Causal Graph Discovery From Self and Mutually Exciting Time Series

Song Wei, Yao Xie, *Member, IEEE*, Christopher S. Josef, and Rishikesan Kamaleswaran

***Abstract*—We present a generalized linear structural causal model, coupled with a novel data-adaptive linear regularization, to recover causal directed acyclic graphs (DAGs) from time series. By leveraging a recently developed stochastic monotone Variational Inequality (VI) formulation, we cast the causal discovery problem as a general convex optimization. Furthermore, we develop a non-asymptotic recovery guarantee and quantifiable uncertainty by solving a linear program to establish confidence intervals for a wide range of non-linear monotone link functions. We validate our theoretical results and show the competitive performance of our method via extensive numerical experiments. Most importantly, we demonstrate the effectiveness of our approach in recovering highly interpretable causal DAGs over Sepsis Associated Derangements (SADs) while achieving comparable prediction performance to powerful "black-box" models such as XGBoost.**

***Index Terms*—Causal structural learning, directed acyclic graph, data-adaptive approach, generalized linear model.**

I. Introduction

CONTINUOUS, automated surveillance systems incorporating machine learning models are becoming increasingly common in healthcare environments. These models can capture temporally dependent changes across multiple patient variables and enhance a clinician's situational awareness by providing an early alarm of an impending adverse event. Among those adverse events, we are particularly interested in sepsis, which is a life-threatening medical condition contributing to one in five deaths globally [1] and stands as one of the most important cases for automated in-hospital surveillance. Recently, many machine learning methods have been developed to predict the onset of sepsis, utilizing electronic medical record (EMR) data [2]. A recent sepsis prediction competition [3] demonstrated the robust performance of XGBoost models [4], [5], [6]; meanwhile, Deep Neural Networks [7] are also commonly used. However, most approaches offer an alert adjudicator very little information pertaining to the reasons for the prediction, leading many to refer to them as "black box" models. Thus, model predictions related to disease identification, particularly for complex diseases, still need to be adjudicated (i.e., interpreted) by a clinician before further action (i.e., treatment) can be initiated. Among the aforementioned works, [6] provided one of the best attempts at identifying causality for their models' predictions by reporting feature importance at a global level for all patients; still, this did not convey which features were most important in arriving at a given prediction for an individual patient. The common lack of interpretability of many clinical models, particularly those related to sepsis, suggests a strong need for principled methods to study the interactions among time series in medical settings.

A natural approach is to model relationships between time series and their effects on sepsis through *Granger causal graphs*. Granger causality assesses whether the history/past of one time series is predictive of another and is a popular notion of causality for time series data. Traditional approaches typically rely on a linear vector autoregressive (VAR) model [8] and consider tests on the VAR coefficients in the bivariate setting. However, it has been recognized that such traditional VAR models have many limitations, including linearity assumption [9] and the absence of directed acyclic graph (DAG) structure, which is essential in causal structural learning [10]. On the one hand, recent advancements in non-linear Granger causality consider Neural Network based approaches coupled with sparsity-inducing penalties [11], [12], but render the optimization problem non-convex. On the other hand, structural vector autoregressive (SVAR) models, which combines the structural causal model (SCM) with the VAR model, leverage DAG-inducing penalties to uncover causal DAGs. Notable contributions include [13], [14], who leveraged adaptive Lasso [15] to recover a Causal DAG, and [16], who applied a recently proposed continuous DAG characterization [17] to encourage such DAG structure. Despite recent advancements, leveraging the well-developed convex optimization techniques to learn a causal DAG remains an open problem. Moreover, the commonly considered DAG structure is less than satisfactory since it cannot capture the lagged self-exciting components, which are important for

Manuscript received 9 January 2023; revised 20 September 2023; accepted 11 December 2023. Date of publication 20 December 2023; date of current version 28 December 2023. This work was supported in part by NSF CAREER under Grant CCF-1650913, Grant NSF DMS-2134037, Grant CMMI-2015787, Grant CMMI-2112533, Grant DMS-1938106, and Grant DMS-1830210; in part by the Coca-Cola Foundation; and in part by the Emory Hospital Grant. *(Corresponding author: Yao Xie.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by NIH Grant, subcontracted through Emory Hospital.

Song Wei and Yao Xie are with the School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: song.wei@gatech.edu; yao.xie@isye.gatech.edu).

Christopher S. Josef is with the Department of Surgery, Emory University School of Medicine, Atlanta, GA 30322 USA (e-mail: cjosef@emory.edu).

Rishikesan Kamaleswaran is with the Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, GA 30322 USA (e-mail: rkamaleswaran@emory.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/JSAIT.2023.3342569, provided by the authors.

Digital Object Identifier 10.1109/JSAIT.2023.3342569

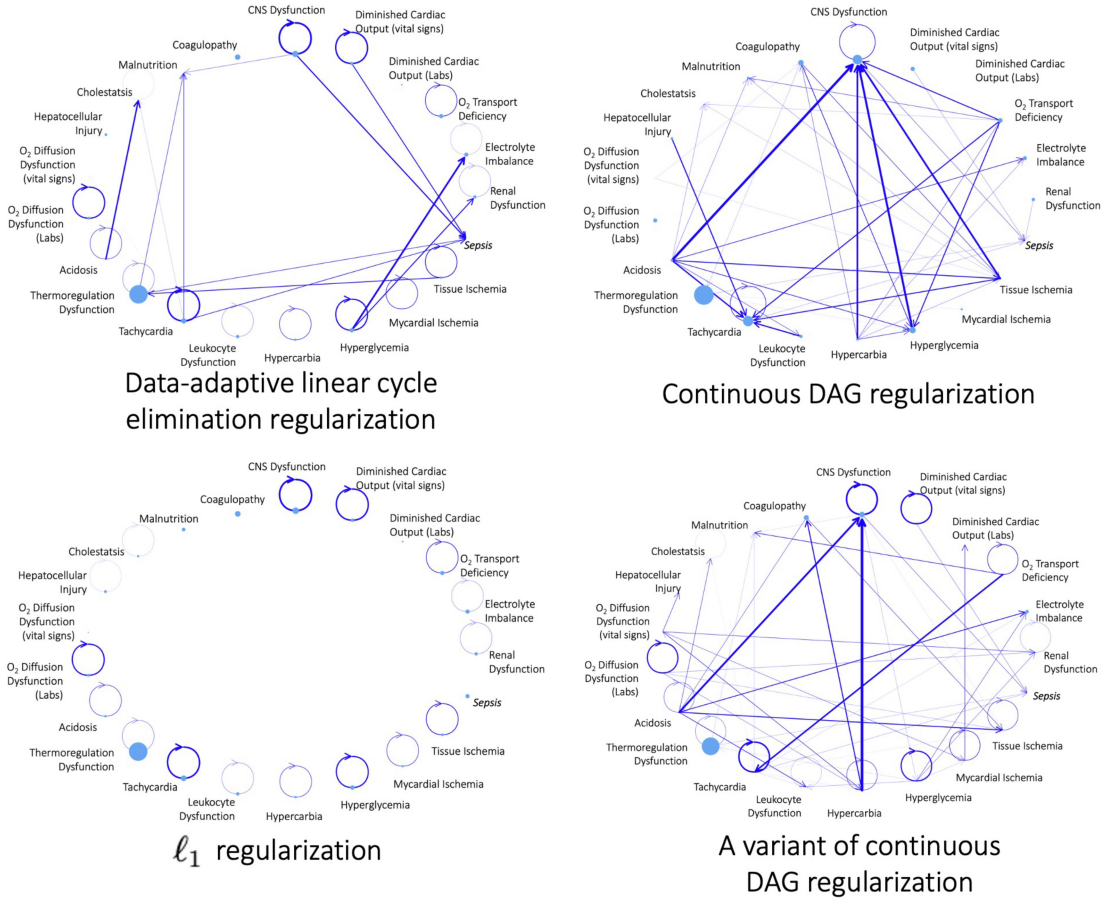

Fig. 1. Causal DAGs for SADs obtained via discrete-time Hawkes network coupled with various types of regularization. The node's size is proportional to the background intensity, and the width of the directed edge is proportional to the exciting effect magnitude. The out-of-sample total CE loss are 2.89 (proposed regularization), 2.75 (ℓ_1 regularization), 3.37 (DAG regularization [17], [20]) and 4.13 (a variant of DAG regularization). Our proposed regularization can help output a DAG with self-exciting edges while achieving the good prediction accuracy; although ℓ_1 regularization achieves the best CE loss, it fails to capture the interactions among SADs, leading to a very uninformative graph.

clinicians to understand how long a node (i.e., a certain type of disease or organ dysfunction) will last once it is triggered.

In this work, we present a generalized linear structural causal model to recover the causal graph from mutually exciting time series, called *discrete-time Hawkes network*. To encourage the desired DAG structure, we propose a novel data-adaptive linear regularizer, enabling us to cast the causal structural learning problem as a convex optimization via a monotone operator Variational Inequality (VI) formulation. Furthermore, performance guarantees are established via recent advances in optimization [18], [19] by developing a non-asymptotic estimation error bound verified by numerical examples. We show the good performance of our proposed method and validate our theoretical findings using extensive numerical experiments. In particular, our real data experiments demonstrate that our proposed method can achieve comparable prediction performance to powerful black-box methods such as XGBoost, while outputting highly interpretable causal DAGs for Sepsis Associated Derangements (SADs), as shown in Figure 1. Although this work only shows the effectiveness of our approach in causal DAG recovery for SADs in medical settings, it can be broadly applicable to other applications.

A. Motivating Application and Dataset

This work is motivated by a real study on Sepsis, which is formally defined as life-threatening organ dysfunction caused by a dysregulated host response to infection [21]. In a recent study of adult sepsis patients [22], each hour of delayed treatment was associated with higher risk-adjusted in-hospital mortality (odds ratio, 1.04 per hour), and thus early recognition of the physiologic aberrations preceding sepsis would afford clinicians more time to intervene and may contribute to improving outcomes. Specifically, we handle a large-scale dataset containing in-hospital EMR derived from the Grady hospital system (an academic, level 1 trauma center located in Atlanta, GA) spanning 2018-2019. The data was collected and analyzed in accordance with Emory Institutional Review Board approved protocol #STUDY00000302. We unitize a retrospective cohort of patients created in our prior work [23], where the patients were included in the Sepsis-3 cohort if they met Sepsis-3 criteria while in the hospital and were admitted for more than 24 hours. The resulting descriptive statistics are provided in Table I. The raw features of each patient include

- *Vital Signs:* In Intensive Care Unit (ICU) environments, vital signs are normally recorded at hourly intervals.

TABLE I
MEDIAN (MED.) AND INTERQUARTILE RANGE (IQR) OF PATIENTS DEMOGRAPHICS. THE UNIT OF TIME MEASUREMENT FOR PATIENT TRAJECTORY LENGTH IS AN HOUR. WE TRUNCATE THE DATA BASED ON EXPERT ADVICE TO ENSURE BOTH PATIENT COHORTS HAVE COMPARABLE SEQUENTIAL ORGAN FAILURE ASSESSMENT (SOFA) SCORES; REFER TO [23] FOR PATIENT COHORT CONSTRUCTION

	Sepsis-3 patients		Non-sepatic patients	
Year	2018 ($n = 409$)	2019 ($n = 454$)	2018 ($n = 960$)	2019 ($n = 1169$) *
Age (Med. and IQR)	58 (38 - 68)	59 (46 - 68)	56 (38 - 67)	55 (37 - 66)
Gender (female percentage)	30.1%	36.6%	37.1%	35.8%
Sofa score (mean)	3.32	3.14	2.18	2.28
Trajectory length (Med. and IQR)	25 (25 - 25)	25 (25 - 25)	17 (13 - 22)	17 (13 - 22)

$\star$ n represents the total number of patients in the corresponding cohort.

However, patients on the floor may only have vital signs measured once every 8 hours.

- *Lab Results:* The laboratory tests are most commonly collected once every 24 hours. However, this collection frequency may change based on the severity of a patient's illness.

The goal is to construct a predictive model for sepsis and an interpretable causal DAG that captures the interactions among those vital signs and Lab results. One difficulty comes from synchronous and continuous-valued observations assumption — in many applications, especially in the medical setting, the observations can be both continuous and categorical-valued. They can also be asynchronous and sampled with different frequencies. For example, vital signs are recorded regularly, whereas Lab results are only ordered when clinically necessary. Since the absence of a lab carries meaning itself, this cannot be simply formulated into a missing data problem. One method to obtain interpretable predictive models is to consider their syndromic nature — there is often a constellation of different physiologic derangements that can combine to create the condition. For example, our prior works [23], [24] leveraged expert opinion to identify the distinct types of measurable, physiologic change that accompanies sepsis-related illness, which is called SADs; see their definition in Table X in Appendix D-A1 in the supplementary material. However, the clinician did not determine the relationship among the SADs; rather, these relationships were an output of model fitting. Although there is a recently proposed principled method to handle such mixed-frequency time series [25], we adopt a similar approach with [23] in this study based on expert advice. In Section VI, we will report the analysis results of this dataset using our proposed method.

B. Literature

We briefly review several closely related areas and defer an extended literature survey to Appendix A in the supplementary material.

