Hawkes Process with Stochastic Triggering Kernel

Feng Zhou[1,4(✉)], Yixuan Zhang[2], Zhidong Li[3,4], Xuhui Fan[1], Yang Wang[3,4], Arcot Sowmya[1], and Fang Chen[3,4]

[1] University of New South Wales, Sydney, Australia
[2] The University of Sydney, Sydney, Australia
[3] University of Technology Sydney, Sydney, Australia
[4] CSIRO DATA61, Sydney, Australia
feng.zhou@data61.csiro.au

Abstract. The impact from past to future is a vital feature in modelling time series data, which has been described by many point processes, e.g. the Hawkes process. In classical Hawkes process, the triggering kernel is assumed to be a deterministic function. However, the triggering kernel can vary with time due to the system uncertainty in real applications. To model this kind of variance, we propose a Hawkes process variant with stochastic triggering kernel, which incorporates the variation of triggering kernel over time. In this model, the triggering kernel is considered to be an independent multivariate Gaussian distribution. We derive and implement a tractable inference algorithm based on variational auto-encoder. Results from synthetic and real data experiments show that the underlying mean triggering kernel and variance band can be recovered, and using the stochastic triggering kernel is more accurate than the vanilla Hawkes process in capacity planning.

Keywords: Hawkes process · Stochastic triggering kernel

1 Introduction

Point process is a common statistical model in describing the pattern of event occurrence in many real world applications, such as a series of earthquakes and the order book in finance. Mutual dependence between events is an important factor in describing the clustering effect in point process. A variety of models are proposed for the dependence, such as Hawkes process (HP) [10] and correcting model [16]. Among those models, HP is the most extensively used one for modelling the self-exciting phenomenon where the influence decays over time.

HP has been used to estimate the intensity (rate of event occurrence) by accumulating the triggering effect from past events. As an intensity estimator, it has been used widely in social networks [18], crime [14] and financial engineering [8]. The triggering kernel in most HP implementations [8] is modelled as a deterministic function. In the real world, however, the actual triggering effect

Q. Yang et al. (Eds.): PAKDD 2019, LNAI 11439, pp. 319–330, 2019.
https://doi.org/10.1007/978-3-030-16148-4_25

from each event can vary because of the system uncertainty and the deterministic triggering kernel is rather limited in capability to model the variation. To model this phenomenon, we introduce variance into the triggering kernel to enable the triggering kernel of HP to be stochastic. We visualize it as a band addition to the triggering kernel (see the example in Fig. 1a).

The importance of the band may be ignored in real applications, because the learned average triggering kernel usually has the largest likelihood to fit the observed data. As a result, when we do prediction, the vanilla HP would eventually be used. However, as we can see later, this band is meaningful for the risk-based planning. For example, when capacity planning is performed in the taxi allocation problem with HP [7], the arriving rate of pickup events is predicted from historic pickups. Based on the prediction, vehicles can be allocated to an area to cover the pickup need (i.e. #pickups $\leq$ #vehicles). If the taxi company uses the intensity $\lceil\lambda\rceil$ learned from vanilla HP as the expected rate of pickups to satisfy, about 50% probability that the pickup need can be satisfied. To plan for a higher probability, more vehicles need to be sent, e.g. for extra probability $P_m = Poisson(x \leq M|\lambda) - Poisson(x \leq \lambda|\lambda)$, extra $m = \lceil M - \lambda\rceil$ vehicles need to be sent. However, in Sect. 6, when there is a significant variance on the triggering effect, sending m vehicles can only satisfy pickup need with extra probability less than P_m, which will lead to a decision with insufficient capacity. Using our stochastic triggering kernel, one can obtain extra information about the distribution of the triggering effect, so the insufficient capacity could be compensated. The similar issue could happen in other HP-based capacity planning applications, as long as there is a significant variance on the triggering kernel.

We propose a HP variant with stochastic triggering kernel (HP-STK), aimed at quantifying the variance of triggering kernel so as to overcome the problem mentioned above. Based on Gaussian white noise, we consider two cases for the variance: homoscedasticity (i.e. constant variance) and heteroscedasticity (i.e. time-varying variance). Then we propose a tractable inference method to replace the original maximum likelihood estimation (MLE) and apply the inference of both cases to the variational auto-encoder (VAE) [11] framework.

To our best knowledge, no work has been done before to model the variance of triggering kernel in HP. Specifically, our work makes the following contributions: (1) we propose a new HP variant named HP-STK, in which the variance of triggering kernel is incorporated to overcome the underestimation problem in capacity planning; (2) two special cases are considered: homoscedasticity and heteroscedasticity; (3) the uniform-trigger-kernel-based MLE is proposed to replace the original MLE and a VAE-based algorithm is used for inference.

