A Refined MISD Algorithm Based on Gaussian Process Regression

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

[1] University of New South Wales, Sydney, NSW 2052, Australia
[2] Data61, CSIRO, 13 Garden St., Eveleigh, NSW, Australia
feng.zhou@data61.csiro.au

Abstract. Time series data is a common data type in real life, and modelling of time series data along with its underlying temporal dynamics is always a challenging job. Temporal point process is an outstanding method to model time series data in domains that require temporal continuity, and includes homogeneous Poisson process, inhomogeneous Poisson process and Hawkes process. We focus on Hawkes process which can explain self-exciting phenomena in many real applications. In classical Hawkes process, the triggering kernel is always assumed to be an exponential decay function, which is inappropriate for some scenarios, so nonparametric methods have been used to deal with this problem, such as model independent stochastic de-clustering (MISD) algorithm. However, MISD algorithm has a strong dependence on the number of bins, which leads to underfitting for some bins and overfitting for others, so the choice of bin number is a critical step. In this paper, we innovatively embed a Gaussian process regression into the iterations of MISD to make this algorithm less sensitive to the choice of bin number.

Keywords: Hawkes process · MISD · Gaussian process
Nonparametric

1 Introduction

In a real application, data is always collected in sequential mode. How to model time series data to discover the underlying temporal dynamics is a challenging problem in this domain. To solve it, different models have been proposed in the past such as recurrent neural network (RNN) [1] and temporal point process [2]. There are many variants of the latter, such as homogeneous Poisson process [3], inhomogeneous Poisson process [4] and Hawkes process [5].

Hawkes process is a self-exciting temporal point process which can explain the self-exciting phenomenon in time series data. In real applications, the occurrence of events in the past will usually have a triggering influence on the future which leads to a clustering effect, for example, in the earthquake domain [6], the crime

D. Phung et al. (Eds.): PAKDD 2018, LNAI 10938, pp. 584–596, 2018.
https://doi.org/10.1007/978-3-319-93037-4_46

domain [7] and the social network domain [8]. In classical Hawkes process, the conditional intensity function can be expressed as:

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

where $\mu > 0$ is the baseline intensity which is a constant, $\{t_i\}$ are the timestamps of observed events before time t indexed by i, and $\gamma(\cdot)$ is the triggering kernel representing the influence from t_i to t. Generally, the triggering kernel $\gamma(t - t_i)$ is always assumed to be an exponential decay function: $\alpha \cdot \exp(-\beta(t-t_i))$, which is inadequate to represent the actual influence in scenarios where it is not like that. Furthermore, in some new fields, there could be lack of prior knowledge about the form of $\gamma(t - t_i)$ or there is no analytic form to describe it [9,10]. In this case, nonparametric methods can be used to estimate the general form of the triggering kernel and the baseline intensity.

An expectation-maximization (EM) algorithm called model independent stochastic de-clustering were proposed to perform nonparametric estimation of the triggering kernel and baseline intensity [11]. Essentially MISD is a histogram density estimator, so there are problems with it: the triggering kernel obtained from MISD is a discrete function and the number of bins used in the model has a vital impact on learning results. It can be seen from the experiments in this paper that the learned triggering kernel is underfitting when fewer bins are used and overfitting when using more. How to determine the optimal number of bins? We can compute the log-likelihood conditioned on bin number M: $\log \mathcal{L}(\{t_i\}|M)$ and compute $\hat{M}$ from maximum likelihood estimation (MLE), or from an un-normalized posterior distribution by multiplying the likelihood with a prior distribution on M such as Poisson distribution[1]. But both these methods will lead to extra computation which is undesirable. Can we propose a refined MISD algorithm which does not depend on the choice of bin number severely? In this paper we innovatively embed a Gaussian process (GP) regression into the iterations of MISD to design a refined algorithm which is less sensitive to the choice of bin number; we call it GP-MISD. In this new method, M can be set to a large number to use over-segmented bins since it can prevent the learning result from overfitting to some extent.

The remainder of the paper is organized as follows: In Sect. 2, we summarize the related work in Hawkes process and its nonparametric estimations. In Sect. 3, we describe the background knowledge about Hawkes process, MISD algorithm and Gaussian process regression and propose our new algorithm GP-MISD. Synthetic data and real data experiments and the detailed discussion are provided in Sects. 4, and 5 concludes this paper.

