SORT 46 (1) January-June 2022, 3-38 DOI: 10.2436/20.8080.02.116

Fifty years later: new directions in Hawkes processes

John Worrall[1,2], Raiha Browning[1,2], Paul Wu[1,2] and Kerrie Mengersen[1,2]

Abstract

The Hawkes process is a self-exciting Poisson point process, characterised by a conditional intensity function. Since its introduction fifty years ago, it has been the subject of numerous research directions and continues to inspire new methodological and theoretical developments as well as new applications. This paper marks half a century of interest in Hawkes processes by presenting a snapshot of four state-of-the-art research directions, categorised as frequentist and Bayesian methods, other modelling approaches and notable theoretical developments. A particular focus is on nonparametric approaches, with advances in kernel estimation and computational efficiencies. A survey of real world applications is provided to illustrate the breadth of application of this remarkable approach.

MSC: *60G55, 62G05.*

Keywords: *Hawkes process, point process, nonparametric.*

1. Introduction

Events occur in the world with frequencies fluctuating over time and space, but often these events are not isolated and their occurrence increases the likelihood of further events. A mathematical model introduced by Hawkes (1971) describes the sequential arrival of these events as a non-Markovian process with a self-exciting nature. The Hawkes process (HP) has wide application in areas such as seismology (Ogata, 1981; Rasmussen, 2013); crime analysis (Yang et al., 2018; Zhuang and Mateu, 2019); traffic incidents (Kalair, Connaughton and Di Loro, 2021; Li, Cui and Chen, 2018); terrorism (Porter and White, 2010; White, Porter and Mazerolle, 2012); finance (Bacry, Mastromatteo and Muzy, 2015); infectious diseases (Kelly et al., 2019; Browning et al., 2021); and social media trends (Hall and Willett, 2016; Zhang, Walder and Rizoiu, 2020b).

[1] School of Mathematical Sciences, Queensland University of Technology, Brisbane, Australia.

[2] QUT Centre for Data Science.

Received: May 2022.

In a HP, the self-exciting nature of the data is modelled through the conditional intensity function which governs the expected arrival rate of events. An important characteristic of this intensity function is the triggering kernel. There has been much research in learning these triggering kernels, including investigation of the underlying assumptions defined through simple parametric functions such as power laws, multiple exponential distributions, and Gaussian, Rayleigh and Weibull functions (Chen, Hawkes and Scalas, 2021; Chiang, Liu and Mohler, 2021). Further studies consider approaches to adapting these parametric models (Kobayashi and Lambiotte, 2016; Du et al., 2016), extending to the multidimensional setting and improving scalability of estimation by low-rank approximation (Zhou, Zha and Song, 2013; Bacry et al., 2020) and mean-field theory (Bacry et al., 2016a).

More flexible approaches to learning the kernel include representing the function as piecewise constant on a finite grid. Seminal work by (Lewis and Mohler (2011); Bacry, Dayri and Muzy (2012) provides a nonparametric framework for estimation that is also being actively explored. Bayesian nonparametric approaches are also emerging, with leading work in the area including (Donnet, Rivoirard and Rousseau, 2020; Zhou et al., 2020b; Zhang et al., 2020b). However, relaxing assumptions and increasing the expressiveness of functions comes at a cost: there is a requirement either for discretisation of the input domain or improved computational requirements to meet increasing practical demands.

These requirements have motivated new research into efficient algorithms for the analysis of HPs, and concomitant investigation of the characteristics of these algorithms. For example Achab et al. (2018) encodes causality of a multivariate process via a moment matching method fitting to second and third order cumulants. Work by Zhang et al. (2020b) takes advantage of latent branching structure and stationarity assumptions to reduce computational complexity and to efficiently infer a flexible representation of the kernel using Gaussian processes. In another very promising direction, Yang et al. (2017) focus on sequential (online) learning by approximating the function in a reproducing kernel Hilbert space. New Bayesian perspectives are lending themselves readily to handling the sheer volume and scalability in online learning (Broderick et al., 2013; Chérief-Abdellatif, Alquier and Khan, 2019; Markwick, 2020).

Another direction for Bayesian nonparametric approaches is in extending HPs to also cluster events via Dirichlet processes (Du et al., 2015). In these examples the form of the triggering kernel is generally parametric, and interest lies in the clustering of the events themselves.

Other recent directions of research into HPs arise from the perspective of graphs (Liu, Yan and Chen, 2018), stochastic differential equations (SDEs) (Lee, Lim and Ong, 2016; Kanazawa and Sornette, 2020) and neural networks (Zhang et al., 2020a; Du et al., 2016). These frameworks aim to provide more flexibility and less bias, while taking advantage of the techniques made available from a rapidly growing statistical data science community.

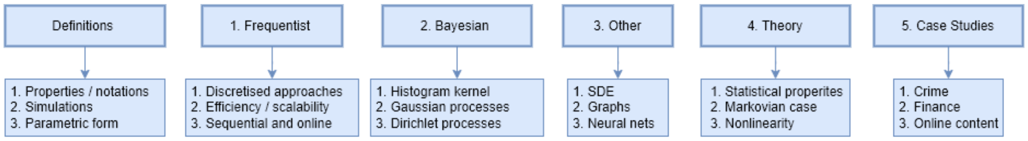

Figure 1. *Structure of paper.*

In addition, there has been considerable and significant research in theoretical properties and guarantees of HPs. Recent bodies of work include advances in estimating higher order statistical properties (Jovanović, Hertz and Rotter, 2015; Cui, Hawkes and Yi, 2020), asymptotic properties of the Markovian class of HPs (Gao and Zhu, 2018b; Zhu, 2015) and developments around nonlinear generalisation (Torrisi, 2016; Gao and Zhu, 2018a; Sulem, Rivoirard and Rousseau, 2021).

This paper aims to provide a review of these new directions in the modelling and analysis of HPs, with an emphasis on nonparametric Bayesian approaches and brief reference to the underpinning theory. We preface the review with a brief overview of notation, definitions and properties of HPs, and close the paper by presenting a survey of recent applications and some substantive applications in crime, finance and social media. The structure of the paper is illustrated in Fig 1.

1.1. Definitions and basic properties

This section provides a brief summary of mathematical definitions, properties and the general form of the HP which will be used throughout the remaining sections.

Definition 1.1 (Poisson process)**.**

A *nonhomogeneous Poisson process* with time varying arrival rate $\lambda(t)$ is defined as a counting process, $N(t) : t \geq\ 0$ which satisfies $t \in \mathbb{R}^+$, with associated history $\mathcal{H}_t : t \geq 0$, such that probability is given by

$$\mathbb{P}(N(t+h) - N(t) = m|\mathcal{H}_t) = \begin{cases} \lambda(t)h + o(h) & \text{m = 1} \\ o(h) & \text{m > 1} \\ 1 - \lambda(t)h + o(h) & \text{m = 0.} \end{cases} \tag{1}$$

Of particular interest in the study of nonhomogeneous Poisson processes is the HP $N(t)$ where $\lambda(t) : \mathbb{R}^+ \to \mathbb{R}^+$.

The time intervals between events (shown in Fig 2 as $t_1, t_2 \ldots, t_7$) are described as inter-arrival event times (Rasmussen, 2018). The point process can be characterised by the distribution function of the next arrival time conditioned on the past. Thus the conditional cumulative density function $F(t|\mathcal{H}_\mu)$ of next arrival time T_{k+1} can be expressed in terms of the conditional density function $f(s|\mathcal{H}_\mu)$,

$$F(t|\mathcal{H}_\mu) = \int_\mu^t \mathbb{P}(T_{k+1} \in [s, s+ds]|\mathcal{H}_\mu)\ ds = \int_\mu^t f(s|\mathcal{H}_\mu) ds.$$

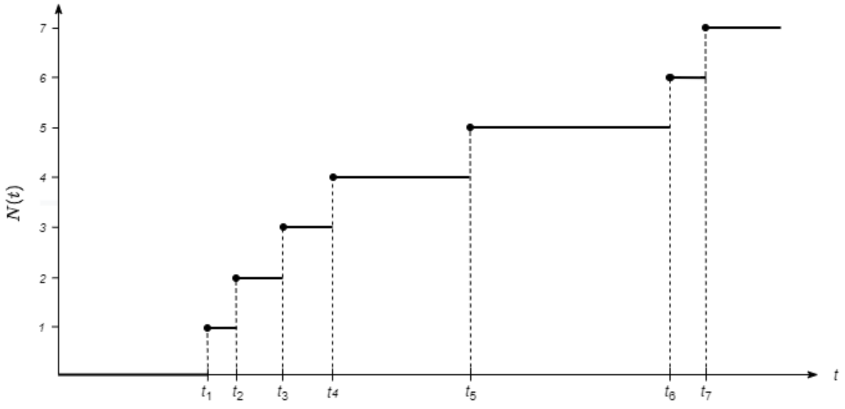

Figure 2. *Point process with stochastic realisation* $\{t_1, t_2 \ldots\}$ *and counting process N(t).*

where $\mathcal{H}_\mu$ is the history of the process until the last arrival (Ozaki, 1979). Where the conditional distribution is given using the law of total probabilities,

$$f(t_1, t_2, \ldots, t_n) = \prod_{i=1}^{n} f(t_i | \mathcal{H}_\mu)).$$

Definition 1.2 (Conditional intensity function)**.**
Let the conditional density be $f(t|\mathcal{H}_{t_n})$ and the corresponding cumulative distribution function $F(t|\mathcal{H}_{t_n})$ for any $t > t_n$. Then $\lambda^*(t)$ is the conditional intensity or hazard function (Cox, 1955). The notation $*$ borrowed from (Daley and Vere-Jones, 2003) is used to represent conditioning on the history up to time t. A more intuitive definition of the conditional intensity function (Daley and Vere-Jones, 2003) is its expected rate of arrivals conditioned on the associated history,

$$\lambda^*(t) = \frac{f^*(t)}{1 - F^*(t)} = \lim_{h \to 0} \frac{\mathbb{E}[N(t+h) - N(t)|\mathcal{H}_t]}{h}.$$

Hawkes (Hawkes, 1971) introduced a class of self-exciting process to model contagious processes, characterised by this conditional intensity function.

