Journal of Machine Learning Research 21 (2020) 1-31 Submitted 11/19; Revised 9/20; Published 12/20

Efficient Inference for Nonparametric Hawkes Processes Using Auxiliary Latent Variables

Feng Zhou[1,2] FENG.ZHOU@DATA61.CSIRO.AU
Zhidong Li[3] ZHIDONG.LI@UTS.EDU.AU
Xuhui Fan[2] XUHUI.FAN@UNSW.EDU.AU
Yang Wang[3] YANG.WANG@UTS.EDU.AU
Arcot Sowmya[2] A.SOWMYA@UNSW.EDU.AU
Fang Chen[3] FANG.CHEN@UTS.EDU.AU

[1] *Data61 CSIRO, 13 Garden Street, Eveleigh, New South Wales, Australia*
[2] *University of New South Wales, Kensington, New South Wales, Australia*
[3] *University of Technology Sydney, Ultimo, New South Wales, Australia*

Editor: Mohammad Emtiyaz Khan

Abstract

The expressive ability of classic Hawkes processes is limited due to the parametric assumption on the baseline intensity and triggering kernel. Therefore, it is desirable to perform inference in a data-driven, nonparametric approach. Many recent works have proposed nonparametric Hawkes process models based on Gaussian processes (GP). However, the likelihood is non-conjugate to the prior resulting in a complicated and time-consuming inference procedure. To address the problem, we present the sigmoid Gaussian Hawkes process model in this paper: the baseline intensity and triggering kernel are both modeled as the sigmoid transformation of random trajectories drawn from a GP. By introducing auxiliary latent random variables (branching structure, Pólya-Gamma random variables and latent marked Poisson processes), the likelihood is converted to two decoupled components with a Gaussian form which allows for an efficient conjugate analytical inference. Using the augmented likelihood, we derive an efficient Gibbs sampling algorithm to sample from the posterior; an efficient expectation-maximization (EM) algorithm to obtain the maximum a posteriori (MAP) estimate and furthermore an efficient mean-field variational inference algorithm to approximate the posterior. To further accelerate the inference, a sparse GP approximation is introduced to reduce complexity. We demonstrate the performance of our three algorithms on both simulated and real data. The experiments show that our proposed inference algorithms can recover well the underlying prompting characteristics efficiently.

Keywords: Hawkes process, Gaussian process, Pólya-Gamma distribution, conjugacy

1. Introduction

The *self-excitation* is a common phenomenon in numerous applications, e.g. in seismology one shock will prompt aftershocks (Hawkes, 1973); in social media a tweet posted by a star may be shared by the followers of the original poster through retweeting (Chen and Tan, 2018). More similar application domains cover crime (Liu et al., 2018), ecosystem (Gupta et al., 2018), transportation (Du et al., 2016) and finance (Bacry et al., 2015). The Hawkes process (Hawkes, 1971) is one important class of point processes which can be utilized to

model the self-exciting phenomenon. An important characteristic of point processes is the conditional intensity: the probability of one event occurring in an infinitesimal time interval given history. Specifically, the conditional intensity of Hawkes process is

$$\lambda(t) = \mu(t) + \int_0^t \phi(t-s) d\mathbb{N}(s) = \mu(t) + \sum_{t_i<t} \phi(t-t_i), \tag{1}$$

where $\mu(t) > 0$ is the baseline intensity, $\{t_i\}$ are timestamps of events occurring before t, $\mathbb{N}(t)$ is the corresponding counting process and $\phi(\tau) > 0$ where $\tau = t - t_i$ is the triggering kernel. Each observation t_i could either arise independently due to the baseline intensity (exogenous) or because of the exciting effect of previous observations via the triggering kernel (endogenous). The summation of triggering kernels explains the nature of self-excitation: events occurring in the past intensify the rate of occurrence in the future.

The classic Hawkes process is supposed to be in a parametric form: the baseline intensity $\mu(t)$ is assumed to be a constant with triggering kernel $\phi(\tau)$ being a parametric function, e.g. exponential decay or power law decay function. However, in reality, the actual exogenous rate μ can change over time due to the varying exterior context; the actual endogenous rate capturing how previous events trigger subsequent ones, which is modeled by $\phi(\tau)$, can be rather flexible among different applications. For example, the exogenous rate of civilian deaths due to insurgent activity is changing over time (Lewis and Mohler, 2011) and the prompting effect of vehicle collision decays periodically and in an oscillatory way (Zhou et al., 2018). Obviously, the models based on the classic Hawkes process tend to be oversimplified or even incapable of capturing the ground truth in numerous scenarios (refer to the real data experiment for an investigation of such an instance). Therefore, it is desirable to estimate the exogenous and endogenous dynamics in a data-driven, nonparametric approach.

A wide variety of nonparametric estimation approaches of Hawkes process have been largely investigated over past few years. From frequentist nonparametric perspective, Marsan and Lengline (2008) proposed to estimate the triggering kernel modeled as a histogram function with an EM algorithm and Lewis and Mohler (2011) extended this method by introducing a smooth regularizer and performed estimation by solving a Euler-Lagrange equation, Zhou et al. (2013) further extended this algorithm to multivariate Hawkes process; Bacry and Muzy (2016) provided an estimation approach that is based on the solution of a Wiener-Hopf equation relating the triggering kernel with the second order statistics; Eichler et al. (2017) and Reynaud-Bouret and Schbath (2010) attempted to minimize a quadratic contrast function with a grid based triggering kernel. From Bayesian nonparametric perspective, most related works are based on Gaussian-Cox processes: the Poisson process with a stochastic intensity modulated by GP. To guarantee the non-negativity of the intensity, trajectories drawn from a GP prior need to be squashed by a link function. For example, a log-Gaussian intensity is utilized by Møller et al. (1998) and Samo and Roberts (2015); Adams et al. (2009) proposed a sigmoid-GP intensity and a tractable Markov chain Monte Carlo (MCMC) algorithm. Lloyd et al. (2015) developed a variational Gaussian approximation algorithm with a square link function. Flaxman et al. (2017) designed a reproducing kernel Hilbert space (RKHS) formulation to estimate the intensity. As far as we know, only a small amount of works attempted to infer the Hawkes process with a GP prior since the Hawkes process is more complicated than the Poisson process. For example,

work	model	nonparametric	link function	conjugacy	inference
Adams et al. (2009)	Poisson	GP for intensity	sigmoid	×	MCMC
Samo and Roberts (2015)	Poisson	GP for intensity	exponential	×	MCMC
Lloyd et al. (2015)	Poisson	GP for intensity	square	×	variational inference
Flaxman et al. (2017)	Poisson	RKHS for intensity	square	-	RKHS
Zhang et al. (2018)	Hawkes	RKHS for only $\phi(\tau)$	square	-	RKHS
Zhang et al. (2019)	Hawkes	GP for only $\phi(\tau)$	square	×	variational inference
Zhou (2019)	Hawkes	GP for both $\mu(t)$ and $\phi(\tau)$	square	×	variational inference
our work	Hawkes	GP for both $\mu(t)$ and $\phi(\tau)$	sigmoid	✓	Gibbs sampler/EM/mean field

Table 1: The differences between some recent Bayesian nonparametric point process models and our model. All works are sorted by time.

Zhou et al. (2018) added an extra GP regression step into the EM algorithm by Marsan and Lengline (2008) to achieve the smoothness and facilitate the choice of hyperparameters; Zhang et al. (2018) extended the approach in Flaxman et al. (2017) to the Hawkes process where the triggering kernel is modeled as the square transformation of the trajectory drawn from the RKHS; Zhang et al. (2019) extended the variational inference algorithm by Lloyd et al. (2015) to the Hawkes process where the triggering kernel is a square transformation of a GP; Zhou (2019) further extended the approach in Lloyd et al. (2015) to model both the flexible baseline intensity and triggering kernel simultaneously. To be more specific, we provide Tab.1 to show the differences between these Bayesian nonparametric models and our model.

All GP modulated intensity models mentioned above have the same issues: **1**) due to the existence of link function, the likelihood of GP variables is non-conjugate to the prior resulting in a non-Gaussian posterior. The non-conjugacy leads to a complicated and time-consuming inference procedure. **2**) Furthermore, in the Hawkes process, the exogenous component (baseline intensity) and the endogenous component (triggering kernel) are coupled in the likelihood, which further hampers the tractability of inference.

To circumvent these problems, we augment the likelihood with auxiliary latent random variables: *branching structure*, *Pólya-Gamma random variables* and *latent marked Poisson processes*. The branching structure of Hawkes process is introduced to decouple $\mu(t)$ and $\phi(\tau)$ to two independent components in the likelihood; inspired by Polson et al. (2013), Donner (2019) and Donner and Opper (2018), we use a sigmoid link function in the model and convert the sigmoid to an infinite mixture of Gaussians involving Pólya-Gamma random variables; the latent marked Poisson processes are augmented to linearize the exponential integral term in likelihood. By augmenting the likelihood in such a way, the likelihood becomes conjugate to the GP prior. With these latent random variables, we use the augmented likelihood to construct three efficient analytical iterative algorithms. The first one is a Gibbs sampler which accurately characterizes the posterior with the second one being an EM algorithm to obtain the MAP estimate; furthermore, we extend the EM algorithm to a mean-field variational inference algorithm that provides an approximated posterior distribution rather than point estimation. It is worth noting that the naïve implementations of three algorithms are time-consuming. To improve the efficiency remarkably, the sparse GP approximation (Titsias, 2009) is introduced.

Specifically, we make the following contributions:

1. We propose the sigmoid Gaussian Hawkes process model wherein the baseline intensity and triggering kernel are both sigmoid-GP rates. The original Hawkes process likelihood is converted to two decoupled factors which are conjugate to GP priors by augmenting the branching structure, Pólya-Gamma random variables and latent marked Poisson processes.

2. Three simple and efficient iterative algorithms: a Gibbs sampler, an EM algorithm and a mean-field variational inference algorithm, are derived with closed-form expressions for the Bayesian nonparametric Hawkes process wherein both baseline intensity and triggering kernel are nonparametric.

3. All three algorithms are efficient because of the closed-form expressions. Moreover, they are further accelerated by the utilization of sparse GP approximation.

Our paper is organized as follows. In Section 2 we present the model of sigmoid Gaussian Hawkes process and how the likelihood is augmented with branching structure, Pólya-Gamma random variables and latent marked Poisson processes. An efficient Gibbs sampler is proposed in Section 3, EM algorithm in Section 4 and mean-field variational inference algorithm in Section 5. In Section 6 we propose some numerical insight on the simulated and real data experiments. In Section 7 we analyze the advantages and disadvantages of each algorithm and the most appropriate application scenario for each algorithm and discuss the relationship among our proposed algorithms, then draw a conclusion and prospect the future research direction in the end.

2. Sigmoid Gaussian Hawkes Process

A Hawkes process is a stochastic process whose realization is a sequence of timestamps $D = \{t_i\}_{i=1}^N \in [0, T]$. Here, t_i stands for the occurrence time of i-th event with T being the observation window. The conditional intensity of Hawkes process is already provided in Eq.(1). Given $\mu(t)$ and $\phi(\tau)$, the Hawkes process likelihood (Daley and Vere-Jones, 2003) is

$$p(D|\mu(t), \phi(\tau)) = \prod_{i=1}^{N} \left[\mu(t_i) + \sum_{t_j<t_i} \phi(t_i - t_j) \right] \cdot \exp\left(- \int_T \left(\mu(t) + \sum_{t_i<t} \phi(t - t_i) \right) dt \right). \quad (2)$$

We propose a GP based Bayesian nonparametric Hawkes process model: sigmoid Gaussian Hawkes process (SGHP) whose baseline intensity and triggering kernel are functions drawn from a GP prior, passed through a sigmoid link function to guarantee non-negativity and then scaled by an upper-bound: $\mu(t) = \lambda_\mu^* \sigma(f(t)), \phi(\tau) = \lambda_\phi^* \sigma(g(\tau))$ where $\sigma(\cdot)$ is the sigmoid function, f and g are two functions drawn from the corresponding GP priors, λ_μ^* and λ_ϕ^* are the upper-bounds of $\mu(t)$ and $\phi(\tau)$.

In a naïve Bayesian framework, the inference of posterior of $\mu(t)$ and $\phi(\tau)$ is non-trivial because of 1) the doubly-intractable problem introduced by Adams et al. (2009) caused by intractable integrals in the numerator and denominator; 2) the posterior has no closed-form solution. However, as we can see later, these two problems can be circumvented by augmenting the likelihood with auxiliary latent random variables. The sigmoid link function is chosen since it can be transformed to infinite mixture of Gaussians; consequently, the augmented likelihood is in a conjugate form allowing for more efficient Gibbs sampling, EM and variational inference with explicit expressions.

2.1 Augmenting Branching Structure

In the above likelihood Eq.(2), the coupling of $\mu(t)$ and $\phi(\tau)$ in the products term leads to inference difficulty. A well-known decoupling method is to incorporate the *branching structure* of Hawkes process (Marsan and Lengline, 2008; Zhou et al., 2013). The branching structure $\mathbf{X}$ is a triangular matrix with Bernoulli variables x_{ij} indicating whether the i-th event is triggered by itself or a previous event j.

$$x_{ii} = \begin{cases} 1 & \text{if event } i \text{ is a background event} \\ 0 & \text{otherwise} \end{cases}$$

$$x_{ij} = \begin{cases} 1 & \text{if event } i \text{ is caused by event } j \\ 0 & \text{otherwise} \end{cases}$$

After augmenting the branching structure $\mathbf{X}$, the joint likelihood has the following representation

$$\begin{aligned} &p(D, \mathbf{X}|\mu(t), \phi(\tau)) \\ &= \underbrace{\prod_{i=1}^{N} \mu(t_i)^{x_{ii}} \exp\left(-\int_T \mu(t)dt\right)}_{\mu(t) \text{ part}} \cdot \underbrace{\prod_{i=2}^{N}\prod_{j=1}^{i-1} \phi(t_i - t_j)^{x_{ij}} \prod_{i=1}^{N} \exp\left(-\int_{T_\phi} \phi(\tau)d\tau\right)}_{\phi(\tau) \text{ part}}, \end{aligned} \tag{3}$$

where we assume the support of triggering kernel is bounded with $[0, T_\phi]$ for the convenience of numerical integral, $\mu(t) = \lambda_\mu^* \sigma(f(t))$, $\phi(\tau) = \lambda_\phi^* \sigma(g(\tau))$. If the branching structure $\mathbf{X}$ is marginalized out in Eq.(3), we get the original likelihood in Eq.(2). After introducing the branching structure, the joint likelihood is decoupled to two independent factors.