2 Related Work

Temporal point process has been used as a continuous mathematical model to reflect temporal dynamics and to predict the arrived time of the next event in

[1] We assume all the bins are equally wide.

many domains such as seismology [12], financial engineering [13], and stock market [14]. Recently, the self-exciting process has become a hot topic for explaining the clustering phenomenon in social networks [15] and crime. The classical self-exciting processes, such as Hawkes process, have a limitation that the latent triggering effect is always assumed to be parametric, which introduces computational convenience but limits the expressive ability of the model. To conquer this problem, various nonparametric methods have been proposed, such as considering the triggering kernel as a linear combination of some kernels [16,17], approximating the triggering kernel by an RNN [18] and empirically estimating the triggering kernel using a histogram density estimator (MISD) where the resolution can be adapted by setting different number of bins for the histogram [19]. Although maximum penalized likelihood estimation (MPLE) has been proposed [19], which is a regularized MISD with an l_2 norm on the gradient to avoid overfitting, the gradient information can only regularize the local variance which limits the use of this method. Based on MISD, the GP-MISD algorithm we propose can produce a continuous triggering kernel function which introduces dependence on all the locations on the triggering kernel. As a result, the method is less likely to be overfitting when the bin number is chosen improperly.

3 Proposed Model

The GP-MISD algorithm is closely related to Hawkes process, MISD and Gaussian process regression, so in Sects. 3.1 and 3.2 the preliminary knowledge about these is provided. Most of the details about MISD are draw from [19]. GP-MISD is formally described in Sect. 3.3.

3.1 Hawkes Process

Temporal point process is a stochastic process, whose realization is a sequence of timestamps $\{t_i\}_{i=1}^N$ in $[0, T]$ where t_i is the occurrence time of i-th event and T is the observation time for the process. In temporal point process, an important characterization is the conditional intensity function $\lambda(t)$ which is defined as:

$$\lambda(t) = \lim_{\delta t \to 0} \frac{P(event\ occurring\ in\ [t, t+\delta t)|\mathcal{H}_t)}{\delta t} \tag{2}$$

where $\mathcal{H}_t = \{t_i | t_i < t\}$ is the history before time t. Different temporal point processes will have different conditional intensity functions to distinguish them. For example, $\lambda(t)$ is a constant for homogeneous Poisson process, a function of time $f(t)$ for inhomogeneous Poisson process, and a function of time and history for Hawkes process. The specific intensity form of Hawkes process is already given in (1). The summation of triggering kernels explains the nature of self-excitation, which is the occurrence of events in the past will intensify events occurring in the future. Given a sequence of observed data $\{t_i\}_{i=1}^n$ in time interval $[0, T]$, the log-likelihood of this list of event times can be expressed as:

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

which can be used in MLE to perform inference for the parameters in the model.

3.2 MISD

Lewis and Mohler [19] provide details on how to use MISD algorithm in one dimension Hawkes, which is an EM-based nonparametric algorithm to ease MLE. Firstly, when the branching structure of a Hawkes process is observable, we can define the following matrix:

$$\begin{aligned} \mathcal{X}_{ij} &= \begin{cases} 1 & \textit{if event } i \textit{ is caused by event } j \\ 0 & \textit{otherwise} \end{cases} \\ \mathcal{X}_{ii} &= \begin{cases} 1 & \textit{if event } i \textit{ is a baseline event} \\ 0 & \textit{otherwise.} \end{cases} \end{aligned} \tag{4}$$

Let us assume baseline intensity μ is a constant and there is no prior knowledge about the form of $\gamma(\cdot)$, so given the branching matrix, the log-likelihood (3) could be decoupled into two independent parts: part μ and part $\gamma(\cdot)$,

$$\begin{aligned} \log \mathcal{L}(\{t_i\}|\mu, \gamma) = &\left[\sum_{i=1}^{n} \mathcal{X}_{ii} \log(\mu)\right] - \mu T \\ &+ \sum_{i=2}^{n} \left[\sum_{j=1}^{i-1} \mathcal{X}_{ij} \log\left(\gamma(t_i - t_j)\right)\right] - \sum_{i=1}^{n} \int_{t_i}^{T} \gamma(t - t_i)dt. \end{aligned} \tag{5}$$

It is straightforward to rewrite this problem into an EM framework, which is the MISD algorithm. When the branching structure is unobservable, the MISD algorithm works by maximizing the expectation of the log-likelihood. Thus $\mathcal{X}_{ij}$ is replaced by p_{ij}, which is the probability of event i caused by event j. The matrix p_{ij} is a lower triangular matrix

$$\begin{bmatrix} p_{11} & & & & \\ p_{21} & p_{22} & & & \\ p_{31} & p_{32} & p_{33} & & \\ & \vdots & & \ddots & \\ p_{n1} & p_{n2} & p_{n3} & \cdots & p_{nn} \end{bmatrix} \tag{6}$$

where $\sum_{j=1}^{i} p_{ij} = 1$, because event i must be caused by previous events or the baseline event.

