

Structural Hawkes Processes for Learning Causal Structure from Discrete-Time Event Sequences

Jie Qiao¹, Ruichu Cai^{1,2*}, Siyu Wu¹, Yu Xiang¹, Keli Zhang³ and Zhifeng Hao⁴

¹School of Computer Science, Guangdong University of Technology, Guangzhou 510006, China

²Peng Cheng Laboratory, Shenzhen 518066, China

³Huawei Noah's Ark Lab, Huawei, Shenzhen 518116, China

⁴College of Science, Shantou University, Shantou 515063, China

qiaojie.chn@gmail.com, cairuichu@gmail.com, fisherwsy@163.com,
thexiang2000@gmail.com, zhangkeli1@huawei.com, haozhifeng@stu.edu.cn

Abstract

Learning causal structure among event types from discrete-time event sequences is a particularly important but challenging task. Existing methods, such as the multivariate Hawkes processes based methods, mostly boil down to learning the so-called Granger causality which assumes that the cause event happens strictly prior to its effect event. Such an assumption is often untenable beyond applications, especially when dealing with discrete-time event sequences in low-resolution; and typical discrete Hawkes processes mainly suffer from identifiability issues raised by the instantaneous effect, i.e., the causal relationship that occurred simultaneously due to the low-resolution data will not be captured by Granger causality. In this work, we propose Structure Hawkes Processes (SHPs) that leverage the instantaneous effect for learning the causal structure among events type in discrete-time event sequence. The proposed method is featured with the minorization-maximization of the likelihood function and a sparse optimization scheme. Theoretical results show that the instantaneous effect is a blessing rather than a curse, and the causal structure is identifiable under the existence of the instantaneous effect. Experiments on synthetic and real-world data verify the effectiveness of the proposed method.

1 Introduction

Learning causal structure among event types on *multi-type event sequences* is an important and challenging task, and has recently found applications in social science [Zhou *et al.*, 2013a], economic [Bacry *et al.*, 2015], network operation maintenance [Cai *et al.*, 2022], etc. Existing methods, such as the multivariate Hawkes processes based methods [Xu *et al.*, 2016; Bhattacharjya *et al.*, 2018; Salehi *et al.*, 2019], mostly boil down to learning the so-called Granger causality [Granger, 1969] which implicitly assumes that all events are recorded instantaneously and accurately such that the cause event happens

strictly prior to its effect event (known as *temporal precedence assumption*). However, due to the limited recording capabilities and storage capacities, retaining event's occurred times with high-resolution is expensive or practically impossible in many real-world applications, and we usually only can access the corresponding discrete-time event sequences. For example, in large wireless networks, the event sequences are usually logged at a certain frequency by different devices whose time might not be accurately synchronized. As a result, low-resolution discrete-time event sequences are obtained and the temporal precedence assumption will be frequently violated in discrete-time event sequences, which raises a serious identifiability issue of causal discovery. For example, as shown in Fig. 1, there are three event sequences produced by three event types v_1 , v_2 , and v_3 , respectively. Let v_1 be the cause of v_2 and v_3 , Fig. 1(a) shows the accurate continuous-time event sequences such that each event occurred time is recorded by t_1, \dots, t_6 . However, such high-resolution sequences are usually not accessible and we can only observe the discrete-time event sequences as shown in Fig. 1(b). As a result, v_1 and v_2 will be considered simultaneous events due to the low-resolution, which violates the temporal precedence assumption. Consequently, many existing point process based methods will fail to capture the causal relationship $v_1 \rightarrow v_2$ as only the events that occur earlier are considered as causes. In contrast, the causal relationship $v_1 \rightarrow v_3$ can still be captured using the point process based method because the cause event v_1 occurs before v_3 . However, one can imagine that as the resolution becomes lower the causal relationship between v_1 and v_3 might no longer be identified as they might become occurs at the same time. Thus, in this paper, we aim to answer the following two questions: 1) How to design and learn a Hawkes process that leverages the instantaneous effect in discrete time? 2) Can we identify the causal relationship in event sequences under the existence of the instantaneous effect?

These two questions have to do with the point processes and the causal discovery, respectively. The former question is about the design of the discrete-time Hawkes processes. The effect of discretization of Hawkes processes has been widely discussed in many ways [Foufoula-Georgiou and Lettenmaier, 1986; Kirchner, 2016; Shlomovich *et al.*, 2022; Jacod and Todorov, 2009]. However, how to design a Hawkes process

*Corresponding author.

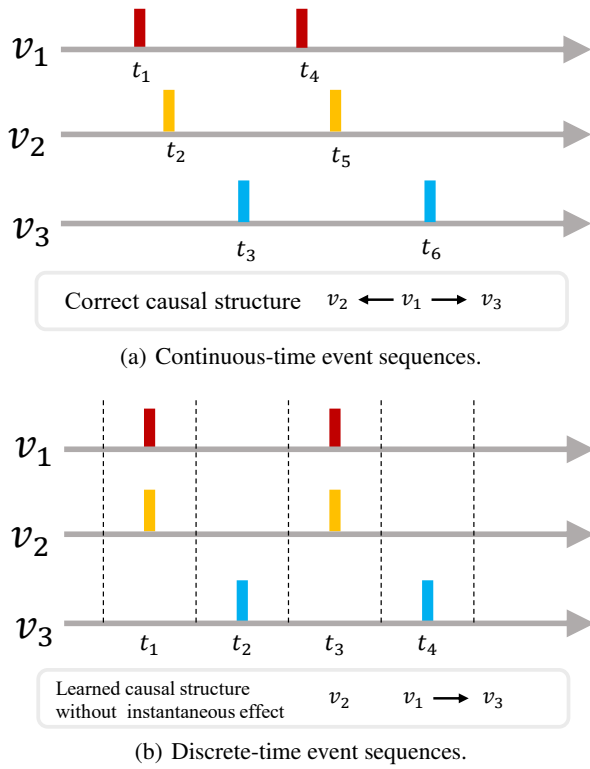


Figure 1: Toy example of the continuous-time and discrete-time event sequences. We consider three types of event v_1, v_2, v_3 and the causal relationship satisfying $v_2 \leftarrow v_1 \rightarrow v_3$. While, in continuous-time, each event’s occur times (t_1, \dots, t_6) can be observed accurately, the discrete-time event sequences have lower resolution, such that types v_1, v_2 will be considered as simultaneous and the Granger-based methods will fail to capture the causal relationship $v_1 \rightarrow v_2$.

that leverages the instantaneous effect in discrete Hawkes process is still unknown. The latter question is related to the causal discovery [Zhang *et al.*, 2018; Qiao *et al.*, 2021; Cai *et al.*, 2018a; Cai *et al.*, 2018b; Yu *et al.*, 2019; Yu *et al.*, 2021; Yang *et al.*, 2021]. Some methods have been developed for continuous-value time series with instantaneous effect. For example, the structural vector autoregression model [Swanson and Granger, 1997] is an extension of the vector autoregression model with instantaneous effect, and it has been shown that under the linear non-Gaussian model, the causal relationship is identifiable under the existence of the instantaneous effect [Hyvärinen *et al.*, 2010]. However, we are not aware of any method that leverages the instantaneous effect to identify the causal relationship in event sequences with point processes. Thus, in this work, we propose Structural Hawkes Processes (SHPs) that leverage the instantaneous effect in discrete-time event sequences. We will proceed as follows. In Section 2, we review the related work. In Section 3.1, we show how to design and learn the structural Hawkes processes. In Section 4, we investigate the identification of structural Hawkes processes and theoretically show that the causal relationship among the instantaneous effect is indeed identifiable. In Section 5, we perform extensive empirical evaluations over both synthetic and real-world data.