2 Related Work

The model proposed in this paper is motivated by the Cox process [2]. The Cox process, also known as the doubly stochastic Poisson process, is a stochastic process which is an extension of a Poisson process where the intensity function

is itself a stochastic process. It has been widely used in many applications, such as astronomy [9] and neuroscience [3]. A common version of Cox process is the Gaussian Cox process [15], where the intensity function is modeled as a Gaussian process. However, the inference is intractable because of non-conjugacy and integration over infinite-dimensional random function. Different inference algorithms based on Markov chain Monte Carlo (MCMC) or Laplace approximation have been proposed in [1,4]. In Cox process, the randomness is added to the intensity, but in this paper the randomness is on triggering kernel to reduce dimensions.

There are also HP extensions to model the randomness of triggering kernel. For example, Dassios [5] proposed a stochastic HP, where jumps in the intensity function are considered to be independent and identically distributed (i.i.d.) random variables. Lee [12] extended all jumps to a stochastic process and solved it using stochastic differential equation. Both works focus on stochastic jumps, but our proposed model considers the whole triggering kernel as a stochastic process which is more generalized.

Another related direction is VAE [11]. VAE has a similar architecture with auto-encoder, but makes an assumption about the distribution of latent variables. VAE is a generative model, which combines ideas from neural network with statistical inference. It can be used to learn a low dimensional representation Z of high dimensional data X. It assumes that the data is generated by a decoder $P(X|Z)$ and the encoder is learning an approximation $Q(Z|X)$ to the posterior distribution $P(Z|X)$. It uses the variational method for latent representation learning, which results in a specific loss function. In this paper we apply the loss of VAE into our model.

3 Proposed Model

3.1 Hawkes Process

A Hawkes process is a stochastic process, whose realization is a sequence of timestamps $\{t_i\} \in [0, T]$. Here, t_i stands for the time of occurrence for the i-th event and T is the observation duration for this process. An important way to characterize a HP is through the definition of a conditional intensity function that captures the temporal dynamics. The conditional intensity function is defined as the probability of event occurring in an infinitesimal time interval $[t, t + dt)$ given the history:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(event\ occurring\ in\ [t, t + \Delta t)|\mathcal{H}_t)}{\Delta t} \tag{1}$$

where $\mathcal{H}_t = \{t_i | t_i < t\}$ are the historical timestamps before time t. Then the specific form of intensity for HP is:

$$\lambda(t) = \mu + \sum_{t_i < t} \gamma^*(t - t_i) \tag{2}$$

where $\mu > 0$ is the baseline intensity which is a constant, and $\gamma^*(\cdot)$ is the triggering kernel. In most cases, the triggering kernel is assumed to be an exponential decay function. The summation of triggering kernels explains the nature of self-excitation, which is the occurrence of events in the past will intensify the intensity of events occurring in the future. Then the log-likelihood function can be expressed using the above conditional intensity as:

$$\log \mathcal{L} = \sum_{i=1}^{n} \log \lambda(t_i) - \int_0^T \lambda(t) dt \tag{3}$$

3.2 HP with Stochastic Triggering Kernel

In HP-STK, we target to introduce variance into the triggering kernel of HP. We define the HP-STK model and see what is the variance of triggering kernel.

Definition 1. *HP-STK is a Hawkes process whose triggering kernel after event t_i can be written as a sample drawn from a stochastic process with $\Delta t \in \mathbb{R}^+$ as:*

$$\gamma_i(\Delta t) = \bar{\gamma}(\Delta t|\boldsymbol{\xi}) + \epsilon_i(\Delta t), \textit{where } \epsilon_i(\Delta t) \sim P(\epsilon(\Delta t)|\boldsymbol{\theta}) \tag{4}$$

where $\gamma_i(\Delta t)$ is the triggering kernel after event t_i, $\bar{\gamma}(\Delta t|\boldsymbol{\xi})$ is a deterministic triggering kernel with parameters $\boldsymbol{\xi}$, $\epsilon_i(\Delta t)$ is a noise function for $\gamma_i(\Delta t)$ and $P(\cdot)$ is a distribution over function with parameters $\boldsymbol{\theta}$.

Naturally, $P(\epsilon(\Delta t)|\boldsymbol{\theta})$ can be defined as a Gaussian process. Here for simplicity $P(\epsilon(\Delta t)|\boldsymbol{\theta})$ is defined as an independent multivariate Gaussian distribution $\boldsymbol{N}(\epsilon(\Delta t)|\mathbf{0}, \sigma^2(\Delta t) \cdot \boldsymbol{I})$ (expressed in finite dimensions) where $\boldsymbol{I}$ is the identity matrix which means there is no covariance. $\bar{\gamma}(\Delta t|\boldsymbol{\xi})$ and $\sigma^2(\Delta t)$ are both defined to be in parametric form. In conclusion, we define $\bar{\gamma}(\Delta t|\boldsymbol{\xi}) = \alpha \exp(-\beta \Delta t)$, $\sigma^2(\Delta t) = \sigma_c^2$ in homoscedastic case and $\sigma^2(\Delta t) = (\alpha_\sigma \exp(-\beta_\sigma \Delta t))^2$ in heteroscedastic case. Here we define the $\sigma(\Delta t)$ to be an exponential decay function because in many scenarios it would be common to have a high variance just after a triggering event and have a lower variance afterwards, but in fact $\sigma(\Delta t)$ can be extended to other cases, e.g. linear decreasing variance or periodic variance. It can be seen that the homoscedasticity is just a special case of heteroscedasticity by setting: $\alpha_\sigma = \sigma_c$ and $\beta_\sigma = 0$.