Definition 1.3 (HP)**.**
Let $D \in \mathbb{N}^+$ and $\{(t_i^j)\}_{j=1,\ldots,D}$ be a D-dimensional point process, with associated counting processes $N_t = (N_t^1, \ldots, N_t^D)$. A multidimensional Hawkes process (MHP) is defined with intensities $\lambda_i^*(t), i = 1, \ldots, D$ given by

$$\lambda_i^*(t) = \mu_i + \sum_{j=1}^{D} \int_0^t \phi_{ij}(t-s) dN_j(s) \tag{2}$$

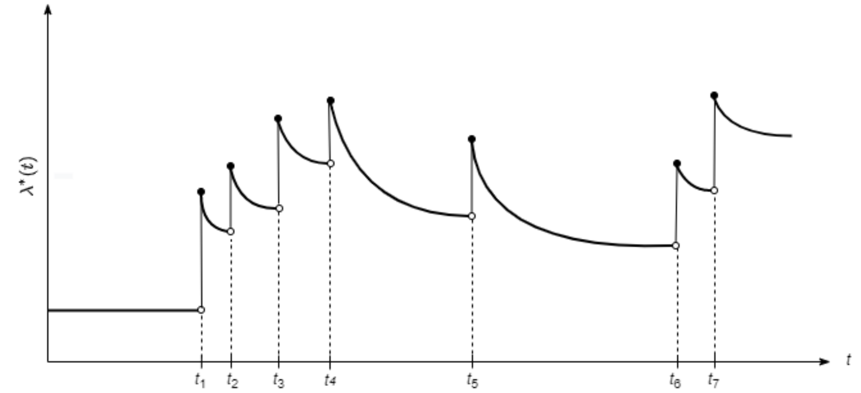

Figure 3. *Conditional intensity function of the HP, with exponential decay.*

where $\mu_i > 0$ is the non-negative background intensity of process i and $\phi_{ij}(\cdot) : (0,\infty) \rightarrow (0,\infty]$ is the excitation function from process j onto process i. When D=1, the univariate HP is expressed as

$$\lambda^*(t) = \mu + \int_0^t \phi\,(t-s)\,dN(s). \tag{3}$$

The self-excitation term within the expression of the HP is designed to capture the influences of all previous events in the current conditional intensity value. In multidimensional cases the self-exciting and mutually-exciting terms are, respectively, $\phi_{ii}(\cdot)$ and $\phi_{ij}(\cdot)$, $i \neq j$. A popular kernel choice is the exponential decay,

$$\phi_{ij}(t-s) = \left(\alpha_{ij} e^{-\beta_{ij}(t-s)}\right)_{i,j=1,\ldots,D} \tag{4}$$

where each arrival in the system instantaneously increases the arrival intensity by α_{ij} and over time the arrivals influence the decay at rate β_{ij} (Fig 3.)

The standard temporal HP can be extended to include spatial dependence, thereby capturing the clustering behaviours of the process through time and space. The process has an analogous description to the temporal HP.

Definition 1.4 (Spatio-temporal HP)**.**
Let $D \in \mathbb{N}^+$. Let $\{(t_i^j),(x_i^j),(y_i^j)\}_{j=1,\ldots,D}$ be a D-dimensional point process, with an associated counting process $N_t = (N_t^1, \ldots, N_t^D)$. A multidimensional spatio-temporal Hawkes process is defined with intensities $\lambda_i^*(\cdot), i = 1, \ldots, D$ given by

$$\lambda_i^*(t,x,y) = \mu_i + \sum_{j=1}^{D} \int_0^t \int_{\mathcal{X}} \int_{\mathcal{Y}} \phi_{ij}(t-s, x-u, y-v)\,dN_j(s)\,dN_j(u)\,dN_j(v). \tag{5}$$

Further generalisations

In the past fifty years, there have been several popular types of generalisations of the conditional density. Three more common approaches are described in reference to the following univariate case equation,

$$g\big(\lambda^*(t)\big) = \mu(t) + \sum_{t>s} \phi(t-s,\xi_i) \tag{6}$$

1. Generalisations of the baseline process, $\mu(t)$, as a function of time effects on exogenous activity;
2. The Marked HP, where marks (ξ_i) associated to events (t_i) have different effects on intensity;
3. Nonlinear processes, where $g\big(\lambda^*(t)\big)$ is a nonlinear function with support in $\mathbb{R}^+$.

Regardless of the assumed background and triggering function forms, the fitness of the HP model is typically measured via the likelihood (Daley and Vere-Jones, 2003).

Definition 1.5 (Likelihood of HP)**.**
Let $N(\cdot)$ be a regular point process on $[0,T]$ for some finite positive T, and let $t_1,...,t_n$ denote a realisation of $N(\cdot)$ over $[0,T]$. Then, the likelihood function L is expressible in the form

$$L(t_1 \ldots, t_n \mid \mu, \phi) = \left[\prod_{i=1}^{n} \lambda^*(t_i)\right] \exp\left(-\int_0^T \lambda^*(u)\,du\right). \tag{7}$$

A condition in ensuring the estimated model is stable and has access to most properties of the HP is stationarity.

Definition 1.6 (Stationarity of HP)**.**
Let $N(\cdot)$ be a multivariate HP on $[0,T]$ for some finite positive T, where $N(\cdot)$ is stationary if a translation in time does not change its distribution. Let Φ be a $D \times D$ matrix with entries given by,

$$\Phi_{ij} = \int_0^\infty \phi_{ij}(u)du.$$

A sufficient condition for stationarity is that $\rho(\Phi) < 1$, where $\rho(\Phi)$ is spectral radius of Φ given as

$$\rho(\Phi) = \max_{x \in \mathcal{S}(\Phi)} |x| \tag{8}$$

where $\mathcal{S}(\Phi)$ is a set of all eigenvalues of Φ.

A number of simulation procedures are available for ensuring stationarity and other stochastic properties of the HP. The generation of synthetic data sets from these methods ensures statistical equivalence to the real population of interest and is an invaluable tool in supporting model design and development.

1.2. Simulating a HP

Concerning the experimental aspects of a self-exciting process, two synthetic generation algorithms are popular.

The first of these is the thinning method, a standard approach to producing nonhomogeneous Poisson processes. The intuition of the algorithm is to combine two generated homogeneous Poisson processes of different rates and to remove points probabilistically, so the remaining points satisfy a time-varying intensity $\lambda(\cdot)$. For the Ogata modified algorithm Ogata (1981), the intensity has no asymptotic upper bound, although it is common to set non-increasing periods without any arrival.

Simulation of a HP may also be represented as an immigration-birth process, leading to a branching simulation procedure (Fig 4). Here immigrants are generated via a homogeneous Poisson rate λ, conditioned on K immigrants with arrival times uniformly i.i.d in time window $(0,T]$. Each immigrant descendant forms a nonhomogeneous Poisson process with intensity (α/β) giving arrival times $[I_i+E_1, I_i+E_2, \ldots, I_i+E_{D_i}]$.

Algorithm 1 Simulating univariate HP by thinning

```
Require: $(\lambda^*(\cdot), T)$
  Initialisation $P \leftarrow [], t \leftarrow 0, \varepsilon \leftarrow 10^{-10}$
  while $t < T$ do
    Set upper bound $M \leftarrow \lambda^*(t+\varepsilon)$
    Generate candidate point $E \leftarrow$ Exp(M)
    $t \leftarrow t+\varepsilon$
    Set with probability $U \sim$ Unif(0,1)
    if $t < T$ and $U \leq \lambda^*(t)$ then
      $P \leftarrow [P,t]$
    end if
  end while
  return P
```

1.3. Parametric models for HPs

There is a rich literature on parametric methods for modelling HPs. These approaches have many uses, particularly when the parametric form of the process is obvious. They are often simple to implement and can provide useful insights into the behaviour of these processes. A brief summary of parametric methods is provided here, covering some of the most popular forms for the triggering kernel and inference techniques for these models.

Algorithm 2 Simulating univariate HP by clusters

```
Require: $(T, \lambda, \alpha, \beta)$
  Initialisation $P \leftarrow [], i \leftarrow 1$
  Generate immigrants $K \sim \text{Pois}(\lambda T)$
  $I_1, I_2, .., I_K \sim \text{Unif}(0, T)$
  Generate descendants $D_1, D_2, .., D_K \sim \text{Pois}(\alpha/\beta)$
  while $i < K$ do
    if $D_i > 0$ then
      $E_1, E_2, .., E_{D_i} \sim \text{Exp}(\beta)$
      $P \leftarrow P \cup [I_i + E_1, I_i + E_2, .., I_i + E_{D_i}]$
    end if
    Set i = i + 1
  end while
  Remove invalid descendants (0,T] $P \leftarrow (P_i : P_i \in P, P_i \leq T)$
  Add immigrants $P \leftarrow \text{Sort}(P \cup [I_1, I_2, .., I_n])$
  return P
```

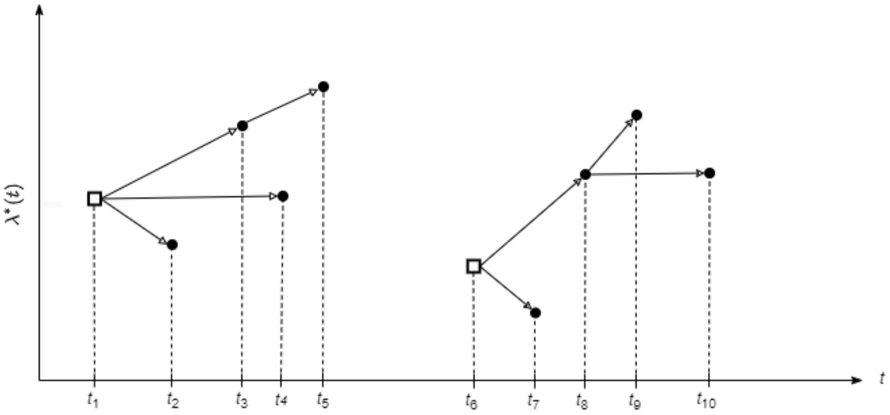

Figure 4. *HP immigrant-birth representation (squares indicate immigrants and circles indicate offspring/descendants).*

1.3.1. Choice of triggering kernel

Although the structure of the conditional intensity function is quite flexible, the most common triggering kernel is parameterised as an exponential decay

$$\phi(t-s) = \alpha e^{-\beta(t-s)}.$$

Here α represents the overall strength of excitation and β denotes the influence decay rate of the arrivals. Hawkes (1971) used this parametric form to derive theoretical properties of the covariance density function and Bartlett spectrum, via the frequency domain. The Laplace transform is given as

$$\mathcal{L}\{\cdot\}(s) = \frac{\alpha\mu\,(2\beta-\alpha)}{2\,(\beta-\alpha)\,(s+\beta-\alpha)} \tag{9}$$

where $s \in \mathbb{C}$. Evaluating the power spectral density (defined in terms of the covariance density) of a HP provides a set of useful tools in discriminating and fitting between models and access to other techniques from the spectral theory field.

In addition, the exponential decay has several other advantageous properties, such as straightforward computation of the expected value of an arbitrary function on $N(t)$, direct simulation, and efficient computation of the likelihood. Most of these properties descend from the Markov property, where the intensity and the pair $(\lambda(t), N(t))$ are Markovian, in the following form,

$$d\lambda(t) = -\beta\lambda(t)dt + \alpha\beta dN(t) \tag{10}$$

Several other parametric kernel forms have also become popular. These include the power law, sinusoidal, Gaussian and rectangular functions supporting different types of interactions among events. In almost all realistic applications, however, it is not obvious which parametric form of the excitation function for HPs is the most appropriate. This has generated a great deal of recent interest in nonparametric specification of the kernel function. Under this representation, traditional assumptions about the triggering kernel can be relaxed to capture the complexities and subtleties of the excitation effects retrieved from the data. Before moving to a more comprehensive discussion of nonparametric directions in Sections 2 and 3, we complete the introduction to HPs with an overview of spatio-temporal approaches and matters of inference.