1) Causal Structural Learning: Structural causal model-based causal discovery methods often boil down to maximizing a score function within the DAG family [26], making efficient DAG learning fundamental to causal discovery. However, learning DAGs from observational data, i.e., the structural learning problem, is NP-hard due to the combinatorial acyclicity constraint [27]. This motivated many research efforts to find efficient approaches for learning DAGs. Recently, [28] proposed an indicator function-based approach to enumerate and eliminate all possible directed cycles; they used truncate ℓ_1-function as a continuous surrogate of indicator function and proposed to use alternating direction method of multipliers to solve the problem numerically. Later on, [29] followed this approach, transferred indicators into binary variables, and leveraged mixed integer programming to solve the problem. In addition, there are also dynamic programming-based approaches, e.g., [30], but they are not scalable in high dimensions unless coupled with a sparsity-inducing penalty, e.g., $A*$ Lasso [31].

One notable recent contribution in structural learning is [17], who formulated the DAG recovery problem as a constrained continuous optimization via a smooth DAG characterization; they applied an augmented Lagrangian method to transfer constraint as penalty and achieved efficient DAG recovery. Later on, [20] proposed to use the non-convex DAG characterization as a penalty directly and showed an asymptotic recovery guarantee for linear Gaussian models. Other notable extensions along this direction include a discrete backpropagation method, exploration of low-rank structure [32] and neural DAG learning [33], [34], [35]. We refer readers to [36], [37], [38], [39] for systematic surveys on structural learning and causal discovery.

We would like to highlight that the DAG structure with lagged self-exciting components considered in this work is new in the literature. Existing works typically allow directed cycles in the adjacency matrices representing the lagged effects [13], [14], [16], [23], [40]. As a question of science, we believe those lagged cycles are less explainable. For example, our prior work [23] discovered a "Renal Dysfunction $\rightarrow$ O_2 Diffusion Dysfunction $\rightarrow$ Renal Dysfunction" cyclic chain pattern, but we believe the "Renal Dysfunction $\rightarrow$ O_2 Diffusion Dysfunction" coupled with the self-exciting pattern of Renal Dysfunction uncovered by our proposed method here is more convincing; see the bottom right panel in Figure 1.

2) Granger Causality for Time Series: One line of research [8] combines SCM and VAR models and develops the so-called structural vector autoregressive models to help uncover the Granger causal graphs with certain desired structures, such as DAG structure. Notable contributions include [13], [14], who applied adaptive Lasso [15] to encourage the DAG structure. Moreover, following [14], [41] extended the finding that the non-Gaussian measurement noise helps the model identifiability to time series setting; later on, [25] further proved identifiability of SVAR models of order one under arbitrary subsampling and mixed frequency scheme. In addition to adaptive Lasso, there are also other approaches to encouraging DAG structure in the SVAR model, such as the aforementioned continuous DAG characterization [16]. As a comparison, our proposed generalized linear model (GLM) can be reformulated into a *stochastic* SCM by using Gumbel-Max trick/technique [42], [43], [44], [45], [46], which is slightly different from the *deterministic* SCMs with measurement noise in SVAR models [13], [14], [16]. Moreover, compared with the commonly adopted DAG-inducing penalties in SVAR models, e.g., the continuous, differentiable but non-convex

DAG characterization [17], our proposed data-adaptive linear method for structural learning approach is not only convex but also flexible in the sense that it can encourage a DAG structure while keeping lagged self-exciting components.

Another line of research focuses on non-linear Granger causality. Common non-linear approaches consider additive non-linear effects from history that decouple across multiple time series, such as [47], which leveraged a separable matrix-valued kernel to infer the non-linear Granger causality. To further capture the potential non-linear interactions between predictors, Neural Networks coupled with sparsity-inducing penalties are adopted [11], [12]. Even though our GLM can be viewed as a Neural Network without a hidden layer, our model is convex, theoretically grounded, and easy to train, which are the major advantages over Neural Network-based methods. In addition, there are also efforts to tackle the high-dimensionality via regularization, such as group Lasso [48], [49] and nuclear norm regularization [50]. For a comprehensive survey on Granger causality, we refer readers to [9].

C. Notations

We use $\mathbb{R}_+$ to denote the collection of non-negative real numbers, i.e., $\mathbb{R}_+ = [0, \infty)$. For integers $0 < m \leq n$, we denote $[m:n] = \{m, \ldots, n\}$; in a special case where $m = 1$, we denote $[n] = \{1, \ldots, n\}$. Superscript $^{\mathrm{T}}$ denotes vector/matrix transpose; column vectors $\mathbf{1}_d = (1, \ldots, 1)^{\mathrm{T}} \in \mathbb{R}^d$, $\mathbf{0}_d = (0, \ldots, 0)^{\mathrm{T}} \in \mathbb{R}^d$, $e_{i,d} \in \mathbb{R}^d$ is the standard basis vector with its i-th element being one and matrix $I_d \in \mathbb{R}^{d \times d}$ denotes the d-by-d identity matrix; $\mathrm{tr}(e^A)$ stands for the trace of the matrix exponential of matrix A. For vectors $a, b \in \mathbb{R}^d$, the comparison $a \leq b$ is element-wise. In addition, we use ∇ to denote the derivative operator; we use $\langle \cdot, \cdot \rangle$ to denote the standard inner product in Euclidean space, $\| \cdot \|_p$ to denote the vector ℓ_p norm and $\| \cdot \|_F$ to denote the matrix F-norm.

II. Discrete-Time Hawkes Network

A. Set-Up and Background

Consider mixed-type observations over a time horizon $T \geq 1$: we observe d_1 sequences of binary time series $\{y_1^{(i)}, \ldots, y_T^{(i)}\}$, $i \in [d_1]$, which represent d_1 type of events' occurrences, d_2 sequences of continuous-values time series $\{x_1^{(i)}, \ldots, x_T^{(i)}\}$, $i \in [d_2]$, and d_3 static variables $z_1, \ldots, z_{d_3}$. In the following, we will refer to the binary variable as node variables, and our primary goal is to recover the graph structure over those d_1 nodes.

Linear multivariate Hawkes process (MHP) models the mutual inter-dependence among variables by considering a conditional intensity of event occurrence, which is jointly determined by a deterministic background and a self-exciting (or inhibiting) term depending on its history observations. Given that the intensity has a natural interpretation as the instantaneous probability and is inspired by linear MHP with the exponential decaying kernel, we model the probability of occurrence for the i-th node variable, $i \in [d_1]$, at time step $t \in [2:T]$ as follows:

$$\mathbb{P}\Big(y_t^{(i)} = 1 | \mathcal{H}_{t-1}\Big) = \nu_i + \sum_{j=1}^{d_3} \gamma_{ij} z_j + \sum_{k=1}^{t-1} \left(\sum_{j=1}^{d_2} \beta_{ij} x_{t-k}^{(j)} e^{-Rk} + \sum_{j=1}^{d_1} \alpha_{ij} y_{t-k}^{(j)} e^{-Rk} \right), \tag{1}$$

where $\mathcal{H}_t$ denotes the history observation up to time t. To ensure the right-hand side (RHS) of the above equation is a valid probability, we add the following linear constraint:

$$0 \leq \nu_i + \sum_{j=1}^{d_3} \gamma_{ij} z_j + \sum_{k=1}^{t-1} \left(\sum_{j=1}^{d_2} \beta_{ij} x_{t-k}^{(j)} e^{-Rk} + \sum_{j=1}^{d_1} \alpha_{ij} y_{t-k}^{(j)} e^{-Rk} \right) \leq 1.$$

For the i-th node variable, $\gamma_{ij} \in \mathbb{R}$ represents the influence from j-th static variable and contributes to the deterministic background intensity together with $\nu_i \in \mathbb{R}_+$; parameter $\alpha_{ij} \in \mathbb{R}$ (or $\beta_{ij} \in \mathbb{R}$) represents the magnitude of the influence from the j-th node variable (or continuous variable) to the i-th node variable, which decays exponentially fast with exponent characterized by $R > 0$ — those parameter interpretations connect (1) with the conditional intensity function of the MHP, e.g., [23]. Moreover, one advantage of the above model is that, as long as the above linear constraint is satisfied, we do not restrict α_{ij} or β_{ij} to be non-negative, meaning that our model can handle both triggering and inhibiting effects.

The matrix $A = (\alpha_{ij}) \in \mathbb{R}^{d_1 \times d_1}$ defines a weighted directed graph $\mathcal{G}(A) = (\mathcal{V}, \mathcal{E})$ on d_1 nodes in the following way: $\mathcal{V}$ is the collection of aforementioned d_1 binary node variables; let $\mathcal{A} \in \{0,1\}^{d_1 \times d_1}$ such that $\mathcal{A}_{ij} = 1$ if $\alpha_{ij} \neq 0$ and zero otherwise, then $\mathcal{A}$ defines the adjacency matrix of a directed graph $\mathcal{G}(A)$, which gives the collection of directed edges $\mathcal{E}$; the weights of the directed edges in $\mathcal{E}$ are defined accordingly by matrix A. In a slight abuse of notation, we will call A the (weighted) adjacency matrix of the graph.

B. Linear Model

One drawback of MHP comes from its scalability; to be precise, considering complete history leads to quadratic complexity with respect to (w.r.t.) the number of events. Since the triggering (or inhibiting) effects from the history observations decay exponentially fast, we typically consider finite memory depth. Similarly, in our discrete-time Hawkes network, we make reasonable simplification by assuming *finite memory depth* $\tau \geq 1$ for both continuous and binary observations. More specifically, consider given history at time $t \in [1-\tau:0]$. At time $t \in [T]$, we use $w_{t-\tau:t-1}$ to denote the observations from $t-\tau$ to $t-1$:

$$w_{t-\tau:t-1} = \Big(1, z_1, \ldots, z_{d_3}, x_{t-1}^{(1)}, \ldots, x_{t-\tau}^{(1)}, \ldots, x_{t-1}^{(d_2)}, \ldots, x_{t-\tau}^{(d_2)}, y_{t-1}^{(1)}, \ldots, y_{t-\tau}^{(1)}, \ldots, y_{t-1}^{(d_1)}, \ldots, y_{t-\tau}^{(d_1)}\Big)^{\mathrm{T}}.$$

To ease the estimation burden, let $\alpha_{ijk} = \alpha_{ij} \exp\{-Rk\}$ and $\beta_{ijk} = \beta_{ij} \exp\{-Rk\}$; in fact, this re-parameterization gives

our model more flexibility and expressiveness. Now, we can rewrite (1) as follows:

$$\mathbb{P}\Big(y_t^{(i)} = 1 \Big| w_{t-\tau:t-1}\Big) = w_{t-\tau:t-1}^{\mathrm{T}}\theta_i,$$
$$\theta_i \in \Theta = \{\theta : 0 \le w_{t-\tau:t-1}^{\mathrm{T}}\theta \le 1,\ t \in [T]\} \subset \mathbb{R}^d, \quad (2)$$

where $d = 1 + d_3 + \tau d_2 + \tau d_1$ is the dimentionality, Θ is the feasible region, and θ_i is the model parameter:

$$\theta_i = \big(\nu_i, \gamma_{i1}, \ldots, \gamma_{id_3}, \beta_{i11}, \ldots, \beta_{i1\tau}, \ldots, \beta_{id_21}, \ldots, \beta_{id_2\tau},$$
$$\alpha_{i11}, \ldots, \alpha_{i1\tau}, \ldots, \alpha_{id_11}, \ldots, \alpha_{id_1\tau}\big)^{\mathrm{T}}.$$

This parameter summarizes the influence from all variables to the i-th node. Before we move on, we want to briefly mention that, as a special case of the GLM, (2) can also be re-parameterized into a causal structural model and its parameters $A_k = (\alpha_{ijk}) \in \mathbb{R}^{d_1 \times d_1}$, $k \in [\tau]$, can be taken as causal graphs under the no unobserved confounder assumption. We will elaborate on these in Section II-C.

