Japanese Journal of Statistics and Data Science
https://doi.org/10.1007/s42081-019-00060-0

ORIGINAL PAPER

Spatial statistics

Estimation, diagnostics, and extensions of nonparametric Hawkes processes with kernel functions

Jiancang Zhuang[1,2,3]

Received: 1 July 2019 / Accepted: 14 October 2019

Abstract
The Hawkes self-exciting model has become one of the most popular point-process models in many research areas in the natural and social sciences because of its capacity for investigating the clustering effect and positive interactions among individual events/particles. This article discusses a general nonparametric framework for the estimation, extensions, and post-estimation diagnostics of Hawkes models, in which we use the kernel functions as the basic smoothing tool.

Keywords Hawkes process · Kernel estimate · ETAS model · Earthquake · Crime · Expectation–maximization algorithm · Stochastic reconstruction

1 Introduction

Analyzing time series data has a core role in analyzing data series with an evolutionary characteristic. However, when we investigate the underlying processes at the microscale, most are continuous processes, point processes, or a mixture. For example, sales at a shop are not daily incomes, but a process of each trade, which includes trading times, amounts, and the types of goods. Point process models are common in research as a natural tool to model the patterns of discrete events that occur in a continuous space, time, or a space–time domain, such as urban fires, wild forest fires, crimes, earthquakes, diseases, tree locations, animal locations, communication network failures, and so on.

Depending on the type of the domain in which the events occur, researchers classify point into two classes: spatial point processes and spatiotemporal/temporal point processes. The difference between these two types of models is that the latter have a

✉ Jiancang Zhuang
zhuangjc@ism.ac.jp

[1] Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan

[2] Department of Statistical Science, The Graduate University for Advance Studies (SOKENDAI), 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan

[3] London Mathematical Laboratory, 8 Margravine Gardens, London W6-8RH, UK

Published online: 04 November 2019

special evolutionary time axis on which researchers can sort events according to their chronological order, and share many common features with time series sequences. Spatial point processes do not have such evolutionary direction and are usually regarded as a permanent pattern of particle locations or a snapshot of a spatiotemporal point process. Spatial point processes are usually modeled using the moment intensities and the Papangelou intensity (e.g., Møller and Waagepetersen 2003). When a property or a characteristic is attached to each event, such as the magnitude of an earthquake or the burned area of a wild fire, the point process is then called a marked point process.

Among the different types of point processes, clustering point processes attracted much interest from mathematicians and statisticians. Typical clustering processes include the Neyman–Scott process (Neyman and Scott 1953, 1958), which has been used to describe the distribution of locations of galaxies in the universe, and the Bartlett–Lewis process to model the rainfall process (Bartlett 1963; Lewis 1964). Many spatiotemporal/temporal clustering point processes can be categorized as a Hawkes self-exciting process (Hawkes 1971a, b; Hawkes and Oakes 1974). In short, a Hawkes process consists of a series of discrete events that each stem from one of two subprocesses: the background subprocess or the clustering subprocess. The former is considered a Poisson process, which can be inhomogeneous in space and/or nonstationary in time, while the latter consists of events from the exciting effect of all of the events that occurred in the past. Equivalently, each event, whether it is a background event or an excited event, triggers ("encourages") the occurrence of the future events according to some probability rules.

The Hawkes excitating model has become one of the most popular models in point process data analysis in both natural and social sciences because of its capacity to investigate the clustering effect (positive interactions) among individual events/particles, and to, thus, help determine the potential causal relationship among them. Nowadays, due to the rapid development of observation and data-storage technologies, big data is also a hot topic in point-process data analysis. With many sequences (datasets) or a long sequence (dataset) containing a huge number of discrete events, a quick tool or general framework to quantify and forecast the clustering or the triggering effect among events is desirable. The Hawkes process model fits this purpose.

In seismological application, researchers use a special form of the Hawkes process, called the epidemic-type aftershock sequence (ETAS) model (e.g., Ogata 1988, 1998; Zhuang et al. 2002, 2004; Console et al. 2003; Helmstetter et al. 2003; Lombardi et al. 2010; Guo et al. 2015) to evaluate the probabilities of future earthquakes as well as analyzing the characteristics of seismicity. It has been adopted by many research institutions or governmental agencies in the United States, Switzerland, Italy, New Zealand, Japan, China, and so on. (Schorlemmer et al. 2018). The ETAS model is now accepted as the standard model for describing earthquake clusters (e.g., Schorlemmer et al. 2018; Huang et al. 2016). Such a model is also used in crime data analysis (see, e.g., Mohler et al. 2011) and in economics, where studies show that the interaction between prices has epidemic features (e.g., Bacry and Muzy 2014). In recent years, researchers applied this model for a data analysis of terrorists' behavior (e.g, Tench et al. 2016), interactions in social networks (e.g., Zipkin et al. 2015), and genomes or neuronal activities (e.g., Truccolo et al. 2005), among others. In all of these areas, a big portion of the theories and methodologies were originally developed using studies of earthquakes as an outcome of studies on the ETAS model. The second biggest source is crime data analysis. The applications of the Hawkes process in other areas are mainly for parameter estimation and results explanation.

The core of spatiotemporal point-process models is the conditional intensity, which gives the expectation of the number of events occurring in a unit time–space range in the near future, under the condition that we know the history of previous events in the process, and/or the history of one or more relevant processes, up to the current time, but not including it. Starting from the conditional intensity, we can conduct a parameter estimation, simulation, forecast, and even control. Many powerful tools associated with the conditional intensity function have been developed for the Hawkes process, such as stochastic de-clustering, stochastic reconstruction, the expectation–maximization (EM) algorithm, first- and higher-order residuals, and Bayesian analysis, as well as the theories associated with asymptotic properties (see a review by Reinhart 2018).

This study focuses on the use of kernel functions to solve the estimation problem of this type of point processes. With these solutions, we provide a ready-to-use tool to perform modeling, analysis, and forecasting for different point-process data in different application areas by using the Hawkes-type point process. In the following, Sect. 2 provides the basic concepts and formulations of the Hawkes process and its variations. Section 3 describes the estimation methods related to parametric Hawkes Models, including the maximum likelihood estimate (MLE), (EM) algorithm, and stochastic declustering. Section 4 introduces the nonparametric kernel estimates of the nonparametric background rate and clustering response components, and Sect. 5 uses two examples to explain how to extend existing models in a data-driven manner.

2 Model and methodology

2.1 Hawkes process

We can determine a point process by its conditional intensity (e.g., Daley and Vere-Jones 2003; Zhuang 2015). For a temporal point process with no overlapping events (a simple temporal point process), the conditional intensity is

$$\lambda(t) = \lim_{\Delta \downarrow 0} \frac{1}{\Delta} \Pr\left\{N([t, t+\Delta)) = 1 \mid \mathcal{H}_t\right\}, \tag{1}$$

where $\mathcal{H}_t$ denotes the σ-algebra generated by the observational N before time t, but not including t. For any measurable set $D \in \mathbb{R}$,

$$\mathbf{E}\left[\int_D f(s)\, N(\mathrm{d}s) - \int_D f(s)\, \lambda(s)\, \mathrm{d}s\right] = 0, \tag{2}$$

where *f*(*t*) is a predictable function; that is, its value at *t* is determined by the observation history of *N* before *t*. This equality and its spatiotemporal (high dimensional) versions are the key to the theories and methods summarized in this article.

