Statistics and Computing (2021) 31:46
https://doi.org/10.1007/s11222-021-10021-x

Efficient EM-variational inference for nonparametric Hawkes process

Feng Zhou[1] · Simon Luo[2,3] · Zhidong Li[4] · Xuhui Fan[5] · Yang Wang[4] · Arcot Sowmya[1] · Fang Chen[4]

Received: 2 May 2020 / Accepted: 18 May 2021 / Published online: 4 June 2021

Abstract
The classic Hawkes process assumes the baseline intensity to be constant and the triggering kernel to be a parametric function. Differently, we present a generalization of the parametric Hawkes process by using a Bayesian nonparametric model called *quadratic Gaussian Hawkes process*. We model the baseline intensity and trigger kernel as the quadratic transformation of random trajectories drawn from a Gaussian process (GP) prior. We derive an analytical expression for the EM-variational inference algorithm by augmenting the latent branching structure of the Hawkes process to embed the variational Gaussian approximation into the EM framework naturally. We also use a series of schemes based on the sparse GP approximation to accelerate the inference algorithm. The results of synthetic and real data experiments show that the underlying baseline intensity and triggering kernel can be recovered efficiently and our model achieved superior performance in fitting capability and prediction accuracy compared to the state-of-the-art approaches.

Keywords Hawkes process · Nonparametric · Gaussian process · Variational inference

1 Introduction

A point process is a stochastic process. It has a wide range of applications. Some examples include seismology (Marsan and Lengline 2008), financial engineering (Hewlett 2006) and epidemics (Rizoiu et al. 2018). The occurrence of an event (e.g. the happening of a disease infection or an earthquake) is treated as a point in a point process on the time axis or the 2-dimensional plane. The Poisson process (Daley and Vere-Jones 2003) and Hawkes process (Hawkes 1971) are two of the most common point processes and they are used to statistically describe the pattern of the occurrence of events.

The Hawkes process can be used as an intensity estimator to model the *self-exciting* phenomenon in a wide range of applications such as traffic accidents (Zhou et al. 2018), criminology (Mohler et al. 2011) and high-frequency financial trades (Bacry et al. 2015). The Hawkes process uses the past events to calculate the probability of the future events occurring. Choosing a function for the baseline intensity and the triggering kernel is a fundamental challenge in Hawkes process. For the classic parametric Hawkes process, the baseline intensity is assumed to be constant and the triggering kernel is assumed to be a parametric function such as an exponential or power-law decay function (Bacry et al. 2015). The parametric assumption leads to convenient inference, however, this assumption is inconsistent with reality in many applications (Mohler et al. 2011; Wheatley et al. 2018). In this situation, the data driven nonparametric approaches are desirable.

The nonparametric Hawkes process is a flexible model that is able to learn the unknown function for the baseline intensity and triggering kernel. For example, Marsan and Lengline (2008) proposed the independent stochastic declustering method to estimate the nonparametric triggering kernel in an EM framework; Lewis and Mohler (2011) extended this algorithm to nonparametric baseline intensity with the Euler-Lagrange equation; Bacry and Muzy (2016) analyzed the relation between the triggering kernel and the second order

✉ Zhidong Li
zhidong.li@uts.edu.au

[1] School of Computer Science and Engineering, University of New South Wales, Kensington, Australia

[2] School of Mathematics and Statistics, The University of Sydney, Sydney, Australia

[3] Data Analytics for Research and Environment, Australian Research Council, Canberra, Australia

[4] Data Science Institute, University of Technology Sydney, Ultimo, Australia

[5] School of Mathematics and Statistics, University of New South Wales, Kensington, Australia

statistics of its counting process to propose an estimation method based on the Wiener–Hopf equation. Eichler et al. (2017) and Reynaud-Bouret et al. (2010) assumed the triggering kernel is a piece-wise constant function and established a quadratic loss to be minimized. Deep learning frameworks have also been explored such as the neural Hawkes process (Mei and Eisner 2017) which used long short-term memory (LSTM) to model the intensity function. However, all previous approaches are frequentist nonparametric algorithms which are based on the likelihood function only. These approaches are problematic because they are prone to overfitting when they do not have the appropriate regularization.

In this paper, we propose a Bayesian nonparametric model for Hawkes process to model a non-constant baseline intensity and a nonparametric triggering kernel with continuous changes. We relieve any parametric assumptions to smooth the baseline intensity and triggering kernel. The Bayesian priors on both components are the quadratic transformation of GP which have theoretical guarantees to be non-negative. In this setting, the inference can be performed without gridding the domain. We utilize the variational Gaussian approximation (Opper and Archambeau 2009) for model inference. However, the inference has two major challenges: (**1**) The baseline intensity is coupled with the triggering kernel in the likelihood function of the Hawkes process, which drastically increases the complexity of performing inference. We address this issue by augmenting the *branching structure* of the Hawkes process to decouple them. The branching structure is a latent variable and is estimated via an expectation-maximization (EM) algorithm (Dempster et al. 1977). The variational Gaussian approximation is embedded into an EM framework naturally. (**2**) In the past, Zhang et al. (2019) have proposed to use a variational Gaussian approximation for Hawkes processes in similar manner, however, the formulation of the baseline intensity was still constant and their inference is performed by high dimensional numerical optimization which is time-consuming let alone embedded into EM iterations. We circumvent this issue by applying the mean-field assumption to derive a closed-form matrix derivative to speed up the inference. Synthetic and real data experimental results show that the flexible baseline intensity and triggering kernel can be recovered and our model is superior w.r.t. fitting capability and prediction accuracy compared to the state-of-the-art techniques. We summarize the contributions presented in the paper as follows:

(1) The baseline intensity and triggering kernel are both relieved to be nonparametric functions that are modulated by a quadratic transformation of a GP.

(2) The variational Gaussian approximation is embedded into an EM framework. The complexity of the EM-variational (EMV) algorithm scales linearly with the number of observations.

(3) We utilize the sparse GP approximation and the mean-field assumption to derive the closed-form matrix derivative of the evidence lower bound (ELBO) to further accelerate EMV to be efficient.

The paper is structured as follows: Sect. 2 presents an overview of the quadratic Gaussian Hawkes process model where we explain how the baseline intensity and triggering kernel are modeled as smooth functions modulated by GP. Section 3 presents the naïve EMV inference algorithm. We then present an accelerated version of the inference algorithm in Sect. 4. The synthetic and real data experiments are summarized in Sect. 5, and conclusions are given in Sect. 6.

2 Quadratic Gaussian Hawkes process

A Hawkes process is a stochastic process that is realized through a sequence of timestamps $D = \{t_i\}_{i=1}^N \in [0, T]$. We denote t_i to be the time of occurrence for the i-th event and T is the observation window of this process. The conditional intensity function is an important way to characterize a Hawkes process so that it is able to capture the temporal dynamics. The conditional intensity function $\lambda(t)$ is defined as the probability of an event occurring in an infinitesimal time interval $[t, t + dt)$ given historical timestamps before t, $\{t_i | t_i < t\}$.

The specific form of the conditional intensity for the Hawkes process is

$$\lambda(t) = \mu(t) + \sum_{t_i < t} \phi(t - t_i) \tag{1}$$

where $\mu(t) > 0$ is the baseline intensity and $\phi(\tau) > 0$ ($\tau = t - t_i$) is the triggering kernel. In the classic Hawkes process, $\mu(t)$ is assumed to be constant and $\phi(\tau)$ is a parametric function such as the exponential decay function. The summation of triggering kernels explains the nature of self-excitation. The non-negativity of $\mu(t)$ and $\phi(\tau)$ guarantees the intensity is non-negative almost surely. Given $\mu(t)$ and $\phi(\tau)$, the Hawkes process likelihood (Daley and Vere-Jones 2003) is written as

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

We propose the quadratic Gaussian Hawkes process (QGHP) which formulates the baseline intensity and triggering kernel as the quadratic transformation of random trajectories drawn from GP priors to guarantee the non-negativity, that is $\mu(t) = f^2(t), \phi(\tau) = g^2(\tau)$ where f

and g are two functions drawn from the corresponding GP prior. It is worth noting that, in this paper, $\mu(t)$ and $\phi(\tau)$ are defined on the support of $[0, T]$ and $[0, T_\phi]$ respectively and the value of $f(t)$ and $g(\tau)$ outside the corresponding intervals will be ignored. The quadratic link function (Flaxman et al. 2017; Lloyd et al. 2015) is used because the inference can be performed in closed form and the variance of the variational distribution is related with the data. For more details about the advantage of utilizing the quadratic link function, please refer to Lloyd et al. (2015, Section 5).