Then the EM iteration is:

(1) E-step: The update for the matrix P:

$$\begin{aligned} p_{ij}^s &= \frac{\gamma^s(t_i - t_j)}{\mu^s + \sum_{j=1}^{i-1} \gamma^s(t_i - t_j)} \\ p_{ii}^s &= \frac{\mu^s}{\mu^s + \sum_{j=1}^{i-1} \gamma^s(t_i - t_j)} \end{aligned} \tag{7}$$

where s is the iteration step.

(2) M-step: The update for baseline intensity:

$$\mu^{s+1} = \frac{1}{T} \sum_{i=1}^{n} p_{ii}^s \tag{8}$$

where T is the observation duration.

Assuming the duration of $\gamma(\varDelta t)$ is limited: $[0, M\delta t]$ where M is the number of bins, δt is the bin width, the update for rates is given by:

$$\gamma_m^{s+1} = \frac{1}{N_m \delta t} \sum_{i,j \in A_m} p_{ij}^s \tag{9}$$

where A_m is the set of pairs of events s.t. $m\delta t \leqslant |t_i - t_j| \leqslant (m+1)\delta t$, $\gamma_m = \gamma(m\delta t)$ where $0 \leqslant m \leqslant M-1$, and N_m is the corresponding normalizing parameter with respect to m-th bin. Equations (8) and (9) are derived from $\frac{\partial}{\partial \mu}\mathbb{E}[\log \mathcal{L}] = 0$ and $\frac{\partial}{\partial \gamma_m}\mathbb{E}[\log \mathcal{L}] = 0$.

3.3 GP-MISD

The key idea in GP-MISD is to embed a Gaussian process regression into the EM iterations, which makes use of those rates learned in each iteration step to perform a regression and get a smooth mean triggering kernel. This smooth mean triggering kernel will be used in the next iteration step, so the iteration goes on.

Gaussian process is an infinite dimensional extension of multivariate normal distribution. In GP, every finite set of points has a multivariate normal distribution, so it can be expressed as a distribution over functions in a continuous domain. GP is specified by the mean function $m(x)$ and covariance kernel $k(x, x')$:

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')) \tag{10}$$

where $f(x)$ is a sample function drawn from GP. Without loss of generality, the prior mean function can be assumed to be zero: $m(x) = 0$, and the only work left is to define the covariance kernel $k(x, x')$. A widely used kernel is squared exponential kernel:

$$k(x_i, x_j) = \theta_0 \exp\left(-\frac{\theta_1}{2}\|x_i - x_j\|^2\right) \tag{11}$$

where θ_0, θ_1 are the hyperparameters.

After getting the observation points $(\gamma_1^s, \gamma_2^s, \cdots, \gamma_M^s)$ in iteration step s in MISD, the GP regression is used to evaluate the posterior mean function $m(x|(\gamma_1^s, \cdots, \gamma_M^s))$ which will be used as the $\gamma(\Delta t)$ in the next iteration step. Specifically, the new algorithm can be divided into three steps:

(1) E-step: The update for the matrix P:

$$\begin{aligned} p_{ij}^s &= \frac{\bar{\gamma}^s(t_i - t_j)}{\mu^s + \sum_{j=1}^{i-1} \bar{\gamma}^s(t_i - t_j)} \\ p_{ii}^s &= \frac{\mu^s}{\mu^s + \sum_{j=1}^{i-1} \bar{\gamma}^s(t_i - t_j)} \end{aligned} \tag{12}$$

