Hawkesian Graphical Event Models

Xiufan Yu XZY22@PSU.EDU
The Pennsylvania State University, University Park, PA, USA

Karthikeyan Shanmugam KARTHIKEYAN.SHANMUGAM2@IBM.COM

Debarun Bhattacharjya DEBARUNB@US.IBM.COM

Tian Gao TGAO@US.IBM.COM

Dharmashankar Subramanian DHARMASH@US.IBM.COM
IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA

Lingzhou Xue LXX6@PSU.EDU
The Pennsylvania State University, University Park, PA, USA

Abstract

Graphical event models (GEMs) provide a framework for graphical representation of multivariate point processes. We propose a class of GEMs named Hawkesian graphical event models (HGEMs) for representing temporal dependencies among different types of events from either a single event stream or multiple independent streams. In our proposed model, the intensity function for an event label is a linear combination of time-shifted kernels where time shifts correspond to prior occurrences of causal event labels in the history, as in a Hawkes process. The number of parameters in our model scales linearly in the number of edges in the graphical model, enabling efficient estimation and inference. This is in contrast to many existing GEMs where the number of parameters scales exponentially in the edges. We use two types of kernels: exponential and Gaussian kernels, and propose a two-step algorithm that combines strengths of both kernels and learns the structure for the underlying graphical model. Experiments on both synthetic and real-world data demonstrate the efficacy of the proposed HGEM, and exhibit expressive power of the two-step learning algorithm in characterizing self-exciting event patterns and reflecting intrinsic Granger-causal relationships.

Keywords: Asymptotic consistency; Event streams; Granger causality; Graphical event models; Hawkes processes; Point processes; Temporal dependencies

1. Introduction

Learning temporal dependencies from streams of different types of events has attracted increasing attention in recent years for a wide range of applications, such as preemptive maintenance for system management (Gunawardana et al., 2011), myocardial infarction prediction for health care (Weiss and Page, 2013), and impact evaluation of social services for social good (Bhattacharjya et al., 2020). An event stream consists of a sequence of timestamps with labels on a common timeline. A label corresponds to an event type and they occur as irregular and asynchronous time arrivals. Exploring the underlying temporal dependencies primarily involves the question: *How does the history of a collection of events statistically affect the arrival time of another event type in the future?* Knowledge of such dependencies can provide data-driven insights for decision-makers to intervene in operations and guide system evolution to achieve a desired goal.

A traditional approach to capture the dynamics of event occurrences is to employ a multivariate point process and use conditional intensity functions. Conditioned on the history of prior event occurrences, the intensities depict the instantaneous rates of future event occurrence. A more challenging task is to model the inter-dependencies of history of multiple event occurrences on the

instantaneous intensities in a compact and interpretable manner. This motivates a graphical representation for the dependence structure in multivariate point processes.

Didelez (2008) introduced local independence graphs – also referred to as graphical event models (GEMs) (Meek, 2014) – to capture dependencies among events, where the intensity function of an event depends only on the history of its parent event labels in the graph. However, in practice it is challenging to consider all possible histories for modeling dependencies. To mitigate this difficulty, Gunawardana et al. (2011) proposed a piece-wise constant intensity model (PCIM) assuming that the intensity function depends only on the projection of histories to pre-determined basis functions. Gunawardana and Meek (2016) slotted the histories of a parent into bins up to a certain maximum time bound in the past and considered a family of models where the multiset of the counts of events in these bins influence the intensity function. Bhattacharjya et al. (2018) simplified by assuming that the intensity function was influenced only by whether or not parent events happened in some recent window, and proposed an algorithm to learn this window without user-provided information.

In this work, we propose a new self-exciting graphical model, named Hawkesian graphical event model (HGEM), to model the temporal dependencies among events given either a single stream of event occurrences or multiple independent streams. One important task in GEMs is to automatically learn the structure of graphical models to find the pattern of dependencies (in the form of a directed graph). Compared with start-of-the-art GEMs, our model greatly reduces the number of parameters that scaled exponentially in the number of edges (like in Bhattacharjya et al. (2018); Gunawardana and Meek (2016)) to a linear scaling. We propose a two-step learning algorithm that first uses exponential kernels to recover the structure and then uses Gaussian kernels to model the intensity functions on learned structure. We prove asymptotic consistency for structure recovery when the true model is a HGEM with exponential kernels. We demonstrate the efficacy of the two-step approach on both synthetic and real datasets comparing with existing baselines.

We consider intensity functions akin to a Hawkes process (Hawkes, 1971). Hawkes models have received considerable interest in many scientific communities such as seismology, criminology, high-frequency financial econometrics, etc. There have been many existing works that studied the estimation of parameters for Hawkes processes (Xu et al., 2016; Chen et al., 2017; Bacry et al., 2020). Our method fundamentally differs from their approach in many aspects. One of the major differences between our work and most prior works in multivariate Hawkes processes is that we take a score-based graph search approach as opposed to penalized matrix estimation for structure learning. In practice, the regularized likelihood optimization requires a carefully chosen tuning parameter to ensure a good performance. In contrast, our approach is entirely data-driven without need of any hyperparameters. Our proposed two-step learning approach combines the strengths of exponential kernels and Gaussian kernels. Xu et al. (2016) also considered Gaussian kernels, but they chose not to vary some parameters of the kernel before optimization, and only a subset of parameters of the kernel function are estimated from the optimization. Our model provides flexibility and automatic learning by parameterizing the kernels. In addition, our model works even when there is only one single event stream available, and we do not require access to identically and independently distributed (i.i.d.) realizations of the process as in Xu et al. (2016).

The rest of this paper is organized as follows. Section 2 sets up the problem framework and presents Hawkesian graphical event models. We then discuss how Granger causality relates to HGEMs. Section 3 introduces our proposed two-step approach for structure learning and parameter learning. Section 4 demonstrates the numerical performance of our proposed models on synthetic datasets as well as on real-world datasets. In Section 5, we conclude and discuss future work.

