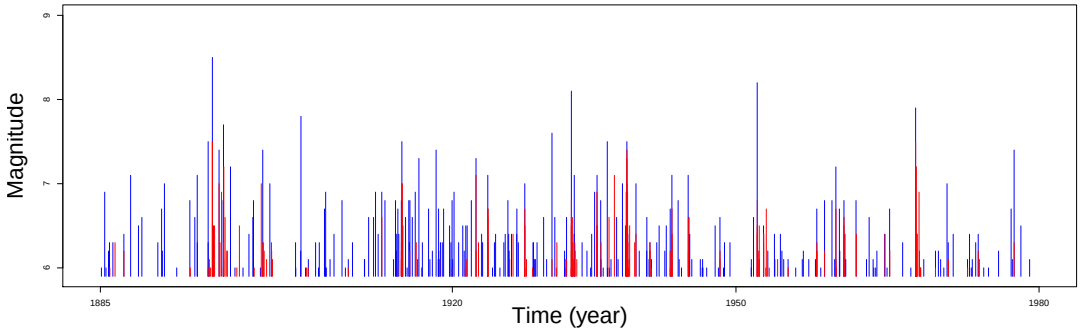

Figure 4.6: Japan earthquake point pattern split into main shocks (blue) and aftershocks (red).

immigrant intensity μ of all models is assigned Exp(139) for the Japan data set and Exp(181) for the USA data set.

All the following inferences are based on $2,000$ posterior samples, obtained after 2000 burn-in and 9 step-sized thinning.

4.5.1 Japan earthquake

The earthquake data set used for unmarked HP models is reused here. This study deals with earthquake occurrence times as well as the magnitudes of earthquakes associated with them. We split the data and targeted earthquakes (118 points) that occurred after 1950/01/01 as a test set for prediction, and fitted the models to the remaining 340 data points (Figure 4.6). The predictive count allows for the evaluation of models by matching the count with the test set and, as a consequence, for model comparison.

According to the prior specification in Section 4.3.3, we assigned the following

priors to the nonparametric model: Lo$(2, 0.03)$ for θ with $L = 30$; Exp(1) and Exp(3) for b_1 and b_2; N$(0, 100)$ for d with $M = 12$; Exp(0.005) for c_0; Exp(1) and Exp(0.2) for a_β and b_β; Exp(139) for μ. Our selection of L and M was based on sensitivity analysis through multiple MCMC runs with increasing parameters. We also consider how much the posterior results on Q-Q plots for the time-rescaling theorem improve with larger L and M. In the prior, $L = 30$ and $M = 12$ are practical choices. For larger L and M, for example, the improvement in the Q-Q plot was indistinguishable while computing time grew. We set the mark space to $\mathcal{K} = (\kappa_0, \kappa_{max}) \approx [6, 8.5]$, the minimum and maximum of observed marks κ_i, for the nonparametric model and $\mathcal{K} = [6, \infty)$ for the ETAS and semiparametric models.

From a modeling perspective, the ETAS and semiparametric models differ in the form for $g_\kappa(x)$. As a result, the semiparametric model has a smaller increasing rate in $\alpha(\kappa)$ and more probability near 0 in the estimated offspring density function than does the ETAS model (Figure 4.7). The mixture-form $\alpha(\kappa)$ of the nonparametric model yields a relatively small increasing rate in $\alpha(\kappa)$, compared to the exponential-form $\alpha(\kappa)$ in the other models. Since the function refers to the number of aftershocks at κ in earthquake applications, the smaller rate of increase implies that the model would have relatively few aftershocks. The mark-dependent offspring density estimates derived from the nonparametric model indicate that the time between an aftershock and its parent gets shorter as the magnitude of the parent increases. Therefore, an earthquake with a large magnitude is more likely to cause aftershocks that occur sooner rather than later. This pattern results in a higher concentration of the offspring density function at 0 for

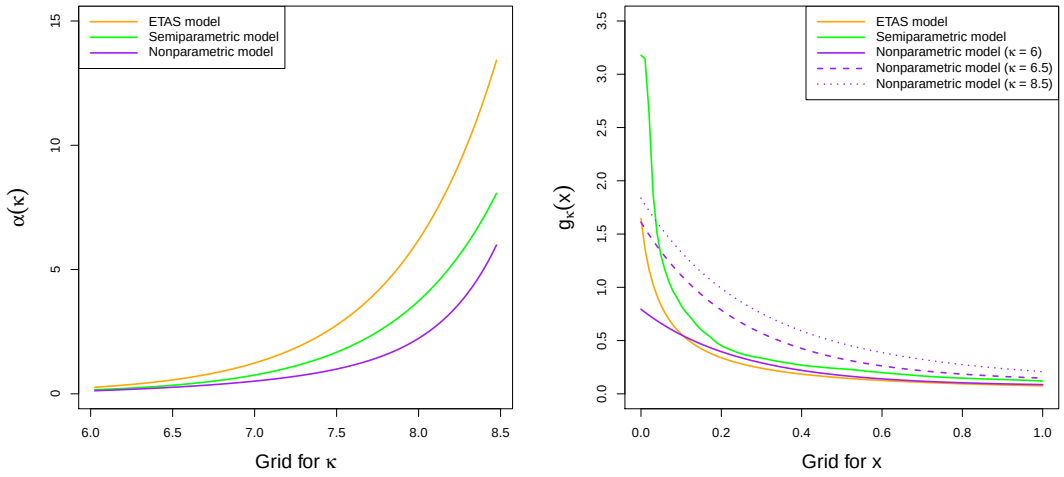

Figure 4.7: Japan earthquakes. The posterior means for: the total offspring intensity function in the left panel; the offspring density function in the right panel under each model.

larger κ.

For model comparison, we apply the misclassification criteria described in Section 3.3.1 and specifically compute the misclassification rate $R = (M_I + M_O)/n$. We found that the nonparametric model has the smallest posterior mean and standard deviation for R: 0.226 and 0.011 for the nonparametric model; 0.232 and 0.023 for the semiparametric model; and 0.265 and 0.025 for the ETAS model.

In this example, we can evaluate the misclassification rate because the information about the type of earthquake (main shock/aftershock) is provided. But, the predictive count can serve as a more general tool for model comparison, since it can be calculated solely from the data without any further information. Denote by $N_{\text{pred}}(B)$ the predictive count in $B \subseteq [1950, 1980] \times [6, 8.5]$. We can obtain $N_{\text{pred}}(B)$ by counting

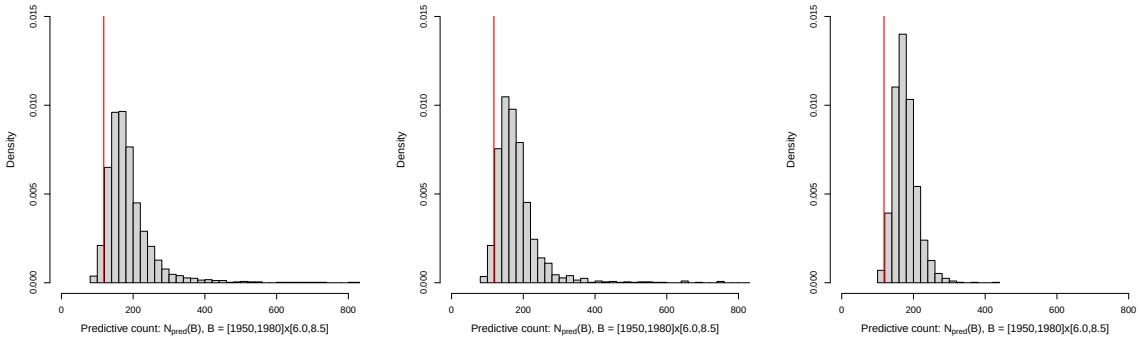

Figure 4.8: Japan earthquakes. Posterior distributions (histogram) for the predictive count, $N_{\text{pred}}(B)$, under the ETAS (left), semiparametric (middle), and nonparametric (right) models, along with the observed count (i.e., the test set size), $N_{\text{obs}}(B) = 118$ (red line).

the number of points of a MHP realization that lie in B at every MCMC iteration. Figure 4.8 displays the posterior distributions of $N_{\text{pred}}(B)$ under each model. The semiparametric model outperforms the ETAS model when it comes to prediction, with the posterior predictive mean being nearer the observed value and with a smaller variance. While all models overpredict the test set size, the nonparametric model yields smaller posterior predictive variance. Furthermore, we conducted the prediction using another training and test sets that were divided by a different time, 1970/01/01. The nonparametric model retained its effectiveness with regard to posterior predictive variability.