2.2 Transformation of Sigmoid Function

We utilize a remarkable representation discovered in the literature of Bayesian inference for logistic regression (Polson et al., 2013) in recent years. Surprisingly, the sigmoid function is redefined as a Gaussian representation

$$\sigma(z) = \int_0^\infty e^{h(\omega,z)} p_{\mathrm{PG}}(\omega|1,0)d\omega, \tag{4}$$

where $h(\omega, z) = z/2 - z^2\omega/2 - \log 2$, $p_{\mathrm{PG}}(\omega|1,0)$ is the Pólya-Gamma distribution with $\omega \in \mathbb{R}^+$. The derivation is shown in Appendix A.

Using Eq.(4), the products of observations $\sigma(f(t_i))$ and $\sigma(g(\tau_{ij}))$ ($\tau_{ij} = t_i - t_j$) in the likelihood Eq.(3) are transformed into a Gaussian form. It is worth noting that we do not need to know the exact form of Pólya-Gamma distribution but only its first order moment.

2.3 Transformation of Exponential Integral

Utilizing Eq.(4) and the sigmoid property $\sigma(z) = 1 - \sigma(-z)$, the exponential integral in the likelihood Eq.(3) can be rewritten as

$$\exp\left(-\int_T \lambda_\mu^* \sigma(f(t))dt\right) = \exp\left(-\int_T \int_{\mathbb{R}^+} \left(1 - e^{h(\omega_\mu, -f(t))}\right) \lambda_\mu^* p_{\mathrm{PG}}(\omega_\mu|1,0)d\omega_\mu dt\right). \tag{5}$$

Donner and Opper (2018) has proved, according to Campbell's theorem (Kingman, 2005) (shown in Appendix B), the right hand side of Eq.(5) is a characteristic functional of a marked Poisson process, so we can rewrite it as

$$\exp\left(-\int_T \lambda_\mu^* \sigma(f(t))dt\right) = \mathbb{E}_{p_{\lambda_\mu}}\left[\prod_{(\omega_\mu,t)\in\Pi_\mu} e^{h(\omega_\mu,-f(t))}\right], \tag{6}$$

where $\Pi_\mu = \{(\omega_{\mu_m}, t_m)\}_{m=1}^{M_\mu}$ denotes a random realization of a marked Poisson process and p_{λ_μ} is the probability measure of the marked Poisson process Π_μ with intensity $\lambda_\mu(t,\omega_\mu) = \lambda_\mu^* p_{\mathrm{PG}}(\omega_\mu|1,0)$. The events $\{t_m\}_{m=1}^{M_\mu}$ follow a Poisson process with rate λ_μ^* and the latent Pólya-Gamma variable ω_{μ_m} denotes the independent *mark* at each location t_m. The detailed derivation can be found in Appendix B. Here, we only discuss the baseline intensity part. All derivation in the triggering kernel part is same as the baseline intensity part except some notations.

2.4 Augmented Likelihood

Substituting Eq.(4) and Eq.(6) into Eq.(3), we obtain the final *factorised and augmented* likelihood

$$\begin{aligned} &p(D, \Pi_\mu, \{\Pi_{\phi_i}\}_{i=1}^N, \boldsymbol{\omega}, \mathbf{X}|\lambda_\mu^*, \lambda_\phi^*, f, g) \\ &= \underbrace{p(D, \Pi_\mu, \boldsymbol{\omega}_{ii}, \mathbf{X}_{ii}|\lambda_\mu^*, f)}_{\mu(t)\text{ part}} \cdot \underbrace{p(D, \{\Pi_{\phi_i}\}_{i=1}^N, \boldsymbol{\omega}_{ij}, \mathbf{X}_{ij}|\lambda_\phi^*, g)}_{\phi(\tau)\text{ part}}, \end{aligned}$$

where

1. the augmented joint likelihood of $\mu(t)$ part (derivation in Appendix C) is

$$p(D, \Pi_\mu, \boldsymbol{\omega}_{ii}, \mathbf{X}_{ii}|\lambda_\mu^*, f) = \prod_{i=1}^N \left(\lambda_\mu(t_i,\omega_{ii}) e^{h(\omega_{ii}, f(t_i))}\right)^{x_{ii}} \cdot p_{\lambda_\mu}(\Pi_\mu|\lambda_\mu^*) \prod_{(\omega_\mu,t)\in\Pi_\mu} e^{h(\omega_\mu,-f(t))} \tag{7}$$

with $\boldsymbol{\omega}_{ii}$ denoting a vector of ω_{ii} on each t_i and $\mathbf{X}_{ii}$ being the diagonal of branching structure $\mathbf{X}$;

2. and the augmented joint likelihood of $\phi(\tau)$ part (derivation in Appendix C) is

$$\begin{aligned} &p(D, \{\Pi_{\phi_i}\}_{i=1}^N, \boldsymbol{\omega}_{ij}, \mathbf{X}_{ij}|\lambda_\phi^*, g) \\ &= \prod_{i=2}^N \prod_{j=1}^{i-1} \left(\lambda_\phi(\tau_{ij},\omega_{ij}) e^{h(\omega_{ij}, g(\tau_{ij}))}\right)^{x_{ij}} \cdot \prod_{i=1}^N \left[p_{\lambda_\phi}(\Pi_{\phi_i}|\lambda_\phi^*) \prod_{(\omega_\phi,\tau)\in\Pi_{\phi_i}} e^{h(\omega_\phi,-g(\tau))}\right], \end{aligned} \tag{8}$$

where p_{λ_ϕ} is the probability measure of the corresponding latent marked Poisson process $\Pi_{\phi_i} = \{(\omega_{\phi_m}, \tau_m)\}_{m=1}^{M_{\phi_i}}$ with intensity $\lambda_\phi(\tau,\omega_\phi) = \lambda_\phi^* p_{\mathrm{PG}}(\omega_\phi|1,0)$, $\boldsymbol{\omega}_{ij}$ denotes the vector of ω_{ij} on each τ_{ij} and $\mathbf{X}_{ij}$ is the entries off the diagonal of branching structure. It is worth noting that there exists N independent latent marked Poisson processes because of the exponential integral product term in Eq.(3).

The motivation of augmenting auxiliary latent random variables should now be clear. The augmented representation of likelihood contains the GP variables $f(\cdot)$ and $g(\cdot)$ only linearly and quadratically in the exponents and is thus conjugate to the GP prior. Our SGHP

model can be considered as an extension of the sigmoid Gaussian Cox process (Adams et al., 2009) in two aspects: **(1)** If the latent Pólya-Gamma random variables in the augmented joint distribution are integrated out ($\mu(t)$ part or $\phi(\tau)$ part), we obtain the likelihood used by Adams et al. (2009) (Eq.4). We utilize the Campbell's theorem to introduce Pólya-Gamma random variables, which results in a likelihood being conjugate to the GP priors. **(2)** The branching structure of Hawkes process is incorporated. This leads to the decoupling of $\mu(t)$ component and $\phi(\tau)$ component, consequently, the solution in Cox process scenario is extended to Hawkes process scenario.

3. Gibbs Sampler

A naïve Gibbs sampler is derived in this section. However, the naïve implementation is time-consuming because of the cubic complexity with respect to (w.r.t.) the number of observations and latent Poisson events when sampling f and g. This issue has been introduced in Adams et al. (2009). To circumvent the problem, we utilize the sparse GP approximation to introduce some inducing points to make the inference efficient.

3.1 Naïve Gibbs Sampler

Incorporating the priors of λ_μ^* and f into Eq.(7), we obtain the joint distribution over all variables of baseline intensity part. Without loss of generality, an improper prior $p(\lambda_\mu^*) = 1/\lambda_\mu^*$ (Bishop, 2006) and a symmetric GP prior $\mathcal{GP}(f|0, K_f)$ are utilized here

$$\begin{aligned} &p(D, \Pi_\mu, \boldsymbol{\omega}_{ii}, \mathbf{X}_{ii}, \lambda_\mu^*, f) \\ &= \prod_{i=1}^{N} \left(\lambda_\mu(t_i, \omega_{ii}) e^{h(\omega_{ii}, f(t_i))}\right)^{x_{ii}} \cdot p_{\lambda_\mu}(\Pi_\mu|\lambda_\mu^*) \prod_{(\omega_\mu, t)\in\Pi_\mu} e^{h(\omega_\mu, -f(t))} \cdot \lambda_\mu^{*-1} \mathcal{GP}(f). \end{aligned} \tag{9}$$

All derivation in the triggering kernel part is same as the baseline intensity part except some notations. The joint distribution over all variables of triggering kernel part is

$$\begin{aligned} &p(D, \{\Pi_{\phi_i}\}_{i=1}^N, \boldsymbol{\omega}_{ij}, \mathbf{X}_{ij}, \lambda_\phi^*, g) \\ &= \prod_{i=2}^{N}\prod_{j=1}^{i-1} \left(\lambda_\phi(\tau_{ij}, \omega_{ij}) e^{h(\omega_{ij}, g(\tau_{ij}))}\right)^{x_{ij}} \cdot \prod_{i=1}^{N} \left[p_{\lambda_\phi}(\Pi_{\phi_i}|\lambda_\phi^*) \prod_{(\omega_\phi,\tau)\in\Pi_{\phi_i}} e^{h(\omega_\phi, -g(\tau))} \right] \cdot \lambda_\phi^{*-1} \mathcal{GP}(g) \end{aligned} \tag{10}$$

with $\mathcal{GP}(g)$ being symmetric $\mathcal{GP}(g|0, K_g)$.

3.1.1 Sampling the Pólya-Gamma Variables

The conditional posteriors of Pólya-Gamma variables $\boldsymbol{\omega}_{ii}$ and $\boldsymbol{\omega}_{ij}$ only depend on the function values f and g at the observations t_i and τ_{ij}

$$\begin{aligned} p(\boldsymbol{\omega}_{ii}|\mathbf{f}) &= \prod_{i=1}^{N} \left(p_{\mathrm{PG}}(\omega_{ii}|1, f(t_i))\right)^{x_{ii}} \\ p(\boldsymbol{\omega}_{ij}|\mathbf{g}) &= \prod_{i=2}^{N}\prod_{j=1}^{i-1} \left(p_{\mathrm{PG}}(\omega_{ij}|1, g(\tau_{ij}))\right)^{x_{ij}}, \end{aligned} \tag{11}$$

where we utilize the tilted Pólya-Gamma distribution $p_{\mathrm{PG}}(\omega|b,c) \propto e^{-c^2\omega/2} p_{\mathrm{PG}}(\omega|b,0)$. The Pólya-Gamma random variable can be efficiently sampled by the method proposed in Polson et al. (2013).

3.1.2 Sampling the Upper Bounds

The conditional posteriors of upper bounds λ_μ^* and λ_ϕ^* depend on the branching structure and latent marked Poisson processes.

$$\begin{aligned} p(\lambda_\mu^*|\mathbf{X}_{ii}, \Pi_\mu) &= \mathrm{Gamma}(\lambda_\mu^*|N_\mu + M_\mu, T) \\ p(\lambda_\phi^*|\mathbf{X}_{ij}, \Pi_\phi) &= \mathrm{Gamma}(\lambda_\phi^*|N_\phi + M_\phi, NT_\phi), \end{aligned} \tag{12}$$

where $N_\mu = \sum_{i=1}^N x_{ii}$, $M_\mu = |\Pi_\mu|$, $N_\phi = \sum_{i=2}^N \sum_{j=1}^{i-1} x_{ij}$ and $M_\phi = \sum_{i=1}^N M_{\phi_i} = \sum_{i=1}^N |\Pi_{\phi_i}|$ with $|\cdot|$ denoting the number of points on a Poisson process.

3.1.3 Sampling the Function Values

Due to the augmentation of Pólya-Gamma random variables, the likelihoods of GP variables $\mathbf{f}_{N_\mu+M_\mu}$ and $\mathbf{g}_{N_\phi+M_\phi}$ are conjugate to the GP priors. Therefore, the conditional posteriors are still Gaussian

$$\begin{aligned} p(\mathbf{f}_{N_\mu+M_\mu}|\boldsymbol{\omega}_{ii}, \Pi_\mu) &= \mathcal{N}(\mathbf{f}_{N_\mu+M_\mu}|\mathbf{m}_{N_\mu+M_\mu}, \boldsymbol{\Sigma}_{N_\mu+M_\mu}) \\ p(\mathbf{g}_{N_\phi+M_\phi}|\boldsymbol{\omega}_{ij}, \{\Pi_{\phi_i}\}_{i=1}^N) &= \mathcal{N}(\mathbf{g}_{N_\phi+M_\phi}|\mathbf{m}_{N_\phi+M_\phi}, \boldsymbol{\Sigma}_{N_\phi+M_\phi}), \end{aligned} \tag{13}$$

with covariance matrix $\boldsymbol{\Sigma}_{N_\mu+M_\mu} = [\mathbf{D}_\mu + \mathbf{K}_{N_\mu+M_\mu}^{-1}]^{-1}$. $\mathbf{D}_\mu$ is a diagonal matrix with its first N_μ entries being $\boldsymbol{\omega}_{ii}$ and the following M_μ entries being $\{\omega_{\mu_m}\}_{m=1}^{M_\mu}$. $\mathbf{K}_{N_\mu+M_\mu}$ is the covariance matrix of GP prior evaluated at the observed points $\{t_i\}_{i=1}^{N_\mu}$ and the latent points $\{t_m\}_{m=1}^{M_\mu}$. The mean $\mathbf{m}_{N_\mu+M_\mu} = \boldsymbol{\Sigma}_{N_\mu+M_\mu} \cdot \mathbf{v}_{N_\mu+M_\mu}$ with the first N_μ entries of $\mathbf{v}_{N_\mu+M_\mu}$ being 0.5 and the following M_μ entries being -0.5. The solution for the mean and covariance matrix of $\mathbf{g}_{N_\phi+M_\phi}$ is the same with the corresponding subscripts being replaced.