2. Hawkesian Graphical Event Models (HGEMs)

2.1 Problem Setup

Let $\mathcal{E}$ be a finite set consisting of different types of event labels with cardinality $|\mathcal{E}| = M$. An event stream is denoted by $\mathcal{D} = \{(t_i, e_i)\}_{i=1}^N$ with $t_1 < \cdots < t_N$, where t_i is the time-stamp of the i-th event and e_i is its event label taking values in $\mathcal{E}$. We further use $t_0 = 0$ to denote the initial time and $T \geq t_N$ to denote the end time, and let $\mathcal{I}(\cdot)$ denote the indicator function.

The event stream can be regarded as a multivariate temporal point process which is commonly characterized using conditional intensity functions (Gunawardana et al., 2011). Let $h_t = \{(t_j, e_j) : t_j < t\}$ represent historical events that happened before time t and $N_e(t)$ be a counting process that records the number of type-e events happening before or at time t. We also define the ending time $t_{\text{end}}(h)$ of a event sequence h as the timestamp of the last event in h, that is $t_{\text{end}}(h) = \max(t : (t, e) \in h))$. In this way, $t_{\text{end}}(h_{t_i}) = t_{i-1}$. The conditional intensity function $\lambda_e(t|h_t)$ describes the expected number of type-e events happening in an infinitesimal interval $[t, t + \Delta t]$ given histories of all other event labels prior to time t, that is, $\lambda_e(t|h_t) = \lim_{\Delta t \to 0} E[N_e(t + \Delta t) - N_e(t)|h_t]/\Delta t$. From a practical perspective, $\lambda_e(t|h_t)$ is assumed to be $\lambda_e(t|h_t) = 0$ for $t \leq t_{\text{end}}(h_t)$ and > 0 for $t > t_{\text{end}}(h_t)$. Then, the log-likelihood of data $\mathcal{D}$ is given by:

$$l(\mathcal{D}) = \sum_{e \in \mathcal{E}} \left(\sum_{i=1}^{N} \mathcal{I}\{e_i = e\} \log(\lambda_e(t_i|h_{t_i})) - \sum_{i=1}^{N} \int_{t_{i-1}}^{t_i} \lambda_e(\tau|h_\tau) d\tau - \int_{t_N}^{T} \lambda_e(\tau|h_\tau) d\tau \right). \quad (1)$$

2.2 Model Description

Following the general idea of GEMs, we introduce a directed graph $\mathcal{G} = (\mathcal{E}, \mathcal{A})$ to represent the dependencies among various types of events in the event stream. The nodes $\mathcal{E}$ in the graph correspond to event labels, and the directed arrows $\mathcal{A}$ represent the dependence of one event label's intensity on histories of other events. For each event label $e \in \mathcal{E}$, its conditional intensity $\lambda_e(t|h_t)$ depends only on historical occurrences of its parent events, that is $\lambda_e(t|h_t) = \lambda_e(t|[h_t]_{\text{Pa}(e)})$, where $\text{Pa}(e) \subseteq \mathcal{E}$ are parents of node e in $\mathcal{G}$ and $[h_t]_{\text{Pa}(e)}$ is the history of events whose labels are in the set $\text{Pa}(e)$.

In order to capture the self-exciting pattern among the events, we proposed a Hawkesian Graphical Event Model (HGEM) $\langle \mathcal{G}, \theta_{\mathcal{G}} \rangle$ by assuming that the conditional intensity function $\lambda_e(t|h_t; \theta_{\mathcal{G}})$ follows a multivariate Hawkes process (Hawkes, 1971), i.e., for each event label $e \in \mathcal{E}$,

$$\lambda_e(t|h_t; \theta_{\mathcal{G}}) = \gamma_e + \sum_{k \in \text{Pa}(e)} \sum_{j: t_j < t} \mathcal{I}\{e_j = k\} \alpha_{ek} \nu_{ek}(t - t_j), \quad (2)$$

where $\gamma_e > 0$ provides a base intensity that is independent of history, $\alpha_{ek} > 0$ measures the magnitude of historical influences of type-k events on type-e events, and $\nu_{ek}(\cdot) > 0$ is a function defined on $\mathbb{R}^+$ that captures the pattern of impacts. For ease of notation, we use $\theta_{\mathcal{G}}$ to denote all the model parameters including γ_e, α_{ek} and those contained in $\nu_{ek}(\cdot)$ under a graph structure $\mathcal{G}$.

Before proceeding with the properties of the model, we consider the following regularity assumptions to ensure that a HGEM defined by conditional intensities in (2) is non-explosive, stationary, and identifiable (Hawkes and Oakes, 1974; Eichler et al., 2017).

Assumption 1 (Non-explosivity) *(1) $\gamma_e > 0$ is lower bounded. (2) The kernel function $\nu_{ek}(z) > 0$ is upper bounded for $z > 0$, and $\nu_{ek}(z) = 0$ for $z \leq 0$.*

Assumption 2 (Stationary & Identifiability) *The spectral norm of the matrix $\mathbf{\Phi} = (\phi_{ek})_{M \times M}$ is less than 1, where $\phi_{ek} = \int_0^\infty \alpha_{ek} \nu_{ek}(z) dz$.*

Assumption 3 (Parameter Space) *Let $\Theta_{\mathcal{G}}$ be the parameter space of $\theta_{\mathcal{G}}$ for a given graph $\mathcal{G}$. $\Theta_{\mathcal{G}}$ is nonempty and compact. Further, there exists an open set $\widetilde{\Theta} \supseteq \Theta_{\mathcal{G}}$ such that $\nu_{ek}(\cdot)$ is differentiable with respect to $\theta_{\mathcal{G}}$ in $\widetilde{\Theta}$.*

Two dominant choices of $\nu_{ek}(\cdot)$ are exponential kernels $\nu^E_{ek}(z) = \mathcal{I}\{z > 0\} \exp(-\beta_{ek} z)$, and Gaussian kernels $\nu^G_{ek}(z) = \mathcal{I}\{z > 0\} \exp\left(-\frac{1}{2} w^2_{ek}(z - \mu_{ek})^2\right)$, where β_{ek}, μ_{ek}, $w_{ek} > 0$ for any $e, k \in \mathcal{E}$. In Section 3, we shall discuss the pros and cons of using the two different kernels and propose a two-step procedure which combines the strengths of utilizing both kernels. From now on, we use superscript E and G to represent the log-likelihood function $l(\theta_{\mathcal{G}}; \mathcal{D})$ defined by exponential kernels and Gaussian kernels respectively. To be more specific, we use $l^E(\theta_{\mathcal{G}}; \mathcal{D})$ to denote the log-likelihood function in (1) for $\nu^E_{ek}(z)$ and $l^G(\theta_{\mathcal{G}}; \mathcal{D})$ with $\nu^G_{ek}(z)$.