2 Related Work

This work is closely related to two topics: point processes and learning causal structure from event sequence.

Point processes. There are mainly two types of point processes for modeling the event sequence. Most existing methods focus on developing the continuous-time Hawkes processes [Hawkes, 1971] which assume that the event in the past can influence the event in the future. Different variants have been developed with different types of intensity in the Hawkes processes, e.g., the parametric functions [Farajtabar *et al.*, 2014; Zhou *et al.*, 2013a; Rasmussen, 2013; Cai *et al.*, 2022], the non-parametric functions [Lewis and Mohler, 2011; Zhou *et al.*, 2013b; Achab *et al.*, 2017], and the recent deep learning based functions [Du *et al.*, 2016; Mei and Eisner, 2017; Shang and Sun, 2019]. Another line of research focuses on the discrete-time Hawkes processes which are more appealing for certain applications. [Seol, 2015] study the limit theorem for the discrete Hawkes processes and it has been extended to the discrete marked Hawkes processes by [Wang, 2022]. [Shlomovich *et al.*, 2022] further discusses the estimation method for 1-dimensional discrete Hawkes processes. However, none of the above methods leverage the instantaneous effect for learning causal structure in discrete-time event sequences.

Learning causal structure from event sequences. Most approaches in learning causal structure from event sequences are based on Granger causality [Granger, 1969]. The basic idea of Granger causality is to constrain that the effect cannot precede the cause and the causal analysis can be conducted using predictability. In particular, many methods have been developed for learning Granger causality from the continuous-time event sequences based on multivariate Hawkes processes. [Xu *et al.*, 2016] proposes a nonparametric Hawkes processes model with group sparsity regularization, while [Zhou *et al.*, 2013a] proposes to use a nuclear and ℓ_1 norm and [Idé *et al.*, 2021] consider an ℓ_0 norm as the sparse regularization. In addition, [Achab *et al.*, 2017] proposes to use cumulant for learning the causal structure without optimization. Recently, some deep point process based methods have been proposed, e.g., [Zhang *et al.*, 2020] introduce an attribution method to uncover the Granger causality. However, the Granger causal relations could be misled in low-resolution which is also pointed out by [Spirtes and Zhang, 2016]. One remedy is to extend Granger causality with instantaneous effects. It has been found that in continuous-value time series, one is able to incorporate the instantaneous effect under the linear relation with non-Gaussian noise [Hyvärinen *et al.*, 2010]. [Runge, 2020] further proposes to use a constraint-based method for learning the causal structure under the instantaneous effects. However, the extensions are only applicable in a restricted case and it is still very challenging to learn causal structure from event sequences with instantaneous effects.

3 Structural Hawkes Processes

We begin with a brief introduction to the general continuous-time multivariate point processes. We then develop the structural Hawkes processes that leverage the instantaneous effect in discrete-time.

3.1 Multivariate Point Processes

A multivariate point process is a random process that can be presented via a $|\mathbf{V}|$ -dimensional counting process $\mathbf{N} = \{N_v(t) | t \in [0, T], v \in \mathbf{V}\}$ where $N_v(t) = N((0, t])$ measures the number of events that occur before time t in the event type v . Each counting process $N_v(t)$ can be characterized by the *conditional intensity function* $\lambda_v(t)$ satisfying:

$$\lambda_v(t)dt = \mathbb{E}[dN_v(t) | \mathcal{F}_{t-}],$$

where $\lambda_v(t)$ characterizes the (conditional) expected number of jumps per unit of time, $dN_v(t) = N_v(t + dt) - N_v(t)$ measures the increment of the jump, $\mathcal{F}_{t-} = \bigcup_{0 \leq s < t, v \in \mathbf{V}} \mathcal{F}_s^v$ in which \mathcal{F}_s^v is the canonical filtration of sub-processes $N_v(t)$ up to time s .

In particular, Hawkes process is a counting process with an intensity function that has the following form:

$$\lambda_v(t) = \mu_v + \sum_{v' \in \mathbf{V}} \int_0^t \phi_{v',v}(t-t') dN_{v'}(t'), \quad (1)$$

where μ_v is the immigration intensity and $\sum_{v' \in \mathbf{V}} \int_0^t \phi_{v',v}(t-t') dN_{v'}(t')$ is the endogenous intensity aiming to model the influence from other event types occurring near time t [Fajtabar *et al.*, 2014]. $\phi_{v',v}(t)$ is a reproduction function characterizing the time variation of the causal influence from event v' to v . Intuitively, one can imagine that events can be seen to arrive either via immigration according to μ_v or via reproduction from past events according to the endogenous intensity.

3.2 Design of Structural Hawkes Processes

To model the instantaneous effect, we begin with extending the continuous-time counting processes into the discrete-time in which the event sequences are observed or collected in $\mathbf{T} = \{\Delta, 2\Delta, \dots, K\Delta\}$, for $K = \lfloor T/\Delta \rfloor$ where $\Delta > 0$ is the length of time interval at each observed time. Then the multivariate counting processes in discrete-time can be defined as $\mathbf{N}^{(\Delta)} = \{N_v^{(\Delta)}(k) | k \in \{0, \dots, K\}, v \in \mathbf{V}\}$, where $N_v^{(\Delta)}(k) = N_v((0, k\Delta])$ measures the number of events that occurs not later than $k\Delta$. We further let $\mathbf{X} = \{X_{v,t} | v \in \mathbf{V}, t \in \{0, \dots, K\}\}$ denote the set of observations at each time interval where $X_{v,t} := N_v(t\Delta) - N_v((t-1)\Delta)$ is an analogy to $dN_v(t)$.

The discrete-time counting processes, however, as we discussed earlier, will ignore the instantaneous effect which could lead to a misleading result. To tackle this issue, we propose structural counting processes with a new type of conditional intensity function that leverages the events that occur at the same period of time.

