Spatio-Temporal Hawkes Point Processes: A Review

Alba BERNABEU, Jiancang ZHUANG, and Jorge MATEU

Hawkes processes are a particularly interesting class of stochastic point processes that were introduced in the early seventies by Alan Hawkes, notably to model the occurrence of seismic events. They are also called self-exciting point processes, in which the occurrence of an event increases the probability of occurrence of another event. The Hawkes process is characterized by a stochastic intensity, which represents the conditional probability density of the occurrence of an event in the immediate future, given the observations in the past. In this paper, we present some background and all major aspects of Hawkes processes, with a particular focus on simulation methods, and estimation techniques widely used in practical modeling aspects. We aim to provide a rich and self-contained overview of these stochastic processes as a way to have an overall vision of Hawkes processes in only one piece of paper. We also discuss possibilities for future research in the area of self-exciting processes.

Key Words: Estimation procedures; Hawkes point processes; Self-exciting processes; Simulation techniques; Spatio-temporal processes.

1. INTRODUCTION

Stochastic point processes frame a versatile approach used to understand the behavior of a discrete number of georeferenced events occurring in a continuous space, time, or a combination of both. Real-data applications span various fields, addressing phenomena such as urban fires, wild forest fires, crimes, earthquakes, diseases, tree locations, animal locations, or communication network failures, to name just a few number of application fields.

In point process modeling framework, three main categories are customized. Spatial point processes focus on events within a defined spatial domain, such as the distribution of trees in a forest, the occurrence of crimes in urban areas, or the locations of earthquakes in a

A. Bernabeu (✉) · J. Mateu, Department of Mathematics, University Jaume I, Castellón, Spain (E-mail: *bernabeu@uji.es; mateu@uji.es*).
J. Zhuang, The Institute of Statistical Mathematics, Research Organisation of Information and Systems, Tokyo, Japan (E-mail: *zhuangjc@ism.ac.jp*). Department of Statistical Science, The Graduate University for Advanced Studies, Tokyo, Japan. London Mathematical Laboratory, London, UK.

Journal of Agricultural, Biological, and Environmental Statistics
https://doi.org/10.1007/s13253-024-00653-7

Published online: 19 September 2024

geographical region (see Illian et al. 2007; Baddeley et al. 2015). Temporal point processes, on the other hand, are concerned with events unfolding over time, without explicitly considering spatial dimensions. Examples include modeling the occurrence of communication network failures or the spread of infectious diseases over time (see Yan et al. 2019; Shchur et al. 2020; Shchur et al. 2021). Spatio-temporal point processes integrate both spatial and temporal dimensions, allowing the analysis of events evolving over both space and time. These models have a dedicated evolutionary time axis, enabling the chronological ordering of events. Examples include studying spatio-temporal patterns of seismic activities, tracking wildfires over both space and time, or understanding the spread of diseases in a geographical area over a specific time frame (see Diggle 2006a, 2013a; González et al. 2016). In brief, spatial point processes capture spatial patterns, temporal point processes focus on temporal dynamics, and spatio-temporal point processes combine both aspects to analyze events evolving in both space and time.

Furthermore, when events are associated with specific properties or features, such as earthquake magnitudes or the extent of wildfire damage, the underlying point process becomes a *marked* point process (Jacobsen 2006; González et al. 2016). This adaptation allows for a more nuanced representation of the data, considering additional attributes associated with each event. This may involve exploring relationships between event magnitudes or sizes, contributing to a more comprehensive understanding of the underlying processes.

In the context of point processes, we also encounter an interesting structure referred to as a *self-exciting process*. These models are of great interest and have already been addressed in the relevant paper of Reinhart (2018). Different from this particular reference, which focuses in the fundamental/conceptual techniques related to the Hawkes process, we have delved deeper and expanded the scope by providing more simulation and estimation methods that are currently of interest and are at the forefront of contemporary research. This type of process, also known as a *Hawkes process*, refers to a stochastic model that describes the occurrence of events over time conditioned on the (past) history. In this case, the likelihood of an event occurring is not solely influenced by external factors, but is also significantly affected by the occurrence of similar events in the past. This characteristic highlights the temporal dependence inherent in self-exciting processes, where the presence of one event can influence and increase the probability of subsequent events; this property renders these processes non-Markovian. Indeed, the Hawkes model, introduced by Alan Hawkes in 1971 (Hawkes and Chen 2021), is commonly associated with the mathematical description of self-exciting processes.

In terms of applications, Hawkes processes find utility in various areas. In data science, they are applied in social network analysis, where an individual interaction can trigger additional interactions (Zipkin et al. 2015). In finance, they are used to model the occurrence of successive financial transactions (Bacry and Muzy 2014). Additionally, in seismology, the Hawkes model has been employed to analyze and model the occurrence of earthquakes (Ogata 1998; Zhuang et al. 2002; Schorlemmer et al. 2018; Kwon et al. 2023). In biology, self-exciting processes are also employed to model the spread of epidemics, where one case may influence the occurrence of future events (Diggle et al. 2010a; Schoenberg 2023). In addition, there is recent literature that extends self-exciting Hawkes processes to more flexible settings, such as that of Cai et al. (2024), which proposes a new approach to learning the latent network structure from multivariate point process data. These authors employ non-stationary Hawkes processes and efficient sparse least squares estimation, demonstrating its effectiveness through simulation studies and neural applications. Another very recent example is the work of Fang et al. (2023), which investigates events among individuals in a network, proposing a group network Hawkes process (GNHP) model with observed and fixed network structure. They introduce a latent group structure to characterize specific user characteristics, and propose a maximum likelihood approach to estimate model parameters and cluster individuals. Finally, we mention the paper by Schatz et al. (2022), which introduces an ARMA point process model combining self-exciting and shot-noise clustering mechanisms, providing flexibility to analyze continuously observed count data.

A Hawkes model includes two components: one is the baseline rate, representing the rate of event occurrence in the absence of previous events, and the other is the excitation component, quantifying how past events affect the current occurrence rate. When extending to spatio-temporal considerations, these models incorporate both spatial and temporal dimensions, providing a comprehensive understanding of how events evolve over both space and time. However, the precise estimation of these parameters can be challenging due to the dependent nature of the process history. Inference in Hawkes processes involves estimating model parameters based on observed data. Two types of approaches can be employed: parametric and non-parametric. The parametric approach assumes a specific structure, such as the functional form of the excitation function and event arrival rates, using techniques such as maximum likelihood or Bayesian methods. In contrast, the non-parametric approach is more flexible and does not impose rigid restrictions on the underlying functional form of the model. Non-parametric methods, such as kernel density estimators and local smoothing, are used to capture temporal dependence without making specific assumptions. The choice between approaches depends on the nature of the process and data availability, aiming for a balance between model fit and interpretative simplicity.

Thus, Hawkes processes are arguable one of the most versatile type of stochastic point process models the literature offers, and we aim to provide an authoritative review in form of state-of-the-art in this field, while focussing on simulation and estimation methods, rather than on deep mathematical developments. The plan of the paper is the following. Section 2 provides a complete theoretical setup of Hawkes processes in time, and then in space-time. We also describe maked spatio-temporal point processes, and the Epidemic-type aftershock sequence (ETAS) model. Section 3 focuses on simulation methods, describing the two main techniques. Section 4 describes a number of estimation methods, going from likelihood and partial likelihood approaches to Expectation-Maximization (EM) techniques and Bayesian inference. Modern machine learning, non-parametric inference and model assessment and diagnostics are also detailed. Finally, Sect. 5 depicts some further challenging issues.

2. HAWKES POINT PROCESSES

2.1. TEMPORAL DOMAIN

A *counting/point process* $\{N(t),\ t \geq 0\}$ is a stochastic process taking values in the set of non-negative integers $\mathbb{N} \cup 0$. It adheres to the following conditions: *(i)* $N(0) = 0$; *(ii)* $N(t)$

is a right-continuous step function with unit increments; *(iii)* $N(T) < \infty$ almost surely if $T < \infty$; *(iv)* $N(t^+) - N(t^-) \leq 1$. Additionally, for a time interval $[0, T)$ where $T < \infty$, the *complete set of observations up to time* T is defined as $\mathscr{H}_T = \{t_i : t_i \in [0, T), \forall i = 1, \ldots, N(T^-)\}$. Here, t_i represents the points in time occurring prior to time T. For a random $t \in [0, T)$, the *history of the process up to time* t is the subset of elements in $\mathscr{H}_T$ recorded strictly before t, denoted as $\mathscr{H}_t = \{t_i \in T : t_i < t\}$.

A univariate *Hawkes process* can be defined as a self-exciting temporal counting/point process N, that can be completely specified by its conditional intensity. Indeed, the self-exciting nature of the data is modeled through the *conditional intensity function* which governs the expected arrival rate of events, assuming simplicity (i.e., no events occur simultaneously). This conditional intensity function $\lambda(t|\mathscr{H}_t)$ is defined as

$$\lambda(t|\mathscr{H}_t) = E[N(dt)|\mathscr{H}_t] = \mu(t) + \sum_{i:t_i<t} \nu(t - t_i), \tag{1}$$

where $\mu(t)$ is the *background rate* of the process N and ν is a function that governs the clustering density of N. The function ν is largely called the *triggering function* (also known as *kernel function* or *exciting function*). Additionally, we can write $\sum_{i:t_i<t}$ as $\sum_{t_i \in \mathscr{H}_t}$. Similarly, some authors denote the conditional intensity function by λ_t and rewrite the sum in (1) as

$$\sum_{i:t_i<t} \nu(t - t_i) = \int_{-\infty}^{t} \nu(t - u)N(du). \tag{2}$$

The validity of the Hawkes model involves critical conditions related to the intensity function $\lambda(t|\mathscr{H}_t)$ and the triggering function $\nu(t)$. It is essential that the intensity is non-negative, $\lambda(t|\mathscr{H}_t) \geq 0$, reflecting the instantaneous rate of events. The triggering function $\nu(t)$ must also be non-negative for $t \geq 0$ and zero for $t < 0$ to maintain causality. Additionally, $\rho = \int_0^{+\infty} \nu(u)du < 1$, where ρ is called the branching ratio. When $\rho < 1$ the process is stationary. Otherwise, the population rate grows to infinity. These conditions are crucial to ensure the coherence and proper interpretation of the Hawkes model in terms of point processes over time.

2.2. Spatio-Temporal Domain

A *spatio-temporal Hawkes process* is a probabilistic model used to analyze and describe the clustering of events in both space and time. It extends the classical Hawkes process by incorporating spatial information, allowing for the modeling of event occurrences in a multidimensional setting. Spatio-temporal models extend the conditional intensity function to predict the rate of events at locations $s = (x, y) \in W \subseteq \mathbb{R}^2$ and times $t \in [0, T)$, $T > 0$. The conditional intensity function is defined in the analogous way to temporal Hawkes processes

$$\lambda(s, t|\mathscr{H}_t) = \mu(s, t) + \sum_{i:t_i<t} g(s - s_i, t - t_i), \tag{3}$$

where $\mu(s,t)$ is the background (or baseline) rate of the process, $g(s,t)$ is the triggering function, $\{s_1 = (x_1, y_1), s_2 = (x_2, y_2), \ldots\}$ represents the observed sequence of event locations, and $\{t_1, t_2, \ldots\}$ denotes the observed event times.

