Benefits from Superposed Hawkes Processes

Hongteng Xu[1,2] **Dixin Luo**[1] **Xu Chen**[3] **Lawrence Carin**[1]
[1]Department of ECE, Duke University [2]InfiniaML, Inc. [3]School of Software, Tsinghua University

Abstract

The superposition of temporal point processes has been studied for many years, although the usefulness of such models for practical applications has not be fully developed. We investigate superposed Hawkes process as an important class of such models, with properties studied in the framework of least squares estimation. The superposition of Hawkes processes is demonstrated to be beneficial for tightening the upper bound of excess risk under certain conditions, and we show the feasibility of the benefit in typical situations. The usefulness of superposed Hawkes processes is verified on synthetic data, and its potential to solve the cold-start problem of recommendation systems is demonstrated on real-world data.

1 Introduction

Given a set of temporal point processes $\{N^m\}_{m=1}^M$, their superposition is a new point process N defined by the sum of counting processes, $i.e.$, $N(t) = \sum_{m=1}^M N^m(t)$, $t \geq 0$. For the superposed point process, its instantiated event sequence is the superposition of the event sequences corresponding to $\{N^m\}_{m=1}^M$. The study of superposed point process has a long history, and many interesting properties have been found (Cox & Smith, 1954; Çinlar & Agnew, 1968; Albin, 1984). However, there exists a marked gap between the study of superposed point processes and practical applications. Existing work mainly focuses on the superposition of simple point processes, $e.g.$, Poisson processes (Çinlar & Agnew, 1968) and renewal processes (Cox & Smith, 1954). These models are oversimplified to describe the mechanism of real-world event sequences, while the properties of the superposition of more complicated point processes are not fully investigated. More essentially, *can we get any benefits from learning superposed point processes? If we can, what would the benefits be in practice?* These are still significant open problems.

Focusing on an important class of point process models, called the Hawkes process (Hawkes, 1971), we give positive answers for the above questions. In particular, we prove that there indeed exist benefits from superposed Hawkes processes in the framework of least-squares learning. Theoretically, for the Hawkes processes with different *exogenous* intensities and shared *endogenous* triggering patterns, we can learn the parameters of the endogenous terms with a tighter bound on excess risk, by superposing the Hawkes processes together. We analyze this superposition-based learning in depth and quantitatively connect its feasibility with the diversity of exogenous intensity. Moreover, we validate the benefits from superposed Hawkes processes on both synthetic and real-world data. We show that learning superposed Hawkes processes is beneficial to solve the cold-start problem of recommendation systems.

2 Related work

2.1 Superposed point processes

As aforementioned, the superposition of temporal point processes has been studied for decades. The early work in (Cox & Smith, 1954) studied the superposition of renewal processes and applied its property to analyze pooling signals in neurophysiology. The work in (Çinlar & Agnew, 1968) analyzed the independence of source processes and the dynamics of the source indicator, for the superposition of Poisson processes and that of renewal processes. This work is further extended to multi-dimensional point processes in (Çinlar, 1968). Recently, the work in (Møller & Berthelsen, 2012) proved that a spatial point process can be transformed to a Poisson process by superposing its observations randomly. From the viewpoint of applications, the superposition of arrival processes is applied to model queue behaviors, which can be learned as a renewal process (Albin, 1984). Additionally, the superposition of arrival processes is used to analyze voice data in (Sriram & Whitt, 1986). More recently, Bayesian-based methods are proposed for classifying source processes from superposed observations (Walsh & Raftery, 2005), and their learning

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84. Copyright 2018 by the author(s).

algorithms can be implemented based on MCMC (Redenbach et al., 2015) or variational inference (Rajala et al., 2016). However, all of these research fruits are based on simple point processes, like Poisson and renewal processes. The superposition of more complicated point processes, *e.g.*, the superposition of (multi-dimensional) Hawkes processes, has not been investigated yet. What is worse, the previous work above always treats superposed point processes as "challenges" in statistical analysis and practical applications. None of them consider potential "benefits" from superposed point processes. Our work fills this gap from the viewpoint of learning Hawkes processes.

2.2 Hawkes processes

Hawkes processes (Hawkes, 1971; Hawkes & Oakes, 1974) are useful tools for modeling and analyzing the mutual-excitation phenomena commonly observed in real-world event sequences, which can be applied to many problems, such as social network analysis (Zhao et al., 2015; Wang et al., 2017) and quantitative finance (Bacry et al., 2012; Hardiman et al., 2013). Many variants of Hawkes processes have been proposed recently, *e.g.*, the mixture of Hawkes processes (Yang & Zha, 2013; Xu & Zha, 2017a), the nonlinear isotonic Hawkes process (Wang et al., 2016) and the time-varying Hawkes process (Xu et al., 2017), which show potential to analyze complicated event sequences. From the perspective of learning methodology, maximum likelihood estimation is one of the most popular approaches for learning Hawkes processes (Lewis & Mohler, 2011; Zhou et al., 2013). Recently, least-squares-based learning methods (Eichler et al., 2017), Wiener-Hopf-based methods (Bacry et al., 2012) and the cumulants-based methods (Achab et al., 2016) are also used to learn and analyze Hawkes processes. However, these methods do not consider the influence of superposition on learning.

3 Superposed Hawkes Processes

3.1 Learning Hawkes processes as linear predictors

Consider D entities with interactions ($e.g.$, D users in a social network). For each entity $d\in\mathcal{D}$, $\mathcal{D}=\{1,...,D\}$, we observe its behaviors at timestamps $\{t_{d,1},t_{d,2},...\}$, which are represented by a counting process $N_d(t)=|\{t_{d,i}|t_{d,i}\leq t,\ i=1,2,...\}|$, $i.e.$, the number of type-d events before and at time t. Here $|\cdot|$ represent the cardinality of a set, and i indicates the index of event in each observed sequence. Accordingly, the sequence of all entities' behaviors, denoted as $\{(t_i,d_i)\}$, can be represented by a D-dimensional counting process $N(t)=[N_d(t)]\in\mathbb{N}^D$. For each $N_d(t)$, the expected instantaneous happening rate of an event is represented by its intensity function:

$$\lambda_d(t)=\frac{\mathbb{E}[dN_d(t)|\mathcal{H}_t]}{dt}, \tag{1}$$

where $\mathcal{H}_t$ contains all historical events happening before or at time t.