The Hawkes process describes the stochastic excitations among a series of events that occur in a continuous time domain or in a spatiotemporal domain. A temporal Hawkes process, supposing its realization $N = \{t_i : i \in \mathbb{Z}\}$ with $\mathbb{Z}$ being the set of all integers, has a conditional intensity in the form

$$\lambda(t) = \mu + \int_{-\infty}^{t} g(t-u)N(\mathrm{d}u) = \mu + \sum_{i:t_i<t} g(t-t_i), \tag{3}$$

where μ is the rate of occurrence of spontaneous events (also called background events or immigrants), and $g(t)$ is the rate of occurrence of the direct offspring generated by an event occurring at 0.

The criticality parameter, which is the average number of direct offspring per ancestor, is

$$\rho = \int_{0}^{\infty} g(u)\,\mathrm{d}u. \tag{4}$$

A stable and stationary Hawkes process requires $\rho < 1$. Otherwise, the rate of occurrence grows to infinity with time. If $\rho < 1$, then this parameter is identical to the branching ratio, which is the proportion of non-spontaneous events in the whole process. In general, these two quantities are different (see Zhuang et al. 2013 for details).

We can extend the Hawkes process easily to the spatiotemporal version

$$\lambda(t,x) = \mu(x) + \int_{(-\infty,t)\times\mathbb{R}^d} g(t-s, x-u)\,\mathrm{d}N(\mathrm{d}s\times\mathrm{d}u), \tag{5}$$

where x denotes the location in the space of $\mathbb{R}^d$. We can also generalize it the multivariate case where, if we have K types events in total, then each type has a conditional intensity of

$$\lambda_k(t,x) = \mu_k(x) + \sum_{\ell=1}^{K} \int_{(-\infty,t)\times\mathbb{R}^d} g_{k\ell}(t-s, x-u)\,\mathrm{d}N_\ell(\mathrm{d}s\times\mathrm{d}u), \tag{6}$$

for $k = 1, \ldots, K$, where $\mu_k(x)$ represents the rate of occurrence of spontaneous events (also called background) for type-k events, and $g_{k\ell}(t-s, x-u)$ is the rate of occurrence of type-k events excited by a type-ℓ event at (s, u).

We can also extend the space–time Hawkes process to cases of marked processes easily:

$$\lambda(t,m) = \mu(x,m) + \int_{(-\infty,t)\times\mathbb{R}^d\times\mathbb{M}} g(t-s, x-u)f(m \mid m')\,\mathrm{d}N(\mathrm{d}s\times\mathrm{d}u\times\mathrm{d}m'), \tag{7}$$

where x and m denote the location in the space of $\mathbb{R}^d$ and the mark in the space of $\mathbb{M}$, respectively, and $f(m \mid m')$ gives the p.d.f. for the magnitudes of direct offspring from an event of magnitude m'. We can regard the multivariate case simply as a marked point process in which the mark takes only a finite number of values. In the above, we assume that the marks of triggered events are location- and time independent.

Linlin model The earliest version of the Hawkes mode used by Hawkes and Oakes (1974) had the following self- and mutually exciting process

$$\lambda(t) = \mu + \int_0^{t_-} g(t-s)\, N(\mathrm{d}s) + \int_0^{t_-} h(t-s)\, X(\mathrm{d}s), \tag{8}$$

where

$$g(t) = \sum_{k=0}^{K} a_k\, t^k\, \mathrm{e}^{-ct} \quad \text{and} \quad h(t) = \sum_{k=0}^{L} b_k\, t^k\, \mathrm{e}^{-ct}, \tag{9}$$

K and L are two given non-negative integers, and $X(t)$ denotes the external process that can trigger events in N, which can be a point process, a continuous process, or mixture of both. Its name, Linlin, came from Ogata's FORTRAN program named "Linlin.f", meaning a linear response effect for both internal and external responses. Ogata and Akaike (1982) use this model to investigate the temporal clustering patterns of seismicity in different regions, and the correlation of seismicity among different regions. A technical problem is to keep $g(t)$ and $h(t)$ positive during the optimization when fitting this model to the data.

The space–time Epidemic-Type Aftershock Sequence (ETAS) model The spatiotemporal ETAS model has been used widely to describe the clustering features of earthquakes in space and time (see Ogata 1998; Zhuang et al. 2002, 2004, 2005; Zhuang and Ogata 2006; Ogata and Zhuang 2006; Console et al. 2003; Sornette and Werner 2005a, b; Helmstetter et al. 2005). The conditional intensity of this model is

$$\lambda(t,x,m) = s(m)\left[\mu(x) + \sum_{j:t_j<t} \kappa(m_j)\, g(t-t_j) f(x-x_j, m_j)\right], \tag{10}$$

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

$$s(m) = \beta \exp[-\beta(m-m_c)], \quad m \ge m_c,$$

represents the probability density of the earthquake magnitude, where m_c is the magnitude threshold of the earthquake, and

$$\kappa(m) = A\, e^{\alpha(m-m_c)}, \quad m \ge m_c, \tag{11}$$

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

$$g(t) = \frac{p-1}{c}\left(1 + \frac{t}{c}\right)^{-p}, \quad t > 0, \tag{12}$$

is the probability density function of the length of the time interval between a child and its parent, and

$$f(x, m) = \frac{q-1}{\pi D e^{\gamma(m-m_c)}}\left(1 + \frac{||x||^2}{D e^{\gamma(m-m_c)}}\right)^{-q} \tag{13}$$

is the probability density function of the relative locations between the parent and children, where m_c is the magnitude threshold.

Mohler et al.'s model for break-in burglary data Mohler et al. (2011) analyze break-in burglary data from the Los Angeles Police Department. Their dataset consisted of 5,376 reported residential burglaries in an 18 km × 18 km region of San Fernando Valley, Los Angeles during 2004–2005. They use a model with a conditional intensity of

$$\lambda(t, x) = \nu(t)\mu(x) + \sum_{k: t_k < t} g(t - t_k, x - x_k) \tag{14}$$

Mohler et al. (2011) assume that the background rate is a function of space and time and use kernel functions to smooth the estimates of both μ and g. In Mohler et al. (2011), $\nu(t)$, $\mu(x, y)$, and $g(t, x, y)$ are all nonparametric functions. Later, Mohler (2014) use an exponential density and a Gaussian density for the decay of occurrence rate of triggered events in time and in 2D space.

Epidemic forecasting In modeling and forecasting routinely collected invasive meningococcal disease (IMD), Meyer et al. (2012, 2016) and Meyer and Held (2014) use the following model

$$\begin{aligned}\lambda(t, x, y) =& \rho_{\tau(t),\xi(x,y)} \exp\left(\beta_0(\kappa) + \boldsymbol{\beta}' \mathbf{z}_{\tau(t),\xi(x,y)}\right) \\ &+ \sum_{j \in I^*(t,x,y)} e^{\eta_j} g(t - t_j \mid \kappa_j) f(x - x_i, y - y_i \mid \kappa_j),\end{aligned} \tag{15}$$

where $\rho_{k,l}$ is the intensity offset in the spatiotemporal grid (k, l); $(\tau(t), \xi(x, y))$ is the grid index in which t, (x, y) is located; $\mathbf{z}_{\tau(t),\xi(x,y)}$ is a linear predictor of endemic covariates on the grid that contains $(t, (x, y))$; η_j is a predictor attached to each infected individual; g and f are the temporal and spatial response functions, respectively; and $I^*(t, x, y) = \{j : t - \epsilon \geq t_j < t \wedge ||(x, y) - (x_j, y_j)|| \leq \delta\}$ with ϵ and δ being positive constants.

Social networks Fox et al. (2016) and Zipkin et al. (2015) use a multivariate Hawkes process to model the mail sent between pairs in a network of officers at the West Point military academy. The difference between these two studies is that the former uses the messages sent by the same sender as a component and the latter uses messages between each pair of officers as a component.

