Sparse Gaussian Process Modulated Hawkes Process

Rui Zhang¹, Christian Walder², and Marian-Andrei Rizoiu³

¹Research School of Computer Science, The Australian National University
²CSIRO, Australia
³The Faculty of Engineering and IT, University of Technology Sydney

Abstract

The Hawkes process has been widely applied to modeling self-exciting events, including neuron spikes, earthquakes and tweets. To avoid designing parametric kernel functions and to be able to quantify the prediction confidence, non-parametric Bayesian Hawkes processes have been proposed. However the inference of such models suffers from unscalability or slow convergence. In this paper, we first propose a new non-parametric Bayesian Hawkes process whose triggering kernel is modeled as a squared sparse Gaussian process. Second, we present the variational inference scheme for the model optimization, which has the advantage of linear time complexity by leveraging the stationarity of the triggering kernel. Third, we contribute a tighter lower bound than the evidence lower bound of the marginal likelihood for the model selection. Finally, we exploit synthetic data and large-scale social media data to validate the efficiency of our method and the practical utility of our approximate marginal likelihood. We show that our approach outperforms state-of-the-art non-parametric Bayesian and non-Bayesian methods.

1 Introduction

The Hawkes process (Hawkes, 1971) is particularly useful to model self-exciting point data – i.e., when the occurrence of a point increases the likelihood of occurrence of new points. The process is parametrised using a background intensity \( \mu \), and a triggering kernel \( \phi \). The Hawkes process can be alternatively represented as a cluster of Poisson processes (Hawkes and Oakes, 1974), where immigrant points are generated following a Poisson process with an intensity \( \mu \) (denoted as \( \text{PP}(\mu) \)), and every existing point triggers new points following a Poisson process \( \text{PP}(\phi) \). Points can therefore be structured into clusters – also known as the branching structure –, where each cluster contains a point and its direct offsprings. The triggering kernel \( \phi \) is shared among all cluster Poisson processes relating to a Hawkes process, and it determines the overall behavior of the process. Consequently, designing the kernel functions is of utmost importance for employing the Hawkes process to a new application, and its study has attracted much attention.

Related Work. In the case in which the optimal triggering kernel for a particular application is unknown, a typical solution is to express it using a non-parametric form such as in the work of Lewis and Mohler (2011); Zhou et al. (2013); Bacry and Muzy (2014); Eichler et al. (2017). These are all frequentist methods which do not quantify the uncertainty of learned triggering kernels. The Bayesian inference for Hawkes processes has also been studied including (Rasmussen, 2013; Linderman and Adams, 2015) which require either constructing a parametric triggering kernels or discretizing the input domain to scale with the data size. The latter approach leads to poor scaling with the dimension of the domain. Donnet et al. (2018) overcome these shortcomings by proposing a continuous non-parametric Bayesian Hawkes process, however they resort to an unscalable Markov chain Monte Carlo (MCMC) estimator to the posterior distribution. More recently, a solution based on Gibbs sampling (Zhang et al., 2019) obtains linear time complexity by exploiting the Hawkes processes’ branching structure and the stationarity of the triggering kernel, but this approach may suffer from slow convergence.
Contributions. In this paper, we address the shortcoming of the previous solutions and we make several contributions as follows. (1) We introduce a new Bayesian non-parametric Hawkes process in which the triggering kernel is modulated by a sparse Gaussian process [Lloyd et al. 2015], and the background intensity is Gamma distributed. We further introduce a variational inference scheme for such a model which allows fast convergence. We denote this approach as Variational Bayesian Hawkes Processes (VBHP). (2) We obtain linear time complexity by utilizing the stationarity of the triggering kernel. To our best knowledge, this is the first sparse Gaussian process modulated Hawkes process which employs the variational inference, which has a linear time complexity and which scales to large real world data. We additionally contribute a tighter lower bound than the evidence lower bound to approximate the marginal likelihood. This tighter lower bound also holds for other graphical models, such as the variational Gaussian mixture models [Attias 1999]. (3) We further show that VBHP provides more accurate prediction than state-of-the-art methods on synthetic data and on two large scale online diffusion datasets. We also validate the efficiency of VBHP, and we demonstrate that the approximate marginal likelihood is effective for performing model selection.

2 Prerequisite

In this section, we review the Hawkes process, the variational inference and the variational Bayesian Poisson process [Lloyd et al. 2015].

2.1 Hawkes Process (HP)

The Hawkes process [Hawkes 1971] is a self-exciting point process, in which the occurrence of a point increases the arrival rate \( \lambda(i) \), i.e. the (conditional) intensity, of new points. Given a set of points \( \{x_i\}_{i=1}^N \), the intensity at \( x \) conditioned on given points is written as:

\[
\lambda(x) = \mu + \sum_{x_i < x} \phi(x - x_i),
\]

where \( \mu > 0 \), considered as a constant, is the background intensity, and \( \phi: \mathcal{T} \to [0, \infty) \) is the triggering kernel defined over the data domain \( \mathcal{T} \). The log-likelihood of \( \{x_i\}_{i=1}^N \) for \( \mu \) and \( \phi \) is given by [Rubin 1972]:

\[
\log p(\{x_i\}_{i=1}^N|\mu, \phi) = \sum_{i=1}^N \log \lambda(x_i) - \int_\mathcal{T} \lambda(x) \, dx.
\]

2.2 Variational Inference (VI)

Consider the latent variable model \( p(x, z|\theta) \) where \( x \) and \( z \) are the observed and the latent variables. The variational approach introduces a variational distribution to approximate the posterior distribution \( q(z|x, \theta') \approx p(z|x, \theta) \) and maximizes a lower bound of the log-likelihood, which can be derived from the non-negative gap perspective:

\[
\log p(x|\theta) = \log \frac{p(x, z|\theta)}{q(z|x, \theta')} - \log \frac{p(z|x, \theta)}{q(z|x, \theta')}
\]

\[= \mathbb{E}_{q(z|x, \theta')} \left[ \log p(x|z, \theta) \right] - \text{KL}(q(z|x, \theta')||p(z|\theta)) + \text{KL}(q(z|x, \theta')||p(z|x, \theta))
\]

\[\geq \text{ELBO}(q(z|x, \theta'), p(x|z, \theta), p(z|\theta)).
\]