The conditional intensity function is defined as the instantaneous rate of events occurring per unit space and time. It captures the influence of past events on the occurrence of future events. In other words, the intensity at a specific location and time depends on the history of events that have occurred in the vicinity. In a similar vein to the temporal case, the intensity formula for a spatio-temporal Hawkes process is built based on several critical assumptions, needing the non-negativity of both the triggering and the base rate. The total intensity depends on a finite sum over past events, with event arrival times considered independent except for their relationship with past events through the kernel g. The branching ratio (ρ) must be within the range (0, 1) for the process to remain stationary. When $\rho = 0$, the process becomes a Poisson process with intensity $\mu(s,t)$, while $\rho \geq 1$ indicates explosive behavior. For further details on the explosive phenomenon and implications of varying ρ, Grimmett and Stirzaker (2001), Asmussen (2003) provide insightful references.

Before proceeding with further details, from now on, we will only work with the spatio-temporal case.

2.3. Particular Details of the Conditional Intensity Function

The intensity function defined in (3) is influenced by several functions and their parameters, including the background rate and the triggering function. The background rate, also known as background function, and denoted by $\mu(s,t)$, represents the expected rate of events occurring in the absence of any triggering events. It can be thought of as the background level of event occurrences. Often, for simplicity, it is assumed to be separable in space and time, such that $\mu(s,t) = \mu_s(s)\mu_t(t)$ (Zhuang and Mateu 2019). Simplified approaches use a constant background function, that is $\mu(s,t) = \mu$ (Schoenberg 2023). On the other hand, the triggering function represents the influence of past events on the intensity function. It characterizes how the occurrence of an event at a particular location and time affects the intensity at subsequent locations and times. The form of the triggering kernel determines the temporal decay and spatial spread of the influence. Different choices of triggering kernels can capture various patterns, such as exponential decay or power-law decay of the influence. The kernel function $g(s,t)$ also follows the same principle: it can be separable or non-separable. In the former case, the kernel function can be decomposed into $g(s,t) = f(s)h(t)$, which is similar to covariance functions in other spatio-temporal models (Miscouridou et al. 2023). However, we can also find non-separable structures in the literature (Kwon et al. 2023; Stindl et al. 2024).

As previously mentioned, it is common in applications to employ separable triggerings in time and space. Hence, we here consider typical choices for the triggering function in both spatial and temporal domains. A commonly used temporal triggering function has an exponential form, expressed as $h(t) = \beta e^{-\beta(t-t_i)}$. This function models the temporal influence of previous events, decreasing exponentially as the time since the last event $(t - t_i)$ increases. Another common choice is the power-law function in time, expressed as $h(t) =$

$k(t - t_i + c)^{-p}$. This function represents a slower temporal decay of influence, where k, c, p are parameters affecting the amplitude and decay rate.

Regarding the spatial domain, spatial triggering functions in Hawkes processes provide distinct approaches to model the influence of nearby and distant events. The commonly used Gaussian function, $f(s = (x, y)) = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{(x-x_i)^2}{2\sigma_x^2} - \frac{(y-y_i)^2}{2\sigma_y^2}\right)$, assigns stronger influence to nearby events, decreasing exponentially as the distance between events increases. Another common choice is the exponential function, expressed as $f(s) = \beta e^{-\beta||s-s_i||}$, which models an exponential decay of spatial influence. It is effective when the influence of nearby events is expected to be significant and decreases rapidly with distance. An alternative option is the power-law function $f(s) = (1 + \frac{||s-s_i||}{\gamma})^{-q}$, where $||s - s_i||$ represents the distance between two spatial points s and s_i, typically the Euclidean distance. This function models a decrease in the spatial influence with distance following a power law. It introduces a characteristic scale γ in the denominator, allowing for adjustments in the amplitude and range of the spatial influence. It is particularly useful when there is a wish for significant variation in influence strength based on the distance between events.

These above-mentioned functions are just a few examples of triggering kernels, and other functional forms can be used depending on the specific problem and the characteristics of the event data. As explained by Laub et al. (2021), the selection of the triggering function is pivotal for the reliability and stability of any parameter estimation procedure for the Hawkes process. For instance, many techniques use normalized triggering functions that integrate to 1 over an infinite domain. The choice between temporal functions and spatial triggering functions in Hawkes processes depends on the specific nature of the data and the anticipated behavior of temporal and spatial dependence within the context of the application. Each function, whether temporal or spatial, comes with its own set of properties, enabling it to capture different patterns of dependence. This selection is crucial for tailoring the model to the characteristics of the observed data, ensuring an accurate representation of the underlying temporal and spatial dynamics in Hawkes processes.

2.4. Marked Spatio-Temporal Point Processes

This concept of a spatio-temporal point process can be extended to the marked spatio-temporal case. Here, a generic observed point is represented as $\mathbf{x}= (t, s, m)$, consisting of a time t, a spatial location $s = (x, y)$, and a marking variable m. The domain is given by $[0, T) \times W \times M$, where $T > 0$, $W \subseteq \mathbb{R}^2$, and $M \subseteq \mathbb{R}$. The value of the counting process at time t denotes the number of events recorded before t (inclusive), with spatial locations in W and marking variables in M. Consequently, the notation for the complete set of observations and the history of the process remains consistent. In this case, the complete set of observations is $\mathscr{H}_T = \{\mathbf{x}_i = (t_i, s_i, m_i) : \mathbf{x}_i \in [0, T) \times W \times M \; \forall i = 1, \ldots, N(T^-)\}$, and the history of the process becomes $\mathscr{H}_t = \{\mathbf{x}_i = (t_i, s_i, m_i) \in \mathscr{H}_T : t_i < t\}$.

These additional details apart from the space-time locations, known as *marks*, can encompass a variety of features or attributes linked to each event, contributing to a more thorough comprehension of the fundamental process. For a *marked spatio-temporal point process*, the intensity measure conditional on the process history $\mathscr{H}_t$ is defined as

$$\lambda(s,t,m|\mathscr{H}_t) = \mu(s,t,m) + \int_{[0,T)\times W\times M} g(t-u, s-v) f(m \mid m')\, dN(du \times dv \times dm'), \tag{4}$$

where the function $f(m|m')$ provides the probability density function for the marks of direct offspring from an event of mark m'. The other functions μ and g are defined analogously to the spatio-temporal case. We can consider the multivariate case simply as a marked point process in which the mark takes only a finite number of values. In the above formula, we assume that the marks of triggered events are independent of location and time. A remarkable model in the realm of Hawkes processes is the one known as the ETAS model (Epidemic-Type Aftershock Sequence). The criticality related to mark-dependent triggering is more complicated (see Zhuang et al. 2013).

Epidemic-Type Aftershock Sequence (ETAS) Model

As previously mentioned, the ETAS model is a popular case of a marked spatio-temporal Hawkes process. The ETAS model in the space-time domain has been extensively employed to characterize the clustering characteristics of earthquakes concerning both their spatial and temporal dimensions (Ogata 1998; Zhuang et al. 2002; Zhuang and Ogata 2004; Zhuang et al. 2005; Zhuang and Ogata 2006; Ogata and Zhuang 2006). The conditional intensity of this model is given by

$$\lambda(s,t,m|\mathscr{H}_t) = h(m)[\mu(s,t) + \sum_{i:t_i<t} k(m_i) g(t-t_i) f(s-s_i, m_i)], \tag{5}$$

where $t, s = (x, y)$ and m represent the time of occurrence, spatial location, and magnitude of the earthquake, respectively. In the formula above

$$h(m) = \beta \exp[-\beta(m_c)], \quad m \geq m_c \tag{6}$$

represents the probability density of the earthquake magnitude, where m_c is the magnitude threshold of the earthquake, defined as the minimum value of seismic magnitude that is considered to include an event in the analysis of seismic activity, and

$$k(m) = A \ \exp[\alpha(m - m_c)], \quad m \geq m_c \tag{7}$$

is the expectation of the number of offspring (productivity), which is a Poisson random variable, from an event of magnitude m. Furthermore,

$$g(t) = \frac{p-1}{c}(1 + \frac{1}{t})^{-p}, \quad t > 0 \tag{8}$$

is the probability density function of the length of the time interval between an offspring and its parent, and

$$f(s,m) = \frac{q-1}{\pi \ D \ \exp[\gamma(m - m_c)]}(1 + \frac{||s||^2}{D \ \exp[\gamma(m - m_c)]})^{-q} \tag{9}$$

is the probability density function of the relative locations between the parent and offspring.

The ETAS model has become widely acknowledged as the standard framework for characterizing earthquake clusters, as put in evidence in studies by Huang et al. (2016), Schorlemmer et al. (2018). Beyond its seismic applications, this model has found utility in diverse fields, including crime data analysis, as observed in the work of Mohler et al. (2011), and in economics, where research demonstrates its relevance to understanding epidemic features in the interaction between prices, as shown by Bacry and Muzy (2014). Notably, in recent years, researchers have extended the application of this model to analyze data related to terrorist behavior (Tench et al. 2016), interactions within social networks (Zipkin et al. 2015), and phenomena such as genomes or neuronal activities (Truccolo et al. 2005), among other domains. Across these diverse areas, a substantial portion of the underlying theories and methodologies has incorporated or adapted the ETAS model to elucidate complex patterns and dynamics (Chiodi and Adelfio 2020; Chiodi et al. 2021; Lo Galbo and Chiodi 2023). In other domains, the Hawkes process is frequently employed in various fields primarily to estimate the parameters of the model and to offer explanations for the observed outcomes. The capacity of the model for parameter estimation and interpretation largely extends its applicability, making it a valuable tool in diverse research areas.

3. SIMULATION TECHNIQUES

Simulating a Hawkes process involves modeling the occurrence of events over both space and time, where the likelihood of an event is influenced by past events in both temporal and spatial dimensions. Two common methods for simulating a Hawkes process follow the *parents and offspring* approach and the *acceptance-rejection* method.

3.1. Parents and Offspring Approach

In the spatio-temporal *parents and offspring* approach, the simulation begins with a set of initial events, or “parents”, distributed in both space and time. These parent events are generated through an initial spatio-temporal intensity function. Subsequently, each parent event may give rise to new events, termed “offspring”, in both space and time, following a conditional spatio-temporal intensity function. The recursive nature of the process allows for the modeling of self-excitation across both spatial and temporal dimensions, capturing how events in one location and time may influence events in nearby locations and subsequent time points.

Explicitly, consider a stochastic process X given by a Poisson cluster process that provides locations $\{(s_i, t_i)\}_{i=1}^{n}$, with $s_i = (x_i, y_i) \in \mathbb{R}^2$ and $t_i \in \mathbb{R}$. The cluster centers of X are given by certain events known as *parents*, while the other events are known as *offspring*. In the context of a spatio-temporal Hawkes process, X satisfies the following rules. Initially, parents are generated according to a Poisson process in space-time. The occurrence rates are determined by the intensity function, which characterizes the frequency with which parental events occur at different space-time locations. Secondly, each parent, located at a specific point in space and time, can generate offspring in both temporal and spatial terms, determined by a space-temporal triggering function or kernel. Thus, each parent initiates the formation of a cluster. These clusters comprise events across various generations. The structure of this branching process entails that an event in a generation can trigger a new Poisson process to generate offspring for the next generation, and so on. The stopping criterion or termination condition of the process could be determined by various factors depending on the specific context of the application. Some possible considerations could include number of generations, thresholds for occurrence rates, convergence criteria, or application objectives. The rate of offspring generation is determined by the intensity function of the process. Thirdly, given the parents, the clusters are independent. Lastly, the complete process is comprised of the combination of all these clusters, including the parents. Examples of this type of simulation method can be found in Diggle et al. (2010b), Mohler et al. (2021). This method is described in Algorithm 1.