3 Parametric estimation

We can classify the forms of Hawkes models into three categories: parametric, nonparametric, and semiparametric. The parametric model can be estimated through the MLE method and the EM algorithm.

3.1 Likelihood function and MLE

Given the observation data of a spatiotemporal parametric Hawkes model in a space–time window $S \times T$, the likelihood function can be written as

$$\log L(\cdot;\theta) = \sum_{i:\,(t_i,x_i)\in S\times T} \log \lambda(t,x;\theta) - \int_T \int_S \lambda(t,x;\theta)\,\mathrm{d}x\,\mathrm{d}t, \tag{16}$$

where $\lambda(t,x;\theta)$ is the conditional intensity of the process and θ denotes the parameter vector in the model (Daley and Vere-Jones 2003). We can estimate the model parameters, supposing that they are regular, by maximizing the likelihood above; that is,

$$\hat{\theta} = \arg_{\theta} \max \log L(\cdot;\theta). \tag{17}$$

Rathbun (1996) discusses the asymptotic normality of the MLE for point processes.

3.2 Decomposing and reassembling the events: stochastic declustering

Consider a Hawkes process with conditional intensity

$$\lambda(t,x) = \mu(t,x) + \sum_{k:t_k<t} g(t-t_k, x-x_k), \tag{18}$$

where $\mu(t,x)$ is the background rate, which is different from the corresponding term in (5) as it can be time dependent, and $g(t,\, x)$ is the rate of occurrence triggered by an event at time 0 and the location at the origin.

The probability that an event, say j, is a background event; that is, the *background probability*, is

$$\varphi_j = \Pr\{\text{Event}\, j \text{ is a background event}\} = \frac{\mu(t_j,x_j)}{\lambda(t_j,x_j)} \tag{19}$$

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

$$\rho_{ij} = \Pr\{\text{Event}\, j \text{ is triggered by } i\} = \frac{g(t_j-t_i, x_j-x_i)}{\lambda(t_j,x_j)}. \tag{20}$$

It is easy to see

$$\varphi_j + \sum_i \rho_{ij} = 1, \text{ for all } j, \tag{21}$$

which implies that an event is always either a background event or is triggered by a previous event. Another explanation for (19) and (20) is that, once an event, say j, occurs at (t, x), we can say that, at (t, x), we observed φ_j background events, and that for each $i = 1, \ldots, j-1$, event i triggers ρ_{ij} direct offspring at (t_j, x_j). In this way, we separate event j into background and offspring events from previous events (Zhuang et al. 2004).

We say that the above probabilities, φ_j, $j = 1, 2, \ldots, n$, are background probabilities because if we select each event j with probability φ_j, we can realize a Poisson process with rate $\mu(t, x)$. To prove this point, we need only to show that the compensator for the resulting process is rate $\mu(u, x)$. For any measurable set $B \in \mathbb{R}^d$,

$$\mathbf{E}\left[\iint_{[0,t]\times B} X(s,x)N(\mathrm{d}s \times \mathrm{d}x)\right] = \mathbf{E}\left[\iint_{[0,t]\times B} \mathbf{E}\left(X(s,x)N(\mathrm{d}s \times \mathrm{d}x) \mid \mathcal{H}_s\right)\right], \tag{22}$$

where $X(t, x)$ is a random field that takes values of 1 and 0 with probabilities $\frac{\mu(t,x)}{\lambda(t,x)}$ and $1 - \frac{\mu(t,x)}{\lambda(t,x)}$ at (t, x), respectively. Since $X(t, x)$ is independent of N conditional on $\mathcal{H}_t$ and $\frac{\mu(t,x)}{\lambda(t,x)}$ is a predictable function, then

$$\begin{aligned}
&\mathbf{E}\left(X(s,x)\,N(\mathrm{d}s \times \mathrm{d}x) \mid \mathcal{H}_s\right)\\
&\quad= \mathbf{E}\left[\mathbf{E}\left(X(s,x)\,N(\mathrm{d}s \times \mathrm{d}x) \mid \mathcal{H}_{s^+}\right) \mid \mathcal{H}_s\right]\\
&\quad= \mathbf{E}\left[\frac{\mu(s,x)}{\lambda(s,x)}\,N(\mathrm{d}s \times \mathrm{d}x) \mid \mathcal{H}_s\right]\\
&\quad= \frac{\mu(s,x)}{\lambda(s,x)}\,\lambda(s,x)\,\mathrm{d}s\,\mathrm{d}x\\
&\quad= \mu(s,x)\,\mathrm{d}s \times \mathrm{d}x.
\end{aligned} \tag{23}$$

In the above, $\mathcal{H}_{s^+} = \cap_{u>s}\mathcal{H}_u$ represents the history of N up to time s and including s. Substituting the above equation into (22), we have

$$\mathbf{E}\left[\iint_{[0,t]\times B} X(s,x)\,N(\mathrm{d}s \times \mathrm{d}x)\right] = \mathbf{E}\left[\int_{[0,t]}\int_B \mu(s,x)\,\mathrm{d}x\,\mathrm{d}s\right]. \tag{24}$$

That is, the resulting process has a compensator with a deterministic rate $\mu(t, x)$. Thus, it is a Poisson process.

3.3 Expectation–maximization algorithm

A direct use of the background and triggering probabilities is to construct an expectation–maximization (EM) algorithm (Veen and Schoenberg 2008; Li et al. 2019). First, we treat the whole process as a missing data problem. The complete observation for each event j is (t_j, x_j, η_j), where η_j takes a value 0 if it is a background event and i if it is triggered by event i. Thus, the complete likelihood for the whole process is

$$\begin{aligned} \log L_{\mathrm{cmplt}} = \sum_{j=1}^{n} \left[I(\eta_j = 0) \log \mu(t_j, x_j) + \sum_{i=1}^{j} I(\eta_j = i) \log \xi(t_j - t_i, x_j - x_i) \right] \\ - \int_T \int_S \lambda(t, x) \, \mathrm{d}t \, \mathrm{d}x \end{aligned} \tag{25}$$

The parameters can be estimated with the following EM algorithm:

E-Step For each step k, calculate $\varphi_j^{(k)}$ and $\rho_{ij}^{(k)}$ for $j = 1, 2, \ldots, n$ and $i = 1, 2, \ldots, j - 1$.

M-Step Maximize the expected log-likelihood:

$$\begin{aligned} \mathbf{E}\left[\log L_{\mathrm{cmplt}}^{(k+1)}\right] = \sum_{j=1}^{n} \left[\varphi_j^{(k)} \log \mu(t_j, x_j) + \sum_{i=1}^{j} \rho_{ij}^{(k)} \log \xi(t_j - t_i, x_j - x_i) \right] \\ - \int_T \int_S \lambda(t, x) \, \mathrm{d}x \, \mathrm{d}t \end{aligned} \tag{26}$$

to obtain the model parameters.

The computational complexity of this algorithm is the same as the original MLE method, proportional to n^2, where n is the number of events in the process.

