Learning Granger Causality for Hawkes Processes

Hongteng Xu HXU42@GATECH.EDU
School of ECE, Georgia Institute of Technology

Mehrdad Farajtabar MEHRDAD@GATECH.EDU
College of Computing, Georgia Institute of Technology

Hongyuan Zha ZHA@CC.GATECH.EDU
College of Computing, Georgia Institute of Technology

Abstract

Learning Granger causality for general point processes is a very challenging task. In this paper, we propose an effective method, *learning Granger causality*, for a special but significant type of point processes — Hawkes process. We reveal the relationship between Hawkes process's impact function and its Granger causality graph. Specifically, our model represents impact functions using a series of basis functions and recovers the Granger causality graph via group sparsity of the impact functions' coefficients. We propose an effective learning algorithm combining a maximum likelihood estimator (MLE) with a sparse-group-lasso (SGL) regularizer. Additionally, the flexibility of our model allows to incorporate the clustering structure event types into learning framework. We analyze our learning algorithm and propose an adaptive procedure to select basis functions. Experiments on both synthetic and real-world data show that our method can learn the Granger causality graph and the triggering patterns of the Hawkes processes simultaneously.

1. Introduction

In many practical situations, we need to deal with a large amount of irregular and asynchronous sequential data observed in continuous time. The applications include the user viewing records in an IPTV system (when and which TV programs are viewed), and the patient records in hospitals (when and what diagnoses and treatments are given), among many others. All of these data can be viewed as event sequences containing multiple event types and modeled via multi-dimensional point processes. A significant task for a multi-dimensional point process is to learn the so-called Granger causality. From the viewpoint of graphical models, it means to construct a directed graph called Granger causality graph (or local independence graph) (Didelez, 2008) over the dimensions (i.e., the event types) of the process. The arrow connecting two nodes indicates that the event of the dimension corresponding to the destination node is dependent on the historical events of the dimension corresponding to the source node. Learning Granger causality for multi-dimensional point processes is meaningful for many practical applications. Take our previous two examples: the Granger causality among IPTV programs helps us to understand users' viewing preferences and patterns, which is important for personalized program recommendation and IPTV system simulation; the Granger causality among diseases helps us to construct a disease network, which is beneficial to predict potential diseases for patients according to their current diagnoses, leading to more effective treatments.

Unfortunately, learning Granger causality for general multi-dimensional point processes is very challenging. Existing works mainly focus on learning Granger causality for time series (Arnold et al., 2007; Eichler, 2012; Basu et al., 2015), where the Granger causality is captured via the so-called vector auto-regressive (VAR) model (Han & Liu, 2013) based on discrete time-lagged variables. For point processes, on the contrary, the event sequence is in continuous time and no fixed time-lagged observation is available. Therefore, it is hard to find a universal and tractable representation of the complicated historical events to describe Granger causality for the process. A potential solution is to construct features for various dimensions from historical events and learn Granger causality via feature selection (Lian et al., 2015). However, this method is highly de-

Preliminary work.

pendent on the specific feature construction method used, resulting in dubious Granger causality.

To make concrete progress, we focus on a special class of point processes called Hawkes processes and their Granger causality. Hawkes processes are widely used and are capable of describing the self-and mutually-triggering patterns among different event types. Applications include neural signal processing (Song et al., 2013), bioinformatics (Carstensen et al., 2010; Reynaud-Bouret et al., 2010), viral diffusion modeling (Yang & Zha, 2013), social network analysis (Zhao et al., 2015; Zhou et al., 2013a;b), financial analysis (Bacry et al., 2013), information system simulation (Luo et al., 2015), etc. Obviously, learning Granger causality will further extend applications of Hawkes processes in many other fields.

Technically, based on the graphical model of point process (Didelez, 2008), the Granger causality of Hawkes process can be captured by its impact functions. Inspired by this fact, we propose a nonparametric model of Hawkes processes, where the impact functions are represented by a series of basis functions, and we discover the Granger causality via group sparsity of impact functions' coefficients. Based on the explicit representation of Granger causality, we propose a novel learning algorithm combining the maximum likelihood estimator with the sparse-group-lasso (SGL) regularizer on impact functions. The pairwise similarity between various impact functions is considered when the clustering structure of event types is available. Introducing these structural constraints enhances the robustness of our method. The learning algorithm applies the EM-based strategy (Lewis & Mohler, 2011; Zhou et al., 2013a) and obtains close-form solutions to update model's parameters iteratively. Furthermore, we discuss the selection of basis function based on sampling theory, and provide a useful guidance for model selection.

Our method captures Granger causality from complicated event sequences in continuous time. Compared with existing learning methods for Hawkes processes (Zhou et al., 2013b; Eichler et al.), our model avoids discretized representation of impact functions and conditional intensity, and considers the induced structures across impact functions. These improvements not only reduce the complexity of the learning algorithm but also improve learning performance. We investigate the robustness of our method to the changes of parameters and the noise of data and test our method on both synthetic and real-world data. Experimental results show that our learning method can indeed reveal the Granger causality of Hawkes processes and obtain superior learning performance compared with other competitors.

2. Related Work

Granger causality. Many efforts have been made to learn the Granger causality of point processes (Meek, 2014). For general random processes, a kernel independence test is developed in (Chwialkowski & Gretton, 2014). Focusing on 1-dimensional point process with simple piecewise constant conditional intensity, a model for capturing temporal dependencies between event types is proposed in (Gunawardana et al., 2011). In (Basu et al., 2015; Song et al., 2013), the inherent grouping structure is considered when learning the Granger casuality on networks from discrete transition process. More recently, the work in (Daneshmand et al., 2014) proposed a continuous-time diffusion network inference method based on parametric cascade generative process. In more general case, a class of graphical models of marked point processes is proposed in (Didelez, 2008) to capture the local independence over various marks. Specializing the work for Hawkes processes, the work in (Eichler et al.) firstly connects Granger causality with impact functions. However, although applying lasso or its variants to capture the intra-structure of nodes (Ahmed & Xing, 2009) is a common strategy, less work has been done on learning causality graph of Hawkes process with sparse-group-lasso as we do, which leads them to be sensitive to noisy and insufficient data.

Hawkes processes. Hawkes processes (Hawkes, 1971) are proposed to model complicated event sequences where historical events have influences on future ones. Its early application focuses on seismic analysis (Daley & Vere-Jones, 2007). Recently, it is extended to multi-dimensional case and applied to more problems, e.g., financial analysis (Bacry et al., 2012; 2013), social network modeling (Farajtabar et al., 2014; 2015; Zhou et al., 2013a;b; Hall & Willett, 2014; Zhao et al., 2015) and bioinformatics (Reynaud-Bouret et al., 2010; Carstensen et al., 2010). Most of existing works use predefined impact function with known parameters, e.g., the exponential functions in (Farajtabar et al., 2014; Rasmussen, 2013; Zhou et al., 2013a; Hall & Willett, 2014) and the power-law functions in (Zhao et al., 2015). For enhancing the flexibility, a nonparametric model of Hawkes process is first proposed in (Lewis & Mohler, 2011) based on ordinary differential equation (ODE) and extended to multi-dimensional case in (Zhou et al., 2013b; Luo et al., 2015). Similarly, the work in (Bacry et al., 2012) proposes a nonparametric estimation of Hawkes processes via solving the Wiener-Hopf equation. Another nonparametric strategy is the contrast function-based estimation in (Reynaud-Bouret et al., 2010; Hansen et al., 2015). It minimizes the estimation error of conditional intensity function and leads to a Least-Squares (LS) problem (Eichler et al.). More recently, (Du et al., 2012; Lemonnier & Vayatis, 2014) decompose impact functions into basis functions to avoid discretization.

The Gaussian process-based methods (Adams et al., 2009; Lloyd et al., 2015; Lian et al., 2015; Samo & Roberts, 2015) have been reported to successfully estimate more general point processes.

3. Basic Concepts

In this section we briefly introduce the basics of point processes and Granger causality and define our problem formally.

