Generalized Hawkes Process: Nonparametric and Nonstationary

Submitted by

Feng Zhou

for the degree of

Doctor of Philosophy

School of Computer Science and Engineering

The University of New South Wales

September 2019

Thesis/Dissertation Sheet

Surname/Family Name	:	**Zhou**
Given Name/s	:	**Feng**
Abbreviation for degree as give in the University calendar	:	**PhD**
Faculty	:	**Engineering**
School	:	**Computer Science and Engineering**
Thesis Title	:	**Generalized Hawkes Process: Nonparametric and Nonstationary**

Abstract 350 words maximum: (PLEASE TYPE)

Point process is a common statistical model used to describe the pattern of event occurrence for many real-world applications, such as earthquake prediction and financial modelling. The prompting characteristics of past events on future ones are a vital factor in the clustering effects in point processes. Hawkes process is the most extensively used point process model for modelling self-exciting phenomena.

One of the key challenges for the Hawkes process is the selection of the function for modelling baseline intensity and the triggering kernel. The vanilla Hawkes process assumes a constant valued function for the baseline intensity and a parametric stationary function for the triggering kernel. The parametric and stationary assumption makes inference convenient but limits the model expression.

To generalize the classical Hawkes process, various nonparametric and nonstationary approaches for the Hawkes process are proposed in this thesis. Specifically, three different nonparametric and nonstationary approaches for Hawkes processes are proposed.

The model independent stochastic declustering (MISD) algorithm is a classical frequentist nonparametric inference algorithm for Hawkes process with triggering kernel and it uses a bin-based histogram function. However, the number of bins, which is fixed manually, usually leads to underfitting or overfitting when improperly chosen. In this thesis, a refined MISD algorithm is proposed to ease the choice of bin number.

Next, a Bayesian nonparametric Hawkes process model is proposed, with Gaussian process as prior for baseline intensity and triggering kernel. Correspondingly, a variational Gaussian approximation, Polya-Gamma based Gibbs sampling, expectation-maximization (EM) and mean-field variational inference algorithms are proposed.

As the next step, nonstationarity is introduced into the classical Hawkes process. A fast cumulant-based multi-resolution segmentation algorithm is proposed that partitions the process into segments to capture the time-varying characteristics.

Finally, future research directions for the nonparametric and nonstationary Hawkes process are discussed.

Declaration relating to disposition of project thesis/dissertation

03/12/2019

……………………………………………………………… ……………………………………………

Signature Date

ORIGINALITY STATEMENT

‘I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.’

Signed ……………………………………………….............

Date 03/12/2019 ……………………………………………….............

COPYRIGHT STATEMENT

Signed ……………………………………………...........................

Date …………10/03/2020……………………………………..............................

AUTHENTICITY STATEMENT

Signed ……………………………………………...........................

Date …………10/03/2020……………………………………..............................

INCLUSION OF PUBLICATIONS STATEMENT

UNSW is supportive of candidates publishing their research results during their candidature as detailed in the UNSW Thesis Examination Procedure.

Publications can be used in their thesis in lieu of a Chapter if:

- The student contributed greater than 50% of the content in the publication and is the "primary author", ie. the student was responsible primarily for the planning, execution and preparation of the work for publication
- The student has approval to include the publication in their thesis in lieu of a Chapter from their supervisor and Postgraduate Coordinator.
- The publication is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in the thesis

Please indicate whether this thesis contains published material or not.

□ *This thesis contains no publications, either published or submitted for publication (if this box is checked, you may delete all the material on page 2)*

☒ *Some of the work described in this thesis has been published and it has been documented in the relevant Chapters with acknowledgement (if this box is checked, you may delete all the material on page 2)*

□ *This thesis has publications (either published or submitted for publication) incorporated into it in lieu of a chapter and the details are presented below*

CANDIDATE'S DECLARATION

I declare that:

- I have complied with the Thesis Examination Procedure
- where I have used a publication in lieu of a Chapter, the listed publication(s) below meet(s) the requirements to be included in the thesis.

Name **Feng Zhou**	**Signature**	**Date (dd/mm/yy)** **03/12/2019**

Postgraduate Coordinator's Declaration (to be filled in where publications are used in lieu of Chapters)

I declare that:

- the information below is accurate
- where listed publication(s) have been used in lieu of Chapter(s), their use complies with the Thesis Examination Procedure
- the minimum requirements for the format of the thesis have been met.

PGC's Name	**PGC's Signature**	**Date (dd/mm/yy)**

To the people who love me and those I love.

Acknowledgments

I would like to express my gratitude to my main supervisors in University of Technology Sydney and Data61 CSIRO including Fang Chen, Yang Wang, Zhidong Li and Chen Cai. All of them give me insightful guidance, valuable suggestions and financial support. I cannot achieve such a smooth completion of degree without your support.

Also, I am very appreciative to my supervisor Arcot Sowmya in University of New South Wales. Professor Arcot Sowmya is an extremely responsible supervisor who treats students very well.

I would also like to thank my cooperator Xuhui Fan in the University of New South Wales who provides me promising ideas and proof reading.

At last, I would like to thank my families for their care and encouragement when I encounter difficulties in the course of my research. Thank you for your company along the way.

Abstract

Point process is a common statistical model used to describe the pattern of event occurrence for many real-world applications, such as earthquake prediction and financial modelling. The prompting characteristics of past events on future ones are a vital factor in the clustering effects in point processes. Hawkes process is the most extensively used point process model for modelling self-exciting phenomena.

One of the key challenges for the Hawkes process is the selection of the function for modelling baseline intensity and the triggering kernel. The vanilla Hawkes process assumes a constant valued function for the baseline intensity and a parametric stationary function for the triggering kernel. The parametric and stationary assumption makes inference convenient but limits the model expression.

To generalize the classical Hawkes process, various nonparametric and nonstationary approaches for the Hawkes process are proposed in this thesis. Specifically, three different nonparametric and nonstationary approaches for Hawkes processes are proposed.

The model independent stochastic declustering (MISD) algorithm is a classical frequentist nonparametric inference algorithm for Hawkes process with triggering kernel and it uses a bin-based histogram function. However, the number of bins, which is fixed manually, usually leads to underfitting or overfitting when improperly chosen. In this thesis, a refined MISD algorithm is proposed to ease the choice of bin number.

Next, a Bayesian nonparametric Hawkes process model is proposed, with Gaussian process as prior for baseline intensity and triggering kernel. Correspondingly, a variational Gaussian approximation, Polya-Gamma based Gibbs sampling, expectation-maximization (EM) and mean-field variational inference algorithms are proposed.

As the next step, nonstationarity is introduced into the classical Hawkes process. A fast cumulant-based multi-resolution segmentation algorithm is proposed that partitions the process into segments to capture the time-varying characteristics.

Finally, future research directions for the nonparametric and nonstationary Hawkes process are discussed.

Contents

List of Figures

List of Tables

List of Abbreviations

ABBREVIATIONS	**FULL EXPERSSIONS**
MISD	Model independent stochastic declustering
EM	Expectation maximization
GP	Gaussian process
MCMC	Markov chain Monte Carlo
MLE	Maximum likelihood estimation
MH	Metropolis-Hasting
KL-divergence	Kullback-Leibler divergence
ELBO	Evidence lower bound
RNN	Recurrent neural network
NYPD	New York City Police Department
QGPHP	Quadratic GP Hawkes process
EMV	EM-variational
MAP	Maximum a posteriori
SGPHP	Sigmoid GP Hawkes process
RKHS	Reproducing kernel Hilbert space
ELL	Expectation of log-likelihood
MF	Mean field
LSTM	Long short-term memory

MRS	Multi-resolution segmentation
NMSE	Normalized mean squared error
GP-MRS	Gaussian process based multi-resolution segmentation

List of Publications

1. F. Zhou, Z. Li, X. Fan, Y. Wang, A. Sowmya, F. Chen, A Refined MISD Algorithm Based on Gaussian Process Regression, Advances in Knowledge Discovery and Data Mining, Springer International Publishing AG, part of Springer Nature 2018, D. Phung et al. (Eds.): PAKDD 2018, LNAI 10938, pp. 584596, 2018.

2. F. Zhou, Y. Zhang, Z. Li, X. Fan, Y. Wang, A. Sowmya and F. Chen, Hawkes Process with Stochastic Triggering Kernel, Advances in Knowledge Discovery and Data Mining, Springer International Publishing AG, part of Springer Nature 2019, Q. Yang et al. (Eds.): PAKDD 2019, LNAI 11439, pp. 319-330, 2019.

3. F. Zhou, Z. Li, X. Fan, Y. Wang, A. Sowmya, F. Chen, Efficient EM-Variational Inference for Hawkes Process, arXiv preprint arXiv:1905.12251. 29 May 2019. Under review by IJCAI 2020.

4. F. Zhou, Z. Li, X. Fan, Y. Wang, A. Sowmya, F. Chen, Scalable Inference for Nonparametric Hawkes Process Using Pólya-Gamma Augmentation. arXiv preprint arXiv:1910.13052. 29 Oct 2019. Under review by Journal of Machine Learning Research.

5. F. Zhou, Z. Li, X. Fan, Y. Wang, A. Sowmya, F. Chen, Fast Multi-resolution Segmentation for Nonstationary Hawkes Process Using Cumulants. arXiv preprint arXiv:1906.02438. 6 Jun 2019. Plan to submit to CIKM 2020.

Chapter 1

Introduction

Point process is a common statistical model used to describe the pattern of event occurrence for many real world applications, such as earthquake prediction [1] and financial modelling [2]. The influence of past events on future ones is a vital factor in the clustering effects of point processes. Many models have been proposed to describe the interactions in a point process. Among those models, the Hawkes process [3] is an important class of point processes that can be utilized to model the *self-exciting* phenomenon in numerous application domains, including criminology [4], ecosystem modelling [5], transport planning [6] and social networks analysis [7].

An important characteristic of point processes is the conditional intensity: the probability of an event occurring in an infinitesimal time interval given the past history. Specifically, the conditional intensity of Hawkes process may be expressed as:

$$\lambda(t) = \mu(t) + \int_0^t \phi(t-s)d\mathbb{N}(s) = \mu(t) + \sum_{t_i<t} \phi(t-t_i), \tag{1.1}$$

where $\mu(t) > 0$ is the baseline intensity, $\{t_i\}$ are timestamps of events occurring before t, $\mathbb{N}(t)$ is the corresponding counting process and $\phi(\tau) > 0$ where $\tau = t - t_i$ is the triggering kernel. The summation of triggering kernels explains the nature of self-excitation: events occurring in the past intensify the rate of occurrence in the future.

The classic Hawkes process is supposed to have a parametric form: the baseline intensity $\mu(t)$ is assumed to be a constant with triggering kernel $\phi(\tau)$ a parametric function, e.g. exponential decay or power law decay function. However, in reality, the actual exogenous rate $\mu(t)$ can change over time due to the varying external context; the actual endogenous rate capturing how previous events trigger posterior ones, modelled by $\phi(\tau)$, can be rather complex among different applications. For example, the exogenous rate of civilian deaths due to insurgent activity is changing over time [8] and the prompting effect of vehicle collision decays periodically and in an oscillatory way [9]. Models based on the classic Hawkes process tend to be oversimplified or even incapable of capturing the ground truth in numerous scenarios. To address this, it is necessary to estimate the exogenous and endogenous dynamics in a data-driven nonparametric way.

A wide variety of nonparametric estimation approaches for Hawkes process have been investigated over the past few years [8, 10, 11, 12]. Basically, there are two categories: frequentist nonparametric, which is based on likelihood only, and Bayesian nonparametric, which needs to incorporate a prior. From the frequentist nonparametric perspective, it was initially proposed [1] to estimate the triggering kernel modelled as a histogram function using an EM algorithm; this was later extended [8] by introducing a smooth regularizer and estimation was performed by solving an Euler-Lagrange equation. This approach was further extended [10] to the multivariate Hawkes process. An estimation approach based on the solution of a Wiener-Hopf equation, relating the triggering kernel with the second order statistics of its counting process was proposed [11]; other efforts [12, 13] attempted to minimize a quadratic contrast function with a grid based triggering kernel.

From the Bayesian nonparametric perspective, most related works are based on Gaussian-Cox processes: the Poisson process with a stochastic intensity modulated by Gaussian process (GP). To guarantee the non-negativity of intensity, trajectories drawn from a GP prior need to be squashed by a link function. For example, a log-Gaussian intensity was utilized [14, 15]; a sigmoid-GP intensity and a tractable Markov chain Monte Carlo (MCMC) algorithm has also been proposed [16]; a variational Gaussian approximation algorithm with a square link function has been developed [17]. As far as is known, only a few works have attempted to infer the nonparametric Hawkes process from a Bayesian perspective, as the Hawkes process is more complex than the Poisson process. For example, a Bayesian nonparametric estimation method for Hawkes process is available [18], in which the baseline intensity is assumed to be a constant and the prior on the triggering kernel is based on a piecewise constant function or a mixture of Beta distributions.

Another issue with the classic Hawkes process is the stationary assumption. It is straightforward to see that the conditional intensity (Eq. 1.1) of the Hawkes process is unchanged over timeshifting if $\mu(t)$ is constant and $\phi(\cdot)$ only depends on $\tau = t - t_i$ not on t, which defines stationarity. The assumption of stationarity leads to reduced model complexity and straightforward inference. However, the point process data generated in many real applications has non-stationary properties, which means that its first, second and higher order cumulants (moments) are changing over time. Applying the vanilla Hawkes process directly to non-stationary data is inappropriate. On the other hand, non-stationarity itself can be an important feature in some applications. For example, in transportation, the influence of the road condition on car accidents changes between day and night and between busy and non-busy hours (see Fig. 1.1).

One of the common methods of analyzing non-stationarity is the use of segmentation. This type of problem is also called a change-point problem that is studied in mathematics [19]. Given a non-stationary point process data, the segmentation algorithm will divide the whole observation period into several non-overlapping contiguous segments in such a way that each segment is more approximately stationary than the original data and can be assumed to be stationary.

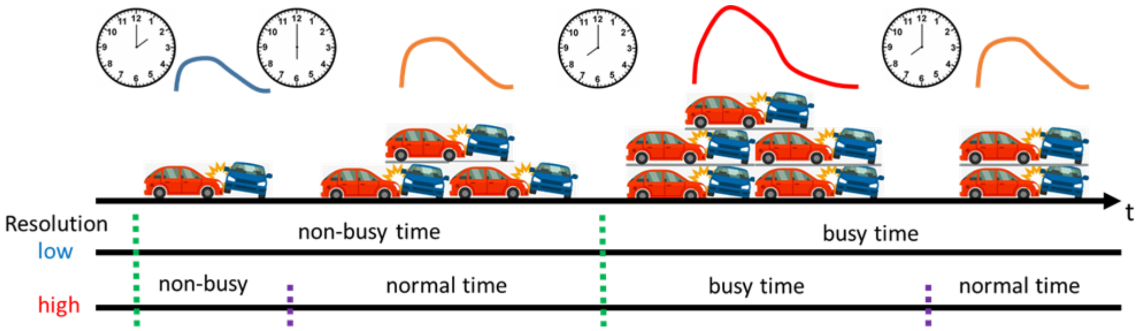


Figure 1.1: Multi-resolution segmentation showing the influence of road condition on car accidents. The influence is low between 2:00-6:00; medium between 6:00 and 8:00, 20:00 and 2:00; high between 8:00 and 20:00. The green dash provides a low-resolution segmentation (2 segments) and together with purple dash provide a high-resolution segmentation (4 segments), providing hierarchical insight into the dynamic time-varying characteristics.

1.1 Contributions

This thesis focuses on the generalization of the Hawkes process to nonparametric and non-stationary scenarios, thereby helping to solve real-world complex problems.

For the nonparametric Hawkes process: from the frequentist perspective, the classic MISD algorithm is modified to ease the choice of hyperparameters; from the Bayesian perspective, GP is utilized to modulate the baseline intensity and triggering kernel such that the two components can be in any form without parametric constraint.

For the non-stationary Hawkes process: the cumulants of the Hawkes process are utilized to propose a fast multi-resolution segmentation algorithm that partitions the process into segments on which stationarity is satisfied.

1.2 Thesis Organization

The rest of the thesis is organized as follows.

In Chapter 2, background knowledge necessary to understand the later chapters is introduced, including Hawkes process, Gaussian process, branching structure and cumulants of Hawkes process, existing parametric and nonparametric estimation methods for Hawkes process, the EM algorithm and Bayesian inference (MCMC and variational inference).

The MISD algorithm [8] is a classic EM algorithm for nonparametric Hawkes process where the triggering kernel is modelled as a histogram function. The number of bins for the histogram function is a hyperparameter for the model that needs to be fixed in advance. A small bin number means an over-simplified model leading to underfitting, while a large number leads to overfitting. In Chapter 3, a Gaussian process regression step is innovatively embedded into the EM iteration. Consequently, the estimation result is stable regardless of the number of bins used in the model.

Unlike the inference approach in Chapter 3, which is based on likelihood only (frequentist method), a Bayesian nonparametric framework for the Hawkes process is proposed in Chapters 4 and 5. In the Bayesian nonparametric framework, the GP prior is incorporated for the baseline intensity and triggering kernel. To guarantee the non-negativity of intensity, trajectories drawn from a GP prior need to be squashed by a link function. Due to the existence of the link function, the likelihood is non-conjugate to the prior, resulting in a complex inference procedure. In Chapter 4, quadratic GP Hawkes process is proposed, whose baseline intensity and triggering kernel are the square transformation of random trajectories from the GP prior. For the quadratic GP Hawkes process model, a variational Gaussian approximation inference method is proposed to approximate the true posterior with a Gaussian distribution.

In Chapter 5, the sigmoid GP Hawkes process is defined, where the link function is a scaled sigmoid function. For the sigmoid GP Hawkes process model, the latent Pólya-Gamma random variables and marked Poisson processes are augmented to convert the likelihood into a conjugate form with the GP prior; consequently, the inference can be much faster and efficient.

In the methods presented in Chapters 3-5, the triggering kernel is assumed to be unchanging over time-shifts, which is named stationarity. However, the triggering influence can change over time in real applications. In Chapter 6, the triggering kernel is assumed to be non-stationary. The key idea is to partition the process into many small sectors on which the stationarity is to be satisfied, then the baseline intensity and triggering kernel are estimated on each sector and these two characteristics are compared for two adjacent sectors. If the characteristics are similar the two sectors belong to the same segment, otherwise the two sectors should be cut apart. However, this naïve idea is impractical, as the estimation of baseline intensity and triggering kernel are time-consuming. Inspired by the Wiener-Hopf inference method [11], the baseline intensity and triggering kernel are replaced with first and second order cumulants of the Hawkes process. The result is a fast multi-resolution segmentation algorithm for non-stationary Hawkes process.

In Chapter 7, the contributions of the thesis are summarised and some promising research directions for future work are discussed.

Chapter 2

Background

In order to understand the subsequent chapters, the background knowledge of Hawkes process, Gaussian process, EM algorithm and Bayesian inference need to be explained. In this chapter, the background on techniques are reviewed. Specifically the background on the Hawkes process is introduced in Sec. 2.1, the Gaussian process in Sec. 2.2, the EM algorithm is discussed in Sec. 2.3, and lastly, Bayesian inference is reviewed in Sec. 2.4. The final section provides a summary of the chapter contents.

2.1 Hawkes Process

The Hawkes process is a self-exciting point process first introduced by Hawkes in 1971 [3]. The future evolution of a self-exciting point process is influenced by the timing of past events. The process is non-Markovian except for some special cases. Thus, the Hawkes process depends on the entire past history and has a long memory. The Hawkes process has wide applications in neuroscience, seismology, finance and many other fields.

An important characteristic that characters a Hawkes process is the conditional intensity function (Eq. 1.1). The likelihood of a Hawkes process is briefly introduced first. A Hawkes process is a stochastic process whose realization is a sequence of timestamps $D = \{t_i\}_{i=1}^{N} \in [0, T]$, where t_i stands for the time of occurrence of the i-th event with T being the observation window. Given $\mu(t)$ and $\phi(\tau)$, the Hawkes process likelihood [20] is defined as

$$p(D|\mu(t),\phi(\tau)) = \prod_{i=1}^{N} \left[\mu(t_i) + \sum_{t_j < t_i} \phi(t_i - t_j) \right] \cdot \exp\left(-\int_T \left(\mu(t) + \sum_{t_i < t} \phi(t - t_i) \right) dt \right). \tag{2.1}$$

2.1.1 Branching Structure of Hawkes Process

The Hawkes process has two equivalent interpretive counterparts. The first is the intensity based version which models the rate of events through conditional intensity function (Eq. 1.1, see Fig. 2.1). The alternative version is the cluster based version where the Hawkes process can be interpreted as a superposition of Poisson processes. Specifically, consider a Poisson cluster process C ([21], also see Fig. 2.1).

1. Let I be a realization of an inhomogeneous Poisson process with rate $\mu(t)$ on the interval $[0, T]$. All the points in I are called immigrants.

2. Each immigrant $x \in I$ will generate an independent cluster of points C_x. Each cluster C_x is generated according to the following branching structure: the cluster C_x consists of generations of offspring of the immigrant x; given the immigrant x and the offspring of generation n, each offspring of generation n will produce its own offspring of generation $n+1$ by generating a realization of an inhomogeneous Poisson process with rate $\phi(t - t_i)$.

3. The superposition of all points in all clusters $C = \sum_x C_x$ will provide a Hawkes process with baseline intensity $\mu(\cdot)$ and triggering kernel $\phi(\cdot)$.

From the Poisson cluster process counterpart, it is straightforward to see the definition of the branching structure: which point is triggered by which point. More

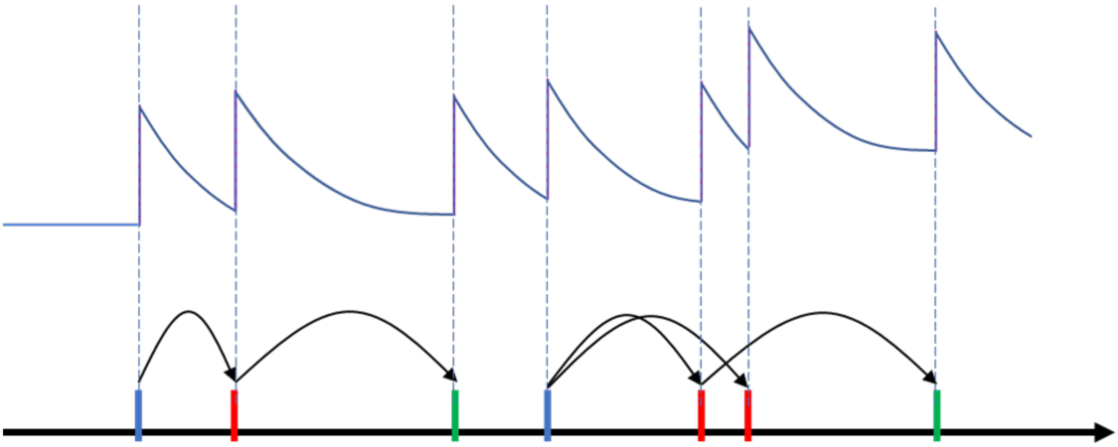


Figure 2.1: The conditional intensity and Poisson cluster interpretation of Hawkes process. The blue solid function is the conditional intensity function. Blue points are the immigrants, with red points being the first generation of offspring and green points the second generation of offspring.

formally, given a realization of Hawkes process with N events, an $N \times N$ lower triangular matrix $\mathbf{B}$ is defined, with binary entry b_{ij} indicating whether the i-th event is triggered by itself or by a previous event j.

$$
\begin{aligned}
b_{ii} &= \begin{cases} 1 & \text{if event } i \text{ is a background event} \\ 0 & \text{otherwise} \end{cases} \\
b_{ij} &= \begin{cases} 1 & \text{if event } i \text{ is caused by event } j \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
\tag{2.2}
$$

For example, the branching structure matrix for Fig. 2.1 is

$$
\begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0
\end{bmatrix}.
$$

Generally speaking, the branching structure is a latent variable for a Hawkes process, which means that it is not included in the observation. As can be seen later, the branching structure is an ingenious latent variable, because the Hawkes process likelihood can be decoupled into two independent factors after introducing it. The joint likelihood of the observation and branching structure is

$$\begin{aligned} p(D, \mathbf{B}|\mu(t), \phi(\tau)) = \prod_{i=1}^{N} \mu(t_i)^{b_{ii}} \exp\left(-\int_T \mu(t)dt\right) \cdot \\ \prod_{i=2}^{N} \prod_{j=1}^{i-1} \phi(t_i - t_j)^{b_{ij}} \prod_{i=1}^{N} \exp\left(-\int_{T_\phi} \phi(\tau)d\tau\right), \end{aligned} \tag{2.3}$$

where T_ϕ is the support of the triggering kernel. It is straightforward to see that the joint likelihood has already been decoupled to two independent components.

2.1.2 Cumulants of Hawkes Process

The cumulants of a Hawkes process are briefly described in this section. The detailed derivation can be found in [11, 22]. Consider a 1-variate Hawkes process $\mathbb{N}(t)$ whose jumps are all of size 1 and whose intensity at time t is $\lambda(t)$. If $\{t_i\}$ denotes the jump times of $\mathbb{N}(t)$, the $\lambda(t)$ can be expressed as Eq. 1.1. If $\mathbb{N}(t)$ is stationary, the following results are obtained. The first order cumulant (mean event rate) is

$$\Lambda dt = \mathbb{E}(d\mathbb{N}_t) = \frac{\mu}{1 - \int \phi(\tau)d\tau} dt. \tag{2.4}$$

The second order cumulant is

$$Cov(d\mathbb{N}_{t_1}, d\mathbb{N}_{t_2}) = \mathbb{E}(d\mathbb{N}_{t_1} d\mathbb{N}_{t_2}) - \mathbb{E}(d\mathbb{N}_{t_1})\mathbb{E}(d\mathbb{N}_{t_2}). \tag{2.5}$$

Because $\mathbb{N}(t)$ is stationary, consequently, $Cov(d\mathbb{N}_{t_1}, d\mathbb{N}_{t_2})$ only depends on $\tau = t_2 - t_1$ and can be expressed as

$$v(\tau)d\tau = \mathbb{E}(d\mathbb{N}_0 d\mathbb{N}_\tau) - \mathbb{E}(d\mathbb{N}_0)\mathbb{E}(d\mathbb{N}_\tau). \tag{2.6}$$

Or, it can be rewritten in terms of conditional expectations

$$g(\tau)d\tau = v(\tau)d\tau/\Lambda = \mathbb{E}(d\mathbb{N}_\tau | d\mathbb{N}_0 = 1) - \Lambda d\tau. \tag{2.7}$$

As proved elsewhere [11], a stationary Hawkes process is uniquely defined by its first and second order cumulants and there is a one-to-one mapping between its second order statistics $g(\tau)$ and the triggering kernel $\phi(\tau)$.

2.1.3 Inference Methods for Hawkes Process

The inference methods used for a Hawkes process are discussed below.

Maximum Likelihood Estimation The classical and easiest method to define inference of a (parametric) Hawkes process is the maximum likelihood estimation (MLE). Normally, a parametric form is assumed for the Hawkes process: constant baseline intensity μ and parametric triggering kernel e.g. exponential decay kernel $\alpha \exp(-\beta\tau)$ ($\alpha > 0$, $\beta > 0$ and $\alpha < \beta$) or power law kernel $\frac{\alpha\beta}{(\beta\tau+1)^{1+\gamma}}$ ($\alpha > 0$, $\beta > 0$ and $\alpha < \gamma$). The likelihood of a Hawkes process is already provided in Eq. 2.1. The optimal parameters can be obtained by maximizing the log-likelihood

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \log P(D|\boldsymbol{\theta}), \tag{2.8}$$

where $\boldsymbol{\theta}$ stands for parameters.

Model Independent Stochastic Declustering The literature [8] provides details on the use of the MISD algorithm in a one dimensional Hawkes process, where MISD is an EM-based nonparametric inference algorithm. Assume that the baseline intensity is a constant μ and there is no prior knowledge about the form of $\phi(\cdot)$. The joint likelihood of the observation and branching structure can be written as

$$p(D, \mathbf{B}|\mu, \phi(\tau)) = \underbrace{\prod_{i=1}^{N} \mu^{b_{ii}} \exp\left(-\mu T\right)}_{} \cdot \underbrace{\prod_{i=2}^{N} \prod_{j=1}^{i-1} \phi(t_i - t_j)^{b_{ij}} \prod_{i=1}^{N} \exp\left(-\int_{T_\phi} \phi(\tau) d\tau\right)}_{}. \tag{2.9}$$

It is easy to see that, after introducing the branching structure, the joint likelihood is decoupled into two independent factors: μ part and $\phi(\cdot)$ part. Correspondingly, the joint log-likelihood is

$$\begin{aligned} &\log p(D, \mathbf{B}|\mu, \phi(\tau)) = \\ &\left[\sum_{i=1}^{N} b_{ii} \log(\mu)\right] - \mu T + \sum_{i=2}^{N} \left[\sum_{j=1}^{i-1} b_{ij} \log\left(\phi(t_i - t_j)\right)\right] - \sum_{i=1}^{N} \int_{T_\phi} \phi(\tau) d\tau. \end{aligned} \tag{2.10}$$

It is straightforward to rewrite this problem into an EM framework, which results in the MISD algorithm. Because the branching structure is a latent variable, the MISD algorithm works by maximizing the expectation of the log-likelihood.

$$\mu^{s+1}, \phi^{s+1} = \operatorname*{argmax}_{\mu, \phi} \mathop{\mathbb{E}}_{\mathbf{B} \sim p(\mathbf{B}|\mu^s, \phi^s)} \log p(D, \mathbf{B}|\mu, \phi(\tau)), \tag{2.11}$$

where s indicates the s-th step in EM iterations. Therefore, b_{ij} in Eq. 2.10 is replaced by p_{ij}, which is the posterior probability of event i caused by event j given μ and

ϕ. The matrix p_{ij} is a lower triangular matrix:

$$\begin{bmatrix} p_{11} & & & & \\ p_{21} & p_{22} & & & \\ p_{31} & p_{32} & p_{33} & & \\ & \vdots & & \ddots & \\ p_{n1} & p_{n2} & p_{n3} & \cdots & p_{nn} \end{bmatrix}, \tag{2.12}$$

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

In summary, the EM iteration is

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

$$\begin{aligned} p_{ij}^s &= \frac{\phi^s(t_i - t_j)}{\mu^s + \sum_{j=1}^{i-1} \phi^s(t_i - t_j)} \\ p_{ii}^s &= \frac{\mu^s}{\mu^s + \sum_{j=1}^{i-1} \phi^s(t_i - t_j)}; \end{aligned} \tag{2.13}$$

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

$$\mu^{s+1} = \frac{1}{T} \sum_{i=1}^{n} p_{ii}^s. \tag{2.14}$$

Assume that $\phi(\tau)$ is an histogram function with the support T_ϕ uniformly divided into M bins (bin width being Δt), then the update for ϕ is given by

$$\phi_m^{s+1} = \frac{1}{|A_m|\Delta t} \sum_{i,j \in A_m} p_{ij}^s, \tag{2.15}$$

where A_m is the set of event-pairs satisfying $m\Delta t \leqslant |t_i - t_j| < (m+1)\Delta t$, ϕ_m is the height of m-th bin where $0 \leqslant m \leqslant M-1$, and $|A_m|$ is the size of A_m. Eq. 2.14 and 2.15 are derived from $\frac{\partial}{\partial \mu}\mathbb{E}_{\mathbf{B}}[\log p(D, \mathbf{B})] = 0$ and $\frac{\partial}{\partial \phi_m}\mathbb{E}_{\mathbf{B}}[\log p(D, \mathbf{B})] = 0$.

Wiener-Hopf Method Based on the literature [11], the following statements may be made:
(1) A 1-variate Hawkes process with stationary increments is uniquely defined by its first-order statistics (i.e. the first-order cumulant Eq. 2.4) and its second-order statistics (given by either its second-order cumulant Eq. 2.6 or equivalently the conditional expectation Eq. 2.7).
(2) The triggering kernel $\phi(\tau)$ and the conditional expectation $g(\tau)$ (Eq. 2.7) satisfy the following Wiener-Hopf equation:

$$g(\tau) = \phi(\tau) + \phi(\tau) * g(\tau), \forall \tau > 0 \tag{2.16}$$

where $*$ stands for convolution. The detailed proof can be found elsewhere [11].

In fact, statement (1) is a direct consequence of Eq. 2.16 which proves that the second-order statistics $g(\tau)$ fully characterize the triggering kernel $\phi(\tau)$, and of Eq. 2.4 which can express μ as a function of $\phi(\tau)$ and the first-order cumulant Λ.

Quadratic Contrast Function Method Other works [12] and [13] attempt to minimize a quadratic contrast (loss) function with a grid based triggering kernel. The former [12] may be elaborated as follows. Given a 1-variate Hawkes process $\mathbb{N}(t)$, the conditional intensity function is once again given by Eq. 1.1 where the triggering kernel $\phi(\tau)$ belongs to a class of non-negative integrable functions. The goal is the nonparametric estimation of the triggering kernel of Hawkes process, which means that no parametric form is assumed for the triggering kernel e.g. exponential decay or power law decay. The real line is divided into intervals of width $h > 0$ and the following is defined:

$$Y_t^h = \mathbb{N}(th) - \mathbb{N}((t-1)h) \tag{2.17}$$

with $t \in \mathbb{Z}$. For every fixed h, Y_t^h represents the number of jumps in time intervals $((t-1)h, th]$. Thus, the random variable Y_t^h is approximately binary with small enough h. If the triggering kernel $\phi(\tau)$ is continuous and h is small enough, $\phi(\tau)$

can be approximated by a piecewise constant function leading to

$$\mathbb{E}(Y_t^h|\mathcal{H}_{h(t-1)}) \approx h\mu + h\sum_{u=1}^{\infty}\phi(uh)Y_{t-u}^h, \tag{2.18}$$

where $\mathcal{H}_{h(t-1)}$ is the history before $(t-1)h$. This naturally suggests a least square estimator for the triggering kernel. Assume there are k bins for $\phi(\tau)$ and zero otherwise, therefore defining a contrast (loss) function

$$\sum_{t=k+1}^{T/h} \|Y_t^h - \mu h - \boldsymbol{\phi}^{h,k}\mathbf{Y}_t^{h,k}\|_2^2, \tag{2.19}$$

where $\boldsymbol{\phi}^{h,k} = (h\phi(h), \ldots, h\phi(kh))^T$ and $\mathbf{Y}_t^{h,k} = (Y_{t-1}^h, \ldots, Y_{t-k}^h)$. Also define

$$\begin{aligned} \boldsymbol{\gamma}_{h,k} &= \text{cov}(Y_t^h, \mathbf{Y}_t^{h,k}) = (\text{cov}(Y_t^h, Y_{t-u}^h))_{u=1,\ldots,k}, \\ \boldsymbol{\Gamma}_{h,k} &= \text{cov}(\mathbf{Y}_t^{h,k}, \mathbf{Y}_t^{h,k}) = (\text{cov}(Y_{t-u}^h, Y_{t-v}^h))_{u,v=1,\ldots,k}. \end{aligned} \tag{2.20}$$

It is straightforward to prove that the above loss function is minimized by

$$\hat{\boldsymbol{\phi}}^{h,k} = \hat{\boldsymbol{\gamma}}_{h,k}^T\hat{\boldsymbol{\Gamma}}_{h,k}^{-1}, \qquad \hat{\mu}^h = \bar{Y}^h - \hat{\boldsymbol{\phi}}^{h,k}\bar{\mathbf{Y}}^{h,k} \tag{2.21}$$

with $\hat{\boldsymbol{\gamma}}_{h,k} = \frac{1}{T/h-k}\sum_{t=k+1}^{T/h}(Y_t^h - \bar{Y}^h)(\mathbf{Y}_t^{h,k} - \bar{\mathbf{Y}}^{h,k})$ being the empirical covariance of Y_t^h and $\mathbf{Y}_t^{h,k}$ and $\hat{\boldsymbol{\Gamma}}_{h,k}$, $\bar{Y}^h$, $\bar{\mathbf{Y}}^{h,k}$ are defined similarly.