The intensity of HP-STK can be written as:

$$\lambda(t) = \mu + \sum_{t_i < t} (\alpha \exp(-\beta(t - t_i)) + \epsilon_i(t - t_i)) \tag{5}$$

To avoid the superposition of $\epsilon_i(t - t_i)$ to explode, $\epsilon_i(\Delta t)$ and $\bar{\gamma}(\Delta t)$ are both defined on the support of $[0, T_\gamma]$ and 0 afterwards. In the theory of point process the intensity has to be positive, so $\lambda(t)$ is restricted to $(\lambda(t))_+$ (i.e. $\lambda(t) = 0$ if $\lambda(t) < 0$). Because the $\gamma_i(\Delta t)$ is subject to Gaussian distribution: see (4), so the $\lambda(t)$ is also subject to Gaussian distribution: see (5) and $(\lambda(t))_+$ is subject to a truncated Gaussian distribution. In real applications, the triggering kernel

variance $\sigma^2(\Delta t)$ is always small compared with the intensity, so the truncated Gaussian distribution can be seen as a Gaussian distribution approximately. As we can see later, using Gaussian to describe $\gamma_i(\Delta t)$ introduces computation convenience to the inference.

3.3 Stability Condition

The stability condition of Hawkes process has been proposed in [10]: the Hawkes process P_t is stable if and only if $\int_0^\infty \gamma^*(\cdot) < 1$.

Because $\gamma_i(\cdot)$ is subject to Gaussian distribution in our model, the $\int_0^\infty \gamma_i(\cdot)$ is also subject to Gaussian distribution:

$$\int_0^\infty \gamma_i(\cdot) \sim N(x|\int_0^\infty \alpha \exp(-\beta t)dt, \int_0^\infty \sigma^2(t)dt) \tag{6}$$

where Δt is replaced by t and x is the integral value.

Definition 2. *HP-STK is probabilistically stable with* $P(\int_0^\infty \gamma_i(\cdot) < 1)$.

For homoscedasticity, in order to avoid the $\int_0^\infty \sigma_c^2 dt$ to explode, $\gamma_i(\Delta t)$ is defined on the support of $[0, T_\gamma]$. Given $\exp(-\beta T_\gamma) \approx 0$, we have the stability probability:

$$P_{homo} = \int_{-\infty}^{1} N(x|\frac{\alpha}{\beta}, \sigma_c^2 T_\gamma)dx \tag{7}$$

For heteroscedasticity, the stability probability is:

$$P_{hetero} = \int_{-\infty}^{1} N(x|\frac{\alpha}{\beta}, \frac{\alpha_\sigma^2}{2\beta_\sigma})dx \tag{8}$$

The probabilistic stability of homoscedastic HP-STK is constrained by T_γ. Therefore, when T_γ is undetermined, heteroscedastic HP-STK is recommended.

4 Inference

4.1 Inference with Uniform Triggering Kernel

Given a set of observed data, the goal of inference is to evaluate these parameters: μ, α, β, σ_c for homoscedastic case, and μ, α, β, α_σ, β_σ for the heteroscedastic case. We use MLE to infer parameters where the log-likelihood is:

$$\begin{aligned} &\log \mathcal{L}(\{t_i\}_{i=1}^n|\mu, \boldsymbol{\Theta}) \\ =&\log \int_{\gamma_n} \cdots \int_{\gamma_2} \int_{\gamma_1} \mathcal{L}(\{t_i\}_{i=1}^n|\mu, \gamma_1, \gamma_2, \cdots, \gamma_n) \\ &P(\gamma_1|\boldsymbol{\Theta})P(\gamma_2|\boldsymbol{\Theta}) \cdots P(\gamma_n|\boldsymbol{\Theta})d\gamma_1 d\gamma_2 \cdots d\gamma_n \end{aligned} \tag{9}$$

where the $\boldsymbol{\Theta}$ stands for $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ in (4). However, this log marginal likelihood is complicated to work out because of multiple integrals. To solve this problem, an

intuitive way is to assume $\gamma_1 = \gamma_2 = \cdots = \gamma_n = \gamma$ (i.e. the uniform triggering kernel). Then the log-likelihood can be rewritten as:

$$\log \mathcal{L}(\{t_i\}|\mu, \boldsymbol{\Theta}) = \log \int_{\gamma} \mathcal{L}(\{t_i\}|\mu, \gamma) \cdot P(\gamma|\boldsymbol{\Theta}) d\gamma \tag{10}$$

It is worth noting that, given a set of observed data, the estimation with the uniform triggering kernel is equivalent to the original one. This is proved by (9) and (10), because we get the same $\log \mathcal{L}(\{t_i\})$ with respect to $\mu, \boldsymbol{\Theta}$.