Definition 1 (Structural counting processes). *A structural counting process is a multivariate counting process $\mathbf{N}^{(\Delta)}$ in discrete-time with the conditional intensity of $N_v^{(\Delta)}$ for each $v \in \mathbf{V}$ satisfying:*

$$\lambda_v(k\Delta)\Delta = \mathbb{E}[X_{v,k} | \mathcal{F}_{(k-1)\Delta} \cup \mathcal{F}_{k\Delta}^{-v}], \quad (2)$$

where $\mathcal{F}_{(k-1)\Delta} = \bigcup_{0 \leq s \leq k-1, v \in \mathbf{V}} \mathcal{F}_s^v$ is the filtration with discrete-time in the past and $\mathcal{F}_{k\Delta}^{-v} := \{\mathcal{F}_{k\Delta}^{v'} | v' \in \mathbf{V} \setminus v\}$ is the filtration that except for type- v event.

Note that the exclusion in $\mathcal{F}_{k\Delta}^{-v}$ is necessary since it makes no sense to use the current number of events to predict itself. Based on the structural counting processes in Definition 1, the structural Hawkes processes can be designed as follows:

Definition 2 (Structural Hawkes processes). *A structural Hawkes process is a structural counting process such that for each $v \in \mathbf{V}$, the intensity of $N_v^{(\Delta)}$ can be written as:*

$$\lambda_v(k\Delta) = \mu_v + \sum_{v' \in \mathbf{V}} \sum_{i=1}^k \phi_{v',v}((k-i)\Delta) X_{v',i}, \quad (3)$$

where $\phi_{v',v}(0) \equiv 0$ ensures the exclusive of type- v event at time $k\Delta$.

We can see that the intensity in Eq. (3) is not only influenced by the events that occur in the past $(k-1)\Delta$ but also the events that occur at the same period of time $k\Delta$.

3.3 Learning of Structural Hawkes Processes

The goal of this work is to identify a proper *directed acyclic graph* (DAG) $\mathcal{G}(\mathbf{V}, \mathbf{E})$ among event types \mathbf{V} such that for a type- v event, the intensity $\lambda_v(t)$ only depends on the event from its cause \mathbf{Pa}_v , where \mathbf{Pa}_v is the parent set of v according to the edge set \mathbf{E} , i.e., $\{v' \rightarrow v | v' \in \mathbf{Pa}_v\} \subseteq \mathbf{E}$. Note that a DAG constraint for the instantaneous causal relation is necessary due to the lack of temporal precedence but it can be easily relaxed by taking lagged and instantaneous relations separately which we omit here for simplicity.

To learn causal structure, we parameterize the causal influence into the reproduction function $\phi_{v',v}(t) = \alpha_{v',v} \kappa(t)$ where $\alpha_{v',v}$ denotes the causal strength of causal relation $v' \rightarrow v$ and $\kappa(t)$ characterizes the time-decay of the causal influence which is usually set as an exponential form $\kappa(t) = \exp(-\beta t)$ for $t \geq 0$ with the hyper-parameter β . That is, the causal relationship can be encoded by the impact function such that for any pairs $v', v \in \mathbf{V}$ if $\phi_{v',v}(t) = 0$ we have $v' \rightarrow v \notin \mathbf{E}$.

Thus, one of the challenges of learning causal structure is to constrain the sparsity of the reproduction function. To this end, we devise the objective function with sparsity constraint using the ℓ_0 norm under the likelihood framework.

Given a collection of discrete-time sequences \mathbf{X} , the log-likelihood of parameters $\Theta = \{\mathbf{A} = [\alpha_{v',v}] \in \mathbb{R}^{|\mathbf{V}| \times |\mathbf{V}|}, \mu = [\mu_v] \in \mathbb{R}^{|\mathbf{V}|}\}$ of SHP can be expressed as follows:

$$\begin{aligned} & \mathcal{L}(\mathcal{G}, \Theta; \mathbf{X}) \\ &= \sum_{v \in \mathbf{V}} \sum_{k=1}^K [-\lambda_v(k\Delta)\Delta + X_{v,k} \log(\lambda_v(k\Delta))] \\ & \quad + \underbrace{\sum_{v \in \mathbf{V}} \sum_{k=1}^K [-\log(X_{v,k}!) + X_{v,k} \log(\Delta)]}_{\text{Const.}} \\ &= \sum_{v \in \mathbf{V}} \sum_{k=1}^K \left[-\left(\mu_v + \sum_{v' \in \mathbf{V}} \sum_{i=1}^k \phi_{v',v}((k-i)\Delta) X_{v',i} \right) \Delta \right. \\ & \quad \left. + X_{v,k} \log \left(\mu_v + \sum_{v' \in \mathbf{V}} \sum_{i=1}^k \phi_{v',v}((k-i)\Delta) X_{v',i} \right) \right] + \text{Const.} \end{aligned} \quad (4)$$

Without further constraint, the log-likelihood function will tend to produce excessive redundant causal edges. Thus, we further penalize the model with ℓ_0 norm to enforce the sparsity and we obtain the objective function as follows:

$$\mathcal{L}_p(\mathcal{G}, \Theta; \mathbf{X}) = \mathcal{L}(\mathcal{G}, \Theta; \mathbf{X}) + \alpha_S \|\mathbf{A}\|_0, \quad (5)$$

where α_S controls the strength of ℓ_0 norm. Although there are different forms of sparse regularization for the constraint, e.g., [Xu *et al.*, 2016] proposed to use the ℓ_1 and ℓ_2 norm, the ℓ_0 can provide better sparsity performance and it is easy to be extended, for example, considering the directed acyclic constraint [Tsamardinos *et al.*, 2006], as many causal structures in real-world are often directed acyclic.

3.4 A Minorization-Maximization-based Algorithm for Learning Causal Structure

However, estimation of the parameters by maximizing the likelihood in Eq. (5) has two obstacles. First, the likelihood of the point processes model is known as flat making the optimization unstable and computationally intensive [Veen and Schoenberg, 2008]. Second, learning DAGs from observational data is a combinatorial problem and, without any assumption, it has been shown to be NP-hard [Chickering *et al.*, 2004]. To tackle the issues above, following [Lewis and Mohler, 2011; Chickering, 2002], we propose to use minorization-maximization (MM) based optimization [Hunter and Lange, 2004] with a two-step causal structure learning algorithm.