Algorithm 1 Spatio-temporal simulation with parents and offspring approach

```
Initialization:
Initialize an empty list of events
Set parameters such as maximum number of iterations and convergence criterion
Generation of parents:
Number of parents $n \sim Po(n_0)$, where $n_0$ is the average rate of the number of parents generated,
  as defined by the user
Assign spatial and temporal locations to these parent events according to a probability distribution
  set by the user
Add the parent events to the list
Generation of offspring:
For each parent:
Number of offspring $n_p \sim Po(k)$, where $k$ is the reproduction number
Assign spatial and temporal locations to the offspring based on the triggering function
Add the offspring to the list
Branching Process:
Initialize a list of events containing all parent and offspring events
Define a stopping criterion for the branching process (e.g., maximum number of events, occurrence
  rate threshold, convergence criterion)
while the stopping criterion is not met and the maximum number of iterations is not reached do
    for each event in the list do
        Determine if the event can trigger the generation of offspring for the next generation
        If so, follow the offspring generation process
        Add the new events to the list
    end for
end while
```

3.2. ACCEPTANCE-REJECTION METHOD

The *acceptance-rejection* method is a commonly used technique for simulating random variables from a given distribution. In the context of simulating a Hawkes process, which is a stochastic self-exciting process, we can adapt this method to generate points according to the intensity of the Hawkes process. A brief description of how this method works for a Hawkes process is the following. We begin by generating points from a homogeneous

Poisson process with a constant rate. Then, we calculate the intensity function of the Hawkes process based on the previously generated points and their times, that is, the conditional space-time intensity function the point pattern should follow. Next, we define an upper bound for this intensity function, which will allow to determine the acceptance probability in the acceptance-rejection method. This upper bound can be calculated by computing the intensity over a grid and using the maximum value obtained. Use the acceptance-rejection method to accept/reject these points according to the relationship between the intensity function of the Hawkes process and the upper bound, also employing a uniform distribution for generating the necessary random numbers in the process. Finally, apply an additional condition (based on a quotient or another conditional function) to filter the generated points and obtain an inhomogeneous Hawkes process that satisfies your specific criteria. Algorithm 2 outlines the steps to follow for simulating spatio-temporal data using the acceptance-rejection method.

Algorithm 2 Spatio-temporal simulation with acceptance-rejection method

```
Generation of homogeneous Poisson point pattern:
Input: Spatial domain $S$, temporal domain $T$, intensity $\lambda_{max}$
Simulate the number of events $n = N(S) \sim Po(\lambda_{max} \times |W \times T|)$
for $i$ in 1 to $n$ do
    Sample a location $(s_i, t_i)$ uniformly from $W \times T$
end for

Thinning of a homogeneous Poisson process:
Input: Intensity function $\lambda(s, t)$
Define an upper bound $\lambda_{max}$ for the intensity function
Simulate a homogeneous Poisson process with intensity $\lambda_{max}$
Thin the process as follows:
for each point $(s, t)$ in the homogeneous Poisson process do
    Calculate $p = \frac{\lambda(s,t)}{\lambda_{max}}$
    Generate a sample $u \sim U(0, 1)$
    if $u \leq p$ then
        Retain the location $(s, t)$
    end if
end for
```

Therefore, in the spatio-temporal variant of the acceptance-rejection approach, the generation of events involves proposing potential spatio-temporal event times and deciding on their acceptance or rejection based on specified criteria, including the conditional spatio-temporal intensity function. Candidate event times are suggested in both spatial and temporal dimensions, and their combined probability of acceptance is compared to a randomly determined acceptance threshold. Accepted spatio-temporal event times contribute to the evolving sequence, and the process continues until a satisfactory spatio-temporal event time is confirmed. This methodology provides flexibility in capturing the intricate interactions and dependencies across both spatial and temporal dimensions within a Hawkes process. Examples of this simulation method can be read in Zhu et al. (2019), Zhu (2019).

3.3. SOME SIMULATED CASES

We consider the following intensity function with a separable kernel, featuring an exponential triggering for time in $T = (0, 100)$, a Gaussian triggering for space in the unit square, and a constant background

$$\lambda(s, t|\mathscr{H}_t) = \mu + \sum_{i:t_i<t} k \cdot w \cdot \exp\{-w(t - t_i)\} \frac{1}{2\pi\sigma^2} \cdot \exp\{-\frac{(x - x_i)^2 + (y - y_i)^2}{2\sigma^2}\}, \tag{10}$$

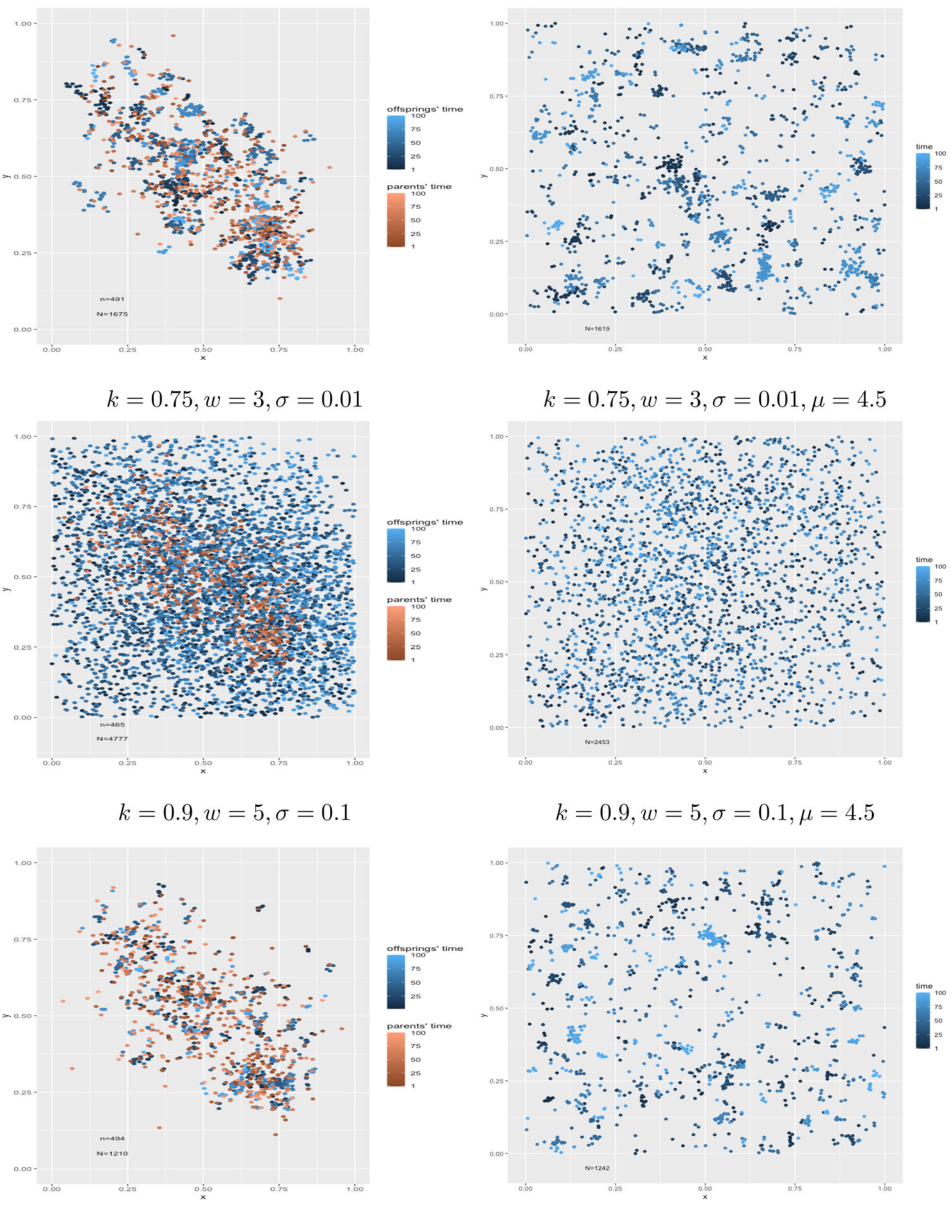
where $s = (x, y)$. In the case of the parents and offspring approach, we fist simulate the number of parents as a Poisson with intensity 500, and simulate the temporal instants as uniform values in T. The parents are spatially distributed following a Gaussian Mixture Model (GMM) with three components, having means $m_1 = [0.3, 0.7], m_2 = [0.5, 0.5], m_3 = [0.7, 0.3]$, diagonal covariance matrices $diag(\Sigma_1) = [0.01, 0.01]$, $diag(\Sigma_2) = [0.025, 0.01]$ and $diag(\Sigma_3) = [0.004, 0.004]$ and mixture probabilities $p_1 = 0.2$, $p_2 = 0.5$ and $p_3 = 0.3$. Offspring events are then iteratively added, where each parent generates a Poisson distributed number of offspring, with spatial and temporal coordinates given by the corresponding triggering effects.

Regarding the acceptance-rejection method, we first simulate a large number of events from a homogeneous Poisson distribution in the considered space-time domain, in our case the unit square and the (0, 100) temporal interval. Then we thin those points with an intensity value less than a uniform number in (0, 1), and retain the others. Those retained follow the aimed Hawkes distribution.

Figure 1 depicts simulated realizations of a Hawkes process using the intensity in (10) with a number of distinct parameters, and using the parents and offspring approach (on the left), and the acceptance-rejection method (on the right). Note that in the parents and offspring method, the parents follow a GMM with three components providing a more clustered structure compared to the acceptance-rejection method for which the parents are randomly scattered in the space-time region. Note that if we were to start off with the uniform distribution of parents, the resulting point patterns from the two approaches would be similar; however, we have chosen to simulate parents from a Gaussian mixture to provide a different approach. In any case, both share the same intensity in (10).

4. ESTIMATION TECHNIQUES

In the realm of Hawkes processes, inference is essential for understanding and modeling the occurrence of events over both time and space. Various statistical methods have been developed to address inference in this context. In this section, we will briefly provide a comprehensive overview of different approaches, primarily focusing on parametric methods such as (full) likelihood, partial likelihood, the Expectation-Maximization (EM) method, Bayesian inference, machine learning techniques, and considerations for non-parametric cases. However, many pieces of research focus on specific cases or approaches due to the



$k = 0.75, w = 3, \sigma = 0.01$

$k = 0.75, w = 3, \sigma = 0.01, \mu = 4.5$

$k = 0.9, w = 5, \sigma = 0.1$

$k = 0.9, w = 5, \sigma = 0.1, \mu = 4.5$

$k = 0.6, w = 1, \sigma = 0.0075$

$k = 0.6, w = 1, \sigma = 0.0075, \mu = 4.5$

Figure 1. Simulated realizations of a Hawkes point process with intensity in (10), and parameters as indicated in each subfigure. Left column shows the cases using a parents and offspring method, while right column shows those obtained with an acceptance-rejection method .

diversity and complexity of the problems they address. The adaptability of a method can largely depend on the specific nature of the problem and the data at hand. More general approaches that aim to provide broad and flexible tools are not always available.