While in this chapter, we work with a constant immigrant intensity to focus on excitation function estimation, the data revealed non-constant intensities for the immigrant point pattern in Section 3.4, where the nonparametric model led to a more accurate prediction result. We thus extend the MHP model using the Erlang mixture immigrant intensity prior from Section 3.1.1. The first panel of Figure 4.9 shows the

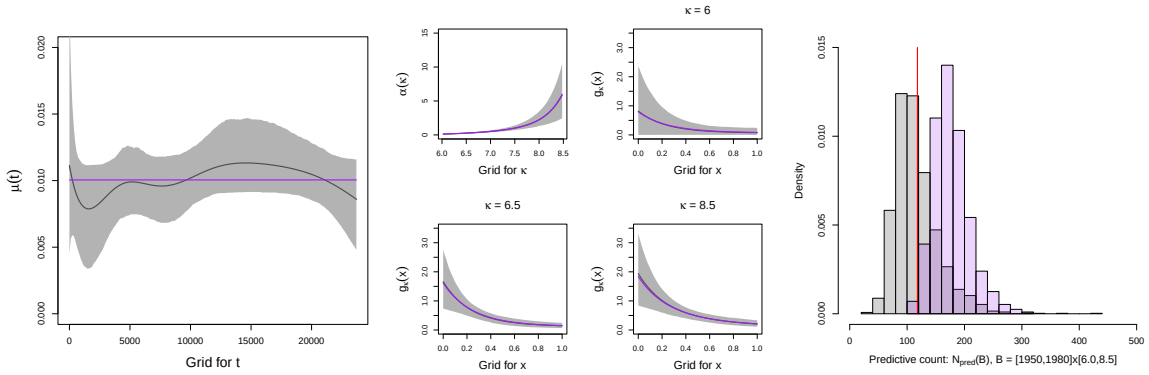

Figure 4.9: Japan earthquakes. Fully nonparametric model with non-constant immigrant intensity. Posterior mean (gray line) and 95% interval (gray shaded area) estimates of: immigrant intensity (left); and $\alpha(\kappa)$ and $g_\kappa(x)$ under different κ values (four panels in the middle). The histogram (gray) in the right panel presents the posterior predictive distribution. For comparison, the results of the nonparametric model with constant immigrant intensity are displayed in purple.

result of the immigrant intensity estimation, which indicates a non-constant shape. The panels in the middle present $\alpha(\kappa)$ and $g_\kappa(x)$ posterior estimates, which are similar to the ones (the posterior means in purple) from the nonparametric model with constant immigrant intensity. The posterior predictive distribution in the last panel highlights the utility of using the Erlang mixture immigrant intensity. The posterior predictive mean is much closer to the observed count, with little increase in the variance. The inferences for the fully nonparametric model are based on 2,000 posterior samples, derived from 40,000 MCMC iterations by 4,000 burn-in and 18 steps thinning.

4.5.2 Southwestern USA (California and Nevada) earthquake

The data set contains 423 earthquakes that occurred from 1921/1/1 to 2020/12/31 in the southwestern USA, with a rectangular region from 32.2° to 41° latitude and from

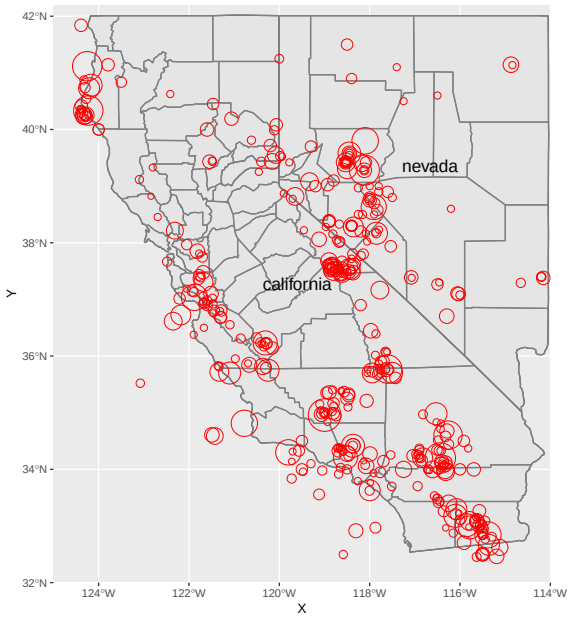

Figure 4.10: Southwestern USA earthquakes (red circle). The radius of the circle indicates the magnitude of the earthquake.

$-125°$ to $-114°$ longitude. The earthquake catalog contains the magnitude, as well as the occurrence time and location. We only consider earthquakes of magnitude greater than 5 in our analysis. Figure 4.10 exhibits the epicenters of earthquakes with their magnitudes shown as radii. As a test set for prediction, we selected the last 170 earthquakes, which occurred within the last 40 years. A model was then fitted to the remaining 253 earthquakes. Our data set comes from a publicly available earthquake repository at `https://earthquake.usgs.gov/earthquakes/search/`.

We considered the following priors for the nonparametric model: $\text{Lo}(2, 0.005)$ for θ with $L = 200$; $\text{Exp}(1)$ and $\text{Exp}(2.5)$ for b_1 and b_2; $\text{N}(0, 100)$ for d with $M = 12$; $\text{Exp}(0.005)$ for c_0; $\text{Exp}(1)$ and $\text{Exp}(0.25)$ for a_β and b_β; $\text{Exp}(173)$ for μ. Considering observed marks with minimum 5.01 and maximum 7.5, we set the mark space to $(5, 7.6)$

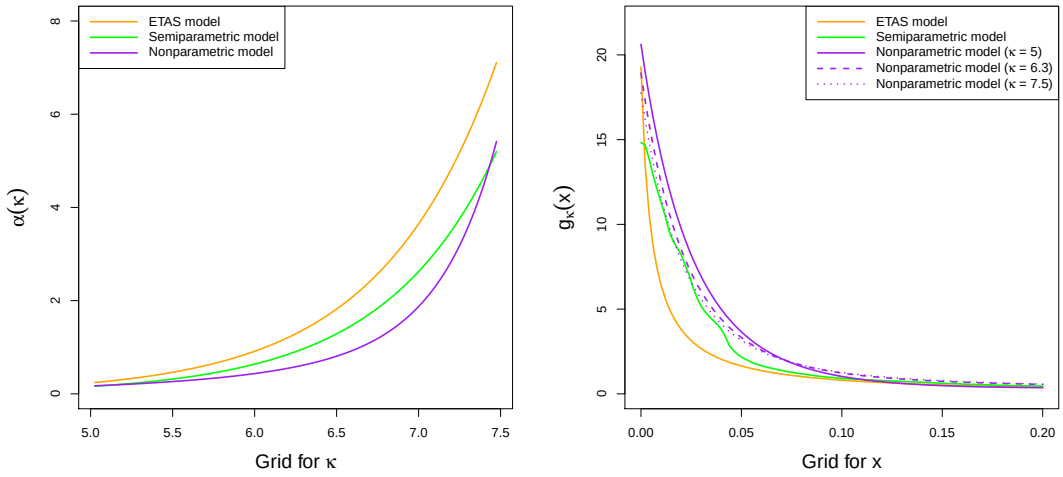

Figure 4.11: Southwestern USA earthquakes. The posterior means for: the total offspring intensity function in the left panel; the offspring density function in the right panel under each model.

for the nonparametric model and $[5, \infty)$ for the alternative models.

In Figure 4.11, the nonparametric model indicates relatively small offspring intensity estimates and, consequently, fewer aftershocks for each κ. Estimates of offspring density from the nonparametric model show small but varying rates of decline in κ. Increasing κ results in a slower rate of decrease with a smaller probability near 0 and a longer tail. This pattern suggests that aftershocks from earthquakes of higher magnitude are more likely to occur over a longer period of time. Further, compared to the Japan earthquakes, the estimated offspring density functions from all models are more centered around 0. From the results, we can infer that southwestern US earthquakes have relatively short intervals between aftershocks and their parents.