Clearly the quadratic contrast function method is related to the Wiener-Hopf equation method, based on the fact that the triggering kernels are both determined by the second-order statistics (covariance or conditional expectation) and the baseline intensities are both determined by the triggering kernel and first-order statistics (mean rate). In fact, the ϕ estimator in Eq. 2.21 is just equivalent to the Nystrom discrete version [23] of the convolution equation (Eq. 2.16) and the μ estimator in Eq. 2.21 is just equivalent to Eq. 2.4.

2.2 Gaussian Process

A continuous stochastic process $f(x)$ is a Gaussian process if and only if for every finite set of indices $x_1, \ldots, x_k$, $\mathbf{f}_{x_1,\ldots,x_k} = (f(x_1), \ldots, f(x_k))$ is a multivariate Gaussian random variable. More formally, GP is specified by the mean function $m(x)$ and covariance kernel $k(x, x')$:

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')), \tag{2.22}$$

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 $k(x, x')$. Unless explicitly stated, the covariance kernel is the squared exponential kernel throughout the thesis:

$$k(x, x') = \theta_0 \exp\left(-\frac{\theta_1}{2}\|x - x'\|^2\right), \tag{2.23}$$

where θ_0, θ_1 are the hyperparameters.

For the GP regression problem [24], assume that the observations are subjected to a normally distributed noise, so the likelihood is in a Gaussian form which is conjugated to the GP prior. Given a set of observations $((x_1, y_1), \ldots, (x_N, y_N))$, the posterior mean and variance are

$$m(x) = \mathbf{k}^T \mathbf{C}_N^{-1} \mathbf{y}, \qquad \sigma^2(x) = c - \mathbf{k}^T \mathbf{C}_N^{-1} \mathbf{k} \tag{2.24}$$

where $\mathbf{k} = (k(x_1, x), \ldots, k(x_N, x))$; $c = k(x, x) + \sigma_\epsilon^2$ and $\mathbf{C}_N$ is the matrix $C(x_n, x_{n'}) = k(x_n, x_{n'}) + \sigma_\epsilon^2 \delta_{nn'}$ with σ_ϵ^2 being the noise variance of observations; $\delta_{nn'} = 1$ when $n = n'$ and 0 otherwise; $n, n' \in N$; $\mathbf{y} = (y_1, \ldots, y_N)$.

For the GP classification problem [24], the inference becomes more complicated because the likelihood is non-Gaussian: a sigmoid transformation of function f (binary classification) or a softmax transformation of f (multi-class classification).

Consequently, there is no analytical expression for the posterior. Various approximate inference algorithms have been proposed, e.g. Laplace approximation or expectation propagation algorithm. More details can be found elsewhere [24].

2.3 EM Algorithm

The EM algorithm is used to find local maximum likelihood (posterior) parameters of a statistical model in cases where the likelihood (posterior) cannot be maximized directly. Typically these models include latent variables, parameters and observations. Usually the model can be formulated more simply by augmentation of latent variables, e.g. the likelihood of a mixture model can be simpler by assuming that each observed data belongs to a mixture component (latent variable).

Finding a maximum likelihood (posterior) solution typically requires the gradient of the likelihood (posterior) w.r.t. all parameters and latent variables. However, this is usually intractable in typical statistical models with latent variables because the result is a set of interlocking equations in which the solution to the parameters requires the values of the latent variables and vice versa. Instead, the EM algorithm uses iterations to find maximum likelihood (posterior) estimates.

Given the observed data D, a set of unobserved latent variables $\mathbf{Z}$ and a vector of parameters $\boldsymbol{\theta}$ along with a likelihood (posterior) $p(D, \mathbf{Z}|\boldsymbol{\theta})$, the EM algorithm seeks to find the maximum of the marginal distribution $p(D|\boldsymbol{\theta}) = \int p(D, \mathbf{Z}|\boldsymbol{\theta}) d\mathbf{Z}$ by iteratively applying expectation (E) and maximization (M) steps:

$$\boldsymbol{\theta}^{s+1} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \underset{\mathbf{Z} \sim p(\mathbf{Z}|\boldsymbol{\theta}^s)}{\mathbb{E}} [\log p(D, \mathbf{Z}|\boldsymbol{\theta})] \tag{2.25}$$

with s indicating iteration steps.

2.4 Bayesian Inference

The two most popular Bayesian inference methods are MCMC [25] and variational inference [26]. The former has the advantage of being asymptotically exact; the latter has the advantage of maximizing an explicit objective and being faster in most cases. In this section, the classical MCMC algorithms are discussed first, including Metropolis-Hasting (MH) sampling [27] and Gibbs sampling [28]; secondly the variational inference algorithms are described, including variational Gaussian approximation [29] and mean-field variational inference [30].

2.4.1 Metropolis-Hasting Sampling

The MH algorithm is an MCMC method for obtaining a sequence of random samples from a probability distribution from which direct sampling is difficult. The MH algorithm works by generating a sequence of samples in such a way that the distribution of samples will closely approximate the desired distribution as more and more samples are generated. These samples are generated recursively with the distribution of the next sample being dependent only on the current sample (Markov chain). Specifically, the algorithm proposes a candidate for the next sample based on the current sample at each iteration. Then, with some probability, the candidate is either accepted or rejected.

Mathematically, given a desired distribution $p(x)$ and a proposal distribution $g(x'|x_t)$, the MH algorithm consists of the following steps:

1. Set an initial sample x_0 and set $t = 0$;
2. Randomly generate a candidate x' from $g(x'|x_t)$;
3. Calculate the acceptance probability $A = \min\left(1, \frac{p(x')g(x_t|x')}{p(x_t)g(x'|x_t)}\right)$;
4. Generate a uniform random variable $u \in [0, 1]$; accept the new sample and set $x_{t+1} = x'$ if $u \leq A$ or reject the new sample and set $x_{t+1} = x_t$ if $u > A$;

set $t = t + 1$ and go back to step 2 until the desired number of samples are obtained.

2.4.2 Gibbs Sampling

The MH sampling algorithm above can only be applied to low dimensional distributions since the acceptance rate will decrease exponentially with the number of dimensions. For high dimensional distributions, Hamiltonian Monte Carlo [31], elliptical slice sampling [32] or Gibbs sampling [28] can be used. Here, only the Gibbs sampling algorithm is discussed, with the alternatives available via references.

Gibbs sampling is an MCMC algorithm for obtaining a sequence of samples which are approximately from a specified multivariate probability distribution when direct sampling is difficult. This sequence can be used to approximate the joint distribution. Gibbs sampling is applicable when the joint distribution is not known explicitly or is difficult to sample from directly, but the conditional distribution of each variable is known and is easier to sample from. The Gibbs sampling algorithm generates a sample from the distribution of each variable in turn, conditional on the current values of other variables. It can be shown that the sequence of samples constitutes a Markov chain and the stationary distribution of that Markov chain is just the target joint distribution.

Mathematically, the aim is to sample from a joint distribution $p(x_1, \ldots, x_N)$. Denote the i-th sample by $\mathbf{X}^i = (x_1^i, \ldots, x_N^i)$. The algorithm follows:

1. Set an initial sample $\mathbf{X}^i$;
2. To sample x_j^{i+1}, update it according to the distribution specified by $p\left(x_j^{i+1} | x_1^{i+1}, \ldots, x_{j-1}^{i+1}, x_{j+1}^i, \ldots, x_N^i\right)$;
3. Repeat until the desired number of samples are obtained.

2.4.3 Variational Gaussian Approximation

To provide an analytical approximation to the posterior, variational Bayes is an alternative to MCMC methods (such as Gibbs sampling) for taking a fully Bayesian approach to statistical inference over complex distributions which are difficult to directly evaluate or sample from. In particular, variational Bayes provides a locally-optimal exact analytical solution to an approximation of the posterior. For many applications, variational Bayes produces solutions of comparable accuracy to Gibbs sampling at greater speed. However, deriving the set of equations used to iteratively update the parameters often requires a large amount of work compared with deriving the comparable Gibbs sampling equations.

In variational inference, the posterior distribution over a set of unobserved variables $\mathbf{Z} = \{Z_1 \dots Z_N\}$ given some data $\mathbf{X}$ is approximated by a variational distribution $q(\mathbf{Z})$: $p(\mathbf{Z} \mid \mathbf{X}) \approx q(\mathbf{Z})$. The distribution $q(\mathbf{Z})$ is restricted to belong to a family of distributions of simpler form than $p(\mathbf{Z} \mid \mathbf{X})$ selected with the intention of making $q(\mathbf{Z})$ similar to the true posterior. The lack of similarity is measured in terms of a dissimilarity function and hence inference is performed by selecting the distribution $q(\mathbf{Z})$ that minimizes the dissimilarity function.

The most common type of variational Bayes uses the Kullback-Leibler divergence (KL-divergence) of p from q as the choice of dissimilarity function. The KL-divergence between $q(\mathbf{Z})$ and $p(\mathbf{Z} \mid \mathbf{X})$ can be written as

$$D_{\mathrm{KL}}(q \parallel p) = \int_{\mathbf{Z}} q(\mathbf{Z}) \left[\log \frac{q(\mathbf{Z})}{p(\mathbf{Z}, \mathbf{X})} + \log p(\mathbf{X}) \right], \tag{2.26}$$

or equivalently expressed as

$$\log p(\mathbf{X}) = D_{\mathrm{KL}}(q \parallel p) - \int_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{q(\mathbf{Z})}{p(\mathbf{Z}, \mathbf{X})} = D_{\mathrm{KL}}(q \parallel p) + \mathcal{L}(q). \tag{2.27}$$

As the log evidence $\log p(\mathbf{X})$ is fixed with respect to q, maximizing the final term $\mathcal{L}(q)$ minimizes the KL divergence of p from q. By appropriate choice of q, $\mathcal{L}(q)$ becomes tractable to compute and to maximize. Hence both an analytical approximation q for the posterior $p(\mathbf{Z} \mid \mathbf{X})$ and a lower bound $\mathcal{L}(q)$ for the evidence $\log p(\mathbf{X})$ are obtained. The term $\mathcal{L}(q)$ is often called the evidence lower bound (ELBO).

The variational Gaussian approximation is a parametric variational inference approach where the variational distribution q is assumed to be a multivariate Gaussian distribution. The advantages of variational Gaussian approximation include the following:
(1) the correlations between variables are incorporated compared to the mean-field method (see Sec. 2.4.4);
(2) generally speaking, due to the Gaussian form, the ELBO or KL-divergence is easier to compute.

More formally, by restricting the variational distribution q to be a multivariate Gaussian distribution with mean $\mathbf{m}$ and covariance Σ,

$$\mathcal{L}(q) = -\int_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{q(\mathbf{Z})}{p(\mathbf{Z}, \mathbf{X})} = \mathbb{E}_q \log p(\mathbf{Z}, \mathbf{X}) + \frac{1}{2} \log(|\Sigma|) + \frac{N}{2} \log(2\pi) + \frac{N}{2} \quad (2.28)$$

where N is the dimension of $\mathbf{Z}$. Hence, setting the derivatives of $\mathcal{L}(q)$ with respect to the variational parameters $\mathbf{m}$ and $\boldsymbol{\Sigma}$ equal to zero leads to

$$\nabla_{\mathbf{m}} \mathbb{E}_q \log p(\mathbf{Z}, \mathbf{X}) = 0, \qquad \boldsymbol{\Sigma}^{-1} = -2 \nabla_{\boldsymbol{\Sigma}} \mathbb{E}_q \log p(\mathbf{Z}, \mathbf{X}). \quad (2.29)$$

This requires the computation of $N(N+1)/2 + N$ variational parameters, which is much larger than the number of parameters required for the mean-field method.

2.4.4 Mean Field Approximation

In contrast to the variational Gaussian approximation, the mean field variational inference is a nonparametric approach where the variational distribution $q(\mathbf{Z})$ is only assumed to factorize over some partition of the latent variables, e.g. $\mathbf{Z}$ is partitioned into $\mathbf{Z}_1 \ldots \mathbf{Z}_M$: $q(\mathbf{Z}) = \prod_{i=1}^{M} q_i(\mathbf{Z}_i \mid \mathbf{X})$. It can be shown using the calculus of variations that the optimal distribution q_j^* for each of the factors minimizing the KL divergence can be expressed as

$$\ln q_j^*(\mathbf{Z}_j \mid \mathbf{X}) = \mathbb{E}_{\mathbf{Z}_{i \neq j}}[\log p(\mathbf{Z}, \mathbf{X})] + \text{constant}. \tag{2.30}$$

The constant in the above expression is the normalizing constant and is usually reinstated by inspection because the rest of the expression can be recognized as being a known type of distribution.

For each partition of variables, by simplifying the expression of the distribution over the variables and examining the functional dependency of the distribution on the variables in question, the family of the distribution can usually be determined. The formula for the distribution parameters will be expressed in terms of expectations of functions of variables in other partitions. In most cases, the distributions of the other variables will be from known families, and the formulas for the relevant expectations can be looked up. However, the formulas depend on the parameters of those distributions, which depend in turn on the expectations of other variables. The result is that the formulas for the parameters of the distribution of each variable can be expressed as a series of interlock equations. Usually, it is not possible to solve this system of equations directly. However, as described above, the dependencies suggest a simple iterative algorithm, which in most cases is guaranteed to converge.

2.5 Summary

In this section, the background knowledge about Hawkes process, Gaussian process, EM algorithm and Bayesian inference is introduced, which is the basis of understanding subsequent chapters. In the next chapter, a frequentist nonparametric inference method of Hawkes process will be discussed.

Chapter 3

A Refined MISD Algorithm for Nonparametric Hawkes Process*

In this chapter, a refined MISD algorithm is proposed. For the classical MISD algorithm discussed in Sec. 2.1.3, the performance has a strong dependence on the number of bins, which needs to be tuned by model selection method e.g. cross validation. To ease the choice of number of bins, an innovative embedding of a Gaussian process regression step into the EM iteration is proposed to obtain a refined MISD algorithm that is less sensitive to the choice of number of bins.

3.1 Overview

In a real application, data is usually collected sequentially. The modelling of such time series data to discover the underlying temporal dynamics is a challenging problem in this domain. To address it, different models have been proposed in the past such as recurrent neural networks (RNN) [33] and temporal point process [34]. There

*Portions of this chapter have been published: F. Zhou, Z. Li, X. Fan, Y. Wang, A. Sowmya, F. Chen, A Refined MISD Algorithm Based on Gaussian Process Regression, Advances in Knowledge Discovery and Data Mining, Springer International Publishing AG, part of Springer Nature 2018, D. Phung et al. (Eds.): PAKDD 2018, LNAI 10938, pp. 584596, 2018 are many variants of the latter, such as homogeneous Poisson process [35], inhomogeneous Poisson process [36] and Hawkes process [3].

The MISD algorithm was proposed to perform nonparametric estimation of the triggering kernel and baseline intensity [1]. Essentially MISD is a histogram density estimator, and the triggering kernel obtained from MISD is a discrete function and the number of bins used in the model has a vital impact on the learning results. It can be seen from the experiments in this chapter that the learned triggering kernel underfits when fewer bins are used, and overfits when more are used. To determine the optimal number of bins, the log-likelihood $\log \mathcal{L}(\{t_i\}|M)$ conditioned on bin number M may be computed as $\hat{M}$ by maximum likelihood estimation, or from an unnormalized posterior distribution by multiplying the likelihood with a prior distribution on M such as a Poisson distribution; it is assumed that all the bins are equally wide. However, both these methods will lead to extra computations which is undesirable. Instead, a refined MISD algorithm which does not depend on the choice of number of bins is proposed in this chapter. A Gaussian process regression is innovatively embedded into the iterations of the MISD algorithm, to produce a refined algorithm which is less sensitive to the choice of number of bins, named GP-MISD. In this new method, M can be set to a large number to use over-segmented bins since it can prevent overfitting to some extent.

The remainder of this chapter is organized as follows: in Sec. 3.2, the new algorithm GP-MISD is proposed. Synthetic and real data experiments and a detailed discussion are provided in Sec. 3.3, the choice and sensitivity of hyperparameters are discussed in Sec. 3.4 and Sec. 3.5 concludes this chapter.

3.2 Proposed Model

The basic MISD algorithm has already been described in Sec. 2.1.3. In this section, the proposed GP-MISD algorithm is presented here. The key idea in GP-MISD is to embed a Gaussian process regression into the EM iterations, which makes use of the rates learned in each iteration step to perform a regression and obtain a smooth mean triggering kernel. This kernel is used in the next iteration step, and the iterations proceed in this fashion.

After obtaining the observation points $(\phi_1^s, \cdots, \phi_M^s)$ in iteration step s in MISD (Eq. 2.15), GP regression is used to evaluate the posterior mean function $m(x|\phi_1^s, \cdots, \phi_M^s)$, which will be used as the $\phi(\tau)$ in the next iteration step. Specifically, the new algorithm can be divided into three steps:

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

$$\begin{aligned} p_{ij}^s &= \frac{\bar{\phi}^s(t_i - t_j)}{\mu^s + \sum_{j=1}^{i-1} \bar{\phi}^s(t_i - t_j)} \\ p_{ii}^s &= \frac{\mu^s}{\mu^s + \sum_{j=1}^{i-1} \bar{\phi}^s(t_i - t_j)}. \end{aligned} \tag{3.1}$$

(2) M-step performs the update for the baseline intensity and triggering kernel, using Eq. 2.14 and Eq. 2.15.

(3) GP-step performs the update for the Gaussian process posterior mean:

$$\bar{\phi}^{s+1}(\tau) = \mathbf{k}^T \mathbf{C}_M^{-1} \boldsymbol{\phi}^{s+1}, \tag{3.2}$$

where $\mathbf{C}_M$ is the matrix with entries $C(\tau_n, \tau_m) = k(\tau_n, \tau_m) + \sigma_\epsilon^2 \delta_{nm}$, $\{\tau_i\}_{i=1}^M$ are the x-values of M triggering kernel points, $k(\cdot)$ is the covariance kernel function, and σ_ϵ^2 is the variance of the noise in observation points, $\mathbf{k} = (k(\tau_1, \tau), k(\tau_2, \tau), \cdots, k(\tau_M, \tau))$, $\boldsymbol{\phi}^{s+1} = (\phi_1^{s+1}, \phi_2^{s+1}, \cdots, \phi_M^{s+1})$ are the y-values of M triggering kernel points on step $s+1$. The final triggering kernel obtained by this algorithm is $\bar{\phi}(\tau)$, and Eq. 3.2 is derived from the standard Gaussian process regression (see Sec. 2.2).

3.3 Experiments

In this section, the synthetic data experiments are presented in Sec. 3.3.1 and real data experiments in Sec. 3.3.2. For evaluation metric, the training error is defined as negative log-likelihood of the training data. Then the model learned is applied to the test data to obtain the test error, which is defined as negative log-likelihood of the test data. For GP hyperparameters, setting the values of θ_0 and θ_1 is also a key step in all GP-based methods. The hyperparameters used to determine the GP kernel implicitly encode information on the flexibility of the GP. The optimization of hyperparameters in GP has been proven to be a non-convex problem [37], which may introduce some difficulty in learning the hyperparameter values. In the experiments, grid search was used to find the optimal hyperparameter values. It was also found that setting the hyperparameter values in a reasonable range does not severely affect the final result.

3.3.1 Synthetic Data

For simplicity, it is assumed that the true triggering kernel is an exponential decay function: $\mu = 1$, $\phi(\tau) = 1 \cdot \exp(-2\tau)$. Two sets of synthetic data are generated from the Hawkes process specified above using the thinning algorithm [38]. For each set, the observation duration T is set to 400, resulting in generation of about 850 events. The first set is used as the training dataset, and the second one as the test dataset.

For the inference, it is assumed that the baseline intensity is constant and the form of the triggering kernel is unknown, so the goal is to infer μ and $\phi(\tau)$. The MISD algorithm is trained the training dataset for different numbers of bins ranging from 3 to 100. $\phi(\tau)$ is assumed to be zero outside the support $[0, 3]$ and the number of iterations is set to 100. The same experimental protocol is also applied to the GP-MISD algorithm. The hyperparameters θ_0, θ_1, σ_ϵ^2 are chosen to be 2.3, 2.3 and 0.01 in the GP step based on grid search.

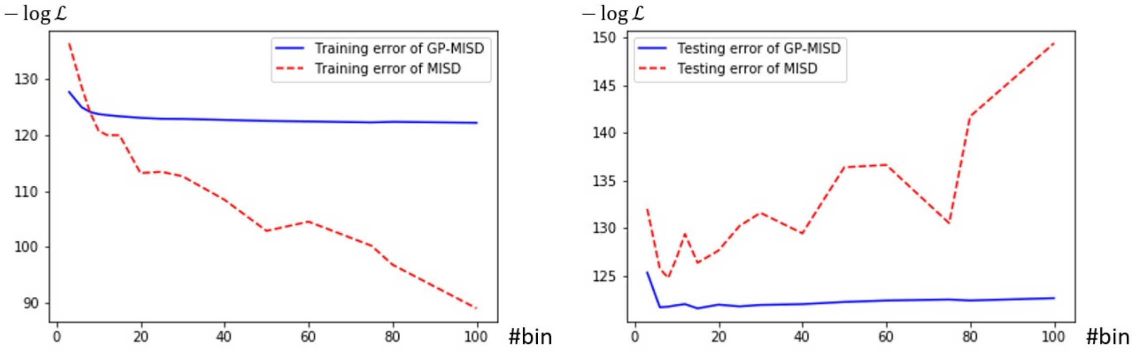


Figure 3.1: The training errors (left) and test errors (right) of MISD and GP-MISD.

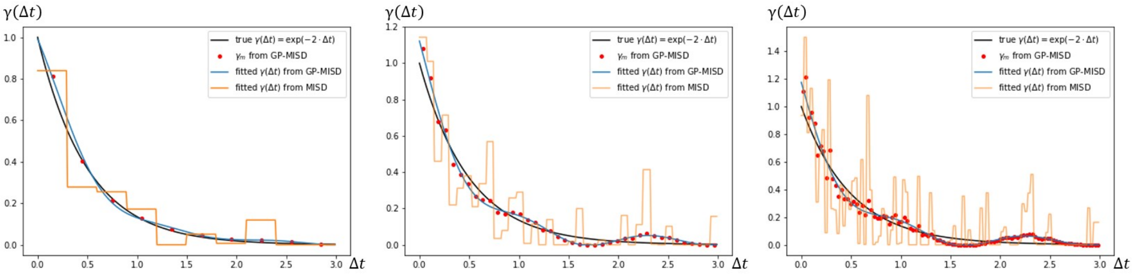


Figure 3.2: The fitting results of $\phi(\tau)$ from MISD and GP-MISD based on 10 bins (left), 40 bins (middle) and 100 bins (right).

The training error and test error for both algorithms appear in Fig. 3.1. It can be seen that as the number of bins increases from 3 to 100, the training error of MISD decreases monotonically, while the test error increases after #bin=8. For GP-MISD, the training error does not decrease rapidly after #bin=8, while the test error is almost constant after #bin=8. These results show that GP-MISD is less sensitive to the choice of number of bins compared to MISD, where the latter is likely to be overfitting when too many bins are used. More importantly, from the test error it is clearly the case that GP-MISD is always superior to MISD no matter how many bins are used, and this can also be found from the fitting results of $\phi(\tau)$ in Fig. 3.2 which is based on #bin=10, 40 and 100. It is clear that the estimation result of $\phi(\tau)$ from GP-MISD matches the ground truth. More importantly, the result is more stable with respect to the number of bins, showing the superiority of GP-MISD on the synthetic datasets.

3.3.2 Real Data

The performances of GP-MISD and MISD are evaluated on real world datasets from two different domains. These datasets are now described.

1. Motor Vehicle Collisions in New York City

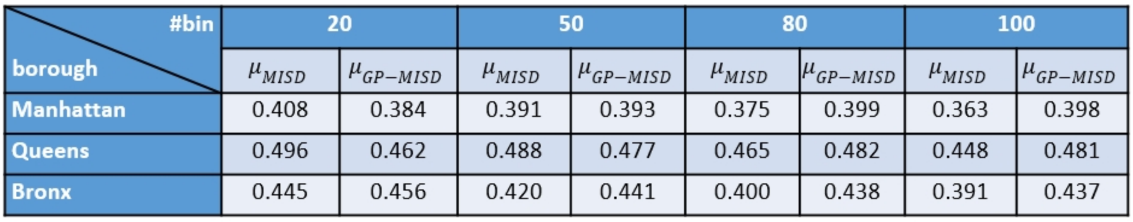
   The motor vehicle collision dataset [39] was made available by the New York City Police Department (NYPD). It contains about 1.05 million vehicle collision records in New York City from July, 2012 to September, 2017. The dataset includes the collision date, time, borough, location, contributing factor and other relevant information. For the proposed model, the most valuable features are the date and time. To avoid the oversize dataset, the collision records in Manhattan, Queens and Bronx caused by 'Alcohol Involvement' were filtered out. For each borough, half of the records are used as the training dataset and the other half as the test dataset. In the dataset, some collisions occur at the same time as the resolution is at minute level, which violates the definition of the temporal point process. To avoid this, a small time interval is added to all the simultaneous records so as to separate them. The hyperparameters θ_0, θ_1, σ_ϵ^2 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 based on grid search. Both algorithms are run for 100 iterations. The support of $\phi(\tau)$ is set to 3.0 so as to be long enough for the triggering effect and the time unit is 1.16 day.

2. NYPD Complaint Data 2017

   This dataset [40] includes all valid felony, misdemeanour and violation crimes reported to the NYPD for all complete quarters at the time of data collection in 2017. It includes 228,000 complaint records in New York City. The columns are complaint number, date, time, offense description, borough and other relevant information. To avoid the oversize dataset, the complaints in Manhattan, Queens and Brooklyn with the offense description of "THEFT-FRAUD" were filtered out. Again, for each borough, half the records are used as the training dataset and the others as the test dataset. A small time interval is added to separate all simultaneously occurring events. The hyperparameters θ_0, θ_1, σ_ϵ^2 are set to 6.45, 6.45, 0.01 for all boroughs based on grid search. Both algorithms are run for 100 iterations. The support of $\phi(\tau)$ is set to 3.0 so as to be long enough for the triggering effect and the time unit is 11.6 days.

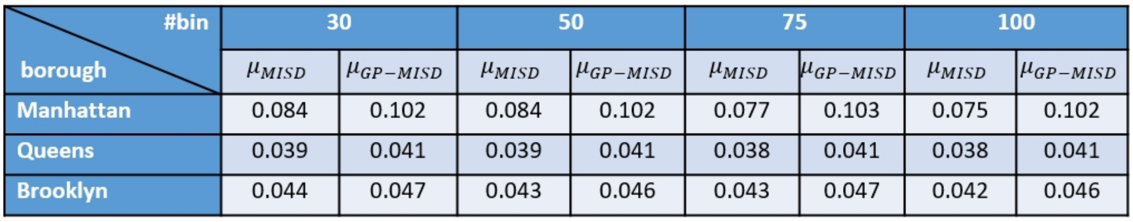
For Motor Vehicle Collisions in New York City, the learned μ, $\phi(\tau)$ and the test errors of both algorithms for #bin=20, 50, 80, 100 are shown in Table 3.1 and Fig. 3.3. For NYPD Complaint Data 2017, the learned μ, $\phi(\tau)$ and the test errors of both algorithms for #bin=30, 50, 75, 100 are shown in Table 3.2 and Fig. 3.4.

Table 3.1: 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

Table 3.2: NYPD Complaint Data 2017: the learned baseline intensity μ from MISD and GP-MISD based on #bin=30, 50, 75, 100.



#bin borough	30		50		75		100	
	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 sets of results, clearly $\phi(\tau)$ of GP-MISD is smoother and more stable than that of MISD and the test error of GP-MISD is always lower than MISD, which is consistent with the synthetic dataset result. GP-MISD effectively avoids the overfitting phenomenon and the algorithm is 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. From the learned triggering kernel in Fig. 3.3, a periodic oscillation is observed. This corresponds to the periodic influence of the initial accident, e.g. a car crash will cause instant traffic jam in the beginning, then the traffic jam on one road will cause further traffic jams on all the connected roads, and the further traffic jams will cause further car crashes in a similar fashion, but overall the influence will become smaller over time. For crime complaints, the triggering patterns in different boroughs are similar and the triggering effect lasts for almost one month, but is significant in the first 10 days. From the learned triggering kernel in Fig. 3.4, a periodic oscillation is not obvious but a peak rate is observed in the first 10 days. This corresponds to the nature of the crime: a criminal case only has a large influence on subsequent cases in an ensuing short period, due to the shortly following police investigation. Moreover, the trend of the triggering kernel is quite dynamic, especially in the short period after the source event has occurred, e.g., within about 0.5 day after initial collision in Fig. 3.3, or about 5 days after the initial complaint in Fig. 3.4. To capture the trend, the #bin must be set to be large enough so that the resolution is sufficient, however, too large a #bin will cause overfitting, such as spikes in the triggering kernel. This is the advantage of GP-MISD, which represents the triggering kernel with continuity and capturing any dynamic trends without overfitting.

3.4 Discussion

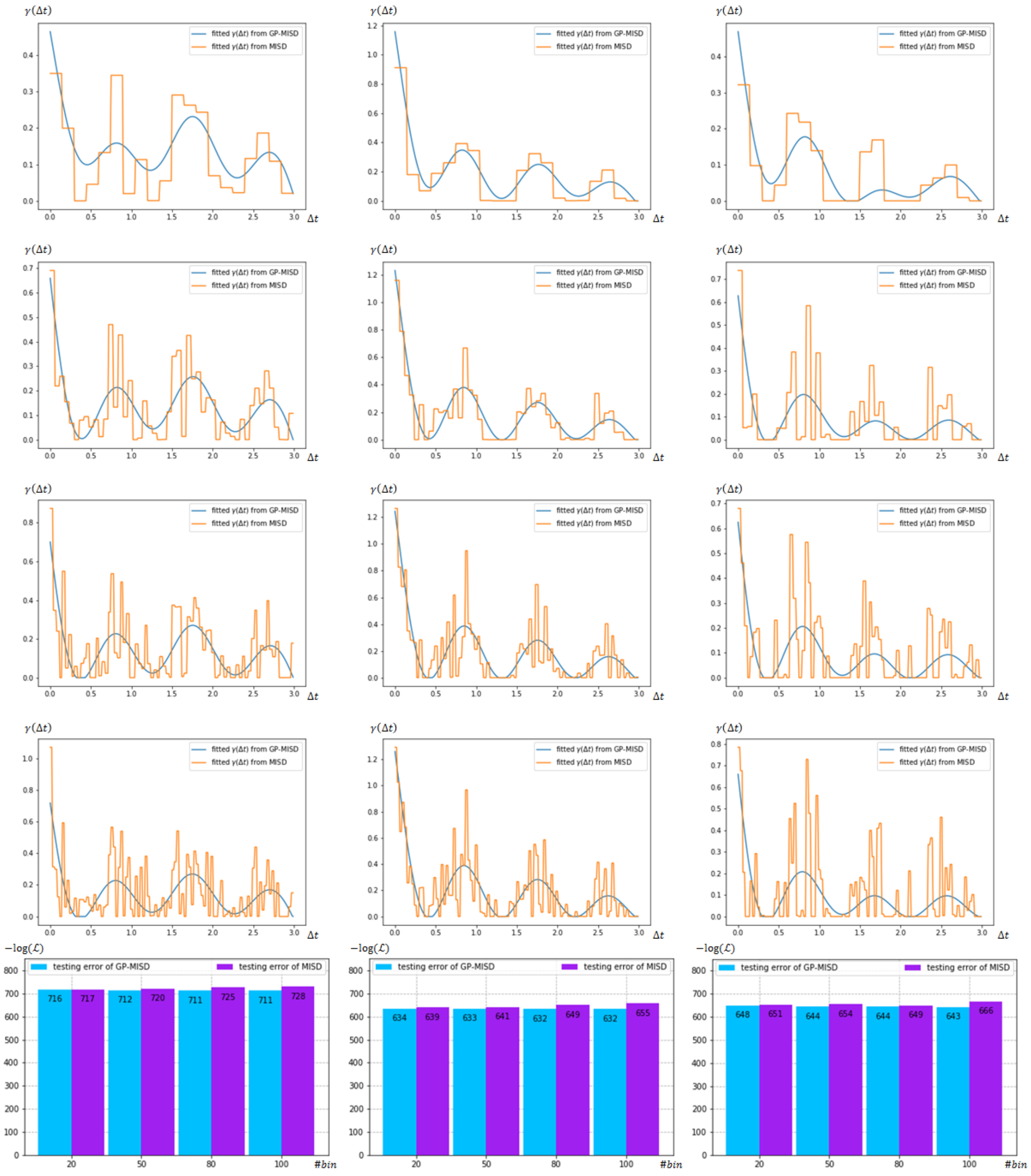
In this section, the choice of and sensitivity of hyperparameters is discussed. In the refined MISD algorithm, the hyperparameters include θ_0, θ_1 and σ_ϵ^2, which are common hyperparameters for GP. Generally speaking, just as in the case of the normal GP, θ_0 and θ_1 have significant influence on the estimation results, but σ_ϵ^2 does not have such a significant effect, because it is a noise variance parameter used to improve the numerical stability and can be set to a small positive value (e.g.

0.01). As stated in the experiments, θ_0 and θ_1 are tuned based on grid search to find the optimal value, and σ_ϵ^2 is set to 0.01.

The advantages of the proposed GP-MISD are already illustrated above: it can smooth the overfitting result caused by the choice of number of bins. Correspondingly, due to the incorporation of GP, a disadvantage is that the hyperparameters have to be chosen carefully. In this work, the hyperparameters are chosen based on grid search, however other more efficient methods can also be applied e.g. gradient based methods.