The D-dimensional counting process may be modeled by a D-dimensional Hawkes process, and the intensity function has the form:

$$\begin{aligned}\lambda_d(t)&=\mu_d(t)+\sum\nolimits_{d'=1}^{D}\int_0^T\phi_{dd'}(t,\ s)dN_{d'}(s)\\&=\mu_d(t)+\sum\nolimits_{(t_i,d_i)\in\mathcal{H}_t}\phi_{dd_i}(t,\ t_i).\end{aligned} \tag{2}$$

where $\boldsymbol{\mu}(t)=[\mu_d(t)]$ corresponds to the background intensity caused by some exogenous factors. Generally, we can model $\boldsymbol{\mu}(t)$ as a D-dimensional homogeneous or inhomogeneous Poisson process. The term $\sum_{d'=1}^{D}\int_0^T\phi_{dd'}(t,s)dN_{d'}(s)$ represents the accumulation of endogenous intensity caused by history (Farajtabar et al., 2014). The impact function (or link function) $\phi_{dd'}(t,s)$, $t\geq s$, represents the influence of the d'-th entity on the d-th entity when their corresponding behaviors (or events) happen at time s and t, respectively. We often assume that the target Hawkes process is shift-invariant: $\phi_{dd'}(t,s)=\phi_{dd'}(t-s)$. For convenience, we represent a Hawkes process as $HP(\boldsymbol{\mu},\boldsymbol{\Phi})$, where $\boldsymbol{\Phi}(t)=[\phi_{dd'}(t)]$, and its instantiated counting process is $N(t)\sim HP(\boldsymbol{\mu},\boldsymbol{\Phi})$.

We may often model $\boldsymbol{\mu}(t)$ as a D-dimensional vector $\boldsymbol{\mu}=[\mu_d]$, and $\phi_{dd'}(t)$ as $a_{dd'}\kappa(t)$ (Zhou et al., 2013), where $\kappa(t)$ is a predefined decay kernel like exponential kernel, $i.e.$, $\kappa(t)=\exp(-wt)$. This model means that the exogenous intensity is a D-dimensional homogeneous Poisson process, while the impact functions $\boldsymbol{\Phi}(t)$ can be parameterized by an infectivity matrix $\boldsymbol{A}=[a_{dd'}]$. In such a situation, the Hawkes process corresponds to a parametric model with $\boldsymbol{\theta}=[\boldsymbol{\mu};\text{vec}(\boldsymbol{A})]\in\mathbb{R}^{D(1+D)}$, where $\text{vec}(\cdot)$ vectorizes its input. Accordingly, its intensity function can be represented as a linear function of $\boldsymbol{\theta}$:

$$\lambda_d(t)=\boldsymbol{x}_d^\top(t)\boldsymbol{\theta}, \tag{3}$$

where $\boldsymbol{x}_d(t)=[\boldsymbol{e}_d;\text{vec}(\boldsymbol{E}(t))]$. $\boldsymbol{e}_d\in\mathbb{R}^D$, whose elements are zeros except the d-th one, which has value 1, corresponds to μ_d. $\boldsymbol{E}(t)=[e_{dd'}(t)]\in\mathbb{R}^{D\times D}$, where $e_{dd'}(t)=\sum_{(t_i,d_i)\in\mathcal{H}_t,\ d_i=d'}\kappa(t-t_i)$, which corresponds to the accumulated decay kernels caused by historical events.

To learn the parameters of the model, we may minimize the squared loss between the counting processes of instantiated event sequences and the integration of intensity function, as in (Eichler et al., 2017; Wang et al., 2016). Specifically, given M event sequences with I events per each, the squared loss is

$$\begin{aligned}&\mathbb{E}\Big[\Big(N(t)-\int_0^t\lambda(s)ds\Big)^2\Big]\\&=\frac{1}{MI}\sum\nolimits_{m=1}^{M}\sum\nolimits_{i=1}^{I}\Big|N_{d_i^m}^m(t_i^m)-\int_0^{t_i^m}\lambda_{d_i^m}(s)ds\Big|^2\\&=\|\boldsymbol{N}-\boldsymbol{X}\boldsymbol{\theta}\|_2^2.\end{aligned} \tag{4}$$

Here, $\boldsymbol{N} = \frac{1}{\sqrt{MI}}[\boldsymbol{N}^1; ...; \boldsymbol{N}^M] \in \mathbb{R}^{MI}$ represents all observed counting processes, in which each $\boldsymbol{N}^m = [N_{d_1^m}^m(t_1^m); ...; N_{d_I^m}^m(t_I^m)]$ contains the number of events with specific types till observed times stamps. Similarly, $\boldsymbol{X} = \frac{1}{\sqrt{MI}}[\boldsymbol{X}^1; ...; \boldsymbol{X}^M] \in \mathbb{R}^{MI\times D(1+D)}$, and each $\boldsymbol{X}^m = [\int_0^{t_1^m} \boldsymbol{x}_{d_1^m}^\top(s)ds; ...; \int_0^{t_I^m} \boldsymbol{x}_{d_I^m}^\top(s)ds] \in \mathbb{R}^{I\times D(1+D)}$, where the integration is an element-wise operation.

It should be noted that in practice it is unnecessary for each sequence to have the same number of events. However, without loss of generality, we assume each event sequence has I events in the following theoretical content for convenience.

From the viewpoint of machine learning, (4) measures the risk that the observed counting processes are different from their expectations estimated by a linear predictor, where $\boldsymbol{N}$ and $\boldsymbol{X}$ are labels and samples, respectively. An ideal predictor can obtain the real expectation, such that (4) corresponds to the variance of counting processes. From the analysis in (Bacry et al., 2012; Hardiman et al., 2013), the variance $\mathbb{E}[(N(t) - \int_0^t \lambda(s)ds)^2] \sim \mathcal{O}(t^2)$. Therefore, differing from a traditional least squares-based linear predictor, the residual errors between sample-label pair (*i.e.*, $N_{d_i^m}(t_i^m) - (\int_0^{t_i^m} \boldsymbol{x}_{d_i^m}^\top(s)ds)\boldsymbol{\theta}$) at different time stamps do not obey the same Gaussian distribution. To solve this problem, we rescale the labels and samples according to the property of the variance mentioned above, and obtain a weighted squared loss (the risk of proposed linear predictor) as:

$$R_{single}(\boldsymbol{\theta}) = \|\boldsymbol{W}(\boldsymbol{N} - \boldsymbol{X}\boldsymbol{\theta})\|_2^2, \tag{5}$$

where the diagonal matrix $\boldsymbol{W} = \text{diag}(\boldsymbol{W}^1, ..., \boldsymbol{W}^M)$ and each $\boldsymbol{W}^m = \text{diag}(\frac{1}{t_1^m}, ...\frac{1}{t_I^m})$. Here we denote the loss as R_{single} because it corresponds to learning a single Hawkes process from M observations.