Optimising the evidence lower bound (ELBO) w.r.t. \( \theta' \) balances between the reconstruction error and the Kullback-Leibler (KL) divergence from the prior. Generally, the conditional \( p(z|x, \theta) \) is known, so is the prior. Thus, for an appropriate choice of \( q \), it is easier to work with this lower bound than the intractable posterior \( p(z|x, \theta) \). We also see that, due to the form of the intractable gap, if \( q \) is a family of distributions which contains elements close to the true unknown posterior, then the variational lower bound will be close to the true likelihood, and \( q \) will be close to the true posterior. An alternative derivation applies Jensen’s inequality:

\[
\log p(x|\theta) = \log \int p(x, z|\theta) \frac{q(z|x, \theta')}{q(z|x, \theta')} \, dz \geq \mathbb{E}_{q(z|x, \theta')} \left[ \log p(x|z, \theta) \right] - \text{KL}(q(z|x, \theta')||p(z|\theta))
\]

\[\equiv \text{ELBO}(q(z|x, \theta'), p(x|z, \theta), p(z|\theta)).
\]
where to make the notation concise, we omit
\[ D \equiv \{ \mathbf{x}_n \}_{n=1}^{N} \]
\[ B, \mu, f, u \]
\[ k_0, c_0, \alpha_1, \alpha_2, \cdots, \alpha_R, \gamma \]
\[ k, c, \mathbf{m}, \mathbf{S}, \alpha_1, \alpha_2, \cdots, \alpha_R, \gamma \]
\[ \mu, \mathbf{S}, \alpha_1, \alpha_2, \cdots, \alpha_R, \gamma \]
\[ \theta \]
\[ \theta' \]

| Table 1: Notations |
|-------------------|
| \( \mathcal{D} \equiv \{ \mathbf{x}_n \}_{n=1}^{N} \) | \( \mathcal{D} \equiv \{ \mathbf{x}_n \}_{n=1}^{N} \) | \( x \) |
| \( B, \mu, f, u \) | \( f, \mu, u \) | \( z \) |
| \( k_0, c_0, \alpha_1, \alpha_2, \cdots, \alpha_R, \gamma \) | \( \alpha_1, \alpha_2, \cdots, \alpha_R, \gamma \) | \( \theta \) |
| \( k, c, \mathbf{m}, \mathbf{S}, \alpha_1, \alpha_2, \cdots, \alpha_R, \gamma \) | \( \mathbf{m}, \mathbf{S}, \alpha_1, \alpha_2, \cdots, \alpha_R, \gamma \) | \( \theta' \) |

\[ \begin{align*}
2.3 \text{ Variational Bayesian Poisson Process (VBPP)}
\end{align*} \]

VBPP \cite{Lloyd et al. 2015} applies the VI to the Bayesian Poisson process, which exploits the sparse Gaussian process (GP) to model the Poisson intensity. Specifically, the sparse GP employs the ARD kernel
\[ K(x, x') \equiv \gamma \prod_{r=1}^{R} \exp \left( -\frac{(x_r - x'_r)^2}{2\alpha_r} \right). \]
(4)

where \( \gamma \) and \( \alpha_r \) are GP hyper-parameters. Let \( u \equiv (f(\mathbf{z}_1), f(\mathbf{z}_2), \cdots, f(\mathbf{z}_M)) \) where \( \mathbf{z}_i \) are inducing points and \( f \) is from the sparse GP. The prior and the approximate posterior distributions of \( u \) are Gaussians \( \mathcal{N}(u|\mathbf{u}, \mathbf{K}_{zz'}) \) and \( q(u|\mathcal{D}, \theta') = \mathcal{N}(u|m, S) \approx p(u|\mathcal{D}, \theta) \), where \( \mathcal{D} \), \( m \) and \( S \) are the dataset, the mean vector and the covariance matrix respectively. Note, both \( u \) and \( f \) employ zero mean priors. Notations of VBPP are connected with those of Section 2.2 in Table 1.

Importantly, the approximate joint distribution of \( f \) and \( u \) uses the exact conditional distribution \( p(f|u, \theta') \), i.e.,
\[ q(f, u|\mathcal{D}, \theta') \equiv p(f|u, \theta')q(u|\mathcal{D}, \theta') \]
(5)

which in turn leads to the posterior GP
\[ q(f|\mathcal{D}, \theta') = \mathcal{N}(f|K_{xx'}K_{zz'}^{-1}m, K_{xx'} - K_{xz}K_{zz'}^{-1}K_{xz'} + K_{xz}K_{zz'}^{-1}S_{zz}K_{zz'}^{-1}K_{xz'}). \]
(6)

Then, the ELBO can be simplified as below
\[ \text{ELBO}(q(f, u), p(\mathcal{D}|f, u), p(f, u)) = \mathbb{E}_{q(f)}[\log p(\mathcal{D}|f)] - \text{KL}(q(u)||p(u)), \]
(7)

where to make the notation concise, we omit \( \mathcal{D} \) and \( \theta' \) in conditions of variational distributions and \( \theta \) in conditions of priors and likelihoods. The first term on r.h.s. is simplified due to conditional independence between \( \mathcal{D} \) and \( u \), while the second term due to the exact conditional \( p(f|u, \theta') \). Notably, the second term is the KL divergence between two multivariate Gaussians, so is available in closed form. The first term can be easily handled for a squared link function mapping \( f \) to the Poisson intensity function \( \lambda(x) = f^2(x) \) and it turns out to be the expectation w.r.t. \( q(f|\mathcal{D}, \theta') \) of the log-likelihood \( \log p(\mathcal{D}|f) = \sum_{n=1}^{N} \log f^2(x_n) - \int_{\mathcal{T}} f^2(x) \ dx \). The expectation of the integral part is relatively straightforward to compute and the expectation of the other (data-dependent) part is available in almost closed-form with a hyper-geometric function making things a little tricky to implement.