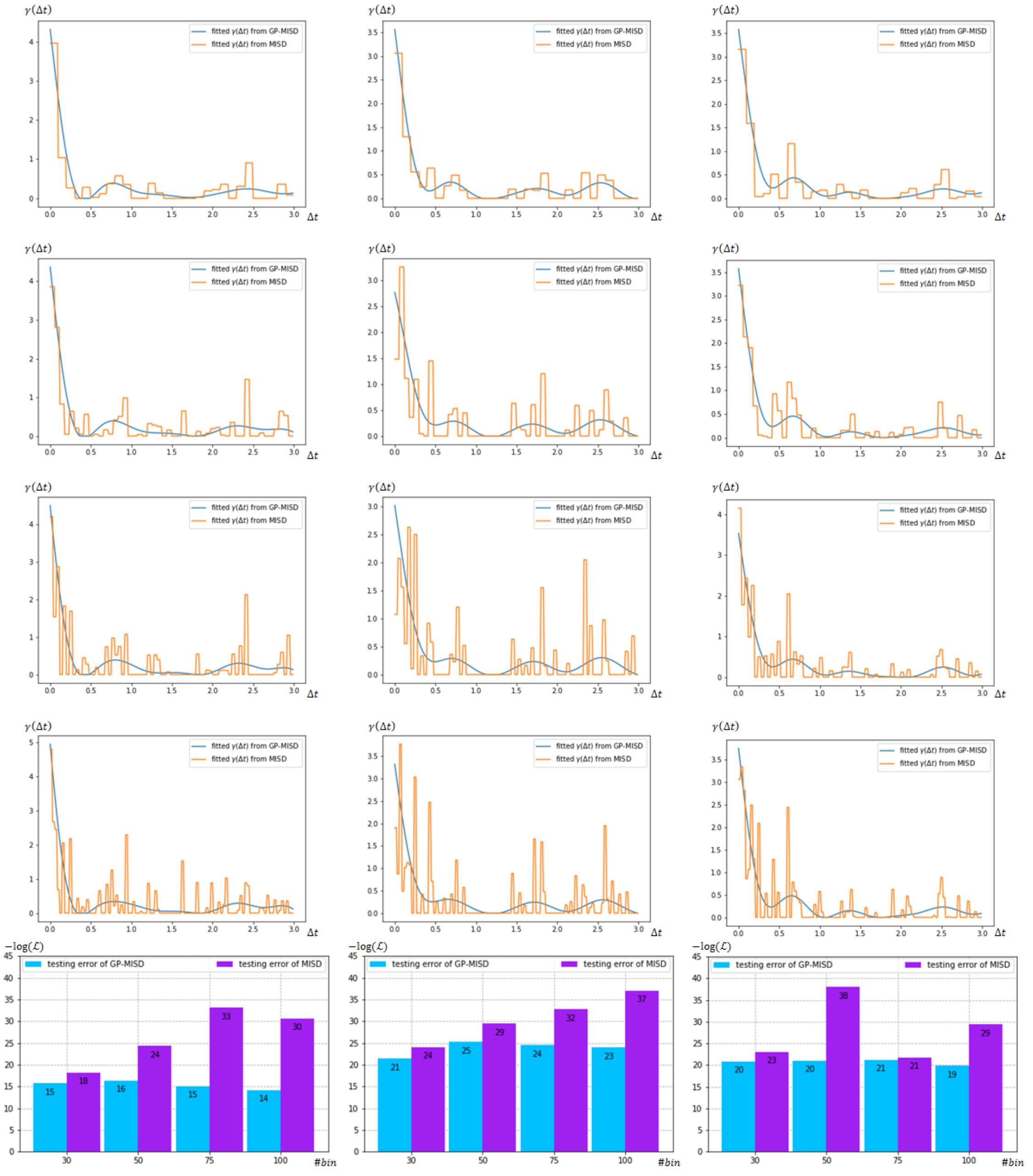
3.5 Summary

In this chapter, a refined MISD algorithm for Hawkes process was proposed, namely the GP-MISD algorithm which can effectively avoid overfitting when more bins are used. The key contribution of embedding a Gaussian process regression into EM iterations actually can be applied to many other algorithms based on bins, resulting in a smoother effect that avoids overfitting. GP-MISD inherits the advantages of MISD of estimating the baseline intensity and triggering kernel without any prior knowledge of the functional form. Experiments were performed on both synthetic and real datasets demonstrating that GP-MISD is less sensitive to the choice of the number of bins and produced consistently superior results to MISD. Although the GP-MISD algorithm is not in itself very complex, it constitutes an important component of the nonstationary Hawkes process presented in Chapter 6.



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

Figure 3.3: Motor Vehicle Collisions in New York City: rows 1-4 show the learned $\phi(\tau)$ from MISD and GP-MISD based on #bin=20, 50, 80, 100 respectively (upper, time unit is 1.16 day), and row 5 shows test errors of both algorithms for #bin=20, 50, 80, 100 (lower). Columns 1-3 stand for Manhattan, Queens, Bronx respectively.



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

Figure 3.4: NYPD Complaint Data 2017: rows 1-4 show the learned $\phi(\tau)$ from MISD and GP-MISD based on #bin=30, 50, 75, 100 respectively (upper, time unit is 11.6 days), and row 5 shows test errors of both algorithms for #bin=30, 50, 75, 100 (lower). Columns 1-3 stand for Manhattan, Queens, Brooklyn respectively.

Chapter 4

Nonparametric Hawkes Process Modulated by Quadratic Gaussian Process*

The GP-MISD algorithm in Chapter 3 provides a frequentist solution to the nonparametric Hawkes process, with the GP regression step used only as a smoother. In this chapter, another type of nonparametric method for the Hawkes process is introduced, namely the Bayesian nonparametric. More specifically, the Gaussian Process modulated Hawkes process is proposed, wherein a GP prior is incorporated for the baseline intensity and triggering kernel.

To guarantee the non-negativity of intensity, trajectories drawn from a GP prior need to be squashed by a link function. Depending on the link function used, two different Bayesian nonparametric Hawkes process models are proposed: (1) quadratic GP Hawkes process where the link function is a square transformation, and (2) sigmoid GP Hawkes process where the link function is a scaled sigmoid function.

*Portions of this chapter have been deposited in arXiv: F. Zhou, Z. Li, X. Fan, Y. Wang, A. Sowmya, F. Chen, Efficient EM-Variational Inference for Hawkes Process, arXiv preprint arXiv:1905.12251. 29 May 2019.

In Sec. 4.1, the issue of Bayesian nonparametric Hawkes process is introduced. The quadratic GP Hawkes process model is described in Sec. 4.2. In Sec. 4.3, an EM based variational Gaussian approximation inference method is proposed to approximate the true posterior with a Gaussian distribution. Some accelerating methods are proposed in Sec. 4.4. Synthetic and real data experiments are discussed in Sec. 4.5. The hyperparameter settings are discussed on Sec. 4.6, with the summary in Sec. 4.7. In Chapter 5, the sigmoid GP Hawkes process will be presented.

4.1 Overview

The Hawkes process has been used as an intensity estimator in a wide range of domains, including social networks [41], criminology [42] and financial engineering [43]. One of the key challenges for the Hawkes process is to select the function for baseline intensity and triggering kernel. The vanilla Hawkes process assumes a constant value for the baseline intensity and parametric function for the triggering kernel e.g. the exponential decay or power-law decay function. The parametric assumption introduces convenience to inference, but is inconsistent with many real applications, e.g. the baseline intensity of civilian deaths due to insurgent activity is changing over time [8] and the triggering kernel of vehicle collision is a periodic decay function [9]. To avoid the necessity for model selection and to model the baseline intensity and triggering kernel with continuous change, a Bayesian nonparametric model for Hawkes process is proposed in this chapter. Any formulated assumption for both baseline intensity and triggering kernel is avoided. The Bayesian priors on both components are some transformation of a GP that guarantees the nonnegativity constraint.

In a naïve Bayesian framework, given the observation D, the posterior of $\mu(t)$

and $\phi(\tau)$ is

$$p(\mu(t), \phi(\tau)|D) = \frac{p(D|\mu(t) = \zeta(f), \phi(\tau) = \zeta(g))\mathcal{GP}(f)\mathcal{GP}(g)}{\iint p(D|\zeta(f), \zeta(g))\mathcal{GP}(f)\mathcal{GP}(g)dfdg}, \tag{4.1}$$

where $\zeta(\cdot)$ is the link function to guarantee the nonnegativity constraint, and f and g are two functions drawn from the corresponding GP priors. In practice, inference of the posterior is non-trivial because of the doubly-intractable problem [16] caused by intractable integrals in the numerator and denominator. This problem is circumvented tactfully in the quadratic GP Hawkes process model.

4.2 Quadratic GP Hawkes Process

To obtain a nonparametric model, a quadratic GP Hawkes process (QGPHP) model is proposed, whose baseline intensity and triggering kernel are the quadratic transformation of random trajectories drawn from GP priors to guarantee non-negativity:

$$p(\mu(t), \phi(\tau)|D) = \frac{p(D|\mu(t) = f^2(t), \phi(\tau) = g^2(\tau))\mathcal{GP}(f)\mathcal{GP}(g)}{\iint p(D|f^2(t), g^2(\tau)))\mathcal{GP}(f)\mathcal{GP}(g)dfdg}, \tag{4.2}$$

where f and g are two functions drawn from the corresponding GP prior. The quadratic link function [17, 44] is used because the inference can be performed in closed form and it retains the connection between the data and the variational uncertainty.

However, the inference is challenging due to two reasons:

1. The baseline intensity is coupled with the triggering kernel in the likelihood of the Hawkes process, which drastically increases the complexity of performing inference. To address this issue, the branching structure is augmented to decouple them. The branching structure is a latent variable and estimated via an EM algorithm.

2. Although variational Gaussian approximation has been used for the Poisson process [17], the inference is performed by high dimensional numerical optimization which is time-consuming.

To circumvent these two problems, it is propose to embed a variational Gaussian approximation into the EM framework, provide some dimensionality reduction methods and derive a closed-form matrix derivative to speed up the inference. Specifically, the following contributions are made by this work:

1. The baseline intensity and triggering kernel are both modelled as nonparametric functions modulated by quadratic transformation of GP.

2. The latent branching structure of the Hawkes process is augmented to decouple the baseline intensity and triggering kernel in the likelihood.

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

4. Sparse GP approximation is utilized to derive a closed-form matrix derivative of ELBO to further accelerate EMV to be efficient.

4.3 Inference

In the inference section, the sparse GP approximation is first introduced to accelerate the inference in Sec. 4.3.1. The branching structure is augmented to decouple baseline intensity part and triggering kernel part in the likelihood in Sec. 4.3.2. The variational Gaussian approximation is proposed for both parts in Sec. 4.3.3. Combining the branching structure and variational Gaussian approximation, the EM-variational algorithm is obtained in Sec. 4.3.4.

4.3.1 Sparse GP Approximation

To improve inference efficiency and avoid the infinite dimensional issue, the sparse GP approximation [45] is used to introduce some inducing points whose definition can be found in [45]. f and g are dependent on their corresponding inducing points $\mathbf{Z}_f = \{z_f^m\}_{m=1}^{M_f}$ 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 have stationary Gaussian distributions $\mathbf{u}_f \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{z_f z_f})$ and $\mathbf{u}_g \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{z_g z_g})$ respectively. Given samples $\mathbf{u}_f$ 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:

$$\begin{aligned} v_f(t) &= \mathbf{k}_{tz_f}\mathbf{K}_{z_f z_f}^{-1}\mathbf{u}_f, \Sigma_f(t, t') = \mathbf{K}_{tt'} - \mathbf{k}_{tz_f}\mathbf{K}_{z_f z_f}^{-1}\mathbf{k}_{z_f t'} \\ v_g(\tau) &= \mathbf{k}_{\tau z_g}\mathbf{K}_{z_g z_g}^{-1}\mathbf{u}_g, \Sigma_g(\tau, \tau') = \mathbf{K}_{\tau\tau'} - \mathbf{k}_{\tau z_g}\mathbf{K}_{z_g z_g}^{-1}\mathbf{k}_{z_g \tau'} \end{aligned} \tag{4.3}$$

with $\mathbf{k}_{tz_f}$ and $\mathbf{k}_{\tau z_g}$ being the kernel vectors w.r.t. observations and inducing points while $\mathbf{K}_{z_f z_f}$, $\mathbf{K}_{z_g z_g}$, $\mathbf{K}_{tt'}$ and $\mathbf{K}_{\tau\tau'}$ are w.r.t. inducing points or observations only. Therefore, the joint distribution is

$$p(D, f, u_f, g, u_g) = p(D|\mu(t) = f^2, \phi(\tau) = g^2)p(f|\mathbf{u}_f)p(g|\mathbf{u}_g)p(\mathbf{u}_f)p(\mathbf{u}_g). \tag{4.4}$$

4.3.2 Augmentation of Branching Structure

For variational inference, the ELBO needs to be obtained, which means f, u_f, g and u_g need to be integrated out in Eq. 4.4. However, performing this procedure directly is difficult because $\mu(t)$ is coupled with $\phi(\tau)$ in the likelihood. To facilitate inference, the branching structure of Hawkes process is introduced to decouple $\mu(t)$

and $\phi(\tau)$. The branching structure $\mathbf{B}$ was already introduced in Sec 2.1.1:

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

After introducing the branching structure, the joint likelihood is obtained (Eq. 2.3). Taking expected value of the log joint likelihood w.r.t. the branching structure, a lower-bound of the log-likelihood is obtained:

$$\begin{aligned} &\mathcal{Q}(\mu(t), \phi(\tau) | \mu^{(s)}(t), \phi^{(s)}(\tau)) \\ &= \mathbb{E}_{\mathbf{B}} \left[\log p(D, \mathbf{B} | \mu(t), \phi(\tau)) \right] \\ &= \underbrace{\left[\sum_{i=1}^{N} p_{ii} \log(\mu(t_i)) \right] - \int_0^T \mu(t) dt}_{\mu(t) \text{ part}} + \\ &\underbrace{\sum_{i=2}^{N} \left[\sum_{j=1}^{i-1} p_{ij} \log\left(\phi(t_i - t_j)\right) \right] - \sum_{i=1}^{N} \int_{t_i}^{t_i + T_\phi} \phi(t - t_i) dt}_{\phi(\tau) \text{ part}} \\ &\triangleq \log \tilde{p}(D | \mu(t), \mathcal{P}_{ii}) + \log \tilde{p}(D | \phi(\tau), \mathcal{P}_{ij}), \end{aligned} \tag{4.5}$$

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

$$\begin{aligned} p_{ij} &= \frac{\phi^{(s)}(\tau_{ij})}{\mu^{(s)}(t_i) + \sum_{j=1}^{i-1} \phi^{(s)}(\tau_{ij})}, \\ p_{ii} &= \frac{\mu^{(s)}(t_i)}{\mu^{(s)}(t_i) + \sum_{j=1}^{i-1} \phi^{(s)}(\tau_{ij})}. \end{aligned} \tag{4.6}$$

Proof of Lower-bound The lower-bound $Q(\mu(t),\phi(\tau)|\mu^{(s)}(t),\phi^{(s)}(\tau))$ in Eq. 4.5 is derived as follows. Based on Jensen's inequality,

$$\begin{aligned}
&\log p(D|\mu(t),\phi(\tau)) = \sum_{i=1}^{N}\log\left(\mu(t_i)+\sum_{j=1}^{i-1}\phi(t_i-t_j)\right)-\int_0^T\left(\mu(t)+\sum_{t_i<t}\phi(t-t_i)\right)dt\\
&\geq \sum_{i=1}^{N}\left(p_{ii}\log\frac{\mu(t_i)}{p_{ii}}+\sum_{j=1}^{i-1}p_{ij}\log\frac{\phi(t_i-t_j)}{p_{ij}}\right)-\int_0^T\mu(t)dt-\sum_{i=1}^{N}\int_{t_i}^{t_i+T_\phi}\phi(t-t_i)dt\\
&=\sum_{i=1}^{N}p_{ii}\log\mu(t_i)-\int_0^T\mu(t)dt+\sum_{i=2}^{N}\sum_{j=1}^{i-1}p_{ij}\log\phi(t_i-t_j)-\sum_{i=1}^{N}\int_{t_i}^{t_i+T_\phi}\phi(t-t_i)dt+C
\end{aligned} \tag{4.7}$$

where C is a constant.

4.3.3 Variational Gaussian Approximation

Due to the decoupling of $\mu(t)$ and $\phi(\tau)$, the inference can be performed for two components independently.

4.3.3.1 Baseline Intensity Part

Consider the $\mu(t)$ part: $\log\tilde{p}(D|\mu(t)=f^2,\mathcal{P}_{ii})$. $\mathcal{P}_{ii}$ means the diagonal entries of $\mathcal{P}=\mathbb{E}(\mathbf{B})$ and $\mathcal{P}_{ij}$ means the others off diagonal. The inducing points $\mathbf{u}_f$ are integrated out using a variational distribution $q(\mathbf{u}_f)=\mathcal{N}(\mathbf{u}_f|\mathbf{m}_f,\mathbf{S}_f)$ where $\mathbf{S}_f$ is positive-semidefinite and symmetric. Jensen's inequality is used to obtain the ELBO of $\mu(t)$ part:

$$\begin{aligned}
&\log \tilde{p}(D|\mathcal{P}_{ii})\\
&= \log\left[\iint \tilde{p}(D|f,\mathcal{P}_{ii})p(f|\mathbf{u}_f)p(\mathbf{u}_f)\frac{q(\mathbf{u}_f)}{q(\mathbf{u}_f)}d\mathbf{u}_f df\right]\\
&\geq \iint p(f|\mathbf{u}_f)q(\mathbf{u}_f)d\mathbf{u}_f \log\tilde{p}(D|f,\mathcal{P}_{ii})df + \iint p(f|\mathbf{u}_f)q(\mathbf{u}_f)df\log\left[\frac{p(\mathbf{u}_f)}{q(\mathbf{u}_f)}\right]d\mathbf{u}_f\\
&= \mathbb{E}_{q(f)}\left[\log\tilde{p}(D|f,\mathcal{P}_{ii})\right] - \mathrm{KL}\left(q(\mathbf{u}_f)||p(\mathbf{u}_f)\right)\\
&\triangleq \mathrm{ELBO}_\mu,
\end{aligned} \tag{4.8}$$

where

$$q(f) = \int p(f|\mathbf{u}_f)q(\mathbf{u}_f)d\mathbf{u}_f = \mathcal{GP}(f|\tilde{v}_f(t),\tilde{\Sigma}_f(t,t')) \tag{4.9}$$

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

$$\begin{aligned}
&\mathbb{E}_{q(f)}\left[\log\tilde{p}(D|f,\mathcal{P}_{ii})\right]\\
&= \sum_{i=1}^{N} p_{ii}\mathbb{E}_{q(f)}\left[\log f^2(t_i)\right] - \int_0^T \left\{\mathbb{E}^2_{q(f)}[f(t)] + \mathrm{Var}_{q(f)}[f(t)]\right\}dt,
\end{aligned} \tag{4.10}$$

where $\mathbb{E}(A^2) = \mathbb{E}^2(A) + \mathrm{Var}(A)$. Eq. 4.10 has an analytical solution which is shown as follows.

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

$$\mathrm{KL}\left(q(\mathbf{u}_f)||p(\mathbf{u}_f)\right) = \frac{1}{2}\left[\mathrm{Tr}(\mathbf{K}_{z_fz_f}^{-1}\mathbf{S}_f) + \log\frac{|\mathbf{K}_{z_fz_f}|}{|\mathbf{S}_f|} - M_f + \mathbf{m}_f^T\mathbf{K}_{z_fz_f}^{-1}\mathbf{m}_f\right], \tag{4.11}$$

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

The last two terms in Eq. 4.10 have analytical solutions [17]:

$$\int_0^T \mathbb{E}^2_{q(f)}[f(t)]dt = \mathbf{m}_f^T \mathbf{K}_{z_f z_f}^{-1} \boldsymbol{\Psi}_f \mathbf{K}_{z_f z_f}^{-1} \mathbf{m}_f, \tag{4.12}$$

$$\int_0^T \mathrm{Var}_{q(f)}[f(t)]dt = \theta_0^f T - \mathrm{Tr}(\mathbf{K}_{z_f z_f}^{-1} \boldsymbol{\Psi}_f) + \mathrm{Tr}(\mathbf{K}_{z_f z_f}^{-1} \mathbf{S}_f \mathbf{K}_{z_f z_f}^{-1} \boldsymbol{\Psi}_f), \tag{4.13}$$

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 [17]:

$$\begin{aligned}&\Psi_f(z_f, z_f') \\ &= -\frac{(\theta_0^f)^2}{2}\sqrt{\frac{\pi}{\theta_1^f}} \exp\left(-\frac{\theta_1^f(z_f - z_f')^2}{4}\right)\left[\mathrm{erf}\left(\sqrt{\theta_1^f}(\bar{z}_f - T)\right) - \mathrm{erf}\left(\sqrt{\theta_1^f}\bar{z}_f\right)\right],\end{aligned} \tag{4.14}$$

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

The first term in Eq. 4.10 also has an analytical solution [17]:

$$\begin{aligned}\mathbb{E}_{q(f)}\left[\log f^2(t_i)\right] &= \int_{-\infty}^{\infty} \log f^2(t_i)\mathcal{N}(f(t_i)|\tilde{v}_f(t_i), \tilde{\sigma}_f^2(t_i))df(t_i) \\ &= -\tilde{G}\left(-\frac{\tilde{v}_f^2(t_i)}{2\tilde{\sigma}_f^2(t_i)}\right) + \log\left(\frac{\tilde{\sigma}_f^2(t_i)}{2}\right) - C_{EM},\end{aligned} \tag{4.15}$$

where $\tilde{\sigma}_f^2(t_i)$ is the diagonal entry of $\tilde{\Sigma}_f(t, t')$ in Eq. 4.9 at t_i, C_{EM} is the Euler-Mascheroni constant 0.57721566 and $\tilde{G}(z)$ is a special case of the partial derivative of the confluent hyper-geometric function ${}_1F_1(a, b, z)$ [17]:

$$\tilde{G}(z) = {}_1F_1{}^{(1,0,0)}(0, 0.5, z). \tag{4.16}$$

It is worth noting that $\tilde{G}(z)$ does not need to be computed for inference. Actually it is sufficient to know that $\tilde{G}(0) = 0$ because $\mathbf{m}_f^* = \mathbf{0}$ (see Sec. 4.4).

4.3.3.2 Triggering Kernel Part

For the $\phi(\tau)$ part: $\log \tilde{p}(D|\phi(\tau) = g^2, \mathcal{P}_{ij})$. Similarly, the inducing points $\mathbf{u}_g$ are integrated out using $q(\mathbf{u}_g) = \mathcal{N}(\mathbf{u}_g|\mathbf{m}_g, \mathbf{S}_g)$ where $\mathbf{S}_g$ is positive-semidefinite and symmetric. The ELBO of $\phi(\tau)$ part is

$$\begin{aligned}\log \tilde{p}(D|\mathcal{P}_{ij}) &= \log \left[\iint \tilde{p}(D|g, \mathcal{P}_{ij}) p(g|\mathbf{u}_g) p(\mathbf{u}_g) \frac{q(\mathbf{u}_g)}{q(\mathbf{u}_g)} d\mathbf{u}_g dg \right] \\ &\geq \mathbb{E}_{q(g)} \left[\log \tilde{p}(D|g, \mathcal{P}_{ij}) \right] - \mathrm{KL}\left(q(\mathbf{u}_g) || p(\mathbf{u}_g) \right) \\ &\triangleq \mathrm{ELBO}_{\phi},\end{aligned} \tag{4.17}$$

where $q(g)$ is Eq. 4.9 with notation f and t replaced by g and τ, respectively. The expectation of log-likelihood over $q(g)$ can be written as

$$\begin{aligned}&\mathbb{E}_{q(g)} \left[\log \tilde{p}(D|g, \mathcal{P}_{ij}) \right] \\ &= \sum_{i=2}^{N} \sum_{j=1}^{i-1} p_{ij} \mathbb{E}_{q(g)} \left[\log g^2(\tau_{ij}) \right] - \sum_{i=1}^{N} \int_0^{T_\phi} \left\{ \mathbb{E}^2_{q(g)}[g(\tau)] + \mathrm{Var}_{q(g)}[g(\tau)] \right\} d\tau.\end{aligned} \tag{4.18}$$

Eq. 4.18 can be solved analytically using the same method as $\mu(t)$ part.

4.3.4 EM-Variational Algorithm

The motivation for augmenting the branching structure should now be clear. By doing so, a surrogate function (lower-bound) is obtained that decouples $\mu(t)$ and $\phi(\tau)$ to two independent components (E step). For each component, variational Gaussian approximation is utilized to derive an ELBO which should be maximized, thus obtaining an optimal variational distribution (M step).

Throughout this work, the squared exponential kernel $k(x, x') = \theta_0 \exp\left(-\frac{\theta_1}{2}\|x - x'\|^2\right)$ is used as the GP covariance kernel. The hyperparameters θ_0 and θ_1 can be optimized by performing maximization of ELBO over $\{\theta_0, \theta_1\}$ using numerical packages. Normally, $\{\theta_0, \theta_1\}$ are updated every 20 iterations. Apart from $\{\theta_0, \theta_1\}$, the hyperparameters left are the number and location of inducing points. Theoretically, the number M and location $\mathbf{Z}$ of inducing points affect the computation complexity and the estimation quality of $\mu(t)$ and $\phi(\tau)$. If M is too large, the inducing points kernel matrix $\mathbf{K}_{zz}$ will be a large matrix which leads to high complexity. If M is too small, the inducing points cannot capture the dynamics of $\mu(t)$ or $\phi(\tau)$.

For fast inference, it is assumed that the inducing points are uniformly located in the domain. Another advantage of uniform location is that the kernel matrix $\mathbf{K}_{zz}$ has Toeplitz structure which means that the matrix inversion can be implemented in $\mathcal{O}(M \log^2 M)$ [46] instead of $\mathcal{O}(M^3)$ in a naïve implementation.

The number of inducing points depends on the application. If $\mu(t)$ or $\phi(\tau)$ is a volatile function, more points are needed to capture the dynamics. In experiments, preliminary runs are performed to gradually increase the number of inducing points and stopped when the resulting $\mu(t)$ or $\phi(\tau)$ is not improved much any more. The pseudo code of naïve EMV is shown in Alg. 4.1.

Algorithm 4.1: Naïve EMV algorithm for QGPHP

```
Result: $\mu(t)$, $\phi(\tau)$
Initialize hyperparameters and $\mathcal{P}$;
for do
    Update $\mathcal{P}$ by Eq. 4.6;
    Update $\mathbf{m}_f^*$, $\mathbf{S}_f^*$, $\mathbf{m}_g^*$ and $\mathbf{S}_g^*$ by $\mathbf{m}_f^*, \mathbf{S}_f^* = \text{argmax}_{\mathbf{m}_f, \mathbf{S}_f} (\text{ELBO}_\mu)$ and
      $\mathbf{m}_g^*, \mathbf{S}_g^* = \text{argmax}_{\mathbf{m}_g, \mathbf{S}_g} (\text{ELBO}_\phi)$;
    Update $\tilde{v}_f^*$, $\tilde{\Sigma}_f^*$, $\tilde{v}_g^*$ and $\tilde{\Sigma}_g^*$ by Eq. 4.9 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 $\tilde{\sigma}_f^{2*}$
      and $\tilde{\sigma}_g^{2*}$ are diagonal entries of $\tilde{\Sigma}_f^*$ and $\tilde{\Sigma}_g^*$;
    Update hyperparameters.
end
```

4.4 Inference Speed Up

The naïve implementation of EMV algorithm in Alg. 4.1 is time-consuming. The bottleneck is the update for $\mathbf{m}_f^*$, $\mathbf{S}_f^*$, $\mathbf{m}_g^*$ and $\mathbf{S}_g^*$ because numerical optimization

has to be performed. 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 $M_f + M_f(M_f + 1)/2$. This is a large space even when M_f is small, and the case of $\mathbf{u}_g$ is the same. Two tricks were developed to speed up the algorithm:
(1) it is shown that $\mathbf{m}^*$ does not need to be inferred, and
(2) the closed-form matrix derivative of ELBO is derived w.r.t. $\mathbf{S}$, which means that a local maximum $\mathbf{S}^*$ can be obtained directly instead of performing numerical optimization.

The transformation function is $\mu(t) = f^2$, which is not a bijection. For every $\mu(t)$, there are 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 it is straightforward to see that the likelihood is symmetric, i.e. the likelihood is the same with f and $-f$. For the prior $\mathcal{GP}(f|\mathbf{u}_f)\mathcal{N}(\mathbf{u}_f|\mathbf{0}, \mathbf{K}_{z_f z_f})$, $\mathbf{u}_f$ can be integrated out and the marginal distribution over f is still Gaussian with mean $\mathbf{0}$. Therefore, the prior of f is also symmetric. Conclusively, the posterior $p(f|D, \mathcal{P}_{ii})$ is symmetric. By variational Gaussian approximation, $p(f|D, \mathcal{P}_{ii})$ is approximated by a normal distribution $q(f) = \mathcal{GP}(f|\tilde{v}_f(t), \tilde{\Sigma}_f(t, t'))$ where $\tilde{v}_f(t) = \mathbf{k}_{tz_f}\mathbf{K}_{z_f z_f}^{-1}\mathbf{m}_f$. Therefore, $\mathbf{m}_f^* = \mathbf{0}$ definitely. This applies to the $\phi(\tau)$ part as well to obtain $\mathbf{m}_g^* = \mathbf{0}$.

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. The closed-form matrix derivative of ELBO is derived over $\mathbf{S}$, which is shown in Sec. 4.4.1.

4.4.1 Matrix Derivative of ELBO

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

$$\begin{aligned}\mathrm{ELBO}_\mu = & -\left(\theta_0^f T - \mathrm{Tr}(\mathbf{K}_{z_f z_f}^{-1}\boldsymbol{\Psi}_f) + \mathrm{Tr}(\mathbf{K}_{z_f z_f}^{-1}\mathbf{S}_f\mathbf{K}_{z_f z_f}^{-1}\boldsymbol{\Psi}_f)\right) \\ & + \sum_{i=1}^{N} p_{ii}\left(-\tilde{G}(0) + \log(\tilde{\sigma}_f^2(t_i)) - \log 2 - C\right) \\ & - \frac{1}{2}\left(\mathrm{Tr}(\mathbf{K}_{z_f z_f}^{-1}\mathbf{S}_f) + \log|\mathbf{K}_{z_f z_f}| - \log|\mathbf{S}_f| - M_f\right).\end{aligned} \tag{4.19}$$

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

$$\begin{aligned}\frac{\partial\mathrm{ELBO}_\mu}{\partial\mathbf{S}_f} = & -(2\mathbf{K}_{z_f z_f}^{-1}\Psi_f\mathbf{K}_{z_f z_f}^{-1} - \mathbf{K}_{z_f z_f}^{-1}\Psi_f\mathbf{K}_{z_f z_f}^{-1}\circ\mathbf{I}) \\ + & \sum_{i=1}^{N} p_{ii}\left(2\mathbf{K}_{z_f z_f}^{-1}\mathbf{k}_{z_f t_i}\mathbf{k}_{t_i z_f}\mathbf{K}_{z_f z_f}^{-1} - \mathbf{K}_{z_f z_f}^{-1}\mathbf{k}_{z_f t_i}\mathbf{k}_{t_i z_f}\mathbf{K}_{z_f z_f}^{-1}\circ\mathbf{I}\right)/\tilde{\sigma}_f^2(t_i) \\ - & \frac{1}{2}\left(2\mathbf{K}_{z_f z_f}^{-1} - \mathbf{K}_{z_f z_f}^{-1}\circ\mathbf{I} - (2\mathbf{S}_f^{-1} - \mathbf{S}_f^{-1}\circ\mathbf{I})\right),\end{aligned} \tag{4.20}$$

where $\mathbf{I}$ means the identity matrix, $\circ$ means 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. 4.9.

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

$$\begin{aligned}\frac{\partial\mathrm{ELBO}_\mu}{\partial\mathbf{S}_f} = & -\mathbf{K}_{z_f z_f}^{-1}\Psi_f\mathbf{K}_{z_f z_f}^{-1}\circ\mathbf{I} \\ & + \sum_{i=1}^{N} p_{ii}\frac{\mathbf{K}_{z_f z_f}^{-1}\mathbf{k}_{z_f t_i}\mathbf{k}_{t_i z_f}\mathbf{K}_{z_f z_f}^{-1}\circ\mathbf{I}}{\tilde{\sigma}_f^2(t_i)} - \frac{1}{2}\left(\mathbf{K}_{z_f z_f}^{-1}\circ\mathbf{I} - \mathbf{S}_f^{-1}\right).\end{aligned} \tag{4.21}$$

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

$$\begin{aligned} \mathrm{ELBO}_\phi = & -\sum_{i=1}^{N}\Big(\theta_0^g T_\phi - \mathrm{Tr}(\mathbf{K}_{z_g z_g}^{-1}\mathbf{\Psi}_g) + \mathrm{Tr}(\mathbf{K}_{z_g z_g}^{-1}\mathbf{S}_g\mathbf{K}_{z_g z_g}^{-1}\mathbf{\Psi}_g)\Big) \\ & + \sum_{i=2}^{N}\sum_{j=1}^{i-1} p_{ij}\Big(-\tilde{G}(0) + \log(\tilde{\sigma}_g^2(\tau_{ij})) - \log 2 - C\Big) \\ & - \frac{1}{2}\Big(\mathrm{Tr}(\mathbf{K}_{z_g z_g}^{-1}\mathbf{S}_g) + \log|\mathbf{K}_{z_g z_g}| - \log|\mathbf{S}_g| - M_g\Big). \end{aligned} \tag{4.22}$$

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

$$\begin{aligned} \frac{\partial \mathrm{ELBO}_\phi}{\partial \mathbf{S}_g} = & -\sum_{i=1}^{N}(2\mathbf{K}_{z_g z_g}^{-1}\Psi_g\mathbf{K}_{z_g z_g}^{-1} - \mathbf{K}_{z_g z_g}^{-1}\Psi_g\mathbf{K}_{z_g z_g}^{-1}\circ\mathbf{I}) \\ + \sum_{i=2}^{N}\sum_{j=1}^{i-1} p_{ij} & \Bigg(2\mathbf{K}_{z_g z_g}^{-1}\mathbf{k}_{z_g\tau_{ij}}\mathbf{k}_{\tau_{ij}z_g}\mathbf{K}_{z_g z_g}^{-1} - \mathbf{K}_{z_g z_g}^{-1}\mathbf{k}_{z_g\tau_{ij}}\mathbf{k}_{\tau_{ij}z_g}\mathbf{K}_{z_g z_g}^{-1}\circ\mathbf{I}\Bigg)/\tilde{\sigma}_g^2(\tau_{ij}) \\ - \frac{1}{2} & \Big(2\mathbf{K}_{z_g z_g}^{-1} - \mathbf{K}_{z_g z_g}^{-1}\circ\mathbf{I} - (2\mathbf{S}_g^{-1} - \mathbf{S}_g^{-1}\circ\mathbf{I})\Big), \end{aligned} \tag{4.23}$$

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

$$\begin{aligned} \frac{\partial \mathrm{ELBO}_\phi}{\partial \mathbf{S}_g} = & -\sum_{i=1}^{N}\mathbf{K}_{z_g z_g}^{-1}\Psi_g\mathbf{K}_{z_g z_g}^{-1}\circ\mathbf{I} \\ & + \sum_{i=2}^{N}\sum_{j=1}^{i-1} p_{ij}\frac{\mathbf{K}_{z_g z_g}^{-1}\mathbf{k}_{z_g\tau_{ij}}\mathbf{k}_{\tau_{ij}z_g}\mathbf{K}_{z_g z_g}^{-1}\circ\mathbf{I}}{\tilde{\sigma}_g^2(\tau_{ij})} - \frac{1}{2}\Big(\mathbf{K}_{z_g z_g}^{-1}\circ\mathbf{I} - \mathbf{S}_g^{-1}\Big). \end{aligned} \tag{4.24}$$

From the proof above, it is straightforward to see that if $\mathbf{S}_f$ is symmetric, $\partial\mathrm{ELBO}_\mu/\partial\mathbf{S}_f = \mathbf{0}$ is a nonlinear system consisting of $M_f(M_f+1)/2$ equations, which is still inefficient due to too many simultaneous equations. To further accelerate the inference, assume $q(\mathbf{u}_f)$ is an independent distribution (mean field approximation [37]) which means $\mathbf{S}_f$ is diagonal. In the diagonal case, $\partial\mathrm{ELBO}_\mu/\partial\mathbf{S}_f = \mathbf{0}$ is a nonlinear system consisting of M_f equations which can be solved faster. In experiments, it was found that this assumption does not make much difference when $\mu(t)$ is not a volatile function. The discussion above applies to the $\phi(\tau)$ part as well. The accelerated EMV is provided in Alg. 4.2.

```
Algorithm 4.2: Accelerated EMV (time-changing $\mu(t)$)
Result: $\mu(t)$, $\phi(\tau)$
Initialize hyperparameters and $\mathcal{P}$;
for do
    Update $\mathcal{P}$ by Eq. 4.6;
    Update $\mathbf{S}_f^*$ and $\mathbf{S}_g^*$ by $\partial\text{ELBO}_\mu/\partial\mathbf{S}_f = \mathbf{0}$ and $\partial\text{ELBO}_\phi/\partial\mathbf{S}_g = \mathbf{0}$;
    Update $\tilde{\Sigma}_f^*$ and $\tilde{\Sigma}_g^*$ by Eq. 4.9 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 $\tilde{\sigma}_f^{2*}$ and $\tilde{\sigma}_g^{2*}$
      are diagonal entries of $\tilde{\Sigma}_f^*$ and $\tilde{\Sigma}_g^*$;
    Update hyperparameters.
end
```

4.4.2 Constant Baseline Intensity

If $\mu(t)$ is constant μ, there is no need to compute the nonlinear system and variables of $\mu(t)$ part, e.g. $\mathbf{k}_{tz_f}$, $\mathbf{K}_{z_f z_f}$, $\mathbf{\Psi}_f$ and $\partial\text{ELBO}_\mu/\partial\mathbf{S}_f = \mathbf{0}$. μ could be estimated by $\mu = \sum_{i=1}^N p_{ii}/T$ in each EM iteration. Consequently, it is faster than the general case. The pseudo code is provided in Alg.4.3.

```
Algorithm 4.3: Accelerated EMV (constant $\mu$)
Result: $\mu$, $\phi(\tau)$
Initialize hyperparameters and $\mathcal{P}$;
for do
    Update $\mathcal{P}$ by Eq. 4.6;
    Update $\mathbf{S}_g^*$ by $\partial\text{ELBO}_\phi/\partial\mathbf{S}_g = \mathbf{0}$;
    Update $\tilde{\Sigma}_g^*$ by Eq. 4.9 with $\mathbf{S}_g^*$;
    Update $\mu$ and $\phi(\tau)$ by $\mu = \sum_{i=1}^N p_{ii}/T$ and $\phi(\tau) = \tilde{\sigma}_g^{2*}$ where $\tilde{\sigma}_g^{2*}$ are
      diagonal entries of $\tilde{\Sigma}_g^*$;
    Update hyperparameters.
end
```

4.4.3 Complexity

An advantage of sparse GP approximation is that the complexity of matrix inversion is fixed at $\mathcal{O}(M_f^3 + M_g^3)$ where M_f (or M_g) $\ll N$. This results in the complexity scaling almost linearly with data size: $\mathcal{O}(NL)$ where $L = \int_{T_\phi} \frac{\mu(t)}{1-\int \phi(\tau)d\tau} dt \ll N$ due to the sparsity of the branching structure: previous points that are more than T_ϕ units away from event i have no influence on event i ($p_{ij} = 0$).

