Journal of Machine Learning Research 25 (2024) 1-56 Submitted 1/24; Revised 9/24; Published 10/24

DAG-Informed Structure Learning from Multi-Dimensional Point Processes

Chunming Zhang CMZHANG@STAT.WISC.EDU
Department of Statistics
University of Wisconsin-Madison
Madison, WI 53706, USA

Muhong Gao GAOMH@AMSS.AC.CN
Academy of Mathematics and System Science
Chinese Academy of Sciences
Beijing 100190, China

Shengji Jia 20200026@LIXIN.EDU.CN
School of Statistics and Mathematics
Shanghai Lixin University of Accounting and Finance
Shanghai, China

Editor: Jie Peng

Abstract

Motivated by inferring causal relationships among neurons using ensemble spike train data, this paper introduces a new technique for learning the structure of a directed acyclic graph (DAG) within a large network of events, applicable to diverse multi-dimensional temporal point process (MuTPP) data. At the core of MuTPP lie the conditional intensity functions, for which we construct a generative model parameterized by the graph parameters of a DAG and develop an equality-constrained estimator, departing from exhaustive search-based methods. We present a novel, flexible augmented Lagrangian (**Flex**-AL) optimization scheme that ensures provable global convergence and computational efficiency gains over the classical AL algorithm. Additionally, we explore causal structure learning by integrating acyclicity-constraints and sparsity-regularization. We demonstrate: (i) in cases without regularization, the incorporation of the acyclicity constraint is essential for ensuring DAG recovery consistency; (ii) with suitable regularization, the DAG-constrained estimator achieves both parameter estimation and DAG reconstruction consistencies similar to the unconstrained counterpart, but significantly enhances empirical performance. Furthermore, simulation studies indicate that our proposed DAG-constrained estimator, when appropriately penalized, yields more accurate graphs compared to unconstrained or unregularized estimators. Finally, we apply the proposed method to two real MuTPP datasets.

Keywords: asymptotic consistency; causal structure; constrained optimization; multivariate counting process; Structural Hamming Distance.

1 Introduction

The multi-dimensional temporal point process (MuTPP) model provides a probabilistic graphical framework for event occurrences observed in continuous time. Its applications

©2024 Chunming Zhang, Muhong Gao, and Shengji Jia.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v25/24-0067.html.

span various domains, including recordings of multiple neuronal spike trains, users' online ratings in social networks, and patients' treatment time records in hospitals, among others. An essential objective involves uncovering the network's structured causal relationships inherent in such MuTPP. Following terminology from causal inference and graphical models, this entails identifying a directed acyclic graph (DAG) $\mathcal{G}$. In this context, the jth node of $\mathcal{G}$ corresponds to the jth dimensional point process data observed within a specific time length, and every directed edge connecting two nodes in $\mathcal{G}$ signifies a particular type of causal effect from the parent node to the child node. The acyclicity assumption of $\mathcal{G}$ ensures that all nodes in the graph can be arranged in a causal order, providing a clear reflection of the causal dependencies among them. See examples in Loh and Bühlmann (2014); Pearl (2009); Shimizu et al. (2006). Practically, learning such a DAG structure from MuTPP yields interpretable insights. In our previous examples, the restored causal structure from neuronal spike train data aids in studying potential neuronal mechanisms in the brain and developing treatment strategies (Dobryakova et al., 2015; Fujisawa et al., 2008). Similarly, the acyclic recommendation propagation tree inferred from online rating data sets represents causal relationships between users, facilitating the design of recommendation systems in social networks (Yang et al., 2012).

In the context of continuous-time MuTPP, the issue of learning DAGs has been relatively unexplored or less understood in existing literature. Although a large body of work exists on causal structure within Gaussian graphical models, structural equation models (Shi and Li, 2022; Zheng et al., 2018), hidden Markov models (Nefian et al., 2002), and linear non-Gaussian acyclic model (Shimizu et al., 2006), these are based on discrete-time indexed observations of multivariate variables, making them unsuitable for modeling time-ordered event occurrence points $\{T_{j,\ell}\}_{\ell=1}^{\mathrm{N}_j}$ at nodes $j = 1, \ldots, d$, whose stochastic mechanism is driven by the 'conditional intensity function' (CIF) in the continuous-time domain, as detailed in Section 2.1. To effectively analyze MuTPP data, which typically lack strict-sense stationary, various continuous-time modeling approaches have emerged, such as the inhomogeneous Poisson process (Rajaram et al., 2005), the Cox model (Perry and Wolfe, 2013), and the Hawkes process (Hawkes, 1971; Reynaud-Bouret and Schbath, 2010; Xu et al., 2016). These methods offer estimated network graphs that reveal potential interactions and causal effects. However, these approaches do not account for the acyclicity assumption, leading to cycles (or self-loops) in the resulting network graphs and creating uncertainties regarding causal connections among nodes in these cycles. Therefore, it is imperative to explore new DAG-informed structure learning techniques aimed at better addressing the challenge of learning causal structure from MuTPP data.

This paper aims to bridge the gap in the literature by striving to achieve this goal. Focusing on the continuous-time CIFs governing the stochastic mechanisms of MuTPP, we build a stochastic generative model (4) in Section 2. This model represents non-linear mappings of historical events. The causal relationships among nodes are portrayed through interaction parameters within the generative model. The sign and magnitude of these parameters indicate the direction and strength of the corresponding effects. The aggregation of interaction parameters with non-zero values configures a causal DAG. For graph parameter estimation, we devise a likelihood-based estimation procedure, which incorporates two distinct types of structural constraints: (i) the DAG constraint, aimed at eliminating cycles in the obtained graph, and (ii) sparsity regularization, promoting the sparsity char-

acteristics of the resulting network. Simultaneously enforcing these restrictions ensures that the recovered network genuinely represents a sparse acyclic graph, which clearly delineates potential causal structures and offers practical guidance for real-world structure learning. Theoretically, we establish two types of consistencies (pertaining to parameter estimation and DAG recovery) for the proposed DAG-constrained regularized estimation method (see Theorem 3). These consistencies hold in a large-dimensional setting where the dimension of the MuTPP is allowed to grow with time length. Furthermore, we demonstrate that both types of consistencies can be achieved even when solely imposing the DAG constraint in estimation, without regularization. This holds true when the true network is sufficiently dense (see Theorem 2). For finite sample scenarios, simulation studies indicate that the proposed DAG-constrained regularized method outperforms the conventional unconstrained method (abbreviated to nonDAG) in terms of reconstruction accuracy (refer to Section 6). These findings provide both theoretical and empirical support for the validity of our proposed method, specifically affirming the reliability of utilizing the DAG constraint.

Enforcing the DAG structural constraint is crucial for accurately recovering the acyclic causal structure. However, simultaneously developing a nonlinear optimization strategy involving the DAG geometry constraint for parameter estimation remains a central challenge. Learning DAGs from data poses an NP-hard problem (Robinson, 1977), primarily due to the challenging enforcement of the combinatorial acyclicity constraint. Consequently, conventional search-based optimization algorithms (Chickering, 2002; Heckerman et al., 1995; Scanagatta et al., 2015) are impractical for large graphs. This work contributes to devising a new non-search-based 'flexible augmented Lagrangian' (**Flex**-AL) algorithm aimed at numerically solving the DAG-constrained optimization problem. Our **Flex**-AL algorithm efficiently addresses the equality-constrained program, converted from the original combinatorial DAG-constrained optimization problem. This derivation utilizes a new characterization result (Zheng et al., 2018) of the acyclicity constraint. Compared to the classical AL algorithm, our **Flex**-AL algorithm not only permits more flexible choices for the explicit forms of Lagrange multiplier and augmentation parameter updates but also exhibits computational efficiency with fewer iterations and reduced computational time. Refer to the discussions in Section 4.2 for further details.

The main contributions of this paper are outlined as follows:

- This paper represents the first systematic development in the literature of DAG-informed modeling, methodology, and theory for the structure learning task of MuTPP. Our proposed DAG-constrained regularized estimation method eliminates cycles in the reconstructed network and outperforms the existing approaches when applied to simulated point process data.

- Addressing computational challenges in the DAG-constrained optimization, we introduce a new efficient non-search-based algorithm termed '**Flex**-AL'. This algorithm is designed with a guaranteed global convergence to the true graph parameters and offers the improved flexibility and computational efficiency over conventional AL algorithms.

- We examine the statistical properties of DAG-constrained and unconstrained estimation methods, both in the presence and absence of regularization. We demonstrate that, in stark contrast to the nonDAG method, the proposed DAG-constrained method

reduces the error bound of 'Structural Hamming Distance' by at least $d(d-1)/2$ in the absence of regularization. Additionally, we establish that the proposed DAG-constrained method achieves consistencies in both parameter estimation and DAG recovery under scenarios (a) without regularization for a sufficiently dense underlying network and (b) with appropriate regularization.

- For the non-stationary MuTPP data driven by the continuous-time CIF model (4), our asymptotic analysis is derived as the continuous time length $\mathrm{T} \to \infty$. This scenario significantly differs from the conventional linear models for stationary or i.i.d. observations where the discrete sample size $n \to \infty$. Our proof techniques combine three distinct elements: (a) analysis of the probabilistic properties of MuTPP, (b) proofs of convergence for penalized M-estimators, and (c) theoretical explorations for the DAG-constraint. Integrating these separate aspects is a novel approach that contributes to the proofs of our main results.

- Our modeling approach and theoretical techniques partly leverage recent findings from Gao et al. (2024), but introduce significant improvements and distinctions. Computationally, addressing the DAG-constrained optimization problem is much more challenging than the unconstrained problem in Gao et al. (2024). However, our **Flex-AL** algorithm effectively handles this challenge, demonstrating superior performance. Theoretically, while Gao et al. (2024) focuses on the asymptotic properties of an unconstrained estimator in a fixed-dimension scenario, our paper provides a more comprehensive analysis. We examine the consistency of estimators with or without DAG constraints across networks with varying levels of sparsity and extend the asymptotic setting from $\mathrm{T} \to \infty$ to the diverging-dimension regime, allowing dimension d to grow with T.

The rest of the paper is organized as follows. Section 2 proposes a new generative model designed to capture causal structures within MuTPP. Section 3 details the proposed DAG-informed structure learning method, which is numerically solved using a new optimization algorithm presented in Section 4. Section 5 explores the theoretical properties of the proposed method. Section 6 provides simulation evaluations aimed at assessing the performance of our method. Section 7 showcases real datasets pertaining to neuronal spike trains and IPTV viewing records. Section 8 briefly concludes. Detailed numerical illustrations and technical derivations are collected in Appendices A, B, and C of a supplemental file.

2 Modeling causal structure in MuTPP

In this section, we aim to introduce the basic setting and notations relevant to multi-dimensional point processes (MuTPP). We proceed by building our generative model designed for learning the DAG-informed structure.

2.1 Multi-dimensional point processes

MuTPP denotes random processes detailing occurrences of specific events (e.g., instances of contagious diseases, neuron spike firing) observed in sequences $\boldsymbol{T}_1, \ldots, \boldsymbol{T}_d$ recorded at d

nodes. Here, each

$$\boldsymbol{T}_j = (T_{j,1},\ \ldots,\ T_{j,\mathrm{N}_j}) \quad \text{with}\, 0 < T_{j,1} < \cdots < T_{j,\mathrm{N}_j} \leq \mathrm{T}, \quad \text{for } j = 1, \ldots, d, \tag{1}$$

represents a sequence of time points $T_{j,\ell}$ corresponding to the ℓ-th event occurring at node j within an experiment of time length T. The associated counting process $N_j(t) = \sum_{\ell \geq 1} \mathrm{I}(0 \leq T_{j,\ell} \leq t)$ tallies the number of events occurring up to and including time t for node j, where $\mathrm{I}(\cdot)$ denotes the indicator operator. An important goal is to uncover the causal structure among nodes from d sequences of time series. To characterize the stochastic nature of event arrival times $\{T_{j,\ell}\}_{\ell=1}^{\mathrm{N}_j}$ in (1), we employ the concept of 'conditional intensity function' (CIF) $\lambda_j(t \mid \mathscr{F}_t)$ (also referenced in Rubin (1972)). This CIF quantifies the instantaneous event occurrence rate at node j. Formally, it is defined as:

$$\lambda_j(t \mid \mathscr{F}_t) = \lim_{\Delta \downarrow 0} \frac{1}{\Delta} \mathrm{P}\big(N_j(t+\Delta) = N_j(t) + 1 \mid \mathscr{F}_t\big) \tag{2}$$

$$= \lim_{\Delta \downarrow 0} \frac{1}{\Delta} \mathrm{P}\big(N_j(t+\Delta) \neq N_j(t) \mid \mathscr{F}_t\big), \quad \text{a.s., for } t \in [0, \mathrm{T}]. \tag{3}$$

Here, $\mathscr{F}_t = \sigma(\{N_j(s) : s \in [0,t],\, j = 1, \ldots, d\})$ represents the sigma-field generated by the event history of all d nodes up to and including time t.

For comparison, the commonly used one-dimensional Poisson process is a specific case within the counting process when $d = 1$ and entails a deterministic CIF. For a more comprehensive overview of point processes and probabilistic properties, refer to Daley and Vere-Jones (2003).

2.2 Modeling causal structure via DAGs

We introduce a general framework to model the dynamics of continuous-time CIFs $\{\lambda_j(t \mid \mathscr{F}_t)\}_{j=1,\ldots,d}$ in (2) and (3) as functions reliant on historical events:

$$\lambda_j(t \mid \mathscr{F}_t) = \exp\big\{w_{0,j} + \textstyle\sum_{i=1}^{d} w_{i,j}\, x_i(t)\big\}, \quad j = 1, \ldots, d, \quad t \in [0, \mathrm{T}]. \tag{4}$$

Here are the explicit interpretations of $w_{0,j}$, $w_{i,j}$, and $x_i(t)$:

- $w_{0,j}$ represents the baseline parameter denoting the background intensity of node j and serves as the intercept within the generative model.
- $w_{i,j}$ denotes the interaction parameter signifying the direction and strength of influence from node i on node j. Specifically, $w_{i,j} > 0$, $w_{i,j} < 0$, and $w_{i,j} = 0$ respectively express *excitatory*, *inhibitory*, and *no* effects. The magnitude of $w_{i,j}$ indicates the strength of this influence.
- The regression covariate $x_i(t)$ acts as the event activity of node i within a short period preceding t. Particularly, we set $x_i(t)$ as:

$$x_i(t) = g\big(N_i((t-\phi, t])/\phi\big), \tag{5}$$

  where $(t-\phi, t]$ denotes the 'lag window' with a width $\phi \in (0, \infty)$. The function $g(\cdot) : [0,\infty) \mapsto [0,\infty)$ is termed the 'shape function', which is non-negative, continuous,

monotonically increasing, and satisfies $g(0) = 0$. In (5), the ratio $N_i((t-\phi, t])/\phi$ represents an empirical estimate of the CIF $\lambda_i(u \mid \mathscr{F}_u)$ during a short interval $u \in (t-\phi, t]$. The shape function $g(\cdot)$ captures the potentially non-linear relationship between the empirical rate and the regression covariate. Practical examples of shape functions include $g(x) = x$ and $g(x) = \log(1+x)$. In our parameter estimation procedure outlined in Section 3, we presume knowledge of both the function $g(\cdot)$ and the constant ϕ. In real applications, $g(\cdot)$ and ϕ could be chosen either through data-driven methods or by referring to specific domain knowledge.

To clarify, our CIF $\lambda_j(t \mid \mathscr{F}_t)$ in model (4) with covariates $x_i(t)$ in (5) is right-continuous in $t \in [0, \infty)$. This aligns with our definition of $\mathscr{F}_t$-measurable CIF $\lambda_j(t \mid \mathscr{F}_t)$ in (2) and (3), where $\mathscr{F}_t$ represents the event history up to and including the current time t. Similar approaches involving right-continuous CIFs can be found in Carstensen et al. (2010); Kass et al. (2014). For modeling MuTPP, there is another line of discretized approaches (Truccolo et al., 2005; Zhang et al., 2016; Zhao et al., 2012) that transform the observed sequence of time points into event counts in discrete time bins and then fit the data via GLM-parameterized Bernoulli or Poisson distributions. In contrast, our approach directly models the continuous-time CIFs associated with the MuTPP, without the need for data transformation or preprocessing. This continuous-time approach has the advantage of preserving the complete information from the original MuTPP data, thus avoiding the information loss caused by approximation error from data transformation; refer to Gao et al. (2024) for more discussions and numerical comparisons.

The weighted adjacency matrix $\mathbf{W} = (w_{i,j}) \in \mathbb{R}^{d \times d}$, derived from model (4), encapsulates the directional causal relationships among nodes. This matrix further induces a causal directed graph $\mathcal{G} = \mathcal{G}(\mathbf{W}) = (\mathcal{V}, \mathcal{E})$, consisting of the node set $\mathcal{V} = \{1, \ldots, d\}$ and the edge set $\mathcal{E} = \mathcal{E}(\mathbf{W}) = \{(i,j) \in \mathcal{V} \times \mathcal{V} : i \neq j;\, w_{i,j} \neq 0\}$ representing directed edges. Each edge (i,j) carries a non-zero causal effect from node i to node j with a specific orientation. Formally, we call $\mathcal{G}(\mathbf{W})$ a directed acyclic graph (DAG) if, for any $k \geq 2$, there exist no k-cycle-inducing edges $(i_1, i_2), \ldots, (i_{k-1}, i_k), (i_k, i_1)$ that all belong to $\mathcal{E}(\mathbf{W})$. Figure 9 (right panel) illustrates an example of a DAG with 8 nodes. Throughout this paper, we assume that $\mathcal{G}(\mathbf{W})$ forms a DAG devoid of directed cycles. In addition, we assume that there is no self-effect, i.e., $w_{i,i} = 0$ for $i = 1, \ldots, d$, because self-effects are typically represented as 'self-loops' in graphical models. The above assumptions are compactly written as $\mathcal{G}(\mathbf{W}) \in \mathbb{D}$, where $\mathbb{D}$ denotes the space of non-self-loop DAGs, in accordance with condition A5 outlined in Section 5. The imposition of acyclicity on graphs is a common practice in literature concerning causal graphical models (e.g., Loh and Bühlmann, 2014; Pearl, 2009; Shimizu et al., 2006). Lemmas C.4 and C.5 in Appendix C.1 verify the identifiability of model (4) with respect to all parameters $\{w_{i,j} : i = 0, 1, \ldots, d;\ j = 1, \ldots, d\}$.

DAGs are fundamental concepts in causal inference. The original definition of causal structures, initially introduced in Pearl (2009) (page 44, Definition 2.2.1), is essentially grounded on DAGs. DAG-based models have been widely used to represent causal or temporal relationships among data variables (Loh and Bühlmann, 2014; Shimizu et al., 2006; Van de Geer and Bühlmann, 2013). DAGs also have specific meaningful interpretations for specific application fields. For example, in the neurology study of brain functional connectivity, DAG models provide a precise depiction of the directed information flow across brain regions (Biswas and Shlizerman, 2022; Zhang et al., 2022b), further helping to better

understand the brain functional integration. In social epidemiology, DAGs are popular models for representing the potential sources and paths of disease transmission in human populations (Ackley et al., 2021; Chen et al., 2021).

3 DAG-informed structure learning

Our objective is to estimate the weighted adjacency matrix $\mathbf{W}$ representing edge parameters from the observed point process data $\{\boldsymbol{T}_j\}_{j=1,\dots,d}$ across d nodes. This estimation will enable the discovery of the causal DAG structure underlying the d-dimensional point processes. For notational clarity, the edge parameter matrix $\mathbf{W}$ is denoted as $\mathbf{W} = (\boldsymbol{w}_{\bullet,1}, \dots, \boldsymbol{w}_{\bullet,d}) \in \mathbb{R}^{d\times d}$, where each column vector is $\boldsymbol{w}_{\bullet,j} = (w_{1,j}, \dots, w_{d,j})^\top$. Consequently, the generative model (4) for the CIF is reformulated as:

$$\lambda_j(t \mid \mathscr{F}_t) = \exp\big\{\widetilde{\boldsymbol{w}}_{\bullet,j}^\top \widetilde{\boldsymbol{x}}(t)\big\}, \quad j = 1, \dots, d, \quad t \in [0, \mathrm{T}]. \tag{6}$$

Here, $\widetilde{\boldsymbol{w}}_{\bullet,j} = (w_{0,j}, w_{1,j}, \dots, w_{d,j})^\top = (w_{0,j}, \boldsymbol{w}_{\bullet,j}^\top)^\top \in \mathbb{R}^{d+1}$ represents the vector of all parameters linked to $\lambda_j(t \mid \mathscr{F}_t)$. Additionally, $\widetilde{\boldsymbol{x}}(t) = (x_0(t), x_1(t), \dots, x_d(t))^\top \in \mathbb{R}^{d+1}$ stands for the vector of regression covariates, where $x_0(t) \equiv 1$. All pertinent parameters are collected in a parameter matrix $\widetilde{\mathbf{W}} = (\widetilde{\boldsymbol{w}}_{\bullet,1}, \dots, \widetilde{\boldsymbol{w}}_{\bullet,d}) \in \mathbb{R}^{(d+1)\times d}$, which can also be represented as $\widetilde{\mathbf{W}} = (\boldsymbol{w}_{0,\bullet}, \mathbf{W}^\top)^\top$, where $\boldsymbol{w}_{0,\bullet} = (w_{0,1}, \dots, w_{0,d})^\top$ denotes the vector of baseline parameters.

We propose a DAG-constrained regularized maximum-likelihood approach for structure learning. Building upon Rubin (1972) and (6), the negative log-likelihood concerning event occurrence times $\{T_{j,\ell}\}_{\ell=1}^{\mathrm{N}_j}$ in (1) can be expressed as:

$$\begin{aligned}
\mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\bullet,j}) &= -\frac{1}{\mathrm{T}} \int_0^{\mathrm{T}} \big[\log\{\lambda_j(t- \mid \mathscr{F}_{t-})\}\, \mathrm{d}N_j(t) - \lambda_j(t \mid \mathscr{F}_t)\, \mathrm{d}t\big] \\
&= -\frac{1}{\mathrm{T}} \int_0^{\mathrm{T}} \big[\widetilde{\boldsymbol{w}}_{\bullet,j}^\top \widetilde{\boldsymbol{x}}(t-)\, \mathrm{d}N_j(t) - \exp\{\widetilde{\boldsymbol{w}}_{\bullet,j}^\top \widetilde{\boldsymbol{x}}(t)\}\, \mathrm{d}t\big],
\end{aligned} \tag{7}$$

where $\lambda_j(t- \mid \mathscr{F}_{t-}) = \lim_{u\uparrow t} \lambda_j(t \mid \mathscr{F}_t)$ and $x(t-) = \lim_{u\uparrow t} x(u)$ represent the left limits. The proposed DAG-constrained regularized-MLE (Maximum Likelihood Estimator) is formulated as:

$$\begin{aligned}
\widehat{\widetilde{\mathbf{W}}} = \quad & \arg\min_{\widetilde{\mathbf{W}}\in\mathbb{R}^{(d+1)\times d}} \quad \mathcal{L}(\widetilde{\mathbf{W}}) + \mathcal{P}(\mathbf{W}; \boldsymbol{\eta}), \\
& \text{subject to} \qquad\quad \mathcal{G}(\mathbf{W}) \in \mathbb{D},
\end{aligned} \tag{8}$$

where

$$\mathcal{L}(\widetilde{\mathbf{W}}) = \textstyle\sum_{j=1}^d \mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\bullet,j}) \tag{9}$$

represents the joint negative log-likelihood function. The term $\mathcal{P}(\mathbf{W}; \boldsymbol{\eta})$ is some regularization term aiming at encouraging sparsity in the estimated network graph. In this context, we utilize the weighted L_1-penalty:

$$\mathcal{P}(\mathbf{W}; \boldsymbol{\eta}) = \sum_{1\le i\ne j\le d}\sum \eta_{i,j}\, |w_{i,j}|, \quad \text{with } \eta_{i,j} \ge 0, \tag{10}$$

where the vector $\boldsymbol{\eta} = ((\eta_{1,2}, \ldots, \eta_{1,d}), (\eta_{2,1}, \ldots, \eta_{2,d}), \ldots, (\eta_{d,1}, \ldots, \eta_{d,d-1}))^\top \in [0, \infty)^{d^2-d}$ contains non-negative regularization parameters. Note that the L_1-penalty $\mathcal{P}(\mathbf{W}; \boldsymbol{\eta}) = \eta \sum \sum_{1 \le i \ne j \le d} |w_{i,j}|$, utilized in prior works (Pamfil et al., 2020; Shi and Li, 2022; Zheng et al., 2018), represents a special case of (10) where $\eta_{i,j} \equiv \eta$ for all i, j. Solving (8) presents significant challenges due to the non-separability of its components regarding individual $\widetilde{\boldsymbol{w}}_{\cdot,j}$, as the DAG constraint pertains to the entire $\mathcal{G}(\mathbf{W})$.

4 Optimization algorithm for DAG-constrained estimator

This section develops our algorithm for solving the constrained optimization problem (8). Many existing methods address the DAG constraint by searching over the $\mathbb{D}$-space, such as order search (Fu and Zhou, 2013; Scanagatta et al., 2015) and greedy search (Chickering, 2002; Heckerman et al., 1995). However, these search-based techniques often encounter computational inefficiencies, particularly with large DAGs, since the search space $\mathbb{D}$ is combinatorial and scales super-exponentially with the number d of nodes (Robinson, 1977). To improve efficiency, we leverage a necessary and sufficient condition (Zheng et al., 2018) that characterizes the acyclicity of $\mathcal{G}(\mathbf{W})$ through a closed-form equality constraint, $h(\mathbf{W}) = 0$, where

$$h(\mathbf{W}) := \operatorname{trace}\{\exp(\mathbf{W} \circ \mathbf{W})\} - d, \tag{11}$$

with the Hadamard product operator '$\circ$' and the matrix exponential '$\exp(A)$' on a matrix A. The DAG-ness function $h(\mathbf{W})$ is non-negative, continuous, and differentiable, quantifying the degree of 'DAG-ness' in the graph $\mathcal{G}(\mathbf{W})$. As $h(\mathbf{W})$ approaches zero, $\mathcal{G}(\mathbf{W})$ more closely resembles a DAG. This insight enables us to transform the original combinatorial optimization problem (8) into an equality-constrained program:

$$\begin{array}{ll} \min\limits_{\widetilde{\mathbf{W}} \in \mathbb{R}^{(d+1)\times d}} & \mathcal{L}(\widetilde{\mathbf{W}}) + \mathcal{P}(\mathbf{W}; \boldsymbol{\eta}), \\ \text{subject to} & h(\mathbf{W}) = 0. \end{array} \tag{12}$$

In the context of the current point process application, both the loss function $\mathcal{L}(\widetilde{\mathbf{W}})$ in (9) and the penalty term $\mathcal{P}(\mathbf{W}; \boldsymbol{\eta})$ in (10) are convex, while the DAG-ness function $h(\mathbf{W})$ is non-convex (refer to Lemma C.9 in Appendix C). Consequently, (12) minimizes a convex function $\mathcal{L}(\widetilde{\mathbf{W}}) + \mathcal{P}(\mathbf{W}; \boldsymbol{\eta})$ over a non-convex set $\{\widetilde{\mathbf{W}} \in \mathbb{R}^{(d+1)\times d} : h(\mathbf{W}) = 0\}$.

For solving general types of equality-constrained optimization problems, the augmented Lagrangian (AL) method (Bertsekas, 2014) has proven successful and has widespread usage in the literature (e.g., Pamfil et al., 2020; Zhang et al., 2022a; Zheng et al., 2018). Specifically, for the DAG equality-constraint in (12), coupled with a non-negative non-convex DAG-ness function $h(\mathbf{W})$, we will devise an enhanced '**Flex**-AL' updating scheme in Section 4.2. This scheme is motivated by the classical AL method ('Clas-AL') reviewed in Section 4.1; an unconstrained subproblem is discussed in Section 4.3.