We also propose an EMV algorithm for the inference: embedding the variational Gaussian approximation into an EM framework. Using a naïve Bayesian framework, the inference of posterior of $\mu(t)$ and $\phi(\tau)$ is non-trivial due to intractable integrals in the numerator and denominator. The doubly-intractable problem has been introduced in Adams et al. (2009). In the following section we present our proposed inference approach that tactfully solves the doubly-intractable problem without gridding the domain.

3 Inference

In this section, we present our key technical contribution on the inference algorithm. We first present using sparse GP approximation to avoid the functional optimization problem. Secondly we augment the branching structure of the Hawkes process to decouple the log-likelihood to two independent components. Thirdly we use a variational Gaussian approximation for the inference in each component. Finally we combine them together to obtain the EMV algorithm.

3.1 Sparse GP approximation

The sparse GP approximation (Titsias 2009) has been used to improve the efficiency and avoid the functional optimization issue. f and g are supposed to be dependent on their corresponding inducing points (definition of inducing points is provided in Titsias (2009)) $\mathbf{Z}_f = \{z_f^m\}_{m=1}^{M_f}$ and $\mathbf{Z}_g = \{z_g^m\}_{m=1}^{M_g}$; the function values of f and g at these inducing points are $\mathbf{u}_f$ and $\mathbf{u}_g$ which are stationary and Gaussian distributed as $\mathbf{u}_f \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{z_f z_f})$ and $\mathbf{u}_g \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{z_g z_g})$. Given a sample $\mathbf{u}_f$ and $\mathbf{u}_g$, f and g are assumed to be $f|\mathbf{u}_f \sim \mathcal{GP}(v_f(t), \Sigma_f(t, t'))$ and $g|\mathbf{u}_g \sim \mathcal{GP}(v_g(\tau), \Sigma_g(\tau, \tau'))$ with mean and covariance

$$v_f(t) = \mathbf{k}_{tz_f}\mathbf{K}_{z_f z_f}^{-1}\mathbf{u}_f,\ \Sigma_f(t, t') = k_{tt'} - \mathbf{k}_{tz_f}\mathbf{K}_{z_f z_f}^{-1}\mathbf{k}_{z_f t'}$$
$$v_g(\tau) = \mathbf{k}_{\tau z_g}\mathbf{K}_{z_g z_g}^{-1}\mathbf{u}_g,\ \Sigma_g(\tau, \tau') = k_{\tau\tau'} - \mathbf{k}_{\tau z_g}\mathbf{K}_{z_g z_g}^{-1}\mathbf{k}_{z_g \tau'}$$

with $\mathbf{k}_{tz_f}$ and $\mathbf{k}_{\tau z_g}$ being the kernel vector w.r.t. observations and inducing points while $\mathbf{K}_{z_f z_f}$, $\mathbf{K}_{z_g z_g}$, $k_{tt'}$ and $k_{\tau\tau'}$ being kernel matrices or values w.r.t. inducing points or observations only. Therefore, the joint distribution of the Hawkes process is

$$\begin{aligned} p(D, f, u_f, g, u_g) =& p(D|\mu(t) = f^2, \phi(\tau) = g^2) \\ & p(f|\mathbf{u}_f)p(g|\mathbf{u}_g)p(\mathbf{u}_f)p(\mathbf{u}_g). \end{aligned} \tag{3}$$

3.2 Augmentation of branching structure

The evidence lower bound needs to be obtained for variational inference (Blei et al. 2017). This means f, u_f, g and u_g need to be integrated out in Eq. (3). However, performing this procedure directly is difficult because $\mu(t)$ is coupled with $\phi(\tau)$ in the likelihood.

The branching structure of Hawkes process (Marsan and Lengline 2008; Zhou et al. 2013) is introduced to facilitate inference by decoupling $\mu(t)$ and $\phi(\tau)$. The branching structure $\mathbf{X}$ is a triangular matrix with Bernoulli variables x_{ij} indicating if the i-th event is triggered by itself or a previous event j.

$$x_{ii} = \begin{cases} 1 & \text{if event } i \text{ is a background event} \\ 0 & \text{otherwise} \end{cases}$$
$$x_{ij} = \begin{cases} 1 & \text{if event } i \text{ is caused by event } j, \quad i \neq j \\ 0 & \text{otherwise} \end{cases}$$

After introducing branching structure, we obtain a lower-bound $\mathcal{Q}(\mu(t), \phi(\tau)|\mu^{(s)}(t), \phi^{(s)}(\tau))$ of the log-likelihood where superscript s denotes the last iteration (proof in "Appendix A").

$$\begin{aligned} &\mathcal{Q}(\mu(t), \phi(\tau)|\mu^{(s)}(t), \phi^{(s)}(\tau)) \\ &= \mathbb{E}_{\mathbf{X}}\left[\log p(D, \mathbf{X}|\mu(t), \phi(\tau))\right] \\ &= \underbrace{\sum_{i=1}^{N} p_{ii}\log(\mu(t_i)) - \int_0^T \mu(t)dt(t)}_{\mu} \text{ part}+ \\ &\quad \underbrace{\sum_{i=2}^{N}\sum_{j=1}^{i-1} p_{ij}\log\left(\phi(t_i - t_j)\right) - \sum_{i=1}^{N}\int_{t_i}^{t_i+T_\phi} \phi(t - t_i)dt(\tau)}_{\phi} \text{ part} \\ &\triangleq \log\tilde{p}(D|\mu(t), \mathcal{P}_{ii}) + \log\tilde{p}(D|\phi(\tau), \mathcal{P}_{ij}), \end{aligned} \tag{4}$$

where $\tilde{p}$ means an unnormalized density; T_ϕ is the support of triggering kernel; we can see the lower-bound is decoupled to two independent parts: $\mu(t)$ part and $\phi(\tau)$ part; $p_{ij} = \mathbb{E}(x_{ij})$ can be understood as the probability that i-th event is affected by a previous event j and p_{ii} is the probability that i-th event is a baseline event. Specifically, it is derived as

$$
\begin{aligned}
p_{ij} &= \frac{\phi^{(s)}(\tau_{ij})}{\mu^{(s)}(t_i) + \sum_{j=1}^{i-1} \phi^{(s)}(\tau_{ij})}, \\
p_{ii} &= \frac{\mu^{(s)}(t_i)}{\mu^{(s)}(t_i) + \sum_{j=1}^{i-1} \phi^{(s)}(\tau_{ij})}.
\end{aligned} \tag{5}
$$

3.3 Variational Gaussian approximation

Now the inference can be performed for two components independently because $\mu(t)$ and $\phi(\tau)$ have been decoupled.

3.3.1 Baseline intensity

For the $\mu(t)$ part: $\log \tilde{p}(D|\mu(t) = f^2, \mathcal{P}_{ii})$. $\mathcal{P}_{ii}$ means the diagonal entries of $\mathcal{P} = \mathbb{E}(\mathbf{X})$. We integrate out inducing points $\mathbf{u}_f$ using a Gaussian variational distribution $q(\mathbf{u}_f) = \mathcal{N}(\mathbf{u}_f|\mathbf{m}_f, \mathbf{S}_f)$. We use Jensen's inequality to obtain the ELBO for the $\mu(t)$ part:

$$
\begin{aligned}
&\log \tilde{p}(D|\mathcal{P}_{ii}) \\
&\quad = \log \left[\iint \tilde{p}(D|f, \mathcal{P}_{ii}) p(f|\mathbf{u}_f) p(\mathbf{u}_f) \frac{q(\mathbf{u}_f)}{q(\mathbf{u}_f)} d\mathbf{u}_f df \right] \\
&\quad \geq \iint p(f|\mathbf{u}_f) q(\mathbf{u}_f) d\mathbf{u}_f \log \tilde{p}(D|f, \mathcal{P}_{ii}) df \\
&\qquad + \iint p(f|\mathbf{u}_f) q(\mathbf{u}_f) df \log \left[\frac{p(\mathbf{u}_f)}{q(\mathbf{u}_f)} \right] d\mathbf{u}_f \\
&\quad = \mathbb{E}_{q(f)} \left[\log \tilde{p}(D|f, \mathcal{P}_{ii}) \right] - \mathrm{KL}\left(q(\mathbf{u}_f) || p(\mathbf{u}_f) \right) \\
&\quad \triangleq \mathrm{ELBO}_{\mu},
\end{aligned} \tag{6}
$$

where

$$
q(f) = \int p(f|\mathbf{u}_f) q(\mathbf{u}_f) d\mathbf{u}_f = \mathcal{GP}(f|\tilde{v}_f(t), \tilde{\Sigma}_f(t, t')) \tag{7}
$$