4 Kernel estimates

4.1 The nonparametric background rate

The EM algorithm is almost easy to implement when μ and ξ are both parametric functions, with only some unknowns parameters. However, in many cases, the explicit form of the background rate μ is usually unknown. Veen and Schoenberg (2008) divide the whole study region into several subregions, each with a constant background rate; that is, they assumed that the background rate was a 2D piecewise function, with all values for the background rate in each subregion being parameters to estimate. Using the MLE method, the background rate μ_k in each subregion can be estimated as follows:

Suppose the background rate

$$\mu(t, x, y) = \mu_k, \quad \text{when } (t, x, y) \in S_k, \quad k = 1, 2, \ldots, K, \tag{27}$$

where K is the total number of subregions and the whole area $A = \cup_{k=1}^{K} S_k$, $S_k \cap S_l = \emptyset$ for $1 \le k \ne l \le K$. Then,

$$\frac{\partial \log L}{\partial \mu_k} = \sum_{i=1}^{N} \frac{I((t_i, x_i, y_i) \in S_k)}{\lambda(t_i, x_i, y_i)} - S_k; \tag{28}$$

and

$$\left.\frac{\partial \log L}{\partial \mu_k}\right|_{\mu_k=\hat{\mu}_k} = 0, \tag{29}$$

yield

$$\sum_{i=1}^{N} \frac{I((t_i, x_i, y_i) \in S_k)}{\hat{\lambda}(t_i, x_i, y_i)} = S_k. \tag{30}$$

Multiplying both sides by $\hat{\mu}_k$ and rearranging the terms, we have

$$\hat{\mu}_k = \frac{1}{S_k} \sum_{i=1}^{N} \frac{\hat{\mu}_k I((t_i, x_i, y_i) \in S_k)}{\hat{\lambda}(t_i, x_i, y_i)} = \frac{1}{S_k} \sum_{i=1}^{N} \hat{\varphi}_i\, I((t_i, x_i, y_i) \in S_k). \tag{31}$$

As discussed in Sect. 3.2, φ_j and ρ_{ij}, $i = 1, 2, \ldots, j-1$, quantify how event j is sliced into background and offspring from previous events. Equation (31) in fact provides a histogram estimator of the background rate function in an iterative manner.

We can modify such histogram estimates easily into kernel estimates. For example, Zhuang et al. (2002, 2004) use a weighted kernel function estimate of $\mu(\cdot, \cdot)$ in combination with variable bandwidths

$$\hat{\mu}(x) = \frac{1}{T} \sum_{j} \varphi_j\, Z(|x - x_j|; h_j), \tag{32}$$

where h_j is the bandwidth for the kernel function corresponding to event j, equal to its distance to the n_pth closest neighboring event, $n_p = 2$–15. When μ is also time dependent, when we can also estimate it using a weighted kernel estimation, for example, as follows

$$\hat{\mu}(t, x) = \sum \varphi_j Z_{h_j}(x - x_j)\, Z^{(\mathrm{t})}_{h_t}(t - t_j), \tag{33}$$

where $Z^{(\mathrm{t})}_{h_t}(\cdot)$ is the temporal kernel with bandwidth h_t.

4.2 The nonparametric triggering term

We can estimate the triggering term $g(\cdot, \cdot)$ by

$$\hat{g}(t, x) = \frac{\sum_{i,j} \rho_{ij}\, I(|t_j - t_i - t| < \delta_t)\, I(|x_j - x_i - x| < \delta_x)}{4\delta_t\, \delta_x\, \sum_{i,j} \rho_{ij}}, \tag{34}$$

where the denominator is for normalizing purposes, Z_h is the Gaussian kernel with bandwidth h, and ρ_{ij} is as defined in (20).

We can verify the estimator above in the following way. First, the spatiotemporal version of (2),

$$\mathbf{E}\left[\int_{[T_1,T_2]\times S} f(t,x)\,\mathrm{d}N(t,x)\right] = \mathbf{E}\left[\int_{T_1}^{T_2}\int_S f(t,x)\,\lambda(t,x)\,\mathrm{d}t\,\mathrm{d}x\right], \tag{35}$$

holds for any predictable process $f(t, x)$, any given time interval $[T_1, T_2]$, and any area S, provided that the integral on either side exists, or that f is nonnegative. Second, let

$$\varrho(s_1,u_1,s_2,u_2) = \begin{cases} g(s_2 - s_1, u_2 - u_1)/\lambda(s_2,u_2), & s_2 \geq s_1; \\ 0, & \text{otherwise.} \end{cases} \tag{36}$$

Regarding t, x, t_i, and x_i as fixed and substituting $f(t_i,x_i,s_2,u_2) = \varrho(t_i,x_i,s_2,u_2,)\,I(s_2 - t_i \in [t-\Delta_t, t+\Delta_t], |u_2 - x_i - x| < \delta_x)$ as a predictable function of s_2 and u_2 into (35) yields

$$\begin{aligned} &\sum_j \varrho(t_i,x_i,t_j,x_j)\,I(t_j - t_i \in [t-\Delta_t, t+\Delta_t], |x_j - x_i - x| < \delta_u) \\ &\approx \int_{T_1}^{T_2}\iint_S \varrho(t_i,x_i,s,u)\,\mathbf{1}(s - t_i \in [t-\Delta_t, t+\Delta_t], |x - x_i - u| < \delta_x)\,\lambda(s,u)\,\mathrm{d}s\,\mathrm{d}u \\ &\approx 2\,g(t,x)\,\Delta_t\,|B(x,\delta_x)| \\ &\propto g(t,x) \end{aligned} \tag{37}$$

if the area of S and the length of $[T_1, T_2]$ are sufficiently large, where δ_x is a small positive real number and $|B(x,\delta_x)|$ is the volume of the ball centered at x with a radius of δ_x. Thus, we can estimate $g(t)$ by

$$\hat{g}(t,x) \propto \sum_{i,j} \rho_{ij}\,I(t_j - t_i \in [t-\Delta_t, t+\Delta_t], |x_j - x_i - x| < \delta_x) \tag{38}$$

where

$$\rho_{ij} = g(t_j - t_i, x_j - x_i)/\lambda(t_j,x_j), \quad i < j. \tag{39}$$

If $g(t, x)$ is separable; that is, $g(t,x) = g_1(t)\,g_2(x)$, then

$$\hat{g}_1(t) \propto \sum_{i,j} \rho_{ij}\,I(t_j - t_i \in [t-\delta_t, t+\delta_t]) \tag{40}$$

$$\hat{g}_2(x) \propto \sum_{i,j} \rho_{ij}\,I(|x_j - x_i - x| < \delta_x), \tag{41}$$

where Δ_x and Δ_y are small positive numbers. These estimates can be revised into their kernel function version correspondingly.

4.3 When both the background rate and triggering term are nonparametric

In the above, when estimating $\mu(t,x)$ and $g(t, x)$, we need to know φ_i and ρ_{ij}, and when estimating φ_i and ρ_{ij}, we need to know μ and g. We can resolve this loop using an iterative algorithm. Given an observed process of events $\{(t_i, x_i) : i = 1, \dots, n\}$ in a time–space window $T \times S$, by guessing some initial value of μ and g, we obtain φ_i and ρ_{ij} for all possible i, j. Then, we estimate the background rate μ and each component in the clustering part g using φ_i and ρ_{ij} with some nonparametric methods, such as kernel estimates or histograms. Once we update μ and g, we go back to the step of calculating $\varphi_{\cdot}$ and $\rho_{\cdot\cdot}$, or stop if convergence is reached. In summary, the iterative estimation procedure includes the following three integrants:

Algorithm 1

1. **Stochastic declustering (Expectation).** Calculate the background probability and triggering probabilities.
2. **Reconstruction (Maximization I).** Estimate the nonparametric function in the model using nonparametric methods such as kernel functions.
3. **Parametrization (Maximization II).** Use the MLE method or EM algorithm to estimate the parameters in the parametric functions.