4.1 Classical AL algorithm for solving (12)

The classical AL method operates as a dual ascent algorithm for solving an unconstrained optimization problem. Corresponding to (12), the unconstrained objective function is ex-

pressed as:

$$L(\widetilde{\mathbf{W}}, \alpha; \rho) = \mathcal{L}(\widetilde{\mathbf{W}}) + \mathcal{P}(\mathbf{W}; \boldsymbol{\eta}) + \alpha\, h(\mathbf{W}) + 2^{-1}\rho\, h^2(\mathbf{W}), \tag{13}$$

where $\alpha \in \mathbb{R}$ is the 'Lagrange multiplier', and $\rho > 0$ denotes the 'augmentation parameter'. With a predetermined ρ, let

$$\widetilde{\mathbf{W}}_\alpha = \arg \min_{\widetilde{\mathbf{W}} \in \mathbb{R}^{(d+1)\times d}} L(\widetilde{\mathbf{W}}, \alpha; \rho). \tag{14}$$

This minimizes (13) over $\widetilde{\mathbf{W}}$, resulting in the dual function of α as:

$$D(\alpha) = L(\widetilde{\mathbf{W}}_\alpha, \alpha; \rho). \tag{15}$$

The conventional AL method tackles (12) by initially seeking the maximizer $\widehat{\alpha}$ of the dual function:

$$\widehat{\alpha} = \arg \max_{\alpha \in \mathbb{R}} D(\alpha). \tag{16}$$

Subsequently, it pursues the minimizer in (14) with α replaced by this $\widehat{\alpha}$, yielding the estimator:

$$\widehat{\widetilde{\mathbf{W}}} = \widetilde{\mathbf{W}}_{\widehat{\alpha}}. \tag{17}$$

It's noteworthy that $L(\widetilde{\mathbf{W}}, \alpha; \rho)$ is linear in α with a straightforward derivative $\frac{\partial L(\widetilde{\mathbf{W}}, \alpha; \rho)}{\partial \alpha} = h(\mathbf{W})$. Due to this linearity, the dual problem (16) can be solved via the gradient ascent update:

$$\alpha \leftarrow \alpha + \beta\, h(\mathbf{W}_\alpha), \quad \text{with a step size } \beta > 0. \tag{18}$$

This AL method, amalgamating the method of Lagrange multipliers (LM) (Bertsekas, 2014) and the penalty method (PM) (Bertsekas, 1975), consolidates their advantages. In contrast to the LM method, the AL method incorporates an augmentation term $\rho/2\, h^2(\mathbf{W})$ in (13), which not only ensures the existence of the minimizer $\widetilde{\mathbf{W}}_\alpha$ for (14) but also accelerates the numerical convergence speed for solving (14). When compared with the PM method, the AL method obviates the necessity of increasing ρ to infinity; this avoidance prevents ill-conditioned solutions $\widetilde{\mathbf{W}}_\alpha$ in (14) due to an excessively large ρ.

Regarding the augmentation parameter ρ in the AL algorithm, the convergence of the AL algorithm can be guaranteed if $\rho > \rho_0$, where $\rho_0 \in (0, \infty)$ is a problem-specific threshold value $\rho_0 \in (0, \infty)$ (Theorem 17.5 in Nocedal and Wright (1999)). In practice, this ρ_0 is unknown. Therefore, it might be more desirable to gradually increase ρ until $\rho > \rho_0$ is achieved. Utilizing this dynamic ρ, the following procedure summarizes the iterative updates for $\widetilde{\mathbf{W}}$, α, and ρ in the AL algorithm, performed at steps $k = 0, 1, \ldots$:

Update $\widetilde{\mathbf{W}}$ by solving the unconstrained subproblem:

$$\widetilde{\mathbf{W}}^{(k+1)} = \arg \min_{\widetilde{\mathbf{W}} \in \mathbb{R}^{(d+1)\times d}} L(\widetilde{\mathbf{W}}, \alpha^{(k)}; \rho^{(k)}), \tag{19}$$

where the kth step solution $\widetilde{\mathbf{W}}^{(k)}$ is used as the initial value.

Update α and ρ according to the rules:

$$\alpha^{(k+1)} = \alpha^{(k)} + \beta_\alpha \, h(\mathbf{W}^{(k+1)}), \tag{20}$$

$$\rho^{(k+1)} = \beta_\rho \, \rho^{(k)}, \tag{21}$$

where $\beta_\alpha > 0$ and $\beta_\rho \geq 1$ are appropriately chosen step sizes.

4.2 Proposed Flex-AL algorithm for solving (12)

The convergence property of the AL algorithm with updates (19), (20), and (21) is established in Nocedal and Wright (1999) under general cases of constraint functions. In the context of our optimization problem (12), the DAG-ness function $h(\mathbf{W})$ is non-negative, resulting in both α and ρ being non-decreasing with iteration steps in the updates (20) and (21). This distinct feature inspires us to investigate the extent to which rules (20) and (21) could be relaxed in updating α and ρ. As we will show in Theorem 1, the convergence of the AL algorithm could be guaranteed with any updates of α and ρ fulfilling the condition:

$$\begin{aligned} &\max\{\alpha^{(k)}, \rho^{(k)}\} \to \infty \text{ as } k \to \infty, \\ \text{and} \quad &\inf_{k\geq 0} \alpha^{(k)} > -\infty, \quad \inf_{k\geq 0} \rho^{(k)} > 0. \end{aligned} \tag{22}$$

In this regard, we devise the 'flexible augmented Lagrangian' (**Flex**-AL) algorithm, referring to the enhanced version of the AL algorithm which offers more flexibility in the updating scheme of α and ρ satisfying condition (22), in contrast to the classical AL ('Clas-AL') updating rules (20) and (21).

Theorem 1 (Global convergence of the Flex-AL algorithm for solving (12)) *Assume that the sequence of Lagrange multipliers $\{\alpha^{(k)}\}_{k\geq 0}$ and the sequence of augmentation parameters $\{\rho^{(k)}\}_{k\geq 0}$ satisfy condition (22). Then we have the following results:*

(i) *Any limit point of the sequence of global minimizers $\{\widetilde{\mathbf{W}}^{(k)}\}_{k\geq 1}$ in (19) is a global minimizer of (12).*

(ii) *Moreover, assume condition A6′ in Section 5. Then $\{\widetilde{\mathbf{W}}^{(k)}\}_{k\geq 1}$ has at least one limit point. Also, if (12) has a unique global minimizer $\widehat{\widetilde{\mathbf{W}}}$, then we have*

$$\widetilde{\mathbf{W}}^{(k)} \to \widehat{\widetilde{\mathbf{W}}} \quad \text{as } k \to \infty.$$

It's worth noticing that our condition (22) is relatively mild and easy to fulfill, granting the **Flex**-AL algorithm more flexibility in choosing updates for α and ρ compared to the Clas-AL algorithm. A concrete example of the **Flex**-AL updates is given by:

$$\alpha^{(k+1)} = \gamma_\alpha \, \alpha^{(k)}, \tag{23}$$

$$\rho^{(k+1)} = \gamma_\rho \, \rho^{(k)}, \tag{24}$$

where $\gamma_\alpha > 1$ and $\gamma_\rho > 1$ are step sizes, and $\alpha^{(0)} > 0$ and $\rho^{(0)} > 0$ are initial values. Particularly, the updated formula (23) for α in the proposed **Flex**-AL algorithm offers several

advantages over the (20) update in the Clas-AL algorithm. Firstly, (23) runs faster, especially in larger networks involving many nodes (d), as it eliminates the costly evaluation of the $d \times d$ matrix $\mathbf{W}$ in the DAG-ness function $h(\mathbf{W})$ present in (20). Simulation experiments in Section 6.5 illustrate that the **Flex**-AL algorithm achieves convergence with fewer iterations and reduced computational time compared to the Clas-AL algorithm. Secondly, the simplicity of (23) simplifies the selection of an appropriate step size γ_α for α, allowing practitioners to balance computational speed and algorithmic stability on a case-by-case basis.

Besides (23)–(24), the **Flex**-AL algorithm accommodates various other update forms that adhere to condition (22). Due to their computational expediency, we will utilize (23)–(24) for numerical experiments in Section 6 and real data analysis in Section 7. It's important to note that Theorem 1 implies that using either (23) or (24) exclusively while keeping the other fixed could also ensure convergence; however, our simulation experiments suggest that using both (23) and (24) concurrently offer faster and more stable performance in practice. Based on our simulations, excessively small values of γ (i.e., γ_α or γ_ρ) in (23) and (24) can reduce the algorithm's convergence speed, while overly large values γ might accelerate α and ρ excessively, potentially leading to ill-conditioned solutions in (19). For practical applications, choosing γ within the range of $[3, 20]$ has proven effective.

Another practical consideration for our **Flex**-AL algorithm involves the stopping criterion. As α and ρ may not reach infinity during practical computations, the iterative sequence $\{\widetilde{\mathbf{W}}^{(k)}\}_{k\geq 1}$ might not precisely converge to the solution of (12), thereby not exactly satisfying the acyclicity constraint $h(\mathbf{W}) = 0$. To address this, we opt to terminate the algorithm when $h(\mathbf{W}^{(k)}) < \epsilon_h$ is attained at a certain step $k = \widehat{k}$, using a small tolerance $\epsilon_h > 0$, close to machine precision (e.g., $\epsilon_h = 10^{-5}$) or specified by the user. Consequently, the resulting weighted adjacency matrix $\mathbf{W}^{(\widehat{k})}$ approximately, albeit not perfectly, represents a DAG. For the final output, a thresholding procedure $\mathbf{W}^{(\widehat{k})} \circ \mathrm{I}(|\mathbf{W}^{(\widehat{k})}| > \omega)$ is applied, with a small constant $\omega > 0$, to eliminate edges inducing cycles. Simulation results indicate that a very small ω, such as $\omega = 0.01$, effectively achieves this objective. The complete procedure of the **Flex**-AL algorithm is summarized in Algorithm 1.

4.3 Proximal quasi-Newton algorithm for solving the unconstrained subproblem (19)

To facilitate the derivation, we convert matrices $\mathbf{W}$ and $\widetilde{\mathbf{W}}$ to the vectorized versions, represented as: $\boldsymbol{w} = (\boldsymbol{w}_{\bullet,1}^\top, \ldots, \boldsymbol{w}_{\bullet,d}^\top)^\top = \mathrm{vec}(\mathbf{W}) \in \mathbb{R}^{d^2}$ and $\widetilde{\boldsymbol{w}} = (\widetilde{\boldsymbol{w}}_{\bullet,1}^\top, \ldots, \widetilde{\boldsymbol{w}}_{\bullet,d}^\top)^\top = \mathrm{vec}(\widetilde{\mathbf{W}}) \in \mathbb{R}^{d^2+d}$, respectively. Functions $\mathcal{L}(\widetilde{\mathbf{W}})$, $h(\mathbf{W})$, and $\mathcal{P}(\mathbf{W}; \boldsymbol{\eta})$ are denoted by $\mathcal{L}(\widetilde{\boldsymbol{w}})$, $h(\boldsymbol{w})$, and $\mathcal{P}(\boldsymbol{w}; \boldsymbol{\eta})$, respectively. The unconstrained subproblem (19) is equivalently rewritten as a composite minimization problem:

$$\min_{\widetilde{\boldsymbol{w}} \in \mathbb{R}^{d^2+d}} f(\widetilde{\boldsymbol{w}}) + \mathcal{P}(\boldsymbol{w}; \boldsymbol{\eta}), \tag{25}$$

where

$$f(\widetilde{\boldsymbol{w}}) = \mathcal{L}(\widetilde{\boldsymbol{w}}) + \alpha\, h(\boldsymbol{w}) + 2^{-1} \rho\, h^2(\boldsymbol{w}) \tag{26}$$

represents the smooth part of the objective function, and $\mathcal{P}(\boldsymbol{w}; \boldsymbol{\eta}) = \sum \sum_{1 \leq i \neq j \leq d} \eta_{i,j} |w_{i,j}|$ is the non-smooth part. Utilizing $\nabla h(\mathbf{W}) = 2 \{\exp(\mathbf{W} \circ \mathbf{W})\}^\top \circ \mathbf{W}$ (Zheng et al., 2018) as

the gradient of $h(\mathbf{W})$ with respect to $\mathbf{W}$, the smooth part $f(\widetilde{\boldsymbol{w}})$ in (26) has the following closed-form gradient vector:

$$\begin{aligned}\nabla f(\widetilde{\boldsymbol{w}}) &= (\nabla \mathcal{L}_{1,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\bullet,1})^\top, \ldots, \nabla \mathcal{L}_{d,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\bullet,d})^\top)^\top \\ &\quad + 2\{\alpha + \rho\, h(\boldsymbol{w})\} \operatorname{vec}\left(\{\exp(\mathbf{W} \circ \mathbf{W})\}^\top \circ \mathbf{W}\right),\end{aligned} \tag{27}$$

where

$$\nabla \mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\bullet,j}) = \frac{1}{\mathrm{T}} \int_0^{\mathrm{T}} \left[\widetilde{\boldsymbol{x}}(t) \exp\{\widetilde{\boldsymbol{w}}_{\bullet,j}^\top \widetilde{\boldsymbol{x}}(t)\}\, \mathrm{d}t - \widetilde{\boldsymbol{x}}(t-)\, \mathrm{d}N_j(t)\right], \quad j = 1, \ldots, d, \tag{28}$$

represents the gradient vector of $\mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\bullet,j})$. However, a closed-form expression for computing the Hessian matrix $\nabla^2 h(\boldsymbol{w})$ of the DAG-ness function $h(\boldsymbol{w})$ in (26) is not available.

Despite significant advancements in devising efficient algorithms for solving a composite minimization problems, they are ineffective for our optimization task (25) due to two primary reasons. (i) The substantial number of parameters (of the order $O(d^2)$) in (25) leads to slow convergence of coordinate descent (CD) based methods (including the stochastic coordinate descent (Shalev-Shwartz and Tewari, 2009) and the blockwise coordinate descent (Simon et al., 2013; Wright, 2012). These methods necessitate repetitive evaluations of the partial derivative at each single coordinate. (ii) Newton-type methods (Hsieh et al., 2011; Yuan et al., 2012), which evaluate the gradient only once in each outer iteration, require the computation of the exact Hessian matrix of $f(\widetilde{\boldsymbol{w}})$. However, this is analytically intractable in our case.

In this paper, we adopt the 'proximal quasi-Newton' (PXQN) method (Zhong et al., 2014) to solve (25). Starting with the iterative solution $\widetilde{\boldsymbol{w}}^{(k)}$ at the kth step, the PXQN algorithm computes the subsequent step update $\widetilde{\boldsymbol{w}}^{(k+1)}$ via the backtracking line search procedure:

$$\widetilde{\boldsymbol{w}}^{(k+1)} = \widetilde{\boldsymbol{w}}^{(k)} + \beta^{(k)}\, \widetilde{\boldsymbol{d}}^{(k)},$$

where the descent direction $\widetilde{\boldsymbol{d}}^{(k)}$ and the step size $\beta^{(k)}$ are determined as follows. The descent direction $\widetilde{\boldsymbol{d}}^{(k)}$ minimizes the regularized quadratic function:

$$\widetilde{\boldsymbol{d}}^{(k)} = \arg\min_{\widetilde{\boldsymbol{d}} \in \mathbb{R}^{d^2+d}} \left\{\nabla f(\widetilde{\boldsymbol{w}}^{(k)})^\top \widetilde{\boldsymbol{d}} + 2^{-1} \widetilde{\boldsymbol{d}}^\top B^{(k)} \widetilde{\boldsymbol{d}} + \mathcal{P}(\boldsymbol{w}^{(k)} + \boldsymbol{d}; \boldsymbol{\eta})\right\}. \tag{29}$$

Here, the matrix $B^{(k)}$ represents the L-BFGS approximation (Nocedal and Wright, 1999) to the Hessian matrix $\nabla^2 f(\widetilde{\boldsymbol{w}}^{(k)})$, and $\nabla f(\widetilde{\boldsymbol{w}}^{(k)})^\top \widetilde{\boldsymbol{d}} + 2^{-1} \widetilde{\boldsymbol{d}}^\top B^{(k)} \widetilde{\boldsymbol{d}}$ is a quadratic approximation of the smooth part $f(\widetilde{\boldsymbol{w}}^{(k)} + \widetilde{\boldsymbol{d}})$. Problem (29) can be efficiently solved by applying an inner coordinate descent algorithm, where the update for each coordinate has a simple closed-form expression. Upon determining the direction vector $\widetilde{\boldsymbol{d}}^{(k)}$, the step size $\beta^{(k)}$ is obtained by searching over grids $\{1, r, r^2, \ldots\}$ for some $r \in (0, 1)$ until the Armijo rule is met:

$$f(\widetilde{\boldsymbol{w}}^{(k)} + \beta^{(k)}\, \widetilde{\boldsymbol{d}}^{(k)}) \leq f(\widetilde{\boldsymbol{w}}^{(k)}) + \beta^{(k)}\, \sigma \Delta^{(k)}, \tag{30}$$

where $\sigma \in (0, 1)$ represents a slope constant, and $\Delta^{(k)} = \nabla f(\widetilde{\boldsymbol{w}}^{(k)})^\top \widetilde{\boldsymbol{d}}^{(k)} + \mathcal{P}(\boldsymbol{w}^{(k)} + \beta^{(k)}\, \boldsymbol{d}^{(k)}; \boldsymbol{\eta}) - \mathcal{P}(\boldsymbol{w}^{(k)}; \boldsymbol{\eta})$. Algorithm 2 outlines the steps of our PXQN algorithm for solving (25).

Due to the non-convexity of $h(\cdot)$, the objective function in (25) is also non-convex. Consequently, our PXQN algorithm may not always converge to the global minimizer of (25) and could end up at a local minimizer. Nevertheless, our simulation experiments in Section 6 show that the PXQN algorithm remains effective, efficiently and accurately solving the large-scale optimization problem (25) despite these non-convexity challenges. In essence, our PXQN algorithm doesn't require recurrent calculations of coordinate-wise gradients, enabling much faster convergence compared to CD-based methods. Unlike Newton-type method reliant on exact Hessian matrix evaluations, our PXQN algorithm utilizes an approximate proxy, denoted as $B^{(k)}$.

4.4 Exact computation of loss function and its gradient

In our PXQN algorithm, the loss function $\mathcal{L}(\widetilde{\boldsymbol{w}})$ and its gradient $\nabla\mathcal{L}(\widetilde{\boldsymbol{w}})$ are iteratively evaluated at $\widetilde{\boldsymbol{w}} = \widetilde{\boldsymbol{w}}_k$ during each step k. Therefore, the algorithm's efficacy depends crucially on the efficient computation of $\mathcal{L}(\widetilde{\boldsymbol{w}})$ and $\nabla\mathcal{L}(\widetilde{\boldsymbol{w}})$. Unlike the simple quadratic loss functions used in Pamfil et al. (2020); Zheng et al. (2018) for learning discrete-time time-series data, our loss function in (7) and (9) for MuTPP involves a complex continuous integral over time t and event counts $N_j(t)$. Accurately and efficiently computing these integrals for MuTPP is more challenging than in the referenced works. Nevertheless, we resolve this challenge by developing an efficient exact computing method for our integral-based loss functions. Referring to a recent work (Gao et al., 2024) (Lemma 3, eq. (22), our CIF $\{\lambda_j(t \mid \mathscr{F}_t)\}_{j=1,\dots,d}$ modeled in (4) are piecewise-constant functions in t, with all discontinuity points listed in the sequence

$$\{\breve{T}_1, \breve{T}_2, \dots, \breve{T}_{\mathrm{N}}\} := \bigcup_{j\in\mathcal{V}} \big\{\{T_{j,\ell}\}_{1\le\ell\le \mathrm{N}_j} \cup \{T_{j,k}+\phi\}_{1\le k\le \mathrm{N}_j}\big\},$$

where $0 < \breve{T}_1 < \breve{T}_2 < \cdots < \breve{T}_{\mathrm{N}} < \mathrm{T}$ are arranged in increasing order. Denote $\breve{T}_0 = 0$ and $\breve{T}_{\mathrm{N}+1} = \mathrm{T}$. By partitioning the time axis into sub-intervals $\{(\breve{T}_\ell, \breve{T}_{\ell+1}]\}_{0\le\ell\le\mathrm{N}}$ according to these discontinuity points, we can recast the integral of loss function (7) or its gradient (28) as the summation of integrals in each sub-interval, namely,

$$\begin{aligned}
\mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\bullet,j}) = &-\mathrm{T}^{-1}\textstyle\sum_{\ell=0}^{\mathrm{N}} \big\{\widetilde{\boldsymbol{w}}_{\bullet,j}^\top \widetilde{\boldsymbol{x}}(\breve{T}_\ell)\cdot \mathrm{I}(\breve{T}_{\ell+1}\in\{T_{j,k}\}_{1\le k\le \mathrm{N}_j}) \\
&- \exp\big\{\widetilde{\boldsymbol{w}}_{\bullet,j}^\top \widetilde{\boldsymbol{x}}(\breve{T}_\ell)\big\}\cdot(\breve{T}_{\ell+1}-\breve{T}_\ell)\big\}, \\
\nabla\mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\bullet,j}) = &-\mathrm{T}^{-1}\textstyle\sum_{\ell=0}^{\mathrm{N}} \widetilde{\boldsymbol{x}}(\breve{T}_\ell)\cdot \big\{\mathrm{I}(\breve{T}_{\ell+1}\in\{T_{j,k}\}_{1\le k\le \mathrm{N}_j}) \\
&- \exp\big\{\widetilde{\boldsymbol{w}}_{\bullet,j}^\top \widetilde{\boldsymbol{x}}(\breve{T}_\ell)\big\}\cdot(\breve{T}_{\ell+1}-\breve{T}_\ell)\big\}.
\end{aligned}$$

Using the above expressions, $\mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\bullet,j})$ and $\nabla\mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\bullet,j})$ can be directly and exactly computed without the need for Monte Carlo approximations for numerical integrals. This further enhances the efficiency and accuracy of implementing the PXQN Algorithm 2 as well as the **Flex**-AL Algorithm 1.

5 Statistical properties of structure learning methods

In this section, we explore the statistical properties of our proposed DAG-constrained estimator (8) for inferring causal structure from MuTPP data, allowing the dimension d to grow with the total time length T. Let $\mathbf{W}^* = (w^*_{i,j}) \in \mathbb{R}^{d\times d}$ and $\widetilde{\mathbf{W}}^* = (\boldsymbol{w}^*_{0,\bullet}, \mathbf{W}^{*\top})^\top \in \mathbb{R}^{(d+1)\times d}$

denote the true weighted adjacency matrix and the true parameter matrix, respectively. We utilize the commonly used metric, 'Structural Hamming Distance (SHD)', to evaluate the performance of network structure recovery (Pamfil et al., 2020; Zhang et al., 2022a; Zheng et al., 2018). In our context, SHD quantifies edge insertions, deletions, or flips needed to transform the estimated network $\mathcal{G}(\widehat{\mathbf{W}})$ into the true network $\mathcal{G}(\mathbf{W}^*)$. It is formally defined as:

$$\begin{aligned}\mathrm{SHD}(\widehat{\mathbf{W}}, \mathbf{W}^*) &= \sum_{1\le i\neq j\le d}\sum \mathrm{I}\big(\mathrm{sign}(\widehat{w}_{i,j}) \neq \mathrm{sign}(w^*_{i,j})\big) \\ &= \mathrm{Ex}(\widehat{\mathbf{W}}, \mathbf{W}^*) + \mathrm{Mi}(\widehat{\mathbf{W}}, \mathbf{W}^*) + \mathrm{Rv}(\widehat{\mathbf{W}}, \mathbf{W}^*), \end{aligned} \tag{31}$$

where the sign function $\mathrm{sign}(x)$ equals $+1$ if $x > 0$, 0 if $x = 0$, and -1 if $x < 0$;

$\mathrm{Ex}(\widehat{\mathbf{W}}, \mathbf{W}^*) = \sum\sum_{1\le i\neq j\le d} \mathrm{I}(\widehat{w}_{i,j} \neq 0,\, w^*_{i,j} = 0)$ counts *extra* edges (or 'false positives'), i.e., detected edges not existing in the true graph;

$\mathrm{Mi}(\widehat{\mathbf{W}}, \mathbf{W}^*) = \sum\sum_{1\le i\neq j\le d} \mathrm{I}(\widehat{w}_{i,j} = 0,\, w^*_{i,j} \neq 0)$ counts *missing* edges (or 'false negatives'), i.e., true edges not detected;

$\mathrm{Rv}(\widehat{\mathbf{W}}, \mathbf{W}^*) = \sum\sum_{1\le i\neq j\le d}\{\mathrm{I}(\widehat{w}_{i,j} > 0,\; w^*_{i,j} < 0) + \mathrm{I}(\widehat{w}_{i,j} < 0,\, w^*_{i,j} > 0)\}$ counts *reversed* edges between detected and true edges.

Two types of consistencies are examined:

$$\text{For parameter estimation: } \|\widehat{\widetilde{\mathbf{W}}} - \widetilde{\mathbf{W}}^*\|_{\mathrm{F}} \xrightarrow{\mathrm{P}} 0 \text{ as } \mathrm{T} \to \infty, \tag{32}$$

$$\text{For DAG recovery: } \quad \mathrm{P}(\mathrm{SHD}(\widehat{\mathbf{W}}, \mathbf{W}^*) = 0) \to 1 \text{ as } \mathrm{T} \to \infty, \tag{33}$$

where '$\|\cdot\|_{\mathrm{F}}$' represents the Frobenius norm of a matrix. In this section, we employ the Frobenius error as the evaluation metric for parameter estimation consistency, due to the correspondence $\|\widetilde{\mathbf{W}}\|_{\mathrm{F}} = \|\widetilde{\boldsymbol{w}}\|_2$, where the Euclidean norm $\|\cdot\|_2$ is a widely-used metric for vector magnitude in statistical literatures.

For brevity and a unified treatment of estimation methods dealing with structural constraints and sparsity features, we introduce an expanded estimator:

$$\widehat{\widetilde{\mathbf{W}}}_{\kappa,\boldsymbol{\eta}} = \arg\min_{\widetilde{\mathbf{W}}\in\mathbb{R}^{(d+1)\times d}:\, h(\mathbf{W})\le\kappa} \mathcal{L}(\widetilde{\mathbf{W}}) + \mathcal{P}(\mathbf{W}; \boldsymbol{\eta}). \tag{34}$$

This estimator includes two subscripts, κ and $\boldsymbol{\eta}$:

the scalar κ, equal to 0 or ∞, acts as a bound parameter for the DAG-ness function $h(\cdot)$,

the regularization parameter vector $\boldsymbol{\eta}$ relates to the weighted L_1-penalty $\mathcal{P}(\mathbf{W}; \boldsymbol{\eta})$ in (10).

In line with (34),

$\kappa = 0$ enforces the DAG constraint in the estimation, whereas $\kappa = \infty$ releases the DAG constraint;

$\boldsymbol{\eta} = \mathbf{0}$ excludes the use of regularization, while $\boldsymbol{\eta} \neq \mathbf{0}$ incorporates regularized estimation.

Specifically: $\widehat{\widetilde{\mathbf{W}}}_{0,\boldsymbol{\eta}}$ coincides with our proposed DAG-constrained regularized MLE estimator $\widehat{\widetilde{\mathbf{W}}}$ in (8), and $\widehat{\widetilde{\mathbf{W}}}_{\infty,\boldsymbol{\eta}}$ reduces to the routine nonDAG regularized MLE estimator.