3.1. Point Processes

A temporal point process is a random process whose realization consists of a list of discrete events in time $\{t_i\}$ with $t_i \in [0, T]$. Here $[0, T]$ is the time interval of the process. It can be equivalently represented as a counting process, $N = \{N(t) | t \in [0, T]\}$, where $N(t)$ records the number of events before time t. A multi-dimensional point process with U types of event is represented by U counting processes $\{N_u\}_{u=1}^{U}$ on a probability space $(\Omega, \mathfrak{F}, \mathbb{P})$. Here, $\Omega = [0, T] \times \mathcal{U}$ is the sample space, where $\mathcal{U} = \{1, ..., U\}$ is the set of event types. $\mathfrak{F} = (\mathfrak{F}(t))_{t\in\mathbb{R}}$ is the filtration representing the set of events sequence the process can realize until time t and $\mathbb{P}$ is the probability measure. The equivalent counting process is $N_u = \{N_u(t) | t \in [0, T]\}$, where $N_u(t)$ is the number of type-u events occurring at or before time t, $u = 1, ..., U$.

A very intuitive way to characterize temporal point processes is via the conditional intensity function. Formally, it is defined as the expected instantaneous rate of happening type-u events given the history:

$$\lambda_u(t) = \lambda_u(t|\mathcal{H}_t^{\mathcal{U}}) = \frac{\mathbb{E}[dN_u(t)|\mathfrak{F}(t)]}{dt}.$$

Here $\mathcal{H}_t^{\mathcal{U}} = \{(t_i, u_i) | t_i < t, u_i \in \mathcal{U}\}$ is the historical record collecting events of all types before time t. The functional form of the intensity is often designed to capture the phenomena of interests.

Hawkes Processes. A multi-dimensional Hawkes process is a counting process who has a particular form of intensity:

$$\lambda_u(t) = \mu_u + \sum\nolimits_{u'=1}^{U} \int_0^t \phi_{uu'}(s) dN_{u'}(t-s), \tag{1}$$

where μ_u is the exogenous base intensity independent of the history while $\sum\nolimits_{u'=1}^{U} \int_0^t \phi_{uu'}(s) dN_{u'}(t-s)$ the endogenous intensity capturing the peer influence (Farajtabar et al., 2014). Function $\phi_{uu'}(t) \geq 0$ is called impact function (or triggering kernel), which measures decay in the influence of historical type-u' events on the subsequent type-u events.

3.2. Granger Causality for Point processes

We are interested in identifying, if possible, a subset of the event types $\mathcal{V} \subset \mathcal{U}$ for the type-u event, such that $\lambda_u(t)$ only depends on historical events of types in $\mathcal{V}$, denoted as $\mathcal{H}_t^{\mathcal{V}}$, and not those of the rest types, denoted as $\mathcal{H}_t^{\mathcal{U}\setminus\mathcal{V}}$. From the viewpoint of graphical model, it is about local independence over the dimensions of the point process — the occurrence of events in $\mathcal{V}$ in the past influences the probability of occurrence of type-u events at present and future while the occurrence of events in $\mathcal{U} \setminus \mathcal{V}$ in the past does not. In order to proceed formally we introduce some notations. For a subset $\mathcal{V} \subset \mathcal{U}$ let $N_{\mathcal{V}} = \{N_u(t) | u \in \mathcal{V}\}$. The filtration $\mathfrak{F}_t^{\mathcal{V}}$ is defined as $\mathfrak{F}_t^{\mathcal{V}} = \sigma\{N_u(s) | s \leq t, u \in \mathcal{V}\}$, i.e., the smallest σ-algebra generated by the random processes. In particular, $\mathfrak{F}_t^{u}$ is the internal filtration of an individual counting process $N_u(t)$ while $\mathfrak{F}_t^{-u}$ is the filtration for the subset $\mathcal{U} \setminus \{u\}$.

Definition 3.1. *(Didelez, 2008). The counting process N_u is locally independent of $N_{u'}$ given $N_{\mathcal{U}\setminus\{u,u'\}}$ if the intensity function $\lambda_u(t)$ is measurable with respect to $\mathfrak{F}_t^{-u'}$ for all $t \in [0, T]$. Otherwise N_u is locally dependent of $N_{u'}$.*

Intuitively, the above definition says that $\{N_{u'}(s) | s < t\}$ does not influence $\lambda_u(t)$, given $\{N_l(s) | s < t,\ l \neq u'\}$. In (Eichler et al.), the notion of Granger non-causality is used, and the above definition is equivalent to saying that type-u' event does not Granger-cause type-u event w.r.t. $\mathfrak{F}_t^{\mathcal{U}}$. Otherwise, we say type-u' event Granger-causes type-u event w.r.t. $\mathfrak{F}_t^{\mathcal{U}}$. With this definition, we can construct the so-called Granger causality graph (or called local independence graph) with the event types as the nodes and the directed edges indicating the causation.

Definition 3.2. *The Granger causality graph of the multi-dimensional point process $\{N_u\}_{u=1}^{U}$ is defined as a graph $G = (\mathcal{U}, \mathcal{E})$ over the dimensions (i.e., event types) of the point process, where a direct edge $u' \to u \in \mathcal{E}$ if type-u' event Granger-causes type-u one.*

Learning Granger causality for a general multi-dimensional point process is a difficult problem. In the next section we introduce an efficient method for learning the Granger causality of the Hawkes process.

4. Proposed Model and Learning Algorithm

In this section, we first generalize a known result for Hawkes process. Then, we propose a model of Hawkes process representing impact functions via a series of basis functions. An efficient learning algorithm combining the MLE with the sparse-group-lasso is applied and analyzed in details. Compared with existing learning algorithms, our algorithm is based on convex optimization and has lower complexity, which learns Granger causality robustly.

4.1. Granger causality of Hawkes Process

Following (Eichler et al.) we generalize their result for learning the Granger causality of multi-dimensional Hawkes process.

Theorem 4.1. *Assume a multi-dimensional Hawkes process with conditional intensity function defined in (1) and Granger causality graph $G(\mathcal{U}, \mathcal{E})$. If the condition $dN_{u'}(t-s) > 0$ for $0 \leq s < t \leq T$ holds, then, $u' \to u \notin \mathcal{E}$ if and only if $\phi_{uu'}(t) = 0$ for $t \in [0, T]$.*

Proof. The conditional intensity $\lambda_u(t)$ can be rewritten as

$$\begin{aligned}\lambda_u(t) =& \mu_u + \sum_{v \neq u'}^{U} \int_0^t \phi_{uv}(s) dN_v(t-s)\\ &+ \int_0^t \phi_{uu'}(s) dN_{u'}(t-s).\end{aligned}$$

The first two terms are $\mathfrak{F}_t^{-u'}$-measurable. Therefore, $\lambda_u(t)$ is $\mathfrak{F}_t^{-u'}$-measurable if and only if the last term is identical to zero, which means the impact function $\phi_{uu'}(t) = 0$ for $t \in [0, T]$. According to Definition 3.1 and 3.2, the proof is finished. □

Theorem 4.1 provides us with an explicit representation of the Granger causality of multi-dimensional Hawkes process — learning whether type-u' event Granger-causes type-u event or not is equivalent to detecting whether the impact function $\phi_{uu'}(t)$ is all-zero or not. In other words, the group sparsity of impact functions along the time dimension indicates the Granger causality graph over the dimensions of Hawkes process. Therefore, for multi-dimensional Hawkes process, we can learn its Granger causality via learning its impact functions instead, which requires tractable and flexible representations of impact functions.

4.2. Learning Task

When we parameterize $\phi_{uu'}(t) = a_{uu'}\kappa(t)$ as (Zhou et al., 2013a) does, where $\kappa(t)$ models time-decay of event's influence and $a_{uu'} \geq 0$ captures the influence of u'-type events on u-type ones, the binarized *infectivity matrix* $A = [\text{sign}(a_{uu'})]$ is the adjacency matrix of the corresponding Granger causality graph. Although such a parametric model simplifies the representation of impact function and reduces the complexity of the model, this achievement comes with the cost of inflexibility of the model — the model estimation will be poor if the data does not conform to the assumptions of the model. To address this problem, we propose a nonparametric model of Hawkes processes, representing the impact function in (1) via a linear combination of basis functions as

$$\phi_{uu'}(t) = \sum\nolimits_{m=1}^{M} a_{uu'}^{m} \kappa_m(t). \tag{2}$$

Here $\kappa_m(t)$ is the m-th basis function and $a_{uu'}^m$ is the coefficient corresponding to $\kappa_m(t)$. The selection of bases will be discussed later in the paper.