1) Estimation: We leverage a recently developed technique [18], [19], which estimates the model parameters by solving stochastic monotone VI, to develop a statistically principled estimator for discrete-time Hawkes network. To be precise, in our linear model (2), for $i \in [d_1]$, we use the weak solution to the following VI as the estimator $\hat{\theta}_i$:

$$\text{find } \hat{\theta}_i \in \Theta : \langle \bar{F}_T^{(i)}(\theta_i), \theta_i - \hat{\theta}_i\rangle \ge 0,\ \forall \theta_i \in \Theta, \qquad \mathrm{VI}[\bar{F}_T^{(i)}, \Theta]$$

where $\bar{F}_T^{(i)}(\theta_i)$ is the empirical vector field defined as follows:

$$\bar{F}_T^{(i)}(\theta_i) = \frac{1}{T}\sum_{t=1}^{T} w_{t-\tau:t-1} w_{t-\tau:t-1}^{\mathrm{T}}\theta_i - \frac{1}{T}\sum_{t=1}^{T} w_{t-\tau:t-1} y_t^{(i)}$$
$$= \mathbb{W}_{1:T}\theta_i - \frac{1}{T}\sum_{t=1}^{T} w_{t-\tau:t-1} y_t^{(i)}, \quad (3)$$

and

$$\mathbb{W}_{1:T} = \frac{1}{T}\sum_{t=1}^{T} w_{t-\tau:t-1} w_{t-\tau:t-1}^{\mathrm{T}} \in \mathbb{R}^{d\times d}. \quad (4)$$

2) Connection to Least Square Estimator: One important observation is that the vector field $\bar{F}_T^{(i)}(\theta_i)$ (3) is indeed the gradient field of the Least Square (LS) objective, meaning that the weak solution to the corresponding VI solves the following LS problem [19]:

$$\min_{\theta_i} \quad \frac{1}{2T}\|\mathbf{w}_{1:T}^{\mathrm{T}}\theta_i - Y_{1:T}^{(i)}\|_2^2,$$
$$\text{s.t.} \quad \mathbf{0}_T \le \mathbf{w}_{1:T}^{\mathrm{T}}\theta_i \le \mathbf{1}_T, \quad (5)$$

where

$$\mathbf{w}_{1:T} = (w_{1-\tau:0}, \ldots, w_{T-\tau:T-1}) \in \mathbb{R}^{d\times T},$$
$$Y_{1:T}^{(i)} = \Big(y_1^{(i)}, \ldots, y_T^{(i)}\Big)^{\mathrm{T}} \in \mathbb{R}^T. \quad (6)$$

One approach to solve (5) is to leverage the well-developed convex optimization tools, such as `CVX` [51] and `Mosek` [52]. An alternative approach is through projected gradient descent (PGD), where the empirical vector field (3) is treated as the gradient field. To be precise, we introduce dual variables $\eta_1 = (\eta_{1,1}, \ldots, \eta_{1,T})^{\mathrm{T}} \in \mathbb{R}_+^T$, $\eta_2 = (\eta_{2,1}, \ldots, \eta_{2,T})^{\mathrm{T}} \in \mathbb{R}_+^T$ and the Lagrangian is given by:

$$L(\theta_i, \eta_1, \eta_2)$$
$$= \frac{1}{2T}\|\mathbf{w}_{1:T}^{\mathrm{T}}\theta_i - Y_{1:T}^{(i)}\|_2^2 + \eta_1^{\mathrm{T}}\big(\mathbf{w}_{1:T}^{\mathrm{T}}\theta_i - \mathbf{1}_T\big) - \eta_2^{\mathrm{T}}\mathbf{w}_{1:T}^{\mathrm{T}}\theta_i.$$

The Lagrangian dual function is $\min_{\theta_i} L(\theta_i, \eta_1, \eta_2)$. As we can see, the Lagrangian above is convex w.r.t. θ_i. By setting $\nabla_{\theta_i} L(\theta_i, \eta_1, \eta_2) = 0$, we have

$$\hat{\theta}_i(\eta_1, \eta_2) = \frac{1}{T}\mathbb{W}_{1:T}^{-1}\Big(\mathbf{w}_{1:T} Y_{1:T}^{(i)}/T - \eta_1 + \eta_2\Big).$$

As pointed out in [19], $\mathbb{W}_{1:T} \in \mathbb{R}^{d\times d}$ (4) will be full rank (and thus invertible) with high probability when T is sufficiently large. By plugging $\hat{\theta}_i(\eta_1, \eta_2)$ into the Lagrangian, we give the dual problem as follows:

$$\max_{\eta_1,\eta_2} \quad L\big(\hat{\theta}_i(\eta_1, \eta_2), \eta_1, \eta_2\big),$$
$$\text{s.t.} \quad \eta_1, \eta_2 \ge \mathbf{0}_T.$$

This problem can be solved by PGD as its projection step simply changes all negative entries to zeros.

C. Generalized Linear Model

As mentioned earlier, the linear assumption is restrictive, and therefore, we consider the following GLM [53] to enhance its expressiveness:

$$\mathbb{P}\Big(y_t^{(i)} = 1 \Big| w_{t-\tau:t-1}\Big) = g\big(w_{t-\tau:t-1}^{\mathrm{T}}\theta_i\big), \quad \theta_i \in \Theta, \quad (7)$$

where $g : \mathbb{R} \to [0, 1]$ is a monotone *link function*. For example, it can be non-linear, such as sigmoid link function $g(x) = 1/(1 + e^{-x})$ on a domain $x \in \mathbb{R}$ and exponential link function $g(x) = 1 - e^{-x}$ on a domain $x \in \mathbb{R}_+$; also, it can be linear $g(x) = x$ on a domain $x \in [0, 1]$, which reduces our GLM (7) to the linear model (2). The feasible region Θ will vary based on the choice of link functions, and we will see two examples later in Section II-C2.

1) Structural Causal Model: One key feature that distinguishes our discrete-time Hawkes network from the existing black-box method is the causal graph under Pearl's framework [10] encoded in the GLM parameters. To be precise, one can uncover the connection between the GLM (7) and the stochastic SCM via the Gumbel-Max technique [42], [43]: Let us denote

$$p_t^{(i)}(1) = \mathbb{P}\Big(y_t^{(i)} = 1 \Big| w_{t-\tau:t-1}\Big) = g\big(w_{t-\tau:t-1}^{\mathrm{T}}\theta_i\big),$$
$$p_t^{(i)}(0) = 1 - p_t^{(i)}(1).$$

Then, our GLM (7) can be reformulated into an SCM as follows:

$$y_t^{(i)} = \arg\max_{y\in\{0,1\}} \Big(\log\Big(p_t^{(i)}(y)\Big) + \epsilon_t^{(i)}\Big), \quad (8)$$

where $\epsilon_t^{(i)}$ is a Gumbel r.v., i.e., $\epsilon_t^{(i)} \sim \mathrm{Gumbel}(0, 1)$. The Gumbel-Max technique tells us that the SCM (8) is equivalent to our GLM (7) in that we still have $\mathbb{P}(y_t^{(i)} = 1 | w_{t-\tau:t-1}) = g(w_{t-\tau:t-1}^{\mathrm{T}}\theta_i)$. Therefore, under standard conditions that there is *no unobserved confounding*, one can see that the adjacency

matrices $A_k = (\alpha_{ijk}) \in \mathbb{R}^{d_1 \times d_1}$, $k \in [\tau]$, represent the causal graph structure over d_1 nodes.

Remark 1 (Connection to Granger Causality): One may find a very close connection between our causal graph with Granger causality in the non-linear autoregressive model [11]; See Appendix B-A in the supplementary material for further details on Granger causality. The key difference is whether or not there is unmeasured confounding: Those two causality notions will overlap in a world where there are no potential causes. However, this is not a very likely setting and a fundamentally untestable one [54]. To understand this, the argument that "Christmas trees sales Granger-cause Christmas" will not hold once one knows that Christmas took place on December 25th for centuries, which can be modeled as a confounding variable that causes both Christmas tree sales and Christmas itself.

2) Estimation With Variational Inequality: Similar to VI[$\bar{F}_T^{(i)}, \Theta$] for the linear model, we use the weak solution to the following VI as the estimator for our GLM (7):

$$\text{find } \hat{\theta}_i \in \Theta : \langle F_T^{(i)}(\theta_i), \theta_i - \hat{\theta}_i \rangle \geq 0, \ \forall \theta_i \in \Theta, \qquad \text{VI}[F_T^{(i)}, \Theta]$$

Parameter θ_i is constrained in a convex set $\Theta \subset \mathbb{R}^d$, which may vary with different non-linear links; we will see two examples later. The main difference from VI[$\bar{F}_T^{(i)}, \Theta$] is the empirical vector field, which is defined as follows:

$$F_T^{(i)}(\theta_i) = \frac{1}{T} \sum_{t=1}^{T} w_{t-\tau:t-1} \Big(g\big(w_{t-\tau:t-1}^{\mathrm{T}} \theta_i\big) - y_t^{(i)} \Big). \tag{9}$$

As we can see, the definition above covers that of $\bar{F}_T^{(i)}$ (3) for linear link case; thus, we will use $F_T^{(i)}$ to denote the empirical vector field for all monotone links in the following. Furthermore, the statistical inference for each node can be *decoupled*, and thus we can perform parallel estimation and simplify the analysis.

The intuition behind this VI-based method is straightforward. Let us consider the global counterpart of the above vector field, whose root is the unknown ground truth $\theta_i^\star$, i.e.,

$$\begin{aligned} F^{(i)}(\theta_i) &= \mathbb{E}_{(w, y^{(i)})} \Big[w \Big(g\big(w^{\mathrm{T}} \theta_i\big) - y^{(i)} \Big) \Big] \\ &= \mathbb{E}_{(w, y^{(i)})} \big[w \big(g\big(w^{\mathrm{T}} \theta_i\big) - g\big(w^{\mathrm{T}} \theta_i^\star\big) \big) \big]. \end{aligned}$$

Although we cannot access this global counterpart, by solving the empirical one VI[$F_T^{(i)}, \Theta$] we could approximate the ground truth very well. We will show how well this approximation can be in Section IV.

Remark 2 (Comparison With the Original Work): As a generalization of the VI-based estimator for binary time series [19], our method can handle mix-type data (i.e., binary and continuous-valued time series and static variables). Furthermore, we show how to leverage regularization in the VI-based estimation framework as well as extend the performance guarantee to general non-linear monotone link functions, on which we will elaborate in Sections III and IV, respectively.

3) Examples for Non-Linear Link Function: Now, we will give two examples of general non-linear monotone links and briefly discuss how to numerically obtain our proposed estimator. Note that the equivalence between our proposed estimator and LS estimator only holds for linear link function since the gradient field of LS objective with general link function will be:

$$\begin{aligned} &\frac{1}{T} \sum_{t=1}^{T} \nabla g\big(w_{t-\tau:t-1}^{\mathrm{T}} \theta_i\big) \Big(g\big(w_{t-\tau:t-1}^{\mathrm{T}} \theta_i\big) - y_t^{(i)} \Big) \\ &\quad = \frac{1}{T} \sum_{t=1}^{T} w_{t-\tau:t-1} g'\big(w_{t-\tau:t-1}^{\mathrm{T}} \theta_i\big) \Big(g\big(w_{t-\tau:t-1}^{\mathrm{T}} \theta_i\big) - y_t^{(i)} \Big), \end{aligned}$$

where g' is the derivative of g. However, in the sigmoid link function case, our proposed estimator reduces to the Maximum Likelihood (ML) estimator for the logistic regression. To be precise, we can show that the empirical vector filed (9) is the gradient field of the objective function of the following ML problem:

$$\begin{aligned} \max_{\theta_i} \quad & \frac{1}{T} \sum_{t=1}^{T} y_t^{(i)} \log g\big(w_{t-\tau:t-1}^{\mathrm{T}} \theta_i\big) \\ & + \Big(1 - y_t^{(i)}\Big) \log\big(1 - g\big(w_{t-\tau:t-1}^{\mathrm{T}} \theta_i\big)\big). \end{aligned}$$

Again, this equivalence between our proposed estimator and ML estimator comes from the fact that $g'(x) = g(x)(1 - g(x))$ for the sigmoid link function and does not hold for other non-linear link functions. One advantage of the sigmoid link function is that we do not need to put additional constraints on the parameter θ_i to ensure $g(w_{t-\tau:t-1}^{\mathrm{T}} \theta_i)$ is a reasonable probability, i.e., the feasible region is $\Theta = \mathbb{R}^d$. To numerically obtain such our proposed estimator, we can use vanilla gradient descent (GD), where the gradient is the empirical vector field (9).

Another non-linear example is the exponential link $g(x) = 1 - e^{-x}$, $x \in \mathbb{R}_+$. Similar to the linear link case, to ensure valid probability, the feasible region is $\Theta = \{\theta : w_{t-\tau:t-1}^{\mathrm{T}} \theta \geq 0, \ t \in [T]\}$. To numerically solve for our proposed estimator, we can again perform PGD on the Lagrangian dual problem. Alternatively, in many real applications where we have prior knowledge that we do not consider inhibiting effect, i.e., the feasible region is $\theta_i \in \mathbb{R}_+^d \subset \Theta$, we can perform PGD on the primal problem.

For general non-linear links, PGD is the most sensible approach to obtain our proposed estimator. However, due the serial correlation in the data, we cannot conduct theoretical convergence analysis as [18] did. Later, we will use numerical simulation to demonstrate the good performance of PGD for all three aforementioned link functions.

III. Data-Adaptive Convex Structural Learning

In causal structural learning [10], it is often of great interest to recover a DAG from observational data. In our analysis, we want a DAG-like structure that additionally keeps the lagged self-exciting components, i.e., length-1 cycles. This is because a stronger self-exciting effect informs the adjudicator that the corresponding node/event can last for a longer time once

triggered. Therefore, our goal is to remove the less explainable *directed cycles with lengths greater than or equal to two* (referred to as *cycles* for brevity) while keeping lagged self-exciting components to improve the result interpretability.

A. Estimation With Data-Adaptive Linear Constraints

1) Existing Characterizations of Acyclicity: The estimation of a DAG structure is challenging due to the combinatorial nature of the acyclicity constraint. One seminal work [17] characterized the acyclicity constraint via the following continuous and differentiable constraint: We consider memory depth $\tau = 1$ and denote $\alpha_{ij} = \alpha_{ij1}$, $A = (\alpha_{ij})$ for brevity (general $\tau \geq 1$ case will be presented later in this subsection); for non-negative weighted adjacency matrix $A \in A \in \mathbb{R}_+^{d_1 \times d_1}$, its induced graph is a DAG if and only if

$$h(A) = \operatorname{tr}\left(e^A\right) - d_1 = \sum_{L=1}^{\infty} \frac{\operatorname{tr}(A^L)}{L!} = 0. \tag{10}$$

The above DAG characterization can be understood as follows: For $A \in \mathbb{R}_+^{d_1 \times d_1}$, $\operatorname{tr}(A^L) \geq 0$, and it will be zero if and only if there does not exist any length-L directed cycle in the induced graph; if $h(A) = 0$, then $\operatorname{tr}(A^L) = 0$ for all $L \geq 1$, implying the induced graph is a DAG.