In experiments on a normal desktop (CPU: i7-6700 with 8GB RAM), the naïve implementation (Alg. 4.1) costs about two hours for $N = 205$, 6 inducing points for both $\mathbf{Z}_f$ and $\mathbf{Z}_g$ and 100 EM iterations. The accelerated algorithms (Alg. 4.2 and 4.3) cost about 4 minutes and 2 minutes respectively in the same setting, which drastically reduces the running time.

4.5 Experimental Results

The performance of EMV is evaluated on both synthetic and real data. Specifically, the accelerated EMV Alg. 4.2 and 4.3 are compared with the following alternatives wherever applicable:

- **Gaussian-Cox (GC) process**: a GP modulated inhomogeneous Poisson process. The inference is performed by a published algorithm [15], which is only applicable to real data.

- **RKHS-Cox (RKHSC) process**: an inhomogeneous Poisson process whose intensity is estimated by a reproducing kernel Hilbert space formulation [44]. It is applicable only to real data.

- **Parametric Hawkes (PH) process**: the vanilla Hawkes process (constant μ and exponential triggering kernel $\alpha \exp(-\beta(t - t_i))$). The inference is performed by MLE.

- **Model Independent Stochastic Declustering (MISD)**: the MISD [8] is an EM based nonparametric algorithm for Hawkes process, where μ is constant and $\phi(\cdot)$ is a histogram function. MISD-# is used (# is the number of bins) to indicate the corresponding model.

- **Wiener-Hopf (WH)**: it is another nonparametric algorithm for Hawkes process where μ is constant and $\phi(\cdot)$ is a continuous function. The inference is based on the solution of a Wiener-Hopf equation [11].

The following metrics are used to evaluate the performance of various methods:

- ***LogLik***: the log-likelihood of test data using the trained model. This is a metric describing the model prediction ability. It is used to measure the performance on synthetic and real data.

- ***EstErr***: estimation error, defined as the integral of squared error between the learned $\hat{\phi}(\tau)$ ($\hat{\mu}(t)$) and the ground truth. It is only used for synthetic data.

- ***Q-Q plot***: quantile-quantile (Q-Q) plot, generated by transforming the real data timestamps by the fitted model according to the time rescaling theorem [47]. Generally speaking, the Q-Q plot visualizes the goodness-of-fit for different models. It is used to measure the performance on real data.

- ***PreAcc***: Given an event sequence $\{t_1, t_2, ..., t_{i-1}\}$, the aim is to predict the time of next event t_i. The t_i has density $P(t_i = t) = \lambda(t) \exp\left(-\int_{t_{i-1}}^{t} \lambda(s)ds\right)$. The expectation of t_i should be $\mathbb{E}[t_i] = \int_{t_{i-1}}^{\infty} tp(t_i = t)dt$. The integral in the equations can be estimated by Monte Carlo method. Multiple timestamps are predicted in a sequence: if the predicted $\hat{t}_i$ is within $t_i \pm \epsilon$ where t_i is the real timestamp and ϵ is an error bound, then it is considered to be a correct prediction; otherwise it is a wrong one. The percentage of correct predictions is defined as the prediction accuracy. It is used to measure the performance on real data.

4.5.1 Experimental Results on Synthetic Data

In synthetic data experiments, the performance of accelerated EMV inference Alg. 4.2 and 4.3 are compared with PH, MISD and WH (GC and RKHSC are excluded because they are Poisson process models and cannot provide μ and ϕ). Four cases are considered:

1. $\mu = 1$ and $\phi(\tau) = 1 \cdot \exp(-2\tau)$;
2. $\mu(t) = \begin{cases} 1 \ \ (0 < t \leq T/2) \\ 2 \ \ (T/2 < t < T) \end{cases}$ and $\phi(\tau) = 1 \cdot \exp(-2\tau)$;
3. $\mu = 1$ and $\phi(\tau) = \begin{cases} 0.25 \sin\tau \ \ (0 < \tau \leq \pi) \\ 0 \ \ (\pi < \tau < T_\phi) \end{cases}$;
4. $\mu(t) = \sin\left(\frac{2\pi}{T} \cdot t\right) + 1 \ (0 < t < T)$ and $\phi(\tau) = 0.3\left(\sin(\frac{2\pi}{3} \cdot \tau) + 1\right) \cdot \exp(-0.7\tau) \ (0 < \tau < T_\phi)$.

The thinning algorithm [38] is used to generate 100 sets of training data and 10 sets of test data with $T = 100$ in four cases. PH, MISD-10, MISD-20, WH and EMV (Alg. 4.2 is used for cases 2 and 4, Alg. 4.3 for case 1 and 3) to perform inference with $T_\phi = 6$. The learned $\hat{\mu}(t)$ and $\hat{\phi}(\tau)$ are shown in Fig. 4.1. The *EstErr* and *LogLik* are shown in Tab. 4.1.

Case 1 is a common one with constant μ and exponential decay $\phi(\tau)$ and 177 points were generated. For hyperparameters, the bandwidth of WH is set to 0.7 and there are 6 inducing points ($M_g = 6$) for EMV. The learned $\hat{\phi}(\tau)$'s are shown in Fig. 4.1. *EstErr* and *LogLik* are shown in Tab. 4.1. PH is the best in this case because the parametric model assumption matches the ground truth.

Case 2 has a non-constant $\mu(t)$ which is a piecewise constant function, and 333 points were generated. The bandwidth of WH is set to 1 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 learned results are shown in Fig. 4.1. *EstErr* and *LogLik* are shown in Tab. 4.1. In this case, EMV is the best for both $\mu(t)$ and $\phi(\tau)$ because other alternatives assume that $\mu(t)$ is a constant, which is inconsistent with the ground truth.

Case 3 has a half sinusoidal triggering kernel, and 181 points were generated. The bandwidth of WH is set to 0.5 and there are 10 inducing points ($M_g = 10$) for EMV. The learned $\hat{\phi}(\tau)$'s are shown in Fig. 4.1. *EstErr* and *LogLik* are shown in Tab. 4.1 in which EMV is still the best for prediction ability although the estimation error is large for μ. The result shows that EMV can learn the correct triggering kernel in non-monotonically decreasing cases.

Case 4 is a general case with time-changing $\mu(t)$ and sinusoidal exponential decay triggering kernel, and 212 points were generated. The bandwidth of WH is 0.9 and there are 6 inducing points on $\phi(\tau)$ ($M_g = 6$) and 8 inducing points on $\mu(t)$ ($M_f = 8$) for EMV. Learned results are shown in Fig. 4.1. *EstErr* and *LogLik* are shown in Tab. 4.1 and EMV is still the best.

Clearly, the EMV algorithm outperforms other alternatives in almost all cases except Case 1. This is because only EMV algorithm is capable of estimating non-parametric $\mu(t)$ and $\phi(\tau)$ concurrently; the reason that PH is the best in Case 1 is the parametric model assumption that matches the ground truth, which is a rare situation in real applications.

4.5.2 Experimental Results on Real Data

The EMV algorithm is applied to two different real datasets and the performance compared to the alternatives. Ground truth information is unavailable for the real world data, so *EstErr* cannot be utilized to measure the performance. As a result, the performance on real data is quantified by the metric of *LogLik*, *Q-Q plot* and *PreAcc*.

Motor Vehicle Collisions in New York City [39]: In this dataset, weekday records in nearly one month (Sep. 18th - Oct. 13th 2017) were filtered out. The

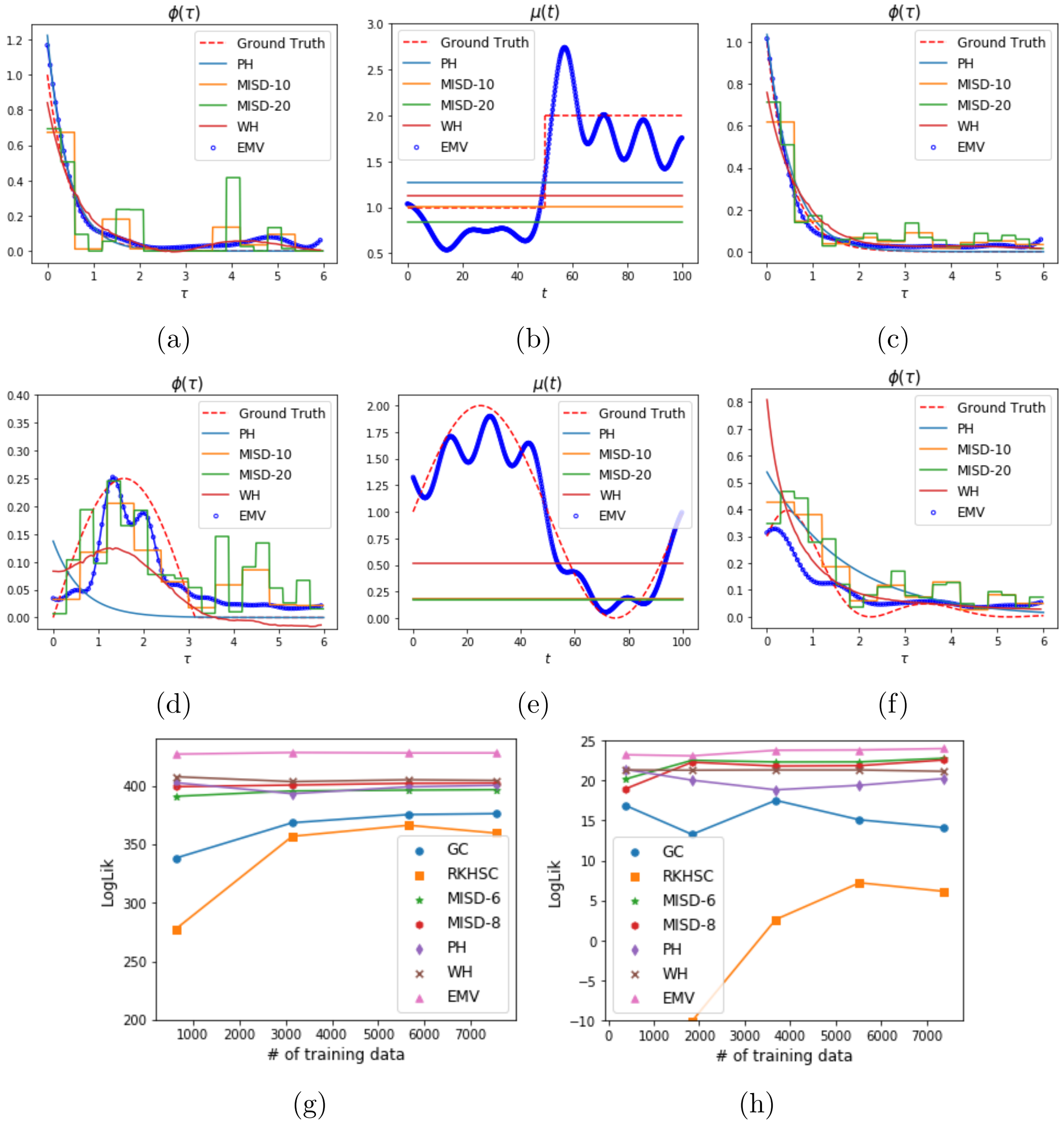


Figure 4.1: Experimental results of synthetic and real data. (a): The estimated $\hat{\phi}(\tau)$ in case 1 (the estimated $\hat{\mu}_{\text{PH}}$=0.973, $\hat{\mu}_{\text{MISD-10}}$=0.698, $\hat{\mu}_{\text{MISD-20}}$=0.620, $\hat{\mu}_{\text{WH}}$=0.762, $\hat{\mu}_{\text{EMV}}$=0.623); (b) and (c): the estimated $\hat{\mu}(t)$ and $\hat{\phi}(\tau)$ in case 2; (d): the estimated $\hat{\phi}(\tau)$ in case 3 (the estimated $\hat{\mu}_{\text{PH}}$=1.199, $\hat{\mu}_{\text{MISD-10}}$=1.039, $\hat{\mu}_{\text{MISD-20}}$=0.861, $\hat{\mu}_{\text{WH}}$=1.357, $\hat{\mu}_{\text{EMV}}$=1.239); (e) and (f): the estimated $\hat{\mu}(t)$ and $\hat{\phi}(\tau)$ in case 4; (g): the *LogLik* of various algorithms over the number of training data for vehicle collision dataset; (h): the *LogLik* for taxi pickup dataset.

Table 4.1: *EstErr* and *LogLik* of synthetic data. $\mu(t)$ is constant in cases 1, 3 and time-changing in cases 2, 4.

		PH	MISD-10	MISD-20	WH	EMV
Case 1	$EstErr(\hat{\mu}, \mu)$	**0.072**	9.116	14.381	5.621	14.196
	$EstErr(\hat{\phi}(\tau), \phi(\tau))$	**0.008**	0.075	0.106	0.009	0.015
	LogLik	**-37.91**	-41.87	-45.13	-38.71	-39.58
Case 2	$EstErr(\hat{\mu}(t), \mu(t))$	64.362	72.847	81.008	83.883	**10.946**
	$EstErr(\hat{\phi}(\tau), \phi(\tau))$	0.015	0.043	0.058	0.013	**0.002**
	LogLik	93.64	91.91	90.93	93.72	**96.85**
Case 3	$EstErr(\hat{\mu}, \mu)$	3.960	**0.155**	1.923	12.745	5.738
	$EstErr(\hat{\phi}(\tau), \phi(\tau))$	0.098	0.021	0.031	0.026	**0.013**
	LogLik	-70.18	-51.66	-51.59	-53.15	**-51.44**
Case 4	$EstErr(\hat{\mu}(t), \mu(t))$	107.951	114.033	118.016	60.106	**10.436**
	$EstErr(\hat{\phi}(\tau), \phi(\tau))$	0.042	0.064	0.165	0.029	**0.018**
	LogLik	6.34	4.13	1.18	0.74	**10.59**

number of collisions on each day is about 600. Records for Sep. 18th - Oct. 6th were used as training data and Oct. 9th - 13th were held out as test data.

In daily transportation, car collisions happening in the past will have a triggering influence on the future because of the traffic congestion caused by the initial accident, so there is a self-exciting phenomenon in this application. Hawkes process has already been applied in the transportation domain in the past. However, even when using nonparametric Hawkes process algorithm like MISD or WH, the baseline intensity is still a constant although the triggering kernel can be nonparametric. This is an inappropriate hypothesis in the vehicle collision application, e.g. the road is quiet at night so the baseline intensity of car accidents is lower than that in the day time, and the traffic is so busy at peak times that the baseline intensity will be increased. Using the EMV inference algorithm, the time-changing baseline intensity and a flexible triggering kernel can be simultaneously learned.

The performance of EMV (Alg. 4.2 and 4.3), WH, 6-bin MISD (MISD-6), 8-bin MISD (MISD-8), PH, RKHSC and GC were compared. The whole observation period T was set to 1440 minutes (24 hours) and the support of triggering kernel

T_ϕ was set to 60 minutes. For hyperparameters, the bandwidth of WH was set to 1.2 and there were 6 inducing points on $\phi(\tau)$ ($M_g = 6$) and 8 inducing points on $\mu(t)$ ($M_f = 8$). The hyperparameters of RKHSC and GC were chosen based on 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.

Green Taxi Pickup in New York City [48]: This dataset includes trip records from all trips completed in green taxis in New York City from January to June 2016. In experiments, the data from Jan. 7th to Feb. 1st were used as training data and Jan. 2nd - 6th were held out as test data. In this period, pickup dates and times were filtered out for long-distance trips ($>$ 15 miles) since long-distance trips usually have different patterns with short ones. The number of pickups each day was about 400.

With a similar setup, the performance of all methods was compared on the taxi pickup dataset. The whole observation period T was set to 24 hours and the support of triggering kernel T_ϕ set to 1 hour.

Results For each dataset, *LogLik* performance of various methods was evaluated when the number of training data varies. *LogLik* of EMV and other alternatives are shown in Fig. 4.1g and Fig. 4.1h. Clearly PH, MISD-6, MISD-8, WH and EMV outperform GC and RKHSC (both are inhomogeneous Poisson processes); this demonstrates the necessity of utilizing Hawkes process to discover the underlying self-exciting phenomenon in both datasets. Besides, the consistent superiority of the EMV algorithm over other Hawkes process inference algorithms (PH, MISD and WH) whose baseline intensity or triggering kernel is too restricted to capture the dynamics proves that the EMV algorithm can describe $\mu(t)$ and $\phi(\tau)$ in a completely flexible manner, leading to better goodness-of-fit.

To further measure performance, the *Q-Q plot* was generated. A sequence of

Table 4.2: 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%
EMV	**71.7%**	**70.4%**

timestamps in the test data was transformed by the fitted model to a set of independent uniform random variables on the interval $(0, 1)$ using the time rescaling theorem [47]. Any statistical assessment that measures agreement between the transformed data and a uniform distribution evaluates how well the fitted model agrees with the test data. Therefore, the *Q-Q plot* of the transformed timestamps with respect to the uniform distribution can be drawn. The perfect model follows a straight line $y = x$. The inhomogeneous Poisson process (GC), nonparametric Hawkes process with constant μ (WH) and nonparametric Hawkes process with time-changing $\mu(t)$ (EMV) were compared in a *Q-Q plot* in Fig. 4.2. It is observed that EMV is generally closer to the straight line, which suggests its better goodness-of-fit than other alternatives.

For the prediction task, *PreAcc* of all alternatives was measured on both datasets. It is assumed that only the top 17% of a sequence is observed ($\epsilon = 0.12$ for vehicle collison and 0.24 for taxi pickup, 500 samples for Monte Carlo integration) and then the time of the next event is predicted, and the real time of the next event, when it occurs, is incorporated into the observed data and then the further next even is predicted and the iteration goes on. Finally, the average *PreAcc* of the test data is computed, which is shown in Tab. 4.2 where it can be observed that EMV is obviously superior to other alternatives.

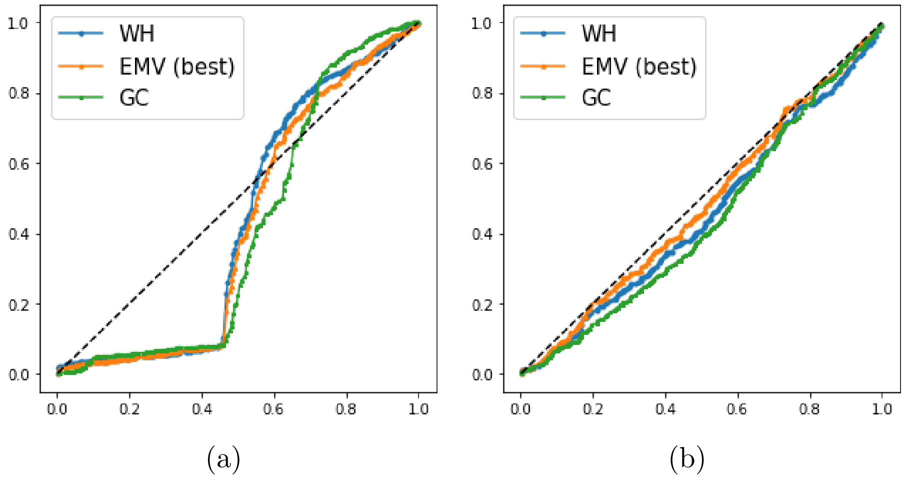


Figure 4.2: *Q-Q plot* of inhomogeneous Poisson process (GC), nonparametric Hawkes process with constant μ (WH), nonparametric Hawkes process with time-changing $\mu(t)$ (EMV). Vehicle collision dataset (left), taxi pickup dataset (right).

4.6 Discussion

In this section, the setting of hyperparameters and the advantages and disadvantages of the proposed algorithm are discussed. As stated in Chapter 3, the hyperparameters θ_0 and θ_1 have a significant effect on the estimation because they define the covariance kernel for the GP which encodes the smoothness of the function space. In order to tune them carefully, the hyperparameters are updated regularly in the EM iterations to maximize the ELBO. The advantage of the proposed algorithm is that the nonparametric baseline intensity and triggering kernel can be obtained with the efficiency largely improved. The disadvantage is that, despite the improved efficiency, the proposed algorithm inherits the disadvantage of variational Gaussian approximation which has a large number of variational parameters. This makes the algorithm still not efficient enough for very large datasets. An extension to the mean-field variational inference is a promising approach to address these problems.

4.7 Summary

In the vanilla 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 capacity for model expression. To further generalize the model and perform inference from a Bayesian perspective, the transformation of GP as prior is applied on the baseline intensity and triggering kernel and solves with an EM-variational inference algorithm. Accelerating methods are provided to make the inference efficient. Experiments show that the EMV inference can provide better results than the alternatives.

Chapter 5

Nonparametric Hawkes Process Modulated by Sigmoid Gaussian Process*

In Chapter 4, the link function was a quadratic transformation to guarantee the non-negativity of intensity. In this chapter, another type of Bayesian nonparametric Hawkes process is introduced, namely, a sigmoid GP Hawkes process where the link function is a scaled sigmoid function.

The posterior of the baseline intensity and triggering kernel with a quadratic link function is non-Gaussian due to the non-conjugacy between the likelihood and prior. As will be seen later, the sigmoid link function has an advantage over the quadratic link function because inference can be performed in a conjugate way. More specifically, for the sigmoid GP Hawkes process, the latent Pólya-Gamma random variables and marked Poisson processes are augmented to convert the likelihood into a conjugate form; consequently, corresponding Gibbs sampling, EM and mean-field

*Portions of this chapter have been deposited in arXiv: F. Zhou, Z. Li, X. Fan, Y. Wang, A. Sowmya, F. Chen, Scalable Inference for Nonparametric Hawkes Process Using Pólya-Gamma Augmentation. arXiv preprint arXiv:1910.13052. 29 Oct 2019 and submitted to the Journal of Machine Learning Research.

variational inference algorithms are proposed.

In the following, an overview of the chapter content is introduced in Sec. 5.1; a sigmoid GP Hawkes process model is proposed in Sec. 5.2; the auxiliary variable augmented likelihood and joint distribution are provided in Sec. 5.3; the Gibbs sampling algorithm is proposed in Sec. 5.4 with EM algorithm in Sec. 5.5 and mean-field variational inference in Sec. 5.6. Discussion is in Sec. 5.7 and summary is in Sec. 5.8.

5.1 Overview

In Chapter 4, a Bayesian nonparametric Hawkes process model was proposed, as a quadratic GP Hawkes process where the link function is chosen to be a square transformation. In the setting with a quadratic link function, an EM based variational Gaussian approximation inference algorithm was proposed. However, the posterior is non-Gaussian since the likelihood of GP variables is non-conjugate to the prior and an approximate inference approach has to be used, which is time consuming.

To circumvent the non-conjugate problem, a new Bayesian nonparametric Hawkes process model is proposed: a sigmoid GP Hawkes process where the link function is chosen to be a scaled sigmoid function. The likelihood is augmented with auxiliary latent random variables: branching structure, Pólya-Gamma random variables and latent marked Poisson processes. The branching structure of Hawkes process is introduced to decouple $\mu(t)$ and $\phi(\tau)$ to two independent components in the likelihood. Inspired by other work [49] and [50], a sigmoid link function is used in the model and the sigmoid converted to an infinite mixture of Gaussians involving Pólya-Gamma random variables. The latent marked Poisson processes are augmented to linearize the exponential integral term in the likelihood. By augmenting the likelihood in such a way, the likelihood becomes conjugate to the GP prior. With these latent random variables, the augmented likelihood is used to construct three efficient analytical iterative algorithms. The first one is a Gibbs sampler to sample from the posterior; the second one is an EM algorithm to obtain the maximum a posteriori (MAP) estimate; furthermore, the EM algorithm is extended to a mean-field variational inference algorithm that is slightly slower but can handle uncertainty with a distribution estimation rather than point estimation. It is worth noting that the naïve implementations of all the algorithms are time-consuming. To improve efficiency, sparse GP approximation [45] is introduced.

Specifically, the contributions made in this work are as follows:

1. By augmenting with latent branching structure, Pólya-Gamma random variables and latent marked Poisson processes, the original Hawkes likelihood is decoupled into two independent parts and both are conjugate to GP priors.

2. Simple and efficient Gibbs sampling, EM and mean-field variational inference algorithms are proposed for sigmoid GP Hawkes process where the baseline intensity and triggering kernel are both scaled sigmoid GP functions.

3. Sparse GP approximation is utilized to incorporate inducing points into the model to drastically reduce complexity.

5.2 Sigmoid GP Hawkes Process

A GP based Bayesian nonparametric Hawkes process model is proposed, namely a sigmoid-GP Hawkes process (SGPHP) whose baseline intensity and triggering kernel are functions drawn from a GP prior, passed through a sigmoid link function and scaled by an upper-bound to guarantee the non-negativity. In a naïve Bayesian framework, the posterior of $\mu(t)$ and $\phi(\tau)$ is

$$p(\mu(t),\phi(\tau)|D)=\frac{p(D|\mu(t)=\lambda_\mu^*\sigma(f),\phi(\tau)=\lambda_\phi^*\sigma(g))\mathcal{GP}(f)\mathcal{GP}(g)}{\iint p(D|\lambda_\mu^*\sigma(f),\lambda_\phi^*\sigma(g))\mathcal{GP}(f)\mathcal{GP}(g)dfdg}, \tag{5.1}$$

where $\sigma(\cdot)$ is the sigmoid function, f and g are two functions drawn from the corresponding GP priors, λ_μ^* and λ_ϕ^* are the upper bounds of $\mu(t)$ and $\phi(\tau)$.

In a naïve Bayesian framework, the inference of posterior of $\mu(t)$ and $\phi(\tau)$ is non-trivial because

1. $\mu(t)$ is coupled with $\phi(\tau)$ in the likelihood;
2. intractable integrals in the numerator and denominator cause a doubly-intractable problem [16];
3. the posterior is non-Gaussian.

However, as is seen later, these problems can be circumvented by augmenting the likelihood with auxiliary latent random variables. The sigmoid link function is chosen since it can be transformed to an infinite mixture of Gaussians; consequently, the augmented likelihood is in a conjugate form allowing for more efficient Gibbs sampling, EM and mean-field variational inference with explicit expressions.

5.3 Likelihood Augmentation

The likelihood augmentation is divided into 3 steps. The branching structure is augmented into likelihood in Sec. 5.3.1. In Sec. 5.3.2, the Pólya-Gamma random variables are augmented with the latent marked Poisson processes augmented in Sec. 5.3.3. The final augmented likelihood and joint density are provided in Sec. 5.3.4.

5.3.1 Augmenting Branching Structure

As before, the branching structure of Hawkes process [1, 10] (see Sec. 2.1.1) is augmented to the Hawkes likelihood (Eq. 2.1) to decouple $\mu(t)$ and $\phi(\tau)$. The joint likelihood with branching structure is Eq. 2.3 with $\mu(t) = \lambda_\mu^*\sigma(f(t))$, $\phi(\tau) = \lambda_\phi^*\sigma(g(\tau))$.

After introducing the branching structure, the joint likelihood is decoupled to two independent factors.

5.3.2 Transformation of Sigmoid Function

To transform the sigmoid function, a remarkable representation discovered in the literature of Bayesian inference for logistic regression [49] is utilized. Surprisingly, the sigmoid function is redefined as a Gaussian representation. It was found that the inverse hyperbolic cosine can be expressed as an infinite mixture of Gaussian densities:

$$\cosh^{-b}(z/2) = \int_0^\infty e^{-z^2\omega/2} p_{\mathrm{PG}}(\omega|b,0)d\omega, \tag{5.2}$$

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

$$\sigma(z) = \frac{e^{z/2}}{2\cosh(z/2)} = \int_0^\infty e^{h(\omega,z)} p_{\mathrm{PG}}(\omega|1,0)d\omega, \tag{5.3}$$

where $h(\omega,z) = z/2 - z^2\omega/2 - \log 2$. Moreover, the posterior Pólya-Gamma distribution also known as (a.k.a.) the tilted Pólya-Gamma distribution which is used later can be expressed as

$$p_{\mathrm{PG}}(\omega|b,c) \propto e^{-c^2\omega/2} p_{\mathrm{PG}}(\omega|b,0). \tag{5.4}$$

Using Eq. 5.3, the products of observations $\sigma(f(t_i))$ and $\sigma(g(\tau_{ij}))$ ($\tau_{ij} = t_i - t_j$) in the likelihood Eq. 2.3 are transformed into a Gaussian form. It is worth noting that the exact form of Pólya-Gamma distribution need not be known but only its first order moment.

5.3.3 Transformation of Exponential Integral

Here, only the baseline intensity part is discussed. All derivations in the triggering kernel part are the same as the baseline intensity part except for some notations. Utilizing Eq. 5.3 and the sigmoid property $\sigma(z) = 1-\sigma(-z)$, the exponential integral in the likelihood Eq. 2.3 can be rewritten as

$$\begin{aligned}&\exp\left(-\int_T \lambda_\mu^* \sigma(f(t))dt\right) = \\ &\exp\left(-\int_T \int_{\mathbb{R}^+} \left(1 - e^{h(\omega_\mu, -f(t))}\right) \lambda_\mu^* p_{\mathrm{PG}}(\omega_\mu|1,0) d\omega_\mu dt\right).\end{aligned} \tag{5.5}$$

According to Campbell's theorem [51] (see below), the right-hand side of Eq. 5.5 is a characteristic functional of a marked Poisson process, and can be rewritten as

$$\exp\left(-\int_T \lambda_\mu^* \sigma(f(t))dt\right) = \mathbb{E}_{p_{\lambda_\mu}}\left[\prod_{(\omega_\mu, t)\in\Pi_\mu} e^{h(\omega_\mu, -f(t))}\right], \tag{5.6}$$

where $\Pi_\mu = \{(\omega_{\mu_m}, t_m)\}_{m=1}^{M_\mu}$ denotes a random realization of a marked Poisson process and p_{λ_μ} is the probability measure of the marked Poisson process Π_μ with intensity $\lambda_\mu(t, \omega_\mu) = \lambda_\mu^* p_{\mathrm{PG}}(\omega_\mu|1,0)$. The events $\{t_m\}_{m=1}^{M_\mu}$ follow a Poisson process with rate λ_μ^* and the latent Pólya-Gamma variable ω_{μ_m} denotes the independent mark at each location t_m. The derivation is shown now.

