UNIVERSITY OF CALIFORNIA
SANTA CRUZ

BAYESIAN NONPARAMETRIC MODELING FOR SPATIAL NONHOMOGENEOUS AND CLUSTERED POINT PATTERN DATA

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

DOCTOR OF PHILOSOPHY

in

STATISTICAL SCIENCE

by

Chunyi Zhao

September 2022

The Dissertation of Chunyi Zhao
is approved:

Professor Athanasios Kottas, Chair

Professor Bruno Sansó

Professor Juhee Lee

Peter Biehl
Vice Provost and Dean of Graduate Studies

Contents

List of Figures

List of Tables

Abstract

Bayesian nonparametric modeling for spatial nonhomogeneous and clustered point pattern data

by

Chunyi Zhao

This work provides a Bayesian nonparametric modeling framework for spatial point processes to account for the irregular domain over which the resulting point pattern occurs in the model formulation while balancing flexible inference with efficient implementation. We start with models for the spatial Poisson process, which assumes independence among points given the number of occurrences, and progress to models for Hawkes processes over space and space-time that capture the self-triggering behaviors and relax the independence assumption. We develop nonparametric Bayesian modeling approaches for Poisson processes using weighted combinations of structured beta densities to represent the point process intensity function. For a regular spatial domain, i.e., the unit square, the model construction implies a Bernstein-Dirichlet prior for the Poisson process density, which supports flexible inference about point process functionals with theoretical guarantees. The key contribution is two classes of flexible and computationally efficient models for spatial Poisson process intensities over irregular domains. We address the choice or estimation of the number of beta basis densities and develop methods for prior specification. For the spatial Hawkes process, we develop a semi-parametric modeling approach, leveraging its clustering representation defined as the superposition of an immigrant Poisson process and several offspring Poisson clustering processes centered on parent points

generated by earlier generations. We apply the model for the Poisson process developed earlier to the latent immigrant Poisson process and complete the hierarchical model for the spatial Hawkes process with parametric formulations for the offspring Poisson processes and a model for the latent branching structure that specifies lineage among points. Finally, we develop a nonparametric model for the spatial offspring Poisson process under the assumption of spatial isotropy, which reduces modeling for the spatial offspring density to that for the spatial offspring-parent distance density. Such construction allows the model to be free from the implied tail behavior constraints imposed by existing parametric options for the offspring density kernel. We incorporate such a method to model for space-time Hawkes processes. For all methods developed in the dissertation, we design posterior simulation algorithms for full inference on key point process functionals and model checking techniques to examine the model fit. Model capacity is demonstrated with numerous simulation studies, and we focus on real data examples using crime point patterns from the city of Boston.

To my family and friends who supported me through the highs and lows

Acknowledgments

I want to extend my gratitude to the people who shaped my statistical education, contributed to the production of this dissertation, and enlightened me in my journey as a statistician.

I am deeply grateful to be advised by Professor Athanasios Kottas. He opened the door of the Bayesian nonparametric world for me, a world I once thought was inaccessible and have been truly privileged to explore for the past five years. Thanasis taught me more than just statistics. Most critically, he gave me a pragmatic attitude toward complicated mathematics. It helped me overcome the imposter syndrome of someone lacking a mathematical background and gave me the courage to learn whatever was necessary for my research. He also taught me humility, so I was not distracted by my ego in pursuing knowledge. I am deeply touched by his compassion toward his colleague, his advisees, and every undergraduate student in his class. He cares about sharing knowledge with anyone interested with a level of clarity that I dream of achieving. He cares about people around him and provides advice, counseling, and encouragement when they are most needed. I will be forever inspired by his work ethic, faith in the quality of work, and dedication to his students.

My path to a Ph.D. in statistics is by no means standard, and many people have contributed to my growth to make it possible. As a sophomore in college, I stumbled into a probability class and became intrigued by the task of quantifying uncertainty ever since. I am forever grateful to Professor Jack O'Brien at Bowdoin College, who gave me a critical introduction to statistics and sparked my interest in exploring the world as a statistician. I am grateful to the professors at UCSC who provided me with a solid statistical education. I especially want to thank

professor Bruno Sansó for his insights and inspiration as a Bayesian statistician specializing in spatial applications, Juhee Lee for her friendship and supervision when I served as the graduate student representative and Raquel Prado for her insights in academic and career development. I am indebted to both Juhee and Bruno, who served as readers on my thesis committee, for their comments and insights that improved this thesis. My achievement is rooted in the knowledge I acquired from learning from all of these professors, and I cannot thank them enough.

I was fortunate to be surrounded by a group of capable and loving friends in the department. Arthur Lui introduced me to the Julia language and provided substantial technical and computational support for my software development. Xiaotian Zheng provided critical advice on the third project and has served as an admirable colleague and a great friend. I thank Isabelle Grenier, Dan Kirsner, Dan Spencer, Kurtis Shuler, Hyotae Kim, Mattew Heiner, Jizhou Kang, Zach Horton, and Yunzhe Li for their support and constructive feedback.

During my Ph.D., I have leaned on people closest to me, and everything I have achieved is only possible through their love. I want to thank my partner, Apoorva Lal, a fellow researcher and Ph.D. student in econometrics, for broadening my statistical horizons and providing unwavering technical and emotional support. And finally, I want to thank my parents for always believing in me and supporting all of my decisions.

Chapter 1

Introduction

Spatial point pattern data is a special case of spatial data for which both the number and locations of events are random. This feature differentiates spatial point pattern data from spatially referenced data where locations, whether in the form of coordinates or areal units, are considered fixed as an index for the corresponding response. Modeling spatial point pattern data from a probabilistic modeling point of view often treats the point pattern as a realization of some underlying point process, whose stochastic mechanism specifies both the total number of events in a subset of the domain and the locations of these events. Illian et al. (2008) provides a good introduction to modeling for various spatial point processes. Specifically, modeling is achieved via the construction of the key functionals that control the point processes, namely the *intensity functions*. The intensity functions are non-negative, locally integrable, and can be informally interpreted as providing the rate of occurrence in an infinitesimal local neighborhood.

When the timestamps of the events are also available, the space-time point pattern $\{x_i, y_i, t_i\}$ is treated as a realization from some space-time point process. Diggle (2017) cov-

ers major methodologies in testing, modeling and estimation approaches for space-time point patterns and related application in biological science and ecology. The chronological order of events allows additional assumptions that enable historical events to impact the current rate via a more general form for the intensity function. When assumed to take an additive form, such intensity function, referred to as the *conditional intensity* function, defines a type of point process that captures so-called *self-exciting* or *self-triggering* behaviors, meaning that the current rate of the event is higher as a result of earlier events.

Modeling the intensity functions can be done via a *parametric* or *nonparametric* approach, where the main difference is whether to assume a particular functional form for the intensities or allow the data to inform the functional form by leveraging flexible function structures. The parametric model can be viewed as a special case for the nonparametric model, as the latter provides structures that can emulate the pre-specified functional form of the former, given sufficient support from the data. Notice that both approaches require a finite number of parameters that specify the intensities, despite the unfortunate misnomer "nonparametric" suggesting no parameters or the widespread belief that the nonparametric methods estimate infinitely many parameters. Learning is achieved by estimating and providing uncertainty quantification for the intensity parameters either by optimizing a certain loss function, here the likelihood function, or a Bayesian approach that learns the distributions of these parameters conditional on observed data.

This work takes a Bayesian approach to achieve inference and prediction. The appeal of the Bayesian approach is that the posterior distribution, i.e., the joint conditional distribution of parameters given data, provides both point estimation and uncertainty quantification in

one step. The posterior distribution often does not have closed-form expression, especially for complex models. Inference is achieved by summarizing samples from the posterior distribution, obtained using the simulation-based algorithm Markov Chain Monte Carlo (MCMC). Such algorithm sequentially traverses the posterior parameter space in a manner that, once convergent, the output of each iteration, i.e., a vector of parameter values, is a draw from the posterior distribution. Prediction is done via simulating from the model for each posterior sample to form the posterior prediction sample, which is then summarized to provide point and interval estimation. The challenge is getting the posterior sample. The MCMC algorithm is often more computationally expensive since the convergence of the chain requires a large number of iterations. The computation demand for Bayesian inference is substantial to the point that it has become a crucial factor in the model formulation and implementation.

The contribution of this work is a Bayesian nonparametric modeling framework to achieve flexible inference for important classes of spatial point processes over the irregular domain. In this chapter, we first introduce the real data example and the substantive questions. We then present the motivations and objectives for each point process model in this dissertation.

We use the crime point pattern data in Boston as the primary real data example in this dissertation. The raw data contains the spatial and time information for over 30 types of crimes from 2017 to 2018 in Boston (Jain, 2018). Fig.1.1 shows the point pattern for vandalism in certain weeks from April 2017 to June 2017. Each point represents an occurrence of vandalism at that location, and the blue shaded area is the city boundary of Boston, which has a complicated shape. We want to understand which neighborhoods in the city have higher crime rate and predict the number of occurrences in a local neighbourhood and a certain time interval. To

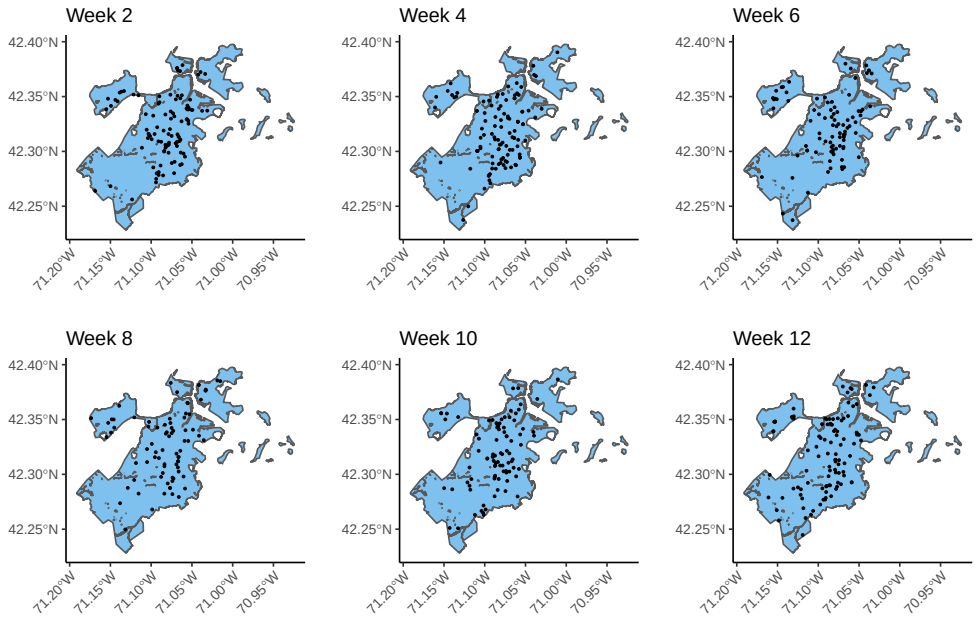


Figure 1.1: Point patterns for Vandalism in certain weeks in the second quarter of 2017 in Boston.

achieve these goals, we develop modeling methodologies for a series of spatial and space-time point processes with increasingly more general assumptions that suit the application better and more complex structures.

A common issue for point pattern modeling is accounting for the effect of the compact observation domain on estimating the intensity. The observation domain $\mathcal{D}$ is the region, along with the time window if event times are recorded, over which the events occur. Spatially, these domains tend to have highly irregular boundaries, just like that of the Boston city. The irregular domain poses challenges for both model formulation and computation. A recurring theme of this dissertation is how to address this issue by constructing models that balance the inference flexibility and computation efficiency given the constraints posed by the irregular domain and

the computation demands of a Bayesian nonparametric approach.

We begin with a modeling framework for the Poisson point process. Theoretical study of Poisson point process can be found in Cressie (1993) and Daley and Vere-Jones (2008), among other reference. In particular, Kingman (1992) provides a more focused study on the properties of Poisson process in one or more dimensions. The conditional intensity of a temporal Poisson process is equivalent to its intensity function, since the Poisson assumption implies independence among the event timestamps given the total number of occurrences. The spatial Poisson process extends such assumption to the two-dimensional space, where the intensity function specifies both the number of events in a compact subset of the observation domain via a Poisson distribution, and the locations of these events via the Poisson process density function $f(\boldsymbol{s}) = \lambda(\boldsymbol{s}) / \int_{\mathcal{D}} \lambda(\boldsymbol{s}) \mathrm{d}\boldsymbol{s}$. The irregular domain $\mathcal{D}$ serves as the support for such Poisson density function. We seek to develop a Bayesian nonparametric prior for the intensity function that respects the shape of $\mathcal{D}$, meaning that the implied prior model for the Poisson density function is a proper density over $\mathcal{D}$. More importantly, such intensity prior model avoids the computation for the normalizing constant, which involves integrating the intensity over $\mathcal{D}$.

The key to achieving these goals is our proposed model for the intensity function as a weighted combination of fixed Beta densities that function like basis functions, and the structured Gamma prior distributions for the random weights. Through the connection between the Poisson density function and the intensity function, our prior for the intensity function suggests a Bernstein-Dirichlet Prior for the Poisson density function, whose theoretical properties support recovery of various functional forms in the posterior. We prefer to work with the prior model for the intensity instead of the density function since the construction leads to a closed-

form expression for the total intensity, which appears in the Poisson process likelihood, and subsequently resulting in conjugate updates for the weights parameters in the posterior simulation.

In this sense, we achieve computational efficiency via fast and easy parameter updates in the MCMC algorithm while maintaining a connection to a rich and flexible Bayesian nonparametric model on the density function. Chapter 2 develops this core idea of intensity modeling for temporal and spatial Poisson processes. We demonstrate how the model handles the irregular spatial domain in its original form (without spatial approximation such as taking the convex hull), without adding too much computational overhead to the algorithm, using synthetic and real data examples.

We apply the spatial Poisson process model to the crime point pattern under the category "Vandalism" in the second quarter of 2017, under the assumption that the point pattern of the same crime in a short time period can be viewed as a realization from a Poisson process. Such a assumption can be questioned easily, as criminal activities are known to have a self-exciting nature, i.e., more crimes tend to happen at locations of previous crimes. This phenomenon inspires us to model point processes that account for self-exciting behavior in the form of allowing spatial clustering among points.

Next, we develop a class of models for spatial Hawkes processes. The Hawkes process, most commonly used as an example of a self-exciting process, is specified by its conditional intensity function that consists of a *background intensity* and a *triggering function*

(Hawkes, 1971). More specifically, the conditional intensity function is defined as

$$\lambda(t \mid \mathcal{H}_t) = \mu(t) + \sum_{i:t_i<t} g(t - t_i)$$

where $\mathcal{H}_t$ denotes the history of events up to time t, $\mu(\cdot)$ is the background intensity function, and $g(\cdot)$ is the triggering function. The triggering function depends on events previous to the current event at t.

The Hawkes process has an equivalent clustering representation (Hawkes and Oakes, 1974), where additional branching structure among the points suggests that the point pattern can be thought of as a realization from a sequence of recursive Poisson cluster processes. Specifically, a Poisson process controlled by the background intensity $\mu(\cdot)$ generates *immigrant points*; subsequent Poisson processes controlled by the triggering function $g(\cdot)$ spawn points centered on the *immigrant points* and their *offspring points*. To model the Hawkes process hierarchically, we model for the immigrant Poisson intensity, the offspring Poisson intensity, and the latent branching structure.

The spatial Hawkes (SH) process is well-defined following the clustering representation as a superposition of several General Shot Noise Cox processes (GSNCPs) with a Poisson immigrant process (Møller and Torrisi, 2005), despite of the lack of natural order in space to imply the order of events. Additional assumptions allow us to model the SH process as the superposition of a Poisson immigrant process and a sequence of offspring Poisson processes centered on parent points identified through the more general latent branching structure for points in space.

We leverage the model for the spatial Poisson process in Chapter 2 as the model for

the immigrant Poisson process intensity, and complete the hierarchical model for the spatial Hawkes process with parametric models for the Poisson offspring processes and a model for the branching structure encoded as a set of trees. Chapter 3 develops a Bayesian modeling framework for a class of SH processes categorized by the immigrant process assumption (homogeneous vs. nonhomogeneous), the domain assumption (irregular vs. unit-square), and the parametric form of the offspring intensity function. The irregular domain serves as both the support for the immigrant and offspring intensities, which poses challenges for the truncation applied to the offspring Poisson density. We develop an efficient Monte Carlo routine that recycles and reuses random samples and caches the computed normalizing constants to gain significant performance improvement. We develop model checking techniques for both the first-order model fit, via predictive residuals over Voronoi tessellation, and second-order model fit, via predictive Ripley's K functions. Finally, we demonstrate the model's capacity using synthetic data.

We revisit the Boston crime data example and apply the SH model with a nonhomogeneous immigrant Poisson process and a bivariate Gaussian offspring kernel allowing spatial skewness to the same vandalism point pattern data, modifying the irregular domain to be the convex hull of the city boundary to speed up computation. We discover a significant amount of self-triggering effect reflected by the *branching ratio*, a parameter that controls the average number of offspring generated in a cluster, far greater than 0. To confirm our discovery, we implement a model for the space-time Hawkes (STH) process with the same formulation for the immigrant and offspring spatial intensity, adding additional assumptions for a time-homogeneous immigrant temporal intensity and an exponentially decaying offspring temporal

intensity. The timestamps facilitate the inference of the branching structure with additional information on the chronological order, which is latent for the space-only process. The inference for the branching ratio under the STH process also suggests significant self-triggering. The inferred immigrant intensities under models for the SH and STH process flag similar regions in Boston to have higher rates of vandalism activities.

We develop models for STH processes focusing on a nonparametric offspring spatial intensity formulation. Reinhart (2018) provides a review of self-exciting spatial-temporal point processes and their applications. The STH process is defined via the space-time conditional intensity:

$$\lambda(x, y, t \mid \mathcal{H}_t) = \mu(x, y, t) + \sum_{i:t_i<t} g(x - x_i, y - y_i, t - t_i)$$

where $\mathcal{H}_t$ denotes the history of events including both space and time information up to time t. The background intensity $\mu(x, y, t)$ controls the rate of occurrence in space and time and is often assumed to be homogeneous in time. The triggering function is often factored into the product of time intensity $g_t(t - t_i)$ and a spatial intensity $g_s(x - x_i, y - y_i)$ as the result of the separability assumption, which suggests that the temporal and spatial triggering effects are independent. Popular parametric methods for $g_s(x - x_i, y - y_i)$ factor it into the product of a total intensity γ_s and a spatial density $f_s(x, y)$ and choose the parametric form of $f_s(x, y)$ between the bivariate Gaussian distribution and the Power Law distribution, which differ on how fast the density function approaches 0 in the tail. In most applications, the two parametric forms are both applied, and extensive literature focuses on comparing the two. A nonparametric formulation for $g_s(x - x_i, y - y_i)$ can capture both shapes under the parametric families and provide more flexible inference without the need to choose between parametric forms.

To achieve a nonparametric formulation, we make an additional assumption of spatial isotropy, an assumption taken by most commonly used parametric families, and reduce modeling a spatial density to modeling a univariate distance density. The distance in our model is that between offsprings and their parent points. The irregular domain here implicitly defines the support of such distance density. We place a scaled Bernstein-Dirichlet prior on the spatial distance density to capture varying tail behavior, no longer confined to the choice between Gaussian and Power Law. Chapter 4 develops models for STH processes under the assumption of offspring spatial isotropy and presents both a parametric formulation under the name ParSTH and a nonparametric formulation under the name NonparSTH for the offspring spatial process, while leveraging the model for nonhomogeneous Poisson process developed in Chapter 2 to model the immigrant Process. We apply both models to the Vandalism data and discover that the NonparSTH model favors a heavier tail for the spatial distance density compared to the ParSTH model.

Chapter 2

Bayesian nonparametric modeling for spatial Poisson processes

We develop nonparametric Bayesian modeling approaches for Poisson processes, using weighted combinations of structured beta densities to represent the point process intensity function. For a regular spatial domain, such as the unit square, the model construction implies a Bernstein-Dirichlet prior for the Poisson process density, which supports general inference for point process functionals. The key contribution of the methodology is two classes of flexible and computationally efficient models for spatial Poisson process intensities over irregular domains. We address the choice or estimation of the number of beta basis densities, and develop methods for prior specification and posterior simulation for full inference about functionals of the point process. The methodology is illustrated with both synthetic and real data sets.

2.1 Introduction

There has been an increasing interest in extracting information from locations in spatial data. For spatial point patterns, both the number and the locations of points are random. Point pattern data is modelled as a realization, within compact domain $\mathcal{D}$, of a point process whose finite dimensional distribution defines the stochastic mechanism for the number and locations of the points. Independent increments along with a Poisson distributional assumption define the Poisson process. A homogeneous Poisson process is equivalent to complete spatial randomness, that is, the point pattern generated is independently and identically uniformly distributed over $\mathcal{D}$. The practically relevant version is the non-homogeneous Poisson process (NHPP), which allows the point process intensity to differ by location. The NHPP is characterized by a non-negative, locally integrable intensity function $\lambda(s)$, such that: for any bounded subset $\mathcal{B}$ of the domain, the number of points in $\mathcal{B}$, $N(\mathcal{B})$, is Poisson$(\int_{\mathcal{B}} \lambda(s)\mathrm{d}s)$ distributed; and, given $N(\mathcal{B})$, the point locations within $\mathcal{B}$ are independent and identically distributed with density $\lambda(s)/\int_{\mathcal{B}} \lambda(u)\mathrm{d}u$. Therefore, the NHPP likelihood corresponding to point pattern $\{s_1, ..., s_n\}$, observed in compact domain $\mathcal{D}$, can be expressed as:

$$p(\{s_1, ..., s_n\}; \lambda(s)) \propto \exp\left(-\int_{\mathcal{D}} \lambda(s)\,\mathrm{d}s\right) \prod_{i=1}^{n} \lambda(s_i) \tag{2.1}$$

where $n \equiv N(\mathcal{D})$. We consider the more common settings where $\mathcal{D} \subset \mathbb{R}$ or $\mathcal{D} \subset \mathbb{R}^2$. We place particular emphasis on spatial NHPPs, and more specifically on building flexible, computationally tractable models for spatial intensities defined over domains with irregular shapes.

Theoretical study of NHPPs can be found in Cressie (1993) and Daley and Vere-Jones (2008), among other references. Diggle (2003) provides background on likelihood and

classical nonparametric inference for spatial NHPPs. Moller and Waagepetersen (2003) discuss simulation-based inference for point processes. Regarding model-based methods for NHPPs, Gelfand and Schliep (2018) categorize the main approaches in two general directions: modeling the trend surface for the intensity function $\lambda(s)$; and, factorizing the intensity function into the total intensity, $\Lambda = \int_{\mathcal{D}} \lambda(s)\mathrm{d}s$, and the NHPP density $f(s) = \lambda(s)/\Lambda$, and modeling each separately.

The early Bayesian nonparametric approaches fall under the first category, focusing on modeling temporal NHPP cumulative intensity functions, $\int_0^t \lambda(s)\mathrm{d}s$, with gamma, beta or general Lévy process priors (Lo, 1982, 1992). The next stage in this line of research involves mixture models for NHPP intensities built from non-negative kernels convolved with weighted gamma processes (Lo and Weng, 1989; Wolpert and Ickstadt, 1998; Ishwaran and James, 2004; Kang et al., 2014). Also in this direction are modeling approaches based on log-Gaussian Cox processes (Moller et al., 1998) under which the logarithm of the intensity function is a realization of a Gaussian process. Adams et al. (2009) proposed a related approach based on a logistic instead of logarithmic transformation to link the Gaussian process with the model for the intensity function. Modeling directly the intensity function $\lambda(s)$ brings computational challenges for full posterior inference due to the likelihood normalizing term, $\exp(-\int_{\mathcal{D}} \lambda(s)\,\mathrm{d}s)$, especially under methods based on Gaussian process priors. Such challenges have been addressed through approximations of the stochastic integral (Brix and Moller, 2001; Brix and Diggle, 2001), data augmentation (Adams et al., 2009), and discretization of the observation domain $\mathcal{D}$ (Illian et al., 2012).

Under the second direction, Kottas (2006) and Kottas and Sansó (2007) proposed an

approach that connects the NHPP intensity function with the density function supported on the observation domain, and models the NHPP density with Dirichlet process mixture priors for density estimation. Taddy and Kottas (2012) extend this modeling approach to marked Poisson processes, and Taddy (2010), Kottas et al. (2012), Xiao et al. (2015) and Rodriguez et al. (2017) develop hierarchical and dynamic models for NHPPs in the context of specific applications. This modeling approach enables an inference framework that builds from well established methods for Dirichlet process mixtures, avoiding the computational challenges due to the NHPP likelihood normalizing component. However, it relies on a potentially restrictive prior structure that models separately the NHPP density and the total intensity over the observation domain.

Inference methods for irregular domain spatial point process intensities have received limited attention in the Bayesian nonparametrics literature. We are only aware of the log-Gaussian Cox process approach of Simpson et al. (2016). Here, the irregular domain adds an extra level of complexity, which has been handled with an approximation to the Gaussian random field, an associated approximation to the NHPP likelihood, and using integrated nested Laplace approximation for fast, but approximate Bayesian inference.

Our main contribution is flexible modeling and computationally efficient inference for NHPPs over spatial domains with irregular shapes. The proposed models do not rely on approximations of the NHPP likelihood and they can be efficiently implemented with standard Markov chain Monte Carlo algorithms for full Bayesian inference and uncertainty quantification. Moreover, in the context of the more commonly studied setting of spatial NHPPs over regular domains, our modeling approach overcomes some of the limitations of existing Bayesian methods, while retaining the feature of flexible inference for general intensity shapes.

We build the model for the NHPP intensity function from weighted combinations of Bernstein polynomial basis functions, that is, beta densities with specified shape parameters. Such parsimonious mixture representation is the key to achieve computationally tractable inference. In Section 2.3, we explore two modeling approaches for spatial Poisson process intensities over irregular domain, taken without loss of generality to be a subset of the unit square. Under the first approach, the representation for the NHPP intensity is motivated by truncating over the irregular domain a NHPP density defined as a weighted combination of Bernstein densities on the unit square. The second approach targets directly the NHPP intensity modeling it as a structured weighted combination of truncated Bernstein densities. The two models offer different benefits while sharing the feature that the total intensity, Λ, can be readily expressed in terms of model parameters. Thus, both models bypass the challenge brought about from the NHPP likelihood normalizing term without separating the total intensity and NHPP density in the prior specification. In the case of regular domain, say the unit square, the two modeling approaches yield the same form for the NHPP intensity which implies a Bernstein-Dirichlet prior for the corresponding NHPP density. To highlight this connection and its implications in posterior simulation, we begin in the next section with the methodology for the simpler setting of temporal NHPPs.

2.2 Methodology for temporal Poisson processes

2.2.1 Model formulation

Here, we focus on modeling one-dimensional NHPPs observed over a bounded domain, taken without loss of generality to be the unit interval. Motivated by Bernstein polynomial priors for densities with bounded support, our model for the intensity function $\lambda(s)$ implies a Bernstein-Dirichlet process prior for the NHPP density, $f(s) = \lambda(s)/\int_0^1 \lambda(u)\mathrm{d}u$, for $s \in [0,1]$.

The Bernstein polynomial prior model for density f on $[0,1]$ is given by $f_K(s \mid F) = \sum_{k=1}^{K} \omega_k \text{ be}(s \mid k, K-k+1)$, where $\text{be}(\cdot \mid a,b)$ is the beta density with mean $a/(a+b)$. The mixture weights are defined through increments of a distribution function F with support on $[0,1]$, such that $\omega_k = F(k/K) - F((k-1)/K)$, for $k = 1,\ldots,K$. A distribution F with flexible shape implies mixture weights that select the appropriate beta basis densities to achieve general shapes for density f. This motivates assigning a nonparametric prior to F, such as the Dirichlet process prior (Ferguson, 1973) which results in the Bernstein-Dirichlet prior for density f (Petrone, 1999a,b). Theoretical support for the Bernstein polynomial model is provided by the fact that, as $K \to \infty$, $f_K(s \mid F)$ converges uniformly to the density of F (Levasseur, 1984); this result is also key to establishing Kullback-Leibler support and posterior consistency of the Bernstein-Dirichlet prior for density estimation (Petrone and Wasserman, 2002). Extensions of Bernstein polynomial prior models include density estimation on higher dimensional spaces (Zheng et al., 2010; Barrientos et al., 2015) and density regression (Barrientos et al., 2017).

Our modeling approach is motivated by the structure of the distribution for the mix-

ture weights, $(\omega_1, ..., \omega_K)$, implied by a Dirichlet process prior, DP(α, F_0), on F, where α is the Dirichlet process precision parameter, and F_0 the centering distribution with support on $[0, 1]$. Based on the Dirichlet process definition, $(\omega_1, ..., \omega_K)$, given α, F_0, and K, follows a Dirichlet $(\alpha A_1, ..., \alpha A_K)$ prior distribution, where $A_k = F_0(k/K) - F_0((k-1)/K)$, for $k = 1, ..., K$. The key observation for the model is that the prior distribution for $(\omega_1, ..., \omega_K)$ can be constructed through independent gamma random variables. In particular, denoting by Ga(a, b) the gamma distribution with mean a/b, we have $\omega_k = V_k / \{\sum_{r=1}^K V_r\}$, where, for $k = 1, ..., K$, the V_k are independently Ga$(\alpha A_k, C)$ distributed, with $C > 0$ a constant.

The proposed model for one-dimensional NHPP intensities is given by:

$$\begin{aligned} \lambda(s) \;&=\; \sum_{k=1}^{K} V_k \;\text{be}(s \mid k, K-k+1), \quad s \in [0,1] \\ V_k \mid \alpha, F_0 \;&\overset{ind.}{\sim}\; \text{Ga}(\alpha\{F_0(k/K) - F_0((k-1)/K)\}, C), \quad k = 1, ..., K. \end{aligned} \tag{2.2}$$

The total intensity over the domain is $\Lambda = \int_0^1 \lambda(u)\mathrm{d}u = \sum_{k=1}^K V_k$, and thus the NHPP density is given by $f(s) = \lambda(s)/\{\int_0^1 \lambda(u)\mathrm{d}u\} = \sum_{k=1}^K \omega_k \,\text{be}(s \mid k, K-k+1)$, where $\omega_k = V_k / \{\sum_{r=1}^K V_r\}$. Hence, the implied model for the NHPP density is the Bernstein-Dirichlet prior model. Based on the Dirichlet process definition, this connection holds true for any K, that is, for any partition $\{S_k = [(k-1)/K, k/K) : k = 1, ..., K\}$ of the unit interval.

Note that, since $\Lambda = \sum_{k=1}^K V_k$, we have $E(\Lambda \mid \alpha) = \alpha/C$, which justifies using a general constant C in the prior for the V_k, rather than taking $C = 1$. That is, we wish to avoid the conflict of large values of α that would be needed under $C = 1$ for large prior expected total intensity versus small values of α favoring non-standard intensity function shapes.

A Ga(a_α, b_α) prior is assigned to α. In terms of model economy, the uniform distribution is an appealing choice for F_0. This choice is sufficiently flexible in practice, as shown

with the data examples of Section 2.2.3, and it also yields a form for the average intensity that facilitates prior specification. With F_0 uniform, the prior mean for the intensity is constant, given by $\mathrm{E}(\alpha)/C$ as proved below, and it does not depend on K.

$$\begin{aligned}
\mathrm{E}(\lambda(s)|\alpha, K) &= \sum_{k=1}^{K} \mathrm{E}(V_k \mid \alpha)\, \mathrm{be}(s|k, K-k+1) \\
&= \frac{\alpha}{C} \sum_{k=1}^{K} \frac{1}{K} \frac{K! s^{k-1}(1-s)^{K-k}}{(k-1)!(K-k)!} \\
&= \frac{\alpha}{C} \sum_{m=0}^{K-1} \frac{(K-1)! s^{m}(1-s)^{K-1-m}}{m!(K-1-m)!} \\
&= \frac{\alpha}{C} \sum_{m=0}^{K-1} \binom{K-1}{m} s^{m}(1-s)^{K-1-m} \\
&= \frac{\alpha}{C}
\end{aligned}$$

using the Binomial theorem. Note that the conditional prior expectation does not depend on K. Finally, $\mathrm{E}(\lambda(s)) = \mathrm{E}(\mathrm{E}(\lambda(s) \mid \alpha)) = \mathrm{E}(\alpha)/C$.

To explore posterior simulation under model (2.2), we consider two equivalent hierarchical model formulations for the observed point pattern $\{0 < s_1 < ... < s_n < 1\}$. As discussed above, there is an one-to-one correspondence between parameter vectors $(V_1, ..., V_K)$ and $\{\Lambda, (\omega_1, ..., \omega_K)\}$, where $\omega_k = F(S_k)$, for $k = 1, ..., K$. The prior distribution for $(V_1, ..., V_K)$ in (2.2) corresponds to a $\mathrm{DP}(\alpha, F_0)$ prior for F, and a $\mathrm{Ga}(\alpha, C)$ prior for Λ. Moreover, the NHPP likelihood in (2.1) can be conveniently expressed in terms of either parameterization:

$$\prod_{k=1}^{K} e^{-V_k} \prod_{i=1}^{n} \left\{ \sum_{k=1}^{K} V_k\, \mathrm{be}(s_i \mid k, K-k+1) \right\} = e^{-\Lambda}\, \Lambda^n \prod_{i=1}^{n} \left\{ \sum_{k=1}^{K} F(S_k)\, \mathrm{be}(s_i \mid k, K-k+1) \right\}.$$

Working with fixed K, the intensity formulation involves parameters $\{(V_1, \ldots, V_K), \alpha\}$. Here, we introduce discrete latent variables $\{\xi_i : i = 1, \ldots, n\}$ indicating basis configuration

for each time event. In a Gibbs sampler setting, the posterior full conditional for each ξ_i is a discrete distribution with support on $\{1, \ldots, K\}$. Most importantly, given $\{\xi_i : i = 1, \ldots, n\}$ and α, each V_k follows a gamma posterior full conditional distribution, independently of $\{V_r : r \neq k\}$. Lastly, α can be sampled using a Metropolis-Hastings step.

Alternatively, the density formulation builds from parameters $\{\Lambda, F, \alpha, K\}$. In this case, we introduce continuous latent variables $\{\theta_i : i = 1, ..., n\}$ to leverage the Dirichlet process mixture representation for the NHPP density function:

$$f(s_i) \equiv f_K(s_i \mid F) = \int \sum_{k=1}^{K} \mathbb{1}_{[\frac{k-1}{K}, \frac{k}{K})}(\theta_i) \, \text{be}(s_i \mid k, K - k + 1) \, \mathrm{d}F(\theta_i). \tag{2.3}$$

A practically important feature of this formulation is that the number of basis densities, K, can be estimated without resorting to trans-dimensional Markov chain Monte Carlo algorithms. Here, the dimension of the parameter space does not change with K because the posterior distribution does not involve the weights ω_k, but rather the random distribution F whose increments define the mixture weights. Posterior simulation proceeds by first sampling from the marginal posterior of $\{(\theta_1, \ldots, \theta_n), \Lambda, \alpha, K\}$, using Markov chain Monte Carlo methods for Dirichlet process mixtures (Escobar and West, 1995; Neal, 2000). We then sample $(\omega_1, \ldots, \omega_K)$, given $(\theta_1, \ldots, \theta_n), \alpha, K$, from the Dirichlet distribution implied by the Dirichlet process conditional posterior distribution for F, given $(\theta_1, \ldots, \theta_n)$ and α. Finally, posterior samples for the NHPP density and intensity can be readily obtained, using their expressions under model (2.2).

We provide the full conditionals used in both posterior simulation algorithms below. Under the intensity formulation given the number of basis K, the Markov Chain Monte Carlo algorithm consists of Gibbs or Metropolis update from the full conditionals for ξ_i, V_k and α.

1. $\xi_i|-$

$$p(\xi_i = j|-) = \frac{V_j \text{be}(s_i|j, K-j+1)}{\sum_{l=1}^{K} V_l \text{be}(s_i|l, K-l+1)}$$

2. $V_k|-$ for $k = 1 \cdots K$, $M_k = \sum_{i=1}^{n} \delta_k(\xi_i)$

$$p(V_k|-) \propto \exp(-V_k) V_k^{M_k} V_k^{\alpha/K-1} \exp(-CV_k) \propto \text{ga}(V_k|M_k + \alpha/K, C+1)$$

3. $\alpha|-$

$$\begin{aligned} p(\alpha|-) &\propto \text{ga}(\alpha|a_\alpha, b_\alpha) \prod_{k=1}^{K} \text{ga}(V_{k_x,k_y}|\alpha K^{-1}, C) \\ &\propto \alpha^{a_\alpha - 1} \exp(-b_\alpha \alpha) C^\alpha \Gamma(\alpha/K)^{-K} \prod_{k=1}^{K} V_k^{\alpha/K} \end{aligned}$$

A Metropolis step is implemented on the log scale with a normal random walk proposal density to sample from this full conditional.

The Markov Chain Monte Carlo algorithm for the density formulation of the temporal Poisson process consists of either a Metropolis or a Gibbs update from the following full-conditionals:

Let $k(s|\theta) = \sum_{k=1}^{K} \mathbb{1}_{((k-1)/K, k/K)}(\theta) \text{be}(s|k, K-k+1)$

1. $\theta_i|\boldsymbol{\theta_{-i}}, -$

$$p(\theta_i|\theta_{-i}, -) = \frac{\alpha q_0}{\alpha q_0 + H} \frac{k(s_i|\theta) f_0(\theta)}{q_0} + \frac{1}{\alpha q_0 + H} \sum_{j=1}^{n^{*-}} k(s_i|\theta_j^{*-}) n_j^- \delta_{\theta_j^{*-}}(\theta_i)$$

$$q_0 = \int k(s_i|\theta) f_0(\theta) \text{d}\theta = \sum_{j=1}^{K} \text{be}(s_i|j, K-j+1)\alpha/K$$

$$H = \sum_{j=1}^{n^{*-}} k(s_i|\theta_j^{*-}) n_j^-$$

where n^{*-} is the number of unique values, $\{\theta_j^{*-} : j = 1 \cdots n^{*-}\}$ is the vector of unique values, and $\{n_j^-, j = 1 \cdots n^{*-}\}$ the vector of the number of observations that take value θ_j^{*-} in the vector $\boldsymbol{\theta_{-i}} = \{\theta_l : l \neq i\}$.

2. $\Lambda|-$

$$\Lambda|- \sim \text{Ga}(\alpha + n, C + 1)$$

where n is the number of points

3. $\alpha|-$

$$\begin{aligned} p(\alpha|-) &\propto (\prod_{m=1}^{n}(\alpha + m - 1))^{-1}\alpha^{n^*}\text{ga}(\Lambda|\alpha, C)\text{ga}(\alpha|a_0, b_0) \\ &\propto (\prod_{m=1}^{n}(\alpha + m - 1))^{-1}\alpha^{n^*}\frac{C^\alpha}{\Gamma(\alpha)}\Lambda^{\alpha-1}\alpha^{a_0-1}\exp(-b_0\alpha) \end{aligned}$$

A Metropolis step is implemented with a normal proposal on the log scale.

4. $K|-$

$$p(K|-) \propto \prod_{i=1}^{n}\left\{\sum_{k=1}^{K}\mathbb{1}_{[\frac{k-1}{K},\frac{k}{K}]}(\theta_i)\text{be}(s_i|k, K-k+1)\right\}\pi(K|\{K_{min}, \cdots, K_{max}\})$$

The full conditional for K is a discrete distribution and can be directly sampled from.

With each draw in the posterior sample for $\{(\theta_1 \cdots \theta_n), \alpha, K\}$, we sample $\{\omega_k : k = 1 \cdots K\}$ from the following Dirichlet distribution

$$\{\omega_k : k = 1 \cdots K\} \sim \text{Dir}(\{\alpha/K + \sum_{i=1}^{n}\mathbb{1}_{[\frac{k-1}{K},\frac{k}{K}]}(\theta_i) : k = 1 \cdots K\})$$

We obtain a draw from the posterior distribution of the intensity function $\lambda(s)$ and the density function $f(s)$ evaluated at location s respectively, given $\{\omega_1, \cdots, \omega_K\}$ via the following functions

$$f(s) = \textstyle\sum_{k=1}^{K}\omega_k\text{be}(s|k, K-k+1)$$

$$\lambda(s) = \Lambda\textstyle\sum_{k=1}^{K}\omega_k\text{be}(s|k, K-k+1)$$

2.2.2 Prior specification

The prior for α and the value for C can be specified using prior guesses at the total intensity, $\hat{\Lambda}$, and an average intensity value, $\hat{\lambda}$, over the observation window. We select b_α to provide a wide range for α, and using $\mathrm{E}(\lambda(s)) = \mathrm{E}(\alpha)/C$, set $\mathrm{E}(\alpha) = a_\alpha/b_\alpha = C\hat{\lambda}$. The marginal prior for the total intensity is $p(\Lambda) = \int \mathrm{Ga}(\Lambda \mid \alpha, C)\, \mathrm{Ga}(\alpha \mid b_\alpha C\hat{\lambda}, b_\alpha)\, \mathrm{d}\alpha$. We use this expression to specify C such that the median of $p(\Lambda)$ is equal to $\hat{\Lambda}$.