3.1.4 Sampling the Latent Marked Poisson Processes

The conditional posterior of the latent marked point process is

$$p(\Pi_\mu|f, \lambda_\mu^*) = \frac{p_{\lambda_\mu}(\Pi_\mu|\lambda_\mu^*) \prod_{(\omega_\mu,t)\in\Pi_\mu} e^{h(\omega_\mu,-f(t))}}{\int p_{\lambda_\mu}(\Pi_\mu|\lambda_\mu^*) \prod_{(\omega_\mu,t)\in\Pi_\mu} e^{h(\omega_\mu,-f(t))} d\Pi_\mu}. \tag{14}$$

As proved by Donner and Opper (2018), this conditional point process is again a marked Poisson process by utilizing the Campbell theorem to calculate its characteristic function. But we provide a more concise proof here: using Eq.(5) and (6) to convert the denominator, Eq.(14) can be rewritten as

$$\begin{aligned} p(\Pi_\mu|f, \lambda_\mu^*) &= \frac{p_{\lambda_\mu}(\Pi_\mu|\lambda_\mu^*) \prod_{(\omega_\mu,t)\in\Pi_\mu} e^{h(\omega_\mu,-f(t))}}{\exp\left(-\iint (1 - e^{h(\omega_\mu,-f(t))}) \lambda_\mu^* p_{\mathrm{PG}}(\omega_\mu|1,0) d\omega_\mu dt\right)} \\ &= \prod_{(\omega_\mu,t)\in\Pi_\mu} \left(e^{h(\omega_\mu,-f(t))} \lambda_\mu^* p_{\mathrm{PG}}(\omega_\mu|1,0)\right) \cdot \exp\left(-\iint e^{h(\omega_\mu,-f(t))} \lambda_\mu^* p_{\mathrm{PG}}(\omega_\mu|1,0) d\omega_\mu dt\right) \end{aligned} \tag{15}$$

It is straightforward to see the above conditional posterior is just in the likelihood form of a marked Poisson process with intensity function

$$\Lambda_\mu(t, \omega_\mu) = e^{h(\omega_\mu, -f(t))}\lambda_\mu^* p_{\mathrm{PG}}(\omega_\mu|1, 0) = \lambda_\mu^* \sigma(-f(t)) p_{\mathrm{PG}}(\omega_\mu|1, f(t)). \tag{16}$$

The derivation of conditional posterior of Π_ϕ is same as Π_μ. It is worth noting that there exists N independent marked Poisson processes with the same intensity function $\Lambda_\phi(\tau, \omega_\phi) = \lambda_\phi^* \sigma(-g(\tau)) p_{\mathrm{PG}}(\omega_\phi|1, g(\tau))$.

For sampling from the posterior marked Poisson processes, we first draw the timestamps t_m (τ_m) with the rate $\lambda_\mu^* \sigma(-f(t))$ ($\lambda_\phi^* \sigma(-g(\tau))$) by using the thinning algorithm (Ogata, 1998), and then draw the marks ω_μ (ω_ϕ) from the conditional distribution $p_{\mathrm{PG}}(\omega_\mu|1, f(t))$ ($p_{\mathrm{PG}}(\omega_\phi|1, g(\tau))$).

3.1.5 Sampling the Branching Structure

After combining Eq.(9) and Eq.(10) and integrating out $\boldsymbol{\omega}_{ii}$ and $\boldsymbol{\omega}_{ij}$, we obtain the conditional posterior of $\mathbf{X}$

$$p(\mathbf{X}|\lambda_\mu^*, \lambda_\phi^*, f, g) \propto \prod_{i=1}^{N} (\mu(t_i))^{x_{ii}} \prod_{i=2}^{N} \prod_{j=1}^{i-1} (\phi(\tau_{ij}))^{x_{ij}},$$

with $\mu(t_i) = \lambda_\mu^* \sigma(f(t_i))$ and $\phi(\tau_{ij}) = \lambda_\phi^* \sigma(g(\tau_{ij}))$. This is a categorical distribution with

$$\begin{aligned} p(x_{ii} = 1) &= \frac{\mu(t_i)}{\mu(t_i) + \sum_{j=1}^{i-1} \phi(\tau_{ij})} \\ p(x_{ij} = 1) &= \frac{\phi(\tau_{ij})}{\mu(t_i) + \sum_{j=1}^{i-1} \phi(\tau_{ij})} \end{aligned} \tag{17}$$

which is a well-known result in Lewis and Mohler (2011) and Zhou et al. (2013).

3.2 Algorithm Speeding Up

The naïve Gibbs sampler above is impractical. The reasons are: **(1)** the bottleneck of the algorithm is the step of sampling function values. Because we have to perform matrix inversion, the complexity is $\mathcal{O}((N_\mu + M_\mu)^3 + (N_\phi + M_\phi)^3)$ where $N_\mu + N_\phi = N$. This means it is non-scalable to even a few hundreds of observations. **(2)** The function values have to be sampled twice in one MCMC loop. Each time when the branching structure or the latent marked Poisson process is updated, the function values have to be updated once in order to avoid dimension mismatch. This slows down the Gibbs sampler even further.

To circumvent these problems, we utilize the sparse GP approximation to introduce some inducing points. f and g are supposed to be dependent on their corresponding inducing points $\{t_s\}_{s=1}^{S_\mu}$ and $\{\tau_s\}_{s=1}^{S_\phi}$ where S_μ and S_ϕ are the number of inducing points for $\mu(t)$ and $\phi(\tau)$; the function values of f and g at these inducing points are $\mathbf{f}_{t_s}$ and $\mathbf{g}_{\tau_s}$. Given a sample $\mathbf{f}_{t_s}$ and $\mathbf{g}_{\tau_s}$, $\mathbf{f}_{N_\mu + M_\mu}$ and $\mathbf{g}_{N_\phi + M_\phi}$ in Eq.(13) are assumed to be the posterior GP mean functions

$$\mathbf{f}_{N_\mu + M_\mu} = \mathbf{K}_{tt_s} \mathbf{K}_{t_s t_s}^{-1} \mathbf{f}_{t_s}, \quad \mathbf{g}_{N_\phi + M_\phi} = \mathbf{K}_{\tau\tau_s} \mathbf{K}_{\tau_s \tau_s}^{-1} \mathbf{g}_{\tau_s}, \tag{18}$$

with $\mathbf{K}_{tt_s}$ and $\mathbf{K}_{\tau\tau_s}$ being the kernel matrixes w.r.t. the observations and inducing points while $\mathbf{K}_{t_st_s}$ and $\mathbf{K}_{\tau_s\tau_s}$ being w.r.t. inducing points only. Now the conditional posteriors of function values are transformed from observations to inducing points

$$\begin{aligned} p(\mathbf{f}_{t_s}|\boldsymbol{\omega}_{ii}, \Pi_\mu) &= \mathcal{N}(\mathbf{f}_{t_s}|\mathbf{m}_{t_s}, \boldsymbol{\Sigma}_{t_s}) \\ p(\mathbf{g}_{\tau_s}|\boldsymbol{\omega}_{ij}, \{\Pi_{\phi_i}\}_{i=1}^N) &= \mathcal{N}(\mathbf{g}_{\tau_s}|\mathbf{m}_{\tau_s}, \boldsymbol{\Sigma}_{\tau_s}) \end{aligned} \tag{19}$$

with $\boldsymbol{\Sigma}_{t_s} = \left[\mathbf{K}_{t_st_s}^{-1}\mathbf{K}_{tt_s}^T\mathbf{D}_\mu\mathbf{K}_{tt_s}\mathbf{K}_{t_st_s}^{-1} + \mathbf{K}_{t_st_s}^{-1}\right]^{-1}$ and $\mathbf{m}_{t_s} = \boldsymbol{\Sigma}_{t_s}\mathbf{K}_{t_st_s}^{-1}\mathbf{K}_{tt_s}^T\mathbf{v}_{N_\mu+M_\mu}$. The solution for $\boldsymbol{\Sigma}_{\tau_s}$ and $\mathbf{m}_{\tau_s}$ is the same with the corresponding subscripts being replaced.

With sparse GP approximation, the complexity of matrix inversion is reduced to $\mathcal{O}(S_\mu^3 + S_\phi^3)$ with $S_\mu \ll N_\mu + M_\mu$ $S_\phi \ll N_\phi + M_\phi$. This results in a complexity scaling *linearly* with data size: $\mathcal{O}(N)$ due to the sparsity of branching structure: each event i is triggered only by a previous event j or itself. What makes this even more remarkable is the fact that the function values only need to be sampled once in one MCMC loop because they only depend on inducing points which are fixed during the sampling process. Moreover, the sampling of latent marked Poisson processes can be parallelized.

3.3 Hyperparameters

Throughout this work, the GP covariance kernel we use is the squared exponential kernel $k(x, x') = \theta_0 \exp\left(-\frac{\theta_1}{2}\|x - x'\|^2\right)$. The hyperparameters θ_0 and θ_1 can be sampled by a Metropolis-Hasting method (Hastings, 1970). Empirically, we find that if we update θ_0 and θ_1 too frequently, the convergence will be slow because the deviation of $\mu(t)$ and $\phi(\tau)$ at the beginning of Gibbs loops will provide poor estimations of θ_0 and θ_1. Therefore, we update them every 20 loops.

Additional hyperparameters are the number and location of inducing points which affect the complexity and estimation quality of $\mu(t)$ and $\phi(\tau)$. A large number of inducing points will lead to high complexity while a small number cannot capture the dynamics. For fast inference, the inducing points are uniformly located on the domain. Another advantage of uniform location is that the kernel matrix has Toeplitz structure (Cunningham et al., 2008) which means the matrix inversion can be implemented more efficiently. The number of inducing points is gradually increased until no more significant improvement. The final pseudo code is provided in Alg.1.

4. EM Algorithm

An EM algorithm is derived to obtain the MAP estimate in this section. With the original likelihood Eq.(2) and GP priors $\mathcal{GP}(f)$ and $\mathcal{GP}(g)$ (symmetric prior $\mathcal{GP}(\cdot|0, K_\cdot)$), the log-posterior corresponds to a penalized log-likelihood. As discussed by Donner (2019) and Rasmussen (2003) for GP models with likelihood depending on finite inputs, the regularizer is given by the squared reproducing kernel Hilbert space (RKHS) norm corresponding to the GP kernel. Therefore, we obtain

$$\hat{\lambda}_\mu^*, \hat{f}, \hat{\lambda}_\phi^*, \hat{g} = \operatorname{argmax}\left\{ \log p(D|\lambda_\mu^*, f, \lambda_\phi^*, g) - \frac{1}{2}\|f\|_{\mathcal{H}_{k_f}}^2 - \frac{1}{2}\|g\|_{\mathcal{H}_{k_g}}^2 \right\}, \tag{20}$$

Algorithm 1: Accelerated Gibbs sampler for SGHP

Result: $\mu(t)$, $\phi(\tau)$
Initialize hyperparameters and $\mathbf{X}$, λ_μ^*, λ_ϕ^*, $\boldsymbol{\omega}_{ii}$, $\boldsymbol{\omega}_{ij}$, $\mathbf{f}_{t_s}$, $\mathbf{g}_{\tau_s}$, Π_μ, $\{\Pi_{\phi_i}\}_{i=1}^N$;
for do
  Sample $\boldsymbol{\omega}_{ii}$ and $\boldsymbol{\omega}_{ij}$ with Eq.(11);
  Sample λ_μ^* and λ_ϕ^* with Eq.(12);
  Sample $\mathbf{f}_{t_s}$ and $\mathbf{g}_{\tau_s}$ with Eq.(19);
  Sample Π_μ and $\{\Pi_{\phi_i}\}_{i=1}^N$ with Eq.(16);
  Sample $\mathbf{X}$ with Eq.(17);
  Sample hyperparameters with Metropolis-Hasting algorithm.
end

where $\hat{\lambda}_\mu^*, \hat{f}, \hat{\lambda}_\phi^*, \hat{g}$ are the MAP estimates, $\|\cdot\|_{\mathcal{H}_k}^2$ is the squared RKHS norm with kernel k. The regularizer is the functional counterpart of log Gaussian prior. Instead of performing direct optimization, we propose an EM algorithm with the augmented auxiliary variables. Specifically, we propose a lower-bound of the log-posterior

$$\begin{aligned}&\mathcal{Q}((\lambda_\mu^*, f, \lambda_\phi^*, g)|(\lambda_\mu^*, f, \lambda_\phi^*, g)_{\text{old}}) = \\ &\mathbb{E}\left[\log p(D, \Pi_\mu, \{\Pi_{\phi_i}\}_{i=1}^N, \boldsymbol{\omega}, \mathbf{X}|\lambda_\mu^*, f, \lambda_\phi^*, g)\right] - \frac{1}{2}\|f\|_{\mathcal{H}_{k_f}}^2 - \frac{1}{2}\|g\|_{\mathcal{H}_{k_g}}^2,\end{aligned} \tag{21}$$

with $\mathbb{E}$ over $p(\Pi_\mu, \{\Pi_{\phi_i}\}_{i=1}^N, \boldsymbol{\omega}, \mathbf{X}|(\lambda_\mu^*, f, \lambda_\phi^*, g)_{\text{old}})$, the subscript "old" means the value from the last iteration.

As we can see later, because of auxiliary variables augmentation, the GP variables are in a quadratic form in the lower-bound, which results in an analytical solution in the M step.

4.1 E Step

In the E step, we first derive the *conditional density* $p(\Pi_\mu, \{\Pi_{\phi_i}\}_{i=1}^N, \boldsymbol{\omega}, \mathbf{X}|(\lambda_\mu^*, f, \lambda_\phi^*, g)_{\text{old}})$ and then compute the *lower-bound* $\mathcal{Q}$.

4.1.1 Conditional Density

The conditional density $p(\Pi_\mu, \{\Pi_{\phi_i}\}_{i=1}^N, \boldsymbol{\omega}, \mathbf{X}|(\lambda_\mu^*, f, \lambda_\phi^*, g)_{\text{old}})$ can be factorized and obtained from Eq.(7) and (8). More specifically, we provide details of these factors.

1. The conditional distributions of Pólya-Gamma variables $\boldsymbol{\omega}_{ii}$ and $\boldsymbol{\omega}_{ij}$ depend on the function values f_{old} and g_{old} at t_i and τ_{ij}

$$\begin{aligned}p(\boldsymbol{\omega}_{ii}|\mathbf{f}_{\text{old}}) &= \prod_{i=1}^N p_{\text{PG}}(\omega_{ii}|1, f_{\text{old}}(t_i)) \\ p(\boldsymbol{\omega}_{ij}|\mathbf{g}_{\text{old}}) &= \prod_{i=2}^N \prod_{j=1}^{i-1} p_{\text{PG}}(\omega_{ij}|1, g_{\text{old}}(\tau_{ij})),\end{aligned} \tag{22}$$

where we marginalize out $\mathbf{X}$ and utilize the tilted Pólya-Gamma distribution $p_{\text{PG}}(\omega|b, c) \propto e^{-c^2\omega/2} p_{\text{PG}}(\omega|b, 0)$ with the first order moment being $\mathbb{E}[\omega] = \frac{b}{2c}\tanh\frac{c}{2}$ (Polson et al., 2013).