1.3.2. Spatio-temporal approaches

A number of authors propose spatio-temporal self-exciting processes. Generally, the triggering kernel is constructed in a separable fashion, where the temporal and spatial dependence can be decomposed (Mohler et al., 2011; Schoenberg, 2016; Reinhart, 2018). A popular parameterisation for the respective kernels is exponential decay in time and Gaussian decay in space. Several Bayesian approaches have also been introduced to model spatio-temporal HPs. These include Holbrook, Ji and Suchard (2022), who model the outbreak of Ebola in West Africa and extend the standard spatio-temporal Hawkes model to learn the evolutionary history of the virus which informs the characteristics for each variant of the virus. Holbrook et al. (2020) also account for uncertainty in the location of events by placing a prior on the spatial position of events.

A popular case of the spatio-temporal HP, originally introduced as a marked, purely temporal process, is an adaption of the Epidemic-Type Aftershock Sequence (ETAS) model (Ogata, 1988) to incorporate spatial dynamics (Ogata, 1998). The spatial ETAS model was introduced in the context of modelling earthquakes through the baseline parameter, and their corresponding aftershocks represented by the triggering kernel. The marks are denoted by m and correspond to the magnitude of the earthquake. Thus, the intensity function can be written in the form,

$$\lambda_m^*(t) = \mu(x,y) + \sum_{t>s} \phi_m(t-s, x-u, y-v) \tag{11}$$

$$\phi_m(t,x,y) = \kappa_m \times \frac{(p-1)c^{p-1}}{(t+c)^p} \times \left(\frac{1}{\pi\sigma_m} \cdot f(\frac{x^2+y^2}{\sigma_m}) \right), \tag{12}$$

where κ_m is the expected number of aftershocks for an earthquake of magnitude m, σ_m is a scale factor, f is a function such that $\int_0^\infty f(x)dx = 1$ holds, and p and c are global constants. The second and third terms of ϕ represent the temporal and spatial decay functions respectively.

1.4. Inference

A range of inference approaches have been used for estimating the parameters of these parametric models. A common procedure is that of maximum likelihood estimation, where the likelihood function is maximised to obtain the set of parameter values that produce the highest likelihood.

Other approaches are based on the branching representation of the HP which allows the likelihood to be decomposed into conditionally independent immigrant and offspring processes. Due to this latent structure, inference methods such as Expectation-Maximisation (EM) and Variational Inference (VI) can be used to integrate over this latent space. A detailed explanation of the EM algorithm for HPs is provided in Laub, Lee and Taimre (2021) and a similar construction is used when performing VI for these models. These inference techniques that utilise the latent structure of HPs are discussed further in the subsequent sections of this review. Efficient Gibbs samplers have also been developed, using the decomposition of the likelihood and placing conjugate priors on the parameters of the model.

We turn now to four general directions of research that illustrate current activity in HPs. These include frequentist nonparametric kernel adaptation and presentation, Bayesian nonparametric approaches, other approaches (stochastic differential equations, graphs and neural networks) and theoretical aspects of HPs. This is intended to be a canvas rather than an exhaustive review of all research directions.

2. Direction 1: Frequentist nonparametric kernel adaptation and estimation

There is now a large literature on various directions of research into nonparametric kernels for HPs. The following discussion focuses on a selection of these directions, based on their novelty, currency and interest to the authors. The focus is initially on several frequentist approaches, followed by efficient estimation methods and finally sequential or online models.

2.1. Discretised scheme

A frequentist approach to estimating the excitation and/or baseline function is defined by approximating the function as a binned grid, where function values are piecewise constant within each bin and the width of each bin is selected optimally to model local variations of the excitation.

2.1.1. Stochastic declustering

Early work by Zhuang, Ogata and Vere-Jones (2002) supports this approach by attempting to differentiate between 'true' background events and triggered events. Such differentiation using the probability for background events, p_{ii}, is called stochastic declustering. Motivated by this work, Model Independent Stochastic Declustering (MISD) was introduced as a nonparametric HP with homogeneous background rate (Marsan and Lengliné, 2008) and later extended for the more general case of varying $\mu(t)$ (Lewis and Mohler, 2011). This method makes use of the branching structure to reduce both baseline and triggering kernel into a density estimation problem. The augmented likelihood of observations D and branching structure B with two independent components is then given by

$$p(D,B|u(t),\phi(\tau)) = \underbrace{\prod_{i=1}^{N} u(t_i)^{b_{ii}} \exp(-uT)}_{u(t)} \cdot \underbrace{\prod_{i=2}^{N}\prod_{j=1}^{i-1} \phi(t_i-t_j)^{b_{ij}} \prod_{i=1}^{N} \exp\left(-\int_0^T \phi(\tau)d\tau\right)}_{\phi(t)} \tag{13}$$

The recovered parameters are then updated via an expectation step, where b_{ij} is replaced by the expectation $\mathbb{E}[b_{ij}] = p_{ij}$, representing the probability that event i is caused by event j

$$p_{ij}^k = \frac{\phi^k(t_i-t_j)}{u^k + \sum_{j=1}^{i-1}\phi^k(t_i-t_j)}, \quad p_{ii}^k = \frac{u^k}{u^k + \sum_{j=1}^{i-1}\phi^k(t_i-t_j)}. \tag{14}$$

This allows for the construction of a matrix P^k, giving events caused by the background rate (diagonal elements) or another event (non-diagonal elements). The maximisation step then updates parameters given the current matrix of probabilities such that,

$$u^{k+1} = \frac{1}{T}\sum_{j=1}^{n} p_{ii}^k, \qquad \phi_m{}^{k+1} = \frac{1}{\delta t}\sum_{i>j\in A_m} p_{ij}^k \tag{15}$$

where δt is a discretisation parameter controlling the bin grid and A_m is the set of pairs of events.

In examining the MISD model, we illustrate in Fig 5 a synthetic exponential kernel (red) compared with the estimated kernel (blue) from the MISD model with varying discretisation parameters.

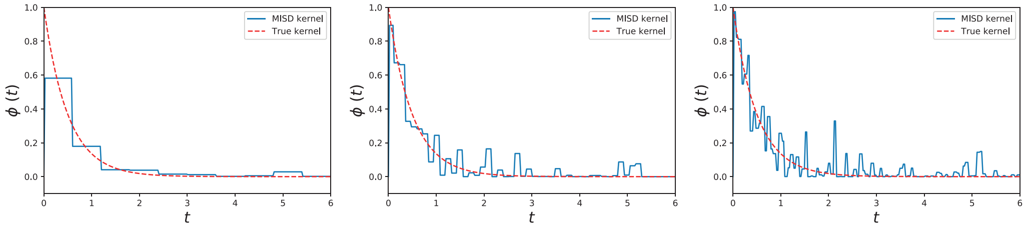

Figure 5. *Kernel estimate (blue) on synthetic exponential kernel, increased bin size 10,50,100 (left to right).*

Results highlight empirically the sensitivity of the chosen discretisation parameter δ to the structure of the kernel ϕ. Incorrect choice of the number of bins leads to underfitting (left) and overfitting (right). This motivates future work that improves on bandwidth choice and boundary effects, which are unavoidable topics of kernel estimation.

Another approach to defining the excitation function on a grid or set of grids is through exploiting relations in the model in the frequency domain between second order statistics and the triggering kernel.

2.1.2. Wiener-Hopf integral

Bacry and Muzy (2016) showed that the kernel matrix of a MHP can be estimated by relating the jump correlation matrix of event processes to a series of Wiener-Hopf equations. This relationship between the first and second order characterisation properties, triggering kernel and background rate of a HP is exploited in the frequency domain to satisfy a unique causal solution. Given this unique solution, the unknown kernel may be solved by a discretised system of linear equations via quadrature and inversion. The triggering kernel matrix function and conditional expectation $g(t)$ satisfy the following Wiener-Hopf equation,

$$g(t) \ = \Phi(t) + \Phi(t) * g(t) \ , \forall t > 0 \tag{16}$$

where $*$ represents convolution (Bacry et al., 2015).

The numerical approximation requires selecting a grid and quadrature scheme for $g(t)$, computing first the estimated $\widetilde{g}$ (Jovanović et al., 2015). Considering the univariate case, where N_t jumps are all size 1 and stationary, the first order property (mean event rate) is

$$\Lambda dt \ = \mathbb{E}(dN_t) = \frac{\mu}{1 - \int \phi \ (\tau) d\tau} dt$$

with second order statistic are summed up by infinitesimal covariances,

$$\text{Cov}\left(dN_{t1}, dN_{t2}\right) = \mathbb{E}\left(dN_{t1} \ dN_{t2}\right) - \mathbb{E}\left(dN_{t1}\right) \mathbb{E}\left(dN_{t2}\right)$$

under assumption N_t has stationary increments, $\operatorname{Cov}(dN_{t1}, dN_{t2})$ only depends on $\tau = t_2 - t_1$ this part of this covariance is can be written as

$$v(\tau)d\tau = \mathbb{E}(dN_0\, dN_\tau) - \mathbb{E}(dN_0)\ \mathbb{E}(dN_\tau).$$

where the second order statistic can be rewritten in terms of conditional expectations,

$$g(\tau)dt\ = v(\tau)\,d\tau\ /\Lambda\ = \mathbb{E}(dN_\tau | dN_0 = 1)\ -\ \Lambda d\tau\ .$$

details of proof in Bacry and Muzy (2016).

Approximation of the equation is commonly given via the Gaussian quadrature method for discretised Wiener-Hopf systems on the interval $[t_{\min},\ t_{\max}]$ and is shown as

$$\widetilde{g}_{ij}(t_n) = \widetilde{\phi}_{ij}(t_n) + \sum_{l=1}^{D}\sum_{k=1}^{K} w_k\ \widetilde{g}_{il}\,(t_n - t_k) w_k \widetilde{\phi}_{ij}(t_n)\ ,\ \forall n \in [0,K], i,j \in [0,D]. \qquad (17)$$

Inverting the obtained linear systems results in estimation of the matrix kernel at quadrature points $\widetilde{\phi}_{ij}$ and lastly estimating μ using the first order cumulant.

In the example below we again consider a simulated exponential decay kernel and approximate Wiener-Hopf equation with optimal bandwidth given by the MSE with respect to grid size.

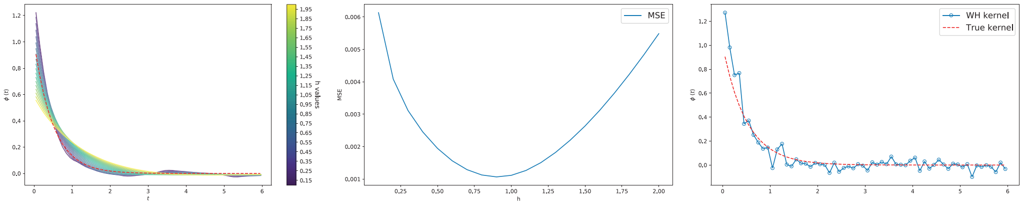

Figure 6. *Kernel estimates with actual in red (left) and mean square error (MSE, middle) with varying width(h). Optimum kernel estimate (right).*

The more expressive bin grid approach compared to parametric methods requires a larger sample size and is restricted to non-Markovian regimes, thus a larger computational cost. This has led to a body of work focusing on computational efficiencies.