After transforming (9) to (10), we can directly infer the parameters using Monte Carlo integration. However, we still need to calculate the likelihood which is not numerically stable. To solve this problem, we propose an inference method based on VAE.

4.2 Inference with VAE

In fact, our proposed model can be considered as a VAE to some extent. So the loss function [11] of VAE can be applied to our model for inference. The loss function of VAE is the negative log-likelihood with a regularizer:

$$L = -\mathbb{E}[\log \mathcal{L}(\{t_i\}|\mu, \gamma(\cdot))] + \kappa \cdot D_{KL}[P(\gamma(\cdot)|\boldsymbol{\Theta}) \| Q(\gamma(\cdot))] \tag{11}$$

where the first term is the expectation of log-likelihood of $\{t_i\}$ given $\gamma(\cdot)$. The expectation is taken with respect to the encoder's distribution over $\gamma(\cdot)$. This term encourages the decoder to learn to construct the observed sequence data. If the decoder's output does not fit the data well, it will incur a large cost in the loss function. The second term is a regularizer with a weight parameter κ. It is the Kullback-Leibler (KL) divergence between the encoder's distribution $P(\gamma(\cdot)|\boldsymbol{\Theta})$[1] and $Q(\gamma(\cdot))$. $Q(\gamma(\cdot))$ is a benchmark distribution and it describes a priori about $\gamma(\cdot)$. This divergence measures how close $P(\gamma(\cdot)|\boldsymbol{\Theta})$ is to $Q(\gamma(\cdot))$.

In the loss function, the first term can be rewritten as $-\int_{\gamma} \log \mathcal{L}(\{t_i\}|\mu, \gamma) \cdot P(\gamma|\boldsymbol{\Theta}) d\gamma$. It is an integral over an infinite-dimensional stochastic function and it has no analytical solution because of non-conjugacy. To solve these problems, we use discretization and Monte Carlo integration to transform the integral into an average of log-likelihood. By putting log into the integration, we avoid the calculation of likelihood by log-likelihood which is more numerically stable. The Monte Carlo integration will produce a volatile loss function which is not differentiable because of the sampling process, and we can use the reparameterization trick [6] in VAE to make it differentiable. The reparameterization trick is as follows: if we have $x \sim \mathcal{N}(m, \sigma^2)$ and then standaridize it to $\mathcal{N}(0, 1)$, we could revert it back to the original distribution by $x = m + x' \cdot \sigma$ where $x' \sim \mathcal{N}(0, 1)$. Now the sampling process is outside the loss function, so the gradient of loss function will not be affected by sampling.

[1] Customarily, $Q(\cdot)$ is used for encoder's distribution in VAE, but here to be consistent with the previous discussion $P(\cdot)$ is used.

The second term is a KL divergence. In VAE, a popular choice of $Q(\cdot)$ is $\mathcal{N}(0,1)$ to express the prior knowledge [6]. In our setting, we select $Q(\gamma(\Delta t)) = \boldsymbol{\mathcal{N}}(\mathbf{0}, \boldsymbol{I})$ for the homoscedastic case, and $Q(\gamma(\Delta t)) = \boldsymbol{\mathcal{N}}(\mathbf{0}, \exp(-\phi \cdot \Delta t) \cdot \boldsymbol{I})$ for the heteroscedastic case, where ϕ is a constant which can be set manually in experiment. Having $Q(\gamma(\Delta t))$ to be a Gaussian distribution also introduces another benefit. Because the $P(\gamma(\cdot)|\boldsymbol{\Theta})$ in our model is also assumed to be Gaussian: see (4), the KL divergence between $P(\gamma(\cdot)|\boldsymbol{\Theta})$ and $Q(\gamma(\Delta t))$ could be computed in closed form. The KL divergence between two Gaussian distributions is:

$$\begin{aligned} D_{KL}[\boldsymbol{\mathcal{N}}(\boldsymbol{m}_1, \boldsymbol{\Sigma}_1) \| \boldsymbol{\mathcal{N}}(\boldsymbol{m}_2, \boldsymbol{\Sigma}_2)] = \frac{1}{2}[\log|\boldsymbol{\Sigma}_2| - \log|\boldsymbol{\Sigma}_1| \\ - k + \mathrm{Tr}\{\Sigma_2^{-1}\Sigma_1\} + (\mathrm{m}_2 - \mathrm{m}_1)^{\mathrm{T}}\Sigma_2^{-1}(\mathrm{m}_2 - \mathrm{m}_1)] \end{aligned} \tag{12}$$

where k is the dimension of Gaussian, $\mathrm{Tr}\{\}$ is the trace of matrix, $|\cdot|$ is the determinant. Both Gaussian distributions in our model are assumed to be independent which means the covariance $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are both diagonal matrices. This independence assumption improves the computational efficiency further.