2.3 Relationship to Granger Causality

Granger causality is a notion that aims to capture temporal dependencies between processes that evolve in time (Granger, 1969). Briefly, if we have a collection of processes V, process $a \in V$ non-Granger causes $b \in V$ if the future of b is independent of the past history of process a given the histories of processes in $V/\{a\}$ (Meek, 2014). Didelez (2008) connected Granger causality with local independence property in GEMs. Meek (2014) further explored assumptions connecting GEMs with causal discovery for point processes, and proposed a new asymmetric graphical separation criterion for directed graphs. Eichler et al. (2017) studied Granger causality of multivariate Hawkes processes. We recall below a key result in this work that establishes the relationship between Granger causality and the conditional intensity functions of various processes.

Lemma 1 (Eichler et al. (2017)) *For a M-dimensional stationary multivariate Hawkes process $N = (N_1, \cdots, N_M)'$, where N_i is a counting process with conditional intensity function*

$$\lambda_i(t|h_t)) = \gamma_i + \sum_{j=1}^{M} \int_0^t \phi_{ij}(u) dN_j(t-u), \tag{3}$$

the process N_j does not Granger-cause N_i with respect to N if and only if $\phi_{ij}(t) = 0$ for all $t > 0$.

Lemma 1 provides an explicit representation of the Granger causal relationships for a multivariate Hawkes process. Connecting the conditional intensity functions of a HGEM model with those defined in (3), for each node in the graph, its intensity function is influenced only by its parent events. As a result, for two event labels $e, k \in \mathcal{E}$, $\phi_{ek}(t) = 0$ for all $t > 0$ if and only if $k \notin \mathrm{Pa}(e)$. Therefore, the graph of the HGEM model $\mathcal{G} = (\mathcal{E}, \mathcal{A})$ is identical to the graph representing Granger causality relations amongst events in $\mathcal{E}$. We summarize this relationship below.

Proposition 1 (Granger-Causality in HGEMs) *For two event labels $e, k \in \mathcal{E}$ in a HGEM $\mathcal{G} = (\mathcal{E}, \mathcal{A})$, label-$k$ event Granger-causes label-e event if and only if k is a parent event of e in $\mathcal{G}$.*

3. Learning HGEMs

3.1 Overview

The learning problem is as follows: given event dataset $\mathcal{D}$, learn HGEM $\langle \mathcal{G}, \theta_{\mathcal{G}} \rangle$, i.e., the structure of the graph, and the conditional intensity parameters for each event label given the structure. One popular approach in point processes literature is via regularized maximum likelihood estimator (MLE) (Xu et al., 2016; Chen et al., 2017; Bacry et al., 2015), which simultaneously achieves parameter learning and structure learning by penalizing some of the coefficients to zeros. However, the

regularization approaches usually require a carefully chosen tuning parameter to control the model complexity. A bad choice of penalty term may resulting in high bias in parameter estimation and severe structure learning errors, leading to spurious edges or miss of important edges.

To avoid the potential issue brought by hyper-parameters, in what follows, we separate the structure learning and parameter learning by defining a score-based criterion to select optimal structure and learning parameters for given structure. To this end, we introduce a two-step learning approach which is specially designed for learning HGEMs. The two-step approach addresses the questions about which kernels to use in HGEMs, and achieves satisfactory performance with respect to both likelihood fitting and structure learning accuracy.

Before proceeding, we first introduce some notation. we use $\subset$ to represent subset relationships between sets and use $\prec$ to represent subset relationships between graphs. For all graph comparisons, we only consider graphs having the same node sets. For two sets S_1 and S_2, $S_1 \subset S_2$ means all elements in S_1 are contained in S_2 and S_2 has at least one element that is not in S_1. For two graphs $\mathcal{G}_1 = (\mathcal{E}, \mathcal{A}_1)$ and $\mathcal{G}_2 = (\mathcal{E}, \mathcal{A}_2)$, $\mathcal{G}_1 \prec \mathcal{G}_2$ means $A_1 \subset \mathcal{A}_2$, in another word, all the arrows in $\mathcal{G}_1$ appear in $\mathcal{G}_2$ and $\mathcal{G}_2$ contains at least one arrow that is not in $\mathcal{G}_1$. We use $\Theta_{\mathcal{G}}$ to denote the parameter space corresponding to $\mathcal{G}$, and use $\theta_{\mathcal{G}} \in \Theta_{\mathcal{G}}$ to denote a parameter contained in $\Theta_{\mathcal{G}}$. Note that for two graphs $\mathcal{G}_1$ and $\mathcal{G}_2$, $\mathcal{G}_1 \prec \mathcal{G}_2$ implies $\Theta_{\mathcal{G}_1} \subset \Theta_{\mathcal{G}_2}$. For a HGEM $\left\langle \mathcal{G}^\star, \theta^\star = \theta^\star_{\mathcal{G}^\star} \right\rangle$, $\mathcal{G}^\star$ and $\theta^\star$ stands for the ground truth graph structure and the ground truth parameters. For any given graph $\mathcal{G}$, $\widehat{\theta}_{\mathcal{G}}$ represent the MLE in parameter space $\Theta_{\mathcal{G}}$, that is

$$\widehat{\theta}_{\mathcal{G}} = \operatorname*{argmax}_{\theta_{\mathcal{G}} \in \Theta_{\mathcal{G}}} l(\theta_{\mathcal{G}}; \mathcal{D}). \tag{4}$$