By enforcing the DAG constraint $h(\mathbf{W}) = 0$, it is expected that falsely-detected cycle-inducing edges can be eliminated, potentially bringing the estimated graph closer to the true DAG. This prompts a natural question: How does the DAG-constrained estimator $\widehat{\widetilde{\mathbf{W}}}_{0,\boldsymbol{\eta}}$ enhance reconstruction accuracy compared to the nonDAG estimator $\widehat{\widetilde{\mathbf{W}}}_{\infty,\boldsymbol{\eta}}$? We explore this inquiry under two scenarios for $\boldsymbol{\eta}$:

(i) $\boldsymbol{\eta} = \mathbf{0}$ in (34), indicating unregularized estimation of the true DAG network $\mathcal{G}(\mathbf{W}^*)$, especially in scenarios involving a large number of edges.

(ii) $\boldsymbol{\eta} \neq \mathbf{0}$ in (34), implying regularized estimation of the true DAG network $\mathcal{G}(\mathbf{W}^*)$, particularly in scenarios involving a small number of edges.

Remark 1 *A recent work* (Chen et al., 2017) *derived similar consistency results for the L_1-regularized non-constrained least-square estimator of network parameters for the non-linear Hawkes process. However, the proof techniques therein are not applicable to our case. Apart from our model and loss function being different from those in* (Chen et al., 2017)*, a major distinctive challenge in our theoretical analysis comes from the use of the* DAG *constraint. Clearly, the* DAG *space $\mathbb{D}$ is a complex, non-convex subspace of $\mathbb{R}^{d\times d}$ with intricate topological properties. In addition, $\widehat{\mathbf{W}}_{0,\boldsymbol{\eta}}$ is not simply a projection of $\widehat{\mathbf{W}}_{\infty,\boldsymbol{\eta}}$ onto $\mathbb{D}$, making the relationship between the estimation errors $\|\widehat{\mathbf{W}}_{0,\boldsymbol{\eta}}-\mathbf{W}^*\|_{\mathrm{F}}$ and $\|\widehat{\mathbf{W}}_{\infty,\boldsymbol{\eta}}-\mathbf{W}^*\|_{\mathrm{F}}$ quite complex. Thus, enforcing the* DAG *constraint essentially influences the theoretical properties of the resulting network estimator and consequently presents significant challenges in the theoretical analysis of statistical consistency. To the authors' knowledge, there is no existing work that applies similar types of structural constraints to multivariate point process models. Our theories in this section are innovative in this research field.*

For clarity, the technical conditions are listed and discussed below.

A1. In multi-dimensional point processes of dimension $d \geq 2$, event occurrence time points $\{T_{j,\ell}\}_{j=1,\ldots,d;\,\ell=1,\ldots,\mathrm{N}_j}$ satisfy $0 < T_{j,1} < \cdots < T_{j,\mathrm{N}_j} \leq \mathrm{T}$ for each $j = 1, \ldots, d$.

A2. For the asymptotic setting in Section 5, the dimension $d = d_{\mathrm{T}}$ is allowed to grow with the time length T. For notational simplicity, we always abbreviate d_{T} as d. Moreover, we assume that $d^4/\mathrm{T} \to 0$ as $\mathrm{T} \to \infty$.

A3. In (5), the 'shape function' $g(\cdot) : [0, \infty) \to [0, \infty)$ is non-negative, continuous, monotonically increasing, with $g(0) = 0$, and bounded from above, i.e., $\sup_{x\in[0,\infty)} g(x) \leq c_0$ for some constant $c_0 \in (0, \infty)$.

A4. The parameter space Θ of $\widetilde{\mathbf{W}}^*$ is compact in $\mathbb{R}^{(d+1)\times d}$. For each $j = 1, \ldots, d$, $\|\widetilde{\boldsymbol{w}}^*_{\cdot,j}\|_1 \leq c_1$ for some constant $c_1 \in (0, \infty)$.

A5. The true weighted adjacency matrix $\mathbf{W}^*$ of the underlying directed graph is a DAG containing no self-loops, i.e., $w^*_{i,i} = 0$ for $i = 1, \ldots, d$. This condition is equivalent to $h(\mathbf{W}^*) = 0$. Also, $\mathcal{E}(\mathbf{W}^*) \neq \varnothing$.

A6. For each $j = 1, \ldots, d$, $\lambda_{\min}(\nabla^2 \mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}^*_{\bullet,j})) \geq C$ for some positive constant $C \in (0, \infty)$, with probability tending to 1.

A6′. For each $j = 1, \ldots, d$, $\inf_{\widetilde{\boldsymbol{w}}_{\bullet,j} \in \mathbb{R}^{d+1}} \lambda_{\min}(\nabla^2 \mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\bullet,j})) \geq C$ for some positive constant $C \in (0, \infty)$.

A7. The true weighted adjacency matrix $\mathbf{W}^*$ satisfies that $\sqrt{\mathrm{T}/d^2} \cdot \min_{(i,j) \in \mathcal{E}(\mathbf{W}^*)} |w^*_{i,j}| \to \infty$, as $\mathrm{T} \to \infty$.

A8. For every $j = 1, \ldots, d$, the $(d+1)$-by-1 random vector $\widehat{\widetilde{\boldsymbol{w}}}_{\bullet,j} = (\widehat{w}_{0,j}, \widehat{w}_{1,j}, \ldots, \widehat{w}_{d,j}) = \arg\min_{\widetilde{\boldsymbol{w}}_{\bullet,j} \in \mathbb{R}^{d+1}} \mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\bullet,j})$ satisfies that $\widehat{w}_{i,j} \neq 0$ with probability one for each $i = 0, 1, \ldots, d$.

Condition A1 relates to the fundamental definition of a multivariate point process. Condition A2 specifies the diverging-dimension setting for our asymptotic results. Conditions A3 and A4 ensure that the true CIF $\lambda^*_j(t \mid \mathscr{F}_t)$ under model (4) is uniformly bounded from above, i.e., $\sup_{j=1,\ldots,d;\, t \in [0,\mathrm{T}]} \lambda^*_j(t \mid \mathscr{F}_t) \leq c$ for some constant $c \in (0, \infty)$ not depending on T or d. Condition A5 relates to the formal definition of a DAG. Conditions A6 resembles the conventional bounded eigenvalue condition (e.g., condition (F) in Fan and Peng (2004), or condition A5 in Zhang et al. (2010)). Conditions A6′ is a sample-level version of condition A6 used for establishing the global convergence of **Flex**-AL algorithm (Theorem 1 (ii)). Condition A7 states that the minimum signal strength $\min_{(i,j) \in \mathcal{E}(\mathbf{W}^*)} |w^*_{i,j}|$ is of higher order than the estimation error rate $O_{\mathrm{P}}(\sqrt{d^2/\mathrm{T}})$. This ensures that the signs of non-zero terms in $\mathbf{W}^*$ could be correctly recovered with probability approaching to 1. Condition A8 implies that the non-regularized non-constrained MLE estimator $\widehat{\widetilde{\boldsymbol{w}}}_{\bullet,j}$ is non-sparse. These conditions are not the weakest possible but facilitate the technical derivations.

We showcase an example of T, d, and $\widetilde{\mathbf{W}}^*$ that satisfy conditions A2, A4, A5, and A7. Let $d = \lceil \mathrm{T}^{1/5} \rceil$,

$$w^*_{0,j} = \mathrm{T}^{-1/5}, \quad w^*_{i,j} = \mathrm{T}^{-1/5} \cdot \mathrm{I}(i > j), \qquad \text{for } i, j = 1, \ldots, d.$$

Then, $d^4/\mathrm{T} = O(\mathrm{T}^{-1/5}) \to 0$ verifies condition A2; $\|\widetilde{\boldsymbol{w}}^*_{\bullet,j}\|_1 \leq (d+1) \cdot \mathrm{T}^{-1/5} = O(1)$ verifies condition A4; $\mathbf{W}^*$ is a non-empty lower triangular matrix satisfying condition A5; and $\sqrt{\mathrm{T}/d^2} \cdot \min_{(i,j) \in \mathcal{E}(\mathbf{W}^*)} |w^*_{i,j}| = O(\mathrm{T}^{1/10}) \to \infty$ verifies condition A7.

5.1 Scenario (i): without regularization ($\boldsymbol{\eta} = \mathbf{0}$ in (34))

In this subsection, we consider the scenario where regularization is absent, i.e., $\boldsymbol{\eta} = \mathbf{0}$ and $\mathcal{P}(\mathbf{W}; \boldsymbol{\eta}) = 0$ in (34). Let $s^* = |\mathcal{E}(\mathbf{W}^*)| = \sum \sum_{1 \leq i \neq j \leq d} \mathrm{I}(w^*_{i,j} \neq 0)$ denote the number of edges in the true edge set $\mathcal{E}(\mathbf{W}^*)$. According to Bang-Jensen and Gutin (2008) (Proposition 2.1.3), any weighted adjacency matrix $\mathbf{W}^*$ of a DAG is a strictly upper triangular matrix, up to a permutation of rows and columns. Therefore, at least half of the off-diagonal elements in $\mathbf{W}^*$ are zeros, implying that $s^* \leq d(d-1)/2$; similar arguments used in (C.20).

Lemma 1 demonstrates that in the absence of regularization, the nonDAG unregularized estimator $\widehat{\widetilde{\mathbf{W}}}_{\infty,\mathbf{0}}$ corresponding to $\kappa = \infty$ and $\boldsymbol{\eta} = \mathbf{0}$ in (34) achieves parameter estimation consistency, but induces inconsistency in DAG recovery.

Lemma 1 (Inconsistency of (33) from nonDAG unregularized $\widehat{\widetilde{\mathbf{W}}}_{\infty,\mathbf{0}}$) *Assume conditions* A1–A6 *and* A7–A8 *in Section* 5. *For $\kappa = \infty$ and $\boldsymbol{\eta} = \mathbf{0}$ in* (34)*, if $d^4/\mathrm{T} \to 0$ as $\mathrm{T} \to \infty$, then there exists a global minimizer $\widehat{\widetilde{\mathbf{W}}}_{\infty,\mathbf{0}}$ of the* nonDAG *unregularized optimization problem* (34)*, such that $\|\widehat{\widetilde{\mathbf{W}}}_{\infty,\mathbf{0}} - \widetilde{\mathbf{W}}^*\|_{\mathrm{F}} = O_{\mathrm{P}}(\sqrt{d^2/\mathrm{T}})$, and the following relations hold with probability tending to* 1*:*

$$\begin{aligned} &\mathrm{Ex}(\widehat{\mathbf{W}}_{\infty,\mathbf{0}}, \mathbf{W}^*) = d(d-1) - s^*, \quad \mathrm{Mi}(\widehat{\mathbf{W}}_{\infty,\mathbf{0}}, \mathbf{W}^*) = 0, \quad \mathrm{Rv}(\widehat{\mathbf{W}}_{\infty,\mathbf{0}}, \mathbf{W}^*) = 0, \\ &\quad \textit{and} \quad \mathrm{SHD}(\widehat{\mathbf{W}}_{\infty,\mathbf{0}}, \mathbf{W}^*) = d(d-1) - s^* \geq 1. \end{aligned} \tag{35}$$

The square-root convergence rate $O_{\mathrm{P}}(\sqrt{d^2/\mathrm{T}})$ is a standard result for the MLE estimators in discrete time-indexed models (e.g., for i.i.d. observations in Fan and Peng (2004); Zhang et al. (2010)). However, such results have not been well studied in the context of continuous-time MuTPP data in (1), which typically do not exhibit strict-sense stationary. Examining (35) reveals that the SHD primarily originates from falsely detected edges, where $\mathrm{P}(\mathrm{SHD}(\widehat{\mathbf{W}}_{\infty,\mathbf{0}}, \mathbf{W}^*) = \mathrm{Ex}(\widehat{\mathbf{W}}_{\infty,\mathbf{0}}, \mathbf{W}^*)) \to 1$ as $\mathrm{T} \to \infty$. Since s^* should never exceed $d(d-1)/2$ for DAGs, result (35) further implies that

$$\mathrm{SHD}(\widehat{\mathbf{W}}_{\infty,\mathbf{0}}, \mathbf{W}^*) = d(d-1) - s^* \geq d(d-1)/2,$$

indicating that at least $d(d-1)/2$ falsely-detected edges in the estimated edge set $\mathcal{E}(\widehat{\mathbf{W}}_{\infty,\mathbf{0}})$. This explains why the nonDAG unregularized estimator $\widehat{\widetilde{\mathbf{W}}}_{\infty,\mathbf{0}}$ fails to achieve DAG recovery consistency.

In stark contrast to $\widehat{\widetilde{\mathbf{W}}}_{\infty,\mathbf{0}}$, the DAG-constrained counterpart $\widehat{\widetilde{\mathbf{W}}}_{0,\mathbf{0}}$ not only reduces the error bound of SHD by at least $d(d-1)/2$ but also maintains both parameter estimation and DAG recovery consistency, especially when the true network is sufficiently dense, as established in Theorem 2.

Theorem 2 (Consistency of (32)–(33) from DAG-constrained unregularized $\widehat{\widetilde{\mathbf{W}}}_{0,\mathbf{0}}$) *Assume conditions* A1–A6 *and* A7 *in Section* 5. *For $\kappa = 0$ and $\boldsymbol{\eta} = \mathbf{0}$ in* (34)*, if $d^4/\mathrm{T} \to 0$ as $\mathrm{T} \to \infty$, then there exists a global minimizer $\widehat{\widetilde{\mathbf{W}}}_{0,\mathbf{0}}$ of the* DAG*-constrained unregularized optimization problem* (34)*, such that $\|\widehat{\widetilde{\mathbf{W}}}_{0,\mathbf{0}} - \widetilde{\mathbf{W}}^*\|_{\mathrm{F}} = O_{\mathrm{P}}(\sqrt{d^2/\mathrm{T}})$, and the following relations hold with probability tending to* 1*:*

$$\begin{aligned} &\mathrm{Ex}(\widehat{\mathbf{W}}_{0,\mathbf{0}}, \mathbf{W}^*) \leq d(d-1)/2 - s^*, \quad \mathrm{Mi}(\widehat{\mathbf{W}}_{0,\mathbf{0}}, \mathbf{W}^*) = 0, \quad \mathrm{Rv}(\widehat{\mathbf{W}}_{0,\mathbf{0}}, \mathbf{W}^*) = 0, \\ &\quad \textit{and} \quad \mathrm{SHD}(\widehat{\mathbf{W}}_{0,\mathbf{0}}, \mathbf{W}^*) \leq d(d-1)/2 - s^*. \end{aligned} \tag{36}$$

Moreover, if the true graph is a dense DAG *satisfying $d(d-1)/2 - s^* = o(1)$, then we have* $\mathrm{P}(\mathrm{SHD}(\widehat{\mathbf{W}}_{0,\mathbf{0}}, \mathbf{W}^*) = 0) \to 1$.

After imposing the DAG constraint $h(\mathbf{W}) = 0$ (i.e., setting $\kappa = 0$ in (34)), the DAG-constrained unregularized estimator $\widehat{\widetilde{\mathbf{W}}}_{0,\mathbf{0}}$ retains the same convergence rate $O_{\mathrm{P}}(\sqrt{d^2/\mathrm{T}})$ as that of the nonDAG counterpart $\widehat{\widetilde{\mathbf{W}}}_{\infty,\mathbf{0}}$ in Lemma 1. Comparing (36) with (35), the reduction of SHD by at least $d(d-1)/2$ (i.e., half of the total number of edges) reflects the significant improvement resulting from enforcing the DAG constraint. Furthermore, $\widehat{\widetilde{\mathbf{W}}}_{0,\mathbf{0}}$ is capable of achieving DAG recovery consistency under sufficiently dense networks, even without utilizing regularization techniques.

5.2 Scenario (ii): with regularization ($\boldsymbol{\eta} \neq \mathbf{0}$ in (34))

Lemma 1 and Theorem 2 in Section 5.1 demonstrate that in the absence of regularization, DAG recovery consistency is only achieved when the true DAG is sufficiently dense. For sparse DAGs, regularization is a common technique to promote sparsity in the structure and ensure model selection consistency. Several recent works (Loh and Bühlmann, 2014; Nandy et al., 2018; Van de Geer and Bühlmann, 2013) have studied the DAG recovery consistency of DAG-constrained regularized estimators for learning linear structural equation models. However, these studies were limited to the context of multivariate Gaussian distribution for observed variables and, therefore, are not directly applicable to MuTPP models for counting process data as described in (7). Another recent work (Pamfil et al., 2020) developed a DAG learning method for SVAR models on time-series data, yet with no theoretical guarantees provided.

In this subsection, we investigate the consistency of our proposed DAG-constrained regularized estimator $\widehat{\widetilde{\mathbf{W}}}$ in (8), or equivalently, $\widehat{\widetilde{\mathbf{W}}}_{0,\boldsymbol{\eta}}$ in (34), which incorporates the weighted L_1-penalty $\mathcal{P}(\mathbf{W};\boldsymbol{\eta}) = \sum\sum_{1\leq i\neq j\leq d}\eta_{i,j}\,|w_{i,j}|$ specified in (10). Let $\mathcal{E}^c(\mathbf{W}) = \{(i,j) \in \mathcal{V}\times\mathcal{V} : i \neq j;\, w_{i,j} = 0\}$ be the complement of the edge set $\mathcal{E}(\mathbf{W})$. We outline some necessary regularity conditions for regularization parameters $\eta_{i,j}$ in $\boldsymbol{\eta}$:

$$\max_{(i,j)\in\mathcal{E}(\mathbf{W}^*)} \eta_{i,j} = O_{\mathrm{P}}(\sqrt{d^2/(s^*\mathrm{T})}), \tag{37}$$

$$\min_{(i,j)\in\mathcal{E}^c(\mathbf{W}^*)} \sqrt{\mathrm{T}/d^3}\,\eta_{i,j} \xrightarrow{\mathrm{P}} \infty \quad \text{as } \mathrm{T}\to\infty, \tag{38}$$

which share similarities with assumptions used in Fan and Peng (2004); Zhang et al. (2010). Lemma 2 presents the consistency results for the nonDAG regularized estimator $\widehat{\widetilde{\mathbf{W}}}_{\infty,\boldsymbol{\eta}}$ with $\kappa = \infty$ in (34).

Lemma 2 (Consistency of (32)–(33) from nonDAG regularized $\widehat{\widetilde{\mathbf{W}}}_{\infty,\boldsymbol{\eta}}$) *Assume conditions* A1–A6 *and* A7 *in Section* 5. *For $\kappa = \infty$ in* (34) *and $\boldsymbol{\eta}$ in* (34) *satisfying conditions* (37) *and* (38), *if $d^4/\mathrm{T} \to 0$ as $\mathrm{T}\to\infty$, then there exists a global minimizer $\widehat{\widetilde{\mathbf{W}}}_{\infty,\boldsymbol{\eta}}$ of the* nonDAG *regularized optimization problem* (34), *such that $\|\widehat{\widetilde{\mathbf{W}}}_{\infty,\boldsymbol{\eta}} - \widetilde{\mathbf{W}}^*\|_{\mathrm{F}} = O_{\mathrm{P}}(\sqrt{d^2/\mathrm{T}})$, and $\mathrm{P}(\mathrm{SHD}(\widehat{\widetilde{\mathbf{W}}}_{\infty,\boldsymbol{\eta}}, \mathbf{W}^*) = 0) \to 1$.*

Lemma 2 demonstrates that the nonDAG regularized estimator $\widehat{\widetilde{\mathbf{W}}}_{\infty,\boldsymbol{\eta}}$ achieves both parameter estimation consistency (32) and DAG recovery consistency (33), if the weighted L_1-penalty is equipped with appropriate regularization parameters in $\boldsymbol{\eta}$. Regarding the DAG-constrained regularized estimator $\widehat{\widetilde{\mathbf{W}}}_{0,\boldsymbol{\eta}}$, Theorem 3 justifies its capability to achieve both consistencies (32)–(33), when using the same weighted L_1-penalty.

Theorem 3 (Consistency of (32)–(33) from DAG-constrained regularized $\widehat{\widetilde{\mathbf{W}}}_{0,\boldsymbol{\eta}}$) *Assume conditions* A1–A6 *and* A7 *in Section* 5. *For $\kappa = 0$ in* (34) *and $\boldsymbol{\eta}$ in* (34) *satisfying conditions* (37) *and* (38), *if $d^4/\mathrm{T} \to 0$ as $\mathrm{T}\to\infty$, then there exists a global minimizer $\widehat{\widetilde{\mathbf{W}}}_{0,\boldsymbol{\eta}}$ of the* DAG-*constrained regularized optimization problem* (34), *such that $\|\widehat{\widetilde{\mathbf{W}}}_{0,\boldsymbol{\eta}} - \widetilde{\mathbf{W}}^*\|_{\mathrm{F}} = O_{\mathrm{P}}(\sqrt{d^2/\mathrm{T}})$, and $\mathrm{P}(\mathrm{SHD}(\widehat{\widetilde{\mathbf{W}}}_{0,\boldsymbol{\eta}}, \mathbf{W}^*) = 0) \to 1$.*

From an asymptotic perspective, Lemma 2 and Theorem 3 affirm that both estimators $\widehat{\widetilde{\mathbf{W}}}_{\infty,\boldsymbol{\eta}}$ and $\widehat{\widetilde{\mathbf{W}}}_{0,\boldsymbol{\eta}}$, with appropriate regularization parameters in $\boldsymbol{\eta}$, can achieve both consistencies in (32) and (33). For finite data samples, simulation experiments in Section 6 demonstrate that the DAG-constrained regularized estimator $\widehat{\widetilde{\mathbf{W}}}_{0,\boldsymbol{\eta}}$ benefits from conceivably better recovery accuracy than the nonDAG counterpart $\widehat{\widetilde{\mathbf{W}}}_{\infty,\boldsymbol{\eta}}$. This lends empirical support to the advantage of utilizing the DAG constraint in exploring the structure.

Our asymptotic theories in Section 5 are based on the diverging-dimension setting, where d grows with T as $\mathrm{T} \to \infty$. Beyond d, other model parameters, such as the true parameter matrix $\widetilde{\mathbf{W}}^*$ and the number s^* of non-zero parameters, also depend on d and thus vary with both d and T. In contrast to the fixed-dimension case in Gao et al. (2024), simultaneously managing multiple varying variables—d, $\widetilde{\mathbf{W}}^*$, and s^*—is far more complex and challenging. To address this, we impose conditions on these variables (see conditions A2, A4, A5, and A7 in Section 5) and employ two main techniques: (i) applying probabilistic results that are non-asymptotic in T and d, unaffected by the diverging-dimension setting (see proofs of our Lemmas C.1–C.4), and (ii) adapting proof techniques from the fixed-dimension setting to our case, such as those used in our Lemmas C.6 and C.7, which modify Lemma B.18 and Theorem 7 of Gao et al. (2024).

6 Simulation studies

To demonstrate the practical utility of our proposed DAG learning method, we conduct simulation experiments varying in scale, sparsity levels of DAGs and time lengths T of the temporal point process data.

6.1 Setup

We consider several types of true underlying DAGs:

Network 1: A small-scale sparse DAG with 10 nodes and 15 edges.

Network 2: A small-scale dense DAG with 10 nodes and 40 edges.

Network 3: A large-scale sparse DAG with 50 nodes and 190 edges.

Network 4: An extra-large sparse DAG with 100 nodes and 392 edges.

Here, 'small/large'-scale refers to the number of nodes in the true graph, and 'sparse/dense' relates to the number of true edges. These DAGs are visualized in Figure 1.

The point process data is generated using model (4) with true covariates:

$$\begin{aligned} x_i^*(t) = g\big(N_i((t-\phi,t])/\phi\big), \quad t \in [0,\mathrm{T}], \quad i = 1,\ldots,d, \\ \text{where } \phi = 1, \text{ and } g(x) = \log\{1+\min(x,10)\}. \end{aligned} \tag{39}$$

For each node $j = 1,\ldots,d$, the true baseline parameters are set as $w_{0,j}^* = -0.8$. For $i,j = 1,\ldots,d$, the true interaction parameters from node i to node j are $w_{i,j}^* = 0.5$ for an excitatory effect, $w_{i,j}^* = -0.5$ for an inhibitory effect, $w_{i,j}^* = 0$ for no effect. The total time length T is selected from grid points ranging between 300 and 1600.

We compare the proposed DAG-constrained method and the unconstrained method nonDAG using three choices of the penalty term: (i) No penalty: $\mathcal{P}(\mathbf{W};\boldsymbol{\eta}) = 0$ (i.e., $\eta_{i,j} \equiv 0$ in (10)). (ii) The L_1-penalty: $\mathcal{P}(\mathbf{W};\boldsymbol{\eta}) = \eta_1 \sum\sum_{1\le i\neq j\le d} |w_{i,j}|$ (i.e., $\eta_{i,j} \equiv \eta_1$ in (10)). (iii) The weighted L_1-penalty in (10):

$$\mathcal{P}(\mathbf{W};\boldsymbol{\eta}) = \sum_{1\le i\neq j\le d}\sum \eta_{i,j}\,|w_{i,j}|, \quad \text{with } \eta_{i,j} = \eta_2\,|\breve{w}_{i,j}|^{-2}, \tag{40}$$

where $\breve{w}_{i,j}$ is the MLE estimator of $w^*_{i,j}$, and $\eta_{i,j}$ in (40) adopts the weight $\eta = \lambda/|\breve{w}_{i,j}|^{\gamma}$ used by the adaptive lasso penalty (Zou, 2006), with $\gamma = 2$. Both tuning parameters η_1 and η_2 are chosen by minimizing the Bayesian Information Criterion (BIC) function (Nishii, 1984), widely used in model selection under large-dimensional settings (e.g., Tang and Li, 2021; Zhao et al., 2012). The methods' abbreviations are summarized in Table 1. The performance evaluation of each method includes error measures: 'false positive (Ex)', 'false negative (Mi)', 'structural hamming distance (SHD)', and 'false discovery rate (FDR)'. The first three error measures are defined in (31), while the FDR of an estimator $\widehat{\mathbf{W}}$ is defined as:

$$\mathrm{FDR}(\widehat{\mathbf{W}}, \mathbf{W}^*) = \{\mathrm{Ex}(\widehat{\mathbf{W}}, \mathbf{W}^*) + \mathrm{Rv}(\widehat{\mathbf{W}}, \mathbf{W}^*)\}/\max\{|\mathcal{E}(\widehat{\mathbf{W}})|,\, 1\}.$$

Here, $|\mathcal{E}(\widehat{\mathbf{W}})| = \sum\sum_{1\le i\neq j\le d} \mathrm{I}(\widehat{w}_{i,j} \neq 0)$, counting the number of identified edges.

Table 1: *Description of estimation methods in numerical studies*

abbreviation	description of estimation methods
'DAG-unreg'	DAG-constrained and unregularized MLE, $\widehat{\widehat{\mathbf{W}}}_{0,\mathbf{0}}$ in (34) without penalty.
'nonDAG-unreg'	nonDAG and unregularized MLE, $\widehat{\widehat{\mathbf{W}}}_{\infty,\mathbf{0}}$ in (34) without penalty.
'DAG-L_1'	DAG-constrained and L_1-regularized MLE, $\widehat{\widehat{\mathbf{W}}}_{0,\eta}$ in (34).
'nonDAG-L_1'	nonDAG and L_1-regularized MLE, $\widehat{\widehat{\mathbf{W}}}_{\infty,\eta}$ in (34).
'DAG-wL_1'	DAG-constrained and weighted-L_1-regularized MLE, $\widehat{\widehat{\mathbf{W}}}_{0,\boldsymbol{\eta}}$ in (34).
'nonDAG-wL_1'	nonDAG and weighted-L_1-regularized MLE, $\widehat{\widehat{\mathbf{W}}}_{\infty,\boldsymbol{\eta}}$ in (34).