After getting the loss function, we can train the model using the generic gradient descent method to optimize the loss with respect to the parameters $\mu, \alpha, \beta, \sigma_c$ in homoscedastic case or $\mu, \alpha, \beta, \alpha_\sigma, \beta_\sigma$ in heteroscedastic case.

5 Synthetic Data Experiment

In synthetic data experiments, we prove that the underlying mean triggering kernel and the corresponding variance parameters can be recovered.

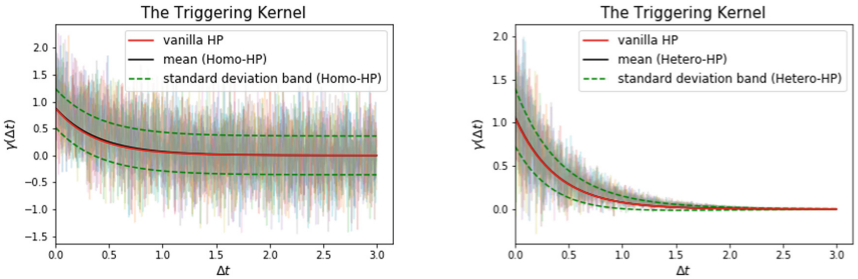
5.1 Homoscedastic Stochastic Triggering Kernel

Based on the thinning algorithm [17], we generate data by setting $\mu = 10$, $\bar{\gamma}(\Delta t) = 1 \cdot \exp(-2 \cdot \Delta t)$, $\sigma_c = 0.5$ and $T_\gamma = 3$. We sampled 10 sets of synthetic data and each of them is a sequence of timestamps in $[0, T]$ where $T = 20$, with a realization of about 400 events.

We use both of the vanilla HP and the homoscedastic HP-STK to recover the parameters for each set of the synthetic data. For both models, the evaluation of parameters is the average of 10 results. For vanilla HP, the final estimations are $\hat{\mu} = 11.04$, $\hat{\alpha} = 0.88$, $\hat{\beta} = 2.71$; for homoscedastic HP-STK, with the configuration of $\kappa = 0.015$, $Q(\gamma(\Delta t)) = \boldsymbol{\mathcal{N}}(\mathbf{0}, \boldsymbol{I})$ and 300 samples from $\boldsymbol{\mathcal{N}}(\mathbf{0}, \boldsymbol{I})$ to perform Monte Carlo integration, the final estimations are $\hat{\mu} = 10.98$, $\hat{\alpha} = 0.88$, $\hat{\beta} = 2.51$, $\hat{\sigma_c} = 0.36$. The learned triggering kernel is shown in Fig. 1a. We can see that the vanilla HP only gives out a deterministic function, while the homoscedastic version gives out an additional variance band.

5.2 Heteroscedastic Stochastic Triggering Kernel

Similarly, in heteroscedastic case, we set $\mu = 2$, $\bar{\gamma}(\Delta t) = 1 \cdot \exp(-2 \cdot \Delta t)$, $\sigma(\Delta t) = 0.5 \cdot \exp(-2 \cdot \Delta t)$ and $T_\gamma = 3$. We generate timestamps in $[0, T]$ where $T = 100$, resulting in a realization of about 400 events. 10 synthetic datasets are generated.



(a) homoscedastic (b) heteroscedastic

Fig. 1. The triggering kernel from vanilla HP and HP-STK (black and red lines overlap), the shade region is the $\gamma_i(\Delta t)$s of 10 sets of synthetic data. (Color figure online)

We use the similar setting for this experiment, except that homoscedastic HP-STK is replaced by heteroscedastic HP-STK. For vanilla HP, the final estimations are $\hat{\mu} = 2.32$, $\hat{\alpha} = 1.05$, $\hat{\beta} = 2.60$; for heteroscedastic HP-STK, with the configuration of $\kappa = 0.015$, $Q(\gamma(\Delta t)) = \boldsymbol{\mathcal{N}}(\mathbf{0}, \exp(-4 \cdot \Delta t) \cdot \boldsymbol{I})$ and 300 samples from $\boldsymbol{\mathcal{N}}(\mathbf{0}, \boldsymbol{I})$ for Monte Carlo integration, the final estimations are $\hat{\mu} = 2.33$, $\hat{\alpha} = 1.06$, $\hat{\beta} = 2.60$, $\hat{\alpha_\sigma} = 0.33$, $\hat{\beta_\sigma} = 1.52$. The learned triggering kernel is shown in Fig. 1b. The vanilla HP only gives out a deterministic function, while the heteroscedastic version gives out an additional time-decreasing variance band.