with the mean $\tilde{v}_f(t) = \mathbf{k}_{tz_f} \mathbf{K}^{-1}_{z_f z_f} \mathbf{m}_f$ and the covariance $\tilde{\Sigma}_f(t, t') = k_{tt'} - \mathbf{k}_{tz_f} \mathbf{K}^{-1}_{z_f z_f} \mathbf{k}_{z_f t'} + \mathbf{k}_{tz_f} \mathbf{K}^{-1}_{z_f z_f} \mathbf{S}_f \mathbf{K}^{-1}_{z_f z_f} \mathbf{k}_{z_f t'}$. The KL $\left(q(\mathbf{u}_f) || p(\mathbf{u}_f) \right)$ term has an analytical solution because both elements are Gaussian distributions. The expectation of log-likelihood over $q(f)$ can be written as

$$
\begin{aligned}
\mathbb{E}_{q(f)} \left[\log \tilde{p}(D|f, \mathcal{P}_{ii}) \right] &= \sum_{i=1}^{N} p_{ii} \mathbb{E}_{q(f)} \left[\log f^2(t_i) \right] \\
&\quad - \int_0^T \left\{ \mathbb{E}^2_{q(f)}[f(t)] + \mathrm{Var}_{q(f)}[f(t)] \right\} dt,
\end{aligned} \tag{8}
$$

where we utilize $\mathbb{E}(A^2) = \mathbb{E}^2(A) + \mathrm{Var}(A)$. Eq. (8) also has an analytical solution shown in "Appendix B".

3.3.2 Triggering kernel

For the $\phi(\tau)$ part: $\log \tilde{p}(D|\phi(\tau) = g^2, \mathcal{P}_{ij})$. $\mathcal{P}_{ij}$ means the entries off diagonal of $\mathcal{P} = \mathbb{E}(\mathbf{X})$. Similarly, we integrate out inducing points $\mathbf{u}_g$ using $q(\mathbf{u}_g) = \mathcal{N}(\mathbf{u}_g|\mathbf{m}_g, \mathbf{S}_g)$. The ELBO for the $\phi(\tau)$ part is

$$
\begin{aligned}
&\log \tilde{p}(D|\mathcal{P}_{ij}) \\
&\quad = \log \left[\iint \tilde{p}(D|g, \mathcal{P}_{ij}) p(g|\mathbf{u}_g) p(\mathbf{u}_g) \frac{q(\mathbf{u}_g)}{q(\mathbf{u}_g)} d\mathbf{u}_g dg \right] \\
&\quad \geq \mathbb{E}_{q(g)} \left[\log \tilde{p}(D|g, \mathcal{P}_{ij}) \right] - \mathrm{KL}\left(q(\mathbf{u}_g) || p(\mathbf{u}_g) \right) \\
&\quad \triangleq \mathrm{ELBO}_{\phi},
\end{aligned} \tag{9}
$$

where $q(g)$ is Eq. (7) with notation f and t replaced by g and τ, respectively. The KL term has an analytical solution and the expectation of log-likelihood over $q(g)$ can be written as

$$
\begin{aligned}
\mathbb{E}_{q(g)} \left[\log \tilde{p}(D|g, \mathcal{P}_{ij}) \right] &= \sum_{i=2}^{N} \sum_{j=1}^{i-1} p_{ij} \mathbb{E}_{q(g)} \left[\log g^2(\tau_{ij}) \right] \\
&\quad - \sum_{i=1}^{N} \int_0^{T_\phi} \left\{ \mathbb{E}^2_{q(g)}[g(\tau)] + \mathrm{Var}_{q(g)}[g(\tau)] \right\} d\tau
\end{aligned} \tag{10}
$$

with analytical solution shown in "Appendix B".

3.4 EM-variational algorithm

In this section we present the EMV inference algorithm to infer $\mu(t)$ and $\phi(\tau)$. By augmenting branching structure, we obtain a surrogate function (lower-bound) decoupling $\mu(t)$ and $\phi(\tau)$ to two independent components (E step). For each component, we utilize variational Gaussian approximation to derive an ELBO which should be maximized, thus obtaining an optimal variational distribution (M step).

The radial basis function kernel is used as the GP covariance kernel throughout this paper. The hyperparameters θ_0 and θ_1 of $k(x, x') = \theta_0 \exp\left(-\frac{\theta_1}{2} \|x - x'\|^2 \right)$ are optimized by performing maximization of ELBO over $\{\theta_0, \theta_1\}$ using numerical packages.

Apart from θ_0 and θ_1, the hyperparameters left are the number and location of inducing points. The number of inducing points M is a trade-off between complexity and accuracy. A large M corresponds to a high dimensional $\mathbf{K}_{zz}$ leading to high complexity, while a small M cannot characterize the dynamics of functions.

In our experiments, we assume the inducing points are uniformly located on the support and gradually increase the number until the resulting $\mu(t)$ or $\phi(\tau)$ is not improved much any more. The pseudocode of naïve EMV is shown in Algorithm 1.

Algorithm 1: Naïve EMV algorithm

```
Result: $\mu(t)$, $\phi(\tau)$
Initialize hyperparameters and $\mathcal{P}$;
for do
    Update $\mathcal{P}$ by Eq. (5);
    Update $\mathbf{m}_f^*$, $\mathbf{S}_f^*$, $\mathbf{m}_g^*$ and $\mathbf{S}_g^*$ by
      $\mathbf{m}_f^*, \mathbf{S}_f^* = \text{argmax}_{\mathbf{m}_f, \mathbf{S}_f}\left(\text{ELBO}_\mu\right)$ and
      $\mathbf{m}_g^*, \mathbf{S}_g^* = \text{argmax}_{\mathbf{m}_g, \mathbf{S}_g}\left(\text{ELBO}_\phi\right)$;
    Update $\tilde{v}_f^*$, $\tilde{\Sigma}_f^*$, $\tilde{v}_g^*$ and $\tilde{\Sigma}_g^*$ by Eq. (7) with $\mathbf{m}_f^*$, $\mathbf{S}_f^*$, $\mathbf{m}_g^*$
      and $\mathbf{S}_g^*$;
    Update $\mu(t)$ and $\phi(\tau)$ by $\mu(t) = (\tilde{v}_f^*)^2 + \tilde{\sigma}_f^{2*}$,
      $\phi(\tau) = (\tilde{v}_g^*)^2 + \tilde{\sigma}_g^{2*}$ where we utilize
      $\mathbb{E}(A^2) = \mathbb{E}^2(A) + \text{Var}(A)$, $\tilde{\sigma}_f^{2*}$ and $\tilde{\sigma}_g^{2*}$ are diagonal
      entries of $\tilde{\Sigma}_f^*$ and $\tilde{\Sigma}_g^*$;
    Update hyperparameters.
end
```

4 Inference acceleration

The naïve implementation of EMV algorithm (Algorithm 1) is computationally expensive. The bottleneck is the update of $\mathbf{m}_f^*, \mathbf{S}_f^*, \mathbf{m}_g^*$ and $\mathbf{S}_g^*$ due to numerical optimization. Supposing the number of inducing points $\mathbf{u}_f$ is M_f, the dimensionality of the search space for optimization over $\mathbf{m}_f$ and $\mathbf{S}_f$ is $M_f + M_f(M_f+1)/2$. The space is large even when M_f is small (the case is the same for $\mathbf{u}_g$). We develop two schemes to speed up the algorithm: **(1)** we prove that the optimal variational mean $\mathbf{m}^*$ is analytically zero, **(2)** the complexity is reduced by using mean-field assumption and we derive the closed-form matrix derivative of ELBO w.r.t. $\mathbf{S}$.

4.1 Optimal variational mean

The transformation function is $\mu(t) = f^2$ and it is not a bijection. For every $\mu(t)$, there will be two positive-negative symmetric $f(t)$'s. The posterior of f can be written as

$$p(f|D, \mathcal{P}_{ii}) \propto$$
$$p(D|\mu(t) = f^2, \mathcal{P}_{ii})\mathcal{GP}(f|\mathbf{u}_f)\mathcal{N}(\mathbf{u}_f|\mathbf{0}, \mathbf{K}_{z_f z_f}),$$

where the likelihood is symmetric with f and $-f$. For the prior $\mathcal{GP}(f|\mathbf{u}_f)\mathcal{N}(\mathbf{u}_f|\mathbf{0}, \mathbf{K}_{z_f z_f})$, we can integrate out $\mathbf{u}_f$ and the marginal distribution over f is still Gaussian with a mean of $\mathbf{0}$. Therefore, the prior of f is also symmetric. Conclusively, the posterior $p(f|D, \mathcal{P}_{ii})$ is symmetric. By using variational Gaussian approximation, we are approximating $p(f|D, \mathcal{P}_{ii})$ by a normal distribution $q(f) = \mathcal{GP}(f|\tilde{v}_f(t), \tilde{\Sigma}_f(t, t'))$, $\tilde{v}_f(t) = \mathbf{k}_{t z_f}\mathbf{K}_{z_f z_f}^{-1}\mathbf{m}_f$. It is easy to see $\mathbf{m}_f^* = \mathbf{0}$ definitely; this applies to the $\phi(\tau)$ part as well to obtain $\mathbf{m}_g^* = \mathbf{0}$.