Plotted in Figure 4.12 are prediction results from each model under three

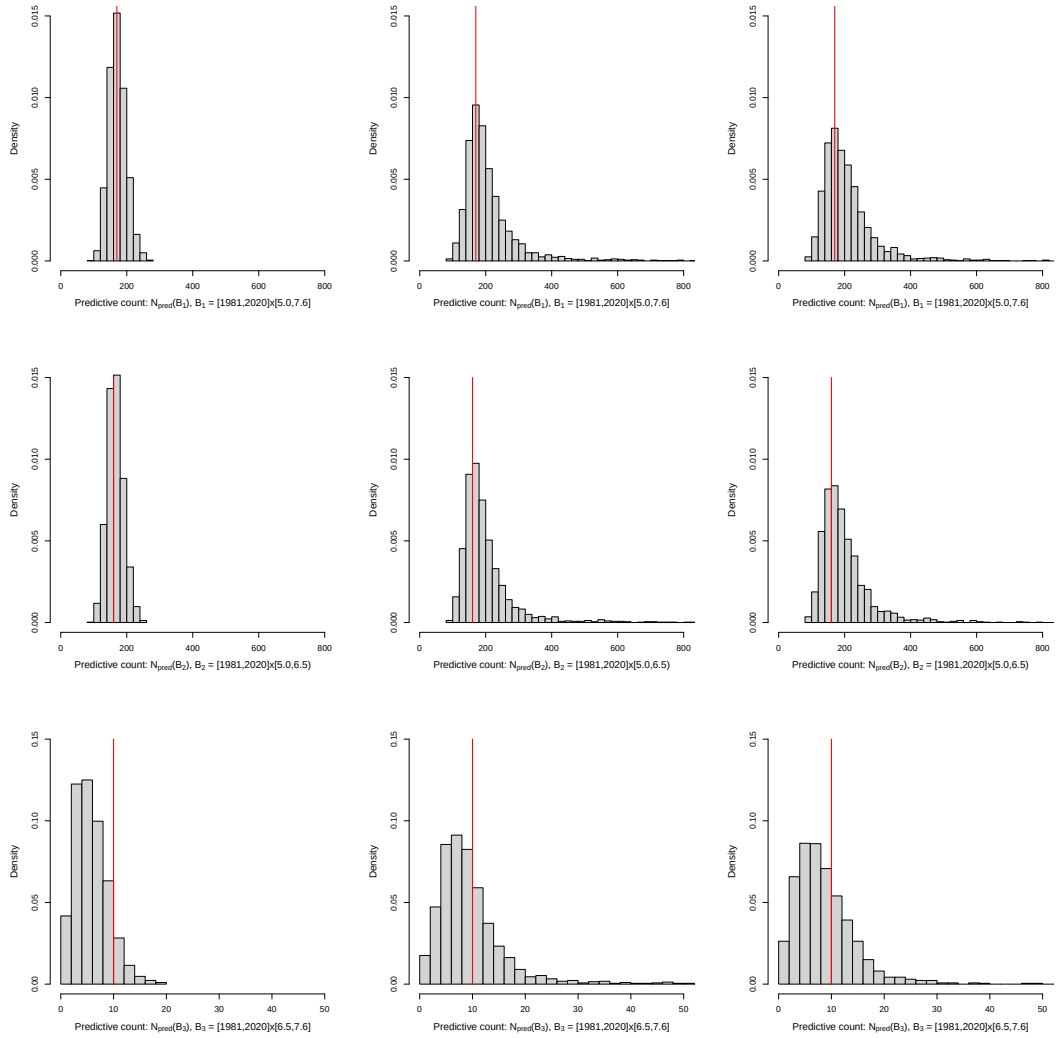

Figure 4.12: Southwestern USA earthquakes. Posterior distributions (histogram) for the predictive count, $N_{\text{pred}}(B)$, under the nonparametric (first column), semiparametric (second column), and ETAS (third column) models, for different domains of B: $B_1 = [1981, 2020] \times [5.0, 7.6]$ in the first row; $B_2 = [1981, 2020] \times [5.0, 6.5)$ in the second row; $B_3 = [1981, 2020] \times [6.5, 7.6]$ in the third row. The red line denotes the observed count in each B.

different prediction domains distinguished by mark intervals (but with a common time interval): the whole grid $[5.0, 7.6]$ for κ in B_1; an initial part of the grid $[5.0, 6.5)$ in B_2;

the remainder of the grid $[6.5, 7.6]$ in B_3.

The nonparametric model outperforms the alternatives with a predictive count corresponding to $N_{\text{obs}}(B)$ with a smaller variance (top row of the figure). For a comprehensive prediction result, we separated the mark grid into sub-domains and calculated the predictive counts in each sub-domain. In nearly all sub-domains, the nonparametric model provides the best performance, except for the tails of the mark grid. In particular, the proposed model underpredicts the count $N_{\text{obs}}(B_3) = 10$ compared to the alternatives (bottom row of the figure).

Furthermore, we have applied the fully nonparametric model in which the Erlang mixture immigrant intensity replaces the constant immigrant intensity (results not shown). The immigrant intensity estimates depart from a constant function less significantly than those for the Japan earthquake data set. Consequently, the use of mixture immigrant intensity does not improve prediction accuracy. The fully nonparametric model may nonetheless be verified as a useful tool for examining the assumption of constant immigrant intensity underpinning the ETAS model.

4.6 Discussion

We have proposed a Bayesian nonparametric model for the MHP excitation function. The nonparametric model is defined as a mixture of basis functions, each of which is composed of an Erlang density for time and a polynomial function for the mark. Mixture weights are defined through increments of a measure, to which we assign

a gamma process prior.

It is a key characteristic of this methodology that it strikes a good balance between model flexibility and computational efficiency. Using the gamma random measure allows flexibility in the excitation function, so the prior model is able to model a wide range of offspring intensity and density functions. Additionally, the use of the gamma process prior enables prior-to-posterior updating of the mixture weight using gamma posterior full conditional distributions. In our modeling approach, the likelihood normalization term is handled by an efficient MCMC algorithm that does not require approximations or advanced computational methods. Moreover, the proposed model has a parsimonious design; all the basis functions of the mixture model have just two common parameters θ and d.

More importantly, the proposed model can accommodate mark-dependence for the offspring density function. The ETAS model in Ogata (1988) assumes a factorization of the excitation function into the two key functions, $\alpha(\kappa)$ and $g(x)$. Spatio-temporal MHP models for earthquake applications also contain structures for the excitation function from which separate functions for the mark and time are derived (e.g., Ogata, 1988; Kagan, 1991; Musmeci and Vere-Jones, 1992; Ogata, 1998; Kumazawa and Ogata, 2014; Nandan et al., 2017). In contrast to existing models, our model results in a mark-dependent offspring density function, allowing us to explore the aftershocks' density in relation to their parents' magnitude. Indeed, we have applied the model to the Japan and California earthquakes and found two different patterns in the density function for the aftershocks. Japan's earthquakes with larger magnitudes have their aftershocks

more likely to occur near the parent, resulting in a more concentrated density around zero. Contrary to this, the larger magnitude earthquakes in California are associated with aftershocks that occur over a longer period of time, resulting in a longer tail density for aftershocks.

We have evaluated the model by forecasting earthquakes using the predictive count. The nonparametric model offers improvement regarding posterior predictive variability. For brevity, we adopt the assumption of a constant immigrant intensity that is part of the ETAS model. However, when the actual immigrant intensity is expected to be non-constant, the fully nonparametric model, which replaces the constant function with the Erlang mixture, would improve prediction accuracy as in the Japan earthquake data set.

Even though the proposed model has a number of positive attributes, there are a few downsides as well. We provide the priors for all parameters of the nonparametric model except for M, which poses a challenge. Accordingly, the selection of M in this study results from sensitivity analysis with increasing M, thereby necessitating multiple MCMC runs. Compared to the ETAS model, the proposed model involves an upper bound for the mark space to transform marks. Although such a bound may be a restriction for general settings, this is not the case for applications to earthquake modeling for which it is natural to set an upper bound on earthquake magnitude.

The focus of this chapter is on modeling the excitation function based on the unpredictable mark assumption, wherein the mark density is independent of the current time and the MHP history. A future research direction involves models with more

general mark density functions that are dependent upon the history and are allowed to be different for main shocks and aftershocks. Model development in this direction is prompted by the finding that main shocks tend to be of greater magnitude than aftershocks. The model should provide better estimates of the total offspring intensity and enhance the prediction of aftershocks.

Chapter 5

Conclusions

The main objective of this dissertation is to provide a Bayesian nonparametric modeling and inference framework for Hawkes processes. The methods extend the inferential scope for the Hawkes process (HP), enabling flexible modeling and tractable inference for its conditional intensity function.

In Chapter 2, we have developed a Bayesian nonparametric model for NHPP intensity functions. The method builds from a mixture representation for the intensity function, for which the basis functions are defined as Erlang densities with common shape parameter. We assign a gamma process prior to the cumulative intensity function the increments of which become the mixture weights. The gamma process prior is the key for model flexibility and computational efficiency in posterior inference. The cumulative intensity can have large prior uncertainty due to the gamma process prior with small precision parameter. The model convergence to the intensity function verifies the model flexibility. The gamma process prior implies independent gamma priors for

the weights. Along with the simple representation of the normalizing constant under the model, the conjugate gamma priors facilitate efficient model implementation for posterior inference.

In Chapter 3, we have proposed Bayesian nonparametric modeling approaches to HP intensities, based on the Erlang mixture model. The HP cluster representation with its factorization of the conditional intensity function allows the flexible prior models to bring their benefits to the HP immigrant intensity and excitation function, including efficient handling of the likelihood normalizing constant. As an alternative method to the offspring Erlang mixture, we have developed a nonparametric mixture model based on the Dirichlet process or the geometric weights prior. The alternative approach focuses on estimating non-increasing offspring densities. In addition to inference methods for model parameters, we provide numerical approaches to inference about the HP first- and second-order intensities. We have considered several criteria for model assessment and comparison, including branching structure estimators and the predictive count for future events.

We have further elaborated the Erlang mixture model for the MHP conditional intensity function in Chapter 4, with focus on MHPs for earthquake modeling. The basis function, whose mixture forms the ground process excitation function, has a multiplicative form consisting of an Erlang density for time and a polynomial function for the mark, which corresponds to earthquake magnitude. The beta density function was chosen as the mark density for its benefit of ready expressions for key functions, such as the total offspring intensity. The mixture weights are defined through incre-

ments of a measure, assigned a gamma process prior for flexible modeling and efficient handling of the likelihood normalizing constant. In addition, since we model the entire excitation function (not separate functions of the excitation as in the ETAS model), the proposed method allows the offspring density to vary with the mark. In the context of earthquake modeling applications, such an offspring density function enables one to explore the density of an aftershock, whose decreasing rate and/or tail behavior change according to earthquake magnitude.

Through the thesis, we have introduced and detailed new modeling methods for HP intensities along with several illustrations with synthetic and real data. A practically important area for future work involves development of comparison tools applicable to real data analyses. Graphical comparison using Q-Q plots for the time-rescaling theorem may not be very informative. For instance, the earthquake data used for model illustration in real data analyses did not provide remarkable differences in the plots. A restriction of our quantitative evaluation methods is that it requires, for model comparison, the immigrant-offspring clustering information for each observed point. Unlike the Japan earthquake data set, the US earthquake catalog does not involve the clustering information, so we compared models only though the predictive count. Developing other comparison criteria will help to study and describe model differences in real data analyses from different perspectives.

Fixed mixture size L of the Erlang mixture models for NHPP and HP intensities can be justified by brevity in modeling framework and by efficiency in model implementation. We also have a strategy to choose a relevant L, which can be further

improved by sensitivity analysis. But, we do not have such a general strategy to specify M (the number of mixture components for the total offspring intensity) of the MHP modeling framework, which is a limitation of our model.