3.2 Parameter Learning

When the graph $\mathcal{G}$ is known, with the parametric assumptions on the condition intensity functions defined in (2), the log-likelihood of dataset $\mathcal{D}$ given a HGEM can be explicitly written as

$$l(\theta_{\mathcal{G}}; \mathcal{D}) = \sum_{e \in \mathcal{E}} \Big(\sum_{i=1}^{N} \mathcal{I}\{e_j = e\} \log(\lambda_e(t_i | h_{t_i}; \theta_{\mathcal{G}})) - T\gamma_e - \sum_{k \in \mathrm{Pa}(e)} \sum_{j=1}^{N} \mathcal{I}\{e_j = k\} \alpha_{ek} F_{ek}(T - t_j) \Big), \tag{5}$$

where $F_{ek}(t) = \int_0^t \nu_{ek}(z) dz$, and $\lambda_e(t_i | h_{t_i}; \theta_{\mathcal{G}}) = \gamma_e + \sum_{k \in \mathrm{Pa}(e)} \sum_{j=1}^{i-1} \mathcal{I}\{e_j = k\} \alpha_{ek} \nu_{ek}(t_i - t_j)$. Since the log-likelihood function $l(\theta_{\mathcal{G}}; \mathcal{D})$ can be decomposed based on each node in the graph, i.e., $l(\theta_{\mathcal{G}}; \mathcal{D}) := \sum_{e \in \mathcal{E}} l_e(\theta_{\mathcal{G}_e}; \mathcal{D})$, we are able to estimate the parameters separately with respect to each event label. The problem of interest becomes for each $e \in \mathcal{E}$, given structure $\mathcal{G}$, find the MLE $\widehat{\theta}_{\mathcal{G}_e} \in \Theta_{\mathcal{G}}$, such that

$$\widehat{\theta}_{\mathcal{G}_e} = \operatorname*{argmax}_{\theta_{\mathcal{G}_e} \in \Theta_{\mathcal{G}}} l_e(\theta_{\mathcal{G}_e}; \mathcal{D}). \tag{6}$$

Therefore, for given structure $\mathcal{G}$, the MLE for $\theta_{\mathcal{G}}$ is directly obtained by $\widehat{\theta}_{\mathcal{G}} = \{\widehat{\theta}_{\mathcal{G}_e}, e \in \mathcal{E}\}$. We implement a nonlinear augmented Lagrange multiplier method (Ye, 1987; Alexios and Stefan, 2015) to numerically solve the optimization problem in (6).

To investigate the asymptotic behavior of a HGEM, we first consider the summability assumption.

Assumption 4 (Summability) *Both $\alpha_{ek}\nu_{ek}(z)$ and $\alpha_{ek} d\nu_{ek}(z)/dz$ are uniformly summable, such that for a time sequence $\{t_i\}_{i=0}^{\infty}$, where $t_0 = 0$ and $\sum_{i=1}^{\infty} (t_i - t_{i-1})^{-1} < \infty$,*

(i) $\sum_{i=1}^{\infty} (t_i - t_{i-1}) \sup_{x \in (t+t_{i-1}, t+t_i]} \alpha_{ek} \left| \frac{d\nu_{ek}(x)}{dx} \right| < C$ *for some constant $C > 0$ and all $t > 0$.*

(ii) $\sum_{i=1}^{\infty} (t_i - t_{i-1}) \sup_{x \in (t+t_{i-1}, t+t_i]} \alpha_{ek} |\nu_{ek}(x)| \to 0$ *as $t \to \infty$.*

Assumption 4 controls the tail behaviors of kernel function and its first-order deviation (Yang et al., 2017). It is trivially satisfied by exponential kernels. Guo et al. (2018) studied consistency of MLE for multivariate Hawkes processes with exponential kernels. As a graphical extension, the MLE consistency remains valid in HGEMs if the graph under consideration contains the ground truth structure. The results are formally presented in Theorem 2.

Theorem 2 *For a HGEM $\left\langle \mathcal{G}^\star, \theta^\star = \theta^\star_{\mathcal{G}^\star} \right\rangle$ generated by exponential kernels, with Assumptions 1-3, as $T \to \infty$, the MLE $\widehat{\theta}_{\mathcal{G}}$ under graph $\mathcal{G}$ is a consistent estimator for $\theta^\star$ if $\mathcal{G} \succ \mathcal{G}^\star$ or $\mathcal{G} = \mathcal{G}^\star$.*

3.3 Structure Learning

The BIC criterion has been widely advocated for structure learning of graphical models (Xue et al., 2012; Gunawardana and Meek, 2016; Bhattacharjya et al., 2018). For a graph $\mathcal{G}$, the BIC score for a HGEM is defined by

$$BIC_T(\mathcal{G}) = -2\, l(\widehat{\theta}_{\mathcal{G}}; \mathcal{D}) + k \log(T), \tag{7}$$

where k is the number of parameters under structure $\mathcal{G}$, $\widehat{\theta}_{\mathcal{G}}$ is the MLE for given $\mathcal{G}$ and $l(\widehat{\theta}_{\mathcal{G}}; \mathcal{D})$ is the corresponding maximum log-likelihood. Following the same notation as in the log-likelihood functions, we use superscript E and G to represent the BIC scores defined by $l^E(\widehat{\theta}_{\mathcal{G}}; \mathcal{D})$ and $l^G(\widehat{\theta}_{\mathcal{G}}; \mathcal{D})$.

As discussed in Section 3.2, the maximum log-likelihood $l(\widehat{\theta}_{\mathcal{G}}; \mathcal{D})$ is decomposable with respect to each node in the graph. In addition, the model complexity penalty k in (7) can also be decomposed as a summation of the total number of parameters in each node. Therefore, the BIC score defined above can be decomposed into

$$BIC_T(\mathcal{G}) = \sum_{e \in \mathcal{E}} BIC_T(\mathcal{G}_e) = \sum_{e \in \mathcal{E}} \left(-2\, l(\widehat{\theta}_{\mathcal{G}_e}; \mathcal{D}) + k_e \log(T) \right). \tag{8}$$

Remark 3 *k_e is the number of parameters corresponding to event label $e \in \mathcal{E}$. Suppose for a given structure $\mathcal{G}_e$, label e has U_e number of parents, then $k_e = 2U_e + 1$ for exponential kernels while $3U_e + 1$ for Gaussian kernels. In comparison with the state-of-art PGEMs (Bhattacharjya et al., 2018), in which $k_e = 2^{U_e}$, our proposed HGEM's parameter complexity scales only linearly in the sparsity of the graphical model.*

Since there are no constraints like acyclicity of the graph, (8) enables the structure learning to be decomposed into learning individual optimal sub-graphs and then combining them to form the global optimal graph. We would like to search for each node to obtain its parent set with the smallest BIC score. Sharing the spirit with many existing structure learning approaches such as Gao and Wei (2018) and Bhattacharjya et al. (2018), we consider the Forward-Backward Search (FBS) for parent search on each node in the graph. The FBS algorithm for HGEM is presented in Algorithm 1.