4.1. LIKELIHOOD

The likelihood is typically expressed as the probability of observing a specific sequence of events given the model parameters. The goal is to estimate the parameters that maximize this likelihood, which in turn provides the best fit of the model to the observed data.

We assume the process follows an orderly pattern (Diggle et al. 2010a,b and references therein). Let $\mathscr{H}_t$ denote the history of the process up to time t, and $\lambda(s, t|\mathscr{H}_t)$ the conditional intensity for an event at location s and time t, given $\mathscr{H}_t$. For observed data $\{(s_i, t_i)\}_{i=1}^{n} \subseteq W \times T$, the likelihood takes the form

$$L(\theta) = (\prod_{i=1}^{n} \lambda(s_i, t_i|\mathscr{H}_{t_i})) \exp\left(-\int_T \int_W \lambda(s, t|\mathscr{H}_t)\, ds\, dt\right). \tag{11}$$

However, for simplicity, it is often preferred to use the log-likelihood function instead

$$\log(L(\theta)) = \sum_{i=1}^{n} \log \lambda(s_i, t_i|\mathscr{H}_{t_i}) - \int_T \int_W \lambda(s, t|\mathscr{H}_t)\, ds\, dt. \tag{12}$$

Likelihood-based inference is relatively straightforward for any model in which the conditional intensity is specified. It is essential that $\lambda(s, t|\mathscr{H}_t)$ is non-negative and integrable over W for any potential history at any time $t \in T$. To draw inferences, two additional assumptions must be met (for an example, see Peng et al. 2005). Firstly, if a model is not directly defined through its conditional intensity, we need to have an explicit expression for $\lambda(s, t|\mathscr{H}_t)$, and this might be challenging, possibly even impossible. Secondly, the integrand of the integral term on the right-hand side of equation (12) is often a complex function with numerous local modes, making the accurate evaluation of the integral intricate. This inherent complexity calls for numerical methods to optimize the likelihood. This optimization depends on several factors, including the choice of the specific triggering kernels, the form of the intensity function used, and the background intensity. Alternatively to maximum likelihood estimation (MLE), Bayesian inference methods based on Markov chain Monte Carlo (MCMC) stand as an accurate framework.

4.2. PARTIAL LIKELIHOOD

We now consider a modification of a statistical method for analyzing survival data, particularly for proportional hazard modeling, that was originally proposed by Cox (1975).This modified variant deals with spatial data and proves to be more convenient for computation, as discussed in Diggle (2006b), Diggle et al. (2010b), Diggle (2013b), Tamayo-Uria et al. (2014). This variant considers a partial likelihood obtained by conditioning on the locations s_i and times t_i, and examining the resulting log-likelihood for the observed time-ordering of the events. Define

$$p_i = \frac{\lambda(s_i, t_i|\mathscr{H}_{t_i})}{\int_W \lambda(s, t_i|\mathscr{H}_{t_i}) ds}, \tag{13}$$

then, the partial log-likelihood is given by

$$L_p(\theta) = \sum_{i=1}^{n} \log \; p_i. \tag{14}$$

The integral in (13) is often challenging to compute analytically. To address this, numerical integration techniques are commonly used to approximate the integral. However, this integral term becomes easy to compute when dealing with spatio-temporal point processes where the region W consists of a finite number of eventual locations s_j (indexed by $j = 1, 2, \ldots, N$), with $N \geq n$ (the number of observed events). In this case, the expression for the partial likelihood $L_p(\theta)$ under these circumstances is defined as follows (see Møller and Sørensen 1994; Diggle 2006b)

$$L_p(\theta) = \sum_{i=1}^{n} \log \; \frac{\lambda(s_i, t_i|\mathscr{H}_{t_i})}{\sum_{j\in\mathscr{R}_i} \lambda(s_j, t_i|\mathscr{H}_{t_i})ds}. \tag{15}$$

The variable $\mathscr{R}_i$, referred to as the *risk set*, is associated with a specific time point t_i within the context of this spatio-temporal point process. Typically, $\mathscr{R}_i$ is defined as a set that includes the indices of potential events at time t_i and extends up to the last event N. In other words, $\mathscr{R}_i$ is usually represented as $\mathscr{R}_i = \{i, i+1, i+2, \ldots, N\}$. This set is essential for calculating the likelihood in the given framework.

As already indicated in Cox (1972), when you maximize the partial likelihood, the corresponding estimates share the general asymptotic properties of maximum likelihood estimators. The partial likelihood method has, however, certain limitations. One of the primary limitations is the loss of efficiency in parameter estimation when compared to full maximum likelihood estimation. The partial likelihood may provide consistent estimates, but they could have larger variances or less precision, leading to wider confidence intervals. Another limitation is the potential for unidentifiability of certain parameters in the model. This means that some parameters may not be uniquely estimable using the partial likelihood approach, which can pose challenges in understanding and interpreting the model. While the loss of identifiability can be advantageous when non-identified parameters are considered nuisance parameters, it can also be a limitation in cases where those parameters are of primary interest or critical to the model interpretation. Regarding complexity, the partial likelihood method may be computationally intensive, particularly in complex models or large datasets, making it less practical in some situations. Related to dependency on model assumptions, like any statistical method, the partial likelihood relies on certain assumptions about the underlying data-generating process. Deviations from these assumptions can affect the reliability and accuracy of the results.

Therefore, while the partial likelihood is a valuable tool in certain statistical analyses, it is essential to be aware of its limitations and consider these factors when choosing the appropriate estimation method for a specific research question or dataset.

4.3. EXPECTATION-MAXIMIZATION (EM)

The Expectation-Maximization (EM) algorithm is a computational approach used to estimate parameters in statistical models when there are unobserved or missing data. It is an iterative algorithm that alternates between the E-step (Expectation step) and the M-step (Maximization step) until convergence.

In the Expectation step (E-step), the algorithm computes the expected values of missing or unobserved data, considering the current estimates of the model parameters. This stage includes determining the posterior probabilities or conditional expectations of latent variables, given both the observed data and current parameter estimates. The Maximization step (M-step) looks for estimating the parameters by maximizing the expected log-likelihood function obtained from the E-step. This involves identifying parameter values that maximize the likelihood of the observed data, while accounting for the expected values of the missing data. A convergence check occurs after each iteration of the E-step and M-step. The algorithm assesses convergence by evaluating a convergence criterion, typically based on changes in the log-likelihood function or parameter estimates between iterations. Each iteration involves repeating the E-step, M-step, and convergence check until the convergence criterion is satisfied.

The EM algorithm is particularly useful when dealing with incomplete data problems, such as clustering, mixture models, or missing data. By iteratively estimating the missing data and updating the model parameters, the algorithm seeks to find the maximum likelihood estimates for the parameters, even in the presence of missing or unobserved data (Liu et al. 2021a; Mohler et al. 2021). It is important to note that the EM algorithm does not guarantee finding the global optimum. It may converge to a local optimum depending on the initial parameter values and the properties of the likelihood function. Therefore, it is often recommended to run the algorithm multiple times with different initializations to increase the chances of finding a good solution. However, a key advantage of the EM algorithm is that convergence to the optimum tends to be numerically more stable and faster than direct likelihood optimization. This is because, in each iteration, the algorithm updates the parameters based on the expectations of the latent variables, which can provide a clearer direction toward the optimum.

In the context of this paper, we consider a self-exciting point process with the following structure

$$\lambda(s, t|\mathscr{H}_t) = \mu(s, t) + \sum_{i:t_i<t} g(s - s_i, t - t_i), \tag{16}$$

where $g(s, t)$ is a triggering kernel modeling the extent to which the risk following an event increases and spreads in space and time, and $\mu(s, t)$ is the background function, which yields the first generation. Then each event (s_i, t_i) triggers a new generation according to $g(s - s_i, t - t_i)$.

The EM algorithm applied to (16) is as follows. Following (12), and using a branching process representation of the model, we can rewrite the log-likelihood as

$$\begin{aligned}\log(L(\theta)) = &\sum_{i\in B\cup G} \log(\mu(s_i,t_i))\mathbb{1}_B(s_i) - \int_T\int_W \mu(s,t)\mathrm{d}s\mathrm{d}t\\ &+\sum_{i\in B\cup G} \log(g(s_i - s_{p(i)}, t_i - t_{p(i)}))\mathbb{1}_G(s_i)\\ &-\int_T\int_W \sum_{i:t_i<t} g(s-s_i, t-t_i)\mathrm{d}s\mathrm{d}t,\end{aligned} \tag{17}$$

where B and G are the sets of background and triggering events, respectively, $\mathbb{1}_B(s_i)$ is an indicator function that takes the value of 1 if $s_i \in B$, and 0 otherwise, and $p(i)$ is defined as the predecessor event of event i. Even though we typically lack knowledge about the precise branching arrangement of the process, we proceed under the assumption that we are able to create a probabilistic branching structure denoted as P, with

$$P_{ij} = \begin{cases}\text{probability that } i \text{ is a background event}, & i = j\\ \text{probability that } i \text{ is triggered by j}, & j < i\end{cases} \tag{18}$$

When we take the expectation of the likelihood with respect to P, we obtain the complete data log-likelihood

$$\begin{aligned}\mathbb{E}[\log(L(\theta))] = &\sum_i P_{ii}\log(\mu(s_i,t_i)) - \int_T\int_W \mu(s,t)\,\mathrm{d}s\,\mathrm{d}t\\ &+\sum_{j<i} P_{ij}\log(g(s_i - s_j, t_i - t_j)) - \sum_j \int_T\int_W g(s-s_j, t-t_j)\,\mathrm{d}s\,\mathrm{d}t.\end{aligned} \tag{19}$$

In this scenario, the estimation process can be divided into two distinct challenges related to density estimation. One involves estimating the background intensity, while the other focuses on estimating the triggering kernel. There are several methods for maximizing the variable $\mathbb{E}[\log(L(\theta))]$ by fitting the parameters involved in the intensity function. These methods include finding partial derivatives with respect to each variable and equating them to zero to derive analytical formulas for optimal values. However, it is important to note that specifying P_{ij} is still necessary. Since a Hawkes Process can be conceptualized as a combination of Poisson processes, this provides insight into how P_{ij} can be determined. Indeed, and denoting the estimators of P by p, in the estimation step, we estimate the probability that event i is a background event via the formula

$$p_{ii} = \frac{\mu(s_i,t_i)}{\lambda(s_i,t_i)}, \tag{20}$$

and the probability that event i is triggered by event j as

$$p_{ij} = \frac{g(s_i - s_j, t_i - t_j)}{\lambda(s_i,t_i)}. \tag{21}$$

We then iterate between the expectation and maximization steps until the estimates stabilize, reaching the specified number of iterations set by the user or until the convergence criterion established by the user is met. (Zhuang et al. 2002; Liu et al. 2021; Lo Galbo and Chiodi 2023).