(2) M-step: The update for baseline intensity and rates is same as before.
(3) GP-step: The update for Gaussian process predictive distribution:

$$\bar{\gamma}^{s+1}(\Delta t) = \boldsymbol{k}^T \boldsymbol{C}_M^{-1} \boldsymbol{\gamma}^{s+1} \tag{13}$$

where $\boldsymbol{C}_M$ is the matrix of $C(\Delta t_n, \Delta t_m) = k(\Delta t_n, \Delta t_m) + \sigma_{noise}^2 \delta_{nm}$, $\{\Delta t_i\}_{i=1}^M$ are the x-values of M rate points, $k(\cdot)$ is the covariance kernel, and σ_{noise}^2 is the variance of observation points' noise, $\boldsymbol{k} = (k(\Delta t_1, \Delta t), k(\Delta t_2, \Delta t), \cdots, k(\Delta t_M, \Delta t))^T$, $\boldsymbol{\gamma}^{s+1} = (\gamma_1^{s+1}, \gamma_2^{s+1}, \cdots, \gamma_M^{s+1})^T$ are the y-values of M rate points on step $s+1$. The final triggering kernel we obtain from this algorithm is $\bar{\gamma}(\Delta t)$. Equation (13) is derived from the standard Gaussian process regression [20].

4 Experiment

4.1 Synthetic Data

For simplicity, we assume the true triggering kernel is an exponential decay function: $\mu = 1$, $\gamma(t - t_i) = 1 \cdot \exp(-2 \cdot (t - t_i))$. Two sets of synthetic data are generated from the Hawkes process specified above using the thinning algorithm [12]. For each set, the observation duration T is set to 400, resulting in a realization of about 850 events. The first set is used as the training data, and the second one is the test data.

For the inference, it is assumed that the baseline intensity is a constant and the form of the triggering kernel is unknown, so the goal is to infer μ and $\gamma(\Delta t)$. For MISD algorithm, we apply the training data for different bin numbers ranging from 3 to 100. $\gamma(\Delta t)$ is assumed to be zero outside the interval $[0, 3]$ and the number of iterations is set to 100. In the evaluation, the training error is defined as $-\log \mathcal{L}$ of the training data. Then the model learned is applied to the test data to get the test error which is defined as $-\log \mathcal{L}$ of the test data. The same process is also applied to the GP-MISD algorithm. The hyperparameters θ_0, θ_1, σ^2_{noise} are set to 2.3, 2.3 and 0.01 in the GP step.

The training error and test error for both algorithms appear in Fig. 1. It can be seen that as the number of bins increases from 3 to 100, the training error of MISD will decrease monotonically, while the test error will increase after #bin = 8. But when we look at GP-MISD, the training error will not decrease rapidly after #bin = 8 and the test error is almost constant after #bin = 8. These results show that GP-MISD is less sensitive to the choice of bin number than MISD which is very likely to be overfitting when too many bins are used. More importantly, from test error we can see that GP-MISD is always superior to MISD no matter how many bins are used, and this can also be found from the fitting result of $\gamma(\Delta t)$ in Fig. 2 which is based on #bin = 10, 40 and 100. It is clear that the $\gamma(\Delta t)$ learned from GP-MISD is closer to the ground truth and more stable, which shows the superiority of GP-MISD.

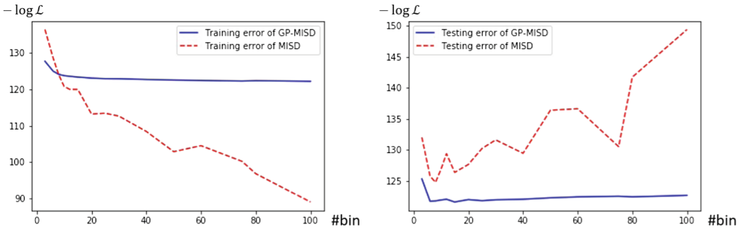


Fig. 1. The training error and test error of MISD and GP-MISD.

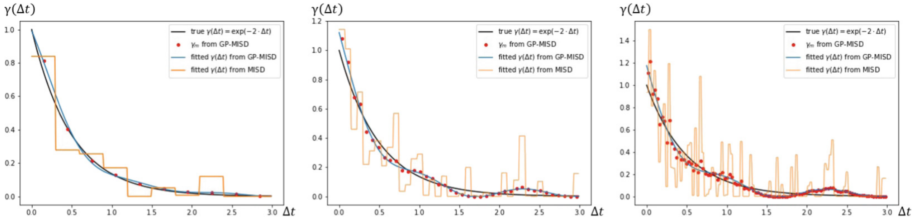


Fig. 2. The fitting result of $\gamma(\Delta t)$ from MISD and GP-MISD based on 10 bins (left), 40 bins (middle) and 100 bins (right).