Theorem 4 (BIC consistency) *For a HGEM generated by exponential kernels, with Assumptions 1-3, we have $\lim_{T\to\infty} P(BIC^E_T(\mathcal{G}) > BIC^E_T(\mathcal{G}^\star)) = 1$ for any $\mathcal{G} \prec \mathcal{G}^\star$ or $\mathcal{G} \succ \mathcal{G}^\star$.*

3.4 Two-Step Learning Approach

In Sections 3.2 and 3.3, we introduce our approach for parameter and structure learning. It remains to discuss the choice of kernel types. As presented in Section 4, we notice that Gaussian kernels yield a better fit to the data with respect to log-likelihood, yet they do not perform well in structure learning. In contrast, exponential kernels are able to recover the structure well, but do not fit likelihood as well as Gaussian kernels.

Algorithm 1 Forward-Backward Search (FBS) for $e \in \mathcal{E}$

Step 1: Initialization
  parent set $\mathrm{Pa}(e) \leftarrow \varnothing$;
  BIC score $S \leftarrow \infty$;
Step 2: Forward Search
repeat
  for each $k \in \mathcal{E} \setminus \mathrm{Pa}(e)$ **do**
    $\mathrm{Pa}_k(e) \leftarrow \mathrm{Pa}(e) \cup \{k\}$;
    $S_k \leftarrow BIC(\mathcal{G}_e)$ for $\mathrm{Pa}_k(e)$;
  end for
  if $\min_k S_k < S$ **then**
    $\mathrm{Pa}(e) \leftarrow \mathrm{Pa}(e) \cup \{k\}$;
    $S \leftarrow \min_k S_k$;
  end if
until $\min_k S_k \geq S$ **or** $\mathrm{Pa}(e)$ equals $\mathcal{E}$
Step 3: Backward Search
repeat
  for each $k \in \mathrm{Pa}(e)$ **do**
    $\mathrm{Pa}_k(e) \leftarrow \mathrm{Pa}(e) \setminus \{k\}$;
    $S_k \leftarrow BIC(\mathcal{G}_e)$ for $\mathrm{Pa}_k(e)$;
  end for
  if $\min_k S_k < S$ **then**
    $\mathrm{Pa}(e) \leftarrow \mathrm{Pa}(e) \cup \{k\}$;
    $S \leftarrow \min_k S_k$;
  end if
until $\min_k S_k \geq S$ **or** $\mathrm{Pa}(e)$ is empty
Return: parent set for node e: $\mathrm{Pa}(e)$

To borrow strengths from both types of kernels, we propose a two-step learning approach illustrated in Algorithm 2. The main idea is to use exponential kernels defined BIC to perform structure learning, and then use Gaussian kernels defined log-likelihood to fit HGEM parameters. The two stages combine the advantages of exponential kernels in structure learning and Gaussian kernels in model fitting, providing a straightforward approach for learning HGEMs in practice.

Algorithm 2 Learning HGEMs

1: **Step 1:** Use Forward-Backward Search (FBS) to find the graph $\widetilde{\mathcal{G}}$ with the smallest BIC score defined by exponential kernels, i.e., $\widetilde{\mathcal{G}} = \underset{\mathcal{G}}{\operatorname{argmin}}\ BIC_T^E(\mathcal{G})$.
2: **Step 2:** Given the structure $\widetilde{\mathcal{G}}$ obtained from Step 1, find the MLE $\widehat{\theta}_{\widetilde{\mathcal{G}}}$ with respect to the log-likelihood defined by Gaussian kernels, that is, $\widehat{\theta}_{\widetilde{\mathcal{G}}} = \underset{\theta_{\widetilde{\mathcal{G}}} \in \Theta_{\mathcal{G}}}{\operatorname{argmax}}\ l^G(\theta_{\widetilde{\mathcal{G}}}; \mathcal{D})$.
3: **Return:** a HGEM $\{\widetilde{\mathcal{G}}, \widehat{\theta}_{\widetilde{\mathcal{G}}}\}$

4. Experimental Results

We evaluate the proposed model through experiments on synthetic as well as real datasets. Our main baseline is the **proximal graphical event model (PGEM)** (Bhattacharjya et al., 2018). In this model, an event label's conditional intensity rate at any time depends only on whether its parent labels have occurred at least once in some recent window(s). As another baseline, we also include the **piece-wise constant intensity model (PCIM)** (Gunawardana et al., 2011; Parikh et al., 2012), which is more general than PGEM but requires the user to specify a set of basis functions in the form of relevant historical time intervals. This model requires domain knowledge to specify the basis functions; we use our judgment while selecting basis functions for the various datasets considered. Since it was not easily evident how to recover the structure from the implementation, we only report the log-likelihood results for PCIM. We also compare with the **sparse-group-lasso regularized maximum likelihood estimation (MLESGL)** (Xu et al., 2016), which serves as a baseline for Hawkes processes. The implementation is conducted via the THAP package (Xu and Zha, 2017).

4.1 Synthetic Data Experiments

We generate datasets from two different graphs, shown in Figure 1. $\mathcal{G}_1$ contains two nodes $(M = 2)$ and both nodes exhibit self-exciting patterns. The arrow from X_1 to X_2 implies occurrences of type-1 events trigger the intensities of type-2 events. $\mathcal{G}_2$ contains five nodes $(M = 5)$ that are all self-excited but mutually non-Granger causal of each other. Essentially, $\mathcal{G}_2$ is a 5-dimensional Hawkes processes in which the occurrences of each label are not affected by other labels.