Since the modern ETAS model involves spatial information in the excitation function (e.g., Ogata, 1998; Ogata and Zhuang, 2006), nonparametric models for spatio-temporal HPs will be a natural and practical extension. In addition, it will be interesting to study nonparametric MHPs with more general mark densities that depend on the history, in particular, allowing main shocks and aftershocks to have different mark distributions for earthquake magnitude.

Bibliography

Adamopoulos, L. (1976). Cluster models for earthquakes: Regional comparisons. *Journal of the International Association for Mathematical Geology*, 8(4):463–475.

Adams, R. P., Murray, I., and MacKay, D. J. C. (2009). Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities. In *Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada.

Andersen, P. K. and Gill, R. (1982). Cox's regression model for counting processes: A large sample study. *The Annals of Statistics*, 10:1100–1120.

Andrews, D. F. and Herzberg, A. M. (1985). *Data, A Collection of Problems from Many Fields for the Student and Research Worker*. Springer-Verlag.

Baddeley, A. and Turner, R. (2005). spatstat: An R package for analyzing spatial point patterns. *Journal of Statistical Software, Articles*, 12(6).

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.

Bocharov, P. P., D'Apice, C., and Pechinkin, A. V. (2011). *Queueing Theory (Reprint)*. De Gruyter.

Brix, A. and Diggle, P. J. (2001). Spatiotemporal prediction for log-Gaussian Cox processes. *Journal of the Royal Statistical Society. Series B*, 63:823–841.

Brix, A. and Møller, J. (2001). Space-time multi type log Gaussian Cox processes with a view to modelling weeds. *Scandinavian Journal of Statistics*, 28:471–488.

Burridge, J. (1981). Empirical Bayes analysis of survival time data. *Journal of the Royal Statistical Society. Series B*, 43:65–75.

Butzer, P. L. (1954). On the extensions of Bernstein polynomials to the infinite interval. *Proceedings of the American Mathematical Society*, 5:547–553.

Cox, D. R. (1955). Some statistical methods connected with series of events. *Journal of the Royal Statistical Society. Series B*, 17:129–164.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

Cressie, N. A. C. (1993). *Statistics for Spatial Data (Revised edition)*. Wiley.

Daley, D. J. and Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes(2nd edition)*. Springer.

Damien, P., Dellaportas, P., Polson, N. G., and Stephens, D. A. (2013). *Bayesian theory and applications*. OUP Oxford.

Diggle, P. (1985). A kernel method for smoothing point process data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 34(2):138–147.

Diggle, P., Rowlingson, B., and Su, T. (2005). Point process methodology for on-line spatio-temporal disease surveillance. *Environmetrics*, 16(5):423–434.

Diggle, P. J. (2014). *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns(3rd edition)*. CRC press.

Diggle, P. J., Moraga, P., Rowlingson, B., and Taylor, B. M. (2013). Spatial and spatio-temporal log-Gaussian Cox processes: Extending the geostatistical paradigm. *Statistical Science*, 28:542–563.

Donnet, S., Rivoirard, V., and Rousseau, J. (2020). Nonparametric bayesian estimation for multivariate hawkes processes. *The Annals of Statistics*, 48(5):2698–2727.

Ebrahimian, H. and Jalayer, F. (2017). Robust seismicity forecasting based on bayesian parameter estimation for epidemiological spatio-temporal aftershock clustering models. *Scientific reports*, 7(1):1–15.

Ebrahimian, H., Jalayer, F., Asprone, D., Lombardi, A. M., Marzocchi, W., Prota, A., and Manfredi, G. (2014). Adaptive daily forecasting of seismic aftershock hazard. *Bulletin of the Seismological Society of America*, 104(1):145–161.

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

Filimonov, V. and Sornette, D. (2015). Apparent criticality and calibration issues in the hawkes self-excited point process model: application to high-frequency financial data. *Quantitative Finance*, 15(8):1293–1314.

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

Fox, E. W., Short, M. B., Schoenberg, F. P., Coronges, K. D., and Bertozzi, A. L. (2016). Modeling e-mail networks and inferring leadership using self-exciting point processes. *Journal of the American Statistical Association*, 111(514):564–584.

Gelfand, A. E. and Schliep, E. M. (2018). Bayesian Inference and Computing for Spatial Point Patterns. *NSF-CBMS Regional Conference Series in Probability and Statistics*, 10:i–125.

Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press.

Grimmett, G. and Stirzaker, D. (2001). *Probability and Random Processes*. Oxford.

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(442).

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*, 33(3):438–443.

Hawkes, A. G. and Oakes, D. (1974). A cluster process representation of a self-exciting process. *Journal of Applied Probability*, 11(3):493–503.

Heikkinen, J. and Arjas, E. (1998). Non-parametric Bayesian estimation of a spatial Poisson intensity. *Scandinavian Journal of Statistics*, 25:435–450.

Heikkinen, J. and Arjas, E. (1999). Modeling a Poisson forest in variable elevations: A nonparametric Bayesian approach. *Biometrics*, 55:738–745.

Helmstetter, A., Kagan, Y. Y., and Jackson, D. D. (2006). Comparison of short-term and time-independent earthquake forecast models for southern california. *Bulletin of the Seismological Society of America*, 96(1):90–106.

Hjort, N. L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *The Annals of Statistics*, 18:1259–1294.

Illian, J. B., Sørbye, S. H., and Rue, H. (2012). A toolbox for fitting complex spatial point process models using integrated nested Laplace approximation (INLA). *The Annals of Applied Statistics*, 6:1499–1530.

Ishwaran, H. and James, L. F. (2004). Computational methods for multiplicative intensity models using weighted gamma processes: Proportional hazards, marked point

processes, and panel count data. *Journal of the American Statistical Association*, 99:175–190.

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

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

Kang, J., Nichols, T. E., Wager, T. D., and Johnson, T. D. (2014). A Bayesian hierarchical spatial point process model for multi-type neuroimaging meta-analysis. *The Annals of Applied Statistics*, 8:1800–1824.

Karlin, S. and Taylor, H. M. (1974). *A First Course in Stochastic Processes*. Academic Press, INC.

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

Kingman, J. F. C. (1993). *Poisson Processes*. Clarendon Press.

Kottas, A. (2006). Dirichlet process mixtures of Beta distributions, with applications to density and intensity estimation. In *Proceedings of the Workshop on Learning with Nonparametric Bayesian Methods, 23rd International Conference on Machine Learning*, Pittsburgh, PA, USA.

Kottas, A., Behseta, S., Moorman, D., Poynor, V., and Olson, C. (2012). Bayesian non-

parametric analysis of neuronal intensity rates. *Journal of Neuroscience Methods*, 203:241–253.

Kottas, A. and Sansó, B. (2007). Bayesian mixture modeling for spatial Poisson process intensities, with applications to extreme value analysis. *Journal of Statistical Planning and Inference*, 137:3151–3163.

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*.

Lawless, J. F. (1987). Regression methods for Poisson process data. *Journal of the American Statistical Association*, 82:808–815.

Lee, S. C. K. and Lin, X. S. (2010). Modeling and evaluating insurance losses via mixtures of Erlang distributions. *North American Actuarial Journal*, 14:107–130.

Lee, S. C. K. and Lin, X. S. (2012). Modeling dependent risks with multivariate erlang mixtures. *ASTIN Bulletin*, 42:153–180.

Lewis, E. and Mohler, G. (2011). A nonparametric em algorithm for multiscale hawkes processes. *Journal of Nonparametric Statistics*, 1(1):1–20.

Lo, A. Y. (1982). Bayesian nonparametric statistical inference for Poisson point processes. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 59:55–66.

Lo, A. Y. (1992). Bayesian inference for Poisson process models with censored data. *Journal of Nonparametric Statistics*, 2:71–80.

Lo, A. Y. and Weng, C. S. (1989). On a class of Bayesian nonparametric estimates: ii. hazard rate estimates. *Annals of the Institute for Statistical Mathematics*, 41:227–245.

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.

Møller, J. and Rasmussen, J. G. (2005). Perfect simulation of hawkes processes. *Advances in applied probability*, 37(3):629–646.

Møller, J. and Rasmussen, J. G. (2006). Approximate simulation of hawkes processes. *Methodology and Computing in Applied Probability*, 8(1):53–64.

Møller, J., Syversveen, A. R., and Waagepetersen, R. P. (1998). Log Gaussian Cox processes. *Scandinavian Journal of Statistics*, 25:451–482.

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.

Petrone, S. (1999a). Bayesian density estimation using Bernstein polynomials. *Canadian Journal of Statistics*, 27:105–126.

Petrone, S. (1999b). Random Bernstein polynomials. *Scandinavian Journal of Statistics*, 26:373–393.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association*, 108(504):1339–1349.

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

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

Reinhart, A. (2018). A review of self-exciting spatio-temporal point processes and their applications (with discussion). *Statistical Science*, 33:299–333.

Rodrigues, A. and Diggle, P. J. (2012). Bayesian estimation and prediction for inhomogeneous spatiotemporal log-Gaussian Cox processes using low-rank models, with application to criminal surveillance. *Journal of the American Statistical Association*, 107:93–101.

Rodriguez, A., Wang, Z., and Kottas, A. (2017). Assessing systematic risk in the S&P500 index between 2000 and 2011: A Bayesian nonparametric approach. *The Annals of Applied Statistics*, 11:527–552.

Ross, G. J. (2021). Bayesian Estimation of the ETAS Model for Earthquake Occurrences. *Bulletin of the Seismological Society of America*, To appear.