4.2 Real Data

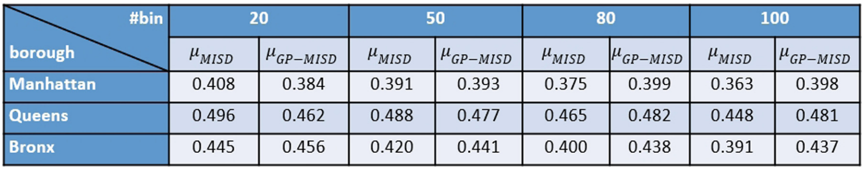
We evaluate the performance of GP-MISD and MISD on real world datasets from two different domains.

Motor Vehicle Collisions in New York City: This motor vehicle collision dataset[2] is provided by the New York City Police Department (NYPD). It contains about 1.05 million vehicle collision records in New York City from July, 2012 to September, 2017. The dataset includes the collision date, time, borough, location, contributing factor and so on. For our model, the most valuable information is the date and time. We filter out the collision records in Manhattan, Queens and Bronx caused by 'Alcohol Involvement'. For each borough, half of the records are used as the training data and the other half as the test data. Just as the synthetic data, we define the test error as $-\log \mathcal{L}$ of the test data. There are some collisions happening at the same time, as the resolution is at minute level, which violates the definition of the temporal point process. To avoid this, we add a small time interval to all the simultaneous records to separate them. The hyperparameters θ_0, θ_1, σ^2_{noise} are set to 3.5, 3.5, 0.01 for Manhattan, 4.5, 4.5, 0.01 for Queens and 3.9, 3.9, 0.01 for Bronx. 100 iterations are performed in both algorithms. The duration of $\gamma(\Delta t)$ is set to 3.0 and the time unit is 1.16 day.

NYPD Complaint Data 2017: This dataset[3] includes all valid felony, misdemeanour and violation crimes reported to the NYPD for all complete quarters so far in 2017. It includes 228 thousand complaint records in New York City. The columns are complaint number, date, time, offense description, Borough etc. We filter out the complaints in Manhattan, Queens and Brooklyn, and the offense description is 'THEFT-FRAUD'. Again, for each borough, half the records are used as training data and the others as test data. Add a small time interval to separate all the simultaneous records. The hyperparameters θ_0, θ_1, σ^2_{noise} are set to 6.45, 6.45, 0.01 for all boroughs. 100 iterations are performed in both algorithms. The duration of $\gamma(\Delta t)$ is set to 3.0 and the time unit is 11.6 days.

Experiment Results: For Motor Vehicle Collisions in New York City, the learned μ, $\gamma(\Delta t)$ and the test errors of both algorithms for #bin = 20, 50, 80, 100 are shown in Table 1 and Fig. 3.