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

$$\mathbb{E}_{\Pi_{\hat{\mathcal{Z}}}}\left[\exp\left(\xi H(\Pi_{\hat{\mathcal{Z}}})\right)\right] = \exp\left[\int_{\hat{\mathcal{Z}}} \left(e^{\xi h(\mathbf{z},\boldsymbol{\omega})} - 1\right) \Lambda(\mathbf{z}, \boldsymbol{\omega}) d\boldsymbol{\omega} d\mathbf{z}\right], \tag{5.7}$$

for any $\xi \in \mathbb{C}$. Eq. 5.7 defines the characteristic functional of a marked Poisson process. This proves Eq. 5.6. The mean and variance are defined as:

$$\begin{aligned}\mathbb{E}_{\Pi_{\hat{\mathcal{Z}}}}\left[H(\Pi_{\hat{\mathcal{Z}}})\right] &= \int_{\hat{\mathcal{Z}}} h(\mathbf{z},\boldsymbol{\omega})\Lambda(\mathbf{z},\boldsymbol{\omega})d\boldsymbol{\omega}d\mathbf{z}\\ \mathrm{Var}_{\Pi_{\hat{\mathcal{Z}}}}\left[H(\Pi_{\hat{\mathcal{Z}}})\right] &= \int_{\hat{\mathcal{Z}}} [h(\mathbf{z},\boldsymbol{\omega})]^2\Lambda(\mathbf{z},\boldsymbol{\omega})d\boldsymbol{\omega}d\mathbf{z}.\end{aligned} \tag{5.8}$$

5.3.4 Augmented Likelihood and Joint Density

Substituting Eq. 5.3 and Eq. 5.6 into Eq. 2.3, the following are obtained:

(1) The augmented joint likelihood of baseline intensity part

$$\begin{aligned}&p(D,\mathbf{B}_{ii}|\lambda_\mu^*,f)\\ &=\prod_{i=1}^N\left(\lambda_\mu^*\sigma(f(t_i))\right)^{b_{ii}}\exp\left(-\int_T\lambda_\mu^*\sigma(f(t))dt\right)\\ &=\prod_{i=1}^N\left(\int_0^\infty\lambda_\mu^* e^{h(\omega_{ii},f(t_i))}p_{\mathrm{PG}}(\omega_{ii}|1,0)d\omega_{ii}\right)^{b_{ii}}\cdot\mathbb{E}_{p_{\lambda_\mu}}\left[\prod_{(\omega_\mu,t)\in\Pi_\mu}e^{h(\omega_\mu,-f(t))}\right]\\ &=\iint\prod_{i=1}^N\left(\lambda_\mu(t_i,\omega_{ii})e^{h(\omega_{ii},f(t_i))}\right)^{b_{ii}}\cdot p_{\lambda_\mu}(\Pi_\mu|\lambda_\mu^*)\prod_{(\omega_\mu,t)\in\Pi_\mu}e^{h(\omega_\mu,-f(t))}d\boldsymbol{\omega}_{ii}d\Pi_\mu\end{aligned} \tag{5.9}$$

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

$$\begin{aligned}&p(D,\Pi_\mu,\boldsymbol{\omega}_{ii},\mathbf{B}_{ii}|\lambda_\mu^*,f)\\ &=\prod_{i=1}^N\left(\lambda_\mu(t_i,\omega_{ii})e^{h(\omega_{ii},f(t_i))}\right)^{b_{ii}}\cdot p_{\lambda_\mu}(\Pi_\mu|\lambda_\mu^*)\prod_{(\omega_\mu,t)\in\Pi_\mu}e^{h(\omega_\mu,-f(t))}.\end{aligned} \tag{5.10}$$

Incorporating the priors of λ_μ^* and f into Eq. 5.10, the joint distribution over all variables is obtained. Without loss of generality, an improper prior $p(\lambda_\mu^*)=1/\lambda_\mu^*$

[37] and a symmetric GP prior $\mathcal{GP}(f|0, K_f)$ are utilized here.

$$
\begin{aligned}
&p(D, \Pi_\mu, \boldsymbol{\omega}_{ii}, \mathbf{B}_{ii}, \lambda_\mu^*, f) \\
&= \prod_{i=1}^{N} \left(\lambda_\mu(t_i, \omega_{ii}) e^{h(\omega_{ii}, f(t_i))}\right)^{b_{ii}} \cdot p_{\lambda_\mu}(\Pi_\mu|\lambda_\mu^*) \prod_{(\omega_\mu, t)\in\Pi_\mu} e^{h(\omega_\mu, -f(t))} \cdot {\lambda_\mu^*}^{-1} \mathcal{GP}(f).
\end{aligned}
\tag{5.11}
$$

(2) The augmented joint likelihood of triggering kernel part

$$
\begin{aligned}
&p(D, \{\Pi_{\phi_i}\}_{i=1}^N, \boldsymbol{\omega}_{ij}, \mathbf{B}_{ij}|\lambda_\phi^*, g) \\
&= \prod_{i=2}^{N}\prod_{j=1}^{i-1} \left(\lambda_\phi(\tau_{ij}, \omega_{ij}) e^{h(\omega_{ij}, g(\tau_{ij}))}\right)^{b_{ij}} \cdot \prod_{i=1}^{N} \left[p_{\lambda_\phi}(\Pi_{\phi_i}|\lambda_\phi^*) \prod_{(\omega_\phi, \tau)\in\Pi_{\phi_i}} e^{h(\omega_\phi, -g(\tau))}\right]
\end{aligned}
\tag{5.12}
$$

and the augmented joint distribution of triggering kernel part is

$$
\begin{aligned}
&p(D, \{\Pi_{\phi_i}\}_{i=1}^N, \boldsymbol{\omega}_{ij}, \mathbf{B}_{ij}, \lambda_\phi^*, g) \\
&= \prod_{i=2}^{N}\prod_{j=1}^{i-1} \left(\lambda_\phi(\tau_{ij}, \omega_{ij}) e^{h(\omega_{ij}, g(\tau_{ij}))}\right)^{b_{ij}} \cdot \\
&\prod_{i=1}^{N} \left[p_{\lambda_\phi}(\Pi_{\phi_i}|\lambda_\phi^*) \prod_{(\omega_\phi, \tau)\in\Pi_{\phi_i}} e^{h(\omega_\phi, -g(\tau))}\right] \cdot {\lambda_\phi^*}^{-1} \mathcal{GP}(g),
\end{aligned}
\tag{5.13}
$$

where $\tau_{ij} = t_i - t_j$, $\mathcal{GP}(g)$ is symmetric $\mathcal{GP}(g|0, K_g)$, p_{λ_ϕ} is the probability measure of the corresponding latent marked Poisson process $\Pi_{\phi_i} = \{(\omega_{\phi_m}, \tau_m)\}_{m=1}^{M_{\phi_i}}$ with intensity $\lambda_\phi(\tau, \omega_\phi) = \lambda_\phi^* p_{\text{PG}}(\omega_\phi|1, 0)$, $\boldsymbol{\omega}_{ij}$ denotes the vector of ω_{ij} on each τ_{ij} and $\mathbf{B}_{ij}$ are the entries off the diagonal of branching structure. It is worth noting that there exist N independent latent marked Poisson processes because of the exponential integral product term in Eq. 2.3. The proof is the same as for the $\mu(t)$ part and is omitted here.

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

5.4 Gibbs Sampler

A naïve Gibbs sampler is derived in this section. However, the naïve implementation is time-consuming because of the cubic complexity w.r.t the number of observations and latent Poisson events when sampling f and g. This issue has been discussed in [16]. To circumvent the problem, sparse GP approximation is utilized to introduce some inducing points to make the inference efficient.

5.4.1 Naïve Gibbs Sampler

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

$$\begin{aligned} p(\boldsymbol{\omega}_{ii}|\mathbf{f}) &= \prod_{i=1}^{N} \left(p_{\mathrm{PG}}(\omega_{ii}|1, f(t_i))\right)^{b_{ii}} \\ p(\boldsymbol{\omega}_{ij}|\mathbf{g}) &= \prod_{i=2}^{N} \prod_{j=1}^{i-1} \left(p_{\mathrm{PG}}(\omega_{ij}|1, g(\tau_{ij}))\right)^{b_{ij}}, \end{aligned} \tag{5.14}$$

where the posterior Pólya-Gamma distribution defined in Eq. 5.4 is utilized. The Pólya-Gamma random variable can be efficiently sampled by a method proposed in [49].

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

$$\begin{aligned} p(\lambda_\mu^*|\mathbf{B}_{ii}, \Pi_\mu) &= \mathrm{Gamma}(\lambda_\mu^*|N_\mu + M_\mu, T) \\ p(\lambda_\phi^*|\mathbf{B}_{ij}, \Pi_\phi) &= \mathrm{Gamma}(\lambda_\phi^*|N_\phi + M_\phi, NT_\phi), \end{aligned} \tag{5.15}$$

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

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

$$\begin{aligned} p(\mathbf{f}_{N_\mu+M_\mu}|\boldsymbol{\omega}_{ii}, \Pi_\mu) &= \mathcal{N}(\mathbf{m}_{N_\mu+M_\mu}, \boldsymbol{\Sigma}_{N_\mu+M_\mu}) \\ p(\mathbf{g}_{N_\phi+M_\phi}|\boldsymbol{\omega}_{ij}, \{\Pi_{\phi_i}\}_{i=1}^N) &= \mathcal{N}(\mathbf{m}_{N_\phi+M_\phi}, \boldsymbol{\Sigma}_{N_\phi+M_\phi}) \end{aligned} \tag{5.16}$$

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

Sampling the latent marked Poisson processes The conditional posterior of the marked point process is

$$p(\Pi_\mu|f, \lambda_\mu^*) = \frac{p_{\lambda_\mu}(\Pi_\mu|\lambda_\mu^*) \prod_{(\omega_\mu,t)\in\Pi_\mu} e^{h(\omega_\mu,-f(t))}}{\int p_{\lambda_\mu}(\Pi_\mu|\lambda_\mu^*) \prod_{(\omega_\mu,t)\in\Pi_\mu} e^{h(\omega_\mu,-f(t))} d\Pi_\mu}. \tag{5.17}$$

As proved elsewhere [50], this conditional point process is again a marked Poisson process by utilizing the Campbell theorem to calculate its characteristic function. A more concise proof is provided here. Using Eq. 5.6 to convert the denominator, Eq. 5.17 can be written as

$$
\begin{aligned}
&p(\Pi_\mu|f,\lambda_\mu^*)\\
&=\frac{p_{\lambda_\mu}(\Pi_\mu|\lambda_\mu^*)\prod_{(\omega_\mu,t)\in\Pi_\mu}e^{h(\omega_\mu,-f(t))}}{\exp\left(-\iint(1-e^{h(\omega_\mu,-f(t))})\lambda_\mu^* p_{\mathrm{PG}}(\omega_\mu|1,0)d\omega_\mu dt\right)}\\
&=\prod_{(\omega_\mu,t)\in\Pi_\mu}\left(e^{h(\omega_\mu,-f(t))}\lambda_\mu^* p_{\mathrm{PG}}(\omega_\mu|1,0)\right)\cdot\exp\left(-\iint e^{h(\omega_\mu,-f(t))}\lambda_\mu^* p_{\mathrm{PG}}(\omega_\mu|1,0)d\omega_\mu dt\right).
\end{aligned}
\tag{5.18}
$$

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

$$\Lambda_\mu(t,\omega_\mu)=e^{h(\omega_\mu,-f(t))}\lambda_\mu^* p_{\mathrm{PG}}(\omega_\mu|1,0)=\lambda_\mu^*\sigma(-f(t))p_{\mathrm{PG}}(\omega_\mu|1,f(t)). \tag{5.19}$$

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

For sampling from the posterior marked Poisson processes, the timestamps t_m (τ_m) are first drawn with the rate $\lambda_\mu^*\sigma(-f(t))$ ($\lambda_\phi^*\sigma(-g(\tau))$) by using the thinning algorithm [38], and then the marks ω_μ (ω_ϕ) are drawn from the conditional distribution $p_{\mathrm{PG}}(\omega_\mu|1,f(t))$ ($p_{\mathrm{PG}}(\omega_\phi|1,g(\tau))$).

Sampling the branching structure After combining Eq. 5.11 and Eq. 5.13 and integrating out $\boldsymbol{\omega}_{ii}$ and $\boldsymbol{\omega}_{ij}$, the conditional posterior of $\mathbf{B}$ is obtained:

$$p(\mathbf{B}|\lambda_\mu^*,\lambda_\phi^*,f,g)\propto\prod_{i=1}^{N}(\mu(t_i))^{b_{ii}}\prod_{i=2}^{N}\prod_{j=1}^{i-1}(\phi(\tau_{ij}))^{b_{ij}} \tag{5.20}$$

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

$$\begin{aligned} p(b_{ii} = 1) &= \frac{\mu(t_i)}{\mu(t_i) + \sum_{j=1}^{i-1} \phi(\tau_{ij})} \\ p(b_{ij} = 1) &= \frac{\phi(\tau_{ij})}{\mu(t_i) + \sum_{j=1}^{i-1} \phi(\tau_{ij})} \end{aligned} \tag{5.21}$$

which is a well-known result [8], [10].

5.4.2 Algorithm Speedup

The naïve Gibbs sampler presented in Sec. 5.4.1 is impractical. The reasons include the following:

1. The bottleneck of the algorithm is in the step of sampling function values. Because matrix inversion has to be performed, the complexity is $\mathcal{O}((N_\mu + M_\mu)^3 + (N_\phi + M_\phi)^3)$ where $N_\mu + N_\phi = N$. This means that it is non-scalable to even a few hundreds of observations.

2. The function values have to be sampled twice in one MCMC loop. Each time that the branching structure or the latent marked Poisson process is updated, the function values have to be updated once in order to avoid dimension mismatch. This slows down the Gibbs sampler even further.

To circumvent these problems, the sparse GP approximation is utilized to introduce some inducing points. f and g are supposed to be dependent on their corresponding inducing points $\{t_s\}_{s=1}^{S_\mu}$ and $\{\tau_s\}_{s=1}^{S_\phi}$; the function values of f and g at these inducing points are $\mathbf{f}_{t_s}$ and $\mathbf{g}_{\tau_s}$. Given a sample $\mathbf{f}_{t_s}$ and $\mathbf{g}_{\tau_s}$, $\mathbf{f}_{N_\mu+M_\mu}$ and $\mathbf{g}_{N_\phi+M_\phi}$ in Eq. 5.16 are assumed to be the posterior GP mean functions:

$$\mathbf{f}_{N_\mu+M_\mu} = \mathbf{K}_{tt_s}\mathbf{K}_{t_st_s}^{-1}\mathbf{f}_{t_s}, \quad \mathbf{g}_{N_\phi+M_\phi} = \mathbf{K}_{\tau\tau_s}\mathbf{K}_{\tau_s\tau_s}^{-1}\mathbf{g}_{\tau_s} \tag{5.22}$$

with $\mathbf{K}_{tt_s}$ and $\mathbf{K}_{\tau\tau_s}$ being the kernel matrixes w.r.t. the observations and inducing points while $\mathbf{K}_{t_s t_s}$ and $\mathbf{K}_{\tau_s \tau_s}$ are w.r.t. inducing points only.

Now the conditional posteriors of function values are transformed from observations to inducing points:

$$\begin{aligned} p(\mathbf{f}_{t_s}|\boldsymbol{\omega}_{ii}, \Pi_\mu) &= \mathcal{N}(\mathbf{m}_{t_s}, \boldsymbol{\Sigma}_{t_s}) \\ p(\mathbf{g}_{\tau_s}|\boldsymbol{\omega}_{ij}, \{\Pi_{\phi_i}\}_{i=1}^N) &= \mathcal{N}(\mathbf{m}_{\tau_s}, \boldsymbol{\Sigma}_{\tau_s}) \end{aligned} \tag{5.23}$$

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

With sparse GP approximation, the complexity is reduced to $\mathcal{O}(S_\mu^3 + S_\phi^3)$ with $S_\mu \ll N_\mu + M_\mu$, $S_\phi \ll N_\phi + M_\phi$. What makes this even more remarkable is the fact that the function values only need to be sampled once in one MCMC loop because they only depend on inducing points which are fixed during the sampling process. Moreover, the sampling of latent marked Poisson processes can be parallelized.

5.4.3 Hyperparameters

Throughout this work, the GP covariance kernel used is the squared exponential kernel $k(x, x') = \theta_0 \exp\left(-\frac{\theta_1}{2}\|x - x'\|^2\right)$. The hyperparameters θ_0 and θ_1 can be sampled by a Metropolis-Hasting method [27]. Normally, they are updated every 20 loops.

Additional hyperparameters are the number and location of inducing points which affect the complexity and estimation quality of $\mu(t)$ and $\phi(\tau)$. A large number of inducing points will lead to high complexity while a small number cannot capture the dynamics. For fast inference, the inducing points are uniformly located on the domain. Another advantage of uniform location is that the kernel matrix has Toeplitz structure [46] which means that the matrix inversion can be implemented more efficiently. The number of inducing points is gradually increased until no more significant improvement occurs. The final pseudo code is provided in Alg. 5.1.

Algorithm 5.1: Accelerated Gibbs sampler for SGPHP

```
Result: $\mu(t)$, $\phi(\tau)$
Initialize hyperparameters and $\mathbf{B}$, $\lambda_\mu^*$, $\lambda_\phi^*$, $\boldsymbol{\omega}_{ii}$, $\boldsymbol{\omega}_{ij}$, $\mathbf{f}_{t_s}$, $\mathbf{g}_{\tau_s}$, $\Pi_\mu$, $\{\Pi_{\phi_i}\}_{i=1}^N$;
for do
    Sample $\boldsymbol{\omega}_{ii}$ and $\boldsymbol{\omega}_{ij}$ with Eq.(5.14);
    Sample $\lambda_\mu^*$ and $\lambda_\phi^*$ with Eq.(5.15);
    Sample $\mathbf{f}_{t_s}$ and $\mathbf{g}_{\tau_s}$ with Eq.(5.23);
    Sample $\Pi_\mu$ and $\{\Pi_{\phi_i}\}_{i=1}^N$ with Eq.(5.19);
    Sample $\mathbf{B}$ with Eq.(5.21);
    Sample hyperparameters with Metropolis-Hasting algorithm.
end
```

5.4.4 Experimental Results

The performance of Gibbs sampler for SGPHP was evaluated on both synthetic and real-world data. Specifically, the Gibbs sampler was compared to the following alternatives:

1. **Hawkes Process (HP)**: the vanilla Hawkes process with constant μ and exponential decay triggering kernel $\alpha \exp(-\beta(t-t_i))$. The inference is performed by MLE.

2. **Wiener-Hopf (WH)**: this is a nonparametric algorithm for Hawkes process where μ is constant and $\phi(\tau)$ is a continuous function. The inference is based on the solution of a Wiener-Hopf equation [11].

3. **Majorization Minimization Euler-Lagrange (MMEL)**: this is another nonparametric algorithm for Hawkes process with constant μ and continuous $\phi(\tau)$. This algorithm similarly utilizes the branching structure and estimates $\phi(\tau)$ by an Euler-Lagrange equation [10].

The following metrics were used to evaluate the performance of the various methods:

- ***TestLL***: the log-likelihood of test data using the trained model. This is a metric describing the model prediction ability. It is used to measure the performance of synthetic and real data.

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

- ***AutCor***: the lag-k autocorrelation between states of Markov chain. It is used for checking the mixing performance of Markov chain. It is used to measure the performance of synthetic data.

- ***PreAcc***: given an event sequence $\{t_1, ..., t_{i-1}\}$, the goal is to predict the time of t_i. The expectation of t_i should be $\mathbb{E}[t_i] = \int_{t_{i-1}}^{\infty} tp(t_i = t)dt$ with $P(t_i = t) = \lambda(t)\exp\left(-\int_{t_{i-1}}^{t}\lambda(s)ds\right)$. The integral can be estimated by Monte Carlo method. Multiple timestamps in a sequence are predicted: if the predicted $\hat{t}_i$ is within an error bound ϵ, then it is considered to be a correct prediction; else it is incorrect. The percentage of correct predictions is defined as the prediction accuracy. It is used to measure the performance of real data.

5.4.4.1 Synthetic Data Experiments

In synthetic data experiments, the thinning algorithm [38] is used to generate 100 sets of training data and 10 sets of test data with $T_\phi = 6$ and $T = 100$ in three cases:

1. $\mu(t) = \begin{cases} 1 & (0 < t \leq T/2) \\ 2 & (T/2 < t < T) \end{cases}$ and $\phi(\tau) = 1 \cdot \exp(-2\tau)$;

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

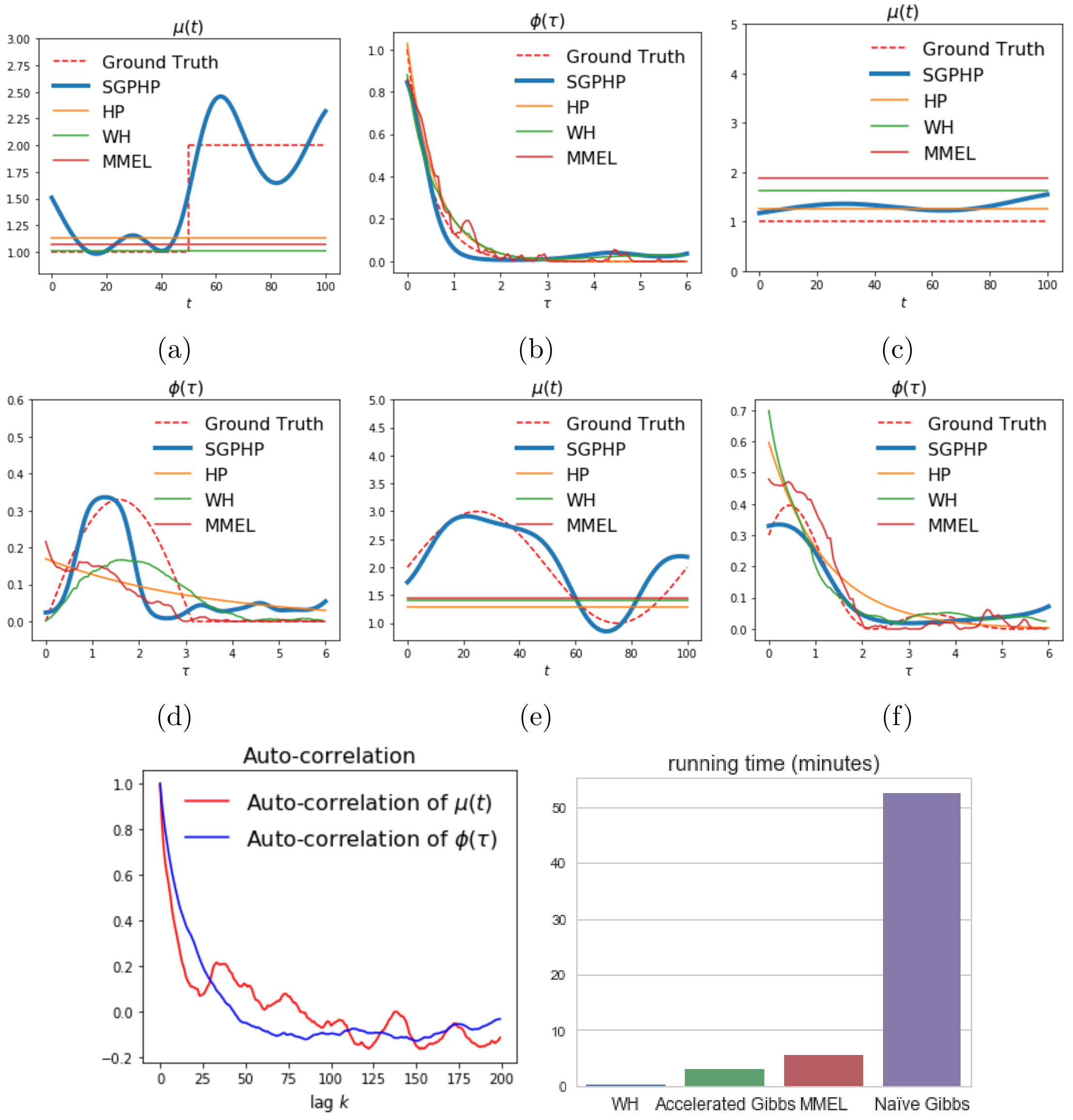
3. $\mu(t) = \sin\left(\frac{2\pi}{T}\cdot t\right) + 2 \quad (0 < t < T)$ and $\phi(\tau) = 0.3\left(\sin(\frac{2\pi}{3}\cdot\tau) + 1\right)\cdot \exp(-0.7\tau) \quad (0 < \tau < T_\phi)$.

In Case1, there is a non-constant $\mu(t)$ which is a piecewise constant function, while Case 2 has a half sinusoidal $\phi(\tau)$. Case 3 is the most general one with time-changing $\mu(t)$ and sinusoidal exponential decay $\phi(\tau)$.

The inducing points and hyperparameters are optimized for inference. The posterior mean $\hat{\mu}(t)$ and $\hat{\phi}(\tau)$ are shown in Fig. 5.1. Clearly, they capture the structure of the underlying rate very well. In Tab. 5.1, the results confirm that the SGPHP model outperforms other alternatives in most cases w.r.t. *TestLL* and *EstErr*. For HP, when the underlying $\mu(t)$ and $\phi(\tau)$ are in complex forms, the estimated baseline intensity and triggering kernel are far away from the ground truth due to parametric constraints. For WH and MMEL, the constant limitation on $\mu(t)$ still exists even though $\phi(\tau)$ has been relaxed to be nonparametric. On the contrary, the SGPHP model provides nonparametric $\hat{\mu}(t)$ and $\hat{\phi}(\tau)$ concurrently.

For Markov chain samplers to be efficient, correlations between samples should decay quickly, and Fig. 5.1g shows that the autocorrelation reaches a plateau of 0 after 80 samples, indicating excellent mixing performance of the sampler.

For efficiency, the running time of the naïve Gibbs sampler is compared to the accelerated Gibbs sampler and MMEL algorithm with the same number of loops, as well as the WH algorithm. In Fig. 5.1h, WH is the fastest algorithm while the accelerated Gibbs sampler is the runner-up. The reason is that WH algorithm is a method based on solving a linear system without the need of iterations, while the Gibbs sampler and MMEL algorithm both need iterative solutions. The naïve Gibbs sampler is the least efficient because of the cubic complexity w.r.t the number of observations and latent Poisson events. In Tab. 5.2 the efficiency of different iterative algorithms is shown on different sizes of data. The result confirms that the accelerated Gibbs sampler scales well enough with data size.



(a) (b) (c)

(d) (e) (f)

(g) (h)

Figure 5.1: The synthetic data experimental results. (a) and (b): estimated $\hat{\mu}(t)$ and $\hat{\phi}(\tau)$ in case 1; (c) and (d) for case 2; (e) and (f) for case 3; (g): autocorrelation of $\mu(t)$ and $\phi(\tau)$ samples in the Markov chain; (h): running time of WH, accelerated Gibbs sampler, MMEL and naïve Gibbs sampler on 225 observation points.

Table 5.1: The *EstErr* and *TestLL* of fictitious data.

		HP	WH	MMEL	SGPHP
Case 1	$EstErr(\hat{\mu}, \mu)$	0.38	0.48	0.43	**0.05**
	$EstErr(\hat{\phi}, \phi)$	**0.0011**	0.0012	0.0022	0.0012
	TestLL	60.61	60.70	59.92	**62.16**
Case 2	$EstErr(\hat{\mu}, \mu)$	**0.26**	0.62	0.88	0.31
	$EstErr(\hat{\phi}, \phi)$	0.0293	0.0071	0.0123	**0.0069**
	TestLL	-8.19	-6.98	-9.74	**-5.48**
Case 3	$EstErr(\hat{\mu}, \mu)$	1.00	0.84	0.81	**0.07**
	$EstErr(\hat{\phi}, \phi)$	0.0040	0.0048	0.0031	**0.0012**
	TestLL	260.93	260.50	259.98	**263.81**

Table 5.2: The running time of different iterative algorithms on different sizes of data. (Time unit: minutes)

# Points	Accelerated Gibbs	MMEL	Naïve Gibbs
100	**0.89**	1.25	6.23
200	**2.93**	5.60	52.53
400	**7.05**	27.92	—

5.4.4.2 Real Data Experiments

The various methods were compared on two real-world crime datasets. In criminology, the presence of a self-exciting phenomenon from past crimes to future ones has been reported [42]. The two datasets both comprise times of security violation or report in a period of several years. The ground truth information is unavailable for the real world data, therefore *EstErr* cannot be utilized to measure the performance. As a result, for each dataset the goodness-of-fit is tested on the test data (*TestLL*) and the time of event occurring in a future time window is predicted (*PreAcc*).

Crime in Vancouver (Canada) [52]

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

NYPD Complaint Data [40]

This dataset includes all valid felony, misdemeanour and violation crimes reported to the New York police department (NYPD) for all complete quarters in 2017. The columns are complaint number, date, time, offense description, borough and other information.

Preliminary Setup For the Crime in Vancouver dataset, the theft records from June to November 2016 occurring in the central business district were filtered out and a small time interval was added to separate all the simultaneously occurring records. For the NYPD Complaint Dataset, the complaints records in Brooklyn and Queens in 2016 with the offense description of petit larceny were filtered out. For each of these datasets, the timestamps of events were split into training and test sets. The precise split varies for each dataset, in particular for Crime in Vancouver, the first 519 data points were selected as training set to train the models, with the rest being test data (time unit: days); for NYPD Complaint Data, the first 324 data points were selected as training set with the rest being test (time unit: days). For the prediction task, it is assumed that only the top 17% of a sequence is observed ($\epsilon = 0.12$ for Crime in Vancouver and 0.89 for NYPD Complaint Data, 400 samples for Monte Carlo integration) and the time of next event is predicted, then the real time of occurrence of the next event is incorporated into the observed data before the next one is predicted, and so on.

Results The *TestLL* of SGPHP and other alternatives are shown in Tab. 5.3. It can be observed that WH, MMEL and SGPHP all outperform HP, which demonstrates the necessity of nonparametric models to capture the underlying dynamic triggering effect. Besides, the consistent superiority of SGPHP over other nonparametric models with constant baseline intensity shows that the SGPHP model can capture the completely flexible $\mu(t)$ and $\phi(\tau)$ concurrently, leading to better goodness-of-fit. The *PreAcc* of all alternatives were measured on both datasets. The average *PreAcc* of the test data is shown in Fig. 5.2 where SGPHP is clearly superior to other alternatives.

Table 5.3: *TestLL* of SGPHP and other alternatives on two real datasets.

Dataset	HP	WH	MMEL	SGPHP
Crime in Vancouver	380.88	400.36	405.22	**428.06**
NYPD Complaint Data	-209.02	-197.21	-198.85	**-194.88**

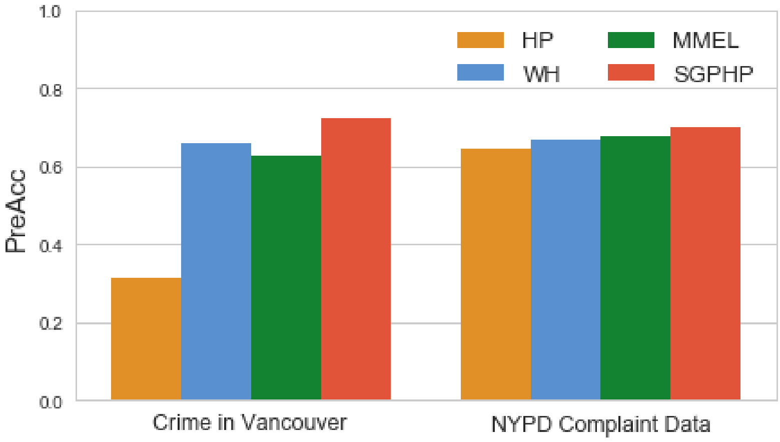


Figure 5.2: The *PreAcc* of SGPHP and other alternatives on two real datasets.

5.5 EM Algorithm

With the original likelihood Eq. 2.1 and GP priors $\mathcal{GP}(f)$ and $\mathcal{GP}(g)$ (symmetric prior $\mathcal{GP}(\cdot|0, K_\cdot)$), the log-posterior corresponds to a penalized log-likelihood. As discussed for GP models with likelihood depending on finite inputs only [24], the regularizer is given by the squared reproducing kernel Hilbert space (RKHS) norm corresponding to the GP kernel. Therefore

$$\hat{\lambda}_{\mu}^{*}, \hat{f}, \hat{\lambda}_{\phi}^{*}, \hat{g} = \operatorname{argmax}\left\{ \log p(D|\lambda_{\mu}^{*}, f, \lambda_{\phi}^{*}, g) - \frac{1}{2}\|f\|_{\mathcal{H}_{k_f}}^2 - \frac{1}{2}\|g\|_{\mathcal{H}_{k_g}}^2 \right\}, \tag{5.24}$$

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

$$\begin{aligned}&\mathcal{Q}((\lambda_{\mu}^{*}, f, \lambda_{\phi}^{*}, g)|(\lambda_{\mu}^{*}, f, \lambda_{\phi}^{*}, g)_{\text{old}})\\ &= \mathbb{E}\left[\log p(D, \Pi_{\mu}, \boldsymbol{\omega}_{ii}, \{\Pi_{\phi_i}\}_{i=1}^{N}, \boldsymbol{\omega}_{ij}, \mathbf{B}|\lambda_{\mu}^{*}, f, \lambda_{\phi}^{*}, g)\right] - \frac{1}{2}\|f\|_{\mathcal{H}_{k_f}}^2 - \frac{1}{2}\|g\|_{\mathcal{H}_{k_g}}^2,\end{aligned} \tag{5.25}$$

with $\mathbb{E}$ over $P(\Pi_{\mu}, \boldsymbol{\omega}_{ii}, \{\Pi_{\phi_i}\}_{i=1}^{N}, \boldsymbol{\omega}_{ij}, \mathbf{B}|(\lambda_{\mu}^{*}, f, \lambda_{\phi}^{*}, g)_{\text{old}})$.

Because of auxiliary variables augmentation, the GP variables are in a quadratic form in the lower-bound, which results in an analytical solution in the M step.

5.5.1 E Step

In the E step, the conditional density $P(\Pi_{\mu}, \boldsymbol{\omega}_{ii}, \{\Pi_{\phi_i}\}_{i=1}^{N}, \boldsymbol{\omega}_{ij}, \mathbf{B}|(\lambda_{\mu}^{*}, f, \lambda_{\phi}^{*}, g)_{\text{old}})$ is derived and then the lower-bound $\mathcal{Q}$ is computed.

5.5.1.1 Conditional Density

The conditional density of Π_{μ}, $\boldsymbol{\omega}_{ii}$, $\{\Pi_{\phi_i}\}_{i=1}^{N}$, $\boldsymbol{\omega}_{ij}$, $\mathbf{B}$ given $(\lambda_{\mu}^{*}, f, \lambda_{\phi}^{*}, g)_{\text{old}}$ can be factorized and obtained from Eq. 5.10 and Eq. 5.12. More specifically, details of these factors are provided.

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

$$\begin{aligned} p(\boldsymbol{\omega}_{ii}|\mathbf{f}_{\text{old}}) &= \prod_{i=1}^{N} p_{\text{PG}}(\omega_{ii}|1, f_{\text{old}}(t_i)) \\ p(\boldsymbol{\omega}_{ij}|\mathbf{g}_{\text{old}}) &= \prod_{i=2}^{N}\prod_{j=1}^{i-1} p_{\text{PG}}(\omega_{ij}|1, g_{\text{old}}(\tau_{ij})), \end{aligned} \tag{5.26}$$

where $\mathbf{B}$ is marginalizes out and the tilted Pólya-Gamma distribution $p_{\text{PG}}(\omega|b,c) \propto e^{-c^2\omega/2}p_{\text{PG}}(\omega|b,0)$ is utilized, with the first order moment being $\mathbb{E}[\omega] = \frac{b}{2c}\tanh\frac{c}{2}$ [49].