2. The conditional density of Π_μ depends on f_{old} and $\lambda^*_{\mu_{\text{old}}}$

$$p(\Pi_\mu|f_{\text{old}},\lambda^*_{\mu_{\text{old}}}) = \frac{p_{\lambda_\mu}(\Pi_\mu|\lambda^*_{\mu_{\text{old}}})\prod_{(\omega_\mu,t)\in\Pi_\mu} e^{h(\omega_\mu,-f_{\text{old}}(t))}}{\int p_{\lambda_\mu}(\Pi_\mu|\lambda^*_{\mu_{\text{old}}})\prod_{(\omega_\mu,t)\in\Pi_\mu} e^{h(\omega_\mu,-f_{\text{old}}(t))}d\Pi_\mu}. \tag{23}$$

Similarly, using Eq.(5) and (6) to convert the denominator, Eq.(23) can be rewritten as

$$\begin{aligned}
p(\Pi_\mu|f_{\text{old}},\lambda^*_{\mu_{\text{old}}}) &= \frac{p_{\lambda_\mu}(\Pi_\mu|\lambda^*_{\mu_{\text{old}}})\prod_{(\omega_\mu,t)\in\Pi_\mu} e^{h(\omega_\mu,-f_{\text{old}}(t))}}{\exp\left(-\iint(1-e^{h(\omega_\mu,-f_{\text{old}}(t))})\lambda^*_{\mu_{\text{old}}}p_{\text{PG}}(\omega_\mu|1,0)d\omega_\mu dt\right)} \\
&= \prod_{\Pi_\mu}\left(e^{h(\omega_\mu,-f_{\text{old}}(t))}\lambda^*_{\mu_{\text{old}}}p_{\text{PG}}(\omega_\mu|1,0)\right)\cdot\exp\left(-\iint e^{h(\omega_\mu,-f_{\text{old}}(t))}\lambda^*_{\mu_{\text{old}}}p_{\text{PG}}(\omega_\mu|1,0)d\omega_\mu dt\right).
\end{aligned}$$

Again, it is straightforward to see the above conditional distribution is in the likelihood form of a marked Poisson process with intensity function

$$\Lambda_\mu(t,\omega_\mu) = e^{h(\omega_\mu,-f_{\text{old}}(t))}\lambda^*_{\mu_{\text{old}}}p_{\text{PG}}(\omega_\mu|1,0) = \lambda^*_{\mu_{\text{old}}}\sigma(-f_{\text{old}}(t))p_{\text{PG}}(\omega_\mu|1,f_{\text{old}}(t)). \tag{24}$$

The derivation of conditional distribution of Π_{ϕ_i} is same as Π_μ with the corresponding subscripts being replaced. It is worth noting that there exists N independent marked Poisson processes $\{\Pi_{\phi_i}\}_{i=1}^N$ with the same intensity function

$$\Lambda_\phi(\tau,\omega_\phi) = \lambda^*_{\phi_{\text{old}}}\sigma(-g_{\text{old}}(\tau))p_{\text{PG}}(\omega_\phi|1,g_{\text{old}}(\tau)). \tag{25}$$

3. Combining Eq.(7) and (8) and marginalizing out $\boldsymbol{\omega}_{ii}$ and $\boldsymbol{\omega}_{ij}$, we obtain the conditional distribution of $\mathbf{X}$

$$p(\mathbf{X}|(\lambda^*_\mu,f,\lambda^*_\phi,g)_{\text{old}}) \propto \prod_{i=1}^N(\mu_{\text{old}}(t_i))^{x_{ii}}\prod_{i=2}^N\prod_{j=1}^{i-1}(\phi_{\text{old}}(\tau_{ij}))^{x_{ij}},$$

with $\mu_{\text{old}}(t_i) = \lambda^*_{\mu_{\text{old}}}\sigma(f_{\text{old}}(t_i))$ and $\phi_{\text{old}}(\tau_{ij}) = \lambda^*_{\phi_{\text{old}}}\sigma(g_{\text{old}}(\tau_{ij}))$. This is a categorical distribution with

$$\begin{aligned}
p(x_{ii}=1) &= \frac{\mu_{\text{old}}(t_i)}{\mu_{\text{old}}(t_i)+\sum_{j=1}^{i-1}\phi_{\text{old}}(\tau_{ij})} \\
p(x_{ij}=1) &= \frac{\phi_{\text{old}}(\tau_{ij})}{\mu_{\text{old}}(t_i)+\sum_{j=1}^{i-1}\phi_{\text{old}}(\tau_{ij})}.
\end{aligned} \tag{26}$$

4.1.2 Lower-bound of Log-posterior

Given those conditional densities above, we can compute the lower-bound $\mathcal{Q}$. The expectation of log-likelihood (ELL) term in Eq.(21) can be rewritten as the summation of baseline intensity part and triggering kernel part. The ELL of baseline intensity part is

$$\begin{aligned}
&\text{ELL}_\mu(\lambda^*_\mu,f) \\
&= \mathbb{E}_{p(\Pi_\mu,\boldsymbol{\omega}_{ii},\mathbf{X}_{ii}|(\lambda^*_\mu,f,\lambda^*_\phi,g)_{\text{old}})}\left[\log p(D,\Pi_\mu,\boldsymbol{\omega}_{ii},\mathbf{X}_{ii}|\lambda^*_\mu,f)\right] \\
&= -\frac{1}{2}\int_T A_\mu(t)f^2(t)dt + \int_T B_\mu(t)f(t)dt \\
&\quad - \lambda^*_\mu T + \left(\sum_{i=1}^N\mathbb{E}(x_{ii}) + \iint\Lambda_\mu(t,\omega_\mu)d\omega_\mu dt\right)\log\lambda^*_\mu
\end{aligned} \tag{27}$$

where

$$A_\mu(t) = \sum_{i=1}^{N} \mathbb{E}[\omega_{ii}]\mathbb{E}[x_{ii}]\delta(t-t_i) + \int_0^\infty \omega_\mu \Lambda_\mu(t,\omega_\mu) d\omega_\mu$$
$$B_\mu(t) = \frac{1}{2}\sum_{i=1}^{N} \mathbb{E}[x_{ii}]\delta(t-t_i) - \frac{1}{2}\int_0^\infty \Lambda_\mu(t,\omega_\mu) d\omega_\mu,$$

with $\mathbb{E}$ over $p(\omega_{ii}|f_{\text{old}}(t_i))$ or $p(x_{ii}|(\lambda_\mu^*, f, \lambda_\phi^*, g)_{\text{old}})$.

Similarly, the ELL of triggering kernel part is written as

$$\begin{aligned}
&\text{ELL}_\phi(\lambda_\phi^*, g) \\
&= \mathbb{E}_{p(\{\Pi_{\phi_i}\}_{i=1}^N, \boldsymbol{\omega}_{ij}, \mathbf{X}_{ij}|(\lambda_\mu^*, f, \lambda_\phi^*, g)_{\text{old}})} \left[\log p(D, \{\Pi_{\phi_i}\}_{i=1}^N, \boldsymbol{\omega}_{ij}, \mathbf{X}_{ij}|\lambda_\phi^*, g)\right] \\
&= -\frac{1}{2}\int_{T_\phi} A_\phi(\tau) g^2(\tau) d\tau + \int_{T_\phi} B_\phi(\tau) g(\tau) d\tau \\
&\quad - N\lambda_\phi^* T_\phi + \left(\sum_{i=2}^{N}\sum_{j=1}^{i-1} \mathbb{E}(x_{ij}) + N\iint \Lambda_\phi(\tau,\omega_\phi) d\omega_\phi d\tau\right) \log \lambda_\phi^*
\end{aligned} \tag{28}$$

where

$$A_\phi(\tau) = \sum_{i=2}^{N}\sum_{j=1}^{i-1} \mathbb{E}[\omega_{ij}]\mathbb{E}[x_{ij}]\delta(\tau-\tau_{ij}) + N\int_0^\infty \omega_\phi \Lambda_\phi(\tau,\omega_\phi) d\omega_\phi$$
$$B_\phi(\tau) = \frac{1}{2}\sum_{i=2}^{N}\sum_{j=1}^{i-1} \mathbb{E}[x_{ij}]\delta(\tau-\tau_{ij}) - \frac{N}{2}\int_0^\infty \Lambda_\phi(\tau,\omega_\phi) d\omega_\phi,$$

with $\mathbb{E}$ over $p(\omega_{ij}|g_{\text{old}}(\tau_{ij}))$ or $p(x_{ij}|(\lambda_\mu^*, f, \lambda_\phi^*, g)_{\text{old}})$.

However, the ELL is intractable for general GP priors by the fact that the ELL is a functional. To circumvent the problem, we utilize the sparse GP approximation to introduce some inducing points. Once again, f and g are supposed to be dependent on their corresponding inducing points $\{t_s\}_{s=1}^{S_\mu}$ and $\{\tau_s\}_{s=1}^{S_\phi}$; the function values of f and g at these inducing points are $\mathbf{f}_{t_s}$ and $\mathbf{g}_{\tau_s}$. Given a sample $\mathbf{f}_{t_s}$ and $\mathbf{g}_{\tau_s}$, $f(t)$ and $g(\tau)$ in Eq.(27) and (28) are assumed to be the posterior mean functions

$$f(t) = \mathbf{k}_{t_s t}^T \mathbf{K}_{t_s t_s}^{-1} \mathbf{f}_{t_s}, \quad g(\tau) = \mathbf{k}_{\tau_s \tau}^T \mathbf{K}_{\tau_s \tau_s}^{-1} \mathbf{g}_{\tau_s}, \tag{29}$$

with $\mathbf{k}_{t_s t}^T$ and $\mathbf{k}_{\tau_s \tau}^T$ being the kernel vector w.r.t. the observations and inducing points while $\mathbf{K}_{t_s t_s}$ and $\mathbf{K}_{\tau_s \tau_s}$ being w.r.t. inducing points only.

Substituting Eq.(29) to Eq.(27) and (28), we obtain

$$\begin{aligned}
&\mathcal{Q}((\lambda_\mu^*, \mathbf{f}_{t_s}, \lambda_\phi^*, \mathbf{g}_{\tau_s})|(\lambda_\mu^*, \mathbf{f}_{t_s}, \lambda_\phi^*, \mathbf{g}_{\tau_s})_{\text{old}}) \\
&= \text{ELL}_\mu(\lambda_\mu^*, \mathbf{f}_{t_s}) + \text{ELL}_\phi(\lambda_\phi^*, \mathbf{g}_{\tau_s}) - \frac{1}{2}\mathbf{f}_{t_s}^T \mathbf{K}_{t_s t_s}^{-1} \mathbf{f}_{t_s} - \frac{1}{2}\mathbf{g}_{\tau_s}^T \mathbf{K}_{\tau_s \tau_s}^{-1} \mathbf{g}_{\tau_s}.
\end{aligned} \tag{30}$$

4.2 M Step

In the M step, we maximize the lower-bound $\mathcal{Q}$. The optimal parameters $\hat{\lambda}_\mu^*, \hat{\mathbf{f}}_{t_s}, \hat{\lambda}_\phi^*, \hat{\mathbf{g}}_{\tau_s}$ can be obtained by setting the gradient of Eq.(30) to zero. Due to auxiliary variables

augmentation, we have analytical solutions

$$\begin{aligned}
\hat{\lambda}_\mu^* &= \left(\sum_{i=1}^{N} \mathbb{E}[x_{ii}] + M_\mu\right)/T \\
\hat{\lambda}_\phi^* &= \left(\sum_{i=2}^{N}\sum_{j=1}^{i-1} \mathbb{E}[x_{ij}] + NM_\phi\right)/(NT_\phi) \\
\hat{\mathbf{f}}_{t_s} &= \mathbf{\Sigma}_{t_s}\mathbf{K}_{t_st_s}^{-1}\int_T B_\mu(t)\mathbf{k}_{t_st}dt \\
\hat{\mathbf{g}}_{\tau_s} &= \mathbf{\Sigma}_{\tau_s}\mathbf{K}_{\tau_s\tau_s}^{-1}\int_{T_\phi} B_\phi(\tau)\mathbf{k}_{\tau_s\tau}d\tau
\end{aligned} \tag{31}$$

where $\mathbf{\Sigma}_{t_s} = \left[\mathbf{K}_{t_st_s}^{-1}\int A_\mu(t)\mathbf{k}_{t_st}\mathbf{k}_{t_st}^T dt\mathbf{K}_{t_st_s}^{-1} + \mathbf{K}_{t_st_s}^{-1}\right]^{-1}$, $M_\mu = \iint \Lambda_\mu(t,\omega_\mu)d\omega_\mu dt$, $\mathbf{\Sigma}_{\tau_s} = \left[\mathbf{K}_{\tau_s\tau_s}^{-1}\int A_\phi(\tau)\mathbf{k}_{\tau_s\tau}\mathbf{k}_{\tau_s\tau}^T d\tau\mathbf{K}_{\tau_s\tau_s}^{-1} + \mathbf{K}_{\tau_s\tau_s}^{-1}\right]^{-1}$, $M_\phi = \iint \Lambda_\phi(\tau,\omega_\phi)d\omega_\phi d\tau$. All intractable integrals are w.r.t. Lebesgue measure, thus can be solved by numerical methods such as Gaussian quadrature (Golub and Welsch, 1969).

4.3 Complexity

The analysis of complexity of EM algorithm is similar with that of Gibbs sampler. By introducing sparse GP approximation, the complexity of matrix inversion is fixed to $\mathcal{O}(S_\mu^3 + S_\phi^3)$ where S_μ (or S_ϕ) $\ll N$. For the fixed $\mu(t)$, $\phi(\tau)$ and T_ϕ, as T increases, the complexity scales *linearly* with data size: $\mathcal{O}(N)$ due to the sparsity of *expectation* of branching structure: previous points that are more than T_ϕ far away from event i have no influence on event i ($\mathbb{E}[x_{ij}] = 0$).

4.4 Hyperparameters

Once again, the GP covariance kernel is the squared exponential kernel. The hyperparameters θ_0 and θ_1 can be optimized by performing maximization of $\mathcal{Q}$ over $\{\theta_0, \theta_1\}$ using numerical packages. Normally, we update $\{\theta_0, \theta_1\}$ every 20 iterations. The number and location of inducing points are optimised utilizing the same method in Section 3.3. The final pseudo code is provided in Alg.2.