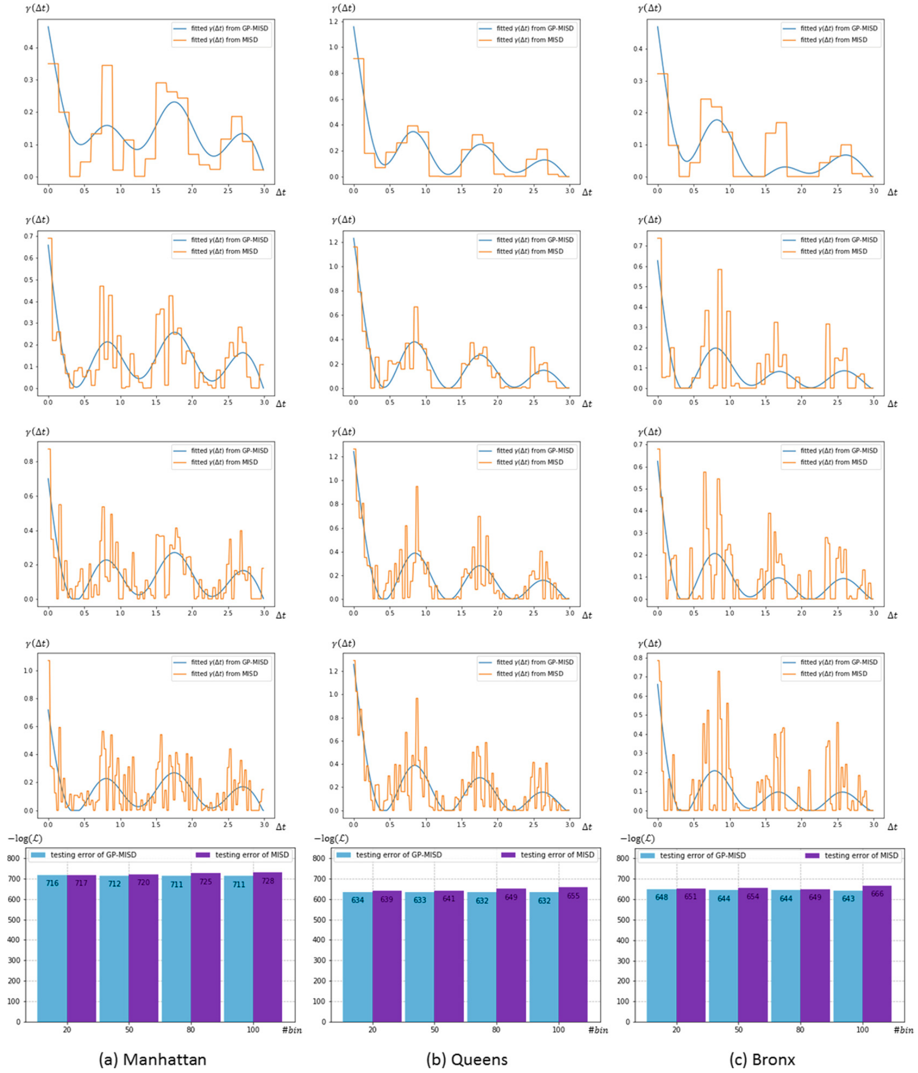
Table 1. Motor Vehicle Collisions in New York City: the learned baseline intensity μ from MISD and GP-MISD based on #bin = 20, 50, 80, 100.



#bin borough	20		50		80		100	
	μ_{MISD}	$\mu_{GP-MISD}$	μ_{MISD}	$\mu_{GP-MISD}$	μ_{MISD}	$\mu_{GP-MISD}$	μ_{MISD}	$\mu_{GP-MISD}$
Manhattan	0.408	0.384	0.391	0.393	0.375	0.399	0.363	0.398
Queens	0.496	0.462	0.488	0.477	0.465	0.482	0.448	0.481
Bronx	0.445	0.456	0.420	0.441	0.400	0.438	0.391	0.437

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

[3] https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Current-YTD/5uac-w243.

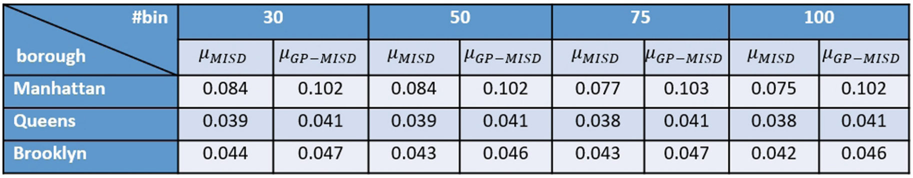


(a) Manhattan (b) Queens (c) Bronx

Fig. 3. Motor Vehicle Collisions in New York City: the learned $\gamma(\varDelta t)$ from MISD and GP-MISD based on #bin = 20, 50, 80, 100 (upper, time unit is 1.16 day), and test errors of both algorithms for #bin = 20, 50, 80, 100 (lower).

For NYPD Complaint Data 2017, the learned μ, $\gamma(\varDelta t)$ and the test errors of both algorithms for #bin = 30, 50, 75, 100 are shown in Table 2 and Fig. 4.