For the first step, the parameters are estimated given a fixed causal graph using the MM algorithm which leverages the additive structure in Eq. (4) to apply Jensen's inequality similar to the EM algorithm and obtain the following lower bound:

$$\begin{aligned} & Q(\Theta | \Theta^{(j)}) \\ &= \sum_{v \in \mathbf{V}} \sum_{k=1}^K \left[- \left(\mu_v + \sum_{v' \in \mathbf{V}} \sum_{i=1}^k \phi_{v',v}((k-i)\Delta) X_{v',i} \right) \Delta \right. \\ & \quad \left. + X_{v,k} \left(q_{v,k}^\mu \log \left(\frac{\mu_v}{q_{v,k}^\mu} \right) \right) \right. \\ & \quad \left. + \sum_{v' \in \mathbf{V}} \sum_{i=1}^k q_{v,k}^\alpha(v', i) \log \left(\frac{\phi_{v',v}((k-i)\Delta) X_{v',i}}{q_{v,k}^\alpha(v', i)} \right) \right] \end{aligned} \quad (6)$$

where $q_{v,k}^\mu = \frac{\mu_v^{(j)}}{\lambda_v^{(j)}(k\Delta)}$ and $q_{v,k}^\alpha(v', i) = \frac{\phi_{v',v}^{(j)}((k-i)\Delta) X_{v',i}}{\lambda_v^{(j)}(k\Delta)}$. $\lambda_v^{(j)}(k\Delta)$ is the conditional intensity function with parameters $\Theta^{(j)}$ in the j -th iteration. Then, by setting $\frac{\partial Q(\Theta | \Theta^{(j)})}{\partial \mu_v} = 0$, $\frac{\partial Q(\Theta | \Theta^{(j)})}{\partial \alpha_{v',v}} = 0$, we obtain the close-form iteration formulas:

$$\begin{aligned} \mu_v^{(j+1)} &= \frac{\sum_{k=1}^K X_{v,k} q_{v,k}^\mu}{K \Delta} \\ \alpha_{v',v}^{(j+1)} &= \begin{cases} \frac{\sum_{k=1}^K \sum_{i=1}^k \kappa((k-i)\Delta) X_{v',i} \Delta}{\sum_{k=1}^K \sum_{i=1}^k q_{v,k}^\alpha(v', i) X_{v,k}} & v' \neq v \\ \frac{\sum_{k=1}^K \sum_{i=1}^{k-1} q_{v,k}^\alpha(v', i) X_{v,k}}{\sum_{k=1}^K \sum_{i=1}^{k-1} \kappa((k-i)\Delta) X_{v',i} \Delta} & v' = v \end{cases} \end{aligned} \quad (7)$$

With the MM algorithm, we then search the causal structure by using a Hill-Climbing-based algorithm as shown in Algorithm 1. It mainly consists of two phases. First, we perform a

Algorithm 1 Learning causal structure using SHP

Input: Data set \mathbf{X}

Output: G^*, Θ^*

- 1: $G^l \leftarrow$ empty graph, $\mathcal{L}_p^* \leftarrow -\infty$
 - 2: **while** $\mathcal{L}_p^*(G^*, \Theta^*; \mathbf{X}) < \mathcal{L}_p^*(G^l, \Theta^l; \mathbf{X})$ **do**
 - 3: $G^*, \Theta^* \leftarrow G^l, \Theta^l$ with largest $\mathcal{L}_p^*(G^l, \Theta^l; \mathbf{X})$
 - 4: **for** every $G^l \in \mathcal{V}(G^*)$ **do**
 - 5: Update Θ^l via iteration in Eq. (7)
 - 6: Record score $\mathcal{L}_p^*(G^l, \Theta^l; \mathbf{X})$
 - 7: **end for**
 - 8: **end while**
 - 9: **return** G^*, Θ^*
-

structure searching scheme by taking one step adding, deleting, and reversing the graph G^* in the last iteration, i.e., in Line 4, $\mathcal{V}(G^*)$ represents a collection of the one-step modified graph of $\mathcal{V}(G^*)$. Furthermore, the acyclic constraint is implemented by eliminating all cyclic causal graphs in $\mathcal{V}(G^*)$. Second, by fixing the graph G^l , we optimize the log-likelihood using the MM algorithm in Line 5. Iterating the two steps above until the likelihood no longer increases.

4 Identifiability

In this section, we aim to answer the question that whether we can identify the causal relationship under the existence of the instantaneous effect. In answering this question, one will need to explore the property of Hawkes processes in discrete-time. Based on the discrete-time likelihood in Eq. (4), each interval is modeled by conditional Poisson distribution with the linear structure in Eq. (3). As such, one may alternatively represent the relation by integer-valued autoregressive (INAR) processes according to [Kirchner, 2016] for analyzing the identification of structural Hawkes Processes. Furthermore, we assume that the causal sufficiency holds, i.e., all relevant variables have been observed [Spirites *et al.*, 2000]. Following [Kirchner, 2016], the INAR(∞) processes can be defined as follows:

Definition 3 (INAR(∞)). For $\theta_k \geq 0, k \in \mathbb{N}_0$, let $\epsilon_t \stackrel{i.i.d.}{\sim} \text{Pois}(\theta_0), t \in \mathbb{N}$, and $\xi_i^{(t,k)} \sim \text{Pois}(\theta_k)$. An Integer-valued autoregressive time series of infinite order (INAR(∞)) process $X_t, t \in \mathbb{N}$ is defined by

$$X_t = \sum_{k=1}^{\infty} \theta_k \circ X_{t-k} + \epsilon_t \quad (8)$$

where \circ is a reproduction operator given by $\theta_k \circ X_{t-k} \equiv \sum_{i=1}^{X_{t-k}} \xi_i^{(t,k)}$ with $\xi_i^{(t,k)}$ be a sequence of i.i.d. non-negative integer-valued random variables that depends on the reproduction coefficients θ_k , $\{\epsilon_t\}_{t \in \mathbb{N}}$ is an i.i.d. integer-valued immigration sequence that are independent of $\{\xi_i^{(t,k)}\}$, and X_{t-k} is independent of ϵ_t for all k .

In general, INAR is a discrete analogy of the continuous autoregressive model. Note that the distribution of the independent variables in INAR could be different and different choices of the distribution would lead to different INAR models [Guerrero *et al.*, 2022]. Here, we use the Poisson choice

to adopt the conditional Poisson distribution in Eq. (4), and it has been shown that such INAR(∞) model will converge to the continuous-time Hawkes processes as the time interval $\Delta \rightarrow 0$:

Theorem 1 ([Kirchner, 2016]). *Let N be a Hawkes process with immigration intensity μ and let $\phi : \mathbb{R} \rightarrow \mathbb{R}_0^+$ be a reproduction intensity that is piecewise continuous with $\phi(t) = 0, t \leq 0$ and $\int \phi(t)dt < 1$. For $\Delta \in (0, \delta)$, let $(X_t^{(\Delta)})$ be an INAR(∞) sequence with immigration parameter $\Delta\mu$ and reproduction coefficients $\Delta\phi(k\Delta), k \in \mathbb{N}$. From the sequences $(X_t^{(\Delta)})_{\Delta \in (0, \delta)}$, we define a family of point processes by*

$$N^{(\Delta)}(A) := \sum_{k:k\Delta \in A} X_k^{(\Delta)}, \quad A \in \mathcal{B}, \Delta \in (0, \delta), \quad (9)$$

where $\mathcal{B} := \mathcal{B}(\mathbb{R})$ is the Borel set in \mathbb{R} . Then, we have that

$$N^{(\Delta)} \xrightarrow{w} N \quad \text{for } \Delta \rightarrow 0. \quad (10)$$

Theorem 1 implies that the properties of INAR can be utilized for analyzing the Hawkes processes in discrete-time. Intuitively, the reproduction function ϕ in the discrete-time Hawkes processes can be represented by a series of reproduction coefficients θ_k for each time period, and the immigration intensity μ in the discrete-time Hawkes processes can be represented by the immigration parameters of ϵ_t .