In an iterative algorithm such as the one above, the feedback should be negative. However, we cannot be assured of negative feedback if we use a nonparametric estimate. For example, consider that we use only

$$\hat{\mu}(t) = \sum_i \varphi_i Z_{h_1}(t - t_i) \quad \text{and} \quad \hat{g}(t) = \sum_{i,j} \rho_{ij} Z_{h_2}(t_j - t_i - t). \tag{42}$$

to estimate μ and g in the conditional intensity

$$\lambda(t) = \mu(t) + \sum_{i: t_i < t} g(t - t_i). \tag{43}$$

If all ρ_{ij}, $j = 1, \dots, N$ and $i < j$ increase at some time, then we obtain a large $\hat{g}(t)$, which yields a larger ρ_{ij} in the next step. This positive feedback finally yields a trivial solution with $\hat{\mu}(t) = 0$. To avoid this positive feedback, Zhuang et al. (2002) and Zhuang and Mateu (2019) introduce relaxation coefficients to prevent positive feedback. Instead of (43), they use

$$\lambda(t) = \nu\mu(t) + A \sum_{i: t_i < t} g(t - t_i) \tag{44}$$

as the conditional intensity function, and the estimates for μ and g become

$$\hat{\mu}(t) \propto \sum_i \varphi_i Z_{h_1}(t - t_i) \quad \text{and} \quad \hat{g}(t) \propto \sum_{i,j} \rho_{ij} Z_{h_2}(t_j - t_i - t). \tag{45}$$

with restrictions $\int_0^T \hat{\mu}(t)\,\mathrm{d}t/T = 1$ and $\int_0^\infty \hat{g}(t)\,\mathrm{d}t = 1$. In the above equations, the parameters ν and A are the relaxation coefficients and are estimated by the MLE or EM method, as given in Sects. 3.1 and 3.3.

Introducing the relaxation coefficients changes Algorithm 1 into the following new algorithm.

Algorithm 2

1. **Stochastic declustering (Expectation).** Calculate the background probability and triggering probabilities.
2. **Reconstruction (Maximization I).** Estimate the nonparametric function in the model using nonparametric methods such as kernel functions.
3. **Parametrization (Maximization II).** Use the MLE method or EM algorithm to estimate the parameters in the parametric functions and the *relaxation coefficients* for the nonparametric functions.

4.4 Choice of bandwidth

Bandwidth selection is always an unavoidable topic with kernel estimation. Many previous works discuss fixed bandwidth, such as Silverman's rule of thumb in Silverman (1986) and Scott (2009). Alternatively and technically, this can be done with the cross-validation method (e.g., Hall and Wehrly 1991) or using the forward predictive likelihood (e.g., Chiodi and Adelfio 2011). In principle, the bandwidth should be selected according to data resolution, which is at the order of 1–10 times of the nearest neighboring distance. Zhuang (2011) reduces n_p to 3 or 4. Xiong et al. (2019) compare two variable bandwidth kernel estimates, with their optimal bandwidth parameters obtained using cross-validation, with two other nonparametric methods, a Bayesian smoothing procedure on a tessellation configuration with smoothness prior, and a newly proposed incomplete centroidal Voronoi tessellation method. They find that the performance differences among the methods are marginal in estimating the seismicity rate.

4.5 Boundary effect correction

The second problem is boundary correction. In most cases, observations lie within some specific range, while the kernel function distributes the mass of an event over a much larger, or even infinite, range. The observation range makes the total weight for an event not equal to 1 and the estimates for the values near the boundary depend on the shape of the boundary. In this study, we adopt the weight-based correction (Hall and Turlach 1999), which can also be called the truncated kernel function. For each event, we use a truncated density $Z_h(\cdot - x_i)/\int_S Z_h(u - x_i)\mathrm{d}u$ instead of $Z_h(\cdot - x_i)$ as the kernel for event i, for which S is the support range of observation.

5 Extending the Hawkes model

We should note that a useful model reflects only partial information about the observation. One of the tasks in statistical analysis is to extend a model such that it includes more predictable information. This requires not only for statistical analysis, but also to obtain a good understanding of the observed phenomena. In this section, we provide two examples to illustrate how to extend the Hawkes model to accommodate the need for better modeling the data.

5.1 Example 1: Building more physics into the model—incorporating earthquake source geometry into seismicity modeling

In the space–time ETAS model, all of the earthquake events are regarded as a point in space–time–magnitude domain. In fact, the rupture of each earthquake has a spatial extension on the earthquake fault, from several kilometers for a M5.5 event up to about 500 km for a M8.0 event. When we regard the focal source of an earthquake as a point and describe the spatial response to the triggering effect an isotropic function of distance from the epicenter of the parent location, as in (13), biased results will be obtained in data analysis, since aftershocks are usually distributed along the rupture fault, especially for large earthquakes. For example, Hainzl et al. (2008) discuss the impact of the rupture extension of the 1992 M7.3 Landers, California earthquake on parameter estimations of the point-process model by comparing the results from space–time and purely temporal ETAS models. They find that ignoring the rupture extensions of earthquakes and assuming an isotropic aftershock response could lead to a significant bias in the parameter estimations, especially an underestimated α value.

Several researchers attempted to correct such biases. As early as in 1998, Ogata (1998) suggested that the aftershock rate is spatially elliptically distributed and that the centroid of the ellipsoid formed from the aftershock cluster should be used as the location of the main shock instead of the initial fracture point. Considering such anisotropy, (13) was replaced by a general bivariate Guassian density and takes different values for different clusters. The aftershock cluster from each main shock was determined by the MBC algorithm (Ogata et al. 1995). The modified elliptic distribution model outputs better results for observed seismicity. However, the MBC algorithm is quite subjective and empirical in terms of dividing aftershock clusters, and the location of the main shock is still a point, either the epicenter or the centroid of the aftershock epicenters. Marsan and Lengliné (2010) and Bach and Hainzl (2012) propose an alternative treatment for the anisotropic response as a function of the distance to earthquake fault, instead of a function of the epicenter distance.

Instead of regarding large earthquakes to have point sources, Guo et al. (2015, 2017) develop a finite-source ETAS model to incorporate the spatial extensions of their ruptures. Each earthquake rupture consists of many small patches, and each patch triggers its own aftershocks isotropically and independently as a usual mainshock. The superposition of triggering effects from all patches produces an anisotropic pattern of the aftershock locations, mainly distributed along the rupturing fault. In mathematics, the spatial response of the production of direct offspring to a large earthquake with source body S_i is

$$f(x;S_i) = \frac{\int_s f(x-u)\, \tau_i(x)\, \ell(\mathrm{d}x)}{\int_S \tau_i(x)\, \ell(\mathrm{d}x)}, \tag{46}$$

where $\tau_i(u)$ is the productivity density at location u in the focal zone and the numerator is used for normalization. The model parameters, the unobserved fault geometry, and the background rate are estimated simultaneously through an EM-type iterative algorithm similar to Algorithm 1. We can use their treatment to invert the earthquake fault from seismicity.

In the estimation, the productivity density $\tau(u)$ is also implemented using stochastic reconstruction, which has a histogram version of estimation of

$$\tau_i(u) \propto \sum_k \tau_{ijk} I(u \in C_j), \tag{47}$$

where

$$\tau_{ijk} = \frac{\kappa(m_i)\, g(t_k - t_i)\, \int_{C_j} f(x_k - u; S_i)\tau_i(u)\mathrm{d}u}{\lambda(t_k, x_k)\, \int_{S_i} \tau_i(u)\mathrm{d}u} \tag{48}$$

represents the probability that event k is triggered by patch C_j, which contains location u in the rupture of event i. Similar to the background rate μ, the function τ can also be estimated iteratively (Guo et al. 2017; Zhuang et al. 2019).