For both structures, we generate synthetic datasets from HGEMs with exponential kernels and Gaussian kernels, respectively. Details of parameter settings are relegated to the Appendix. In each setup, we generate 20 event streams. We conduct various learning approaches using the event stream one at a time, and evaluate model performances averaged across the streams. We use negative log-likelihood to evaluate model fitting, and accuracy for structure recovery. We compare performances of six different approaches, including our proposed two-step learned HGEMs, HGEMs with exponential as well as Gaussian kernels. PGEM, PCIM and MLESGL serve as baselines. For ease of notation, we refer to our proposed methods using acronyms as follows: (i) HGEM (T): HGEM fit with two-step approach (Algorithm 2); (ii) HGEM (E): HGEM fit with exponential kernels; (iii) HGEM (G): HGEM fit with Gaussian kernels.

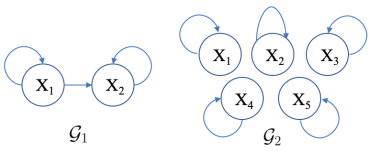

Figure 1: Graphs for Synthetic Datasets

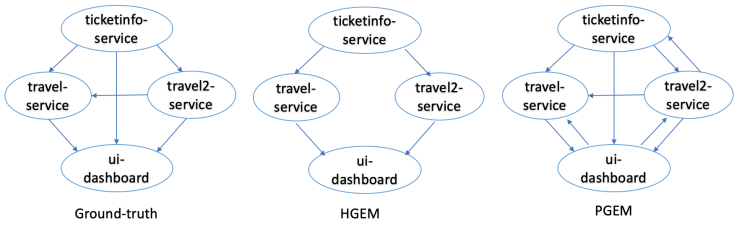

Figure 2: Learned Graphs from Microservice Dataset

Tables 1 reports the negative log-likelihood (neg-LL) and structure learning accuracy to compare the six approaches using four HGEMs generated from the aforementioned data generating process. From the results, we can see that HGEM(E) yields slightly better performance than the baselines with respect to both log-likelihood and structure learning accuracy. HGEM(G) exhibits large improvements in model fitting as it has much smaller negative likelihood than that of HGEM(E) and the baselines. However, it has a large drawback in terms of structure learning accuracy. It is exciting to see that our proposed two-step learned HGEM(T) shows a much better performance compared to the other five methods. As can be seen from the Table, HGEM(T) reveals its ability in achieving a high structure learning accuracy and small negative log-likelihood simultaneously.

	HGEM($\mathcal{G}_1^E$)		HGEM($\mathcal{G}_1^G$)		HGEM($\mathcal{G}_2^E$)		HGEM($\mathcal{G}_2^G$)	
	neg-LL	Accuracy	neg-LL	Accuracy	neg-LL	Accuracy	neg-LL	Accuracy
HGEM (T)	480.33	98.8 %	160.57	97.5 %	634.52	98.0 %	-303.93	99.6 %
HGEM (E)	1787.46	98.8 %	1660.97	97.5 %	2162.79	98.0 %	1796.74	99.6 %
HGEM (G)	422.72	73.8 %	131.11	73.8 %	349.45	75.6 %	-672.76	77.2 %
PGEM	1807.50	95.0 %	1712.04	96.3 %	2195.53	97.6 %	1942.96	99.1 %
PCIM	1818.44	-	1725.46	-	2180.02	-	1929.49	-
MLESGL	1907.88	96.3 %	1633.02	93.8 %	2461.01	94.8 %	1897.66	94.0 %

Table 1: Model Performances on Synthetic Datasets

In summary, the experimental results coincide with our statements in Section 3 that HGEMs with exponential kernels tend to perform well in structure learning while HGEMs using Gaussian kernels tend to achieve better likelihood in modeling fitting. Our propose two-step learning algorithm combines the strength of the two types of kernels in one HGEM. Inherited from the structure learning outcomes of utilizing exponential kernels in the first step, the HGEM(T) retains high accuracy in structure recovery. After obtaining the structure, using Gaussian kernels in the second step helps improve the model fitting with respect to log-likelihood.

4.2 Real Data Experiments

For our real data analysis, we use the same datasets considered in Bhattacharjya et al. (2018). The first involves real-world political event streams from the Integrated Crisis and Early Warning System (ICEWS) dataset, which was constructed by machine-generated event detection over streaming news articles (O'Brien, 2010). ICEWS involves events where an actor performs an action on another actor, for instance 'Police (Mexico) Fight Citizen (Mexico)'. The second dataset includes selected words that are treated as events, from two books in the SPMF data mining library (Fournier-Viger et al., 2014). We ignore the top 100 most frequent words to remove the stop-words and pay attention to the next most frequent M words. Each word is labeled as an event type and its index in the book is encoded as the occurrence time.

	HGEM	PGEM	PCIM
Argentina	5090	6090	5931
Brazil	5392	7047	6605
Colombia	1332	1495	1493
Mexico	2054	2794	2726
Venezuela	1988	2380	2265

Table 2: neg-LL on ICEWS Dataset

	HGEM	PGEM	PCIM
BIBLE (M=10)	65269	72013	72801
BIBLE (M=20)	123990	138254	140327
LEVIATHAN (M=10)	17287	18870	19237
LEVIATHAN (M=20)	32174	35179	36055

Table 3: neg-LL on Books Dataset

Tables 2 and 3 compare the negative log-likelihood for the HGEM (two-step) with the PGEM and PCIM learner baselines. We observe that HGEM fits the data better than the baselines for all datasets in ICEWS, hinting that political event datasets may involve historical dependencies that are more amenable to the spikes and decays of Hawkes-like intensity rates. We also see HGEM fits the data substantially much better on Books dataset.

In addition, we examine the structure learning performance of our proposed approach using the train ticket microservice data (Zhou et al., 2018). Figure 2 plots the learned graph from HGEM and PGEM along with the ground-truth graph. Even though HGEM misses some edges compared with the ground truth, it correctly reflects the Granger-causal relationship. On the contrary, the PGEM gives a lot of spurious edges, which is less desirable in causal analysis.

5. Conclusion and Future Work

In this work, we propose the Hawkesian graphical event model (HGEM), a new class of graphical event models for learning temporal dependencies among different types of events in an event stream. From a modeling perspective, our proposed model captures the self-exciting patterns inherently. Connecting the multivariate Hawkes process with graphical representations, the proposed model provides a more interpretable model to reveal temporal dependencies. More importantly, benefiting from the relationships between the Granger causality and intensity functions in a multivariate Hawkes process, a HGEM acquires the ability of explicitly implying the Granger causal

relationships among event labels in the model. We also propose a two-step algorithm for learning HGEMs. The proposed approach combines the strengths of two popular kernel functions, resulting in substantial improvements in both model fitting and structure learning. In addition, the proposed approach is data-driven, which makes HGEMs practically convenient. We demonstrate the expressive power of HGEMs in model fitting and structure learning on both synthetic and real datasets.