Suppose we have a set of event sequences $\mathcal{S} = \{s_c\}_{c=1}^{C}$. $s_c = \{(t_i^c, u_i^c)\}_{i=1}^{N_c}$, where t_i^c is the time stamp of the i-th event of s_c and $u_i^c \in \{1, ..., U\}$ is the type of the event. Thus, the log-likelihood of model parameters $\Theta = \{\boldsymbol{A} = [a_{uu'}^m] \in \mathbb{R}^{U \times U \times M}, \boldsymbol{\mu} = [\mu_u] \in \mathbb{R}^U\}$ can be expressed as:

$$\begin{aligned}\mathcal{L}_\Theta =& \sum_{c=1}^{C} \left\{ \sum_{i=1}^{N_c} \log \lambda_{u_i^c}(t_i^c) - \sum_{u=1}^{U} \int_0^{T_c} \lambda_u(s) ds \right\}\\ =& \sum_{c=1}^{C} \left\{ \sum_{i=1}^{N_c} \log \left(\mu_{u_i^c} + \sum_{j=1}^{i-1} \sum_{m=1}^{M} a_{u_i^c u_j^c}^m \kappa_m(\tau_{ij}^c) \right) \right.\\ &\left. - \sum_{u=1}^{U} \left(T_c \mu_u + \sum_{i=1}^{N_c} \sum_{m=1}^{M} a_{uu_i^c}^m K_m(T_c - t_i^c) \right) \right\},\end{aligned} \tag{3}$$

where $\tau_{ij}^c = t_i^c - t_j^c$, $K_m(t) = \int_0^t \kappa_m(s) ds$. For constructing Granger causality accurately and robustly, we consider the following three types of regularizers:

Local Independence. According to Theorem 4.1, the u'-type event has no influence on the u-type one (i.e., directed edge $u' \to u \notin \mathcal{E}$) if and only if $\phi_{uu'}(t) = 0$ for all $t \in \mathbb{R}$, which requires $a_{uu'}^m = 0$ for all m. Therefore, we use group-lasso (Yang et al., 2010; Simon et al., 2013; Song et al., 2013) to regularize the coefficients of impact functions, denoted as $\|\boldsymbol{A}\|_{1,2} = \sum_{u,u'} \|a_{uu'}\|_2$, where $a_{uu'} = [a_{uu'}^1, ..., a_{uu'}^M]^\top$. It means that along the time dimension the coefficients' tensor $\boldsymbol{A}$ should yield to the constraint of group sparsity.

Temporal Sparsity. A necessary condition for the stationarity of Hawkes process is $\int_0^\infty \phi_{ij}(s) ds < \infty$, which means $\lim_{t\to\infty} \phi_{ij}(t) \to 0$. Therefore, we add sparsity constraints to the coefficients of impact functions, denoted as $\|\boldsymbol{A}\|_1 = \sum_{u,u',m} |a_{uu'}^m|$.

Pairwise Similarity. Event types of Hawkes process may exhibit clustering structure. For example, if u and u' are similar event types, their influences on other event types should be similar (i.e., $\phi_{\cdot u}(t)$ are close to $\phi_{\cdot u'}(t)$) and the influences of other event types on them should be similar as well (i.e., $\phi_{u\cdot}(t)$ are close to $\phi_{u'\cdot}(t)$). When the clustering structure is (partially) available, we add constraints of pairwise similarity on the coefficients of corresponding impact functions as follows

$$E(\boldsymbol{A}) = \sum\nolimits_{u=1}^{U} \sum\nolimits_{u' \in \mathcal{C}_u} \|a_{u\cdot} - a_{u'\cdot}\|_F^2 + \|a_{\cdot u'} - a_{\cdot u}\|_F^2.$$

$\mathcal{C}_u$ contains the event types within the cluster that the event of u type resides. $a_{u\cdot} \in \mathbb{R}^{U\times M}$ is the slice of $\boldsymbol{A}$ with row index u, and $a_{\cdot u} \in \mathbb{R}^{U\times M}$ is the slice with column index u. In summary, the learning problem of the Hawkes process is

$$\min_{\Theta\geq\mathbf{0}} -\mathcal{L}_\Theta + \alpha_S\|\boldsymbol{A}\|_1 + \alpha_G\|\boldsymbol{A}\|_{1,2} + \alpha_P E(\boldsymbol{A}). \tag{4}$$

Here α_S, α_G and α_P control the influences of the regularizers. The nonnegative constraint guarantees the model being physically-meaningful.

4.3. An EM-based Algorithm

Following (Lewis & Mohler, 2011; Zhou et al., 2013b), we propose an EM-based learning algorithm for solving optimization problem (4) iteratively. Specifically, given current parameters $\Theta^{(k)}$, we first apply the Jensen's inequality and construct a tight upper-bound of log-likelihood function appeared in (3) as follows:

$$\begin{aligned}
&Q_{\Theta|\Theta^{(k)}} = \\
&\sum_{c=1}^{C}\Bigg\{\sum_{i=1}^{N_c}\Bigg(p_{ii}\log\frac{\mu_{u_i^c}}{p_{ij}} + \sum_{j=1}^{i-1}\sum_{m=1}^{M} p_{ij}^m \log\frac{a_{u_i^c u_j^c}^m \kappa_m(\tau_{ij}^c)}{p_{ij}^m}\Bigg) \\
&-\sum_{u=1}^{U}\Bigg(T_c\mu_u + \sum_{i=1}^{N_c}\sum_{m=1}^{M} a_{uu_i^c}^m K_m(T_c - t_i^c)\Bigg)\Bigg\},
\end{aligned}$$

$p_{ii} = \mu_{u_i^c}^{(k)}/\lambda_{u_i^c}^{(k)}(t_i^c)$ and $p_{ij}^m = a_{u_i^c u_j^c}^{m,(k)}\kappa_m(\tau_{ij}^c)/\lambda_{u_i^c}^{(k)}(t_i^c)$. $\lambda_u^{(k)}(t)$ is the conditional intensity function computed with current parameters. When there is pairwise similarity constraint, we rewrite $E(\boldsymbol{A})$ given current parameters as

$$E_{\Theta|\Theta^{(k)}}(\boldsymbol{A}) = \sum_{u=1}^{U}\sum_{u'\in\mathcal{C}_u}\|a_{u\cdot} - a_{u'\cdot}^{(k)}\|_F^2 + \|a_{\cdot u'} - a_{\cdot u}^{(k)}\|_F^2.$$

Replacing $\mathcal{L}_\Theta$ and $E(\boldsymbol{A})$ with $Q_{\Theta|\Theta^{(k)}}$ and $E_{\Theta|\Theta^{(k)}}(\boldsymbol{A})$ respectively, we decouple parameters and obtain the surrogate objective function $F = -Q_{\Theta|\Theta^{(k)}} + \alpha_S\|\boldsymbol{A}\|_1 + \alpha_G\|\boldsymbol{A}\|_{1,2} + \alpha_P E_{\Theta|\Theta^{(k)}}(\boldsymbol{A})$. Then, we update each individual parameter via solving $\frac{\partial F}{\partial\Theta} = \mathbf{0}$, and obtain the following closed form updates:

$$\mu_u^{(k+1)} = (\sum\nolimits_{c=1}^{C}\sum\nolimits_{u_i^c=u} p_{ii})/(\sum\nolimits_{c=1}^{C} T_c), \tag{5}$$

$$a_{uu'}^{m,(k+1)} = (-B + \sqrt{B^2 - 4AC})/(2A), \tag{6}$$

where

$$\begin{aligned}
A &= \frac{\alpha_G}{\|a_{uu'}^{(k)}\|_2} + 2(|\mathcal{C}_u| + |\mathcal{C}_{u'}|)\alpha_P',\ \ \alpha_P' = \begin{cases}\alpha_P, & u'\in\mathcal{C}_u\\ 0, & \text{others}\end{cases}\\
B &= \sum\nolimits_{c=1}^{C}\sum\nolimits_{u_i^c=u'} K_m(T_c - t_i^c) + \alpha_S\\
&\quad - 2\alpha_P'(\sum\nolimits_{v\in\mathcal{C}_u} a_{vu'}^{m,(k)} + \sum\nolimits_{v'\in\mathcal{C}_{u'}} a_{uv'}^{m,(k)}),\\
C &= -\sum\nolimits_{c=1}^{C}\sum\nolimits_{u_i^c=u}\sum\nolimits_{u_j^c=u'} p_{ij}^m.
\end{aligned}$$