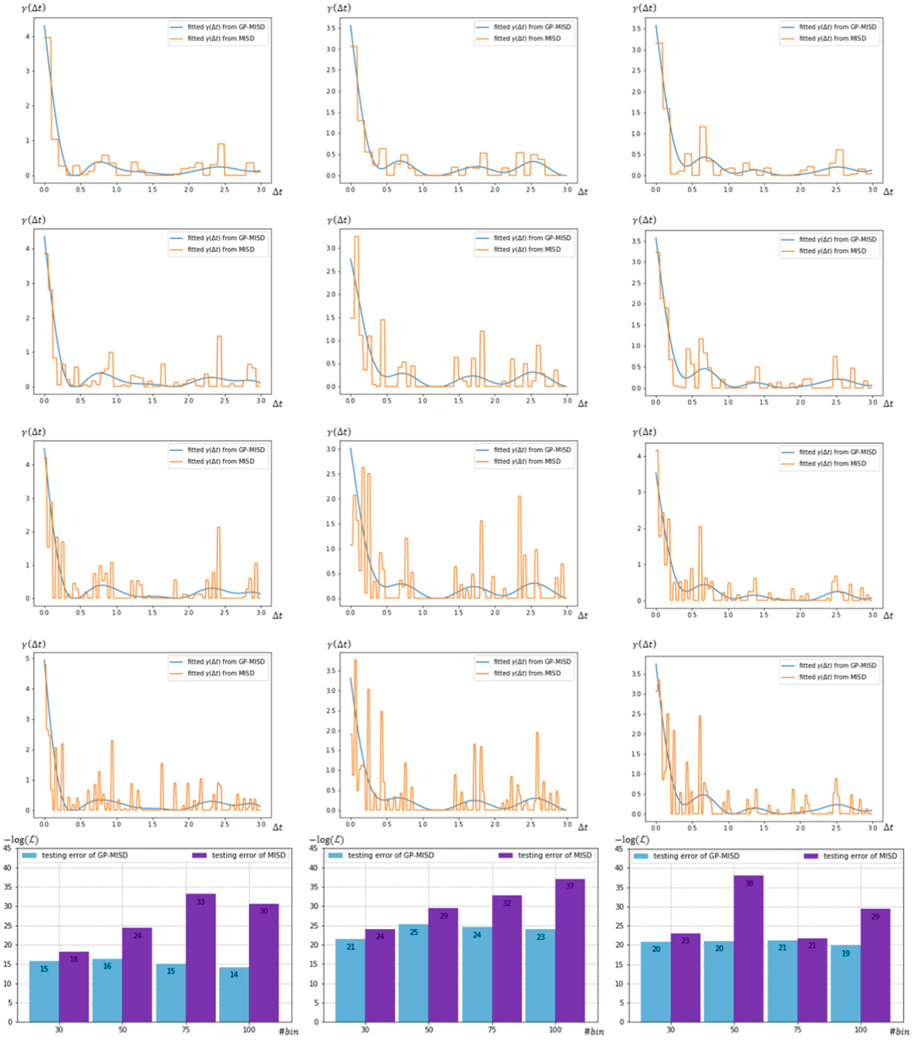
Table 2. NYPD Complaint Data 2017: the learned baseline intensity μ from MISD and GP-MISD based on #bin = 30, 50, 75, 100.



#bin	30		50		75		100	
borough	μ_{MISD}	$\mu_{GP-MISD}$	μ_{MISD}	$\mu_{GP-MISD}$	μ_{MISD}	$\mu_{GP-MISD}$	μ_{MISD}	$\mu_{GP-MISD}$
Manhattan	0.084	0.102	0.084	0.102	0.077	0.103	0.075	0.102
Queens	0.039	0.041	0.039	0.041	0.038	0.041	0.038	0.041
Brooklyn	0.044	0.047	0.043	0.046	0.043	0.047	0.042	0.046

From both experimental results, we can see that $\gamma(\Delta t)$ from GP-MISD is smoother and more stable than that from MISD and the test error of GP-MISD is always lower than MISD, which is consistent with the synthetic data result: the former effectively avoids the overfitting phenomenon and makes this algorithm less sensitive to the choice of #bin. For vehicle collision, the triggering patterns in different boroughs are similar and the triggering effect lasts for about 4.5 days; for crime complaint, the triggering patterns in different boroughs are similar and the triggering effect lasts for almost one month, but significant in the first 10 days. Moreover, we can see that the trend of triggering kernel is quite dynamic, especially in the short period after the source event happened, e.g., within about 0.5 day after the initial collision in Fig. 3, or about 5 days after the initial complaint in Fig. 4. To capture the trend, the #bin must be set to be large enough so that the resolution is high, however, too large a #bin will cause overfitting, such as spikes in the triggering kernel. This is the advantage of GP-MISD to represent the triggering kernel with continuity, capturing any dynamic trends without overfitting.

Setting hyperparameters θ_0 and θ_1 is also a key step in all GP-based methods. The hyperparameters used to determine the GP kernel values implicitly encode the information on how flexible the GP could be. The optimization of hyperparameters in GP has been proved to be a non-convex problem [20], which may introduce some difficulty in learning hyperparameters. In our experiments, we use grid search to find the optimal hyperparameters and find that setting the hyperparameters in a reasonable range does not severely affect the final result.