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

$$p(\Pi_\mu|f_{\text{old}}, \lambda^*_{\mu_{\text{old}}}) = \frac{p_{\lambda_\mu}(\Pi_\mu|\lambda^*_{\mu_{\text{old}}})\prod_{(\omega_\mu,t)\in\Pi_\mu} e^{h(\omega_\mu,-f_{\text{old}}(t))}}{\int p_{\lambda_\mu}(\Pi_\mu|\lambda^*_{\mu_{\text{old}}})\prod_{(\omega_\mu,t)\in\Pi_\mu} e^{h(\omega_\mu,-f_{\text{old}}(t))}d\Pi_\mu}. \tag{5.27}$$

Using Eq. 5.5 and Eq. 5.6 to convert the denominator, Eq. 5.27 can be rewritten as

$$\begin{aligned} &p(\Pi_\mu|f_{\text{old}}, \lambda^*_{\mu_{\text{old}}}) \\ &= \frac{p_{\lambda_\mu}(\Pi_\mu|\lambda^*_{\mu_{\text{old}}})\prod_{(\omega_\mu,t)\in\Pi_\mu} e^{h(\omega_\mu,-f_{\text{old}}(t))}}{\exp\left(-\iint(1-e^{h(\omega_\mu,-f_{\text{old}}(t))})\lambda^*_{\mu_{\text{old}}}p_{\text{PG}}(\omega_\mu|1,0)d\omega_\mu dt\right)} \\ &= \prod_{(\omega_\mu,t)\in\Pi_\mu}\left(e^{h(\omega_\mu,-f_{\text{old}}(t))}\lambda^*_{\mu_{\text{old}}}p_{\text{PG}}(\omega_\mu|1,0)\right) \\ &\cdot\exp\left(-\iint e^{h(\omega_\mu,-f_{\text{old}}(t))}\lambda^*_{\mu_{\text{old}}}p_{\text{PG}}(\omega_\mu|1,0)d\omega_\mu dt\right). \end{aligned} \tag{5.28}$$

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

$$\begin{aligned} \Lambda_\mu(t,\omega_\mu) &= e^{h(\omega_\mu,-f_{\text{old}}(t))}\lambda^*_{\mu_{\text{old}}}p_{\text{PG}}(\omega_\mu|1,0) \\ &= \lambda^*_{\mu_{\text{old}}}\sigma(-f_{\text{old}}(t))p_{\text{PG}}(\omega_\mu|1,f_{\text{old}}(t)). \end{aligned} \tag{5.29}$$

The derivation of the conditional distribution of Π_{ϕ_i} is the same as that of Π_μ, with the corresponding subscripts being replaced. It is worth noting that there exist

N independent marked Poisson processes $\{\Pi_{\phi_i}\}_{i=1}^N$ with the same intensity function

$$\Lambda_\phi(\tau, \omega_\phi) = \lambda^*_{\phi_{\text{old}}} \sigma(-g_{\text{old}}(\tau)) p_{\text{PG}}(\omega_\phi | 1, g_{\text{old}}(\tau)). \tag{5.30}$$

3. Combining Eq. 5.10 and Eq. 5.12 and marginalizing out $\boldsymbol{\omega}_{ii}$ and $\boldsymbol{\omega}_{ij}$, the conditional distribution of $\mathbf{B}$ is obtained:

$$p(\mathbf{B} | (\lambda^*_\mu, f, \lambda^*_\phi, g)_{\text{old}}) \propto \prod_{i=1}^N (\mu_{\text{old}}(t_i))^{b_{ii}} \prod_{i=2}^N \prod_{j=1}^{i-1} (\phi_{\text{old}}(\tau_{ij}))^{b_{ij}}, \tag{5.31}$$

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

$$\begin{aligned} p(b_{ii} = 1) &= \frac{\mu_{\text{old}}(t_i)}{\mu_{\text{old}}(t_i) + \sum_{j=1}^{i-1} \phi_{\text{old}}(\tau_{ij})} \\ p(b_{ij} = 1) &= \frac{\phi_{\text{old}}(\tau_{ij})}{\mu_{\text{old}}(t_i) + \sum_{j=1}^{i-1} \phi_{\text{old}}(\tau_{ij})}. \end{aligned} \tag{5.32}$$

5.5.1.2 Lower-bound of Log-posterior

Given those conditional densities above, the lower-bound $\mathcal{Q}$ can be computed. The expectation of log-likelihood (ELL) term in Eq. 5.25 can be rewritten as the summation of the baseline intensity part and the triggering kernel part. The ELL of baseline intensity part is

$$\begin{aligned} \text{ELL}_\mu(\lambda^*_\mu, f) &= \mathbb{E}_{P(\Pi_\mu, \boldsymbol{\omega}_{ii}, \mathbf{B}_{ii} | (\lambda^*_\mu, f, \lambda^*_\phi, g)_{\text{old}})} \left[\log p(D, \Pi_\mu, \boldsymbol{\omega}_{ii}, \mathbf{B}_{ii} | \lambda^*_\mu, f) \right] \\ &= -\frac{1}{2} \int_T A_\mu(t) f^2(t) dt + \int_T B_\mu(t) f(t) dt - \lambda^*_\mu T \\ &\left(\sum_{i=1}^N \mathbb{E}(b_{ii}) + \iint \Lambda_\mu(t, \omega_\mu) d\omega_\mu dt \right) \log \lambda^*_\mu, \end{aligned} \tag{5.33}$$

where

$$\begin{aligned} A_\mu(t) &= \sum_{i=1}^{N} \mathbb{E}[\omega_{ii}]\mathbb{E}[b_{ii}]\delta(t-t_i) + \int_0^\infty \omega_\mu \Lambda_\mu(t,\omega_\mu)d\omega_\mu \\ B_\mu(t) &= \frac{1}{2}\sum_{i=1}^{N} \mathbb{E}[b_{ii}]\delta(t-t_i) - \frac{1}{2}\int_0^\infty \Lambda_\mu(t,\omega_\mu)d\omega_\mu, \end{aligned} \tag{5.34}$$

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

Similarly, the ELL of triggering kernel part is written as

$$\begin{aligned} \text{ELL}_\phi(\lambda_\phi^*, g) =& \mathbb{E}_{P(\{\Pi_{\phi_i}\},\boldsymbol{\omega}_{ij},\mathbf{B}_{ij}|(\lambda_\mu^*,f,\lambda_\phi^*,g)_{\text{old}})}\left[\log p(D,\{\Pi_{\phi_i}\},\boldsymbol{\omega}_{ij},\mathbf{B}_{ij}|\lambda_\phi^*,g)\right] \\ &= -\frac{1}{2}\int_{T_\phi} A_\phi(\tau)g^2(\tau)d\tau + \int_{T_\phi} B_\phi(\tau)g(\tau)d\tau - N\lambda_\phi^* T_\phi \\ &\left(\sum_{i=2}^{N}\sum_{j=1}^{i-1}\mathbb{E}(b_{ij}) + N\iint \Lambda_\phi(\tau,\omega_\phi)d\omega_\phi d\tau\right)\log\lambda_\phi^*, \end{aligned} \tag{5.35}$$

where

$$\begin{aligned} A_\phi(\tau) =& \sum_{i,j} \mathbb{E}[\omega_{ij}]\mathbb{E}[b_{ij}]\delta(\tau-\tau_{ij}) + N\int_0^\infty \omega_\phi \Lambda_\phi(\tau,\omega_\phi)d\omega_\phi \\ B_\phi(\tau) =& \frac{1}{2}\sum_{i,j} \mathbb{E}[b_{ij}]\delta(\tau-\tau_{ij}) - \frac{N}{2}\int_0^\infty \Lambda_\phi(\tau,\omega_\phi)d\omega_\phi, \end{aligned} \tag{5.36}$$

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

However, the computation of ELL is intractable for general GP priors due to the fact that ELL is a functional. To circumvent the problem, the sparse GP approximation is utilized to introduce some inducing points. f and g are supposed to be dependent on their corresponding inducing points $\{t_s\}_{s=1}^{S_\mu}$ and $\{\tau_s\}_{s=1}^{S_\phi}$; the function values of f and g at these inducing points are $\mathbf{f}_{t_s}$ and $\mathbf{g}_{\tau_s}$. Given a sample $\mathbf{f}_{t_s}$ and $\mathbf{g}_{\tau_s}$, $f(t)$ and $g(\tau)$ in Eq. 5.33 and Eq. 5.35 are assumed to be the posterior mean functions

$$f(t) = \mathbf{k}_{t_s t}^T \mathbf{K}_{t_s t_s}^{-1}\mathbf{f}_{t_s}, \quad g(\tau) = \mathbf{k}_{\tau_s \tau}^T \mathbf{K}_{\tau_s \tau_s}^{-1}\mathbf{g}_{\tau_s} \tag{5.37}$$

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

Substituting Eq. 5.37 to Eq. 5.33 and Eq. 5.35,

$$\begin{aligned}&\mathcal{Q}((\lambda_\mu^*, \mathbf{f}_{t_s}, \lambda_\phi^*, \mathbf{g}_{\tau_s})|(\lambda_\mu^*, \mathbf{f}_{t_s}, \lambda_\phi^*, \mathbf{g}_{\tau_s})_{\text{old}})\\&= \text{ELL}_\mu(\lambda_\mu^*, \mathbf{f}_{t_s}) + \text{ELL}_\phi(\lambda_\phi^*, \mathbf{g}_{\tau_s}) - \frac{1}{2}\mathbf{f}_{t_s}^T\mathbf{K}_{t_st_s}^{-1}\mathbf{f}_{t_s} - \frac{1}{2}\mathbf{g}_{\tau_s}^T\mathbf{K}_{\tau_s\tau_s}^{-1}\mathbf{g}_{\tau_s}.\end{aligned} \tag{5.38}$$

5.5.2 M Step

In the M step, the lower-bound $\mathcal{Q}$ is maximized. The optimal parameters $\hat{\lambda}_\mu^*, \hat{\mathbf{f}}_{t_s}, \hat{\lambda}_\phi^*, \hat{\mathbf{g}}_{\tau_s}$ can be obtained by setting the gradient of Eq. 5.38 to zero. Due to auxiliary variables augmentation, analytical solutions are obtained:

$$\begin{aligned}\hat{\lambda}_\mu^* &= \left(\sum_{i=1}^N p_{ii} + M_\mu\right)/T\\ \hat{\lambda}_\phi^* &= \left(\sum_{i=2}^N \sum_{j=1}^{i-1} p_{ij} + NM_\phi\right)/(NT_\phi)\\ \hat{\mathbf{f}}_{t_s} &= \mathbf{\Sigma}_{t_s}\mathbf{K}_{t_st_s}^{-1}\int_T B_\mu(t)\mathbf{k}_{t_st}dt\\ \hat{\mathbf{g}}_{\tau_s} &= \mathbf{\Sigma}_{\tau_s}\mathbf{K}_{\tau_s\tau_s}^{-1}\int_{T_\phi} B_\phi(\tau)\mathbf{k}_{\tau_s\tau}d\tau\end{aligned} \tag{5.39}$$

where $\mathbf{\Sigma}_{t_s} = \left[\mathbf{K}_{t_st_s}^{-1}\int A_\mu(t)\mathbf{k}_{t_st}\mathbf{k}_{t_st}^T dt\mathbf{K}_{t_st_s}^{-1} + \mathbf{K}_{t_st_s}^{-1}\right]^{-1}$, $M_\mu = \iint \Lambda_\mu(t,\omega_\mu)d\omega_\mu dt$, $\mathbf{\Sigma}_{\tau_s} = \left[\mathbf{K}_{\tau_s\tau_s}^{-1}\int A_\phi(\tau)\mathbf{k}_{\tau_s\tau}\mathbf{k}_{\tau_s\tau}^T d\tau\mathbf{K}_{\tau_s\tau_s}^{-1} + \mathbf{K}_{\tau_s\tau_s}^{-1}\right]^{-1}$, $M_\phi = \iint \Lambda_\phi(\tau,\omega_\phi)d\omega_\phi d\tau$, $p_{ii} = p(b_{ii} = 1)$, $p_{ij} = p(b_{ij} = 1)$. All intractable integrals can be solved by Gaussian quadrature.

5.5.3 Complexity

Another advantage of sparse GP approximation is that the complexity of matrix inversion is fixed at $\mathcal{O}(S_\mu^3 + S_\phi^3)$ where S_μ (or S_ϕ) $\ll N$. This results in a complexity that scales almost linearly with data size: $\mathcal{O}(NL)$ where $L = \int_{T_\phi}\frac{\mu(t)}{1-\int\phi(\tau)d\tau}dt \ll N$ due to the sparsity of expectation of the branching structure: previous points that are more than T_ϕ far away from event i have no influence on event i ($\mathbb{E}[b_{ij}] = 0$).

5.5.4 Hyperparameters

Throughout this chapter, the GP covariance kernel is the squared exponential kernel $k(x, x') = \theta_0 \exp\left(-\frac{\theta_1}{2}\|x - x'\|^2\right)$. The hyperparameters θ_0 and θ_1 can be optimized by performing maximization of $\mathcal{Q}$ over $\boldsymbol{\theta} = \{\theta_0, \theta_1\}$ using numerical packages. Normally, $\boldsymbol{\theta}$ is updated every 20 iterations.

Additional hyperparameters are the number and location of inducing points which affect the complexity and estimation quality of $\mu(t)$ and $\phi(\tau)$. A large number of inducing points will lead to high complexity, while a small number cannot capture the dynamics. For fast inference, the inducing points are uniformly located on the domain. Another advantage of uniform location is that the kernel matrix has Toeplitz structure [46] which means that matrix inversion can be implemented more efficiently. The number of inducing points is gradually increased until no more significant improvement occurs. The final pseudo code is provided in Alg. 5.2.

Algorithm 5.2: EM algorithm for SGPHP

```
Result: $\mu(t) = \lambda_\mu^* \sigma(f(t))$, $\phi(\tau) = \lambda_\phi^* \sigma(g(\tau))$
Initialize hyperparameters and $\mathbf{B}$, $\lambda_\mu^*$, $\lambda_\phi^*$, $\boldsymbol{\omega}_{ii}$, $\boldsymbol{\omega}_{ij}$, $\mathbf{f}_{t_s}$, $\mathbf{g}_{\tau_s}$, $\Pi_\mu$, $\{\Pi_{\phi_i}\}_{i=1}^N$;
for do
    Update the posterior of $\boldsymbol{\omega}_{ii}$ and $\boldsymbol{\omega}_{ij}$ by Eq. 5.26;
    Update intensities of $\Pi_\mu$ and $\{\Pi_\phi\}$ by Eq. 5.29 and 5.30;
    Update the posterior of $\mathbf{B}$ by Eq. 5.32;
    Update $\lambda_\mu^*, \mathbf{f}_{t_s}, \lambda_\phi^*$ and $\mathbf{g}_{\tau_s}$ by Eq. 5.39;
    Update hyperparameters.
end
```

5.5.5 Experimental Results

The synthetic and real data experimental results of EM are shown together with that of the mean-field variational inference in Sec. 5.6.3.

5.6 Mean-field Variational Inference

The EM algorithm is now extended to a mean-field variational inference [37] algorithm which solves the inference problem slightly slower than EM, but can provide uncertainty with a distribution estimation rather than point estimation.

In variational inference, the posterior distribution over latent variables is approximated by a variational distribution. The optimal variational distribution is chosen by minimising the KL divergence or equivalently by maximizing the ELBO. A common approach is the mean-field method where the variational distribution is assumed to factorize over some partition of latent variables.

5.6.1 Optimal Variational Distributions

For the problem at hand, the joint distribution over all variables is shown in Eq. 5.11 and 5.13. It is assumed that the variational distribution q can be factorized as

$$\begin{aligned} &q(\Pi_\mu, \boldsymbol{\omega}_{ii}, \{\Pi_{\phi_i}\}_{i=1}^N, \boldsymbol{\omega}_{ij}, \mathbf{B}, \lambda_\mu^*, f, \lambda_\phi^*, g) = \\ &q_1(\Pi_\mu, \boldsymbol{\omega}_{ii}, \{\Pi_{\phi_i}\}_{i=1}^N, \boldsymbol{\omega}_{ij}, \mathbf{B}) q_2(\lambda_\mu^*, f, \lambda_\phi^*, g). \end{aligned} \tag{5.40}$$

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

$$\begin{aligned} &\log q_1(\Pi_\mu, \boldsymbol{\omega}_{ii}, \{\Pi_{\phi_i}\}_{i=1}^N, \boldsymbol{\omega}_{ij}, \mathbf{B}) = \\ &\quad \mathbb{E}_{q_2}[\log p(\Pi_\mu, \boldsymbol{\omega}_{ii}, \{\Pi_{\phi_i}\}_{i=1}^N, \boldsymbol{\omega}_{ij}, \mathbf{B}, \lambda_\mu^*, f, \lambda_\phi^*, g)] + C_1, \\ &\log q_2(\lambda_\mu^*, f, \lambda_\phi^*, g) = \\ &\quad \mathbb{E}_{q_1}[\log p(\Pi_\mu, \boldsymbol{\omega}_{ii}, \{\Pi_{\phi_i}\}_{i=1}^N, \boldsymbol{\omega}_{ij}, \mathbf{B}, \lambda_\mu^*, f, \lambda_\phi^*, g)] + C_2. \end{aligned} \tag{5.41}$$

Substituting Eq. 5.11 and 5.13 into Eq. 5.41, the optimal distribution for each factor can be obtained as follows. What is worth noting is that $\boldsymbol{\omega}_{ii}$ and $\boldsymbol{\omega}_{ij}$ are coupled with branching structure $\mathbf{B}$. $\boldsymbol{\omega}$ is marginalized out when solving $q_1(\mathbf{B})$ and vice versa.

Optimal Density of Pólya-Gamma Variables

$$
\begin{aligned}
q_1(\boldsymbol{\omega}_{ii}) &= \prod_{i=1}^{N} p_{\mathrm{PG}}(\omega_{ii}|1, \tilde{f}(t_i)) \\
q_1(\boldsymbol{\omega}_{ij}) &= \prod_{i=2}^{N}\prod_{j=1}^{i-1} p_{\mathrm{PG}}(\omega_{ij}|1, \tilde{g}(\tau_{ij})),
\end{aligned}
\tag{5.42}
$$

where $\tilde{f}(t_i) = \sqrt{\mathbb{E}(f^2(t_i))}$ and $\tilde{g}(\tau_{ij}) = \sqrt{\mathbb{E}(g^2(\tau_{ij}))}$ which is computed utilizing $\mathbb{E}(C^2) = \mathbb{E}^2(C) + \mathrm{Var}(C)$.

Optimal Marked Poisson Processes

$$
\begin{aligned}
\Lambda_\mu^1(t, \omega_\mu) &= \tilde{\lambda}_\mu^* \sigma(-\tilde{f}(t)) p_{\mathrm{PG}}(\omega_\mu|1, \tilde{f}(t)) e^{(\tilde{f}(t)-\bar{f}(t))/2} \\
\Lambda_\phi^1(\tau, \omega_\phi) &= \tilde{\lambda}_\phi^* \sigma(-\tilde{g}(\tau)) p_{\mathrm{PG}}(\omega_\phi|1, \tilde{g}(\tau)) e^{(\tilde{g}(\tau)-\bar{g}(\tau))/2},
\end{aligned}
\tag{5.43}
$$

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

Optimal Density of Intensity Upper-bounds

$$
\begin{aligned}
q_2(\lambda_\mu^*) &= \mathrm{Gamma}(\lambda_\mu^*|\alpha_\mu, \beta_\mu) \\
q_2(\lambda_\phi^*) &= \mathrm{Gamma}(\lambda_\phi^*|\alpha_\phi, \beta_\phi),
\end{aligned}
\tag{5.44}
$$

where $\alpha_\mu = \sum_{i=1}^N \mathbb{E}(b_{ii}) + \iint \Lambda_\mu^1(t, \omega_\mu) dt d\omega_\mu$, $\beta_\mu = T$, $\alpha_\phi = \sum_{i=1}^N \sum_{j=1}^{i-1} \mathbb{E}(b_{ij}) + N \iint \Lambda_\phi^1(\tau, \omega_\phi) d\tau d\omega_\phi$, $\beta_\phi = NT_\phi$ and all intractable integrals can be solved by Gaussian quadrature. This provides the required expectation for Eq. 5.43 by $\mathbb{E}(\lambda^*) = \alpha/\beta$ and $\mathbb{E}(\log \lambda^*) = \psi(\alpha) - \log \beta$ where $\psi(\cdot)$ is the digamma function.

Optimal Sparse Gaussian Process

$$\begin{aligned} q_2(\mathbf{f}_{t_s}) &= \mathcal{N}(\mathbf{f}_{t_s}|\tilde{\mathbf{m}}_{t_s}, \tilde{\mathbf{\Sigma}}_{t_s}) \\ q_2(\mathbf{g}_{\tau_s}) &= \mathcal{N}(\mathbf{g}_{\tau_s}|\tilde{\mathbf{m}}_{\tau_s}, \tilde{\mathbf{\Sigma}}_{\tau_s}), \end{aligned} \tag{5.45}$$

where $\tilde{\mathbf{\Sigma}}_{t_s} = \left[\mathbf{K}_{t_s t_s}^{-1} \int \tilde{A}_\mu(t)\mathbf{k}_{t_s t}\mathbf{k}_{t_s t}^T dt \mathbf{K}_{t_s t_s}^{-1} + \mathbf{K}_{t_s t_s}^{-1}\right]^{-1}$, $\tilde{\mathbf{m}}_{t_s} = \tilde{\mathbf{\Sigma}}_{t_s}\mathbf{K}_{t_s t_s}^{-1} \int \tilde{B}_\mu(t)\mathbf{k}_{t_s t} dt$ with $\tilde{A}_\mu(t) = \sum_{i=1}^N \mathbb{E}[\omega_{ii}]\mathbb{E}[b_{ii}]\delta(t-t_i) + \int_0^\infty \omega_\mu \Lambda_\mu^1(t, \omega_\mu) d\omega_\mu$ and $\tilde{B}_\mu(t) = \frac{1}{2}\sum_{i=1}^N \mathbb{E}[b_{ii}]\delta(t-t_i) - \frac{1}{2}\int_0^\infty \Lambda_\mu^1(t, \omega_\mu) d\omega_\mu$; $\tilde{\mathbf{\Sigma}}_{\tau_s} = \left[\mathbf{K}_{\tau_s \tau_s}^{-1} \int \tilde{A}_\phi(\tau)\mathbf{k}_{\tau_s \tau}\mathbf{k}_{\tau_s \tau}^T d\tau \mathbf{K}_{\tau_s \tau_s}^{-1} + \mathbf{K}_{\tau_s \tau_s}^{-1}\right]^{-1}$, $\tilde{\mathbf{m}}_{\tau_s} = \tilde{\mathbf{\Sigma}}_{\tau_s}\mathbf{K}_{\tau_s \tau_s}^{-1} \int \tilde{B}_\phi(\tau)\mathbf{k}_{\tau_s \tau} d\tau$ with $\tilde{A}_\phi(\tau) = \sum_{i,j} \mathbb{E}[\omega_{ij}]\mathbb{E}[b_{ij}]\delta(\tau-\tau_{ij}) + N\int_0^\infty \omega_\phi \Lambda_\phi^1(\tau, \omega_\phi) d\omega_\phi$ and $\tilde{B}_\phi(\tau) = \frac{1}{2}\sum_{i,j} \mathbb{E}[b_{ij}]\delta(\tau - \tau_{ij}) - \frac{N}{2}\int_0^\infty \Lambda_\phi^1(\tau, \omega_\phi) d\omega_\phi$. All intractable integrals are solved by Gaussian quadrature. Note also the similarity to EM algorithm in Eq. 5.39.

Optimal Density of Branching Structure

$$\begin{aligned} q_1(b_{ii} = 1) &= \frac{\tilde{\mu}(t_i)}{\tilde{\mu}(t_i) + \sum_{j=1}^{i-1} \tilde{\phi}(\tau_{ij})} \\ q_1(b_{ij} = 1) &= \frac{\tilde{\phi}(\tau_{ij})}{\tilde{\mu}(t_i) + \sum_{j=1}^{i-1} \tilde{\phi}(\tau_{ij})}, \end{aligned} \tag{5.46}$$

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

5.6.2 Hyperparameters

Similarly, the hyperparameters θ_0 and θ_1 can be optimized by performing maximization of ELBO over $\{\theta_0, \theta_1\}$ using numerical packages. The optimization of number and location of inducing points is the same as for the EM algorithm. The final pseudo code is provided in Alg. 5.3.

Algorithm 5.3: Mean-field algorithm for SGPHP

```
Result: $\mu(t) = \lambda_\mu^* \sigma(f(t))$, $\phi(\tau) = \lambda_\phi^* \sigma(g(\tau))$
Initialize hyperparameters and variational distributions of $\mathbf{B}$, $\lambda_\mu^*$, $\lambda_\phi^*$, $\boldsymbol{\omega}_{ii}$, $\boldsymbol{\omega}_{ij}$,
  $\mathbf{f}_{t_s}$, $\mathbf{g}_{\tau_s}$, $\Pi_\mu$, $\{\Pi_{\phi_i}\}_{i=1}^N$;
for do
    Update $q_1$ of $\boldsymbol{\omega}_{ii}$ and $\boldsymbol{\omega}_{ij}$ by Eq. 5.42;
    Update $\Lambda^1$ of $\Pi_\mu$ and $\{\Pi_\phi\}$ by Eq. 5.43;
    Update $q_2$ of $\lambda_\mu^*$ and $\lambda_\phi^*$ by Eq. 5.44;
    Update $q_2$ of $\mathbf{f}_{t_s}$ and $\mathbf{g}_{\tau_s}$ by Eq. 5.45;
    Update $q_1$ of $\mathbf{B}$ by Eq. 5.46;
    Update hyperparameters.
end
```

5.6.3 Experimental Results

The performance of the proposed EM and mean-field (MF) algorithms are evaluated on both synthetic and real-world datasets. Specifically, the proposed algorithms are compared to the following alternatives.

1. **Maximum Likelihood Estimation (MLE)**: the vanilla Hawkes process with constant μ and exponential decay triggering kernel;

2. **Wiener-Hopf (WH)**: a nonparametric algorithm for Hawkes process where μ is constant and $\phi(\tau)$ is a nonparametric function [11];

3. **Majorization Minimization Euler-Lagrange (MMEL)**: another nonparametric algorithm for Hawkes process with constant μ and smooth $\phi(\tau)$, which similarly utilizes the branching structure and estimates $\phi(\tau)$ by an Euler-Lagrange equation [10].

The long short-term memory (LSTM) based neural Hawkes process [53] was also tried, but it was hard to converge at least on the datasets used in this work. On the contrary, the proposed algorithms are easier to converge due to the fact that there are fewer parameters to tune, which constitutes another advantage.

The following metrics were used to evaluate the performance of the various methods:

1. ***TestLL***: the log-likelihood of hold-out data using the trained model. This is a metric describing the model prediction ability. It is used to measure the performance of synthetic and real data.
2. ***EstErr***: the mean squared error between the estimated $\hat{\mu}(t)$, $\hat{\phi}(\tau)$ and the ground truth. It is only used for synthetic data.
3. ***RunTime***: the running time of various methods w.r.t. the training dataset size. It is used for synthetic data.
4. ***Q-Q plot***: the plot visualizes the goodness-of-fit for different models using time rescaling theorem [47]. It is used for the real data.

Table 5.4: *EstErr* and *TestLL* for synthetic and real datasets.

		MLE	WH	MMEL	EM	MF
Case 1	$EstErr(\hat{\mu}, \mu)$	0.236	0.228	0.173	**0.137**	0.223
	$EstErr(\hat{\phi}, \phi)$	0.0289	0.0039	0.0053	0.0016	**0.0005**
	TestLL	31.21	33.72	32.78	33.87	**33.98**
Case 2	$EstErr(\hat{\mu}, \mu)$	1.141	0.706	1.135	0.134	**0.099**
	$EstErr(\hat{\phi}, \phi)$	0.0076	0.0086	0.0082	**0.0011**	0.0020
	TestLL	27.48	22.97	26.35	**33.18**	32.58
Collision	*TestLL*	420.41	439.67	470.56	470.34	**494.46**
Crime	*TestLL*	371.59	400.36	381.42	**520.22**	375.29

5.6.3.1 Synthetic Data Experiments

In synthetic data experiments, the thinning algorithm [38] was used to generate 100 sets of training data and 10 sets of hold-out data with $T_\phi = 6$ and $T = 100$ in two different cases: (1) $\mu(t)$ is a constant and (2) $\mu(t)$ is changing over time.

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

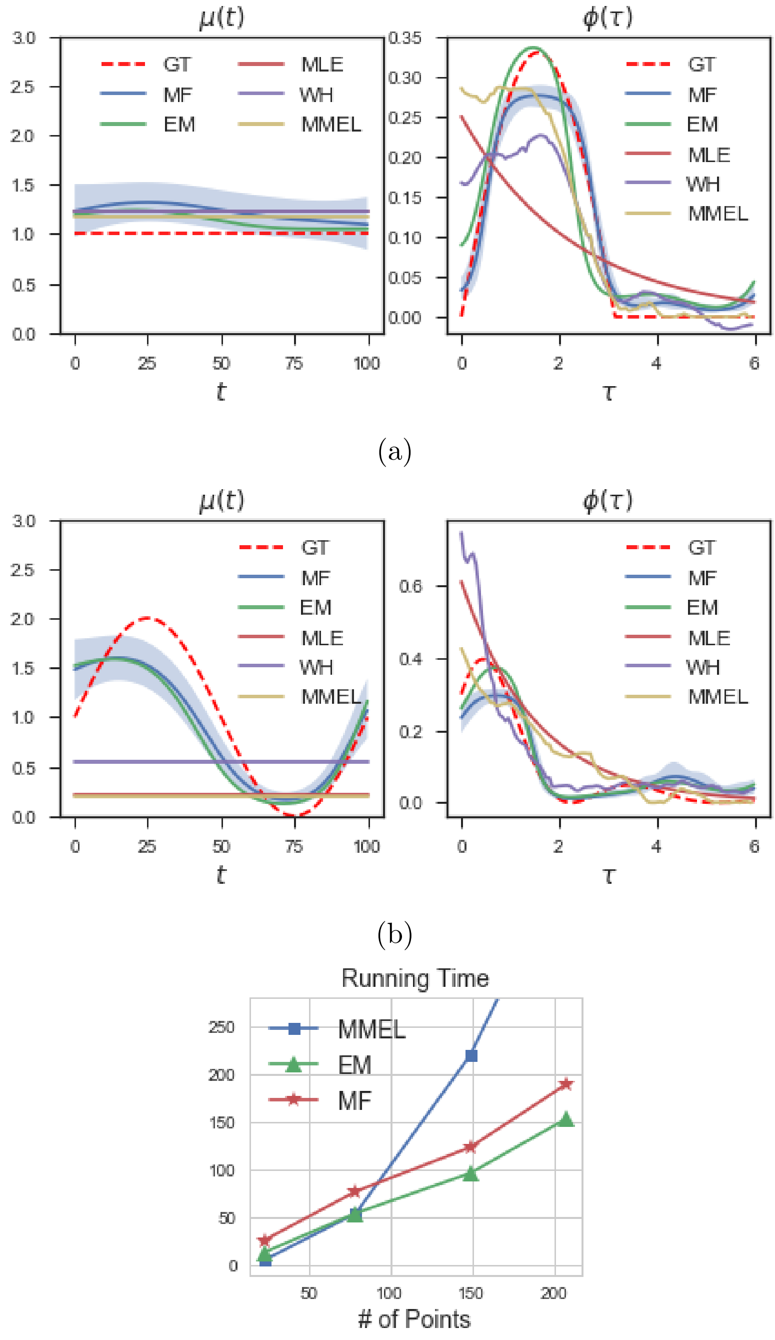
2. $\mu(t) = \sin\left(\frac{2\pi}{T}\cdot t\right) + 1 \quad (0 < t < T)$ and $\phi(\tau) = 0.3\left(\sin(\frac{2\pi}{3}\cdot\tau) + 1\right)\cdot\exp(-0.7\tau) \quad (0 < \tau < T_\phi)$.

Case 1 has constant $\mu(t)$ and a half sinusoidal triggering kernel $\phi(\tau)$. The bandwidth of WH is set to 0.3 and there are 10 inducing points on both $\mu(t)$ and $\phi(\tau)$ for EM and MF. Case 2 is more general, with time-changing $\mu(t)$ and a sinusoidal exponential decay triggering kernel. The bandwidth of WH is 0.1 and there are 10 inducing points on both $\mu(t)$ and $\phi(\tau)$ for EM and MF. The inducing points and hyperparameters are optimized for inference. The estimated $\hat{\mu}(t)$ and $\hat{\phi}(\tau)$ are shown in Fig. 5.3a and 5.3b. From an accuracy perspective, the results in Tab. 5.4 confirm that the EM and MF algorithms outperform the alternatives in all cases w.r.t. *TestLL* and *EstErr* (mean for MF). For MLE, the estimated results are far from the ground truth due to parametric constraints. For WH and MMEL, the constant limitation on $\mu(t)$ still exists even though $\phi(\tau)$ has been relaxed. On the contrary, the proposed EM and MF algorithms provide nonparametric $\hat{\mu}(t)$ and $\hat{\phi}(\tau)$ concurrently. From an efficiency perspective, both EM and MF algorithms were compared to the other iterative nonparametric algorithm MMEL with the same number of iterations. EM and MF's *RunTime* scales linearly with the number of observations in Fig. 5.3c, which establishes their superior efficiency.

5.6.3.2 Real Data Experiments

In the real data section, both the EM and MF algorithms were applied to two different datasets. The ground truth information is not available for the real world data, therefore the *EstErr* cannot be utilized to measure the performance. As a result, the performance of real data is quantified by the metrics *TestLL* and *Q-Q plot*.

Motor Vehicle Collisions in New York City [39]: I filter out weekday



(a)

(b)

(c)

Figure 5.3: Synthetic data experimental results. Estimated $\hat{\mu}(t)$ and $\hat{\phi}(\tau)$ for (a) Case 1 (with shading being one standard deviation for MF) (b): Case 2. Clearly both EM and MF algorithms capture the structure of underlying rates better than other alternatives. (c): Running time (seconds) of different iterative nonparametric algorithms on varying # of observations. Both EM and MF algorithms scale linearly with # of observations, which is more efficient than MMEL. (GT=Ground Truth)

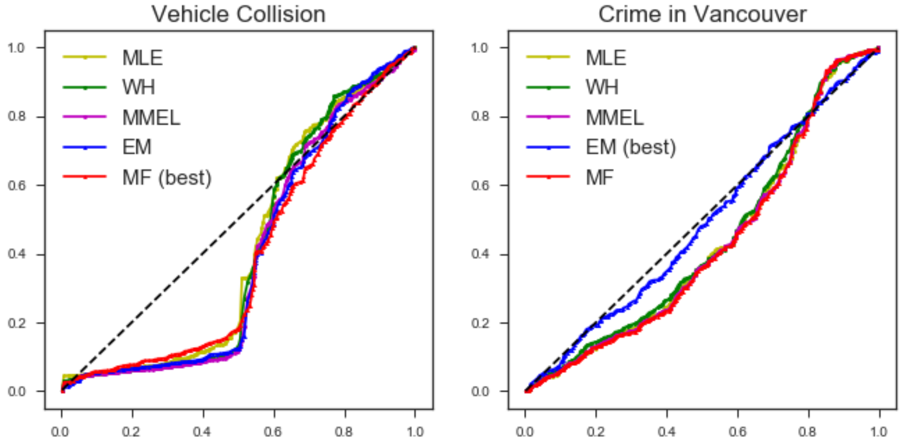
records in nearly one month (Sep.18th-Oct.13th 2017). The number of collisions in each day is about 600. Records in Sep.18th-Oct.6th are used as training data and Oct.9th-13th are held out as test data.

The performance of EM, MF and other alternatives were compared on this dataset. The whole observation period T was set to 1440 minutes (24 hours) and the support of the triggering kernel T_ϕ was set to 60 minutes. For hyperparameters, the bandwidth of WH was set to 0.3 and there are 10 inducing points on $\mu(t)$ and $\phi(\tau)$ for EM and MF with hyperparameters $\{\theta_0, \theta_1\}$ optimized for inference. The final result is the average of learned $\mu(t)$ or $\phi(\tau)$ of each day.

Crime in Vancouver [52]: I filter out the crime records from 2016-06-01 to 2016-08-31 as training data and 2016-09-01 to 2016-11-30 as test data, add a small time interval to separate all the simultaneous records and delete some records with empty values.

The whole observation period T was set to 92 days and the support of $\phi(\tau)$ was set to 6 days. For hyperparameters, the bandwidth of WH was set to 0.5 and there are still 10 inducing points on $\mu(t)$ and $\phi(\tau)$ for EM and MF with hyperparameters $\{\theta_0, \theta_1\}$ optimized for inference.

The performance of the proposed algorithms was compared to the alternatives. For each inference algorithm, its predictive performance was evaluated using *TestLL*. The *TestLL* of EM, MF and other alternatives are shown in Tab. 5.4. The proposed EM and MF algorithms are consistently superior over the alternatives whose baseline intensity or triggering kernel is too restricted to capture the dynamics. To further measure performance, the *Q-Q plot* was generated. A sequence of timestamps in the hold-out data was transformed by the fitted model to a set of independent uniform random variables on the interval $(0, 1)$. The result is shown in Fig. 5.4. All experimental results establish that the proposed algorithms can not only describe $\mu(t)$ and $\phi(\tau)$ in a completely flexible manner leading to better goodness-of-fit, but also with superior efficiency.



(a) (b)

Figure 5.4: *Q-Q plot* of EM, MF and other alternative inference algorithms for the real datasets.