Note the connection between α and K in controlling the shape of prior realizations for the NHPP intensity: for fixed α, increasing K results in intensities with larger number of modes and more local features; and, for fixed K, decreasing α favors more variability and more localized structure in the intensities. In practice, it may suffice to estimate only α keeping K fixed at sufficiently large values. Note that the beta densities in model (2.2) play the role of basis functions rather than of kernel densities in finite mixture models. Also key is the Dirichlet process underlying the prior for the weights V_k, which select the subset of beta densities that contribute more to the intensity representation. As illustrated with simulated data in Section 2.2.3, the discrete nature of the Dirichlet process prior can effectively guard against over-fitting if one conservatively chooses a larger value for K than may be necessary for a particular point pattern.

A possible approach to specify K involves prior information on the peak of the intensity, $\hat{\lambda}_{\max}$, without necessarily knowing where in the observation window the peak occurs. The idea is to find K such that $\hat{\lambda}_{\max}$ matches a percentile of the prior distribution of $b^* V_{\max}$, where $V_{\max} = \max\{V_k : k = 1, ..., K\}$, and b^* is the modal value of the beta$(2, K-1)$ density,

Table 2.1: Illustration of the prior specification strategy for K. $Q_{0.9}^{V_{\max}}$ denotes the 90th percentile of the marginal prior distribution for $\max\{V_k : k = 1, ..., K\}$, and b^* the modal value of the beta$(2, K-1)$ basis density.

K	$Q_{0.9}^{V_{\max}}$	b^*	$b^* \times Q_{0.9}^{V_{\max}}$
20	232.34	7.56	1755.85
30	208.18	11.23	2338.0
50	181.36	18.58	3370.34
100	167.38	36.97	6188.82

that is, the first member of the Bernstein polynomial basis with a unimodal density. Under the uniform F_0 distribution, the V_k are independently and identically gamma distributed, and thus the prior distribution of $V_{\max}$ is analytically available given α; the marginal prior for $V_{\max}$ can also be readily explored through simulation. Table 2.1 provides an illustration, using the 90th percentile of the marginal prior distribution for $V_{\max}$, under a Ga$(2.53, 0.1)$ prior for α, and with values for the peak intensity that are relevant to one of the data examples of Section 2.2.3.

As discussed in Section 2.2.1, using the intensity formulation, with fixed K, allows for a particularly simple and efficient method to implement model (2.2). The more general version of the model with random K can be implemented at the expense of somewhat more complex Markov chain Monte Carlo algorithms for Dirichlet process mixtures. A discrete uniform or a truncated Poisson distribution with support on $[K_{\min}, K_{\max}]$ are possible priors for K.

2.2.3 Synthetic data examples for the temporal NHPP model

We consider two synthetic data sets generated from NHPPs with bimodal intensities. For the first example, the intensity is $\lambda(s) = 700\,\text{be}(s \mid 3, 18) + 300\,\text{be}(s \mid 13, 8)$; this can be viewed as a special case of model (2.2) with $K = 20$, although our prior model does not allow for zero weights. The second data set is obtained by logit-transforming points generated from a weighted combination of normal densities, $\lambda(s) = 400\,\text{N}(s \mid -2.2, 1.0) + 600\,\text{N}(s \mid 0.3, 0.8)$. We take large sizes for the simulated point patterns – $n = 993$ for the first, and $n = 1037$ for the second example – to ensure a meaningful comparison of posterior estimates with the true intensities.

We follow the approach of Section 2.2.2 to specify $C = 0.023$ and a $\text{Ga}(2.53, 0.1)$ prior for α, using for both data examples 1000 as the prior estimate for the total intensity, and 1100 for the average intensity. For the first example, we take $K = 20$, as well as $K = 40$ to study the implication of using a number of basis densities that is twice as large as what should suffice. For the second example, assume we are told that the peak of the intensity has a value around 2300. Then, referring to Table 2.1, $K = 30$ can be taken as the number of basis densities, or, more conservatively, as a lower bound. We consider again a larger value, $K = 50$, to check sensitivity of posterior inference results. We also implemented the density formulation for the second example, with a uniform prior on $[20, 60]$ assigned to K.

As shown in Fig. 2.1, the model is effective in estimating the weights that drive the bimodal intensity shape of the two-component beta mixture. Under $K = 20$, it gives most weight to V_3 and V_{13}, that correspond to basis densities $\text{be}(s \mid 3, 18)$ and $\text{be}(s \mid 13, 8)$, whereas

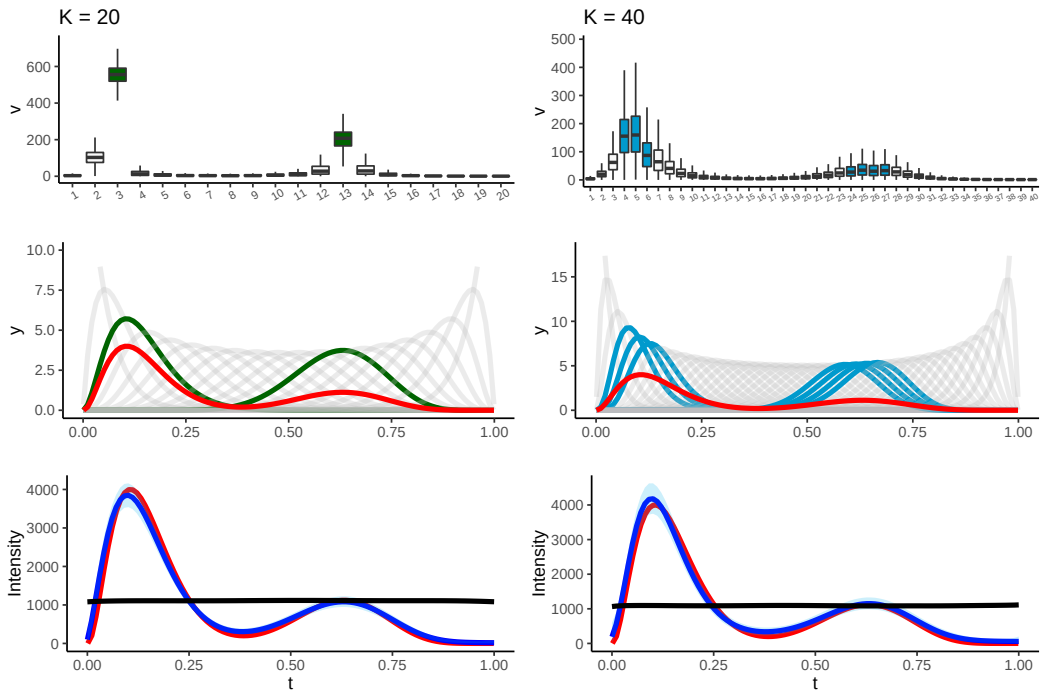


Figure 2.1: Beta mixture synthetic data example. Results under the intensity formulation with $K = 20$ (left column) and $K = 40$ (right column). Boxplots of posterior samples for the weights V_k (first row), the beta basis densities corresponding to the largest V_k (second row), and posterior mean (blue line) and 95% interval estimates (light blue shaded bands) for the intensity function (third row). In the second and third rows, the red line denotes the true density and intensity, respectively. In the third row, the black line indicates the prior mean for the intensity function.

when $K = 40$, the model favors 6-7 basis densities with peaks in the same range as the two modes of the underlying intensity. Hence, the model is able to achieve sparsity in estimation of the mixture weights when a surplus of basis densities are used, even though F_0 is a uniform distribution. Moreover, with the exception of some increase in the width of posterior uncertainty bands, inference results for the intensity function are similar under the two different choices for K.

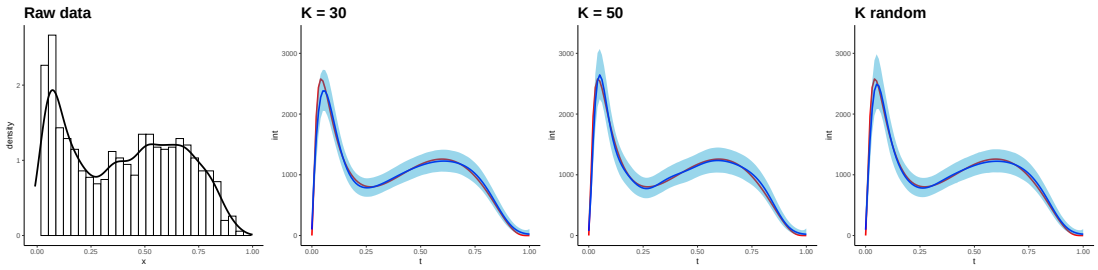


Figure 2.2: Logit-normal mixture synthetic data example. From left to right, histogram of the simulated time points, and posterior mean (blue line) and 95% interval estimates (light blue shaded bands) for the intensity function under $K = 30$, $K = 50$, and K random. The red line in the last three panels denotes the true intensity.

This is also the case with the posterior inference results for the logit-normal mixture data example; see Fig. 2.2. Under the density formulation, the posterior median for K is 36, with the 95% credible interval given by $[22, 56]$. The intensity function under random K has similar point estimate and a slightly tighter uncertainty band compared to that under $K = 50$.

2.3 Modeling approaches for spatial Poisson processes

We begin with the case of a regular domain for the spatial NHPP, taken without loss of generality to be the unit square, such that $s \equiv (x, y) \in [0, 1]^2$. The extension of the Bernstein polynomial basis consists of products of beta densities. More specifically, the basis density with index (k_x, k_y), for $k_x, k_y = 1, ..., K$, is defined as

$$\phi_{k_x,k_y}(x, y) = \text{be}\,(x \mid k_x, K - k_x + 1)\,\text{be}\,(y \mid k_y, K - k_y + 1)\,, \quad (x, y) \in [0, 1]^2. \qquad (2.4)$$

Although the number of basis densities may be different in the x and y dimensions, we use the more parsimonious form with $K_x = K_y = K$.

Then, we can extend model (2.2) to the following model for spatial NHPP intensities over $[0,1]^2$:

$$\begin{aligned} \lambda(x,y) &= \sum_{k_x,k_y=1}^{K} V_{k_x,k_y}\, \phi_{k_x,k_y}(x,y), \quad (x,y) \in [0,1]^2 \\ V_{k_x,k_y} \mid \alpha, F_0 &\overset{ind.}{\sim} \text{Ga}(\alpha F_0(S_{k_x,k_y}), C), \quad k_x, k_y = 1, ..., K \end{aligned} \tag{2.5}$$

where $S_{k_x,k_y} = [(k_x-1)/K, k_x/K) \times [(k_y-1)/K, k_y/K)$, and $F_0(S_{k_x,k_y})$ is the probability of S_{k_x,k_y} under a specified distribution F_0 on $[0,1]^2$; in particular, $F_0(S_{k_x,k_y}) = 1/K^2$ under the uniform distribution for F_0.

Again, the total intensity over the domain is readily obtained as $\Lambda = \int_0^1 \int_0^1 \lambda(x,y)\,\mathrm{d}x\mathrm{d}y = \sum_{k_x,k_y=1}^{K} V_{k_x,k_y}$, and the NHPP density is given by $f(x,y) = \sum_{k_x,k_y=1}^{K} \omega_{k_x,k_y}\, \phi_{k_x,k_y}(x,y)$, where $\omega_{k_x,k_y} = V_{k_x,k_y} / \{\sum_{k_x,k_y=1}^{K} V_{k_x,k_y}\}$. The implied prior distribution for the mixture weights $\{\omega_{k_x,k_y}\}$ corresponds to constructing them through $\omega_{k_x,k_y} = F(S_{k_x,k_y})$, where F is a random distribution on $[0,1]^2$ assigned a DP(α, F_0) prior.

We thus retain the connection between the intensity prior model in (2.5) and the two-dimensional Bernstein-Dirichlet prior model for the NHPP density, as well as the equivalent hierarchical model formulations for the data. Again, the implied Ga(α, C) prior for Λ ensures the coherence between the intensity and density prior models, the latter comprising parameters $\{\Lambda, F, \alpha, K\}$. Extending the approaches outlined in Section 2.2.1, posterior simulation can be implemented using either the intensity or density formulation. The prior mean intensity is $\mathrm{E}(\lambda(x,y)) = \mathrm{E}(\alpha)/C$ as proved below, and thus the prior specification approach of Section 2.2.2 can be extended to model (2.5).

$$
\begin{aligned}
\mathrm{E}(\lambda(x,y) \mid \alpha, K) &= \sum_{k_x=1}^{K}\sum_{k_y=1}^{K} \mathrm{E}(V_{k_x,k_y} \mid \alpha)\, \mathrm{be}(x|k_x, K-k_x+1)\mathrm{be}(y|k_y, K-k_y+1) \\
&= \frac{\alpha}{C}\frac{1}{K^2}\sum_{k_x=1}^{K}\sum_{k_y=1}^{K} \mathrm{be}(x|k_x, K-k_x+1)\mathrm{be}(y|k_y, K-k_y+1) \\
&= \frac{\alpha}{C}
\end{aligned}
$$

using the fact that $K^{-1}\sum_{m=1}^{K} \mathrm{be}(s|m, K-m+1) = 1$, which is essentially a restatement of the Binomial theorem.

To achieve our main objective of flexible inference for NHPP spatial intensities recorded over irregular domain $\mathcal{D} \subset [0,1]^2$, we propose two different modeling approaches. Under the first model, presented in Section 2.3.1, the intensity formulation is motivated by truncating over $\mathcal{D}$ the NHPP density $f(x,y)$ defined on $[0,1]^2$. The second model, developed in Section 2.3.2, builds the basis representation for the intensity through the corresponding density which is defined as a mixture of truncated beta densities over $\mathcal{D}$ with weights induced by a random distribution F on $\mathcal{D}$. In both cases, the Bernstein polynomial prior structure is especially attractive to model spatial point process intensities over irregular domains, a practically relevant problem that, arguably, has not been fully addressed in the Bayesian nonparametrics literature.

2.3.1 The intensity model

Under the first modeling perspective, the representation for the NHPP intensity $\lambda_{\mathcal{D}}(x,y)$ over irregular domain $\mathcal{D}$ is revealed by the expression for $f_{\mathcal{D}}(x,y)$, the NHPP density truncated

on $\mathcal{D}$. In particular,

$$f_{\mathcal{D}}(x,y) = \frac{f(x,y)}{\int\int_{\mathcal{D}} f(u,v)\,\mathrm{d}u\mathrm{d}v} = \sum_{k_x,k_y=1}^{K} \frac{V_{k_x,k_y}B_{k_x,k_y}}{\sum_{k_x,k_y=1}^{K} V_{k_x,k_y}B_{k_x,k_y}}\,\phi^*_{k_x,k_y}(x,y), \;\; (x,y)\in\mathcal{D} \tag{2.6}$$

where $B_{k_x,k_y} = \int\int_{\mathcal{D}} \phi_{k_x,k_y}(x,y)\,\mathrm{d}x\mathrm{d}y$, $\phi^*_{k_x,k_y}(x,y) = \phi_{k_x,k_y}(x,y)/B_{k_x,k_y}$ are the basis densities truncated on $\mathcal{D}$, and we have used the fact that $\omega_{k_x,k_y}B_{k_x,k_y}/\{\sum_{k_x,k_y=1}^{K}\omega_{k_x,k_y}B_{k_x,k_y}\} = V_{k_x,k_y}B_{k_x,k_y}/\{\sum_{k_x,k_y=1}^{K}V_{k_x,k_y}B_{k_x,k_y}\}$. The implied model for the intensity function is:

$$\lambda_{\mathcal{D}}(x,y) = \sum_{k_x,k_y=1}^{K} V_{k_x,k_y}B_{k_x,k_y}\,\phi^*_{k_x,k_y}(x,y), \quad (x,y)\in\mathcal{D} \tag{2.7}$$

where $V_{k_x,k_y} \mid \alpha \overset{ind.}{\sim} \text{Ga}(\alpha/K^2, C)$, for $k_x, k_y = 1, ..., K$, taking the uniform distribution for F_0, and placing a $\text{Ga}(a_\alpha, b_\alpha)$ prior on α.

Evidently, (2.5) and (2.7) agree when $\mathcal{D}$ is the unit square. Note that B_{k_x,k_y} will be small for basis densities with significant mass outside $\mathcal{D}$. Hence, although model (2.7) uses all K^2 basis densities, the constants B_{k_x,k_y} provide an additional adjustment to the one applied by the random coefficients V_{k_x,k_y}. The overhead cost of computing the normalizing constants B_{k_x,k_y} is very small, since, with fixed K, they need to be computed only once.

For posterior simulation, we introduce a pair of latent variables, (ξ_i, η_i), for each point in the spatial point pattern, $\{(x_i, y_i) : i = 1, ..., n\}$, to identify the corresponding basis density. Then, the hierarchical model for the data can be written as:

$$\{(x_i,y_i)\} \mid \boldsymbol{V}, \{(\xi_i,\eta_i)\} \sim \left(\prod_{k_x,k_y=1}^{K}\exp(-V_{k_x,k_y}B_{k_x,k_y})\right)\prod_{i=1}^{n}\Lambda_{\mathcal{D}}\,\phi^*_{\xi_i,\eta_i}(x_i,y_i)$$

$$(\xi_i,\eta_i) \mid \boldsymbol{V} \overset{i.i.d.}{\sim} \sum_{k_x,k_y=1}^{K}\frac{V_{k_x,k_y}B_{k_x,k_y}}{\Lambda_{\mathcal{D}}}\,\delta_{(k_x,k_y)}(\xi_i,\eta_i), \quad i = 1, ..., n$$

$$\alpha, \boldsymbol{V} \sim \text{Ga}(\alpha \mid a_\alpha, b_\alpha)\prod_{k_x,k_y=1}^{K}\text{Ga}(V_{k_x,k_y} \mid \alpha K^{-2}, C)$$

where $V = \{V_{k_x,k_y} : k_x, k_y = 1, ..., K\}$, and $\Lambda_{\mathcal{D}}$ is the total intensity over the irregular domain, $\Lambda_{\mathcal{D}} = \int\int_{\mathcal{D}} \lambda_{\mathcal{D}}(x, y)\, \mathrm{d}x\mathrm{d}y = \sum_{k_x,k_y=1}^{K} V_{k_x,k_y} B_{k_x,k_y}$.

As with models (2.2) and (2.5), the form of the NHPP likelihood normalizing term implied by the intensity model (2.7) results in efficient posterior simulation with remarkably simple updates for parameters $\{V_{k_x,k_y}\}$; given the (ξ_i, η_i) and α, the V_{k_x,k_y} are conditionally independent and gamma distributed. $\{B_{k_x,k_y}, k_x, k_y = 1 \cdots K\}$ can be computed given K and $\mathcal{D}$ before running the Markov Chain Monte Carlo algorithm to save computation time.

Given the number of basis K, the Markov Chain Monte Carlo algorithm consists of Gibbs or Metropolis update from the full conditionals for $\{\xi_i, \eta_i\}, V_{k_x,k_y}$ and α:

1. The full conditional for $\{\xi_i, \eta_i\}$ are discrete distributions

$$p(\xi_i = m, \eta_i = n|-) = \frac{V_{m,n}\text{be}(x_i|m, K-m+1)\text{be}(y_i|n, K-n+1)}{\sum_{p,q=1}^{K} V_{p,q}\text{be}(x_i|p, K-p+1)\text{be}(y_i|q, K-q+1)}$$

2. The full conditional for $V_{k_x,k_y}, k_x, k_y = 1 \cdots K$ are independent Gamma distributions, which can be sampled directly from in a vectorized fashion. Let M_{k_x,k_y} be the number of latent variable pairs (ξ_i, τ_i) in step 1 that take value (k_x, k_y).

$$\begin{aligned} p(V_{k_x,k_y}|-) &\propto \exp(-V_{k_x,k_y}B_{k_x,k_y}) \prod_{i=1}^{n} \Lambda_{\mathcal{D}} \sum_{k_x,k_y=1}^{K} \frac{V_{k_x,k_y}B_{k_x,k_y}}{\Lambda_{\mathcal{D}}} \delta_{(k_x,k_y)}(\xi_i, \eta_i)\phi^*_{k_x,k_y}(x_i, y_i) \\ &\quad \times \text{ga}(V_{k_x,k_y}|\alpha/K^2, C) \\ &\propto \exp(-V_{k_x,k_y}B_{k_x,k_y}) \prod_{i=1}^{n} (V_{k_x,k_y}B_{k_x,k_y})^{\delta_{(k_x,k_y)}(\xi_i,\eta_i)} \text{ga}(V_{k_x,k_y}|\alpha/K^2, C) \\ &\propto \text{ga}(V_{k_x,k_y}|M_{k_x,k_y} + \alpha/K^2, C + B_{k_x,k_y}) \end{aligned}$$

3. The full conditional for α is

$$p(\alpha|-) \propto \text{ga}(\alpha|a_\alpha, b_\alpha) \prod_{k_x,k_y=1}^{K} \text{ga}(V_{k_x,k_y}|\alpha K^{-2}, C)$$

$$\propto \alpha^{a_\alpha - 1} \exp(-b_\alpha \alpha) C^\alpha \Gamma(\alpha/K^2)^{-K^2} \prod_{k_x,k_y=1}^{K} V_{k_x,k_y}^{\alpha K^{-2}}$$

A Metropolis step is implemented on the log scale with a normal random walk proposal density to sample from this full conditional.

In contrast to models (2.2) and (2.5), the NHPP density in (2.6) does not follow the Bernstein-Dirichlet prior. Consequently, we do not have a Dirichlet process mixture representation for the hierarchical model for the data, which allows estimating K without trans-dimensional posterior simulation algorithms. Therefore, practical implementation of model (2.8) requires specifying K. In practice, we recommend sensitivity analysis for the value of K. With K selected, the approach of Section 2.2.2 can be used to specify the prior for α and the value for C. The prior mean of the intensity function is again given by $\text{E}(\lambda_{\mathcal{D}}(x,y)) = \text{E}(\alpha)/C$ as shown below, and, although $\Lambda_{\mathcal{D}}$ no longer follows a gamma prior distribution, given α, its marginal prior can be easily developed by simulation.

$$\begin{aligned}
\text{E}(\lambda_{\mathcal{D}}(x,y) \mid \alpha) &= \sum_{k_x=1}^{K} \sum_{k_y=1}^{K} B_{k_x,k_y} \text{E}(V_{k_x,k_y} \mid \alpha) \left(B_{k_x,k_y}^{-1} \text{be}(x|k_x, K-k_x+1) \text{be}(y|k_y, K-k_y+1) \right) \\
&= \sum_{k_x=1}^{K} \sum_{k_y=1}^{K} \text{E}(V_{k_x,k_y} \mid \alpha) \, \text{be}(x|k_x, K-k_x+1) \text{be}(y|k_y, K-k_y+1) \\
&= \frac{\alpha}{C}
\end{aligned}$$

2.3.2 The density model

Here, we seek to develop a model for the irregular domain intensity that corresponds to a Bernstein-Dirichlet prior for the associated density, in the spirit of models (2.2) and (2.5). To this end, we define directly the density $f_{\mathcal{D}}(x, y)$ as a mixture of truncated beta basis densities:

$$f_{\mathcal{D}}(x, y) \;=\; \sum_{(k_x,k_y)\in J_K} \omega^*_{k_x,k_y}\, \phi^*_{k_x,k_y}(x, y), \quad (x, y) \in \mathcal{D} \tag{2.8}$$

where $J_K = \{(k_x, k_y) : S_{k_x,k_y} \cap \mathcal{D} \neq \emptyset\}$ is the index set for all non-empty intersections, $S^*_{k_x,k_y} = S_{k_x,k_y} \cap \mathcal{D}$, of the unit square partitioning sets $\{S_{k_x,k_y} : k_x, k_y = 1, ..., K\}$ with $\mathcal{D}$. The mixture weights are defined as $\omega^*_{k_x,k_y} = F(S^*_{k_x,k_y})$, where F is a random distribution on $\mathcal{D}$ following a DP(α, F_0) prior, with F_0 taken to be the uniform distribution on $\mathcal{D}$.

We now define the model for the irregular domain spatial intensity as

$$\begin{aligned} \lambda_{\mathcal{D}}(x, y) \;&=\; \sum_{(k_x,k_y)\in J_K} V^*_{k_x,k_y}\, \phi^*_{k_x,k_y}(x, y), \quad (x, y) \in \mathcal{D} \\ V^*_{k_x,k_y} \mid \alpha \;&\overset{ind.}{\sim}\; \text{Ga}(\alpha F_0(S^*_{k_x,k_y}), C), \quad (k_x, k_y) \in J_K \end{aligned} \tag{2.9}$$

such that the density $f_{\mathcal{D}}(x, y) = \lambda_{\mathcal{D}}(x, y)/\{\int\int_{\mathcal{D}} \lambda_{\mathcal{D}}(u, v)\, \mathrm{d}u\mathrm{d}v\}$ follows the prior model in (2.8). Again, the key link between parameterizations $\{V^*_{k_x,k_y} : (k_x, k_y) \in J_K\}$ and $\{\Lambda_{\mathcal{D}}, \{\omega^*_{k_x,k_y} : (k_x, k_y) \in J_K\}\}$ is the practical expression for the total intensity $\Lambda_{\mathcal{D}} = \int\int_{\mathcal{D}} \lambda_{\mathcal{D}}(x, y)\, \mathrm{d}x\mathrm{d}y = \sum_{(k_x,k_y)\in J_K} V^*_{k_x,k_y}$, and its Ga$(\alpha, C)$ prior implied by (2.9).

For a spatial point pattern $\{(x_i, y_i) : i = 1, ..., n\}$ recorded over $\mathcal{D}$, we can write the NHPP likelihood in terms of either the intensity of density formulation:

$$\begin{aligned} &\exp\left(-\textstyle\sum_{(k_x,k_y)\in J_K} V^*_{k_x,k_y}\right) \prod_{i=1}^{n} \left\{\textstyle\sum_{(k_x,k_y)\in J_K} V^*_{k_x,k_y}\, \phi^*_{k_x,k_y}(x_i, y_i)\right\} \\ &= \exp(-\Lambda_{\mathcal{D}})\, \Lambda^n_{\mathcal{D}} \prod_{i=1}^{n} \left\{\textstyle\sum_{(k_x,k_y)\in J_K} F(S^*_{k_x,k_y})\, \phi^*_{k_x,k_y}(x_i, y_i)\right\}. \end{aligned}$$

To explore the posterior distribution for $\{\Lambda_{\mathcal{D}}, F, \alpha, K\}$ under the density formulation, we introduce bivariate continuous latent variables $\{z_i\}$ to write the hierarchical model for the data:

$$\begin{gathered}
\{(x_i, y_i)\} \mid \{\boldsymbol{z}_i\}, \Lambda_{\mathcal{D}}, K \sim \exp(-\Lambda_{\mathcal{D}}) \prod_{i=1}^{n} \Lambda_{\mathcal{D}} \sum_{k_x, k_y \in J_k} \mathbb{1}_{S^*_{k_x,k_y}}(\boldsymbol{z_i}) \phi^*_{k_x,k_y}(x_i, y_i) \\
(x_i, y_i), \boldsymbol{z_i} \in \mathcal{D}, \quad \boldsymbol{z_i} \mid F \overset{i.i.d}{\sim} F, \;\; i = 1 \cdots, n \\
F \mid \alpha \sim \text{DP}(\alpha, F_0) \;\; F_0(\cdot) \equiv \text{Unif}(\mathcal{D}) \\
\Lambda_{\mathcal{D}} \mid \alpha \sim \text{Ga}(\alpha, C) \;\; \alpha \sim \text{Ga}(\alpha \mid a_\alpha, b_\alpha) \;\; K \sim \pi(K \mid \{K_{\min}, \cdots, K_{\max}\})
\end{gathered} \tag{2.10}$$

where $\Lambda_{\mathcal{D}} \mid \alpha \sim \text{Ga}(\alpha, C)$, with a $\text{Ga}(a_\alpha, b_\alpha)$ prior placed on α, and with a discrete uniform or a truncated Poisson prior distribution for K with support on $[K_{\min}, K_{\max}]$. The posterior simulation method is more involved than the one for the intensity model of Section 2.3.1, but it allows for estimation of K without trans-dimensional computational techniques.

The Markov Chain Monte Carlo algorithm consists of either Metropolis or Gibbs update from the following full-conditionals:

1. $\boldsymbol{z_i} | \boldsymbol{z_{-i}}, \alpha, K$, where $\boldsymbol{z_i}$ is the bivariate continuous latent variable.

   Let $k^*(\boldsymbol{s_i} | \boldsymbol{z_i}) = \sum_{(k_x,k_y) \in J_k} \mathbb{1}_{S^*_{k_x,k_y}}(\boldsymbol{z_i}) W^*_{k_x,k_y,i}$, where $W^*_{k_x,k_y,i} = \phi^*_{k_x,k_y}(x_i, y_i)$ is a constant.

$$p(\boldsymbol{z_i}|\boldsymbol{z_{-i}}, \boldsymbol{s_i}) = \frac{\alpha q_0}{\alpha q_0 + H} \frac{k^*(\boldsymbol{s_i}|\boldsymbol{z})f_0(\boldsymbol{z})}{q_0} + \frac{1}{\alpha q_0 + H} \sum_{j=1}^{n^{*-}} k^*(\boldsymbol{s_i}|\boldsymbol{z_j^{*-}}) n_j^- \delta_{\boldsymbol{z_j^{*-}}}(\boldsymbol{z_i})$$

$$= \frac{\alpha q_0}{\alpha q_0 + H} q(\boldsymbol{z}|\boldsymbol{s_i}) + \frac{1}{\alpha q_0 + H} \sum_{j=1}^{n^{*-}} k^*(\boldsymbol{s_i}|\boldsymbol{z_j^{*-}}) n_j^- \delta_{\boldsymbol{z_j^{*-}}}(\boldsymbol{z_i})$$

$$q_0 = \int k^*(\boldsymbol{s_i}|\boldsymbol{z}) f_0(\boldsymbol{z}) \mathrm{d}\boldsymbol{z} = \sum_{(k_x,k_y)\in J_k} W^*_{k_x,k_y,i\cdot} \mid S^*_{k_x,k_y} \mid / \mid \mathcal{D} \mid$$

$$q(\boldsymbol{z}|\boldsymbol{s_i}) = \sum_{(k_x,k_y)\in J_k} W^*_{k_x,k_y,i} q_0^{-1} \mathbb{1}_{S^*_{k_x,k_y}}(\boldsymbol{z}) \mid \mathcal{D} \mid^{-1}$$

$$= \sum_{(k_x,k_y)\in J_k} \frac{W^*_{k_x,k_y,i}|S^*_{k_x,k_y}|}{\sum_{m,n} W^*_{m,n,i}|S^*_{k_x,k_y}|} \mathbb{1}_{S^*_{k_x,k_y}}(\boldsymbol{z}) \mid S^*_{k_x,k_y} \mid^{-1}$$

$$H = \sum_{j=1}^{n^{*-}} k^*(\boldsymbol{s_i}|z_j^{*-}) n_j^-$$

where n^{*-} is the number of unique values, $\{\boldsymbol{z_j^{*-}} : j = 1 \cdots n^{*-}\}$ is the vector of unique values, and $\{n_j^-, j = 1 \cdots n^{*-}\}$ is the vector of the number of observations that take value z_j^{*-} in the vector $\boldsymbol{z_{-i}} = \{\boldsymbol{z_l} : l \neq i\}$.

2. $\Lambda_{\mathcal{D}}|-$

$$\Lambda_{\mathcal{D}}|- \sim \text{Ga}(\alpha + n, C + 1)$$

where n is number of points.

3. $\alpha|-$

$$p(\alpha|-) \propto (\prod_{m=1}^{n} (\alpha + m - 1))^{-1} \alpha^{n^*} \text{ga}(\Lambda_{\mathcal{D}}|\alpha, C) \text{ga}(\alpha|a_0, b_0)$$

$$\propto (\prod_{m=1}^{n} (\alpha + m - 1))^{-1} \alpha^{n^*} \frac{C^\alpha}{\Gamma(\alpha)} \Lambda_{\mathcal{D}}^{\alpha-1} \alpha^{a_0-1} \exp(-b_0\alpha)$$

A Metropolis step is implemented with a normal proposal on the log scale.

4. $K|-$

$$p(K|-) \propto \prod_{i=1}^{n} \left\{ \sum_{k_x,k_y\in J_K} \mathbb{1}_{S^*_{k_x,k_y}}(\boldsymbol{z_i}) W^*_{k_x,k_y,i} \right\} \pi(K|\{K_{min}, \cdots, K_{max}\})$$

The full conditional for K is a discrete distribution and can be directly sample from.

With each draw in the posterior sample for $\{\Lambda_{\mathcal{D}}, \{\boldsymbol{z_i}\}, \alpha, K\}$, we obtain a draw from the posterior distribution of $\{\omega^*_{k_x,k_y} : (k_x, k_y) \in J_K\}$ by sampling from the following Dirichlet distribution:

$$\{\omega^*_{k_x,k_y} : k_x, k_y = 1 \cdots K\} \sim \text{Dir}(\{\alpha/ \mid S^*_{k_x,k_y} \mid + \sum_{i=1}^{n} \mathbb{1}_{S^*_{k_x,k_y}}(\boldsymbol{z_i}) : k_x, k_y = 1 \cdots K\})$$

We obtain a draw from the posterior distribution of the intensity function $\lambda_{\mathcal{D}}(\boldsymbol{s})$ and the density function $f_{\mathcal{D}}(\boldsymbol{s})$ evaluated at location $\boldsymbol{s} = (x, y)$ respectively, given $\{\omega^*_{k_x,k_y} : k_x, k_y = 1 \cdots K\}$, via the following function

$$f_{\mathcal{D}}(x, y) = \textstyle\sum_{(k_x,k_y)\in J_K} \omega^*_{k_x,k_y} \phi^*_{k_x,k_y}(x, y)$$

$$\lambda_{\mathcal{D}}(x, y) = \Lambda_{\mathcal{D}} \textstyle\sum_{(k_x,k_y)\in J_K} \omega^*_{k_x,k_y} \phi^*_{k_x,k_y}(x, y)$$

The marginal prior for the total intensity is $p(\Lambda_{\mathcal{D}}) = \int \text{Ga}(\Lambda_{\mathcal{D}} \mid \alpha, C)\, \text{Ga}(\alpha \mid a_\alpha, b_\alpha)\, \text{d}\alpha$. Under model (2.9), there is no closed-form expression for $\text{E}(\lambda_{\mathcal{D}}(x, y))$, but $\text{E}(\alpha)/C$ is an approximate lower bound for the prior mean intensity as illustrated below.

$$\begin{aligned}
\text{E}(\lambda_{\mathcal{D}}(x, y) \mid \alpha, K) &= \sum_{(k_x,k_y)\in J_K} \frac{\alpha F_0(S^*_{k_x,k_y})}{C} \phi^*_{k_x,k_y}(x, y) \\
&\approx \frac{\alpha}{C} \sum_{(k_x,k_y)\in J_K} \frac{1}{K^2} \phi^*_{k_x,k_y}(x, y) \\
&\geq \frac{\alpha}{C} \sum_{(k_x,k_y)\in J_K} \frac{1}{K^2} B_{k_x,k_y} \phi^*_{k_x,k_y}(x, y) \\
&\approx \frac{\alpha}{C} \sum_{k_x=1}^{K} \sum_{k_y=1}^{K} \frac{1}{K^2} B_{k_x,k_y} \phi^*_{k_x,k_y}(x, y) \\
&= \frac{\alpha}{C} \sum_{k_x=1}^{K} \sum_{k_y=1}^{K} \frac{1}{K^2} \text{be}(x|k_x, K - k_x + 1)\text{be}(y|k_y, K - k_y + 1) = \frac{\alpha}{C}
\end{aligned}$$

In step 2, we use the fact that $S^*_{k_x,k_y}$ is the overlap between the $K \times K$ unit square partition set S_{k_x,k_y} and the irregular domain $\mathcal{D}$, and will have area either exactly equal to $1/K^2$, when

$S^*_{k_x,k_y} = S_{k_x,k_y}$, or area that can be approximated by $1/K^2$. In step 4, we use the fact that $B_{k_x,k_y} \approx 0$ for $(k_x, k_y) \notin J_K$. With this caveat, the approach of Section 2.2.2 can be used to specify the prior hyperparameters for α and the value for C. The earlier approach to specify K can be used here to guide the choice of the support for the prior on K.

2.4 Synthetic data examples

2.4.1 Examples over regular domain

We study the inference results under the intensity model over the unit-square to examine the model's capacity to capture the so-called "banana-shaped" spatial density which demonstrates spatial correlation between the X and Y dimension. The spatial Bernstein density is a product of two independent univariate beta densities. Therefore, these spatial Bernstein densities do not account for the spatial correlation when used as basis functions. Our hope is that with enough such basis functions and random weights informed by the data, our model can capture spatial correlation using the mixture.

We design a synthetic example where the true NHPP density is a mixture of two bivariate Gaussian densities with positive correlation coefficient parameters, transformed to the logit scale. The total intensity over the unit-square is 2000. We apply the intensity model with a Ga$(0.1, 0.005)$ prior for α, $C = 0.01$ and $K = 20, 30, 40$ respectively. Following the prior specification strategy in Section 2.2.2, we compare the prior guess for the intensity mode

$\hat{\lambda} = 16000$ to the quantity $b^{*2}Q_{0.9}^{max}$ to calibrate the value for K. We conclude that a good prior guess for K is 30.

The model is capable of capturing the banana shape overall but performs best when $K = 30$. Fig. 2.3 shows the comparison between the posterior mean estimate for the density function and the true density used in simulation under different values of K. The point estimate under $K = 20$ does not capture the elliptical contour near $(0.25, 0.25)$ well but shows great improvement under $K = 30$. Under $K = 40$, we observe clear diagonal elliptical contour near $(0.25, 0.25)$ in the point estimate. However it also includes additional local modes that might suggest over-fitting. Following our prior specification strategy, we conclude that $K = 30$ is enough to capture the spatial correlation in the NHPP density shown in this example.

2.4.2 Examples over irregular domain

We study inference results under both the intensity and density model, using point patterns generated under three different scenarios for the irregular shape of the spatial NHPP. The synthetic point patterns are plotted in Fig. 2.4, and the true intensities, as well as their corresponding polygonal domain, are shown in Fig. 2.5. For cases (a) and (b), the true NHPP density is a mixture of two bivariate logit-normal densities, truncated over the respective domain, which results in a unimodal intensity. Case (c) arises from truncating a mixture of bivariate beta densities that accumulates most of its mass at the $(0, 1)$ and $(1, 0)$ corners of the unit square.

For all three cases, the intensity model (2.8) is implemented with $C = 0.05$, a $\text{Ga}(2, 0.01)$ prior for α, and with $K = 20$. The posterior mean and uncertainty estimates reported in Fig. 2.5 demonstrate that the model recovers well the underlying intensity shapes

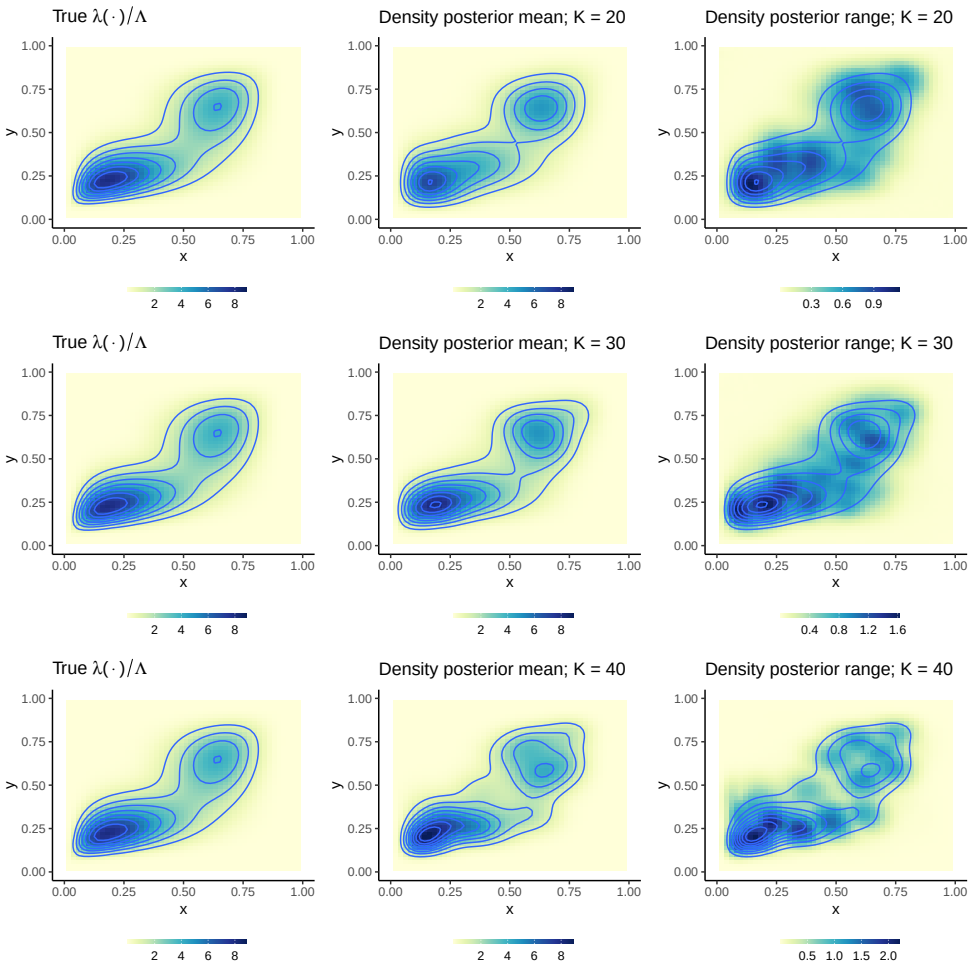


Figure 2.3: Results for synthetic data example over regular domain with $K = 20$ in the first row, $K = 30$ in the second row and $K = 40$ in the third row; the first column shows the true NHPP density, second column the posterior mean for the NHPP density and third column the 95% credible interval length.