The idea of HGEMs can be further extended to model situations where occurrences of various types of events influence the evolution of a set of state variables which reflect a system's status. The framework of modeling the dynamics of an event-driven system was first introduced by Bhattacharjya et al. (2020). In the future, we plan to extend our work to an "event-state" system to modeling the impacts of events on certain state variables as well as learning the complex temporal dependencies and Granger causality among states.

Acknowledgments

This work was conducted under the auspices of the IBM Science for Social Good initiative. Xiufan Yu and Lingzhou Xue were supported in part by NSF grants DMS–1953189 and CCF–2007823.

Appendix A. Proofs

A.1 Proof of Theorem 2

Proof Guo et al. (2018) studied the MLE estimator of multivariate Hawkes processes with decaying kernels. They proved that under the regularity conditions on parameter space, stationary and identifiability of processes, and summability of decaying kernels, $\widehat{\theta}_{MLE} = \underset{\theta\in\Theta}{\operatorname{argmax}}\, l^E(\theta;\mathcal{D})$ consistently estimates $\theta^\star \in \Theta$ as $T \to \infty$. By definition, a HGEM with exponential kernels naturally satisfies all of their assumptions. For any $\mathcal{G}$ such that $\mathcal{G}^\star \prec \mathcal{G}$ or $\mathcal{G}^\star = \mathcal{G}$, $\theta^\star \in \Theta_{\mathcal{G}^\star} \subseteq \Theta_{\mathcal{G}}$. As a result, $\widehat{\theta}_{\mathcal{G}} = \underset{\theta_{\mathcal{G}}\in\Theta_{\mathcal{G}}}{\operatorname{argmax}}\, l^E(\theta_{\mathcal{G}};\mathcal{D})$ is a consistent estimator of $\theta^\star$ as $T \to \infty$. ■

A.2 Proof of Theorem 4

Proof If $\mathcal{G} \succ \mathcal{G}^\star$, then $\Theta_{\mathcal{G}} \supset \Theta_{\mathcal{G}^\star}$ and $k_{\mathcal{G}} \geq k_{\mathcal{G}^\star}$. Ogata (1978) proved under mild conditions, $l^E(\widehat{\theta}_{\mathcal{G}};D) - l^E(\theta^\star_{\mathcal{G}};D) = O_p(1)$. Note that $\theta^\star_{\mathcal{G}} = \theta^\star_{\mathcal{G}^\star}$ for $\mathcal{G} \succ \mathcal{G}^\star$. Then,

$$P(BIC^E_T(\mathcal{G}) - BIC^E_T(\mathcal{G}^\star) > 0) \geq\ P(O_p(1) + (k_{\mathcal{G}} - k_{\mathcal{G}^\star})\log(T) > 0) \to 1 \quad \text{as } T \to \infty.$$

If $\mathcal{G} \prec \mathcal{G}^\star$, then $\Theta_{\mathcal{G}} \subset \Theta_{\mathcal{G}^\star}$ and $k_{\mathcal{G}} \leq k_{\mathcal{G}^\star}$. Building on results of Ogata (1978), Guo et al. (2018) further proved that, for the true parameter $\theta^\star \in \Theta_{\mathcal{G}^\star}$, an arbitrary open neighborhood U around $\theta^\star$, there exists an $\epsilon > 0$, such that:

$$\lim_{T\to\infty} P\left(\sup_{\theta\in U\subseteq\Theta_{\mathcal{G}^\star}} l^E(\theta;\mathcal{D}) \geq \sup_{\theta\in\Theta_{\mathcal{G}^\star}\setminus U} l^E(\theta;D) + \epsilon T\right) = 1. \tag{9}$$

Since U is an arbitrary neighborhood we choose it as follows: Define $u : \Theta_{\mathcal{G}} \to \Theta_{\mathcal{G}^\star}$ where the lifting function u zero-pads for the parameters that are due to the extra edges in $\mathcal{G}^\star$ but not in $\mathcal{G}$ and for all other edges that are shared, u is an identity function. Since, $\mathcal{G}$ is missing an edge that is in $\mathcal{G}^\star$ associated with non trivial parameters, we have that $\|\theta^\star - u(\widehat{\theta}_{\mathcal{G}})\|_2 > \delta$ for any T. We take a small open neighborhood around $\theta^\star$ that excludes $u(\widehat{\theta}_{\mathcal{G}})$. We have, $\sup_{\theta\in\Theta_{\mathcal{G}^\star}\setminus U} l^E(\theta;D) \geq l^E(u(\widehat{\theta}_{\mathcal{G}});D)$ and $l^E(\widehat{\theta}_{\mathcal{G}^\star};D) \geq \sup_{\theta\in U} l^E(\theta;D)$. Thus, $\{\mathcal{E} : \sup_{\theta\in U\subseteq\Theta_{\mathcal{G}^\star}} l^E(\theta;D) \geq \sup_{\theta\in\Theta_{\mathcal{G}^\star}\setminus U} l^E(\theta;D) + \epsilon T\}$ implies $\{\widetilde{\mathcal{E}} : l^E(\widehat{\theta}_{\mathcal{G}^\star};D) \geq l^E(u(\widehat{\theta}_{\mathcal{G}});D) + \epsilon T\}$. Equation (9) implies that $P(\mathcal{E}) \to 1$. This implies that $P(\tilde{\mathcal{E}}) \to 1$. Hence $P\left(l^E(\widehat{\theta}_{\mathcal{G}^\star};\mathcal{D}) \geq l^E(\widehat{\theta}_{\mathcal{G}};D) + \epsilon T\right) \to 1$. Therefore,

$$P(BIC_T^E(\mathcal{G}) - BIC_T^E(\mathcal{G}^\star) > 0) = P(2(l^E(\widehat{\theta}_{\mathcal{G}^\star}; \mathcal{D}) - l^E(\widehat{\theta}_{\mathcal{G}}; D)) + (k_{\mathcal{G}} - k_{\mathcal{G}^\star}) \log(T) > 0)$$
$$\geq P(\tilde{\mathcal{E}}) P(2\epsilon T - (k_{\mathcal{G}^\star} - k_{\mathcal{G}}) \log(T) > 0) \to 1 \quad \text{as } T \to \infty.$$