2.2. Improved estimation scalability and efficiency

Achab et al. (2018) decreases computational costs by replacing estimation of kernels through matching cumulants (or moments). This strategy relates the branching structure of an MHP to Granger causality, estimating cumulative values to quantify the causal relationship among each node by estimating the matrix,

$$\int \Phi(t)\ dt = \int_0^\infty \phi_{ij}(t)\ dt\ \geq 0\ \text{ for } 1 \leq\ i,j \leq d. \qquad (18)$$

It first computes from sequences moments, up to the third estimates, $\hat{\mathcal{M}}$ and minimises the L^2 error between these estimates and actual moments $\mathcal{M}^{\text{true}}$ (uniquely determined from $||\Phi(t)||$) where

$$||\hat{\Phi}(t)|| = \arg \min_{||\Phi(t)||} ||M(||\Phi(t)||) - \hat{M}||^2, \tag{19}$$

where the matrix $\mathcal{R}$ is given as

$$\mathcal{R} = (\mathbb{I}^d - \hat{\Phi}())^{-1}.$$

The explicit relationships between the matrix and cumulants are then defined as the following identities Λ, C, K. Estimation is given from general formulae for the integral of cumulants of an MHP in Jovanović et al. (2015), where 3rd order statistics are connected to skewness of N_t (Achab et al., 2018), shown as

$$\begin{aligned} K^{ijk}dt = \mathbb{E}\left(dN_t^i(\Delta_H N_t^j - 2H\Lambda^j)(\Delta_H N_t^k - 2H\Lambda^k)\right) \\ - dt\ \Lambda^i \mathbb{E}\left((\Delta_H N_t^j - 2H\Lambda^j)(\Delta_H N_t^k - 2H\Lambda^k)\right) \end{aligned} \tag{20}$$

where $\Delta_H N_t^i = N_{t+H}^i - N_{t-H}^i$ with first and second moments given as

$$\begin{aligned} \Lambda^i dt &= \mathbb{E}(dN_t^i) = \lim_{n\to\infty} \frac{1}{T}\ \lambda_i{}^*(t) = (\mathbb{I} - ||\Phi||)^{-1}\mu_i \\ C^{ij}dt\ &= \mathbb{E}(dN_t^i(\Delta_H N_t^j - 2H\Lambda^j)) \end{aligned}$$

and with $\hat{\Lambda}, \hat{C}\hat{K}$ to be incorporated in the estimator $\hat{\mathcal{R}} = \arg\min_{\mathcal{R}} L(\mathcal{R}))$ such that

$$L(\mathcal{R}) = (1-k)||K^c(\mathcal{R} - \hat{K}^c)||_2^2 + k||C(\mathcal{R}) - \hat{C}||_2^2,$$

where K^c is the tensor contraction of tensor K, and the coefficient k is used to scale the two terms

$$k = \frac{||\hat{K}^c||_2^2}{||\hat{K}^c||_2^2 + ||\hat{C}||_2^2}.$$

Inverting (19) leads to the recovered matrix

$$\hat{\Phi}(t) = (\mathbb{I}^d - \hat{\mathcal{R}})^{-1}.$$

In the univariate case, the $\hat{\Phi}$ can be estimated from the second order statistics, whereas in higher dimensions the third order or skewness is required for unique $\hat{\Phi}(t)$.

The nonparametric cumulant method outperforms the previously discussed MISD and Wiener-Hopf algorithms, given its reduced complexity. The recovered matrix also provides a quantifiable degree of endogeneity in a system and the causality structure of a network.

Another approach to improving the computational bin grid process is by updating parameters in a single pass.

2.3. Sequential and online approaches

Yang et al. (2017) proposed an online procedure where the triggering function belongs to a Reproducing Kernel Hilbert Space (RKHS). Assume that there exists a function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that there is a positive definite kernel,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K(x_i, x_j) \geq 0$$

where $n \in \mathbb{N}, c \in \mathbb{R}$ and that for any $x \in \mathcal{X}$, the evaluation functional is bounded as

$$f(x) = \langle f, K(x, \cdot)\rangle_H \leq C||f||_H$$

for some constant C. Suppose that $f(x)$ satisfies the decreasing tail property with tail function $\varepsilon_f(\cdot)$ if

$$\sum_{k=m}^{\infty}(t_k - t_{k-1}) \sup_{x\in(t_k, t_{k-1}]} |f(x)| \leq \varepsilon_f(t_{m-1}), \forall m > 0, \tag{21}$$

where $\varepsilon_f(\cdot)$ is a bounded and continuous function such that $\lim_{t\to\infty} \varepsilon_f(t) = 0$. Then the assumed triggering function belongs to a RKHS where similarities among high-dimensional and complex distributions are mapped onto lower-dimensional ones. The process then takes the log-likelihood function and optimises over a discretised version,

$$\begin{aligned} L_i(\lambda) =& \sum_{d=1}^{D}\sum_{k=1}^{M(t)} \int_{\chi_{k-1}}^{\chi_k} \lambda_d(s)ds - y_{d,k}\ log\lambda_d(t_k) \\ =& \sum_{d=1}^{D} \Delta L_{d,i}(\lambda_d) \end{aligned}$$

where partitioning $\{0, \chi_1 ... \chi_{M(t)}\}$ on the interval $[0, T]$ is defined as

$$\chi_{k+1} = \min_{t_i > \chi_k}\{\iota * \lfloor \chi_k/\iota \rfloor + \iota, t_i\}$$

for some small $\iota > 0$. The discretised version is then expressed as

$$\begin{aligned} L(\lambda) =& \sum_{d=1}^{D}\sum_{k=1}^{M(t)} \int_{\chi_{k-1}}^{\chi_k} (\chi_k - \chi_{k-1}) - \lambda_d(\chi_k) - y_{d,k}\log(\lambda_i(\chi_k)) \\ =& \sum_{d=1}^{D} \Delta L_{d,t}(\lambda_d). \end{aligned} \tag{22}$$

To perform fast evaluation, the optimisation algorithm processes each partition and employs the following three properties. The first is a truncation of the intensity function that considers arrivals within a recent window $[t - z, t)$ as

$$\lambda_i^z(t) = \mu_i \sum_{i=1}^{p} \int_0^t \mathbb{1}\{t - \tau < z\} f_{ij}(t - \tau) dN_j(\tau). \tag{23}$$

The second is Tikhonov regularisation over the baseline and triggering kernel, adding weight terms to the loss function and keeping the resultant values small. The third is enforcing positivity through the projection steps in the optimised triggering function part. By reducing complexity through the RKHS and by exploiting inter-arrival MHP properties, the resulting nonparametric algorithm recovers estimates over a single pass with comparable computation cost of alternative parametric online learning algorithms.

2.4. Summary

Nonparametric HPs comprise a major direction of current research. Notwithstanding the advantages of a nonparametric framework, such as the enablement of a more expressive triggering function, the approach induces a number of challenges. Firstly, the discretisation approach employed in fitting the nonparametric model requires a larger sample size compared to more traditional parametric methods that fit better on shorter and fewer arrival sequences. However, they may underfit on longer sequences. This is easily seen when relating the bin division grid concept to a histogram of inter-arrival times. Secondly, computational loads for estimation and inference become much larger, given the above-mentioned sample size requirements and as the binning process is a sequential process that cannot take account of the Markovian property given in the exponential function.

Several research directions to address these challenges have been discussed. The first is an improved computational estimation method in matching cumulants. The second is a reduction in computational complexity though some assumptions on the kernel (belonging to RKHS) to estimate parameters given a single pass on some discretised time domain.

3. Direction 2: Bayesian nonparametric approaches

Bayesian approaches inherit the usual benefits of more flexible hierarchical modelling and probabilistic inference. A number of Bayesian nonparametric approaches to modelling HPs have been proposed in the literature. In particular, the majority of methods discussed in this section estimate either the baseline rate, triggering kernel, or both, using either a nonparametric histogram kernel or Gaussian processes. Also of interest in the Bayesian nonparametric literature for HP is using the self-exciting properties of HPs to determine the clustering of events using Dirichlet processes.

3.1. Histogram kernel

A drawback of the binning based methods for estimating a histogram function discussed in Section 2 is that they require *a priori* selection of the grid size. This often leads to models that are either overfitted or underfitted. Donnet et al. (2020) propose a Bayesian nonparametric approach for modelling HP that eases this choice.

The authors derive posterior concentration rates for HPs, and exemplify these results through a nonparametric histogram representation of the triggering kernel. This form of kernel is motivated in the neuroscience context, mimicking the behaviour of action potentials to model the interaction between neurons in the brain. The resulting histogram kernel defined on the compact set $(0,A)$ with J components, change points at $s = (s_0 = 0, s_1, ..., s_{J-1}, s_J = A)$ and respective heights $w = (w_1, ..., w_J)$ such that $\sum_{j=1}^{J} w_j = 1$, has the form,

$$\phi(t|J,w,s) = \delta \sum_{j=1}^{J} \frac{w_j}{s_j - s_{j-1}} \mathbb{1}_{t \in (s_{j-1}, s_j)}. \tag{24}$$

where $\delta \sim \text{Bern}(p)$ is an indicator variable that determines whether the histogram function is active with probability p, or whether the heights for all components is zero. The parameters of this model are inferred using Reversible-jump Markov chain Monte Carlo (RJMCMC), which proves to be a costly procedure. RJMCMC (Green, 1995) is a trans-dimensional approach to Bayesian inference that allows the model to move between different parameter spaces. In this example the various parameter spaces are determined by the possible values of J. A drawback of RJMCMC, as found in this study, is that it is computationally expensive and experiences slower mixing that more standard approaches.

3.2. Gaussian processes

To circumvent the issue of slow inference, several authors have proposed efficient algorithms by instead estimating the model parameters as flexible functions using Gaussian processes. A common feature in all of these approaches is the augmentation of the branching structure to decompose the likelihood function into conditionally independent processes.

Zhang et al. (2019) suggest a practical direction for improvement by proposing a flexible triggering kernel represented as a quadratic transformation of a Gaussian process $f(\cdot)$ given by,

$$\phi(t) = \frac{1}{2} f^2(t). \tag{25}$$

This form is selected as it has certain analytical advantages. With a conjugate Gamma prior on the baseline parameter μ, the adapted Laplace method (Walder and Bishop, 2017) approximates the posterior conditioning on the branching structure, resulting in scalable estimates of theoretical linear time complexity, $O(n)$. An approximated sampling structure (Halpin, 2013) is used to reduce computation by considering only high-probability triggering relationships; this is achieved by assuming that the probability for extremely unlikely triggering relationships is very close to zero. The model is estimated through an EM implementation of both a block Gibbs sampler and MAP estimator. The approach maintains conjugacy relationships due to the decomposition of the likelihood to facilitate a closed form in computation of sequential updates to the model.