4.4. BAYESIAN APPROACH

Bayesian inference for a spatio-temporal Hawkes process involves the use of Bayesian statistical methods to estimate the process parameters and make predictions based on observed data. Suppose that observations $Y = (Y_1, \dots, Y_n)$ have been generated by a probability model $p(Y_1, \dots, Y_n|\theta)$ with θ representing an unknown parameter vector. In Bayesian statistics, the analysis starts with a prior distribution $\pi(\theta)$ that encapsulates existing knowledge about θ derived from prior studies. In situations where we aim to minimize the influence of prior knowledge on our analysis, a non-informative choice for $\pi(\theta)$ is possible. Then the posterior distribution $p(\theta|Y_1, \dots, Y_n)$ combine information about θ from both the prior and the data

$$p(\theta|Y) = \frac{p(Y_1, \dots, Y_n|\theta)\pi(\theta)}{\int p(Y_1, \dots, Y_n|\theta)\pi(\theta)\mathrm{d}\theta}. \tag{22}$$

Understanding the posterior distribution facilitates the derivation of point estimates for θ, similar to the maximum likelihood framework. Additionally, it enables the representation of all uncertainties surrounding θ. This uncertainty can be quantified, providing a comprehensive understanding of the variability of the parameters through credible intervals, and facilitating model comparison and hypothesis testing within the Bayesian framework.

For a Hawkes process with conditional intensity $\lambda(s, t|\mathscr{H}_t)$, and background and triggering functions $\mu(s, t)$ and $g(s, t)$, let θ be the vector of involved parameters. The likelihood function given by (11) is then specified to represent the probability of observing data $Y = \{(s_i, t_i)\}_{i=1}^n$ given the parameters, i.e., $L(\theta) = L(\theta|Y)$. The posterior distribution will be $p(\theta|Y) \propto L(\theta|Y) \cdot \pi(\theta)$. Posterior distributions are analyzed to understand parameter estimates and uncertainties. Model validation involves comparing predictions with observed data, and predictions for future events are made based on parameter estimates. The strength of Bayesian inference lies in its ability to provide parameter estimates with accompanying measures of uncertainty, facilitating informed decision-making and improving prediction accuracy in spatio-temporal event applications.

Serafini et al. (2023) suggest an efficient method for approximately inferring the Bayesian parameters of a Hawkes process. This is achieved by employing Integrated Nested Laplace Approximation (INLA) methodology to estimate the posterior distribution of the Hawkes process parameters. The approximation relies on the linear decomposition of the likelihood. INLA is proposed as an alternative to MCMC methods, with a focus on analytically approximating the posterior distribution through deterministic techniques, leveraging numerical integration. They offer a technique to perform approximate Bayesian inference of Hawkes process parameters based on the use of the R-package `inlabru`. INLA and the `inlabru` package are efficient tools for Bayesian parameter inference in Hawkes processes, offering analytical approximation methods for computational speed and stability. Their deterministic techniques simplify the inference process, providing ease of use and reproducibility, especially with integration into the R programming language. Focused on the temporal dimension of Hawkes models, they excel in modeling temporal dependencies in point processes. The INLA methodology, combined with `inlabru`, brings additional advantages such as automatic model selection and handling missing data. While they may have limitations, their versatility and active user communities make them valuable for researchers and practitioners in Bayesian parameter inference for Hawkes processes. However, as mentioned, INLA and the `inlabru` package have limitations. Their efficiency may diminish in complex models, and deterministic approximations might lack the flexibility of stochastic methods, limiting their capability to capture certain posterior distributions. Model success depends on accurate specifications, and mis-specification could lead to biased results. Adjusting parameters may be necessary, and scalability issues could arise in large or computationally intensive datasets. Results interpretation might be less intuitive, and the effectiveness of the tools can be influenced by updates and ongoing development. Despite these limitations, INLA and inlabru remain valuable, with suitability depending on the specific modeling context and data characteristics. Moreover, the `inlabru` package is specifically applied to the temporal dimension of Hawkes models, providing an efficient and analytical approximation of the posterior distribution.

Jones-Todd and Helsdingen (2022) propose a Bayesian approach using the R-package `stelfi`, which employs Template Model Builder (TMB) (Kristensen et al. 2016) to fit various Hawkes process models. TMB is a C++ template library designed for the efficient implementation of statistical models. Its key features include automatic differentiation capability, efficiency in handling large datasets, and seamless integration with R, particularly using likelihood-based approaches. Users can express their statistical models in C++, and TMB automatically generates the necessary code to calculate both likelihoods and gradients. The versatility of TMB allows users to tackle complex statistical models while benefiting from community support. Despite its advantages, it comes with significant limitations. Programming in C++ may prove challenging for those unfamiliar with the language, and the inherent complexity of models can lead to intricate code. The integration with R, although designed to be seamless, may pose challenges for some users, especially those less acquainted with R. Additionally, the stability and availability of new features may depend on the development status and update frequency of TMB. It is noteworthy that, within the TMB package, only one scenario with a specific kernel is addressed, and the functions are tailored for a particular set of circumstances, limiting their applicability in broader contexts.

Lekha et al. (2021) rely on a customized Markov chain Monte Carlo (MCMC) method to estimate the model parameters. They formulate a flexible spatio-temporal Hawkes model to study extreme terrorist insurgencies and estimate parameters within a Bayesian hierarchical framework using a hybrid Metropolis-within-Gibbs particle Markov Chain Monte Carlo (MCMC) algorithm. Using the particle method, a set of particles approximates the posterior distribution of parameters, enabling efficient exploration of the parameter space. Subsequently, the Metropolis-within-Gibbs method is applied to perform more local and refined updates of model parameters. This combination aims to leverage the strengths of both methods, particularly beneficial in high-dimensional scenarios or when dealing with complex and multimodal posterior distributions. It enhances exploration of the parameter space and achieves accurate convergence in Bayesian parameter estimation for the model.

Nevertheless, the use of a customized MCMC method can be computationally demanding and resource-intensive, especially in high-dimensional scenarios. The effectiveness and accuracy of the estimation may also rely on the accurate specification of the model, with the possibility of inaccurate approximations in the case of formulation errors.

Another study that uses MCMC is Ebrahimian and Jalayer (2017). The method employed is a simulation-based approach designed to offer a reliable assessment of the spatial distribution of events within a specified forecast timeframe following the primary event. This technique considers the uncertainty associated with the ETAS model parameters by treating them as the posterior joint probability distribution conditioned on the events that have occurred. An MCMC simulation scheme is utilized to directly sample from the posterior probability distribution for ETAS model parameters. The study has limitations, including assuming a constant seismicity rate, which may not be valid universally. It relies on the ETAS model, which has constraints and may not fully capture all aspects of seismic activity. The assumption of time-independent spatial event distribution may not always align with reality. Additionally, the simulation-based approach used may be computationally intensive and impractical for real-time forecasting in some scenarios. The study proposes an advantageous method for short-term earthquake forecasting through an adaptive Metropolis-Hastings algorithm and the application of the ETAS model. The methodology, validated in past events, stands out for its robust approach and the ability to consider uncertainty in model parameters. The method shows potential for emergency decision-making and risk mitigation.

Ross (2021) introduces a Bayesian estimation method for the Epidemic-Type Aftershock Sequence (ETAS) model in earthquake forecasting. It compares the performance of a direct MCMC scheme and a latent variable scheme, demonstrating that the latent variable scheme is more efficient. The paper emphasizes the importance of incorporating parameter uncertainty in earthquake predictions and validates the accuracy of the predictions made by the Bayesian ETAS model. The advantages of the latent variable scheme for estimating the ETAS model include computational efficiency, improved convergence, and practical applicability to large earthquake catalogs, making it a superior choice for seismic forecasting. However, while the latent variable scheme offers significant advantages in terms of computational efficiency and convergence, its complexity, potential additional computational burdens, and the need for expertise in Bayesian inference and MCMC methods could limit its widespread adoption and practical implementation in seismic prediction and related fields.

In Molkenthin et al. (2022), the GP-ETAS model, a Bayesian approach for modeling spatio-temporal seismic occurrences, is analyzed. It employs a Gaussian process to represent the background intensity and triggering functions, incorporating latent variables to simplify the inference problem. The paper presents experimental results comparing GP-ETAS with standard ETAS models, discusses challenges in inference, and provides insights into parameter estimation and computational efficiency. The GP-ETAS model differs from standard ETAS models by incorporating a non-parametric Bayesian approach to model the background intensity using a Gaussian Process, while maintaining a classical parametric form for the triggering function. This results in a more flexible, robust, and data-driven representation of the spatio-temporal ETAS model. Challenges in inferring GP-ETAS model parameters include the complex model form, the need for comprehensive uncertainty quantification, and robust recovery of background intensity and triggering function parameters.

Bayesian inference is performed using Monte Carlo sampling techniques, such as MCMC. One limitation is the need for a more informed choice of prior for the Gaussian process, requiring further research to incorporate spatial information about fault zones and geological features. Despite limitations, the Bayesian approach in the GP-ETAS model offers key advantages, including robust uncertainty quantification, improved reliability compared to maximum likelihood estimation methods, and enhanced adaptability to diverse data scenarios. This framework also allows the incorporation of prior information, enriching the flexibility and applicability of the model in modeling spatio-temporal seismic events.

4.5. MACHINE LEARNING

In the context of Hawkes processes, machine learning plays a crucial role in modeling, parameter estimation, and predicting future events. Using algorithms such as support vector machines, neural networks, and regression methods, more complex models are designed to capture subtle patterns in the data. Hyperparameter optimization, often necessary in Hawkes models, is performed using machine learning techniques such as grid or random search. These models are employed to forecast future events, and supervised learning enhances these predictions through training on historical data. Machine learning metrics, such as precision and F1-score, evaluate model performance. In environments with large temporal datasets, machine learning facilitates efficient data manipulation and the implementation of Hawkes algorithms. Overall, machine learning enhances the ability to understand the temporal dynamics of events in Hawkes processes.

Muir and Ross (2023) focus on using deep Gaussian processes (DGP) to model seismicity background rates in earthquake catalogs. It introduces the deep-GP-ETAS model, designed for capturing time-varying seismicity rates, particularly in seismic swarms dominated by aseismic tectonic processes. Gaussian Processes (GPs) serve as prior distributions over the space of functions (in their case, the space of continuous functions within the interval $[0, T]$). They are characterized by the property that the distribution $\{f(x_i)\}$ of a GP f, evaluated at any finite set of points $\{x_i\}$, follows a multivariate normal distribution. This distribution is determined by a mean function $m(x)$ and a covariance function $C(x, x')$, where x, x' represent points in the function space:

$$\begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_n) \end{bmatrix} \sim N\left(\begin{bmatrix} m(x_1) \\ m(x_2) \\ \vdots \\ m(x_n) \end{bmatrix}, \begin{bmatrix} C(x_1, x_1) & C(x_1, x_2) & \dots & C(x_1, x_n) \\ C(x_2, x_1) & C(x_2, x_2) & \dots & C(x_2, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ C(x_n, x_1) & C(x_n, x_2) & \dots & C(x_n, x_n) \end{bmatrix}\right). \tag{23}$$

The GP is fully characterized by its mean and covariance functions. The pivotal component in GP modeling is the covariance function, which delineates the smoothness and characteristic length scales of the GP. An illustrative example of this is the covariance function known as squared exponential, $C(x, x') = \sigma \exp\left(-\frac{\|x-x'\|^2}{2l^2}\right)$, which generates functions f that are infinitely differentiable and have characteristic length-scale l and amplitude σ, but there are other several options.