Specifically, given the property of INAR processes, the analysis of the identifiability of structural Hawkes processes can be typically performed in two folds—the identifiability of the temporal structural Hawkes processes and the instantaneous structural Hawkes processes [Hyvärinen *et al.*, 2010]. For the former, the temporal resolution of the measurement that is high enough to capture the former and latter relationship of the events, the identifiability can be derived by local independence, which has been well explored by [Mogensen and Hansen, 2020]. For example, in Fig. 1, the causal relationship $v_1 \rightarrow v_3$ can be simply identified by the independence. Thus, in this work, we are more interested in the instantaneous structural Hawkes processes—the measurements have lower time resolution such that the causal influences are instantaneous. In such a case, one can use a model in which the influences are instantaneous, leading to Bayesian networks (BNs) or structural equation models (SEMs), e.g., the causal relationship $v_1 \rightarrow v_2$ in Fig. 1 belong to this class.

Thus, to analyze the identification of SHP, based on the INAR model, we consider the instantaneous causal structure in the structural Hawkes process:

Definition 4 (Instantaneous causal structure in structural Hawkes process). *Let $\epsilon_{v,t} \stackrel{i.i.d.}{\sim} \text{Pois}(\mu_v)$, and $\xi_i^{(v',v)} \sim \text{Pois}(\alpha_{v',v})$. The instantaneous causal structure in the structural Hawkes process consists of a set of equations of the form*

$$X_{v,t} = \sum_{v' \in \mathbf{V}} \alpha_{v',v} \circ X_{v',t} + \epsilon_v, \quad v \in \mathbf{V}, \quad (11)$$

where $\alpha_{v',v} > 0$ for $v' \in \mathbf{Pa}_v$ and $\alpha_{v',v} = 0$ for $v' \notin \mathbf{Pa}_v$, with $\alpha_{v',v} \circ X_{v',t} \equiv \sum_{i=1}^{X_{v',t}} \xi_i^{(v',v)}$, and $\mu_v > 0$ if v is a root

variable i.e., the variable whose parent set is empty $\mathbf{Pa}_v = \emptyset$. The random variables $X_{v',t}$, $\xi_i^{(v',v)}$, and ϵ_v are independent of each other.

The identifiability of the instantaneous causal structure, however, is much more difficult compared with the temporal causal relationship and most of them suffer from lack of identifiability. For example, without any constraint, the instantaneous causal structure may not be identified due to the Markov equivalence class [Pearl, 2009] in which all graphs in the equivalence class encode the same conditional independence. In such a case, different model with different causal graph and parameter, $M_1 = (\mathcal{G}_1, \Theta_1)$, $M_2 = (\mathcal{G}_2, \Theta_2)$ will produce the same distribution. Such non-identifiability also exists even considering an additional constraint, e.g., the linear Gaussian SEM is also non-identifiable [Spirites *et al.*, 2000].

Thus, an essential goal of this work is to investigate the identifiability of the instantaneous causal structure, i.e., whether there exists another causal structure that entails the same distribution to the underlying causal model. Such identifiability basically boils down to bivariate cases that whether there exists a backward model that is distribution equivalent to the causal direction. It can be easily extended to multivariate cases by incorporating the conditional independent constraint in the causal structure with the causal faithfulness assumption, i.e., the causal graph faithfully displays every dependency [Pearl, 1988]. Specifically, let X be the cause variable and Y be the effect variable, and surprisingly, the following theorem shows that the bivariate instantaneous causal pair is indeed identifiable:

Theorem 2. *Let $X \rightarrow Y$ be the correct causal direction that follows*

$$Y = \sum_{i=1}^X \xi_i + \epsilon, \quad X, \xi_i, \text{ and } \epsilon \text{ are independent}, \quad (12)$$

where $\xi_i \sim \text{Pois}(\alpha_{X,Y})$, $\epsilon \sim \text{Pois}(\mu_Y)$, $X \sim \text{Pois}(\mu_X)$. Then, there does not exist a backward model that admits the following equation:

$$X = \sum_{i=1}^Y \hat{\xi}_i + \hat{\epsilon}, \quad Y, \hat{\xi}_i, \text{ and } \hat{\epsilon} \text{ are independent}, \quad (13)$$

where $\hat{\xi}_i \sim \text{Pois}(\hat{\alpha}_{Y,X})$, $\hat{\epsilon} \sim \text{Pois}(\hat{\mu}_X)$, $Y \sim \text{Pois}(\hat{\mu}_Y)$.

To better understand the identifiability, we provide two alternative proofs for this theorem. The details of the proofs are given in the supplementary material [Qiao *et al.*, 2023]. The first proof shows that it is impossible to have a backward model that the probability distribution of the two directions is equivalence. While the second proof shows that the distribution of Y must not admit the Poisson distribution and therefore the distribution is not equivalent. We also provide an empirical study of the identifiability for the bivariate causal pair in Section 5. Finally, we generalize the result of Theorem 2 to the multivariate causal structure and show that the multivariate instantaneous causal structure is also identifiable:

Theorem 3. *With the causal faithfulness assumption and causal sufficiency assumption, the multivariate instantaneous causal structure is identifiable.*

The main idea behind the proof is that for any causal structure (\mathcal{G}, Θ) . there does not exist another causal structure

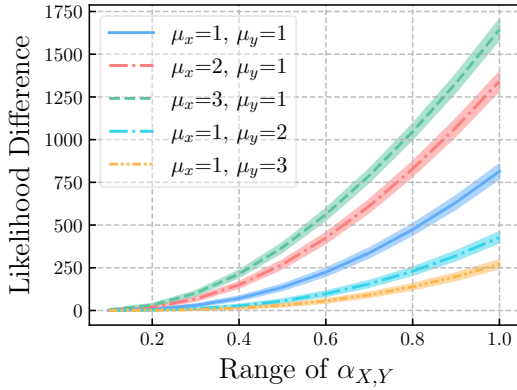


Figure 2: The likelihood difference between two causal pairs in Theorem 2 with different causal influence and immigration intensity.