4.2 Optimal variational covariance

With the setting of $\mathbf{m}^* = \mathbf{0}$, the update for $\mathbf{m}_f^*, \mathbf{S}_f^*, \mathbf{m}_g^*$ and $\mathbf{S}_g^*$ becomes the maximization of ELBO over $\mathbf{S}$ only. We derive the closed-form matrix derivative of ELBO over $\mathbf{S}$ (proof in "Appendix C")

$$\begin{aligned}
\frac{\partial \text{ELBO}_\mu}{\partial \mathbf{S}_f} &= -(2\mathbf{K}_{z_f z_f}^{-1}\Psi_f\mathbf{K}_{z_f z_f}^{-1} - \mathbf{K}_{z_f z_f}^{-1}\Psi_f\mathbf{K}_{z_f z_f}^{-1} \circ \mathbf{I}) \\
&+ \sum_{i=1}^{N} p_{ii}\bigg(2\mathbf{K}_{z_f z_f}^{-1}\mathbf{k}_{z_f t_i}\mathbf{k}_{t_i z_f}\mathbf{K}_{z_f z_f}^{-1} \\
&- \mathbf{K}_{z_f z_f}^{-1}\mathbf{k}_{z_f t_i}\mathbf{k}_{t_i z_f}\mathbf{K}_{z_f z_f}^{-1} \circ \mathbf{I}\bigg)/\tilde{\sigma}_f^2(t_i) \\
&- \frac{1}{2}\left(2\mathbf{K}_{z_f z_f}^{-1} - \mathbf{K}_{z_f z_f}^{-1} \circ \mathbf{I} - (2\mathbf{S}_f^{-1} - \mathbf{S}_f^{-1} \circ \mathbf{I})\right), \\
\frac{\partial \text{ELBO}_\phi}{\partial \mathbf{S}_g} &= -\sum_{i=1}^{N}(2\mathbf{K}_{z_g z_g}^{-1}\Psi_g\mathbf{K}_{z_g z_g}^{-1} - \mathbf{K}_{z_g z_g}^{-1}\Psi_g\mathbf{K}_{z_g z_g}^{-1} \circ \mathbf{I}) \\
&+ \sum_{i=2}^{N}\sum_{j=1}^{i-1} p_{ij}\bigg(2\mathbf{K}_{z_g z_g}^{-1}\mathbf{k}_{z_g \tau_{ij}}\mathbf{k}_{\tau_{ij} z_g}\mathbf{K}_{z_g z_g}^{-1} \\
&- \mathbf{K}_{z_g z_g}^{-1}\mathbf{k}_{z_g \tau_{ij}}\mathbf{k}_{\tau_{ij} z_g}\mathbf{K}_{z_g z_g}^{-1} \circ \mathbf{I}\bigg)/\tilde{\sigma}_g^2(\tau_{ij}) \\
&- \frac{1}{2}\left(2\mathbf{K}_{z_g z_g}^{-1} - \mathbf{K}_{z_g z_g}^{-1} \circ \mathbf{I} - (2\mathbf{S}_g^{-1} - \mathbf{S}_g^{-1} \circ \mathbf{I})\right),
\end{aligned} \tag{11}$$

where $\mathbf{I}$ denotes the identity matrix, $\circ$ denotes the Hadam-ard (elementwise) product and $\tilde{\sigma}_f^2(t_i) = \theta_0^f - \mathbf{k}_{t_i z_f}\mathbf{K}_{z_f z_f}^{-1}\mathbf{k}_{z_f t_i} + \mathbf{k}_{t_i z_f}\mathbf{K}_{z_f z_f}^{-1}\mathbf{S}_f\mathbf{K}_{z_f z_f}^{-1}\mathbf{k}_{z_f t_i}$ is the diagonal of $\tilde{\Sigma}_f(t, t')$ and $\tilde{\sigma}_g^2(\tau_{ij}) = \theta_0^g - \mathbf{k}_{\tau_{ij} z_g}\mathbf{K}_{z_g z_g}^{-1}\mathbf{k}_{z_g \tau_{ij}} + \mathbf{k}_{\tau_{ij} z_g}\mathbf{K}_{z_g z_g}^{-1}\mathbf{S}_g\mathbf{K}_{z_g z_g}^{-1}\mathbf{k}_{z_g \tau_{ij}}$ is the diagonal of $\tilde{\Sigma}_g(\tau, \tau')$.

Intuitively, by setting Eq. (11) to 0, the optimal variational covariance can be obtained. However, it is still inefficient because $M_f(M_f+1)/2$ equations are in the nonlinear system. To further accelerate the inference algorithm, $\mathbf{S}_f$ and $\mathbf{S}_g$ are assumed to be diagonal (mean field approximation (Bishop 2007)) so that Eq. (11) can be further simplified. We derive the simplified matrix derivative in the diagonal case (proof in "Appendix C")

$$\begin{aligned}
\frac{\partial \text{ELBO}_\mu}{\partial \mathbf{S}_f} &= -\mathbf{K}_{z_f z_f}^{-1}\Psi_f\mathbf{K}_{z_f z_f}^{-1} \circ \mathbf{I} \\
&+ \sum_{i=1}^{N} p_{ii}\frac{\mathbf{K}_{z_f z_f}^{-1}\mathbf{k}_{z_f t_i}\mathbf{k}_{t_i z_f}\mathbf{K}_{z_f z_f}^{-1} \circ \mathbf{I}}{\tilde{\sigma}_f^2(t_i)} \\
&-\frac{1}{2}\left(\mathbf{K}_{z_f z_f}^{-1} \circ \mathbf{I} - \mathbf{S}_f^{-1}\right), \\
\frac{\partial \text{ELBO}_\phi}{\partial \mathbf{S}_g} &= -\sum_{i=1}^{N}\mathbf{K}_{z_g z_g}^{-1}\Psi_g\mathbf{K}_{z_g z_g}^{-1} \circ \mathbf{I}
\end{aligned}$$

$$+\sum_{i,j} p_{ij} \frac{\mathbf{K}_{z_g z_g}^{-1}\mathbf{k}_{z_g \tau_{ij}}\mathbf{k}_{\tau_{ij} z_g}\mathbf{K}_{z_g z_g}^{-1} \circ \mathbf{I}}{\tilde{\sigma}_g^2(\tau_{ij})}$$
$$-\frac{1}{2}\left(\mathbf{K}_{z_g z_g}^{-1} \circ \mathbf{I} - \mathbf{S}_g^{-1}\right). \qquad (12)$$

In practice, Eq.(12) can be used as a gradient expression for any gradient-based optimization packages. In experiments, we find the diagonal assumption does not make much difference when the underlying function value ($\mu(t)$ or $\phi(\tau)$) does not change drastically. The accelerated EMV is provided in Algorithm 2.

Algorithm 2: Accelerated EMV

```
Result: $\mu(t), \phi(\tau)$
Initialize hyperparameters and $\mathcal{P}$;
for do
    Update $\mathcal{P}$ by Eq. (5);
    Update $\mathbf{S}_f^*$ and $\mathbf{S}_g^*$ by $\partial\text{ELBO}_\mu/\partial\mathbf{S}_f = \mathbf{0}$ and
      $\partial\text{ELBO}_\phi/\partial\mathbf{S}_g = \mathbf{0}$ using Eq. (12);
    Update $\tilde{\Sigma}_f^*$ and $\tilde{\Sigma}_g^*$ by Eq. (7) with $\mathbf{S}_f^*$ and $\mathbf{S}_g^*$;
    Update $\mu(t)$ and $\phi(\tau)$ by $\mu(t) = \tilde{\sigma}_f^{2*}$ and $\phi(\tau) = \tilde{\sigma}_g^{2*}$
      where we utilize $\mathbb{E}(A^2) = \mathbb{E}^2(A) + \text{Var}(A)$, $\tilde{\sigma}_f^{2*}$ and $\tilde{\sigma}_g^{2*}$
      are diagonal entries of $\tilde{\Sigma}_f^*$ and $\tilde{\Sigma}_g^*$;
    Update hyperparameters.
end
```

4.3 Complexity

The complexity of matrix inversion is reduced to $\mathcal{O}(M_f^3 + M_g^3)$ where M_f (or M_g) $\ll N$ because of the sparse GP approximation. This results in a complexity scaling linearly with data size: $\mathcal{O}(N)$ multiplied by a constant L where $L = \int_{T_\phi} \frac{\mu(t)}{1-\int\phi(\tau)d\tau}dt \ll N$ because of the sparsity of branching structure $p_{ij} = 0$: previous points more than T_ϕ far away from i-th point have no influence on i-th point.