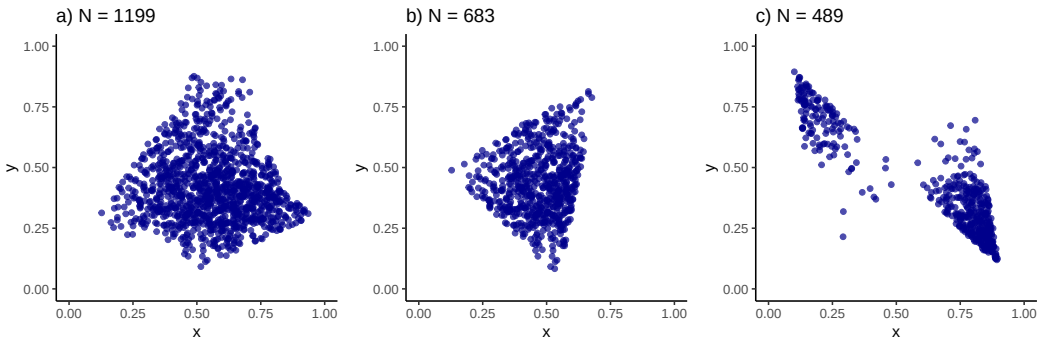


Figure 2.4: Synthetic spatial point patterns for the irregular domain simulation study. The size of each point pattern is shown in the corresponding panel.

over the different polygons.

We also applied the density model (2.10), using for all three data sets, a Ga$(5, 0.1)$ prior for α, $C = 0.01$, and a discrete uniform prior on $[5, 25]$ for K. The posterior probability for K at its posterior mode was: Pr$(K = 13 \mid \text{data}) = 0.89$ in case (a), Pr$(K = 12 \mid \text{data}) = 0.99$ in case (b), and Pr$(K = 9 \mid \text{data}) = 0.81$ in case (c). The posterior mean and uncertainty estimates under the density model were similar to the ones reported in Fig. 2.5 under the intensity model.

As an additional illustration, we consider a point pattern of size $n = 303$ drawn from a NHPP with density $0.7\,\text{be}(x \mid 4, 17)\text{be}(y \mid 10, 11) + 0.3\,\text{be}(x \mid 12, 9)\text{be}(y \mid 4, 17)$ truncated to the triangle with vertices $\{(0.01, 0.01), (0.2, 0.9), (0.9, 0.1)\}$, and with total intensity 300. Here, the truth is designed to resemble the intensity model with $K = 20$, and we test the performance of the density model in estimating K and other NHPP functionals.

Model (2.10) is implemented with a Ga$(2, 0.01)$ prior for α, $C = 0.01$, and a discrete uniform prior for K with support on $[15, 25]$. The posterior mean and uncertainty estimates in

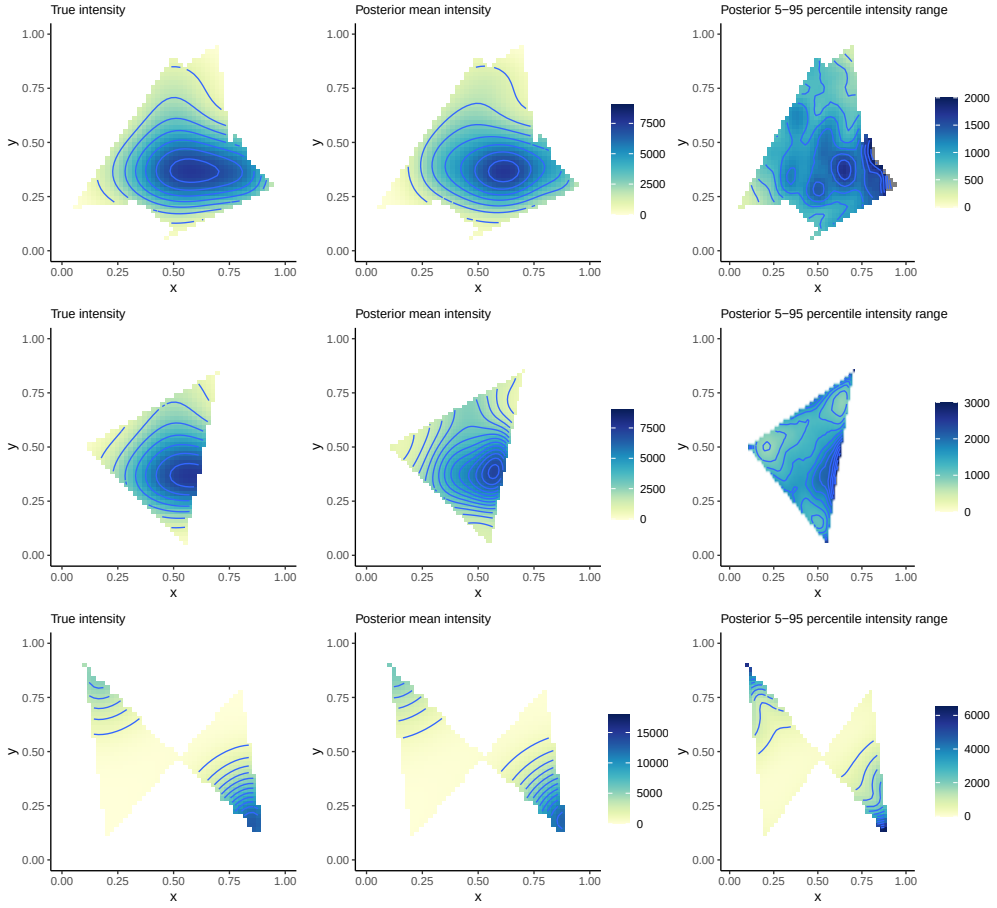


Figure 2.5: Results for the data in Fig. 2.4 under the intensity model. The left panel shows the true intensity function, the middle panel the posterior mean intensity estimate, and the right panel a posterior uncertainty estimate in the form of the difference between the 95th and 5th percentiles of the posterior distribution for the intensity function.

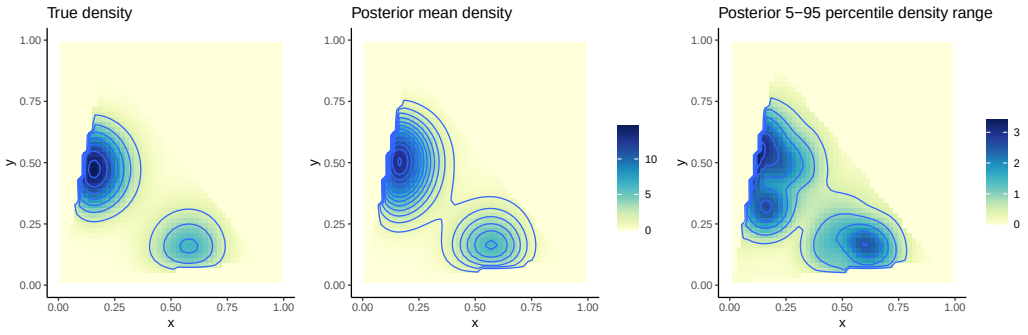


Figure 2.6: Results for the synthetic spatial point pattern generated from NHPP density $0.7\,\text{be}(x \mid 4, 17)\text{be}(y \mid 10, 11) + 0.3\,\text{be}(x \mid 12, 9)\text{be}(y \mid 4, 17)$ truncated to the triangle with vertices $\{(0.01, 0.01), (0.2, 0.9), (0.9, 0.1)\}$. The left panel includes the true density. Based on the density model, the middle panel plots the posterior mean density estimate, and the right panel an uncertainty estimate given by the difference between the 95th and 5th percentiles of the posterior distribution of the density function.

Fig. 2.6 show that the underlying bimodal density shape is recovered well, taking into account the moderate size of the point pattern. The posterior mean for the total intensity is 301.1, and the 95% posterior credible interval is given by $(267.3, 334.3)$. The 95% posterior credible interval for K is $[19, 25]$, and the posterior mode is 20, with $\Pr(K = 20 \mid \text{data}) = 0.46$. We note that increasing the size of the simulated point pattern results in posterior distributions for K that are more concentrated around $K = 20$.

2.5 Boston crime data analysis

For an illustration with real data, we consider the point pattern of $n = 1251$ locations in the city of Boston where vandalism occurred during the second quarter of year 2017; see the top left panel of Fig. 2.7. In general, spatial point patterns of crime depict more clustering than

what a NHPP can model. However, we use such data here to illustrate the spatial NHPP model over a non-trivial irregular domain, including model checking of the NHPP assumption.

The Boston City crime data and the Boston city boundary shape file in longitude and latitude format are publicly available online (Jain, 2018; BostonGIS, 2018). We use the `R rmapshaper` package (Teucher et al., 2021) to smooth this complicated boundary while retaining its key spatial topology. The simplified boundary in the form of Multipolygons is then mapped to a subset of the unit square. To process the raw data, we remove entries with geo-location as NAs, project the vandalism incidence locations from longitude and latitude into Northing and Easting, and finally map the crime locations and city boundary points to the unit square.

We focus on inference results under the density model, implemented with $C = 0.01$, a Ga$(5, 0.1)$ prior for α, and a truncated Poisson prior for K with mean 20 and support on $[20, 60]$. Fig. 2.7 plots posterior mean and uncertainty estimates for the intensity of vandalism incidences. The posterior mean for the total intensity of vandalism in the second quarter of 2017 is 1234, with the 95% posterior credible interval given by $(1167, 1303)$. The posterior distribution for K has effective support on $[36, 52]$ and posterior mode at 40 with Pr$(K = 40 \mid$ data$) = 0.34$.

For graphical model checking, we consider predictive residuals (Leininger and Gelfand, 2017a), defined as $N_{\text{pred}}(\mathcal{B}) - N_{\text{obs}}(\mathcal{B})$, where $N_{\text{obs}}(\mathcal{B})$ and $N_{\text{pred}}(\mathcal{B})$ are respectively the observed and predicted number of points in $\mathcal{B}$, a subset of the spatial point process domain $\mathcal{D}$. To sample from the posterior distribution of $N_{\text{pred}}(\mathcal{B})$, we draw from the Poisson$(\int\int_{\mathcal{B}} \lambda_{\mathcal{D}}(x, y)\, \mathrm{d}x\mathrm{d}y)$ distribution for each posterior realization of $\lambda_{\mathcal{D}}(x, y)$.

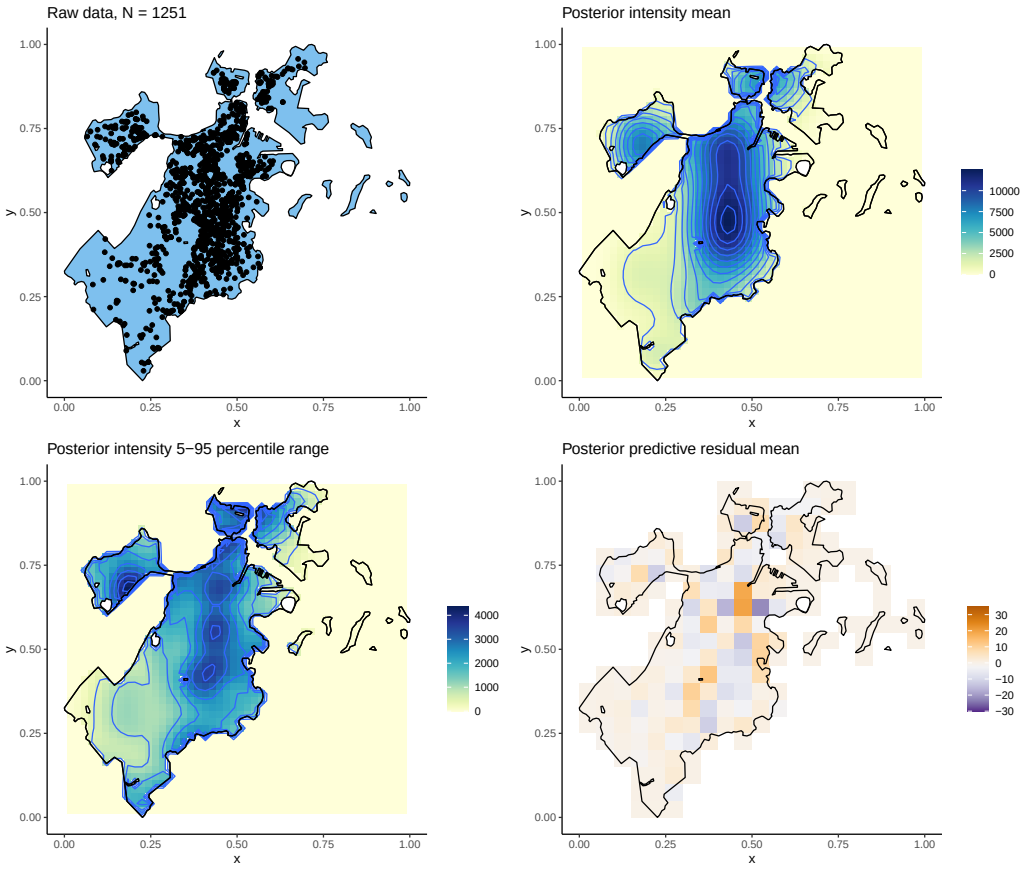


Figure 2.7: Boston crime data: vandalism in the second quarter of 2017. The observed point pattern is shown in the top left panel. Under the density model, the top right panel plots the posterior mean intensity estimate, and the bottom left panel the difference between the 95th and 5th percentile of the posterior distribution for the intensity function. The bottom right panel plots the posterior mean estimates for the predictive residuals.

We use the predictive residuals mainly as a graphical diagnostic tool to identify regions with potential misfit, and provide only qualitative comparison between models. Baddeley et al. (2005) defines a class of classical residuals for spatial point process and provides the corresponding means and variances, which can be used to calibrate such residuals. They define the raw residual, referred to later as the realized residual by Leininger and Gelfand (2017a),

as $N_{\text{obs}}(\mathcal{B}) - \int_{\mathcal{B}} \lambda(x, y)\mathrm{d}x\mathrm{d}y$. We follow the advice by Leininger and Gelfand (2017a) and use the predictive residuals defined above instead of the realized residuals, since the credible interval for the latter is not expected to achieve empirical coverage of 0. Notice that the raw residual essentially compares the observed count to the expectation of it in $\mathcal{B}$ and does not account for the additional variation from sampling. The predictive residuals however achieves an apples-to-apples comparison between observed counts and predicted counts and provides natural interpretation.

In general, lack of fit may be due to the NHPP assumption for the point process that generates the particular point pattern and/or the model used for the NHPP intensity. A flexible prior probability model for the NHPP intensity is practically useful in that it allows focusing discrepancies in the residuals on the NHPP assumption.

To implement model checking with predictive residuals, we create a 20×20 grid over the unit square and select the subset of these 400 square regions that overlap with the Boston city boundary $\mathcal{D}$ as the target regions to cover the entire Boston city. The bottom right panel of Fig. 2.7 plots the posterior mean estimates for the predictive residuals. The residuals in regions near the city boundary are evaluated based on only the subsets that overlap with $\mathcal{D}$. This residual analysis suggests a decent fit of the NHPP model. It is perhaps not surprising that the sub-regions with the more pronounced non-zero residual estimates correspond to parts of the city where the data suggest clustering, for which a more general point process than the NHPP would be expected to provide better model fit.

2.6 Discussion

We have presented two models for spatial NHPP intensities over domains with irregular shapes. To our knowledge, this is the first treatment of this practically relevant problem with methodology that supports general intensity function shapes and allows for full Bayesian inference, while avoiding any type of approximation of the NHPP likelihood.

In the more commonly studied setting of a regular domain, the two modeling approaches result in the same formulation for the NHPP intensity, which corresponds to a Bernstein-Dirichlet prior for the associated NHPP density. Hence, as a useful byproduct of the methodology, we establish a connection between density and intensity estimation under Bernstein-Dirichlet priors. Relative to existing approaches that model directly the intensity function over regular domain, the proposed method arguably offers a substantially more practical inference framework. The prior model for the intensity function can be equivalently represented in terms of a prior for the total intensity over the observation domain and a prior for the density function. In contrast with related existing methods, the priors for the NHPP density and the total intensity are guaranteed to be compatible with the prior for the NHPP intensity.

The two proposed models for spatial NHPPs over irregular domains $\mathcal{D}$, the intensity model (2.7) and density model (2.9), arise from different perspectives. The former model builds from truncating the Bernstein-Dirichlet density model over $\mathcal{D}$, whereas the latter constructs the irregular domain density as a mixture of truncated beta basis densities. The intensity model uses all K^2 basis densities $\{\phi^*_{k_x,k_y}\}$ and relies on random weights, further adjusted by the normalizing constants B_{k_x,k_y}, to select appropriate basis members in constructing the intensity

functional form. The density model is generally more efficient in the intensity representation, as it utilizes a subset of the K^2 basis densities $\{\phi^*_{k_x,k_y}\}$, the size of such subset determined by the particular domain $\mathcal{D}$. For settings where a value for K can be specified, possibly appealing to empirical experience with synthetic data examples, the intensity model offers the benefit of particularly simple and efficient model fitting. The density model affords more generality in the inference scheme by allowing uncertainty with respect to the number of basis densities, at the cost of a more involved posterior simulation method, which however does not require complex trans-dimensional computational techniques. For both models, the intensity representation through beta densities with specified parameters is essential for the practicality and computational efficiency of the inference methods for spatial NHPPs over irregular domains.

The two proposed models for the spatial NHPPs use only the event point pattern data and can be extended to incorporate extra information. Here we focus on two types of additional variables that potentially enrich the modeling: covariates and marks. The covariates under the context of point process modeling are often spatially dependent, but are treated as fixed variable within a pre-specified tract. For example, the population density varies over the country but is fixed for a given county. A general approach to incorporate covariates for NHPP models is to let the intensity function be dependent on the covariates, such that the new intensity function takes the following form:

$$\lambda_{\mathcal{D}}(\boldsymbol{s}, x(\boldsymbol{s}) \mid \boldsymbol{\beta}) = \lambda_{\mathcal{D}}(\boldsymbol{s}) \exp(x(\boldsymbol{s})^T \boldsymbol{\beta})$$

where $x(\boldsymbol{s})$ is a set of spatially varying covariates and $\boldsymbol{\beta}$ is the corresponding regression coefficients. $\lambda_{\mathcal{D}}(\boldsymbol{s})$ can be formulated using either the intensity or density model introduced in this

chapter. On the log-scale, the nonparametric intensity function $\lambda_{\mathcal{D}}(\boldsymbol{s})$ can be thought as an error term that accounts for spatial heterogeneity while the mean level of occurrence is specified by the regression term. This approach is similar to the log Gaussian Cox process where the intensity on the log scale has a mean dependent on the covariates and an error controlled by the Gaussian process.

The challenge with this approach is to compute the normalizing constants. Certain approximation is required for such computation since the total intensity, now defined as $\Lambda_{\mathcal{D}} = \int_{\mathcal{D}} \lambda_{\mathcal{D}}(\boldsymbol{s}) \exp(x(\boldsymbol{s})^T \beta) \mathrm{d}\boldsymbol{s}$, no longer has a closed-form expression. One could discretize the nonparametric intensity $\lambda_{\mathcal{D}}(\boldsymbol{s})$ over a fine grid over $\mathcal{D}$ and use Monte Carlo integration by evaluating $\lambda_{\mathcal{D}}(\boldsymbol{s}) \exp(x(\boldsymbol{s})^T \boldsymbol{\beta})$ at the centroids of the grid cells.

A mark can be thought of as a random variable that is only generated because the event occurs. It can be modeled with a distribution whose density factors into the NHPP intensity function. For the spatial Poisson process, consider random marks $y_i \equiv y_{\boldsymbol{s_i}} \in \mathcal{M}$ associated with events at $\boldsymbol{s_i}$. The marked NHPP is then characterized by the intensity function $\lambda^*_{\mathcal{D}}(\boldsymbol{s_i}, y_i) = \lambda_{\mathcal{D}}(\boldsymbol{s_i}) m_{\boldsymbol{s}}(y_i)$, where $m_{\boldsymbol{s}}(\cdot)$ is the density function for the marks. The likelihood for the observed point pattern $\{(\boldsymbol{s_i}, y_i) : i = 1, \cdots, n\}$ can be written as

$$L(\{(\boldsymbol{s_i}, y_i) : i = 1, \cdots, n\}) = \exp(\int_{\mathcal{D}} \lambda_{\mathcal{D}}(\boldsymbol{s}) \mathrm{d}\boldsymbol{s}) \prod_{i=1}^{n} \lambda_{\mathcal{D}}(\boldsymbol{s_i})) \prod_{i=1}^{n} m_{\boldsymbol{s}}(y_i)$$

where the normalizing constant is a result of simplification since $\int_{\mathcal{D}} \int_{\mathcal{M}} \lambda^*_{\mathcal{D}}(\boldsymbol{s}, \boldsymbol{u}) \mathrm{d}\boldsymbol{u} \mathrm{d}\boldsymbol{s} = \int_{\mathcal{D}} \lambda_{\mathcal{D}}(\boldsymbol{s}) \mathrm{d}\boldsymbol{s}$, since $m_{\boldsymbol{s}}(y_{\boldsymbol{s}})$ is a density. One can combine our proposed nonparametric model for the NHPP intensity $\lambda_{\mathcal{D}}(\boldsymbol{s})$ and a model for the marks distribution to model marked NHPP, as discussed in Taddy and Kottas (2012).

The intensity model proposed in this chapter serves as the building block in models for the more general spatial point processes, the spatial Hawkes and space-time Hawkes process. From now on, we will refer to the intensity model for NHPP as BPNHPP. The next two chapters explore Bayesian semiparametric modeling for the spatial and space-time Hawkes processes to account for potential self-triggering effects of the criminal activities.

Chapter 3

Bayesian semi-parametric modeling for spatial Hawkes processes

3.1 Introduction and motivation

The temporal Hawkes process, first introduced by Hawkes (1971) as a self-exciting process, is characterized by the conditional intensity function, which represents the rate of event at time t as the sum of the rate of background events and rates of events triggered by previous events up to time t. Let $\mathcal{H}_t$ denote the history of events up to time t, the intensity function at time t conditional on the history is $\lambda(t|\mathcal{H}_t) = \mu(t) + \sum_{i:t_i<t} g(t-t_i)$, where $\mu(t)$ is the background event intensity function, and $g(t-t_i)$ is the trigger function that depends on events prior to t.

The temporal Hawkes process has an equivalent representation as the superposition of many generations of Poisson cluster processes conditional on the branching structure (Hawkes

and Oakes, 1974). The Poisson cluster process, which is a special case of the general Cox process, is defined by a parent Poisson process that describes the point pattern that serves as the cluster centers, and an offspring Poisson process that describes the point pattern within each cluster that is centered on points from the parent process (Daley and Vere-Jones, 2003). The Hawkes process takes a step further and allows the offspring in each cluster to generate their own offspring as a new Poisson cluster process. The resulting point pattern taken as the superposition of point patterns across all generations is equivalent to a realization from a point process defined by the Hawkes process conditional intensity function.

Specifically, the hierarchical model starts with a Poisson process with intensity $\mu(t)$ that generates the first generation of points that serve as the initial cluster centers. Then offspring points are generated according to offspring Poisson processes with intensity function $g(t - t_i)$, for each parent point t_i. Such intensity function is often referred to as the *triggering function*. The branching structure, which describes the generation and parentage for each point, can be modeled as augmented variables in the hierarchical model. These latent variables indicate the index for each point's parent and represent the full branching information. Introducing these latent variables reduces modeling the Hawkes process through the conditional intensity function to modeling a series of Poisson processes.

Certain restriction on the triggering function is required for the Hawkes process to have finite realization, or from the simulation point view, to stop generating further offsprings and not to blow up. In particular, the triggering function is required to satisfy the following condition: $0 < \gamma = \int_0^\infty g(u)\mathrm{d}u < 1$, where γ is referred to as the *branching ratio* and controls the average number of offspring points generated in a family.

The space-time self-exciting process provides an intermediate step to extend the Hawkes process from temporal to spatial applications; see the review by (Reinhart, 2018). The space-time Hawkes process is defined via a conditional intensity function

$$\lambda(\boldsymbol{s}, t|\mathcal{H}_t) = \mu(\boldsymbol{s}, t) + \sum_{i:t_i<t} g(\boldsymbol{s} - \boldsymbol{s_i}, t - t_i).$$

The background intensity function $\mu(\boldsymbol{s}, t)$ controls the immigrant Poisson process; the triggering function $g(\boldsymbol{s}, t)$ controls the offspring Poisson processes for events triggered by previous events. Such conditional intensity function incorporates spatial proximity in the triggering function in a fashion that only partially impacts the branching structure. Just like in the temporal Hawkes process, the order of event time in the space-time Hawkes process defines the parameter space of the branching structure since only the events prior to time t can serve as parent for event at t. This class of model has limitations therefore in a spatial applications where the time of event is not available, or spatial proximity directly impacts the branching structure.

Such limitations inspire the study of the spatial Hawkes process, where the branching structure is purely based on spatial proximity without information on event time. The lack of temporal information makes it less obvious how to define the $\mathcal{H}_t$ in the conditional intensity, though later we will explain that certain constraints on the branching structure still exist.

The clustering representation of the spatial Hawkes process provides better understanding of the process and naturally inspires a Bayesian hierarchical modeling approach. Proposed by Møller and Torrisi (2007) when modeling a general form of clustering Point process, the General Shot Noise Cox Process (GSNCP), the clustering representation of the spatial Hawkes process is formalized to be a superposition of a Poisson process and many later gener-

ations of GSNCPs clustering on points in previous generations.

To our knowledge, there is no work in the literature on modeling and inference for the spatial Hawkes processes. There is however, such literature for space-time Hawkes process. The frequentist inference methods for space-time Hawkes processes utilize the conditional intensity function that leads to either the likelihood (e.g. Veen and Schoenberg, 2008; Peng et al., 2005) or the partial likelihood (e.g. Diggle, 2006). A technique called stochastic declustering infers and labels each point as either from the immigrant process or an offspring process in an iterative fashion. This technique has inspired a series of methods that construct nonparametric estimators for the immigrant process intensity function from the immigrant points and a parametric estimator for the offspring density (Zhuang et al., 2002; Chiodi and Adelfio, 2011), or model the offspring density with a data-driven approach (Marsan and Lengliné, 2008). The likelihood computation involves an integral with respect to the conditional intensity function that can be numerically unstable, as suggested by Veen and Schoenberg (2008). Instead they propose an Expectation maximization (EM) algorithm with the complete data likelihood defined by the hierarchical formulation with the latent branching structure which proves to be numerically more stable than maximum likelihood estimation. Rasmussen (2013) proposed two Bayesian estimation methods for the temporal Hawkes process, using MCMC on the complete data likelihood, with a Metropolis update on the branching structure within a Gibbs sampler, and a model with augmented parameters to apply a Gibbs sampler step to the latent branching structure. From this brief review, we notice that most approaches avoid using the conditional intensity directly due to computation complexity and instead utilize the clustering representation to achieve inference in a hierarchical or at least iterative fashion.

We show that the clustering representation also exists for the spatial Hawkes process conditional on the latent branching structure with modified constraints on its parameter space due to the lack of natural order in space. We obtain a hierarchical formulation of the spatial Hawkes process using its connection to the GSNCP to propose a Bayesian model framework that incorporates the latent branching structure as part of the model to provide easy updates in the posterior simulation (Section 3.2). Under this Bayesian framework, we present a class of models that are permuted from three aspects of modeling choices: the shape of observation window, the type of immigrant process, and the functional form of the parametric offspring density (Section 3.3). We discuss model checking and model comparison techniques (Section 3.4) and demonstrate model capacity with synthetic data (Section 3.5, Section 3.6) and a real data example on crime in Boston city (Section 3.7).

3.2 Spatial Hawkes Processes

The spatial Hawkes process (Møller and Torrisi, 2007) is defined as the superposition of countable generations of point processes, $X = \bigcup_{n=0}^{\infty} G_n, G_n \in \mathbb{R}^2$, such that the point process for generation G_{n+1}, conditional on the previous generation G_n, is a Poisson process on $\mathbb{R}^2$ with the following intensity function:

$$\lambda_{n+1}(\boldsymbol{s}) = \sum_{\boldsymbol{s_j} \in G_n} \boldsymbol{\gamma}(\boldsymbol{s} - \boldsymbol{s_j}) \tag{3.1}$$

where $\boldsymbol{s}$ is the location for a point in G_{n+1} and $\boldsymbol{s_j}$ is the location of a point in G_n. Here, $\boldsymbol{\gamma}(\cdot)$ is a spatial intensity function for any location shift $\boldsymbol{s} - \boldsymbol{s_j}$. The points in G_0 are referred to as the *immigrants*, and those in later generations, $G_n, n > 1$, as the *offsprings*. The point process that

generates G_0, called the immigrant process, is a Poisson process on $\mathbb{R}^2$ with intensity function $\mu(\boldsymbol{s})$.

Just like the temporal Hawkes process defined as a multi-generation Poisson cluster process, the spatial Hawkes process can be considered as the superposition of a series of Cox processes. The Cox process is a Poisson process with a random intensity function. Møller and Torrisi (2005) defines the point process with intensity function in the form of (3.1) as a special case of the GSNCP. In GSNCP, the random intensity function $\lambda(\boldsymbol{\epsilon})$ is defined through a point process Φ on $\mathbb{R}^d \times (0,\infty), \times(0,\infty)$. Let $\boldsymbol{\nu_i}$ be the set of points in Φ, and $\boldsymbol{\epsilon}$ be the location of a point from a realization from the GSNCP. The random intensity function $\lambda(\boldsymbol{\epsilon})$ is defined as follows

$$\lambda(\boldsymbol{\epsilon}) = \sum_{(\boldsymbol{\nu_i},\gamma_i,b_i)\in\Phi} \gamma_i k_{b_i}(\boldsymbol{\nu_i},\boldsymbol{\epsilon}) \tag{3.2}$$

where $\gamma_i > 0$ is the total intensity for the Poisson process and $k_{b_i}(\cdot,\cdot)$ is a kernel density function with bandwidth b_i.

The GSNCP can be viewed as a Cox process where each point $\boldsymbol{\nu_i}$ generates offspring independently following a Poisson point process with intensity $\gamma_i k_{b_i}(\boldsymbol{\nu_i},\boldsymbol{\epsilon})$. Møller and Torrisi (2005) establish the connection between the spatial Hawkes Process and the GSNCP by recognizing that $G_{n+1}|G_n$ in a spatial Hawkes process is a GSNCP where the bandwidth parameter b_i is fixed and identical for all i, and Φ is the point process for G_n in (3.2). We follow such construction and assume that the offspring density kernel is controlled by the same set of parameters $\boldsymbol{\theta_o}$. We thus omit the subscript b_i from the offspring density function and rewrite it as $k(\boldsymbol{\epsilon}|\boldsymbol{\nu_i},\boldsymbol{\theta_o})$. We assume that the total intensity γ_i is the same for all offspring Poisson processes and denote this shared parameter as γ. This total offspring intensity is also known

as the *branching ratio* and it directly controls how many sub-branches are spawn from a single parent point. We change notation such that $\boldsymbol{s}$ denotes the location of a point in G_{n+1} and $\boldsymbol{s_j}$ the location of the j-th point in G_n. Under these assumptions, the intensity function of $G_{n+1}|G_n$ in (3.1) is simply

$$\lambda_{n+1}(\boldsymbol{s}) = \sum_{\boldsymbol{s_j} \in G_n} \lambda(\boldsymbol{s}|\boldsymbol{s_j}, \gamma, \boldsymbol{\theta_o}) = \sum_{\boldsymbol{s_j} \in G_n} \gamma k(\boldsymbol{s}|\boldsymbol{s_j}, \boldsymbol{\theta_o}) \tag{3.3}$$

Furthermore, by the superposition theorem for Poisson processes (Kingman, 1992), G_{n+1} with the intensity λ_{n+1} is the superposition of N_n independent Poisson processes, where N_n is the number of points in G_n.

Let $\mathcal{T}$ be the full branching structure that specifies the family tree across all generations. Let $\boldsymbol{\theta} = \{\boldsymbol{\theta_I}, \boldsymbol{\theta_o}\}$ be the vector of parameters for the immigrant and offspring Poisson process intensities. The hierarchical formulation for the spatial Hawkes process X given the branching structure is the following:

$$\begin{gathered} X|\mathcal{T} \equiv \textstyle\bigcup_{n=0}^{\infty} G_n \\ G_n|G_{n-1}, \mathcal{T} \equiv \textstyle\bigcup_j \text{NHPP}(s \in O_j|\boldsymbol{\theta_o}) \quad s \in G_n, s_j \in G_{n-1} \\ G_0 \equiv \text{NHPP}(\boldsymbol{\theta_I}) \end{gathered} \tag{3.4}$$

where O_j is the set of points in G_n generated by the Poisson process centered on location $\boldsymbol{s_j}$ in G_{n-1}. Conditional on the latent parent label j for each point in G_{n+1}, the complete data likelihood for G_{n+1} is the product of N_n Poisson process likelihoods terms, each with intensity $\gamma k(\boldsymbol{s} \mid \boldsymbol{s_j}, \boldsymbol{\theta_o})$, where $s_j \in G_n$. Recall that the spatial Poisson process likelihood, based on point pattern $\{\boldsymbol{s_1}, \cdots, \boldsymbol{s_n}\}$ observed in $\mathcal{D}$, is fully specified by the intensity function $\lambda(\boldsymbol{s})$. We

denote such likelihood by $\text{PP}(\cdot \mid \lambda(\boldsymbol{s}))$,

$$\text{PP}(\{\boldsymbol{s_1}, \cdots, \boldsymbol{s_n}\}; \lambda(\boldsymbol{s})) \propto \exp\Big(-\int_{\mathcal{D}} \lambda(\boldsymbol{s}) \mathrm{d}\boldsymbol{s}\Big) \prod_{i=1}^{n} \lambda(\boldsymbol{s_i})$$

We then obtain the likelihood for points in G_{n+1} as a product of Poisson process likelihoods as follows:

$$p(G_{n+1}, G_n) = p(G_{n+1}|G_n)p(G_n) = \left\{ \prod_{\boldsymbol{s_j} \in G_n} \text{PP}\Big(\{\boldsymbol{s} \in O_j\}; \lambda(\boldsymbol{s}) = \gamma k(\boldsymbol{s}|\boldsymbol{s_j}, \boldsymbol{\theta_o})\Big) \right\} p(G_n) \tag{3.5}$$

We use this formulation to obtain the complete-data likelihood for the spatial Hawkes process conditional on the latent branching structure constructed in a similar fashion as in the temporal case (Rasmussen, 2013). The key is to recognize that the branching structure information required by the likelihood is fully specified by the parent label of each point, with the special case of immigrants with no parents. Therefore, we use $[i]$ to denote the index for the parent point of i, which indicates that point i is generated from the Poisson process centered on point $[i]$, and let $[i] = 0$ for immigrant points. Let the spatial point pattern data be denoted as the set of locations $\boldsymbol{S} = \{\boldsymbol{s_1}, \cdots, \boldsymbol{s_n}\}$. By applying (3.5) recursively, and using O to denote the collection of all offspring points such that $O = \cup_{j=1}^{n} O_j$, we obtain the complete-data likelihood conditional on parent label $[i]$ as the following

$$p(\boldsymbol{S}|\ [i]) = \left\{ \prod_{j=1}^{n} \text{PP}\Big(\{\boldsymbol{s_i} : [i] = j\}; \lambda(\boldsymbol{s_i}) = \gamma k(\boldsymbol{s_i}|\boldsymbol{s_{[i]}}, \boldsymbol{\theta_o})\Big) \right\} \cdot \text{PP}(\{\boldsymbol{s_i} : [i] = 0\}; \mu(\boldsymbol{s_i})) \tag{3.6}$$

Notice that the last generation of points, which are the leaf nodes in the branching trees, will have no children in reality, meaning that the offspring Poisson processes centered on these

points only contribute their normalizing constants to the likelihood.

There is a subtle difference between the branching structure in the spatial Hawkes process and that in the temporal one. In the temporal case, there exists a natural order of events implied by the timestamp of the points, since only points that occurred prior to time t can be the parent for point at t. Such natural order of events makes it easy to determine the valid parent set for a given point according to chronological order, but it does not exist in spatial point processes. The issue is that the spatial location of an event does not directly restrict the lineage among points, which makes the parameter space for the branching structure more complex compared to its counterpart in the temporal case. However, some restrictions on the valid parent set of point i still exist when we represent the lineage using a set of trees. Each tree starts with a specific immigrant point as the root node and branches off hierarchically to a set of offspring points as the descendants of such immigrant point until reaching the leaf nodes which have no children. Based on this representation, the parent of an offspring point is the point one generation closer to the immigrant on the same branch in the latent family tree set $\mathcal{T}$. And the immigrant points have no parent.

Inference on this latent structure requires a valid proposal to mutate the tree structure. Conditional on the current branching structure, such proposal needs to satisfy the following condition: a point in the proposed tree structure cannot be the parent of any points that descend from itself in the current structure. Fig 3.1 illustrates a comparison of valid proposal for branching structures in the temporal and spatial Hawkes processes. The observed point locations for the temporal Hawkes process are the timestamps over the positive real line. Here the timestamp is indexed by point index, not by chronological order. The coordinates of nodes in the

right panel are the locations for the observed realization from the spatial Hawkes process. The nodes and arrows illustrate the current latent branching structures. In the temporal case, a valid proposal for the branching structure needs to respect the chronological order. For example, the node 2 cannot be the offspring of node 6, since $t_2 < t_6$. In the spatial case however, node 2 can be proposed to be the offspring of point 6, since point 6 is not a descendent of point 2 based on the current branching structure. In the temporal case, we can easily get the available parent set for a point i as the set of points j with $t_j < t_i$. In the spatial case, we need the complete family tree $\mathcal{T}$ for each point in G_0 and choose parents for i from the points that are not direct descendants from i.

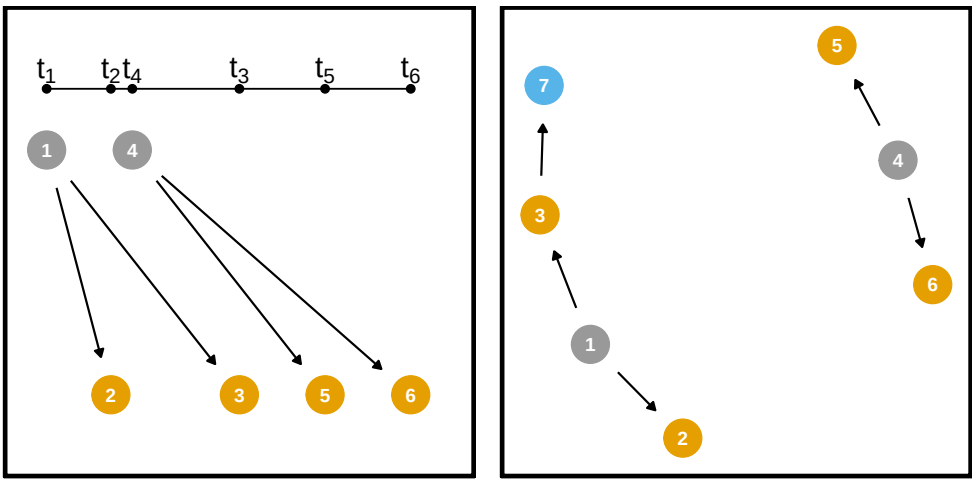


Figure 3.1: Branching structure in temporal (left panel) and spatial (right panel) point processes. In both panels, the node's color indicates generation (gray = G_0, yellow = G_1, blue = G_2); the arrows suggest parent-children relationship. The left panel shows the node location over the positive real line; the right panel over the unit square.