Algorithm 2: EM algorithm for SGHP

Result: $\mu(t) = \lambda_\mu^*\sigma(f(t))$, $\phi(\tau) = \lambda_\phi^*\sigma(g(\tau))$

Initialize hyperparameters and $\mathbf{X}$, λ_μ^*, λ_ϕ^*, $\boldsymbol{\omega}_{ii}$, $\boldsymbol{\omega}_{ij}$, $\mathbf{f}_{t_s}$, $\mathbf{g}_{\tau_s}$, Π_μ, $\{\Pi_{\phi_i}\}_{i=1}^N$;

for do

- Update the posterior of $\boldsymbol{\omega}_{ii}$ and $\boldsymbol{\omega}_{ij}$ by Eq.(22);
- Update intensities of Π_μ and $\{\Pi_\phi\}$ by Eq.(24), (25);
- Update the posterior of $\mathbf{X}$ by Eq.(26);
- Update $\lambda_\mu^*, \mathbf{f}_{t_s}, \lambda_\phi^*$ and $\mathbf{g}_{\tau_s}$ by Eq.(31);
- Update hyperparameters.

end

5. Mean-field Variational Inference

In this section, we extend the EM algorithm to a mean-field variational inference (Bishop, 2006) algorithm which solves the inference problem slightly slower than EM, but can provide uncertainty with a distribution estimation rather than point estimation.

In variational inference, the posterior distribution over latent variables is approximated by a variational distribution. The optimal variational distribution is chosen by minimising the Kullback-Leibler (KL) divergence or equivalently maximizing the evidence lower bound (ELBO). A common approach is the mean-field method where the variational distribution is assumed to factorize over some partition of latent variables. The mean-field variational inference algorithm can be seen as an extension of the EM algorithm from a MAP point estimation to a fully Bayesian estimation which approximates the posterior distribution of all variables.

For the problem at hand, after incorporating priors of λ_μ^*, f, λ_ϕ^* and g into Eq.(7) and (8), we obtain the joint distribution over all variables in Eq.(9) and (10). Because of the mean-field assumption, we assume the variational distribution q can factorize as

$$q(\Pi_\mu, \{\Pi_{\phi_i}\}_{i=1}^N, \boldsymbol{\omega}, \mathbf{X}, \lambda_\mu^*, f, \lambda_\phi^*, g) = q_1(\Pi_\mu, \{\Pi_{\phi_i}\}_{i=1}^N, \boldsymbol{\omega}, \mathbf{X}) q_2(\lambda_\mu^*, f, \lambda_\phi^*, g).$$

A standard derivation in the variational mean-field approach shows that the optimal distribution for each factor maximizing the ELBO is given by

$$\begin{aligned} \log q_1(\Pi_\mu, \{\Pi_{\phi_i}\}_{i=1}^N, \boldsymbol{\omega}, \mathbf{X}) &= \mathbb{E}_{q_2}[\log p(\Pi_\mu, \{\Pi_{\phi_i}\}_{i=1}^N, \boldsymbol{\omega}, \mathbf{X}, \lambda_\mu^*, f, \lambda_\phi^*, g)] + C_1 \\ \log q_2(\lambda_\mu^*, f, \lambda_\phi^*, g) &= \mathbb{E}_{q_1}[\log p(\Pi_\mu, \{\Pi_{\phi_i}\}_{i=1}^N, \boldsymbol{\omega}, \mathbf{X}, \lambda_\mu^*, f, \lambda_\phi^*, g)] + C_2 \end{aligned} \tag{32}$$

Substituting Eq.(9) and (10) into Eq.(32), we obtain the optimal distribution for each factor maximizing the ELBO. What is worth noting is that, to circumvent the functional problem caused by f and g, we again utilize the sparse GP approximation to introduce inducing points. Once again, given a sample $\mathbf{f}_{t_s}$ and $\mathbf{g}_{\tau_s}$, f and g in Eq.(32) are assumed to be the posterior mean functions in Eq.(29).

5.1 Optimal Density of Pólya-Gamma Variables

$$\begin{aligned} q_1(\boldsymbol{\omega}_{ii}) &= \prod_{i=1}^N p_{\mathrm{PG}}(\omega_{ii}|1, \tilde{f}(t_i)) \\ q_1(\boldsymbol{\omega}_{ij}) &= \prod_{i=2}^N \prod_{j=1}^{i-1} p_{\mathrm{PG}}(\omega_{ij}|1, \tilde{g}(\tau_{ij})), \end{aligned} \tag{33}$$

where we marginalize out $\mathbf{X}$ and $\tilde{f}(t_i) = \sqrt{\mathbb{E}(f^2(t_i))}$ and $\tilde{g}(\tau_{ij}) = \sqrt{\mathbb{E}(g^2(\tau_{ij}))}$ which can be computed utilizing $\mathbb{E}(C^2) = \mathbb{E}^2(C) + \mathrm{Var}(C)$.

5.2 Optimal Marked Poisson Processes

$$\begin{aligned} \Lambda_\mu^1(t, \omega_\mu) &= \tilde{\lambda}_\mu^* \sigma(-\tilde{f}(t)) p_{\mathrm{PG}}(\omega_\mu|1, \tilde{f}(t)) e^{(\tilde{f}(t) - \bar{f}(t))/2} \\ \Lambda_\phi^1(\tau, \omega_\phi) &= \tilde{\lambda}_\phi^* \sigma(-\tilde{g}(\tau)) p_{\mathrm{PG}}(\omega_\phi|1, \tilde{g}(\tau)) e^{(\tilde{g}(\tau) - \bar{g}(\tau))/2}, \end{aligned} \tag{34}$$

where $\tilde{\lambda}_\mu^* = e^{\mathbb{E}(\log \lambda_\mu^*)}$, $\bar{f}(t) = \mathbb{E}(f(t))$, $\tilde{\lambda}_\phi^* = e^{\mathbb{E}(\log \lambda_\phi^*)}$ and $\bar{g}(\tau) = \mathbb{E}(g(\tau))$.

5.3 Optimal Density of Intensity Upper-bounds

$$\begin{aligned} q_2(\lambda_\mu^*) &= \text{Gamma}(\lambda_\mu^*|\alpha_\mu, \beta_\mu) \\ q_2(\lambda_\phi^*) &= \text{Gamma}(\lambda_\phi^*|\alpha_\phi, \beta_\phi), \end{aligned} \tag{35}$$

with $\alpha_\mu = \sum_{i=1}^N \mathbb{E}(x_{ii}) + \iint \Lambda_\mu^1(t, \omega_\mu) dt d\omega_\mu$, $\alpha_\phi = \sum_{i=2}^N \sum_{j=1}^{i-1} \mathbb{E}(x_{ij}) + N \iint \Lambda_\phi^1(\tau, \omega_\phi) d\tau d\omega_\phi$, $\beta_\mu = T$, $\beta_\phi = NT_\phi$ and all intractable integrals can be solved by Gaussian quadrature. This provides the required expectation for Eq.(34) by $\mathbb{E}(\lambda^*) = \alpha/\beta$ and $\mathbb{E}(\log \lambda^*) = \psi(\alpha) - \log \beta$ where $\psi(\cdot)$ is the digamma function. Note also the similarity to EM algorithm in Eq.(31).

5.4 Optimal Sparse Gaussian Process

$$\begin{aligned} q_2(\mathbf{f}_{t_s}) &= \mathcal{N}(\mathbf{f}_{t_s}|\tilde{\mathbf{m}}_{t_s}, \tilde{\mathbf{\Sigma}}_{t_s}) \\ q_2(\mathbf{g}_{\tau_s}) &= \mathcal{N}(\mathbf{g}_{\tau_s}|\tilde{\mathbf{m}}_{\tau_s}, \tilde{\mathbf{\Sigma}}_{\tau_s}), \end{aligned} \tag{36}$$

where

$$\tilde{\mathbf{\Sigma}}_{t_s} = \left[\mathbf{K}_{t_s t_s}^{-1} \int \tilde{A}_\mu(t) \mathbf{k}_{t_s t} \mathbf{k}_{t_s t}^T dt \mathbf{K}_{t_s t_s}^{-1} + \mathbf{K}_{t_s t_s}^{-1} \right]^{-1}, \quad \tilde{\mathbf{m}}_{t_s} = \tilde{\mathbf{\Sigma}}_{t_s} \mathbf{K}_{t_s t_s}^{-1} \int \tilde{B}_\mu(t) \mathbf{k}_{t_s t} dt$$

with

$$\begin{aligned} \tilde{A}_\mu(t) &= \sum_{i=1}^N \mathbb{E}[\omega_{ii}] \mathbb{E}[x_{ii}] \delta(t - t_i) + \int_0^\infty \omega_\mu \Lambda_\mu^1(t, \omega_\mu) d\omega_\mu \\ \tilde{B}_\mu(t) &= \frac{1}{2} \sum_{i=1}^N \mathbb{E}[x_{ii}] \delta(t - t_i) - \frac{1}{2} \int_0^\infty \Lambda_\mu^1(t, \omega_\mu) d\omega_\mu \end{aligned}$$

and

$$\tilde{\mathbf{\Sigma}}_{\tau_s} = \left[\mathbf{K}_{\tau_s \tau_s}^{-1} \int \tilde{A}_\phi(\tau) \mathbf{k}_{\tau_s \tau} \mathbf{k}_{\tau_s \tau}^T d\tau \mathbf{K}_{\tau_s \tau_s}^{-1} + \mathbf{K}_{\tau_s \tau_s}^{-1} \right]^{-1}, \quad \tilde{\mathbf{m}}_{\tau_s} = \tilde{\mathbf{\Sigma}}_{\tau_s} \mathbf{K}_{\tau_s \tau_s}^{-1} \int \tilde{B}_\phi(\tau) \mathbf{k}_{\tau_s \tau} d\tau$$

with

$$\begin{aligned} \tilde{A}_\phi(\tau) &= \sum_{i=2}^N \sum_{j=1}^{i-1} \mathbb{E}[\omega_{ij}] \mathbb{E}[x_{ij}] \delta(\tau - \tau_{ij}) + N \int_0^\infty \omega_\phi \Lambda_\phi^1(\tau, \omega_\phi) d\omega_\phi \\ \tilde{B}_\phi(\tau) &= \frac{1}{2} \sum_{i=2}^N \sum_{j=1}^{i-1} \mathbb{E}[x_{ij}] \delta(\tau - \tau_{ij}) - \frac{N}{2} \int_0^\infty \Lambda_\phi^1(\tau, \omega_\phi) d\omega_\phi. \end{aligned}$$

All intractable integrals are w.r.t. Lebesgue measure, thus can be solved by Gaussian quadrature. Note also the similarity to EM algorithm in Eq.(31).

5.5 Optimal Density of Branching Structure

$$\begin{aligned} q_1(x_{ii} = 1) &= \frac{\tilde{\mu}(t_i)}{\tilde{\mu}(t_i) + \sum_{j=1}^{i-1} \tilde{\phi}(\tau_{ij})} \\ q_1(x_{ij} = 1) &= \frac{\tilde{\phi}(\tau_{ij})}{\tilde{\mu}(t_i) + \sum_{j=1}^{i-1} \tilde{\phi}(\tau_{ij})}, \end{aligned} \tag{37}$$

where we marginalize out $\boldsymbol{\omega}$ and $\tilde{\mu}(t_i) = \tilde{\lambda}_\mu^* e^{\mathbb{E}(\log \sigma(f(t_i)))}$, $\tilde{\phi}(\tau_{ij}) = \tilde{\lambda}_\phi^* e^{\mathbb{E}(\log \sigma(g(\tau_{ij})))}$. The $\mathbb{E}(\log \sigma(\cdot))$ term can be solved by Gaussian quadrature.

5.6 Complexity

The analysis of complexity of mean-field approach is similar with that of EM algorithm. For the fixed $\mu(t)$, $\phi(\tau)$ and T_ϕ, as T increases, the complexity scales *linearly* with data size: $\mathcal{O}(N)$ due to the sparsity of expectation of branching structure. Since the mean-field approach computes not only the mean but also the variance, it is slightly slower than the EM algorithm.

5.7 Hyperparameters

Similarly, the hyperparameters θ_0 and θ_1 can be optimized by performing maximization of ELBO over $\{\theta_0, \theta_1\}$ using numerical packages. The optimization of number and location of inducing points is same as EM algorithm. The final pseudo code is provided in Alg.3.

Algorithm 3: Mean-field algorithm for SGHP

Result: $\mu(t) = \lambda_\mu^* \sigma(f(t))$, $\phi(\tau) = \lambda_\phi^* \sigma(g(\tau))$
Initialize hyperparameters and variational distributions of $\mathbf{X}$, λ_μ^*, λ_ϕ^*, $\boldsymbol{\omega}_{ii}$, $\boldsymbol{\omega}_{ij}$, $\mathbf{f}_{t_s}$, $\mathbf{g}_{\tau_s}$, Π_μ, $\{\Pi_{\phi_i}\}_{i=1}^N$;
for do
  Update q_1 of $\boldsymbol{\omega}_{ii}$ and $\boldsymbol{\omega}_{ij}$ by Eq.(33);
  Update Λ^1 of Π_μ and $\{\Pi_\phi\}$ by Eq.(34);
  Update q_2 of λ_μ^* and λ_ϕ^* by Eq.(35);
  Update q_2 of $\mathbf{f}_{t_s}$ and $\mathbf{g}_{\tau_s}$ by Eq.(36);
  Update q_1 of $\mathbf{X}$ by Eq.(37);
  Update hyperparameters.
end

6. Experiments

We evaluate the performance of our proposed Gibbs sampler, EM and mean-field (MF) algorithms on both simulated and real-world data. Specifically, we compare our proposed algorithms to the following alternatives.

- *Sigmoid Gaussian Cox Process (SGCP)*: an inhomogeneous Poisson process where the intensity is modeled as a scaled sigmoid transformation of a GP (Adams et al., 2009). This baseline is used for real data only as the ground truth in simulated data is fixed to Hawkes process.

- *Maximum Likelihood Estimation (MLE)*: the vanilla Hawkes process with constant μ and exponential decay triggering kernel $\alpha \exp(-\beta(t-t_i))$. The inference is performed by MLE (Ozaki, 1979).

- *Wiener-Hopf (WH)*: a nonparametric algorithm for Hawkes process where μ is a constant and $\phi(\tau)$ is a continuous nonparametric function. The inference is based on the solution of a Wiener-Hopf equation (Bacry and Muzy, 2016).

- *Majorization Minimization Euler-Lagrange (MMEL)*: a nonparametric algorithm for the Hawkes process with constant μ and smooth $\phi(\tau)$, which similarly utilized the branching structure and estimated $\phi(\tau)$ by an Euler-Lagrange equation (Zhou et al., 2013).

We also tried to compare to the long short-term memory (LSTM) based neural Hawkes process (Mei and Eisner, 2017) but found it hard to converge at least on our data. On the contrary, our proposed algorithms are easier to converge due to the fact that there are fewer parameters to tune, which constitutes another advantage.

We use the following metrics to evaluate the performance of various methods:

- *TestLL*: the log-likelihood of hold-out data using the trained model. This is a metric describing the model prediction ability.