Our experiment is conducted on a desktop computer (CPU: i7-6700 with 8GB RAM). The runtime of the naïve implementation (Algorithm 1) is more than 2 h with $N = 205$, 6 inducing points for both $\mathbf{Z}_f$ and $\mathbf{Z}_g$ and 100 EM iterations. The accelerated Algorithm 2 costs only 4 min in the same setting, which drastically reduces the running time.

5 Experimental results

We evaluate the fitting and prediction ability of our proposed QGHP model in both synthetic and real data experiments. We compare our approach with the following baseline models:

- Gaussian-Cox (GC) process: a GP modulated inhomogeneous Poisson process. The inference is performed by the algorithm in Samo and Roberts (2015). It is only for real data.
- RKHS-Cox (RKHSC) process: an inhomogeneous Poisson process whose intensity is estimated by a reproducing kernel Hilbert space formulation (Flaxman et al. 2017). It is only for real data.
- Parametric Hawkes (PH) process: the classic Hawkes process with constant baseline intensity and exponential decay triggering kernel. The inference is performed by maximum likelihood estimation.
- Model independent stochastic declustering (MISD): the MISD (Lewis and Mohler 2011) is an EM-based nonparametric algorithm for Hawkes process, where the baseline intensity is constant and the triggering kernel is discretized to be a histogram function. We use *MISD-#* (*#* is the number of bins) to indicate the corresponding model.
- Wiener–Hopf (WH): a nonparametric Hawkes process inference algorithm with constant baseline intensity and smooth triggering kernel. The inference is performed by the solution of a Wiener–Hopf equation (Bacry and Muzy 2016).
- Variational Bayesian Hawkes process (VBHP): a Bayesian nonparametric Hawkes process with constant baseline intensity and smooth triggering kernel. The inference is performed by variational inference (Zhang et al. 2019).

5.1 Synthetic data experiments

In our synthetic data experiment, we compare the fitting and prediction ability of our model to that of PH, MISD-10, MISD-20, WH and VBHP. It is worth noting that GC and RKHSC are excluded because they are Poisson process models which cannot provide the baseline intensity and triggering kernel.

We consider a general scenario with time changing baseline intensity and non-exponential triggering kernel

$$\mu(t) = \sin\left(\frac{2\pi}{T}\cdot t\right) + 1 \;\; (0 < t < T)$$
$$\phi(\tau) = \begin{cases} 0.25\sin\tau & (0 < \tau \le \pi) \\ 0 & (\pi < \tau < T_\phi) \end{cases}$$

and use the thinning algorithm (Ogata 1998) to generate two sets of observations with one being the training data and the other one the test data. With $T_\phi = 6$ and $T = 400$, 1716 points are obtained for the training and test data which is shown in Fig. 1a. Our goal is to estimate the baseline intensity and triggering kernel based on the observations.

Metrics To compare the fitting and prediction ability of different models, we utilize multiple metrics including training and test log-likelihood (*LogLik*), and estimation error

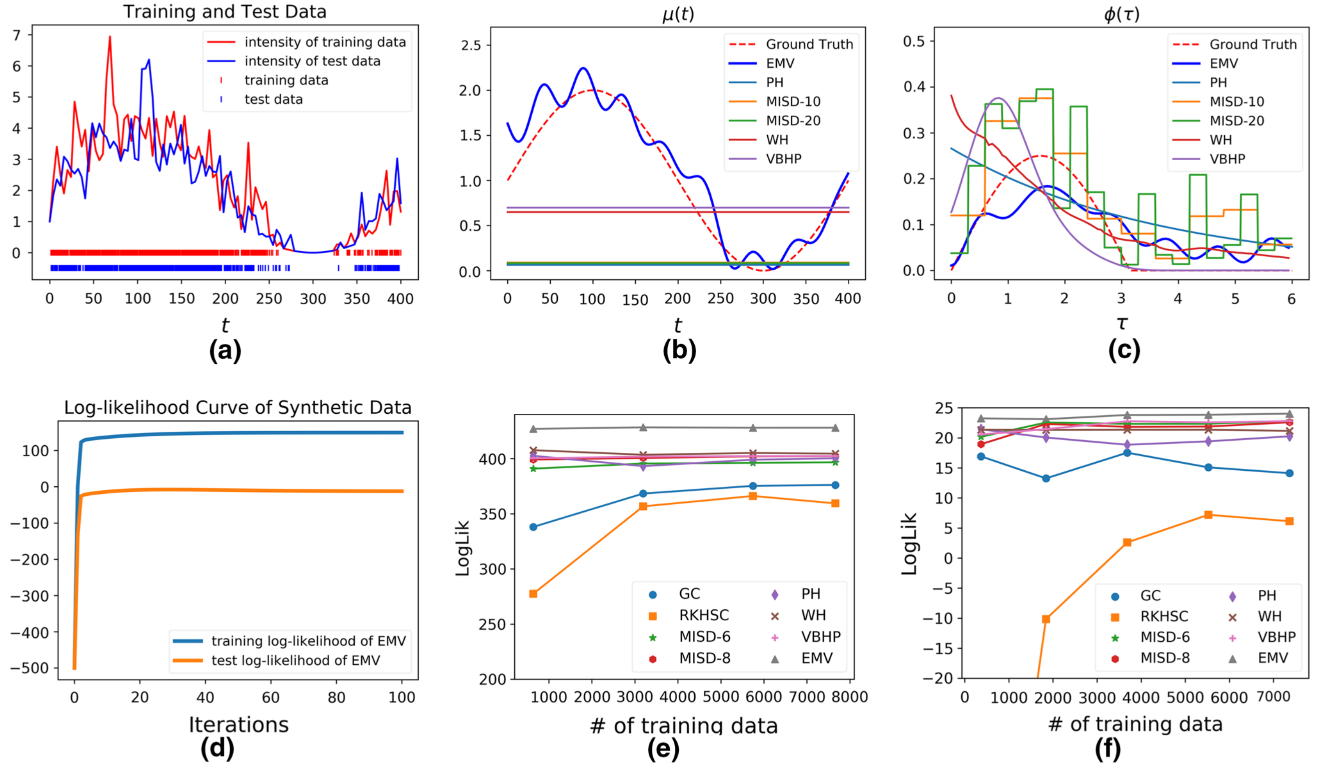

Fig. 1 Experimental results of synthetic and real data. **a** The simulated points and intensity function of training and test datasets. **b** and **c** The estimated $\hat{\mu}(t)$ and $\hat{\phi}(\tau)$ of the synthetic data of all models. **d**0 The training and test log-likelihood curve of EMV algorithm w.r.t. iterations for the synthetic data. **e** and **f** The training *LogLik* of various models over the number of training data for vehicle collision and taxi pickup, respectively

(*EstErr*) which is defined as the integral mean squared error between the estimated $\hat{\phi}(\tau)$ ($\hat{\mu}(t)$) and the ground truth.

Results In experiments, the hyperparameters in all models are fine tuned to obtain the optimal test log-likelihood. For our EMV algorithm, 10 and 8 inducing points are uniformly located on the support of $\mu(t)$ and $\phi(\tau)$ by cross validation. The estimated $\hat{\mu}(t)$ and $\hat{\phi}(\tau)$ of all models are shown in Fig. 1b and c where the result obtained from EMV algorithm is the closest to the ground truth. The training and test *LogLik*, and *EstErr* results are shown in Table 1 where we can see our QGHP model outperforms the alternatives w.r.t. all metrics. This is because only our QGHP model is capable of estimating nonparametric $\mu(t)$ and $\phi(\tau)$ simultaneously, which leads to a better goodness-of-fit. As shown in Fig. 1d, our EMV algorithm converges fast with only 10 iterations needed to reach a plateau.

5.2 Real data experiments

In our real data experiment, we apply our QGHP model to two different datasets, estimate the baseline intensity and triggering kernel based on observations using EMV algorithm and compare its performance to the state-of-the-art approaches.

Vehicle Collisions[1] The vehicle collision dataset is from New York City Police Department. We filter out weekday records in almost one month (Sep.18th–Oct.13th 2017). The number of collisions in each day is about 600. Records in Sep.18th-Oct.6th are used as training data and Oct.9th–13th are held out as test data.