3 Variational Bayesian Hawkes Process (VBHP)

Notations To extend VBPP to HP, we introduce a latent branching structure \( B \) and a latent background intensity \( \mu \). The same priors are adopted for \( f \) and \( u \), namely, \( u \sim \mathcal{N}(0, \mathbf{K}_{zz'}) \) and \( f \sim \mathcal{N}(0, \mathbf{K}_{xx'}) \), and the same approximate joint posterior on \( f \) and \( u \) as [Eq. (5)]. We assume that the prior distribution of \( \mu \) is a Gamma distribution, i.e., \( \mu \sim \text{Gamma}(k_0, c_0) \)\(^1\) and that the posterior distribution is approximated by another Gamma distribution, \( q(\mu|\mathcal{D}, \theta') = \text{Gamma}(\mu|\mathbf{k}, \mathbf{c}) \approx p(\mu|\mathcal{D}, \theta') \). Notations of VBHP are summarized in Table 1.

\(^1\)\text{Gamma}(\mu|k_0, c_0) = \frac{1}{\Gamma(k_0)c_0^k_0} \mu^{k_0-1} e^{-\mu/c_0}
We further employ the approximate posterior distribution of the branching structure, $q(B|\mathcal{D})$, to complete our variational distribution, so that

$$q(B, \mu, f, u|\mathcal{D}, \theta') \equiv q(B|\mathcal{D}, \theta')q(\mu|\mathcal{D}, \theta')p(f|u, \theta')q(u|\mathcal{D}, \theta').$$

(8)

Based on this, the VBHP ELBO can be available in closed form.

**ELBO** First, we concisely derive the VBHP ELBO as below (see details in Appendix A.1 (App. 2019)):

$$\text{ELBO}(q(B, \mu, f, u), p(D|B, \mu, f, u), p(B, \mu, f, u))$$

$$= \mathbb{E}_{q(B, \mu, f)}[\log p(D|B, \mu, f)] - \text{KL}(q(B)||p(B)) - \text{KL}(q(\mu)||p(\mu)) - \text{KL}(q(u)||p(u))$$

$$= \mathbb{E}_{q(B, \mu, f)}\left[ \log \frac{p(D, B|f, \mu)}{q(B)} \right] - \text{KL}(q(\mu)||p(\mu)) - \text{KL}(q(u)||p(u)),$$

(9)

where the second line is obtained by simplifications analogous to those of Eq. (7) and the third line is due to merging the reconstruction term with the KL term of $B$. The KL terms w.r.t. $\mu$ and $u$ are KL divergences between gamma distributions and Gaussian distributions respectively, which have closed-form expressions:

$$\text{KL}(q(\mu)||p(\mu)) = (k-k_0)\psi(k) - k_0 \log c - k - \log \frac{\Gamma(k)}{\Gamma(k_0)} + \frac{ek}{c},$$

(10)

$$\text{KL}(q(u)||p(u)) = \frac{1}{2} \left[ \text{Tr}(K_{zz'}^{-1}S) + \log \frac{|K_{zz'}^{-1}|}{|S|} - M + m^T K_{zz'}^{-1} m \right],$$

(11)

where $\Gamma$ and $\psi$ are Gamma and Digamma functions. Now, we are left with the problem of computing the data-dependent expectation of Eq. (9)

**Data Dependent Expectation (DDE)** The derivation of DDE has two major steps. First, we show the expectation of the log-likelihood $\log p(D, B|f, \mu)$ w.r.t. the variational distribution of the branching structure $q(B)$, which is given as below (see details in Appendix A.2 (App. 2019)):

$$\mathbb{E}_{q(B)}[\log p(D, B|f, \mu)] = \sum_{i=1}^{N} \left( q_{i,0} \log \mu + \sum_{j=1}^{i-1} q_{i,j} \log f_{i,j}^2 - \int_{\mathcal{G}} f_{i,j}^2 \, \text{d}x \right) - \mu|\mathcal{T}|,$$

(12)

where $q_{i,j}$ is the approximate posterior distribution of $x_j$ triggering $x_i$ ($x_0$ represents the background) and for notational simplification, we introduce $f_{i,j} \equiv f(x_i - x_j)$ and $f_{i} \equiv f(x - x_i)$. Interestingly, the branching structure decomposes the Hawkes log-likelihood into weighted Poisson log-likelihoods which allows existing Bayesian Poisson process models to be applied to Hawkes processes.

Second, due to the independence among variational $B$, $\mu$ and $f$, plugging Eq. (12) into the DDE term of Eq. (9) results in separate expectations over $B$, $\mu$ and $f$:

$$\text{DDE} = \sum_{i=1}^{N} \left( q_{i,0}(\psi(k) + \log c) + \sum_{j=1}^{i-1} q_{i,j} \mathbb{E}_{q(f)}(\log f_{i,j}^2) - \int_{\mathcal{G}} \mathbb{E}_{q(f)}(f_{i,j}^2) \, \text{d}x \right) + H_B - kc|\mathcal{T}|,$$

(13)

where $H_B = - \sum_{i=1}^{N} \sum_{j=0}^{i-1} p_{ij} \log p_{ij}$ is the entropy of the branching structure variable and $q(f)$ is identical to the counterpart of VBPP (as Eq. (6)). Maximizing the VBHP ELBO makes a balance among three parts: the expected log joint likelihood of point data and the branching structure, the diversity of the branching structure and KL divergences over the background intensity and over the inducing points.

As in VBPP, the term $\mathbb{E}_{q(f)}(\log f_{i,j}^2)$ is available in the closed form with a hyper-geometric function:

$$\mathbb{E}_{q(f)}(\log f_{i,j}^2) = - \tilde{G}\left( - \frac{\nu_{ij}}{2\Sigma_{ij}} \right) + \log \left( \frac{\Sigma_{ij}}{2} \right) - C,$$

(14)

where $\nu_{ij} = \nu(x_i - x_j)$, $\Sigma_{ij} = \Sigma(x_i - x_j, x_i - x_j)$ ($\nu$ and $\Sigma$ are defined in Eq. (6)). $C \approx 0.57721566$ is the Euler-Mascheroni constant and $\tilde{G}$ is defined as $\tilde{G}(z) = {}_1 F_1^{(1,0,0)}(0, 1/2, z)$, i.e., the partial
derivative of the confluent hyper-geometric function \( F \) w.r.t. the first argument. The computation of \( \bar{G} \) can use the method of [Acaranii and Gasaneo (2008)]. We implement \( G \) and \( \bar{G} \) by linear interpolation of a lookup table (see the demo from supplementary material).