Intuitively, cycles with length $L \geq d_1$ do not contribute to the DAG characterization, and thus one can truncate the infinite series (10) [33]. Indeed, one can always apply topological ordering to get a lower triangle adjacency matrix $\tilde{A}$ for a DAG, which is nilpotent such that $\tilde{A}^{d_1} = 0$; such a topological reorder of nodes corresponds to applying permutation matrix P to the original adjacency matrix A, i.e., $\tilde{A} = PAP^{\mathrm{T}}$, and one still has $A^{d_1} = 0$ (since a permutation matrix satisfies $P^{-1} = P^{\mathrm{T}}$); see [55, Proposition 1] for an equivalent characterization of DAG as $A^{d_1} = 0$. Quite contrary to the work by [56] which put emphasis on long cycles, [55] proposed to truncate the series (10) to

$$h_{\text{trunc}}(A) = \sum_{L=1}^{k} \operatorname{tr}\left(A^L\right) = 0, \tag{11}$$

where $k < d_1$ since they observed that "higher-order terms that are close to zero".

2) Motivation: Following [55], we propose to apply "soft" linear constraint to encourage acyclicity while maintaining the convexity. Specifically, for $L \in [2:k]$, we relax the strict characterization $\operatorname{tr}(A^L) = 0$ by constraining the weighted sum for all possible length-L cycles: for $i_L \to i_{L-1} \to \cdots \to i_1 \to i_L$:

$$\alpha_{i_1 i_2} + \alpha_{i_2 i_3} + \cdots + \alpha_{i_{L-1} i_L} + \alpha_{i_L i_1} \leq \delta. \tag{12}$$

Notice that we do not put constraint on $L = 1$ case since the self-exciting effects carry meaning and are desirable in our analysis.

One simple estimation method would be to include the above linear constraints into the feasible region, which will not change the convexity since the intersection of two convex sets is still convex. However, the number of linear constraints will be on the order of d_1^k, and the constraint hyperparameter $\delta \geq 0$ should also vary for different length-L cycles, depending on the ground truth weight parameters in the corresponding cycle. Fortunately, due to the consistency result (to be presented in the next section), we can address the above issues by obtaining data-adaptive linear constraints. To be precise, as illustrated in Figure 2, the recovery guarantee implies that the existence of edge $3 \to 2$ might be a result of noise in the observations, and the VI solution tends to output such a false discovery. Therefore, one can simply add the following data-adaptive constraints

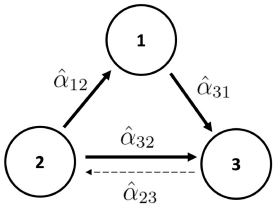

Fig. 2. Illustration of the estimated graph without regularization. Existence of an edge represents the corresponding estimated weight is greater than zero and dashed edge $3 \to 2$ indicates that its weight is very small, which could be a result of noisy observations.

$$\alpha_{23} + \alpha_{32} \leq \hat{\alpha}_{32}, \quad \alpha_{12} + \alpha_{23} + \alpha_{31} \leq \hat{\alpha}_{12} + \hat{\alpha}_{31},$$

into the feasible region when solving the VI. Moreover, in the above illustrative example, oftentimes imposing the first constraint suffices to remove edge $3 \to 2$, meaning that removing short cycles suffices to remove long cycles; in both our simulation and real experiments, we have demonstrated that it suffices to consider $k = 3$ in moderate-sized graph (around 20 nodes) setting; indeed, we only discover short cycles in our real data example (see Figure 5). We will formally state our method with $k = 3$ in the following.

3) Proposed Constraint: Consider the causal graphs induced by the estimated adjacency matrices $\hat{A}_\ell = (\hat{\alpha}_{ij\ell}) \in \mathbb{R}^{d_1 \times d_1}$, $\ell \in [\tau]$, using the VI-based estimator $\mathrm{VI}[F_T^{(i)}, \Theta]$. As mentioned earlier, cycles in those graphs are undesirable, and we want to remove them. Let us begin with formally defining cycles: for positive integer $L \geq 2$, if there exist $\ell \in [\tau]$ and mutually different indices $i_1, \ldots, i_L \in [d_1]$ such that

$$\hat{\alpha}_{i_1 i_L \ell} > 0, \quad \hat{\alpha}_{i_{k+1} i_k \ell} > 0, \quad k \in [L-1],$$

then we say there exists a length-L (directed) cycle in the directed graphs induced by $\hat{A}_\ell$'s.

We consider all possible length-2 and length-3 cycles in those graphs, whose indices are as follows:

$$I_{2,\ell} = \left\{(i,j) : i \neq j,\ \hat{\alpha}_{ij\ell} > 0,\ \hat{\alpha}_{ji\ell} > 0\right\}, \quad \ell \in [\tau],$$

$$I_{3,\ell} = \left\{(i,j,k) : i,j,k \text{ mutually different},\ \hat{\alpha}_{ij\ell} > 0,\ \hat{\alpha}_{jk\ell} > 0,\ \hat{\alpha}_{ki\ell} > 0\right\}, \quad \ell \in [\tau].$$

As illustrated in Figure 2, in each cycle, the edge with the least weight could be caused by noisy observation, meaning that we should remove such edges to eliminate the corresponding cycle. To do so, we impose the following *data-adaptive linear*

cycle elimination constraints to shrink the weights of those "least important edges":

$$\begin{aligned}\alpha_{ij\ell}+\alpha_{ji\ell}&\le\delta_{2,\ell}(i,j),\quad (i,j)\in I_{2,\ell},\quad \ell\in[\tau],\\ \alpha_{ij\ell}+\alpha_{jk\ell}+\alpha_{ki\ell}&\le\delta_{3,\ell}(i,j,k),\quad (i,j,k)\in I_{3,\ell},\quad \ell\in[\tau],\end{aligned}\tag{13}$$

where, for $\ell\in[\tau]$, the *data-adaptive regularization strength parameters* $\delta_{2,\ell}(i,j),(i,j)\in I_{2,\ell}$ and $\delta_{3,\ell}(i,j,k),(i,j,k)\in I_{3,\ell}$ are defined as follows:

$$\begin{aligned}\delta_{2,\ell}(i,j)&=\hat\alpha_{ij\ell}+\hat\alpha_{ji\ell}-\min\{\hat\alpha_{ij\ell},\hat\alpha_{ji\ell}\}=\max\{\hat\alpha_{ij\ell},\hat\alpha_{ji\ell}\},\\ \delta_{3,\ell}(i,j,k)&=\hat\alpha_{ij\ell}+\hat\alpha_{jk\ell}+\hat\alpha_{ki\ell}-\min\{\hat\alpha_{ij\ell},\hat\alpha_{jk\ell},\hat\alpha_{ki\ell}\}.\end{aligned}\tag{14}$$

4) Constrained Joint VI-Based Estimation: Different from the aforementioned decoupled learning approach in Section II-C2, here we need to estimate parameters $\theta_1,\dots,\theta_{d_1}$ jointly to remove cycles and encourage our desired DAG structure. We concatenate the parameter vectors into a matrix, i.e., $\theta=(\theta_1,\dots,\theta_{d_1})\in\mathbb{R}^{d\times d_1}$, and the feasible region of the concatenated parameter is then defined as:

$$\tilde\Theta=\{\theta=(\theta_1,\dots,\theta_{d_1}):\theta_i\in\Theta,\ i\in[d_1]\}.\tag{15}$$

The joint estimator coupled with the data-adaptive linear cycle elimination constraint is defined as the weak solution to the following VI:

$$\text{find }\hat\theta\in\Theta^{\text{DAL}}:\langle\text{vec}(F_T(\theta)),\text{vec}\Big(\theta-\hat\theta\Big)\rangle\ge0,\quad\forall\theta\in\Theta^{\text{DAL}},\qquad \text{VI}[F_T,\Theta^{\text{DAL}}]$$

where $\text{vec}(A)$ is the vector of columns of A stacked one under the other, the empirical "vector" field is

$$F_T(\theta)=\Big(F_T^{(1)}(\theta_1),\dots,F_T^{(d_1)}(\theta_{d_1})\Big)\in\mathbb{R}^{d\times d_1},\tag{16}$$

and vector field $F_T^{(i)}(\theta_i)\in\mathbb{R}^d$ is defined in (9). Moreover, the convex set Θ^{DAL} incorporates the above data-adaptive linear constraints (13) and is defined as follows:

$$\begin{aligned}\Theta^{\text{DAL}}=\Big\{&\theta:\theta\in\tilde\Theta,\ e_{f_{j,\ell},d}^{\mathrm T}\theta e_{i,d_1}+e_{f_{i,\ell},d}^{\mathrm T}\theta e_{j,d_1}\le\delta_{2,\ell}(i,j),\\ &(i,j)\in I_{2,\ell},\ \ell\in[\tau],\\ &e_{f_{j,\ell},d}^{\mathrm T}\theta e_{i,d_1}+e_{f_{k,\ell},d}^{\mathrm T}\theta e_{j,d_1}+e_{f_{i,\ell},d}^{\mathrm T}\theta e_{k,d_1}\le\delta_{3,\ell}(i,j,k),\\ &(i,j,k)\in I_{3,\ell},\ \ell\in[\tau]\Big\},\end{aligned}$$

where regularization strength parameters $\delta_{2,\ell}(i,j),\delta_{3,\ell}(i,j,k)$ are defined in (14) and $f_{j,\ell}=1+d_3+\tau d_2+(j-1)\tau+\ell$ such that $e_{f_{j,\ell},d}^{\mathrm T}\theta e_{i,d_1}=\alpha_{ij\ell}$.

5) A Special Case: Linear Link Function: Now, we elaborate on our proposed regularization on a special linear link case. The vector field $F_T(\theta)$ (16) can be expressed as follows:

$$F_T(\theta)=\frac1T\mathbf{w}_{1:T}\mathbf{w}_{1:T}^{\mathrm T}\theta-\frac1T\mathbf{w}_{1:T}Y=\mathbb{W}_{1:T}\theta-\frac1T\mathbf{w}_{1:T}Y,$$

where $Y=(Y_{1:T}^{(1)},\dots,Y_{1:T}^{(d_1)})\in\mathbb{R}^{T\times d_1}$ and $Y_{1:T}^{(i)}$, $\mathbf{w}_{1:T}\in\mathbb{R}^{d\times T}$ are defined in (6). Similar to the linear model example in Section II-B, the above vector field is the gradient field of the least square objective, and our proposed estimator VI$[F_T,\Theta^{\text{DAL}}]$ boils down to the LS estimator, which solves the following constrained optimization problem:

$$\begin{aligned}\min_\theta\quad&\frac{1}{2T}\sum_{i=1}^{d_1}\|\mathbf{w}_{1:T}^{\mathrm T}\theta_i-Y_{i,1:T}\|_2^2=\frac{1}{2T}\|\mathbf{w}_{1:T}^{\mathrm T}\theta-Y\|_F^2,\\ \text{s.t.}\quad&\mathbf{0}_T\le\mathbf{w}_{1:T}^{\mathrm T}\theta_i\le\mathbf{1}_T,\ i\in[d_1],\\ &e_{f_{j,\ell},d}^{\mathrm T}\theta e_{i,d_1}+e_{f_{i,\ell},d}^{\mathrm T}\theta e_{j,d_1}\le\delta_{2,\ell}(i,j),\\ &(i,j)\in I_{2,\ell},\ \ell\in[\tau],\\ &e_{f_{j,\ell},d}^{\mathrm T}\theta e_{i,d_1}+e_{f_{k,\ell},d}^{\mathrm T}\theta e_{j,d_1}+e_{f_{i,\ell},d}^{\mathrm T}\theta e_{k,d_1}\le\delta_{3,\ell}(i,j,k),\\ &(i,j,k)\in I_{3,\ell},\ \ell\in[\tau].\end{aligned}\tag{17}$$

Similarly, (17) is convex and can be efficiently solved by a well-develop toolkit such as `Mosek`.