The proof of Theorem 2 is complete. ■

Appendix B. Parameters of HGEMs for Synthetic Datasets

HGEM # 1: $\gamma_1 = 0.4$, $\alpha_{11} = 0.2$, $\beta_{11} = 0.8$, $\gamma_2 = 0.5$, $\alpha_{21} = \alpha_{22} = 0.3$, $\beta_{21} = 0.8$, $\beta_{22} = 1$.
HGEM # 2: $\gamma_1 = 0.4$, $\alpha_{11} = 0.2$, $\mu_{11} = 0.2$, $w_{11} = 2$, $\gamma_2 = 0.5$, $\alpha_{21} = \alpha_{22} = 0.3$, $\mu_{21} = 0$, $\mu_{22} = 0.3$, $w_{21} = w_{22} = 1$.
HGEM # 3: $\gamma_1 = \alpha_{11} = 0.25$, $\gamma_2 = \alpha_{22} = 0.30$, $\gamma_3 = \alpha_{33} = 0.35$, $\gamma_4 = \alpha_{44} = 0.40$, $\gamma_5 = \alpha_{55} = 0.45$, $\beta_{11} = \beta_{22} = \beta_{33} = \beta_{44} = \beta_{55} = 1$.
HGEM # 4: $\gamma_1 = \alpha_{11} = \mu_{11} = 0.25$, $\gamma_2 = \alpha_{22} = \mu_{22} = 0.30$, $\gamma_3 = \alpha_{33} = \mu_{33} = 0.35$, $\gamma_4 = \alpha_{44} = \mu_{44} = 0.40$, $\gamma_5 = \alpha_{55} = \mu_{55} = 0.45$, $w_{11} = w_{22} = w_{33} = w_{44} = w_{55} = 1$.

References

G. Alexios and T. Stefan. *Rsolnp: General Non-linear Optimization Using Augmented Lagrange Multiplier Method*, 2015. R package version 1.16.

E. Bacry, I. Mastromatteo, and J.-F. Muzy. Hawkes processes in finance. *Market Microstructure and Liquidity*, 1(01):1550005, 2015.

E. Bacry, M. Bompaire, S. Gaiffas, and J.-F. Muzy. Sparse and low-rank multivariate Hawkes processes. *Journal of Machine Learning Research*, 21:1–32, 2020.

D. Bhattacharjya, D. Subramanian, and T. Gao. Proximal graphical event models. In *Advances in Neural Information Processing Systems*, pages 8136–8145, 2018.

D. Bhattacharjya, K. Shanmugam, T. Gao, N. Mattei, K. R. Varshney, and D. Subramanian. Event-driven continuous time Bayesian networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

S. Chen, D. Witten, and A. Shojaie. Nearly assumptionless screening for the mutually-exciting multivariate Hawkes process. *Electronic Journal of Statistics*, 11(1):1207, 2017.

V. Didelez. 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.

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

P. Fournier-Viger, A. Gomariz, T. Gueniche, A. Soltani, C.-W. Wu, and V. S. Tseng. SPMF: A Java open-source pattern mining library. *The Journal of Machine Learning Research*, 15(1): 3389–3393, 2014.

T. Gao and D. Wei. Parallel Bayesian network structure learning. In *International Conference on Machine Learning*, pages 1685–1694, 2018.

C. W. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969.

A. Gunawardana and C. Meek. Universal models of multivariate temporal point processes. In *Artificial Intelligence and Statistics*, pages 556–563, 2016.

A. Gunawardana, C. Meek, and P. Xu. A model for temporal dependencies in event streams. In *Advances in neural information processing systems*, pages 1962–1970, 2011.

X. Guo, A. Hu, R. Xu, and J. Zhang. Consistency and computation of regularized MLES for multivariate Hawkes processes. *arXiv preprint arXiv:1810.02955*, 2018.

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

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

C. Meek. Toward learning graphical and causal process models. In *Proceedings of the UAI 2014 Conference on Causal Inference: Learning and Prediction-Volume 1274*, pages 43–48. CEUR-WS. org, 2014.

S. P. O'Brien. Crisis early warning and decision support: Contemporary approaches and thoughts on future research. *International Studies Review*, 12:87–104, 2010.

Y. Ogata. The asymptotic behaviour of maximum likelihood estimators for stationary point processes. *Annals of the Institute of Statistical Mathematics*, 30(1):243–261, 1978.

A. P. Parikh, A. Gunawardana, and C. Meek. Conjoint modeling of temporal dependencies in event streams. In *Proceedings of Uncertainty in Artificial Intelligence Workshop on Bayesian Modeling Applications*, August 2012.

J. C. Weiss and D. Page. Forest-based point process for event prediction from electronic health records. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 547–562. Springer, 2013.

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

H. Xu, M. Farajtabar, and H. Zha. Learning Granger causality for Hawkes processes. In *International Conference on Machine Learning*, pages 1717–1726, 2016.

L. Xue, H. Zou, and T. Cai. Nonconcave penalized composite conditional likelihood estimation of sparse ising models. *The Annals of Statistics*, 40(3):1403–1429, 2012.

Y. Yang, J. Etesami, N. He, and N. Kiyavash. Online learning for multivariate Hawkes processes. In *Advances in Neural Information Processing Systems*, pages 4937–4946, 2017.

Y. Ye. *Interior Algorithms for Linear, Quadratic, and Linearly Constrained Non-Linear Programming*. PhD thesis, Department of ESS, Stanford University, 1987.

X. Zhou, X. Peng, T. Xie, J. Sun, C. Xu, C. Ji, and W. Zhao. Poster: Benchmarking microservice systems for software engineering research. In *2018 IEEE/ACM 40th International Conference on Software Engineering: Companion (ICSE-Companion)*, pages 323–324. IEEE, 2018.