$(\hat{\mathcal{G}}, \hat{\Theta})$ that is distribution equivalent. Specifically, with the causal faithfulness and causal sufficiency assumption, the Markov equivalent class is identified. Since all Markov equivalent classes share the same skeleton, we only need to show that each causal pair will not admit another causal structure that has the reversed causal direction while having the same likelihood (i.e., distribution equivalent).

5 Experiments

In this section, we test the proposed SHP and the baselines on both synthetic and real-world data. The baseline methods include ADM4 [Zhou *et al.*, 2013a], NPHC [Achab *et al.*, 2017], MLE_SGL [Xu *et al.*, 2016], and PCMCi Plus [Runge *et al.*, 2019]. We further develop SHP_NH, an ablation study of SHP that removes the hill-climb based searching scheme and uses a threshold to determine the causal structure following the work of ADM4 and MLE_SGL. In all following experiments, Recall, Precision, and F1 are used as the evaluation metrics. The results of recall and precision are provided in the supplementary material.

5.1 Synthetic Experiments

In this part, we design extensive control experiments using synthetic data to test the correctness of the theory and the sensitivity of sample size, length of time interval, number of event types, and different ranges of α and μ . In the sensitivity experiment, we synthesize data with fixed parameters while traversing the target parameter as shown in Fig. 3. The default settings are listed below, sample size=20000, time interval=5, number of event types=20, range of $\alpha \in [0.3, 0.5]$, range of $\mu \in [0.0005, 0.0001]$. All experimental results are averaged over 100 randomly generated causal structures.

The generating process proceeds as follows: 1) randomly generate a directed causal graph \mathcal{G} with a certain average in-degree; 2) generate the events according to randomly generated parameters $\alpha_{v',v}, \mu_v$ from \mathcal{G} , and 3) aggregate the number of counts at each interval for each event according to the length of time interval to synthesize the discrete-time process.

Two-variable case. We first conduct a simple two-variable experiment to verify the identifiability in Theorem 2, by com-

puting the likelihood difference between the causal direction and the reversed direction on a simulated causal pair with different causal influence $\alpha_{X,Y}$. That is, we simulated data using the model $Y_t = \alpha_{X,Y} \circ X_t + \epsilon_t$ with $X_t \sim \text{Pois}(\mu_x)$ and $\epsilon_t \sim \text{Pois}(\mu_y)$. Each experiment is conducted 100 times with random causal pairs. As shown in Fig. 2, we can see that the likelihood difference is always greater than zero which means that it is always identifiable unless the degenerate cases and it verifies the correctness of Theorem 2. In addition, as the causal influence $\alpha_{X,Y}$ decreases, the likelihood difference also decreases and tends to zero. This is reasonable as the causal influence becomes zero, and the two variables will also become independent making the model non-identifiable. Moreover, the level of immigration intensity also would affect the likelihood difference, which is reasonable as the lower the μ_x , the weaker the causal relation. Similarly, the higher the μ_y , the stronger the noise making the causal relation weaker. The more general multivariate case experiments will be implied in the following sensitivity analysis.

Sensitivity analysis. As shown in Fig. 3, we conduct six different control experiments for SHP. In general, our proposed SHP method outperforms all the baseline methods in all six control experiments.

In the control experiments of time interval given in Fig. 3(a), as the time interval controls the temporal resolution of the measurement sequence, the larger the time interval, the lower the temporal resolution, and the more instantaneous causal influences will occur leading to the decrease or insensitive of performance of the baseline methods. In the contract, SHP keeps giving the best results at all intervals. We also notice that the performance will decrease as the time interval increase because it reduces the sample size and the sequence become less informative. In addition, the performance of SHP_NH stresses the effectiveness of the searching algorithm which has lower precision and stability than SHP but also outperforms the baseline methods if the event sequences contain sufficient information on the instantaneous effect, which also verifies the ability to capture the instantaneous effect in SHP.

For the controlled experiments of causal strength and immigration intensity in Fig. 3(b) and Fig. 3(c), we can see that a reasonable causal strength and immigration intensity are required for the baseline methods while SHP is insensitive to both of them which shows the robustness of SHP.

In the sample size controlled experiments given in Fig. 3(d), all methods are robust to the sample size, and in particular, a 1000 sample size is enough to produce a reasonable performance for SHP, which demonstrates the practicality of SHP.

For the causal structure controlled experiments given in Fig. 3(e) and Fig. 3(f), SHP performs well in both experiments, and the average indegree is also important as it increases the performance of most methods decreases but SHP has the smallest decline.

5.2 Real World Experiments

We also test the proposed SHP on a very challenging real-world dataset¹ from real telecommunication networks. The