To sample from the spatial Hawkes process, we will follow the hierarchical model formulation and sample G_{n+1} conditional on G_n from independent Poisson processes. The

simulation follows steps as follows:

1. Sample G_0 from a Poisson process with intensity function $\mu(\boldsymbol{s})$. We first sample N_0, the total number of points in G_0, from a Poisson distribution with mean Λ, where $\Lambda = \int_{\mathcal{D}} \mu(\boldsymbol{s}) d\boldsymbol{s}$ for NHPP or simply $\mu \mid \mathcal{D} \mid$ for HPP. Next, we sample N_0 points i.i.d from the density function $f_0(\boldsymbol{s}) = \mu(\boldsymbol{s})/\Lambda$.

2. Sample G_{n+1} conditional on points in G_n from N_n independent Poisson processes for $n = 0, 1, \cdots$. The Poisson process centered on $\boldsymbol{s_i}, i = 1, \cdots, N_n$, has total intensity γ and density $k(\boldsymbol{s}|\boldsymbol{s_i}, \boldsymbol{\theta_o})$. The union of realizations from the N_n Poisson processes forms G_{n+1} with size N_{n+1}.

3. Repeat step 2 until no further offspring is generated. For the Hawkes process to be stable, or stop generating offspring within finite generations, the branching ratio γ needs to satisfy the condition $0 < \gamma < 1$.

3.3 Bayesian semi-parametric modeling framework for spatial Hawkes processes

We provide model formulations for a class of spatial Hawkes processes that are assumed to have strict support over some compact domain $\mathcal{D}$. Such assumption implies that all parents generate offspring points only within $\mathcal{D}$. Equivalently, the parent of any offspring point is among the observed point. Such assumption is valid for most applications where the self-exciting behavior is effective in a small range from the parent event. For application in crime

forecast, we can safely assume that similar crimes tend to occur in specific neighborhoods and have limiting effects outside the city boundary.

In this section, we discuss modeling choices from the following three aspects: the shape of the support $\mathcal{D}$, the type of the immigrant process, and the choice of offspring density. We describe the model formulation under each configuration and the corresponding posterior simulation details.

3.3.1 Modeling for point patterns over an irregular domain $\mathcal{D}$

The default choice for the domain $\mathcal{D}$ for a spatial point pattern is the unit square. However, for most applications, events only occur in bounded regions within the unit square, and such regions' boundaries are often highly irregular. We refer to the unit-square as the regular domain and any strict subset of the unit-square as the irregular domain. The shape of $\mathcal{D}$ has significant implications for modeling the spatial Hawkes process since taking $\mathcal{D}$ to be the unit-square in specific applications may violate the strict support assumption. One example is modeling tree locations in a forest, which often has a boundary that maps to a subset of the unit square. Choosing the boundary to be the unit-square allows offspring points to occur outside of the forest boundary where no trees can survive in reality. On the contrary, the crime pattern in certain cities can have support over the unit square, especially when the city boundary is almost rectangular. From the modeling perspective, the choice of $\mathcal{D}$ poses constraints for the immigrant and offspring processes model, since both immigrant and offspring Poisson densities have the same support $\mathcal{D}$. We will address how our models account for irregular domain $\mathcal{D}$ as a key contribution of this work.

3.3.2 Model for the immigrant process

The choice of the immigrant process intensity largely depends on the application. The homogeneous Poisson process is suitable for applications where the inhomogeneity in the point pattern is assumed to be mainly driven by offspring clustering. In these cases, it is reasonable to assume that the immigrant point pattern is simple and sparse, whereas the offspring point pattern is a complex result of the evolution of many generations. One example of such a use case is the point pattern of a forest. The first generation of trees may have appeared in the forest in a completely random fashion. The clusters of trees as observed today result from many generations of reproduction centered on parent trees that have appeared earlier.

Alternatively, the NHPP is more suitable when the immigrant point pattern is complex, and the offspring process consists of fewer generations with sparse patterns to account for a small amount of local clustering. A good example is the locations of crimes in a city in a given period since it is a common practice to profile crime patterns via hot spot analysis, which suggests varied levels of crime rate dependent on location. The initial crimes are treated as a realization from a NHPP whose intensity function captures these hot spots well. Furthermore, the triggering effect of these initial crimes tends to fade quickly and therefore it produces fewer generations of triggered crimes as offspring.

The ideal NHPP model for the immigrant process in a spatial Hawkes process should provide both flexible inference and computational efficiency while accounting for the irregular observation window in the model construction. The computational requirements come from the fact that the NHPP serves as a latent layer in the full hierarchical model that relies on

the Markov Chain Monte Carlo algorithm for posterior inference. Fast posterior sampling for the latent NHPP parameters reduces the computational cost for each iteration of the posterior simulation. Based on these considerations, we choose the intensity formulation from the model proposed in Section 2.3.1. The model is constructed on the connection between a simple prior placed on a representation of the intensity function as a weighted sum of beta density functions to a Bernstein-Dirichlet process prior placed on the NHPP density function. Such connection is the key for the model to achieve efficient posterior simulation for full Bayesian inference while providing rich prior support for the intensity function. Lastly, the model accounts for the irregular domain as part of the model construction at minimal additional computation cost.

Under the intensity formulation, we represent the immigrant intensity function as a weighted combinations of the spatial Bernstein densities, $\phi_{k_x,k_y}(x,y) = \text{be}(x|k_x, K-k_x+1)\text{be}(y|k_y, K-k_y+1)$, where $(k_x,k_y) \in \{(1,1),(1,2),\cdots,(K,K)\}$.The model for the NHPP intensity function is the following:

$$\begin{gathered}\mu(x,y) = \textstyle\sum_{k_x,k_y=1}^{K} V_{k_x,k_y}\phi_{k_x,k_y}(x,y), (x,y) \in [0,1]^2 \\ V_{k_x,k_y}|\alpha \overset{ind.}{\sim} \text{Ga}(\alpha/K^2, C), k_x, k_y = 1\cdots K\end{gathered} \tag{3.7}$$

This intensity model implies the following mixture representation for the NHPP density function $f(x,y) = \mu(x,y)/\int\int_{[0,1]^2}\mu(x,y)\text{d}x\text{d}y = \sum_{k_x,k_y=1}^{K}\omega_{k_x,k_y}\phi_{k_x,k_y}(x,y)$, where $\{\omega_{k_x,k_y}\} \sim \text{Dir}(\{\alpha_k = \alpha/K^2, k = 1\cdots K^2\})$. The particular Dirichlet distribution connects this mixture representation to a Bernstein-Dirichlet prior placed directly on the NHPP density function. This connection achieves flexible inference using the corresponding intensity model (3.7), since the Bernstein-Dirichlet prior has nice properties such as uniform convergence (Levasseur, 1984) and posterior consistency for density estimation (Petrone and Wasserman, 2002).

To accommodate for the irregular domain, we truncate the NHPP density function over the unit-square to $\mathcal{D}$: $f_{\mathcal{D}} = f(x,y)/\int\int_{\mathcal{D}} f(u,v)\mathrm{d}u\mathrm{d}v = \sum_{k_x,k_y}^{K} V_{k_x,k_y} B_{k_x,k_y} \Lambda_{\mathcal{D}}^{-1} \phi^*_{k_x,k_y}(x,y)$, where $\Lambda_{\mathcal{D}} = \sum_{k_x,k_y}^{K} V_{k_x,k_y} B_{k_x,k_y}$ and ϕ^* denotes the truncated Bernstein densities with normalizing constants B_{k_x,k_y}. Conditional on the parent label $[i]$, we model the collection of immigrant points for which $[i] = 0$ with the following NHPP intensity function

$$\mu(x_i, y_i) = \sum_{k_x,k_y=1}^{K} V_{k_x,k_y} B_{k_x,k_y} \phi^*_{k_x,k_y}(x_i, y_i) \quad (x_i, y_i) \in \mathcal{D}, [i] = 0 \tag{3.8}$$

Incorporating this formulation to model the immigrant intensity in the spatial Hawkes process is straightforward. The only difference between modeling the immigrant Poisson process and an actual Poisson process is that the observed locations in the former come from a random subset of a larger point pattern and can change in the posterior simulation depending on the current latent branching structure. The varying number of observations across MCMC iterations means that the number of latent index parameters (ξ_i, η_i) also changes. This does not cause issue for the intensity formulation, since (ξ_i, η_i) are independent for i conditional on V_{k_x,k_y} in the posterior. Alternatively, applying the density formulation becomes problematic because the latent parameters $\boldsymbol{z_i}$ are conditionally dependent in the posterior. Sampling $\boldsymbol{z_i}$ from the posterior full conditional depends on values for all $\boldsymbol{z}_{-i}$. There is no default value for $\boldsymbol{z}_i$ in the current posterior sample when i is not an immigrant in this iteration but becomes an immigrant in the next. As a result, we apply the intensity formulation to the subset of point patterns with immigrant identity in each posterior simulation iteration and define immigrant parameter set $\boldsymbol{\theta_I}$ to be $\{V_{k_x,k_y}, \alpha\}$ and specify values for K and C.

3.3.3 Offspring density choices

We consider two parametric forms for the offspring density: a bivariate beta density over the unit-square or a bivariate normal density truncated to $\mathcal{D}$. Given the strict support assumption, the bivariate beta density is only suitable for regular domain $\mathcal{D}$, whereas the truncated bivariate normal density can be applied to any compact domain $\mathcal{D}$ when proper truncation is enforced.

The bivariate beta distribution is constructed as a product of two independent beta densities in the x and y dimensions, each parameterized by a mode and a dispersion parameter. Such parameterization allows the offspring density to be centered at its parent location $(x_{[i]}, y_{[i]})$ by defining the modes of the two univariate beta densities accordingly: $k(\boldsymbol{s_i}|\boldsymbol{s_{[i]}}, \tau) = \text{beta}(x_i|x_{[i]}, \tau)\ \text{beta}(y_i|y_{[i]}, \tau)$ where the density function $\text{beta}(\cdot|m, \tau)$ has mode m and dispersion parameter τ. The new parameters m and τ map to the conventional beta parameterization $\text{beta}(a, b)$ with expectation $a/(a+b)$ via $a = m(\tau - 2) + 1, b = (1 - m)(\tau - 2)$.

The offspring density under such formulation has variance dependent on the parent location. The variance of the marginal beta distribution, $\mathbb{V}(x) = m(1-m)(\tau-2)^2 + \tau - 1/(\tau^2(\tau+1))$, depends on the parent's location which defines the mode m. The variance of the marginal density in the x or y dimension achieves a maximum at $m = 0.5$ and decreases as m approaches 0 or 1. Thus, the effective density range is small near the boundary but large at the center. Simulation suggests that given a small dispersion parameter, the shape of the offspring density does not vary significantly across space.

More generally, we adopt a bivariate normal density centered at the parent location

and truncated to the compact domain $\mathcal{D}$ for the offspring density. We specify the covariance matrix Σ with three parameters, σ_x, σ_y, ρ, to control the spread in the x and y direction and the spatial correlation. The offspring kernel conditional on the parent location is thus $k(\boldsymbol{s_i}|\boldsymbol{s_{[i]}}, \Sigma) \propto I_{[\boldsymbol{s_i} \in \mathcal{D}]} \mathrm{N}_2(\boldsymbol{s_i}|\boldsymbol{s_{[i]}}, \Sigma)$. Computing the normalizing constant in this truncated distribution can be expensive, especially because such computation happens for all families at each iteration of the posterior simulation. We design a special Monte Carlo routine to cache some of the computations, which improves the algorithm's speed remarkably. We will discuss the details of such implementation in Section 3.3.5.

3.3.4 Model formulations

We first describe the complete data likelihood for the spatial Hawkes process conditional on the branching structure. Let $\{[i]\}$ be the set of latent variables to denote the index of parent for point i, $i = 1 \cdots, n$. Let $k(\boldsymbol{s_i}|\boldsymbol{s_{[i]}}, \boldsymbol{\theta_o})$ be the general form of a parametric offspring density function for observation i with mode at its parent location $\boldsymbol{s_{[i]}}$ and the offspring parameter vector $\boldsymbol{\theta_o}$. Let γ denote the branching ratio, i.e., the total intensity for any individual offspring Poisson process. The complete data likelihood conditional on parent information $[i]$ is given by

$$
\begin{aligned}
&L(\boldsymbol{s}|\{[i]\}, \boldsymbol{\theta_I}, \boldsymbol{\theta_o}) \\
&\quad = \exp(-\int_{\mathcal{D}} \mu(\boldsymbol{s}|\boldsymbol{\theta_I}) d\boldsymbol{s} - \gamma \sum_{j=1}^{N} \int_{\mathcal{D}} k(\boldsymbol{u}|\boldsymbol{s_j}, \boldsymbol{\theta_o}) d\boldsymbol{u}) \prod_{i:[i]=0} \mu(\boldsymbol{s_i} \mid \boldsymbol{\theta_I}) \prod_{i:\boldsymbol{s_i} \in O} \gamma k(\boldsymbol{s_i}|\boldsymbol{s_{[i]}}, \boldsymbol{\theta_o})
\end{aligned}
\tag{3.9}
$$

where $\mathcal{D}$ is the observation window, and O is the set of offspring points.

Immigrant Process	Offspring Kernel	$\mathcal{D}$	Model Name
HPP	bivariate beta	unit-square	HPP-Bibeta
HPP	truncated bivariate Normal	irregular domain	HPP-Tbinorm
NHPP	bivariate beta	unit-square	BPNHPP-Bibeta
NHPP	truncated bivariate Normal	unit-square	BPNHPP-Tbinorm
NHPP	truncated bivariate Normal	irregular domain	BPNHPP-Ireg-Tbinorm

Table 3.1: Model configurations for SH processes.

The permutation of domain shape, immigrant process type, and offspring kernel choice render five models as listed in Table 3.1. The following section will describe the hierarchical model under each configuration and discuss their posterior simulation details. We will start with a simple model, HPP-Bibeta, to demonstrate posterior updates for the latent branching structure, then proceed to describe the similar model HPP-Tbinorm with the alternative offspring kernel. Next, we describe the augmented hierarchical model BPNHPP-Bibeta with the immigrant process modeled as an NHPP and the offspring kernel as a bivariate beta density. Lastly, we describe the most complicated but most general model, BPNHPP-Ireg-Tbinorm. Notice that when $\mathcal{D}$ is the unit-square, the BPNHPP-Ireg-Tbinorm is equivalent to BPNHPP-Tbinorm. We thus skip details on this model since the immigrant and offspring process components appear in previous configurations.

The complete data likelihood suggests that we can model the immigrant points and offspring points separately conditional on $[i]$ since the contributions from both processes are separable as multiplicative terms in the likelihood. The hierarchical model can be therefore decomposed into three parts: the model for immigrant points with $[i] = 0$ controlled by the

immigrant intensity function $\mu(\boldsymbol{s})$ with immigrant parameter $\boldsymbol{\theta_I}$, the model for offspring points with $[i] \neq 0$ controlled by branching ratio γ and offspring kernel parameter $\boldsymbol{\theta_o}$, and lastly, the prior for the latent variables $[i]$. Notice that updating $[i]$ for $i = 1 \cdots, N$ is equivalent to sampling from the posterior full-conditional of the branching structure since the full branching structure $\mathcal{T}$ can be mapped one-to-one to the set of latent variables $[i]$.

For models with homogeneous Poisson immigrant processes, we focus on the third part of the hierarchical model since the immigrant process is simple and controlled by one parameter $\mu(\boldsymbol{s}) = \mu$. We now describe the prior for the branching structure $\mathcal{T}$ in terms of summaries of $[i]$. Let $\boldsymbol{C_i}$ denote the set of points that descend from point i according to the branching structure $\mathcal{T}$; let $\boldsymbol{C_i^c}$ denote the complement of $\boldsymbol{C_i}$, i.e., the set of valid parent points for point i. A discrete uniform prior on the union of valid parent point set and 0, $\{0, \boldsymbol{C_i^c}\}$, ensures a proper prior for the branching structure, which assumes an equal chance for a point i to be an immigrant point or a child of points in set $\boldsymbol{C_i^c}$.

The hierarchical model for a spatial Hawkes process with homogeneous Poisson immigrant process follows the form below with $k(\boldsymbol{s_i}|\boldsymbol{s_{[i]}}, \boldsymbol{\theta_o})$ as the general form for any offspring kernel, and $F_\gamma(a_g, b_g)$ as the prior for the branching ratio γ and $F_{\boldsymbol{\theta_o}}$ as the prior for the immigrant parameters $\boldsymbol{\theta_o}$:

$$\begin{gathered}\{\boldsymbol{s_i}\}|[i], \mu, \gamma, \{[i]\} \sim \exp(-\mu \mid \mathcal{D} \mid -N\gamma) \prod_{\{i:[i]=0\}} \mu \text{Unif}(\boldsymbol{s_i}|\mathcal{D}) \prod_{\{i:\boldsymbol{s_i} \in O\}} \gamma k(\boldsymbol{s_i}|\boldsymbol{s_{[i]}}, \boldsymbol{\theta_o}) \\ \gamma \sim F_\gamma(a_g, b_g), \quad \boldsymbol{\theta_o} \sim F_{\boldsymbol{\theta_o}}, \quad [i] \overset{ind.}{\sim} \text{Uniform}(\{0, \boldsymbol{C_i^c}\}).\end{gathered} \tag{3.10}$$

The nature of the branching ratio requires its prior support to be $[0, 1]$ for the spatial Hawkes process to be stable. We consider two choices for $F_\gamma(a_g, b_g)$: the beta distribution and

the truncated gamma distribution, where a_g and b_g are the two shape parameters for the former, and shape and rate parameter for the latter. The following section presents the MCMC detail under each prior. We use the truncated gamma distribution in the simulation study and the real data example due to better mixing behavior under such prior.

We now describe the parametric model for the offspring density $k(\boldsymbol{s_i}|\boldsymbol{s_j}, \boldsymbol{\theta_o})$ under the bivariate beta kernel and the truncated bivariate Normal kernel. The domain $\mathcal{D}$ is restricted to the unit-square when using the bivariate beta kernel controlled by $\boldsymbol{\theta_o} = \{\tau\}$. We place a gamma prior on the dispersion parameter τ. Alternatively, the truncated bivariate Normal kernel applies to any compact observation window $\mathcal{D}$ and is controlled by $\boldsymbol{\theta_o} = \{\sigma_x, \sigma_y, \rho\}$. We place inverse-Gamma priors on σ_x and σ_y with hyperparameters a_x, b_x, a_y, b_y. We place a beta prior on the transformation of the correlation coefficient ρ, $h(\rho) = (\rho + 1)/2$ to map ρ from $[-1, 1]$ to $[0, 1]$.

The posterior full conditional for the latent variable $[i]$ is the following discrete distribution:

$$\pi([i] = j|-) = \begin{cases} \frac{\gamma k(\boldsymbol{s_i}|s_j, \boldsymbol{\theta_o})}{\mu + \gamma \sum\limits_{l \in \boldsymbol{C_i^c}} k(\boldsymbol{s_i}|\boldsymbol{s_l}, \boldsymbol{\theta_o})} & j \in \boldsymbol{C_i^c} \\ \\ \frac{\mu}{\mu + \gamma \sum\limits_{l \in \boldsymbol{C_i^c}} k(\boldsymbol{s_i}|\boldsymbol{s_l}, \boldsymbol{\theta_o})} & j = 0 \end{cases} \tag{3.11}$$

Sampling from such discrete distribution can be easily achieved by evaluating both the immigrant intensity μ and offspring intensity $\gamma k(\boldsymbol{s_i}|\boldsymbol{s_j}, \boldsymbol{\theta_o})$ for all valid parents with index $j \in \boldsymbol{C_i^c}$ for i. Concretely for each $[i]$, a routine will first define $\boldsymbol{C_i}$ by tracing down the branching structure recursively and adding point located above i in the tree structure until reaching the node that contains i. Notice that i is included in $\boldsymbol{C_i}$, and $\boldsymbol{C_i^c}$ is the complement taken with

respect to the entire point pattern. The offspring intensity function is then evaluated at $\boldsymbol{s_i}$ given parent points indexed by elements in $\boldsymbol{C_i^c}$. We leave the posterior updates for $\boldsymbol{\mu}, \gamma$ and $\boldsymbol{\theta_o}$ to the next section where we discuss implementation details.

Next we describe the models with nonhomogeneous Poisson immigrant processes. We will use the intensity model discussed in Section 3.3.2 with parameter set $\{\{V_{k_x,k_y}\}, \alpha\}$ and pre-specified constants C and K. We augment the parameter space to achieve easy posterior updates with the set of latent variables $\{\xi_j, \eta_j\}$ as the basis index pair for point with index $j \in \{j : [j] = 0\}$. We use $\phi(\boldsymbol{s_j})$ to denote $\phi(x_j, y_j)$, where $\boldsymbol{s_j} = \{x_j, y_j\}$. The hierarchical model for the BPNHPP-Bibeta configuration is the following:

$$\begin{gathered}
\{\boldsymbol{s_i}\}|\boldsymbol{V}, \{\xi_j, \eta_j\}, \gamma, \{[i]\} \sim \exp(-\Lambda - N\gamma) \prod_{\{i:[i]=0\}} \Lambda\phi_{\xi_j,\eta_j}(\boldsymbol{s_i}) \prod_{i:\boldsymbol{s_i}\in O} \gamma k(\boldsymbol{s_i}|\boldsymbol{s_{[i]}}, \tau) \\
\pi(\xi_i = k_x, \eta_i = k_y|K, \boldsymbol{V}, [i] = 0) = V_{k_x,k_y}/\Lambda \quad [i] \overset{ind.}{\sim} \text{Uniform}(\{0, \boldsymbol{C_i^c}\}) \\
V_{k_x,k_y}|\alpha, K \sim \text{Ga}(\alpha/K^2, C) \quad \alpha \sim \text{Ga}(a_\alpha, b_\alpha) \quad \gamma \sim F_\gamma(a_g, b_g) \quad \tau \sim \text{Ga}(a_g, b_g)
\end{gathered} \tag{3.12}$$

where $\boldsymbol{V} = \{V_{k_x,k_y}\}$, and $\Lambda = \int\int_{[0,1]^2} \mu(x, y)\mathrm{d}x\mathrm{d}y = \sum_{k_x=1}^K \sum_{k_y=1}^K V_{k_x,k_y}$.

The posterior updates for $\{V_{k_x,k_y}\}$ and $\{\xi_i, \eta_i\}$ conditional on the immigrant index set $\{i : [i] = 0\}$ are simple: given $\{\xi_i, \eta_i\}$ and α, the V_{k_x,k_y} are conditionally independent and gamma-distributed; given $\{V_{k_x,k_y}\}$ and point location $\boldsymbol{s_i}$, the full-conditional for the latent variables $\{\xi_i, \eta_i\}$ is a discrete distribution. The update for the branching structure is again achieved by updating $[i]$ for $i = 1 \cdots, N$ sequentially. Here the full conditional for $[i]$ is also a discrete distribution with the same form as (3.11) with the immigrant intensity $\mu(\boldsymbol{s_i}) = \sum_{k_x,k_y=1}^K V_{k_x,k_y}\phi(\boldsymbol{s_i})$ in the place of μ. The updates for α, γ and τ will be discussed in the next section.

Finally we describe the most general model BPNHPP-Ireg-TBinorm, where 'Ireg' in the configuration name emphasizes that the domain $\mathcal{D}$ can be of any shape. We skip the BPNHPP-TBinorm configuration, which can be considered either as similar to (3.12) with modified offspring kernel, or the special case for BPNHPP-Ireg-TBinorm applied to the regular domain. To accommodate the irregular observation window $\mathcal{D}$, we represent the intensity function as a weighted sum of truncated Bernstein densities $\phi^*_{\xi_i,\eta_i}(\boldsymbol{s_i}) = B^{-1}_{\xi_i,\eta_i}\phi_{\xi_i,\eta_i}(\boldsymbol{s_i})$ and adjust the weights V_{k_x,k_y} by normalizing constant B_{k_x,k_y} for the Bernstein density with index (k_x, k_y). The hierarchical model is given below:

$$
\begin{aligned}
\{\boldsymbol{s_i}\}|\boldsymbol{V}, \{\xi_j, \eta_j\}, \gamma, \{[i]\} &\sim \exp(-\Lambda_{\mathcal{D}} - N\gamma) \prod_{\{i:[i]=0\}} \Lambda_D \phi^*_{\xi_i,\eta_i}(\boldsymbol{s_i}) \prod_{i:\boldsymbol{s_i}\in O} \gamma k(\boldsymbol{s_i}|\boldsymbol{s_{[i]}}, \Sigma) \\
\pi(\xi_i = k_x, \eta_i = k_y|K, \{V_{k_x,k_y}\}, [i] = 0) &= V_{k_x,k_y}/\Lambda_{\mathcal{D}} \quad [i] \overset{ind.}{\sim} \text{Uniform}(\{0, \boldsymbol{C_i^c}\}) \\
V_{k_x,k_y}|K &\sim \text{Ga}(\alpha/K^2, C) \quad \alpha \sim \text{Ga}(a_\alpha, b_\alpha) \quad \gamma \sim F_\gamma(a_g, b_g) \\
\Sigma \equiv \{\sigma_x^2, \sigma_y^2, \rho\} &\sim \text{inv-Ga}(a_x, b_x)\text{inv-Ga}(a_y, b_y)\text{Be}((\rho+1)/2|a_\rho, b_\rho)
\end{aligned}
\tag{3.13}
$$

where $\Lambda_{\mathcal{D}} = \int\int_{\mathcal{D}} \mu(x,y)\mathrm{d}x\mathrm{d}y = \sum_{k_x,k_y=1}^{K} V_{k_x,k_y}B_{k_x,k_y}$. The updates for $\boldsymbol{V}$ and $\{\xi_i, \eta_i\}$ are again simple with conditionally independent gamma full-conditionals for V_{k_x,k_y} and discrete full-conditionals for $\{\xi_i, \eta_i\}$. The updates for $[i]$ assumes the same form as (3.11) with the modification of replacing μ everywhere with the nonhomogeneous immigrant intensity $\mu(\boldsymbol{s}) = \sum_{k_x,k_y} V_{k_x,k_y}B_{k_x,k_y}\phi^*_{k_x,k_y}(\boldsymbol{s})$. The updates for the offspring parameters $\{\sigma_x^2, \sigma_y^2, \rho\}$ will be discussed in the next section.

3.3.5 Posterior simulation

In this section, we first discuss the posterior full-conditional updates for the branching ratio γ and offspring parameter(s) $\boldsymbol{\theta_o}$, where $\theta_o \equiv \tau$ for the bivariate beta kernel and $\boldsymbol{\theta_o} \equiv \{\sigma_x^2, \sigma_y^2, \rho\}$ for the truncated bivariate Normal kernel.

The posterior update for the branching ratio requires a Metropolis step under a beta prior and a rejection sampling step under a truncated Gamma prior. The full-conditional under beta prior is $\pi(\gamma|-) \propto \exp\{-N\gamma\}\gamma^{|\boldsymbol{O}|}\gamma^{a_g-1}(1-\gamma)^{b_g-1}$, where $|\boldsymbol{O}|$ is the number of offspring points. Sampling can be achieved by a Metropolis step with proposal for the logit-transformed γ or introducing an auxiliary variable ζ such that $\pi(\gamma, \zeta|-) \propto I_{(\zeta<\exp\{-N\gamma\})}\gamma^{|O|}$ $\gamma^{a_g-1}(1-\gamma)^{b_g-1}$ The full conditional for ζ is $\pi(\zeta|\gamma) \sim \text{Uniform}(0, \exp(-N\gamma))$ and the new full-conditional for γ is $\pi(\gamma|\zeta, -) \sim \text{beta}(|O| + a_g, b_g)I_{[\gamma<-\frac{log(\zeta)}{N})]}$. We sample ζ from a uniform distribution and γ conditional on ζ from a truncated beta distribution. We obtain poor mixing for γ with such sampling approach. Alternatively, the full conditional under truncated Gamma prior is $\pi(\gamma|-) \propto \gamma^{|O|+a_g-1}\exp\{-(N+b_g)\gamma\}I_{[\gamma\in[0,1]]}$, which can be sampled by rejecting samples from the Gamma distribution with shape $|\boldsymbol{O}| + a_g$ and rate $N + b_g$ that do not fall in $[0, 1]$.

The updates for offspring kernel parameters are achieved by Metropolis steps. For bivariate beta kernel, the dispersion parameter τ has full-conditional $\pi(\tau|-) \propto \prod_{i\in O} k(\boldsymbol{s_i}|\boldsymbol{s_{[i]}}, \tau)\cdot$ gamma$(\tau|a_\tau, b_\tau)$. The update is achieved through a Metropolis step with proposal distribution as a normal centered on the current log-transformed τ value. For the truncated bivariate Normal density, we jointly update $\{\sigma_x^2, \sigma_y^2, \rho\}$ through the transformation $\{\log(\sigma_x^2), \log(\sigma_y^2),$

$\text{logit}((\rho + 1/2)\}$ with a normal proposal density with covariance matrix $\Sigma_{tune} = \boldsymbol{I_3}\sigma^2_{tune}$.

The normalizing constant for the bivariate Normal density is computed numerically using Monte Carlo approximation and therefore poses a significant computational challenge. When evaluating the offspring density, computing the normalizing constant occurs and requires updates when the parent location and the offspring parameters change. Two types of computation require computing the normalizing constant by Monte Carlo approximation: 1) when updating the parent index $[i]$ for point i, the offspring kernel is evaluated $N - |\boldsymbol{C_i}|$ times with parent locations in $\boldsymbol{C_i^c}$; 2) when updating the offspring kernel parameter Σ, with fixed parent locations, the offspring kernel is evaluated for all points with current and proposed Σ. We implement two tricks to reduce computational cost in these two situations: in 1), we create only one Monte Carlo sample when Σ is given from a bivariate Normal distribution centered at $\mathbf{0}$ with covariance matrix Σ. For a kernel with parent location $\boldsymbol{s_{[i]}}$, we shift the entire MC sample by such parent location. In 2), we cache the normalizing constant for points with the same parent, so the normalizing constant is only computed once. These tricks turn out to be necessary and efficient. Together they lead to a 300 times speed up compared to an algorithm that computes the normalizing constant on the fly using a naive MC implementation.

Next, we describe the posterior updates for parameters of the NHPP model $\{V_{k_x,k_y}\}, \alpha$ and the latent variables ξ_j, η_j for $[j] = 0$. In the most general case where $\mathcal{D}$ is a subset of the unit square, the full-conditional for $\{V_{k_x,k_y}\}$ is a gamma distribution with shape parameter $\sum_{i=1}^{N} I_{[\xi_i=k_x,\tau_i=k_y,[i]=0]} + \alpha/K^2$ and rate parameter $C + B_{k_x,k_y}$. Notice that when $\mathcal{D}$ is the unit-square, the normalizing constant B_{k_x,k_y} is 1 for the corresponding spatial Bernstein density, thus we obtain the simplified posterior update for BPNHPP over unit square with-

out using B_{k_x,k_y}. The full conditional for ξ_j, τ_j are independent discrete distributions with $p(\xi_j = m, \eta_j = n|-) \propto V_{m,n}\phi_{m,n}(\boldsymbol{s_j})$ for $[j] = 0$. And finally, α is updated by a Metropolis step with Normal proposal on the log-scale.

3.4 Model checking and comparison

3.4.1 Predictive residuals over the Voronoi tessellation

We perform the posterior predictive residual check in the simulation study as the primary method to examine the first moment inference. The predictive residual is defined as the difference between the actual number of observations and the predicted number of observations over a partition of the domain $\mathcal{D}$. Leininger and Gelfand (2017b) suggests that the predictive residual is preferable to the raw residual since the credible interval of the latter is not expected to cover 0. We obtain a sample from the posterior distribution of the predictive residual by simulating a point pattern replicate for each posterior sample. For each replicate, we compute the predictive residual with the bth posterior sample of the model parameters for $b = 1 \cdots, B$ and obtain a posterior sample of predictive residual of size B.

As recommended by Bray et al. (2014), we apply the Voronoi tessellation to partition the observation window for residual inference to avoid bias introduced by an arbitrary grid: the distribution of the expected number of events per grid cell is skewed if the grid cell is too small, but the over- and under-estimation can cancel out if such cell is too large. The Voronoi tessellation partitions the observation window using a set of points such that each subset contains only one of these points which has the shortest distance to any points in the subset. We

report the posterior mean of the predictive residuals in each Voronoi tessellation partition in a map. For synthetic data, the choice of points to perform the Voronoi tessellation can simply be the immigrant points that are available through simulation. For real data where the identity of immigrant point is unknown, we can use a clustering algorithm to identify local cluster centers.

3.4.2 Ripley's K Function

We use the Ripley's K function (Ripley, 1976) to examine the second moment inference under the proposed models. More specifically, we use the estimate of the generalization of the original K function for inhomogeneous point patterns developed by Baddeley et al. (2000) as a summary statistic to perform model checking and comparison. The inhomogeneous K function is defined as

$$K_{\text{inhom}}(r) = \frac{1}{\mid B \mid}\text{E}\Big(\sum_{\boldsymbol{s_i} \in \boldsymbol{S} \cap B} \sum_{\boldsymbol{s_j} \in \boldsymbol{S} \setminus \{\boldsymbol{s_i}\}} \frac{\mathbb{1}_{[||\boldsymbol{s_i} - \boldsymbol{s_j}|| \leq r]}}{\lambda(\boldsymbol{s_i})\lambda(\boldsymbol{s_j})} \Big), r \geq 0$$

for any subset $B \subset \mathcal{D}$, where $\mathbb{1}_{[\cdot]}$ is the indicator function, $|B|$ is the area of B and $\lambda(\cdot)$ is the non-constant first-order intensity function for the finite point pattern $\boldsymbol{S} = \{\boldsymbol{s_1}, \cdots, \boldsymbol{s_n}\} \subset \mathbb{R}^2$. We obtain a posterior sample for such K function by simulating a point pattern replicate for each posterior sample of model parameters, and apply the inhomogeneous K function estimator to the replicate. We use the inhomogeneous K function estimate using the observed point pattern as the "truth" to benchmark the posterior mean and interval estimate of the K function. Additionally, we examine data variability by simulating multiple point pattern replicates based on the posterior mean estimate of the model parameters, and obtain K function estimates for these replicates. We check whether the posterior credible interval for the K function covers the

"truth" for goodness of fit, and compare such coverage across multiple models applied to the same data to select the better performing model.

We use the K function more as a summary statistic than a hypothesis testing tool. The goal for such comparison between K function estimates using real data vs. simulated replicates is not to test whether the HPP or NHPP assumption is valid for the point process, but rather how similar these estimates are under the same assumption. The K function estimates of the point pattern replicates serve as summaries of the second moment inference for different models and can be compared against the summary based on the observed data to examine model fit. The reasons that lead to difference in K function estimates between the replicated data under a certain model and the observed data may come from different aspects of modeling: the model assumes the wrong stochastic process; the model assumes the right stochastic process but performs the estimation poorly, etc. Thus, we do not use such comparison to test whether the observed data is generated from the point process assumed by the model. The K function is implemented through accessing the `Kest.fft` and `Kinhom` function from the `spatstat` package in `R` (Baddeley and Turner, 2005).

3.5 Simulation Study for the BPNHPP-Bibeta model

In this section, we present two simulation studies: the first one with HPP as the true G_0, and the second one with NHPP as the true G_0 with a unimodal logit-transformed bivariate Gaussian immigrant density. The true offspring densities in both cases take the form of aforementioned bivariate beta centered on the parent location with parameter τ that controls

dispersion. We consider three models: BPNHPP , HPP-Bibeta model, and BPNHPP-Bibeta formulated in (3.12). The BPNHPP model refers to the intensity formulation for the spatial NHPP model. The first scenario with HPP G_0 truth is designed to explore the identifiability of the BPNHPP-Bibeta under a HPP-Bibeta truth where the offspring density functionals are the same in the truth and model specification. The fact that BPNHPP-Bibeta is able to recover the homogeneous G_0 would provide empirical evidence to the model's identifiability since it can distinguish between the immigrants and the offspring. In the second scenario, we want to test the BPNHPP-Bibeta model capacity against a BPNHPP model and compare the inference of the target point pattern's first and second moment properties. We choose the BPNHPP model as a reference model since our previous research showed that such model provides a flexible prior to capture the variety of intensity shapes of a NHPP.

3.5.1 True immigrant process as HPP

In the first scenario, we simulate three synthetic datasets by specifying Λ the intensity of G_0, γ the branching ratio, and τ the offspring dispersion parameter: 1) $\Lambda = 800, \gamma = 0.2, \tau = 200.0$, 2) $\Lambda = 500, \gamma = 0.5, \tau = 200.0$, 3) $\Lambda = 500, \gamma = 0.5, \tau = 300.0$, shown in Fig 3.2. The number of immigrants and total observations in the three cases are 833/1062, 526/1127, 526/1099 respectively.

We recognize that the model will have significant difficulty in distinguishing the immigrants from the offspring if weakly informative prior is given. We first fix the constant C used in the BPNHPP specification for G_0 to be 1, which drives the dispersion parameter α for the underlying DP prior to be close the number of immigrants in expectation. A larger value

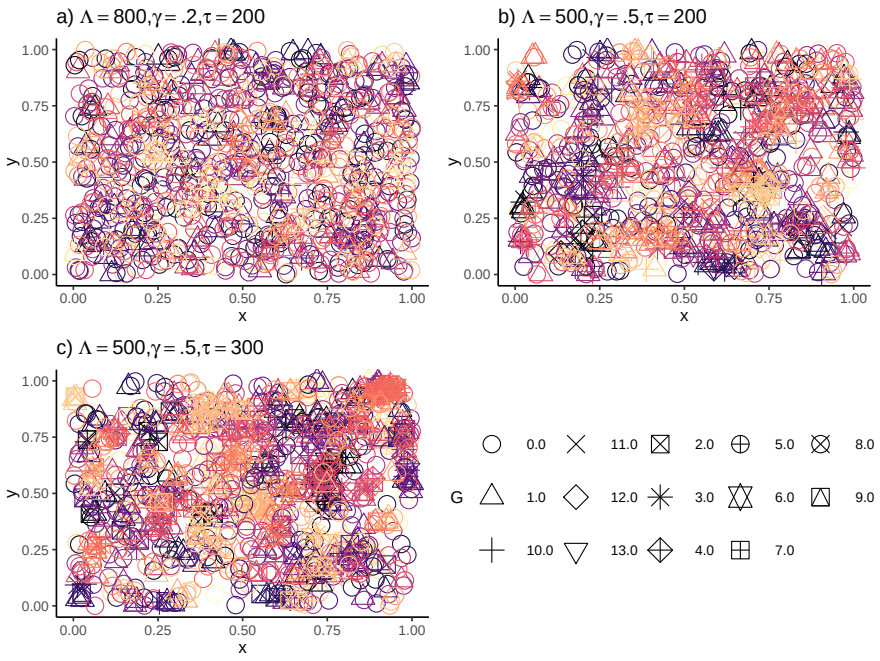


Figure 3.2: Simulated point pattern data in case a), b) and c) under the HPP-Bibeta process; the shape of the point indicates generation; points with the same color belong to the same family.

for α leads to the realized G_0 density to be closer to the baseline uniform density. We then give α an informative prior, $\alpha \sim \text{gamma}(8000, 10)$, with prior 95% range $[782, 817]$ in case 1) and $\text{gamma}(5000, 10)$ with prior 95% range $[485, 513]$ in case 2) and 3). K is set to 10 so that we use 100 basis functions in all three cases for the inference of G_0. Lastly τ is given informative prior $\text{gamma}(200.0, 1)$ with prior 95% range $[173, 226]$ and $\text{gamma}(300.0, 1)$ with prior 95% range $[260, 333]$.

Fig. 3.3 shows the posterior mean estimate of the G_0 intensity function, which can be compared to the total intensity Λ everywhere in the unit square given the HPP truth. In case a), the posterior mean intensity appears to be close to 800 except for some local variations with an uncertainty band with width 200 almost everywhere except for the boundaries. Similar patterns

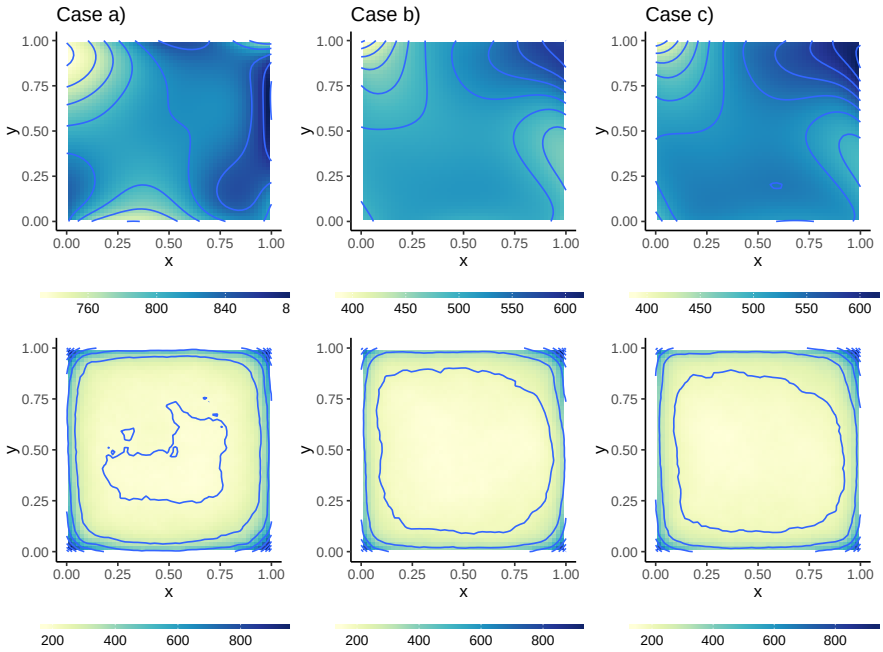


Figure 3.3: Posterior intensity point and interval estimation for G_0: posterior mean intensity function in the first row, the difference between the 90th and 10th percentile of the posterior distribution for the intensity function on the second row; The constant G_0 intensity Λ is 800 for case and 500 for case b) and c).

appear in case b) and c) as well. These observations indicate that with an informative prior on DP precision parameter α, the model is able to capture the homogeneous nature of G_0.