In many practical situations, the systems described by Hawkes processes are endogenously stationary, and their fluctuations are caused by the changes of exogenous intensities. For example, the interactions between users (*i.e.*, entities) in social networks can be modeled by Hawkes processes. In practice, the event sequences we observed may share the same endogenous triggering patterns, while having different exogenous intensities, because the infectivity among different users (*i.e.*, the impact functions $\boldsymbol{\Phi}(t)$ or their infectivity matrix $\boldsymbol{A}$) is stationary in a long time range (Zhou et al., 2013), but these sequences of users' behaviors may be driven by different information sources and contents (Farajtabar et al., 2014).

In the case of multiple sources, we require M Hawkes processes to model the M event sequences. The processes have individual $\{\boldsymbol{\mu}^m\}_{m=1}^M$ and shared infectivity $\boldsymbol{A}$. We can still learn these models jointly by solving a least squares problem, in which the parameter $\boldsymbol{\theta}_{multi} = [\boldsymbol{\mu}^1; ...; \boldsymbol{\mu}^M; \text{vec}(\boldsymbol{A})] \in \mathbb{R}^{D(M+D)}$ and the squared loss is

$$R_{multi}(\boldsymbol{\theta}_{multi}) = \|\boldsymbol{W}(\boldsymbol{N} - \boldsymbol{X}_{multi}\boldsymbol{\theta}_{multi})\|_2^2, \tag{6}$$

where $\boldsymbol{X}_{multi} = [\boldsymbol{X}_\mu, \boldsymbol{X}_A] \in \mathbb{R}^{MI\times D(M+D)}$. The last D^2 columns of $\boldsymbol{X}_{multi}$ (*i.e.*, $\boldsymbol{X}_A$) are the same with those of $\boldsymbol{X}$ in (4), while the first MD columns (*i.e.*, $\boldsymbol{X}_\mu$) are sparse, which corresponds to different $\boldsymbol{\mu}^m$. Specifically, for the sample $N_{d_i^m}(t_i^m)$, the $(I(m-1)+i)$-th row of $\boldsymbol{X}_\mu$ are all zeros except the $(D(m-1)+d_i^m)$-th element, with value $\frac{1}{\sqrt{MI}}$.

Equation (6) may be viewed as a special case of multi-task learning of Hawkes processes in (Luo et al., 2015), in which Hawkes processes with different exogenous intensities share the same endogenous triggering patterns. If the exogenous intensities have certain structures, *e.g.*, the $\boldsymbol{\mu}^m$'s are sparse or they are grouped and low-rank, we can further impose some regularizers in (6). However, learning multiple Hawkes processes jointly (*i.e.*, $\min_{\boldsymbol{\theta}_{multi}} R_{multi}$) is harder than learning a single one (*i.e.*, $\min_{\boldsymbol{\theta}_{single}} R_{single}$), which has more parameters and requires more observations. In the following subsection we will show that by superposing the Hawkes processes, we can obtain better learning results with fewer observations, especially for the endogenous impact functions and the corresponding infectivity matrix.

3.2 Benefits from superposition

For independent Hawkes processes with shared impact functions, their superposition has the following property:

Theorem 3.1. *For M independent Hawkes processes with shared impact functions, where $N^m(t) \sim HP(\boldsymbol{\mu}^m(t), \boldsymbol{\Phi})$ and $m = 1, ..., M$, their superposition is still a Hawkes process, i.e., $N(t) = \sum_{m=1}^M N^m(t)$ and $N(t) \sim HP(\sum_{m=1}^M \boldsymbol{\mu}^m(t), \boldsymbol{\Phi})$.*

Proof. Because $N(t) = \sum_{m=1}^M N^m(t)$, for $d \in \mathcal{D}$, its intensity is

$$\begin{aligned}\lambda_d(t) =& \frac{\mathbb{E}[dN_d(t)|\mathcal{H}_t]}{dt} = \sum\nolimits_{m=1}^M \frac{\mathbb{E}[dN_d^m(t)| \cup_{l=1}^M \mathcal{H}_t^l]}{dt}\\ =& \sum\nolimits_{m=1}^M \frac{\mathbb{E}[dN_d^m(t)|\mathcal{H}_t^m]}{dt} = \sum\nolimits_{m=1}^M \lambda_d^m(t).\end{aligned}$$

Here $\mathcal{H}_t = \cup_{m=1}^M \mathcal{H}_t^m$ contains all historical events in the superposed process. For the Hawkes processes with shared impact functions, we have

$$\begin{aligned}\lambda_d(t) =& \sum\nolimits_{m=1}^M \Big(\mu_d^m(t) + \sum\nolimits_{(t_i^m, d_i^m)\in\mathcal{H}_t^m} \phi_{dd_i^m}(t - t_i^m)\Big)\\ =& \sum\nolimits_{m=1}^M \mu_d^m(t) + \sum\nolimits_{(t_i, d_i)\in\mathcal{H}_t} \phi_{dd_i}(t - t_i),\end{aligned}$$

for $d \in \mathcal{D}$. According to the definition of Hawkes process in (2), we have $N(t) \sim HP(\sum_{m=1}^M \boldsymbol{\mu}^m(t), \boldsymbol{\Phi})$. □

This property implies that when we aim to learn the impact functions of the target Hawkes processes, besides learning from independent observations, we can learn from the superposed observations of the Hawkes processes. In particular, given $\{N^m(t)\}_{m=1}^M$, the aforementioned traditional strategy learns multiple Hawkes processes jointly by $\min_{\boldsymbol{\theta}_{multi}} R_{multi}(\boldsymbol{\theta}_{multi})$. The optimal solution $\hat{\boldsymbol{\theta}}_{multi} = [\hat{\boldsymbol{\mu}}^1; ...; \hat{\boldsymbol{\mu}}^M; \text{vec}(\widehat{\boldsymbol{A}})]$. By contrast, our strategy first obtains a superposed Hawkes process $N_t = \sum_m N^m(t)$, and then learns a single Hawkes process by $\min_{\boldsymbol{\theta}_{super}} R_{super}(\boldsymbol{\theta}_{super})$, where

$$\begin{aligned} R_{super}(\boldsymbol{\theta}_{super}) &= \frac{1}{M^2} R_{single}(\boldsymbol{\theta}_{super}) \\ &= \Big\| \frac{1}{M} \boldsymbol{W}(\boldsymbol{N} - \boldsymbol{X}\boldsymbol{\theta}_{super}) \Big\|_2^2. \end{aligned} \tag{7}$$

The scaling constant $\frac{1}{M}$ in the last term ensures that the dynamic range of the superposed counting process $N(t)$ is approximately the same as that of a single counting process $N^m(t)$, $m = 1, ..., M$. Its optimal solution $\hat{\boldsymbol{\theta}}_{super} = [\sum_{m=1}^M \hat{\boldsymbol{\mu}}^m; \text{vec}(\widehat{\boldsymbol{A}})]$. Given $\widehat{\boldsymbol{A}}$, we can further estimate $\boldsymbol{\mu}^m$ by solving M independent least squares problem.