(a) Manhattan (b) Queens (c) Brooklyn

Fig. 4. NYPD Complaint Data 2017: the learned $\gamma(\Delta t)$ from MISD and GP-MISD based on #bin = 30, 50, 75, 100 (upper, time unit is 11.6 days), and test errors of both algorithms for #bin = 30, 50, 75, 100 (lower).

5 Conclusion

To conclude, in this paper we propose a refined MISD algorithm for Hawkes process: GP-MISD algorithm which can effectively avoid overfitting when more bins are used. The key thought of embedding a Gaussian process regression into the EM iterations actually can be applied to most algorithms based on bins, resulting in a smooth effect to avoid overfitting. GP-MISD inherits the advantage from MISD to predict the baseline intensity and triggering kernel without any prior knowledge of the function form of latent triggering kernel. We have performed experiments on both synthetic and real datasets demonstrating that GP-MISD is less sensitive to the choice of #bin and has consistent superiority to MISD.

References

1. Mikolov, T., Karafit, M., Burget, L., Cernock, J., Khudanpur, S.: Recurrent neural network based language model. In: Interspeech, vol. 2, p. 3 (2010)
2. Schoenberg, F.P., Brillinger, D.R., Guttorp, P.: Point processes, spatialtemporal. In: Encyclopedia of Environmetrics (2002)
3. Thompson Jr., W.A.: Homogeneous Poisson processes. In: Point Process Models with Applications to Safety and Reliability, pp. 21–31. Springer, Boston (1988). https://doi.org/10.1007/978-1-4613-1067-9_3
4. Weinberg, J., Brown, L.D., Stroud, J.R.: Bayesian forecasting of an inhomogeneous Poisson process with applications to call center data. J. Am. Stat. Assoc. **102**(480), 1185–1198 (2007)
5. Hawkes, A.G.: Spectra of some self-exciting and mutually exciting point processes. Biometrika **58**(1), 83–90 (1971)
6. Vere-Jones, D.: Stochastic models for earthquake occurrence. J. Roy. Stat. Soc. Ser. B (Methodological) **32**, 1–62 (1970)
7. Short, M.B., Mohler, G.O., Brantingham, P.J., Tita, G.E.: Gang rivalry dynamics via coupled point process networks. Discret. Contin. Dyn. Syst. Ser. B **19**(5), 1459–1477 (2014)
8. Mitchell, L., Cates, M.E.: Hawkes process as a model of social interactions: a view on video dynamics. J. Phys. Math. Theor. **43**(4), 045101 (2009)
9. 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)
10. Lewis, E., Mohler, G., Brantingham, P.J., Bertozzi, A.L.: Self-exciting point process models of civilian deaths in Iraq. Secur. J. **25**(3), 244–264 (2012)
11. Marsan, D., Lengline, O.: Extending earthquakes reach through cascading. Science **319**(5866), 1076–1079 (2008)
12. Ogata, Y.: Space-time point-process models for earthquake occurrences. Ann. Inst. Stat. Math. **50**(2), 379–402 (1998)
13. Bacry, E., Jaisson, T., Muzy, J.F.: Estimation of slowly decreasing Hawkes kernels: application to high-frequency order book dynamics. Quant. Financ. **16**(8), 1179–1201 (2016)
14. Hardiman, S., Bercot, N., Bouchaud, J.P.: Critical reflexivity in financial markets: a Hawkes process analysis. Eur. Phys. J. B **86**, 442 (2013)
15. Zhou, K., Zha, H., Song, L.: Learning social infectivity in sparse low-rank networks using multi-dimensional Hawkes processes. In: Artificial Intelligence and Statistics, pp. 641–649 (2013)
16. Zhou, K., Zha, H., Song, L.: Learning triggering kernels for multi-dimensional Hawkes processes. In: Proceedings of the 30th International Conference on Machine Learning, pp. 1301–1309 (2013)
17. Du, N., Song, L., Yuan, M., Smola, A.J.: Learning networks of heterogeneous influence. In: Advances in Neural Information Processing Systems, pp. 2780–2788 (2012)

18. 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 (2016)
19. Lewis, E., Mohler, G.: A nonparametric EM algorithm for multiscale Hawkes processes. J. Nonparametric Stat. **1**(1), 1–20 (2011)
20. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, New York (2006)