This model has been applied to seismicity in different regions, such as Southwestern China (Guo et al. 2015), Japan (Guo et al. 2017), and Italy (Zhuang et al. 2019). Figure 1 shows the surface projection of the spatial variation of the productivity density along the Nocia earthquake (2016-10-30 08:40 local time, 13.11°E, 42.83°N, Mw 6.5) rupture fault obtained by Zhuang et al. (2019). The main coseismic slip area is to the updip from the hypocenter. The major parts of the direct aftershocks are located on the north and south of the area with biggest coseismic slip. These patches of high productivity situated are close to but do not overlap with the area of the high coseismic slips. The feature has been observed for many large earthquakes.

5.2 Example 2: Complexity in the background rate: crime modeling

This example is related to the model used by Molher and others in modeling crime behavior. Studies using crime data based on point processes do not consider the periodic components in the background rate. Since criminals are also human beings, their behaviors should be controlled by their biological clock and could be influenced by periodic social activity (Felson and Boba 2010). Thus, studies should account for periodicity, for instance, daily periodicity and weekly periodicity, to build a more precise model.

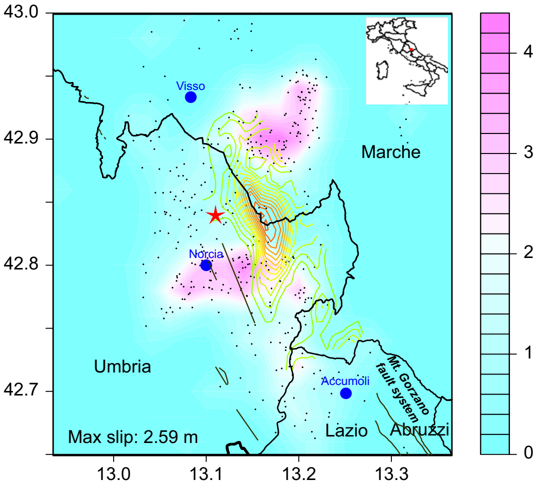

Fig. 1 (c.f. Zhuang et al. 2019) Comparison between the pattern of the productivity of direct offspring along the rupture areas (contour images) inferred by the finite-source ETAS model and coseismic slip (contour lines) for the Norcia earthquake (2016-10-30, Mw 6.5). The values of the coseismic slip from zero to the maximum for each event are indicated by contour lines from green to red in rainbow colors. The red stars represent the epicenter of the corresponding major earthquake, and the blue dots represent the locations of towns and cities in the area. The small black dots mark the locations of small events that occurred shortly after the corresponding major events. The traces of active faults are also plotted in black lines. The kinematic model adopted for the earthquake of Norcia (2016-10-30) is based on Chiaraluce et al. (2017)

Zhuang and Mateu (2019) use the following space–time point process model to describe the robbery data in Castollen, Spain, which they specify completely using a conditional intensity function

$$\lambda(t,x) = \mu_0\, \mu_{\mathrm{t}}(t)\, \mu_{\mathrm{d}}(t)\, \mu_{\mathrm{w}}(t)\, \mu_{\mathrm{b}}(x) + A \int_{-\infty}^{t} \int_{S} g_1(t-s)\, h(x-u)\, \mathrm{d}N(s,u), \quad (49)$$

where A and μ_0 are the relaxation coefficients to estimate, the average values of the trend term $\mu_{\mathrm{t}}(t)$, daily periodicity $\mu_{\mathrm{d}}(t)$, weekly periodicity $\mu_{\mathrm{w}}(t)$, and spatial background heterogeneity $\mu_{\mathrm{b}}(x,y)$ are all normalized to 1, and the temporal response g_1 and the spatial response g_2 are normalized as p.d.f.s.

Though we cannot estimate the periodic components of the background rate in our model formulation directly using the stochastic reconstruction method, we can solve this problem using the spatiotemporal version of (2). Given a realization of the point process $\{(t_i, x_i) : i = 1, 2, \ldots, n\}$ in a time–space range $[T_1, T_2] \times S$, where t and $x = (x^{(1)}, x^{(2)})$ denote time and location, respectively, the long-term trend term $\mu_{\mathrm{t}}(t)$ in the background component can be reconstructed in the following way. Let

$$w^{(\mathrm{t})}(t,x) = \mu_{\mathrm{t}}(t)\, \mu_{\mathrm{b}}(x) / \lambda(t,x)$$

and $f(t,x) = w^{(\mathrm{t})}(t,x)$, and substitute f into (35). Then, assuming that μ_{t} is smooth enough,

$$\begin{aligned}
&\sum_i w^{(\mathrm{t})}(t_i, x_i)\,\mathbf{1}(t_i \in [t-\Delta_t,\ t+\Delta_t]) \\
&\quad\approx \int_{T_1}^{T_2} \int_S w^{(\mathrm{t})}(s,x)\lambda(s,x)\,\mathbf{1}(s \in [t-\Delta_t, t+\Delta_t])\,\mathrm{d}s\,\mathrm{d}x \\
&\quad= \int_{t-\Delta_t}^{t+\Delta_t} \mu_{\mathrm{t}}(s)\,\mathrm{d}s \int_S \mu_{\mathrm{b}}(x)\,\mathrm{d}x \\
&\quad\propto \int_{t-\Delta_t}^{t+\Delta_t} \mu_{\mathrm{t}}(s)\,\mathrm{d}s \\
&\quad\approx 2\mu_{\mathrm{t}}(t)\,\Delta_t,
\end{aligned} \tag{50}$$

where Δ_t is a small positive number. For ease of writing, define

$$w_i^{(\mathrm{t})} = \mu_{\mathrm{t}}(t_i)\,\mu_{\mathrm{b}}(x_i, y_i)/\lambda(t_i, x_i, y_i); \tag{51}$$

Then,

$$\hat{\mu}_{\mathrm{t}}(t) \propto \sum_i w_i^{(\mathrm{t})}\,\mathbf{1}(t_i \in [t-\Delta_t,\ t+\Delta_t]). \tag{52}$$

Similarly, we can reconstruct the other components in the background rate as follows

$$\hat{\mu}_{\mathrm{d}}(t) \propto \sum_i w_i^{(\mathrm{d})}\, I\left(t_i \in \bigcup_{k\in\mathbb{Z}} [t+k-\Delta_t,\ t+k+\Delta_t] \right), \quad t \in [0,1], \tag{53}$$

$$\hat{\mu}_{\mathrm{w}}(t) \propto \sum_i w_i^{(\mathrm{w})}\, I\left(t_i \in \bigcup_{k\in\mathbb{Z}} [t+7k-\Delta_t,\ t+7k+\Delta_t] \right), \quad t \in [0,7], \tag{54}$$

and

$$\hat{\mu}_{\mathrm{b}}(x,y) \propto \sum_i \varphi_i\,\mathbf{1}(x_i \in [x-\Delta_x,\ x+\Delta_x])\,\mathbf{1}(y_i \in [y-\Delta_y,\ x+\Delta_y]), \tag{55}$$

where

$$w_i^{(\mathrm{d})} = \mu_{\mathrm{d}}(t_i)\,\mu_{\mathrm{b}}(x_i, y_i)/\lambda(t_i, x_i, y_i), \tag{56}$$

$$w_i^{(\mathrm{w})} = \mu_{\mathrm{w}}(t_i)\,\mu_{\mathrm{b}}(x_i, y_i)/\lambda(t_i, x_i, y_i), \tag{57}$$