Muir and Ross (2023) detail the process of generating white noise, grid discretization, and the use of DGP to approximate multilayer models. It discusses model priors, sampling schemes, and provides examples of synthetic earthquake catalogs. The results highlight the effectiveness of DGP in capturing variable background rates in seismicity. The deep-GP-ETAS model extends the ETAS model, employing a robust deep Gaussian process formulation for the background rate. This probabilistic approach adjusts its structure to match data constraints, efficiently sampled using a Metropolis-within-Gibbs scheme and stochastic partial differential equation (SPDE) approximation for Matérn Gaussian processes.

The limitations of Gaussian process (GP) models in large-scale inverse problem settings, such as poor scalability, are addressed in the paper. The proposed solution introduces an approximation method for GP models using SPDE. Nevertheless, the DGP-ETAS model stands out for its ability to analyze seismicity background rates in earthquake sequences. It offers a detailed quantitative analysis characterizing time-varying seismicity rates under general assumptions, incorporates prior information allowing the inclusion of a priori knowledge about ETAS parameters, essential for robust seismicity analysis, and provides accurate uncertainty estimates for parameters, crucial for assessing hazard probabilities accurately in forecasting scenarios. Its flexibility allows adaptation to various cases, making it a valuable tool for understanding seismicity in different situations.

Stockman et al. (2023) discuss the development and application of neural point processes for earthquake forecasting. The prevalent approach in neural point processes involves obtaining a concise representation of event history through the use of a Recurrent Neural Network (RNN) (Du et al. 2016). In this methodology, the input representing inter-event times ($\tau_i = t_i - t_{i-1}$), is initially fed into the RNN. The hidden state of the RNN (h_i) undergoes an update using learnable parameters (W^h, w^τ, b^h) and an activation function (σ),

$$h_i = \sigma(W^h h_{i-1} + w^\tau \tau_i + b^h). \tag{24}$$

The conditional intensity function is then expressed as a function of the elapsed time from the most recent event, dependent on the hidden state of the RNN,

$$\lambda(t|\mathscr{H}_t) = \phi(t - t_i|h_i). \tag{25}$$

Here, ϕ is a non-negative function known as the hazard function, and t_i denotes the time of the most recent event. To circumvent direct numerical integration of the intensity function, the integral of the hazard function is modeled using a fully connected neural network (Omi et al. 2019)

$$\Phi(\tau|h_i) = \int_0^\tau \phi(s|h_i)\,\mathrm{d}s. \tag{26}$$

With this model construction, the log-likelihood of observing a sequence of event times (t_i) can be expressed as

$$\log(\{t_i\}) = \sum_i \left[\log \frac{\partial}{\partial \tau}\Phi(\tau_i|h_i) - \Phi(\tau_i|h_i)\right]. \tag{27}$$

Stockman et al. (2023) compare the performance of the neural model to the ETAS model, and highlights the advantages of the neural model, such as robustness to missing data and computational efficiency. It also explores the use of smaller events to forecast larger earthquakes and discusses the limitations and future directions for the model. The paper provides detailed descriptions of the model components and methodology, as well as access to the data set and models used in the study. Regarding the modeling of magnitudes through a completely unconstrained density that is also time-history dependent, the model may struggle with simulating events into the future, as it can only leverage the intensity function at a given point in time with a thinning algorithm, unlike the ETAS model which can be simulated due to its equivalent branching process formulation. However, the advantages of using smaller events to forecast larger earthquakes lie in their potential to provide valuable information about seismic activity, and the likelihood of observing target events based on the intensity function is calculated using a machine learning variant of point processes for short-term earthquake forecasting.

Chen et al. (2021) propose a new class of parameterizations for spatio-temporal point processes that leverage neural ordinary differential equations (Neural ODEs) as a computational method, enabling flexible, high-fidelity models of discrete events localized in continuous time and space. They focus on combining these Neural ODEs and continuous normalizing flows (CNFs). In this way, these authors work on three models: time-varying CNF, which allows a spatial distribution to vary over time, without necessarily depending on previous events in the spatial domain; jump CNF, where the dynamics of the spatial distribution are conditioned on the continuous hidden state, allowing instantaneous updates of the distribution after each new observed event; and attentive CNF, which, using a transformer architecture, considers conditional dependencies between previous events and the current spatial distribution, allowing for continuous and parallel updates of the distribution. The combination of Neural ODEs and CNFs enables modeling complex distributions in the spatial and temporal domain, providing flexibility to adapt to a variety of data and contexts.

Zuo et al. (2020) propose a transformer Hawkes process (THP) model, which is a model to capture long-term temporal dependencies in event sequences. This model combines elements of two approaches: Hawkes processes, which model the occurrence of time-dependent events, and transformers, which are machine learning models designed to capture long-term relationships in sequential data. THP utilizes the self-attention mechanism of transformers to capture long-term dependencies among events in a sequence. This allows the model to analyze and learn complex patterns of temporal interaction between events. One of the advantages of THP is its computational efficiency, making it suitable for handling large volumes of event sequence data. Additionally, numerical experiments have shown that THP outperforms existing models in terms of prediction accuracy and its ability to capture the underlying temporal structure in the data.

The model proposed by Zuo et al. (2020) is characterized by its self-attention module, which plays a fundamental role. Unlike recurrent neural networks (RNNs), which utilize a recurrent structure, the attention mechanism in THP does not have this characteristic. However, since their model still needs to capture the temporal information of the inputs, i.e., the time-stamps, they propose a temporal encoding procedure defined by

$$[z(t_j)]_i = \begin{cases} \cos\left(t_j/10000^{\frac{i-1}{M}}\right) & \text{if } i \text{ es odd,} \\ \sin\left(t_j/10000^{\frac{i}{M}}\right) & \text{if } i \text{ es even.} \end{cases} \tag{28}$$

So, by using trigonometric functions to define a temporal encoding for each time-stamp, i.e., for each t_j, they deterministically compute $z(t_j) \in \mathbb{R}^M$, where M is the dimension of the encoding.

4.6. SEMI-PARAMETRIC AND NON-PARAMETRIC APPROACHES

A semi-parametric approach in the context of Hawkes processes involves combining parametric and non-parametric elements when modeling the conditional intensity of the process. In a purely parametric approach, a specific form for the conditional intensity would be assumed, implying detailed assumptions about the structure of the process. On the other hand, a non-parametric approach would avoid such specific assumptions but may face challenges in terms of estimation and generalization. The semi-parametric approach seeks a balance by combining parametric and non-parametric aspects. In the case of Hawkes processes, this might involve assuming a parametric form for certain components of the model, such as the background rate, while other components, like the clustering response, are modeled in a non-parametric fashion.

For non-parametric estimation, it is common to discuss stochastic declustering. Consider a Hawkes process characterized by a conditional intensity given by (3), where $\mu(s, t)$ denotes the background rate and $g(s, t)$ represents the rate of occurrences triggered by an event at time 0 and location at the origin. As seen previously, the probability of an event being a background event (background probability) is given in (20) and (21), where p_{ij} and p_{jj} are defined.

Another interpretation for this equation is that, once an event has occurred at (s, t), we can state that it is a background event with probability p_{jj}. Furthermore, for each $i = 1, \ldots, j-1$, event i triggers p_{ij} direct offspring at (s_j, t_j). In this way, event j is splitted into background and offspring from preceding events, as explained by Zhuang and Ogata (2004). Consequently, the aforementioned approach offers a non-parametric method for estimating functions $\mu(\cdot, \cdot)$ and $g(\cdot, \cdot)$. For example, $g(\cdot, \cdot)$ can be estimated by

$$\hat{g}(s, t) = \frac{\sum_{i,j} p_{ij} \cdot \mathbb{I}\left(|t_j - t_i - t| < \delta_t\right) \mathbb{I}(|s_j - s_i - s| < \delta_s)}{4\delta_t \delta_s \sum_{i,j} p_{ij}}. \tag{29}$$

In this context, the denominator serves the purpose of normalization, and δ_t and δ_s represent two small positive values. Additionally, an alternative approach for estimating $\mu(\cdot, \cdot)$ involves techniques such as weighted kernel estimation, as exemplified below

$$\hat{\mu}(s, t) = \sum p_{ii} Z_{h_s}(s - s_i) Z_{h_t}(t - t_i). \tag{30}$$

In this scenario, Z_h represents the Gaussian kernel with bandwidth h, where h_s and h_t are the bandwidths used for spatial and temporal smoothing, respectively.

For the analysis, during the estimation of $\mu(s,t)$ and $g(s,t)$, it becomes necessary to have information about p_{ii} and p_{ij}. Conversely, when estimating p_{ii} and p_{ij}, knowledge of μ and g is required. This cyclic dependency is resolved through an iterative algorithm. By starting with an initial assumption for μ and g based on an observed sequence of events $\{(s_i,t_i)\}_{i=1}^{n}$ within a space-time window $W \times T$, we compute p_{ii} and p_{ij} for all possible i and j. Subsequently, we estimate the background rate μ and each component in the clustering part g using p_{ii} and p_{ij}, employing non-parametric methods such as kernel estimation or histogram. Once μ and g have been updated, the process iterates by recalculating p, continuing until convergence is achieved or until the stopping criteria are met.

Zhuang and Mateu (2019) apply a semi-parametric approach to an extended Hawkes model, where the periodic components are separated from the long-term trend in the background rate. The introduction of two relaxation parameters for the background rate and clustering effect adds flexibility to the model. Furthermore, the fact that these relaxation parameters are estimated using maximum likelihood implies a semi-parametric approach in the estimation of the model parameters. This approach allows for capturing both the parametric structure and the non-parametric complexities of the process, providing greater flexibility and adaptability to the model.

In the context of Hawkes processes, a non-parametric approach involves avoiding specific assumptions about the functional form of the conditional intensity of the process. Instead of imposing a predetermined parametric structure, this approach seeks to model the intensity in a more flexible manner, allowing the data to dictate the shape of the function. This is achieved by directly estimating the conditional intensity from observed data using methods such as kernel estimation or histogram-based methods. While flexibility is a strength, non-parametric methods may face challenges in terms of precise estimation, especially when data is limited. Additionally, extrapolation beyond the observed range may be more uncertain compared to well-specified parametric models. However, a non-parametric approach may better adapt to data with complex or unusual patterns, as it does not impose specific restrictions on the form of the conditional intensity. The choice between parametric and non-parametric approaches will depend on the nature of the data and the amount of available information.

In Zhuang (2020), the main focus is on using kernel functions to address the inherent challenges in estimating these types of point processes. The proposed solutions provide an immediately applicable tool for modeling, analyzing, and forecasting in various applications using different point process data, employing the Hawkes-type point process. This paper discusses techniques for using the Hawkes process and exploring the encouraging causal correlation among discrete events. Once the direction of the model extension is determined, they construct, estimate, and diagnose a new model using stochastic reconstruction techniques along with some non-parametric estimation methods, among which the kernel function is efficient, straightforward, and easy to implement.