6.2 Network 1 and Network 2

In this subsection, we present simulation results for the small-scale DAGs Networks 1 and 2 (illustrated in Figure 1).

Figure 2 displays the simulation results for Network 1. In learning this sparse network, the unregularized methods, namely 'nonDAG-unreg' and 'DAG-unreg', exhibit significantly inferior performance compared to the regularized methods, primarily due to a larger number of false positives (Ex). Therefore, Figure 2 uses log(Ex+1), log(Mi+1), and log(SHD+1) on the y-axes for various methods. Across all penalty choices, the DAG-constrained method outperforms its nonDAG counterpart concerning Ex, SHD, and FDR. However, for Mi,

the impact of adding the DAG constraint seems negligible. This indicates that the cyclicity constraint effectively reduces SHD and FDR, primarily by reducing the number of Ex. Comparing the two penalties, the weighted L_1-penalty generally exhibits better performance than the L_1-penalty in terms of SHD, while for Ex and FDR, the same conclusion holds under large time lengths T $\in \{900, 1200, 1500\}$. Notably, for 'nonDAG-wL_1' and 'DAG-wL_1' methods, SHDs approach zero with increasing total time length T, in line with our theoretical results (Lemma 2 and Theorem 3) on DAG recovery consistency in Section 5.2. While both the regularization technique and the DAG constraint can improve estimation accuracy, they do so through different structural impacts on the estimated network: regularization enhances the sparsity of the network, whereas the DAG constraint enforces the acyclicity property in the resulting causal graph. In practice, the weighted L_1-penalty and the DAG constraint are both indispensable and need to be used together to achieve the best result. As seen from Figure 2, 'DAG-wL_1' demonstrates the best overall performance, evidenced by the lowest SHD across all methods and various time lengths T.

For the dense Network 2, as depicted in Figure 3, the conclusions parallel those of Network 1. Irrespective of the sparsity level in the true graph, the DAG-constrained method consistently outperforms its nonDAG counterpart. Notably, the performance of the DAG-constrained unregularized method, 'DAG-unreg', significantly diminishes (in terms of Ex, SHD, and FDR) compared to the nonDAG counterpart, 'nonDAG-unreg'. Its performance becomes comparable to that of the regularized methods ('nonDAG-L_1', 'DAG-L_1', 'nonDAG-wL_1', and 'DAG-wL_1'). These results align with the theoretical findings (Lemma 1 and Theorem 2) from Section 5.1, indicating that the 'DAG-unreg' method achieves DAG recovery consistency for sufficiently dense networks. In contrast, the 'nonDAG-unreg' method consistently falls short of reaching this goal. Thus, imposing the DAG constraint significantly enhances DAG learning accuracy, particularly for dense DAGs.

6.3 Network 3

In Network 3, represented by a large sparse DAG with 50 nodes (as depicted in Figure 1), the simulation results are summarized in Figure 4. Similar to Network 1, the unregularized methods, namely 'nonDAG-unreg' and 'DAG-unreg', demonstrate significantly less effective performance compared to the regularized methods and are therefore omitted.

The overall conclusion, as depicted in Figure 4, largely concurs with the findings from Figures 2 and 3. The DAG-constrained methods, 'DAG-L_1' and 'DAG-wL_1', consistently outperform the unconstrained methods, 'nonDAG-L_1' and 'nonDAG-wL_1', in terms of SHD and FDR, particularly with a large time length T. This shows that incorporating the DAG constraint consistently enhances the accuracy of structure learning across networks of varying scales. An interesting finding from Figure 4 is the sharp decrease in SHD observed in the L_1-regularized methods, 'nonDAG-L_1' and 'DAG-L_1', from T = 800 to T = 1200. In contrast, the SHDs for 'nonDAG-wL_1' and 'DAG-wL_1' decrease steadily with increasing time length T. This observation suggests that the weighted L_1-penalty performs better and exhibits more stable behavior compared to the L_1-penalty, especially for large networks.

Figure 5 offers a comprehensive comparison of the estimated networks generated from all four methods for a randomly selected dataset with a time length T = 1200. Clearly, the DAG-constrained methods, 'DAG-L_1' and 'DAG-wL_1', produce sparser estimated graphs

with significantly smaller SHDs compared to those obtained from the nonDAG methods, 'nonDAG-L_1' and 'nonDAG-wL_1'. Similar comparative plots for Network 1 and Network 2 are provided in Figures 11 and 12 of supplementary Appendix A. These results confirm that applying the DAG constraint effectively reduces falsely-detected cycle-inducing edges, resulting in a more accurate reconstruction of the true DAG.

6.4 Network 4

To evaluate the scalability of our proposed method with larger network sizes, we conducted simulation experiments on Network 4, an extra-large sparse DAG comprising 100 nodes as depicted in Figure 1. The simulation results are summarized in Figure 6. Note that the number of model parameters is $100 \times 101 = 10,100$, while the sample size (the total number of time stamps $\sum_{j=1}^{d} \mathrm{N}_j$) for $\mathrm{T} \in [600, 2400]$ is typically between $30,000$ and $1,000,000$, exceeding the number of parameters. The conclusions drawn from Figure 6 align consistently with those derived from Networks 1, 2, and 3. The DAG-constrained methods, namely 'DAG-L_1' and 'DAG-wL_1', consistently outperform the unconstrained methods 'nonDAG-L_1' and 'nonDAG-wL_1' in terms of SHD and FDR, especially evident in scenarios with a larger time length T. This outcome validates the effectiveness of our proposed DAG-constrained method, showcasing its capability in achieving both estimation accuracy and computational efficiency, even when dealing with very large DAGs comprising approximately $10,000$ network parameters.

6.5 Computational time of **Flex**-AL algorithm

To evaluate the computational efficiency of our proposed **Flex**-AL algorithm in Section 4.2, we conducted additional simulation experiments comparing its runtime against the classical Clas-AL algorithm in Section 4.1. For each simulation replication, both algorithms are employed using the 'DAG-wL_1' and 'DAG-unreg' methods to estimate Network 3. The total time length of the synthetic point process data was fixed at $\mathrm{T} = 1200$. To ensure a fair comparison, identical step sizes $\{\beta_\alpha, \beta_\rho, \gamma_\alpha, \gamma_\rho\}$ were utilized in both algorithms (refer to (20), (21), (23), and (24)), specifically set to $\beta_\alpha = \beta_\rho = \gamma_\alpha = \gamma_\rho = 5$. Both algorithms adopted the same stopping rule described in Section 4.2, terminating once $h(\mathbf{W}^{(k)}) < \epsilon_h$ at a ceratin iteration step $k = \widehat{k}$, with $\epsilon_h = 10^{-5}$. Figure 7 presents the results derived from 100 replications.

The left panels display the Frobenius norm of the difference between the iterative update $\widetilde{\mathbf{W}}^{(k)}$ and the final output solutions $\widehat{\widetilde{\mathbf{W}}} = \widetilde{\mathbf{W}}^{(\widehat{k})}$ averaged across 100 replicate samples, against the iteration step k. It's evident that the **Flex**-AL algorithm achieves convergence in fewer iterations (averaging within 5 steps) compared to the Clas-AL algorithm (averaging more than 10 steps). The right panels depict boxplots showcasing the runtime for the final solutions $\widehat{\widetilde{\mathbf{W}}} = \widetilde{\mathbf{W}}^{(\widehat{k})}$ across 100 replications. Clearly, the **Flex**-AL algorithm achieves convergence faster than the Clas-AL algorithm. Comparable plots for Network 1 and Network 2 are given in Figures 13 and 14 in the supplementary appendix. These results offer convincing evidence of the computational efficiency of our **Flex**-AL algorithm, reinforcing the conclusions outlined in Section 4.2. To provide further numerical evidence, we consider an additional simulation scenario where the step size is fixed at $\gamma_\alpha = 5$ in **Flex**-AL, while

varying the step sizes $\beta_\alpha \in \{1, 5, 50, 500\}$ for the Clas-AL algorithm. Figure 15 displays the results for Network 2. It can be seen that **Flex**-AL achieves faster convergence with fewer iterations and less computational time compared to Clas-AL across all selected step sizes β_α. This result underscores the advantage of **Flex**-AL over Clas-AL, regardless of the step size values.

7 Real data analyses

In this section, we demonstrate the applications of our proposed method to two distinct real-world MuTPP datasets: one involving neuronal spike train data, and the other concerning an Internet Protocol television viewing record dataset.

7.1 Neuronal spike train data

We analyze the neuronal spike train dataset in Fujisawa et al. (2008), comprising multi-neuron recordings obtained from rats performing a working memory task. This dataset, available at `http://crcns.org/data-sets/pfc/pfc-2/about-pfc-2`, includes 89 recording sessions, each corresponding to an experimental period spanning approximately 23.37 hours. For consistency with previous findings, we select the session 'EE.188' used in Fujisawa et al. (2008). This session contains spike train data recorded from 117 isolated neurons in the rat medial prefrontal cortex and intermediate CA1 area of the hippocampus during a 46.87-minute working memory task. Before applying our method, we conduct a data cleaning procedure. Among the 117 neurons, 15 neurons (ID numbers: 4, 11, 15, 19, 25, 27, 46, 93, 126, 129, 151, 178, 209, 285, 290) either initiate spiking too late (after $t = 100$ sec) or cease spiking too early (before $t = 2600$ sec) and are thus excluded from our analysis. Subsequently, the cleaned dataset comprises a total of $629,800$ spike time stamps from the remaining 102 neurons within the time interval $[100, 2600]$ seconds, totally 2500 seconds in duration. Following the cleaning process, we partition the data into two sets: the training set encompassing spikes within the time range of $[100, 1600]$ seconds, and the testing set containing spikes between $[1600, 2600]$ seconds.

We employ both the 'nonDAG-wL_1' and 'DAG-wL_1' methods (as described in Table 1) to the training dataset. For both methods, we use a lag-width $\phi = 1$, considering that a neuron's spiking activity may influence other neurons within a short period of time-lag of up to one second, as discussed in Truccolo et al. (2005). Both methods utilize the shape function $g(x) = \log\{1 + \min(x, 10)\}$ and the weighted L_1-penalty as described in (39) and (40). Figure 8 presents the estimated network graphs alongside the corresponding heat maps. As anticipated, both methods tend to identify more excitatory effects than inhibitory ones. This aligns with the typical physiological configuration where the proportion of excitatory (pyramidal) neurons often exceeds that of inhibitory interneurons (refer to Fujisawa et al. (2008); Zhao et al. (2012)). Notably, specific neurons (ID number: 30, 52, 120, 135, 156, 201, 245, 295) exhibit significantly more excitatory effects from other neurons, hinting at their probable characterization as interneurons. This observation primarily echoes the outcomes presented in Figures 2 and 3 of Fujisawa et al. (2008). Comparing the heat map presented in Figure 3b of Fujisawa et al. (2008), where most identified effects are close to the diagonal, our heat maps in Figure 8 display a more evenly distribution of detected effects.

This disparity suggests that our methods might capture a broader spectrum of interactions within the neuron ensemble.

In Figure 8, the comparison between the 'nonDAG-wL_1' and 'DAG-wL_1' methods reveals that 'DAG-wL_1' yields a sparser graph compared to 'nonDAG-wL_1'. Due to the acyclicity constraint, the 'DAG-wL_1' method eliminates cycle-inducing edges, resulting in a network with fewer edges than the one obtained from the 'nonDAG-wL_1' method. To illustrate this contrast further, Figure 9 showcases the estimated sub-networks involving 8 selected neurons (ID numbers 52, 54, 72, 73, 135, 186, 245, 276). Clearly, the 'nonDAG-wL_1' method detects a number of cycles formed by mutually excitatory or inhibitory neurons (e.g., $52 \rightleftarrows 72$, $72 \rightleftarrows 135$, $52 \rightleftarrows 245$, $54 \rightleftarrows 276$), indicating potential 'co-firing' neurons; in contrast, the 'DAG-wL_1' method eliminates these cycles and obtains a DAG. Note that the co-firing pattern typically represents a high level of cross-correlation (Roux et al., 2022) between neurons, but does not imply any causal relationship. Thus, removing one or two edges from the mutual connections between these co-firing neurons may better identify the acyclic causal effect as well as the direction of information flow. A closer inspection reveals that the 'DAG-wL_1' method, in most times, removes the weaker effect and retains the stronger ones (e.g., $245 \to 186$, $135 \to 72$, $73 \to 245$). In some instances, the 'DAG-wL_1' method removes both effects between neurons (e.g., 52 and 72). By eliminating these cycles, 'DAG-wL_1' identifies causal chains among these neurons. For instance, $135 \to 72 \to 245 \to 186$ represents one such chain, demonstrating that causal effects propagate along this sequence. Note that in such causal chains, the detected causal effect flows from the neuron on the left (chain head) to the one on the right (chain tail), but not in the reverse direction. This demonstrates the directionality of causal relationships and the potential information flow within this neuron ensemble. It is important to note that our detected causal network basically represents the statistical causal relationships, which are different concepts from the actual neuronal connections in the brain. In this regard, our estimation results cannot be directly validated by biological facts. Nevertheless, our inferred causal network is valuable for assisting further research on neuronal behaviors and information transmission mechanisms in the brain.

To assess the goodness-of-fit, we use the predictive log-likelihood (PLL) metric, calculated as the log-likelihood of the estimated model fitted to the testing dataset. Table 2 displays the results for both methods, showing that the 'DAG-wL_1' method achieves a higher PLL compared to the 'nonDAG-wL_1' method. This indicates that employing the DAG constraint enhances model fitting and prediction performance. In summary, these findings underscore the advantage and utility of our proposed 'DAG-wL_1' method.

7.2 IPTV viewing record data

We test our proposed method on the IPTV (Internet Protocol television) viewing record dataset, which is publicly available and accessible at `https://ieee-dataport.org`. This dataset consists of logs of TV channel watching events from $13,246$ IPTV viewers in Guangzhou, P.R. China, during a one-month period in August 2014. Each data point corresponds to one watching session of one user, containing the information of the user ID (in the range of $[1, 13246]$), the channel ID (157 channel ID numbers in total, in the range of $[1, 817]$), the starting time point, and duration of this watching session. In the user viewing behavior study (e.g., Luo et al., 2015; Xu et al., 2016), it is commonly assumed that a user's

watching one type of channel would either prompt or inhibit the viewing of another channel a short period afterward. For example, viewing a sports channel may trigger users to watch a fitness channel afterwards, but might reduce the users' interests in watching a children's channel (since sport programs are typically watched by adults). Motivated by this aspect, we aim to explore the causal effects of viewing behaviors among different channels. To extract the MuTPP from the IPTV data, we first exclude those data points of watching sessions with durations less than 10 minutes, since a too-short duration is typically caused by some random switches of channels and may not reflect meaningful information about the user behavior. We then extract from the remaining data the ordered time sequence $\boldsymbol{T}_j = (T_{j,1}, T_{j,2}, \ldots, T_{j,\mathrm{N}_j})$ for each channel j, with each $T_{j,\ell} \in [0, 768]$ (hours) being the starting time point of one watching session. Among the 157 channels, 76 of them are considered as 'rarely watched', with less than 1000 data points in the extracted sequence $\boldsymbol{T}_j$ (i.e., with N_j less than 1000), and thus are excluded from our analysis. After this data cleaning and preprocessing procedure, we obtain our MuTPP data $\{\boldsymbol{T}_j\}_{j=1,\ldots,d}$ with a total number of $888,820$ timestamps, extracted from the remaining $d = 81$ channels in a period of length 768 hours. This cleaned data is split into the training set (containing timestamps in the first 400 hours) and the testing set (containing timestamps in the following 368 hours).

As a comparison, both the 'nonDAG-wL_1' and 'DAG-wL_1' methods are applied to the training data using the same covariates and weighted L_1-penalty as defined in (39) and (40), with $\phi = 1$ and $g(x) = \log\{1 + \min(x, 10)\}$. Figure 10 exhibits the estimated network graphs and the corresponding heat maps. It is observed from the top panels of Figure 10 that most of the detected effects in the two estimated networks are excitatory effects, while the number of detected inhibitory effects is much fewer. This is reasonable according to our conventional knowledge, as watching a certain type of TV program would more often trigger watching other related TV programs, but less often cause any inhibition. Among the 81 selected channels, 3 channels (with ID numbers 20, 78, 136) produce notably larger numbers of excitatory effects, that is, each of them triggers more than 10 other channels, as observed from both of the two detected networks. Therefore, it is anticipated that these 3 channels may broadcast some high-quality programs, e.g., popular live TV shows or newly released TV series, which trigger users to view the related replays, recaps, or news later on. For the comparison between the two estimation methods, 'DAG-wL_1' detects a sparser and clearer acyclic network, while 'nonDAG-wL_1' gives a relatively denser and more tangled network, in which a number of cycles or mutually connected pairs of nodes could be observed. These cycles may cause ambiguities as for which channels in the cycle are the authentic hosts of those high-quality programs that trigger users to watch other programs or enhance the TV rating points. In contrast, our proposed 'DAG-wL_1' method is able to eliminate these cycles in the detected network, and provides a more convincing inference result on the causal structure among the selected channels. Our detected DAG causal network provides various practical guidelines e.g., for TV directors to improve program qualities, and for ad agencies to find the best option of TV channels to advertise. Table 2 presents the PLL results of the two methods on the testing data to evaluate their model fitting performance. The results show that 'DAG-wL_1' outperforms 'nonDAG-wL_1' in terms of PLL, indicating its superior ability for model fitting and prediction. In summary, our proposed 'DAG-wL_1' method has demonstrated stronger interpretability, accuracy, and practicality in this case study on real-world IPTV data.

8 Discussion

This paper presents a novel investigation into the causal structure of continuous-time MuTPP, with large scales in dimensions and duration. Our approach operates under the belief that amidst all possible interactions between nodes, only a small fraction represent genuinely significant causal effects, forming a sparse DAG that encapsulates the complete causal structure. Our proposed method integrates the DAG structural constraint with sparsity-inducing regularization in the estimation procedure. This dual imposition ensures the estimated network is both acyclic and guarantees the insignificance of certain interactions by setting their parameters to precisely zero. We develop the computationally efficient **Flex**-AL algorithm designed to solve the resultant DAG equality-constrained optimization problem. Additionally, our work offers new theoretical insights into the proposed estimator's capabilities for graph parameter estimation and DAG reconstruction. Beyond illustrating its efficacy with examples from neuronal spike train data and the IPTV viewing record data, our method could be applicable to various real-world MuTPP datasets where a causal structure underlies node interactions.

Several challenging issues need further exploration: **(a)** The non-convex nature of the DAG-ness function $h(\boldsymbol{w})$ leads to a non-convex $f(\widetilde{\boldsymbol{w}})$ in the subproblem (25). The global convergence property of the PXQN algorithm under such conditions remains largely unclear in the existing literature. It is also pertinent to investigate whether and under what conditions Algorithm 2, aimed at solving the subproblem (25), maintains a super-linear convergence rate with an increasing number of iterations. **(b)** The combination of imposing the DAG constraint with the weighted L_1-penalty could enhance structure learning accuracy, as evidenced both empirically in the simulation studies of Section 6 and theoretically in the asymptotic results of Section 5, across multiple useful scenarios. Moreover, it is worth evaluating the impact of the DAG constraint in scenarios where the regularity conditions (37) and (38) for weights are relaxed or violated. Additionally, exploring other types of penalties beyond the weighted L_1-penalty in regularization could be insightful. **(c)** Our DAG causal graph is based on the classic setup of acyclic causal structure established in Pearl (2009), but such types of DAG assumptions may not necessarily hold for all actual situations. In certain real applications, special types of cyclic casual effects may also appear, e.g., the feedback cycles in economic processes (Spirtes, 2013). In such cases, it is of interest to relax the DAG constraint appropriately so that these special types of cyclic causal effects can also be accommodated. **(d)** This work establishes a new framework that links causal graphical models to multivariate point process data. As far as we know, no existing work has considered the influence of confounders on the point process data. To deal with potential confounders which interact with the point process data, a possible strategy is to include confounder variables in the model, combined with reasonable assumptions of parametric or non-parametric distributions for these variables. These pursuits are left as our future research.

Acknowledgments and Disclosure of Funding

The authors thank the Action Editor and three reviewers for insightful comments that significantly improved the paper's presentation. Zhang was supported by the U.S. National

Science Foundation grants DMS-2013486 and DMS-1712418, and provided by the University of Wisconsin-Madison Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation. Gao's research was partly supported by the China Postdoctoral Science Foundation (grant E2909328). Jia's research was partly supported by the National Philosophy and Social Science Foundation of China (No. 23BTJ046). Conflict of interest/Competing interests: Not applicable.

Appendix A. Supplementary results for numerical experiments

A.1 Algorithms 1 and 2

Algorithm 1 *Flexible Augmented Lagrangian* (**Flex**-AL) *algorithm for solving* (12)

Input: Point process data $\{\boldsymbol{T}_j\}_{j=1,\ldots,d}$, initial guess $\widetilde{\mathbf{W}}^{(0)} = \mathbf{0}_{(d+1)\times d}$, $\alpha^{(0)} = 1$, $\rho^{(0)} = 1$, step sizes $\gamma_\alpha > 1$ and $\gamma_\rho > 1$, tolerance $\epsilon_h > 0$, threshold $\omega > 0$.

1: **for** $k = 0, 1, 2, \ldots$ **do**
2: Solve program (19): $\widetilde{\mathbf{W}}^{(k+1)} \leftarrow \arg\min_{\widetilde{\mathbf{W}} \in \mathbb{R}^{(d+1)\times d}} L(\widetilde{\mathbf{W}}, \alpha^{(k)}; \rho^{(k)})$.
3: Update α and ρ via (23)–(24): $\alpha^{(k+1)} \leftarrow \gamma_\alpha \, \alpha^{(k)}$, $\rho^{(k+1)} \leftarrow \gamma_\rho \, \rho^{(k)}$.
4: Check stopping criteria: if $h(\mathbf{W}^{(k+1)}) < \epsilon_h$, set $\widehat{k} = k + 1$ and terminate.
5: **end for**

Output: Thresholded estimator $\mathbf{W}^{(\widehat{k})} \circ \mathrm{I}(|\mathbf{W}^{(\widehat{k})}| > \omega)$ of the weighted adjacency matrix.

Algorithm 2 *Proximal Quasi-Newton* (PXQN) *algorithm for solving* (25)

Input: Point process data $\{\boldsymbol{T}_j\}_{j=1,\ldots,d}$, initial guess $\widetilde{\boldsymbol{w}}^{(0)} = \mathbf{0}$, tolerance $\epsilon_g > 0$.

1: **for** $k = 0, 1, 2, \ldots$ **do**
2: Compute the current gradient $\nabla f(\widetilde{\boldsymbol{w}}^{(k)})$ via (27).
3: Check stopping criteria: if $\|\nabla f(\widetilde{\boldsymbol{w}}^{(k)})\|_2 < \epsilon_g$, set $\widetilde{\boldsymbol{w}} = \widetilde{\boldsymbol{w}}^{(k)}$ and terminate.
4: Update the L-BFGS approximation $B^{(k)}$ of Hessian matrix $\nabla^2 f(\widetilde{\boldsymbol{w}}^{(k)})$.
5: Compute the descent direction $\widetilde{\boldsymbol{d}}^{(k)}$ by solving (29) via the coordinate descent algorithm.
6: Line search for the step size $\beta^{(k)}$ until the Armijo rule (30) is satisfied.
7: Generate the new iterate $\widetilde{\boldsymbol{w}}^{(k+1)} \leftarrow \widetilde{\boldsymbol{w}}^{(k)} + \beta^{(k)} \, \widetilde{\boldsymbol{d}}^{(k)}$.
8: **end for**

Output: The solution vector $\widetilde{\boldsymbol{w}}$ of (25).

A.2 Figures of simulation studies in Section 6

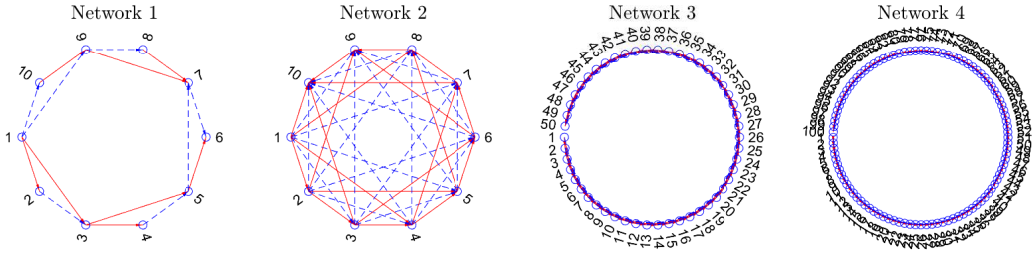


Figure 1: **(Network 1, Network 2, Network 3, and Network 4)** Network 1*: a small sparse* DAG *of* 10 *nodes, with* 8 *excitatory and* 7 *inhibitory effects.* Network 2*: a small dense* DAG *of* 10 *nodes, with* 20 *excitatory and* 20 *inhibitory effects.* Network 3*: a large sparse* DAG *of* 50 *nodes, with* 96 *excitatory and* 94 *inhibitory effects.* Network 4*: an extra-large sparse* DAG *of* 100 *nodes, with* 197 *excitatory and* 195 *inhibitory effects. Red solid lines with arrows represent excitatory effects; blue dashed lines with arrows denote inhibitory effects.*

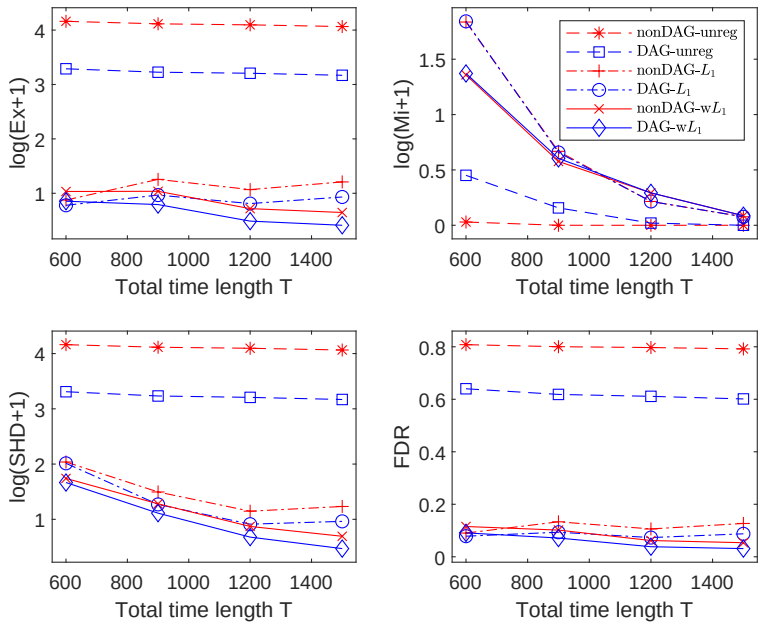


Figure 2: **(Network 1 – small sparse DAG: simulation results)** *Compare methods in Table* 1*. Top left:* 'log(Ex + 1)'*; top right:* 'log(Mi + 1)'*; bottom left:* 'log(SHD + 1)'*; bottom right:* 'FDR'*. Results are averaged over* 100 *replicate samples for each time length* $T \in \{600, 900, 1200, 1500\}$.