3.5.2 True immigrant process as NHPP

In the second scenario, we simulate three cases by varying the total intensity Λ, the branching ratio γ and the offspring density dispersion parameter τ. a) $\Lambda = 500, \gamma = 0.4, \tau = 200.0$ b) $\Lambda = 80, \gamma = 0.7, \tau = 100.0$ c) $\Lambda = 80, \gamma = 0.7, \tau = 200.0$, shown in Fig 3.4. The immigrants vs total observations in the three scenarios are 526/863, 90/373 and 90/334 respectively.

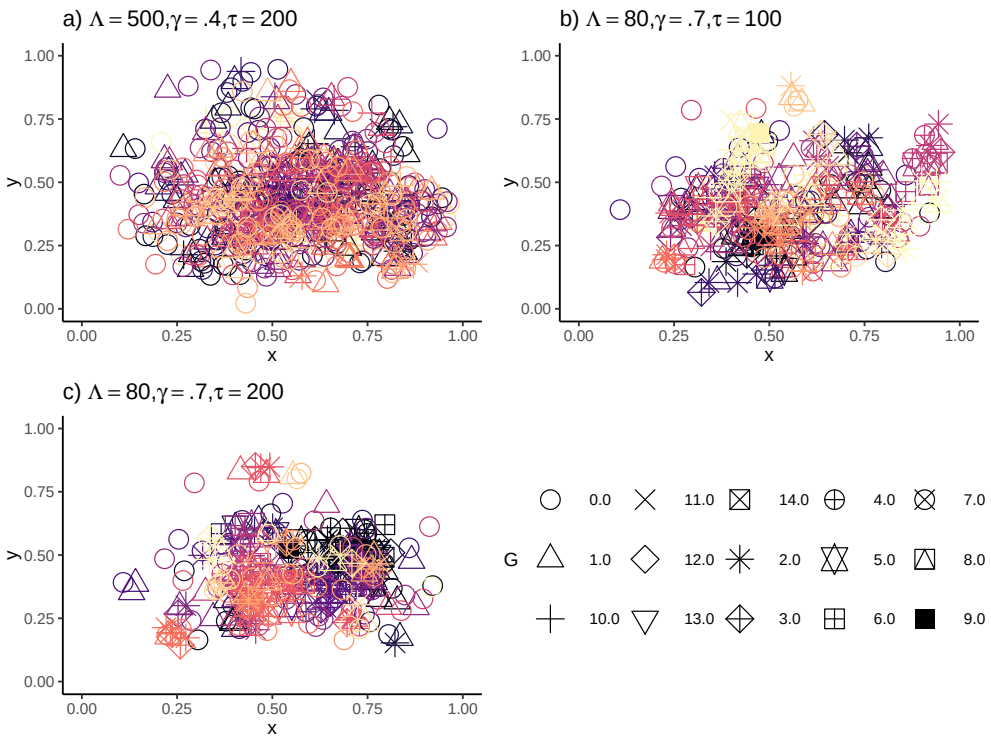


Figure 3.4: Simulated data for case a), b) and c) under NHPP-Hawkes truth; the immigrant density function is a logit-transformed bivariate normal density that is independent in x and y dimension; the point shape indicates its generation; points with the same color belong to the same family.

We set $C = 0.1$ for all cases, as we expect the immigrant intensity function to deviate from uniform drastically. α is given weakly informative prior gamma$(25, 1)$ in case a) and gamma$(10, 1)$ in case b) and c). K is set to 20 so that we use 400 basis functions to infer G_0. τ is given informative priors gamma$(200, 1)$ in case a) and c) and gamma$(100, 1)$ in case 2) . The Gibbs sampler in all cases were run for 10000 iterations with the first 5000 discarded as burn-in and the rest thinned by 3.

The first moment inference for the three scenarios is illustrated in Fig. 3.5. Across the three cases, the posterior mean residual results appear to be similar under BPNHPP and

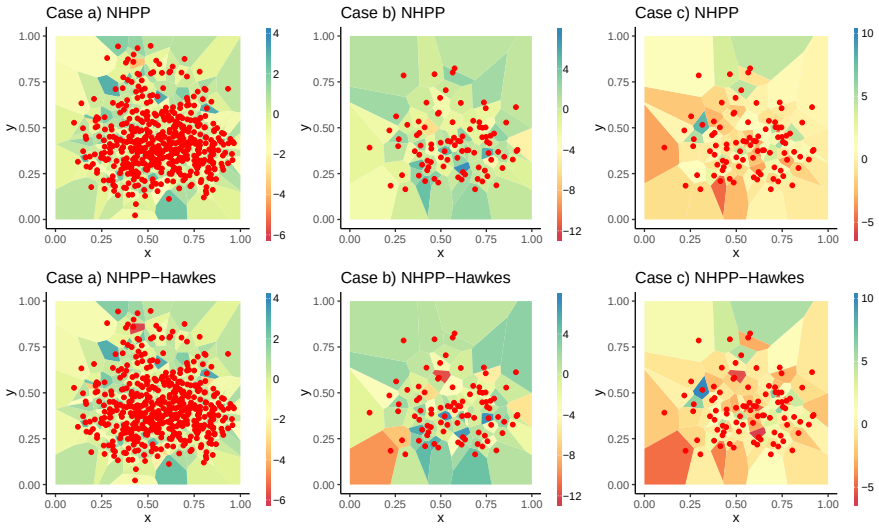


Figure 3.5: Posterior predictive residuals mean estimate over Voronoi tessellation based on true immigrant points: the red dots are immigrant points under the simulated truth.

BPNHPP-Bibeta, with the latter showing more local variance. These results are consistent across other simulated truth with the same G_0 density function and different branching ratio γ. Comparing case a) to case b) and c), we aim to detect the relationship between immigrants and offspring ratio and the accuracy of capturing the first moment property under the two models. However, based on the point estimate of the first moment shown in Fig. 3.5, the two models seem to capture the first moment equally well even under the scenarios where the ratio between offspring and immigrant counts is high.

We use Ripley's K function to examine the second moment inference under the BPNHPP and BPNHPP-Bibeta model. Fig. 3.6 shows the posterior mean and 95 % credible interval of K function estimate with border correction. BPNHPP-Bibeta provides wider uncertainty band for the K function estimates and is able to capture the truth in all three cases. BPNHPP provides

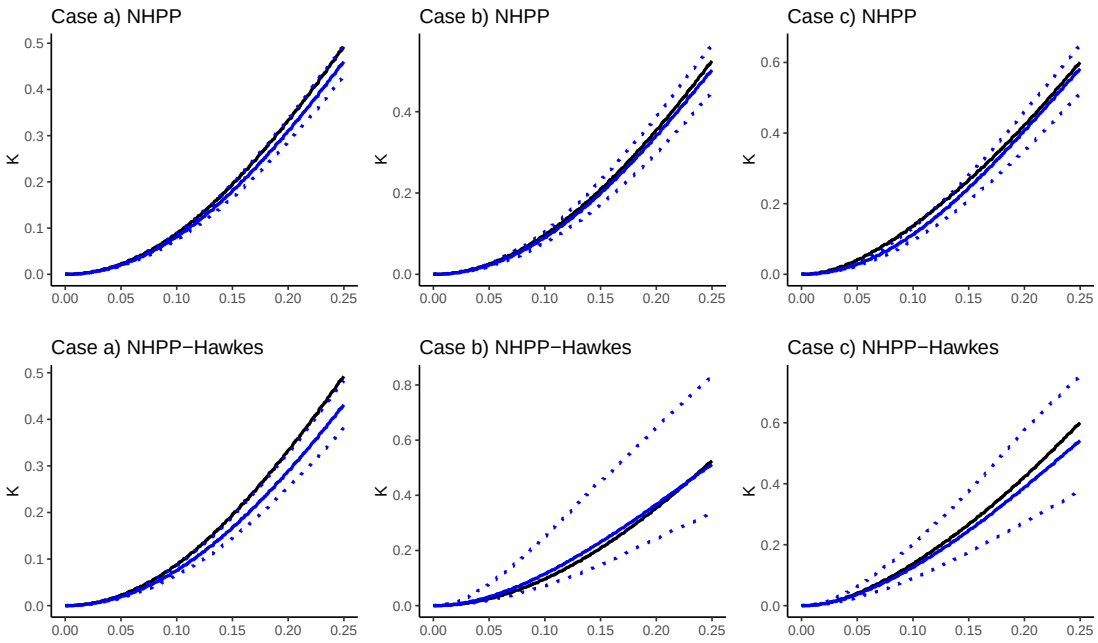


Figure 3.6: Empirical (black solid line), posterior mean (blue solid line), and 95% credible interval (blue dotted line) estimation for HPP K function with boarder correction.

similar point estimates with smaller credible interval that may miss the truth occasionally (especially in case c)). These differences could be due to stochastic variations caused by simulation, since in this analysis only one copy of data is simulated under each posterior sample. We then use bootstrap method to examine data variation: Fig 3.7 shows the inhomogeneous K function estimates applied to 40 data sets simulated under the posterior mean estimates of parameters under BPNHPP and BPNHPP-Bibeta. In case a), both models failed to capture the truth when the offspring to immigrants ratio is low, while in b) and c), BPNHPP-Bibeta performs noticeably better than BPNHPP when the offspring to immigrants ratio is higher.

Lastly, we show the inference for G_0 intensity function under the BPNHPP-Bibeta model that can be compared to the simulation truth in Fig. 3.8. The intensity point estimate for

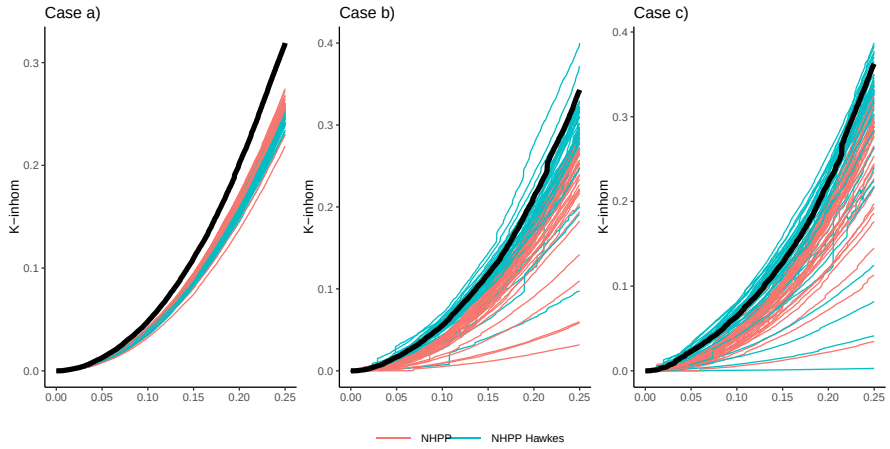


Figure 3.7: K functions realizations for point pattern replicates based on posterior mean estimates of the model parameters: 40 replicates are simulated using the posterior mean parameter estimates for each model and the K function estimates for each are shown as separate curves. The black solid curve is the K function estimated based on the observed point pattern.

Case a) in panel (2, 1) compared to the true intensity in panel (1, 2) is able to capture the overall unimodal shape of the intensity function albeit missing the mode location slightly. When the immigrants are fewer, capturing the immigrant intensity is much harder as shown in case b) (panel 2, 2) and c) (panel 2,3). A less dispersed offspring intensity helps the model to capture the immigrant intensity better since in case c) the posterior mean intensity function captures the shape better than in case b). Looking at the simulated data under Case b) in Fig. 3.4, we notice that there is a big cluster near (0.8, 0.5) that causes the skewness in the posterior intensity estimate of the immigrant intensity. The model also misses the mode intensity in case b) by a lot. These two observations together suggest that higher dispersion in offspring locations makes it more difficult for the model to identify the immigrants.

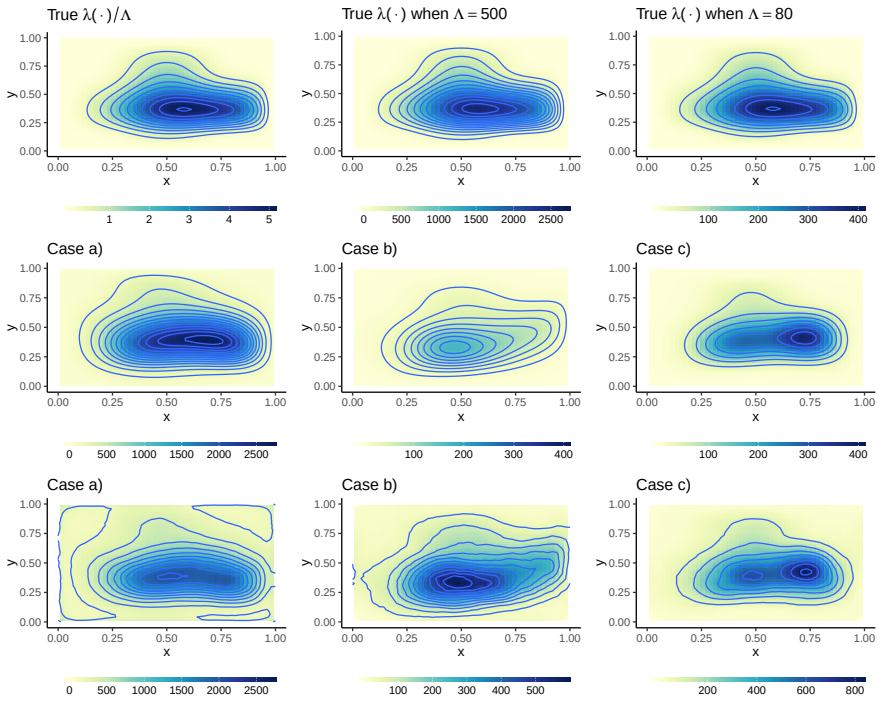


Figure 3.8: Posterior intensity mean and 95% credible interval length estimates for the immigrant generation G_0: the first row shows simulation truth; the second row shows posterior mean estimate of density/intensity functions, and the third row shows the difference between 95th and 5th percentile of the posterior distribution for the density/intensity function.

3.6 Simulation study for the BPNHPP-Ireg-Tbinorm model

3.6.1 Simulation over synthetic irregular domain

The simulation study is designed to demonstrate the model's capacity under situations where the immigrant intensity functional form, the branching ratio and offspring density differ. We use two densities to simulate the immigrant density: 1) a mixture of bivariate beta independent in x, y dimension with five well separated modes 2) a mixture of logit normal. The irregular domains are designed to take significant area of the unit square (otherwise one

can imagine re-scaling the point pattern and the boundary to fit the unit square). Finally, these densities are chosen such that the density mass over the irregular domain is greater than 0.96 since we assume the event that drives the point pattern mainly happen over the irregular domain and is not impacted by events outside the boundary. For case 1), the irregular domain is a spade in the unit square; for case 2), the irregular domain is a diamond shape polygon.

The main levers of tuning the model are the choice of constants K and C, whether to fix α or model α as random and the choice of hyper-parameters for the prior distributions for α, γ, and ρ. The specification for α and C are jointly determined with a prior understanding of the total immigrant intensity. By construction, $E(\alpha)/C$ is the expected total intensity for the immigrant process over the unit square. One practical way to specify the total intensity for the immigrant process is to come up with a prior guess of the proportion of immigrant points in the point pattern in order to estimate the number of immigrant points $|G_0|$. The prior expectation of $|G_0|$ is $E(\alpha)/(C) \times |\mathcal{D}|$, where $\mathcal{D}$ is the area of the irregular domain. We can therefore estimate the ratio $E(\alpha)/C$. In the simulation study, we observe that larger value of α leads to poor model fit where the immigrant process dominates and drives the branching ratio to 0, thus underestimating the K function as a result. We discover that fixing α at the prior expectation can facilitate the convergence of the MCMC algorithm, since the chain for α is usually sticky and can get stuck at higher values. The key then is to choose an appropriate C value so that α can be fixed at a relatively small value while matching α/C to the prior guess.

We adopt three sets of priors for the branching ratio: a uninformative prior as a uniform distribution over $[0, 1]$, a weakly informative prior with decreasing density over $[0, 1]$, and a informative prior with density mass concentrated around the true value for γ. Specifically, the

uniform distribution is approximated by a truncated gamma distribution with both shape and rate parameters close to 0. The informative priors are truncated gamma distribution with the prior expectation set equal to the true γ and different variance to control the density function.

We place priors on the transformation of the offspring parameter ρ: $(1+\rho)/2$. The uninformative prior is set to be the uniform distribution and the informative prior is set to have prior expectation at the truth.

In the first scenario, we focus on comparing the inference results where a) both priors for γ and ρ are informative and b) both are uniform while keeping the priors for other parameters the same. The conclusion is that the model is not sensitive to the prior specification in this case and recovers the truth in both cases. The posterior mean and 95% credible intervals for model parameters are presented in Table 3.2, where γ is the branching ratio, $\Lambda_{\mathcal{D}}$ is the total immigrant intensity over the irregular domain, and (σ_x, σ_y, ρ) is the truncated normal covariance parameters. Specifically, $\Lambda_{\mathcal{D}}$ is computed via $\Lambda_D = \sum_{k_x,k_y=1}^{K} B_{k_x,k_y} V_{k_x,k_y}$ and is interpreted as the expected average number of immigrant points over the irregular domain. Λ_D is negatively correlated with γ since fewer immigrant points would lead to more offspring points to achieve the overall population. The simulation truth used in the first scenario is $\gamma = 0.3, \Lambda_D = 290.34, \sigma_x = \sigma_y = 0.02, \rho = 0$, which are covered by the credible interval in both cases. The point and interval estimation for most parameters are similar under both prior specifications except for ρ. Under informative prior for ρ, the credible interval is smaller and the point estimate is closer to truth.

We present posterior inference on predictive residuals over the Voronoi partition based on immigrant points, predictive K function, and posterior immigrant density for model under

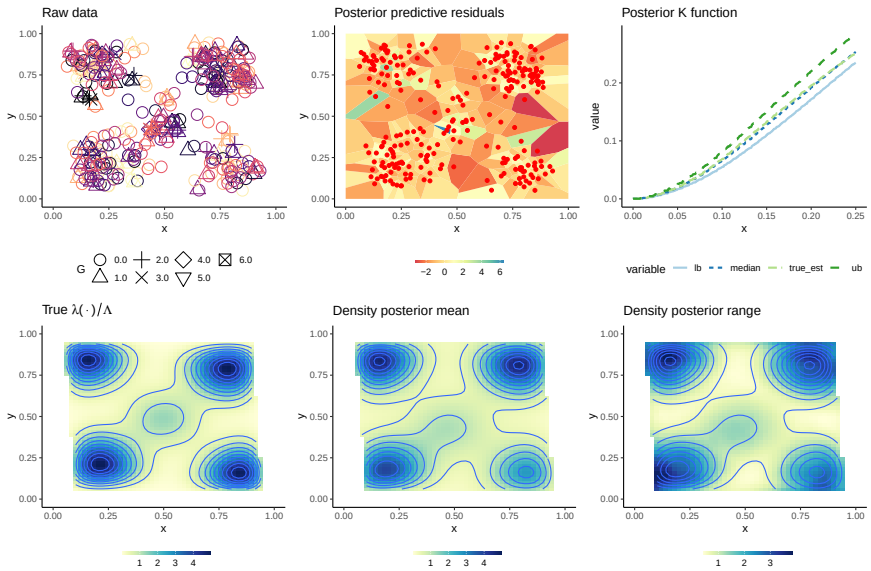


Figure 3.9: Posterior inference under informative priors for data simulated with a bivariate beta mixture as the true immigrant density.

informative prior in Fig. 3.9. The inferences under the uninformative prior are similar to the those in Fig. 3.9 and are skipped here. The posterior predictive K function inference suggests that the estimation captures the second-moment property of the underlying point process well. The residuals shows largest over-estimation in area around $(0.8, 0.5)$ where there is no points simulated and largest under-estimation near $(0.5, 0.45)$ where a couple of immigrant points partition a small region. Overall, the residuals are close to 0 as suggested by the dominating light orange color in the residual plots. Finally, the density estimation recovers the five modes in the immigrant density function in terms of both magnitude and position. The posterior 95% credible intervals for the immigrant densities cover the truth in both cases.

The second scenario is particularly challenging for the parametric form of the BPNHPP-Ireg-TBinorm model since in the simulation truth both the immigrant and offspring density ker-

Cases	γ	$\Lambda_{\mathcal{D}}$	σ_x	σ_y	ρ
pbeta a)	0.363 (**0.3**)	254.794 (**290.34**)	0.019 (**0.02**)	0.019 (**0.02**)	-0.136 (**0**)
	(0.230, 0.524)	(189.347, 315.857)	(0.012, 0.029)	(0.013, 0.029)	(-0.479, 0.213)
pbeta b)	0.362 (**0.3**)	252.919 (**290.34**)	0.019 (**0.02**)	0.021 (**0.02**)	-0.340 (**0**)
	(0.227, 0.559)	(173.350, 316.723)	(0.012, 0.032)	(0.013, 0.034)	(-0.801, 0.292)
snail a)	0.169 (**0.3**)	349.162 (**288.81**)	0.018 (**0.02**)	0.018 (**0.02**)	0.001(**0**)
	(0.066 , 0.322)	(281.708, 410.757)	(0.010 , 0.036)	(0.010 , 0.037)	(-0.417 , 0.430)
snail b)	0.162 (**0.3**)	355.692 (**288.81**)	0.016 (**0.02**)	0.019 (**0.02**)	0.248 (**0**)
	(0.059, 0.342)	(281.005, 413.633)	(0.010, 0.035)	(0.010, 0.037)	(-0.554, 0.951)
snail c)	0.125 (**0.3**)	371.192 (**288.81**)	0.017 (**0.02**)	0.017 (**0.02**)	-0.014 (**0**)
	(0.0002, 0.356)	(282.026, 441.546)	(0.009, 0.035)	(0.009, 0.045)	(-0.449, 0.404)

Table 3.2: Simulation study results for point patterns simulated over synthetic boundaries: pbeta refers to an immigrant density as a mixture of four bivariate beta densities, snail refers to an immigrant density as a mixture of bivariate normal densities on the logit scale. Each entry shows the posterior mean, the true values in bold and parenthesis on top, and the posterior 95% credible interval at the bottom.

nel are Gaussian (or a transformation of a Gaussian density) and are therefore difficult for the model to distinguish apart. We adopt a weakly informative prior $\text{Ga}(0.3, 1)$ for the branching ratio γ with true $\gamma = 0.3$ so that the prior expectation matches with the truth. However, such prior has median at 0.053 and prior 95% credible interval $[2.39 \times 10^{-6}, 0.839]$ and therefore favors a process with few offspring generations. The informative prior $\text{Ga}(3, 10)$ also has prior expectation matching the truth, a mode at 0.2, and a tighter prior credible interval $[0.061, 0.7]$. We find the model converges to a process where $\gamma = 0$ when using a uniform prior for γ. We suspect the particular functional form of the immigrant density function is the culprit since the model is able recover γ under uniform prior in the first scenario.

We report the inference under the following three cases: a) both informative prior on γ and ρ, b) informative prior on γ and uniform prior on ρ, and c) weakly informative prior on γ and informative prior on ρ. The comparison between case a), b) and case a), c) show model sensitivity with respect to prior specification of ρ and γ respectively. The posterior inferences for case a)-c) under the second scenario are presented in the last three rows in Table 3.2. The truth for model parameters in this scenario is $\gamma = 0.3, \Lambda_{\mathcal{D}} = 288.81, \sigma_x = \sigma_y = 0.02, \rho = 0$. Comparing a) and b) suggests that the point estimation is closer to truth and the credible interval is tighter under the informative prior for γ. Comparing a) and c) suggests that the informative prior has limited impact on the posterior point and interval estimates for γ. We notice that the point estimates tend to be lower than the truth for γ and higher than the truth for $\Lambda_{\mathcal{D}}$ across three cases. This confirms that the model has difficulty distinguishing the immigrant points from the offspring points and favors a process with more immigrant points and fewer offspring points than the truth in the posterior. The posterior credible intervals cover the truth for all parameters

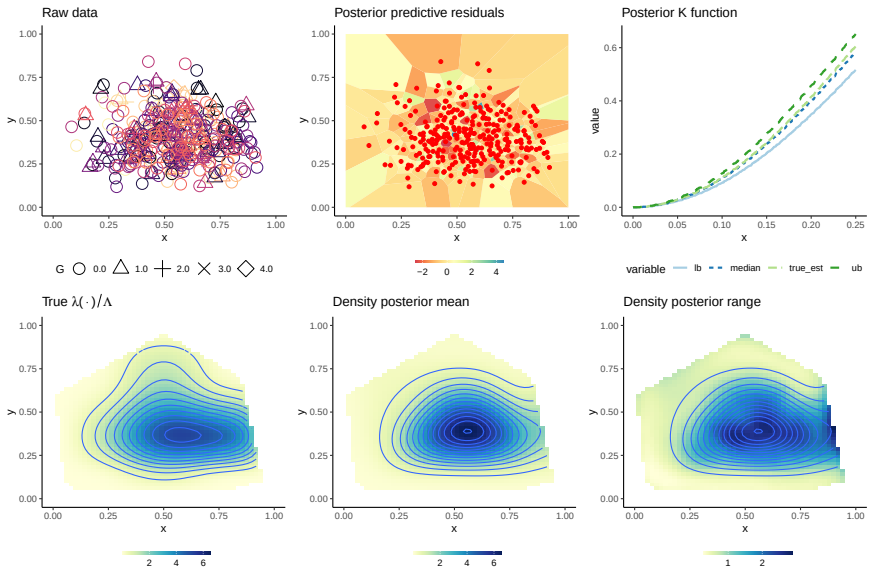


Figure 3.10: Posterior inference under informative priors for data simulated with a logit Normal mixture as the true immigrant density.

across the three cases.

We present the posterior inference for the predictive residuals, predictive K functions and point and interval estimation for the immigrant density under case a) in Fig.3.10. The results for all three cases are very similar.

3.6.2 Simulation over Boston city boundary

We process the Boston city boundary using the `sf` package in `R`. We took the Boston city boundary shape file and simplify it using the standard method in `sf`, and exclude the East Boston area from our study since it is disconnected from the rest of the city by water. The result is a concave polygon consisting of 31 points, referred to as Boston concave boundary (see the shaded areas in the second row of Fig. 3.12) . This boundary is further simplified by

Cases	γ	$\Lambda_{\mathcal{D}}$	σ_x	σ_y	ρ
qbeta a)	0.331 (**0.3**)	336.652 (**382.92**)	0.017 (**0.02**)	0.021(**0.02**)	-0.010 (**0**)
	(0.188, 0.495)	(251.892, 418.118)	(0.012, 0.026)	(0.012, 0.033)	(-0.368, 0.345)
qbeta b)	0.325(**0.3**)	338.001 (**382.92**)	0.018 (**0.02**)	0.021 (**0.02**)	0.121 (**0**)
	(0.168, 0.524)	(247.390, 424.505)	(0.011, 0.028)	(0.012, 0.035)	(-0.573, 0.711)
hnorm	0.395(**0.3**)	564.4 (**654.88**)	0.045 (**0.05**)	0.036 (**0.05**)	0.328 (**0.3**)
	(0.290, 0.509)	(445.46, 683.83)	(0.032, 0.065)	(0.024, 0.056)	(-0.137, 0.809)

Table 3.3: Posterior inference for model parameters for synthetic data over simplified Boston boundaries. Each entry shows the posterior mean, the true values in bold and parenthesis on top, and the posterior 95% credible interval at the bottom.

taking the convex hull which results in a concave polygon, referred to as the Boston convex boundary (see the shaded areas in second row of Fig. 3.11). We notice that the run speed of the algorithm depends on the irregular domain shape mainly due to the Monte Carlo computation of the normalizing constant for the offspring normal kernel. We therefore prefer simpler boundary to realistic ones to reduce computation cost. In this simulation study, we present two simulated cases using these two boundaries respectively for a comparison in run speed.

We design the first case to have a truth close to the model over the simpler Boston convex boundary and aim to test prior sensitivity. The data generating process for the G_0 is a mixture of 4 bivariate beta densities outside of the Bernstein polynomial family, such that the mixture has 3 well separated modes. The branching ratio is 0.3, the total intensity over the irregular domain is 382.92, and the offspring kernel parameters is $(\sigma_x, \sigma_y, \rho) = (0.02, 0.02, 0)$. We apply the model with the following two sets of priors: a) both informative prior on the

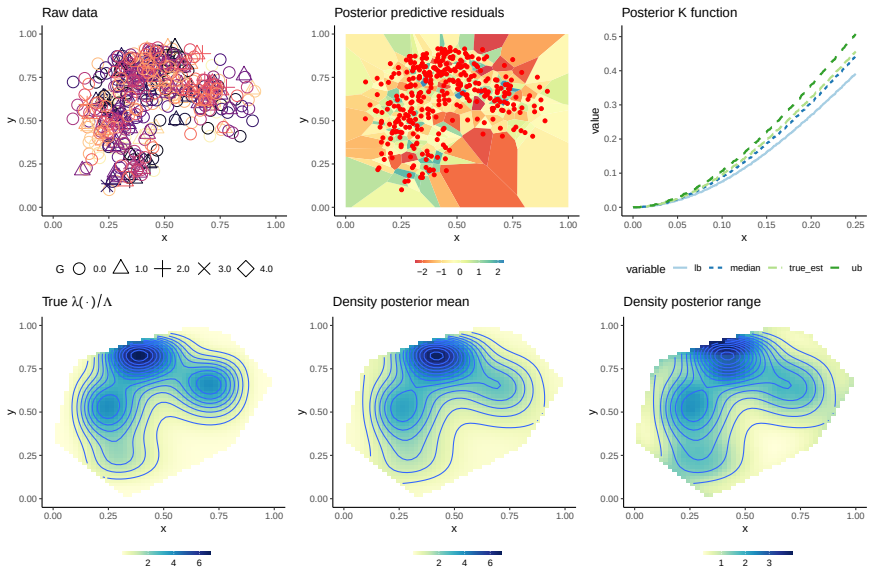


Figure 3.11: Posterior inference under informative priors for data simulated with a mixture of four bivariate beta density as the true immigrant density over Boston convex hull.

branching ratio and the correlation, and b) weakly informative prior on the branching ratio and uniform prior on the correlation. The inference on model parameters are presented in Table 3.3 under case names qbeta a) and qbeta b). All parameters are covered by the posterior 95% credible interval. The branching ratio is not sensitive to the prior, since the point and interval estimates are similar under informative and weakly informative priors. The correlation is sensitive to the prior, with a much wider interval under the uniform prior. The posterior predictive inference for the Voronoi residuals, K function and point and interval estimation for the immigrant density is presented in Fig. 3.11.

We design the second case to test the limit of the model's capacity with a truth as a mixture of six bivariate log-normal distributions that results in highly localized immigrant density pattern over the more complex Boston concave boundary. The mixture density has five

well-separated modes that resembles the potential local hot spots for crime application. We discover that an informative prior for the branching ratio leads to better inference of the K-function and proper coverage of the true branching ratio value in the posterior 95% credible interval. Under uninformative prior for γ, the model tends to overestimate the branching ratio and underestimate the total immigrant intensity. We suspect such behavior is a result of the similar Gaussian functional form of the immigrant and offspring density and the highly localized immigrant density which makes it difficult for the model to distinguish between immigrants and offsprings. We also discover that fixing the precision parameter α of the Dirichlet process prior leads to faster and more stable convergence. We choose α based on prior knowledge on the total immigrant intensity over the unit-square according to prior expectation results in **??**. The posterior inference on model parameters are presented in Table 3.2 with row name hnorm and the 95% credible intervals cover the true parameter values. The posterior predictive inference on K-function and the immigrant density inference in presented in Fig. 3.12.

3.7 Real data example: Boston city crime

We apply the model for spatial Hawkes process with truncated Bivariate Normal offspring kernel to the same Boston city crime data used in Section 2.5, with some modifications to satisfy additional modeling assumptions. Based on the assumption that the point process is defined over a compact irregular domain, we choose to exclude the points from the north east region and the islands in the east that are disconnected from the rest of the city by water. We choose to simplify the main city boundary to a concave polygon since the simulation study in

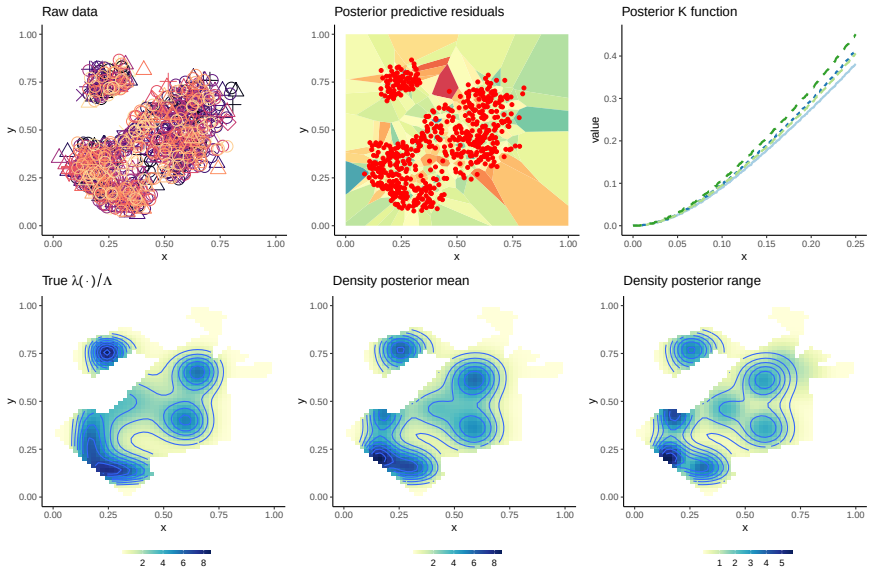


Figure 3.12: Posterior inference for mixture of six log-normal densities over Boston concave boundary.

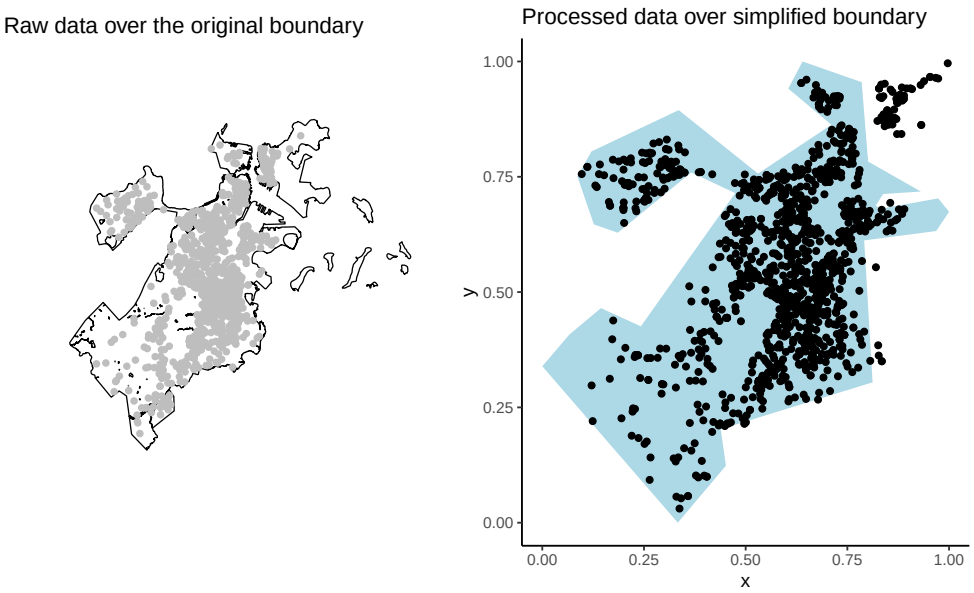


Figure 3.13: Vandalism point pattern in Boston city from April to June 2017 over original Boston city boundary on the left, mapped to the unit-square defined by the bounding box of the simplified boundary on the right.

Section 3.6 suggests that the model run-time scales with the complexity of the irregular boundary. We use the bounding box of the main city region as reference for the map between Northing and Easting to x, y in the unit square. We again focus on modeling vandalism in 2017 from April to June. Fig. 3.13 shows the 1251 raw data points over the original city boundary in the left panel. We treat the point pattern in the three month time period as a realization from a spatial Hawkes process defined over the main city region shown in the right panel in Fig. 3.13. When fitting the model, we exclude points outside of the simplified irregular domain and obtained 1180 points.

We fit the model assuming low branching ratio and high background intensity and assign informative priors according to these assumptions. We start with an uninformative prior for the offspring kernel parameters and run the MCMC for 1000 iterations for a warm start. The result suggests σ_x, σ_y smaller than 0.03 and a ρ around 0.15. We propose two priors for the branching ratio concentrated at 0.3, referred to as prior a), and 0.1, referred to as prior b) respectively. The prior expectation for the total intensity for the immigrant process over the unit-square is set to 1700. We incorporate these information in the prior specification and start the chain at $(0.03, 0.03, 0.15)$ for the offspring kernel parameters for a final run of 15000 iterations.

The posterior inference contradicts our prior assumption that the branching ratio is high and the background intensity is low, despite informative prior that suggests the opposite. The posterior means for the branching ratio are larger than the prior mean in both prior scenarios: 0.67 compared to 0.3 in case a) and 0.522 compared to 0.1 in case b). The spatial kernel parameters suggests a small positive correlation between the y and x coordinates and very small triggering range with σ_x, σ_y close to 0.01 in both cases. Fig 3.14 and Fig 3.15 present the pos-

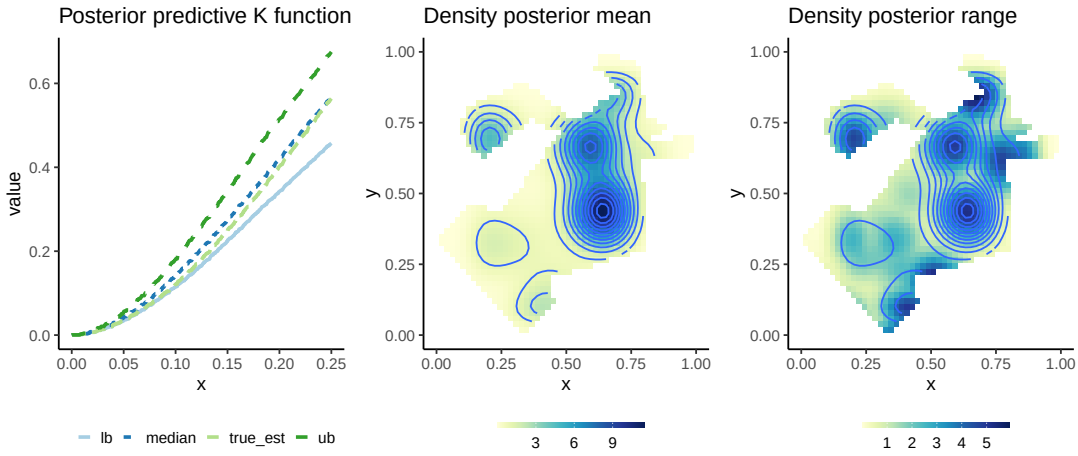


Figure 3.14: Posterior predictive inference on the K function and posterior inference on the immigrant process density under the SH model with $\mathrm{E}(\gamma) = 0.3$.

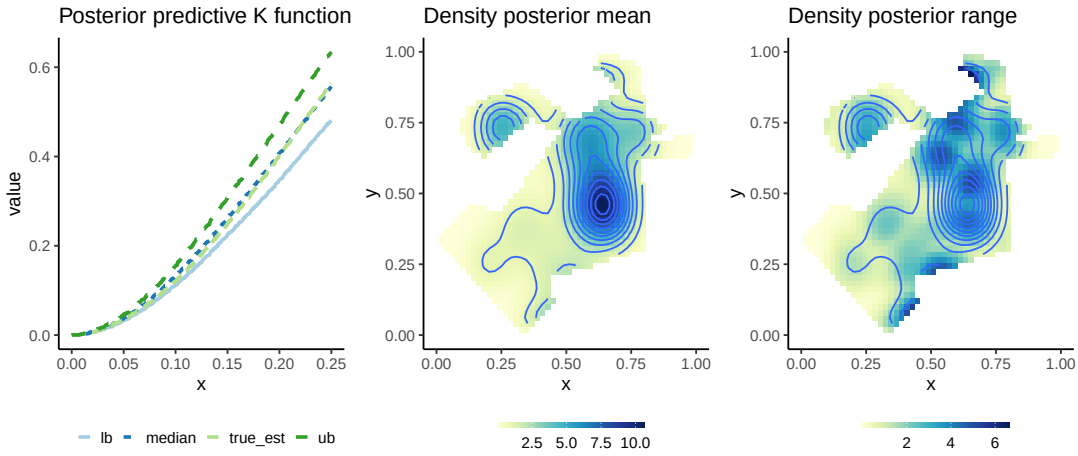


Figure 3.15: Posterior predictive inference on the K function and posterior inference on the immigrant process density under the SH model with $\mathrm{E}(\gamma) = 0.1$.