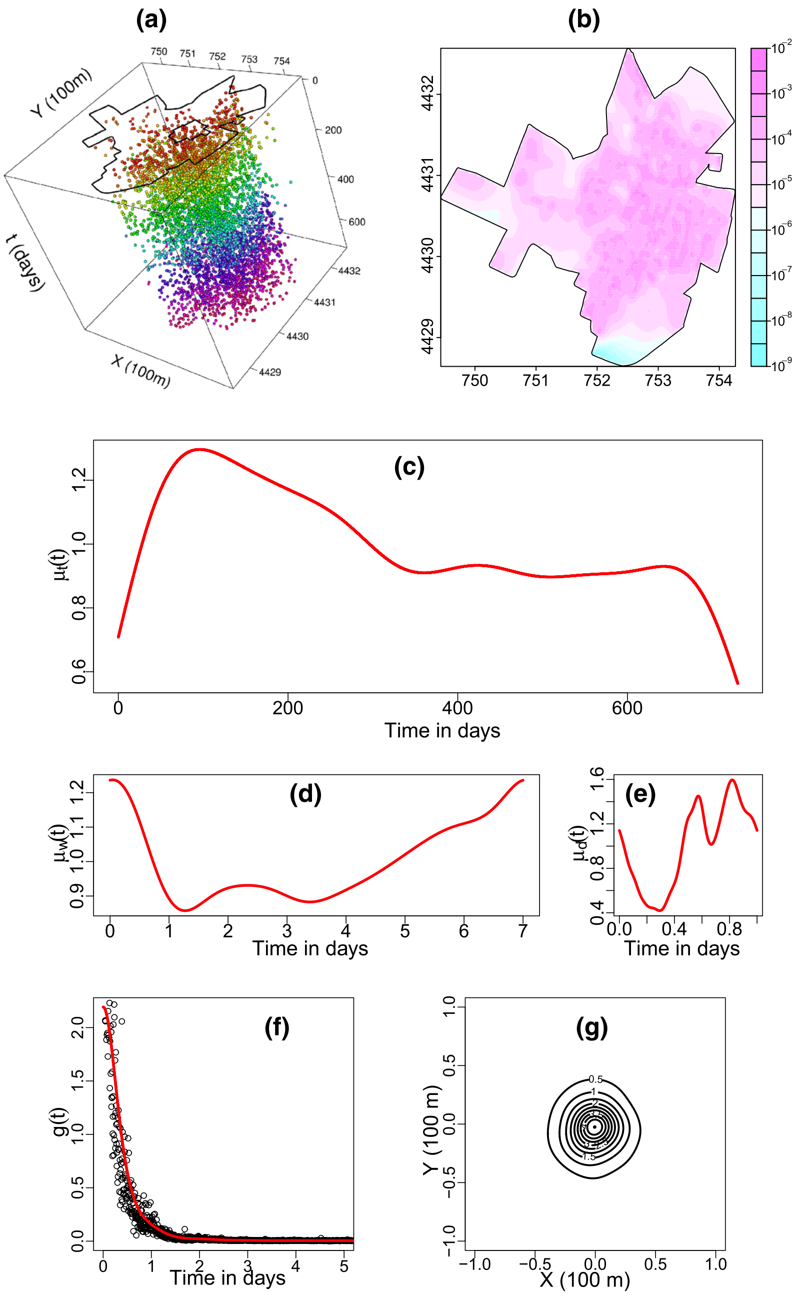

Fig. 2 (Modified from Zhuang and Mateu 2019) **a** A 3D plot of robbery-related violence in Castellon, Spain, 2012–2013. The rainbow colors show the times of occurrence, with red-colored points representing the earliest events and magenta ones the latest. **b** Estimated spatial background rate $\mu_b(x, y)$. **c** Estimated trend function. **d** Estimated weekly periodicity. **e** Estimated daily periodicity. **f** Estimated temporal response function. **g** Estimated spatial response function

$$\varphi_i = \mu_0\, \mu_{\mathrm{t}}(t_i)\, \mu_{\mathrm{d}}(t_i)\, \mu_{\mathrm{w}}(t_i)\, \mu_{\mathrm{b}}(x_i, y_i)/\lambda(t_i, x_i, y_i), \tag{58}$$

and Δ_t, Δ_x, and Δ_y are small positive numbers.

We can improve the above estimates, as well as the histogram estimate of the excitation terms, by introducing kernel smoothing with a correction of the edge effect. The details are available in Zhuang and Mateu (2019) and omitted here.

Zhuang and Mateu (2019) analyze robbery crimes in Castellon, Spain from 2012 to 2013 and disentangle the different background components using the method described in this subsection, as in Fig. 2. Their results show that robbery crime is highly influenced by daily life rhythms, revealed by its daily and weekly periodicity, and that clustering can explain about 3% of such crimes.

6 Conclusion

In summary, this article discusses techniques for using the Hawkes process to investigate the causal encouraging correlation among discrete events. We can divide the whole process into 4 steps:

1. Model design. Design the model according to the features of the observation data, specifically the particular mathematical form of the Hawkes model (parametric, nonparametric, or semiparametric), which depend on the available empirical knowledge of the studied process.
2. Estimation design. Design the estimation according to the types of model formation, use the MLE method or the EM algorithm to estimate parametric model, and use stochastic reconstruction or Equation (2) to reconstruct the nonparametric components.
3. Improvement. Improve the estimation using kernel estimates or the Bayesian method.
4. Diagnosing the new model. The reconstruction method can be naturally used as a diagnostic tool to check whether it is possible to improve the model or not.

From the two examples in Sect. 5, we can see that a good rule to extend a Hawkes model is to account for the data features and physical mechanisms of each specific individual process. Once the direction of the model extension is determined, we can construct, estimate, and diagnose a new model using the stochastic reconstruction techniques together with some nonparametric estimation methods, among which the kernel function is efficient, straightforward, and easy to implement.

Acknowledgements This research is partially supported by Grants-in-Aid no. 19H04073 for Scientific Research from the Japan Society for the Promotion of Science. The author is grateful to D. Vere-Jones and Y. Ogata for their long-time support and encouragement. I also would like to thank the Editor and two reviewers for their helpful comments.

References

Bach, C., & Hainzl, S. (2012). Improving empirical aftershock modeling based on additional source information. *Journal of Geophysical Research Solid Earth*, *117*(B4), B04312.

Bacry, E., & Muzy, J.-F. (2014). Hawkes model for price and trades high-frequency dynamics. *Quantitative Finance*, *14*(7), 1147–1166.

Bartlett, M. S. (1963). The spectral analysis of point processes. *Journal of the Royal Statistical Society Series B (Methodological)*, *25*(2), 264–296.

Chiaraluce, L., Di Stefano, R., Tinti, E., Scognamiglio, L., Michele, M., Casarotti, E., et al. (2017). The 2016 central Italy seismic sequence: A first look at the mainshocks, aftershocks, and source models. *Seismological Research Letters*, *88*(3), 757–771.

Chiodi, M., & Adelfio, G. (2011). Forward likelihood-based predictive approach for space-time point processes. *Environmetrics*, *22*(6), 749–757.

Console, R., Murru, M., & Lombardi, A. M. (2003). Refining earthquake clustering models. *Journal of Geophysical Research*, *108*(B10), 2468.

Daley, D. D., & Vere-Jones, D. (2003). *An introduction to theory of point processes: Volume 1: Elementary theory and methods* (2nd ed.). New York: Springer.

Felson, M., & Boba, R. (2010). *Crime and everyday life*. Thousand Oaks: SAGE Publications.