For the integral \( \int_{\mathcal{T}} \mathbb{E}_{q(f)}(f^2_i) \, dx \), [Lloyd et al. (2015)] give analytical expressions and we adapt them here as:

\[
\int_{\mathcal{T}} \mathbb{E}_{q(f)}(f^2_i) \, dx = \int_{\mathcal{T}} \mathbb{E}_{q(f)}[f(x)] \, dx + \int_{\mathcal{T}} \text{Var}_{q(f)}[f(x)] \, dx,
\]

where \( \mathcal{T}_i \equiv \times_{r=1}^{\mathcal{R}} [T_r^{\text{min}}, T_r^{\text{max}} - x_{i,r}^\ast] \) is the sampling region for \( x_i \) truncated by \( \mathcal{T} \) and each integral term has the closed form (see derivations in Appendices A.3 to A.5 (App., 2019)):

\[
\int_{\mathcal{T}_i} \mathbb{E}_{q(f)}(f(x)) \, dx = m^T K^{-1} \Psi_t K^{-1} m,
\]

\[
\int_{\mathcal{T}_i} \text{Var}_{q(f)}(f(x)) \, dx = \gamma |\mathcal{T}_i| - \text{Tr}(K^{-1} \Psi_t) + \text{Tr}(K^{-1} S K^{-1} \Psi_t),
\]

\[
\Psi_t(z, z') = \gamma^2 \prod_{r=1}^{\mathcal{R}} -\frac{\pi \alpha_r}{2} \exp \left( -\frac{(z_r - z_r')^2}{4 \alpha_r} \right) \left[ \text{erf} \left( \frac{z_r - T_r^{\text{max}} + x_{i,r}}{\sqrt{\alpha_r}} \right) - \text{erf} \left( \frac{z_r - T_r^{\text{min}}}{\sqrt{\alpha_r}} \right) \right],
\]

where \( z_r = (z_r + z_r')/2 \).

**Optimisation** To optimise the model parameters, we employ the EM algorithm. Specifically, \( q_{ij} \)'s are optimised in the E step and \( m, S, k \) and \( c \) are updated to increase ELBO in the M step. The updating equations for \( q_{ij} \) can be derived by the Lagrange multiplier method, which are directly given as below:

\[
q_{ij} = \begin{cases} 
\frac{\exp(\mathbb{E}_{q(f)}(\log f_{ij}^2))}{\theta \exp(\psi(k)) + \sum_{j=1}^{r-1} \exp(\mathbb{E}_{q(f)}(\log f_{ij}^2))}, & j > 0; \\
\frac{\theta \exp(\psi(k))}{\theta \exp(\psi(k)) + \sum_{j=1}^{r-1} \exp(\mathbb{E}_{q(f)}(\log f_{ij}^2))}, & j = 0.
\end{cases}
\]  

Similarly to VBPP, we fix inducing points on a regular grid over \( \mathcal{T} \), which reduces the function variance at locations far from data. The selection of GP kernel hyper-parameters is based on maximizing the marginal likelihood which is approximated by a tighter lower bound given in the following paragraph. Despite the observation that more inducing points lead to better fitting accuracy (Lloyd et al. 2015 [Nelson and Ghahramani, 2006]), in the case of our more complex VBHP, more inducing points may cause slow convergence (Fig. 4a (App., 2019)) under some hyper-parameters and therefore poor performance in limited iterations. Besides, more inducing points need more fit time, so the suitable number of inducing points makes a balance between accuracy and the time cost.

**Marginal Likelihood** A common approximation to the log marginal likelihood is the ELBO. For our VBHP, we derive a tighter approximation to the marginal likelihood (see the derivation in Appendix B (App., 2019)):

\[
\log p(D | \theta) \geq \mathbb{E}_{q(B, f, \mu, \nu | D, \theta')} \left[ \log p(D | \theta, B, f, \mu) \right],
\]  

which actually is the data-dependent expectation of the ELBO. This approximation is tighter because it doesn’t subtract non-negative KL divergences over the background intensity and over the inducing points. However, it doesn’t guarantee minimizing the intractable KL divergences between approximate and true posterior background intensity and inducing points, that is, the intractable non-negative gap in Eq. (3). This is an important reason why tighter variational bounds are not necessarily better (Rainforth et al., 2018) and we therefore don’t use it in place of the ELBO as the objective function. Some of other graphical models, such as the variational Gaussian mixture model (Attias, 1999), have the similar tighter lower bounds.
Time Complexity and Acceleration  In the E step of model optimisation, updating $q_{ij}$’s requires computing the mean and the variance of all $f_{ij}$’s, which both take $O(M^3 + M^2N^2)$. Here we omit the data dimension $R$ since normally $M > R$ for regular-grid-located inducing points. Similarly in the M step, computing the hyper-geometric term needs the mean and the variance of all $f_{ij}$’s. Further computation of the integral terms takes $O(M^3N)$. Thus, the total time complexity per iteration is $O(M^3N + M^2N^2)$.

To accelerate VBHP, we utilise the stationarity of the triggering kernel as (Zhang et al., 2019), which says for input larger than some point, the kernel has negligible values. This is rational as a HP with finite expected size has the integral of the triggering kernel over $[0, \infty)$ less than 1. Our work scales the domain to some $T$ without breaking the stationarity. Specifically, we assume a region $S$ as the support of the triggering kernel and there is $\tilde{f}(x_i - x_j) = 0$ for any $x_i - x_j \notin S$, which reduces possible parents of a point from the whole history to local points in the support. As a result, the total time complexity can be reduced to $O(M^3N)$.