The car collision can be modelled as a self-exciting phenomenon because there will be triggering influence on subsequent accidents caused by the traffic congestion caused by the initial accident. In the past, the nonparametric Hawkes processes have been applied to the transportation domain so that the triggering kernel is relieved to be nonparametric, but the baseline intensity is still a constant in approaches such as MISD or WH. This is an inappropriate hypothesis in the vehicle collision application because there are many situations where the baseline is not constant. For example the road is quiet at night so the baseline intensity of car accidents is lower than that in the normal time and the traffic is so busy at peak times that the baseline intensity will be increased. Our proposed QGHP provides a solution that is able to learn both

[1] https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95.

Table 1 Training and test *LogLik* (larger values indicate better performance), and *EstErr* (smaller values indicate better performance) of synthetic data of all models

	PH	MISD-10	MISD-20	WH	VBHP	EMV
Training *LogLik*	127.84	130.48	134.95	90.78	132.75	149.29
Test *LogLik*	– 12.11	– 16.13	– 24.23	– 45.43	– 14.64	– 9.17
$EstErr(\hat{\mu}(t), \mu(t))$	549.69	531.09	535.54	248.79	235.99	15.98
$EstErr(\hat{\phi}(\tau), \phi(\tau))$	0.116	0.051	0.076	0.061	0.052	0.018

the time-changing baseline intensity and a flexible triggering kernel simultaneously.

We compare the fitting and prediction ability of EMV, VBHP, WH, MISD-6, MISD-8, PH, RKHSC and GC. The whole observation period T is set to 1440 min (24 h) and the support of triggering kernel T_ϕ is set to 60 min. For the hyperparameters, the bandwidth of WH is selected to be 1.2 using cross-validation, and there are 6 inducing points on $\phi(\tau)$ ($M_g = 6$) and 8 inducing points on $\mu(t)$ ($M_f = 8$) for EMV. The hyperparameters for RKHSC, GC and VBHP are chosen based on a grid search to minimize the error between the integral of learned intensity and the average number of timestamps on each sequence. The final result is the average of learned $\hat{\mu}(t)$ or $\hat{\phi}(\tau)$ of each day.

Taxi Pickup[2] This dataset includes trip records from all trips completed in green taxis in New York City from January to June 2016. For the experiments, we select the data from Jan.7th to Feb.1st to be the training data and Jan.2nd–6th is held out as test data. In this period, we filter out pickup dates and times for long-distance trips (> 15 miles) since long-distance trips usually have different patterns compared to shorter trips. The average number of pickups each day is about 400.

We also compare the fitting and prediction ability of all models on the taxi pickup dataset with the same setup with vehicle collision. The whole observation period T is set to 24 h and the support of triggering kernel T_ϕ is set to be 1 h.

Metrics The *EstErr* cannot be used as a metric because the ground truth is unknown for real world data. Instead, we use the training *LogLik* and prediction accuracy (*PreAcc*) as metrics to measure the fitting and prediction ability respectively. Given an event sequence $\{t_1, ..., t_{i-1}\}$, the mean arrival time of next point t_i is estimated as $\mathbb{E}[t_i] = \int_{t_{i-1}}^{\infty} tp(t_i = t)dt$ where $p(t_i = t) = \lambda(t)\exp\left(-\int_{t_{i-1}}^{t} \lambda(s)ds\right)$. The intractable integral can be estimated using Monte Carlo methods. The *PreAcc* is defined as the percentage of points whose predicted arrival time is within an error bound of the true arrival time.

Results For the fitting task, we evaluate the training *LogLik* of various models when the number of training data varies for each dataset. The training *LogLik* of EMV and other baseline models for both real datasets are shown in Fig. 1e and f. We observe that PH, MISD-6, MISD-8, WH, VBHP and EMV outperform GC and RKHSC (which are both inhomogeneous Poisson processes); this demonstrates the necessity of using Hawkes process to discover the underlying self-exciting phenome-non in both real datasets. In addition, our EMV algorithm is consistently superior to all other Hawkes process inference algorithms (PH, MISD, WH and VBHP) which have assumptions on the baseline intensity or triggering kernel restricting their ability to capture the dynamics. This demonstrates that our EMV algorithm possesses the better goodness-of-fit because it has the flexibility to describe the nonparametric $\mu(t)$ and $\phi(\tau)$ simultaneously. The learned baseline intensities for two datasets are shown in Fig. 2, which provides support for our speculation that the baseline intensity at midnight (2:00 a.m.–4:00 a.m.) is much lower than that in the daytime (6:00 a.m.–18:00 p.m.) for car accidents and taxi pickups.

In the prediction task, we measure the *PreAcc* of all alternatives on both datasets. We assume only the top 17% of a sequence is observed (the error bound $\epsilon = 0.12$ for vehicle collision and 0.24 for taxi pickup, 500 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. The average *PreAcc* of the test data is shown in Table 2 where we can observe that EMV is superior to the alternatives.

6 Conclusion

In the classic Hawkes process, the baseline intensity and triggering kernel are assumed to be a constant and a parametric function respectively, which is convenient for inference but leads to limited capability for model expression. To further generalize the model and perform inference from a Bayesian perspective, we apply the quadratic transformation of GP as prior on the baseline intensity and triggering kernel and solve it with an EM-variational inference algorithm. We use the sparse GP approximation and the mean-field assumption to derive a closed-form matrix derivative of the ELBO to make the inference more efficient. Our experimental results have show that our model can achieve a better fitting and

[2] https://data.cityofnewyork.us/Transportation/2016-Green-Taxi-Trip-Data/hvrh-b6nb.

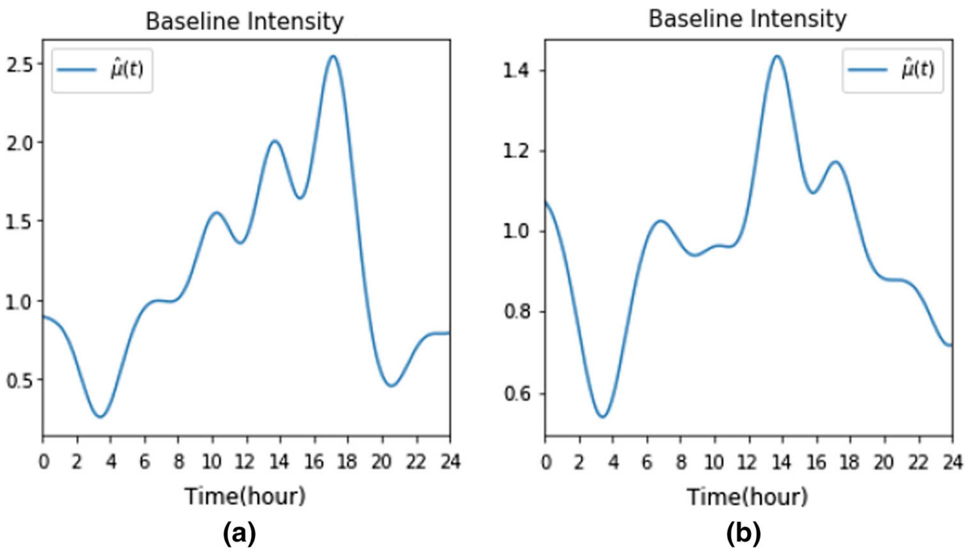

Fig. 2 The estimated baseline intensity $\hat{\mu}(t)$ for two real datasets, **a** vehicle collisions; **b** taxi pickup

Table 2 The *PreAcc* of all alternatives on both real datasets

	Vehicle collision (%)	Taxi pickup (%)
GC	17.3	53.8
RKHSC	29.2	64.0
PH	60.6	67.1
MISD-6	67.6	68.3
MISD-8	67.6	67.9
WH	67.3	67.5
VBHP	67.8	67.7
EMV	71.7	70.4

prediction performance than the state-of-the-art approaches for both synthetic and real datasets. Further investigation includes the extension to multivariate Hawkes process with sharing properties on the triggering kernels and the more general spatial-temporal process model where the triggering kernel is defined on a multi-dimensional space.