Most applications, including our motivating example, only considers triggering effect, meaning that one can replace $\mathbf{w}_{1:T}^{\mathrm T}\theta_i\ge\mathbf{0}_T$ with $\theta_i\ge\mathbf{0}_d$ as a relaxation. In addition, since the prediction of the i-th event's occurrence at time t is by comparing the estimated probability $w_{t-\tau:t-1}^{\mathrm T}\theta_i$ with a threshold selected using the validation dataset, we can further relax the constraint $\mathbf{w}_{1:T}^{\mathrm T}\theta_i\le\mathbf{1}_T$ and treat $w_{t-\tau:t-1}^{\mathrm T}\theta_i$ as a "score" instead of a probability. Thus, we can adopt the following penalized form:

$$\begin{aligned}\min_\theta\quad&\frac{1}{2T}\|\mathbf{w}_{1:T}^{\mathrm T}\theta-Y\|_F^2\\ &+\sum_{\ell=1}^{\tau}\Bigg(\sum_{(i,j)\in I_{2,\ell}}\frac{\lambda\Big(e_{f_{j,\ell},d}^{\mathrm T}\theta e_{i,d_1}+e_{f_{i,\ell},d}^{\mathrm T}\theta e_{j,d_1}\Big)}{\delta_{2,\ell}(i,j)}\\ &+\sum_{(i,j,k)\in I_{3,\ell}}\frac{\lambda\Big(e_{f_{j,\ell},d}^{\mathrm T}\theta e_{i,d_1}+e_{f_{k,\ell},d}^{\mathrm T}\theta e_{j,d_1}+e_{f_{i,\ell},d}^{\mathrm T}\theta e_{k,d_1}\Big)}{\delta_{3,\ell}(i,j,k)}\Bigg),\\ \text{s.t.}\quad&\theta_i\ge\mathbf{0}_T,\ i\in[d_1],\end{aligned}\tag{18}$$

where λ is a hyperparameter that controls the strength of regularization. The data-adaptive regularization strength parameters $\delta_{2,\ell}(i,j),\delta_{3,\ell}(i,j,k)$ appear in the denominator since smaller $\delta_{2,\ell}(i,j),\delta_{3,\ell}(i,j,k)$ imply stronger penalty, which closely resembles adaptive Lasso [15]. Most importantly, (18) can be solved efficiently using PGD, where at each iteration, the update rule is as follows:

$$\begin{aligned}\hat\theta\leftarrow\hat\theta-\eta\Bigg(&F_T\Big(\hat\theta\Big)+\sum_{\ell=1}^{\tau}\Bigg(\sum_{(i,j)\in I_{2,\ell}}\frac{\lambda\Big(e_{f_{j,\ell},d}e_{i,d_1}^{\mathrm T}+e_{f_{i,\ell},d}e_{j,d_1}^{\mathrm T}\Big)}{\delta_{2,\ell}(i,j)}\\ &+\sum_{(i,j,k)\in I_{3,\ell}}\frac{\lambda\Big(e_{f_{j,\ell},d}e_{i,d_1}^{\mathrm T}+e_{f_{k,\ell},d}e_{j,d_1}^{\mathrm T}+e_{f_{i,\ell},d}e_{k,d_1}^{\mathrm T}\Big)}{\delta_{3,\ell}(i,j,k)}\Bigg)\Bigg),\end{aligned}\tag{19}$$

where η is the step size/learning rate hyperparameter and empirical field $F_T(\cdot)$ is given in (16). After the above update in each iteration, the projection onto the feasible region $\mathbb{R}_+^{d\times d_1}$ can be simply done by replacing all negative entries in $\hat\theta$ with zeros.

B. Penalized Joint VI-Based Estimation

As previously discussed in Section II-C2, the VI$[F_T^{(i)}, \Theta]$ can be solved by PGD as the feasible region Θ is a convex set. However, VI$[F_T, \Theta^{\mathrm{DAL}}]$ additionally incorporates the data-adaptive linear constraints into its feasible region Θ^{DAL} to encourage a DAG structure with desired lagged self-exciting components, making the projection step harder to implement. Alternatively, it will be much easier if we can transfer the constraints into the penalty. Inspired by the penalized form for the linear link special case (18) (which is very similar to adaptive Lasso [15]), we propose a *data-adaptive linear penalized VI-based estimator*, which is the weak solution to the following VI:

$$\text{find } \hat{\theta} \in \tilde{\Theta} : \left\langle \mathrm{vec}\left(F_T^{\mathrm{DAL}}(\theta)\right), \mathrm{vec}\left(\theta - \hat{\theta}\right)\right\rangle \geq 0, \quad \forall \theta \in \tilde{\Theta}, \qquad \mathrm{VI}[F_T^{\mathrm{DAL}}, \tilde{\Theta}]$$

where the feasible region $\tilde{\Theta}$ is defined in (15) and the *data-adaptive linear penalized vector filed* $F_T^{\mathrm{DAL}}(\theta)$ is defined as follows:

$$F_T^{\mathrm{DAL}}(\theta) = F_T(\theta) + \sum_{\ell=1}^{\tau}\left(\sum_{(i,j)\in I_{2,\ell}} \frac{\lambda\left(e_{f_{j,\ell},d} e_{i,d_1}^{\mathrm{T}} + e_{f_{i,\ell},d} e_{j,d_1}^{\mathrm{T}}\right)}{\delta_{2,\ell}(i,j)} + \sum_{(i,j,k)\in I_{3,\ell}} \frac{\lambda\left(e_{f_{j,\ell},d} e_{i,d_1}^{\mathrm{T}} + e_{f_{k,\ell},d} e_{j,d_1}^{\mathrm{T}} + e_{f_{i,\ell},d} e_{k,d_1}^{\mathrm{T}}\right)}{\delta_{3,\ell}(i,j,k)}\right). \tag{20}$$

Here, λ is a tunable penalty strength hyperparameter, $F_T(\theta) = (F_T^{(1)}(\theta_1), \ldots, F_T^{(d_1)}(\theta_{d_1})) \in \mathbb{R}^{d\times d_1}$ is the concatenated field (16) and vector field $F_T^{(i)}(\theta_i) \in \mathbb{R}^d$ is defined in (9). Compared with VI$[F_T, \Theta^{\mathrm{DAL}}]$, it is much easier to solve VI$[F_T^{\mathrm{DAL}}, \tilde{\Theta}]$ using PGD. For example, in the exponential link function case, if we restrict our consideration to triggering effect only, we can use (19) as the update rule in PGD and zero out all negative entries after each update as the projection step in each iteration.

Remark 3: One advantage of our data-adaptive linear regularization is its flexibility, and it is the user's choice to decide which potential cycle should be included in the constraint. Please refer to our work [57] for more numerical examples of recovering strict DAGs (i.e., without any lagged self-exciting components) by additionally including length-1 cycles in our data-adaptive linear constraints.

Remark 4: The above idea to transfer constraint into a penalty by adding the penalty's derivative to the empirical vector field opens up possibilities to consider various types of penalties to encourage desired structures when using our proposed VI-based estimator, e.g., the continuous DAG characterization [17] and the adaptive Lasso [15]; one can see Section V below for more details on our proposed VI-based estimator coupled with DAG regularization (21) and ℓ_1 regularization (22).

IV. Non-Asymptotic Performance Guarantee

In this section, we will show our proposed estimator has nice statistical properties, i.e., it is unique and consistent; the proof is deferred to Appendix B-B in the supplementary material due to space consideration. In addition, we will also derive a linear program (LP) based confidence interval (CI) of parameters θ_i's, which we defer to Appendix B-C in the supplementary material. One pitfall of our theoretical analysis is the lack of guarantee for the proposed data-adaptive linear method and we leave this topic for future discussion. We begin with two necessary model assumptions:

Assumption 1: The link function $g(\cdot)$ is continuous and monotone, and the vector field $G(\theta) = \mathbb{E}_w[w g(w^{\mathrm{T}}\theta)]$ is well defined (and therefore monotone along with g). Moreover, g is differentiable and has uniformly bounded first order derivative $m_g \leq |g'| \leq M_g$ for $0 < m_g \leq M_g$.

Assumption 2: The observations (static, binary, and continuous) are bounded almost surely: there exists $M_w > 0$ such that at any time step t, we have $\|w_{t-\tau:t-1}\|_\infty \leq M_w$ with probability one.

Theorem 1 (Upper Bound on $\|\hat{\theta}_i - \theta_i^\star\|_2$): Under Assumptions 1 and 2, for $i \in [d_1]$ and any $\varepsilon \in (0,1)$, with probability at least $1-\varepsilon$, the ℓ_2 distance between ground truth $\theta_i^\star$ and the weak solution $\hat{\theta}_i$ to VI$[F_T^{(i)}, \Theta]$ can be upper bounded as follows:

$$\|\hat{\theta}_i - \theta_i^\star\|_2 \leq \frac{M_w}{m_g \lambda_1}\sqrt{\frac{d\log(2d/\varepsilon)}{T}},$$

where λ_1 is the smallest eigenvalue of $\mathbb{W}_{1:T} = \sum_{t=1}^{T} w_{t-\tau:t-1} w_{t-\tau:t-1}^{\mathrm{T}}/T$ (4).

The above theorem is an extension to the general link function case with mixed-type data of [19, Th. 1]. As pointed out in [19], $\mathbb{W}_{1:T} \in \mathbb{R}^{d\times d}$ will be full rank for sufficiently large T, i.e., λ_1 will be a positive constant with high probability.

Remark 5 (Identifiablility): The uniqueness, or rather, the identifiability, comes from the nice property of the underlying vector field. To be precise, in the proof of the above theorem (see Appendix B-B in the supplementary material), we have shown the vector field $F_T^{(i)}(\theta_i)$ is monotone modulus $m_g\lambda_1$ under Assumption 1. Then, the following lemma tells us that our proposed estimator is unique:

Lemma 1 [18, Lemma 3.1]: Let Θ be a convex compact set and H be a monotone vector field on Θ with monotonicity modulus $\kappa > 0$, i.e.,

$$\forall\, z, z' \in \Theta, \left[H(z) - H(z')\right]^{\mathrm{T}}(z - z') \geq \kappa\|z - z'\|_2^2.$$

Then, the weak solution $\bar{z}$ to VI$[H, \Theta]$ exists and is unique. It satisfies:

$$H(z)^{\mathrm{T}}(z - \bar{z}) \geq \kappa\|z - \bar{z}\|_2^2.$$

Next, we will use both simulations and real examples to show the good performance of our method for causal structural learning.

V. Numerical Simulation

In this section, we conduct numerical simulations to show the good performance of VI-based estimator VI$[F_T^{(i)}, \Theta]$.

We will show the competitive performance of our VI-based method compared with benchmark methods such as Neural Network based method [12], even under the model mis-specification setting. Importantly, we also show that our proposed data-adaptive linear regularization outperforms other DAG-inducing regularization approaches in structural learning. Due to space consideration, the complete experimental configurations and the comparison between VI-based estimator and under the model mis-specification setting is deferred to Appendices C-A and C-B.

1) Evaluation Metrics: We consider a simple $\tau = 1$ case in our simulations, and we are interested in the estimation of model parameters: (i) background intensity $\nu = (\nu_1, \ldots, \nu_{d_1})^{\mathrm{T}}$ and (ii) self- and mutual-exciting matrix $A_1 = (\alpha_{ij1})$; for brevity, we drop the last subscript "1" and denote the adjacency matrix by $A = (\alpha_{ij})$. We consider (i) the ℓ_2 norm of the background intensity estimation error $\|\hat{\nu} - \nu\|_2$ (ν err.) and (ii) matrix F-norm of the self- and mutual-exciting matrix estimation error $\|\hat{A} - A\|_F$ (A err.). Additionally, we report the Structural Hamming Distance (SHD) between $\hat{A}$ and A, which reflects how close the recovered graph is to the ground truth. This is the primary quantitative metric in the following experiment. SHD is the number of edge flips, insertions, and deletions in order to transform between two graphs. In particular, when edge $i \to j$ is in the true graph, i.e., $\alpha_{ji} > 0$, whereas edge $i \leftarrow j$ is in the estimated graph, i.e., $\hat{\alpha}_{ij} > 0$, the SHD is increased by 1 via edge flip instead of 2 by edge insertion and deletion. In addition, since we are interested in DAG structure with self-exciting components, we also consider a measure of "DAG-ness" on the recovered adjacency matrix (after zeroing out the diagonal entries of $\hat{A}$), denoted by $h(A_0)$ (10). We need to mention that small $h(A_0)$ with large SHD means we recover a DAG which is not close to the ground truth and this does not imply good structure recovery.

2) Benchmark Regularization Approaches: Let us first formally introduce several benchmark regularization approaches. Recall that we use $A = (\alpha_{ij})$ to denote $A_1 = (\alpha_{ij1})$ in $\tau = 1$ case for brevity:

- *Continuous DAG Regularization and a Proposed Variant:* Recall the continuous and differentiable (but not convex) characterization by [17] in (10), which can measure the DAG-ness of A. Most importantly, this DAG characterization has closed from derivative, i.e., $\nabla h(A) = (e^A)^{\mathrm{T}}$. Inspired by [20], we use this characterization as a penalty directly. We take advantage of its differentiability and add its derivative to the concatenated field $F_T(\theta)$ (16), which will be treated as the gradient field in PGD. More specifically, let $J = (\mathbf{0}_{d_1}, I_{d_1}) \in \mathbb{R}^{d_1 \times d}$ and we will have $J\theta = A^{\mathrm{T}}$. Then, the vector field coupled with DAG regularization $F_T^{\mathrm{DAG}}(\cdot)$ is defined as follows:

$$F_T^{\mathrm{DAG}}(\theta) = F_T(\theta) \ + \lambda J^{\mathrm{T}} \nabla h(J\theta) = F_T(\theta) \ + \lambda J^{\mathrm{T}} e^A, \tag{21}$$

  where tunable hyperparameter λ controls the penalty strength. The PGD update rule is given by:

$$\hat{\theta} \leftarrow \hat{\theta} - \eta F_T^{\mathrm{DAG}}\left(\hat{\theta}\right),$$

  where η is the learning rate hyperparameter and is also tunable. One drawback of the aforementioned DAG regularization is that it removes not only cycles but also lagged self-exciting components; this is evidenced in Figure 3. To keep those informative lagged self-exciting components, we simply zero out the diagonal elements in DAG regularization derivative $\nabla h(J\theta)$ in (21). Thus, the PGD update will not shrink the diagonal elements.
- *ℓ_1 Regularization and Adaptive Lasso:* We adopt the ℓ_1 penalty as another benchmark, which encourages a sparse graph structure and, in turn, eliminates cycles. The ℓ_1 penalized vector filed is defined as follows:

$$F_T^{\ell_1}(\theta) = F_T(\theta) + \lambda J^{\mathrm{T}} \nabla(|J\theta|_1), \tag{22}$$

  where $|\cdot|_1$ is the summation of the absolute values of all entries. Similarly, the VI-based estimator can be efficiently solved by PGD using the following update rule:

$$\hat{\theta} \leftarrow \hat{\theta} - \eta F_T^{\ell_1}\left(\hat{\theta}\right).$$

  As a variant of ℓ_1 regularization, adaptive ℓ_1 regularization (or adaptive Lasso [15]) replaces $\lambda|\alpha_{ij}|$ with $\frac{\lambda}{\hat{\alpha}_{ij}}|\alpha_{ij}|$ in (22); for $\hat{\alpha}_{ij} = 0$ case, we adopt a simple remedy by adding penalty term $10^3\lambda|\alpha_{ij}|$ to enforce α_{ij} to be zero.