In a similar style, Zhou et al. (2021) uses Gaussian processes to represent both the baseline rate and triggering kernel. They also perform a quadratic transformation of a

Gaussian process and further employ a sparse GP approximation (Titsias, 2009) to reduce complexity and avoid costly optimisation procedures. These authors also decouple the baseline rate and triggering kernel by augmenting the branching structure, thereby introducing a fast EM style mean-field variational Bayes algorithm.

Zhou et al. (2020a) instead adopt a sigmoid transformation of a Gaussian process for the baseline rate and triggering kernel, again using a sparse GP approximation. These functions have the form,

$$\mu(t) = \lambda_\mu^* \sigma(f_1(t)), \quad \phi(t) = \lambda_\phi^* \sigma(f_2(t)) \tag{26}$$

where λ_μ^* and λ_ϕ^* are upper bounds of μ and ϕ respectively, $\sigma(\cdot)$ is the sigmoid function and $f_1(\cdot)$ and $f_2(\cdot)$ are generated from a Gaussian process. In addition to augmenting the branching structure, several other processes are introduced to allow for conjugate inference. Several efficient inference schemes are also proposed, namely a Gibbs sampler, an EM algorithm and a mean-field variational inference algorithm. In this experiment all three algorithms performed comparatively well. In this work the sigmoid function is defined as a Gaussian representation, with Polya-Gamma augmentation (Polson, Scott and Windle, 2013),

$$\sigma(z) \;= \frac{e^{z/2}}{2\,\cosh(z/2)} = \int_0^\infty e^{h(\omega,z)} p_{PG}(\omega|b,0)\, d\omega \tag{27}$$

where $h(\omega, z) = z/2 - z^2\omega/2 \;-\; \log 2$ and PG is the Polya-Gamma distribution.

Data augmentation provides a mechanism that eliminates the need to evaluate the high dimensional integral, allowing for efficient conjugate inference. This augmentation strategy with the polya-gamma technique is an interesting development as the likelihood becomes conjugate for the GP prior, thereby allowing for speed compared to other augmentation techniques, and it is an effective method for posterior inference. Malemshinitski, Ojeda and Opper (2022) extend the process further by allowing nonlinear and inhibitory effects in the kernel, ensuring that the intensity is non-negative via a sigmoidal link function. This approach is also computationally efficient, given the new likelihood form and mean-field variational inference; however, it does not rely on the commonly used branching structure.

3.3. Dirichlet approaches

Yet another popular direction Bayesian nonparametric modelling is in incorporating the dynamics of HPs into Dirichlet processes to inform event clustering. This enables capture of the diversity of event types, while the self-exciting process describes the temporal dynamics. The framework of the Dirichlet process means that the number of clusters grows as the complexity of the data increases.

An example of this is the Dirichlet-Hawkes (DHP) model proposed by Du et al. (2015). The authors cluster streams of data, such as news articles and social media content, using a Dirichlet process augmented with a temporal HP to determine the intensity of arrivals. The overarching idea of this work is to determine related actions of media platforms given a particular occurrence of a highly impactful event, through word content and time of occurrence.

The model is a generalisation of the Dirichlet process. Generally, the probability of joining an existing cluster or a new cluster is proportional to the number of observations currently in each cluster or the concentration parameter respectively. The authors instead specify these probabilities as counts that are temporally weighted by the triggering kernel $\phi_{\theta_k}(t-t_i)$ for each existing cluster, or the baseline rate μ for new customers. Hence the baseline rate acts as the concentration parameter in the Dirichlet process.

Let θ_k be the parameters of the bag-of-words document content model for the kth cluster and w_n^v be the vth word in the nth document. Then the model is given by,

$$\begin{aligned} w_n^v &\sim \text{Multinomial}(\theta_k) \\ \theta_k &\sim DHP(\mu, G_0) \\ G_0 &\sim \text{Dirichlet}(\theta_0) \end{aligned}$$

where θ_0 is the concentration parameter for the base distribution in the Dirichlet process.

The choice of algorithm for parameter inference is motivated by the streaming context. A Sequential Monte Carlo framework is used which allows the authors to reuse previous samples. When necessary, duplicate timestamps are resampled as this is a violation of the assumptions of a point process. A Gibbs sampler similar to Neal (2000) is embedded within this framework to sample the cluster labels in the following way. For event at time t_n with cluster allocation s_n,

- Remove t_n from cluster s_n.
- Calculate the probability of t_n belonging to cluster j,

$$p(s_n = j|t_n, \text{rest}) = \begin{cases} \frac{\phi_{\theta_k}(t_n-t_i)}{\mu+\sum_{t_n>t_i}\phi_{\theta_k}(t_n-t_i)} & \text{if } j \text{ occupied,} \\ \frac{\mu}{\mu+\sum_{t_n>t_i}\phi_{\theta_k}(t_n-t_i)} & \text{otherwise.} \end{cases} \tag{28}$$

- Sample cluster allocation for t_n from (28). If j is unoccupied draw θ_j from G_0.

Blundell, Heller and Beck (2012) present another extension on Dirichlet processes for HP. The authors combine HPs with the infinite relational model (IRM) (Xu et al., 2006; Kemp et al., 2006), a graph based approach to modelling the relationship between entities given previously declared relationships. In this model events are represented as vertices on a graph and are clustered according to a Chinese restaurant process (CRP) (Aldous, 1985). Each pair of clusters (in both directions) has a corresponding HP with a parametric form for the conditional intensity function. Let V be the set of events or vertices, π denote the partition of events, and n_j be the number of immigrant events in cluster n_j. For $\lambda^*_{pq}(t)$ the model then has the form,

$$\begin{aligned}
&\pi \sim CRP(\alpha) \\
&\lambda^*_{pq}(t) = \mu_{pq} n_p n_q + \int_0^t \phi_{pq}(t-s) dN_j(s) \; \forall p,q \in \text{range}(\pi) \\
&N_{pq}(\cdot) \sim \text{HP}(\lambda^*_{pq}(\cdot)) \\
&N_{uv}(\cdot) \sim \text{Thinning}(N_{\pi(u)\pi(v)}(\cdot))
\end{aligned}$$

where α is the concentration parameter of the CRP and the thinning process $N_{uv}(\cdot)$ determines the edges of the directed graph by thinning, or distributing, all events in both clusters among the edges such that $N_{pq} = \sum_{u,v} N_{uv}(\cdot)$. Parameter inference is performed using Markov chain Monte Carlo methods as there is no conjugate prior available for this likelihood. The partition of the individuals in the model is updated via a Gibbs sampler, in a similar fashion to Du et al. (2015) with modifications for their model. The remaining model parameters are updated using a slice sampler.

A natural extension of the above approaches is the hierarchical Dirichlet. The inclusion of hierarchies facilitates description of a wide range of phenomena in the data and system under inspection. For example, a hierarchical Dirichlet HP proposed by Markwick (2020) is applied to 5 minute foreign exchange trade data, that is grouped daily for individual day HPs whilst allowing pooling of information where there is less data. The model is able to learn seasonality in trading events simultaneously, with the nonparametric background rate shown as,

$$\begin{aligned}
\mu_d(t) &\sim \mu_0 \cdot f_D(t)\,, \\
f_D(t) &\sim \int k(t|\theta) dG_D(\theta), \\
G_D &\sim DP(\alpha_D, G_0) \\
G_0 &\sim DP(\nu, H)
\end{aligned} \tag{29}$$

where μ_0 and f_D are the amplitude and density of controlling events on a day, respectively, with individual days d grouped by Days, D. The individual Dirichlet process model G_D with base measure for mixing kernel k (beta distributions) is,

$$k(y_i|\theta) = \text{Beta}(y_i|\mu, \upsilon, T) = \frac{y_i^{\frac{\mu\upsilon}{T}-1}(T-y_i)^{\upsilon(1-\frac{\mu}{T}-1)}}{B(\frac{\mu}{T}, \upsilon(1-\frac{\mu}{T}))T^{\upsilon-1}}$$

with non-conjugate prior for the mixture kernel,

$$G_0(\mu, \upsilon|T, \alpha_0, \beta_0) = U(\mu|[0,T])\text{Inv-Gamma}(\upsilon|\alpha_0, \beta_0)$$

and the global Dirichlet process learnt from the data. Augmenting the latent structure and selecting conjugate priors for the model parameters lead to a fully-Gibbs sampling algorithm. The model benefits from the ability of trades being updated in real time (online) and modelling the days of week's trades whilst sharing data amongst all groups with dynamic forecasts.

3.4. Summary

A number of non-parametric Bayesian inference procedures for the HP were reviewed. Computational improvements were highlighted with approximation and estimation strategies giving linear time complexity. Other approaches provided important improvements in flexibility and uncertainty in modelling the kernel. Finally, a scalable online clustering method in Dirichlet Process allows for the number of samples to grow with the HP, while the hierarchical approach supports pooling information and aiding where limited data size is available.

4. Direction 3: Other approaches

This section presents a brief review of three other directions in HP research. These include stochastic differential equations, graphs and neural networks.

In the first direction, Lee et al. (2016) extended the HP model to include randomness of the triggering kernel and introduced contagion parameters to control the levels of excitation. Each level of the excitation function is a stochastic process and is solved using a stochastic differential equation that follows a Geometric Brownian Motion and Exponential Langevin dynamics, inferred through Bayesian methods. The model attempts to better approximate applications where self-excitation intensities are accelerated with correlated levels of contagion.

The second direction points to graph-based approaches. This allows the user to determine the interaction between components within multivariate HPs by recovering the latent network structure. Generally, this is achieved by estimating the infectivity matrix, for which the ijth element describes the expected number of offspring events expected in dimension i given an event in dimension j. Several authors have introduced sparse and low-rank approximations to the matrix to control interactions within the network and improve computational efficiency.

An early example is given by Linderman and Adams (2014). The authors combine HPs with random graph models by decomposing the infectivity matrix into a binary adjacency matrix representing network sparsity, and a weight matrix to model interaction strength. A parallelisable Gibbs sampler is used to infer the model parameters. Guo et al. (2015) augment this approach for uncovering the latent network with a new Bayesian language model to study the evolution of dialogue within a social network over time. Linderman, Wang and Blei (2017) focus on inferring the latent structure of a social network when the data is not fully observed, where several types of missing data are considered. A new sequential Monte Carlo approach is proposed to recover the missing data. Liu et al. (2018) exploit MHP spatio-temporal properties by introducing a graph regularisation method, in which a penalisation term from the proximity of the infectivity matrix to a spatial connection matrix learns the influence among MHP characteristics.

Bacry et al. (2015) provide an extension by introducing a sparsity and low-rank induced penalisation, resulting in an excitation matrix of few non-zero and independent rows. This enhances scalability and improves estimation of the kernel. In a similar vein, Zhou et al. (2013); Bacry et al. (2020) perform inference for higher dimensional HPs by modelling the excitation function as a low-rank approximation with regularised objective functions. The sparsity introduced in the infectivity matrix ensures that individuals are only impacted by a small number of users in the network while a small fraction of hubs can have wide-spread influence. Similarly, the Mean-Field Hypothesis as described by Bacry et al. (2016a) improves computational efficiency when recovering parameters in higher dimensions, given fluctuations of stochastic intensity are small.