5.7 Discussion

In this section, the setting of hyperparameters and the advantages and disadvantages of the proposed algorithms are discussed. As in the previous chapters, the hyperparameters θ_0 and θ_1 have a significant effect on the estimation. In order to tune them carefully, the hyperparameters are updated regularly in the Gibbs sampler, EM iterations or mean-field iterations from its posterior distribution, to maximize the surrogate function or to maximize the ELBO.

The advantage of the Gibbs sampler is that the derivation is relatively easy compared with EM and MF and the estimated posterior is accurate, however the disadvantage is that the Gibbs sampler is relatively slow because the latent Poisson process has to be sampled, which is a time-consuming procedure. On the contrary, the EM algorithm is the most efficient and can provide the exact MAP estimate, but the disadvantage is that it can only provide a point estimation and not a posterior distribution representing the uncertainty. The MF algorithm is intermediate, achieving good efficiency and also providing uncertainty estimation, but the drawback is that the estimation is an approximation.

5.8 Summary

In this chapter, efficient Gibbs sampling, EM and MF variational inference algorithms are proposed for the sigmoid GP Hawkes process. By augmenting the branching structure, Pólya-Gamma random variables and latent marked Poisson processes, the inference can be performed efficiently in a conjugate way. Furthermore, by introducing a sparse GP approximation, the proposed algorithms scale linearly with the number of observations. The experimental results on synthetic and real datasets confirm that the accuracy and efficiency of the proposed algorithms are superior to state-of-the-art alternatives.

Chapter 6

Fast Multi-resolution Segmentation for Nonstationary Hawkes Process[*]

As discussed in Sec. 2.1, the conditional intensity of the Hawkes process is unchanged over timeshifting because μ is a constant and $\phi(\cdot)$ only depends on $\tau = t - t_i$ and not on t, which provides stationarity. The assumption of stationarity leads to easy inference, but it is inconsistent and not very useful with real applications. Applying the vanilla Hawkes process directly to the nonstationary data is inappropriate. On the other hand, nonstationarity itself may be an important feature. Discovering the underlying nonstationarity will provide more valuable temporal information that stationary models cannot provide.

One of the common methods of analyzing nonstationary time series is to use segmentation. This type of problem is also called a change-point problem and is well studied [19]. Given a nonstationary point process data, a segmentation algorithm will divide the whole observation period into several non-overlapping contiguous

[*] Portions of this chapter have been deposited in arXiv: F. Zhou, Z. Li, X. Fan, Y. Wang, A. Sowmya, F. Chen, Fast Multi-resolution Segmentation for Nonstationary Hawkes Process Using Cumulants. arXiv preprint arXiv:1906.02438. 6 Jun 2019.

segments such that each segment is more approximately stationary than the original data and can be assumed to be stationary.

The goal of this chapter is to propose a multi-resolution segmentation algorithm for the nonstationary Hawkes process which can reveal the optimal partition structure at different resolutions. Multi-resolution segmentation is meaningful in real data applications. For example, when the traffic data is analysed later in Sec. 6.4, the lower resolution partition (e.g. two segments) corresponds to a "coarser" distinction such as between day and night, while the higher resolution partition (e.g. four or six segments) corresponds to a "finer" distinction such as between alternating busy and non-busy hours. This will help provide more insights into the nonstationary structure of point process data.

Intuitively, the most appropriate segmentation may be found using the estimated baseline intensities and triggering kernels at different times. However, the estimation of μ and $\phi(\cdot)$ is time consuming. For computational efficiency, μ and $\phi(\cdot)$ are not estimated directly, instead cumulants are used, which can be estimated quickly. Consequently, segmentation can be fast to implement. Details are discussed in Sec. 6.1.

After segmentation, the baseline intensity and triggering kernel are learned piecewise on each segment. Specifically, the Wiener-Hopf method [11] is used. Consequently, the learned triggering kernel is nonparametric. Overall, this defines a nonstationary and nonparametric Hawkes process model.

The performance of the segmentation algorithm depends on the choice of hyperparameters, which is investigated in Sec. 6.2.4. To ease the choice of hyperparameters, the idea of GP-MISD (see Chapter 3) is used to make the algorithm more robust. Overall, this chapter makes the following contributions:

1. A multi-resolution segmentation (MRS) algorithm is proposed that can partition the nonstationary Hawkes process into a desired output resolution.

2. The MRS algorithm does not directly depend on the estimation of μ and $\phi(\cdot)$

which is time consuming, but instead on the estimation of cumulants of the Hawkes process which are fast to compute. Consequently, the MRS algorithm is fast, with linear time computation complexity.

3. The idea of GP-MISD is incorporated into the MRS algorithm in order to ease the choice of the hyperparameter.

In the following, the Naïve MRS algorithm is proposed in Sec. 6.1. In Sec. 6.2, synthetic data experiments are performed for constant and non-constant baseline intensity cases; also the complexity and the influence of hyperparameters are analysed. In Sec. 6.3, a novel GP based MRS algorithm is proposed to solve the problem of choosing hyperparameters. The real traffic data is analysed in Sec. 6.4. The approach is discussed in Sec 6.5, and remarks are provided in Sec. 6.6.

6.1 Multi-resolution Segmentation

Assume that there is a set of observations $\{t_i\}_{i=1}^N$ on $[0,T]$ from a nonstationary Hawkes process, where the baseline intensity μ is constant (the case with nonconstant $\mu(t)$ is later shown in Sec. 6.2.2) but the triggering kernel $\phi(\tau)$ is changing over time t. Given M, the basic idea of MRS is to uniformly divide the observation period $[0,T]$ into M sectors at the highest resolution, e.g. $s_1, s_2, ..., s_M$, where $\{s_j\}_{j=1}^M$ are sectors and $|s_j|$ is the width of the sector. In each s_j, the point process is assumed to be stationary.

Intuitively, the triggering kernel $\phi(\tau)$ can be estimated in each sector, and adjacent pairs compared directly to find the maximum possible partitioning positions. However, the estimation of $\phi(\tau)$ is time consuming, whether it is performed in parametric way (MLE) or nonparametric way (MISD algorithm, Wiener-Hopf method), and estimating on all sectors is even more time consuming. In order to increase computational efficiency, $\phi(\tau)$ is not estimated in each sector directly, instead the second order statistics $g_j(\tau)$ which can be estimated fast are used. The second order statistics $g_j(\tau)$ in each sector can be empirically estimated using the empirical version of Eq. 2.7.

The reason $\phi(\tau)$ in each sector can be replaced with $g_j(\tau)$ is that there is a one-to-one mapping between them, so the difference between two adjacent $g_j(\tau)$ stands for the nonstationarity of $\phi(\tau)$. The difference of two adjacent $g_j(\tau)$ is written as a normalized mean squared error (NMSE):

$$NMSE = \mathbb{E}_\tau \left((\frac{g_j(\tau)}{\int g_j(\tau)d\tau} - \frac{g_{j+1}(\tau)}{\int g_{j+1}(\tau)d\tau})^2 \right). \tag{6.1}$$

In most of cases, $g_j(\tau)$ is an even function for 1-variate Hawkes process when $\tau \to \pm\infty$ $g_j(\tau) \to 0$. If $g_j(\tau)$ is expressed as a histogram function $g_j(\tau) = \sum_{k=1}^{K}(g_j^k \delta_{kh})$ where $\delta_{kh}(\tau) = 1$ if $(k-1)h \leq \tau < kh$ and 0 otherwise, h is the bin-width and $g_j(\tau)$ is 0 beyond the support of Kh, $g_j(\tau)$ can be written as a vector $\mathbf{g}_j = [g_j^k]_{k=1}^K$ (see Fig. 6.1). Eq. 6.1 also can be converted to a discrete version:

$$NMSE = \frac{\sum_{k=1}^{K} \left((\frac{g_j^k}{2h\sum_{k=1}^{K} g_j^k} - \frac{g_{j+1}^k}{2h\sum_{k=1}^{K} g_{j+1}^k})^2 \right)}{K}. \tag{6.2}$$

Then a desired number of segments (the desired output resolution) R is set. Given NMSE on all the candidate cutting positions, the largest $R-1$ cutting positions are picked, which produce the segmentation. Alternatively, a threshold corresponding to R can be used: if the NMSE between $g_j(\tau)$ and $g_{j+1}(\tau)$ is smaller than the threshold, the two adjacent sectors are considered to be approximately stationary and are not partitioned; otherwise the partition occurs at the current candidate position. The MRS scheme is shown in Fig. 6.2.

By adjusting the desired output resolution R, the multi-resolution segmentation algorithm will output segments at different resolutions. For example, when $R = M$ the partitioner will output the highest resolution (cutting at all candidate positions), and as R becomes smaller the output resolution will be lower (fewer segments will be created) until there is no partition at all.

After performing the segmentation, the second order statistics $g(\tau)$ on each segment can be used to nonparametrically solve the triggering kernel $\phi(\tau)$ on each segment. As proved elsewhere [11], $\phi(\tau)$ and $g(\tau)$ satisfy the Wiener-Hopf equation (Eq. 2.16). In most cases, this equation cannot be solved analytically, but there exist many methods [54, 55] to solve it numerically. A common one is the Nystrom method [23]. Its basic idea is to use the Gaussian quadrature method to numerically approximate the convolution in Eq. 2.16, consequently, Eq. 2.16 is converted to a standard linear system that can be easily solved by inversion. It is worth mentioning that Eq. 2.16 only gives out $\phi(\tau)$, and to estimate μ, the first order cumulant Eq. 2.4 has to be used.

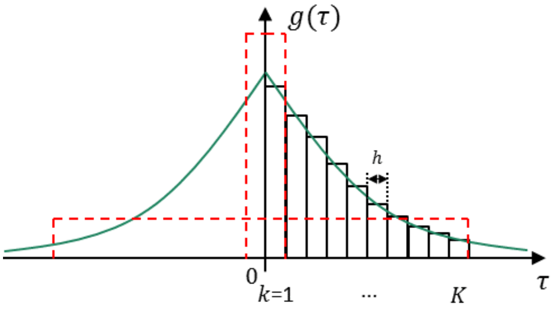


Figure 6.1: $g(\tau)$ is expressed as a histogram. Red dashed lines correspond to two extreme cases.

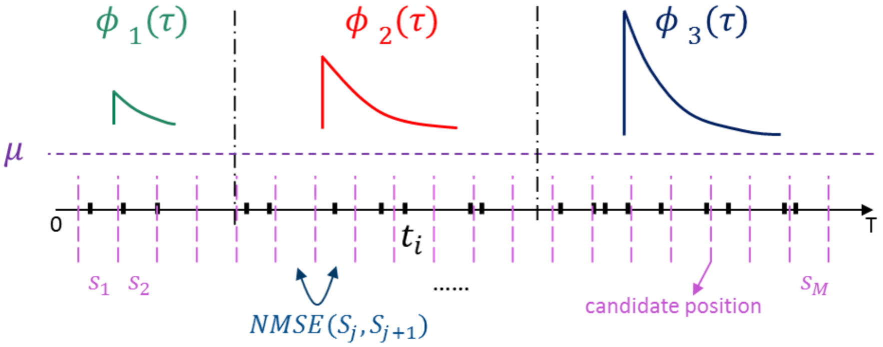


Figure 6.2: Multi-resolution segmentation scheme, for simplicity μ is assumed to be constant and that there are only 3 different $\phi(\tau)$'s distributed on $[0, T]$.

The pseudocode for the MRS algorithm and the estimation of μ and $\phi(\tau)$ is formally presented in Alg. 6.1.

```
Algorithm 6.1: MRS Algorithm and estimation of $\mu$'s and $\phi(\tau)$'s
  Input: $\{t_i\}_{i=1}^N$, $T$, $R$, $M$, $K$
  Output: partition positions, $\mu$ and $\phi(\tau)$ on each segment
  Function
      Uniformly divide $[0, T]$ into sectors $\{s_j\}_{j=1}^M$;
      Estimate the second order statistics $\mathbf{g}_j = [g_j^k]_{k=1}^K$ on each $s_j$ using Eq. 2.7;
      Compute the NMSE between two adjacent $\mathbf{g}_j$ using Eq. 6.2;
      Set a desired output resolution $R$ to obtain the partition positions;
      After segmentation, estimate the second order statistics $g(\tau)$ on each
        segment using Eq. 2.7;
      Estimate $\mu$ and $\phi(\tau)$ on each segment using Eq. 2.16 and Eq. 2.4;
      Return partition positions, $\mu$'s and $\phi(\tau)$'s.
  end
```

6.2 Synthetic Data Experiments

6.2.1 Constant Baseline Intensity

In this part, the thinning algorithm [38] is used to independently generate 40 sets of observations $\{\{t_i\}_{i=1}^{N_l}\}_{l=1}^{40}$ on $[0, 1000]$ from a nonstationary Hawkes process, where N_l is the number of points on l-th observation, μ is assumed to be constant 1, there are 3 different triggering kernels $\phi_1(\tau) = 1 \cdot \exp(-2\tau)$, $\phi_2(\tau) = 2 \cdot \exp(-3\tau)$ and $\phi_3(\tau) = 3 \cdot \exp(-4\tau)$ distributing on $[0, 200]$, $[200, 600]$ and $[600, 1000]$ respectively (see Fig. 6.2).

The goal is to find the underlying partition structure and estimate μ and $\phi(\tau)$'s in a nonparametric way. The highest resolution is set to be $M = 10$ ($|s_j| = 100$), $g_j(\tau)$ is expressed as a histogram function $g_j(\tau) = \sum_{k=1}^K (g_j^k \delta_{kh})$ where $h = 0.75$ and $K = 8$.

It is worth mentioning that the difference between two adjacent estimated $g_j(\tau)$ and $g_{j+1}(\tau)$ comes from two sources: the first is the nonstationarity of $\phi(\tau)$ which is desired, and the second is the randomness of $g_j(\tau)$ induced by the estimation variance. As the highest resolution M becomes higher ($|s_j|$ becomes smaller), there will be fewer observation points in each sector and, consequently, the variance of estimated $g_j(\tau)$ will be larger and the difference coming from the second source will be larger, which will lead to a misidentified segmentation. This problem is due to the influence of the hyperparameter M which will be discussed in Sec. 6.2.4. For now, the estimated $g_j(\tau)$ is averaged over 40 sets of independent observations to eliminate the variance as far as possible. Similarly, as K becomes larger (h becomes smaller), there will be fewer observation points in each bin and, consequently, the estimation variance of $g_j(\tau)$ will be larger (see Fig. 6.8) and the difference from the second source will also be larger, which will lead to a misidentified segmentation. This problem is due to the influence of the hyperparameter K, which will be discussed in Sec. 6.2.4. For now, $K = 8$, which was chosen empirically.

In Tab. 6.1, the multi-resolution segmentation results on the synthetically generated dataset are shown as R becomes larger from 1 to the highest resolution $M = 10$. As discussed, when R is 1, there is no segmentation at all; as R becomes larger the output resolution increases, and when R is 3, the partition positions totally match the ground truth; when $R = 10$ the algorithm segments at every candidate position (the highest resolution). To quantify the difference caused by estimation variance, the proportion of the minimum threshold (corresponding to R) over the maximum NMSE (see Fig. 6.6) is shown in Tab. 6.1. Clearly the difference between two adjacent estimated $g_j(\tau)$ induced by estimation variance is below 50% (the last correct segmentation is "600" which corresponds to 50.75%), which means that the MRS algorithm is robust enough to produce the correct segmentation.

Setting $R = 3$, the correct segmentation $[0, 200]$, $[200, 600]$, $[600, 1000]$ is obtained. The next step is to infer μ and $\phi(\tau)$ on each segment. The second order statistics $g(\tau)$ are empirically estimated on each segment and the Wiener-Hopf equation (Eq. 2.16) solved. The $\hat{\mu}$'s estimated in three segments are averaged to obtain the final $\hat{\mu}$. The final estimated $\hat{\mu} = 0.89$ and the estimated $\hat{\phi}(\tau)$'s are shown in

Table 6.1: Multi-resolution segmentation results (μ is constant), the “new position” is the newly added partition position

R	1	2	3	4	5
New Position	$\emptyset$	200	600	100	800
$\frac{\text{Min(Threshold)}}{\text{Max(NMSE)}}$	100%	82.61%	50.75%	47.12%	44.47%
R	6	7	8	9	10
New Position	500	900	400	700	300
$\frac{\text{Min(Threshold)}}{\text{Max(NMSE)}}$	34.65%	33.83%	22.19%	6.72%	0%

Fig. 6.3. It can be seen that the estimation results match the ground truth.

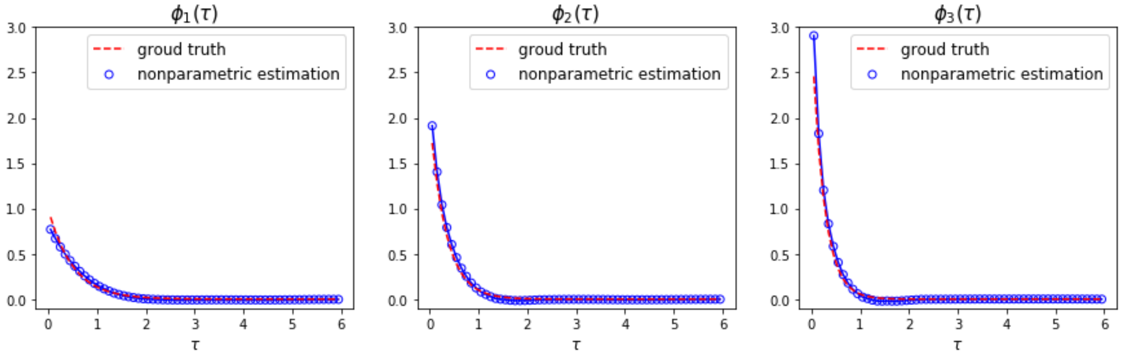


Figure 6.3: μ is constant: the estimated $\hat{\phi}_1(\tau)$, $\hat{\phi}_2(\tau)$ and $\hat{\phi}_3(\tau)$ distributed on $[0, 200]$, $[200, 600]$ and $[600, 1000]$. The ground truth are $1 \cdot \exp(-2\tau)$, $2 \cdot \exp(-3\tau)$ and $3 \cdot \exp(-4\tau)$, respectively.

6.2.2 Nonconstant Baseline Intensity

The μ was constant in Sec. 6.2.1. In this section, μ is relaxed to be nonconstant where μ is 2, 1.5 and 1 on $[0, 200]$, $[200, 600]$ and $[600, 1000]$ respectively. The thinning algorithm is used to independently generate 40 sets of observations on $[0, 1000]$, with triggering kernels $\phi_1(\tau) = 1 \cdot \exp(-2\tau)$, $\phi_2(\tau) = 2 \cdot \exp(-4\tau)$ and $\phi_3(\tau) = 3 \cdot \exp(-4\tau)$ distributing on $[0, 200]$, $[200, 600]$ and $[600, 1000]$, respectively.

The multi-resolution segmentation results are shown in Tab. 6.2. The result is similar to that in Tab. 6.1: the algorithm does not cut when $R = 1$ and has only one cut at 600 when $R = 2$; it provides the ground truth partition when $R = 3$

and the highest resolution when $R = 10$. From the proportions in the third and sixth rows, clearly the difference induced by estimation variance is smaller than Tab. 6.1: it is below 11% (the last correct cutting is "200" which corresponds to 11.41%). Actually, this is a direct consequence of more observation points because the baseline intensity is higher in this case.

Table 6.2: Multi-resolution segmentation results (μ is nonconstant), the "new position" is the newly added partition position

R	1	2	3	4	5
New Position	$\emptyset$	600	200	500	900
$\frac{\text{Min(Threshold)}}{\text{Max(NMSE)}}$	100%	88.45%	11.41%	8.05%	7.15%
R	6	7	8	9	10
New Position	400	300	700	800	100
$\frac{\text{Min(Threshold)}}{\text{Max(NMSE)}}$	6.88%	5.17%	2.24%	1.25%	0%

As before, μ and $\phi(\tau)$ are inferred on each segment after obtaining the correct segmentation $[0, 200]$, $[200, 600]$ and $[600, 1000]$. The estimated $\hat{\mu}_1 = 2.05$, $\hat{\mu}_2 = 1.64$, $\hat{\mu}_3 = 1.01$ and $\hat{\phi}(\tau)$'s are shown in Fig. 6.4. It can be seen that the estimation result matches with the ground truth.

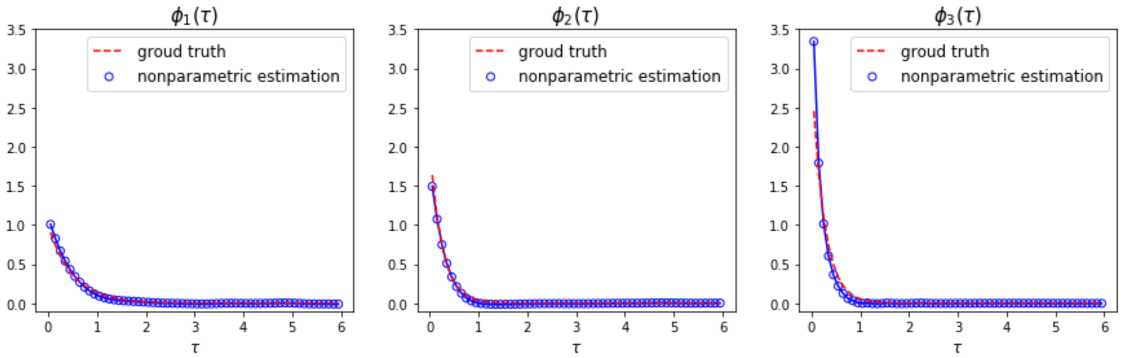


Figure 6.4: μ is nonconstant: the estimated $\hat{\phi}_1(\tau)$, $\hat{\phi}_2(\tau)$ and $\hat{\phi}_3(\tau)$ distributing on $[0, 200]$, $[200, 600]$ and $[600, 1000]$. The ground truth are $1 \cdot \exp(-2\tau)$, $2 \cdot \exp(-4\tau)$ and $3 \cdot \exp(-4\tau)$, respectively.

6.2.3 Computational Complexity

There are mainly two steps involved in the estimation procedure: the first step is to estimate the optimal segmentation and the second step is to estimate μ and $\phi(\tau)$ on each segment. In this chapter, the reference to the MRS algorithm is only to the first segmentation searching step. The computational complexity of the MRS algorithm is analysed now. The second step is the Wiener-Hopf equation method whose complexity can be found elsewhere [11]. It will be shown that the MRS algorithm has linear time computational complexity, which means that it is practical in real applications.

It can be seen later that the complexity of MRS mainly depends on two parameters: the highest resolution M and the total number of points on all sets of independent observations multiplying the number of bins on $g_j(\tau)$: $\mathcal{N}K$ where $\mathcal{N} = \sum_l N_l$.

The complexity of estimation of second order statistics $g_j(\tau)$ is approximately $\mathcal{O}(K)$, so the complexity of estimation of $g_j(\tau)$ on each sector is approximately $\mathcal{O}(n_jK)$ where n_j is the number of points in s_j, so the complexity of all $g_j(\tau)$ on l-th observation is $\mathcal{O}(N_lK)$, consequently, the complexity of all $g_j(\tau)$ on all sets of independent observations is $\mathcal{O}(\mathcal{N}K)$. The complexity of averaging the estimated $g_j(\tau)$ over independent observations on M sectors is $\mathcal{O}(M)$ and the complexity of NMSE between two adjacent $g_j(\tau)$ over M sectors is $\mathcal{O}(M-1)$. So the final complexity of MRS is $\mathcal{O}(\mathcal{N}K + M)$.

For a fixed M e.g. $M = 10$, the optimal segmentations for different sizes of $\mathcal{N}K$ are searched and the experimental results are shown in Fig. 6.5 left. It can be seen that the computational time is linear with $\mathcal{N}K$. For fixed $\mathcal{N}K$ e.g. $\mathcal{N}K = 31,547 \times 8$, the optimal segmentations for different sizes of M are then searched and the experiment results are shown on the right. The computational time is linear with the highest resolution M.

Also MRS is fast because it only takes about 10 seconds for about 120,000 observation points when $K = 8$ and $M = 10$ on a normal desktop (CPU: i7-6700 with

8GB RAM), which proves its practicability in real problems. The computational time for estimation of $\phi(\tau)$ and $g(\tau)$ were also compared: given 1,896 observation points, the computational time of $g(\tau)$ is only 0.5 second, while that of $\phi(\tau)$ is 38.4 seconds, which shows that replacing $\phi(\tau)$ with $g(\tau)$ saves time.

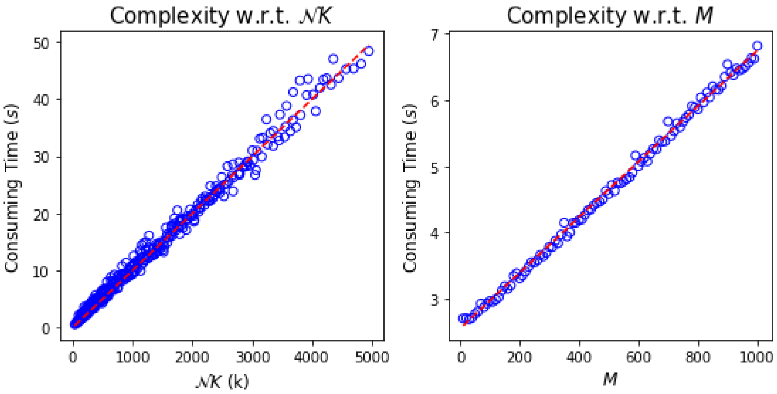


Figure 6.5: Computational time of MRS w.r.t. (left) $\mathcal{N}K$ for $M = 10$ and (right) w.r.t. M for $\mathcal{N}K = 31,547 \times 8$.

6.2.4 Influence of Hyperparameters

As stated in Sec. 6.2.3, the difference between two adjacent estimated $g_j(\tau)$ and $g_{j+1}(\tau)$ is made up of two sources: the first source is the difference between $\mathbb{E}(g_j(\tau))$ and $\mathbb{E}(g_{j+1}(\tau))$ which is the nonstationarity of $\phi(\tau)$ that is desired, and the second source is the estimation variance of $g_j(\tau)$ induced by the choice of hyperparameters. Given a sufficient number of observation points, there are two hyperparameters affecting the performance of MRS: M and K. In this section, the same experiment as in Sec. 6.3.1 is repeated for different values of M and K to observe the influence of each hyperparameter.

Hyperparameter M Intuitively, the highest resolution M should not be too small or too large. If too small, there are very candidate partition positions, and consequently, the segmentation result from MRS will not be good. If too large, there will be fewer points in each sector s_j given finite observation points on $[0, T]$, which means that the estimated $g_j(\tau)$ is far from the $\mathbb{E}(g_j(\tau))$, and consequently the segmentation result from MRS will not be good either.

Experiments were performed for fixed K e.g. $K = 8$, and increasing M from 3 to 20. The results with $R = 3$ are shown in Tab. 6.3 and the corresponding NMSE results are shown in Fig. 6.6. As can be seen when M is in the range $[10, 16]$ the segmentation result from MRS is close to the ground truth, and when M is larger than 20 the first source difference has been flooded by the second source (the estimation variance) and, consequently, the partition positions are misidentified.

Table 6.3: Segmentation results w.r.t. M

Highest Resolution M	3	8	10
Partition Positions	333.33,666.66	125,250	200,600
Highest Resolution M	12	16	20
Partition Positions	166.67,666.67	187.5,687.5	150,350

Hyperparameter K As will be found, given an appropriate highest resolution M, the performance of MRS is also affected by the hyperparameter K. The reason behind this phenomenon is that as K becomes larger, there are more bins on $g_j(\tau)$ and the estimated $\mathbf{g}_j = [g_j^k]_{k=1}^K$ is overfitted. To demonstrate this, experiments were performed with the highest resolution $M = 10$, and $K = 10, 40$ and 100. The estimated $\mathbf{g}_1$ when $K = 10, 40$ and 100 is shown in Fig. 6.8 (only the positive half is shown because it is an even function). It is clear that the result when $K = 100$ is overfitted, since there are many spikes up and down.

The more bins used, the larger the estimation variance of $g_j(\tau)$, and consequently the second source difference is even larger than the first source, leading to misidentified segmentation. To show this, the segmentation and NMSE results when $K = 10, 20, 30$ and 40 with $R = 3$ are shown in Tab. 6.4 and Fig. 6.7. As can be seen, when $K \geq 30$ the segmentation obtained from MRS no longer matches the ground truth.

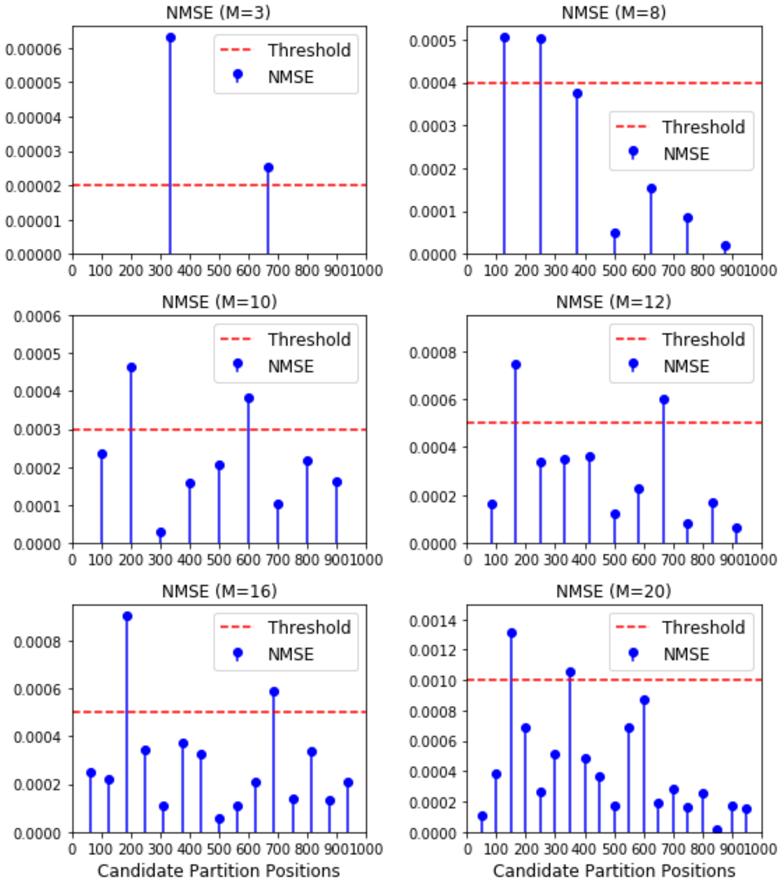


Figure 6.6: For $K = 8$, NMSE of MRS w.r.t. M. The threshold corresponds to $R = 3$.

Table 6.4: Segmentation results w.r.t. K

K	10	20	30	40
Partition Positions	200,600	200,600	100,200	100,200

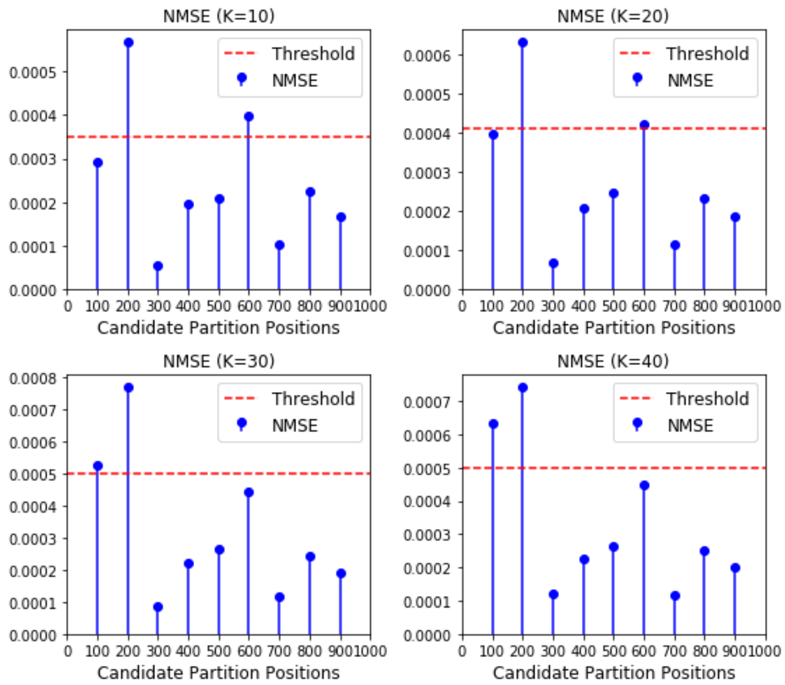


Figure 6.7: For $M = 10$, NMSE of MRS w.r.t. K. The threshold corresponds to $R = 3$.

6.3 Choice of Hyperparameters

To obtain the optimal hyperparameters M and K, intuitively, experiments can be performed on different values of M and K to check which one is better. Nevertheless, for a more robust model, in this section, an idea similar to GP-MISD (see Chapter 3) is used to propose a refined MRS algorithm, namely the Gaussian process based MRS (GP-MRS), so that an optimal value of K need not be chosen. As shown later, a large K can be arbitrarily set, as GP-MRS can prevent it from overfitting, and consequently the new algorithm is more robust.

The key idea of GP-MRS is to use a standard Gaussian process regression to smooth the vector $\mathbf{g}_j = [g_j^k]_{k=1}^K$ in each sector. The posterior mean function $\overline{g}_j(\tau)$ obtained from Gaussian process regression is used to replace the directly estimated $\mathbf{g}_j$ in NMSE (Eq. 6.2). By using GP-MRS the difference between adjacent $g_j(\tau)$ induced by the overfitting of $g_j(\tau)$ can be effectively eliminated when K is large (compare Fig. 6.7 and Fig. 6.9).

It is worth noting that the same Gaussian process based idea cannot be applied to optimize the value of M, because too large an M will lead to a sparse problem in each sector where the Gaussian process regression cannot provide a true posterior mean function. To obtain the optimal hyperparameter M, an empirical formula may be used:

$$M \approx \frac{\mathcal{N}}{250L}, \tag{6.3}$$

where L is the number of independent observations.

6.3.1 Synthetic Data Experiment on GP-MRS

The GP-MRS algorithm was applied to the same synthetic dataset used in Sec. 6.2.4. The GP hyperparameters were tuned carefully. The estimated $\overline{g}_1(\tau)$ when $K = 10, 40$ and 100 is shown in Fig. 6.8 (only the positive half is shown). It is clear that the $\overline{g}_j(\tau)$ from GP-MRS is stable, no matter what K is. The segmentation results and NMSE with $R = 3$ are shown in Tab. 6.5 and Fig. 6.9. It can be seen that the segmentation produced by and NMSE of GP-MRS are both stable, no matter the number of bins on $g_j(\tau)$. It can produce the correct segmentation even in cases where MRS does not work.

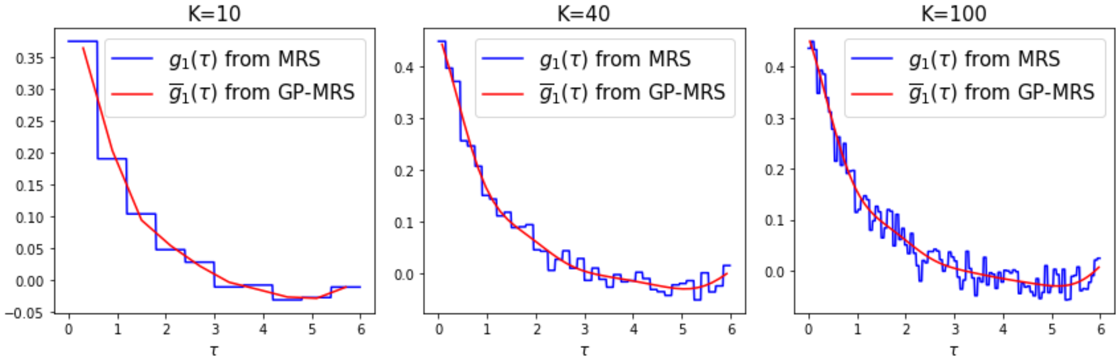


Figure 6.8: For $M = 10$, estimated $g_1(\tau)$ from MRS and GP-MRS when $K = 10, 40$ and 100.