¹<https://competition.huaweicloud.com/informations/mobile/1000041487/dataset>

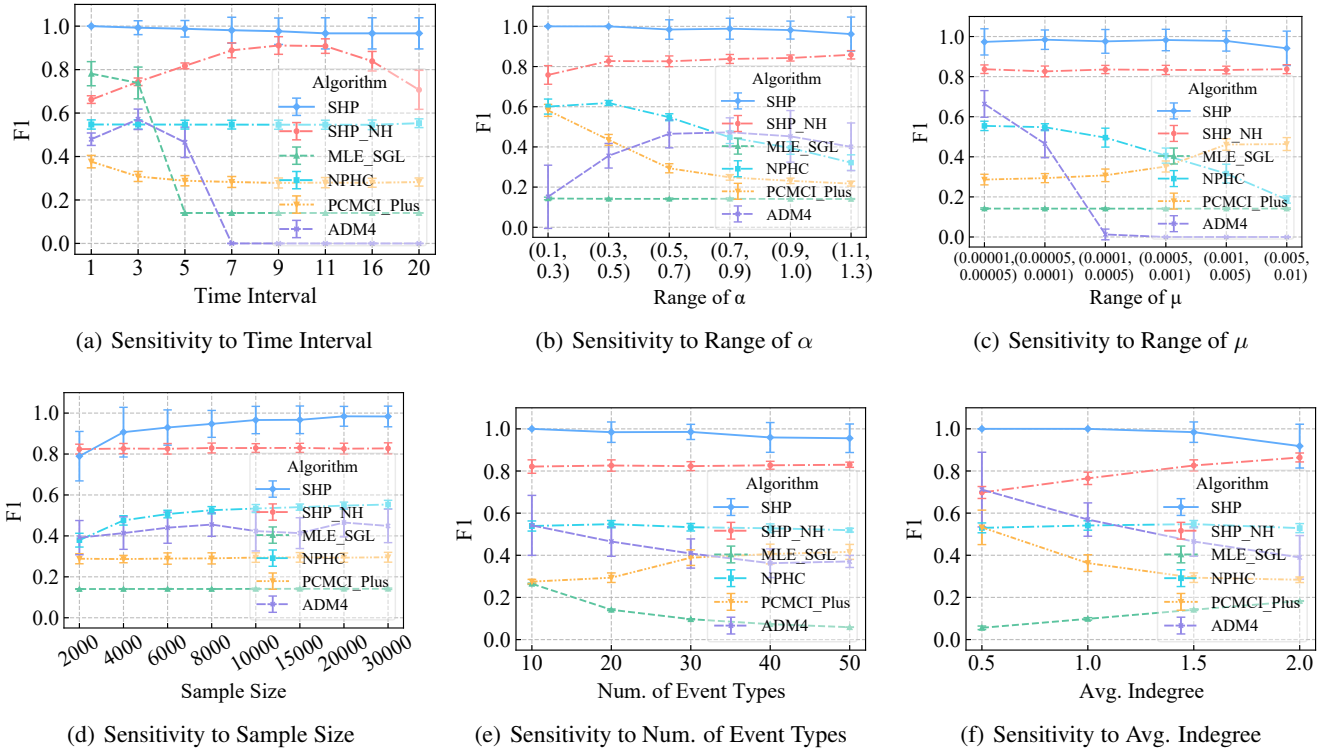


Figure 3: F1 in the Sensitivity Experiments

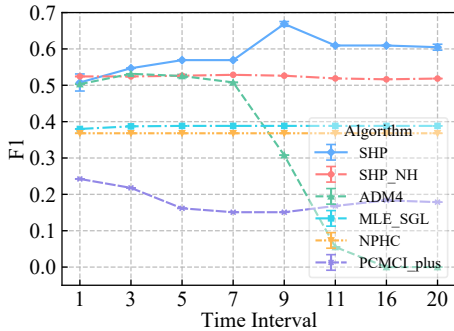


Figure 4: Real-world experiment on different temporal resolutions.

dataset records eight months of alarms that occurred in a real metropolitan cellular network. The alarms are generated by fifty-five devices in the metropolitan cellular network, which consists of eighteen types of alarms. Our goal is to learn the causal structure among the alarm types. Note that the causal impact among alarms is not deterministic but probabilistic in this system, which is similar to our model setting. All experiments have been conducted with different random seeds and the results are significant according to Wilcoxon signed-rank test. In addition, the ground truths of the causal relationships among alarm types are provided by domain experts.

As shown in Fig. 4, we conduct the real-world experiments by manually setting different temporal resolutions of the ob-

served sequence (i.e., from one second to nine seconds) and the original temporal resolution is one second. Interestingly, as the time interval increases, the performance of our method also increases instead of decreasing, while the baseline methods all decrease. The reason is that in the real-world scenario, there exists a communication latency in the logging system such that the recorded timestamp might be not fully accurate, and this error can be mitigated by decreasing the temporal resolution. In contrast, after decreasing the temporal resolution, other methods fail to capture the causal relationship under the instantaneous effect. This verifies the effectiveness of SHP and stresses the importance of the instantaneous effect.

6 Conclusion

In this work, we study how to model and leverage the instantaneous effect for learning causal structure in discrete-time event sequences. We propose structural Hawkes processes that leverage the instantaneous effects and a practical algorithm for learning the causal structure among event types. Theoretical results show that the instantaneous causal structure in structural Hawkes processes is indeed identifiable. To the best of our knowledge, this is the first causal structure learning method for event sequences with instantaneous effects. The success of SHP not only provides an effective solution for learning causal structure from real-world event sequences but also shows a promising direction for causal discovery from the discrete-time event sequences. In future, we plan to extend our work to a point process with a more general intensity function.

Acknowledgements

This research was supported in part by National Key R&D Program of China (2021ZD0111501), National Science Fund for Excellent Young Scholars (62122022), Natural Science Foundation of China (61876043, 61976052), the major key project of PCL (PCL2021A12). We appreciate the comments from anonymous reviewers, which greatly helped to improve the paper.

References

- [Achab *et al.*, 2017] Massil Achab, Emmanuel Bacry, Stéphane Gaïffas, Iacopo Mastromatteo, and Jean-François Muzy. Uncovering causality from multivariate Hawkes integrated cumulants. *Journal of Machine Learning Research*, 18(1):6998–7025, 2017.
- [Bacry *et al.*, 2015] Emmanuel Bacry, Iacopo Mastromatteo, and Jean-François Muzy. Hawkes processes in finance. *Market Microstructure and Liquidity*, 1(01):1550005, 2015.
- [Bhattacharjya *et al.*, 2018] Debarun Bhattacharjya, Dharmashankar Subramanian, and Tian Gao. Proximal graphical event models. In *Advances in Neural Information Processing Systems*, pages 8136–8145, 2018.
- [Cai *et al.*, 2018a] Ruichu Cai, Jie Qiao, Kun Zhang, Zhenjie Zhang, and Zhifeng Hao. Causal discovery from discrete data using hidden compact representation. In *Advances in Neural Information Processing Systems*, pages 2666–2674, 2018.
- [Cai *et al.*, 2018b] Ruichu Cai, Jie Qiao, Zhenjie Zhang, and Zhifeng Hao. SELF: Structural equational embedded likelihood framework for causal discovery. In *AAAI*, 2018.
- [Cai *et al.*, 2022] Ruichu Cai, Siyu Wu, Jie Qiao, Zhifeng Hao, Keli Zhang, and Xi Zhang. THPs: Topological Hawkes Processes for Learning Causal Structure on Event Sequences. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.
- [Chickering *et al.*, 2004] Max Chickering, David Heckerman, and Chris Meek. Large-sample learning of Bayesian networks is NP-hard. *Journal of Machine Learning Research*, 5:1287–1330, 2004.
- [Chickering, 2002] David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.
- [Du *et al.*, 2016] Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel Gomez-Rodriguez, and Le Song. Recurrent marked temporal point processes: Embedding event history to vector. In *International Conference on Knowledge Discovery and Data Mining*, pages 1555–1564, 2016.
- [Farajtabar *et al.*, 2014] Mehrdad Farajtabar, Nan Du, Manuel Gomez Rodriguez, Isabel Valera, Hongyuan Zha, and Le Song. Shaping social activity by incentivizing users. In *Advances in Neural Information Processing Systems*, pages 2474–2482, 2014.
- [Foufoula-Georgiou and Lettenmaier, 1986] Efi Foufoula-Georgiou and Dennis P Lettenmaier. Continuous-time versus discrete-time point process models for rainfall occurrence series. *Water Resources Research*, 22(4):531–542, 1986.
- [Granger, 1969] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.
- [Guerrero *et al.*, 2022] Matheus B Guerrero, Wagner Barreto-Souza, and Hernando Ombao. Integer-valued autoregressive processes with prespecified marginal and innovation distributions: A novel perspective. *Stochastic Models*, 38(1):70–90, 2022.
- [Hawkes, 1971] Alan G Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971.
- [Hunter and Lange, 2004] David R Hunter and Kenneth Lange. A tutorial on MM algorithms. *The American Statistician*, 58(1):30–37, 2004.
- [Hyvärinen *et al.*, 2010] Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11(5), 2010.
- [Idé *et al.*, 2021] Tsuyoshi Idé, Georgios Kollias, Dzung Phan, and Naoki Abe. Cardinality-Regularized Hawkes-Granger Model. *Advances in Neural Information Processing Systems*, 34:2682–2694, 2021.
- [Jacod and Todorov, 2009] Jean Jacod and Viktor Todorov. Testing for common arrivals of jumps for discretely observed multidimensional processes. *The Annals of Statistics*, 37(4):1792–1838, 2009.
- [Kirchner, 2016] Matthias Kirchner. Hawkes and INAR (∞) processes. *Stochastic Processes and their Applications*, 126(8):2494–2525, 2016.
- [Lewis and Mohler, 2011] Erik Lewis and George Mohler. A nonparametric EM algorithm for multiscale Hawkes processes. *Journal of Nonparametric Statistics*, 1(1):1–20, 2011.
- [Mei and Eisner, 2017] Hongyuan Mei and Jason M Eisner. The neural hawkes process: A neurally self-modulating multivariate point process. In *Advances in Neural Information Processing Systems*, pages 6754–6764, 2017.
- [Mogensen and Hansen, 2020] Søren Wengel Mogensen and Niels Richard Hansen. Markov equivalence of marginalized local independence graphs. *The Annals of Statistics*, 48(1):539–559, 2020.
- [Pearl, 1988] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan kaufmann, 1988.
- [Pearl, 2009] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge university press, 2009.
- [Qiao *et al.*, 2021] Jie Qiao, Ruichu Cai, Kun Zhang, Zhenjie Zhang, and Zhifeng Hao. Causal discovery with confounding cascade nonlinear additive noise models. *ACM*