Furthermore, for solving sparse-group-lasso (SGL), we apply the soft-thresholding method in (Simon et al., 2013) to shrink the updated parameters. Specifically, we set $a_{uu'}^{(k+1)}$ to all-zero if the following condition is holds:

$$\|S_{\eta\alpha_S}(a_{uu'}^{(k+1)} - \eta\nabla_{a_{uu'}}Q|_{a_{uu'}^{(k)}})\|_2 \leq \eta\alpha_G, \tag{7}$$

where $S_\alpha(z) = sign(z)(|z| - \alpha)_+$ achieves soft-thresholding for each element of input. $\nabla_x f|_{x_0}$ is the subgradient of function f at x_0 w.r.t. variable x. We have $Q = -Q_{\Theta|\Theta^{(k)}} + \alpha_P E(\boldsymbol{A})$, and η is a small constant. For the $a_{uu'}^{(k+1)}$ unsatisfying (7), we shrink it as

$$\begin{aligned}
a_{uu'}^{(k+1)} &= \left(1 - \frac{\eta\alpha_G}{\|S_{\eta\alpha_S}(a_{uu'}^{(k+1)} - \eta\nabla_{a_{uu'}}Q|_{a_{uu'}^{(k)}})\|_2}\right)_+ \\
&\quad\times S_{\eta\alpha_S}(a_{uu'}^{(k+1)} - \eta\nabla_{a_{uu'}}Q|_{a_{uu'}^{(k)}})
\end{aligned} \tag{8}$$

In summary, Algorithm 1 gives the scheme of our MLE-based algorithm with sparse-group-lasso and pairwise similarity constraints, which is called MLE-SGLP for short. The detailed derivation is given in the appendix.

Algorithm 1 Learning Hawkes Processes (MLE-SGLP)

```
Input: Event sequences $\mathcal{S} = \{s_c\}_{c=1}^{C}$, parameters $\alpha_S$,
  $\alpha_G$, (optional) clustering structure and $\alpha_P$.
Output: Parameters of model, $\boldsymbol{\mu}$ and $\boldsymbol{A}$.
Initialize $\boldsymbol{\mu} = [\mu_u]$ and $\boldsymbol{A} = [a_{uu'}^m]$ randomly.
repeat
  repeat
    Update $\boldsymbol{\mu}$ and $\boldsymbol{A}$ via (5) and (6), respectively.
  until convergence
  for $u, u' = 1 : U$
    if (7) holds, $a_{uu'} = \mathbf{0}$; else, update $a_{uu'}$ via (8).
until convergence
```

4.4. Adaptive Selection of Basis Functions

Although the nonparametric models in (Lemonnier & Vayatis, 2014; Zhou et al., 2013b) represent impact functions as we do via a set of basis functions, they do not provide guidance for the selection process of basis functions. A contribution of our work is proposing a method of selecting basis functions founded on sampling theory (Alan et al., 1989). Specifically, we focus on the impact functions satisfying following assumptions.

Assumption 4.1. *(i) $\phi(t) \geq 0$, and $\int_0^\infty \phi(t)dt < \infty$. (ii) For arbitrary $\epsilon > 0$, there always exists a ω_0, such that $\int_{\omega_0}^\infty |\hat{\phi}(\omega)|d\omega \leq \epsilon$. $\hat{\phi}(\omega)$ is the Fourier transform of $\phi(t)$.*

The assumption (i) guarantees the existence of $\hat{\phi}(\omega)$, while the assumption (ii) means that we can find a function with a bandlimit, denoted as $\frac{\omega_0}{2\pi}$, to approximate the target impact function with bounded residual. Based on these two assumptions, the representation of impact function in (2) can be explained as a sampling process. The $\{a_{uu'}^m\}_{m=1}^M$ can be viewed as the discretized samples of $\phi_{uu'}(t)$ in $[0, T]$ and $\kappa_m(t) = \kappa_\omega(t, t_m)$ is sampling function (i.e., sinc or Gaussian function[1]) corresponding to a low-pass filter with cut-off frequency ω. t_m is the sampling location corresponding to $a_{uu'}^m$ and the sampling rate is $\frac{\omega}{\pi}$. The Nyquist-Shannon theorem requires us to have $\omega = \omega_0$, at least, such that the sampling rate is high enough (i.e., $\frac{\omega_0}{\pi}$, twice bandlimit) to approximate the impact function. Accordingly, the number of samples is $M = \lceil \frac{T\omega_0}{\pi} \rceil$, where $\lceil x \rceil$ returns the smallest integer larger than or equal to x.

Based on the above argument, the core of selecting basis functions is estimating ω_0 for impact functions. It is hard because we cannot observe impact functions directly. Fortunately, based on (1) we know that the bandlimits of impact functions cannot be larger than that of conditional intensity functions $\lambda(t) = \sum_{u=1}^U \lambda_u(t)$. When sufficient training sequences $\mathcal{S} = \{s_c\}_{c=1}^C$ are available, we can estimate $\lambda(t)$ via a Gaussian-based kernel density estimator:

$$\lambda(t) = \sum\nolimits_{c=1}^{C} \sum\nolimits_{i=1}^{N_c} G_h(t - t_i^c). \tag{9}$$

Here $G_h(\cdot)$ is a Gaussian kernel with the bandlimit h. Applying Silverman's rule of thumb (Silverman, 1986), we set optimal $h = (\frac{4\hat{\sigma}^5}{3\sum_c N_c})^{0.2}$, where $\hat{\sigma}$ is the standard deviation of time stamps $\{t_i^c\}$. Therefore, given the upper bound of residual ϵ, we can estimate ω_0 from the Fourier transformation of $\lambda(t)$, which actually does not require us to compute $\lambda(t)$ via (9) directly. In summary, we propose Algorithm 2 to select basis functions and more detailed analysis is given in the appendix.

4.5. Properties of The Proposed Method

Compared with existing state-of-art methods, e.g., the ODE-based algorithm in (Zhou et al., 2013b) and the Least-Squares (LS) algorithm in (Eichler et al.), our algorithm has following advantages.

Algorithm 2 Selecting basis functions

```
Input: Event sequences $\mathcal{S} = s_c{}_{c=1}^{C}$, upper bound of residual $\epsilon$.
Output: Basis functions $\{\kappa_{\omega_0}(t, t_m)\}_{m=1}^M$.
Compute $\left(\sum_{c=1}^C N_c \sqrt{2\pi h^2}\right) e^{-\frac{\omega^2 h^2}{2}}$ to bound $|\hat{\lambda}(\omega)|$.
Find the smallest $\omega_0$ satisfying $\int_{\omega_0}^\infty |\hat{\lambda}(\omega)|d\omega \leq \epsilon$.
The proposed basis functions $\{\kappa_{\omega_0}(t, t_m)\}_{m=1}^M$ are selected, where $\omega_0$ is the cut-off frequency of basis function and $t_m = \frac{(m-1)T}{M}$, $M = \lceil \frac{T\omega_0}{\pi} \rceil$.
```

Computational complexity: Given a training sequence with N events, the ODE-based algorithm in (Zhou et al., 2013b) represents impact functions by M basis functions, where each basis function is discretized to L points. It learns basis functions and coefficients via alternating optimization — coefficients are updated via the MLE given basis functions, and then, the basis functions are updated via solving M Euler-Lagrange equations. The complexity of the ODE-based algorithm is $\mathcal{O}(MN^3U^2 + ML(NU + N^2))$. The LS algorithm in (Eichler et al.) directly discretizes the timeline into L small intervals, ensuring that there is at most one event in each interval. In such a situation, impact functions are also discretized to L points. The computational complexity of the algorithm is $\mathcal{O}(NU^3L^3)$. In contrast, our algorithm is based on known basis functions and does not estimate impact function via discretized points. The computational complexity of our algorithm is $\mathcal{O}(MN^3U^2)$. For getting accurate estimation, the two competitors requires $L \gg N$. Therefore, the computational complexity of the LS algorithm is the highest among the the three, and our complexity is at least comparable to that of the ODE-based algorithm.