Although our superposition-based strategy cannot learn exogenous intensities simultaneously with endogenous impact functions, it provides benefits for learning impact functions. Specifically, given observed samples we can obtain a tighter bound on the excess risk under a certain condition:

Theorem 3.2. *Suppose that we have M independent and stationary D-dimensional Hawkes processes with shared impact functions, i.e., $\{HP(\boldsymbol{\mu}^m, \boldsymbol{A})\}_{m=1}^M$, where the parameters are bounded as $\|\boldsymbol{\mu}^m\|_2^2 \leq B_\mu$ and $\|vec(\boldsymbol{A})\|_2^2 \leq B_A$. Each of them has an observed event sequence with I events. Then the bound on the excess risk $\mathbb{E}[R_{super}(\hat{\boldsymbol{\theta}}_{super}) - R_{super}(\boldsymbol{\theta}^*_{super})]$ is tighter than that of $\mathbb{E}[R_{multi}(\hat{\boldsymbol{\theta}}_{multi}) - R_{multi}(\boldsymbol{\theta}^*_{multi})]$ when the upper bound of $\|\sum_{m=1}^M \boldsymbol{\mu}^m\|_2^2$, denoted as $B_{\Sigma\mu}$, satisfies*

$$\begin{aligned} B_{\Sigma\mu} \leq & M B_\mu + D(M+D) B_\mu \log\Big(1 + \frac{MI}{D(M+D)}\Big) \\ & - D(1+D) B_\mu \log\Big(1 + \frac{MI}{D(1+D)}\Big). \end{aligned} \tag{8}$$

Here $\boldsymbol{\theta}^*$ represents the ground truth of parameters.

Proof. The heart of the proof is the upper bound on the excess risk of linear predictor derived by (Shamir, 2015). In particular, for the linear predictor $\hat{\boldsymbol{\theta}}$ learned by minimizing the squared loss $R(\boldsymbol{\theta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2$, where $\boldsymbol{\theta} \in \{\boldsymbol{\theta} \in \mathbb{R}^C : \|\boldsymbol{\theta}\|_2^2 \leq B\}$ and the M observations $\boldsymbol{y} = [y_1; ...; y_M]$ satisfy $y_i \in \{y : |y| \leq Y\}$, we have

$$\mathbb{E}[R(\hat{\boldsymbol{\theta}}) - R(\boldsymbol{\theta}^*)] \leq \mathcal{O}\Big(\frac{B + CY^2 \log(1 + \frac{M}{C})}{M}\Big). \tag{9}$$

For the loss functions in (6, 7), the observations $\boldsymbol{N}$ is rescaled by $\frac{1}{M}\boldsymbol{W}$ and $\boldsymbol{W}$, respectively. Additionally, the analysis in (Zhu, 2013) shows that $\lim_{t\rightarrow\infty} \frac{N_d(t)}{t} = \frac{\mu_d}{1 - \|\boldsymbol{\Phi}\|}$ for $d = 1, ..., D$. Because of the stationarity of Hawkes processes, we have $1 - \|\boldsymbol{\Phi}\| \gg 0$, and accordingly, the range of the scaled observations should be in the same order of magnitude with the exogenous intensities. Therefore, the Y^2 in (9) can be replaced by $\mathcal{O}(B_\mu)$ in our work.

According to the analysis above, we apply (9) to the two learning strategies mentioned above and obtain

$$\begin{aligned} &\mathbb{E}[R_{multi}(\hat{\boldsymbol{\theta}}_{multi}) - R_{multi}(\boldsymbol{\theta}^*_{multi})] \\ &\leq \mathcal{O}\Big(\frac{B_A + MB_\mu + D(M+D)B_\mu \log(1 + \frac{MI}{D(M+D)})}{MI}\Big), \\ &\mathbb{E}[R_{super}(\hat{\boldsymbol{\theta}}_{super}) - R_{super}(\boldsymbol{\theta}^*_{super})] \\ &\leq \mathcal{O}\Big(\frac{B_A + B_{\Sigma\mu} + D(1+D)B_\mu \log(1 + \frac{MI}{D(1+D)})}{MI}\Big). \end{aligned}$$

Here, both of these two strategies use MI samples in their loss functions. However, the dimension of $\boldsymbol{\theta}_{multi}$ is $D(M+D)$ for learning multiple Hawkes processes while the dimension of $\boldsymbol{\theta}_{super}$ is just $D(1+D)$ for learning a superposed Hawkes process. $B_A + MB_\mu$ is the upper bound of $\|\boldsymbol{\theta}_{multi}\|_2^2$ and $B_A + B_{\Sigma\mu}$ is the upper bound of $\|\boldsymbol{\theta}_{super}\|_2^2$. Clearly, $\mathbb{E}[R_{super}(\hat{\boldsymbol{\theta}}_{super}) - R_{super}(\boldsymbol{\theta}^*_{super})] \leq \mathbb{E}[R_{multi}(\hat{\boldsymbol{\theta}}_{multi}) - R_{multi}(\boldsymbol{\theta}^*_{multi})]$ when

$$\begin{aligned} & B_{\Sigma\mu} + D(1+D)B_\mu \log(1 + \frac{MI}{D(1+D)}) \\ & \leq MB_\mu + D(M+D)B_\mu \log(1 + \frac{MI}{D(M+D)}). \end{aligned} \tag{10}$$

This completes the proof. □

Theorem 3.2 means that under a certain condition learning a superposed Hawkes process can get better convergence of the loss function with fewer samples compared with learning multiple Hawkes processes, which is meaningful to improve the robustness of learning endogenous impact functions. Note that this benefit from superposition is only available when the condition in (8) is satisfied. Based on Theorem 3.2, we can show that the diversity of the exogenous intensities has a large influence on the feasibility of our superposition-based learning strategy. In particular, we have the following two lemmas.