Predictive Distribution of the Triggering Kernel  The predictive distribution of $f(x)$ depends on the posterior $u$ and we assume that the optimal variational distribution of $u$ approximates the true posterior distribution, namely $q(u|\mathcal{D}, \theta^*) = \mathcal{N}(u|m^*, S^*)$ $\approx p(u|\mathcal{D}, \theta)$. Therefore, there is $q(f|\mathcal{D}, \theta^*) \approx p(f|\mathcal{D}, \theta)$, i.e., the approximate predictive $f(\bar{x}) \sim \mathcal{N}(K_{\bar{x}x}K_{xx}^{-1}m^*, K_{\bar{x}\bar{x}} - K_{\bar{x}x}K_{xx}^{-1}K_{x\bar{x}} + K_{x\bar{x}}K_{xx}^{-1}S^*S^{-1}K_{x\bar{x}})$ $\equiv \mathcal{N}(\tilde{\nu}, \tilde{\sigma}^2)$. Given the relation $\phi(\cdot) = f^2(\cdot)$, it is straightforward to derive the corresponding $\phi(\tilde{z}) \sim \text{Gamma}(\tilde{k}, \tilde{c})$ where the shape $k = (\tilde{\nu}^2 + \tilde{\sigma}^2)^2/[2\tilde{\sigma}^2(2\tilde{\nu}^2 + \tilde{\sigma}^2)]$ and the scale $c = 2\tilde{\sigma}^2(2\tilde{\nu}^2 + \tilde{\sigma}^2)/(\tilde{\nu}^2 + \tilde{\sigma}^2)$.

Extension to the Multivariate Hawkes Process  The above derivations are based on the univariate Hawkes process, which is easy to understand compared with the counterpart for the multivariate Hawkes process, and the latter can be obtained following the same derivations.
4 Experiments

Evaluation We employ two metrics: the $L_2$ distance (for ground truth known only) to measure the difference between predictive and true models, formulated as $L_2(\phi_{\text{pred}}, \phi_{\text{true}}) = \left( \int_{T} (\phi_{\text{pred}}(x) - \phi_{\text{true}}(x))^2 \, dx \right)^{0.5}$ and $L_2(\mu_{\text{pred}}, \mu_{\text{true}}) = |\mu_{\text{pred}} - \mu_{\text{true}}|$; the held-out log likelihood (HLL, shown as Eq. (2)) to describe how well the predictive model fits data. More specifically for HLL, points in each original datum are randomly assigned to either training or test data with equal probabilities and evaluations over a number of test data are averaged as summarisation. To reduce the impact of the sequence length, we divide the HLL by the number of points.

Prediction For two metrics employed above, we use the pointwise mode of the posterior triggering kernel as the prediction because it is computationally intractable to find the posterior mode at multiple point locations (Zhang et al., 2019). Besides, we exploit the mode of the posterior background intensity as the predictive background intensity.

Baselines We use the following models as baselines. (1) A parametric Hawkes process equipped with the sum of exponential (SumExp) triggering kernel $\phi(t) = \sum_{i=1}^{K} a_1^i a_2^i \exp(-a_2^i t)$ and the constant background intensity. (2) The ordinary differential equation (ODE) based non-parametric non-bayesian Hawkes process (Zhou et al., 2013). The code is publicly available (Bacry et al., 2017). (3) Wiener-Hopf (WH) equation based non-parametric non-bayesian Hawkes process (Zhou et al., 2013). The code is publicly available (Bacry et al., 2017). (4) The gibbs sampling based Bayesian non-parametric Hawkes process (Gibbs Hawkes) (Zhang et al., 2019). For fairness, the ARD kernel is exploited and corresponding eigenfunctions are approximated by Nyström method (Williams and Rasmussen, 2006), where regular grid points are used as VBHP. Different from Zhang et al., 2019, Gibbs Hawkes is trained on single sequences.

4.1 On Synthetic Data

Our synthetic data are generated from three Hawkes processes over $T = [0, \pi]$, whose triggering kernels are Sine (sin, Eq. (21)), Cosine (cos, Eq. (22)) and exponential (exp, Eq. (23)) functions respectively, and whose background intensities are the same, $\mu = 10$.

\[
\phi_{\text{sin}}(t) = 0.9 \sin(3t) + 0.9, \quad t \in [0, \pi/2]; \quad \text{otherwise}, \quad 0; \quad (21)
\]
\[
\phi_{\text{cos}}(t) = \cos(2t) + 1, \quad t \in [0, \pi/2]; \quad \text{otherwise}, \quad 0; \quad (22)
\]
\[
\phi_{\text{exp}}(t) = 5 \exp(-5t), \quad t \in [0, \infty). \quad (23)
\]

Model Selection As the marginal likelihood $p(D|\theta)$ is a key advantage of our method over non-Bayesian approaches (Zhou et al., 2013; Bacry and Muzy, 2014), we investigate its efficacy for model selection. Fig. 1b shows the contour plot of the approximate log marginal likelihood of a single sample. The observation is that the contour plot of the approximate marginal likelihood has a similar shape to that of the contour plots of $L_2(\phi)$ (Fig. 1c) and of HLL (Fig. 1d) — GP hyper-parameters with relatively high marginal likelihoods have relatively low $L_2$ errors and relatively high HLL. Fig. 1a plots the posterior triggering kernel corresponding to the maximal approximate marginal likelihood. This demonstrates the practical utility of both the marginal likelihood itself and our approximation to it. Gibbs Hawkes has a similar observation as Fig. 5 (App., 2019).

Evaluation To evaluate VBHP on synthetic data, 20 sequences are drawn from each model and 100 pairs of train and test sequences drawn from each sample to compute the HLL. We select GP hyper-parameters of Gibbs Hawkes and of VBHP by maximizing the approximate marginal likelihood. Table 2 shows evaluations for baselines and VBHP (using 10 inducing points for trade-off between accuracy and time, so does Gibbs Hawkes) in both $L_2$ and HLL. VBHP achieves the best performance but is two orders of magnitudes slower than Gibbs Hawkes (shown as Figs. 2a and 2b). Interestingly, SumExp fits $\phi_{\text{exp}}$ best by using a single exponential function in $L_2$ distance while due to learning on single sequences, the background intensity has relatively high errors.

4.2 Real World Data

We exploit two large scale tweet datasets to further conclude. ACTIVE (Rizoiu et al., 2018) is a tweet dataset which was collected in 2014 and contains around 41k (re)tweet temporal point processes with
links to Youtube videos. Each sequence contains at least 20 (re)tweets. SEISMIC (Zhao et al., 2015) is large scale tweet dataset which was collected from October 7 to November 7, 2011 and contains around 166k (re)tweet temporal point processes. Each sequence contains at least 50 (re)tweets.