Appendices

A Proof of lower-bound

The lower-bound $Q(\mu(t), \phi(\tau)|\mu^{(s)}(t), \phi^{(s)}(\tau))$ in Eq. (4) is induced as follows. Based on Jensen's inequality, we have

$$\log p(D|\mu(t), \phi(\tau)) = \sum_{i=1}^{N} \log \left(\mu(t_i) + \sum_{j=1}^{i-1} \phi(t_i - t_j) \right) - \int_0^T \left(\mu(t) + \sum_{t_i<t} \phi(t - t_i) \right) dt$$

$$\geq \sum_{i=1}^{N} \left(p_{ii} \log \frac{\mu(t_i)}{p_{ii}} + \sum_{j=1}^{i-1} p_{ij} \log \frac{\phi(t_i - t_j)}{p_{ij}} \right) - \int_0^T \mu(t)dt - \sum_{i=1}^{N} \int_{t_i}^{t_i+T_\phi} \phi(t - t_i)dt$$

$$= \sum_{i=1}^{N} p_{ii} \log \mu(t_i) - \int_0^T \mu(t)dt + \sum_{i=2}^{N} \sum_{j=1}^{i-1} p_{ij} \log \phi(t_i - t_j) - \sum_{i=1}^{N} \int_{t_i}^{t_i+T_\phi} \phi(t - t_i)dt + C \tag{13}$$

where C is a constant because p_{ii} and p_{ij} are given in the E-step.

B Analytical solution of ELBO

The KL $\left(q(\mathbf{u}_f)||p(\mathbf{u}_f)\right)$ can be written as

$$\text{KL}\left(q(\mathbf{u}_f)||p(\mathbf{u}_f)\right) = \frac{1}{2}\left[\text{Tr}(\mathbf{K}_{z_f z_f}^{-1} \mathbf{S}_f) + \log \frac{|\mathbf{K}_{z_f z_f}|}{|\mathbf{S}_f|} - M_f + \mathbf{m}_f^T \mathbf{K}_{z_f z_f}^{-1} \mathbf{m}_f \right], \tag{14}$$

where $\text{Tr}(\cdot)$ means trace, $|\cdot|$ means determinant and M_f is the dimensionality of $\mathbf{u}_f$.

The last two terms in Eq. (8) have analytical solutions (Lloyd et al. 2015)

$$\int_0^T \mathbb{E}_{q(f)}^2[f(t)]dt = \mathbf{m}_f^T \mathbf{K}_{z_f z_f}^{-1} \mathbf{\Psi}_f \mathbf{K}_{z_f z_f}^{-1} \mathbf{m}_f, \tag{15}$$

$$\int_0^T \mathrm{Var}_{q(f)}[f(t)]dt = \theta_0^f T - \mathrm{Tr}(\mathbf{K}_{z_f z_f}^{-1}\boldsymbol{\Psi}_f) + \mathrm{Tr}(\mathbf{K}_{z_f z_f}^{-1}\mathbf{S}_f\mathbf{K}_{z_f z_f}^{-1}\boldsymbol{\Psi}_f), \tag{16}$$

where $\Psi_f(z_f, z'_f) = \int_0^T k(z_f, t)k(t, z'_f)dt$. For the squared exponential kernel, Ψ_f can be written as (Lloyd et al. 2015)

$$\Psi_f(z_f, z'_f) = -\frac{(\theta_0^f)^2}{2}\sqrt{\frac{\pi}{\theta_1^f}}\exp\left(-\frac{\theta_1^f(z_f - z'_f)^2}{4}\right)\left[\mathrm{erf}\left(\sqrt{\theta_1^f}(\bar{z}_f - T)\right) - \mathrm{erf}\left(\sqrt{\theta_1^f}\bar{z}_f\right)\right], \tag{17}$$

where $\mathrm{erf}(\cdot)$ is Gauss error function and $\bar{z}_f = (z_f + z'_f)/2$.

The first term in Eq. (8) also has an analytical solution (Lloyd et al. 2015)

$$\begin{aligned}\mathbb{E}_{q(f)}\left[\log f^2(t_i)\right] &= \int_{-\infty}^{\infty} \log f^2(t_i)\mathcal{N}(f(t_i)|\tilde{v}_f(t_i), \tilde{\sigma}_f^2(t_i))df(t_i) \\ &= -\tilde{G}\left(-\frac{\tilde{v}_f^2(t_i)}{2\tilde{\sigma}_f^2(t_i)}\right) + \log\left(\frac{\tilde{\sigma}_f^2(t_i)}{2}\right) - C,\end{aligned} \tag{18}$$

where $\tilde{\sigma}_f^2(t_i)$ is the diagonal entry of $\tilde{\Sigma}_f(t, t')$ in Eq. (7) at t_i, C is the Euler-Mascheroni constant 0.57721566 and $\tilde{G}(z)$ is a special case of the partial derivative of the confluent hypergeometric function ${}_1F_1(a, b, z)$ (Lloyd et al. 2015)

$$\tilde{G}(z) = {}_1F_1{}^{(1,0,0)}(0, 0.5, z). \tag{19}$$

It is worth noting that $\tilde{G}(z)$ does not need to be computed for inference. Actually we only need to know $\tilde{G}(0) = 0$ because $\mathbf{m}_f^* = \mathbf{0}$ as we have shown in the section of inference speed up.

C Matrix derivative of ELBO

Given $\mathbf{m}_f = \mathbf{0}$, ELBO_μ can be written as

$$\begin{aligned}&\mathrm{ELBO}_\mu \\ &= -\left(\theta_0^f T - \mathrm{Tr}(\mathbf{K}_{z_f z_f}^{-1}\boldsymbol{\Psi}_f) + \mathrm{Tr}(\mathbf{K}_{z_f z_f}^{-1}\mathbf{S}_f\mathbf{K}_{z_f z_f}^{-1}\boldsymbol{\Psi}_f)\right) \\ &+ \sum_{i=1}^N p_{ii}\left(-\tilde{G}(0) + \log(\tilde{\sigma}_f^2(t_i)) - \log 2 - C\right) \\ &- \frac{1}{2}\left(\mathrm{Tr}(\mathbf{K}_{z_f z_f}^{-1}\mathbf{S}_f) + \log|\mathbf{K}_{z_f z_f}| - \log|\mathbf{S}_f| - M_f\right).\end{aligned} \tag{20}$$

If $\mathbf{S}_f$ is symmetric, $\partial\mathrm{ELBO}_\mu/\partial\mathbf{S}_f$ can be written as

$$\begin{aligned}\frac{\partial\mathrm{ELBO}_\mu}{\partial\mathbf{S}_f} &= -(2\mathbf{K}_{z_f z_f}^{-1}\Psi_f\mathbf{K}_{z_f z_f}^{-1} - \mathbf{K}_{z_f z_f}^{-1}\Psi_f\mathbf{K}_{z_f z_f}^{-1} \circ \mathbf{I}) \\ &+ \sum_{i=1}^N p_{ii}\Big(2\mathbf{K}_{z_f z_f}^{-1}\mathbf{k}_{z_f t_i}\mathbf{k}_{t_i z_f}\mathbf{K}_{z_f z_f}^{-1} \\ &- \mathbf{K}_{z_f z_f}^{-1}\mathbf{k}_{z_f t_i}\mathbf{k}_{t_i z_f}\mathbf{K}_{z_f z_f}^{-1} \circ \mathbf{I}\Big)/\tilde{\sigma}_f^2(t_i) \\ &- \frac{1}{2}\left(2\mathbf{K}_{z_f z_f}^{-1} - \mathbf{K}_{z_f z_f}^{-1} \circ \mathbf{I} - (2\mathbf{S}_f^{-1} - \mathbf{S}_f^{-1} \circ \mathbf{I})\right),\end{aligned} \tag{21}$$

where $\mathbf{I}$ denotes the identity matrix, $\circ$ denotes the Hadamard (elementwise) product and $\tilde{\sigma}_f^2(t_i) = \theta_0^f - \mathbf{k}_{t_i z_f}\mathbf{K}_{z_f z_f}^{-1}\mathbf{k}_{z_f t_i} + \mathbf{k}_{t_i z_f}\mathbf{K}_{z_f z_f}^{-1}\mathbf{S}_f\mathbf{K}_{z_f z_f}^{-1}\mathbf{k}_{z_f t_i}$ is the diagonal entry of $\tilde{\Sigma}_f(t, t')$ in Eq. (7).