6 Applications

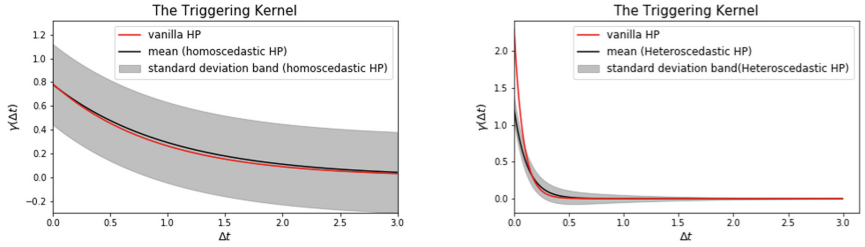
To evaluate the effectiveness of our model, we conduct experiments on two real datasets, taxi pickup and crime. We discuss the results and show how HP-STK outperforms vanilla HP in the application of decision on capacity planning.

6.1 Datasets and Experiment Setting

Green Taxi Pickup in New York City: This dataset includes trip records from all trips completed in green taxis in New York City from January to June in 2016. In the experiment, the data from January 1st to 15th is used. We filter out pick-up dates and times for all long-distance trips (>15 miles), since the long distance trips usually have different patterns with short ones [13]. In addition, we pre-process the data by adding a small time interval to separate all the simultaneous records. As a result, we obtain 6223 pickups for 15 days, and the observed variance is 50.39 given 1 h as time interval. This means the actual number of pickups in short periods can be very unstable, so we model it with the homoscedastic HP-STK.

We apply both of the vanilla HP and homoscedastic HP-STK to model the triggering effect of pickups. We assume the triggering kernels are independent for different days. The support of $\gamma(\Delta t)$ is $[0, 3]$ and the time unit is 1 h.

The evaluation of parameters is the average of 15 training results. For vanilla HP, the final estimations are $\hat{\mu} = 5.23$, $\hat{\alpha} = 0.78$, $\hat{\beta} = 1.09$; for homoscedastic HP-STK, with $\kappa = 0.015$, $Q(\gamma(\Delta t)) = \boldsymbol{\mathcal{N}}(\mathbf{0}, \boldsymbol{I})$ which is consistent with the synthetic experiment, the final estimations are $\hat{\mu} = 5.25$, $\hat{\alpha} = 0.78$, $\hat{\beta} = 0.98$, $\hat{\sigma_c} = 0.34$. The learned $\gamma(\cdot)$ is shown in Fig. 2a. It can be seen that the mean $\gamma(\cdot)$ from homoscedastic version is close to the vanilla result, but it gives out an additional variance band. The corresponding intensity of January 4th is plotted in Fig. 3a. The black solid line is the intensity learned from vanilla HP, the gray band corresponds to the variance band of intensity with $\pm\sigma_\lambda(t)$.



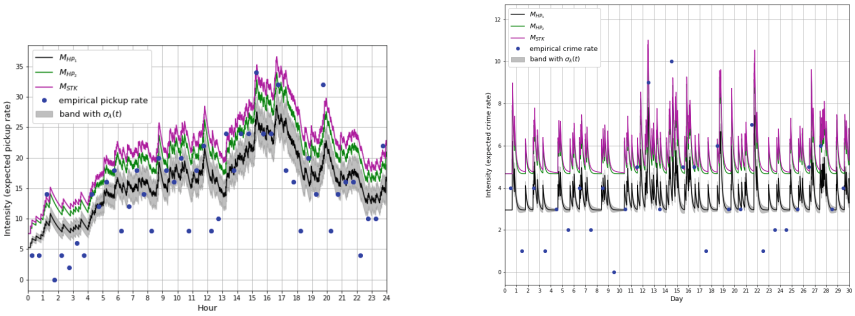
(a) Taxi: homoscedasticity (b) Crime: heteroscedasticity

Fig. 2. Trigger kernels learned from two real datasets for vanilla HP and HP-STK.

Theft of Vehicle in Vancouver: The data of crimes in Vancouver comes from the Vancouver Open Data Catalogue. It is extracted on 2017-07-18 and it includes all valid felony, misdemeanour and violation crimes from 2003-01-01 to 2017-07-13. We filter out the records of which the crime type is 'Theft of Vehicle' from 2012 to 2016. As a result, we obtain 6320 records for 5 years and the observed variance is 4.29 given 1 day as the time interval. This is stabler than taxi pickups, therefore we model it with the heteroscedastic HP-STK.

We apply both the vanilla HP and heteroscedastic HP-STK to model the triggering effect in crime. We assume the triggering kernels are independent for different years. The support of $\gamma(\Delta t)$ is $[0, 3]$ and the time unit is 1 day.