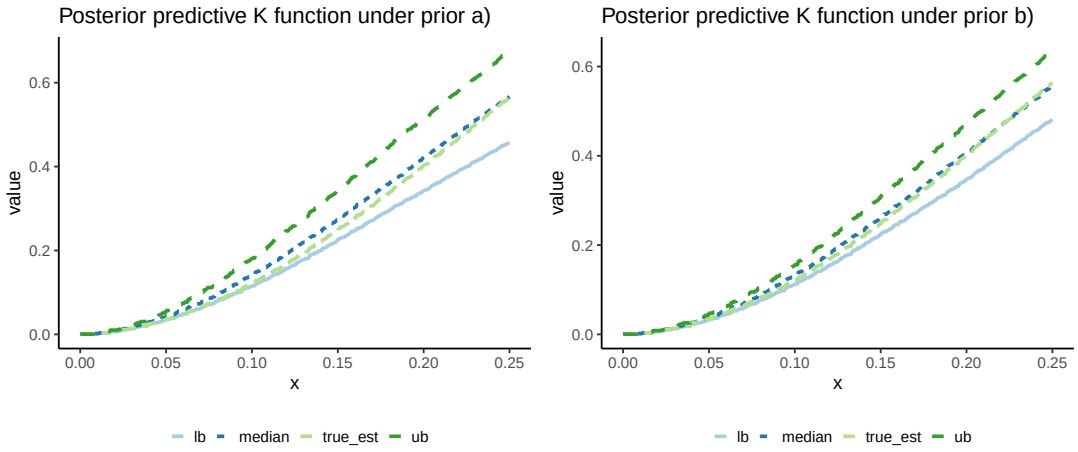


Figure 3.16: Comparison of the Posterior predictive inference on the K function under the two prior scenarios for the SH model.

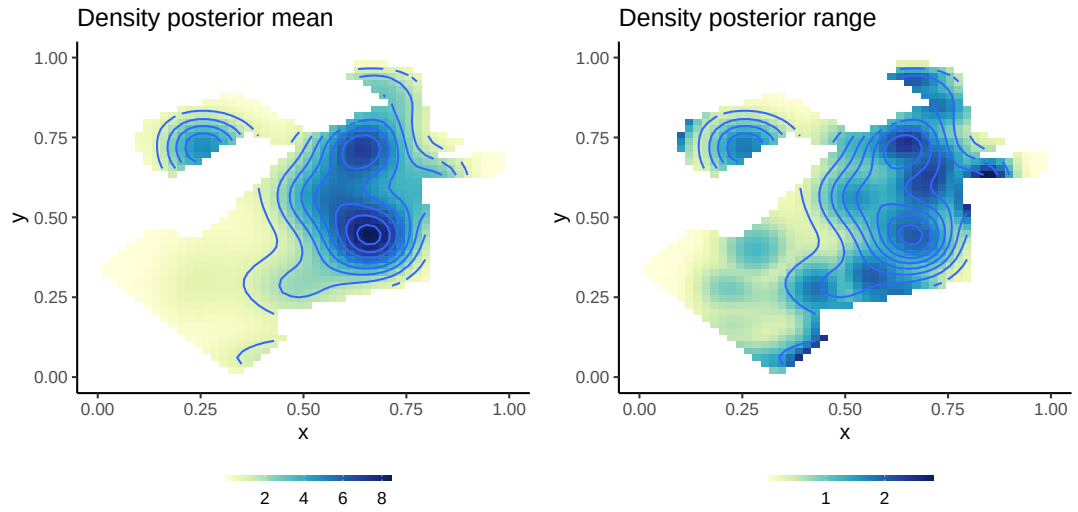


Figure 3.17: NHPP density under the BP-NHPP model applied to the vandalism point pattern in Q2 2017.

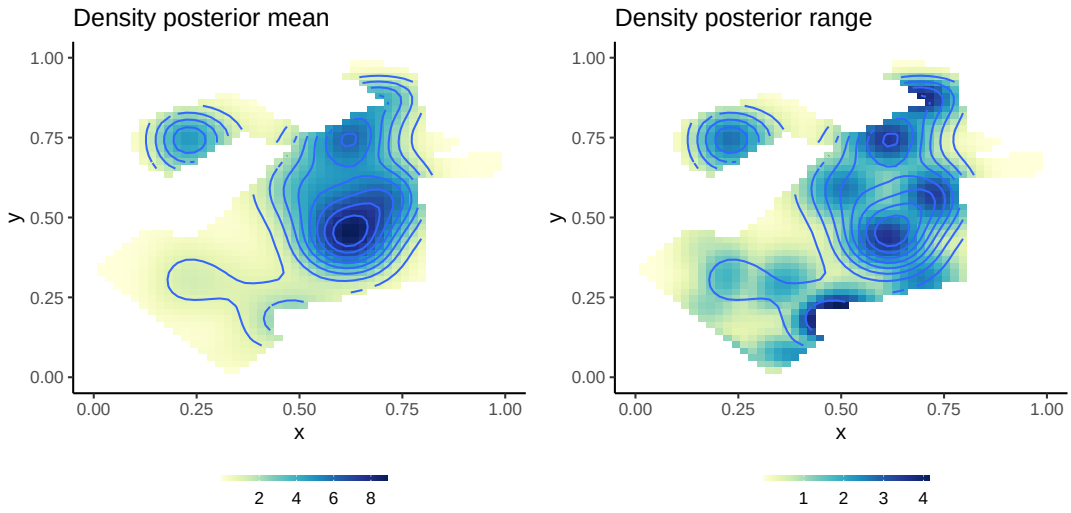


Figure 3.18: Posterior inference for the G_0 immigrant density under the ParSTH model.

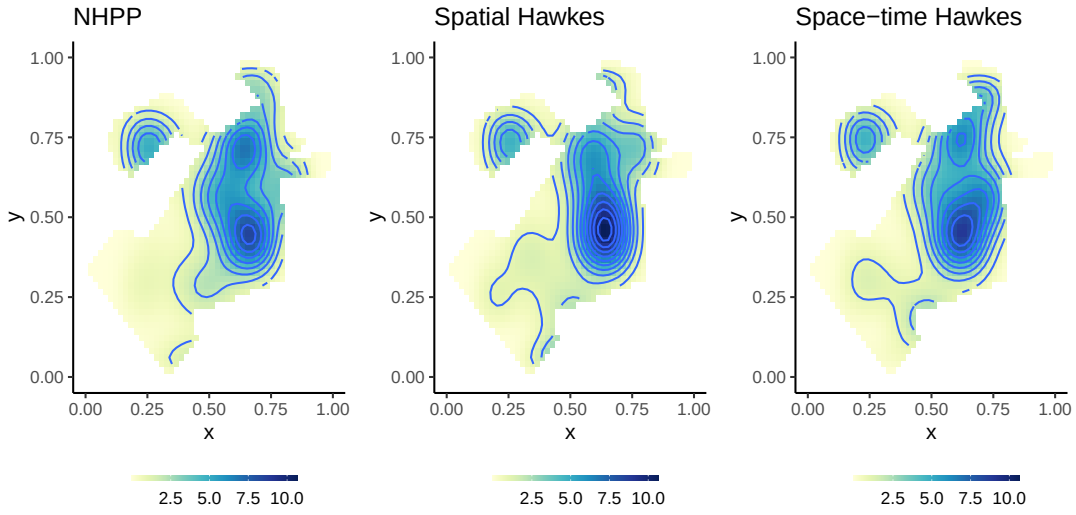


Figure 3.19: Posterior mean estimate for the G_0 density under the NHPP, spatial Hawkes process and the space-time Hawkes process from left to right.

terior predictive K function, the Poisson density posterior mean for the immigrant process, and its posterior range for case a) and b). At a closer look, Fig 3.16 suggests better fit under the case b), as the posterior predictive K function mean tracks the truth more closely. Recall that the true estimation here is a nonparametric estimation for the K function based on the observed data, and is the same across the two prior scenarios. We observe that under both priors, the 95% posterior predictive intervals for the K function cover the "truth".

We find such findings less surprising after fitting the same data augmented with temporal information over the same boundary to the space time Hawkes process with the same spatial offspring kernel and an exponential kernel for the time component. We will discuss in more detail the space-time Hawkes process and the parametric offspring intensity model in the next chapter. The key difference between the space-time Hawkes process (referred to as ParSTH) and the spatial Hawkes process models is that ParSTH poses constraints on the branching structure following the event time. The conditional intensity function for the space-time Hawkes process is $\lambda(x, y, t) = \mu(x, y) + \sum_{t_i:t_i<t} h(x - x_i, y - y_i, t - t_i)$, where $\mu(x, y)$ is the immigrant intensity function, which constant over time and $h(\cdot, \cdot, \cdot)$ is the triggering function. We use the intensity formulation over the irregular domain for the immigrant process. Here we treat the ParSTH as the golden standard for model validation since with additional information on event time, ParSTH model has the best chance to distinguish between offsprings and immigrants. The result from ParSTH also suggests a high branching ratio and low immigrant total intensity scenario. Notice that here the branching ratio and the immigrant intensity are only meaningful qualitatively and are not directly comparable between SH and ParSTH.

We present a comparison of G0 density based on NHPP, SH and ParSTH. The NHPP

model applied to the entire point pattern can be viewed as a special case of the spatial Hawkes process where the branching ratio is 0. The NHPP model assumes independence among points given the total number of occurrences, which can be restrictive especially in potentially self-exciting activities such as crimes and earthquakes. We relax such assumption with a model for the spatial Hawkes process, which captures the dependence among points through the self-exciting behavior captured by the offspring processes, thus effectively leading to local clustering patterns. We can also compare the spatial Hawkes process and the space-time Hawkes process since they both capture the self-exciting behavior based on information in different dimensions. With additional information on event time, the space-time Hawkes process model explores a smaller parameter space for the latent branching structure since the natural order in time puts clear restriction on the available parent for any given point. We discover that despite the different assumptions in each model, the inferences on the immigrant density show very similar pattern as suggested by the comparison of the posterior mean estimate for the immigrant density in Fig. 3.19: all suggest two modes in the main city area and one additional mode in the top left corner. We notice that the uncertainty band for G0 density is the smallest for the BPNHPP model (shown in Fig. 3.17) and become larger for the STH (see Fig. 3.18) and SH model. Such finding aligns with the model assumption, since the immigrant point pattern is latent for the latter two models; the STH model has smaller uncertainty with additional information from the time dimension.

Model	γ	$\Lambda_{\mathcal{D}}$	σ_x	σ_y	ρ
SH	0.67	371.90	0.016	0.012	0.160
E(γ) = 0.3	(0.58, 0.75)	(260.41, 494.5)	(0.013, 0.020)	(0.010, 0.015)	(-0.076, 0.32)
SH	0.522	522.19	0.012	0.010	0.222
E(γ) = 0.1	(0.44, 0.62)	(401.20, 627.28)	(0.0095, 0.0154)	(0.0079, 0.012)	(-0.071, 0.473)
ParSTH	0.55	46.43	0.015	0.013	0.111
by Week	(0.464 , 0.630)	(39.698 , 53.121)	(0.012 , 0.019)	(0.010 , 0.015)	(-0.095 , 0.311)

Table 3.4: Posterior inference for model parameters in the real data example.

3.8 Conclusion

We proposed a Bayesian hierarchical model for the spatial Hawkes process, leveraging its clustering representation and the flexible and efficient nonparametric Bayesian model for the Poisson process developed in Chapter 2 as an important building block. We explored options for both the immigrant Process and the parametric forms for the offspring triggering function, and provided strategies to choose the appropriate immigrant-offspring process combination given context of the application. We discovered the additional computation overhead introduced by truncating the offspring density function to the irregular domain, and developed algorithm routines to improve performance. Finally, we discovered that the crime point pattern in Boston presents significant self-exciting tendencies by applying the most flexible SH model to the Vandalism point pattern. The model for SH process developed in this chapter accounts for more general assumptions for the underlying point process, and contains the model for the NHPP as a special case for $\gamma = 0$. Meanwhile, we acknowledge its limitation brought by

the parametric form of the offspring density and the related computation dependency on the irregular domain $\mathcal{D}$. We address these issues in the next chapter.

Chapter 4

Bayesian semi-parametric modeling for space-time Hawkes Processes

The STH process can be viewed as the extension of the original self-triggering process proposed by Hawkes (1971) that incorporates spatial information in the conditional intensity to describe the self-triggering behavior of the point process. Such extension leads to useful modeling development in applications such seismology (Ogata, 1988, 1998), crime analysis (Mohler et al., 2011), network analysis (Linderman and Adams, 2014) and finance (Bacry et al., 2015).

Earlier space-time self-triggering models, especially the Epidemic Type Aftershock Sequence models in seismology proposed by Ogata (1988), assume a homogeneous background process and certain families of parametric triggering functions that entail certain density function tail behaviors. Both assumptions are restrictive under the context of broader application scenarios. Later methods relax one or both assumptions by either proposing a nonparamet-

ric background intensity and parametric triggering function (Adelfio and Chiodi, 2015; Mohler et al., 2011), a constant background intensity with a nonparametric histogram estimator for the triggering kernel (Marsan and Lengliné, 2008), or fully nonparametric forms for both (Fox et al., 2016; Yuan et al., 2018). Specifically, popular models for the background intensity include kernel smoothing (Ogata and Katsura, 1988; Zhuang et al., 2002); the triggering function is estimated using histogram estimator for the distance between parent and offspring points identified by the latent branching process (Marsan and Lengliné, 2008; Fox et al., 2016; Yuan et al., 2018). Estimation is achieved via Maximum Likelihood Estimation (MLE) (Reinhart, 2018) or EM-type algorithm (Marsan and Lengliné, 2008; Fox et al., 2016). A comprehensive review of frequentist approach for modeling space-time Hawkes process can be found in Reinhart (2018).

Bayesian methods for space-time Hawkes process address the uncertainty quantification issue that posed challenge under the frequentist scheme. Many Bayesian methods were proposed for modeling the univariate or multivariate Hawkes process over time with either parametric formulation for the conditional intensity (Rasmussen, 2013; Ross, 2016; Linderman and Adams, 2014) or nonparametric versions (e.g., Linderman and Adams, 2015). Recent development starts to focus on modeling for space-Hawkes process and tackle the related computation issues. Holbrook et al. (2021) develops a scalable and parallelizable Bayesian inference scheme based on approximated likelihood using a conditional intensity formulation, which consists of a kernel smoother for the background intensity and a triggering function that's exponential in time and Gaussian in space. Alternatively Kolev and Ross (2020) and Molkenthin et al. (2022) both utilize the clustering representation of the space-time Hawkes process with augmented la-

tent branching structure and a semi-parametric conditional intensity function. Kolev and Ross (2020) factors the background intensity as the product of a total intensity term and spatial density, modelled as a Dirichlet Process mixture model with bivariate Gaussian kernels. Molkenthin et al. (2022) models the background intensity with a Sigmoidal-Cox Gaussian process Adams et al. (2009) with proper augmentation schemes to account for the finite spatial domain. Our approach is similar to this line of work in the sense that we also deploy augmented latent branching structure and nonparametric immigrant intensity.

Our contribution in this work is a Bayesian nonparametric prior for the spatial triggering function that improves the existing nonparametric approaches by explicitly incorporating the irregular domain in the model formulation and providing full inference for the point process functionals. We focus on modeling STH processes over compact yet irregular domains, with assumptions motivated directly by crime point pattern modeling and forecasting. We choose to model the triggering function as the product of separable intensity functions for the spatial and temporal components. We present two formulations for the spatial triggering function: a parametric bivariate Gaussian model that accounts for spatial skewness and a nonparametric model in spatial distance assuming spatial isotropy. Both models incorporate the irregular domain explicitly in the model formulation with proper truncation in the spatial distributions. We achieve inference by augmenting the parameter space with latent branching structure in the hierarchical model that allows easy updates in the posterior simulation.

In this chapter, we first introduce the STH process and the hierarchical modeling approach utilizing the clustering representation of the STH process. Throughout the chapter, we model the immigrant intensity nonparametrically with the intensity formulation for NHPP intro-

duced in Chapter 2. We then present the model formulation for the fully parametric triggering function and the semiparametric triggering function with a parametric temporal component and nonparametric spatial component. The chapter concludes with a real data example based on crimes in Boston city.

4.1 Space-time Hawkes process

We assume that the space-time point pattern $\{(\boldsymbol{s_i}, t_i) : i = 1 \cdots, N\}$ where $\boldsymbol{s_i} = (x_i, y_i)$ is a realization from the underlying space-time Hawkes process X over the compact domain $\mathcal{D} \times (0, T] \subset \mathbb{R}^2 \times \mathbb{R}$. The irregular spatial domain $\mathcal{D}$ is often a subset of the unit-square as the spatial coordinates of (x_i, y_i) can be mapped to the unit-square without loss of generality. A space-time point process is defined via the conditional intensity function that characterizes the intensity at time t given the history prior to t. Let $\mathcal{H}_t$ be the history up to time t that includes information on the location and time for events prior to t, the conditional intensity function $\lambda(\boldsymbol{s}, t \mid \mathcal{H}_t)$ is defined as

$$\lambda(\boldsymbol{s}, t \mid \mathcal{H}_t) = \mu(\boldsymbol{s}, t) + \sum_{i:t_i<t} g(\boldsymbol{s} - \boldsymbol{s_i}, t - t_i)$$

The clustering representation of the STH process consists of an immigrant or background process and many offspring processes triggered by prior events. The immigrant process is defined by the background intensity function $\mu(\boldsymbol{s}, t)$ that controls the occurrence rate for immigrant points. The immigrant points then trigger follow-up events with a rate defined by the triggering function $g(s - s_i, t - t_i)$ that temporarily increases the rate of occurrence at time t depending on events prior to t. Often referred to as the offspring points, these follow-up points

continue to spawn further offspring points. Each point in the Hawkes process belongs to either an immigrant or offspring process that depends on its parent point. Such identity of immigrant vs. offspring and the parent index of offspring points are encoded in the latent branching structure, which is a realization from a branching process.The triggering function $g(\boldsymbol{s} - \boldsymbol{s_i}, t - t_i)$ controls how the self-exciting behavior decays in time and relative spatial location. The natural order of events implied by the time dimension of the point pattern restricts the branching structure by only allowing points prior to t to trigger point at t.

Conditional on the latent branching structure $[i]$ that identifies the immigrant points and the parents of offspring points, the space-time Hawkes process can be represented by a superposition of a immigrant Poisson process that generates the subset of points $\{(\boldsymbol{s_i}, t_i) : [i] = 0\}$ with intensity $\mu(\boldsymbol{s}, t)$ and offspring Poisson processes $\{(\boldsymbol{s_i}, t_i) : [i] = j\}$ with intensity $g(\boldsymbol{s_i} - \boldsymbol{s_j}, t_i - t_j)$. We use $\boldsymbol{\theta_I}$ to denote the set of parameters for the immigrant intensity function and $\boldsymbol{\theta_o}$ the set of parameters for the triggering function. Such representation naturally inspires a Bayesian hierarchical modeling approach since the immigrant and offspring processes are independent Poisson processes conditional on the latent branching structure. By taking a Bayesian approach, we avoid a computationally intensive routine to evaluate the likelihood function defined via the conditional intensity but rather update parameters in simpler models for Poisson processes iteratively. The conditional independence leads to separable contribution from the immigrant and offspring processes in the complete-data likelihood. Let I be the collection of immigrant points where $[i] = 0$ and O_j be the collection of points that are offspring of a point j, $j = 1 \cdots N$. Notice that many O_j will be empty since each point has a nonzero probability of generating 0 offspring. The complete data likelihood is

$$L(\{\boldsymbol{s_i}, t_i\} \mid [i], \boldsymbol{\theta_I}, \boldsymbol{\theta_o}) = p(I \mid [i] = 0, \mu(\boldsymbol{s_i} \mid \boldsymbol{\theta_I})) \prod_{i=1}^{n} p(O_j \mid [i] = j, g(\boldsymbol{s_i} - \boldsymbol{s_j}, t_i - t_j \mid \boldsymbol{\theta_o})$$

$$= \exp(-\int_{\mathcal{D}} \int_0^T \mu(\boldsymbol{s} \mid \boldsymbol{\theta_o}) \mathrm{d}\boldsymbol{s}\mathrm{d}t) \prod_{i \in I} \mu(\boldsymbol{s_i} \mid \boldsymbol{\theta_I}) \times$$

$$\prod_{j=1}^{n} \exp(-\int_{\mathcal{D}} \int_0^T g(\boldsymbol{s} - \boldsymbol{s_j}, t - t_j \mid \boldsymbol{\theta_o}) \mathrm{d}\boldsymbol{s}\mathrm{d}t) \prod_{i \in O_j} g(\boldsymbol{s} - \boldsymbol{s_j}, t - t_j \mid \boldsymbol{\theta_o}) \tag{4.1}$$

Bayesian inference can be achieved by modeling the latent branching structure as part of the hierarchical model that includes a model for the immigrant Poisson process and a model for the offspring Poisson process. The latent branching structure is fully specified by $[i]$ and we can treat $[i]$ to be the missing data and estimate $[i]$ and parameters $(\boldsymbol{\theta_I}, \boldsymbol{\theta_o})$ together. Estimating $[i]$ involves identifying a set of possible parent points for point i, which is readily available since the chronological order of the events dictates that only points before point i can be the parent of point i. We complete the hierarchical model with priors on the immigrant and offspring parameters.

Two sets of assumptions about the underlying point process can be made to simplify the modeling approach: the immigrant intensity function $\mu(\boldsymbol{s}, t)$ can be assumed to be time-homogeneous, i.e. $\mu(t, x, y) = \mu(x, y)$; the triggering function $g(t - t_j, x - x_j, y - y_j)$ can be assumed to be separable such that $g(t - t_j, x - x_j, y - y_j) = g_t(t - t_i) g_s(x - x_i, y - y_i)$, where $g_t(\cdot)$ and $g_s(\cdot, \cdot)$ are intensity functions for the time and space. These two assumptions are almost always made in applications with STH processes. Additional assumptions are required depending on whether to model either the immigrant intensity function, the offspring intensity function, or both nonparametrically.

We propose to model a class of space-time Hawkes processes that have background

intensity $\mu(\boldsymbol{s}, t)$ that is homogeneous in time but nonhomogeneous in space, i.e. $\mu(\boldsymbol{s}, t) = \mu(\boldsymbol{s})$, and an triggering function that is separable in time and space. Our key contribution is explicitly incorporating the observed point pattern's irregular domain in the point process model and handles truncation properly in the space dimension for both the immigrant and offspring processes. Motivated by applications in crime forecasting, we assume that the underlying point process occurs strictly over the compact irregular domain $\mathcal{D}$ such as the boundary of a city and ignore potential triggering from and toward points unobserved outside of the irregular domain. It is a reasonable assumption for the crime application since the triggering effect wears off quickly as offspring shift further away from their parents.We propose two models for the triggering functions with a parametric form for $g_t(t)$ in both and a parametric and a nonparametric model for $g_s(\boldsymbol{s})$ respectively. The fully parametric offspring model makes no approximation in the proposed process and allows anisotropy in the spatial kernel at a high computational cost that scales with the complexity of the irregular domain shape. The semi-parametric offspring model assumes spatial isotropy, proposes a stochastic process close but not identical to the space-time Hawkes process, and gains an advantage in computational speed up and flexible inference for the $g_s(\cdot)$.

4.2 Model for the immigrant process

The immigrant Poisson process can be modeled independently conditional on the branching structure. Specifically, the branching structure identifies the immigrant point set I as the points with $[i] = 0$ as the observations for the immigrant Poisson process. Under the

modeling assumption that the immigrant process is time-homogeneous, the immigrant Poisson process can be viewed as a spatial Poisson process over $\mathcal{D}$ with intensity function $\mu(\boldsymbol{s} \mid \boldsymbol{\theta_I})$ and each spatial location is augmented with a timestamp uniformly distributed over $[0, T]$. The challenge for modeling the immigrant process is capturing the spatial inhomogeneity among points in set I, the elements in which vary across the MCMC iterations depending on the estimation of the parent label $[i]$.

We propose to model the immigrant Poisson process with the BPNHPP model under the intensity formulation. As discussed in Chapter 2, this model can capture local trends in the point pattern over an irregular domain in a spatial NHPP with efficient use of beta densities as basis functions that lead to fast posterior updates. More importantly, the conditional independence of the latent parameters (ξ_i, η_i) in the posterior bypasses the issue that there is a varying number of immigrant points across MCMC iterations and retains the conditional independence between the immigrant parameters and the branching structure.

We adapt the BPNHPP model for a NHPP over space to a model for the time-homogeneous space-time Poison process, with the following formulation for the intensity function $\mu(\boldsymbol{s})$:

$$\mu(\boldsymbol{s} \mid \{V_{k_x,k_y}\}) = \sum_{k_x=1}^{K} \sum_{k_y=1}^{K} V_{k_x,k_y} B_{k_x,k_y} \phi^*_{k_x,k_y}(x, y) \tag{4.2}$$

where $\{V_{k_x,k_y}\}$ are structured weights, $\{\phi^*_{k_x,k_y}(x, y)\}$ are the Bernstein spatial basis functions over the irregular domain $\mathcal{D}$ such that $\int_{\mathcal{D}} \phi^*_{k_x,k_y}(x, y)\mathrm{d}x\mathrm{d}y = 1$, and B_{k_x,k_y} is the normalizing constant for the Bernstein spatial basis function with index (k_x, k_y) over $\mathcal{D}$. The total intensity over the space-time domain $\mathcal{D} \times (0, T]$ is

$$\int_{\mathcal{D}} \int_0^T \mu(\boldsymbol{s} \mid \{V_{k_x,k_y}\})\mathrm{d}\boldsymbol{s}\mathrm{d}t = T \sum_{k_x=1}^{K} \sum_{k_y=1}^{K} V_{k_x,k_y} B_{k_x,k_y} \tag{4.3}$$

The space-time Poisson process likelihood then can be written as

$$p(I \mid \{V_{k_x,k_y}\}) = \exp(-T \sum_{k_x=1}^{K} \sum_{k_y=1}^{K} V_{k_x,k_y} B_{k_x,k_y}) \prod_{i \in I} \mu(\boldsymbol{s_i} \mid \{V_{k_x,k_y}\})$$

We place structured Gamma prior on $V_{k_x,k_y} \sim \text{Ga}(\alpha/K^2, C)$ where K is the number of basis and C is a constant. We discussed the connection between such structured Gamma prior to a Bernstein-Dirichlet Prior on the Poisson process density in Chapter 2. We augment the parameter space with latent basis label (ξ_i, η_i) for each immigrant point i such that its prior follows a discrete distribution that depends on $\{V_{k_x,k_y}\}$. The hierarchical model for the immigrant Point process conditional on $I = \{i : [i] = 0\}$ is then:

$$\begin{gathered}
\{(x_i, y_i, t_i)\} \mid V_{k_x,k_y}, \{\xi_i, \eta_i\} \sim \exp(-T \textstyle\sum_{k_x,k_y=1}^{K} V_{k_x,k_y} B_{k_x,k_y}) \prod_{i=1}^{n} \Lambda \phi^*_{\xi_i,\eta_i}(x_i, y_i) \\
(\xi_i, \eta_i) \mid V_{k_x,k_y} \overset{i.i.d}{\sim} \textstyle\sum_{k_x,k_y=1}^{K} \frac{V_{k_x,k_y} B_{k_x,k_y}}{\Lambda} \delta_{k_x,k_y}(\xi_i, \eta_i) \\
V_{k_x,k_y} \overset{ind.}{\sim} \text{Ga}(\alpha/K^2, C) \quad \alpha \sim \text{Ga}(\alpha_a, \alpha_b)
\end{gathered} \tag{4.4}$$

The posterior updates for $\{V_{k_x,k_y}\}, \{(\xi_i, \eta_i)\}$ and α are similar to those described in Chapter 2. The modification brought by the change in likelihood presents in the full-conditionals for $\{V_{k_x,k_y}\}$, the posterior full conditionals for which become $\text{Ga}(\alpha/K^2 + M_{k_x,k_y}, C + T \cdot B_{k_x,k_y})$, where M_{k_x,k_y} is the number of immigrant points with $(\xi_i, \eta_i) = (k_x, k_y)$.

4.3 A fully parametric model for offspring processes

Conditional on the latent parent label $[i]$, the set of offspring points is a superposition of realizations from N offspring processes, each centered on the observed point i. In the branching process, every point in the observed point pattern has a nonzero probability to gener-

ate offspring points. In reality, many points have zero offspring points. We model the offspring processes via a triggering function specified by the same set of offspring parameters and the locations of the parent points respectively over the spatial support $\mathcal{D}$ and temporal support $(0, T]$. The triggering function controls the branching ratio, the total intensity of the offspring process, and the distribution of offspring points over space and time relative to their parents. We first present a parametric model for the triggering function. Let γ be the branching ratio of any offspring process, and j be the latent parent index, we assume that the offspring spatial location follows a truncated bivariate Gaussian distribution centered on the parent location $\boldsymbol{s_j}$ with covariance Σ; the offspring time location follows an exponential distribution with mean $1/\omega$. The triggering function is

$$g(\boldsymbol{s} - \boldsymbol{s_j}, t - t_j) = \gamma \, \mathrm{TN}_2(\boldsymbol{s} - \boldsymbol{s_j} \mid \mathbf{0}, \Sigma, \mathcal{D}) \, \mathrm{Exp}(t - t_j \mid \omega)$$

Let O be the collection of points for which $[i] \neq 0$, $F_t(\cdot)$ is the c.d.f. of the exponential distribution, the likelihood for the offspring processes is given below:

$$p(O \mid \Sigma, \omega) = \exp\{-\sum_{j=1}^{N} \gamma \cdot F_t(T - t_j)\} \prod_{i \in O} \gamma \, \mathrm{TN}_2(\boldsymbol{s_i} - \boldsymbol{s_{[i]}} \mid \mathbf{0}, \Sigma, \mathcal{D}) \, \mathrm{Exp}(t_i - t_{[i]} \mid \omega)$$

Notice that no approximation is required to evaluate the likelihood since we assume the spatial density to be a truncated density over $\mathcal{D}$, meaning that $\int_{\mathcal{D}} \mathrm{TN}_2(\boldsymbol{s} - \boldsymbol{s_j} \mid \mathbf{0}, \Sigma, \mathcal{D}) \mathrm{d}\boldsymbol{s} = 1$. The advantage of such an approach is that the normalizing constant for the Poisson process becomes easy to compute; the disadvantage is that we need to evaluate the normalizing constant for a bivariate Gaussian density over the irregular domain $\mathcal{D}$. We continue to use the computation tricks introduced in Section 3.3.5 to speed up the evaluation of these normalizing constants via Monte Carlo.

We use the variances in the x and y dimension, σ_x^2 and σ_y^2, and the correlation ρ to specify the covariance matrix $\Sigma = [\sigma_x^2 \quad \sigma_x\sigma_y\rho; \sigma_x\sigma_y\rho \quad \sigma_y^2]$. We place inverse-gamma priors on σ_x^2 and σ_y^2, and a beta prior on the transformation of ρ, $h(\rho) = (1+\rho)/2$. We place a gamma prior on the exponential distribution parameter ω. Inference for $\{\sigma_x^2, \sigma_y^2, \rho\}$ and ω is achieved via separate metropolis-hasting updates in the Gibbs sampler for the full model.

4.4 A Semi-parametric model for the offspring Processes

We present a separable triggering function model with a nonparametric formulation for the spatial component and a parametric formulation for the temporal component. In addition to separability, we assume that the offspring spatial density is isotropic, meaning the density only depends on the spatial distance between the parents and their children. The isotropy assumption allows us to model the spatial component in the triggering function via a univariate offspring-parent distance distribution instead of a bivariate offspring spatial density centered on the parent. More crucially, such a dimension reduction largely reduces the computation cost of the MCMC algorithm since the normalizing constants based on the univariate densities can be directly evaluated using integration instead of Monte Carlo approximation.

The most commonly used isotropic kernel is the bivariate Gaussian kernel with a covariance matrix $\boldsymbol{I_2}\sigma^2$ centered on the parent location (x_j, y_j). The kernel can be re-written as a function of the distance $r = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$ between the offspring location (x_i, y_i) and parent location (x_j, y_j). Another widely-used spatial kernel for self-exciting processes is the Power Law distribution, which specifies the spatial distribution in terms of distance directly.

Choosing between the Gaussian and the Power Law kernel relies on a prior understanding of the tail behavior of the offspring process, which is difficult to acquire in practice since the offspring processes are latent. As a result, extensive discussion on model comparison between the different parametric spatial kernels is unavoidable.

More crucially, the spatial distance distribution within the compact irregular domain $\mathcal{D}$ is usually not analytically available due to truncation even when a certain family of the spatial kernel is assumed. These two factors motivate us to place a flexible nonparametric prior on the spatial distance density to capture its varying functional form while respect the irregular domain $\mathcal{D}$ as its support. Given our strict assumption of the compact domain $\mathcal{D}$, we conclude that any parent-offspring distance arising from the observed point pattern should be bounded from above. Such bounded support for the spatial distance distribution allows us to represent the spatial distance density as a mixture of scaled univariate Bernstein densities. Chapter 2 introduced a model for the univariate Poisson process with a prior model for the intensity function that implies a Bernstein-Dirichlet prior for the NHPP density function. Here we borrow this idea to model the offspring spatial intensity, with modified Bernstein densities scaled to a support over $(0, R_{max})$, where R_{max} is the upper bound for possible parent-offspring distances.

4.4.1 Spatial distance distribution

For an isotropic spatial distribution centered on a given location (x_i, y_i), we define the distribution for the distance between any point that follows such spatial distribution and the center location (x_i, y_i) as the spatial distance distribution. The isotropy assumption allows

us to express the spatial distribution $f_s(x, y)$ in terms of the distance of the spatial shift $r = \sqrt{(x - x_i)^2 + (y - y_i)^2}$ such that $f_s(x, y) = f_s(r)$. The spatial distance distribution $f_s^*(r)$ is the marginal distribution in polar coordinates for the point-center distance under the assumption that the distribution for the angle is uniform over $[0, 2\pi]$.

The map between $f_s^*(r)$ and $f_s(r)$ is $f_s^*(r) = 2\pi r f_s(r)$. To see this, notice that since $f_s(r)$ is a proper density, i.e., $\int_0^{2\pi} \int_0^{\infty} f_s(r) r \mathrm{d}r \mathrm{d}\theta = 1$, the joint distribution for (r, θ) in polar coordinates can be factored as

$$f^*(r, \theta) = 2\pi r f_s(r) \cdot \frac{1}{2\pi} \mathbb{1}_{(0,2\pi)}(\theta)$$

Therefore, the marginal distribution for r is

$$f_s^*(r) = 2\pi r f_s(r)$$

Here we discuss two classes of parametric isotropic spatial distributions and their corresponding spatial distance distributions. For the isotropic Gaussian distribution centered on (x_i, y_i), the spatial density $f_s(x, y) = \mathrm{N}(x|x_i, \sigma^2)\mathrm{N}(x|y_i, \sigma^2)$ can be equivalently expressed as $f_s(r) = (2\pi\sigma^2)^{-1} \exp\{-r^2/(2\sigma^2)\}$ in terms of $r = \sqrt{(x - x_i)^2 + (y - y_i)^2}$. The spatial distance distribution has the following density $f_s^*(r) = r/\sigma^2 \exp(-r^2/(2\sigma^2))$, which is the density for a Weibull distribution with shape parameter 2 and scale parameter $\sigma\sqrt{2}$. Another popular isotropic spatial distribution is the Power Law distribution which allows the tail of the distribution to decay at polynomial rate. Under the Power Law distribution, $f_s(x, y) = \pi^{-1}\alpha\beta^{\alpha}(\beta + x^2 + y^2)^{-(\alpha+1)}$, and the spatial distance density is $f^*(r) = 2\alpha\beta^{\alpha} r(\beta + r^2)^{-(\alpha+1)}$.

The spatial distance distribution provides another way to simulate from the corresponding isotropic spatial distribution. We first generate the distance to center r according to

$f^*(r)$, then the angle θ in polar coordinates from a uniform distribution over $[0, 2\pi]$, and finally derive the spatial location as $(x, y) = (x_i + r\cos(\theta), y_i + r\sin(\theta))$. We use the inverse-c.d.f. method to sample from $f^*(r)$ in the case of the Power Law distribution. The c.d.f. of the spatial density distribution is $F^*(r) = 1 - \beta^\alpha(r^2 + \beta)^{-\alpha}$

4.4.2 A nonparametric spatial distance intensity model

Under the separable assumption, the triggering function $g(\boldsymbol{s} - \boldsymbol{s_i}, t - t_i)$ can be factored into two intensity functions in space and time respectively:

$$g(\boldsymbol{s} - \boldsymbol{s_i}, t - t_i) = h_s(\boldsymbol{s} - \boldsymbol{s_i}, t - t_i)h_t(t - t_i) = h_s(r)h_t(t - t_i)$$

Instead of assuming a parametric functional form for the space intensity function $h_s(\boldsymbol{s} - \boldsymbol{s_i})$, we rely on the isotropy assumption and model the spatial component via the spatial distance density $f_s^*(r)$. Specifically we propose a nonparametric model for the spatial distance intensity function $h_s^*(r)$. Such intensity function is the spatial distance density function $f_s^*(r)$ scaled by a total intensity γ_s, i.e., $h_s^*(r) = \gamma_s f_s^*(r)$. The prior model we propose for $h_s^*(r)$ implies a Bernstein-Dirichlet prior scaled to a support of $(0, R_{max})$ for the spatial distance density $f_s^*(r)$. It follows that $h_s^*(r) = 2\pi r h_s(r)$ since $f_s^*(r) = 2\pi r f_s(r)$. We model the spatial component $h_s(r)$ nonparametrically via a prior on its scaled form $h_s^*(r)$. The time intensity function $h_t(t - t_i)$ follows the parametric form such that $h_t(t - t_i \mid \gamma_t, \omega) = \gamma_t \text{Exp}(t - t_i \mid \omega)$, where ω is the rate parameter for the exponential distribution.

Given the strict support assumption that the point process occurs over a compact domain $\mathcal{D}$, the support for the spatial distance intensity function is bounded from above by some

real value R_{max} that depends on the shape of the irregular domain $\mathcal{D}$. The default choice of R_{max} is the largest pair-wise distance among any point in $\mathcal{D}$. $(0, R_{max}]$ is the shared support for any offspring process that occurs within $\mathcal{D}$. The actual observed distance interval for a certain offspring process depends on the parent location. Similar to the parametric model, the offspring points are centered on the location of the parent point. Depending on the specific parent location (x_j, y_j), the offspring-parent distance can be only observed on a truncated interval $(0, R_j)$ where R_j, $R_j < R_{max}$, is the maximum distance between the parent location (x_j, y_j) to the boundary $\mathcal{D}$. Given its bounded support, we propose to model the spatial intensity function $h_s^*(r)$ as a weighted combination of beta densities scaled to $(0, R_{max})$ with structured gamma priors for the weights $\{V_l\}$:

$$\begin{gathered} h_s^*(r) = \sum_{l=1}^{L} V_l \tfrac{1}{R_{max}} \text{be}(r/R_{max} \mid l, L-l+1) \quad r \in (0, R_{max}] \\ V_l \overset{ind.}{\sim} \text{Ga}(\alpha_L/L, C_L) \end{gathered} \tag{4.5}$$

Such construction implies a Bernstein-Dirichlet Prior scaled to $(0, R_{max}]$ for the spatial distance density $f_s^*(r)$. The spatial distance density $f_s^*(r)$ as the result of the prior model for $h_s^*(r)$ is a mixture of the same set of beta densities. The normalized weights $\omega_l = \sum_l V_l / (\sum_p V_p)$ have a Dirichlet distribution with concentration parameters $(\alpha_L/L, \cdots, \alpha_L/L)$.

$$\begin{gathered} f_r^*(r) = \sum_{l=1}^{L} \omega_l \tfrac{1}{R_{max}} \text{be}(r/R_{max} | l, L-l+1) \\ \{\omega_l : l = 1 \cdot L\} \sim \text{Dir}(\{\alpha_o/L, \cdots, \alpha_L/L\}) \end{gathered} \tag{4.6}$$

These results are equivalent to a Bernstein-Dirichlet prior for $f_s^*(r)$ scaled to $[0, R_{max}]$, which has a uniform baseline distribution F_0 over $[0, R_{max}]$ and scaled Bernstein beta densities.