Convexity: Both LS algorithm and ours are convex and can achieve global optimum solution. The ODE-based algorithm, however, learns basis functions and coefficients alternatively. It is not convex and is prune to a local optima.

Inference of Granger causality: Neither the ODE-based algorithm nor the LS algorithm considers to infer the Granger causality graph of process when learning model. Without suitable regularizers on impact functions, the impact functions learned by these two algorithms are nonzero generally, which cannot indicate the Granger causality graph exactly. What is worse, the LS algorithm even may obtain physically-meaningless impact functions with negative values. To the best of our knowledge, our algorithm is the first attempt to solving this problem via combining MLE of the Hawkes process with sparse-group-lasso,

[1] For Gaussian filter $\kappa_\omega(t, t_m) = \exp(-(t - t_m)^2/(2\sigma^2))$, its bandlimit is defined as $\omega = \sigma^{-1}$.

which learns the Granger causality graph robustly, especially in the case having few training sequences.

5. Experiments

For demonstrating the feasibility and the efficiency of our algorithm (MLE-SGLP), we test our it on both synthetic and real-world data, and compare it with the state-of-art methods, including the ODE-based method in (Zhou et al., 2013b), the Least-Squares (LS) method in (Eichler et al.). We also investigate the influences of regularizers via comparing our algorithm with its variants, including the pure MLE without any regularizer (MLE), the MLE with group-lasso (MLE-GL), and the MLE with sparse regularizer (MLE-S). For evaluating various algorithms comprehensively, given estimate $\tilde{\Theta} = \{\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{A}}\}$, we apply the following measurements:

1) *Loglike*: The log-likelihood of testing data.

2) $e_\mu = \frac{\|\tilde{\boldsymbol{\mu}}-\boldsymbol{\mu}\|_2}{\|\boldsymbol{\mu}\|_2}$, the relative error of $\boldsymbol{\mu}$.

3) $e_\phi = \frac{1}{U^2}\sum_{u,u'}\frac{\int_0^T|\tilde{\phi}_{uu'}(t)-\phi_{uu'}(t)|dt}{\int_0^T\phi_{uu'}(t)dt}$, the relative error of $\Phi(t) = [\phi_{uu'}(t)]$.

4) *Sparsity of impact function*: Impact functions are visualized and the Granger causality graph is indicated via all-zero impact functions.

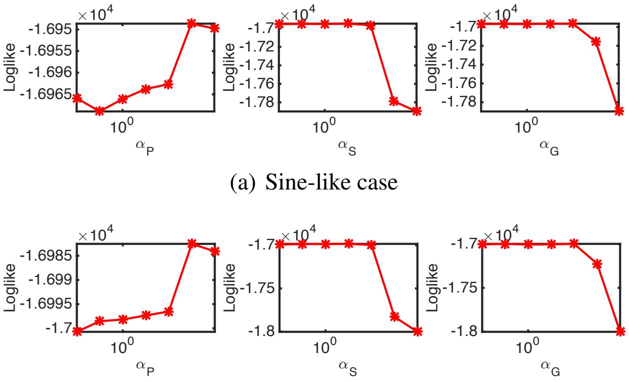
5.1. Synthetic Data

We generate two synthetic data sets using sine-like impact functions and piecewise constant impact function respectively. Each of them contains 500 event sequences with time length $T = 50$ generated via a Hawkes process with $U = 5$. The exogenous base intensity of each event type is uniformly sampled from $[0, \frac{1}{U}]$. The sine-like impact functions are generated as

$$\phi_{uv}(t) = \begin{cases} b_{uv}(1-\cos(\omega_{uv}t-\pi s_{uv})), & t\in[0, \frac{2-s_{uv}}{4\pi\omega_{uv}}],\\ 0, & \text{otherwise,}\end{cases}$$

where $\{b_{uv}, \omega_{uv}, s_{uv}\}$ are set as $\{0.05, 0.6\pi, 1\}$ when $u, v \in \{1, 2, 3\}$, $\{0.05, 0.4\pi, 0\}$ when $u, v \in \{4, 5\}$, $\{0.02, 0.2\pi, 0\}$ when u (or v) $= 4$, v (or u) $\in \{1, 2, 3\}$. The piecewise constant impact functions are the truncated results of above sine-like ones.

We test various learning algorithms on each of the two data sets with 10 trials, respectively. In each trial, $C = \{50, ..., 250\}$ sequences are chosen randomly as training set while the rest 250 sequences are chosen as testing set. In all trials, Gaussian basis functions are used, whose number and bandlimit are decided by Algorithm 2. With the help of cross validation, we test our algorithm with various parameters in a wide range, where $\alpha_P, \alpha_S, \alpha_G \in [10^{-2}, 10^4]$. According to the measure *Loglike*, we set $\alpha_S = 10$, $\alpha_G = 100$, $\alpha_P = 1000$. The curves of *Loglike* w.r.t. the three parameters are shown in Fig. 1. We can find that the learning result is relatively stable when changing the parameters in a wide range.



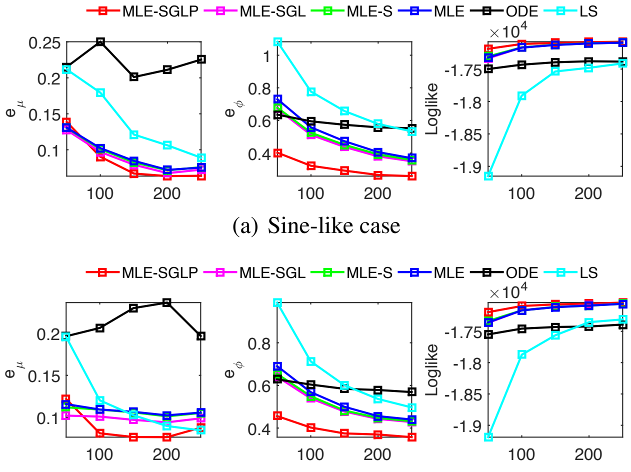
(a) Sine-like case

(b) Piecewise constant case

Figure 1. The curves of *Loglike* w.r.t. the change of α_P, α_G and α_S are shown. In each subfigure, left: $\alpha_G = 100$, $\alpha_S = 10$, $\alpha_P \in [10^{-2}, 10^4]$; middle: $\alpha_G = 100$, $\alpha_P = 1000$, $\alpha_S \in [10^{-2}, 10^4]$; right: $\alpha_P = 1000$, $\alpha_S = 10$, $\alpha_G \in [10^{-2}, 10^4]$. The number of training sequence is 250.

The testing results are shown in Fig. 2. We can find that our learning algorithm performs better than other competitors on both data sets, i.e., higher *Loglike*, lower e_μ and e_ϕ, w.r.t. various C. Especially when having few training sequences, the ODE-based and the LS algorithm need to learn too many parameters from insufficient samples so they are inferior to our MLE-SGLP algorithm and its variants because of the over-fitting problem. Additionally, by increasing the number of training sequences, the performance of the ODE-based algorithm does not improve a lot — the nature of non-convexity may lead the ODE-based algorithm to fall into local optimal. All MLE-based algorithms are superior to the ODE-based algorithm and the LS algorithm, and the regularizers proposed in this paper indeed help to improve learning results of MLE. Specifically, if the clustering structure is available, our MLE-SGLP algorithm will obtain the best results. Otherwise, our MLE-SGL algorithm will be the best.