- *EstErr*: the mean squared error between the estimated $\hat{\mu}(t)$, $\hat{\phi}(\tau)$ and the ground truth. It is only used for *simulated data*.

- *PreAcc*: given an event sequence $\{t_n\}_{n=1}^{i-1}$, we wish to predict the time of t_i. The expectation of t_i is $\mathbb{E}[t_i] = \int_{t_{i-1}}^{\infty} tp(t_i = t)dt$ with $P(t_i = t) = \lambda(t) \exp\left(-\int_{t_{i-1}}^{t} \lambda(s)ds\right)$. The integral can be estimated by Monte Carlo method. We predict multiple timestamps in a sequence: if the predicted $\hat{t}_i$ is within an error bound ϵ, then it is considered to be a correct prediction; or it is a wrong one. The percentage of correct prediction is defined as the prediction accuracy. It is only used for *real data*.

- *RunTime*: the running time of various methods w.r.t. the number of training data.

6.1 Simulated Data Experiments

In simulated data experiments, we use the thinning algorithm (Ogata, 1998) to generate 100 sets of training data and 10 sets of test data with $T_\phi = 6$ and $T = 100$ in three cases:

1. $\mu(t) = 1$ and $\phi(\tau) = 1 \cdot \exp(-2\tau)$;

2. $\mu = 1$ and $\phi(\tau) = \begin{cases} 0.33 \sin \tau & (0 < \tau \leq \pi) \\ 0 & (\pi < \tau < T_\phi) \end{cases}$;

3. $\mu(t) = \sin\left(\frac{2\pi}{T} \cdot t\right) + 1 \quad (0 < t < T)$ and $\phi(\tau) = 0.3\left(\sin(\frac{2\pi}{3} \cdot \tau) + 1\right) \cdot \exp(-0.7\tau) \quad (0 < \tau < T_\phi)$.

The first case is the traditional case with a constant μ and an exponential decay $\phi(\tau)$ while the second case has a half sinusoidal $\phi(\tau)$. The third case is the most general one with time-changing baseline intensity and sinusoidal exponential decay $\phi(\tau)$. The first case is used to show that our SGHP model can work well for the classic scenario; the second case is to show SGHP can recover a constant baseline intensity with a non-exponential-decay triggering kernel; the third case is the most general case which demonstrates the powerful fitting ability of SGHP fitting the time-changing $\mu(t)$ and flexible non-exponential-decay $\phi(\tau)$ simultaneously.

The inducing points and hyperparameters are optimized for inference. The estimated $\hat{\mu}(t)$ and $\hat{\phi}(\tau)$ are shown in Fig1, 2 and 3. The learned results from Gibbs, EM and MF are similar with each other with Gibbs providing an accurate posterior, EM providing a MAP estimate and MF providing an approximated posterior. The posterior variance obtained

from MF is relatively smaller than that of Gibbs, which is a well known result in Blei et al. (2017).

For three cases, we compare the alternative baseline inference algorithms to our Gibbs sampler in Fig.1, EM algorithm in Fig.2 and MF approach in Fig.3. The corresponding *TestLL* and *EstErr* are computed in Tab.2. For Gibbs and MF, multiple trajectories are sampled from the posterior distribution and *TestLL* and *EstErr* are reported with mean and standard deviation. To show the convergence of three algorithms, the training loglikelihood curves are plotted for Gibbs, EM and MF of three cases in Fig.1d, 2d and 3d. It is observed that the loglikelihood of three algorithms in three cases reaches a plateau after 50 loops indicating excellent convergence.

From accuracy perspective, the result in Tab.2 confirms that our Gibbs, EM and MF algorithms outperform alternatives in most cases except the first case. For the first case, the MLE is the champion w.r.t. *TestLL* because the parametric assumption coincides with the ground truth which is a rare scenario in real applications, but other algorithms are still competitive. It is worth noting that because the decay parameter β is not fixed in our case, the MLE is a non-convex optimization but the Gibbs sampler can avoid the local maximum and that may be the reason the *EstErr* of Gibbs is better. For the second case, the MLE estimation is far away from the ground truth due to parametric constraints on both $\mu(t)$ and $\phi(\tau)$; our SGHP model achieves comparable performance with WH and MMEL because all of them can model the constant baseline intensity and flexible triggering kernel. For the third case, the MLE estimation still deviates severely from the ground truth; for WH and MMEL, the $\mu(t)$ is limited to be constant which in turn affects the estimation of $\phi(\tau)$; on the contrary, our Gibbs, EM and MF algorithms provide the most flexible estimation of both $\hat{\mu}(t)$ and $\hat{\phi}(\tau)$, that is the reason their estimation result is closest to the ground truth. The EM and MF algorithms are in general slightly outperformed by the Gibbs sampler perhaps because the Gibbs sampler can characterize the true posterior more accurately than EM and MF.

From efficiency perspective, the *RunTime* of Gibbs, EM and MF algorithms are compared to the other iterative nonparametric algorithm MMEL (WH is excluded as it is based on the solution of a linear system without the need of iteration, SGCP is excluded as it is a Poisson process model) with the same number of iterations. In Fig.4, we can see Gibbs, EM and MF have superior efficiency to MMEL with complexity scaling linearly with observation as we analysed in the section of complexity. The reason our proposed algorithms are efficient is: on one side, the sparse GP is utilized to reduce the complexity of matrix inversion; on the other side, all proposed algorithms have explicit closed-form expressions due to the conjugacy induced by the augmentation of auxiliary latent variables. Furthermore, Gibbs sampler is less efficient than EM and MF because the sampling procedure of latent Poisson processes is computation intensive; MF is slightly slower than EM due to the extra computation of variance.

6.2 Real Data Experiments

We compare various methods on two real-world data sets of crime. In criminology, there is a self-exciting phenomenon from past crimes to future ones which is reported in Mohler et al. (2011). The two data sets both comprise times of security violation or report in a

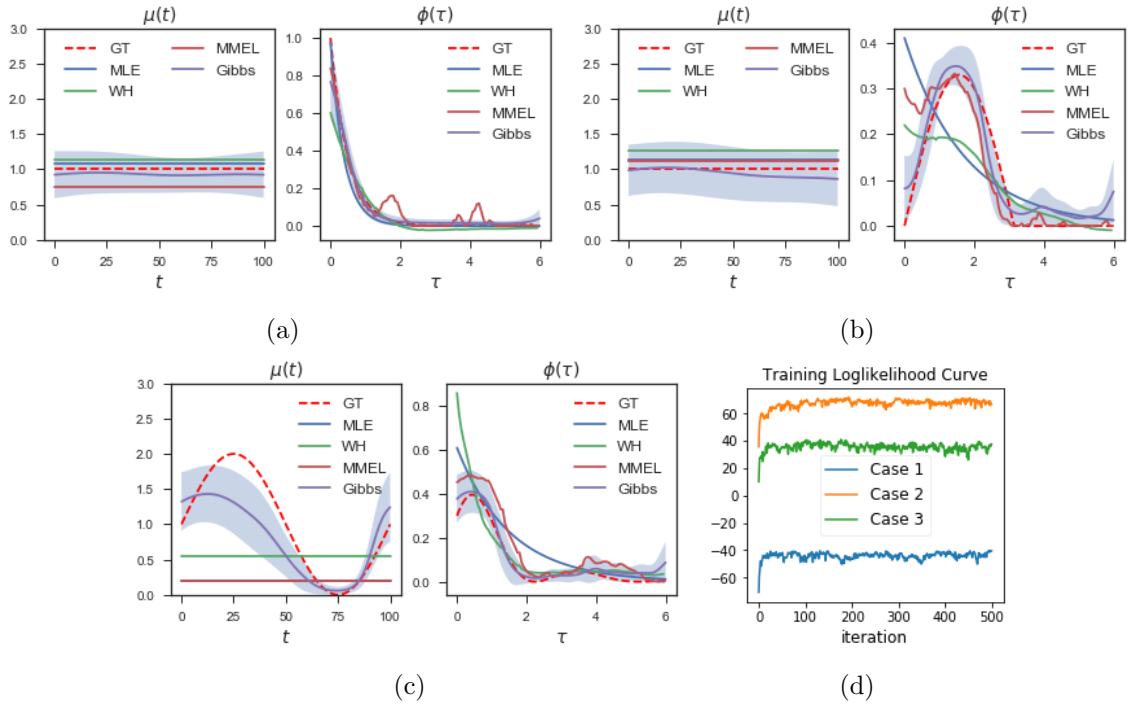


Figure 1: Gibbs sampler: simulated data experimental results. (a): The estimated $\hat{\mu}(t)$ and $\hat{\phi}(\tau)$ (with shading being one standard deviation) for Case 1; (b): for Case 2; (c): for Case 3; (d): the training loglikelihood of Gibbs sampler in three cases. (GT=Ground Truth)

period of several years. For each data set, we aim to test the goodness-of-fit on test data (*TestLL*) and predict the time of event happening in the future time window (*PreAcc*).

6.2.1 Crime in Vancouver (Canada)

The data of crimes in Vancouver[1] comes from the Vancouver Open Data Catalogue. It includes miscellaneous crimes from 2003-01-01 to 2017-07-13. The columns are crime type, year, month, day, hour, minute, block, neighbourhood, latitude, longitude etc.

6.2.2 NYPD Complaint Data

This data set[2] includes all valid felony, misdemeanour and violation crimes reported to the New York police department (NYPD) for all complete quarters so far in 2017. The columns are complaint number, date, time, offense description, borough etc.

1. `https://www.kaggle.com/wosaku/crime-in-vancouver`
2. `https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Current-YTD/5uac-w243`

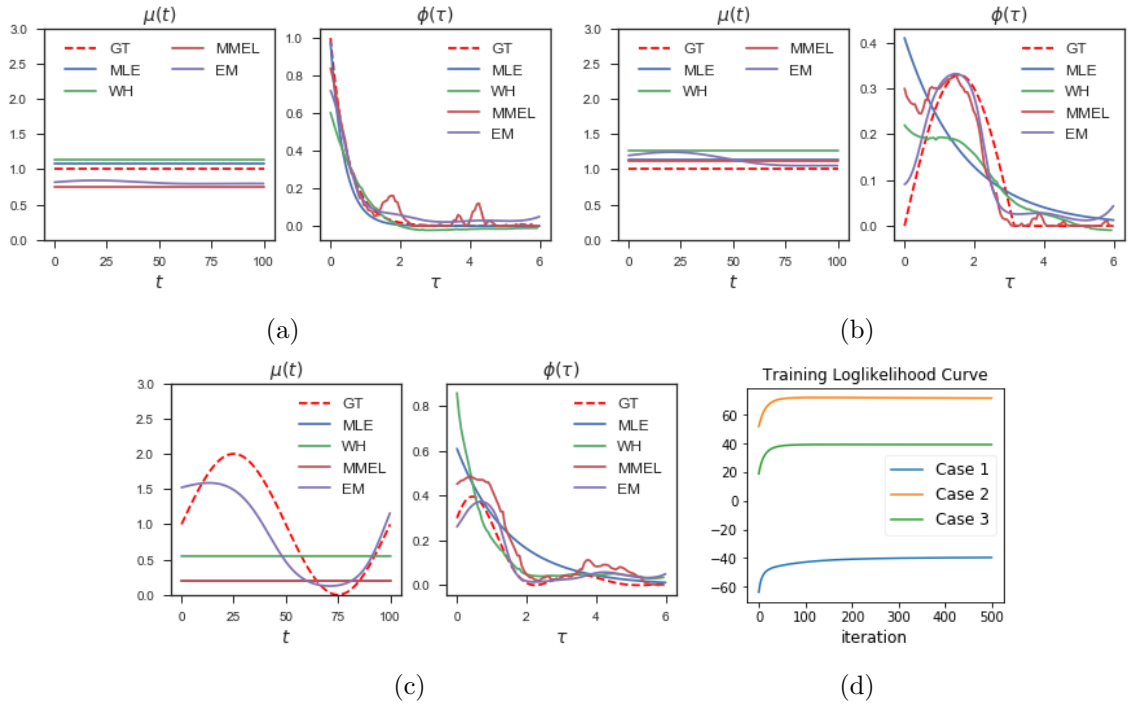


Figure 2: EM algorithm: simulated data experimental results. (a): The estimated $\hat{\mu}(t)$ and $\hat{\phi}(\tau)$ for Case 1; (b): for Case 2; (c) for Case 3; (d): the training loglikelihood of EM in three cases.(GT=Ground Truth)

6.2.3 Preliminary Setup

For Crime in Vancouver, we filter out the theft records from June to November in 2016 happening in central business district and add a small time interval to separate all the simultaneous records. For NYPD Complaint Data, we filter out the complaints records in Brooklyn and Queens in 2016 with the offense description being petit larceny. For each of these data sets, we split the timestamps of events into a train and test set. The precise split scheme varies for each data set as follows: for Crime in Vancouver, the first 519 data points are selected as training set to train the models with the rest being test data (time unit: days); for NYPD Complaint Data, the first 324 data points are selected as training set with the rest being test (time unit: days). For the prediction task, we assume the top 17% of a sequence is observed ($\epsilon = 0.14$ for Crime in Vancouver and 1 for NYPD Complaint Data where the choice of ϵ only affects the absolute magnitude of prediction accuracy but not the relative magnitude, 400 samples for Monte Carlo integration) and then predict the time of next event, and then the real time of next event is incorporated into the observed data and then predict the further next one and the iteration goes on.