In a third direction, the nonlinearity of the intensity function can be modelled as a neural network. Recurrent neural networks encode sequences of input states and output states, where each state is determined by the preceding state and the hidden state captures other past states. The parameters are fitted by an optimisation procedure on a nonlinear function, such as a sigmoidal or hyperbolic tangent.

Improving on the recurrent neural network issues, long short-term memory (LSTM) architecture mitigates the vanishing gradient problem, extending memory by modelling HP intensities of multiple events trained through ‘forget gates’ to control influences of past events on the current state (Mei and Eisner, 2017). Some other neural network approaches are the self-attentive/transformer models (Zuo et al., 2020; Zhang et al., 2020a) and graph convolution networks (Shang and Sun, 2019), showing computational efficiencies and improved prediction accuracy.

Several approaches have also been proposed to model spatio-temporal HPs using neural networks. Okawa et al. (2021) construct the intensity function for HPs to accept images as input by combining convolutional neural networks with continuous convolution kernels to output a multiplicative factor that influences the process in addition to the standard temporal and spatial triggering kernels. An alternative model that relies on neural networks to approximate the conditional intensity function of the HP is proposed by Du et al. (2021). The authors introduce a framework that learns the graph structure of the process which is then combined with temporal and spatial information. There are numerous other variations in neural network approaches as this is a significant body of active research in this area.

4.1. Summary

In this section we presented and discussed three further approaches, namely stochastic differential equations, graphs and neural networks. Although these related fields do not conveniently fit in the previous sections, they highlight the breadth of HPs in different research areas and show significant recent growth.

5. Direction 4: Theoretical guarantees and statistical properties

There is an emerging deep literature on theoretical aspects of HPs. Here we touch on three of these, namely developments in statistical properties, the special case of Markovian HPs, and nonlinear representation of self exciting processes.

With respect to developments in statistical properties, definitions past the first and second order statistics are possible given weakly stationary state conditions (Daley and Vere-Jones, 2003), but they become less intuitive and tractable as their statistical order increases. By introducing a combinatorial formula, Jovanović et al. (2015) allow the integral of cumulants (and consequent moments) to be calculated of arbitrary order for HPs. Specifically, given a set of $s \in \{1,...,D\}$ components and one of times $t_s = \{t_1,...,t_{|s|}\}$, the cumulant density of a HP is defined as

$$k(N^s) = dt^{-|s|}\sum_{\pi}(|\pi|-1!)(-1)^{|\pi|-1}\prod_{B\in\pi}\left\langle \prod_{i\in B} dN_t^i \right\rangle, \tag{30}$$

where the sum runs over all partitions $|\pi|$ in s, $|\cdot|$ denotes the number of blocks and B labels individually the blocks of π. Moments in terms of cumulants are expressed as

$$\left\langle \prod_{i\in s} dN_{t_i}^i \right\rangle dt^{-|s|} = \sum_{\pi}\prod_{B\in\pi} k(N^B). \tag{31}$$

Jovanović et al. (2015) represent HPs as a cluster process, showing how to express (30) as a sum of integral terms by enumerating all possible rooted trees. The contribution of enumerating these ‘family trees’ that represent the complex interactions between point events, can be performed systematically and thus ease computational costs.

In an alternative approach to finding moments, Cui et al. (2020) used elementary derivations of self-exciting processes, setting the objective function to evaluate probabilistic arguments that yield a differential equation for the required moment.

Some other progress made in the direction of asymptotic results is in the study of a special class of HPs that is Markovian. For instance, when the exciting function is exponential, the joint process $(N_t, \lambda t)$ is then Markovian (10). In the paper by Gao and Zhu (2018b), the functional law of large numbers and central limit theorems are derived for the linear HP where the initial intensity and time are large, defined as

$$\lambda(t) := \int_{-\infty}^{t} = \alpha e^{-\beta(t-s)} dN(s) = \lambda_0 \cdot e^{-\beta t} + \int_{0}^{t} \alpha e^{-\beta(t-s)} dN(s) \;\text{ as }\; \lambda_0 = n \to \infty$$

where the process λ is Markovian given $d\lambda(t) = -\beta\lambda(t)dt + \alpha dN(t)$. Such limit theorems (details in Gao and Zhu (2018b)) provide insight into macroscopic behavior of large initial intensity asymptotics of HPs. Furthering the Markovian HPs towards the nonlinear case, a proof for large deviation was obtained by Zhu (2015), where the exciting function is both exponential and a sum of exponentials. More recent work by Kanazawa and Sornette (2020) provides a theoretical framework to embed non-Markovian kernels as Markovian, with the aim of tackling more general and complex derived HP models. This process of introducing auxiliary field variables via a master equation provides a formulation in terms of linear stochastic partial differential equations that are Markovian.

Another direction in theoretical work is the study of nonlinear HPs. For example, Torrisi (2016, 2017) derive explicit bounds in the Gaussian and Poisson approximations on nonlinear HPs using Stein's method and Malliavin calculus. Gao and Zhu (2018a) present a study of a new asymptotic regime and its relation to the mean field limit for higher dimensions. Finally, from the perspective of asymptotic frequentist properties of Bayesian estimators, Donnet et al. (2020) consider nonparametric MHP posterior concentration rate ε_t around the true parameter θ^*,

$$\mathbb{E}_{\theta^*}\left(\prod(d(\theta,\theta^*) > \varepsilon_t|N_t)\right) = o(1) \text{ as } T \to \infty \tag{32}$$

in understanding influential features of the prior. The prior models are defined as a piecewise constant function and a mixture of Beta distributions that is given by,

$$\phi_{ij}(\cdot) = \rho_{ij}\left(\int_0^1 g_{\alpha_{ij}\varepsilon} dM_{ij}(\varepsilon)\right),\ g_{\alpha\varepsilon}(x) = \frac{\Gamma(\alpha/\ (\varepsilon(1-\varepsilon)))}{\Gamma(\alpha/\varepsilon)\ \Gamma(\alpha/(1-\varepsilon))} x^{\frac{\alpha}{1-\varepsilon}-1}(1-x)^{\frac{\alpha}{\varepsilon}\ -1} \tag{33}$$

where M_{ij} are bounded signed measures on $[0,1]$ such that $|M_{kl}| = 1$. The asymptotic posterior concentration rates are derived in stochastic terms and $\mathbb{L}_1$ distances $d(\theta,\theta^*)$. Sulem et al. (2021) furthers theoretical guarantees on estimation methods by considering nonlinear and inhibition effects of MHPs, obtaining the concentration rates of the posterior distribution on the parameters.

5.1. Summary

The theory of HPs is extensive with numerous areas of development. It is not our goal to give a detail account here, rather to provide the reader with three interesting current challenges that researchers are tackling. First we show approaches to nth order cumulant density formula derived in terms of Poisson cluster processes, secondly a number of derived theorems from a special class of the HPs (Markovian) and finally explicit bounds and posterior concentration rates for nonlinear HPs.

6. Real-world cameos

As noted in the Introduction, a key aspect of HP modelling is its suitability to real world applications. Many of the papers discussed in this review motivated and illustrated their methods with substantial examples. Tables 1 and 2 provide a scan of these applications and the corresponding findings, categorised by the estimation and numerical methods described in previous sections. A small number of cameos are described in further detail below.

6.1. Cameo 1: Crime

The issue of refining the parametric form of the triggering kernel is circumvented by a nonparametric approach to parameter estimation. Mohler et al. (2011) introduce a spatio-temporal model for burglaries in Los Angeles. The model, inspired by the ETAS model developed to model seismic activity, is given by

$$\lambda(t,x,y) = \mu_t(t)\mu_b(x,y) + \int_{-\infty}^{t}\int_{\mathcal{X}}\int_{\mathcal{Y}} \phi(t-s,x-u,y-v)dN(s)dN(u)dN(v) \tag{34}$$

where $\mu_t(t)$ and $\mu_b(x,y)$ are temporal and spatial baseline functions, respectively. Model parameters are estimated via variable-bandwidth Kernel Density Estimation (KDE).

A recent extension of this work is the semi-parametric spatiotemporal model employed by Zhuang and Mateu (2019), which describes complexities of criminal behaviors by incorporating their biological clock and periodic social activity. The conditional intensity is defined as

$$\begin{aligned}\lambda(t,x,y) = {} & \mu_0\mu_t(t)\mu_d(t)\mu_w(t)\mu_b(x,y) + \\ & A\int_{-\infty}^{t}\int_{\mathcal{X}}\int_{\mathcal{Y}} \phi_1(t-s)\phi_2(x-u,y-v)dN(s)dN(u)dN(v)\end{aligned} \tag{35}$$

where relaxation coefficients A and μ_0 stabilise the estimation process via maximisation likelihood, giving the model a semiparametric component. The other terms extend the nonparametric MISD model, where the baseline periodicity is estimated via residual analysis with daily/weekly terms μ_d and μ_w, average trend μ_t and spatial background $\mu_b(x,y)$ all normalised to 1. The triggering kernels, both temporal ϕ_1 and spatial ϕ_2, are then normalised as density functions. The introduction of periodic terms and estimation of their relative contributions is used to model crime rates in Castellon, Spain. In addition to uncovering daily and weekly patterns in robberies, the authors' analysis reveals the high influence of the background rate compared to the clustering effect which explains roughly 3% of the overall intensity.

6.2. Cameo 2: Finance

Kirchner (2017) shows the close relation of HPs to an Integer Auto-Regression (INAR) where the distribution of the resulting bin count sequence is approximated as a multivariate INAR(p). Fitting a mutually exciting bivariate HP to trades and limit orders on S&P 500, Kirchner (2017) determines an asymmetric relationship between both incoming orders exciting limit order and market orders, and finds that market order has barely an effect on incoming limit order.

In further support of high frequency applications, the Hawkes Graphs approach by Embrechts and Kirchner (2018) efficiently fits dozens of event streams. This method also provides a natural approach to studying connectivity and causality.

The suitability of the nonparametric HP method to very large datasets was also demonstrated by Bacry, Jaisson and Muzy (2016b). In the approach taken by these authors, a series of Wiener-Hopf equations is solved by Gaussian quadrature to estimate the kernel matrix, where market orders of two future assets on EUREX were shown to closely fit a power law function. Rambaldi, Bacry and Lillo (2017) couples this nonparametric kernel estimation with a MHP to successfully show the complex interactions between time of arrival of orders in limit order books (LOB) and their size. Their work highlights the fact that high frequency orders on EUREX exchange are not suitable to be described with a simple model assuming independence between volume and time.

6.3. Cameo 3: Online content

The model proposed by Du et al. (2015), summarised in Section 3.3, was applied by the authors to a stream of news articles for a 35 day period at the beginning of 2011. The aim of the model is to identify emerging news stories by clustering related news articles based on the terms used in each article.

To determine the words included in the vocabulary of the model, named entities are identified and words that do not add information to the text are pruned, leaving a vocabulary of terms consisting largely of named entities, nouns, verbs and adjectives. The triggering kernel is made up of a linear combination of known radial basis function kernels. These kernels assign mass to the excitation function based on the distance between particular reference time points and the time elapsed for a pair of events. In this study the reference time points range from 30 minutes to 168 hours, capturing a range of both short and long time excitation effects. A number of meaningful news stories were identified as clusters, including the 2011 shooting in Tuscan, the release of the film 'Dark Knight Rises', the space shuttle Endeavour's last mission and cyclone Yasi in Queensland, Australia. A key outcome for this work is the ability to track the trend of each of these stories through examining both the form of the triggering kernel and the level of overall excitation.