Lemma 3.3. *For the Hawkes processes with the same exogenous intensity and endogenous impact functions, the superposition-based learning strategy is inefficient, i.e., $\mathbb{E}[R_{super}(\hat{\boldsymbol{\theta}}_{super}) - R_{super}(\boldsymbol{\theta}^*_{super})] \geq \mathbb{E}[R_{single}(\hat{\boldsymbol{\theta}}) - R_{single}(\boldsymbol{\theta}^*)]$.*

Proof. In this case, $\boldsymbol{\mu}^1 = ... = \boldsymbol{\mu}^M = \boldsymbol{\mu}$ and $\|\sum_{m=1}^M \boldsymbol{\mu}^m\|_2^2 = M^2\|\boldsymbol{\mu}\|_2^2 \leq M^2 B_\mu = B_{\Sigma\mu}$. Instead of learning multiple Hawkes processes, we only need to learn a single Hawkes process from multiple independent event sequences or from a superposition of them. Similar to the proof of Theorem 3.2, when we minimize $R_{single}(\boldsymbol{\theta})$, its excess risk is bounded as

$$\begin{aligned}&\mathbb{E}[R_{single}(\hat{\boldsymbol{\theta}}) - R_{single}(\boldsymbol{\theta}^*)]\\ &\leq \mathcal{O}\Big(\frac{B_A + B_\mu + D(1+D)B_\mu \log(1+\frac{MI}{D(1+D)})}{MI}\Big).\end{aligned}$$

Based on the relationship that $B_{\Sigma\mu} = M^2 B_\mu$, when we minimize $R_{super}(\boldsymbol{\theta}_{super})$, its excess risk is bounded as

$$\begin{aligned}&\mathbb{E}[R_{super}(\hat{\boldsymbol{\theta}}_{super}) - R_{super}(\boldsymbol{\theta}^*_{super})]\\ &\leq \mathcal{O}\Big(\frac{B_A + M^2 B_\mu + D(1+D)B_\mu \log(1+\frac{MI}{D(1+D)})}{MI}\Big).\end{aligned}$$

Because $M^2 B_\mu > B_\mu$ for $M > 1$, $\mathbb{E}[R_{super}(\hat{\boldsymbol{\theta}}_{super}) - R_{super}(\boldsymbol{\theta}^*_{super})] \geq \mathbb{E}[R_{single}(\hat{\boldsymbol{\theta}}) - R_{single}(\boldsymbol{\theta}^*)]$. □

This Lemma reflects the reason why the superposition of point processes is treated as a challenge in a lot of previous work. It means that if the event sequences are generated by a single Hawkes process, learning from the superposition of the sequences suffers a higher excess risk. What is worse, the more event sequences are superposed, the higher risk we will have in the learning result. In such a situation, we need to avoid the superposition of Hawkes processes.

Fortunately, it is hard to describe real-world data using a single Hawkes process model. To suppress the risk of model misspecification, we can use multiple Hawkes processes with shared impact functions and different exogenous intensities to describe the data generated by an endogenously-stationary system with various exogenous fluctuations. In this situation, applying the superposition-based learning strategy can be efficient, especially in the following case:

Lemma 3.4. *For the Hawkes processes with complementary exogenous intensities, i.e., $\{HP(\boldsymbol{\mu}^m, \boldsymbol{\Phi})\}_{m=1}^M$ and supp$(\boldsymbol{\mu}^m) \cap$ supp$(\boldsymbol{\mu}^{m'}) = \emptyset$ for all $m \neq m'$, the superposition-based learning strategy always provides us with benefits on efficiency, i.e., $\mathbb{E}[R_{super}(\hat{\boldsymbol{\theta}}_{super}) - R_{super}(\boldsymbol{\theta}^*_{super})] \leq \mathbb{E}[R_{multi}(\hat{\boldsymbol{\theta}}_{multi}) - R_{multi}(\boldsymbol{\theta}^*_{multi})]$.*

Proof. Here, the supp$(\cdot)$ returns to the set of the indices of nonzero elements. Because supp$(\boldsymbol{\mu}^m)\cap$supp$(\boldsymbol{\mu}^{m'}) = \emptyset$ for all $m \neq m'$, we have $\|\sum_{m=1}^M \boldsymbol{\mu}^m\|_2^2 = \sum_{m=1}^M \|\boldsymbol{\mu}^m\|_2^2 \leq MB_\mu = B_{\Sigma\mu}$. Plugging the upper bound into the condition (8), we have

$$\begin{aligned}MB_\mu \leq &MB_\mu + D(M+D)B_\mu \log\Big(1+\frac{MI}{D(M+D)}\Big)\\ &- D(1+D)B_\mu \log\Big(1+\frac{MI}{D(1+D)}\Big).\end{aligned}$$

This inequality always holds because $D(M+D)\log(1+\frac{MI}{D(M+D)}) \geq D(1+D)\log(1+\frac{MI}{D(1+D)})$ for $M > 1$, $D \geq 1$ and $I \geq 1$. □

3.3 The cost of the benefits

It should be noted that when we apply the superposition-based learning strategy, we have to increase the computational complexity to obtain the tighter bound of excess risk. In particular, the matrix $\boldsymbol{X}$ in R_{multi} contains the accumulation of integral decay kernels corresponding to M event sequences with I events per each. The computational complexity for getting the $\boldsymbol{X}$ is $\mathcal{O}(MI^2)$. When we superpose the M event sequences together, we obtain a denser superposed sequence with MI events, so the computational complexity of the matrix $\boldsymbol{X}$ in R_{super} is $\mathcal{O}(M^2I^2)$.

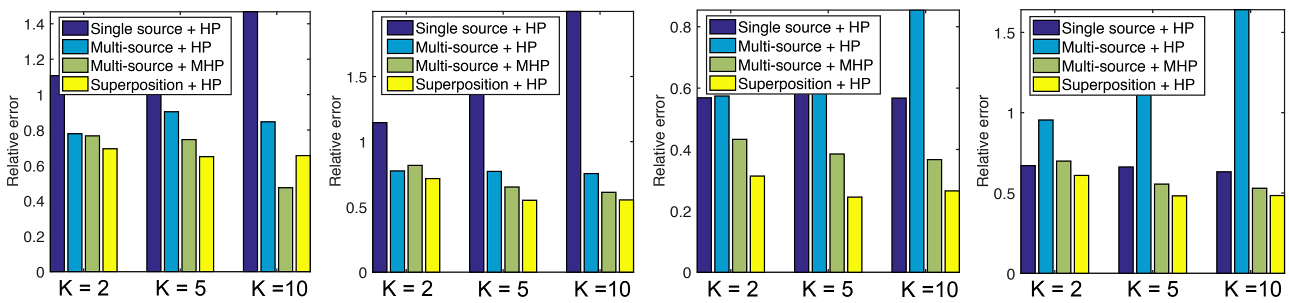
Additionally, after applying the superposition-based learning strategy, we cannot learn the different exogenous intensities simultaneously with the impact functions because they have been accumulated by the superposition operation.