Schoenberg, F. P., Hoffmann, M., and Harrigan, R. J. (2019). A recursive point process model for infectious diseases. *Annals of the Institute of Statistical Mathematics*, 71(5):1271–1287.

Simpson, D., Illian, J. B., Lindgren, F., Sørbye, S. H., and Rue, H. (2016). Going off grid:

computationally efficient inference for log-Gaussian Cox processes. *Biometrika*, 103:49–70.

Taddy, M. (2010). Autoregressive mixture models for dynamic spatial Poisson processes: Application to tracking the intensity of violent crime. *Journal of the American Statistical Association*, 105:1403–1417.

Taddy, M. A. and Kottas, A. (2012). Mixture modeling for marked Poisson processes. *Bayesian Analysis*, 7:335–362.

Taylor, B., Davies, T., Rowlingson, B., and Diggle, P. (2013). lgcp: An R package for inference with spatial and spatio-temporal log-Gaussian Cox processes. *Journal of Statistical Software*, 52:1–40.

Taylor, B. and Diggle, P. (2014). INLA or MCMC? A tutorial and comparative evaluation for spatial prediction in log-Gaussian Cox processes. *Journal of Statistical Computation and Simulation*, 84(10):2266–2284.

Taylor, B. M., Davies, T. M., Rowlingson, B. S., and Diggle, P. J. (2015). Bayesian inference and data augmentation schemes for spatial, spatiotemporal and multivariate log-Gaussian Cox processes in R. *Journal of Statistical Software*, 63:1–48.

Tijms, H. (1994). *Stochastic Models: An Algorithm Approach*. John Wiley.

Vargas, N. and Gneiting, T. (2012). Bayesian point process modelling of earthquake occurrences. Technical report, Technical Report, Ruprecht-Karls University Heidelberg, Heidelberg, Germany ....

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*, 103:614–624.

Venturini, S., Dominici, F., and Parmigiani, G. (2008). Gamma shape mixtures for heavy-tailed distributions. *The Annals of Applied Statistics*, 2:756–776.

Wolpert, R. L., Clyde, M. A., and Tu, C. (2011). Stochastic expansions using continuous dictionaries: Lévy adaptive regression kernels. *The Annals of Statistics*, 39(4):1916–1962.

Wolpert, R. L. and Ickstadt, K. (1998). Poisson/gamma random field models for spatial statistics. *Biometrika*, 85:251–267.

Xiao, S., Kottas, A., and Sansó, B. (2015). Modeling for seasonal marked point processes: An analysis of evolving hurricane occurrences. *The Annals of Applied Statistics*, 9:353–382.

Xiao, S., Kottas, A., Sansó, B., and Kim, H. (2021). Nonparametric Bayesian modeling and estimation for renewal processes. *Technometrics*, 63:100–115.

Zhang, R., Walder, C., Rizoiu, M.-A., and Xie, L. (2018). Efficient non-parametric bayesian hawkes processes. *arXiv preprint arXiv:1810.03730*.

Zhao, C. and Kottas, A. (2021). Modeling for Poisson process intensities over irregular spatial domains. *arXiv:2106.04654 [stat.ME]*.

Zhou, F., Li, Z., Fan, X., Wang, Y., Sowmya, A., and Chen, F. (2019). Efficient em-variational inference for hawkes process. *arXiv preprint arXiv:1905.12251*.

Zhou, F., Li, Z., Fan, X., Wang, Y., Sowmya, A., and Chen, F. (2020). Efficient inference for nonparametric hawkes processes using auxiliary latent variables. *Journal of Machine Learning Research*, 21(241):1–31.

Zhou, K., Zha, H., and Song, L. (2013). Learning triggering kernels for multi-dimensional hawkes processes. In *International Conference on Machine Learning*, pages 1301–1309. PMLR.

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

Appendix A

Computational performance of the NHPP Erlang mixture model

To report on computing time for implementing the Erlang mixture model, we consider the three synthetic data examples of Section 2.2, which allows us to study the effect of the point pattern size (n) and the number of Erlang basis densities (J). Tables A.1 and A.2 include computing times (in minutes) for 70,000 MCMC posterior samples. We also provide estimates for the effective sample size (ESS), that is, the MCMC sample size adjusted for autocorrelation, thus estimating how many uncorrelated samples the posterior samples are equivalent to. The ESS was computed using function **effectiveSize** of the **R coda** package. Results for the ESS are based on 60,000 posterior samples obtained after discarding the first 10,000 samples. The "mean ESS" reported in the tables is the average of 51 effective sample sizes for $\lambda(t_k)$, for $k = 1, \ldots, 51$, where t_k are equally-spaced points on a grid over $(0, T)$. All MCMC posterior simulations were

	$J = 50$	$J = 75$	$J = 100$
Computing Time	19.6	25.8	33.9
Mean ESS	3391	3065	2984

Table A.1: Synthetic data from temporal NHPP with bimodal intensity (Section 2.2.3). Computing time (in minutes) for 70,000 MCMC iterations, and average of effective sample sizes for the intensity evaluated at 51 grid points in $(0, T)$, under three different values of J.

	Bimodal ($n = 112$)	Decreasing ($n = 491$)	Increasing ($n = 565$)
Computing Time	19.6	28.5	29.5
Mean ESS	3391	5792	2122

Table A.2: Synthetic data from temporal NHPP with bimodal/decreasing/increasing intensity (Section 2.2). Computing time (in minutes) for 70,000 MCMC iterations, and average of effective sample sizes for the intensity evaluated at 51 grid points in $(0, T)$, under $J = 50$ for all three data examples.

performed on a laptop with an Intel i5-8250U 1.6GHz (8 CPUs) processor.

Increasing J increases the dimension of the parameter space. Table A.1 indicates the corresponding rate of increase in computing time, and decrease in mean ESS. Under $J = 50$ for the three synthetic data examples of Section 2.2, Table A.2 shows how computing time increases with the point pattern size. Note that there is no evident relationship between mean ESS and the point pattern size.

Figure A.1 provides an illustration of autocorrelation in the MCMC posterior samples. Plotted for the synthetic data of Sections 2.2.1 and 2.2.2 are averages of autocorrelation functions (at 48 lags) for the intensity function evaluated at the 51 equally-spaced grid points in $(0, T)$. The difference in the autocorrelations in the two

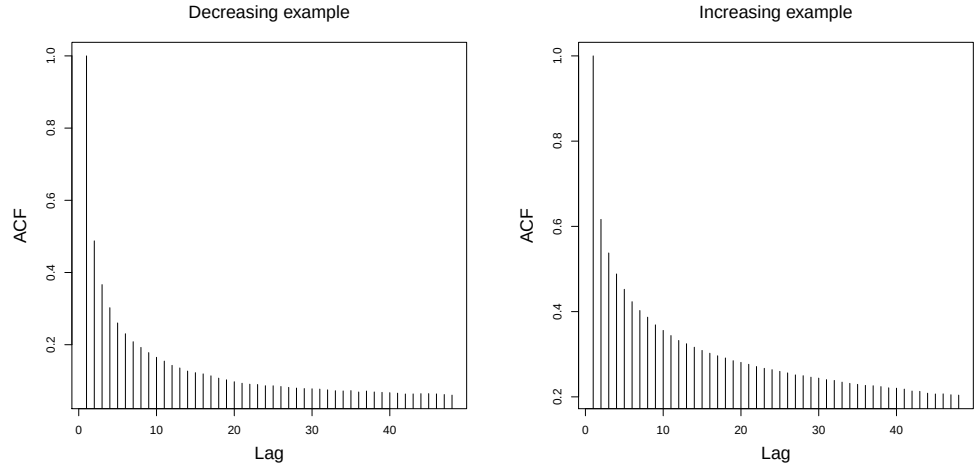

Figure A.1: Synthetic data from temporal NHPP with decreasing/increasing intensity (Sections 2.2.1 and 2.2.2). Average of autocorrelation functions for the intensity evaluated at 51 grid points in $(0, T)$.

panels of Figure A.1 is compatible with the corresponding mean ESS given in Table A.2.

Convergence and mixing of the MCMC algorithm can also be assessed graphically through trace plots of the intensity function evaluated at specific time points within the observation window. An example is given in Figure A.2 for the synthetic data of Section 2.2.1. The plots in Figure A.2 are representative of intensity trace plots obtained for all other data examples.

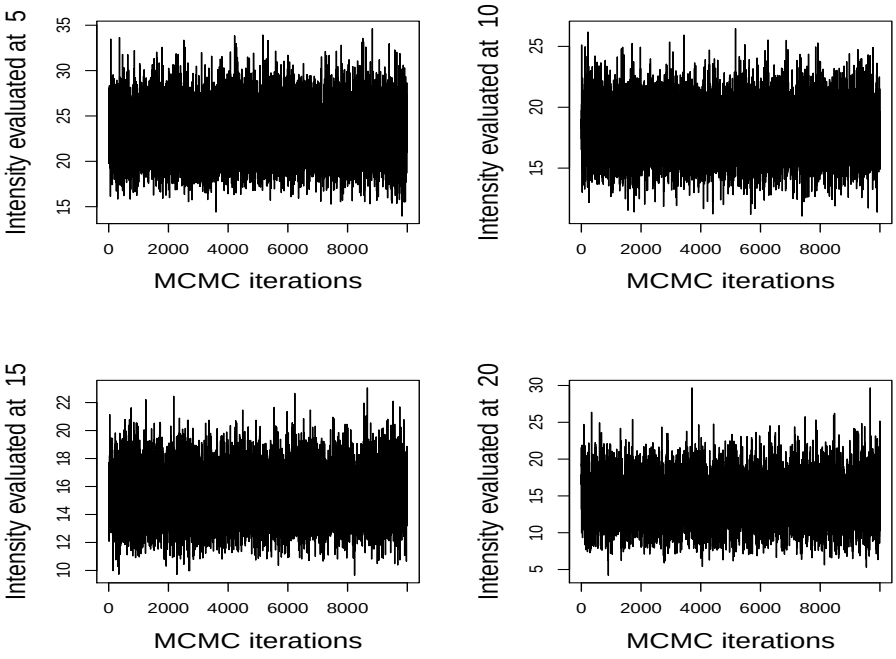

Figure A.2: Synthetic data from temporal NHPP with decreasing intensity (Section 2.2.1). Trace plots of posterior samples for the intensity function evaluated at time points $t = 5$, 10, 15, 20.