‘

Table 1. *Recent research applications in nonparametric HPs. i.e. traffic incidents, financial markets, crime, memes and epidemiology.*

Researchers	Date	Application	Type/Method	Results
* Zhou F, Luo S, Li Z et al.	2021	NY vehicle incidents	HP, MISD, EM, VI	Flexible baseline and kernel avoids overfitting and improves on vehicle collision predictions.
** Kalair K, Connaughton C and Di Loro P	2021	UK traffic incidents	HP, SP, MISD	Self-excitation shown to account for 6-7% of observed secondary incidents.
* Markwick D	2020	FX trades	HP, EM, MCMC	Hierarchical model accounts natural daily trading, with real time updating and dynamic forecasting.
* Donnet S, Rivoirard V and Rousseau J	2020	Neuronal	MHP, MCMC	Recovers interaction graphs of neurons activity, simulating action potentials.
** Park J, Chaffeea A, Harrigan R, Schoenberg F	2020	Ebola spread west Africa	HP, MISD	Utility of HPs as alternative approach in forecasting epidemic spread, showing improved RMSE on SEIR models.
** Zhuang J and Mateu J	2019	Spain's crime	HP, SP, MISD	3% of crime explained through clustering and daily/weekly periodicity activity.
* Zhang R, Walder C, Rizoiu M and Xie L	2019	Twitter meme	HP, MCMC, EM	Captures and compares categories of contents given decaying tweets and measure diffusion in linear time complexity.
** Schoenberg F, Gordon J and Harrigan R	2018	US plague spread	HP, MISD	Estimated contagion time 0-7.5 days with fitted model improving on computation and kernel estimates.
** Nichols K, Trevino E, Ikeda N et al.	2018	California earthquakes	HP, EM, MISD, ETAS	Adapted model for smaller magnitude earthquakes, accounting for more varied spatial and temporal features.
** Achab M, Bacry E, Gaiffas S et al.	2018	Meme tracking	MHP, MO	Shows an improvement on existing methods in relative error (7%), rank correlation and computing time.
** Embrechts P and Kirchner M	2018	FX trades	MHP, AR	The Hawkes Graph models supports multi-type high frequency streaming data types and causality analysis.
* Seonwoo Y, Park S and Oh A	2018	New York Times articles	HP, SMC	Texts in news articles are reconstructed for narratives and thread structures from the New York times, highlighting algorithm's performance efficiency.

Intensity functions categories
* Bayesian nonparametric
** Frequentist nonparametric

Type/Method Acronyms
(HP)Univariate Hawkes process, (MHP)Multivariate Hawkes process, (SP)Spatial, (MISD)Model Independent Stochastic Declustering, (WH)Wiener-Hopf, (ETAS) Epidemic Type Aftershock Shock, (AR)Auto Regressive, (MO)Moments, (GD)Gradient descent, (MAP)Maximum a posterior, (EM)Expectation Maximisation, (VI)Variational Inference, (MCMC) Markov chain Monte Carlo, (SMC)Sequential Monte Carlo

Table 2. *Continued. Recent research applications in nonparametric HPs.*

Researchers	Date	Application	Type/Method	Results
∗∗ Yang Y, Etesami J, He N and Kiyavash N	2017	Meme tracking	MHP, RKHS	Popular phrases on news agencies are model in a scalable and online learning efficient algorithm, O(log T).
∗∗ Kirchner M	2017	FX trades	MHP, AR	Fits bivariate HPs on trades and limit order to LOB E-mini S&P 500, asymmetric relation shown between components.
∗∗ Rambaldi M, Bacry E and Lillo F	2017	FX trades	MHP, WH	Analysing interaction between time and arrival of orders (in LOB) and their size, shows unsuitability of independence between volume and time with simpler models.
∗ Xu H and Zha H	2017	ICU signal analysis	MHP, EM, MCMC	HP sequence clustering on pattern recognition and signal processing for clinical decision-making in the intensive care unit.
∗∗ Bacry E, Jaisson T and Muzy J	2016b	FX trades	MHP, WH	Models market orders arrivals on EUREX exchange showing power-law like shape and suitability to larger datasets.
∗∗ Hall E and Willett R	2016	Meme tracking	MHP, GD	Estimates underlying network of posts with scalable and online algorithm of O(n^2) complexity.
∗ Mavroforakis C, Valera I and Gomez-Rodriguez M	2016	Online user activity	HP, SMC	Learns users patterns on Stack Overflow through tracing online activity, grouped streaming data with hierarchical Dirichlet HPs.
∗ Fox EW, Schoenberg, FP and Gordon JS	2016	Japan's earthquakes	HP, SP, ETAS	Incorporated histogram estimator and variable bandwidth kernel improves seismicity model and provides support as a powerful exploratory tool.
∗ Linderman S and Adams R	2015	Chicago gang violence	HP, SP, MCMC	Spatial Gaussian mixture model to predict gang related homicide that exhibit mutually exciting properties.
∗ Du N, Farajtabar M, Ahmed A et al.	2015	News	HP, MCMC, MAP	Uncovering topic specific clusters with learned lantent temporal dynamics showing an intuitive way in tracking trends online.
∗ Blundell C, Heller K and Beck J	2012	Social networks	MHP, MCMC	Inferred social network structure based on email interaction threads.

Intensity functions categories
∗ Bayesian nonparametric
∗∗ Frequentist nonparametric

Type/Method Acronyms
(HP)Univariate Hawkes process, (MHP)Multivariate Hawkes process, (SP)Spatial,
(MISD)Model Independent Stochastic Declustering, (WH)Wiener-Hopf,
(ETAS) Epidemic Type Aftershock Shock, (AR)Auto Regressive, (MO)Moments,
(GD)Gradient descent, (MAP)Maximum a posterior, (EM)Expectation Maximisation,
(VI)Variational Inference, (MCMC) Markov chain Monte Carlo, (SMC)Sequential Monte Carlo

7. Conclusion and Challenges

The past fifty years has seen the HP embedded as a staple methodology in the statistical literature. The growth in research directions inspired by the HP is itself a HP! Even after half a century, this pursuit continues through new theoretical, methodological and computational developments and new applications. The papers referenced in this review were selected to highlight some of the current directions in these areas and to provide a broad overview for new readers in the field. A range of research directions, in particular parametric, nonparametric, online and Bayesian approaches, were highlighted along with a number of real-world applications. The quantity and quality of the work reviewed here, and the large body of literature that was unfortunately not included, are a portent for another fifty years of exciting research related to HPs.

Acknowledgement

This research was supported by the ARC Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS) and the Australian Research Council (ARC) Laureate Fellowship Program under the project 'Bayesian Learning for Decision Making in the Big Data Era' (ID: FL150100150).

References

Achab, M., Bacry, E., Gaiffas, S., Mastromatteo, I. and Muzy, J. F. (2018). Uncovering causality from multivariate hawkes integrated cumulants. *Journal of Machine Learning Research*, *18* (2016), 1–28.

Aldous, D. J. (1985). Exchangeability and related topics. *École d'Été de Probabilités de Saint-Flour XIII—1983 1–198. Lecture Notes in Math.* 1117. Springer, Berlin.

Bacry, E., Dayri, K. and Muzy, J. F. (2012). Non-parametric kernel estimation for symmetric Hawkes processes. Application to high frequency financial data. *European Physical Journal B*, *85* (5).

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

Bacry, E. and Muzy, J. F. (2016). First- and second-order statistics characterization of hawkes processes and non-parametric estimation. *IEEE Transactions on Information Theory*, 62 (4), 2184–2202.

Bacry, E., Gaïffas, S., Mastromatteo, I. and Muzy, J. F. (2016a). Mean-field inference of Hawkes point processes. *Journal of Physics A: Mathematical and Theoretical*, *49* (17).

Bacry, E., Jaisson, T. and Muzy, J. F. (2016b). Estimation of slowly decreasing Hawkes kernels: application to high-frequency order book dynamics. *Quantitative Finance*, *16* (8), 1179–1201.

Bacry, E., Bompaire, M., Gaïffas, S. and Muzy, J. F. (2020). Sparse and low-rank multivariate Hawkes processes. *Journal of Machine Learning Research*, *21*, 1–32.

Blundell, C., Heller, K. A. and Beck, J. M. (2012). Modelling reciprocating relationships with Hawkes processes. *Advances in Neural Information Processing Systems*, *4*, 2600–2608.

Broderick, T., Boyd, N., Wibisono, A., Wilson, A. C. and Jordan, M. I. (2013). Streaming variational Bayes. *Advances in Neural Information Processing Systems*, 1–9.

Browning, R., Sulem, D., Mengersen, K., Rivoirard, V. and Rousseau, J. (2021). Simple discrete-time self-exciting models can describe complex dynamic processes: A case study of COVID-19. *PLOS ONE*, *16 (4 April),* 1–28.

Chen, J., Hawkes, A. G. and Scalas, E. (2021). A Fractional Hawkes Process. *SEMA SIMAI Springer Series*, *26*, 121–131.

Chérief-Abdellatif, B. E., Alquier, P. and Khan, M. E. (2019). A generalization bound for online variational inference. *Asian Conference on Machine Learning*, 662–677.

Chiang, W. H., Liu, X. and Mohler, G. (2021). Hawkes process modeling of COVID-19 with mobility leading indicators and spatial covariates. *International Journal of Forecasting*, (40).

Cox, D. R. (1955). Some statistical methods connected with series of events. *Journal of the Royal Statistical Society: Series B (Methodological)*, *17* (2), 129–157.

Cui, L., Hawkes, A. and Yi, H. (2020). An elementary derivation of moments of Hawkes processes. *Advances in Applied Probability*, *52* (1), 102–137.

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

Donnet, S., Rivoirard, V. and Rousseau, J. (2020). Nonparametric Bayesian estimation for multivariate Hawkes processes. *The Annals of statistics*, 2698–2727.

Du, H., Zhou, Y., Ma, Y. and Wang, S. (2021). Astrologer: Exploiting graph neural Hawkes process for event propagation prediction with spatio-temporal characteristics. *Knowledge-Based Systems*, *228*, 107247.

Du, N., Dai, H., Trivedi, R., Upadhyay, U., Gomez-Rodriguez, M. and Song, L. (2016). Recurrent marked temporal point processes: Embedding event history to vector. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17 August*, 1555–1564.

Du, N., Farajtabar, M., Ahmed, A., Smola, A. J. and Song, L. (2015). Dirichlet-Hawkes processes with applications to clustering continuous-time document streams. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *2015-August*, 219–228.

Embrechts, P. and Kirchner, M. (2018). Hawkes graphs. *Theory of Probability and its Applications*, *62* (1), 132–155.

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

Gao, F. and Zhu, L. (2018a). Some asymptotic results for nonlinear Hawkes processes. *Stochastic Processes and their Applications*, *128* (12), 4051–4077.