4 Experiments

4.1 Validations based on synthetic data

To verify the benefits from superposed Hawkes processes, we first test our superposition-based learning strategy on a synthetic data set and compare it with its competitors on learning errors of impact functions. The synthetic data set is generated as follows: Given K D-dimensional Hawkes process models, we generate 20 event sequences for each. Each sequence has about 50 events in the time window $[0, 100]$. These Hawkes processes share the same impact functions, which are parameterized as an infectivity matrix $\boldsymbol{A} \in \mathbb{R}^{D\times D}$ and a predefined decay kernel $\exp(-t)$. The matrix $\boldsymbol{A}$ is random with spectral norm $\|\boldsymbol{A}\|_2 = 0.5$. The exogenous intensity $\boldsymbol{\mu}$ of each Hawkes process is a sparse vector, in which only one element is nonzero. The location of the nonzero element is randomly selected and the value is uniformly sampled from the interval $[0, 1]$. Given the parameters, all the event sequences are simulated by the branching process-based method in (Møller & Rasmussen, 2006). The number of models, K, is set to be 2, 5 and 10, respectively. The dimension D is set to be 5 or 10.

This synthetic data set and its simulation process imitate the behaviors in social networks. The users in the network correspond to the dimensions of Hawkes process and their relationships are captured by the infectivity matrix. The information sources in the network are different in different situations. Each event sequence records the behaviors of the users when one of them releases some information. The information releasing behaviors of the source user are modeled by a homogeneous Poisson process. The following behaviors of other users are triggered accordingly, which are



(a) Least squares, $D = 5$ (b) Least squares, $D = 10$ (c) MLE, $D = 5$ (d) MLE, $D = 10$

Figure 1: Estimation errors of impact functions obtained by various methods.

modeled by inhomogeneous Poisson processes. The superposition of all these Poisson processes is a Hawkes process. More details can be found in (Møller & Rasmussen, 2006; Farajtabar et al., 2014).

We can learn the infectivity matrix based on the following four strategies:

1. Single source + HP: Select the event sequences corresponding to a specific kind of Hawkes process and learn a single Hawkes process.

2. Multi-source + HP: Ignore the diversity of the exogenous intensities and learn a single Hawkes process from all event sequences.

3. Multi-source + MHP: Consider the diversity of the exogenous intensities and learn multiple Hawkes processes from all event sequences.

4. Proposed Superposition + HP: Learn a single Hawkes process from the superposition of the event sequences.

The superposition is achieved as follows: 20 superposed event sequences are obtained. Each of them is the superposition of K event sequences and the k-th sequences is one of the 20 sequences corresponding to the k-th Hawkes processes.

Using the least-squares-based learning method mentioned in Section 3.1, we learn the infectivity matrix based on the strategies above, respectively. Figure 1(a) and Figure 1(b) compare these strategies on the relative estimation error of the infectivity matrix ($i.e.$, $\frac{\|\widehat{\boldsymbol{A}}-\boldsymbol{A}^*\|_F}{\|\boldsymbol{A}^*\|_F}$) when $D = 5$ and 10. The results are the average of 10 trials. We can find that the "Single source + HP" is the worst because it only takes advantage of the information of one source. Although it may avoid the problem of model misspecification, using much fewer samples makes it suffer from over-fitting and leads to large estimation errors. The "Multi-source + HP" strategy improves the learning results because it fully uses all event sequences. However, the improvement is limited because it ignores the diversity of the exogenous intensities and only learns a misspecified model. The "Multi-source + MHP" strategy is much better than the previous two strategies because it captures the diversity of the exogenous intensities by introducing more parameters. In the case with $K = 10$ and $D = 5$, it is even better than the proposed "Superposition + HP" strategy. Finally, the proposed "Superposition + HP" strategy achieves the smallest learning errors in most situations, which means that this strategy indeed obtains some benefits from the superposed Hawkes process, as shown in Lemma 3.4.

In our view, the reason for the inferiority of the proposed "Superposition + HP" strategy in the case with $K = 10$ and $D = 5$ is due to the lack of diversity of exogenous intensities. For $k = 1, ..., K$, the exogenous intensity vector of the k-th Hawkes process model has only one nonzero element. When $D < K$, there are some Hawkes process models having similar exogenous intensity vectors – the locations of their nonzero elements are the same with each other. According to Lemma 3.3, the superposition of these similar Hawkes processes does harm to the learning results. In other words, this failure case actually verifies our study results. When we increase the dimension of the model to $D = 10$, we can find in Figure 1(b) that the performance of the proposed "Superposition + HP" strategy is consistently better than its competitors.

Another interesting phenomenon is that when we apply the maximum likelihood estimation (MLE) method (Lewis & Mohler, 2011; Zhou et al., 2013) to learn the Hawkes processes, the superposition-based strategy also outperforms to other strategies on the estimation error of the infectivity matrix. Figure 1(c) and Figure 1(d) visualize the comparisons. We can find that the proposed "Superposition + HP" strategy achieves much better learning results in all cases and the learning results obtained by MLE are better than those obtained by the least squares method. Additionally, different from the cases applying least squares method, when applying MLE, the "Multi-source + HP" is the worst strategy. These observations reveal that 1) the MLE method has better sample complexity than the least squares method; 2) compared to over-fitting, the MLE method is more sensitive to the problem of model misspec-

Table 1: Statistics of our data set.

Category	#Users	#Items	#Ratings
Baby	1240	658	3142
Garden	650	466	1522
Pet	2128	958	5240

ification. The theoretical analysis of the benefits from superposed Hawkes processes in the framework of MLE is an interesting problem, which is left for our future work.

4.2 Applications to the cold-start problem of recommendation systems

A potential application of our work is solving the cold-start problem of recommendation systems. Specifically, the cold-start of a recommendation system means recommending certain items to the users having extremely few recorded behaviors, which is significant for practical recommendation systems. Traditional solutions of the cold-start problem rely on the side information like users' profiles, but they ignore the fact that for the users having few records the side information of them is likely to be limited as well. Therefore, how to solve the cold-start problem without side information is an important but challenging problem.

Recently, Hawkes process-based recommendation systems have been proposed (Du et al., 2015), which inspires us to solve the cold-start problem from the viewpoint of learning Hawkes processes. Generally, for each user her/his buying behavior of an item may trigger her/his following purchases of other items. The infectivity between different items is stationary, which is the endogenous nature of the recommendation system. The preference of each individual user corresponds to the personalized exogenous fluctuation of the system, and her/his purchases can be formulated as an event sequence, which can be captured by a Hawkes process. The event sequences of different users are instances of the Hawkes processes having the same impact functions (and infectivity matrix) but different exogenous intensities. If we can learn the infectivity matrix $\boldsymbol{A}$ of all D items, we can recommend items for each user at time t according to her/his historical buying behaviors $\mathcal{H}_t$ by finding the item with the highest endogenous intensity:

$$d_{next} = \arg\max_{d\in\mathcal{D}} \sum\nolimits_{(t_i,d_i)\in\mathcal{H}_t} a_{dd_i}\exp(-w(t-t_i)).$$

In the cold-start problem, the event sequences are extremely short, so it is hard to learn a reliable infectivity matrix. With the help of the superposition-based learning strategy, we can increase the robustness of learned infectivity matrix and recommend items with higher accuracy.