Appendix B

Comparison with log-Gaussian Cox process models for spatial NHPP intensity function

Here, we compare our model with Bayesian nonparametric models based on Gaussian process (GP) priors for logarithmic or logit transformations of the NHPP intensity (e.g., Møller et al., 1998; Brix and Diggle, 2001; Adams et al., 2009). Software in the form of **R** packages is available for spatial and spatio-temporal log-Gaussian Cox process (LGCP) models (Baddeley and Turner, 2005; Taylor et al., 2013), although its output is limited in terms of inferences that are of interest in our setting. Therefore, for the spatial NHPP case, this section focuses on graphical comparison of point estimates for the intensity surface in the context of the synthetic data considered in Section 2.3.3.

The LGCP model for spatial NHPP intensities can be expressed as $\lambda(\boldsymbol{s}) =$

$\tilde{\lambda}(\boldsymbol{s})\exp(g(\boldsymbol{s}))$, where function g is assigned a GP prior with isotropic correlation function, and variance σ^2. The correlation function includes parameter $\phi > 0$, which controls the rate at which correlation decreases with distance, and it may contain additional parameters (as in, e.g., the Matérn case). The mean of the GP prior is set to $-\sigma^2/2$, which implies $\mathrm{E}(\exp(g(\boldsymbol{s}))) = 1$, and thus $\mathrm{E}(\lambda(\boldsymbol{s})) = \tilde{\lambda}(\boldsymbol{s})$.

To obtain point estimates for the spatial intensity function under the LGCP model, one can use two **R** packages: **spatstat** (Baddeley and Turner, 2005) for parameter estimation, and **lgcp** (Taylor et al., 2013) for intensity estimation.

The function **lgcpPredictSpatial** of the **lgcp** package performs posterior inference for the intensity, obtaining posterior samples for a discretized version of function g through Metropolis-adjusted Langevin algorithms (Taylor and Diggle, 2014). However, to use **lgcpPredictSpatial**, values for σ^2 and ϕ, and function $\tilde{\lambda}(\boldsymbol{s})$ need to be provided. Since the **lgcp** package does not contain any functions for inference about the LGCP model hyperparameters, we use the **spatstat** package, which provides non-Bayesian estimates for σ^2, ϕ, and $\tilde{\lambda}(\boldsymbol{s})$. In particular, **spatstat** function **density.ppp** computes a nonparametric kernel intensity estimator, which can be used to estimate $\tilde{\lambda}(\boldsymbol{s})$. Moreover, function **lgcp.estpcf** yields estimates for σ^2 and ϕ through a nonparametric kernel estimator for the pair correlation function.

Evidently, the approach described above is not fully Bayesian. And, since it involves a two-stage estimation procedure, comparison with the Erlang mixture model in terms of uncertainty estimates is not particularly meaningful. We thus consider graphical comparison of the Erlang mixture model posterior mean intensity estimate with

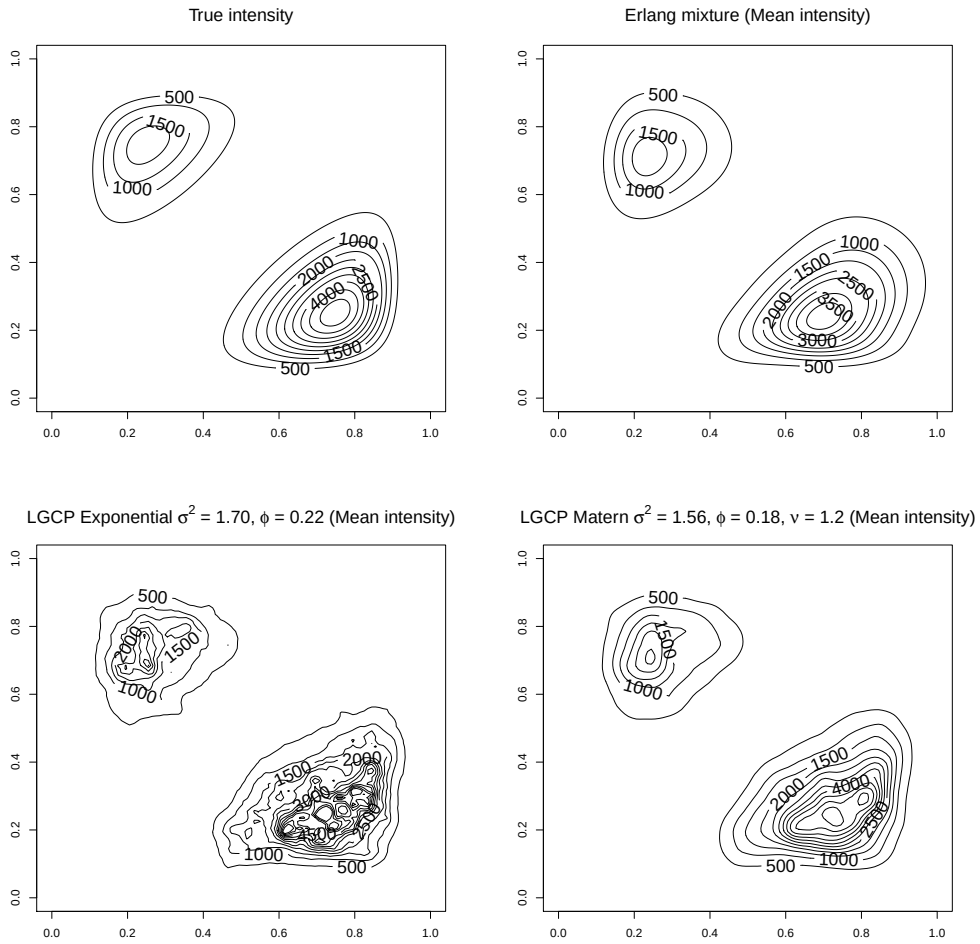

Figure B.1: Synthetic data from spatial NHPP with bimodal intensity defined through a two-component mixture of bivariate logit-normal densities (Section 2.3.3). The top left panel plots the true intensity function, the top right panel the posterior mean intensity under the Erlang mixture model, and the two bottom panels the point estimate for the intensity under the LGCP model with exponential and Matérn GP correlation function.

the point estimates obtained from the LGCP model under two different GP correlation functions. The comparison is based on the synthetic data example presented in Section 2.3.3, for which we have the true underlying intensity as the point of reference.

The top row of Figure B.1 shows the true intensity, and the posterior mean

intensity under the Erlang mixture model, obtained under the prior specification discussed in Section 2.3.3. Note that the Erlang mixture model estimates are fairly robust to the prior choice for the model hyperparameters, indeed, there is substantial learning for the hyperparameters under very dispersed priors (refer to Figure 7 of the paper).

The bottom row of Figure B.1 plots the LGCP model point estimates for the intensity, using either an exponential or a Matérn GP correlation function. We note that a further challenge with the use of LGCP models is that the **spatstat** package does not provide estimation for the additional parameters of correlation functions more general than the exponential, in particular, it does not provide an estimate for the smoothness parameter (ν) of the Matérn correlation function. As suggested in Taylor et al. (2013), we selected the value for ν by comparing graphically the pair correlation function estimator and the covariance function (with σ^2 and ϕ estimated).

The LGCP model estimates retrieve the bimodal global pattern of the true intensity, but with localized behavior (especially in the case of the exponential correlation function) that is not present in the underlying intensity surface. In addition to the challenge of obtaining full inference with appropriate uncertainty quantification, the sensitivity of the point estimates to the choice of the GP correlation function is evident.

Appendix C

MCMC posterior simulation for the immigrant Erlang mixture model

To explore the posterior distribution of parameters of the immigrant Erlang mixture model, we use Gibbs sampling. Here, we describe MCMC details to complete the posterior inference outlined in Section 3.2.1.1.