Evaluation Similarly to synthetic experiments, we evaluate the fitting performance by averaging HLL of 20 test sequences randomly drawn from each original datum. We scale all original data to [0, π) and use 10 inducing points for VBHP and Gibbs Hawkes in terms of time and accuracy. The model selection is still by maximizing the approximate marginal likelihood. The results are shown in Table 2 and again we observe similar predictive performance of VBHP which performs best. This additionally demonstrates the marginal likelihood and our approximation are useful in practice.

Fit Time We further evaluate the fitting speed\footnote{The CPU we use is Intel(R) Core(TM) i7-7700 CPU @ 3.60GHz and the language is Python 3.6.5.} of VBHP and Gibbs Hawkes on sample synthetic and real world point processes, summarized in Figs. 2a and 2b. The fit time is averaged over iterations and it is observed that the increasing trends with the number of inducing points and with the data size are similar between Gibbs Hawkes and VBHP. Although VBHP is significantly slower than Gibbs Hawkes, VBHP converges in 10~20 iterations [Fig. 4 (App., 2019)], giving an average convergence time of 549 seconds for a sequence of 1000 events, compared to 699 seconds for Gibbs Hawkes. The slope of VBHP in Fig. 2b (log-scale) is 1.04 and the correlation coefficient is 0.96, so the fit time is linear to the data size.

5 Conclusions

In this paper, we provided the VI scheme for a new Bayesian non-parametric Hawkes process whose triggering kernel is modulated by a sparse Gaussian process and background intensity is Gamma distributed. To address the difficulty with scaling with the data size, we utilize the stationarity of the triggering kernel to reduce the number of possible parents of each point. We additionally contribute a tighter lower bound of the marginal likelihood which is empirically shown to be effective for model selection. On synthetic data and two large Twitter diffusion datasets, VBHP enjoys linear fit time with the data size and fast convergence rate, and provides more accurate prediction than those of state-of-the-art approaches.

Table 2: Results on Synthetic and Real World Data (Mean ± One Standard Variance)

| Measure | Data  | SumExp | ODE   | WH    | Gibbs Hawkes | VBHP  |
|---------|-------|--------|-------|-------|--------------|-------|
| L₂      | Sin   | \(\phi:0.693\pm0.028\) | 0.665±0.121 | 2.463±0.145 | 0.408±0.198 | 0.183±0.897 |
|         | \(\mu:2.968\pm1.640\) | 4.514±3.808 | 6.794±5.054 | 4.108±3.949 | 0.579±0.523 |
|         | Cos   | \(\phi:0.473\pm0.102\) | 0.697±0.056 | 1.743±0.083 | 0.666±0.586 | 0.292±0.196 |
|         | \(\mu:2.751\pm1.902\) | 7.030±5.662 | 6.099±4.613 | 4.685±4.421 | 0.515±0.293 |
|         | Exp   | \(\phi:0.133\pm0.138\) | 1.835±0.539 | 2.254±2.042 | 0.676±0.233 | 0.235±0.102 |
|         | \(\mu:3.290\pm1.991\) | 8.969±8.604 | 16.66±20.95 | 7.648±9.647 | 0.486±0.418 |
| HLL     | Sin   | 3.490±0.400 | 3.489±0.413 | 3.233±0.273 | 3.492±0.406 | 3.497±0.406 |
|         | Cos   | 3.874±0.544 | 3.872±0.552 | 3.613±0.373 | 3.871±0.562 | 3.878±0.548 |
|         | Exp   | 2.825±0.481 | 2.822±0.496 | 2.782±0.490 | 2.820±0.492 | 2.829±0.487 |
|         | ACTIVE| 1.692±1.371 | 0.880±2.716 | 0.710±0.943 | 1.323±2.160 | 1.867±1.181 |
|         | SEISMIC| 2.943±0.959 | 2.582±1.665 | 1.489±1.796 | 3.110±1.251 | 3.164±0.843 |
References

Ancarani, L. U. and Gasaneo, G. (2008). Derivatives of any order of the confluent hypergeometric function $f_1 1 (a, b, z)$ with respect to the parameter a or b. Journal of Mathematical Physics, 49(6):063508.

App. (2019). Appendix: Variational inference for sparse gaussian process modulated hawkes process. In the supplementary material.

Attias, H. (1999). Inferring parameters and structure of latent variable models by variational bayes. In Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence, pages 21–30. Morgan Kaufmann Publishers Inc.

Bacry, E., Bompaire, M., Găifias, S., and Poulsen, S. (2017). tick: a Python library for statistical learning, with a particular emphasis on time-dependent modeling. ArXiv e-prints.

Bacry, E. and Muzy, J.-F. (2014). Second order statistics characterization of hawkes processes and non-parametric estimation. arXiv preprint arXiv:1401.0903.

Donnet, S., Rivoirard, V., and Rousseau, J. (2018). Nonparametric bayesian estimation of multivariate hawkes processes. arXiv preprint arXiv:1802.05975.

Eichler, M., Dahlhaus, R., and Dueck, J. (2017). Graphical modeling for multivariate hawkes processes with nonparametric link functions. arXiv preprint arXiv:1401.0903.

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

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

Lewis, E. and Mohler, G. (2011). A nonparametric em algorithm for multiscale hawkes processes. Journal of Nonparametric Statistics, 1(1):1–20.

Linderman, S. W. and Adams, R. P. (2015). Scalable bayesian inference for excitatory point process networks. arXiv preprint arXiv:1507.03228.

Lloyd, C., Gunter, T., Osborne, M., and Roberts, S. (2015). Variational inference for gaussian process modulated poisson processes. In ICML, pages 1814–1822.

Rainforth, T., Kosiorek, A. R., Le, T. A., Maddison, C. J., Igl, M., Wood, F., and Teh, Y. W. (2018). Tighter variational bounds are not necessarily better. arXiv preprint arXiv:1802.04537.

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

Rizoiu, M.-A., Mishra, S., Kong, Q., Carman, M., and Xie, L. (2018). Sir-hawkes: Linking epidemic models and hawkes processes to model diffusions in finite populations. In Proceedings of the 27th International Conference on World Wide Web, pages 419–428. International World Wide Web Conferences Steering Committee.