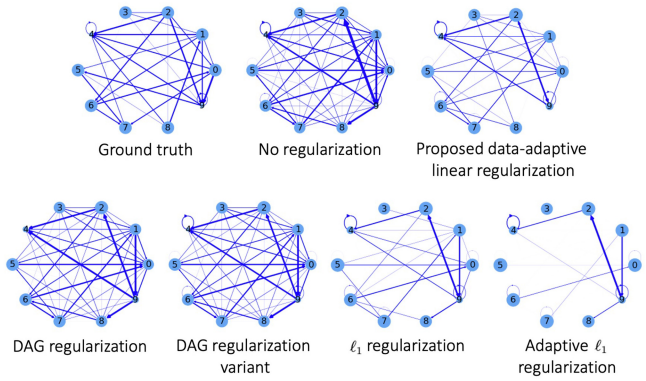

Fig. 3. Simulated example: demonstration of the effectiveness of our proposed data-adaptive linear regularization. We visualize the recovered graph structures for a $d_1 = 10$ and $T = 500$ illustrative example using our proposed VI-based estimator coupled with various types of regularization (specified on top of each panel). We can observe that our proposed VI-based estimator coupled with our proposed data-adaptive linear constraint can return the closest graph structure to the ground truth; see quantitative evaluation metrics, such as SHD, in Table II.

3) Results: We first demonstrate the competitive performance of our proposed data-adaptive linear method on a $d_1 = 10$ illustrative example, where we adopt an exponential link function. We visualize the recovered graphs using our VI-based estimator with exponential link coupled with various types of regularization in Figure 3, and we report all aforementioned quantitative metrics in Table II; additionally, we report the relative errors in the illustrative example in Table VIII in the supplementary material. We observe that our proposed data-adaptive linear regularization can achieve the best weight recovery accuracy (in terms of ν. err. and A err.) and structure recovery accuracy (in terms of SHD) compared with all benchmark methods.

TABLE II
SIMULATED EXAMPLE: QUANTITATIVE METRICS OF THE EXAMPLE IN FIGURE 3. WE CAN OBSERVE THAT OUR PROPOSED VI-BASED ESTIMATOR, COUPLED WITH OUR PROPOSED DATA-ADAPTIVE LINEAR CONSTRAINT, CAN ACHIEVE BETTER ESTIMATION ACCURACY WHILE ENCOURAGING A DESIRED DAG STRUCTURE. BESIDES, OUR PROPOSED METHOD ALSO GIVES THE BEST STRUCTURE RECOVERY, I.E., THE SMALLEST SHD; ALTHOUGH THE ADAPTIVE ℓ_1 APPROACH ACHIEVES THE BEST DAG-NESS, IT ACHIEVES SO BY REMOVING MANY IMPORTANT EDGES AND CANNOT OUTPUT A CORRECT GRAPH STRUCTURE (AS EVIDENCED IN FIGURE 3)

Regularization	None	Proposed	DAG	DAG-Variant	ℓ_1	Ada. ℓ_1
A err.	0.3874	**0.2094**	0.3541	0.2949	*0.2501*	0.3022
ν err.	0.1175	**0.0775**	0.0895	*0.0841*	0.0884	0.1251
$h(A_0)$	0.1223	0.0308	0.0337	*0.0242*	0.0274	**0.0232**
SHD	41	**25**	32	34	41	*29*

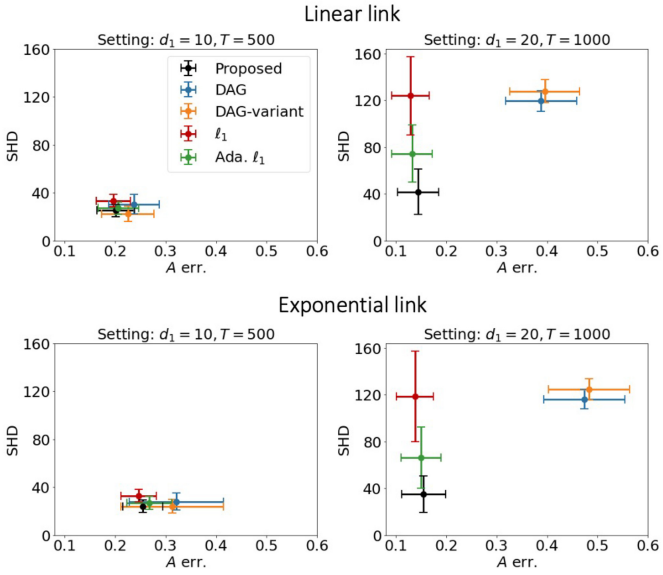

Fig. 4. Simulation: mean (dot) and standard deviation (error bar) of matrix F-norm of the self- and mutual-exciting matrix estimation error (A err.) and Structural Hamming Distance (SHD) over 100 independent trials for various types of regularization. For each regularization, the closer it is to the origin, the better it is. We can observe that our proposed data-adaptive linear regularization performs the best (especially in higher dimensional cases) in terms of structure recovery while achieving almost the same weight recovery accuracy with the best result.

To further validate the good performance of our proposed data-adaptive linear method, we run 100 independent trials for $d_1 = 10, T = 500$ and $d_1 = 20, T = 1000$ cases as well as linear link and exponential link functions cases. We plot the mean and standard deviation of A err. and SHD in Figure 4; for completeness, we also report the raw values of the mean and standard deviation of all four aforementioned metrics in Table IX in the supplementary material. The complete details, including random DAG generation, PGD to solve for the estimators, and additional results (Tables VIII and IX in the supplementary material), are deferred to Appendix C-C in the supplementary material.

Figure 4 shows that, in low dimensional (i.e., $d_1 = 10$) case, ℓ_1 regularization does well in weight recovery but fails in structure recovery, whereas our proposed variant of DAG regularization prioritizes the structure recovery but performs poorly in weight recovery. As a comparison, our proposed data-adaptive linear regularization achieves comparable weight and structure recovery accuracy to ℓ_1 regularization and the proposed variant of DAG regularization, respectively, suggesting that it can balance the weight recovery accuracy and the structure recovery accuracy in low dimensional case. In the higher dimensional (i.e., $d_1 = 20$) case, our proposed approach achieves the best structure recovery accuracy while maintaining nearly the same weight recovery accuracy with the best result (achieved by ℓ_1 regularization-based method). It is interesting to observe that adaptive ℓ_1 regularization's performance lies between ℓ_1 regularization and our proposed regularization. In addition, our proposed regularization has a dominating performance over DAG regularization-based approaches in terms of both structure recovery accuracy and weight recovery accuracy. These observations are also validated by Table IX in the Appendix in the supplementary material.

VI. REAL DATA EXAMPLE

In this section, we demonstrate the usefulness of our proposed method in a real study. We perform a train-validation-test split to select the models and their hyperparameters based on the predictive performance on the held-out test dataset. We show that the proposed discrete-time Hawkes network with a linear link function achieves the best performance. To enhance interpretability, we perform Bootstrap uncertainty quantification to remove false discoveries in the causal graphs. Importantly, our proposed DAG-encouraging regularization can further boost both the predictive performance and the causal graph interpretability.

A. Settings

1) Dataset and Sepsis Associated Derangements: This real study targets a short time window right after the SOFA score turns 2 for ICU patients; see patient demographics in Table I, Section I-A. To reduce the complexity of the computations due to high-dimensional raw features (i.e., vital signs and Lab results), expert (i.e., clinician) opinion is utilized to identify common and clinically relevant SADs that could be detected using structured EMR data. In particular, those Labs and vital signs are all converted into binary SADs, representing nodes in the graph; As all those raw features are used in SADs' construction, they are not input to the model to avoid undesired correlation amongst nodes. Please find further details in Table X in Appendix D-A1 in the supplementary material.

2) Evaluation Metrics: The primary quantitative evaluation metric is Cross Entropy (CE) loss, as our model outputs predicted probabilities for binary SADs sequentially. As SADs do not occur very often for each patient, we also use Focal loss as an alternative metric to account for such class imbalance issues. Furthermore, we focus on the interpretability of the resulting causal DAG by (i) counting the number of undesirable length-L cycles, $L \in \{2, 3, 4, 5\}$, and (ii) studying whether or not the inferred interactions align

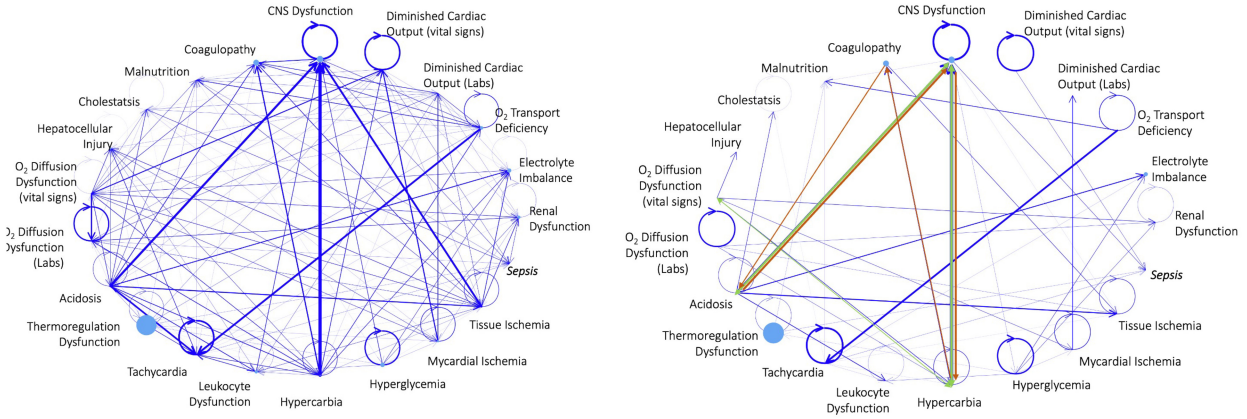

Fig. 5. Causal graphs recovered by our proposed discrete-time Hawkes network using a linear link without (left) and with (right) Bootstrap uncertainty quantification. We remove the directed edges whose 95% Bootstrap confidence intervals contain zero. We can observe that, by comparing both graphs, the graph with BP has much fewer edges (which are caused by noisy observation) and is thus more interpretable. However, there still exist three length-2 cycles (green) and one length-4 cycle (orange) in this graph.

with well-known physiologic relationships. Please find further details in Appendix D-A2 in the supplementary material.

3) Training Details: We perform train-validation-test data split to fine-tune the hyperparameters: we use the 2018 data as the training dataset, and we randomly split the 2019 dataset by half (for both sepsis and non-sepsis patient cohorts) into validation and testing datasets; we will select the hyperparameters based on the CE loss on the validation dataset and demonstrate its performance on the test dataset. We fit the candidate models using the 2018 real data and select the hyperparameters based on the total CE loss on the 2019 validation dataset. Please find further details in Appendix D-A3 in the supplementary material.

B. Model Comparison

1) Candidate Models: We compare the VI-based estimation $\mathrm{VI}[F_T^{(i)}, \Theta]$ with the black-box methods — we choose XGBoost over other models, e.g., Neural Networks, since it outperformed other candidate models in the 2019 Physionet Challenge on sepsis prediction [58]. For the VI-based estimation, we consider linear, sigmoid, and exponential link functions. Furthermore, we incorporate expert opinion/prior medical knowledge that no inhibiting effect exists: on one hand, it remained unclear how to interpret inhibiting effects among SADs; on the other hand, restricting our consideration to triggering effects reduces the feasible region to $\theta \in \mathbb{R}_+^{d \times d_1}$, which enables us to leverage PGD to numerically obtain the estimators.

2) Results: We visualize the recovered causal graphs of the SADs in Figure 5 and report the out-of-sample testing metrics in Table III. We can observe that our proposed model with linear link function achieves the best performance for predicting most SADs; in particular, it has the smallest out-of-sample CE loss for sepsis prediction. Although the exponential link function outperforms the linear link for some SADs, the improvements are negligible; in addition, it performs rather poorly for CNS Dysfunction and Tachycardia predictions. Therefore, we will continue our real data analysis using the VI-based estimation coupled with the linear link function. For completeness, further comparisons with XGBoost, VI-based estimation coupled with non-linear links (such as Figure 12 in the supplementary material), and VI-based estimation without prior medical knowledge (i.e., with potential inhibiting effects) are deferred to Appendices D-B1, D-B2, and D-B3 in the supplementary material, respectively.