Under such prior model for $h_s^*(r)$, and $h_s(r) = h_s^*(r)/2\pi$, the offspring Poisson process intensity function, .i.e. the triggering function for an offspring process with parent

location (x_j, y_j, t_j) is defined as the follows:

$$g(r_{i,j}, t_{i,j}) = \frac{1}{2\pi r_{i,j}} \gamma_t \text{Exp}(t_{i,j} \mid \omega) \sum_{l=1}^{L} V_l \phi_l(r, R_{max}) \tag{4.7}$$

where $\phi_l(r, R_{max}) = R_{max}^{-1} \text{be}(r/R_{max} \mid l, L-l+1)$ is the l-th Bernstein basis function scaled to $(0, R_{max}]$.

To evaluate the normalizing constants for offspring Poisson processes, we use the following approximation to simplify the evaluation:

$$\begin{aligned}
\int_{\mathcal{D}} g(\boldsymbol{s} - \boldsymbol{s_j}, t - t_j) \mathrm{d}\boldsymbol{s} &\approx \int_0^{R_j} \int_0^{2\pi} \gamma_s f_s(r) \gamma_t f_t(t - t_j) r \mathrm{d}r \mathrm{d}\theta \\
&= \int_0^{R_j} 2\pi r h_s(r) \gamma_t f_t(t - t_j) \mathrm{d}r = \int_0^{R_j} h_s^*(r) \gamma_t f_t(t - t_j) \mathrm{d}r \\
&= \gamma_t f_t(t - t_j) \sum_{l=1}^{L} V_l B_{l,j}
\end{aligned}$$

where $B_{l,j} = \int_0^{R_j} \phi_l(r, R_{max}) \mathrm{d}r$. We use the circle centered on (x_j, y_j) with radius R_j as the region for integration, which always covers $\mathcal{D}$, as an approximation for the integration region $\mathcal{D}$. Fig. 4.1 shows the relationship between R_j and R_{max} over a hypothetical irregular domain $\mathcal{D}$. Such approximation works well when the effective range of $h_s^*(r)$ is way smaller than R_j, and the integrand is effectively 0 outside of $\mathcal{D}$.

The complete data likelihood for the offspring points O conditional on the latent parent label $[i]$ is approximately

$$p(O \mid \{V_l\}, \gamma_t, \omega) \approx \exp(-\sum_{j=1}^{N} \gamma_t F_t(T - t_j \mid \omega) \sum_{l=1}^{L} V_l B_{l,j}) \prod_{i \in O} g(r_{i,[i]}, t_{i,[i]})$$

where the approximation comes from the approximation in the integral term.

To achieve easier inference, we introduce latent basis label l_i for every offspring point $i \in O$ to indicate the Bernstein basis function that the offspring-parent distance $r_{i,[i]}$ is associ-

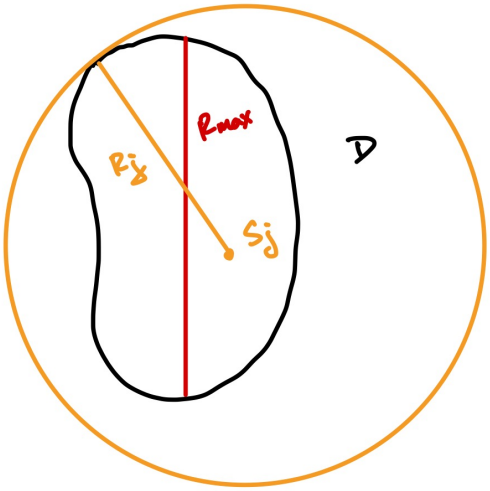


Figure 4.1: Illustration of R_j and R_{max} over $\mathcal{D}$: the orange circle is centered on a point $\boldsymbol{s_j}$ with radius R_j.

ated with. The hierarchical model for the offspring process is therefore

$$\begin{gathered}\{(x_i, y_i, t_i)\} \mid \{V_l\}, \{l_i\} \sim \exp(-\textstyle\sum_{j=1}^{N} \gamma_t H_j \sum_{l=1}^{L} V_l B_{l,j}) \prod_{i \in O} \frac{1}{2\pi r_{i,[i]}} \gamma_s \phi_{l_i}(r_{i,[i]}) \gamma_t \text{Exp}(t_{i,[i]}) \\ l_i \mid [i] = j \overset{i.i.d.}{\sim} \textstyle\sum_{l=1}^{L} \frac{V_l}{\gamma_s} \delta_l(l_i) \quad V_l \overset{ind.}{\sim} \text{Ga}(\alpha_L / L, C_L)\end{gathered} \tag{4.8}$$

where $H_j = \int_0^{T-t_j} \text{Exp}(u \mid \omega) du$ and $\phi_{l_i}(r_{i,[i]})$ implicitly depends on R_{max}.

4.4.3 Specification for R_{max} and R_j

R_{max} is the maximum distance between any possible pair of points in $\mathcal{D}$. The complexity of computing R_{max} is mostly determined by the complexity of the irregular domain boundary. We find that such computation can be much simplified when $\mathcal{D}$ is a convex polygon for which R_{max} is achieved at the largest distance among any pair of boundary nodes. When $\mathcal{D}$ is convex, the line that connects any pair of boundary points is located inside of $\mathcal{D}$ and is longer than any lines that connect the points along with it. Therefore, R_{max} should be achieved

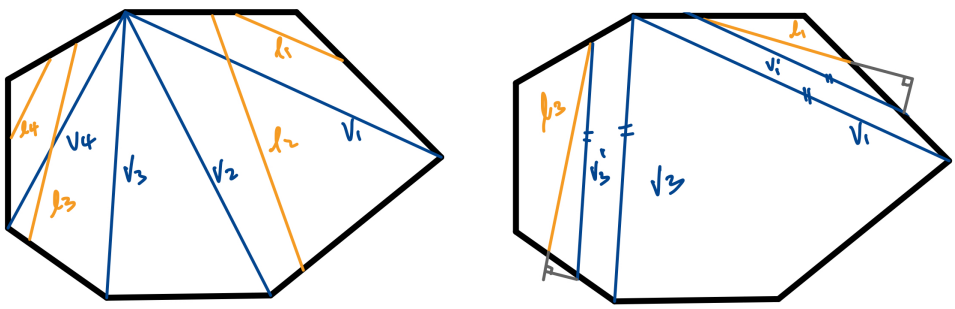


Figure 4.2: Illustration of R_{max} when $\mathcal{D}$ is a convex polygon.

at the distance of a line that connects boundary points. Fig. 4.2 demonstrates how the orange lines (l_1, l_2, l_3, l_4) that connect general boundary points will be shorter than some blue lines $(v_1, v_2, v3, v_4)$ that connects the boundary nodes. The right panel illustrates two possible scenarios. We can prove that $||\ l_1\ ||<||\ v_1\ ||$ by adding v_1', a line that starts with one end of l_1 and is parallel to v_1. It is easy to see that $||\ v_1\ ||>||\ v_1'\ ||>||\ l_1\ ||$. Similarly adding v_3' parallel to v_3 makes it easy to see that $||\ v_3\ ||>||\ v_3'\ ||>||\ l_3\ ||$. Such a method applies in general, and we can always find a line connecting two nodes longer than the line connecting a certain pair of boundary points.

R_j is the maximum distance between a point $\boldsymbol{s_j}$ and the boundary $\mathcal{D}$. R_j can be easily computed when $\mathcal{D}$ is convex. Fig. 4.3 shows that for any orange line that goes from $\boldsymbol{s_j}$ to a boundary point that's not the node, some blue line that connects such point and a node will be longer than such orange line. Therefore, R_j can be computed as the maximum distance between $\boldsymbol{s_j}$ and boundary nodes given $\mathcal{D}$.

Notice that the definitions of R_{max} and R_j only depend on $\mathcal{D}$ and can be computed

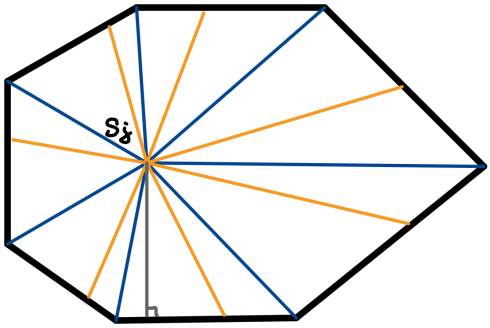


Figure 4.3: Illustration for R_j when $\mathcal{D}$ is a convex polygon.

as constants once given $\mathcal{D}$. The normalizing constants in the Bernstein basis functions for the offspring spatial distance density $B_{l,j} = \int_0^{R_j} \phi_l(u, R_{max})\mathrm{d}u$ depends both on R_j and R_{max} and can be computed before the posterior simulation. Thus, the computation in the MCMC algorithm is not dependent on the shape of the irregular domain shape once these constants are computed using $\mathcal{D}$. Such a feature is an improvement compared to the parametric model, where the complexity of $\mathcal{D}$ directly determines the cost of computing normalizing constants that occurs multiple times in each MCMC iteration. For the semi-parametric model, as long as these constants can be computed before the running the MCMC algorithm, the model scales well to point patterns over more complex boundaries.

4.4.4 Posterior simulation

The posterior simulation for the full model involves the following three steps: update the latent parent label $[i]$ for each point and identify the set of immigrants I and offspring points O, update the immigrant parameter $\boldsymbol{\theta_I}$ with points in I as data and then update the offspring

parameter $\boldsymbol{\theta_o}$ based on points in O. We skip the posterior updates for the immigrant parameters since they are slight modifications to those shown in Chapter 2. We focus on the updates for the parent label and offspring parameters.

The full-conditional for the latent parent label is a discrete distribution with support as the union of all available parent set C_i and 0 which suggests that the point is an immigrant. The available parent set C_i is the index of points that occurs prior to point i such that $C_i = \{j : t_j < t_i\}$. The full-conditional of $[i] = j$ is proportional to either the immigrant intensity evaluated at $(\boldsymbol{s_i}, t_i)$ when $j = 0$ or the triggering function evaluated at $(\boldsymbol{s_i}, t_i)$ given the parent of i is j.

$$p([i] = j \mid -) \propto \begin{cases} \sum_{k_x,k_y=1}^{K} V_{k_x,k_y} B_{k_x,k_y} \phi^*_{k_x,k_y}(x_i, y_i) & j = 0 \\ \frac{1}{2\pi r_{i,j}} \gamma_t \text{Exp}(t_{i,j} \mid \omega) \sum_{l=1}^{L} V_l \phi_l(r, R_{max}) & j \neq 0 \end{cases}$$

The offspring parameter set $\boldsymbol{\theta_o} = \{\{V_l\}, \{l_i\}, \gamma_t, \omega\}$ is updated in the following steps: first update $\{l_i\}$ from a discrete full-conditional distribution where $p(l_i \mid [i] = j) \propto V_l \phi_l(r_{ij}, R_{max})$; update $\{V_l\}$ from conditionally independent Gamma distribution Ga($\alpha_o/L + M_l, C_o + \gamma_t \sum_{j=1}^{n} H_j B_{l,j}$), where M_l is the number of offspring-parent distances associated with basis l; then update γ_t from a truncated Gamma distribution over $[0, 1]$ with shape parameter $\alpha_{\gamma_t} + |O|$, where $\mid O \mid$ is the number of offspring points, and rate parameter $\sum_{j=1}^{N} H_j \sum_{l=1}^{L} V_l B_{l,j}$; finally update ω with a Metropolis-Hastings step.

4.5 Forecast

We present two kinds of forecast approaches for the space-time Hawkes model: the *marginal* forecast and the *conditional* forecast. Let the point pattern be observed in the time window $(0, T]$ and let h be the forecast period length. The marginal forecast simulates a realization from the underlying process over $\mathcal{D} \times (0, T + h]$ without using observed data points in such simulation. Alternatively, the conditional model uses the set of inferred immigrant points in $(0, T]$ under each posterior sample and combines them with forecast immigrant points in $(T, T + h]$ to be the complete immigrant sets in $(0, T + h]$. These immigrant points then continue to produce offspring points until no more offspring is generated. Both approaches rely on a routine to simulate the immigrant point locations under the BPNHPP model over $\mathcal{D}$ and timestamps uniformly in a specific time window, and a routine to simulate offspring points based on a set of immigrant points. The subset of simulated point patterns that fall in the time window $(T, T + h]$ is the forecast point pattern under both approaches.

For the marginal model, the immigrant points are simulated as a realization from the BPNHPP model over irregular domain $\mathcal{D}$ over $(0, T + h]$. We simulate the total number of immigrant points from a Poisson distribution with mean $(T + h) \sum_{k_x,k_y} V_{k_x,k_y} B_{k_x,k_y}$, the basis labels (ξ_i, η_i) associated with each observation from $\{(1, 1), \cdots, (K, K)\}$ with weights proportional to $V_{k_x,k_y} B_{k_x,k_y}$ for $k_x, k_y = 1 \cdots, K$, and finally the actual locations (x_i, y_i) according to the Bernstein spatial basis function $\phi^*_{\xi_i,\eta_i}(x, y)$. For the conditional model, the immigrant points in $(0, T)$ are observed points with parent label $[i] = 0$. Then the immigrant points in $(T, T + h)$ are a realization from the BPNHPP model over $\mathcal{D}$ in $(T, T + h)$.

Given a certain immigrant point, the offspring points are then generated recursively centered on the initial immigrant and all its offsprings. A particular family given the parent point (t_j, x_j, y_j) is generated from a Poisson process with intensity defined as the triggering function. We first generate the number of children n_j from a Poisson process with mean $\gamma_t H_j \sum_{l=1}^{L} V_l B_{l,j}$. To obtain offspring's timestamps, we simulate n_j time shifts δt from a truncated exponential distribution with mean $1/\omega$. The offspring location shifts $(\delta x, \delta y)$ are obtained by first simulating n_j spatial distances r, n_j polar coordinate angles θ uniformly over $(0, 2\pi)$, and mapping (r, θ) to $(\delta x, \delta y)$ via $\delta x = r\cos\theta$ and $\delta y = r\sin\theta$. To simulate from the nonparametric spatial distance density, we first sample n_j basis label l_i from $\{1, \cdots, L\}$ with weights proportional to V_l and then the distances r from the Bernstein density over $(0, R_{max})$ indexed by l_i. We map these location and time shifts to the offspring locations via $(x_i, y_i, t_j) = (x_j + \delta x, y_j + \delta y, t_j + \delta t)$. In the process of generating offspring points, We ensure proper truncation by rejecting sampling. Specifically, we obtain a sample of (x_i, y_i) of size 5 to 10 times of n_j, and keep only the first n_j of them that are inside $\mathcal{D}$.

The simulation stops when no more offspring points are generated. The subset of the resulting point pattern which occurs in $(T, T + h]$ is the forecast point pattern.

4.6 Simulation study

We simulate the synthetic data from parametric space-time Hawkes processes with Gaussian and Power Law offspring densities to demonstrate the model's capacity to capture different tail behaviors introduced by the parametric form of the offspring density.

We first describe the simulation algorithm for the space-time Hawkes process, given the irregular domain $\mathcal{D}$, the total immigrant intensity Λ over space and time, the observed time window T, the branching ratio $\gamma < 1$, the immigrant spatial density function $f_I(x, y)$, and the offspring density function $f_o(x, y, t) = f_s(x, y)f_t(t)$. Notice that $f_s(x, y)$ and $f_I(x, y)$ are both proper densities defined over $\mathcal{D}$.

1. Simulate N_o immigrants locations (x_i, y_i), where $N_0 \sim$ Poi(Λ), i.i.d. from the immigrant spatial density function $f_I(x, y)$; simulate their timestamps t_i uniformly over $(0, T]$.

2. For each immigrant point j, generate n_j offspring points, where $n_j \sim$ Poi(γ), with location (x_i, y_i) following the offspring spatial density $f_s(x, y)$ centered on their parent locations (x_j, y_j) and time increment δ_t following the offspring temporal density $f_t(t)$ truncated over $(0, T - t_j)$. The event time for the offspring $t_i = t_j + \delta t$.

3. Repeat step 2 for each offspring point and their offspring points until no more offspring points are generated.

4. The simulated point pattern is the union of all immigrant points and offspring points.

4.6.1 Sensitivity analysis

We design the following simulation study where the point pattern is generated over the unit-square with an immigrant density $f_o(x, y)$ as a mixture of four bivariate beta densities, which together produce three distinct density modes, an offspring temporal density of an exponential distribution with mean 0.2, and a branching ratio $\gamma = 0.3$. The simulation scenarios are created based two time horizons $T = 20, 400$ and the following offspring densities:

$\text{N}(\mathbf{0}, \boldsymbol{I}_2 0.05^2)$, $\text{N}(\mathbf{0}, \boldsymbol{I}_2 0.1^2)$, and $\text{PL}(2, 0.05)$. The total immigrant intensity over space and time is set to 800 for all cases, resulting in point patterns with around 800 immigrant points and 300 offspring points. The goal here is to examine whether the model can recover the spatial distance density from different parametric families and test the prior sensitivity for hyperparameter values.

The model configuration consists of the following aspects: immigrant hyperparameters α and C, offspring hyperparameters α_L, C_L, and prior for the temporal branching ratio γ_t. For the immigrant parameters, we set C to value $(0.1, 0.5)$, and choose α by setting α/C to the total intensity at any time, i.e. Λ/T. Additionally, we set α to be $\Lambda/T * C * 0.9$ to create a scenario where the prior total intensity over space at any time is under-estimated. We set the ratio between offspring hyperparameters α_L/C_L to 1 in the hope that γ_t will be centered on γ. The prior for the spatial branching ratio $\gamma_s = \sum_l V_l$ is $\text{Ga}(\alpha_L, C_L)$, and has expectation α_L/C_L. The model can only jointly identify the the total branching ratio $\gamma = \gamma_t * \sum_{V_l}$. By setting α_L/C_L to 1, we can specify the prior for γ_t according to the prior knowledge of γ. We choose α_L, C_L values from $(0.01, 0.01)$ and $(1.0, 1.0)$ with the latter suggesting stronger prior belief that the spatial branching ratio γ_s is centered around 1. We choose the prior for γ_t from two truncated Gamma distributions centered on the branching ratio γ with varying amounts of prior precision. In total, there are 16 model configurations initially.

We apply these 16 configurations to data simulated from space-time Hawkes processes with $\text{PL}(2, 0.05)$ over the time horizon $T = 800$ and $T = 1000$. The extended timeline makes it easier for the model to identify the latent branching structure, since the clusters of points are well separated in the time dimension. We discover that when offspring hy-

perparameters (α_L, C_L) is set to (1.0, 1.0), the model produces 95% credible intervals that cover the true branching ratio $\gamma = 0.3$ across other aspects of model configuration. When $(\alpha_L, C_L) = (0.01, 0.01)$, the model tends to under-estimate the branching ratio and over-estimate the total immigrant intensity over space-time.

We then fix $(\alpha_L, C_L) = (1.0, 1.0)$ and apply eight configurations to data generated with isotropic Normal densities with standard deviations of 0.05 and 0.1, and time horizons $T = 20$ and $T = 400$. Together we have 32 cases across these 4 data sets. Fig.4.4 presents the posterior mean and 95% credible interval for the number of offspring points, the branching ratio, and the total immigrant intensity over space-time across these 32 cases. One noticeable trend is that when the time horizon is closer, i.e., $T = 20$ in case 1 to 16, the credible interval tends to be wider, suggesting more uncertainty. Such behavior suggests difficulties for the model to distinguish between immigrant and offspring points when the points are dense in the time dimension. With a further time horizon, i.e., $T = 400$ case 17-32, the inference for the immigrant total intensity and number of offspring points is relatively stable across different model configurations. Therefore we conclude that the model is not sensitive to immigrant hyperparameter miss-specification and prior choice for γ_t. In all 32 cases, the 95% credible interval covers the actual immigrant total intensity of 800 and branching ratio of 0.3, and in most cases, covers the true number of offspring points of 330.

We then apply the eight configurations to data simulated with offspring from PL$(2, 0.05)$ over time horizons $T = 20$ and $T = 400$, which leads to 16 cases. We observe a similar trend that a shorter time horizon produces larger credible intervals. The estimation of the total immigrant intensity and number of offspring points are similar across the 16 cases. We observe a

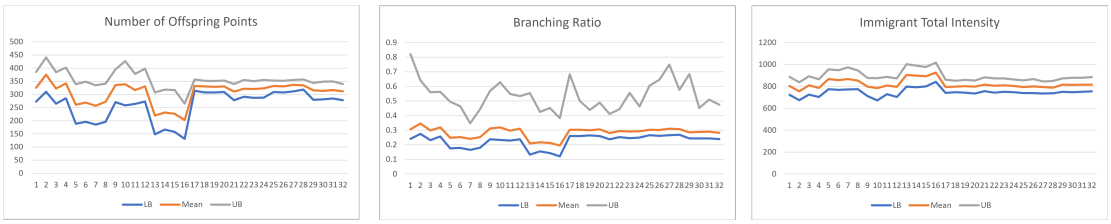


Figure 4.4: Posterior inference for the number of offspring points, the branching ratio and the immigrant total intensity across model configurations using synthetic data with Gaussian offspring kernel.

systematic overestimation of the total intensity and underestimation of the number of offspring. Further investigation suggests that the model tends to misidentify offspring points further away as immigrant points since such Power Law offspring density results in a long tail for the spatial distance density.

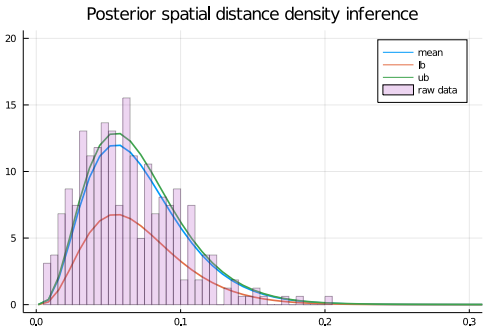
Additionally, we study the model performance under the PL truth with the following offspring densities: $PL(2, 0.01), PL(4, 0.05)$ and $PL(3, 0.1)$ within time window $(0, 20]$ and $(0, 400]$. The total intensity over space and time is again set to 800, resulting in point patterns around 1200 points. We apply the model under the default setting: informative prior on the branching ratio γ_t, $(\alpha_L, C_L) = (1.0, 1.0)$ and $C = 0.1$. We observe that the model performs better in inference in these scenarios. The 95% credible intervals for the branching ratio and total intensity cover the truth in all six scenarios. The 95% credible intervals for the number of offspring cover the truth in all scenarios except when $T = 20$ and offspring density truth is PL$(3, 0.1)$.

4.6.2 Inference on the spatial distance density

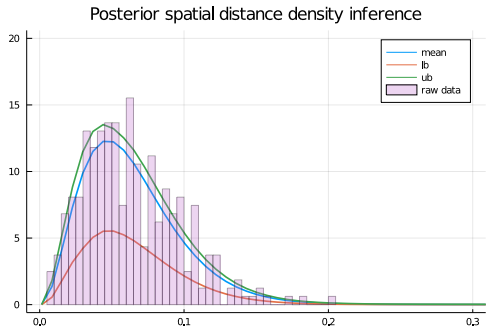
Section 4.4.1 derives the spatial distance density $f_s^*(r)$ for the isotropic Gaussian and Power Law distribution without any spatial truncation. Here we emphasize that given our simulation method, we generate the offspring points from a parametric spatial density truncated over the domain $\mathcal{D}$. Therefore, the corresponding $f_s^*(r)$ over $\mathcal{D}$ depends on $\mathcal{D}$ and does not have a tractable analytical form. Instead, we use the histogram of the actual offspring-parent distance as the benchmark to examine the inference of the spatial distance density. Given the offspring set O and parent label $[i]$, the distances $r_{i,[i]}, i \in O$ can be easily computed. We plot the true offspring-parent distance histograms and present the posterior mean and 95% interval of the offspring spatial distance density for Gaussian cases in Fig. 4.5 - Fig. 4.6, and for Power Law cases in Fig. 4.7. We observe that the model performs better when the timeline is longer ($T = 400$), especially at capturing the tail behavior when the effective range of the offspring spatial density is large. The model captures the mode and tail of the spatial densities from both Gaussian and Power Law families with varying levels of success depending on the time horizon.

4.6.3 Forecast

We examine the model capacity to forecast in a future time window that is of 0.5 length of the observed timeline for the following spatial distance density truth: $N(\mathbf{0}, 0.05\boldsymbol{I_2})$, $N(\mathbf{0}, 0.1\boldsymbol{I_2})$, $PL(2, 0.05)$, $PL(2, 0.01)$, $PL(4, 0.05$, $PL(3, 0.1)$ over two time horizons $T = 20, T = 40$. To compare the model forecasts, we simulate the observed data in the time window $(0, 1.5T)$, fit the model on the training point pattern in $(0, T)$, and hold out the point pattern in $(T, 1.5T)$ to provide the forecast truth for validation. In the training and holdout data, we

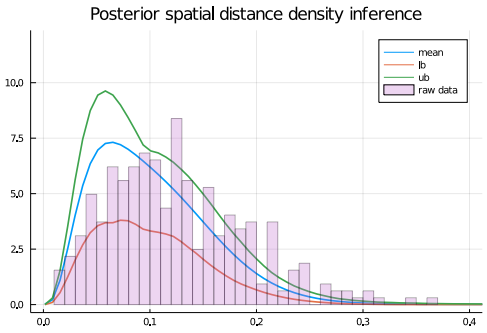


(a) $T = 20$

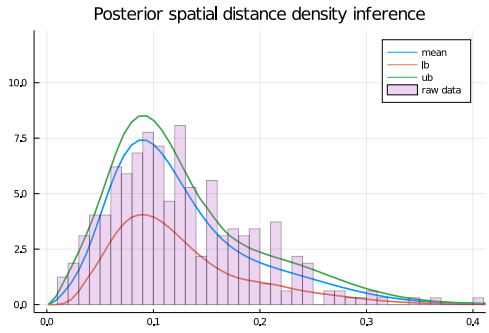


(b) $T = 400$

Figure 4.5: Posterior inference for the offspring spatial distance density under Normal$(0, 0.05)$ truth.



(a) $T = 20$



(b) $T = 400$

Figure 4.6: Posterior inference for the offspring spatial distance density under Normal$(0, 0.1)$ truth.

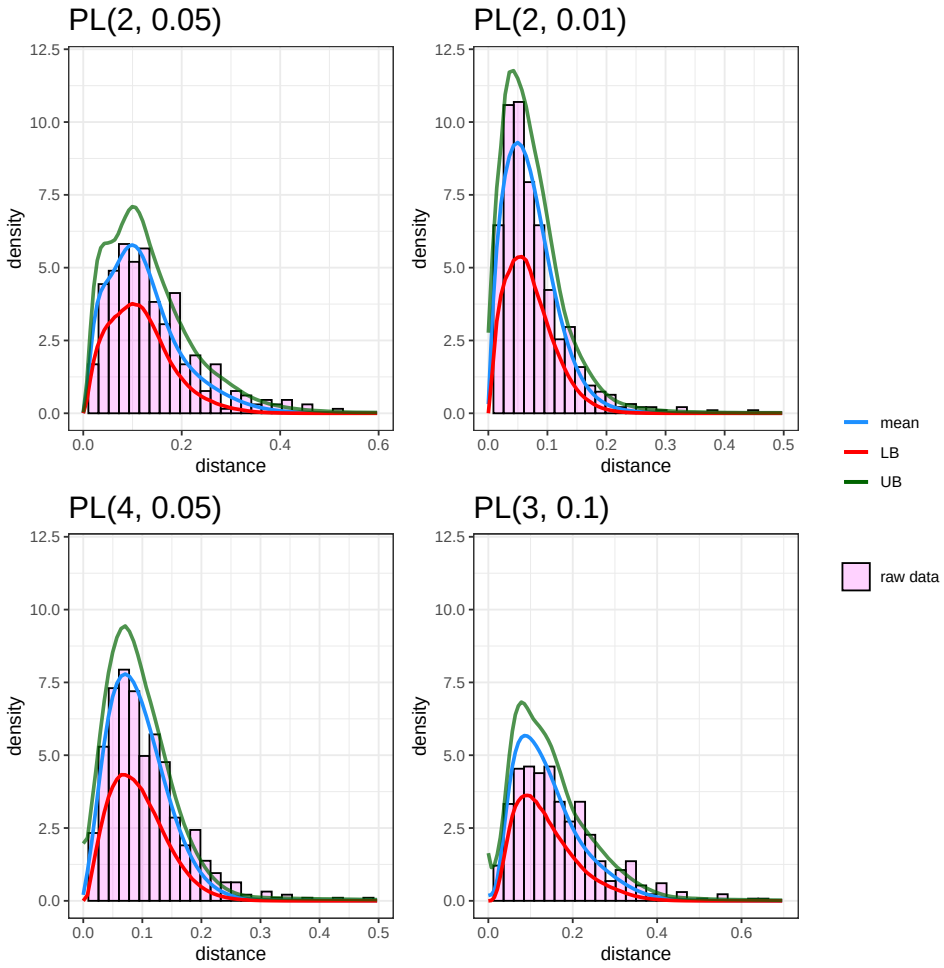


Figure 4.7: Posterior inference for the offspring spatial distance density under Power Law distributions $\mathrm{PL}(2, 0.05)$, $\mathrm{PL}(2, 0.01)$, $\mathrm{PL}(4, 0.05)$, $\mathrm{PL}(3, 0.1)$ when $T = 400$.

	Truth	Full Forecast	Conditional Forecast
Immigrant	399	418.65 (361, 478)	420.08 (362, 477.75)
Offspring	170	144.71 (95, 204.75)	144.76 (97.25 206)
Total	569	563.36 (490, 645)	564.84 (490.25, 642)

Table 4.1: Forecast result for data simulated under Gaussian offspring kernel $\mathrm{N}(\mathbf{0}, 0.05\boldsymbol{I_2})$ and $T = 20$. The forecast columns shows posterior mean and 95% forecast interval.

compute the number of immigrants, the number of offspring, and the total number of points as metrics for comparison. To obtain the posterior forecast interval for these metrics, we perform both the full and conditional forecasts based on each posterior sample and count the number of immigrants, offspring, and total points in the forecast point patterns. Table 4.1 shows the forecast results for data simulated with Gaussian offspring density with standard deviation 0.05 and over time horizon $(0, 20)$. The forecast is performed over the time window $(20, 30)$, and the true counts are shown in the first column. We observe that the full and conditional forecasts result in similar counts for the three metrics.

We examine the forecast performance across the simulation truth and time horizon. Fig 4.8 presents the point estimates and 95% credible intervals for the immigrant, offspring, and total counts across simulation scenarios with both Gaussian and PL offspring truth. The actual counts in the holdout point pattern are marked with triangles and the posterior mean with circles. We label the Gaussian cases in the format of "σ - T", e.g., "05-20" refers to the case where the Gaussian offspring kernel has a standard deviation of 0.05 and the time horizon T is 20. Similarly, we label the PL cases in the format of "α-β-T", e.g., "2-05-20" is the case with

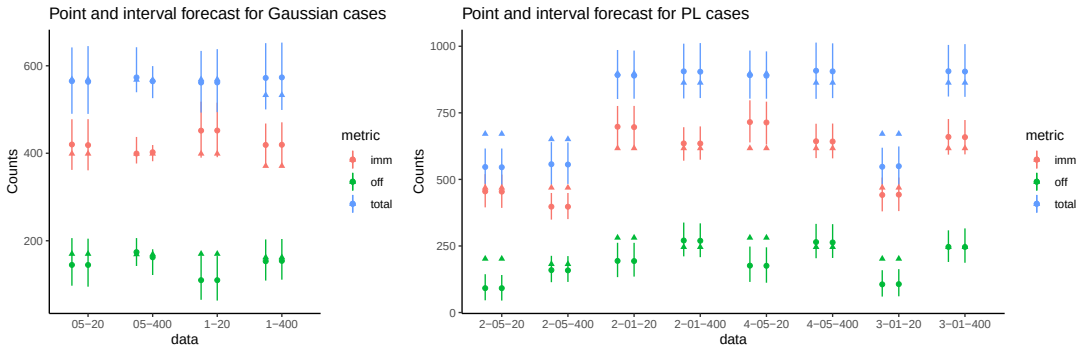


Figure 4.8: Posterior point and interval forecasts for immigrant, offspring, and total event counts in the holdout period under different simulation scenarios.

PL$(2, 0.05)$ offspring density and time horizon 20.

We observe that the forecast intervals cover the true counts for immigrants, offspring, and total counts in almost all Gaussian cases. The offspring count for the scenario "0.1-20" $(\sigma = 0.1, T = 20)$ is underestimated with the true count slightly larger than the forecast upper bound. The forecasts for the PL cases have mixed results depending on the truth and time horizon. We observe that for the $PL(2, 0.01)$, $PL(4, 0.05)$ and $PL(3, 0.01)$ cases, the forecasts under longer timeline $T = 400$ recover the truth for all three metrics. Under a shorter timeline, the model underestimates the offspring counts for all four cases.

4.7 Real data example: Boston city crime revisited

We apply the space-time Hawkes process models, the nonparametric offspring spatial model (NonparSTH), and the isotropic Gaussian offspring model (ParSTH) to the crime point pattern data in Boston city. We choose to model the Vandalism point pattern from April to

June in 2017, taking the first ten weeks to be the training set, and holding out the last three weeks as the testing set. We compare the spatial distance inference under the two models and their forecasting performance in both the temporal and spatial aspects. We perform both full and conditional forecasts with the NonparSTH and ParSTH models. Under each approach, we obtain a replicated point pattern over 13 weeks for each posterior sample and filter such point pattern to the last three weeks for validation. We compare the predicted number of events to the observed counts in the hold-out period to check temporal forecast accuracy. We obtain predictive residuals by filtering the predicted point pattern in the validation set to 9 by 9 grid cells over the unit square to subtract the observed counts from the predicted counts per cell. Finally, we use the Posterior predictive loss criteria treating the count of events in each grid cell as the observation unit to perform an informal model comparison.

Self-exciting point process model have been applied to crime modeling and forecast. The substantive problem is as follows: given the history of past events' location and time, rank a set of pre-specified spatial regions in a city according to the risk of future crimes. The spatial map that are color coded based on risk of criminal activity is called the crime hotspot map. To achieve this goal using a point process modeling approach, Mohler et al. (2011) developed a fully-nonparametric model for space-time Hawkes process to forecast burglary and it was later extended to a marked point process model that incorporate multiple crime types in Mohler (2014). Both compare point process based approach and other hotspot maps approaches based on a predictive accuracy measure using hotspot ranking. We adopt this approach to perform model comparison between NonparSTH and ParSTH.

We choose the central city region (excluding the northeast region and islands in the

east) to be the irregular domain $\mathcal{D}$, the same region we used in Chapter 3. We further simplify the domain to be the convex hull, making it easier to define R_{max} and R_j for the NonparSTH model. Spatially, we map the locations of points to the unit square and exclude any points outside the convex hull. We take the timestamp of each point and convert it to the number of weeks since 00:00:00 04-01-2017. The test period is the first ten weeks and the testing period is weeks 11 to 13. We discover some edge cases where the spatial locations of the point patterns are identical while the timestamps are different. This is potentially due to multiple crimes in the small vicinity of a specific location at different times, and the logging protocol registers the location to the same longitude and latitude. Such edge cases introduce estimation issues for the NonparSTH model. The first Bernstein density function in the offspring spatial distance density mixture, be$(r \mid 1, L)$, is a monotonic decreasing density with an asymptote at 0. Having multiple points at the same location in the data causes the NonparSTH model always to favor the first basis and results in biased estimations of the offspring spatial distance shape. To address this data issue, we identify pairs of points with a pair-wise distance equal to 0 and remove the point with a later timestamp. Such procedure removes about 15% of the points, and we obtain a point pattern of size 1018 over 13 weeks.

We apply the NonparSTH model with configurations consists of four prior on γ_t, two centered on 0.6 (Ga$(7, 10)$ and Ga$(61, 100)$) and another two centered on 0.2 (Ga$(2, 10)$ and Ga$(21, 100)$, three choices for number of offspring Bernstein densities used L: $L = 80, 100$ or 120. The inference for γ_t is not sensitive to the prior specification: posterior inference for γ_t are similar under the four priors with posterior mean for γ_t around 0.6, 2.5 posterior percentile around 0.5, and 97.5 percentile around 0.8. The offspring spatial density shows similar trend

under different choice of L with a density mode around 30 near $r = 0.02$. We show result under the NonparSTH model with prior Ga$(61, 100)$ for γ_t, Ga$(11, 10)$ for ω, and $L = 120$ here. We apply the ParSTH model with a Ga$(7, 10)$ prior on γ and Ga$(11, 10)$ prior on ω. We fix the offspring Gaussian parameter ρ to be 0 and set $\sigma = \sigma_x = \sigma_y$ under the isotropic assumption. We place a IG$(2, 0.0001)$ prior on σ^2. In both NonparSTH and ParSTH model, we set $K = 40$ and $C = 0.0005$ for the immigrant NHPP model with a fixed value for α. The MCMC chain for both models is run for 10,000 iterations, with the first 2000 iterations discarded and the rest of the chain thinned by 4.

We compare the inference for the offspring spatial distance density under the two models by summarizing the empirical density of the inferred parent-offspring distances. For each posterior sample of the latent branching structure, we compute the parent-offspring distances and apply a histogram estimator to such distances sample with 30 equally spaced bins over $(0, 0.15)$. We summarize the posterior empirical densities with posterior mean and 95% credible interval for the density in each bin. We then plot the posterior mean and interval estimate for both NonparSTH in blue and ParSTH in red in the right panel in Fig. 4.9. We observe that both models have a similar estimate for the offspring spatial density mode around 0.025. However, the NonparSTH model has a heavier tail. The left panel shows the inference for spatial distance density under the NonparSTH model. The middle panel shows the posterior inference of Weibull density functions with shape parameter two and scale parameter $\sigma^{(b)}\sqrt{2}$ for $b = 1 \cdots B$ in the posterior sample of σ. Such Weibull density is the spatial distance density for isotropic Gaussian kernel over $\mathbb{R}^2$ with standard deviation σ. Since we cannot obtain the closed-form expression for the spatial distance density over the irregular domain $\mathcal{D}$, these

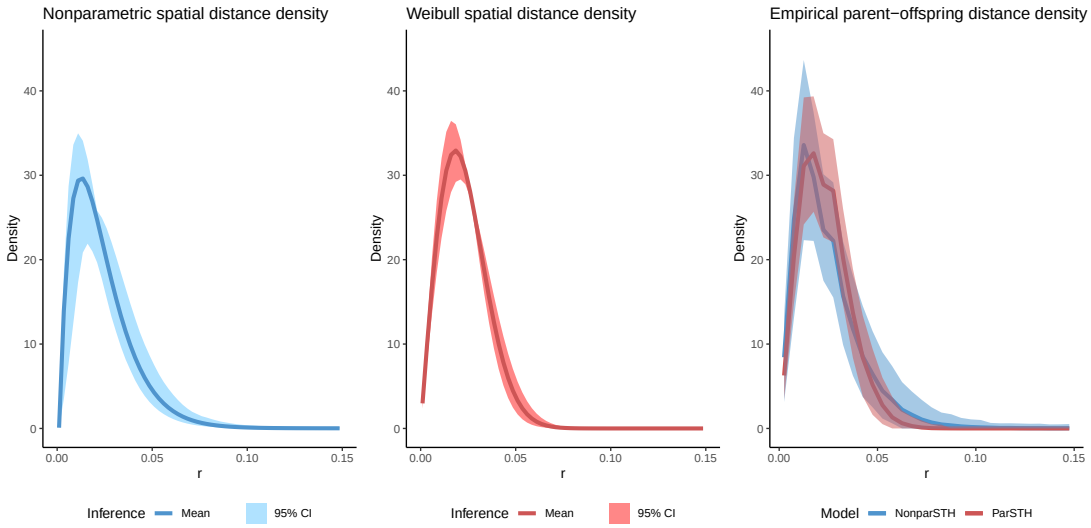


Figure 4.9: Posterior inference of offspring spatial distance density.

Weibull densities serve as crude approximations.

We compare the inference for the immigrant spatial density under the two models shown in Fig.4.10. The two point estimates for the density share a similar mode around (0.75, 0.5), and differs slightly on the location for the other mode around (0.6, 0.7). The difference is likely induced by different immigrant set I under the two models. Overall, these trends are consistent with what we see from results in Chapter 2 and Chapter 3.