The evaluation of parameters is the average of 5 training results. For vanilla HP, the final estimations are $\hat{\mu} = 2.80$, $\hat{\alpha} = 2.29$, $\hat{\beta} = 12.21$; for heteroscedastic HP-STK, with the configuration of $\kappa = 0.015$, $Q(\gamma(\Delta t)) = \boldsymbol{\mathcal{N}}(\mathbf{0}, \exp(-4 \cdot \Delta t) \cdot \boldsymbol{I})$ which is consistent with the synthetic experiment, the final estimations are $\hat{\mu} = 2.95$, $\hat{\alpha} = 1.21$, $\hat{\beta} = 8.31$, $\hat{\alpha_\sigma} = 0.17$, $\hat{\beta_\sigma} = 1.25$. The learned $\gamma(\cdot)$ is shown in Fig. 2b. It can be seen that the mean $\gamma(\cdot)$ from heteroscedastic version is close to the vanilla result, but it gives out an additional variance band. The corresponding intensity of 2016-year crime is plotted in Fig. 3b. The black solid line is the intensity learned from vanilla HP, the gray band corresponds to the variance band of intensity with $\pm\sigma_\lambda(t)$.



(a) (b)

Fig. 3. (a): M_{HP_1}, M_{HP_2} and M_{STK} of 4th Jan. in taxi dataset based on vanilla HP and homoscedastic HP-STK. Blue points are empirically estimated pickup rates in every 30 min. (b): M_{HP_1}, M_{HP_2} and M_{STK} of 2016 year crime in Vancouver based on vanilla HP and heteroscedastic HP-STK. Blue points are empirically estimated crime rates in each day. (Only 30 days are shown). (Color figure online)

6.2 Use Case for HP-STK

We examine the use case based on the variance of triggering kernel for HP-STK. It is discussed with the comparison with vanilla HP and applied to both datasets.

Decision for Capacity Planning: In the taxi dataset, the taxi company needs to decide the number $(M(t))$ of taxis to meet the pickup need on time t. We omit t in the following discussion for simplicity. If the company uses intensity λ_{HP} learned from vanilla HP, and send $M_{HP_1} = \lambda_{HP}$ (black line in Fig. 3a) taxis to satisfy the pickup need, about 50% probability[2] that all pickups can be satisfied. To plan for a higher probability, the planner needs to send more taxis. So if the variance of Poisson distribution is taken into consideration, we let $M_{HP_2} = \lambda_{HP} + \sqrt{\lambda_{HP}}$ (green line in Fig. 3a), then theoretically extra 29.7% $(Poisson(x < \lambda_{HP} + \sqrt{\lambda_{HP}}) - Poisson(x < \lambda_{HP})$ where x is real pickup need) probability should be added to satisfy the need, given that the average intensity of 15 days is about $\overline{\lambda}_{HP} = 17$ pickups per hour. To empirically estimate this probability, we compute pickup rates in each 0.5 h (blue points in Fig. 3a). The probability is defined as the number of blue points under the corresponding intensity line divided by the total number, which is shown in Table 1. However, in Table 1, only about 23.2% probability has been added using M_{HP_2} comparing with using M_{HP_1} by averaging the probabilities of 15 days. The difference between theoretical and empirical results means that using vanilla HP underestimates the uncertainty while our method can provide more accurate one.

To demonstrate the superiority of our model, here we also show the M_{STK} got from homoscedastic HP-STK. After we learn the variance of triggering kernel σ_c, we can get the variance of intensity σ_λ (gray band in Fig. 3a) using (5). Then we sample $\{\lambda^i_{STK}\}_{i=1}^{100}$ from the Gaussian distribution $N(\overline{\lambda_{STK}}, \sigma^2_\lambda)$,

[2] Based on the Poisson process, the probability could be larger when intensity is low.

Table 1. The probability of satisfying pickup need for M_{HP_1}, M_{HP_2} and M_{STK} of Jan. 1st to 15th taxi pickup in NYC.

Probability	Jan. 1st	Jan. 2nd	Jan. 3th	Jan. 4th	Jan. 5th	Jan. 6th	Jan. 7th	Jan. 8th	Jan. 9th	Jan. 10th	Jan. 11th	Jan. 12th	Jan. 13th	Jan. 14th	Jan. 15th	Average
M_{HP1}	50.0%	62.5%	47.9%	54.2%	56.3%	72.9%	66.7%	52.1%	56.3%	52.1%	64.6%	72.9%	62.5%	64.6%	52.1%	59.2%
M_{HP2}	64.6%	83.3%	83.3%	83.3%	89.6%	87.5%	85.4%	81.3%	83.3%	85.4%	83.3%	91.7%	77.1%	81.3%	75.0%	82.4%
M_{STK}	72.9%	85.4%	89.6%	89.6%	93.8%	89.6%	89.6%	87.5%	85.4%	89.6%	89.6%	93.8%	89.6%	91.7%	81.3%	87.9%

Table 2. The probability of satisfying security need for M_{HP_1}, M_{HP_2} and M_{STK} of 2012 to 2016 crime in Vancouver.