TABLE III

COMPARISON BETWEEN THE VI-BASED ESTIMATION COUPLED WITH VARIOUS LINK FUNCTIONS AND BLACK-BOX XGBOOST: WE REPORT THE AVERAGE AND STANDARD DEVIATION OF CROSS ENTROPY LOSS OVER ALL PATIENTS IN THE 2019 TEST DATASET FOR ALL METHODS. THE BEST RESULTS (BEFORE WE ROUND THE NUMBER) ARE HIGHLIGHTED. WE CAN OBSERVE THE VI-BASED ESTIMATION COUPLED WITH LINEAR LINK FUNCTION OUTPERFORMS OTHER CANDIDATE METHODS WHEN PREDICTING MOST SADS. ALTHOUGH THE EXPONENTIAL LINK FUNCTION PERFORMS THE BEST FOR MANY SADS, ITS IMPROVEMENTS COMPARED WITH LINEAR LINK ARE MARGINAL, LET ALONE IT IS NOT ROBUST IN THE SENSE THAT IT PERFORMS POORLY FOR CNSDYS AND TACHY PREDICTIONS

	Linear link	Exponential link	Sigmoid link	XGBoost
RenDys	**0.1246** (0.2591)	0.1246 (0.2586)	0.6198 (1.5365)	0.3799 (0.2207)
LyteImbal	0.1781 (0.2623)	**0.1780** (0.2605)	1.0075 (1.9436)	0.4511 (0.1644)
O2TxpDef	**0.1921** (0.2224)	0.1922 (0.2295)	1.1341 (1.8396)	0.4515 (0.1964)
DCO (L)	0.0129 (0.0902)	**0.0128** (0.0896)	0.0399 (0.3269)	0.0847 (0.0939)
DCO (V)	**0.1461** (0.2909)	0.1491 (0.2934)	0.9865 (2.7843)	0.3522 (0.3263)
CNSDys	0.2604 (0.4068)	1.0357 (3.7069)	0.6446 (1.1268)	**0.2491** (0.3492)
Coag	0.2441 (0.2034)	**0.2441** (0.2002)	1.5349 (1.8136)	0.5254 (0.1670)
MalNut	0.1253 (0.2048)	**0.1252** (0.2036)	0.6470 (1.3905)	0.3746 (0.2091)
Chole	**0.0495** (0.1712)	0.0495 (0.1710)	0.1985 (0.8692)	0.2420 (0.2142)
HepatoDys	0.0839 (0.2061)	**0.0838** (0.2051)	0.3943 (1.1954)	0.3207 (0.1898)
O2DiffDys (L)	**0.0081** (0.1482)	0.0095 (0.1833)	0.0141 (0.2543)	0.0181 (0.0455)
O2DiffDys (V)	**0.1494** (0.3069)	0.1524 (0.3105)	1.018 (2.7288)	0.3503 (0.3170)
Acidosis	0.0609 (0.1850)	**0.0608** (0.1841)	0.3347 (1.3496)	0.2191 (0.1457)
ThermoDys	**0.0053** (0.0470)	0.0306 (0.4780)	0.0306 (0.4780)	0.4426 (0.1440)
Tachy	**0.3911** (0.4710)	2.4593 (4.4066)	0.5624 (0.6535)	0.3967 (0.3082)
LeukDys	0.1243 (0.2368)	**0.1243** (0.2360)	0.6385 (1.515)	0.3887 (0.2047)
HypCarb	**0.0511** (0.1773)	0.0513 (0.1782)	0.2442 (1.0791)	0.2220 (0.1911)
HypGly	**0.2849** (0.2766)	0.2879 (0.2773)	2.274 (3.5716)	0.5141 (0.2314)
MyoIsch	**0.0587** (0.2313)	0.0591 (0.2331)	0.2746 (1.265)	0.2268 (0.2191)
TissueIsch	0.0739 (0.1814)	**0.0739** (0.1799)	0.4139 (1.4378)	0.2664 (0.1593)
SEP3	**0.1318** (0.1803)	0.1319 (0.1794)	0.6915 (1.2001)	0.4155 (0.2680)

C. Causal DAG Discovery

Table III shows the competitive performance of the VI-based estimation coupled with the linear link function, and thus we continue our analysis with such mode choice. The

TABLE IV
COMPARISON AMONG VARIOUS TYPES OF REGULARIZATION: WE REPORT THE NUMBER OF CYCLES FOR VARIOUS METHODS. WE CAN SEE BOOTSTRAP CAN REMOVE MOST OF THE CYCLES BY REMOVING THE "LESS IMPORTANT EDGES" IN THE GRAPH; SEE THE COMPARISON BETWEEN BOTH GRAPHS IN FIGURE 5 FOR A GRAPHICAL ILLUSTRATION. MOREOVER, BUILT ON TOP OF THE BOOTSTRAP UNCERTAINTY QUANTIFICATION, PROPER DAG-INDUCING REGULARIZATION CAN COMPLETELY REMOVE CYCLES AND ENCOURAGE A DESIRED "DAG WITH SELF-EXCITING COMPONENTS" STRUCTURE

	Linear link	Exponential link	Sigmoid link	Linear link (BP)	Linear link (BP + proposed)	Linear link (BP + ℓ_1)	Linear link (BP + DAG)	Linear link (BP + DAG-variant)
Num. of Len-2 Cycles	39	45	0	3	0	0	0	3
Num. of Len-3 Cycles	136	192	0	0	0	0	0	0
Num. of Len-4 Cycles	680	1023	0	1	0	0	0	1
Num. of Len-5 Cycles	3310	5419	0	0	0	0	0	0

TABLE V
COMPARISON AMONG VARIOUS TYPES OF REGULARIZATION: WE REPORT THE AVERAGE AND STANDARD DEVIATION OF CROSS ENTROPY LOSS OVER ALL PATIENTS IN THE 2019 TEST DATASET. THE BEST RESULTS (BEFORE WE ROUND THE NUMBER) ARE HIGHLIGHTED. WE CAN OBSERVE OUR PROPOSED DATA-ADAPTIVE LINEAR REGULARIZATION CAN ACHIEVE THE BEST PERFORMANCE FOR MOST SADS COMPARED WITH OTHER DAG-INDUCING REGULARIZATIONS; MOREOVER, BY COMPARING THIS TABLE WITH TABLE III, WE CAN SEE IT ACHIEVES ALMOST THE SAME PERFORMANCE AS THE BEST ACHIEVABLE PERFORMANCE

	Linear link (BP)	Linear link (BP + proposed)	Linear link (BP + ℓ_1)	Linear link (BP + DAG)	Linear link (BP + DAG-variant)
RenDys	0.2342 (0.6967)	0.1253 (0.2655)	**0.1240**(0.2586)	0.1261 (0.2686)	0.2342 (0.6967)
LyteImbal	0.1801 (0.2793)	0.1801 (0.2741)	**0.1792**(0.2709)	0.1794 (0.2790)	0.1801 (0.2793)
O2TxpDef	0.5044 (0.8425)	**0.1920**(0.2310)	0.1921 (0.2320)	0.1970 (0.2511)	0.5044 (0.8425)
DCO (L)	0.0308 (0.2553)	0.0129 (0.0967)	**0.0128**(0.0898)	0.0320 (0.2615)	0.0308 (0.2553)
DCO (V)	0.3142 (0.7536)	**0.1451**(0.2892)	0.1453 (0.2852)	0.1784 (0.4096)	0.3142 (0.7536)
CNSDys	0.2583 (0.4001)	0.2589 (0.3810)	0.2617 (0.3882)	0.4609 (0.2358)	**0.2583**(0.4001)
Coag	0.2461 (0.2236)	0.2454 (0.2177)	0.2451 (0.2138)	**0.2450**(0.2141)	0.2461 (0.2236)
MalNut	**0.1270**(0.2209)	0.1290 (0.2343)	0.1290 (0.2304)	0.1337 (0.2466)	0.1270 (0.2209)
Chole	0.0689 (0.2679)	0.0689 (0.2679)	**0.0496**(0.1890)	0.0976 (0.3813)	0.0689 (0.2679)
HepatoDys	0.3155 (0.9563)	0.0888 (0.2393)	**0.0860**(0.2232)	0.0887 (0.2379)	0.3155 (0.9563)
O2DiffDys (L)	0.0075 (0.1495)	**0.0051**(0.0883)	0.0052 (0.0876)	0.0113 (0.2034)	0.0075 (0.1495)
O2DiffDys (V)	0.3370 (0.9067)	0.1507 (0.3177)	**0.1504**(0.3146)	0.1849 (0.4181)	0.3370 (0.9067)
Acidosis	0.0696 (0.2577)	0.1123 (0.4437)	**0.0611**(0.1864)	0.0779 (0.2809)	0.0696 (0.2577)
ThermoDys	0.0066 (0.0476)	0.0057 (0.0478)	**0.0055**(0.0480)	0.0098 (0.1183)	0.0064 (0.0476)
Tachy	0.4835 (0.7772)	0.3736 (0.3732)	**0.3730**(0.3725)	0.5379 (0.1770)	0.4835 (0.7772)
LeukDys	0.1688 (0.4368)	0.1261 (0.2506)	**0.1255**(0.2452)	0.1271 (0.2538)	0.1688 (0.4368)
HypCarb	0.0512 (0.1814)	0.0513 (0.1791)	0.0514 (0.1801)	0.0593 (0.2290)	**0.0512**(0.1815)
HypGly	0.4121 (0.6075)	**0.2853**(0.2793)	0.2853 (0.2785)	0.3189 (0.3336)	0.4121 (0.6075)
MyoIsch	0.1040 (0.5856)	0.1295 (0.6292)	**0.0585**(0.2312)	0.0687 (0.2952)	0.1040 (0.5856)
TissueIsch	0.0741 (0.1936)	0.0743 (0.1891)	0.0745 (0.1857)	0.1067 (0.3488)	**0.0741**(0.1936)
SEP3	0.1323 (0.1848)	0.1323 (0.1848)	0.1349 (0.1794)	**0.1320**(0.1810)	0.1323 (0.1848)

objective now is to improve the result interpretability by considering causal structural learning.

1) Bootstrap UQ: We report the number of length-L cycles for $L \in \{2, 3, 4, 5\}$ in Table IV. As we can see from that table, the recovered graph without uncertainty quantification and DAG-inducing regularization contains many cycles, making the results less explainable.

The left panel in Figure 5 shows many edges with very small weights, meaning that the existence of such an edge might be a result of noisy observations. For example, although the edge from Diminished Cardiac Output (vital signs) to sepsis events agrees with the well-known causal relationships in sepsis-related illness, its weight is too small to convince the clinician that such a triggering effect is statistically significant. Thus, before applying regularization, we first perform Bootstrap UQ, and the existence of an edge is determined by its Bootstrap confidence interval: we assign zero weight to that edge if its 95% CI contains zero; otherwise, we use the median of the Bootstrap results as the weight. Here, we obtain the CI based on 1500 Bootstrap trails; complete details are deferred to Appendix D-A4 in the supplementary material. The resulting graph is reported in the right panel in Figure 5 and the CE loss is reported in Table V.

The results in Table IV and Figure 5 show that BP can remove a substantive amount of cycles. Importantly, it is good to observe that the well-known triggering effect from Diminished Cardiac Output (vital signs) to sepsis events is statistically significant; see Figure 5. However, as evidenced by Tables III and V, performing Bootstrap UQ leads to much worse CE loss for almost all SADs To improve its prediction performance to make the interpretable graphs more convincing, and to remove the remaining cycles highlighted in Figure 5, we consider causal structural learning via our penalized VI-based estimation such as VI[F_T^{DAL}, $\tilde{\Theta}$].

2) Causal DAG Recovery via Regularization: We adopt the regularization approaches described in Sections III-A and V; again, for each regularization, we perform Bootstrap UQ with 1500 trials and 95% confidence level; we select the regularization strength hyperparameter λ using grid search based on the validation total CE loss. We report the CE loss on the test dataset for each regularization (with corresponding selected λ's) in Table V; the resulting graphs are visualized at the beginning of this paper in Figure 1. From Table V, we can observe that our proposed data-adaptive linear regularization can not only remove cycles while keeping the lagged self-exciting components but also

reduce the out-of-sample prediction CE loss. This suggests that Bootstrap coupled with our proposed DAG-inducing regularization outputs a highly interpretable causal DAG (Figure 1 and Table IV) while achieving almost identical out-of-sample prediction performance (Tables III and V). Additionally, we report an additional metric — Focal loss — in Table XIII, Appendix D-B4 in the supplementary material, and hyperparameter λ selection table in Table XIV, Appendix D-B5 in the supplementary material, re-affirming the aforementioned findings.

3) Interpretation: Figure 1 elucidates which relationships are most important in the graph, which is an essential aspect of interpretability. For example, the triggering effect from Diminished Cardiac Output (vital signs) to sepsis events remains significant after the Bootstrap UQ and cycle elimination; in fact, nearly all triggering effects of sepsis events remain significant. We provide the top causes of sepsis discovered in Figure 1, showing their similarity to the results of XGBoost [6] on clinically published data [3]. Due to space consideration, one can find those results in Appendix D-B6 in the supplementary material.