- Transactions on Intelligent Systems and Technology (TIST)*, 12(6):1–28, 2021.
- [Qiao *et al.*, 2023] Jie Qiao, Ruichu Cai, Siyu Wu, Yu Xiang, Keli Zhang, and Zhifeng Hao. Structural Hawkes processes for learning causal structure from discrete-time event sequences. *arXiv:2305.05986*, 2023.
- [Rasmussen, 2013] Jakob Gulddahl Rasmussen. Bayesian inference for Hawkes processes. *Methodology and Computing in Applied Probability*, 15(3):623–642, 2013.
- [Runge *et al.*, 2019] Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 5(11):eaau4996, 2019.
- [Runge, 2020] Jakob Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Uncertainty in Artificial Intelligence*, pages 1388–1397. PMLR, 2020.
- [Salehi *et al.*, 2019] Farnood Salehi, William Trouleau, Matthias Grossglauser, and Patrick Thiran. Learning Hawkes processes from a handful of events. In *Advances in Neural Information Processing Systems*, pages 12694–12704, 2019.
- [Seol, 2015] Youngsoo Seol. Limit theorems for discrete Hawkes processes. *Statistics & Probability Letters*, 99:223–229, 2015.
- [Shang and Sun, 2019] Jin Shang and Mingxuan Sun. Geometric Hawkes processes with graph convolutional recurrent neural networks. In *AAAI Conference on Artificial Intelligence*, volume 33, pages 4878–4885, 2019.
- [Shlomovich *et al.*, 2022] Leigh Shlomovich, Edward AK Cohen, Niall Adams, and Lekha Patel. Parameter estimation of binned Hawkes processes. *Journal of Computational and Graphical Statistics*, pages 1–11, 2022.
- [Spirtes and Zhang, 2016] Peter Spirtes and Kun Zhang. Causal discovery and inference: concepts and recent methodological advances. In *Applied Informatics*, volume 3, pages 1–28. SpringerOpen, 2016.
- [Spirtes *et al.*, 2000] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT press, 2000.
- [Swanson and Granger, 1997] Norman R Swanson and Clive WJ Granger. Impulse response functions based on a causal approach to residual orthogonalization in vector autoregressions. *Journal of the American Statistical Association*, 92(437):357–367, 1997.
- [Tsamardinos *et al.*, 2006] Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78, 2006.
- [Veen and Schoenberg, 2008] Alejandro Veen and Frederic P Schoenberg. Estimation of space-time branching process models in seismology using an em-type algorithm. *Journal of the American Statistical Association*, 103(482):614–624, 2008.
- [Wang, 2022] Haixu Wang. Limit theorems for a discrete-time marked Hawkes process. *Statistics & Probability Letters*, 184:109368, 2022.
- [Xu *et al.*, 2016] Hongteng Xu, Mehrdad Farajtabar, and Hongyuan Zha. Learning granger causality for Hawkes processes. In *International Conference on Machine Learning*, pages 1717–1726, 2016.
- [Yang *et al.*, 2021] Shuai Yang, Kui Yu, Fuyuan Cao, Lin Liu, Hao Wang, and Jiuyong Li. Learning causal representations for robust domain adaptation. *IEEE Transactions on Knowledge and Data Engineering*, 2021.
- [Yu *et al.*, 2019] Kui Yu, Lin Liu, Jiuyong Li, Wei Ding, and Thuc Duy Le. Multi-source causal feature selection. *IEEE transactions on pattern analysis and machine intelligence*, 42(9):2240–2256, 2019.
- [Yu *et al.*, 2021] Kui Yu, Lin Liu, and Jiuyong Li. A unified view of causal and non-causal feature selection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(4):1–46, 2021.
- [Zhang *et al.*, 2018] Kun Zhang, Bernhard Schölkopf, Peter Spirtes, and Clark Glymour. Learning causality and causality-related learning: some recent progress. *National science review*, 5(1):26–29, 2018.
- [Zhang *et al.*, 2020] Wei Zhang, Thomas K. Panum, Somesh Jha, Prasad Chalasani, and David Page. CAUSE: Learning granger causality from event sequences using attribution methods. In *International Conference on Machine Learning*, 2020.
- [Zhou *et al.*, 2013a] Ke Zhou, Hongyuan Zha, and Le Song. Learning social infectivity in sparse low-rank networks using multi-dimensional Hawkes processes. In *Artificial Intelligence and Statistics*, pages 641–649, 2013.
- [Zhou *et al.*, 2013b] Ke Zhou, Hongyuan Zha, and Le Song. Learning triggering kernels for multi-dimensional Hawkes processes. In *International Conference on Machine Learning*, pages 1301–1309, 2013.