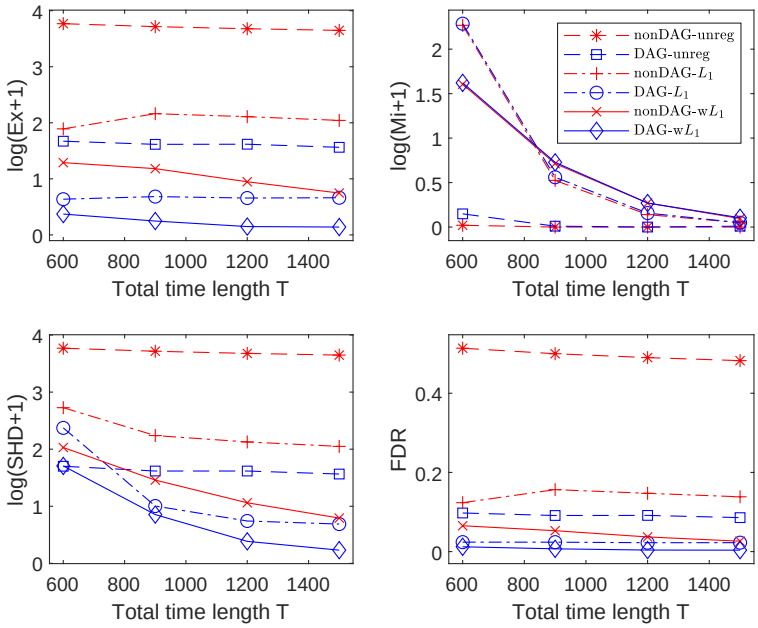


Figure 3: **(Network 2 – small dense DAG: simulation results)** *The caption is similar to that of Figure 2, except for* Network 2.

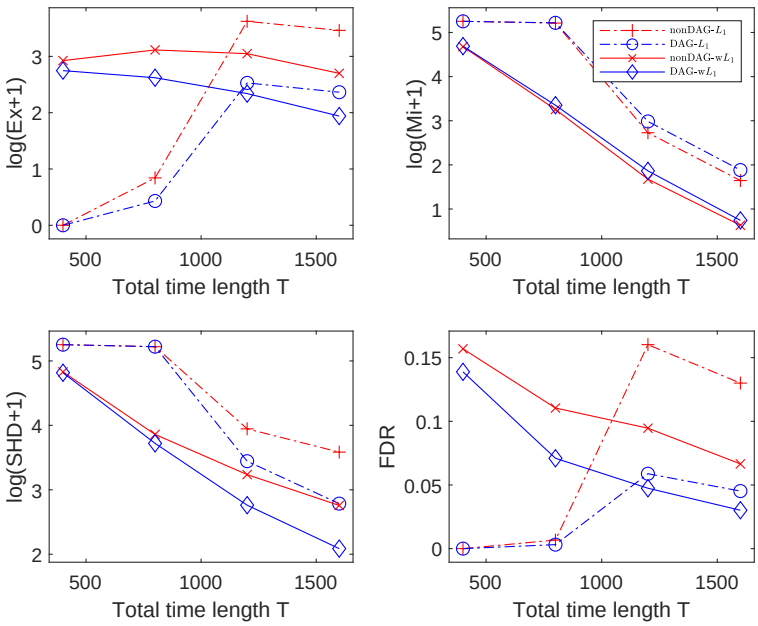


Figure 4: **(Network 3 – large sparse DAG: simulation results)** *The caption is similar to that of Figure 2, except for* Network 3, *with time length* T $\in \{400, 800, 1200, 1600\}$.

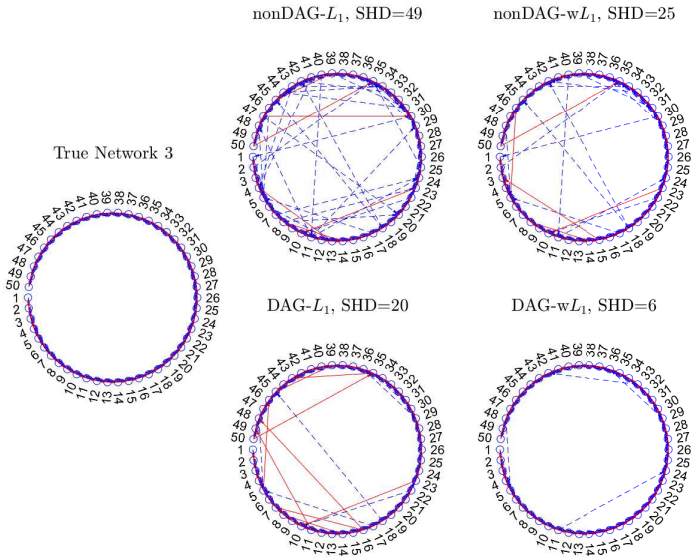


Figure 5: **(Network 3: estimated networks)** *Left column: true* Network 3*; middle and right columns: estimated networks for one simulated dataset using methods* 'nonDAG-L_1', 'DAG-L_1', 'nonDAG-wL_1', *and* 'DAG-wL_1' *from Table* 1*, with* T = 1200.

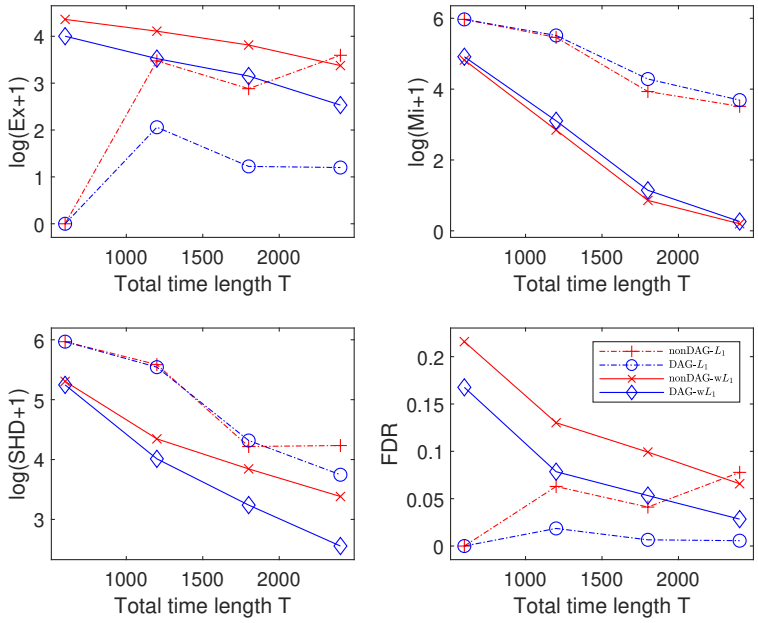


Figure 6: **(Network 4 – extra-large sparse DAG: simulation results)** *The caption is similar to that of Figure* 4*, except for* Network 4*, with time length* T $\in$ {600, 1200, 1800, 2400}.

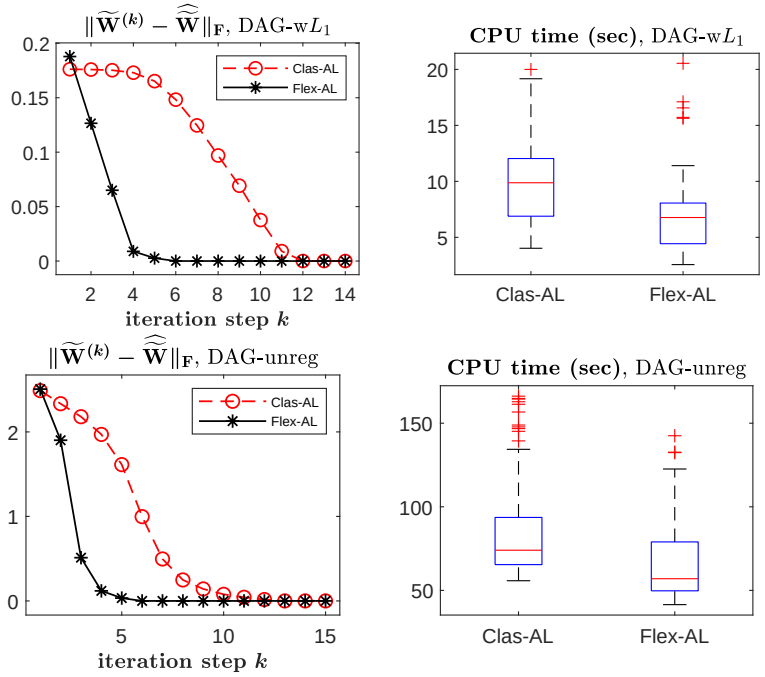


Figure 7: **(Network 3: comparing computational times of algorithms)** *Left panel: average approximation error* $\|\widetilde{\mathbf{W}}^{(k)} - \widehat{\widetilde{\mathbf{W}}}\|_{\mathrm{F}}$ *over* 100 *replicate samples versus the iteration step* k*; right panel: boxplots of runtime for estimators* $\widehat{\widetilde{\mathbf{W}}}$ *across replications. Top row: 'DAG-*wL_1*' method; bottom row: 'DAG-unreg' method.* $\mathrm{T} = 1200$.

A.3 Figures and tables of real data analysis in Section 7

Table 2: **(Real datasets)** *Predictive log-likelihood* (PLL) *for each method.*

Data	nonDAG-wL_1	DAG-wL_1
Neuron spike train data in Section 7.1	188.59	189.02
IPTV viewing record data in Section 7.2	2637.0	2671.2

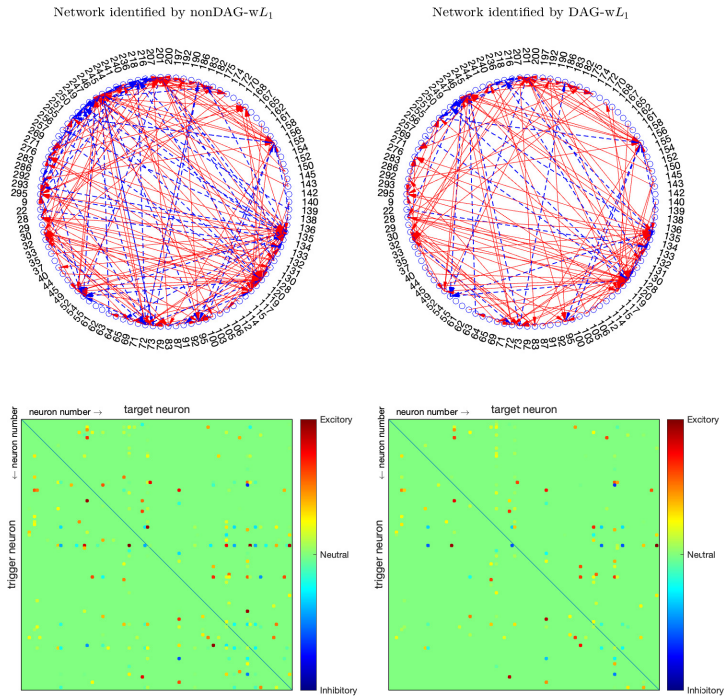


Figure 8: **(Real spike train data: estimated networks)** *Estimated network graphs and heat maps using the* 'nonDAG-wL_1' *method* (*left panel*) *and* 'DAG-wL_1' *method* (*right panel*). *In the network graphs, the numbers represent neuron* ID *numbers. Red solid lines with arrows: excitatory effects; blue dashed lines with arrows: inhibitory effects. In the heat maps, red, blue, and background green colors correspond to excitatory, inhibitory, and no effect on the target neuron from the trigger neuron, respectively. The exact color is determined by the interaction strength.*

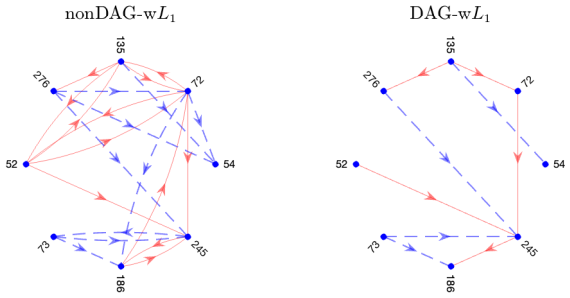


Figure 9: **(Real spike train data: estimated sub-networks)** *Left panel: estimated sub-network graph using the* 'nonDAG-wL_1' *method; right panel: estimated sub-network graph using the* 'DAG-wL_1' *method. The numbers represent neuron* ID *numbers. Red solid lines with arrows: excitatory effects; blue dashed lines with arrows: inhibitory effects.*

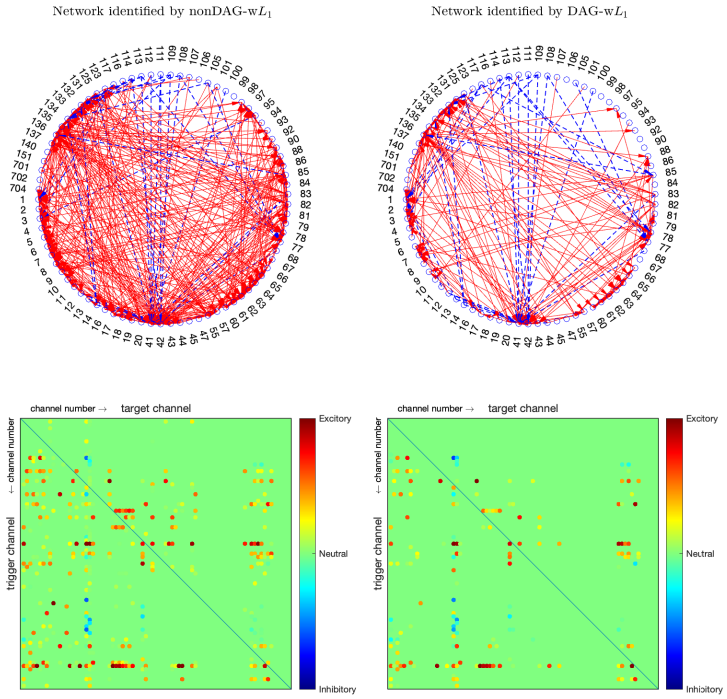


Figure 10: (IPTV **viewing record data: estimated networks**) *The caption is similar to that of Figure* 8*, except for the* IPTV *viewing record data.*

A.4 Additional illustrations for simulation studies in Section 6

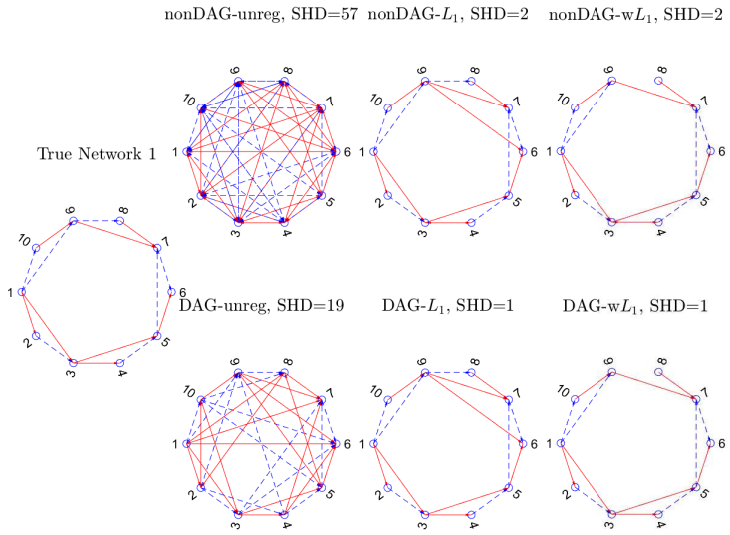


Figure 11: **(Network 1: estimated networks)** *The caption is similar to that of Figure* 5, *except for* Network 1 *with* T = 1200.

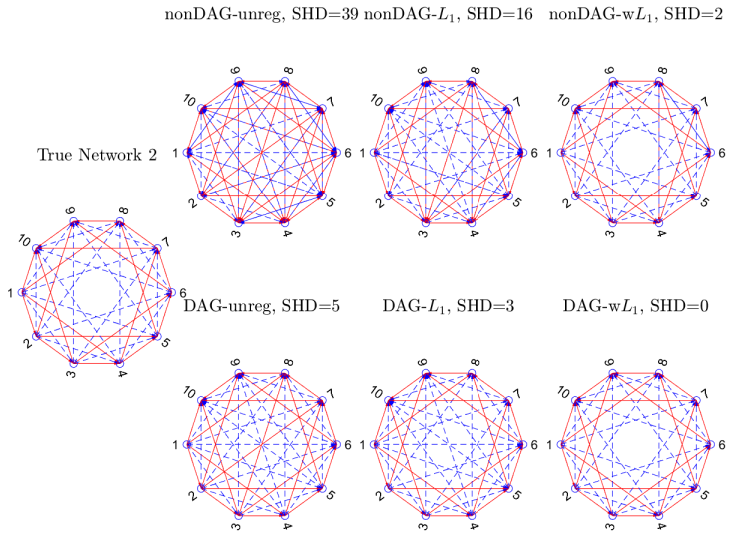


Figure 12: **(Network 2: estimated networks)** *The caption is similar to that of Figure* 5, *except for* Network 2 *with* T = 1200.

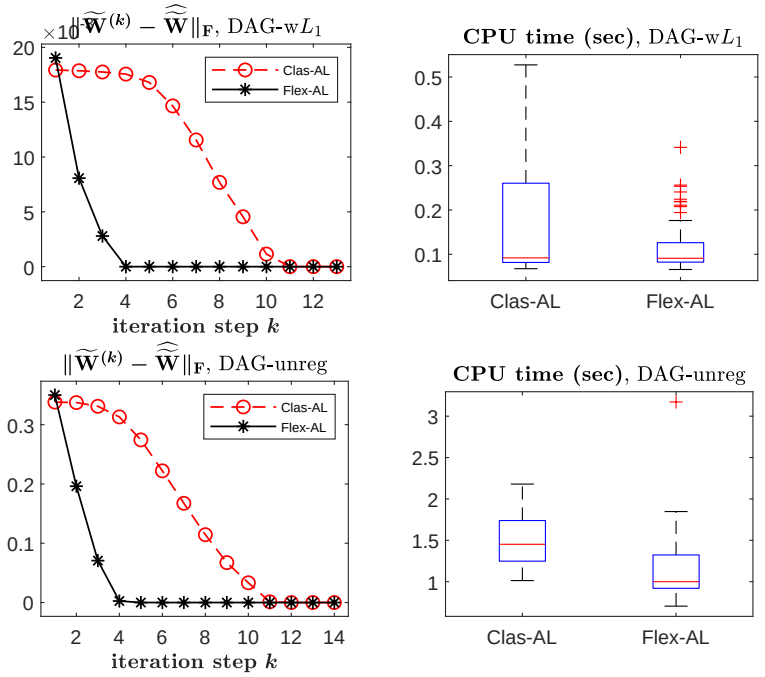


Figure 13: **(Network 1: comparing computational times of algorithms)** *The caption is similar to that of Figure* 7*, except for* Network 1 *with* T = 1200.

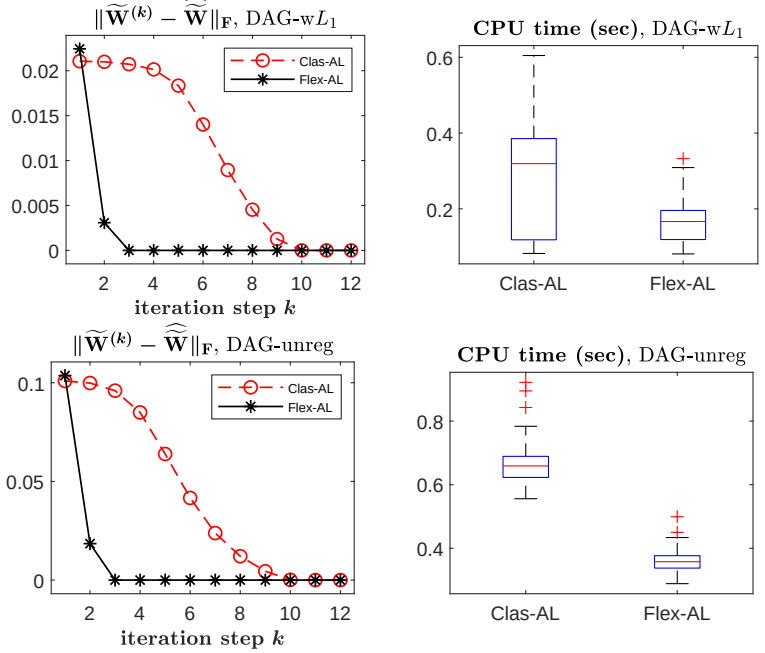


Figure 14: **(Network 2: comparing computational times of algorithms)** *The caption is similar to that of Figure* 7*, except for* Network 2 *with* T = 1200.

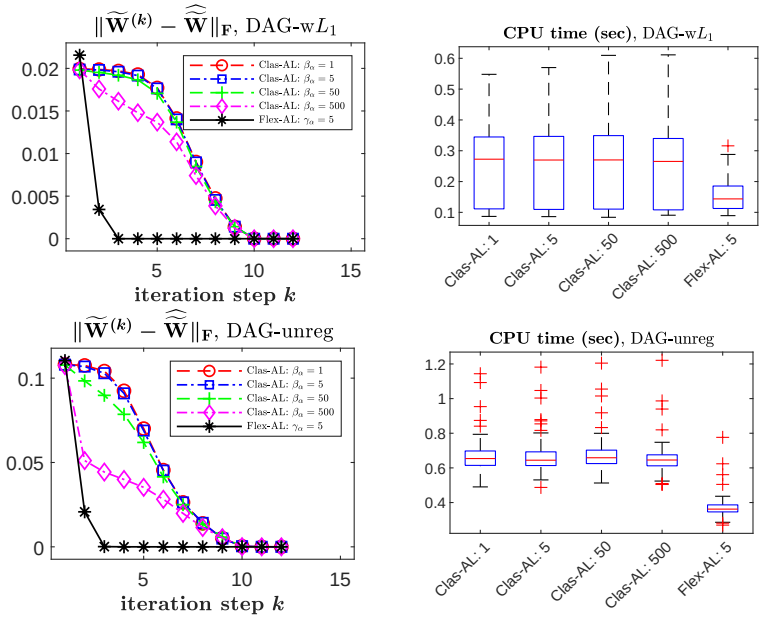


Figure 15: **(Network 2: comparing computational times of the** Clas-AL **algorithm using different step sizes** $\beta_\alpha \in \{1, 5, 50, 500\}$ **with the Flex-**AL **algorithm using step size** $\gamma_\alpha = 5$**)** *Left panel: average approximation error* $\|\widetilde{\mathbf{W}}^{(k)} - \widehat{\widetilde{\mathbf{W}}}\|_{\mathrm{F}}$ *over* 100 *replicate samples versus the iteration step* k*; right panel: boxplots of runtime for estimators* $\widehat{\widetilde{\mathbf{W}}}$ *across replications. Top row: 'DAG-*wL_1*' method; bottom row: 'DAG-unreg' method.* $\mathrm{T} = 1200$.

Appendix B. Notations

Notations in the proof: Let $\mathbf{0}$ represent a vector of zeros, $\mathbf{0}_p$ denote a square p-by-p zero matrix, and $\mathbf{0}_{n\times p}$ denote the n-by-p zero matrix. For a vector $\boldsymbol{u} = (u_1, \ldots, u_m)^\top \in \mathbb{R}^m$, $\|\boldsymbol{u}\|_1 = \sum_{i=1}^m |u_i|$ denotes the L_1 norm of $\boldsymbol{u}$, $\|\boldsymbol{u}\|_2 = \sqrt{\sum_{i=1}^m u_i^2}$ denotes the L_2 norm of $\boldsymbol{u}$, and $\|\boldsymbol{u}\|_\infty = \max_{1\le i\le m} |u_i|$ denotes the L_∞ norm of $\boldsymbol{u}$. For a function $F(\cdot)$ of $\boldsymbol{u}$, $\nabla F(\cdot) = (\partial F(\cdot)/\partial u_1, \ldots, \partial F(\cdot)/\partial u_m)^\top$ denotes the gradient vector, and $\nabla^2 F(\cdot) = (\partial^2 F(\cdot)/(\partial u_i\, \partial u_j)) \in \mathbb{R}^{m\times m}$ represents the Hessian matrix of $F(\cdot)$. For a matrix $A = (a_{i,j}) \in \mathbb{R}^{m\times n}$, define $|A| = (|a_{i,j}|) \in \mathbb{R}^{m\times n}$, and let $\mathrm{vec}(A) = (a_{1,1}, \ldots, a_{m,1}, \ldots, a_{1,n}, \ldots, a_{m,n})^\top \in \mathbb{R}^{m\cdot n}$ denote the vectorization of A. Let $\|A\|_{\mathrm{F}} = \|\mathrm{vec}(A)\|_2 = \sqrt{\sum_{i=1}^m \sum_{j=1}^n a_{i,j}^2}$ denote the Frobenius norm of A, and $\|A\|_1 = \|\mathrm{vec}(A)\|_1 = \sum_{i=1}^m \sum_{j=1}^n |a_{i,j}|$ denote the L_1 norm of $\mathrm{vec}(A)$. For matrices $A = (a_{i,j})$ and $B = (b_{i,j})$ of the same dimensions, $A \circ B = (a_{i,j} b_{i,j})$ is the Hadamard product. The minimum eigenvalue of a matrix A is denoted by $\lambda_{\min}(A)$.

For a weighted adjacency matrix $\mathbf{W} = (\boldsymbol{w}_{\cdot,1}, \ldots, \boldsymbol{w}_{\cdot,d}) \in \mathbb{R}^{d\times d}$ and a parameter matrix $\widetilde{\mathbf{W}} = (\widetilde{\boldsymbol{w}}_{\cdot,1}, \ldots, \widetilde{\boldsymbol{w}}_{\cdot,d}) = (\boldsymbol{w}_{0,\cdot}, \mathbf{W}^\top)^\top \in \mathbb{R}^{(d+1)\times d}$, their vectorized versions are denoted by $\boldsymbol{w} = (\boldsymbol{w}_{\cdot,1}^\top, \ldots, \boldsymbol{w}_{\cdot,d}^\top)^\top = \mathrm{vec}(\mathbf{W}) \in \mathbb{R}^{d^2}$, and $\widetilde{\boldsymbol{w}} = (\widetilde{\boldsymbol{w}}_{\cdot,1}^\top, \ldots, \widetilde{\boldsymbol{w}}_{\cdot,d}^\top)^\top = \mathrm{vec}(\widetilde{\mathbf{W}}) \in \mathbb{R}^{d^2+d}$ respectively. Likewise, functions $\mathcal{P}(\mathbf{W}; \boldsymbol{\eta})$, $\mathcal{L}(\widetilde{\mathbf{W}})$, and $h(\mathbf{W})$ are represented by $\mathcal{P}(\boldsymbol{w}; \boldsymbol{\eta})$, $\mathcal{L}(\widetilde{\boldsymbol{w}})$, and $h(\boldsymbol{w})$, respectively. Notations $\mathbf{W}^*$ and $\widetilde{\mathbf{W}}^*$ are defined in the first paragraph of Section 5.1. For $j = 1, \ldots, d$, and $t \in [0, \mathrm{T}]$, let $\lambda_j^*(t \mid \mathscr{F}_t) = \exp\{w_{0,j}^* + \sum_{i=1}^d w_{i,j}^*\, x_i(t)\}$ denote the true CIF modeled by (4) with the true parameters $\{w_{i,j}^*\}_{i=0,1,\ldots,d}$, to distinct from $\lambda_j(t \mid \mathscr{F}_t)$ in (4) with generic parameters $\{w_{i,j}\}_{i=0,1,\ldots,d}$.

Appendix C. Proofs of main results

The proofs are divided into three parts. Appendix C.1 presents preliminary Lemmas C.1–C.5, which demonstrate the fundamental properties of our MuTPP model (4) as well as the loss function (9). Appendix C.2 presents supporting Lemmas C.6–C.14 related to the statistical and algorithmic properties of our proposed DAG learning method. Appendix C.3 proves lemmas and theorems in the main text.