The branching structure latent variables, y_i, $i = 1, \ldots, n$, have the discrete uniform prior with the Dirac delta function for y_1. Then, the posterior full conditional distribution for the latent variables is derived as

$$\Pr(y_i = k|\boldsymbol{\nu}, \theta, \gamma, \alpha, \boldsymbol{t}) = \begin{cases} \dfrac{\sum_{l=1}^{L} \nu_l \mathrm{Ga}(t_i|l, \theta^{-1})}{\sum_{l=1}^{L} \nu_l \mathrm{Ga}(t_i|l, \theta^{-1}) + \sum_{r=1}^{i-1} \gamma \mathrm{Exp}(t_i - t_r|\alpha)}, & k = 0; \\ \\ \dfrac{\gamma \mathrm{Exp}(t_i - t_k|\alpha)}{\sum_{l=1}^{L} \nu_l \mathrm{Ga}(t_i|l, \theta^{-1}) + \sum_{r=1}^{i-1} \gamma \mathrm{Exp}(t_i - t_r|\alpha)}, & k = 1, \ldots, i-1. \end{cases}$$

We introduce latent variables $\{\xi_i : t_i \in I\}$ for the hierarchical model representation in (3.8). The posterior full conditional of the latent variables is a discrete

distribution on $\{1, \ldots, L\}$, such that $\Pr(\xi_i = l | \boldsymbol{\nu}, \theta, \boldsymbol{y}, \boldsymbol{t}) \propto \nu_l \mathrm{Ga}(t_i | l, \theta^{-1}),\ \ l = 1, \ldots, L$.

Let $n_l = |\{\xi_i = l : t_i \in I\}|$ for $l = 1, \ldots, L$. Under the $\mathrm{Ga}(c_0\theta/b, c_0)$ prior, induced by the gamma process prior, the posterior full conditional distribution of ν_l is given by $\mathrm{Ga}(n_l + c_0\theta/b, \int_0^T \mathrm{Ga}(u|l, \theta^{-1})du + c_0)$.

The branching ratio, γ, has a gamma conjugate prior, given by $\mathrm{Ga}(a_\gamma, b_\gamma)$. Denote by n_O the offspring cluster size, defined in Section 3.3.1. The posterior full conditional distribution is available in closed-form, $\mathrm{Ga}(n_O + a_\gamma, \sum_{t_i \in \boldsymbol{t}} \int_0^{T-t_i} \mathrm{Exp}(s|\alpha)ds + b_\gamma)$.

Finally, parameters θ and α and hyperparameters c_0 and b_{G_0} of the gamma process prior are updated with Metropolis-Hastings (M-H) algorithms, using log-normal proposal distributions.

Appendix D

MCMC posterior simulation for the offspring Erlang mixture model

Gibbs sampling plays a central role to explore the posterior distribution of model parameters. Latent variables of the branching structure, assigned a Dirac delta function and discrete uniform distributions, as in the immigrant Erlang mixture model, have the following posterior full conditional distribution

$$\Pr(y_i = k|\mu, \boldsymbol{\nu}, \theta, \boldsymbol{t}) = \begin{cases} \dfrac{\mu}{\mu + \sum_{r=1}^{i-1}\sum_{l=1}^{L}\nu_l \mathrm{Ga}(t_i - t_r|l, \theta^{-1})}, & k = 0; \\ \\ \dfrac{\sum_{l=1}^{L}\nu_l \mathrm{Ga}(t_i - t_k|l, \theta^{-1})}{\mu + \sum_{r=1}^{i-1}\sum_{l=1}^{L}\nu_l \mathrm{Ga}(t_i - t_r|l, \theta^{-1})}, & k = 1, \ldots, i-1. \end{cases}$$

Under the assumption of constant immigrant intensity, the $\mathrm{Exp}(a_\mu)$ prior provides the posterior full conditional distribution, $\mathrm{Ga}(n_I + 1, T + a_\mu)$, for μ, where n_I is the immigrant cluster size, defined in Section 3.3.1.

Latent variables $\{\xi_i : t_i \in O\}$, used for the hierarchical representation appear-

ing in (3.9), have the full conditional distribution in the form of the discrete distribution on $\{1, \ldots, L\}$, such that $\Pr(\xi_i = l | \boldsymbol{\nu}, \theta, \boldsymbol{y}, \boldsymbol{t}) \propto \nu_l \mathrm{Ga}(t_i | l, \theta^{-1})$ for $l = 1, \ldots, L$.

Under the gamma prior distribution, $\mathrm{Ga}(A_l, \eta)$, for the mixture weight ν_l, the posterior full conditional distribution is given in closed-form as follows $\mathrm{Ga}(\tilde{a}, \tilde{b})$, where $\tilde{a} = n_l + A_l$, $\tilde{b} = \int_0^T \mathrm{Ga}(u | l, \theta^{-1}) du + \eta$, and $n_l = |\{\xi_i = l : t_i \in I\}|$ for $l = 1, \ldots, L$.

Finally, we use M-H steps with log-normal proposal distributions to sample parameter θ and hyperparameter b_{F_0} of the gamma process prior.

Appendix E

MCMC posterior simulation for uniform-mixture-based models

Posterior inference method for the models is based on the blocked Gibbs sampler, involving additional updating steps for the branching structure, y_i, and parameters μ and γ.

With the priors of a Dirac delta function for y_1 and the discrete uniform distribution y_i, $i = 2, \ldots, n$, the posterior full conditional distribution for the latent variables is defined as

$$\Pr(y_i = k|\mu, \boldsymbol{\omega}, \boldsymbol{Z}, \gamma, \boldsymbol{t}) = \begin{cases} \dfrac{\mu}{\mu + \gamma \sum_{r=1}^{i-1} \sum_{l=1}^{L} \omega_l \frac{1}{Z_l} 1_{(0,Z_l)}(t_i - t_r)}, & k = 0; \\ \\ \dfrac{\gamma \sum_{l=1}^{L} \omega_l \frac{1}{Z_l} 1_{(0,Z_l)}(t_i - t_k)}{\mu + \gamma \sum_{r=1}^{i-1} \sum_{l=1}^{L} \omega_l \frac{1}{Z_l} 1_{(0,Z_l)}(t_i - t_r)}, & k = 1, \ldots, i-1 \end{cases}$$

The constant immigrant intensity assumption and the $\text{Exp}(a_\mu)$ prior for μ

yields the posterior full conditional distribution, $\text{Ga}(n_I + 1, T + a_\mu)$.

With the $\text{Ga}(a_\gamma, b_\gamma)$ prior for γ, we can derive the posterior full conditional distribution as $\text{Ga}(n_O + a_\gamma, \sum_{l=1}^{L} \omega_l K(Z_l) + b_\gamma)$, where n_O denotes the offspring cluster size, defined in Section 3.3.1.

The exponential hyperprior distribution, $\text{Exp}(a_\beta)$, for scale β of the inverse gamma prior for Z_l provides ready prior-to-posterior updating with the posterior full conditional distribution, $\text{Ga}(3L + 1, \sum_{l=1}^{L} Z_l^{-1} + a_\beta)$.

Latent variables $\{\xi_i : t_i \in O\}$ for the hierarchical representation of uniform-mixture-based models have the posterior full conditional distribution derived as $\Pr(\xi_i = l|\boldsymbol{\omega}, \boldsymbol{Z}, \boldsymbol{y}, \boldsymbol{t}) \propto \omega_l \frac{1}{Z_l} 1_{(0,Z_l)}(t_i - t_{y_i})$ for $l = 1, \ldots, L$.

We assign the $\text{IG}(3, \beta)$ prior to scale Z_l of the uniform mixtures, which provides the piecewise truncated inverse gamma distribution as the posterior full conditional distribution. Denote by ξ_s^* distinct values of latent variables $\{\xi_i : i \in O\}$ for $s = 1, \ldots, n_O^*$ with the size, n_O^*, of the distinct values. Posterior updating Z_l is conditioned on the result, whether $l = \xi_s^*$ for any s or not, as follows

- If $l \neq \xi_s^*$ for all s,

$$Z_l|\boldsymbol{\omega}, \boldsymbol{\xi}, \gamma, \beta, \boldsymbol{y}, \boldsymbol{t} \overset{ind.}{\sim} \begin{cases} \mathrm{IG}(Z_l|3, \gamma\omega_l(\sum_{j=1}^n (T-t_j)) + \beta), \\ \qquad Z_l \in (T-t_1, \infty) \text{ w/ prob. } \frac{c_0}{\sum_{k=0}^n c_k}; \\ \mathrm{IG}(Z_l|3, \gamma\omega_l(\sum_{j=r+1}^n (T-t_j)) + \beta), \\ \qquad Z_l \in (T-t_{r+1}, T-t_r], \text{ w/ prob. } \frac{c_r}{\sum_{k=0}^n c_k}; \\ \mathrm{IG}(Z_l|3, \beta), \ Z_l \in (0, T-t_n], \text{ w/ prob. } \frac{c_n}{\sum_{k=0}^n c_k} \end{cases}$$

$$c_0 = \frac{\beta^3}{\Gamma(3)} \frac{\Gamma(3)}{(\gamma\omega_l \sum_{k=1}^n (T-t_k) + \beta)^3} \Big(\int_{T-t_1}^{\infty} \mathrm{IG}(s|3, \gamma\omega_l \sum_{k=1}^n (T-t_k) + \beta) ds \Big)$$

$$c_r = \frac{\beta^3}{\Gamma(3)} \frac{\Gamma(3)}{(\gamma\omega_l \sum_{k=r+1}^n (T-t_k) + \beta)^3} \times \Big(\int_{T-t_{r+1}}^{T-t_r} \mathrm{IG}(s|3, \gamma\omega_l \sum_{k=r+1}^n (T-t_k) + \beta) ds \Big) \exp\{-\gamma\omega_l r\}, \quad r = 1, \ldots, n-1$$

$$c_n = \frac{\beta^3}{\Gamma(3)} \frac{\Gamma(3)}{\beta^3} \Big(\int_0^{T-t_n} \mathrm{IG}(s|3, \beta) ds \Big) \exp\{-\gamma\omega_l n\}$$