If $\mathbf{S}_f$ is diagonal, $\partial\mathrm{ELBO}_\mu/\partial\mathbf{S}_f$ can be further simplified as

$$\begin{aligned}\frac{\partial\mathrm{ELBO}_\mu}{\partial\mathbf{S}_f} &= -\mathbf{K}_{z_f z_f}^{-1}\Psi_f\mathbf{K}_{z_f z_f}^{-1} \circ \mathbf{I} \\ &+ \sum_{i=1}^N p_{ii}\frac{\mathbf{K}_{z_f z_f}^{-1}\mathbf{k}_{z_f t_i}\mathbf{k}_{t_i z_f}\mathbf{K}_{z_f z_f}^{-1} \circ \mathbf{I}}{\tilde{\sigma}_f^2(t_i)} \\ &- \frac{1}{2}\left(\mathbf{K}_{z_f z_f}^{-1} \circ \mathbf{I} - \mathbf{S}_f^{-1}\right).\end{aligned} \tag{22}$$

Similarly given $\mathbf{m}_g = \mathbf{0}$, ELBO_ϕ can be written as

$$\begin{aligned}&\mathrm{ELBO}_\phi \\ &= -\sum_{i=1}^N\left(\theta_0^g T_\phi - \mathrm{Tr}(\mathbf{K}_{z_g z_g}^{-1}\boldsymbol{\Psi}_g) + \mathrm{Tr}(\mathbf{K}_{z_g z_g}^{-1}\mathbf{S}_g\mathbf{K}_{z_g z_g}^{-1}\boldsymbol{\Psi}_g)\right) \\ &+ \sum_{i=2}^N\sum_{j=1}^{i-1} p_{ij}\left(-\tilde{G}(0) + \log(\tilde{\sigma}_g^2(\tau_{ij})) - \log 2 - C\right) \\ &- \frac{1}{2}\left(\mathrm{Tr}(\mathbf{K}_{z_g z_g}^{-1}\mathbf{S}_g) + \log|\mathbf{K}_{z_g z_g}| - \log|\mathbf{S}_g| - M_g\right).\end{aligned} \tag{23}$$

If $\mathbf{S}_g$ is symmetric, $\partial\mathrm{ELBO}_\phi/\partial\mathbf{S}_g$ can be written as

$$\begin{aligned}\frac{\partial\mathrm{ELBO}_\phi}{\partial\mathbf{S}_g} &= -\sum_{i=1}^N(2\mathbf{K}_{z_g z_g}^{-1}\Psi_g\mathbf{K}_{z_g z_g}^{-1} - \mathbf{K}_{z_g z_g}^{-1}\Psi_g\mathbf{K}_{z_g z_g}^{-1} \circ \mathbf{I}) \\ &+ \sum_{i=2}^N\sum_{j=1}^{i-1} p_{ij}\Big(2\mathbf{K}_{z_g z_g}^{-1}\mathbf{k}_{z_g \tau_{ij}}\mathbf{k}_{\tau_{ij} z_g}\mathbf{K}_{z_g z_g}^{-1} \\ &- \mathbf{K}_{z_g z_g}^{-1}\mathbf{k}_{z_g \tau_{ij}}\mathbf{k}_{\tau_{ij} z_g}\mathbf{K}_{z_g z_g}^{-1} \circ \mathbf{I}\Big)/\tilde{\sigma}_g^2(\tau_{ij}) \\ &- \frac{1}{2}\left(2\mathbf{K}_{z_g z_g}^{-1} - \mathbf{K}_{z_g z_g}^{-1} \circ \mathbf{I} - (2\mathbf{S}_g^{-1} - \mathbf{S}_g^{-1} \circ \mathbf{I})\right),\end{aligned} \tag{24}$$

where $\tilde{\sigma}_g^2(\tau_{ij}) = \theta_0^g - \mathbf{k}_{\tau_{ij} z_g}\mathbf{K}_{z_g z_g}^{-1}\mathbf{k}_{z_g \tau_{ij}} + \mathbf{k}_{\tau_{ij} z_g}\mathbf{K}_{z_g z_g}^{-1}\mathbf{S}_g\mathbf{K}_{z_g z_g}^{-1}\mathbf{k}_{z_g \tau_{ij}}$ is the diagonal entry of $\tilde{\Sigma}_g(\tau, \tau')$.

If $\mathbf{S}_g$ is diagonal, $\partial \mathrm{ELBO}_\phi / \partial \mathbf{S}_g$ can be further simplified as

$$\begin{aligned}\frac{\partial \mathrm{ELBO}_\phi}{\partial \mathbf{S}_g} &= -\sum_{i=1}^{N} \mathbf{K}_{z_g z_g}^{-1} \Psi_g \mathbf{K}_{z_g z_g}^{-1} \circ \mathbf{I} \\ &+ \sum_{i=2}^{N} \sum_{j=1}^{i-1} p_{ij} \frac{\mathbf{K}_{z_g z_g}^{-1} \mathbf{k}_{z_g \tau_{ij}} \mathbf{k}_{\tau_{ij} z_g} \mathbf{K}_{z_g z_g}^{-1} \circ \mathbf{I}}{\tilde{\sigma}_g^2(\tau_{ij})} \\ &- \frac{1}{2} \left(\mathbf{K}_{z_g z_g}^{-1} \circ \mathbf{I} - \mathbf{S}_g^{-1} \right). \end{aligned} \tag{25}$$

References

Adams, R.P., Murray, I., MacKay, D.J.: Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities. In: Proceedings of the 26th Annual International Conference on Machine Learning, pp. 9–16. ACM (2009)

Bacry, E., Muzy, J.F.: First-and second-order statistics characterization of Hawkes processes and non-parametric estimation. IEEE Trans. Inf. Theory **62**(4), 2184–2202 (2016)

Bacry, E., Mastromatteo, I., Muzy, J.F.: Hawkes processes in finance. Mark. Microstruct. Liq. **1**(01), 1550005 (2015)

Bishop, C.: Pattern recognition and machine learning (Information Science and Statistics), 1st edn. 2006. corr. 2nd printing edn. Springer, New York (2007)

Blei, D.M., Kucukelbir, A., McAuliffe, J.D.: Variational inference: a review for statisticians. J. Am. Stat. Assoc. **112**(518), 859–877 (2017)

Daley, D.J., Vere-Jones, D.: An Introduction to the Theory of Point Processes. Probability and Its Applications, vol. I. Springer, Berlin (2003)

Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B (Methodol.) **39**(1), 1–22 (1977)

Eichler, M., Dahlhaus, R., Dueck, J.: Graphical modeling for multivariate Hawkes processes with nonparametric link functions. J. Time Ser. Anal. **38**(2), 225–242 (2017)

Flaxman, S., Teh, Y.W., Sejdinovic, D., et al.: Poisson intensity estimation with reproducing kernels. Electron. J. Stat. **11**(2), 5081–5104 (2017)

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

Hewlett, P.: Clustering of order arrivals, price impact and trade path optimisation. In: Workshop on Financial Modeling with Jump processes, Ecole Polytechnique, pp 6–8 (2006)

Lewis, E., Mohler, G.: A nonparametric EM algorithm for multiscale Hawkes processes. J. Nonparametr. Stat. **1**(1), 1–20 (2011)

Lloyd, C., Gunter, T., Osborne, M., Roberts, S.: Variational inference for Gaussian process modulated Poisson processes. In: International Conference on Machine Learning, pp. 1814–1822 (2015)

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

Mei, H., Eisner, J.M.: The neural Hawkes process: a neurally self-modulating multivariate point process. In: Advances in Neural Information Processing Systems, pp. 6754–6764 (2017)

Mohler, G.O., Short, M.B., Brantingham, P.J., Schoenberg, F.P., Tita, G.E.: Self-exciting point process modeling of crime. J. Am. Stat. Assoc. **106**(493), 100–108 (2011)

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

Opper, M., Archambeau, C.: The variational Gaussian approximation revisited. Neural Comput. **21**(3), 786–792 (2009)

Reynaud-Bouret, P., Schbath, S., et al.: Adaptive estimation for Hawkes processes; application to genome analysis. Ann. Stat. **38**(5), 2781–2822 (2010)

Rizoiu, M.A., Mishra, S., Kong, Q., Carman, M., Xie, L.: SIR-Hawkes: linking epidemic models and Hawkes processes to model diffusions in finite populations. In: Proceedings of the 2018 World Wide Web Conference, pp. 419–428 (2018)

Samo, Y.L.K., Roberts, S,: Scalable nonparametric Bayesian inference on point processes with Gaussian processes. In: International Conference on Machine Learning, pp. 2227–2236 (2015)

Titsias, M.: Variational learning of inducing variables in sparse Gaussian processes. In: Artificial Intelligence and Statistics, pp. 567–574 (2009)

Wheatley, S., Schatz, M., Sornette, D.: The ARMA point process and its estimation. arXiv preprint arXiv:1806.09948 (2018)

Zhang, R., Walder, C., Rizoiu, M.A.: Variational inference for sparse Gaussian process modulated Hawkes process. arXiv preprint arXiv:1905.10496v2 (2019)

Zhou, K., Zha, H., Song, L.: Learning triggering kernels for multi-dimensional Hawkes processes. In: International Conference on Machine Learning, pp. 1301–1309 (2013)

Zhou, F., Li, Z., Fan, X., Wang, Y., Sowmya, A., Chen, F.: A refined MISD algorithm based on Gaussian process regression. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp 584–596. Springer (2018)