C.1 Preliminary lemmas on MuTPP model (4)

Lemma C.1 (Stability of MuTPP) *Assume conditions* A1, A3, *and* A4 *in Section* 5. *For each* $j = 1, \ldots, d$, *we have*

$$\mathrm{P}(\lambda_j^*(t \mid \mathscr{F}_t) < \infty) = 1, \text{ and } \mathrm{P}(N_j(t) < \infty) = 1, \text{ for any } t \in [0, \mathrm{T}].$$

Proof Conditions A3 and A4 guarantee that $\sup_{j=1,\ldots,d;\, t\in[0,\mathrm{T}]} \lambda_j^*(t \mid \mathscr{F}_t) \le c$, for some constant $c \in (0, \infty)$. Using this and Gao et al. (2024) (Theorem 2, eq. (40) and eq. (41)), we have $\mathrm{E}\{N_j(t)\} \le c\, t$ and $\mathrm{P}(N_j(t) < \infty) = 1$ for any $t \in [0, \mathrm{T}]$. This completes the proof. ■

Lemma C.2 (Non-stationarity of MuTPP) *Assume conditions* A1, A3, A4, *and* A5 *in Section* 5. *There exists some node* $j_0 \in \{1, \ldots, d\}$ *such that the counting process* $N_{j_0}(t)$ *is non-stationary, i.e., the distribution of* $\lambda_{j_0}^*(t \mid \mathscr{F}_t)$ *varies with respect to* $t \in [0, \mathrm{T}]$.

Proof Condition A5 ensures that the true network $\mathcal{E}(\mathbf{W}^*) \neq \varnothing$. Then Lemma C.2 is proved by applying Gao et al. (2024) (Lemma B.9). ■

Lemma C.3 (Bounded second moment of $\partial\mathcal{L}(\widetilde{\boldsymbol{w}})/\partial w_{i,j}|_{\widetilde{\boldsymbol{w}}=\widetilde{\boldsymbol{w}}^*}$) *Assume conditions* A1–A4 *in Section* 5. *There exists some constant $c \in (0,\infty)$ such that*

$$\mathrm{E}[\{\partial\mathcal{L}(\widetilde{\boldsymbol{w}})/\partial w_{i,j}|_{\widetilde{\boldsymbol{w}}=\widetilde{\boldsymbol{w}}^*}\}^2] \leq c/\mathrm{T}, \tag{C.1}$$

for all $i = 0, 1, \ldots, d$ and $j = 1, \ldots, d$. Moreover, we have

$$\|\nabla\mathcal{L}(\widetilde{\boldsymbol{w}}^*)\|_2 = O_{\mathrm{P}}(\sqrt{d^2/\mathrm{T}}), \quad \textit{as } \mathrm{T} \to \infty. \tag{C.2}$$

Proof According to (7) and (28), for $i = 0, 1, \ldots, d$ and $j = 1, \ldots, d$, we have $\partial\mathcal{L}(\widetilde{\boldsymbol{w}})/\partial w_{i,j} = \mathrm{T}^{-1}\int_0^{\mathrm{T}}\{x_i(t)\lambda_j(t \mid \mathscr{F}_t)\,\mathrm{d}t - x_i(t-)\,\mathrm{d}N_j(t)\}$. Conditions A3 and A4 guarantee two inequalities $\sup_{i=0,1,\ldots,d;\, t\in[0,\mathrm{T}]} x_i(t) \leq c_1$ and $\sup_{j=1,\ldots,d;\, t\in[0,\mathrm{T}]} \lambda_j^*(t \mid \mathscr{F}_t) \leq c_2$ for some constant $c_1, c_2 \in (0,\infty)$, which and Gao et al. (2024) (Theorem 2, eq. (42)) give

$$\begin{aligned} &\mathrm{E}\{\partial\mathcal{L}(\widetilde{\boldsymbol{w}})/\partial w_{i,j}|_{\widetilde{\boldsymbol{w}}=\widetilde{\boldsymbol{w}}^*}\} = 0, \\ &\mathrm{var}\{\partial\mathcal{L}(\widetilde{\boldsymbol{w}})/\partial w_{i,j}|_{\widetilde{\boldsymbol{w}}=\widetilde{\boldsymbol{w}}^*}\} = \frac{1}{\mathrm{T}^2}\cdot\mathrm{E}\Big\{\int_0^{\mathrm{T}} x_i^2(t)\lambda_j^*(t \mid \mathscr{F}_t)\,\mathrm{d}t\Big\} \leq c_1^2\, c_2/\mathrm{T}. \end{aligned} \tag{C.3}$$

This proves (C.1). It follows that

$$\mathrm{E}\big\{\|\nabla\mathcal{L}(\widetilde{\boldsymbol{w}}^*)\|_2^2\big\} = \textstyle\sum_{i=0}^{d}\sum_{j=1}^{d}\mathrm{E}\Big[\Big\{\dfrac{\partial\mathcal{L}(\widetilde{\boldsymbol{w}})}{\partial w_{i,j}}\Big|_{\widetilde{\boldsymbol{w}}=\widetilde{\boldsymbol{w}}^*}\Big\}^2\Big] \leq (d+1)\,d\cdot c_1^2 c_2/\mathrm{T} = O(d^2/\mathrm{T}).$$

This, together with Markov's inequality, proves (C.2). ■

Lemma C.4 (Strict convexity of $\mathrm{E}\{\mathcal{L}(\widetilde{\boldsymbol{w}})\}$) *Assume conditions* A1, A3, *and* A4 *in Section* 5. *Then $\mathrm{E}\{\mathcal{L}(\widetilde{\boldsymbol{w}})\}$ is strictly convex in $\widetilde{\boldsymbol{w}} \in \mathbb{R}^{d^2+d}$. Moreover, it has a unique global minimizer $\widetilde{\boldsymbol{w}} = \widetilde{\boldsymbol{w}}^*$.*

Proof Note that $\mathrm{E}\{\mathcal{L}(\widetilde{\boldsymbol{w}})\} = \sum_{j=1}^{d}\mathrm{E}\{\mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\boldsymbol{\cdot},j})\}$. By (28) and (C.3), for each $j = 1, \ldots, d$, we have $\nabla\mathrm{E}\{\mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}^*_{\boldsymbol{\cdot},j})\} = \mathbf{0}$, and $\nabla^2\mathrm{E}\{\mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\boldsymbol{\cdot},j})\} = \mathrm{T}^{-1}\cdot\mathrm{E}[\int_0^{\mathrm{T}}\widetilde{\boldsymbol{x}}(t)\,\widetilde{\boldsymbol{x}}(t)^\top\cdot\exp\{\widetilde{\boldsymbol{w}}_{\boldsymbol{\cdot},j}^\top\widetilde{\boldsymbol{x}}(t)\}\,\mathrm{d}t]$ for $\widetilde{\boldsymbol{w}}_{\boldsymbol{\cdot},j} \in \mathbb{R}^{d+1}$. It suffices to show that the matrix $f(\widetilde{\boldsymbol{v}}) := \mathrm{E}[\int_0^{\mathrm{T}}\widetilde{\boldsymbol{x}}(t)\,\widetilde{\boldsymbol{x}}(t)^\top\cdot\exp\{\widetilde{\boldsymbol{v}}^\top\widetilde{\boldsymbol{x}}(t)\}\,\mathrm{d}t]$ is strictly positive-definite for any $\widetilde{\boldsymbol{v}} \in \mathbb{R}^{d+1}$.

Using Gao et al. (2024) (Lemma B.10, eq. (116); and Lemma B.17, eq. (149)), for each $j = 1, \ldots, d$, there exist some time points $0 < t_1 < t_0 < \mathrm{T}$, such that

$$\mathrm{P}(N_j(t_0) = 0, \text{ for all } j = 1, \ldots, d) > 0, \tag{C.4}$$

$$\mathrm{P}(N_j(t_1) = N_j(t_0) = 1, \text{ and } N_k(t_0) = 0 \text{ for all } k \in \{1, \ldots, d\} \setminus \{i\}) > 0. \tag{C.5}$$

For $j = 1, \ldots, d$, denote by A_j the event in the probability in (C.5). When A_j occurs, for any $t \in [t_1, t_0]$, (5) implies that $x_j(t) = g(1/\phi)$ and $x_k(t) = 0$ for all $k \in \{1, \ldots, d\} \setminus \{i\}$. Also,

denote by A_0 the event in (C.4). When A_0 occurs, we have $x_j(t) = 0$ for all $j = 1, \ldots, d$, and $t \in [t_1, t_0]$. Thus, for any $\widetilde{\boldsymbol{u}} \in \mathbb{R}^{d+1}$ with $\|\widetilde{\boldsymbol{u}}\|_2 > 0$, we have

$$\begin{aligned}
&\widetilde{\boldsymbol{u}}^\top f(\widetilde{\boldsymbol{v}})\widetilde{\boldsymbol{u}} \\
\geq &\mathrm{E}\Big[\int_{t_1}^{t_0} (\widetilde{\boldsymbol{u}}^\top \widetilde{\boldsymbol{x}}(t))^2 \exp\{\widetilde{\boldsymbol{v}}^\top \widetilde{\boldsymbol{x}}(t)\}\,\mathrm{d}t\Big] \\
\geq &\mathrm{P}(A_0) \cdot u_0^2 \cdot \exp(v_0) + \textstyle\sum_{j=1}^d \mathrm{P}(A_j) \cdot \{u_0 + u_j \cdot g(1/\phi)\}^2 \cdot \exp\{v_0 + v_j \cdot g(1/\phi)\} > 0.
\end{aligned}$$

This proves that $f(\widetilde{\boldsymbol{v}})$ is strictly positive-definite. ■

Lemma C.5 (Identifiability of model (4)) *Assume conditions* A1, A3, *and* A4 *in Section* 5. *Then model* (4) *is identifiable with respect to the parameters* $\{w_{i,j} \in \mathbb{R} : i = 0, 1, \ldots, d; j = 1, \ldots, d\}$.

Proof For a parameter matrix $\widetilde{\mathbf{W}} \in \mathbb{R}^{(d+1)\times d}$ which collects all parameters $\{w_{i,j} \in \mathbb{R} : i = 0, 1, \ldots, d; j = 1, \ldots, d\}$, denote by $\mathscr{P}(\widetilde{\mathbf{W}})$ the probability distribution of the MuTPP specified in model (4). To prove the identifiability of model (4), it suffices to show that the mapping from $\widetilde{\mathbf{W}}$ to $\mathscr{P}(\widetilde{\mathbf{W}})$ is one-to-one.

Suppose that there exist distinct $\widetilde{\mathbf{W}}_1, \widetilde{\mathbf{W}}_2 \in \mathbb{R}^{(d+1)\times d}$, such that $\mathscr{P}(\widetilde{\mathbf{W}}_1) = \mathscr{P}(\widetilde{\mathbf{W}}_2)$. Then the negative log-likelihood function $\mathcal{L}(\cdot)$ in (9) satisfies $\mathcal{L}(\widetilde{\mathbf{W}}_1) = \mathcal{L}(\widetilde{\mathbf{W}}_2)$, and hence the equality $\mathrm{E}_{\widetilde{\mathbf{W}}_1}\{\mathcal{L}(\widetilde{\mathbf{W}}_1)\} = \mathrm{E}_{\widetilde{\mathbf{W}}_1}\{\mathcal{L}(\widetilde{\mathbf{W}}_2)\}$, where $\mathrm{E}_{\widetilde{\mathbf{W}}_1}(\cdot)$ denotes the expectation operator with respect to the probability distribution $\mathscr{P}(\widetilde{\mathbf{W}}_1)$. By Lemma C.4, $\widetilde{\mathbf{W}}_1$ is the unique global minimizer of function $\mathrm{E}_{\widetilde{\mathbf{W}}_1}\{\mathcal{L}(\cdot)\}$. This, combined with the above equality, implies that $\widetilde{\mathbf{W}}_2$ is another distinct global minimizer of $\mathrm{E}_{\widetilde{\mathbf{W}}_1}\{\mathcal{L}(\cdot)\}$, which obviously contradicts. The proof is completed. ■

C.2 Supporting lemmas on estimation and algorithms

Lemmas C.6–C.12 aid in proving lemmas and theorems (consistency) in Section 5, where the asymptotic setting is established by letting the time length $\mathrm{T} \to \infty$. Lemmas C.13–C.14 assist in proving Theorem 1 (algorithmic convergence) in Section 4, where the convergence is based on letting the number of iteration steps $k \to \infty$.

Lemma C.6 *Assume conditions* A1–A6 *in Section* 5. *Assume that* $\boldsymbol{\eta}$ *in* (34) *satisfies condition* (37). *Assume* $\kappa = \infty$ *in* (34). *If* $d^4/\mathrm{T} \to 0$ *as* $\mathrm{T} \to \infty$, *then there exists a local minimizer* $\widehat{\widetilde{\mathbf{W}}}_{\infty,\boldsymbol{\eta}}$ *of the* nonDAG *optimization problem* (34), *such that* $\|\widehat{\widetilde{\mathbf{W}}}_{\infty,\boldsymbol{\eta}} - \widetilde{\mathbf{W}}^*\|_{\mathrm{F}} = O_{\mathrm{P}}(\sqrt{d^2/\mathrm{T}})$.

Proof Denote $Q(\widetilde{\boldsymbol{w}}) = \mathcal{L}(\widetilde{\boldsymbol{w}}) + \mathcal{P}(\boldsymbol{w}; \boldsymbol{\eta})$ as the objective function in (34). Let $r_{\mathrm{T}} = \sqrt{d^2/\mathrm{T}}$. Following the proofs in Fan and Peng (2004); Zhang et al. (2010), it suffices to show that for any given $\epsilon > 0$, there exists a large constant C_ϵ, such that

$$\mathrm{P}\Big(\inf_{\widetilde{\boldsymbol{u}} \in \mathbb{R}^{d^2+d}: \|\widetilde{\boldsymbol{u}}\|_2 = C_\epsilon} Q(\widetilde{\boldsymbol{w}}^* + r_{\mathrm{T}}\,\widetilde{\boldsymbol{u}}) > Q(\widetilde{\boldsymbol{w}}^*)\Big) \geq 1 - \epsilon \tag{C.6}$$

for all sufficiently large T.

In accordance with the notation $\widetilde{\boldsymbol{w}}$, we write the vector $\widetilde{\boldsymbol{u}} = (\widetilde{\boldsymbol{u}}_{\bullet,1}^\top, \ldots, \widetilde{\boldsymbol{u}}_{\bullet,d}^\top)^\top$ with $\widetilde{\boldsymbol{u}}_{\bullet,j} = (u_{0,j}, u_{1,j}, \ldots, u_{d,j}) \in \mathbb{R}^{d+1}$, $j = 1, \ldots, d$. As $w_{i,j}^* = 0$ for $(i,j) \in \mathcal{E}^c(\mathbf{W}^*)$, we obtain:

$$\begin{aligned}
&Q(\widetilde{\boldsymbol{w}}^* + r_{\mathrm{T}}\,\widetilde{\boldsymbol{u}}) - Q(\widetilde{\boldsymbol{w}}^*) \\
=&\mathcal{L}(\widetilde{\boldsymbol{w}}^* + r_{\mathrm{T}}\widetilde{\boldsymbol{u}}) - \mathcal{L}(\widetilde{\boldsymbol{w}}^*) + \sum_{1\le i\ne j\le d}\sum \eta_{i,j}\big\{|w_{i,j}^* + r_{\mathrm{T}}u_{i,j}| - |w_{i,j}^*|\big\} \\
\ge&\mathcal{L}(\widetilde{\boldsymbol{w}}^* + r_{\mathrm{T}}\widetilde{\boldsymbol{u}}) - \mathcal{L}(\widetilde{\boldsymbol{w}}^*) + \sum_{(i,j)\in\mathcal{E}(\mathbf{W}^*)}\sum \eta_{i,j}\big\{|w_{i,j}^* + r_{\mathrm{T}}u_{i,j}| - |w_{i,j}^*|\big\} \\
\equiv&Q_{\mathrm{I}} + Q_{\mathrm{II}}.
\end{aligned}$$

We first focus on the term Q_{I}. Observe that:

$$Q_{\mathrm{I}} = \mathcal{L}(\widetilde{\boldsymbol{w}}^* + r_{\mathrm{T}}\widetilde{\boldsymbol{u}}) - \mathcal{L}(\widetilde{\boldsymbol{w}}^*) = \sum_{j=1}^{d}\Big\{\mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\bullet,j}^* + r_{\mathrm{T}}\widetilde{\boldsymbol{u}}_{\bullet,j}) - \mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\bullet,j}^*)\Big\}. \tag{C.7}$$

For each $j = 1, \ldots, d$, using Taylor's expansion of $\mathcal{L}_{j,\mathrm{T}}(\cdot)$ in (7) around $\widetilde{\boldsymbol{w}}_{\bullet,j}^*$, we get:

$$\mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\bullet,j}^* + r_{\mathrm{T}}\widetilde{\boldsymbol{u}}_{\bullet,j}) - \mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\bullet,j}^*) \equiv I_{j,1} + I_{j,2} + I_{j,3},$$

with

$$\begin{aligned}
I_{j,1} &= \nabla\mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\bullet,j}^*)^\top r_{\mathrm{T}}\,\widetilde{\boldsymbol{u}}_{\bullet,j} \\
&= \frac{1}{\mathrm{T}}\int_0^{\mathrm{T}}\Big[\widetilde{\boldsymbol{x}}(t)^\top r_{\mathrm{T}}\,\widetilde{\boldsymbol{u}}_{\bullet,j}\exp\big\{\widetilde{\boldsymbol{w}}_{\bullet,j}^{*\top}\widetilde{\boldsymbol{x}}(t)\big\}\,\mathrm{d}t - \widetilde{\boldsymbol{x}}(t-)^\top r_{\mathrm{T}}\,\widetilde{\boldsymbol{u}}_{\bullet,j}\,\mathrm{d}N_j(t)\Big], \\
I_{j,2} &= \frac{r_{\mathrm{T}}^2}{2}\widetilde{\boldsymbol{u}}_{\bullet,j}^\top\nabla^2\mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\bullet,j}^*)\widetilde{\boldsymbol{u}}_{\bullet,j} \\
&= \frac{1}{2\mathrm{T}}\int_0^{\mathrm{T}}\{\widetilde{\boldsymbol{x}}(t)^\top r_{\mathrm{T}}\,\widetilde{\boldsymbol{u}}_{\bullet,j}\}^2\exp\big\{\widetilde{\boldsymbol{w}}_{\bullet,j}^{*\top}\widetilde{\boldsymbol{x}}(t)\big\}\,\mathrm{d}t, \\
I_{j,3} &= \frac{1}{6\mathrm{T}}\int_0^{\mathrm{T}}\{\widetilde{\boldsymbol{x}}(t)^\top r_{\mathrm{T}}\,\widetilde{\boldsymbol{u}}_{\bullet,j}\}^3\exp\big\{\widetilde{\boldsymbol{w}}_{\bullet,j}^{\star\star\top}\widetilde{\boldsymbol{x}}(t)\big\}\,\mathrm{d}t,
\end{aligned} \tag{C.8}$$

where $\widetilde{\boldsymbol{w}}_{\bullet,j}^{\star\star}$ lies between $\widetilde{\boldsymbol{w}}_{\bullet,j}^*$ and $\widetilde{\boldsymbol{w}}_{\bullet,j}$. Using (C.7)–(C.8), we obtain:

$$Q_{\mathrm{I}} = \sum_{j=1}^{d}\big\{\mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\bullet,j}) - \mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\bullet,j}^*)\big\} \equiv I_1 + I_2 + I_3,$$

where $I_1 = \sum_{j=1}^d I_{j,1}$, $I_2 = \sum_{j=1}^d I_{j,2}$, and $I_3 = \sum_{j=1}^d I_{j,3}$.

For the term I_1, an application of Lemma C.3 gives that:

$$\begin{aligned}
|I_1| &= \Big|\sum_{j=1}^{d} I_{j,1}\Big| = \Big|\sum_{j=1}^{d}\nabla\mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\bullet,j}^*)^\top r_{\mathrm{T}}\,\widetilde{\boldsymbol{u}}_{\bullet,j}\Big| = |\nabla\mathcal{L}(\widetilde{\boldsymbol{w}}^*)^\top r_{\mathrm{T}}\,\widetilde{\boldsymbol{u}}| \\
&\le r_{\mathrm{T}}\|\nabla\mathcal{L}(\widetilde{\boldsymbol{w}}^*)\|_2\|\widetilde{\boldsymbol{u}}\|_2 = r_{\mathrm{T}}\,O_{\mathrm{P}}\big(\sqrt{d^2/\mathrm{T}}\big)\,\|\widetilde{\boldsymbol{u}}\|_2 = r_{\mathrm{T}}^2\,O_{\mathrm{P}}(1)\,\|\widetilde{\boldsymbol{u}}\|_2.
\end{aligned} \tag{C.9}$$

For the term I_2, by condition A6, we have:

$$I_2 = \sum_{j=1}^{d} I_{j,2} = \frac{r_{\mathrm{T}}^2}{2} \sum_{j=1}^{d} \widetilde{\boldsymbol{u}}_{\boldsymbol{\cdot},j}^{\top} \nabla^2 \mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\boldsymbol{\cdot},j}^{*}) \widetilde{\boldsymbol{u}}_{\boldsymbol{\cdot},j}$$
$$\geq (C/2) r_{\mathrm{T}}^2 \sum_{j=1}^{d} \|\widetilde{\boldsymbol{u}}_{\boldsymbol{\cdot},j}\|_2^2 \cdot (1 + o_{\mathrm{P}}(1)) = (C/2) r_{\mathrm{T}}^2 \|\widetilde{\boldsymbol{u}}\|_2^2 \cdot (1 + o_{\mathrm{P}}(1)),$$

for some constant $C \in (0, \infty)$. For the term I_3, condition A3 implies that each component of $\widetilde{\boldsymbol{x}}(t)$ is bounded from above. This, together with condition A4, yields that the term $\exp\{\widetilde{\boldsymbol{w}}_{\boldsymbol{\cdot},j}^{\star\star\top} \widetilde{\boldsymbol{x}}(t)\}$ in (C.8) is also bounded from above. Hence, there exists a constant $c \in (0, \infty)$, such that: $|I_{j,3}| \leq c\, r_{\mathrm{T}}^3 \sum_{k=0}^{d} |u_{k,j}|^3$ for all $j = 1, \ldots, d$. It follows that:

$$|I_3| = \Big| \sum_{j=1}^{d} I_{j,3} \Big| \;\leq\; c\, r_{\mathrm{T}}^3 \sum_{j=1}^{d} \sum_{k=0}^{d} |u_{k,j}|^3 \;\leq\; c\, r_{\mathrm{T}}^3 \|\widetilde{\boldsymbol{u}}\|_2^3.$$

We next consider the term Q_{II}. Using the triangle inequality and condition (37), we have:

$$\begin{aligned}
|Q_{\mathrm{II}}| &= \Big| \sum\sum_{(i,j) \in \mathcal{E}(\mathbf{W}^*)} \eta_{i,j} \left(|w_{i,j}^* + r_{\mathrm{T}} u_{i,j}| - |w_{i,j}^*| \right) \Big| \\
&\leq \sum\sum_{(i,j) \in \mathcal{E}(\mathbf{W}^*)} \eta_{i,j}\, r_{\mathrm{T}} \, |u_{i,j}| \\
&\leq r_{\mathrm{T}} \max_{(i,j) \in \mathcal{E}(\mathbf{W}^*)} \eta_{i,j} \sum\sum_{(i,j) \in \mathcal{E}(\mathbf{W}^*)} |u_{i,j}| \\
&\leq r_{\mathrm{T}} \max_{(i,j) \in \mathcal{E}(\mathbf{W}^*)} \eta_{i,j}\, \sqrt{s}\, \|\widetilde{\boldsymbol{u}}\|_2 \\
&= r_{\mathrm{T}}\, O_{\mathrm{P}}(\sqrt{d^2/(s\mathrm{T})})\, \sqrt{s}\, \|\widetilde{\boldsymbol{u}}\|_2 \\
&= r_{\mathrm{T}}^2\, O_{\mathrm{P}}(1)\, \|\widetilde{\boldsymbol{u}}\|_2. \qquad \text{(C.10)}
\end{aligned}$$

Note that the condition $d^4/\mathrm{T} \to 0$ implies that $r_{\mathrm{T}} \to 0$ as $\mathrm{T} \to \infty$. Using this and (C.9)–(C.10), the terms I_1, I_3, and Q_{II} are dominated by the positive term I_2 for sufficiently large $\|\widetilde{\boldsymbol{u}}\|_2$. This proves (C.6). ■

Lemma C.7 *Assume conditions* A1–A6 *in Section* 5. *Assume that* $\boldsymbol{\eta}$ *in* (34) *satisfies condition* (38). *Assume* $\kappa = \infty$ *in* (34). *If* $d^4/\mathrm{T} \to 0$ *as* $\mathrm{T} \to \infty$, *then for any local minimizer* $\widehat{\widetilde{\mathbf{W}}}_{\infty,\boldsymbol{\eta}}$ *of* (34) *satisfying* $\|\widehat{\widetilde{\mathbf{W}}}_{\infty,\boldsymbol{\eta}} - \widetilde{\mathbf{W}}^*\|_{\mathrm{F}} = O_{\mathrm{P}}(\sqrt{d^2/\mathrm{T}})$, *we have*

$$\mathrm{P}\big(\mathrm{Ex}(\widehat{\mathbf{W}}_{\infty,\boldsymbol{\eta}}, \mathbf{W}^*) = 0\big) \to 1 \quad \textit{as } \mathrm{T} \to \infty.$$

Proof Letting $\widehat{\widetilde{\boldsymbol{w}}} = \mathrm{vec}(\widehat{\widetilde{\mathbf{W}}}_{\infty,\boldsymbol{\eta}}) \in \mathbb{R}^{d^2+d}$ denote the vectorization of the matrix $\widehat{\widetilde{\mathbf{W}}}_{\infty,\boldsymbol{\eta}}$ and letting $r_{\mathrm{T}} = \sqrt{d^2/\mathrm{T}}$, the condition $\|\widehat{\widetilde{\mathbf{W}}}_{\infty,\boldsymbol{\eta}} - \widetilde{\mathbf{W}}^*\|_{\mathrm{F}} = O_{\mathrm{P}}(\sqrt{d^2/\mathrm{T}})$ could be rewritten as $\|\widehat{\widetilde{\boldsymbol{w}}} - \widetilde{\boldsymbol{w}}^*\|_2 = O_{\mathrm{P}}(r_{\mathrm{T}})$. This implies that for any $\epsilon > 0$, there exists a constant C_ϵ, such that

$$\mathrm{P}\big(\|\widehat{\widetilde{\boldsymbol{w}}} - \widetilde{\boldsymbol{w}}^*\|_2 \leq r_{\mathrm{T}} C_\epsilon\big) > 1 - \epsilon \qquad \text{(C.11)}$$

for all sufficiently large T. Note that for any $i, k = 0, 1, \ldots, d$, and $j = 1, \ldots, d$,

$$\frac{\partial^2 \mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\bullet,j})}{\partial w_{i,j} \partial w_{k,j}} = \frac{1}{\mathrm{T}} \int_0^{\mathrm{T}} x_i(t) x_k(t) \exp\{\widetilde{\boldsymbol{w}}_{\bullet,j}^\top \widetilde{\boldsymbol{x}}(t)\} \, \mathrm{d}t.$$