Table 6.5: Segmentation results of GP-MRS w.r.t. K

K	10	20	30
Partition Positions	200,600	200,600	200,600
K	40	100	200
Partition Positions	200,600	200,600	200,600

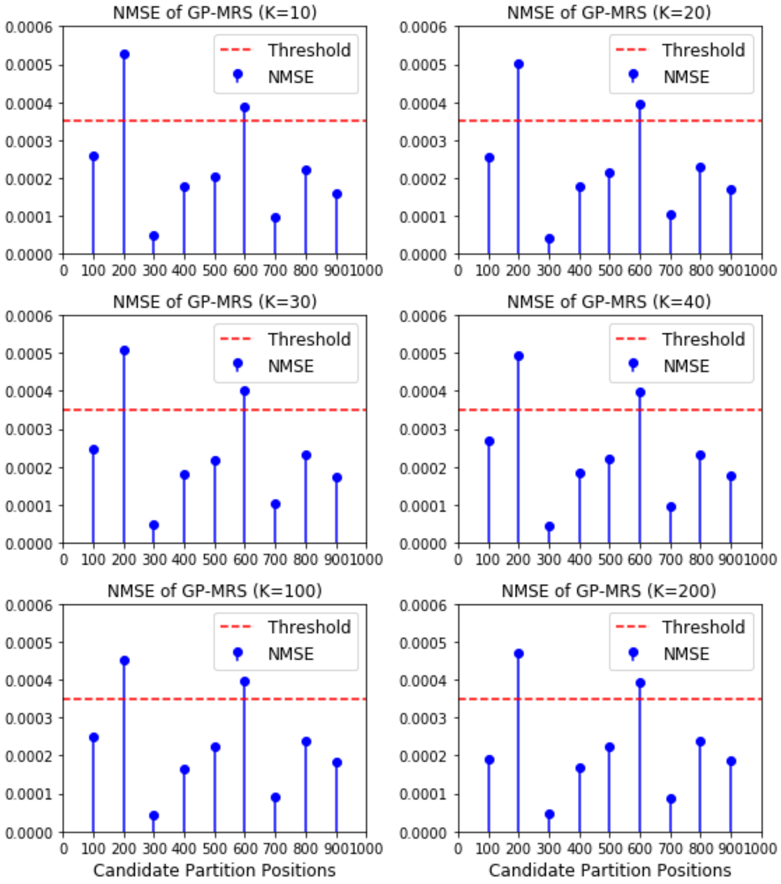


Figure 6.9: For $M = 10$, NMSE of GP-MRS w.r.t. K. The threshold corresponds to $R = 3$.

6.3.2 Computational Complexity of GP-MRS

In a standard GP implementation, the computationa complexity is $\mathcal{O}(K^3)$ when calculating K training points ($\mathbf{g}_j = [g_j^k]_{k=1}^K$). Theoretically, this is due to the need to invert an $K \times K$ covariance matrix. The final complexity of GP-MRS is approximately $\mathcal{O}(\mathcal{N}K + MK^3)$. Unavoidably, the introduction of GP regression into MRS make the algorithm slower. The same experiments as in Sec. 6.2.3 were performed. For $\mathcal{N} = 120,000$ and $M = 10$, the computational time of GP-MRS when $K = 40$ is 48.64 seconds on a normal desktop. This is still acceptable when K is not too large.

6.4 Real Data Experiment

The GP-MRS algorithm was applied to a real world dataset of vehicle collisions to discover the underlying time-varying characteristics. As ground truth information is unavailable for the real world data, two metrics were used to measure the performance: 1) the physical meaning of segmentation to check whether it is consistent with real traffic conditions, and 2) the log-likelihood of test data to check whether the learned result fits the test data better.

6.4.1 Vehicle Collisions in New York City

In daily transportation, the stationarity of Hawkes process is not satisfied for vehicle collision records. For example, at night the traffic condition is not so busy that the triggering effect is lower than in the daytime; or at peak time the traffic condition is very busy so that the triggering effect must be higher than that in the normal time. As can be seen later, the hierarchical time-varying characteristics of triggering kernel and baseline intensity over 24 hours can be discovered using the MRS algorithm.

Weekdays I filter out the collision records in New York City [39] on all weekdays from May 1st 2017 to June 30th 2017. Assuming that the observations every day are independent of each other, there are 45 sets of independent observations. Totally, there are 137,578 observation points. GP-MRS is used for segmentation, which is fast enough in this case. The whole observation period T is set to 1440 minutes (24 hours a day). The support of $\phi(\tau)$ is set to 8 minutes. The hyperparameters of GP-MRS are tuned carefully; K is arbitrarily set to 20 and M is set to 12 by using Eq. 6.3, which means that the size of the sector is 120 minutes (2 hours).

Segmentation searching was performed. Because the period of one day can be considered as a cycle, segmentation can be seen as cutting on a circle. In previous experiments, R corresponds to $R-1$ cutting positions, whereas here R corresponds to R cutting positions. When the desired output resolution is $R = 2$, the computational time of GP-MRS is about 10 seconds and the cutting positions are at 2:00 and 8:00. The segmentation results are shown in Fig. 6.10 left and can be interpreted as busy time and non-busy time.

After segmentation, μ and $\phi(\tau)$ are estimated on each segment. The estimated μ's, where μ is assumed to be nonconstant, are $\mu_1 = 0.317$ and $\mu_2 = 0.127$, and the estimated $\phi(\tau)$'s are shown in Fig. 6.11. Both μ_1 and $\phi_1(\tau)$ are larger than μ_2 and $\phi_2(\tau)$, which is consistent with common sense because the traffic condition is more crowded in the busy time, and consequently the baseline intensity and triggering effect of vehicle collisions are both increased. Additionally, the nonparametrically learned triggering kernel is not strictly monotonically decreasing: there is a small bump around 5 minutes after the initial collision, which is different from the parametric result (shown in Fig. 6.15) and this shows the superior flexibility of nonparametric estimation.

To show the multi-resolution property of GP-MRS, the desired output resolution R is increased to 4 and a finer segmentation on 24 hours is obtained. The computational time of GP-MRS in this case is also about 10 seconds and the cut-

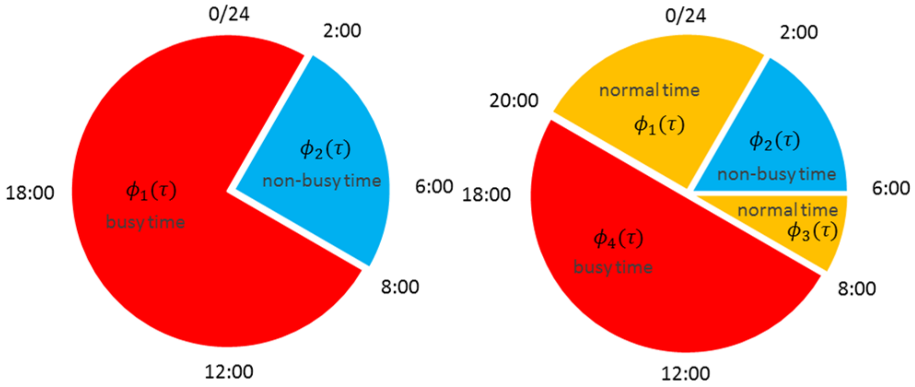


Figure 6.10: Weekdays: 24-hour segmentation result of Vehicle Collisions in New York City, 2 segments on the left and 4 segments on the right.

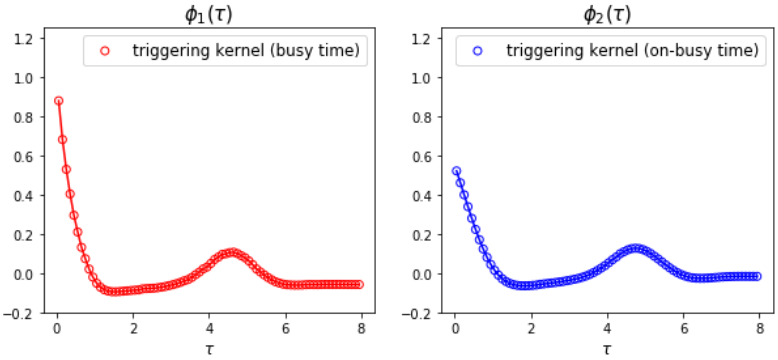


Figure 6.11: Weekdays: Estimated $\phi(\tau)$ of Vehicle Collisions in New York City for busy time and non-busy time.

ting positions are at 2:00, 6:00, 8:00 and 20:00. The segmentation results are shown in Fig. 6.10 right. The segmentations can be interpreted as normal time, busy time and non-busy time. Late night is between 2:00 and 6:00 which are non-busy hours; the after-work entertainment hours (from 20:00 to 2:00) together with morning commute hours (from 6:00 to 8:00) are the normal time; daytime (from 8:00 to 20:00) is the busy time. The estimated μ's are $\mu_1 = 0.32$, $\mu_2 = 0.12$, $\mu_3 = 0.29$ and $\mu_4 = 0.59$. The estimated $\phi(\tau)$'s are shown in Fig. 6.12. Both the baseline intensity and triggering kernel of the busy time are larger than those of normal time and of non-busy time. The baseline intensity and triggering kernel of the two normal times are similar to each other.

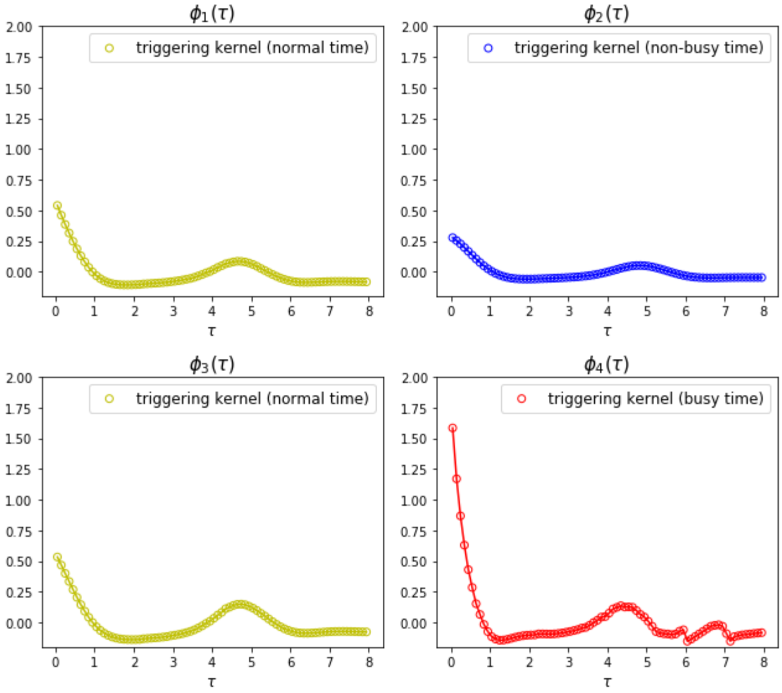


Figure 6.12: Weekdays: Estimated $\phi(\tau)$ of Vehicle Collisions in New York City for normal time, busy time and non-busy time.

Weekends The collision records on all weekends from February 1st 2017 to August 31th 2017 were filtered out. Totally, there are 122,782 observation points. All the settings of the experiment are the same as the weekdays experiment. As before, segmentation search is performed. When the desired output resolution is $R = 2$, the cutting positions are at 2:00 and 8:00 which are the same as for weekdays. The segmentation is shown in Fig. 6.13 left. When R is 3, a finer segmentation is obtained, with the cutting positions at 2:00, 8:00 and 12:00. The segmentation is shown in Fig. 6.13 right. The segmentation can be interpreted as normal time, busy time and non-busy time. The difference between weekends and weekdays is obvious: people wake up late at weekends so the non-busy time is from 2:00 to 8:00; from 8:00 to 12:00, people start to go out of their homes for shopping or other activities so it becomes busy; from 12:00 to 2:00 at weekends, it is normal time which is not as busy as the weekdays.

The estimated $\phi(\tau)$'s are shown in Fig. 6.14 (only the 3-segment result is shown); the estimated μ's are $\mu_1 = 0.33$, $\mu_2 = 0.14$ and $\mu_3 = 0.47$. The baseline intensity and triggering kernel of busy time are larger than those of normal time and also of non-busy time. Also, μ and $\phi(\tau)$ of busy time at weekends are smaller than those of weekdays, which is consistent with common sense because there are fewer vehicles on the road at weekends.

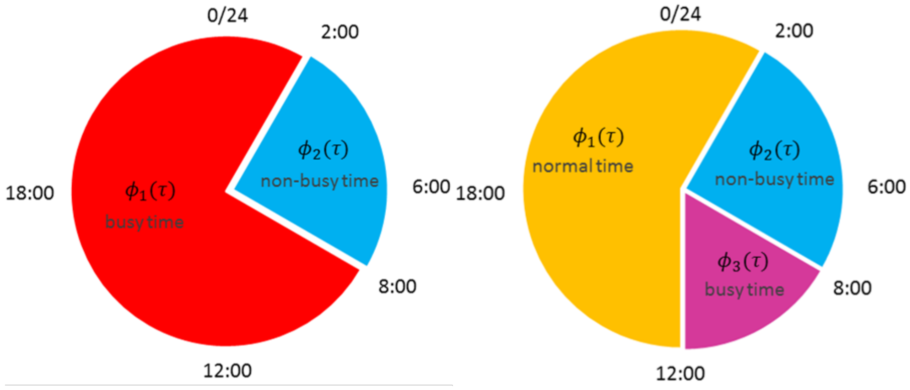


Figure 6.13: Weekends: 24-hour segmentation result of Vehicle Collisions in New York City, 2 segments on the left and 3 segments on the right.

6.4.2 Comparison with Classical Models

To show the superiority of the proposed model, the learned results of stationary parametric Hawkes process (vanilla version), stationary nonparametric Hawkes pro-

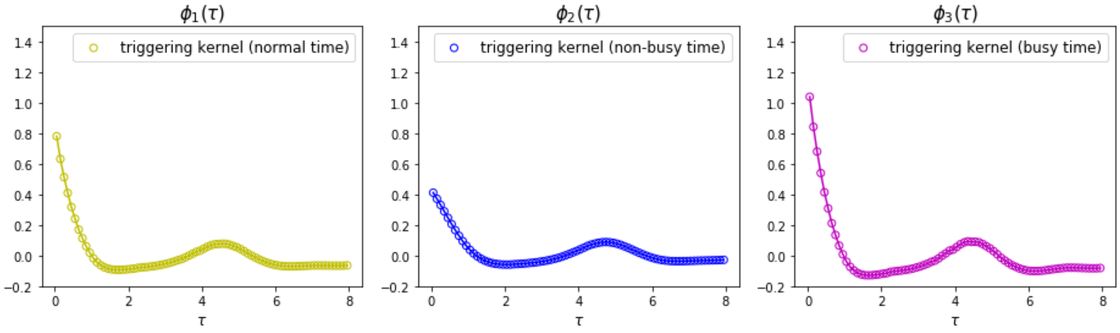


Figure 6.14: Weekends: Estimated $\phi(\tau)$ of Vehicle Collisions in New York City for normal time, busy time and non-busy time.

cess (nonparametric version) and nonstationary nonparametric Hawkes process (proposed version) are compared. The learned baseline intensity and triggering kernel of the first two models are shown below. Also, the negative log-likelihood of the three models on the test dataset are compared. As can be seen, the proposed version fits the test data better.

For the stationary parametric Hawkes process, assume that the baseline intensity μ is constant and the triggering kernel is an exponential decay function: $\phi(\tau) = \alpha \cdot \exp(-\beta\tau)$. The goal of inference is to infer parameters μ, α and β. Inference can be performed by using MLE. The learned $\mu = 0.22$ and $\phi(\tau)$ are shown in Fig. 6.15.

For the stationary nonparametric Hawkes process, assume that the baseline intensity μ is constant and there is only one kind of triggering kernel which is in nonparametric form. The goal of inference is to infer μ and the triggering kernel $\phi(\tau)$. Inference can be performed by using the Wiener-Hopf equation method. The learned $\mu = 0.26$ and $\phi(\tau)$ are shown in Fig. 6.15.

The negative log-likelihood ($-logL$) of the three models on test data (weekdays) are compared. The results are also shown in Fig. 6.15. The proposed model, namely the nonstationary nonparametric version with 4 segments, fits the data best. Followed by the stationary nonparametric version and the stationary parametric version.

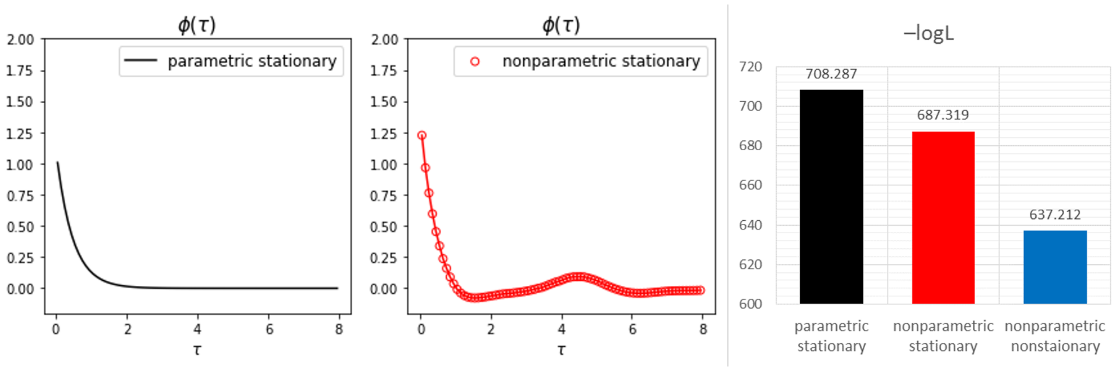


Figure 6.15: Estimated $\phi(\tau)$ of Vehicle Collisions in New York City of (left) stationary parametric Hawkes process and (middle) stationary nonparametric Hawkes process and (right) $-logL$ of stationary parametric version, stationary nonparametric version and nonstationary nonparametric version.

6.5 Discussion

In this section, the setting of hyperparameters and the advantages and disadvantages of the proposed algorithm are discussed. The hyperparameters that have a significant effect on the estimation are K and M. As stated in Sec 6.2.4, an excessively large or small value of K or M hampers correct estimation. In this work, the hyperparameters K and M are chosen based on grid search. A better method to choose hyperparameter values would be useful in the future.

The advantage of the proposed MRS and GP-MRS algorithm is that it can provide a hierarchical insight into the temporal dynamics of the nonstationary Hawkes process, which is more general than the classical stationary Hawkes process, and is scalable to large datasets. The disadvantage is that, based on the current version, the hyperparameters have a large influence on the segmentation result, which can be further improved in the future.

6.6 Remarks

The vanilla Hawkes process assumes stationarity which introduces computation convenience to inference but limits model flexibility. In this chapter, a fast multiresolution segmentation algorithm was proposed to partition the time axis into many segments, and the number of segments depends on the desired output resolution which depends on how fine grained the operator wants the segmentation to be. In this way, the underlying time-varying characteristics of a nonstationary Hawkes process can be discovered. Because second order statistics are utilized, the segmentation algorithm is fast. After segmentation, the Wiener-Hopf method is applied to each segment to estimate the baseline intensity and nonparametric triggering kernel. Overall, the output is a nonstationary and nonparametric Hawkes process. To ease the choice of hyperparameter K, GP-MRS is proposed at the cost of lower computational efficiency, which is still acceptable. The experimental results show the superiority of the proposed model.

Chapter 7

Conclusions and Future Work

In this thesis, a set of novel nonparametric (frequentist or Bayesian) and nonstationary Hawkes process models were developed; concurrently, the corresponding inference algorithms were proposed. The proposed models generalize the classical Hawkes process with fewer constraints in terms of parametric form and stationarity. In this chapter, the main contributions are summarized and potential future research directions are discussed.

In Chapter 1, the background of Hawkes process and the main contributions are introduced. Some preliminary knowledge about the following chapters are provided in Chapter 2. A frequentist nonparametric algorithm: Refined MISD algorithm is proposed in Chapter 3. Two Bayesian nonparametric Hawkes process models are proposed in Chapter 4 and 5 with quadratic link function in Chapter 4 and sigmoid link function in Chapter 5. A fast multi-resolution segmentation algorithm is provided for nonstationary Hawkes process in Chapter 6.

7.1 Thesis Summary and Contributions

In general, the thesis demonstrates that the classical Hawkes process suffers from some constraints when applied to real applications, and the focus in this thesis has been on the parametric and stationary constraints. In attempting to generalize the Hawkes process in these two directions, several advances have been made, which are now summarized.

Refined MISD algorithm: The MISD algorithm is a classical frequentist nonparametric EM algorithm for Hawkes process. One issue with MISD is that the number of bins have to be chosen, making it a hyperparameter affecting the estimated result severely. To avoid tedious model selection e.g. cross validation, a refined MISD algorithm was proposed in Chapter 3. A GP regression step was innovatively embedded into the EM iteration, and consequently the number of bins can be arbitrarily set to a large number without the danger of overfitting. The proposed GP-MISD algorithm is a contribution to the frequentist nonparametric inference for Hawkes process and is also relevant to the fast segmentation algorithm for nonstationary Hawkes process that is discussed in Chapter 6.

Gaussian Process Modulated Hawkes Process (Quadratic Link Function): To model the baseline intensity and triggering kernel of the Hawkes process using nonparametric functions, the Bayesian nonparametric framework with GP prior is a good choice. In order to guarantee the non-negativity of rates, the GP function has to be passed through a link function. The issues with the GP based Bayesian nonparametric model are

1. the coupling of the baseline intensity and triggering kernel in the likelihood of the Hawkes process,
2. the intractable integrals in the numerator and denominator of Bayes rule, and
3. the likelihood being non-conjugate to the GP prior and the posterior being non-Gaussian, due to the existence of link function.

In Chapter 4, the link function was chosen to be a quadratic transformation. The branching structure of the Hawkes process was augmented to decouple the baseline intensity and triggering kernel. Although the posterior is still non-Gaussian, variational Gaussian approximation was utilized so that a Gaussian distribution could approximate the true posterior. As a result, the variational Gaussian approximation was embedded into an EM framework to propose the EMV algorithm. The advantage of the EMV algorithm is that the ELBO has an analytical solution. Moreover, the dimension of the searching space was reduced and the closed-form matrix derivative derived to accelerate the inference.

Gaussian Process Modulated Hawkes Process (Sigmoid Link Function): To further circumvent the non-conjugacy problem, in Chapter 5 the link function was chosen to be the scaled sigmoid function. After augmentation of the branching structure, Pólya-Gamma random variables and latent marked Poisson processes, the likelihood of Hawkes process was converted to two independent factors that are conjugate to the GP priors. Due to the conjugacy, inference can be performed more efficiently.

Depending on the augmented likelihood or joint distribution, a Gibbs sampling algorithm was developed to perform MCMC inference, an EM algorithm to obtain the MAP estimate and a mean-field variational algorithm to perform variational inference. To circumvent the infinite dimensional functional issue, sparse GP approximation was also introduced, consequently all the algorithms are linearly scalable. Experimental results show that the proposed model and inference are superior to other state-of-the-art.

Fast Multi-resolution Segmentation for Nonstationary Hawkes Process: In Chapter 6, the Hawkes process was generalized for nonstationarity. There exist various methods to deal with nonstationary stochastic processes. In this thesis, the focus is mainly on the segmentation method: divide the process into small sectors, estimate the desired statistics and compute the proper partition positions. The issue with the naïve segmentation method is that it is time consuming if based on the estimation of the baseline intensity and triggering kernel. Instead, in Chapter 6 it is proposed to replace it with the estimation of cumulants which is faster to implement. Adjusting the desired resolution, the segmentation algorithm will output partitions at different resolutions. This is the reason why it is named multi-resolution segmentation. Furthermore, the performance of the segmentation algorithm heavily depends on the choice of hyperparameters in experiments. To facilitate the choice of hyperparameters, an approach similar to that in Chapter 3 is adopted to propose a GP-MRS algorithm. The experiments on the vehicle collision dataset demonstrate that the proposed segmentation algorithm can reveal the underlying time-varying characteristics at different resolutions efficiently.

7.2 Limitations and Future Work

In this thesis, the focus is mainly on the 1-variate and 1-dimension Hawkes process. However, the relative models and inference algorithms can be extended to multi-variate and multi-dimensional Hawkes process in future work, e.g. the more general spatial-temporal process model where the triggering kernel is defined on a multi-dimensional space.

Another limitation is that in this thesis, the triggering kernel had to be positive to guarantee the non-negativity of intensity function. Therefore, a new meaningful generalization of the Hawkes process is the nonlinear Hawkes process:

$$\lambda(t) = \zeta\left(\mu(t) + \sum_{t_i<t} \phi(t-t_i)\right) \tag{7.1}$$

where $\zeta(\cdot)$ is a nonlinear function that guarantees the non-negativity of intensity. In Chapters 4 and 5, the intrinsic reason that the GP function had to be passed through a link function was the linearity of Hawkes process. With the generalization to nonlinearity, the triggering kernel can be modelled with negative values. This is especially meaningful in some application areas, e.g. neuroscience, because the excitory and inhibitive prompting [56] between different neurons needs to be incorporated into the model simultaneously. The relative inference algorithms, e.g. the methodology for sigmoid GP Hawkes process, can hopefully be applied to the nonlinear Hawkes process model in future work.

A third limitation is the efficiency. Although efficiency is one of the key points that are considered in this thesis. For example, the conjugacy is incorporated into the model in Chapter 5 to make inference faster than that in Chapter 4. However, all our methods are offline methods which means they cannot be applied to very large dataset. In future work, some online learning algorithms can be proposed for the nonparametric or nonstationary Hawkes process.

7.3 Concluding Remarks

This thesis attempts to generalize the classical Hawkes process in two tracks: nonparametric and nonstationary. Frequentist and Bayesian approaches are considered respectively. The presentation of this thesis is in a progressive structure providing a step-by-step understanding of generalized Hawkes process. This thesis establishes an important foundation for researchers in Hawkes process area but there are still some interesting limitations which could be lucubrated in future work.

Bibliography

[1] David Marsan and Olivier Lengline, "Extending earthquakes' reach through cascading," *Science*, vol. 319, no. 5866, pp. 1076–1079, 2008.

[2] Patrick Hewlett, "Clustering of order arrivals, price impact and trade path optimisation," in *Workshop on Financial Modeling with Jump processes, Ecole Polytechnique*, 2006, pp. 6–8.

[3] Alan G Hawkes, "Spectra of some self-exciting and mutually exciting point processes," *Biometrika*, vol. 58, no. 1, pp. 83–90, 1971.

[4] Yanchi Liu, Tan Yan, and Haifeng Chen, "Exploiting graph regularized multi-dimensional Hawkes processes for modeling events with spatio-temporal characteristics.," in *IJCAI*, 2018, pp. 2475–2482.

[5] Amrita Gupta, Mehrdad Farajtabar, Bistra Dilkina, and Hongyuan Zha, "Discrete interventions in Hawkes processes with applications in invasive species management.," in *IJCAI*, 2018, pp. 3385–3392.

[6] Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel Gomez-Rodriguez, and Le Song, "Recurrent marked temporal point processes: embedding event history to vector," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 1555–1564.

[7] Julio Cesar Louzada Pinto, Tijani Chahed, and Eitan Altman, "Trend detection in social networks using Hawkes processes," in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. ACM, 2015, pp. 1441–1448.

[8] Erik Lewis and George Mohler, "A nonparametric EM algorithm for multiscale Hawkes processes," *Journal of Nonparametric Statistics*, vol. 1, no. 1, pp. 1–20, 2011.

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

[10] Ke Zhou, Hongyuan Zha, and Le Song, "Learning triggering kernels for multi-dimensional Hawkes processes," in *International Conference on Machine Learning*, 2013, pp. 1301–1309.

[11] Emmanuel Bacry and Jean-François Muzy, "First-and second-order statistics characterization of Hawkes processes and non-parametric estimation," *IEEE Transactions on Information Theory*, vol. 62, no. 4, pp. 2184–2202, 2016.

[12] Michael Eichler, Rainer Dahlhaus, and Johannes Dueck, "Graphical modeling for multivariate Hawkes processes with nonparametric link functions," *Journal of Time Series Analysis*, vol. 38, no. 2, pp. 225–242, 2017.

[13] Patricia Reynaud-Bouret, Sophie Schbath, et al., "Adaptive estimation for Hawkes processes; application to genome analysis," *The Annals of Statistics*, vol. 38, no. 5, pp. 2781–2822, 2010.

[14] Jesper Møller, Anne Randi Syversveen, and Rasmus Plenge Waagepetersen, "Log Gaussian Cox processes," *Scandinavian journal of statistics*, vol. 25, no. 3, pp. 451–482, 1998.

[15] Yves-Laurent Kom Samo and Stephen Roberts, "Scalable nonparametric Bayesian inference on point processes with Gaussian processes," in *International Conference on Machine Learning*, 2015, pp. 2227–2236.

[16] Ryan Prescott Adams, Iain Murray, and David JC MacKay, "Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities," in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 9–16.

[17] Chris Lloyd, Tom Gunter, Michael Osborne, and Stephen Roberts, "Variational inference for Gaussian process modulated Poisson processes," in *International Conference on Machine Learning*, 2015, pp. 1814–1822.

[18] Judith Rousseau, Sophie Donnet, and Vincent Rivoirard, "Nonparametric Bayesian estimation of multivariate Hawkes processes," *arXiv preprint arXiv:1802.05975*, 2018.

[19] Edward G Carlstein, Hans-Georg Müller, and David Siegmund, "Change-point problems," IMS, 1994.

[20] Daryl J Daley and David Vere-Jones, "An introduction to the theory of point processes. vol. i. probability and its applications," 2003.

[21] Jakob Gulddahl Rasmussen, "Bayesian inference for Hawkes processes," *Methodology and Computing in Applied Probability*, vol. 15, no. 3, pp. 623–642, 2013.

[22] Stojan Jovanović, John Hertz, and Stefan Rotter, "Cumulants of hawkes point processes," *Physical Review E*, vol. 91, no. 4, pp. 042802, 2015.

[23] Evert J Nyström, "Über die praktische Auflösung von Integralgleichungen mit Anwendungen auf Randwertaufgaben," *Acta Mathematica*, vol. 54, no. 1, pp. 185–204, 1930.

[24] Carl Edward Rasmussen, "Gaussian processes in machine learning," in *Summer School on Machine Learning*. Springer, 2003, pp. 63–71.

[25] Bernd A Berg, *Markov chain Monte Carlo simulations and their statistical analysis: with web-based Fortran code*, World Scientific Publishing Company, 2004.

[26] Charles W Fox and Stephen J Roberts, "A tutorial on variational bayesian inference," *Artificial intelligence review*, vol. 38, no. 2, pp. 85–95, 2012.

[27] WK Hastings, "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, pp. 97–109, 1970.

[28] Stuart Geman and Donald Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on pattern analysis and machine intelligence*, , no. 6, pp. 721–741, 1984.

[29] Manfred Opper and Cédric Archambeau, "The variational Gaussian approximation revisited," *Neural computation*, vol. 21, no. 3, pp. 786–792, 2009.

[30] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul, "An introduction to variational methods for graphical models," *Machine learning*, vol. 37, no. 2, pp. 183–233, 1999.

[31] Radford M Neal et al., "MCMC using Hamiltonian dynamics," *Handbook of markov chain monte carlo*, vol. 2, no. 11, pp. 2, 2011.

[32] Iain Murray, Ryan Prescott Adams, and David JC MacKay, "Elliptical slice sampling," 2010.

[33] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur, "Recurrent neural network based language model," in *Eleventh annual conference of the international speech communication association*, 2010.

[34] Frederic Paik Schoenberg, David R Brillinger, and Peter Guttorp, "Point processes, spatial-temporal," *Wiley StatsRef: Statistics Reference Online*, 2014.

[35] W Thompson, *Point process models with applications to safety and reliability*, Springer Science & Business Media, 2012.

[36] Jonathan Weinberg, Lawrence D Brown, and Jonathan R Stroud, "Bayesian forecasting of an inhomogeneous Poisson process with applications to call center data," *Journal of the American Statistical Association*, vol. 102, no. 480, pp. 1185–1198, 2007.

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

[38] Yosihiko Ogata, "Space-time point-process models for earthquake occurrences," *Annals of the Institute of Statistical Mathematics*, vol. 50, no. 2, pp. 379–402, 1998.

[39] NYC Open Data, "Motor Vehicle Collisions-Crashes," https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95.

[40] NYC Open Data, "NYPD Complaint Data Historic," https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i.

[41] Manuel Gomez Rodriguez, David Balduzzi, and Bernhard Schölkopf, "Uncovering the temporal dynamics of diffusion networks," *arXiv preprint arXiv:1105.0697*, 2011.

[42] George O Mohler, Martin B Short, P Jeffrey Brantingham, Frederic Paik Schoenberg, and George E Tita, "Self-exciting point process modeling of crime," *Journal of the American Statistical Association*, vol. 106, no. 493, pp. 100–108, 2011.

[43] Paul Embrechts, Thomas Liniger, and Lu Lin, “Multivariate Hawkes processes: an application to financial data,” *Journal of Applied Probability*, vol. 48, no. A, pp. 367–378, 2011.

[44] Seth Flaxman, Yee Whye Teh, Dino Sejdinovic, et al., “Poisson intensity estimation with reproducing kernels,” *Electronic Journal of Statistics*, vol. 11, no. 2, pp. 5081–5104, 2017.

[45] Michalis Titsias, “Variational learning of inducing variables in sparse Gaussian processes,” in *Artificial Intelligence and Statistics*, 2009, pp. 567–574.

[46] John P Cunningham, Krishna V Shenoy, and Maneesh Sahani, “Fast Gaussian process methods for point process intensity estimation,” in *Proceedings of the 25th international conference on Machine learning.* ACM, 2008, pp. 192–199.

[47] F Papangelou, “Integrability of expected increments of point processes and a related random change of scale,” *Transactions of the American Mathematical Society*, vol. 165, pp. 483–506, 1972.

[48] NYC Open Data, “2016 Green Taxi Trip Data,” https://data.cityofnewyork.us/Transportation/2016-Green-Taxi-Trip-Data/hvrh-b6nb.

[49] Nicholas G Polson, James G Scott, and Jesse Windle, “Bayesian inference for logistic models using Pólya-Gamma latent variables,” *Journal of the American statistical Association*, vol. 108, no. 504, pp. 1339–1349, 2013.

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

[51] John Frank Charles Kingman, “Poisson processes,” *Encyclopedia of biostatistics*, vol. 6, 2005.

[52] Vancouver Open Data Catalogue, “Crime in Vancouver,” https://www.kaggle.com/wosaku/crime-in-vancouver.

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

[54] Benjamin Noble and George Weiss, "Methods based on the Wiener-Hopf technique for the solution of partial differential equations," *Physics Today*, vol. 12, pp. 50, 1959.

[55] Kendall Atkinson, "A survey of numerical methods for the solution of Fredholm integral equations of the second kind," 1976.

[56] Shizhe Chen, Ali Shojaie, Eric Shea-Brown, and Daniela Witten, "The multivariate Hawkes process in high dimensions: Beyond mutual excitation," *arXiv preprint arXiv:1707.04928*, 2017.