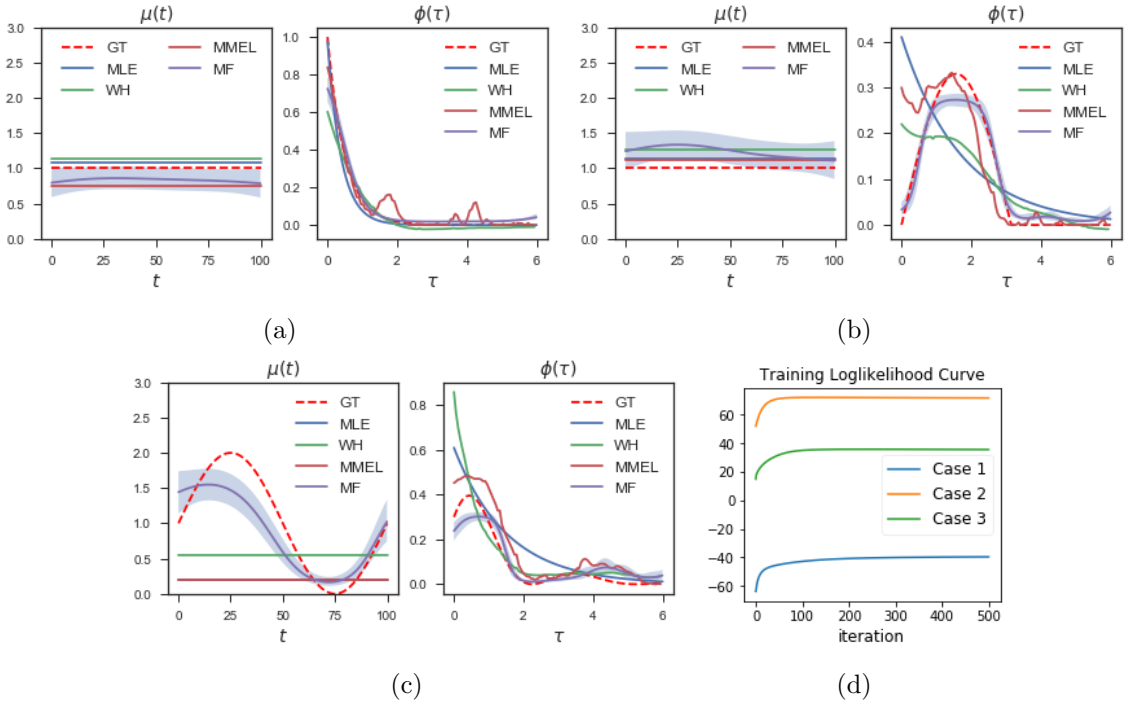
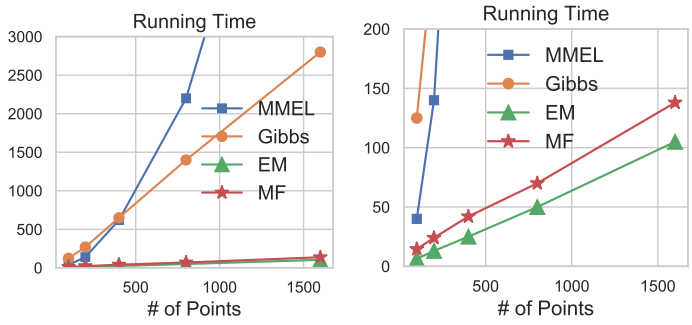


Figure 3: MF approach: simulated data experimental results. (a): The estimated $\hat{\mu}(t)$ and $\hat{\phi}(\tau)$ (with shading being one standard deviation) for Case 1; (b): for Case 2; (c) for Case 3; (d): the training loglikelihood of MF in three cases where the mean is used. (GT=Ground Truth)

		MLE	WH	MMEL	Gibbs	EM	MF
Case 1	$EstErr(\hat{\mu}, \mu)$	0.077	0.142	0.253	**0.068** (0.045)	0.186	0.165 (0.031)
	$EstErr(\hat{\phi}, \phi)$	0.0012	0.0045	0.0021	**0.0008** (0.0032)	0.0017	0.0013 (0.0015)
	TestLL	**-55.75**	-56.04	-56.77	-56.12 (2.05)	-56.53	-56.51 (1.51)
Case 2	$EstErr(\hat{\mu}, \mu)$	0.144	0.263	0.123	**0.053** (0.134)	0.141	0.236 (0.035)
	$EstErr(\hat{\phi}, \phi)$	0.0288	0.0065	0.0053	0.0012 (0.0151)	0.0016	**0.0005** (0.0071)
	TestLL	27.92	30.13	30.21	**31.89** (1.52)	30.96	31.33 (1.06)
Case 3	$EstErr(\hat{\mu}, \mu)$	1.141	0.701	1.153	0.165 (0.042)	0.134	**0.112** (0.023)
	$EstErr(\hat{\phi}, \phi)$	0.0076	0.0093	0.0058	**0.0008** (0.0125)	0.0011	0.0019 (0.0068)
	TestLL	32.93	28.66	32.26	**38.94** (2.31)	38.21	37.43 (1.22)

Table 2: *EstErr* and *TestLL* for simulated data sets. For distribution estimation algorithms Gibbs and MF, the mean and standard deviation (in brackets) are computed after running each experiment 5 times.



(a) (b)

Figure 4: The running time (seconds) of different iterative nonparametric algorithms on varying # of observations with 100 iterations. Gibbs, EM and MF algorithms scale linearly with observation, which is more efficient than MMEL. (b) is the zoom in of the bottom of (a).

6.2.4 Results

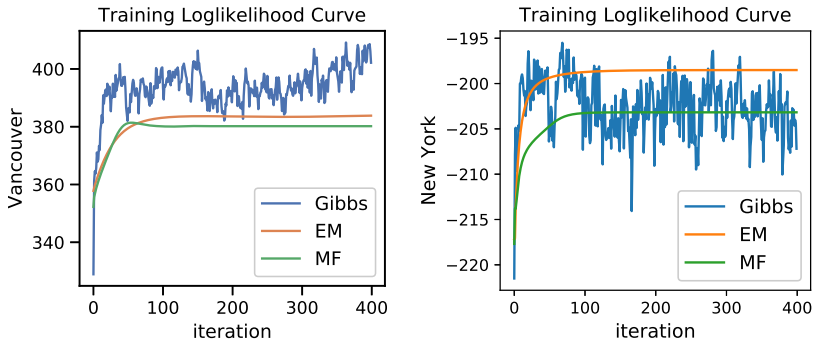
To access the convergence of our proposed algorithms, the training loglikelihood curves of three algorithms are plotted in Fig.5 for both data sets. We can see all three algorithms reach a plateau after 100 loops. For crime in Vancouver, the Gibbs sampler achieves a higher training loglikelihood compared to EM and MF; this might be the result of the non-convexity of loss which we will discuss in Section 7. For NYPD complaint data, the Gibbs sampler and EM are higher than MF, this could be due to the fact that the MF provides an approximated posterior whose mean deviates from the mode of true posterior.

The *TestLL* of our proposed algorithms and alternatives are shown in Tab. 3. We can see all Hawkes-based models outperform SGCP; this demonstrates the necessity of utilizing Hawkes process to model the self-exciting phenomenon in crime domain. Also, WH, MMEL, Gibbs, EM and MF all outperform MLE, which demonstrates the necessity of nonparametric models to capture the underlying dynamic triggering effect. Besides, Gibbs, EM and MF's consistent superiority over other nonparametric models with constant baseline intensity demonstrates that the time-changing baseline intensity does provide a better fitting capability. Our SGHP model has a natural advantage because it not only captures the completely flexible $\mu(t)$ and $\phi(\tau)$ leading to better goodness-of-fit but also has the superior efficiency.

We also measure the *PreAcc* of all alternatives on both data sets. The average *PreAcc* of test data is shown in Fig.6. In general, the *PreAcc* result is consistent with *TestLL*. From the results of *PreAcc*, the Hawkes-based models outperforming SGCP demonstrates the self-exciting effect in both real data sets; all nonparametric models outperforming MLE demonstrates the better fitting capability of nonparametric models; the performance of Gibbs, EM and MF is comparable with that of other nonparametric models. We can see that the self-exciting phenomenon in the data of crime in Vancouver is more obvious than that in the NYPD complaint data because the Hawkes-based models outperform SGCP with

TestLL	SGCP	MLE	WH	MMEL	Gibbs	EM	MF
Vancouver	299.77 (13.11)	380.88	400.36	386.66	430.43 (11.47)	**458.27**	453.11 (8.94)
New York	-292.30 (4.97)	-276.00	-225.93	-198.80	**-193.76** (4.59)	-195.08	-200.70 (3.32)

Table 3: *TestLL* for real data sets. For distribution estimation algorithms SGCP, Gibbs and MF, the mean and standard deviation (in brackets) are computed after running each experiment 5 times.



(a) (b)

Figure 5: The training loglikelihood curves of Gibbs, EM and MF for real data (a): crime in Vancouver; (b): NYPD complaint data.

a larger magnitude in the data of crime in Vancouver than that in the NYPD complaint data.

7. Discussions and Conclusions

In this section, we analyze the advantages and disadvantages of each algorithm and the most appropriate application scenario for each algorithm and discuss the relationship among our proposed algorithms, then draw a conclusion and discuss some future research directions in the end.

7.1 Which Algorithm to Use

We proposed three inference algorithms in this paper, so a natural question is: which algorithm should be used or which one is better? The experimental results have verified that the estimation results of three algorithms are close to each other and it is difficult to say which one is definitely better than the others. Which one to use depends on what sort of statistics you desire to estimate (approximate v.s. exact; distribution v.s. point) in the application. Gibbs sampler provides an asymptotically exact estimation of the posterior but is the least efficient algorithm among the three; EM is the fastest algorithm and provides an exact MAP estimate, but it cannot characterize the uncertainty as a point estimator and

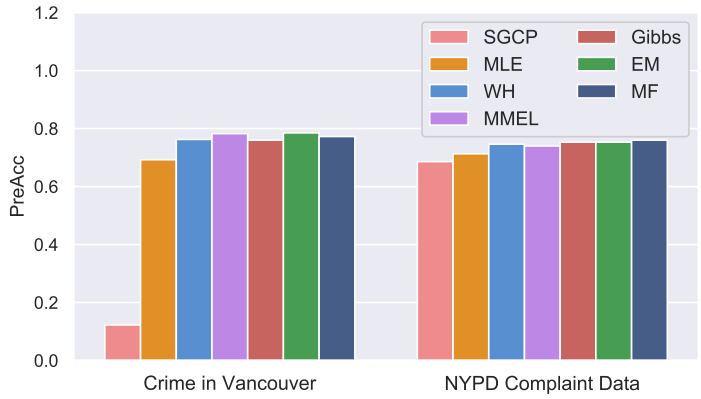


Figure 6: The *PreAcc* of Gibbs, EM, MF and alternatives on two real data sets.

may fall into a local maximum; MF combines the advantages of speed and uncertainty but the estimated posterior distribution is an approximation and may also get a locally optimal solution. To summarize, if the efficiency is the first consideration, EM and MF should be given the priority; if the uncertainty is necessary, Gibbs and MF are recommended; if the accuracy is the most crucial, Gibbs and EM are superior.

7.2 Relation Among Gibbs, EM and MF

To some extent, the Gibbs sampler of our SGHP model can be considered as a stochastic EM (SEM) algorithm (Celeux, 1985) for Hawkes process. In deterministic EM Hawkes process algorithms e.g. our proposed EM, Lewis and Mohler (2011) and Zhou et al. (2013), the basic idea is to compute the branching structure in probabilistic counterpart and maximize the corresponding surrogate (penalized) likelihood. Instead of computing the surrogate likelihood based on probabilistic branching structure, we incorporate a stochastic branching structure sampling step between E and M steps. This results in a much easier update of the parameters based on the pseudo-completed data. Moreover, unlike the deterministic EM trying to maximize the surrogate likelihood, we sample the conditional posterior in M step. This prompts the transformation between EM-Hawkes and Gibbs-Hawkes. Although the Gibbs sampler is time-consuming compared to EM and MF, an advantage of Gibbs-Hawkes is, with a complex surrogate function, the random samplings prevent converging to saddle points or local maxima while the EM-Hawkes and MF-Hawkes cannot avoid. As we stated earlier, the MF-Hawkes can be treated as an extension of the EM-Hawkes from the most probable value of each parameter (MAP estimate) and true posterior distribution of latent variables to Bayesian estimation which estimates an approximated posterior distribution of the parameters and latent variables.

7.3 Conclusions and Prospects

To conclude, we propose the sigmoid Gaussian Hawkes process model which nonparametricly represents the baseline intensity and triggering kernel as the scaled sigmoid trans-

formation of Gaussian processes. By augmenting the branching structure, Pólya-Gamma random variables and latent marked Poisson processes, the likelihood is transformed to a conjugate form with the prior. Three efficient inference algorithms: Gibbs sampler, EM algorithm and mean-filed variational inference, are proposed. The choice of inference algorithms depends on the requirement in application. Moreover, by introducing sparse GP approximation, our proposed algorithms are further accelerated. The simulated and real data experimental results confirm that the fitting capability of our proposed algorithms is superior to the state-of-the-art alternatives with better efficiency at the same time.

In this paper, we considered the single-variate and single-dimensional Hawkes process for convenience. Future research can be done on the extension to multivariate or multidimensional Hawkes processes. Multidimensional Hawkes processes are the ideal model for modeling the self-exciting characteristics of events occurring in a high dimensional space, e.g. the 2D planar space, which is suitable for some specific applications such as in the earth quake domain. Multivariate Hawkes processes are designed to characterize the mutually-exciting phenomenon among multiple instances. Both aforementioned models require more complex computation which raises even more challenging problems.

Appendix A. Proof of Transformation of Sigmoid Function

Polson et al. (2013) found that the inverse hyperbolic cosine can be expressed as an infinite mixture of Gaussian densities

$$\cosh^{-b}(z/2) = \int_0^\infty e^{-z^2\omega/2} p_{\mathrm{PG}}(\omega|b,0)d\omega, \tag{38}$$

where $p_{\mathrm{PG}}(\omega|b,0)$ is the Pólya-Gamma distribution with $\omega \in \mathbb{R}^+$. As a result, the sigmoid function can be defined as a Gaussian representation

$$\sigma(z) = \frac{e^{z/2}}{2\cosh(z/2)} = \int_0^\infty e^{h(\omega,z)} p_{\mathrm{PG}}(\omega|1,0)d\omega, \tag{39}$$

where $h(\omega,z) = z/2 - z^2\omega/2 - \log 2$. This proves Eq.(4) in the main paper.