The primary outcome of interest for this work was sepsis. Meanwhile, as demonstrated in Figure 1, the causal relationship between any node pair can be estimated: Indeed, we can identify several strong triggering effects shared by both graphs, which are commonly recognized (though in other types of patients). For example, the exciting effect from Hyperglycemia to Electrolyte Imbalance is commonly seen in type 2 diabetes patients [59], and the observation that Acidosis precedes Cholestatsis is common for patients with pregnancy [60]. Meanwhile, our model can predict all SAD events in the graph, and this gives clinician users insight into the probability of observing subsequent SADs after sepsis. Even though our prediction of sepsis events is not perfect, the ability to predict other SADs that are on the path to sepsis or identify different potential pathways to an adverse event is also very important for clinicians to respond to those potential adverse events accordingly. Overall, the fact that identified triggering effects agree with the well-known physiologic relationships and the satisfying predictive performance affirm the usefulness of our proposed method.

VII. Conclusion

In this work, we present a GLM for causal DAG discovery from mutually exciting time series data. Most importantly, our proposed data-adaptive linear DAG-inducing regularization helps formulate the model estimation as a convex optimization problem. Furthermore, we establish a non-asymptotic estimation error upper bound for the GLM, which is verified numerically; we also give a confidence interval by solving linear programs. Both our numerical simulation and real data example show the good performance of our proposed method, making its future adoption in conducting continuous surveillance under medical settings and other similar problems much more likely. Meanwhile, there are a few interesting topics that the current work does not cover. For example, the convexity inherent in our proposed data-adaptive linear causal discovery method opens up the possibility of establishing performance guarantees, which we leave for future discussion.

References

[1] *Global Report on the Epidemiology and Burden of Sepsis: Current Evidence, Identifying Gaps and Future Directions*, World Health Org., Geneva, Switzerland, 2020.

[2] L. M. Fleuren et al., "Machine learning for the prediction of sepsis: A systematic review and meta-analysis of diagnostic test accuracy," *Intensive Care Med.*, vol. 46, no. 3, pp. 383–400, 2020.

[3] M. A. Reyna et al., "Early prediction of sepsis from clinical data: The PhysioNet/computing in cardiology challenge 2019," in *Proc. Comput. Cardiol. (CinC)*, 2019, pp. 1–8.

[4] J. A. Du, N. Sadr, and P. de Chazal, "Automated prediction of sepsis onset using gradient boosted decision trees," in *Proc. Comput. Cardiol. (CinC)*, 2019, pp. 1–4.

[5] M. Zabihi, S. Kiranyaz, and M. Gabbouj, "Sepsis prediction in intensive care unit using ensemble of XGboost models," in *Proc. Comput. Cardiol. (CinC)*, 2019, pp. 1–4.

[6] M. Yang et al., "An explainable artificial intelligence predictor for early detection of sepsis," *Crit. Care Med.*, vol. 48, no. 11, pp. e1091–e1096, 2020.

[7] S. P. Shashikumar, C. S. Josef, A. Sharma, and S. Nemati, "DeepAISE—An interpretable and recurrent neural survival model for early prediction of sepsis," *Artif. Intell. Med.*, vol. 113, Mar. 2021, Art. no. 102036.

[8] H. Lütkepohl, *New Introduction to Multiple Time Series Analysis*. Heidelberg, Germany: Springer, 2005.

[9] A. Shojaie and E. B. Fox, "Granger causality: A review and recent advances," *Annu. Rev. Stat. Appl.*, vol. 9, pp. 289–319, Mar. 2022.

[10] J. Pearl, *Causality*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[11] A. Tank, I. Covert, N. Foti, A. Shojaie, and E. Fox, "Neural Granger causality for nonlinear time series," Feb. 2018, *arXiv:1802.05842v2*.

[12] S. Khanna and V. Y. F. Tan, "Economy statistical recurrent units for inferring nonlinear Granger causality," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–19.

[13] K. Zhang and A. Hyvärinen, "Causality discovery with additive disturbances: An information-theoretical perspective," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Disc. Databases*, 2009, pp. 570–585.

[14] A. Hyvärinen, K. Zhang, S. Shimizu, and P. O. Hoyer, "Estimation of a structural vector autoregression model using non-Gaussianity," *J. Mach. Learn. Res.*, vol. 11, no. 5, pp. 1709–1731, 2010.

[15] H. Zou, "The adaptive lasso and its oracle properties," *J. Amer. stat. Assoc.*, vol. 101, no. 476, pp. 1418–1429, 2006.

[16] R. Pamfil et al., "DYNOTEARS: Structure learning from time-series data," in *Proc. Int. Conf. Artif. Intell. Stat.*, 2020, pp. 1595–1605.

[17] X. Zheng, B. Aragam, P. K. Ravikumar, and E. P. Xing, "DAGs with NO TEARS: Continuous optimization for structure learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–22.

[18] A. B. Juditsky and A. Nemirovski, "Signal recovery by stochastic optimization," *Autom. Remote Control*, vol. 80, no. 10, pp. 1878–1893, 2019.

[19] A. Juditsky, A. Nemirovski, L. Xie, and Y. Xie, "Convex parameter recovery for interacting marked processes," *IEEE J. Sel. Areas Inf. Theory*, vol. 1, no. 3, pp. 799–813, Nov. 2020.

[20] I. Ng, A. Ghassami, and K. Zhang, "On the role of sparsity and DAG constraints for learning linear dags," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 17943–17954.

[21] M. Singer et al., "The third international consensus definitions for sepsis and septic shock (sepsis-3)," *JAMA*, vol. 315, no. 8, pp. 801–810, 2016.

[22] C. W. Seymour et al., "Time to treatment and mortality during mandated emergency care for sepsis," *New Engl. J. Med.*, vol. 376, no. 23, pp. 2235–2244, 2017.

[23] S. Wei, Y. Xie, C. S. Josef, and R. Kamaleswaran, "Granger causal chain discovery for sepsis-associated derangements via continuous-time Hawkes processes," in *Proc. 29th ACM SIGKDD Conf. Knowl. Disc. Data Min.*, 2023, pp. 2536–2546.

[24] S. Wei, Y. Xie, C. S. Josef, and R. Kamaleswaran, "Causal graph recovery for sepsis-associated derangements via interpretable Hawkes networks," in *Proc. Int. Conf. Mach. Learn. (IMLH)*, 2021, pp. 1–72.

[25] A. Tank, E. B. Fox, and A. Shojaie, "Identifiability and estimation of structural vector autoregressive models for subsampled and mixed-frequency time series," *Biometrika*, vol. 106, no. 2, pp. 433–452, 2019.

[26] C. Glymour, K. Zhang, and P. Spirtes, "Review of causal discovery methods based on graphical models," *Front. Genet.*, vol. 10, p. 524, 2019.
[27] M. Chickering, D. Heckerman, and C. Meek, "Large-sample learning of Bayesian networks is NP-hard," *J. Mach. Learn. Res.*, vol. 5, pp. 1287–1330, Oct. 2004.
[28] Y. Yuan, X. Shen, W. Pan, and Z. Wang, "Constrained likelihood for reconstructing a directed acyclic Gaussian graph," *Biometrika*, vol. 106, no. 1, pp. 109–125, 2019.
[29] H. Manzour, S. Küçükyavuz, H.-H. Wu, and A. Shojaie, "Integer programming for learning directed acyclic graphs from continuous data," *INFORMS J. Optim.*, vol. 3, no. 1, pp. 46–73, 2021.
[30] P.-L. Loh and P. Bühlmann, "High-dimensional learning of linear causal networks via inverse covariance estimation," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3065–3105, 2014.
[31] J. Xiang and S. Kim, "A* Lasso for learning a sparse Bayesian network structure for continuous variables," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 26, 2013, pp. 1–9.
[32] Z. Fang, S. Zhu, J. Zhang, Y. Liu, Z. Chen, and Y. He, "On low-rank directed acyclic graphs and causal structure learning," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, May 22, 2023, doi: 10.1109/TNNLS.2023.3273353.
[33] Y. Yu, J. Chen, T. Gao, and M. Yu, "DAG-GNN: DAG structure learning with graph neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7154–7163.
[34] N. R. Ke et al., "Learning neural causal models from unknown interventions," 2019, *arXiv:1910.01075*.
[35] S. Lachapelle, P. Brouillard, T. Deleu, and S. Lacoste-Julien, "Gradient-based neural DAG learning," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–23.
[36] M. Scanagatta, A. Salmerón, and F. Stella, "A survey on Bayesian network structure learning from data," *Progr. Artif. Intell.*, vol. 8, no. 4, pp. 425–439, 2019.
[37] B. Schölkopf et al., "Toward causal representation learning," *Proc. IEEE*, vol. 109, no. 5, pp. 612–634, May 2021.
[38] N. K. Kitson, A. C. Constantinou, Z. Guo, Y. Liu, and K. Chobtham, "A survey of Bayesian network structure learning," 2021, *arXiv:2109.11415*.
[39] M. J. Vowels, N. C. Camgoz, and R. Bowden, "D'ya like DAGs? A survey on structure learning and causal discovery," *ACM Comput. Surveys*, vol. 55, no. 4, pp. 1–36, 2022.
[40] A. Tong, L. Atanackovic, J. Hartford, and Y. Bengio, "Bayesian dynamic causal discovery," in *Proc. Causal View Dyn. Syst. NeurIPS Workshop*, 2022, pp. 1–14.
[41] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen, "A linear non-Gaussian acyclic model for causal discovery," *J. Mach. Learn. Res.*, vol. 7, no. 10, pp. 2003–2030, 2006.
[42] C. Maddison, A. Mnih, and Y. Teh, "The concrete distribution: A continuous relaxation of discrete random variables," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–20.
[43] E. Jang, S. Gu, and B. Poole, "Categorical reparametrization with Gumbel-softmax," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–13.
[44] M. Oberst and D. Sontag, "Counterfactual off-policy evaluation with Gumbel-max structural causal models," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 4881–4890.
[45] G. Lorberbom, D. D. Johnson, C. J. Maddison, D. Tarlow, and T. Hazan, "Learning generalized Gumbel-max causal mechanisms," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 26792–26803.
[46] K. Noorbakhsh and M. Rodriguez, "Counterfactual temporal point processes," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 24810–24823.
[47] V. Sindhwani, H. Q. Minh, and A. C. Lozano, "Scalable matrix-valued kernel learning for high-dimensional nonlinear multivariate regression and Granger causality," in *Proc. 29th Conf. Uncertain. Artif. Intell.*, 2013, pp. 586–595.
[48] A. Bolstad, B. D. Van Veen, and R. Nowak, "Causal network inference via group sparse regularization," *IEEE Trans. Signal Process.*, vol. 59, no. 6, pp. 2628–2641, Jun. 2011.
[49] S. Basu, A. Shojaie, and G. Michailidis, "Network Granger causality with inherent grouping structure," *J. Mach. Learn. Res.*, vol. 16, no. 1, pp. 417–453, 2015.
[50] S. Basu, X. Li, and G. Michailidis, "Low rank and structured modeling of high-dimensional vector autoregressions," *IEEE Trans. Signal Process.*, vol. 67, no. 5, pp. 1207–1222, Mar. 2019.
[51] M. Grant and S. Boyd. "CVX: MATLAB software for disciplined convex programming, version 2.1." 2014. [Online]. Available: http://cvxr.com/cvx/
[52] (Mosek ApS, Copenhage, Denmark). *The MOSEK Optimization Toolbox for Python Manual. Version 10.0*. (2019). [Online]. Available: https://docs.mosek.com/latest/pythonapi/index.html
[53] B. Efron, *Exponential Families in Theory and Practice*. Cambridge, U.K.: Cambridge Univ. Press, 2022.
[54] M. Lechner, "The relation of different concepts of causality used in time series and microeconometrics," *Economet. Rev.*, vol. 30, no. 1, pp. 109–127, 2010.
[55] Z. Zhang et al., "Truncated matrix power iteration for differentiable DAG learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 18390–18402.
[56] K. Bello, B. Aragam, and P. Ravikumar, "DAGMA: Learning dags via m-matrices and a log-determinant acyclicity characterization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 8226–8239.
[57] S. Wei and Y. Xie, "Causal structural learning from time series: A convex optimization approach," 2023, *arXiv:2301.11336*.
[58] M. A. Reyna et al., "Early prediction of sepsis from clinical data: The PhysioNet/computing in cardiology challenge 2019," *Crit. Care Med.*, vol. 48, no. 2, pp. 210–217, Feb. 2020.
[59] R. N. Khan, F. Saba, S. F. Kausar, and M. H. Siddiqui, "Pattern of electrolyte imbalance in type 2 diabetes patients: Experience from a tertiary care hospital," *Pakistan J. Med. Sci.*, vol. 35, no. 3, p. 797, 2019.
[60] K. Sterrenburg, W. Visser, L. Smit, and J. Cornette, "Acidosis: A potential explanation for adverse fetal outcome in intrahepatic cholestasis of pregnancy. A case report," *Obstetr. Med.*, vol. 7, no. 4, pp. 177–179, 2014.