Other articles that explore the implementation of a non-parametric framework for the analysis of Hawkes models include Mohler et al. (2011), Fox et al. (2016), Yuan et al. (2019). Moreover, non-parametric Hawkes can be broadly categorized into two primary approaches: frequentist and Bayesian. The frequentist approach to Hawkes process modeling and inference involves assuming that the excitation function (or matrix) can be defined on a binned grid (or a set of grids). Within each bin, the values of the functions are taken as piecewise constant. The bin width is chosen to be expressive enough to model the local variations of the self-excitation effect. Some references are Lewis and Mohler (2011), Bacry and Muzy (2016). On the other hand, in Bayesian non-parametric treatment of Hawkes processes, the assumption revolves around modeling the triggering kernel and background rates using distributions (or mixtures of distributions) from the “Exponential Family”. These distributions, through their conjugacy relationships, allow for closed-form computations of sequential updates in the model. We can cite the research of Donnet et al. (2018), Zhuang and Mateu (2019), Zhang et al. (2019).

4.7. MODEL ASSESSMENT AND DIAGNOSTICS

Residual analysis methods are essential for assessing the adequacy of point process models. These methods enable the detection of shortcomings in the model fitting, and provide suggestions for improvements. Reinhart (2018) already made some reviews on this. Among the main residual analysis methods used in evaluating Hawkes point process models, we have the idea of rescaled residuals, which is a technique that transforms events from the point process so that, under a good model fit, they follow a homogeneous Poisson process. This transformation can reveal deviations from the model, such as temporal and spatial dependence among events. And we also have the thinning residual method, which involves removing events from the original point process based on their conditional intensity, allowing for the identification of regions where the model has a poor estimation of the intensity.

These methods have been fundamental in the evaluation of point process models, as exemplified in works such as Ogata (1988), which is a fundamental contribution to the field of residual analysis applied to point process models, specifically in seismology. Ogata develops statistical models for earthquake occurrences and proposes residual analysis methods to assess the fit of these models. This author uses rescaled residuals to identify deficiencies in the ETAS model, such as temporal and spatial dependence among events. Ogata demonstrates how to transform event times and locations to follow a homogeneous Poisson process, allowing for the detection of model deviations. The general procedure (based on Meyer 1971) involved a collection $\{N_i\}$ of completely simple univariate point processes on the real half-line, provided that $\int_0^\infty \lambda(t,i)\mathrm{d}t = \infty$ for each i. By rescaling the points, moving each point (t,i) to the point $(\int_0^t \lambda(t',i)\mathrm{d}t', i)$, one obtains a sequence of independent homogeneous Poisson processes with unit rate. The application of this method to earthquake data provides information about the accuracy of the ETAS model, and suggests improvements in modeling seismic activity. Moreover, Schoenberg (2003) presents methods to evaluate the fit of multidimensional point process models through residual analysis, and applies these methods to models of space-time-magnitude distribution of earthquakes, using the multidimensional version of the ETAS model. Indeed, Schoenberg (2003) employs two residual analysis methods to assess the fit of the multidimensional ETAS model. The first method involves rescaled residuals, transforming points along a coordinate to form a homogeneous Poisson process within an irregular and random boundary. This author also extended the previous approach to marked space-time point processes, that is, by transforming each point (t_i, x_i, y_i, m_i) to $\left(\int_0^t \lambda(t',i)\,\mathrm{d}t', x_i, y_i, m_i\right)$, one obtains again an independent sequence of homogeneous

Poisson processes with unit rate. The second method thins the point process according to the conditional intensity, creating a homogeneous Poisson process in the original space. Indeed, given a space-time-magnitude point process N with conditional intensity λ, one can obtain a homogeneous Poisson process with rate b by independently retaining each point (t_i, x_i, y_i, m_i) in the original point process with probability $b/\lambda(t_i, x_i, y_i, m_i)$, where b is the minimum of $\lambda(t, x, y, m)$ over the entire observation region. When this thinning is done using the estimated conditional intensity $\hat{\lambda}$ in place of λ, the remaining points, which we call thinned residual points, can be examined for their uniformity. Another significant contribution in this regard is Zhuang (2006), which details the second-order residual analysis, which is particularly useful when investigating second-order properties, such as clustering or inhibition. Given a second-order $\mathscr{F}$-predictable function $H(t, x; t', x')$, which can be generated from processes of the form $h_1(t, x) \cdot h_2(t', x')$, where h_1 and h_2 are first-order predictable processes, through linear combinations and monotone limits, the second-order innovation with respect to $H(t, x; t', x')$ for a domain D is defined as

$$\begin{aligned} V_2(D, H, \lambda) = &\iint_{D \diagdown \mathrm{diag}_2(D)} H(t, x; t', x') N(dt' \times dx') N(dt \times dx) \\ &- \iint_D H(t, x; t', x') \lambda(t, x) \lambda(t', x') \mu(dt' \times dx') \mu(dt \times dx), \end{aligned} \tag{31}$$

where $\mathrm{diag}_2(D) = \{(t, x, t', x') \in D\}$. The second-order residual can be defined now as $R_2(D, \hat{H}\hat{\lambda}) = V_2(D, \hat{H}, \hat{\lambda})$. Lastly, the book Zhuang and Ogata (2006) introduces the underlying theory of point processes and residual analysis methods, providing a solid theoretical foundation for the application of these methods. All of these works have been seminal in residual analysis.

However, more general ideas on residual analysis are in Zhuang (2015), which discusses not only conditional intensity models but also other types of point process models. Based on the technique of residual analysis, they propose weighted likelihood estimators for temporal, spatial, and spatio-temporal point processes, through the weighted likelihood of the form,

$$\log \mathrm{WL} = \sum_{x_i \in B \cap N} h(x_i - x_0) \log \chi(x_i) - \int_B h(x - x_0) \chi(x) \ell(\mathrm{d}x). \tag{32}$$

In (32), the term χ represents a function that is specific to the context of the model being considered. Specifically, by replacing χ with μ, λ, or λ_p, we can obtain the weighted versions of the Poisson likelihood (WLPoisson), the likelihood (WL), and the pseudo-likelihood (WLpseudo) for moment intensity models, conditional intensity models, and Papangelou intensity models, respectively. Other innovative process-based diagnostics include Gordon et al. (2012), who employed Voronoi residuals, deviations, super-thinning, and graphical residual methods, among other residual analysis methods. Also, Bray et al. (2014) proposed a new residual analysis method for spatial or spatio-temporal point processes involving inspecting the differences between the modeled conditional intensity and the observed number of points over the Voronoi cells generated by the observations. The resulting residuals can be used to construct diagnostic methods with greater statistical power than residuals based on rectangular grids. The rescaled Voronoi residuals take the form $\hat{R}_V(C_i) = \frac{1-\int \hat{\lambda}(t,x,y)\,\mathrm{d}t\,\mathrm{d}x\,\mathrm{d}y}{\sqrt{\int \hat{\lambda}(t,x,y)\,\mathrm{d}t\,\mathrm{d}x\,\mathrm{d}y}}$. These residuals provide a measure of deviation from the expected intensity distribution within each Voronoi cell C_i. For further details, please refer to the respective cited articles.

5. SOME FURTHER CHALLENGING ISSUES

An important aspect of a Hawkes model is capturing the triggering effect of the event on its subsequent events. Since the distribution of point processes is completely specified by the conditional intensity function (the occurrence rate of events conditioning on the history), such triggering effect has been captured by an influence kernel function embedded in the conditional intensity. In statistical literature, the kernel function usually assumes a parametric form. For example, the original work by Hawkes (Hawkes 1971) considers an exponential decaying influence function over time, and the seminal work (Ogata 1998) introduces an ETAS model, which considers an influence function that exponentially decays over space and time.

With the increasing complexity of modern applications, there has been much recent effort in developing recurrent neural network (RNN)-based point processes, leveraging the rich representation power of RNNs (Du et al. 2016; Mei and Eisner 2017; Xiao et al. 2017).

However, there are several limitations of existing RNN-based models. First, such models typically do not consider the kernel function, and thus, the RNN approach does not enjoy the interpretability of the kernel function based models. Second, the popular RNN models such as Long Short-Term Memory (LSTM) still implicitly discounts the influence of events over time (due to their recursive structure). Such assumptions may not hold in many real-world applications. Third, a majority of the existing works mainly focus on one-dimensional temporal point processes.

Although there are works on marked point processes (Du et al. 2016; Mei and Eisner 2017; Reinhart 2018), they are primarily based on simplifying assumptions that the marks are conditionally independent of the event's time and location, which is equivalent to assuming the kernel is separable; these assumptions may fail to capture some complex non-stationary, time- and location-dependent triggering effects for various types of events, as observed for many real-world applications

One modern (and still open for research) contribution would be presenting a novel general non-stationary point process model based on neural networks, referred to as the neural spectral marked point process (NSMPP). The key component is a new powerful representation of the kernel function using neural networks, which enables us to go beyond stationarity (and thus go beyond standard Hawkes processes) and has the capacity to model high-dimensional marks. Understanding and characterizing what kernels can be learned through such an approach is left for future studies. Also, when extending to high-dimensional mark space, more efficient algorithm is needed for the computation of the log-likelihood.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Declarations

Conflict of interest As the corresponding author, I declare that there is no Conflict of interest for any of the authors regarding the publication of this manuscript.

[Received March 2024. Revised August 2024. Accepted August 2024.]

REFERENCES

Asmussen S (2003) Applications of mathematics: stochastic modelling and applied probability, applied probability and queues. Springer, Oxford University Press

Bacry E, Muzy J (2016) First- and second-order statistics characterization of hawkes processes and non-parametric estimation. IEEE Trans Inf Theory 62:2184–2202

Bacry E, Muzy JF (2014) Hawkes model for price and trades high-frequency dynamics. Quant Finance 14(7):1147–1166

Baddeley A, Rubak E, Turner R (2015) Spatial point patterns: methodology and applications with R. Chapman and Hall/CRC Press, London

Bray A, Wong K, Barr CD, Schoenberg FP (2014) Voronoi cell based residual analysis of spatial point process models with applications to Southern California earthquake forecasts. Ann Appl Stat 8(4):2247–2267

Cai X, Zhang W, Guan Y (2024) Latent network structure learning from high-dimensional multivariate point processes. J Am Stat Assoc 119(545):95–108

Chen RT, Amos B, Nickel M (2021) Neural spatio-temporal point processes. In: International conference on learning representations (ICLR)

Chiodi M, Adelfio G (2020) Including covariates in the etas model triggered seismicity. d/SEAS Working Paper. Forthcoming

Chiodi M, Nicolis O, Adelfio G, DàAngelo N, GonzÃ lez A (2021) Etas space-time modeling of chile triggered seismicity using covariates: Some preliminary results. Appl Sci 11(19)

Cox DR (1972) Regression models and life tables (with discussion). J R Stat Soc B 34:187–220

Cox DR (1975) Partial likelihood. Biometrika 62(2):269–276

Diggle PJ (2006a) Spatio-temporal point processes: methods and applications. Number 107 in Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, Boca Raton, FL