- If $l = \xi_s^*$ for any s, $s = 1, \ldots, n_O^*$,

$$Z_l|\boldsymbol{\omega}, \boldsymbol{\xi}, \gamma, \beta, \boldsymbol{y}, \boldsymbol{t} \overset{ind.}{\sim} \begin{cases} \mathrm{IG}(Z_l|3 + n_l, \gamma\omega_l(\sum_{j=1}^n (T-t_j)) + \beta), \\ \qquad Z_l \in (T-t_1, \infty) \text{ w/ prob. } \frac{c_0}{\sum_{k=0}^n c_k}; \\ \mathrm{IG}(Z_l|3 + n_l, \gamma\omega_l(\sum_{j=r+1}^n (T-t_j)) + \beta), \\ \qquad Z_l \in (T-t_{r+1}, T-t_r] \text{ w/ prob. } \frac{c_r}{\sum_{k=0}^n c_k}; \\ \mathrm{IG}(Z_l|3 + n_l, \beta), \ Z_l \in (0, T-t_n], \text{ w/ prob. } \frac{c_n}{\sum_{k=0}^n c_k} \end{cases}$$

$$c_0 = \frac{\beta^3}{\Gamma(3)} \frac{\Gamma(3+n_l)}{(\gamma\omega_l \sum_{k=1}^{n}(T-t_k)+\beta)^{3+n_l}} \Big(\int_{b_0}^{\infty} \mathrm{IG}(s|3+n_l, \gamma\omega_l \sum_{k=1}^{n}(T-t_k)+\beta) ds \Big)$$

$$c_r = \frac{\beta^3}{\Gamma(3)} \frac{\Gamma(3+n_l)}{(\gamma\omega_l \sum_{k=r+1}^{n}(T-t_k)+\beta)^{3+n_l}}$$
$$\times \Big(\int_{b_r}^{T-t_r} \mathrm{IG}(s|3+n_l, \gamma\omega_l \sum_{k=r+1}^{n}(T-t_k)+\beta) ds \Big) \exp\{-\gamma\omega_l r\}, \quad r = 1, \ldots, n-1$$

$$c_n = \frac{\beta^3}{\Gamma(3)} \frac{\Gamma(3+n_l)}{\beta^{3+n_l}} \Big(\int_{b_n}^{T-t_n} \mathrm{IG}(s|3+n_l, \beta) ds \Big) \exp\{-\gamma\omega_l n\}$$

where $n_l = |\{\xi_j : \xi_j = l, \ \ t_j \in O\}|$, $b_0 = T - t_1$, $b_r = \min(T-t_r, \max(T-t_{r+1}, \max(t_j - t_{y_j})))$, and $b_n = \min(T-t_n, \max(t_j - t_{y_j}))$.

Under the Dirichlet process prior for the mixing distribution, the posterior full conditional distribution for ω_l is proportional to $\exp\{-\sum_{l=1}^{L} \gamma\omega_l K(Z_l)\} \prod_{l=1}^{L} \omega_l^{n_l} \prod_{l=1}^{L-1}(1-\nu_l)^{\alpha-1}$. We can represent the distribution through latent variables $\nu_1, \ldots, \nu_{L-1}$ as follows

$$p(\boldsymbol{\nu}|\boldsymbol{\xi}, \boldsymbol{Z}, \gamma, \alpha) \propto \exp\{-\gamma(\nu_1 K(Z_1) + \sum_{l=2}^{L-1} \nu_l (\prod_{r=1}^{l-1}(1-\nu_r)) K(Z_l) + \prod_{l=1}^{L-1}(1-\nu_l) K(Z_L))\}$$
$$\times \nu_1^{n_1} \Big[\prod_{l=2}^{L-1} \Big(\nu_l (\prod_{r=1}^{l-1}(1-\nu_r)) \Big)^{n_l} \Big] \Big(\prod_{l=1}^{L-1}(1-\nu_l) \Big)^{n_L} \prod_{l=1}^{L-1}(1-\nu_l)^{\alpha-1}$$

We draw the posterior sample of ν_l from the distribution using slice sampling. It allows us to sample ν_l through well-known distributions, the beta and exponential distributions.

The precision parameter, α, of the Dirichlet process, under the $\mathrm{Ga}(a_\alpha, b_\alpha)$ prior, has the full conditional distribution, $\mathrm{Ga}(a_\alpha + L - 1, b_\alpha - \sum_{l=1}^{L-1} \log(1-\nu_l))$.

The geometric weights prior model defines the mixture weight using a single latent variable ζ. The posterior full conditional distribution for ζ is proportional to $\exp\{-\sum_{l=1}^{L-1} \gamma K(Z_l)(1-\zeta)\zeta^{l-1} - \gamma K(Z_L)\zeta^{L-1}\} \zeta^{\sum_{l=1}^{L}(l-1)n_l + a_\zeta - 1} (1-\zeta)^{\sum_{l=1}^{L-1} n_l + b_\zeta - 1}$. We use the M-H algorithm to sample ζ with a Beta proposal distribution to which a tuning parameter, η, is added for better mixing.

Appendix F

Derivation of the asymptotic expected offspring density function under the offspring Erlang mixture model

Under the offspring Erlang mixture model, taking the expectation of the excitation function $h(x)$ over the weight yields the expected excitation function $\mathrm{E}(h(x)) = \sum_{l=1}^{L} A_l \mathrm{ga}(x|l, \theta^{-1})/\eta$ and thus the expected offspring density function

$$\begin{aligned}\mathrm{E}(g(x)) &= \sum_{l=1}^{L} A_l \mathrm{ga}(x|l, \theta^{-1})/\alpha_0 \\ &= \sum_{l=1}^{L-1} [F_0(l\theta) - F_0((l-1)\theta)]\mathrm{ga}(x|l, \theta^{-1}) + [1 - F_0((L-1)\theta)]\mathrm{ga}(x|L, \theta^{-1}),\end{aligned}$$

where $A_l = \alpha_0[F_0(l\theta) - F_0((l-1)\theta)]$, $l = 1, \ldots, L-1$ and $A_L = \alpha_0[1 - F_0((L-1)\theta)]$. With an exponential distribution for F_0 such that $F_0(x) = 1 - \exp\{-x/b_{F_0}\}$, the expected

offspring density function can be expressed as

$$\begin{aligned}
\mathrm{E}(g(x)) &= \sum_{l=1}^{L-1}[\exp\{-l\theta/b_{F_0}\}(\exp\{\theta/b_{F_0}\}-1)]\mathrm{ga}(x|l,\theta^{-1}) \\
&\quad + [\exp\{-L\theta/b_{F_0}\}\exp\{\theta/b_{F_0}\}]\mathrm{ga}(x|L,\theta^{-1}) \\
&= \sum_{l=1}^{L}[\exp\{-l\theta/b_{F_0}\}\exp\{\theta/b_{F_0}\}]\mathrm{ga}(x|l,\theta^{-1}) - \sum_{l=1}^{L-1}[\exp\{-l\theta/b_{F_0}\}]\mathrm{ga}(x|l,\theta^{-1}).
\end{aligned}$$

Therefore, the asymptotic density function, as L tends infinity, is derived as

$$\begin{aligned}
\lim_{L\to\infty}\mathrm{E}(g(x)) &= \theta^{-1}\exp\{x/\theta(\exp\{-\theta/b_{F_0}\}-1)\} \\
&\quad - \theta^{-1}\exp\{-\theta/b_{F_0}\}\exp\{x/\theta(\exp\{-\theta/b_{F_0}\}-1)\} \\
&= [(1-\exp\{-\theta/b_{F_0}\})/\theta]\exp\{x/\theta(\exp\{-\theta/b_{F_0}\}-1)\},
\end{aligned}$$

that is, the exponential density function with rate $(1-\exp\{-\theta/b_{F_0}\})/\theta$. Denote by X the random distance between an offspring point and its parent. The distance X is distributed according to the offspring density function of HPs. Therefore, the mean distance under the asymptotic density function is $\mathrm{E}(X) = \theta/(1-\exp\{-\theta/b_{F_0}\})$.

Appendix G

Derivation of the mean distance for uniform-mixture-based models

Let F be a random mixing distribution with a DP / GW prior with mean F_0. The expected offspring density over F is given by

$$\mathrm{E}(g(x)) = \mathrm{E}\Big(\int \theta^{-1} \mathbf{1}_{[0,\theta)}(x) dF(\theta) \Big) = \int \theta^{-1} \mathbf{1}_{[0,\theta)}(x) dF_0(\theta).$$

Denote by X the random distance between an offspring point and its parent. With an inverse gamma distribution $IG(\theta|3,\beta)$ with shape 3 and scale β for the centering function F_0, the mean distance is derived as

$$\begin{aligned}
\mathrm{E}(X) &= \int x \Big(\int \theta^{-1} \mathbf{1}_{[0,\theta)}(x) \mathrm{IG}(\theta|3,\beta) d\theta \Big) dx = \int \theta^{-1} \Big(\int x \mathbf{1}_{[0,\theta)}(x) dx \Big) \mathrm{IG}(\theta|3,\beta) d\theta \\
&= \int \theta^{-1} (\theta^2/2) \beta^3 / \gamma(3) \theta^{-(3+1)} \exp\{-\beta/\theta\} d\theta = \beta^3/\gamma(3)/2 \int \theta^{-(3+1)+1} \exp\{-\beta/\theta\} d\theta \\
&= \beta^3/\gamma(3)/2 \times \gamma(2)/\beta^2 = \beta/4.
\end{aligned}$$