Rubin, I. (1972). Regular point processes and their detection. IEEE Transactions on Information Theory, 18(5):547–557.

Snelson, E. and Ghahramani, Z. (2006). Sparse gaussian processes using pseudo-inputs. In Advances in neural information processing systems, pages 1257–1264.

Williams, C. K. and Rasmussen, C. E. (2006). Gaussian processes for machine learning, volume 2. MIT Press Cambridge, MA.

Zhang, R., Walder, C., Rizoiu, M.-A., and Xie, L. (2019). Efficient Non-parametric Bayesian Hawkes Processes. In International Joint Conference on Artificial Intelligence (IJCAI’19), Macao, China.
Zhao, Q., Erdogdu, M. A., He, H. Y., Rajaraman, A., and Leskovec, J. (2015). Seismic: A self-exciting point process model for predicting tweet popularity. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1513–1522. ACM.

Zhou, K., Zha, H., and Song, L. (2013). Learning triggering kernels for multi-dimensional hawkes processes. In International Conference on Machine Learning, pages 1301–1309.
As per Eq. (3), there is

Accompanying the submission Variational Inference for Sparse Gaussian Process Modulated Hawkes Process.

A Omitted Derivations

A.1 Deriving Eq. (9)

As per Eq. (3) there is

\[ \text{ELBO}(q(B, \mu, f, u), p(D|B, \mu, f, u), p(B, \mu, f, u)) = \mathbb{E}_{q(B, \mu, f)}[\log p(D|B, \mu, f)] - KL(q(B, \mu, f,u)||p(B, \mu, f, u)) \]

where the KL term can be simplified as

\[ KL(q(B, \mu, f,u)||p(B, \mu, f, u)) \]

Thus, the ELBO is finally written as

\[ \text{ELBO} = \mathbb{E}_{q(B, \mu, f)}[\log \frac{p(D, B|\mu, f)}{q(B)}] - KL(q(\mu)||p(\mu)) - KL(q(u)||p(u)). \]

A.2 Deriving Eq. (12)

First, note that the branching structure $B$ has the tree structure and is a collection of triggering relationships between paired points. Each point $x_i$ has a parent $x_{pa_i}$ ($0 \leq pa_i \leq i - 1$, $x_0$ represents the background) and the triggering events are independent, so the probability of the branching structure is the product of the probabilities of triggering events, that is

\[ q(B) = \prod_{i=1}^{N} q_{i,pa_i}, \quad (24) \]
where \( q_{i,pa_i} \) is the approximate posterior distribution of the triggering event where \( x_{pa_i} \), triggering \( x_i \).

According to the cluster representation of the Hawkes process, the branching structure decomposes the process into a cluster of Poisson process. Thus, the log-likelihood \( \log p(D, B | f, \mu) \) turns out to be a sum of log Poisson likelihoods:

\[
\log p(D, B | f, \mu) = \sum_{i=1}^{N} \left( \log f_{i,pa_i}^2 - \int_{\mathcal{T}} f_i^2 \, dx \right) - \mu |T|,
\]

where for notational simplification, we introduce \( f_{i0}^2 \equiv \mu \), \( f_{i,pa_i} \equiv f(x_i - x_{pa}) \) for \( pa_i > 0 \) and \( f_i \equiv f(x - x_i) \). Furthermore, the expectation of the above log-likelihood w.r.t. \( q(B) \) is derived as below

\[
\mathbb{E}_{q(B)} \left[ \log p(D, B | \mu, f, \theta) \right] = \sum_{i=1}^{N} \left( \log f_{i,pa_i}^2 - \int_{\mathcal{T}} f_i^2 \, dx \right) - \mu |T| \quad \text{(Eq. (25))}
\]

\[
= \sum_{i=1}^{N} \left( \mathbb{E}_{q(B)} \left( \log f_{i,pa_i}^2 \right) - \int_{\mathcal{T}} f_i^2 \, dx \right) - \mu |T| \quad \text{(simplification)}
\]

\[
= \sum_{i=1}^{N} \left( \sum_{\{pa_i\}_{j=1}^{N}} \left( \prod_{j=1}^{N} q_{j,pa_j} \right) \log f_{i,pa_i}^2 - \int_{\mathcal{T}} f_i^2 \, dx \right) - \mu |T| \quad \text{(split summations)}
\]

\[
= \sum_{i=1}^{N} \left( \sum_{\{pa_i\}_{j=1}^{N}} \prod_{j=1, j \neq i}^{N} q_{j,pa_j} \log f_{i,pa_i}^2 - \int_{\mathcal{T}} f_i^2 \, dx \right) - \mu |T| \quad \text{(distributive law of multiplication)}
\]

\[
= \sum_{i=1}^{N} \left( \sum_{\{pa_i\}_{j=0}^{i-1}} q_{i,pa_i} \log f_{i,pa_i}^2 - \int_{\mathcal{T}} f_i^2 \, dx \right) - \mu |T|
\]

\[
= \sum_{i=1}^{N} \left( q_{i,0} \log \mu + \sum_{j=1}^{i-1} q_{i,j} \log f_{i,j}^2 - \int_{\mathcal{T}} f_i^2 \, dx \right) - \mu |T|
\]

A.3 Deriving \[\text{Eq. (16)}\]

\[
\int_{\mathcal{T}_i} \mathbb{E}_{q(f)}^2[f(x)] \, dx = \int_{\mathcal{T}_i} m^T K_{zz}^{-1} K_{xz} K_{xx}^{-1} m \, dx
\]

\[
= m^T K_{zz}^{-1} \int_{\mathcal{T}_i} K_{xz} K_{xx} \, dx K_{xx}^{-1} m
\]

\[
= m^T K_{zz}^{-1} \Psi_i K_{xx}^{-1} m.
\]