Probability	2012	2013	2014	2015	2016	Average
M_{HP1}	64.2%	69.6%	55.1%	53.7%	44.3%	57.4%
M_{HP2}	88.3%	91.2%	87.7%	79.2%	75.7%	84.4%
M_{STK}	91.0%	92.6%	88.2%	82.7%	79.8%	86.9%

which are samples larger than the mean. We set the expected rate of pickups as $M_{STK} = \frac{1}{100}\sum_{i=1}^{100}(\lambda^i_{STK} + \sqrt{\lambda^i_{STK}})$ (magenta line in Fig. 3a). We also compute the probability of satisfying pickup need of M_{STK} which is shown in Table 1. It can be seen that about 28.7% probability has been added using M_{STK} comparing with using M_{HP_1} by averaging the probabilities of 15 days. This result is close to the theoretical result 29.7%, which means HP-STK is more accurate.

Similarly, the capacity planning decision task is also performed in crime dataset using same definition for M_{HP_1}, M_{HP_2} and M_{STK} (black, green and magenta lines in Fig. 3b, respectively). In the dataset, $\bar{\lambda}_{HP} = 4$, therefore theoretically we should observe 26.05% difference between M_{HP_2} and M_{HP_1}. To empirically estimate this probability, we compute crime occurrence rates in each day which are shown as blue points in Fig. 3b. The probability result is shown in Table 2 which shows that the difference between M_{HP_2} and M_{HP_1} is 27.0% that is close to the theoretical one. This means the variance of triggering kernel is quite small, which is consistent with the result in Fig. 2b. In such case, the vanilla HP is good enough for capacity planning and there is no need to use HP-STK because the magenta line is very close to the green line (see Fig. 3b).

7 Conclusion

We extended HP with stochastic triggering kernel and considered both the homoscedastic and heteroscedastic cases. Our proposed model can provide the variance of triggering kernel, so allow us to overcome the underestimation problem in capacity planning. Along with the model, we also propose a tractable inference based on VAE loss function. Results from synthetic data show that the HP-STK model can recover the underlying mean triggering kernel and the corresponding variance. The usage of HP-STK in taxi allocation discloses that the taxi pickup has a highly stochastic triggering kernel. Vanilla HP will underestimate the expected pickup rate. Without misleading the taxi dispatcher, HP-STK could provide a more accurate rate. Furthermore, another case in crime with a

stabler triggering kernel is used to test that HP-STK could disclose the data stability as expected. There is also freedom to maneuver the stochastic triggering kernel to adapt to other real-life applications or to invent nonparametric stochastic triggering kernels.

References

1. 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)
2. Cox, D.R.: Some statistical methods connected with series of events. J. Roy. Stat. Soc. Ser. B (Methodol.) **17**, 129–164 (1955)
3. Cunningham, J.P., Byron, M.Y., Shenoy, K.V., Sahani, M.: Inferring neural firing rates from spike trains using Gaussian processes. In: Advances in Neural Information Processing Systems, pp. 329–336 (2008)
4. Cunningham, J.P., Shenoy, K.V., Sahani, M.: Fast Gaussian process methods for point process intensity estimation. In: Proceedings of the 25th International Conference on Machine Learning, pp. 192–199. ACM (2008)
5. Dassios, A., Zhao, H., et al.: Exact simulation of Hawkes process with exponentially decaying intensity. Electron. Commun. Probab. **18**, 1–13 (2013)
6. Doersch, C.: Tutorial on variational autoencoders. arXiv preprint arXiv:1606.05908 (2016)
7. Du, N., Dai, H., Trivedi, R., Upadhyay, U., Gomez-Rodriguez, M., Song, L.: Recurrent marked temporal point processes: embedding event history to vector. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1555–1564. ACM (2016)
8. Embrechts, P., Liniger, T., Lin, L.: Multivariate Hawkes processes: an application to financial data. J. Appl. Probab. **48**(A), 367–378 (2011)
9. Gregory, P., Loredo, T.J.: A new method for the detection of a periodic signal of unknown shape and period. Astrophys. J. **398**, 146–168 (1992)
10. Hawkes, A.G.: Spectra of some self-exciting and mutually exciting point processes. Biometrika **58**(1), 83–90 (1971)
11. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
12. Lee, Y., Lim, K.W., Ong, C.S.: Hawkes processes with stochastic excitations. In: International Conference on Machine Learning, pp. 79–88 (2016)
13. Menon, A.K., Lee, Y.: Predicting short-term public transport demand via inhomogeneous Poisson processes. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp. 2207–2210. ACM (2017)
14. 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)
15. Møller, J., Syversveen, A.R., Waagepetersen, R.P.: Log gaussian cox processes. Scand. J. Stat. **25**(3), 451–482 (1998)
16. Ogata, Y., Vere-Jones, D.: Inference for earthquake models: a self-correcting model. Stoch. Process. Appl. **17**(2), 337–347 (1984)
17. Ogata, Y.: On lewis' simulation method for point processes. IEEE Trans. Inf. Theory **27**(1), 23–31 (1981)
18. Rodriguez, M.G., Balduzzi, D., Schölkopf, B.: Uncovering the temporal dynamics of diffusion networks. arXiv preprint arXiv:1105.0697 (2011)