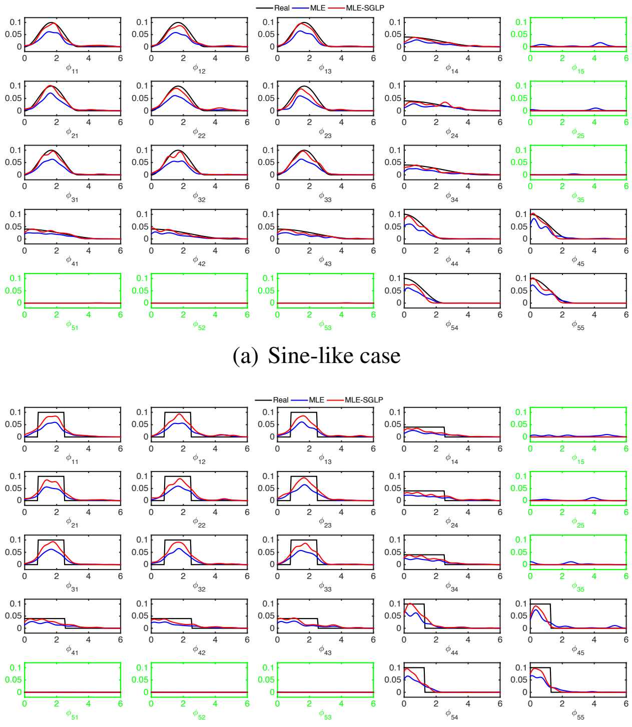
For demonstrating the importance of the sparse-group-lasso regularizer to learning Granger causality graph, Fig. 3 visualizes the estimates of impact functions obtained by various methods. The Granger causality graph of the target Hawkes process is learned by finding those all-zero impact functions (the green subfigures). In both cases, our MLE-SGLP algorithm can obtain right all-zero impact functions while the pure MLE algorithm sometimes fails because of the lack of sparse-related regularizer. It means that introducing sparse-group-lasso into the framework of MLE is



(a) Sine-like case

(b) Piecewise constant case

Figure 2. In each subfigure, the comparisons on e_μ, e_ϕ, and *Loglike* for various methods are shown.



(a) Sine-like case

(b) Piecewise constant case

Figure 3. Contributions of regularizers: comparisons of impact functions obtained via MLE-SGLP and pure MLE using 500 training sequences. The green subfigures contain the all-zero impact functions. The black lines are real impact functions, the blue lines are the estimates from pure MLE and the red ones are proposed estimates from MLE-SGLP.

necessary for learning Granger causality. Note that, even if the basis functions we select do not match well with the real case, i.e., the Gaussian basis functions are not suitable for piecewise constant impact functions, our MLE-SGLP algorithm can still learn the Granger causality graph of the Hawkes process with high accuracy. As Fig. 2(b) shows, although the estimates of non-zero impact functions based on Gaussian basis functions do not fit the ground truth well, the all-zero impact functions are learned exactly via our MLE-SGLP algorithm.

5.2. Real-world Data

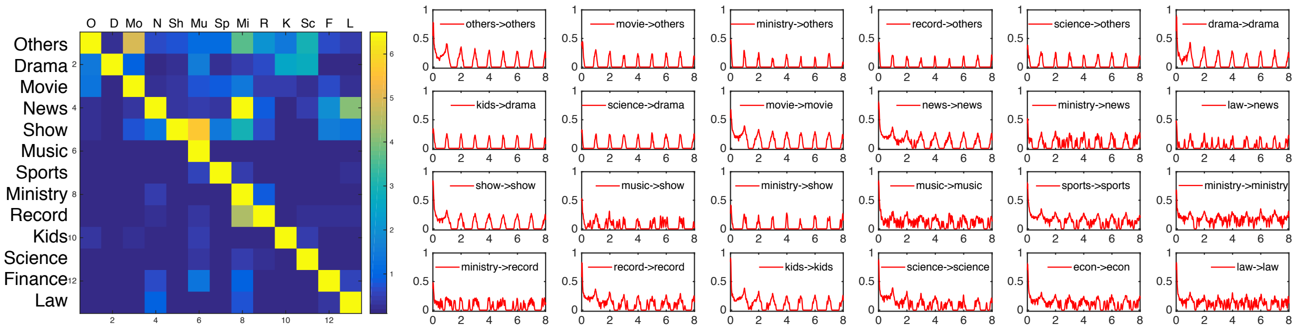
We test our algorithm on the IPTV viewing record data set (Luo et al., 2015). The data set records the viewing behavior of 7100 users, i.e., what and when they watch, in the IPTV system from January to November 2012. $U(=13)$ categories of TV programs are predefined. Similar to (Luo et al., 2015), we model users' viewing behavior via a Hawkes process, in which the TV programs' categories exist self-and mutually-triggering patterns. For example, viewing an episode of a drama would lead to viewing the following episodes (self-triggering) and related news of actors (mutually-triggering). Therefore, the causality among various categories is dependent not only on the predetermined displaying schedule but also on users' viewing preferences.

We capture the Granger causality graph of programs' categories via learning impact functions. In this case, the pairwise sparsity is not applied because the clustering structure is not available. The training data is the viewing behavior in the first 10 months and testing data is the viewing behavior in the last month. Considering the fact that many TV programs are daily or weekly periodic and the time length of most TV programs is about 20-40 minutes, we set the time length of impact function to be 8 days (i.e., the influence of a program will not exist over a week) and the number of samples $M = 576$ (i.e., one sample per 20 minutes). The cut-off frequency of sampling function is $w_0 = \pi M/T$, where T is the number of minutes in 8 days. Table. 1 gives *Loglike* for various methods. Even using a PC with 16GB memory, the LS algorithm runs out-of-memory in this case because it requires to discretize long event sequences with dense samples. Compared with the ODE-based algorithm and pure MLE algorithm, the MLE with regularizers has better *Loglike* and our MLE-SGL algorithm obtains the best result.

Table 1. Loglike ($\times 10^6$) for various methods

ALG.	LS	ODE	MLE	MLE-S	MLE-SGL
Loglike	—	-1.919	-1.876	-1.874	**-1.872**

We define the infectivity of the u'-th TV program category on the u-th one as $\int_0^\infty \phi_{uu'}(s)ds$, which is shown in Fig. 4(a). It can be viewed as an adjacency matrix of the Granger causality graph. Additionally, by ranking the in-



(a) Infectivity matrix

(b) Top 24 impact functions

Figure 4. (a) The infectivity matrix for various TV programs. The element in the u-th row and the u'-th column is $\int_0^\infty \phi_{uu'}(s)ds$. (b) Estimates of nonzero impact functions for the IPTV data. By ranking the infectivity $\int_0^\infty \phi_{uu'}(s)ds$ from high to low, the top 24 impact functions are shown. For visualization, $\phi_{uu'}^{0.25}(t)$ is shown in each subfigure.

fectivity from high to low, the top 24 impact functions are selected and shown in Fig. 4(b). We think our algorithm works well because the following reasonable phenomena are observed in our learning results:

1) All TV program categories have obvious self-triggering patterns because most of TV programs display periodically. Viewers are likely to watch them daily at the same time. Our learning results reflect these phenomena: the main diagonal elements of the infectivity matrix in Fig. 4(a) are much larger than other ones, and the estimates of impact functions in Fig. 4(b) have clear daily-periodic pattern.

2) Some popular categories having a large number of viewers and long displaying time, e.g., “drama”, “movie”, “news” and “talk show”, are likely to be triggered by others, while the other unpopular ones having relative fewer but fixed viewers and short displaying time, e.g., “music”, “kids’ program”, “science”, are mainly triggered by themselves. It is easy to find that the infectivity matrix we learned reflects these patterns — the non-diagonal elements involving those unpopular categories are very small or zero. In Fig. 4(b) the non-zero impact functions mainly involve popular categories. Additionally, because few viewing events about these categories are observed in the training data, the estimates of the impact functions involving unpopular categories are relatively noisy.

In summary, our algorithm performs better on the IPTV data set than other competitors. The learning results are reasonable and interpretable, which prove the rationality and the feasibility of our algorithm to some degree.

6. Conclusion

In this paper, we learn the Granger causality of Hawkes processes according to the relationship between the Granger causality and impact functions. Combining the MLE with the sparse-group-lasso, we propose an effective algorithm to learn the Granger causality graph of the target process and captures its temporal dynamics simultaneously. We demonstrate the robustness and the rationality of our work on both synthetic and real-world data. In the future, we plan to extend our work and analyze the Granger causality of general point processes.

Appendix

Derivation of Surrogate Function

Using the Jensen’s inequality, we have following inequality for all c and i:

$$\log\left(\mu_{u_i^c} + \sum\nolimits_{m=1}^{M} \sum\nolimits_{j=1}^{i-1} a^m_{u_i^c u_j^c}\kappa(\tau^c_{ij})\right)$$
$$\geq p_{ii}\log\left(\frac{\mu_{u_i^c}}{p_{ii}}\right) + \sum_{m=1}^{M}\sum_{j=1}^{i-1} p^m_{ij}\log\left(\frac{a^m_{u_i^c u_j^c}\kappa(\tau^c_{ij})}{p^m_{ij}}\right).$$

The equation holds if and only if $\mu_u = \mu_u^{(k)}$ and $a^m_{uu'} = a^{m,(k)}_{uu'}$. Therefore, we have $Q_{\Theta|\Theta^{(k)}} \geq \mathcal{L}_\Theta$ and $Q_{\Theta^{(k)}|\Theta^{(k)}} = \mathcal{L}_{\Theta^{(k)}}$.