The training and testing data used in this work are from the Amazon product data set (APD) provided by (He & McAuley, 2016). The APD contains millions of buying-and-rating behaviors to the items grouped into 24 categories. For each categories, millions of user-item pairs spanning from May 1996 to July 2014 and their time stamps are recorded. Focusing on the cold-start problem, we select the users having extremely fewer purchases to recommend items. Specifically, we preprocess the buying-and-rating behaviors of three categories ($i.e.$, "Baby", "Patio, Lawn and Garden", and "Pet Supplies") as follows. For the items having more than 40 rating behaviors, we select their users satisfying three conditions: 1) the number of behaviors of the users spanning from January 2014 to April 2014 is no more than 3; 2) the scores they gave to these items are 4 or 5; 3) they bought and rated at least one item from April 2014 to July 2014. After the preprocessing above, we obtain a subset of the APD to train and test different methods in the cold-start situation. The statistics of our data set is given in Table 1. According to users' buying-and-rating behaviors from January 2014 to April 2014, we aim to predict (recommend) items for them. Because during this period only one or two buying behaviors happened, this is a typical cold-start problem.

To demonstrate the usefulness of our strategy, we compare the "Superposition + HP" strategy with its most powerful competitor "Multi-source + MHP". Additionally, we consider three popular recommendation methods, including recommending the most popular item to all users (MostPopular), the Bayesian personalized ranking (BPR) in (Rendle et al., 2009) and the factorization of personalized Markov chains (FPMC) in (Rendle et al., 2010).

We evaluate the recommendation results achieved by various methods and analyze their performance in the cold-start situation. For each method, we define the generated recommendation list for user m, $m = 1, ..., M$, as $\boldsymbol{r}^m = \{d_1^m, d_2^m, \cdots, d_N^M\}$, where N is the number of recommended items, $d_i^m \in \mathcal{D}$ is ranked at the i-th position in $\boldsymbol{r}^m$. Suppose the real set of the items that user m will buy is $\boldsymbol{t}^m$, we thus use the top-N precision ($P@N$), recall ($R@N$) and F_1-score ($F_1@N$) as the measurements, which are defined as:

$$P@N = \frac{1}{M}\sum_m P_m@N = \frac{1}{M}\sum_m \frac{|\boldsymbol{r}^m \cap \boldsymbol{t}^m|}{|\boldsymbol{r}^m|} \times 100\%$$

$$R@N = \frac{1}{M}\sum_m R_m@N = \frac{1}{M}\sum_m \frac{|\boldsymbol{r}^m \cap \boldsymbol{t}^m|}{|\boldsymbol{t}^m|} \times 100\%$$

$$F_1@N = \frac{1}{M}\sum_m F_{1m}@N = \frac{1}{M}\sum_m \frac{2\cdot P_m@N \cdot R_m@N}{P_m@N + R_m@N}$$

In this experiment, we set $N = 5$, 10 and 20, respectively.

Table 2 summarizes the performance of various methods on the three categories listed in Table 1. For the categories "Garden" and "Pet", applying the proposed "Superposition + HP" startegy improves the performance of recommendation systems in the cold-start situation. The gains obtained

Table 2: Summary of the performance for various methods.

Method		MostPopular			BPR			FPMC			Multi-source+MHP			Superposition+HP		
Metric		$P@N$	$R@N$	$F_1@N$	$P@N$	$R@N$	$F_1@N$	$P@N$	$R@N$	$F_1@N$	$P@N$	$R@N$	$F_1@N$	$P@N$	$R@N$	$F_1@N$
Top5	Baby	0.145	0.726	0.242	0.306	1.532	0.511	**0.484**	**2.419**	**0.806**	0.339	1.694	0.565	0.306	1.532	0.511
	Garden	0.277	1.385	0.462	0.646	3.231	1.077	0.277	1.385	0.462	0.739	3.692	1.231	**1.046**	**5.231**	**1.744**
	Pet	0.517	2.585	0.862	0.526	2.632	0.877	0.517	2.585	0.862	0.780	3.900	1.300	**0.864**	**4.323**	**1.441**
Top10	Baby	0.234	2.339	0.425	**0.379**	**3.790**	**0.689**	0.307	3.065	0.557	0.218	2.177	0.396	0.282	2.822	0.513
	Garden	0.246	2.462	0.448	0.431	4.308	0.783	0.308	3.077	0.559	0.646	6.461	1.174	**0.800**	**8.000**	**1.454**
	Pet	0.371	3.712	0.675	0.428	4.276	0.778	0.470	4.700	0.854	0.549	5.498	1.000	**0.630**	**6.297**	**1.145**
Top20	Baby	0.335	6.694	0.638	0.294	5.887	0.561	**0.339**	**6.774**	**0.645**	0.194	3.871	0.369	0.254	5.081	0.484
	Garden	0.369	7.385	0.703	0.431	8.615	0.821	0.300	6.000	0.571	0.439	8.769	0.835	**0.508**	**10.154**	**0.967**
	Pet	0.374	7.472	0.712	0.465	9.305	0.886	0.371	7.425	0.707	0.338	6.767	0.645	**0.489**	**9.774**	**0.931**

by our method on the three measurements are more than 10% consistently compared with other methods. These results verify the potential of our method to solve the cold-start problem of recommendation systems.

However, for the category "Baby", we can find that our method is inferior to the BPR or the FPMC method. According to our analysis, a possible reason for this phenomenon is that the buying-and-rating behaviors for the items of "Baby" category may not obey the Hawkes process model. Evidence supporting this explanation is that both "Multi-source + MHP" and "Superposition + HP" obtain unsatisfying recommendation results for this category. It should be noted that even if the results provided by the superposition-based learning strategy are not optimal, it still outperforms the "Multi-source + MHP" strategy, which means that the benefits from superposed Hawkes process for learning infectivity matrix are still available.

The synthetic and real-world experiments mentioned above are implemented based on our MATLAB Hawkes process toolbox "THAP" (Xu & Zha, 2017b), which can be found at `https://github.com/HongtengXu/Hawkes-Process-Toolkit`. In particular, for each real-world data set involving about one thousand users and hundreds of items, the runtime of our method is less than 1 minute and without any acceleration.