Gao, X. and Zhu, L. (2018b). Limit theorems for Markovian Hawkes processes with a large initial intensity. *Stochastic Processes and their Applications*, *128* (11), 3807–3839.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, *82*, 711–732.

Guo, F., Blundell, C., Wallach, H. and Heller, K. (2015). The Bayesian echo chamber: Modeling social influence via linguistic accommodation. *Proceedings of the 18*$^{\text{th}}$ *International Conference on Artificial Intelligence and Statistics (AISTATS)2015, San Diego, CA, USA. JMLR: W&CP, volume 38*.

Hall, E. C. and Willett, R. M. (2016). Tracking dynamic point processes on networks. *IEEE Transactions on Information Theory*, *62* (7), 4327–4346.

Halpin, P. F. (2013). An EM algorithm for Hawkes process. *New Developments in Quantitative Psychology: Proceedings of the 77th International Meeting of the Psychometric Society*, (212).

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

Holbrook, A. J., Ji, X. and Suchard, M. A. (2022). From viral evolution to spatial contagion: a biologically modulated Hawkes model. *Bioinformatics (Oxford, England)*, *38* (7), 1846–1856.

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

Jovanović, S., Hertz, J. and Rotter, S. (2015). Cumulants of Hawkes point processes. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, *91* (4).

Kalair, K., Connaughton, C. and Di Loro, P. A. (2021). A non-parametric Hawkes process model of primary and secondary accidents on a UK smart motorway. *Journal of the Royal Statistical Society Series C, vol. 70* (1), 80-97.

Kanazawa, K. and Sornette, D. (2020). Field master equation theory of the self-excited Hawkes process. *Physical Review Research*, *2* (3), 33442.

Kelly, J. D., Park, J., Harrigan, R. J., Hoff, N. A., Lee, S. D., Wannier, R., Selo, B., Mossoko, M., Njoloko, B., Okitolonda-Wemakoy, E., Mbala-Kingebeni, P., Rutherford, G. W., Smith, T. B., Ahuka-Mundeke, S., Muyembe-Tamfum, J. J., Rimoin, A. W. and Schoenberg, F. P. (2019). Real-time predictions of the 2018–2019 Ebola virus disease outbreak in the Democratic Republic of the Congo using Hawkes point process models. *Epidemics*, *28*, 100354.

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T. and Ueda, N. (2006). Learning systems of concepts with an infinite relational model. *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*, 381–388.

Kirchner, M. (2017). An estimation procedure for the Hawkes process. *Quantitative Finance*, *17* (4), 571–595.

Kobayashi, R. and Lambiotte, R. (2016). TiDeH: Time-dependent Hawkes process for predicting retweet dynamics. *Tenth International AAAI Conference on Web and Social Media*.

Laub, P. J., Lee, Y. and Taimre, T. (2021). The elements of Hawkes processes. *Springer International Publishing. Cham.*

Lee, Y., Lim, K. W. and Ong, C. S. (2016). Hawkes processes with stochastic excitations. *33^{rd} International Conference on Machine Learning, ICML 2016*, *1*, 132–145.

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

Li, Z., Cui, L. and Chen, J. (2018). Traffic accident modelling via self-exciting point processes. *Reliability Engineering and System Safety*, *180* (July), 312–320.

Linderman, S. W., and Adams, R. P. (2014). Discovering latent network structure in point process data. *31st International Conference on Machine Learning, ICML 2014*, *4*, 3268–3281.

Linderman, S. W., and Adams, R. P. (2015). Scalable Bayesian inference for excitatory point process networks. *Computer Science*. 1–16.

Linderman, S. W., Wang, Y., and Blei, D. M. (2017). Bayesian inference for latent Hawkes processes. *Advances in Approximate Bayesian Inference Workshop at the 31st Conference on Neural Information Processing Systems.*

Liu, Y., Yan, T. and Chen, H. (2018). Exploiting graph regularized multi-dimensional Hawkes processes for modeling events with spatio-temporal characteristics. *IJCAI International Joint Conference on Artificial Intelligence*, *2018-July*, 2475–2482.

Malem-shinitski, N., Ojeda, C. and Opper, M. (2022). Variational Bayesian inference for non-linear Hawkes process with Gaussian process self-effects. *Entropy, 24(3):* 356.

Markwick, D. (2020). Bayesian nonparametric Hawkes processes with applications. Doctoral thesis (Ph.D), UCL (University College London).

Marsan, D. and Lengliné, O. (2008). Extending earthquakes' reach through cascading. *Science*, *319* (5866), 1076–1079.

Mavroforakis, C., Valera, I. and Rodriguez, M. G. (2017). Modeling the dynamics of online learning activity. *In Proceedings of the 26th International World Wide Web Conference, 2017.*

Mei, H. and Eisner, J. (2017). The neural Hawkes process: A neurally self-modulating multivariate point process. *Advances in Neural Information Processing Systems*, *2017- Decem* (Nips), 6755–6765.

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

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, *9* (2), 249–265.

Nichols, K., Trevino, E., Ikeda, N., Philo, D., Garcia, A. and Bowman, D. (2018). Interdependency amongst earthquake magnitudes in Southern California. *Journal of Applied Statistics*, *45* (4), 763–774.

Ogata, Y. (1981). On Lewis' simulation method for point processes. *IEEE Transactions on Information Theory*, *27* (1), 23–31.

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of Computational and Graphical Statistics*, *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.

Okawa, M., Iwata, T., Tanaka, Y., Toda, H., Kurashima, T. and Kashima, H. (2021). Dynamic Hawkes processes for discovering time-evolving communities' states behind diffusion processes. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21).*

Ozaki, T. (1979). Maximum likelihood estimation of hawkes' self-exciting point processes. *Annals of the Institute of Statistical Mathematics*, *31* (1), 145–155.

Park, J., Chaffee, A. W., Harrigan, R. J. and Schoenberg, F. P. (2020). A non-parametric Hawkes model of the spread of Ebola in west Africa. *Journal of Applied Statistics, 49* (3), 621-6371.

Polson, N. G., Scott, J. G. and Windle, J. (2013). Bayesian inference for logistic models using Pólya-Gamma latent variables. *Journal of the American Statistical Association*, *108* (504), 1339–1349.

Porter, M. D. and White, G. (2010). Self-exciting hurdle models for terrorist activity. *Annals of Applied Statistics*, *4* (1), 106–124.

Rambaldi, M., Bacry, E. and Lillo, F. (2017). The role of volume in order book dynamics: a multivariate Hawkes process analysis. *Quantitative Finance*, *17* (7), 999–1020.

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

Rasmussen, J. G. (2018). Lecture Notes: Temporal point processes and the conditional intensity function. *arXiv:1806.00221v1.*

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

Schoenberg, F. P. (2016). A note on the consistent estimation of spatial-temporal point process parameters. *Statistica Sinica*, *26*, 861-879.

Schoenberg, F. P., Gordon, J. S. and Harrigan, R. J. (2018). Analytic computation of nonparametric Marsan–Lengliné estimates for Hawkes point processes. *Journal of Nonparametric Statistics*, *30* (3), 742–757.

Seonwoo, Y., Oh, A. and Park, S. (2018). Hierarchical Dirichlet Gaussian marked Hawkes process for narrative reconstruction in continuous time domain. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3316–3325.

Shang, J. and Sun, M. (2019). Geometric Hawkes processes with graph convolutional recurrent neural networks. *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, 4878–4885.

Sulem, D., Rivoirard, V. and Rousseau, J. (2021). Bayesian estimation of nonlinear Hawkes process. https://arxiv.org/abs/2103.17164v2

Titsias, M. K. (2009). Variational learning of inducing variables in sparse Gaussian processes. *Artificial intelligence and statistics*, *5*, 567–574.

Torrisi, G. L. (2016). Gaussian approximation of nonlinear Hawkes processes. *Annals of Applied Probability*, *26* (4), 2106–2140.

Torrisi, G. L. (2017). Poisson approximation of point processes with stochastic intensity, and application to nonlinear Hawkes processes. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, *53* (2), 679–700.

Walder, C. J. and Bishop, A. N. (2017). Fast Bayesian intensity estimation for the permanental process. *34th International Conference on Machine Learning, ICML 2017*, *7*, 5459– 5471.

White, G., Porter, M. D. and Mazerolle, L. (2012). Terrorism Risk, Resilience and Volatility: A Comparison of Terrorism Patterns in Three Southeast Asian Countries. *Journal of Quantitative Criminology*, *29* (2), 295–320.

Xu, H. and Zha, H. (2017). A Dirichlet mixture model of Hawkes processes for event sequence clustering. *Advances in Neural Information Processing Systems*, *2017-December* (Nips), 1355–1364.

Xu, Z., Tresp, V., Yu, K. and Kriegel, H.-P. (2006). Infinite hidden relational models. *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, 544–551.

Yang, Y., Etesami, J., He, N. and Kiyavash, N. (2017). Online learning for multivariate Hawkes processes. *Advances in Neural Information Processing Systems*, *2017-December* (2), 4938–4947.

Yang, Y., Etesami, J., He, N. and Kiyavash, N. (2018). Nonparametric Hawkes processes: online estimation and generalization bounds. (Nips 2017), 1–39.

Zhang, R., Walder, C., Rizoiu, M. A. and Xie, L. (2019). Efficient non-parametric Bayesian Hawkes processes. *IJCAI International Joint Conference on Artificial Intelligence*, *2019-Augus*, 4299–4305.

Zhang, Q., Lipani, A., Kirnap, O. and Yilmaz, E. (2020a). Self-attentive Hawkes process. *37th International Conference on Machine Learning, ICML 2020*, *PartF16814*, 11117– 11127.

Zhang, R., Walder, C. and Rizoiu, M.-A. (2020b). Variational inference for sparse Gaussian process modulated Hawkes process. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34* (04), 6803–6810.

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

Zhou, F., Li, Z., Fan, X., Wang, Y., Sowmya, A. and Chen, F. (2020b). Fast multiresolution segmentation for nonstationary Hawkes process using cumulants. *International Journal of Data Science and Analytics*, *10* (4), 321–330.

Zhou, F., Luo, S., Li, Z., Fan, X., Wang, Y., Sowmya, A. and Chen, F. (2021). Efficient EM-variational inference for nonparametric Hawkes process. *Statistics and Computing*, *31* (4), 1–11.

Zhou, K., Zha, H. and Song, L. (2013). Learning social infectivity in sparse low-rank networks using multi-dimensional Hawkes processes. *Journal of Machine Learning Research*, *31*, 641–649.

Zhu, L. (2015). Large deviations for Markovian nonlinear Hawkes processes. *Annals of Applied Probability*, *25* (2), 548–581.

Zhuang, J., Ogata, Y. and Vere-Jones, D. (2002). Stochastic declustering of space-time earthquake occurrences. *Journal of the American Statistical Association*, *97* (458), 369–380.

Zhuang, J. and 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), 919–942.

Zuo, S., Jiang, H., Li, Z., Zhao, T. and Zha, H. (2020). Transformer Hawkes process. *37th International Conference on Machine Learning, ICML 2020*, *PartF16814*, 11628–11638.