Derivation of Algorithm 1

We have surrogate objective function $F = -Q_{\Theta|\Theta^{(k)}} + \alpha_S\|\boldsymbol{A}\|_1 + \alpha_G\|\boldsymbol{A}\|_{1,2} + \alpha_P E_{\Theta|\Theta^{(k)}}(\boldsymbol{A})$, where $Q = -Q_{\Theta|\Theta^{(k)}} + +\alpha_P E_{\Theta|\Theta^{(k)}}(\boldsymbol{A})$ is the data fidelity term. Similar to (Simon et al., 2013), we choose a group $a_{uu'} = [a^1_{uu'}, ..., a^M_{uu'}]^\top$ to minimize and fix other parameters. Given current estimate $a^{(k)}_{uu'}$, we majorize Q as

$$Q \leq Q|_{a^{(k)}_{uu'}} + (a_{uu'} - a^{(k)}_{uu'})\nabla_{a_{uu'}} Q|_{a^{(k)}_{uu'}} + \frac{1}{2\eta}\|a_{uu'} - a^{(k)}_{uu'}\|_2^2. \tag{10}$$

Introducing (10) to the surrogate objective function, we rewrite the optimization problem as

$$\begin{aligned}\min_{a_{uu'}\geq\mathbf{0}} & Q|_{a_{uu'}^{(k)}} + (a_{uu'} - a_{uu'}^{(k)})\nabla_{a_{uu'}} Q|_{a_{uu'}^{(k)}} \\ & + \frac{1}{2\eta}\|a_{uu'} - a_{uu'}^{(k)}\|_2^2 + +\alpha_S\|a_{uu'}\|_1 \\ & + \alpha_G\|a_{uu'}\|_2.\end{aligned} \tag{11}$$

Because both $Q|_{a_{uu'}^{(k)}}$ and $\nabla_{a_{uu'}} Q|_{a_{uu'}^{(k)}}$ are known, we add $\frac{\eta}{2}\|\nabla_{a_{uu'}} Q|_{a_{uu'}^{(k)}}\|_2^2$ to the objective function of (11) and reduce $Q|_{a_{uu'}^{(k)}}$ from it, and obtain an equivalent optimization problem

$$\begin{aligned}\min_{a_{uu'}\geq\mathbf{0}} & \frac{1}{2\eta}\|a_{uu'} - (a_{uu'}^{(k)} - \eta\nabla_{a_{uu'}} Q|_{a_{uu'}^{(k)}})\|_2^2 \\ & + \alpha_S\|a_{uu'}\|_1 + \alpha_G\|a_{uu'}\|_2.\end{aligned} \tag{12}$$

The objective function in (12) is convex, so the optimal solution is characterized by the subgradient equations.

$$a_{uu'}^{(k)} - \eta\nabla_{a_{uu'}} Q|_{a_{uu'}^{(k)}} - a_{uu'} = \eta\alpha_S\gamma + \eta\alpha_G\beta. \tag{13}$$

$\gamma = [\gamma_1, ..., \gamma_M]^\top$, where $\gamma_m = 1$ if $a_{uu'}^m > 0$, and in $[0, 1]$ otherwise. $\beta = \frac{a_{uu'}}{\|a_{uu'}\|_2}$ if $a_{uu'} \neq \mathbf{0}$, and in the set $\{x|\|x\|_2 \leq 1\}$ otherwise. Combining the subgradient equations with the basic algebra in (Simon et al., 2013), we get that $a_{uu'} = \mathbf{0}$ if (7) holds, otherwise $a_{uu'}$ satisfies

$$\left(1 + \frac{\eta\alpha_G}{\|a_{uu'}\|_2}\right) a_{uu'} = S_{\eta\alpha_S}(a_{uu'}^{(k)} - \eta\nabla_{a_{uu'}} Q|_{a_{uu'}^{(k)}}). \tag{14}$$

Taking the norm on both sides, $\|a_{uu'}\|_2$ can be replaced by

$$(\|S_{\eta\alpha_S}(a_{uu'}^{(k)} - \eta\nabla_{a_{uu'}} Q|_{a_{uu'}^{(k)}})\|_2 - t\eta\alpha_G)_+. \tag{15}$$

Replacing the $\|a_{uu'}\|_2$ in (14) with (15), we obtain the generalized gradient step in (8).

6.1. Details of Basis Function Selection

In our model, the intensity function of Hawkes process over all dimensions is:

$$\begin{aligned}\lambda(t) &= \sum_{u=1}^{U} \lambda_u(t) \\ &= \sum_{u=1}^{U}\left(\mu_u + \sum\nolimits_{u'=1}^{U}\int_0^t \phi_{uu'}(s)dN_{u'}(t-s)\right) \\ &= \sum_{u=1}^{U}\mu_u + \sum_{u=1}^{U}\sum_{t_i<t}\phi_{uu_i}(t-t_i) \\ &= \sum_{u=1}^{U}\mu_u + \sum_{u=1}^{U}\sum_{t_i<t}\sum_{m=1}^{M} a_{uu_i}^m \kappa_m(t-t_i).\end{aligned} \tag{16}$$

Applying Fourier transform, we have

$$\begin{aligned}\hat{\lambda}(\omega) = & \sum_{u=1}^{U}\mu_u\sqrt{2\pi}\delta(\omega) \\ & + \sum_{u=1}^{U}\sum_{t_i<t}\sum_{m=1}^{M} a_{uu_i}^m e^{-j\omega t_i}\hat{\kappa}_m(\omega).\end{aligned} \tag{17}$$

In other words, the spectral of $\lambda(t)$ is the weighted sum of those of basis functions. Therefore, the cut-off frequency of basis function is bounded by that of intensity function.

As we show in our paper, given training sequences $\mathcal{S} = \{s_c\}_{c=1}^{C}$, , where $s_c = \{(t_i^c, u_i^c)\}_{i=1}^{N_c}$, we can estimate $\lambda(t)$ empirically via a Gaussian-based kernel density estimator:

$$\lambda(t) = \sum\nolimits_{c=1}^{C}\sum\nolimits_{i=1}^{N_c} G_h(t - t_i^c). \tag{18}$$

Here t_i^c is the time stamp of the i-th event at the c-th sequence. $G_h(t - t_i^c) = \exp(-\frac{(t-t_i^c)^2}{2h^2})$ is a Gaussian kernel with the bandwidth h.

Because we only care about the selection of basis functions, we just need to estimate the spectral of $\lambda(t)$ rather than compute (9) directly. Specifically, applying Silverman's rule of thumb (Silverman, 1986), we first set optimal $h = (\frac{4\hat{\sigma}^5}{3\sum_c N_c})^{0.2}$, where $\hat{\sigma}$ is the standard deviation of time stamps $\{t_i^c\}$. Applying Fourier transform, we compute an upper bound for the spectral of $\lambda(t)$ as

$$\begin{aligned}|\hat{\lambda}(\omega)| &= \left|\int_{-\infty}^{\infty}\lambda(t)e^{-j\omega t}dt\right| \\ &= \left|\sum_{c=1}^{C}\sum_{i=1}^{N_c}\int_{-\infty}^{\infty} e^{-\frac{(t-t_i^c)^2}{2h^2}} e^{-j\omega t}dt\right| \\ &\leq \sum_{c=1}^{C}\sum_{i=1}^{N_c}\left|\int_{-\infty}^{\infty} e^{-\frac{(t-t_i^c)^2}{2h^c}} e^{-j\omega t}dt\right| \\ &= \sum_{c=1}^{C}\sum_{i=1}^{N_c}\left|e^{-j\omega t_i^c} e^{-\frac{\omega^2 h^2}{2}}\sqrt{2\pi h^2}\right| \\ &\leq \sum_{c=1}^{C}\sum_{i=1}^{N_c}\left|e^{-j\omega t_i^c}\right|\left|e^{-\frac{\omega^2 h^2}{2}}\sqrt{2\pi h^2}\right| \\ &= \left(\sum_{c=1}^{C} N_c\sqrt{2\pi h^2}\right) e^{-\frac{\omega^2 h^2}{2}}.\end{aligned} \tag{19}$$