A.4 Deriving \[\text{Eq. (17)}\]

\[
\int_{\mathcal{T}_i} \text{Var}^2_{q(f)}[f(x)] \, dx
\]

\[
= \int_{\mathcal{T}_i} K_{xx} - K_{xx} K_{zz}^{-1} K_{xz} + K_{xx} K_{zz}^{-1} SK_{zz}^{-1} K_{xx} \, dx
\]

\[
= \int_{\mathcal{T}_i} \gamma - \text{Tr}(K_{zz}^{-1} K_{xz} K_{xx}) + \text{Tr}(K_{zz}^{-1} S K_{zz}^{-1} K_{xx} K_{xx}) \, dx
\]
\[= \gamma |T_i| - \text{Tr}(K_{zz'}^{-1} \int_{T_i} K_{xx} K_{z} \, dx) + \text{Tr}(K_{zz'}^{-1} S K_{zz'}^{-1} \int_{T_i} K_{xx} K_{z} \, dx)\]
\[= \gamma |T_i| - \text{Tr}(K_{zz'}^{-1} \Psi_i) + \text{Tr}(K_{zz'}^{-1} S K_{zz'}^{-1} \Psi_i).\]

A.5 Deriving Eq. (18)

\[\Psi_i(z, z') = \int_{T_i} K_{xx} K_{z} \, dx\]
\[= \int_{T_i} \gamma^2 R \prod_{r=1}^{R} \exp \left( -\frac{(x_r - z_r')^2}{2\alpha_r} \right) \exp \left( -\frac{(z_r' - x_r)^2}{2\alpha_r} \right) \, dx\]
\[= \int_{T_i} \gamma^2 R \prod_{r=1}^{R} \exp \left( -\frac{(z_r - z_r')^2}{4\alpha_r} \right) \exp \left( -\frac{(z_r' - x_r)^2}{\alpha_r} \right) \, dx, \quad (y_r = (\bar{z}_r - x_r)/\alpha_r)\]
\[= \gamma^2 \prod_{r=1}^{R} \sqrt{\frac{\alpha_r}{2}} \exp \left( -\frac{1}{4\alpha_r} \right) \int_{T_{i,r}} \exp \left( -\frac{(\bar{z}_r - \bar{z}_r')^2}{\alpha_r} \right) \, dy_r\]
\[= \gamma^2 \prod_{r=1}^{R} \exp \left( -\frac{(z_r - z_r')^2}{4\alpha_r} \right) \int_{T_{i,r}} \exp \left( -\frac{(\bar{z}_r - x_r)^2}{\alpha_r} \right) \, dy_r\]
\[= \gamma^2 \prod_{r=1}^{R} \exp \left( -\frac{(z_r - z_r')^2}{4\alpha_r} \right) \int_{T_{i,r}} \exp \left( -\frac{(\bar{z}_r - x_r)^2}{\alpha_r} \right) \, dy_r\]
\[= \frac{1}{2} \exp \left(-\frac{(\bar{z}_r - x_r)^2}{\alpha_r} \right) \left[ \text{erf} \left( \frac{\bar{z}_r - \bar{T}_{max}}{\sqrt{\alpha_r}} \right) - \text{erf} \left( \frac{\bar{z}_r - \bar{T}_{min}}{\sqrt{\alpha_r}} \right) \right].\]

B Deriving Eq. (20)

To make derivations concise, we omit \(D\) and \(\theta'\) from the conditions of variational distributions \(q(\cdot)\) and \(\theta\) from priors and likelihoods. Ste-by-step derivations are shown as below

\[\log p(D) = \log \sum_{B} p(D, B) \frac{q(B)}{q(B)} \quad \text{(Bayes’ rules)}\]
\[\geq \sum_{B} q(B) \log p(D, B) + H(B) \quad \text{(Jensen’s inequality)}\]
\[= \sum_{B} \lim_{N \to \infty} \frac{N_B}{N} \log p(D, B) + H(B) \quad \text{(definition of probability)}\]
\[= \lim_{N \to \infty} \frac{1}{N} \sum_{B} N_B \log p(D, B) + H(B)\]
\[= \lim_{N \to \infty} \frac{1}{N} \log p(\{D, B_i\}_{i=1}^{N}) + H(B) \quad \text{(Bayes’ rules)}\]
\[= \lim_{N \to \infty} \frac{1}{N} \log \int \int p(\{D, B_i\}_{i=1}^{N} | \mu, f) p(\mu) p(f|u) p(u) \frac{q(\mu) q(u)}{q(\mu) q(u)} \, df \, du + H(B)\]
\[\geq \lim_{N \to \infty} \frac{1}{N} \int \int q(\mu) q(u) p(f|u) \log \left[ p(\{D, B_i\}_{i=1}^{N} | \mu, f) \frac{p(\mu) p(u)}{q(\mu) q(u)} \right] \, df \, du \, d\mu + H(B)\]
\[= \lim_{N \to \infty} \frac{1}{N} \int \int q(\mu) q(u) p(f|u) \log p(\{D, B_i\}_{i=1}^{N} | \mu, f) \, df \, du \, d\mu\]
\[+ \lim_{N \to \infty} \frac{1}{N} \int q(u) \log \frac{p(u)}{q(u)} \, du + \lim_{N \to \infty} \frac{1}{N} \int q(\mu) \log \frac{p(\mu)}{q(\mu)} \, d\mu + H(B)\]
\[= 0 + 0 + H(B)\]
\[
= \lim_{N \to \infty} \sum_{B} \frac{N_B}{N} \int \int q(\mu) p(f|u) p(u) \log p(D, B|\mu, f) \ df \ d\mu \ du + H(B)
\]
\[(\text{Simplification})\]
\[
= \mathbb{E}_{q(B, f, \mu)} \left[ \log \frac{p(D, B|f, \mu)}{q(B)} \right]
\]

(split the log likelihood)

(26)

C Additional Experiment Results

Figure 3: Posterior Triggering Kernels Inferred By VBHP and Gibbs Hawkes.

Figure 4: Convergence Rate of VBHP and Gibbs Hawkes with Different Numbers of Inducing Points. VBHP and Gibbs Hawkes measure respectively the relative error of the approximate marginal likelihood and of the posterior distribution of the Gaussian process.
Figure 5: Gibbs Hawkes Estimated on the Same HP (with $\phi_{\sin}$) Sample of Fig. 1. The GP hyperparameters correspond to the maximal approximate log marginal likelihood is marked as red star. 10 inducing points are employed.