Fox, E. W., Schoenberg, F. P., & Gordon, J. S. (2016). Spatially inhomogeneous background rate estimators and uncertainty quantification for nonparametric hawkes point process models of earthquake occurrences. *The Annals of Applied Statistics*, *10*(3), 1725–1756.

Guo, Y., Zhuang, J., Hirata, N., & Zhou, S. (2017). Heterogeneity of direct aftershock productivity of the main shock rupture. *Journal of Geophysical Research Solid Earth*, *122*(7), 5288–5305. 2017JB014064.

Guo, Y., Zhuang, J., & Zhou, S. (2015). An improved space-time ETAS model for inverting the rupture geometry from seismicity triggering. *Journal of Geophysical Research Solid Earth*, *120*(5), 3309–3323. 2015JB011979.

Hainzl, S., Christophersen, A., & Enescu, B. (2008). Impact of earthquake rupture extensions on parameter estimations of point-process models. *Bulletin of the Seismological Society of America*, *98*(4), 2066–2072.

Hall, P., & Turlach, B. A. (1999). Reducing bias in curve estimation by use of weights. *Computational Statistics and data Analysis*, *30*, 67–86.

Hall, P., & Wehrly, T. E. (1991). A geometrical method for removing edge effects from kernel-type nonparametric regression estimators. *Journal of the American Statistical Association*, *86*, 665–672.

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

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

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

Helmstetter, A., Kagan, Y. Y., & Jackson, D. D. (2005). Importance of small earthquakes for stress transfers and earthquake triggering. *Journal of Geophysical Research*, *2005*, 110.

Helmstetter, A., Sornette, D., & Grasso, J.-R. (2003). Mainshocks are aftershocks of conditional foreshocks: How do foreshock statistical properties emerge from aftershock laws? *Journal of Geophysical Research*, *108*, 2046.

Huang, Q., Gerstenberger, M., & Zhuang, J. (2016). Current challenges in statistical seismology. *Pure and Applied Geophysics*, *173*(1), 1–3.

Lewis, P. A. W. (1964). A branching poisson process model for the analysis of computer failure patterns. *Journal of the Royal Statistical Society Series B (Methodological)*, *26*(3), 398–456.

Li, C., Song, Z., & Wang, W. (2019). Space-time inhomogeneous background intensity estimators for semi-parametric space-time self-exciting point process models. *Annals of the Institute of Statistical Mathematics*, *2019*, 1–13.

Lombardi, A. M., Cocco, M., & Marzocchi, W. (2010). On the increase of background seismicity rate during the 1997–1998 Umbria-Marche, central Italy, sequence: apparent variation or fluid-driven triggering? *Bulletin of the Seismological Society of America*, *100*(3), 1138–1152.

Marsan, D., & Lengliné, O. (2010). A new estimation of the decay of aftershock density with distance to the mainshock. *Journal of Geophysical Research Solid Earth*, *115*, B9.

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

Meyer, S., & Held, L. (2014). Power-law models for infectious disease spread. *Annals of Applied Statistics*, *8*(3), 1612–1639.

Meyer, S., Warnke, I., Rossler, W., & Held, L. (2016). Model-based testing for space-time interaction using point processes: An application to psychiatric hospital admissions in an urban area. *Spatial and Spatio-temporal Epidemiology*, *17*, 15–25.

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

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., & Tita, G. E. (2011). Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, *106*(493), 100–108.

MØller, J., & Waagepetersen, R. P. (2003). *Statistical inference and simulation for spatial point processes*. London: Chapman and Hall.

Neyman, J. E., & Scott, E. L. (1953). Frequency of separation and interlocking of clusters of galaxies. *Proceedings of the National Academy of Sciences of the United States of America*, *39*, 737–743.

Neyman, J. E., & Scott, E. L. (1958). A statistical approach to problems of cosmology. *Journal of the Royal Statistical Society Series B (Methodological)*, *20*, 1–43.

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

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

Ogata, Y., & Akaike, H. (1982). *On linear intensity model for mixed doubly stochastic Poisson and self-excting point processes*. New York: Springer.

Ogata, Y., Utsu, T., & Katsura, K. (1995). Statistical features of foreshocks in comparison with other earthquake clusters. *Geophysical Journal International*, *121*(1), 233–254.

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

Rathbun, S. L. (1996). Asymptotic properties of the maximum likelihood estimator for spatio-temporal point processes. *Journal of Statistical Planning and Inference*, *51*, 55–74.

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

Schorlemmer, D., Werner, M., Marzocchi, W., Jordan, T., Ogata, Y., Jackson, D., et al. (2018). The collaboratory for the study of earthquake predictability: Achievements and priorities. *Seismological Research Letters*, *89*(4), 1305–1313.

Scott, D. W. (2009). *Multivariate density estimation: Theory, practice, and visualization*, (Vol. 383). New York: Wiley.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London: Chapman and Hall.

Sornette, D., & Werner, M. J. (2005a). Apparent clustering and apparent background earthquakes biased by undetected seismicity. *Journal of Geophysiscal Reseach*, *110*, B09303.

Sornette, D., & Werner, M. J. (2005b). Constraints on the size of the smallest triggering earthquake from the epidemic-type aftershock sequence model, Båth's law, and observed aftershock sequences. *Journal of Geophysical Research*, *110*, B08304.

Tench, S., Fry, H., & Gill, P. (2016). Spatio-temporal patterns of IED usage by the Provisional Irish Republican Army. *European Journal of Applied Mathematics*, *27*(3), 377–402.

Truccolo, W., Eden, U. T., Fellows, M. R., Donoghue, J. P., & Brown, E. N. (2005). A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *Journal of Neurophysiology*, *93*, 1074–1089.

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

Xiong, Z., Zhuang, J., & Zhou, S. (2019). Long-term earthquake risk in north China estimated from a modern catalogue. *Bulletin of the Seismological Society of America***(submitted)**.

Zhuang, J. (2011). Next-day earthquake forecasts by using the ETAS model. *Earth Planet and Space*, *63*, 207–216.

Zhuang, J. (2015). Weighted likelihood estimators for point processes. *Spatial Statistics*, *14*(B), 166–178.

Zhuang, J., Chang, C. P., Ogata, Y., & Chen, Y. I. (2005). A study on the background and clustering seismicity in the Taiwan region by using a point process model. *Journal of Geophysical Research*, *110*, B05S13.

Zhuang, J., & Mateu, J. (2019). A semiparametric spatiotemporal Hawkes-type point process model with periodic background for crime data. *Journal of the Royal Statistical Society Series A (Statistics in Society)*, *182*(3), 919–42.

Zhuang, J., Murru, M., Falcone, G., & Guo, Y. (2019). An extensive study of clustering features of seismicity in Italy from 2005 to 2016. *Geophysical Journal International*, *216*(1), 302–318.

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

Zhuang, J., Ogata, Y., & Vere-Jones, D. (2002). Stochastic declustering of space-time earthquake occurrences. *Journal of the American Statistical Association*, *97*(3), 369–380.

Zhuang, J., Ogata, Y., & Vere-Jones, D. (2004). Analyzing earthquake clustering features by using stochastic reconstruction. *Journal of Geophysical Research*, *109*(3), B05301.

Zhuang, J., Werner, M. J., & Harte, D. S. (2013). Stability of earthquake clustering models: Criticality and branching ratios. *Physical Review E*, *88*, 062109.

Zipkin, J. R., Schoenberg, F. P., Coronges, K., & Bertozzi, A. L. (2015). Point-process models of social network interactions: Parameter estimation and missing data recovery. *European Journal of Applied Mathematics*, *FirstView*, 1–28.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.