Diggle PJ (2006b) Spatio-temporal point processes, partial likelihood, foot and mouth disease. Stat Methods Med Res 15:325–336

Diggle PJ (2013) Statistical analysis of spatial and spatio-temporal point patterns. Stat Sci 28(3):436–455

Diggle PJ (2013b) Statistical analysis of spatial and spatio-temporal point patterns. Chapman & Hall Monographs on Statistics & Applied Probability. CRC Press, Boca Raton, Florida, 3 edition

Diggle PJ, Guan Y, Hart A, Paize F, Stanton M (2010) Estimating individual-level risk in spatial epidemiology using spatially aggregated information on the population at risk. J Am Stat Assoc 105(492):1394–1402

Diggle PJ, Kaimi I, Abellana R (2010) Partial-likelihood analysis of spatio-temporal point-process data. Biometrics 66(2):347–354

Donnet S, Rivoirard V, Rousseau J (2018) Nonparametric bayesian estimation of multivariate hawkes processes

Du N, Dai H, Trivedi R, Upadhyay U, Gomez-Rodriguez M, Song L (2016) Recurrent marked temporal point processes: Embedding event history to vector. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 1555–1564

Ebrahimian H, Jalayer F (2017) Robust seismicity forecasting based on bayesian parameter estimation for epidemiological spatio-temporal aftershock clustering models. Sci Rep 7:9803

Fang S, Xu Y, Xu Z, Zhu C, Guan Y (2023) Group network hawkes process. J Am Stat Assoc:1–17

Fox EW, Schoenberg FP, Gordon JS (2016) Spatially inhomogeneous background rate estimators and uncertainty quantification for nonparametric hawkes point process models of earthquake occurrences. Ann Appl Stat 10(3):1725–1756

González JA, Rodríguez-Cortés FJ, Cronie O, Mateu J (2016) Spatio-temporal point process statistics: a review. Spat Stat 18:505–544

Gordon JS, Clements RA, Schoenberg FP, Schorlemmer D (2015) Voronoi residuals and other residual analyses applied to csep earthquake forecasts. Spat Stat 14b:133–150

Grimmett G, Stirzaker D (2001) Probability and random processes. Oxford University Press

Hawkes A (1971) Spectra of some self-exciting and mutually exciting point processes. Biometrika 58(1):83–90

Hawkes A, Chen J (2021) A personal history of hawkes process. Proc Inst Stat Math 69(2):123–143

Huang Q, Gerstenberger M, Zhuang J (2016) Current challenges in statistical seismology. Pure Appl Geophys 173(1):1–3

Illian J, Penttinen A, Stoyan H, Stoyan D (2007) Statistical analysis and modelling of spatial point patterns. Wiley, Chichester, UK

Jacobsen M (2006) Point process theory and applications: marked point and piecewise deterministic processes. Birkhäuser

Jones-Todd CM, van Helsdingen A (2022) stelfi: an R package for fitting hawkes and log-gaussian cox point process models. Methods in Ecology and Evolution

Kristensen K, Nielsen A, Berg CW, Skaug H, Bell BM (2016) TMB: automatic differentiation and laplace approximation. J Stat Softw 70(5):1–21

Kwon J, Zheng Y, Jun M (2023) Flexible spatio-temporal hawkes process models for earthquake occurrences. Spat Stat 54:100728

Laub PJ, Lee Y, Taimre T (2021) The elements of hawkes processes. Springer

Lekha LS, Tucker JD, Huerta G (2021) Spatio-temporal extreme event modeling of terror insurgencies

Lewis E, Mohler G (2011) A nonparametric em algorithm for multiscale hawkes processes. J Nonparam Stat 1:1–20

Liu X, Carter J, Ray B, Mohler G (2021a) Point process modeling of drug overdoses with heterogeneous and missing data. Ann Appl Stat

Liu X, Carter J, Ray B, Mohler G (2021) Point process modeling of drug overdoses with heterogeneous and missing data. Ann Appl Stat 15(1):88–101

Lo Galbo G, Chiodi M (2023) Severe convective storms' reproduction: Empirical analysis from the marked self-exciting point processes point of view. Research Square. PREPRINT (Version 1)

Mei H, Eisner J (2017) The neural hawkes process: A neurally self-modulating multivariate point process

Meyer PA (1971) Demonstration simplifiée d'un théorème de knight. In Séminaire de Probabilités V, pages 191–195. Université Strasbourg, Lecture Notes in Mathematics, 191

Miscouridou X, Bhatt S, Mohler G, Flaxman S, Mishra S (2023) Cox-hawkes: doubly stochastic spatiotemporal poisson processes. Trans Mach Learn Res

Mohler G, Mishra S, Ray B, Magee L, Huynh P, Canada M, O'Donnell D, Flaxman S (2021) A modified two-process knox test for investigating the relationship between law enforcement opioid seizures and overdoses. Proc R Soc A

Mohler G, Short MB, Brantingham PJ, Schoenberg FP, Tita GE (2011) Self-exciting point process modeling of crime. J Am Stat Assoc 106(493):100–108

Molkenthin C, Donner C, Reich S (2022) Gp-etas: semiparametric bayesian inference for the spatio-temporal epidemic type aftershock sequence model. Stat Comput 32:29

Møller J, Sørensen M (1994) Statistical analysis of a spatial birth-and-death process model with a view to modelling linear dune fields. Scand J Stat 21(1):1–19

Muir JB, Ross ZE (2023) A deep gaussian process model for seismicity background rates. Geophys J Int 234(1):427–438

Ogata Y (1988) Statistical models for earthquake occurrences and residual analysis for point processes. J Am Stat Assoc 83:9–27

Ogata Y (1998) Space-time point-process models for earthquake occurrences. Ann Inst Stat Math 50(2):379–402

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

Omi T, Ueda N, Aihara K (2019) Fully neural network based model for general temporal point processes. In: Wallach H, Larochelle H, Beygelzimer A, d' Alché-Buc F, Fox E, Garnett R (eds) Advances in neural information processing systems, volume 32, pp 2122–2132. Curran Associates, Inc

Peng RD, Schoenberg FP, Woods JA (2005) A spacetime conditional intensity model for evaluating a wildfire hazard index. J Am Stat Assoc 100(469):26–35

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

Ross GJ (2021) Bayesian estimation of the etas model for earthquake occurrences. Bull Seismol Soc Am 111(3):1473–1480

Schatz M, Wheatley S, Sornette D (2022) The arma point process and its estimation. Econom Stat 24:164–182

Schoenberg F (2023) Estimating covid-19 transmission time using hawkes point processes. Ann Appl Stat 17(4):3349–3362

Schoenberg FP (2003) Multidimensional residual analysis of point process models for earthquake occurrences. J Am Stat Assoc 98(464):789–795

Schorlemmer D, Werner M, Marzocchi W, Jordan T, Ogata Y, Jackson D (2018) The collaboratory for the study of earthquake predictability: Achievements and priorities. Seismol Res Lett 89(4):1305–1313

Serafini F, Lindgren F, Naylor M (2023) Approximation of bayesian hawkes process with inlabru. Wiley

Shchur O, Biloš M, Günnemann S (2020) Intensity-free learning of temporal point processes

Shchur O, Türkmen AC, Januschowski T, Günnemann S (2021) Neural temporal point processes: A review

Stindl T, Kwan J, Chen F, Yongtao G (2024) Modelling gunfire in Washington. DC using a spatiotemporal hawkes process with nonseparable contagious gunfire intensity, Manuscript

Stockman S, Lawson DJ, Werner MJ (2023) Forecasting the 2016–2017 central apennines earthquake sequence with a neural point process. Earths Fut 11:e2023EF003777

Tamayo-Uria I, Mateu J, Diggle PJ (2014) Modelling of the spatio-temporal distribution of rat sightings in an urban environment. Spat Stat 9:192–206

Tench S, Fry H, Gill P (2016) Spatio-temporal patterns of ied usage by the provisional irish republican army. Eur J Appl Math 27(3):377–402

Truccolo W, Eden UT, Fellows MR, Donoghue JP, Brown EN (2005) A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. J Neurophysiol 93:1074–1089

Xiao S, Yan J, Chu SM, Yang X, Zha H (2017) Modeling the intensity function of point process via recurrent neural networks

Yan J, Xu H, Li L (2019) Modeling and applications for temporal point processes. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. Association for Computing Machinery, pp 3227–3228

Yuan B, Li H, Bertozzi AL, Brantingham PJ, Porter MA (2019) Multivariate spatiotemporal hawkes processes and network reconstruction. SIAM J Math Data Sci 1(2):356–382

Zhang R, Walder CJ, Rizoiu M, Xie L (2019) Efficient non-parametric bayesian hawkes processes. In: Proceedings of the twenty-eighth international joint conference on artificial intelligence, IJCAI 2019, Macao, China, pp 4299–4305

Zhu S (2019) Spatio temporal point-process simulator

Zhu S, Li S, Peng Z, Xie Y (2019) Imitation learning of neural spatio-temporal point processes. Submitted on 13 Jun 2019 (v1), last revised 22 Jan 2021 (this version, v4)

Zhuang J (2006) Second-order residual analysis of spatiotemporal point processes and applications in model evaluation. J R Stat Soc Ser B Stat Methodol 68(4):635–653

Zhuang J (2015) Weighted likelihood estimators for point processes. Spat Stat 14:166–178

Zhuang J (2020) Estimation, diagnostics, and extensions of nonparametric hawkes processes with kernel functions. Jpn J Stat Data Sci 3:391–412

Zhuang J, Chang CP, Ogata Y, Chen YI (2005) A study on the background and clustering seismicity in the taiwan region by using a point process model. J Geophys Res 110:B05S13

Zhuang J, Mateu J (2019) A semiparametric spatiotemporal hawkes-type point process model with periodic background for crime data. J R Stat Soc Ser A Stat Soc 182(3):919–942

Zhuang J, Ogata Y (2006) Properties of the probability distribution associated with the largest event in an earthquake cluster and their implications to foreshocks. Phys Rev E 73:046134

Zhuang J, Ogata Y, Vere-Jones D (2002) Stochastic declustering of space-time earthquake occurrences. J Am Stat Assoc 97(458):369–380

Zhuang J, Ogata Y (2004) Analyzing earthquake clustering features by using stochastic reconstruction. J Geophys Res 109(3):B05301

Zhuang J, Ogata Y (2006) Diagnostic analysis of space-time branching processes for earthquakes. In: Baddeley A, Gregori P, Mateu J, Stoica R, Stoyan D (eds) Case Studies in Spatial Point Process Modeling, Lecture Notes in Statistics, vol 185. New York, NY, Springer, pp 185–203

Zhuang J, Werner MJ, Harte DS (2013) Stability of earthquake clustering models: Criticality and branching ratios. Phys Rev E 88(6):062109

Zipkin JR, Schoenberg FP, Coronges K, Bertozzi AL (2015) Point-process models of social network interactions: Parameter estimation and missing data recovery. EurJournal of Applied Mathematics. FirstView, pp 1–28

Zuo S, Jiang H, Li Z, Zhao T, Zha H (2020) Proceedings of the 37th international conference on machine learning. In: Proceedings of the 37th international conference on machine learning, volume PMLR 119, pp 11692–11702