Appendix B. Campbell's Theorem

Let $\Pi_{\hat{\mathcal{Z}}} = \{(\mathbf{z}_n, \boldsymbol{\omega}_n)\}_{n=1}^N$ be a marked Poisson process on the product space $\hat{\mathcal{Z}} = \mathcal{Z} \times \Omega$ with intensity $\Lambda(\mathbf{z},\boldsymbol{\omega}) = \Lambda(\mathbf{z})p(\boldsymbol{\omega}|\mathbf{z})$. $\Lambda(\mathbf{z})$ is the intensity for the unmarked Poisson process $\{\mathbf{z}_n\}_{n=1}^N$ with $\boldsymbol{\omega}_n \sim p(\boldsymbol{\omega}_n|\mathbf{z}_n)$ being an independent mark drawn at each $\mathbf{z}_n$. Furthermore, we define a function $h(\mathbf{z},\boldsymbol{\omega}) : \mathcal{Z} \times \Omega \to \mathbb{R}$ and the sum $H(\Pi_{\hat{\mathcal{Z}}}) = \sum_{(\mathbf{z},\boldsymbol{\omega})\in\Pi_{\hat{\mathcal{Z}}}} h(\mathbf{z},\boldsymbol{\omega})$. If $\Lambda(\mathbf{z},\boldsymbol{\omega}) < \infty$, then

$$\mathbb{E}_{\Pi_{\hat{\mathcal{Z}}}}\left[\exp\left(\xi H(\Pi_{\hat{\mathcal{Z}}})\right)\right] = \exp\left[\int_{\hat{\mathcal{Z}}} \left(e^{\xi h(\mathbf{z},\boldsymbol{\omega})} - 1\right)\Lambda(\mathbf{z},\boldsymbol{\omega})d\boldsymbol{\omega}d\mathbf{z}\right],$$

for any $\xi \in \mathbb{C}$. The above equation defines the characteristic functional of a marked Poisson process. This proves Eq.(6) in the main paper. The mean and variance are

$$\mathbb{E}_{\Pi_{\hat{\mathcal{Z}}}}\left[H(\Pi_{\hat{\mathcal{Z}}})\right] = \int_{\hat{\mathcal{Z}}} h(\mathbf{z},\boldsymbol{\omega})\Lambda(\mathbf{z},\boldsymbol{\omega})d\boldsymbol{\omega}d\mathbf{z}$$
$$\mathrm{Var}_{\Pi_{\hat{\mathcal{Z}}}}\left[H(\Pi_{\hat{\mathcal{Z}}})\right] = \int_{\hat{\mathcal{Z}}} [h(\mathbf{z},\boldsymbol{\omega})]^2\Lambda(\mathbf{z},\boldsymbol{\omega})d\boldsymbol{\omega}d\mathbf{z}$$

Appendix C. Proof of Augmented Likelihood

Substituting Eq.(4) and (6) into Eq.(3) in the main paper, we obtain the augmented joint likelihood of baseline intensity part

$$\begin{aligned}
&p(D, \mathbf{X}_{ii}|\lambda_\mu^*, f)\\
&= \prod_{i=1}^{N} \left(\lambda_\mu^* \sigma(f(t_i))\right)^{x_{ii}} \exp\left(-\int_T \lambda_\mu^* \sigma(f(t))dt\right)\\
&= \prod_{i=1}^{N} \left(\int_0^\infty \lambda_\mu^* e^{h(\omega_{ii},f(t_i))} p_{\mathrm{PG}}(\omega_{ii}|1,0)d\omega_{ii}\right)^{x_{ii}} \cdot \mathbb{E}_{p_{\lambda_\mu}}\left[\prod_{(\omega_\mu,t)\in\Pi_\mu} e^{h(\omega_\mu,-f(t))}\right]\\
&= \iint \prod_{i=1}^{N} \left(\lambda_\mu(t_i,\omega_{ii}) e^{h(\omega_{ii},f(t_i))}\right)^{x_{ii}} \cdot p_{\lambda_\mu}(\Pi_\mu|\lambda_\mu^*) \prod_{(\omega_\mu,t)\in\Pi_\mu} e^{h(\omega_\mu,-f(t))} d\boldsymbol{\omega}_{ii} d\Pi_\mu.
\end{aligned}$$

with $\boldsymbol{\omega}_{ii}$ denoting a vector of ω_{ii} and $\lambda_\mu(t_i,\omega_{ii}) = \lambda_\mu^* p_{\mathrm{PG}}(\omega_{ii}|1,0)$. Therefore, the augmented joint likelihood is

$$p(D, \Pi_\mu, \boldsymbol{\omega}_{ii}, \mathbf{X}_{ii}|\lambda_\mu^*, f) = \prod_{i=1}^{N} \left(\lambda_\mu(t_i,\omega_{ii}) e^{h(\omega_{ii},f(t_i))}\right)^{x_{ii}} \cdot p_{\lambda_\mu}(\Pi_\mu|\lambda_\mu^*) \prod_{(\omega_\mu,t)\in\Pi_\mu} e^{h(\omega_\mu,-f(t))}$$

This proves Eq.(7) in the main paper. The proof of Eq.(8) in the main paper is same and omitted here.

Appendix D. Comparison with Another Related Work

In this section, we provide some comparison results with another related work by Zhou (2019) which also utilized GP to model the nonparametric Hawkes process. The model in Zhou (2019) is a Hawkes process in which both the baseline intensity and triggering kernel are modeled by the square transformation of GPs to guarantee the nonnegativity, which is the key difference with our model (sigmoid link function) in this paper. The comparison in this section illustrates how the specific link function (sigmoid vs. square) impacts the inference procedure. Generally speaking, both Zhou (2019) and our proposed method use the sparse GP to accelerate the inference and provide the flexible $\mu(t)$ and $\phi(\tau)$ simultaneously, but the inference algorithm in Zhou (2019) is computation intensive compared with our method because it performs optimization to find the optimal covariance matrix of variational distribution wherein each optimization loop requires a complex matrix computation.

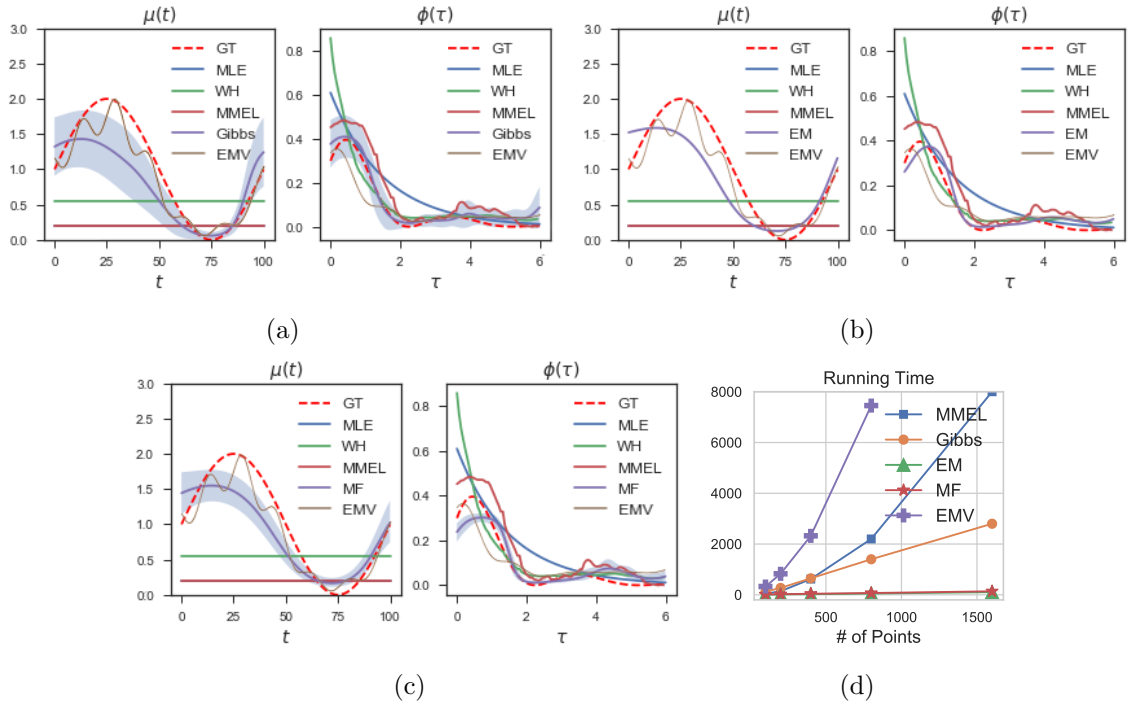


Figure 7: The estimated $\hat{\mu}(t)$ and $\hat{\phi}(\tau)$ of our proposed algorithms, EMV and alternatives for the simulated data in Case 3. (a): for Gibbs sampler; (b) for EM; (c) for MF. (with shading being one standard deviation, GT=Ground Truth); (d): the running time (seconds) of different iterative nonparametric algorithms on varying # of observations with 100 iterations.

	MLE	WH	MMEL	EMV	Gibbs	EM	MF
$EstErr(\hat{\mu}, \mu)$	1.141	0.701	1.153	**0.046**	0.165 (0.042)	0.134	0.112 (0.023)
$EstErr(\hat{\phi}, \phi)$	0.0076	0.0093	0.0058	0.0039	**0.0008** (0.0125)	0.0011	0.0019 (0.0068)
TestLL	32.93	28.66	32.26	38.88	**38.94** (2.31)	38.21	37.43 (1.22)

Table 4: The *EstErr* and *TestLL* results of our methods, EMV and alternatives for Case 3. For the distribution estimation algorithms Gibbs and MF, the mean and standard deviation (in brackets) are computed after running each experiment 5 times.

We perform experiments to compare the inference algorithm EM-variational (denoted by EMV) in Zhou (2019) with our three inference algorithms and alternatives w.r.t. the estimation accuracy (*EstErr*), test loglikelihood (*TestLL*) and running time for the synthetic data in Case 3. The estimated $\hat{\mu}(t)$ and $\hat{\phi}(\tau)$ are shown in Fig.7. To compare the estimation results quantitatively, the *EstErr* and *TestLL* are shown in Tab.4 where our proposed methods are competitive with EMV. Also, the running time w.r.t. different number of observations is shown in Fig.7d where our proposed methods are more efficient than EMV.

References

Ryan Prescott Adams, Iain Murray, and David JC MacKay. Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities. In *International Conference on Machine Learning*, pages 9–16. ACM, 2009.

Emmanuel Bacry and Jean-François Muzy. First-and second-order statistics characterization of Hawkes processes and non-parametric estimation. *IEEE Transactions on Information Theory*, 62(4):2184–2202, 2016.

Emmanuel Bacry, Iacopo Mastromatteo, and Jean-François Muzy. Hawkes processes in finance. *Market Microstructure and Liquidity*, 1(01):1550005, 2015.

Christopher M Bishop. *Pattern Recognition and Machine Learning*. springer, 2006.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Gilles Celeux. The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly*, 2:73–82, 1985.

Feng Chen and Wai Hong Tan. Marked self-exciting point process modelling of information diffusion on twitter. *The Annals of Applied Statistics*, 12(4):2175–2196, 2018.

John P Cunningham, Krishna V Shenoy, and Maneesh Sahani. Fast Gaussian process methods for point process intensity estimation. In *International Conference on Machine Learning*, pages 192–199. ACM, 2008.

Daryl J Daley and David Vere-Jones. *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*. Springer Science & Business Media, 2003.

Christian Donner. *Bayesian Inference of Inhomogeneous Point Process Models*. PhD thesis, Technische Universität Berlin, 2019.

Christian Donner and Manfred Opper. Efficient Bayesian inference for a Gaussian process density model. *arXiv preprint arXiv:1805.11494*, 2018.

Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel Gomez-Rodriguez, and Le Song. Recurrent marked temporal point processes: embedding event history to vector. In *International Conference on Knowledge Discovery and Data Mining*, pages 1555–1564. ACM, 2016.

Michael Eichler, Rainer Dahlhaus, and Johannes Dueck. Graphical modeling for multivariate Hawkes processes with nonparametric link functions. *Journal of Time Series Analysis*, 38(2):225–242, 2017.

Seth Flaxman, Yee Whye Teh, and Dino Sejdinovic. Poisson intensity estimation with reproducing kernels. *Electronic Journal of Statistics*, 11(2):5081–5104, 2017.

Gene H Golub and John H Welsch. Calculation of Gauss quadrature rules. *Mathematics of Computation*, 23(106):221–230, 1969.

Amrita Gupta, Mehrdad Farajtabar, Bistra Dilkina, and Hongyuan Zha. Discrete interventions in Hawkes processes with applications in invasive species management. In *International Joint Conferences on Artificial Intelligence*, pages 3385–3392, 2018.

WK Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, pages 97–109, 1970.

AG Hawkes. Cluster models for earthquakes-regional comparisons. *Bulletin of the International Statistical Institute*, 45(3):454–461, 1973.

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

John Frank Charles Kingman. Poisson processes. *Encyclopedia of Biostatistics*, 6, 2005.

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

Yanchi Liu, Tan Yan, and Haifeng Chen. Exploiting graph regularized multi-dimensional Hawkes processes for modeling events with spatio-temporal characteristics. In *International Joint Conferences on Artificial Intelligence*, pages 2475–2482, 2018.

Chris Lloyd, Tom Gunter, Michael Osborne, and Stephen Roberts. Variational inference for Gaussian process modulated Poisson processes. In *International Conference on Machine Learning*, pages 1814–1822, 2015.

David Marsan and Olivier Lengline. Extending earthquakes' reach through cascading. *Science*, 319(5866):1076–1079, 2008.

Hongyuan Mei and Jason M Eisner. The neural Hawkes process: A neurally self-modulating multivariate point process. In *Advances in Neural Information Processing Systems*, pages 6754–6764, 2017.

George O Mohler, Martin B Short, P Jeffrey Brantingham, Frederic Paik Schoenberg, and George E Tita. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493):100–108, 2011.

Jesper Møller, Anne Randi Syversveen, and Rasmus Plenge Waagepetersen. Log Gaussian Cox processes. *Scandinavian Journal of Statistics*, 25(3):451–482, 1998.

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

Tohru Ozaki. Maximum likelihood estimation of Hawkes' self-exciting point processes. *Annals of the Institute of Statistical Mathematics*, 31(1):145–155, 1979.

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

Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer, 2003.

Patricia Reynaud-Bouret and Sophie Schbath. Adaptive estimation for Hawkes processes; application to genome analysis. *The Annals of Statistics*, 38(5):2781–2822, 2010.

Yves-Laurent Kom Samo and Stephen Roberts. Scalable nonparametric Bayesian inference on point processes with Gaussian processes. In *International Conference on Machine Learning*, pages 2227–2236, 2015.

Michalis Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *Artificial Intelligence and Statistics*, pages 567–574, 2009.

Rui Zhang, Christian Walder, Marian-Andrei Rizoiu, and Lexing Xie. Efficient non-parametric Bayesian Hawkes processes. *arXiv preprint arXiv:1810.03730*, 2018.

Rui Zhang, Christian Walder, and Marian-Andrei Rizoiu. Variational inference for sparse Gaussian process modulated Hawkes process. *arXiv preprint arXiv:1905.10496v2*, 2019.

Feng Zhou. Efficient EM-variational inference for Hawkes process. *arXiv preprint arXiv:1905.12251*, 2019.

Feng Zhou, Zhidong Li, Xuhui Fan, Yang Wang, Arcot Sowmya, and Fang Chen. A refined MISD algorithm based on Gaussian process regression. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 584–596. Springer, 2018.

Ke Zhou, Hongyuan Zha, and Le Song. Learning triggering kernels for multi-dimensional Hawkes processes. In *International Conference on Machine Learning*, pages 1301–1309. ACM, 2013.