By conditions A3 and A4, we observe that both $x_i(t)$ and $\sup_{\widetilde{\boldsymbol{w}}: \|\widetilde{\boldsymbol{w}} - \widetilde{\boldsymbol{w}}^*\|_2 \leq r_{\mathrm{T}} C_\epsilon} \exp\{\widetilde{\boldsymbol{w}}_{\bullet,j}^\top \widetilde{\boldsymbol{x}}(t)\}$ are bounded from above. Therefore, there exists a constant $c \in (0, \infty)$, such that

$$\sup_{\widetilde{\boldsymbol{w}}: \|\widetilde{\boldsymbol{w}} - \widetilde{\boldsymbol{w}}^*\|_2 \leq r_{\mathrm{T}} C_\epsilon} \Big| \frac{\partial^2 \mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\bullet,j})}{\partial w_{i,j} \partial w_{k,j}} \Big| \leq c \tag{C.12}$$

for any $i, k = 0, 1, \ldots, d$ and any $j = 1, \ldots, d$. By Taylor's expansion, (C.12), and Lemma C.3, we obtain

$$\begin{aligned}
&\sup_{i,j=1,\ldots,d} \Big\{ \sup_{\widetilde{\boldsymbol{w}}: \|\widetilde{\boldsymbol{w}} - \widetilde{\boldsymbol{w}}^*\|_2 \leq r_{\mathrm{T}} C_\epsilon} \Big| \frac{\partial \mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\bullet,j})}{\partial w_{i,j}} \Big| \Big\} \\
\leq &\sup_{i,j=1,\ldots,d} \Big\{ \Big| \frac{\partial \mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\bullet,j})}{\partial w_{i,j}} \Big|_{\widetilde{\boldsymbol{w}}_{\bullet,j} = \widetilde{\boldsymbol{w}}^*_{\bullet,j}} \Big| + \sup_{\widetilde{\boldsymbol{w}}: \|\widetilde{\boldsymbol{w}} - \widetilde{\boldsymbol{w}}^*\|_2 \leq r_{\mathrm{T}} C_\epsilon} \sum_{k=0}^{d} \Big| \frac{\partial^2 \mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\bullet,j})}{\partial w_{i,j} \partial w_{k,j}} \Big| \, |w_{k,j} - w^*_{k,j}| \Big\} \\
\leq &\sup_{i,j=1,\ldots,d} \Big\{ \Big| \frac{\partial \mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\bullet,j})}{\partial w_{i,j}} \Big|_{\widetilde{\boldsymbol{w}}_{\bullet,j} = \widetilde{\boldsymbol{w}}^*_{\bullet,j}} \Big| + \sup_{\widetilde{\boldsymbol{w}}: \|\widetilde{\boldsymbol{w}} - \widetilde{\boldsymbol{w}}^*\|_2 \leq r_{\mathrm{T}} C_\epsilon} c \sqrt{d+1} \, \|\widetilde{\boldsymbol{w}}_{\bullet,j} - \widetilde{\boldsymbol{w}}^*_{\bullet,j}\|_2 \Big\} \\
\leq &\|\nabla \mathcal{L}(\widetilde{\boldsymbol{w}}^*)\|_2 + c \sqrt{d+1} \, r_{\mathrm{T}} C_\epsilon \\
= &O_{\mathrm{P}}(\sqrt{d^2/\mathrm{T}}) + O(\sqrt{d^3/\mathrm{T}}) \\
= &O_{\mathrm{P}}(\sqrt{d^3/\mathrm{T}}).
\end{aligned}$$

Using this and condition (38), the following inequalities hold with probability tending to 1 as $\mathrm{T} \to \infty$:

$$\begin{aligned}
&\inf_{\widetilde{\boldsymbol{w}}: \|\widetilde{\boldsymbol{w}} - \widetilde{\boldsymbol{w}}^*\|_2 \leq r_{\mathrm{T}} C_\epsilon} \Big\{ \inf_{w_{i,j} > 0 : (i,j) \in \mathcal{E}^c(\mathbf{W}^*)} \frac{\partial Q(\widetilde{\boldsymbol{w}})}{\partial w_{i,j}} \Big\} \\
= &\inf_{\widetilde{\boldsymbol{w}}: \|\widetilde{\boldsymbol{w}} - \widetilde{\boldsymbol{w}}^*\|_2 \leq r_{\mathrm{T}} C_\epsilon} \Big(\inf_{w_{i,j} > 0 : (i,j) \in \mathcal{E}^c(\mathbf{W}^*)} \Big\{ \frac{\partial \mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\bullet,j})}{\partial w_{i,j}} + \eta_{i,j} \operatorname{sign}(w_{i,j}) \Big\} \Big) \\
\geq &- \sup_{i,j=1,\ldots,d} \Big\{ \sup_{\widetilde{\boldsymbol{w}}: \|\widetilde{\boldsymbol{w}} - \widetilde{\boldsymbol{w}}^*\|_2 \leq r_{\mathrm{T}} C_\epsilon} \Big| \frac{\partial \mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\bullet,j})}{\partial w_{i,j}} \Big| \Big\} + \min_{(i,j) \in \mathcal{E}^c(\mathbf{W}^*)} \eta_{i,j} \\
> &0,
\end{aligned} \tag{C.13}$$

and

$$\begin{aligned}
&\sup_{\widetilde{\boldsymbol{w}}: \|\widetilde{\boldsymbol{w}} - \widetilde{\boldsymbol{w}}^*\|_2 \leq r_{\mathrm{T}} C_\epsilon} \Big\{ \sup_{w_{i,j} < 0 : (i,j) \in \mathcal{E}^c(\mathbf{W}^*)} \frac{\partial Q(\widetilde{\boldsymbol{w}})}{\partial w_{i,j}} \Big\} \\
= &\sup_{\widetilde{\boldsymbol{w}}: \|\widetilde{\boldsymbol{w}} - \widetilde{\boldsymbol{w}}^*\|_2 \leq r_{\mathrm{T}} C_\epsilon} \Big(\sup_{w_{i,j} < 0 : (i,j) \in \mathcal{E}^c(\mathbf{W}^*)} \Big\{ \frac{\partial \mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\bullet,j})}{\partial w_{i,j}} + \eta_{i,j} \operatorname{sign}(w_{i,j}) \Big\} \Big) \\
\leq &\sup_{i,j=1,\ldots,d} \Big\{ \sup_{\widetilde{\boldsymbol{w}}: \|\widetilde{\boldsymbol{w}} - \widetilde{\boldsymbol{w}}^*\|_2 \leq r_{\mathrm{T}} C_\epsilon} \Big| \frac{\partial \mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\bullet,j})}{\partial w_{i,j}} \Big| \Big\} - \min_{(i,j) \in \mathcal{E}^c(\mathbf{W}^*)} \eta_{i,j} \\
< &0,
\end{aligned} \tag{C.14}$$

where $Q(\widetilde{\boldsymbol{w}}) = \mathcal{L}(\widetilde{\boldsymbol{w}}) + \mathcal{P}(\boldsymbol{w};\boldsymbol{\eta})$ denotes the objective function in (34). Thus, (C.13) and (C.14) together state that the following result holds with probability tending to 1 as $\mathrm{T} \to \infty$:

$$\begin{array}{l} \text{for all } \widetilde{\boldsymbol{w}} \text{ in the ball } \{\widetilde{\boldsymbol{w}} : \|\widetilde{\boldsymbol{w}} - \widetilde{\boldsymbol{w}}^*\|_2 \le r_{\mathrm{T}} C_\epsilon\}, \\ \frac{\partial Q(\widetilde{\boldsymbol{w}})}{\partial w_{i,j}} \text{ and } w_{i,j} \text{ have the same sign, for all } (i,j) \in \mathcal{E}^c(\mathbf{W}^*). \end{array} \tag{C.15}$$

Using (C.15), (C.11) and $\widehat{\widetilde{\boldsymbol{w}}}$ being the local minimizer of $Q(\widetilde{\boldsymbol{w}})$, for all sufficiently large T, we have

$$\mathrm{P}\big(\widehat{w}_{i,j} = 0, \text{ for all } (i,j) \in \mathcal{E}^c(\mathbf{W}^*)\big) \ge 1 - 2\epsilon.$$

Since ϵ is arbitrary, letting $\epsilon \to 0$ gives that

$$\mathrm{P}\big(\widehat{w}_{i,j} = 0, \text{ for all } (i,j) \in \mathcal{E}^c(\mathbf{W}^*)\big) \to 1 \quad \text{as } \mathrm{T} \to \infty.$$

The proof follows from the definition of 'Ex' in (31). ■

Lemma C.8 *Assume condition* A7 *in Section* 5. *If* $d^4/\mathrm{T} \to 0$ *as* $\mathrm{T} \to \infty$, *then for any estimator* $\widehat{\widetilde{\mathbf{W}}}$ *satisfying* $\|\widehat{\widetilde{\mathbf{W}}} - \widetilde{\mathbf{W}}^*\|_{\mathrm{F}} = O_{\mathrm{P}}(\sqrt{d^2/\mathrm{T}})$, *we have*

$$\mathrm{P}(\mathrm{Mi}(\widehat{\mathbf{W}}, \mathbf{W}^*) = 0) \to 1, \quad \textit{and} \quad \mathrm{P}(\mathrm{Rv}(\widehat{\mathbf{W}}, \mathbf{W}^*) = 0) \to 1.$$

Proof Note that

$$\max_{(i,j)\in\mathcal{E}(\mathbf{W}^*)} |\widehat{w}_{i,j} - w^*_{i,j}| \le \|\widehat{\widetilde{\boldsymbol{w}}} - \widetilde{\boldsymbol{w}}^*\|_\infty \le \|\widehat{\widetilde{\boldsymbol{w}}} - \widetilde{\boldsymbol{w}}^*\|_2 = \|\widehat{\widetilde{\mathbf{W}}} - \widetilde{\mathbf{W}}^*\|_{\mathrm{F}} = O_{\mathrm{P}}(\sqrt{d^2/\mathrm{T}}).$$

Condition A7 states that $\sqrt{\mathrm{T}/d^2} \cdot \min_{(i,j)\in\mathcal{E}(\mathbf{W}^*)} |w^*_{i,j}| \to \infty$. Combining these two results, we have that as $\mathrm{T} \to \infty$,

$$\mathrm{P}\Big(\max_{(i,j)\in\mathcal{E}(\mathbf{W}^*)} |\widehat{w}_{i,j} - w^*_{i,j}| < \min_{(i,j)\in\mathcal{E}(\mathbf{W}^*)} |w^*_{i,j}|\Big) \to 1,$$

which further implies that

$$\mathrm{P}\big(\mathrm{sign}(\widehat{w}_{i,j}) = \mathrm{sign}(w^*_{i,j}), \text{ for all } (i,j) \in \mathcal{E}(\mathbf{W}^*)\big) \to 1.$$

By definitions of 'Mi' and 'Rv' in (31), Lemma C.8 directly follows. ■

Lemma C.9 *Both the loss function* $\mathcal{L}(\widetilde{\mathbf{W}})$ *in* (9) *and the weighted* L_1*-penalty function* $\mathcal{P}(\mathbf{W};\boldsymbol{\eta})$ *in* (10) *are convex, but the* DAG*-ness function* $h(\mathbf{W})$ *in* (11) *is non-convex.*

Proof Note that $\mathcal{L}(\widetilde{\mathbf{W}}) = \sum_{j=1}^d \mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\boldsymbol{\cdot},j})$, where each component $\mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\boldsymbol{\cdot},j})$ has a positive semi-definite Hessian matrix

$$\nabla^2 \mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\boldsymbol{\cdot},j}) = \frac{1}{\mathrm{T}} \int_0^{\mathrm{T}} \widetilde{\boldsymbol{x}}(t)\, \widetilde{\boldsymbol{x}}(t)^\top \, \exp\{\widetilde{\boldsymbol{w}}_{\boldsymbol{\cdot},j}^\top \widetilde{\boldsymbol{x}}(t)\}\, \mathrm{d}t,$$

indicating the convexity of $\mathcal{L}(\widetilde{\mathbf{W}})$. Also, it is clearly seen that the penalty function $\mathcal{P}(\mathbf{W};\boldsymbol{\eta}) = \sum\sum_{1\le i\neq j\le d}\eta_{i,j}\,|w_{i,j}|$ in (10) is convex.

Next, we prove that the DAG-ness function $h(\mathbf{W})$ is non-convex. For

$$\mathcal{D} = \{\mathbf{W}\in\mathbb{R}^{d\times d} : h(\mathbf{W}) = 0\}, \tag{C.16}$$

which collects $\mathbf{W}$ corresponding to weighted adjacency matrices $\widetilde{\mathbf{W}}$ of DAGs, it suffices to show that $\mathcal{D}$ is a non-convex set in $\mathbb{R}^{d\times d}$. Let $A = (a_{i,j}) \in \mathbb{R}^{d\times d}$ be the weighted adjacency matrix with $a_{1,2} = 1$, and all other entries equal to zero, $B = (b_{i,j}) \in \mathbb{R}^{d\times d}$ be the weighted adjacency matrix with $b_{2,1} = 1$, and all other entries equal to zero, and $E = (e_{i,j}) \in \mathbb{R}^{d\times d} = (A + B)/2$. Clearly, both A and B represent DAGs, and therefore $A \in \mathcal{D}$ and $B \in \mathcal{D}$. However, since $e_{1,2} = e_{2,1} = 1/2$ are non-zero, E induces a graph that contains a cycle formed by edges $(1, 2)$ and $(2, 1)$, and thus $E \notin \mathcal{D}$. This completes the proof. ■

Lemma C.10 *Assume condition* A5 *in Section* 5. *Let* $Q(\widetilde{\mathbf{W}}) = \mathcal{L}(\widetilde{\mathbf{W}}) + \mathcal{P}(\mathbf{W};\boldsymbol{\eta})$ *be the objective function in* (12). *If*

$$\inf_{\widetilde{\mathbf{W}}\in\mathbb{R}^{(d+1)\times d}:\|\widetilde{\mathbf{W}}-\widetilde{\mathbf{W}}^*\|_{\mathrm{F}}=r} Q(\widetilde{\mathbf{W}}) > Q(\widetilde{\mathbf{W}}^*) \tag{C.17}$$

for some $r \in (0,\infty)$, *then there exists a global minimizer* $\widehat{\widetilde{\mathbf{W}}}_{0,\boldsymbol{\eta}}$ *of the* DAG-*constrained optimization problem* (34) (*with* $\kappa = 0$), *such that* $\|\widehat{\widetilde{\mathbf{W}}}_{0,\boldsymbol{\eta}} - \widetilde{\mathbf{W}}^*\|_{\mathrm{F}} < r$.

Proof Note that when $\kappa = 0$, the constrained optimization problem (34) is identical to (12). It suffices to show that there exists a global minimizer $\widehat{\widetilde{\mathbf{W}}}$ of (12), such that $\|\widehat{\widetilde{\mathbf{W}}}-\widetilde{\mathbf{W}}^*\|_{\mathrm{F}} < r$. Since $h(\cdot)$ is continuous, we have that $\{\widetilde{\mathbf{W}} \in \mathbb{R}^{(d+1)\times d} : h(\mathbf{W}) = 0\}$ is a closed set, further implying that $\{\widetilde{\mathbf{W}} \in \mathbb{R}^{(d+1)\times d} : \|\widetilde{\mathbf{W}} - \widetilde{\mathbf{W}}^*\|_{\mathrm{F}} \le r,\, h(\mathbf{W}) = 0\}$ is a compact set in $\mathbb{R}^{(d+1)\times d}$. Then by the extreme value theorem and the fact that the function $Q(\widetilde{\mathbf{W}})$ is continuous, there exists a global minimizer $\widehat{\widetilde{\mathbf{W}}}$ of $Q(\widetilde{\mathbf{W}})$ over $\{\widetilde{\mathbf{W}} \in \mathbb{R}^{(d+1)\times d} : \|\widetilde{\mathbf{W}} - \widetilde{\mathbf{W}}^*\|_{\mathrm{F}} \le r,\ h(\mathbf{W}) = 0\}$, i.e.,

$$\widehat{\widetilde{\mathbf{W}}} = \arg\min_{\widetilde{\mathbf{W}}:\|\widetilde{\mathbf{W}}-\widetilde{\mathbf{W}}^*\|_{\mathrm{F}}\le r,\ h(\mathbf{W})=0} Q(\widetilde{\mathbf{W}}).$$

Condition A5 gives that $\widetilde{\mathbf{W}}^* \in \{\widetilde{\mathbf{W}} : \|\widetilde{\mathbf{W}}-\widetilde{\mathbf{W}}^*\|_{\mathrm{F}} \le r,\ h(\mathbf{W}) = 0\}$, implying that $Q(\widetilde{\mathbf{W}}^*) \ge Q(\widehat{\widetilde{\mathbf{W}}})$. Therefore, we obtain

$$\inf_{\widetilde{\mathbf{W}}:\|\widetilde{\mathbf{W}}-\widetilde{\mathbf{W}}^*\|_{\mathrm{F}}=r} Q(\widetilde{\mathbf{W}}) > Q(\widetilde{\mathbf{W}}^*) \ge Q(\widehat{\widetilde{\mathbf{W}}}), \tag{C.18}$$

which proves that $\widehat{\widetilde{\mathbf{W}}}$ minimizes $Q(\widetilde{\mathbf{W}})$ over $\{\widetilde{\mathbf{W}} : \|\widetilde{\mathbf{W}} - \widetilde{\mathbf{W}}^*\|_{\mathrm{F}} < r,\ h(\widetilde{\mathbf{W}}) = 0\}$.

The remaining proof is to show that $\widehat{\widetilde{\mathbf{W}}}$ is a global minimizer of (12). If this is violated, then there exists a matrix $\widetilde{\mathbf{W}}^\dagger \in \mathbb{R}^{(d+1)\times d}$, satisfying $\|\widetilde{\mathbf{W}}^\dagger - \widetilde{\mathbf{W}}^*\|_{\mathrm{F}} > r$ and $h(\mathbf{W}^\dagger) = 0$,

but $Q(\widetilde{\mathbf{W}}^\dagger) < Q(\widehat{\widetilde{\mathbf{W}}})$. Since $\|\widetilde{\mathbf{W}}^\dagger - \widetilde{\mathbf{W}}^*\|_{\mathrm{F}} > r > \|\widehat{\widetilde{\mathbf{W}}} - \widetilde{\mathbf{W}}^*\|_{\mathrm{F}}$, there exists a matrix $\widetilde{\mathbf{W}}^\ddagger$ on the line segment between $\widetilde{\mathbf{W}}^\dagger$ and $\widehat{\widetilde{\mathbf{W}}}$, with $\|\widetilde{\mathbf{W}}^\ddagger - \widetilde{\mathbf{W}}^*\|_{\mathrm{F}} = r$. By (C.18), we have

$$Q(\widetilde{\mathbf{W}}^\ddagger) \geq \inf_{\widetilde{\mathbf{W}}:\|\widetilde{\mathbf{W}}-\widetilde{\mathbf{W}}^*\|_{\mathrm{F}}=r} Q(\widetilde{\mathbf{W}}) > Q(\widehat{\widetilde{\mathbf{W}}}) > Q(\widetilde{\mathbf{W}}^\dagger),$$

implying that $Q(\widetilde{\mathbf{W}}^\ddagger)$ exceeds both $Q(\widehat{\widetilde{\mathbf{W}}})$ and $Q(\widetilde{\mathbf{W}}^\dagger)$. This contradicts the convexity of $Q(\widetilde{\mathbf{W}})$, a fact verified by Lemma C.9. ■

Lemma C.11 *If an estimator $\widehat{\mathbf{W}}$ of $\mathbf{W}^*$ satisfies that $\mathcal{G}(\widehat{\mathbf{W}}) \in \mathbb{D}$, $\mathrm{Mi}(\widehat{\mathbf{W}}, \mathbf{W}^*) = 0$, and $\mathrm{Rv}(\widehat{\mathbf{W}}, \mathbf{W}^*) = 0$, then we have*

$$\mathrm{Ex}(\widehat{\mathbf{W}}, \mathbf{W}^*) \leq d(d-1)/2 - s^*.$$

Proof Applying Proposition 2.1.3 of Bang-Jensen and Gutin (2008) and the fact that $\mathcal{G}(\widehat{\mathbf{W}}) \in \mathbb{D}$, there exists a permutation $\boldsymbol{\pi} = (\pi(1), \ldots, \pi(d))$ of $\{1, \ldots, d\}$, such that the permuted weighted adjacency matrix $\widehat{\mathbf{W}}^{(\boldsymbol{\pi})} = (\widehat{w}_{i,j}^{(\boldsymbol{\pi})}) \in \mathbb{R}^{d\times d}$ with entries $\widehat{w}_{i,j}^{(\boldsymbol{\pi})} = \widehat{w}_{\pi(i),\pi(j)}$ is a strictly upper triangular matrix, satisfying

$$\widehat{w}_{i,j}^{(\boldsymbol{\pi})} = 0, \quad \text{for any } i \geq j, \quad i, j = 1, \ldots, d. \tag{C.19}$$

This tells that

$$|\mathcal{E}(\widehat{\mathbf{W}})| = |\mathcal{E}(\widehat{\mathbf{W}}^{(\boldsymbol{\pi})})| \leq d(d-1)/2. \tag{C.20}$$

Let $C = \sum\sum_{1\leq i\neq j\leq d} \mathrm{I}(w_{i,j}^* > 0,\, \widehat{w}_{i,j} > 0) + \sum\sum_{1\leq i\neq j\leq d} \mathrm{I}(w_{i,j}^* < 0,\, \widehat{w}_{i,j} < 0)$. Then

$$\begin{aligned}
\sum_{1\leq i\neq j\leq d}\sum \mathrm{I}(\widehat{w}_{i,j} \neq 0) &= \mathrm{Ex}(\widehat{\mathbf{W}}, \mathbf{W}^*) + \mathrm{Rv}(\widehat{\mathbf{W}}, \mathbf{W}^*) + C \\
&= \mathrm{Ex}(\widehat{\mathbf{W}}, \mathbf{W}^*) + C, \\
\sum_{1\leq i\neq j\leq d}\sum \mathrm{I}(w_{i,j}^* \neq 0) &= \mathrm{Mi}(\widehat{\mathbf{W}}, \mathbf{W}^*) + \mathrm{Rv}(\widehat{\mathbf{W}}, \mathbf{W}^*) + C \\
&= C,
\end{aligned}$$

under the assumptions $\mathrm{Mi}(\widehat{\mathbf{W}}, \mathbf{W}^*) = 0$ and $\mathrm{Rv}(\widehat{\mathbf{W}}, \mathbf{W}^*) = 0$. It follows that

$$\begin{aligned}
\mathrm{Ex}(\widehat{\mathbf{W}}, \mathbf{W}^*) &= \sum_{1\leq i\neq j\leq d}\sum \mathrm{I}(\widehat{w}_{i,j} \neq 0) - \sum_{1\leq i\neq j\leq d}\sum \mathrm{I}(w_{i,j}^* \neq 0) \\
&\leq d(d-1)/2 - s^*,
\end{aligned}$$

where $\sum\sum_{1\leq i\neq j\leq d} \mathrm{I}(\widehat{w}_{i,j} \neq 0) = |\mathcal{E}(\widehat{\mathbf{W}})| \leq d(d-1)/2$ utilizes (C.20) and $\sum\sum_{1\leq i\neq j\leq d} \mathrm{I}(w_{i,j}^* \neq 0) = s^*$. This completes the proof. ■

Lemma C.12 *Assume condition* A5 *in Section* 5. *If there exists a global minimizer* $\widehat{\widetilde{\mathbf{W}}}_{\infty,\boldsymbol{\eta}}$ *of* (34) (*with* $\kappa = \infty$) *that achieves* $\mathrm{P}(\mathrm{SHD}(\widehat{\mathbf{W}}_{\infty,\boldsymbol{\eta}}, \mathbf{W}^*) = 0) \to 1$ *as* $\mathrm{T} \to \infty$, *then there exists a global minimizer* $\widehat{\widetilde{\mathbf{W}}}_{0,\boldsymbol{\eta}}$ *of* (34) (*with* $\kappa = 0$) *that satisfies* $\mathrm{P}(\widehat{\widetilde{\mathbf{W}}}_{0,\boldsymbol{\eta}} = \widehat{\widetilde{\mathbf{W}}}_{\infty,\boldsymbol{\eta}}) \to 1$ *as* $\mathrm{T} \to \infty$ *and, therefore, achieves* $\mathrm{P}(\mathrm{SHD}(\widehat{\mathbf{W}}_{0,\boldsymbol{\eta}}, \mathbf{W}^*) = 0) \to 1$ *as* $\mathrm{T} \to \infty$.

Proof Let $Q(\widetilde{\mathbf{W}}) = \mathcal{L}(\widetilde{\mathbf{W}}) + \mathcal{P}(\mathbf{W}; \boldsymbol{\eta})$ be the objective function in (34). For an estimator $\widehat{\widetilde{\mathbf{W}}}_{\infty,\boldsymbol{\eta}}$ with $\mathrm{P}(\mathrm{SHD}(\widehat{\mathbf{W}}, \mathbf{W}^*) = 0) \to 1$, we construct an estimator $\widehat{\widetilde{\mathbf{W}}}_{\mathrm{D}}$ as follows:

$$\widehat{\widetilde{\mathbf{W}}}_{\mathrm{D}} = \begin{cases} \widehat{\widetilde{\mathbf{W}}}_{\infty,\boldsymbol{\eta}}, & \text{if } \mathrm{SHD}(\widehat{\mathbf{W}}_{\infty,\boldsymbol{\eta}}, \mathbf{W}^*) = 0, \\ \widehat{\widetilde{\mathbf{W}}}, & \text{if } \mathrm{SHD}(\widehat{\mathbf{W}}_{\infty,\boldsymbol{\eta}}, \mathbf{W}^*) > 0, \end{cases} \tag{C.21}$$

where $\widehat{\widetilde{\mathbf{W}}}$ is an arbitrary global minimizer of (8), satisfying $h(\widehat{\widetilde{\mathbf{W}}}) = 0$. By condition A5, the true weighted adjacency matrix $\mathbf{W}^*$ represents a DAG. If $\mathrm{SHD}(\widehat{\mathbf{W}}_{\infty,\boldsymbol{\eta}}, \mathbf{W}^*) = 0$ holds, then $\widehat{\mathbf{W}}_{\infty,\boldsymbol{\eta}}$ also represents a DAG, satisfying $h(\widehat{\mathbf{W}}_{\infty,\boldsymbol{\eta}}) = 0$. Therefore, $\widehat{\widetilde{\mathbf{W}}}_{\mathrm{D}}$ in either case of (C.21) satisfies the DAG constraint $h(\widehat{\widetilde{\mathbf{W}}}_{\mathrm{D}}) = 0$, and minimizes the objective function $Q(\widetilde{\mathbf{W}})$. This implies that $\widehat{\widetilde{\mathbf{W}}}_{\mathrm{D}}$ defined by (C.21) is also a global minimizer of (34) with $\kappa = 0$.