We compare the two models in terms of forecast performance in the hold-out period. Temporally, we compare the forecast number of points in the hold-out period under both models to the truth. Such counts, simulated via the branching process that decides the number of generations and points in each family, are controlled by the total intensity of the immigrant process and the branching ratio. Since we do not have the information on immigrant vs. offspring identity in the hold-out set, we can only compare the total forecast counts. Fig. 4.11 shows the

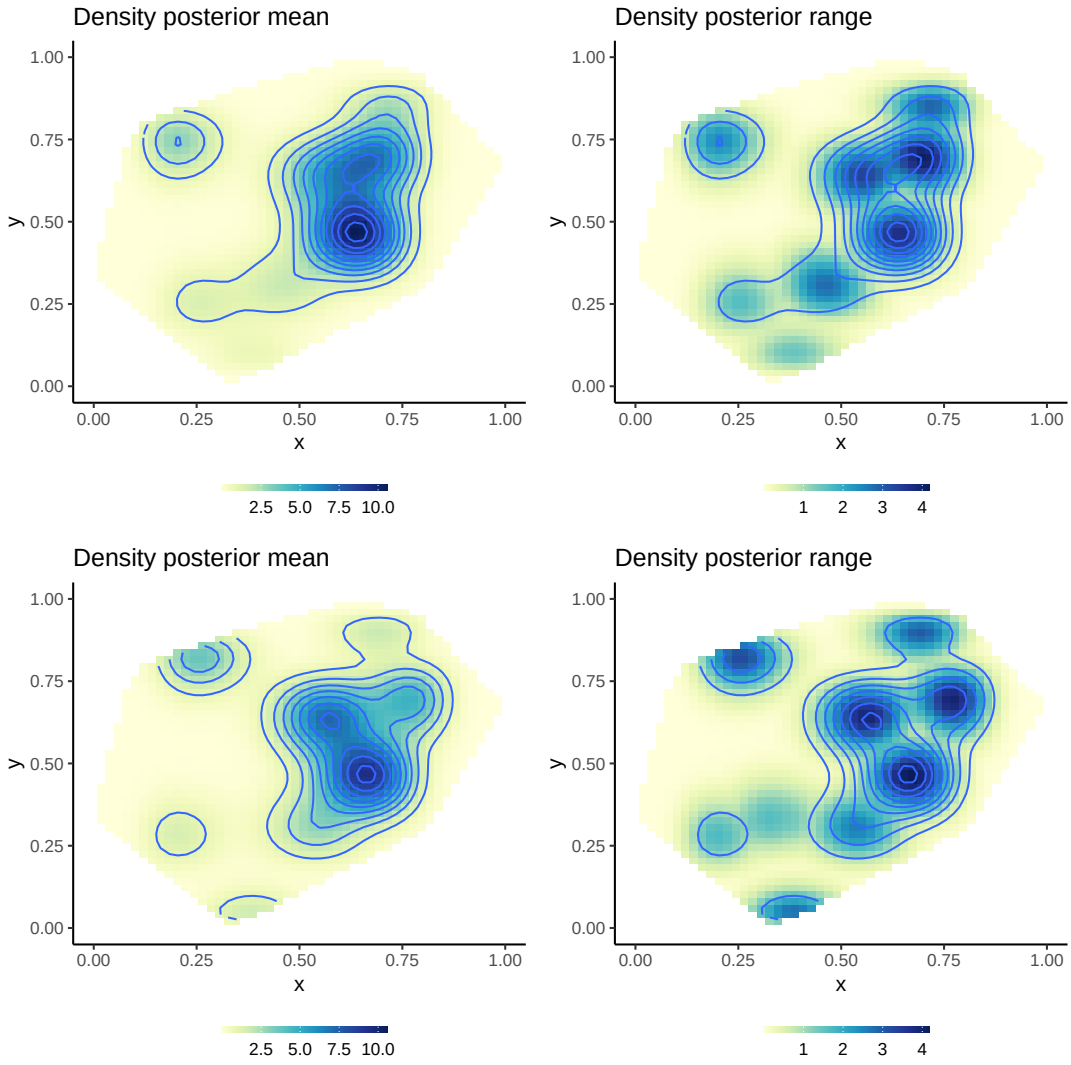


Figure 4.10: Posterior inference for G_0 spatial density under NonparSTH (first row) and ParSTH (second row).

comparison among the two NonparSTH models and the ParSTH model. NonparSTH achieves a slightly better forecast with point estimates closer to the truth for both the full and conditional forecasts.

Spatially, we use the posterior predictive residuals over grid cells to compare local forecast accuracy. Fig. 4.12 shows the predictive mean residuals for NonparSTH in the first row and ParSTH in the second row, the full forecast in the first column, and the conditional forecast in the second column. Table 4.2 shows the summary of the predicted residuals across grid cells, mean (min, max), for both models under both forecast methods. The average residual for the NonparSTH model is closer to 0 than that for the ParSTH model, with a smaller range using both the full and conditional forecast methods.

To come up with a comprehensive comparison metrics, we summarize the spatial forecasts with the predictive loss criterion. We treat the number of points in each of the 9 by 9 grid cells as individual observations, and use the sum of squares of difference between posterior mean estimates and observations as a measure for goodness of fit and the sum of posterior variance for each observation as a measure of penalty. Together these two terms provide a holistic comparison between models. We use this measure only as an informal model comparison scheme, since the independent observations assumption for posterior predictive loss criterion does not hold under the underlying space-time Hawkes process, which allows dependence among points located in different grid cells.

Table 4.2 shows the posterior predictive loss criterion using the full and conditional forecasts in the last column. The conditional forecast achieves lower loss for both models, and the NonparSTH performs better than the ParSTH under this criterion. Overall, we observe

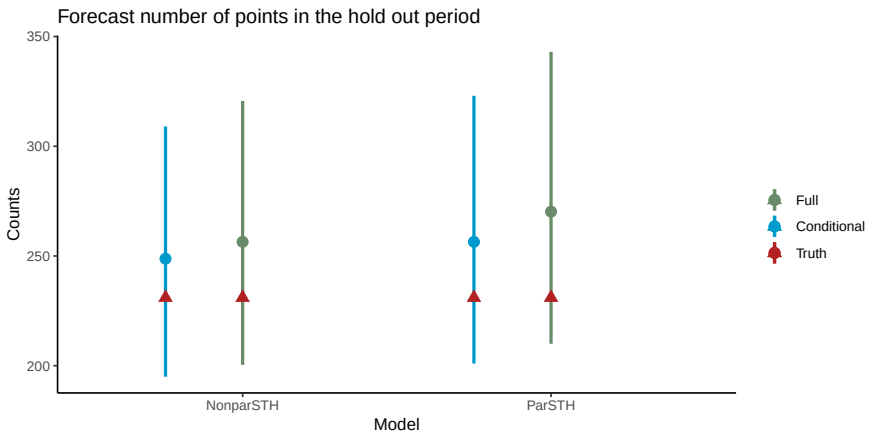


Figure 4.11: Forecast number of points in the hold out period under the two models.

Model	Full forecast	Conditional forecast	Predictive loss criterion
NonparSTH	0.314 (-7.369 , 7.542)	0.219 (-6.780, 6.954)	770.58 / 696.00
ParSTH	0.472 (-8.471,7.446)	0.327 (-7.727, 6.569)	894.70 / 776.28

Table 4.2: Posterior predictive performance of NonparSTH and ParSTH using full and conditional forecast method.

that the conditional forecast method shows better performance both temporally and spatially. NonparSTH achieves a better forecast.

We produce crime hotspot visualizations based on the posterior forecast in the holdout period under NonparSTH using the conditional forecast method to demonstrate the practical utility of our proposed methods. In each map from Fig. 4.13, the color for grid cells correspond to the observed event counts in the holdout period that are positive and the red highlights indicate the cells with top M forecast counts predicted by the model. Comparing the hotspot cells flagged by the model against those flagged by observed counts provides an informal check for

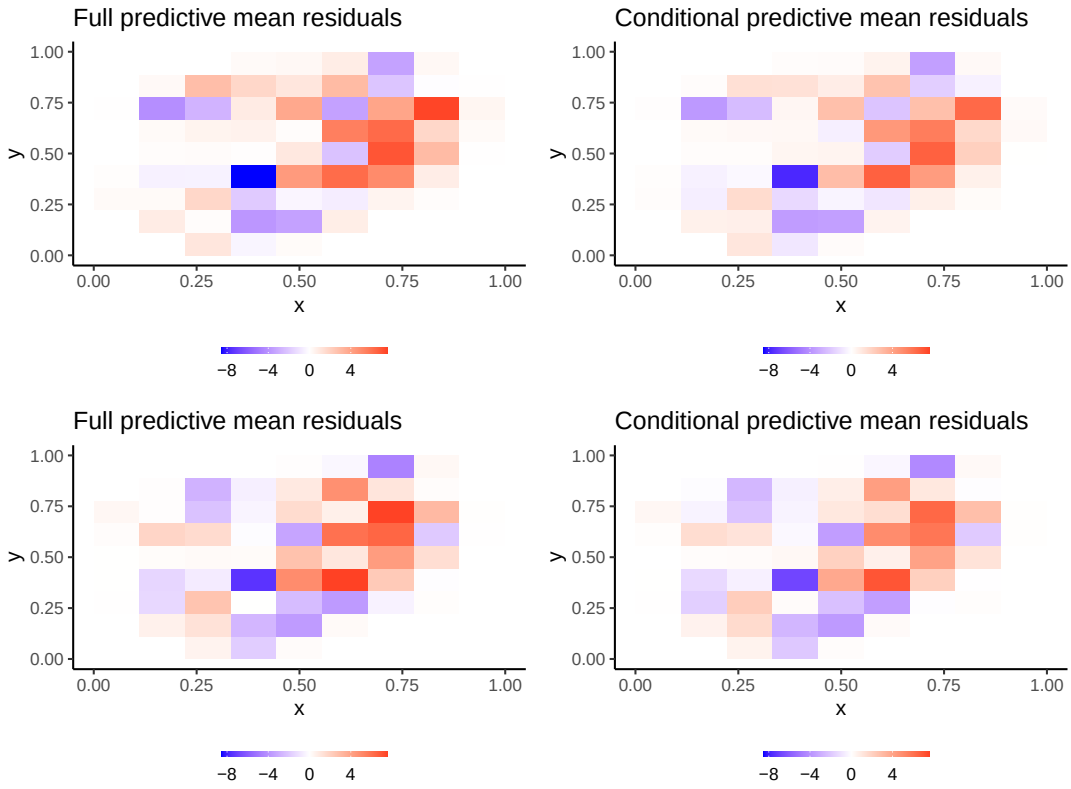


Figure 4.12: The posterior mean for the predictive residuals over a 9×9 grid in the unit square under NonparSTH (first row) and ParSTH (second row).

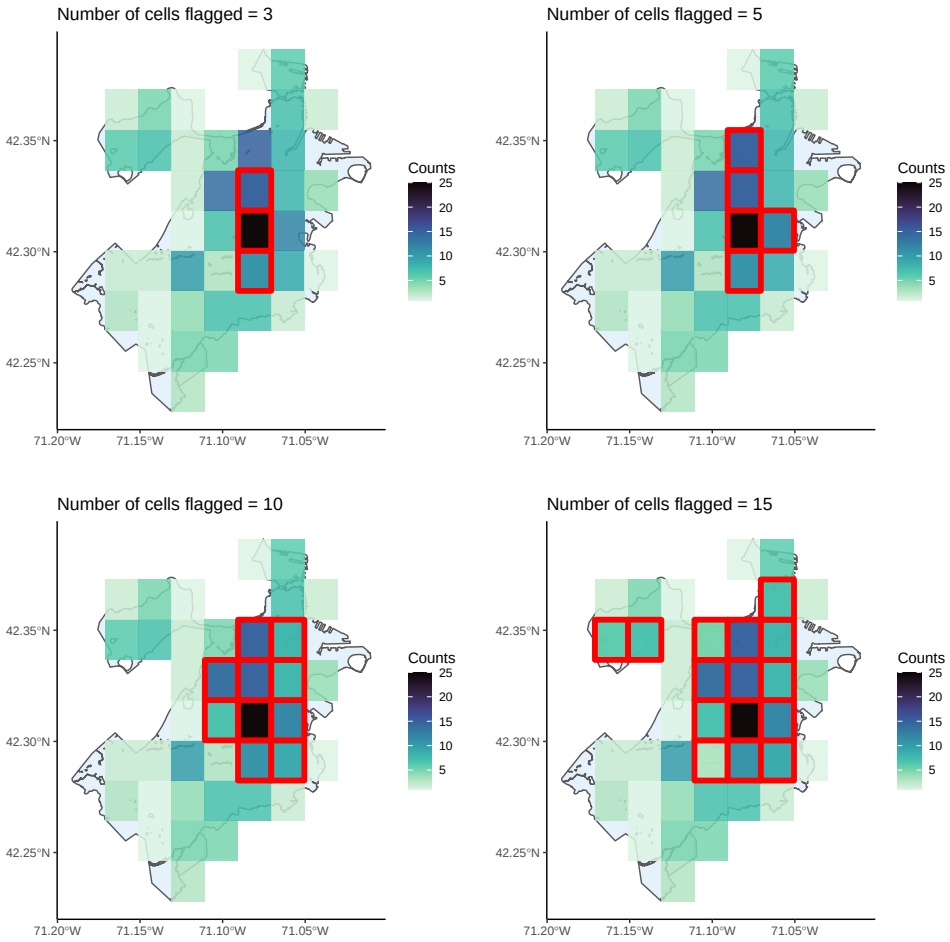


Figure 4.13: Vandalism hotspot maps under NonparSTH conditional forecast over 9×9 partition of the unit-square where the cells are colored by observed number of events in the holdout period, and the cells highlighted by red are chosen by the model with the most forecast number of events.

prediction accuracy. We observe that the model is able to flag the high risk areas with 10 cells, and is able to capture the second local mode in the north-west region with 15 cells. This map can be used to advice police intervention, as highlighted areas represent regions with higher risks for future crimes. Our approach can provide more information since the rank is based on actual forecast counts instead of an abstract measure of risks and is therefore more informative of the real-world consequences.

Finally, we compare two models using an empirical accuracy measure of a hotspot ranking method, which computes the percentage of observed crimes in the holdout period that fall into areas flagged for intervention according to predicted risk of future crimes. This metric is adopted by Mohler et al. (2011) and Mohler (2014) to compare models and resembles an ROC curve in the sense that larger area under the curve means higher prediction accuracy overall. Specifically, we rank the 81 cells based on the posterior mean predicted number of events within each cell over the holdout period and take the top M cells to flag as regions for intervention. Then we compute the fraction of actual crimes over the city in the holdout period that fall in the flagged regions. Fig. 4.14 shows such fraction against the corresponding percentage of cells flagged. The NonparSTH performs slightly better than the ParSTH model with larger area under the curve. Notice that the top 1/4 cells captures more than 75% of actual crimes in the holdout period, which provides evidence for accurate prediction under both models.

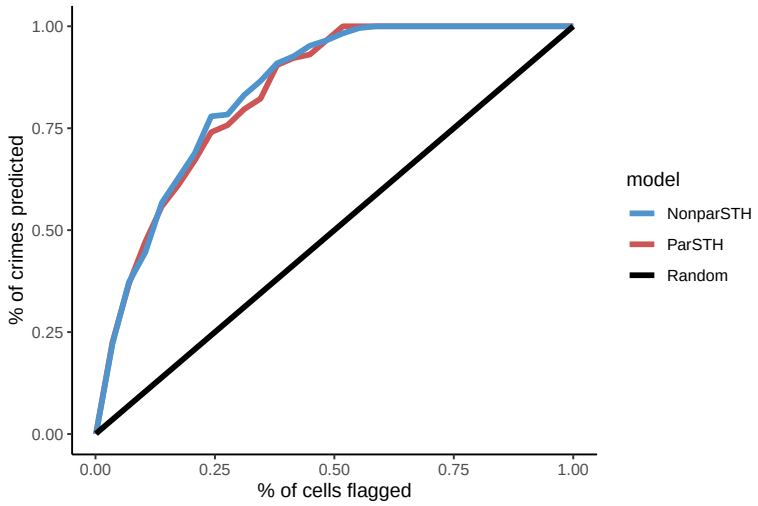


Figure 4.14: Percentage of crimes predicted over holdout period against percentage of cells flagged for intervention

4.8 Conclusion

We propose a semi-parametric space-time Hawkes process model with augmented branching structure that allows flexible inference and efficient simulation. Our approach for the immigrant intensity has the following comparative advantage compared to existing approaches: no approximation in the likelihood with proper treatment of the normalizing constants for both the immigrant and offspring Poisson processes; simple augmentation scheme compared to a Gaussian Process based approach. It is common to assume the spatial region of observation to be $\mathbb{R}^2$ for both the immigrant process and offspring process, which greatly simplify the intractable normalizing constants in the Poisson process likelihood. Such treatment for the offspring process introduces small often negligible bias, especially when the spatial triggering effect decays within short range. However, for the immigrant process this simplification often introduces

bias that requires boundary correction. Without such approximation, the DPMM-based approach for the immigrant process Kolev and Ross (2020) will be computationally intractable when evaluating the random normalizing constant that depends on the Dirichlet Process mixture. Molkenthin et al. (2022) avoids such approximation for the immigrant process by creating an augmented Poisson process realization $\tilde{I}$, so that the union of $\tilde{I}$ and the original immigrant process realization I comes from a homogeneous Poisson process over the finite domain $\mathcal{D}$. The computational cost of such augmentation is high, partially due to the intrinsic cost of the Gaussian process prior, which results in a $\mathcal{O}((N_{I\cup\tilde{I}})^3)$ complexity, where $N_{I\cup\tilde{I}}$ is the number of points in both observed and latent immigrant processes. Our approach has a computation complexity of $\mathcal{O}(K^2 \cdot N_I)$, where K is the number of Bernstein densities no greater than 60 for practical purpose and N_I is the number of observed immigrant points. Our approach therefore can be much more efficient.

Incorporating covariates in the space-time Hawkes process model is an important extension, especially under the context of crime modeling where covariate such as population density is known to impact criminal activity. Incorporating covariates in Hawkes process is less straightforward, since the covariates can impact the immigrant process, the offspring process and the branching structure based on different assumption of the underlying data generating process. For crime modeling, one could argue that the covariates have larger impact on the immigrant intensity since the rate of occurrence for immigrant events are mainly driven by external factors such as population density and demographic information. Reinhart and Greenhouse (2018) formulates the immigrant intensity in a space-time Hawkes process as a piece-wise constant function, where the constant intensity within a region is defined via a regression form

depending on the spatial varying covariates. Based on this approach, we can apply the extension for NHPP model with spatially varying covariates discussed in Section 2.6 for the immigrant Poison process and allow the nonparametric term in the intensity function to explain the spatial heterogeneity not captured by the covariates. Liang et al. (2014) took this approach for the modeling of a space-time Poisson process that incorporates both spatial and temporal covariates into the intensity function, where the baseline intensity is modelled nonparametrically via process convolution. The challenge of this approach, however, is the computational cost due to the lack of closed-form expression for the total intensity.

Here we discuss a particular approach to incorporate spatial covariates such as the population density in the STH process model under the context of crime modeling. Population density often comes in as summary statistics at the county or zip code level. Here we treat population density, denoted by $p(\boldsymbol{s})$, as a discrete variable whose levels $p = 1, \cdots, P$ map one-to-one to a set of regions $\{\mathcal{B}_1, \cdots, \mathcal{B}_P\}$ which forms a partition of $\mathcal{D}$. We formulate the immigrant Poisson intensity such that points within a partition share the same set of random weights $V_{k_x,k_y}(\boldsymbol{s}) \equiv V_{k_x,k_y}^{(p(\boldsymbol{s}))}$, indexed by the population density level within that partition:

$$\mu(x, y, t) = \sum_{k_x=1}^{K} \sum_{k_y=1}^{K} V_{k_x,k_y}^{(p(\boldsymbol{s}))} \phi_{k_x,k_y}^{*}(x, y)$$

This is similar to a spatial fixed effect model that uses the covariates as spatial index and fits a mean to each region. In our formulation we still allow local heterogeneity within the region but implicitly assume the weights within region are on the same level of magnitude which reflects the impact of the population density. One might want to impose some structures in the priors for these P sets of weights to induce smoothness, such that weights in adjacent regions are more

similar to each other.

The nice feature of this formulation is that we can still have tractable normalizing constant for the immigrant Poisson process. The total intensity over space-time is the following

$$\int_{\mathcal{D}}\int_0^T \mu(x,y,t)\mathrm{d}x\mathrm{d}y\mathrm{d}t = T\cdot\sum_{p=1}^{P}\sum_{k_x=1}^{K}\sum_{k_y=1}^{K} V_{k_x,k_y}^{(p)}\int_{\mathcal{B}_p}\phi^*(x,y)\mathrm{d}x\mathrm{d}y$$

where the integral in the right hand side is easy to compute.

Finally, we discuss how to account for the type of crime in the STH process model. Unlike the approach in Reinhart and Greenhouse (2018) that treats the type of crime as a discrete covariate that informs the immigrant intensity, we prefer to formulate the problem as modeling for a multivariate STH process where the subprocesses are indexed by the type of crime. Let the multivariate STH process be denoted as a collection of subprocesses $\{N_1,\cdots,N_U\}$ where each subprocess N_u, i.e., the point process for crime type u, has conditional intensity $\lambda_u(t,\boldsymbol{s}) = \mu_u(t,\boldsymbol{s}) + \sum_{t_k<t} G_{u_k,u}(t-t_k,\boldsymbol{s}-\boldsymbol{s_k})$. The observed N points can be labeled $(t_k,\boldsymbol{s_k},u_k)$ for $k=1,\cdots,N$, where u_k indicates to which subprocess point k belongs. $G_{u_k,u}(\cdot,\cdot)$ is the triggering function defined by the parent point k that belongs to subprocess N_{u_k}. The triggering matrix $K\in\mathbb{R}^{U\times U}$ such that $K_{u,v}$ is the branching ratio defined according to $G_{u,v}(\cdot,\cdot)$. Yuan et al. (2021) provides a fast and accurate estimation method for such multivariate STH process.

One could consider extending the definition of the triggering function defined in this chapter to the multivariate case by formulating $G_{u_k,u}(t-t_k,\boldsymbol{s}-\boldsymbol{s_k}) = h^t_{u_k,u}(t-t_k)h^{\boldsymbol{s}}_{u_k,u}(\boldsymbol{s}-\boldsymbol{s_k})$ as product of separable time and space intensity functions. Specifically the time intensity function $h^t_{u_k,u}(t-t_k)$ can be factored into a branching ratio $K^t_{u_k,u}$ and a temporal density $f_t(t-t_k)$, where $K^t_{u_k,u}$ is the (u_k,u) element of the temporal triggering matrix K^t. The

spatial intensity can be modeled via the spatial distance intensity $h^r_{u_k,u}(r) = V_{u_k,u,l}\phi_l(r)$ as weighted combination of scaled Bernstein densities as discussed in this chapter. The weights $V_{u_k,u,l}$ are indexed by the subprocesses to which the parent and offspring points belongs. The (u, v) element in the spatial triggering matrix $K^{\boldsymbol{s}}$ is the branching ratio between subprocess N_u and N_v, and is expressed as $\sum_{l=1}^{L} V_{u,v,l}$. For the multivariate STH process to be stationary, $K \in \mathbb{R}^{U\times U}$ needs to satisfy $||\ K\ ||< 1$, where $||\ K\ ||$ is the spectral norm of K. In this formulation, $K = K^t \otimes K^{\boldsymbol{s}}$, where $\otimes$ denotes element-wise product. One can center each element in $K^{\boldsymbol{s}}$ around 1 and impose constraints on K^t such that $||\ K^t\ ||< 1$.

Chapter 5

Conclusion

We conclude this dissertation with discussions of the unifying theme of the three projects and some possible extension from both methodology and application perspectives. In this dissertation we carefully handled the issue of spatial irregular domain and its implication for both inference and computation. In the intensity model for NHPP, we obtain a closed-form expression for total intensity over $\mathcal{D}$, $\Lambda_{\mathcal{D}}$, the evaluation of which relies on accurate numerical approximation of the normalizing constants B_{k_x,k_y}. Such computation is more numerically stable and therefore introduce less bias than the numerical integration required under a model such as the log Gaussian Cox process where there are two folds of approximation. It first requires a discretization of the underlying Gaussian process over an finite grid, and then Monte Carlo integration which evaluating the intensity function at the centroids of the grid cells. More importantly, the computational complexity of our proposed model is independent from the complexity of the irregular domain boundary, since the irregular domain $\mathcal{D}$ only enters the computation via the evaluation of the B_{k_x,k_y}s. Inference-wise, our construction leads to a proper NHPP density

over the irregular domain, which does not require bias correction to account for boundary effect when using a maximum likelihood estimation approach (Reinhart, 2018). Such theme extends to the modeling for the SH and STH processes, where the model needs to respect $\mathcal{D}$ in both the immigrant and offspring processes.

For a SH process model, the parametric offspring kernel has a support over $\mathcal{D}$ and requires truncation when the kernel is bivariate Gaussian. Such truncation is necessary as earlier exploration suggests unavoidable bias in the kernel parameter estimation when fitting a untruncated kernel to the observed point pattern. The implication of such truncation is the evaluation of the normalizing constants, which is sped up by the efficient Monte Carlo routines we designed. For prediction, the truncated offspring kernel leads to rejection sampling for the offspring locations which have support over $\mathcal{D}$ instead of $\mathbb{R}^2$. The computation bottleneck for both routine is an algorithm that detects whether a location is inside of the irregular domain $\mathcal{D}$ and is implemented using a Julia package `Luxor`. As a result, the computational complexity depends on the irregular domain boundary, which imposes prohibitive computational cost when using a boundary in its raw form consisted of thousands of boundary points. The Boston city boundary in chapter 3 is simplified using `sf` package in `R` to reduce the number of boundary points while preserving its geometry. We suspect that the computational efficiency of the posterior simulation algorithm will be improved if more efficient implementation of the "point-inside-boundary" algorithm is used.

Such computational dependency on the boundary complexity motivates the development of the nonparametric distance-based offspring kernel in Chapter 4. Here the truncation is handled with R_j as the realized range for a cluster of offspring points centered on $\boldsymbol{s_j}$. The irreg-

ular domain $\mathcal{D}$ therefore enters the computation via both the B_{k_x,k_y}s for the immigrant process and the $B_{l,j}$s for the offspring process. Both sets of constants can be computed efficiently with high accuracy. We acknowledge the approximation in the domain shape introduced by modeling a univariate density instead of a spatial kernel, but want to emphasize the computation gain from such approximation. In Chapter 4, we defined the constant R_{max} and R_j based on a simplified boundary scenario with polygon approximation of $\mathcal{D}$ by taking its convex hull, where the specification of R_j and R_max is straightforward. More complex boundary can be used at the cost of more computational costly algorithms to define these constants.

We developed a Bayesian modeling framework for spatial point process by leveraging a flexible representation of the intensity function for the NHPP as a building block and building hierarchical models for the more general SH and STH processes, while accounting for the irregular domain with extra care in both model formulation and computation. Such framework is designed to be applicable to a variety of problems where point pattern data tends to be non-homogeneous and clustered. All models inference, prediction and forecast are implemented in `Julia` and will be made available as an open-source software maintained by me in the future.

Bibliography

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 Annual International Conference on Machine Learning*, New York, NY, USA: ACM, ICML '09, pp. 9–16.

Adelfio, G. and Chiodi, M. (2015), "Alternated estimation in semi-parametric space-time branching-type point processes with application to seismic catalogs," *Stochastic environmental research and risk assessment: research journal*, 29, 443–450.

Bacry, E., Mastromatteo, I., and Muzy, J.-F. (2015), "Hawkes Processes in Finance," *Market Microstructure and Liquidity*, 01, 1550005.

Baddeley, A. and Turner, R. (2005), "spatstat: An R Package for Analyzing Spatial Point Patterns," *Journal of Statistical Software, Articles*, 12, 1–42.

Baddeley, A., Turner, R., Moller, J., and Hazelton, M. (2005), "Residual analysis for spatial point processes (with discussion)," *Journal of the Royal Statistical Society, Series B*, 67, 617–666.

Baddeley, A. J., Møller, J., and Waagepetersen, R. (2000), "Non- and semi-parametric estimation of interaction in inhomogeneous point patterns," *Statistica Neerlandica*, 54, 329–350.

Bahraoui, Z., Bolancé, C., Pelican, E., and Vernic, R. (2015), "On the bivariate Sarmanov distribution and copula. An application on insurance data using truncated marginal distributions," `https://www.idescat.cat/sort/sort392/39.2.3.bahraoui-etal.pdf`.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2003), *Hierarchical Modeling and Analysis for Spatial Data*, Chapman and Hall/CRC.

Barrientos, A. F., Jara, A., and Quintana, F. A. (2015), "Bayesian density estimation for compositional data using random Bernstein polynomials," *Journal of Statistical Planning and Inference*, 166, 116–125.

— (2017), "Fully Nonparametric Regression for Bounded Data Using Dependent Bernstein Polynomials," *Journal of the American Statistical Association*, 112, 806–825.

BostonGIS (2018), "City of Boston Boundary [Shapefile] "Boston Boundary"," https://www.arcgis.com/home/item.html?id=734463787ac44a648fe9119af4e98cae, 2019-9-19.

Bray, A., Wong, K., Barr, C. D., and Schoenberg, F. P. (2014), "Voronoi residual analysis of spatial point process models with applications to California earthquake forecasts," *The annals of applied statistics*, 8, 2247–2267.

Brix, A. (1999), "Generalized Gamma Measures and Shot-Noise Cox Processes," *Advances in applied probability*, 31, 929–953.

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 Moller, J. (2001), "Space-time Multi Type Log Gaussian Cox Processes with a View to Modelling Weeds," *Scandinavian Journal of Statistics*, 28, 471–488.

Carlin, B. P. and Chib, S. (1995), "Bayesian Model Choice via Markov Chain Monte Carlo Methods," *Journal of the Royal Statistical Society, Series B*, 57, 473–484.

Chiodi, M. and Adelfio, G. (2011), "Forward likelihood-based predictive approach for space-time point processes," *Environmetrics*, 22, 749–757.

Cressie, N. A. C. (1993), *Statistics for spatial data*, New York: Wiley.

Daley, D. J. and Vere-Jones, D. (2003), *An Introduction to the Theory of Point Processes Volume 1: Elementary Theory and Method*, vol. 1, Springer, New York, NY.

— (2008), *An Introduction to the Theory of Point Processes: Volume II: General Theory and Structure*, Springer, New York, NY.

Diggle, P. (2003), *Statistical Analysis of Spatial Point Patterns*, Arnold.

Diggle, P. J. (2006), "Spatio-temporal point processes, partial likelihood, foot and mouth disease," *Statistical methods in medical research*, 15, 325–336.

— (2017), *Statistical analysis of spatial and spatio-temporal point patterns*, Chapman and Hall/CRC.

Dykstra, R. L. and Laud, P. (1981), "A Bayesian Nonparametric Approach to Reliability," *Annals of statistics*, 9, 356–367.

Escobar, M. D. and West, M. (1995), "Bayesian Density Estimation and Inference Using Mixtures," *Journal of the American Statistical Association*, 90, 577–588.

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

Ferraccioli, F., Arnone, E., Finos, L., Ramsay, J. O., and Sangalli, L. M. (2021), "Nonparametric density estimation over complicated domains," *Journal of the Royal Statistical Society, Series B*, 83, 346–368.

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

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. (2001), "Convergence rates for density estimation with Bernstein polynomials," *Annals of statistics*, 29, 1264–1280.

Godsill, S. J. (2001), "On the Relationship Between Markov chain Monte Carlo Methods for Model Uncertainty," *Journal of computational and graphical statistics: a joint publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America*, 10, 230–248.

Green, P. J. and Hastie, D. I. (2009), "Reversible jump MCMC," *Genetics*.

Hawkes, A. G. (1971), "Spectra of some self-exciting and mutually exciting point processes," *Biometrika*, 58, 83–90.

Hawkes, A. G. and Oakes, D. (1974), "A Cluster Process Representation of a Self-Exciting Process," *Journal of applied probability*, 11, 493–503.

Holbrook, A. J., Loeffler, C. E., Flaxman, S. R., and Suchard, M. A. (2021), "Scalable Bayesian inference for self-excitatory stochastic processes applied to big American gunfire data," *Statistics and computing*, 31, 4.

Huber, S. and Held, M. (2012), "A FAST STRAIGHT-SKELETON ALGORITHM BASED ON GENERALIZED MOTORCYCLE GRAPHS," *International journal of computational geometry & applications*, 22, 471–498.

Illian, J., Penttinen, A., Stoyan, H., and Stoyan, D. (2008), *Statistical Analysis and Modelling of Spatial Point Patterns*, John Wiley & Sons.

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," *Journal of the American Statistical Association*, 99, 175–190.

Jain, A. (2018), "Crimes in Boston," https://www.kaggle.com/ankkur13/boston-crime-data, 2019-7-10.

Jones, M. C. (2002), "On Khintchine's Theorem and its Place in Random Variate Generation," *The American statistician*, 56, 304–307.

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.

Kingman, J. F. C. (1977), "Remarks on the spatial distribution of a reproducing population," *Journal of applied probability*, 14, 577–583.

— (1992), *Poisson Processes*, Clarendon Press.

Kolev, A. A. and Ross, G. J. (2020), "Semiparametric Bayesian Forecasting of Spatial Earthquake Occurrences," .

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 nonparametric 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 in-

tensities, with applications to extreme value analysis," *Journal of statistical planning and inference*, 137, 3151–3163.

Kuo, L. and Ghosh, S. K. (1997), "Bayesian nonparametric inference for nonhomogeneous Poisson processes," Tech. rep., University of Connecticut, Department of Statistics.

Leininger, T. J. and Gelfand, A. E. (2017a), "Bayesian Inference and Model Assessment for Spatial Point Patterns Using Posterior Predictive Samples," *Bayesian Analysis*, 12, 1–30.

— (2017b), "Bayesian Inference and Model Assessment for Spatial Point Patterns Using Posterior Predictive Samples," *Bayesian analysis*, 12, 1–30.

Levasseur, K. M. (1984), "A probabilistic proof of the Weierstrass approximation theorem," *The American Mathematical Monthly*, 91, 249–250.

Lewis, P. A. W. and Shedler, G. S. (1979), "Simulation of nonhomogeneous Poisson processes by thinning," *Naval Research Logistics*, 26, 403–413.

Liang, W. W. J., Colvin, J. B., Sansó, B., and Lee, H. K. H. (2014), "Modeling and Anomalous Cluster Detection for Point Processes Using Process Convolutions," *Journal of computational and graphical statistics: a joint publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America*, 23, 129–150.

Linderman, S. and Adams, R. (2014), "Discovering Latent Network Structure in Point Process Data," in *Proceedings of the 31st International Conference on Machine Learning*, eds. Xing, E. P. and Jebara, T., Bejing, China: PMLR, vol. 32 of *Proceedings of Machine Learning Research*, pp. 1413–1421.

Linderman, S. W. and Adams, R. P. (2015), "Scalable Bayesian Inference for Excitatory Point Process Networks," .

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

— (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 of Statistical Mathematics*, 41, 227–245.

Mariathasan, S., Turley, S. J., Nickles, D., Castiglioni, A., Yuen, K., Wang, Y., Kadel, III, E. E., Koeppen, H., Astarita, J. L., Cubas, R., Jhunjhunwala, S., Banchereau, R., Yang, Y., Guan, Y., Chalouni, C., Ziai, J., Şenbabaoğlu, Y., Santoro, S., Sheinson, D., Hung, J., Giltnane, J. M., Pierce, A. A., Mesh, K., Lianoglou, S., Riegler, J., Carano, R. A. D., Eriksson, P., Höglund, M., Somarriba, L., Halligan, D. L., van der Heijden, M. S., Loriot, Y., Rosenberg, J. E., Fong, L., Mellman, I., Chen, D. S., Green, M., Derleth, C., Fine, G. D., Hegde, P. S., Bourgon, R., and Powles, T. (2018), "TGFβ attenuates tumour response to PD-L1 blockade by contributing to exclusion of T cells," *Nature*, 554, 544–548.

Marsan, D. and Lengliné, O. (2008), "Extending earthquakes' reach through cascading," *Science*, 319, 1076–1079.

Mohler, G. (2014), "Marked point process hotspot maps for homicide and gun crime prediction in Chicago," *International journal of forecasting*, 30, 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, 100–108.

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

Møller, J. (2003), "Shot Noise Cox Processes," *Advances in applied probability*, 35, 614–640.

— (2005), "Properties of Spatial Cox Process Models," *Journal of Statistical Research of Iran*, 2, 89–106.

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

Møller, J. and Torrisi, G. L. (2005), "Generalised shot noise Cox processes," *Advances in applied probability*, 37, 48–74.

— (2007), "The pair correlation function of spatial Hawkes processes," *Statistics & probability letters*, 77, 995–1003.

Moller, J. and Waagepetersen, R. P. (2003), *Statistical Inference and Simulation for Spatial Point Processes*, Chapman and Hall/CRC.

Müller, P. and Mitra, R. (2013), "Bayesian Nonparametric Inference – Why and How," *Bayesian analysis*, 8, 269–302.

Murray, I., Ghahramani, Z., and MacKay, D. J. C. (2006), "MCMC for doubly-intractable distributions," in *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, Arlington, Virginia, USA: AUAI Press, UAI'06, pp. 359–366.

Neal, R. (2000), "Markov Chain Sampling Methods for Dirichlet Process Mixture Models," *Journal of Computational and Graphical Statistics*, 9, 249–265.

Nelsen, R. B. (2006), *An Introduction to Copulas*, Springer Series in Statistics, Springer-Verlag New York, 2nd ed.

Neyman, J. and Scott, E. L. (1958), "Statistical Approach to Problems of Cosmology," *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 20, 1–29.

Ogata, Y. (1988), "Statistical Models for Earthquake Occurrences and Residual Analysis for Point Processes," *Journal of the American Statistical Association*, 83, 9–27.

— (1998), "Space-Time Point-Process Models for Earthquake Occurrences," *Annals of the Institute of Statistical Mathematics*, 50, 379–402.

Ogata, Y. and Katsura, K. (1988), "Likelihood analysis of spatial inhomogeneity for marked point patterns," *Annals of the Institute of Statistical Mathematics*, 40, 29–39.

Olkin, I. and Trikalinos, T. A. (2015), "Constructions for a bivariate beta distribution," *Statistics & probability letters*, 96, 54–60.

Paez, M. S. and Walker, S. G. (2018), "Modeling with a large class of unimodal multivariate distributions," *Journal of applied statistics*, 45, 1823–1845.

Pelican, E. and Vernic, R. (2013), "Maximum-likelihood estimation for the multivariate Sarmanov distribution: simulation study," *International journal of computer mathematics*, 90, 1958–1970.

Peng, R. D., Schoenberg, F. P., and Woods, J. A. (2005), "A Space–Time Conditional Intensity Model for Evaluating a Wildfire Hazard Index," *Journal of the American Statistical Association*, 100, 26–35.

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

— (1999b), "Random Bernstein Polynomials," *Scandinavian Journal of Statistics*, 26, 373–393.

Petrone, S. and Wasserman, L. (2002), "Consistency of Bernstein Polynomial Posteriors," *Journal of the Royal Statistical Society, Series B*, 64, 79–100.

Rasmussen, J. G. (2013), "Bayesian Inference for Hawkes Processes," *Methodology and computing in applied probability*, 15, 623–642.

Reinhart, A. (2018), "A Review of Self-Exciting Spatio-Temporal Point Processes and Their Applications," *Statistical science: a review journal of the Institute of Mathematical Statistics*, 33, 299–318.

Reinhart, A. and Greenhouse, J. (2018), "Self-exciting point processes with spatial covariates: modelling the dynamics of crime," *Journal of the Royal Statistical Society. Series C, Applied statistics*, 67, 1305–1329.

Ripley, B. D. (1976), "The Second-Order Analysis of Stationary Point Processes," *Journal of applied probability*, 13, 255–266.

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. (2016), "Bayesian estimation of the ETAS model for earthquake occurrences," *Preprint*.

Sethuraman, J. (1994), "A constructive definition of Dirichlet priors," *Statistica Sinica*, 4, 639–650.

Shepp, L. A. (1962), "Symmetric Random Walk," *Transactions of the American Mathematical Society*, 104, 144–153.

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.

Sudyanti, P. A. and Rao, V. (2018), "Flexible Mixture Modeling on Constrained Spaces," .

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.

Tenbusch, A. (1994), "Two-dimensional Bernstein polynomial density estimators," *Metrika*, 41, 233–253.

Teucher, A., Russell, K., and Bloch, M. (2021), *rmapshaper: Client for 'mapshaper' for 'Geospatial' Operations*, https://cran.r-project.org/package=rmapshaper.

Ting Lee, M.-L. (1996), "Properties and applications of the sarmanov family of bivariate distributions," *Communications in Statistics - Theory and Methods*, 25, 1207–1222.

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.

Vitale, R. A. (1975), "A Bernstein Polynomial Approach to Density Function Estimation," in *Statistical Inference and Related Topics*, ed. Puri, M. L., Academic Press, pp. 87–99.

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.

Yuan, B., Li, H., Bertozzi, A. L., Jeffrey Brantingham, P., and Porter, M. A. (2018), "Multivariate Spatiotemporal Hawkes Processes and Network Reconstruction," .

Yuan, B., Schoenberg, F. P., and Bertozzi, A. L. (2021), "Fast estimation of multivariate spatiotemporal Hawkes processes and network reconstruction," *Annals of the Institute of Statistical Mathematics*, 73, 1127–1152.

Zheng, Y., Zhu, J., and Roy, A. (2010), "Nonparametric Bayesian inference for the spectral density function of a random field," *Biometrika*, 97, 238–245.

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.