Furthermore, we can compute the upper bound of the absolute sum of the spectral higher than ω_0 as

$$\begin{aligned}
&\int_{\omega_0}^{\infty} |\hat{\lambda}(\omega)| d\omega \\
\leq &\left(\sum_{c=1}^{C} N_c \sqrt{2\pi h^2}\right) \int_{\omega_0}^{\infty} e^{-\frac{\omega^2 h^2}{2}} d\omega \\
=&2\pi \left(\sum_{c=1}^{C} N_c\right) \int_{\omega_0}^{\infty} \frac{h}{\sqrt{2\pi}} e^{-\frac{\omega^2 h^2}{2}} d\omega \\
=&2\pi \left(\sum_{c=1}^{C} N_c\right) \left(\frac{1}{2} - \int_{0}^{\omega_0} \frac{h}{\sqrt{2\pi}} e^{-\frac{\omega^2 h^2}{2}} d\omega\right) \\
=&2\pi \left(\sum_{c=1}^{C} N_c\right) \left(\frac{1}{2} - \frac{1}{2}\int_{-\omega_0}^{\omega_0} \frac{h}{\sqrt{2\pi}} e^{-\frac{\omega^2 h^2}{2}} d\omega\right) \\
=&\pi \left(\sum_{c=1}^{C} N_c\right) \left(1 - \frac{1}{\sqrt{2}} \mathrm{erf}(\omega_0 h)\right),
\end{aligned} \tag{20}$$

where $\mathrm{erf}(x) = \frac{1}{\sqrt{\pi}} \int_{-x}^{x} e^{-t^2} dt$.

Therefore, give a bound of residual ϵ, we can find an ω_0 guaranteeing $\int_{\omega_0}^{\infty} |\hat{\lambda}(\omega)| d\omega \leq \epsilon$, or $\mathrm{erf}(\omega_0 h) \geq \sqrt{2} - \frac{\sqrt{2}\epsilon}{\pi \sum_{c=1}^{C} N_c}$. The proposed basis functions $\{\kappa_{\omega_0}(t, t_m)\}_{m=1}^{M}$ are selected, where ω_0 is the cut-off frequency of basis function and $t_m = \frac{(m-1)T}{M}$, $M = \lceil \frac{T\omega_0}{\pi} \rceil$.

References

Adams, Ryan Prescott, Murray, Iain, and MacKay, David JC. Tractable nonparametric bayesian inference in poisson processes with gaussian process intensities. In *ICML*, 2009.

Ahmed, Amr and Xing, Eric P. Recovering time-varying networks of dependencies in social and biological studies. *Proceedings of the National Academy of Sciences*, 106(29):11878–11883, 2009.

Alan, V Oppenheim, Ronald, W Schafer, and John, RB. Discrete-time signal processing. *New Jersey, Printice Hall Inc*, 1989.

Arnold, Andrew, Liu, Yan, and Abe, Naoki. Temporal causal modeling with graphical granger methods. In *KDD*, 2007.

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*, 85(5):1–12, 2012.

Bacry, Emmanuel, Delattre, Sylvain, Hoffmann, Marc, and Muzy, Jean-Francois. Some limit theorems for hawkes processes and application to financial statistics. *Stochastic Processes and their Applications*, 123(7):2475–2499, 2013.

Basu, Sumanta, Shojaie, Ali, and Michailidis, George. Network granger causality with inherent grouping structure. *Journal of Machine Learning Research*, 16:417–453, 2015.

Carstensen, Lisbeth, Sandelin, Albin, Winther, Ole, and Hansen, Niels R. Multivariate hawkes process models of the occurrence of regulatory elements. *BMC bioinformatics*, 11(1):456, 2010.

Chwialkowski, Kacper and Gretton, Arthur. A kernel independence test for random processes. In *ICML*, 2014.

Daley, Daryl J and Vere-Jones, David. *An introduction to the theory of point processes: volume II: general theory and structure*, volume 2. Springer Science & Business Media, 2007.

Daneshmand, Hadi, Gomez-Rodriguez, Manuel, Song, Le, and Schoelkopf, Bernhard. Estimating diffusion network structures: Recovery conditions, sample complexity & soft-thresholding algorithm. In *ICML*, 2014.

Didelez, Vanessa. Graphical models for marked point processes based on local independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):245–264, 2008.

Du, Nan, Song, Le, Yuan, Ming, and Smola, Alex J. Learning networks of heterogeneous influence. In *NIPS*, 2012.

Eichler, Michael. Graphical modelling of multivariate time series. *Probability Theory and Related Fields*, 153(1-2): 233–268, 2012.

Eichler, Michael, Dahlhaus, Rainer, and Dueck, Johannes. Graphical modeling for multivariate hawkes processes with nonparametric link functions. *Preprint*.

Farajtabar, M., Wang, Y., Gomez-Rodriguez, M., Li, S., Zha, H., and Song, L. Coevolve: A joint point process model for information diffusion and network coevolution. In *NIPS*, 2015.

Farajtabar, Mehrdad, Du, Nan, Gomez-Rodriguez, Manuel, Valera, Isabel, Zha, Hongyuan, and Song, Le. Shaping social activity by incentivizing users. In *Advances in neural information processing systems*, pp. 2474–2482, 2014.

Gunawardana, Asela, Meek, Christopher, and Xu, Puyang. A model for temporal dependencies in event streams. In *NIPS*, 2011.

Hall, Eric C and Willett, Rebecca M. Tracking dynamic point processes on networks. *arXiv preprint arXiv:1409.0031*, 2014.

Han, Fang and Liu, Han. Transition matrix estimation in high dimensional time series. In *ICML*, 2013.

Hansen, Niels Richard, Reynaud-Bouret, Patricia, Rivoirard, Vincent, et al. Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli*, 21(1):83–143, 2015.

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

Lemonnier, Remi and Vayatis, Nicolas. Nonparametric markovian learning of triggering kernels for mutually exciting and mutually inhibiting multivariate hawkes processes. In *Machine Learning and Knowledge Discovery in Databases*, pp. 161–176. 2014.

Lewis, Erik and Mohler, George. A nonparametric em algorithm for multiscale hawkes processes. *Journal of Nonparametric Statistics*, 2011.

Lian, Wenzhao, Henao, Ricardo, Rao, Vinayak, Lucas, Joseph, and Carin, Lawrence. A multitask point process predictive model. In *ICML*, 2015.

Lloyd, Chris, Gunter, Tom, Osborne, Michael A, and Roberts, Stephen J. Variational inference for gaussian process modulated poisson processes. In *ICML*, 2015.

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

Meek, Christopher. Toward learning graphical and causal process models. In *UAI Workshop Causal Inference: Learning and Prediction*, 2014.

Rasmussen, Jakob Gulddahl. Bayesian inference for hawkes processes. *Methodology and Computing in Applied Probability*, 15(3):623–642, 2013.

Reynaud-Bouret, Patricia, Schbath, Sophie, et al. Adaptive estimation for hawkes processes; application to genome analysis. *The Annals of Statistics*, 38(5):2781–2822, 2010.

Samo, Yves-Laurent Kom and Roberts, Stephen. Scalable nonparametric bayesian inference on point processes with gaussian processes. In *ICML*, 2015.

Silverman, Bernard W. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.

Simon, Noah, Friedman, Jerome, Hastie, Trevor, and Tibshirani, Robert. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

Song, Dong, Wang, Haonan, Tu, Catherine Y, Marmarelis, Vasilis Z, Hampson, Robert E, Deadwyler, Sam A, and Berger, Theodore W. Identification of sparse neural functional connectivity using penalized likelihood estimation and basis functions. *Journal of computational neuroscience*, 35(3):335–357, 2013.

Yang, Haiqin, Xu, Zenglin, King, Irwin, and Lyu, Michael R. Online learning for group lasso. In *ICML*, 2010.

Yang, Shuang-Hong and Zha, Hongyuan. Mixture of mutually exciting processes for viral diffusion. In *ICML*, 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 multidimensional hawkes processes. In *AISTATS*, 2013a.

Zhou, Ke, Zha, Hongyuan, and Song, Le. Learning triggering kernels for multi-dimensional hawkes processes. In *ICML*, 2013b.