5 Conclusions and Future Work

We have studied the properties of superposed Hawkes processes and have explored the potential benefits provided by the superposition operation for learning Hawkes process models. We demonstrate that with the help of superposition we can estimate the impact functions (or infectivity matrix) of the target Hawkes processes with lower excess risk in the framework of least squares-based learning. The typical feasible and infeasible conditions are given as well. We verify our theoretical results on synthetic data and show the potential of superposition-based learning strategy to solve the cold-start problem of recommendation systems.

The experimental results in Figure 1(c) and Figure 1(d) imply that the benefits from superposed Hawkes processes are also available in the framework of maximum likelihood estimation. In the future, we plan to analyze the influence of superposition on the maximum likelihood estimation of Hawkes processes in theory. Additionally, studying the superposition of other kinds of point process models is also our future work.

Acknowledgments

This work was supported in part by DARPA, DOE, NIH, NSF and ONR. We also thank reviewers for their insightful and helpful comments.

References

Achab, Massil, Bacry, Emmanuel, Gaïffas, Stéphane, Mastromatteo, Iacopo, and Muzy, Jean-Francois. Uncovering Causality from multivariate Hawkes integrated Cumulants. *arXiv preprint arXiv:1607.06333*, 2016.

Albin, Susan L. Approximating a point process by a renewal process, II: Superposition arrival processes to queues. *Operations Research*, 32(5):1133–1162, 1984.

Bacry, Emmanuel, Dayri, Khalil, and Muzy, Jean-François. Non-parametric kernel estimation for symmetric Hawkes processes: Application to high frequency financial data. *The European Physical Journal B-Condensed Matter and Complex Systems*, 85(5):1–12, 2012.

Çinlar, Erhan. On the superposition of m-dimensional point processes. *Journal of Applied Probability*, 5(1):169–176, 1968.

Çinlar, Erhan and Agnew, RA. On the superposition of point processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 576–581, 1968.

Cox, DR and Smith, Walter L. On the superposition of renewal processes. *Biometrika*, 41(1-2):91–99, 1954.

Du, Nan, Wang, Yichen, He, Niao, Sun, Jimeng, and Song, Le. Time-sensitive recommendation from recurrent user activities. In *NIPS*, 2015.

Eichler, Michael, Dahlhaus, Rainer, and Dueck, Johannes. Graphical modeling for multivariate Hawkes processes with nonparametric link functions. *Journal of Time Series Analysis*, 38(2):225–242, 2017.

Farajtabar, Mehrdad, Du, Nan, Rodriguez, Manuel Gomez, Valera, Isabel, Zha, Hongyuan, and Song, Le. Shaping social activity by incentivizing users. In *NIPS*, 2014.

Hardiman, Stephen, Bercot, Nicolas, and Bouchaud, Jean-Philippe. Critical reflexivity in financial markets: a Hawkes process analysis. *The European Physical Journal B: Condensed Matter and Complex Systems*, 86(10): 1–9, 2013.

Hawkes, Alan G. Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 438–443, 1971.

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

He, Ruining and McAuley, Julian. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *WWW*, 2016.

Lewis, Erik and Mohler, George. A nonparametric EM algorithm for multiscale Hawkes processes. *Journal of Nonparametric Statistics*, 1(1):1–20, 2011.

Luo, Dixin, Xu, Hongteng, Zhen, Yi, Ning, Xia, Zha, Hongyuan, Yang, Xiaokang, and Zhang, Wenjun. Multi-task multi-dimensional Hawkes processes for modeling event sequences. In *IJCAI*, 2015.

Møller, Jesper and Berthelsen, Kasper K. Transforming spatial point processes into poisson processes using random superposition. *Advances in Applied Probability*, 44 (1):42–62, 2012.

Møller, Jesper and Rasmussen, Jakob G. Approximate simulation of Hawkes processes. *Methodology and Computing in Applied Probability*, 8(1):53–64, 2006.

Rajala, Tuomas, Redenbach, Claudia, Särkkä, Aila, and Sormani, Martina. Variational Bayes approach for classification of points in superpositions of point processes. *Spatial Statistics*, 15:85–99, 2016.

Redenbach, Claudia, Särkkä, Aila, and Sormani, Martina. Classification of points in superpositions of strauss and poisson processes. *Spatial Statistics*, 12:81–95, 2015.

Rendle, Steffen, Freudenthaler, Christoph, Gantner, Zeno, and Schmidt-Thieme, Lars. BPR: Bayesian personalized ranking from implicit feedback. In *UAI*, 2009.

Rendle, Steffen, Freudenthaler, Christoph, and Schmidt-Thieme, Lars. Factorizing personalized markov chains for next-basket recommendation. In *WWW*, 2010.

Shamir, Ohad. The sample complexity of learning linear predictors with the squared loss. *Journal of Machine Learning Research*, 16:3475–3486, 2015.

Sriram, Kotikalapudi and Whitt, Ward. Characterizing superposition arrival processes in packet multiplexers for voice and data. *Journal on selected areas in communications*, 4(6):833–846, 1986.

Walsh, Daniel CI and Raftery, Adrian E. Classification of mixtures of spatial point processes via partial bayes factors. *Journal of Computational and Graphical Statistics*, 14(1):139–154, 2005.

Wang, Yichen, Xie, Bo, Du, Nan, and Song, Le. Isotonic Hawkes processes. In *ICML*, 2016.

Wang, Yichen, Ye, Xiaojing, Zhou, Haomin, Zha, Hongyuan, and Song, Le. Linking micro event history to macro prediction in point process models. In *AISTATS*, 2017.

Xu, Hongteng and Zha, Hongyuan. A Dirichlet mixture model of Hawkes processes for event sequence clustering. In *NIPS*, 2017a.

Xu, Hongteng and Zha, Hongyuan. THAP: A Matlab toolkit for learning with Hawkes processes. *arXiv preprint arXiv:1708.09252*, 2017b.

Xu, Hongteng, Luo, Dixin, and Zha, Hongyuan. Learning Hawkes processes from short doubly-censored event sequences. In *ICML*, 2017.

Yang, Shuang-Hong and Zha, Hongyuan. Mixture of mutually exciting processes for viral diffusion. In *International Conference on Machine Learning*, pp. 1–9, 2013.

Zhao, Qingyuan, Erdogdu, Murat A, He, Hera Y, Rajaraman, Anand, and Leskovec, Jure. SEISMIC: A self-exciting point process model for predicting tweet popularity. In *KDD*, 2015.

Zhou, Ke, Zha, Hongyuan, and Song, Le. Learning social infectivity in sparse low-rank networks using multi-dimensional Hawkes processes. In *AISTATS*, 2013.

Zhu, Lingjiong. Ruin probabilities for risk processes with non-stationary arrivals and subexponential claims. *Insurance: Mathematics and Economics*, 53(3):544–550, 2013.