It follows from (C.21) and $\mathrm{P}(\mathrm{SHD}(\widehat{\mathbf{W}}_{\infty,\boldsymbol{\eta}}, \mathbf{W}^*) = 0) \to 1$ that $\mathrm{P}(\widehat{\widetilde{\mathbf{W}}}_{\mathrm{D}} = \widehat{\widetilde{\mathbf{W}}}_{\infty,\boldsymbol{\eta}}) \to 1$ and $\mathrm{P}(\mathrm{SHD}(\widehat{\mathbf{W}}_{\mathrm{D}}, \mathbf{W}^*) = 0) \to 1$ as $\mathrm{T} \to \infty$. Setting $\widehat{\widetilde{\mathbf{W}}}_{0,\boldsymbol{\eta}} = \widehat{\widetilde{\mathbf{W}}}_{\mathrm{D}}$ completes the proof. ■

Lemma C.13 *Assume that the sequences* $\{\alpha^{(k)}\}_{k\geq 0}$ *and* $\{\rho^{(k)}\}_{k\geq 0}$ *satisfy condition* (22). *Let* $\{\mathbf{W}^{(k)}\}_{k\geq 0}$ *be a sequence of weighted adjacency matrices. Then we have:*

(a)

$$\limsup_{k\to\infty} \{\alpha^{(k)}\, h(\mathbf{W}^{(k)}) + 2^{-1}\rho^{(k)}\, h^2(\mathbf{W}^{(k)})\} \geq 0. \tag{C.22}$$

(b) *Moreover, if* $\limsup_{k\to\infty} \{\alpha^{(k)}\, h(\mathbf{W}^{(k)}) + 2^{-1}\rho^{(k)}\, h^2(\mathbf{W}^{(k)})\} < \infty$, *then*

$$h(\mathbf{W}^{(k)}) \to 0 \quad \text{as } k \to \infty. \tag{C.23}$$

Proof Note that the condition $\inf_{k\geq 0} \alpha^{(k)} > -\infty$ in (22) implies the existence of a constant $c \in \mathbb{R}$ such that $\alpha^{(k)} \geq -|c|$ for all $k \geq 0$.

We first prove part (a). If $\{\alpha^{(k)}\}_{k\geq 0}$ has a non-negative subsequence, then (C.22) obviously holds. Now, consider the case where $-|c| \leq \alpha^{(k)} < 0$ for all sufficiently large k. The condition $\max\{\alpha^{(k)}, \rho^{(k)}\} \to \infty$ in (22) implies that $\rho^{(k)} \to \infty$.

(i) In the case where $h(\mathbf{W}^{(k)}) \to 0$ holds, the fact that $-|c| \leq \alpha^{(k)} < 0$ for all large k gives:

$$\limsup_{k\to\infty} \{\alpha^{(k)}\, h(\mathbf{W}^{(k)}) + 2^{-1}\rho^{(k)}\, h^2(\mathbf{W}^{(k)})\} = \limsup_{k\to\infty} \{2^{-1}\rho^{(k)}\, h^2(\mathbf{W}^{(k)})\} \geq 0.$$

(ii) In the case where $h(\mathbf{W}^{(k)}) \to 0$ fails, $\rho^{(k)} \to \infty$ implies that $\limsup_{k\to\infty}\{2^{-1}\rho^{(k)}\, h(\mathbf{W}^{(k)})\} = \infty$. Therefore, there exists a subsequence $\{k_\ell\}_{\ell\geq 1}$ of $\{1, 2, \ldots\}$, such that $\alpha^{(k_\ell)} + 2^{-1}\rho^{(k_\ell)}\, h(\mathbf{W}^{(k_\ell)}) > 0$ for all $\{k_\ell\}_{\ell\geq 1}$. It follows that for all $\ell \geq 1$:

$$\alpha^{(k_\ell)}\, h(\mathbf{W}^{(k_\ell)}) + 2^{-1}\rho^{(k_\ell)}\, h^2(\mathbf{W}^{(k_\ell)}) = \{\alpha^{(k_\ell)} + 2^{-1}\rho^{(k_\ell)}\, h(\mathbf{W}^{(k_\ell)})\}h(\mathbf{W}^{(k_\ell)}) \geq 0,$$

which completes the proof of (C.22).

Next, we prove part (b). If $h(\mathbf{W}^{(k)}) > 1$, then $h^2(\mathbf{W}^{(k)}) > h(\mathbf{W}^{(k)})$, implying:

$$\alpha^{(k)}\, h(\mathbf{W}^{(k)}) + 2^{-1}\rho^{(k)}\, h^2(\mathbf{W}^{(k)}) \geq \{\alpha^{(k)} + 2^{-1}\rho^{(k)}\}h(\mathbf{W}^{(k)}).$$

If $0 \leq h(\mathbf{W}^{(k)}) \leq 1$, then $0 \leq h^2(\mathbf{W}^{(k)}) \leq h(\mathbf{W}^{(k)}) \leq 1$. Together with the fact that $\alpha^{(k)} + |c| \geq 0$, we get:

$$\begin{aligned}
&\alpha^{(k)}\, h(\mathbf{W}^{(k)}) + 2^{-1}\rho^{(k)}\, h^2(\mathbf{W}^{(k)}) \\
=&\{\alpha^{(k)} + |c|\}\, h(\mathbf{W}^{(k)}) + 2^{-1}\rho^{(k)}\, h^2(\mathbf{W}^{(k)}) - |c|\, h(\mathbf{W}^{(k)}) \\
\geq&\{\alpha^{(k)} + |c| + 2^{-1}\rho^{(k)}\}\, h^2(\mathbf{W}^{(k)}) - |c|\, h(\mathbf{W}^{(k)}) \\
\geq&\{\alpha^{(k)} + 2^{-1}\rho^{(k)}\}\, h^2(\mathbf{W}^{(k)}) - |c|.
\end{aligned}$$

Combining the above two inequalities, we obtain:

$$\alpha^{(k)}\, h(\mathbf{W}^{(k)}) + 2^{-1}\rho^{(k)}\, h^2(\mathbf{W}^{(k)}) \geq \{\alpha^{(k)} + 2^{-1}\rho^{(k)}\}\, \min\{h^2(\mathbf{W}^{(k)}), h(\mathbf{W}^{(k)})\} - |c|.$$

This, along with the condition $\limsup_{k\to\infty}\{\alpha^{(k)}\, h(\mathbf{W}^{(k)}) + 2^{-1}\rho^{(k)}\, h^2(\mathbf{W}^{(k)})\} < \infty$, implies:

$$\begin{aligned}
\infty >& \limsup_{k\to\infty}\{\alpha^{(k)}\, h(\mathbf{W}^{(k)}) + 2^{-1}\rho^{(k)}\, h^2(\mathbf{W}^{(k)})\} + |c| \\
\geq& \limsup_{k\to\infty}\{\alpha^{(k)} + 2^{-1}\rho^{(k)}\}\, \min\{h^2(\mathbf{W}^{(k)}), h(\mathbf{W}^{(k)})\}. \qquad \text{(C.24)}
\end{aligned}$$

Note that conditions $\rho^{(k)} \geq 0$, $\alpha^{(k)} \geq -|c|$, and $\max\{\alpha^{(k)}, \rho^{(k)}\} \to \infty$ yield $\alpha^{(k)} + 2^{-1}\rho^{(k)} \to \infty$. Using this and (C.24), we conclude that $\min\{h^2(\mathbf{W}^{(k)}), h(\mathbf{W}^{(k)})\} \to 0$ as $k \to \infty$, from which (C.23) follows directly. ■

Lemma C.14 *Assume condition* A6′ *in Section* 5. *Suppose that* $\inf_{k\geq 0}\alpha^{(k)} > -\infty$ *and* $\inf_{k\geq 0}\rho^{(k)} > 0$. *Then, the optimization problem* (19) *has at least one global minimizer for each integer* $k \geq 0$. *Moreover, there exists a constant* $r \in (0, \infty)$, *such that any global minimizer* $\widetilde{\mathbf{W}}^{(k+1)}$ *of* (19) *satisfies* $\|\widetilde{\mathbf{W}}^{(k+1)}\|_{\mathrm{F}} \leq r$.

Proof To prove Lemma C.14, it suffices to show the existence of a constant $r \in (0, \infty)$ such that, for all integers $k \geq 0$, the inequality

$$\inf_{\widetilde{\boldsymbol{w}}:\|\widetilde{\boldsymbol{w}}\|_2 > r} L(\widetilde{\boldsymbol{w}}, \alpha^{(k)}; \rho^{(k)}) > L(\mathbf{0}, \alpha^{(k)}; \rho^{(k)}) \qquad \text{(C.25)}$$

holds.

Consider two vectors $\widetilde{\boldsymbol{w}} = (\widetilde{\boldsymbol{w}}_{\bullet,1}^\top, \ldots, \widetilde{\boldsymbol{w}}_{\bullet,d}^\top)^\top \in \mathbb{R}^{d^2+d}$ and $\widetilde{\boldsymbol{w}}^\dagger = (\widetilde{\boldsymbol{w}}_{\bullet,1}^{\dagger\top}, \ldots, \widetilde{\boldsymbol{w}}_{\bullet,d}^{\dagger\top})^\top \in \mathbb{R}^{d^2+d}$. Employing Taylor's expansion, we express the difference in the loss function as

$$\begin{aligned}
&\mathcal{L}(\widetilde{\boldsymbol{w}}) - \mathcal{L}(\widetilde{\boldsymbol{w}}^\dagger) \\
=&\sum_{j=1}^{d} \{\mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\bullet,j}) - \mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\bullet,j}^\dagger)\} \\
=&\sum_{j=1}^{d} \{\nabla \mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\bullet,j}^\dagger)^\top (\widetilde{\boldsymbol{w}}_{\bullet,j} - \widetilde{\boldsymbol{w}}_{\bullet,j}^\dagger) + 2^{-1} (\widetilde{\boldsymbol{w}}_{\bullet,j} - \widetilde{\boldsymbol{w}}_{\bullet,j}^\dagger)^\top \nabla^2 \mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\bullet,j}^\ddagger)(\widetilde{\boldsymbol{w}}_{\bullet,j} - \widetilde{\boldsymbol{w}}_{\bullet,j}^\dagger)\}, \quad \text{(C.26)}
\end{aligned}$$

where $\widetilde{\boldsymbol{w}}_{\bullet,j}^\ddagger$ lies between $\widetilde{\boldsymbol{w}}_{\bullet,j}$ and $\widetilde{\boldsymbol{w}}_{\bullet,j}^\dagger$. Utilizing (C.26) and condition A6′, we obtain

$$\begin{aligned}
\mathcal{L}(\widetilde{\boldsymbol{w}}) - \mathcal{L}(\widetilde{\boldsymbol{w}}^\dagger) &\geq \sum_{j=1}^{d} \{\nabla \mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\bullet,j}^\dagger)^\top (\widetilde{\boldsymbol{w}}_{\bullet,j} - \widetilde{\boldsymbol{w}}_{\bullet,j}^\dagger) + 2^{-1} C \|\widetilde{\boldsymbol{w}}_{\bullet,j} - \widetilde{\boldsymbol{w}}_{\bullet,j}^\dagger\|_2^2\} \\
&= \nabla \mathcal{L}(\widetilde{\boldsymbol{w}}^\dagger)^\top (\widetilde{\boldsymbol{w}} - \widetilde{\boldsymbol{w}}^\dagger) + 2^{-1} C \|\widetilde{\boldsymbol{w}} - \widetilde{\boldsymbol{w}}^\dagger\|_2^2, \quad \text{(C.27)}
\end{aligned}$$

where $C \in (0, \infty)$ is a constant. Letting $\widetilde{\boldsymbol{w}}^\dagger = \mathbf{0}$ in (C.27) and using Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}
\mathcal{L}(\widetilde{\boldsymbol{w}}) &\geq \mathcal{L}(\mathbf{0}) + \nabla \mathcal{L}(\mathbf{0})^\top \widetilde{\boldsymbol{w}} + 2^{-1} C \|\widetilde{\boldsymbol{w}}\|_2^2 \\
&\geq \mathcal{L}(\mathbf{0}) - \|\nabla \mathcal{L}(\mathbf{0})\|_2 \, \|\widetilde{\boldsymbol{w}}\|_2 + 2^{-1} C \|\widetilde{\boldsymbol{w}}\|_2^2. \quad \text{(C.28)}
\end{aligned}$$

On the other hand, the conditions $\inf_{k \geq 0} \alpha^{(k)} > -\infty$ and $\inf_{k \geq 0} \rho^{(k)} > 0$ imply the existence of constants $C_1 \in \mathbb{R}$ and $C_2 \in (0, \infty)$ such that $\alpha^{(k)} \geq C_1$ and $\rho^{(k)} \geq C_2$. Combining these facts with $h(\boldsymbol{w}) \geq 0$, it is easy to verify that for all integers $k \geq 0$, we have

$$\alpha^{(k)} h(\boldsymbol{w}) + 2^{-1} \rho^{(k)} h^2(\boldsymbol{w}) \geq C_3, \quad \text{(C.29)}$$

for a negative constant $C_3 = -C_1^2/(2C_2) < 0$. By (C.28), (C.29), and the fact that $\mathcal{P}(\boldsymbol{w}; \boldsymbol{\eta}) \geq 0$, we obtain

$$\begin{aligned}
L(\widetilde{\boldsymbol{w}}, \alpha^{(k)}; \rho^{(k)}) &= \mathcal{L}(\widetilde{\boldsymbol{w}}) + \mathcal{P}(\boldsymbol{w}; \boldsymbol{\eta}) + \alpha^{(k)} h(\boldsymbol{w}) + 2^{-1} \rho^{(k)} h^2(\boldsymbol{w}) \\
&\geq \mathcal{L}(\mathbf{0}) - \|\nabla \mathcal{L}(\mathbf{0})\|_2 \, \|\widetilde{\boldsymbol{w}}\|_2 + 2^{-1} C \|\widetilde{\boldsymbol{w}}\|_2^2 + C_3.
\end{aligned}$$

Applying the quadratic formula yields

$$\inf_{\widetilde{\boldsymbol{w}}: \|\widetilde{\boldsymbol{w}}\|_2 > r} \{-\|\nabla \mathcal{L}(\mathbf{0})\|_2 \, \|\widetilde{\boldsymbol{w}}\|_2 + 2^{-1} C \|\widetilde{\boldsymbol{w}}\|_2^2 + C_3\} > 0,$$

where $r = \{\|\nabla \mathcal{L}(\mathbf{0})\|_2 + \sqrt{\|\nabla \mathcal{L}(\mathbf{0})\|_2^2 - 2C\, C_3}\}/C + 1 \in (0, \infty)$. Combining these inequalities, we get

$$\begin{aligned}
\inf_{\widetilde{\boldsymbol{w}}: \|\widetilde{\boldsymbol{w}}\|_2 > r} L(\widetilde{\boldsymbol{w}}, \alpha^{(k)}; \rho^{(k)}) &\geq \mathcal{L}(\mathbf{0}) + \inf_{\widetilde{\boldsymbol{w}}: \|\widetilde{\boldsymbol{w}}\|_2 > r} \{-\|\nabla \mathcal{L}(\mathbf{0})\|_2 \, \|\widetilde{\boldsymbol{w}}\|_2 + 2^{-1} C \|\widetilde{\boldsymbol{w}}\|_2^2 + C_3\} \\
&> \mathcal{L}(\mathbf{0}) \\
&= L(\mathbf{0}, \alpha^{(k)}; \rho^{(k)}),
\end{aligned}$$

which proves (C.25). ■

C.3 Proofs of main Theorems and Lemmas

Proof of Theorem 1. Let $Q(\widetilde{\mathbf{W}}) = \mathcal{L}(\widetilde{\mathbf{W}}) + \mathcal{P}(\mathbf{W};\boldsymbol{\eta})$ denote the objective function in (12), and define

$$Q_{\inf} = \inf_{\widetilde{\mathbf{W}}\in\mathbb{R}^{(d+1)\times d}: h(\mathbf{W})=0} Q(\widetilde{\mathbf{W}})$$

as the infimum value of (12).

We first establish result (i). For any $\widetilde{\mathbf{W}} \in \mathbb{R}^{(d+1)\times d}$ satisfying $h(\mathbf{W}) = 0$, we have

$$L(\widetilde{\mathbf{W}}, \alpha; \rho) = Q(\widetilde{\mathbf{W}}) + \alpha\, h(\mathbf{W}) + 2^{-1}\rho\, h^2(\mathbf{W}) = Q(\widetilde{\mathbf{W}}).$$

For any given $\alpha \in \mathbb{R}$ and $\rho > 0$, we have

$$\inf_{\widetilde{\mathbf{W}}\in\mathbb{R}^{(d+1)\times d}: h(\mathbf{W})=0} L(\widetilde{\mathbf{W}}, \alpha; \rho) = \inf_{\widetilde{\mathbf{W}}\in\mathbb{R}^{(d+1)\times d}: h(\mathbf{W})=0} Q(\widetilde{\mathbf{W}}) = Q_{\inf}. \tag{C.30}$$

Since $\widetilde{\mathbf{W}}^{(k+1)}$ in (19) minimizes the function $L(\widetilde{\mathbf{W}}, \alpha^{(k)}; \rho^{(k)})$ over $\widetilde{\mathbf{W}} \in \mathbb{R}^{(d+1)\times d}$, we get

$$L(\widetilde{\mathbf{W}}^{(k+1)}, \alpha^{(k)}; \rho^{(k)}) \le L(\widetilde{\mathbf{W}}, \alpha^{(k)}; \rho^{(k)}), \quad \text{for all } \widetilde{\mathbf{W}} \in \mathbb{R}^{(d+1)\times d}. \tag{C.31}$$

Combining (C.31) and (C.30), we obtain the relation for any integer $k > 0$:

$$\begin{aligned} & Q(\widetilde{\mathbf{W}}^{(k+1)}) + \alpha^{(k)}\, h(\mathbf{W}^{(k+1)}) + 2^{-1}\rho^{(k)}\, h^2(\mathbf{W}^{(k+1)}) \\ =& L(\widetilde{\mathbf{W}}^{(k+1)}, \alpha^{(k)}; \rho^{(k)}) \\ \le& \inf_{\widetilde{\mathbf{W}}\in\mathbb{R}^{(d+1)\times d}: h(\mathbf{W})=0} L(\widetilde{\mathbf{W}}, \alpha^{(k)}; \rho^{(k)}) = Q_{\inf}. \end{aligned} \tag{C.32}$$

For any limit point $\widetilde{\mathbf{W}}^{\dagger}$ of the sequence $\{\widetilde{\mathbf{W}}^{(k)}\}_{k\ge1}$, there exists a subsequence $\{\widetilde{\mathbf{W}}^{(k_\ell+1)}\}_{\ell\ge1}$ of $\{\widetilde{\mathbf{W}}^{(k)}\}_{k\ge1}$ satisfying $\widetilde{\mathbf{W}}^{\dagger} = \lim_{\ell\to\infty} \widetilde{\mathbf{W}}^{(k_\ell+1)}$. By (C.32), we have

$$\begin{aligned} Q_{\inf} &\ge \limsup_{\ell\to\infty} \big\{ Q(\widetilde{\mathbf{W}}^{(k_\ell+1)}) + \alpha^{(k_\ell)}\, h(\mathbf{W}^{(k_\ell+1)}) + 2^{-1}\rho^{(k_\ell)}\, h^2(\mathbf{W}^{(k_\ell+1)}) \big\} \\ &= Q(\widetilde{\mathbf{W}}^{\dagger}) + \limsup_{\ell\to\infty} \big\{ \alpha^{(k_\ell)}\, h(\mathbf{W}^{(k_\ell+1)}) + 2^{-1}\rho^{(k_\ell)}\, h^2(\mathbf{W}^{(k_\ell+1)}) \big\}. \end{aligned} \tag{C.33}$$

This implies

$$\limsup_{\ell\to\infty} \big\{ \alpha^{(k_\ell)}\, h(\mathbf{W}^{(k_\ell+1)}) + 2^{-1}\rho^{(k_\ell)}\, h^2(\mathbf{W}^{(k_\ell+1)}) \big\} < \infty.$$

By (C.23) in Lemma C.13, it follows that $h(\mathbf{W}^{(k_\ell+1)}) \to 0$ as $\ell \to \infty$. Using the facts that $h(\mathbf{W}^{(k_\ell+1)}) \to 0$, $\mathbf{W}^{(k_\ell+1)} \to \mathbf{W}^{\dagger}$, and the continuity of the function $h(\cdot)$, we have $h(\mathbf{W}^{\dagger}) = 0$. Therefore, $\widetilde{\mathbf{W}}^{\dagger}$ satisfies the equality constraint $h(\mathbf{W}^{\dagger}) = 0$, implying that

$$Q(\widetilde{\mathbf{W}}^{\dagger}) \ge \inf_{\widetilde{\mathbf{W}}\in\mathbb{R}^{(d+1)\times d}: h(\mathbf{W})=0} Q(\widetilde{\mathbf{W}}) = Q_{\inf}.$$

On the other hand, by (C.22) and (C.33), we have

$$Q(\widetilde{\mathbf{W}}^{\dagger}) \le Q(\widetilde{\mathbf{W}}^{\dagger}) + \limsup_{\ell\to\infty}\{\alpha^{(k_\ell)}\, h(\mathbf{W}^{(k_\ell+1)}) + 2^{-1}\rho^{(k_\ell)}\, h^2(\mathbf{W}^{(k_\ell+1)})\} \le Q_{\inf}.$$

Combining the above two results gives that $Q(\widetilde{\mathbf{W}}^{\dagger}) = Q_{\inf}$, which implies that $\widetilde{\mathbf{W}}^{\dagger}$ is a global minimizer of (12). This proves result (i).

Next, we prove result (ii). If condition A6′ is satisfied, then Lemma C.14 gives that $\{\|\widetilde{\mathbf{W}}^{(k)}\|_{\mathrm{F}}\}_{k\ge1}$ is bounded. Using this and the accumulative point principle, we conclude that $\{\widetilde{\mathbf{W}}^{(k)}\}_{k\ge1}$ has at least one limit point. Moreover, if (12) has a unique global minimizer $\widehat{\widetilde{\mathbf{W}}}$, then result (i) implies that $\{\widetilde{\mathbf{W}}^{(k)}\}_{k\ge1}$ precisely has one unique limit point $\widetilde{\mathbf{W}}^{\dagger} = \widehat{\widetilde{\mathbf{W}}}$. Combining this with the fact that $\{\|\widetilde{\mathbf{W}}^{(k)}\|_{\mathrm{F}}\}_{k\ge1}$ is bounded, we have $\widetilde{\mathbf{W}}^{(k)} \to \widehat{\widetilde{\mathbf{W}}}$ as $k \to \infty$. The proof is completed. ■

Proof of Lemma 1. Note that $\mathcal{P}(\mathbf{W};\boldsymbol{\eta}) = 0$ is a special case of the weighted L_1-penalty, with regularization parameters in $\boldsymbol{\eta} = \mathbf{0}$ satisfying condition (37). By Lemma C.6, there exists a local minimizer $\widehat{\widetilde{\mathbf{W}}}_{\infty,\mathbf{0}}$ of (34) satisfying $\|\widehat{\widetilde{\mathbf{W}}}_{\infty,\mathbf{0}} - \widetilde{\mathbf{W}}^*\|_{\mathrm{F}} = O_{\mathrm{P}}(\sqrt{d^2/\mathrm{T}})$, where the objective function in (34) reduces to $\mathcal{L}(\widetilde{\mathbf{W}})$, a convex function verified by Lemma C.9. Hence, any local minimizer $\widehat{\widetilde{\mathbf{W}}}_{\infty,\mathbf{0}}$ must also be a global minimizer of (34). Write $\widehat{\widetilde{\mathbf{W}}}_{\infty,\mathbf{0}} = (\widehat{\widetilde{\boldsymbol{w}}}_{\bullet,1},\ldots,\widehat{\widetilde{\boldsymbol{w}}}_{\bullet,d})$, with $\widehat{\widetilde{\boldsymbol{w}}}_{\bullet,j} = (\widehat{w}_{0,j},\widehat{w}_{1,j},\ldots,\widehat{w}_{d,j})^{\top} \in \mathbb{R}^{d+1}$. Moreover, it follows from Lemma C.8 that $\mathrm{P}(\mathrm{Mi}(\widehat{\mathbf{W}}_{\infty,\mathbf{0}},\mathbf{W}^*) = 0$ and $\mathrm{Rv}(\widehat{\mathbf{W}}_{\infty,\mathbf{0}},\mathbf{W}^*) = 0) \to 1$.

We next show that $\mathrm{P}(\mathrm{Ex}(\widehat{\mathbf{W}}_{\infty,\mathbf{0}},\mathbf{W}^*) = d(d-1) - s^*) = 1$. By the definition of 'Ex' in (31), we have

$$\begin{aligned}\mathrm{Ex}(\widehat{\mathbf{W}}_{\infty,\mathbf{0}},\mathbf{W}^*) &= \sum_{1\le i\ne j\le d}\sum \mathrm{I}\big(\widehat{w}_{i,j}\ne 0, w^*_{i,j} = 0\big)\\ &= \sum_{(i,j)\in\mathcal{E}^c(\mathbf{W}^*)}\sum \mathrm{I}(\widehat{w}_{i,j}\ne 0)\\ &\le |\mathcal{E}^c(\mathbf{W}^*)| = d(d-1) - s^*,\end{aligned}$$

where the inequality becomes an equality if and only if $\widehat{w}_{i,j} \ne 0$ for all $(i,j) \in \mathcal{E}^c(\mathbf{W}^*)$, namely,

$$\begin{aligned}&\mathrm{Ex}(\widehat{\mathbf{W}}_{\infty,\mathbf{0}},\mathbf{W}^*) = d(d-1) - s^*\\ \text{is equivalent to}\quad &\widehat{w}_{i,j}\ne 0 \text{ for all } (i,j)\in\mathcal{E}^c(\mathbf{W}^*).\end{aligned} \tag{C.34}$$

Since $\widehat{\widetilde{\mathbf{W}}}_{\infty,\mathbf{0}}$ minimizes $\mathcal{L}(\widetilde{\mathbf{W}}) = \sum_{j=1}^d \mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\bullet,j})$, it follows that each $\widehat{\widetilde{\boldsymbol{w}}}_{\bullet,j}$ minimizes $\mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\bullet,j})$, i.e., $\widehat{\widetilde{\boldsymbol{w}}}_{\bullet,j} = \arg\min_{\widetilde{\boldsymbol{w}}_{\bullet,j}\in\mathbb{R}^{d+1}} \mathcal{L}_{j,\mathrm{T}}(\widetilde{\boldsymbol{w}}_{\bullet,j})$. By condition A8, we have $\mathrm{P}(\widehat{w}_{i,j} = 0) = 0$, for all $i = 0,1,\ldots,d$ and $j = 1,\ldots,d$. Using this and (C.34) gives that

$$\begin{aligned}&\mathrm{P}\big(\mathrm{Ex}(\widehat{\mathbf{W}}_{\infty,\mathbf{0}},\mathbf{W}^*) = d(d-1) - s^*\big)\\ =&\mathrm{P}\big(\widehat{w}_{i,j}\ne 0 \text{ for all } (i,j)\in\mathcal{E}^c(\mathbf{W}^*)\big)\\ \ge& 1 - \sum_{(i,j)\in\mathcal{E}^c(\mathbf{W}^*)}\sum \mathrm{P}(\widehat{w}_{i,j} = 0) = 1,\end{aligned}$$

where $d(d-1) - s^* \ge 1$ comes from the fact that s^* is bounded by $d(d-1)/2$ for DAGs, as shown in (C.20). The proof is completed. ■

Proof of Theorem 2. Let $r_{\mathrm{T}} = \sqrt{d^2/\mathrm{T}}$. In the proof of Lemma C.6, we showed that for any given $\epsilon > 0$, there is a large constant C_ϵ, such that

$$\mathrm{P}\Big(\inf_{\widetilde{\boldsymbol{u}}:\|\widetilde{\boldsymbol{u}}\|_2 = C_\epsilon} Q(\widetilde{\boldsymbol{w}}^* + r_{\mathrm{T}}\, \widetilde{\boldsymbol{u}}) > Q(\widetilde{\boldsymbol{w}}^*) \Big) \geq 1 - \epsilon$$

for all sufficiently large T. Using this, (C.17), and Lemma C.10, there exists a global minimizer $\widehat{\widetilde{\mathbf{W}}}_{0,\mathbf{0}}$ of (34) (with $\kappa = 0$ and $\boldsymbol{\eta} = \mathbf{0}$), such that

$$\mathrm{P}\big(\|\widehat{\widetilde{\mathbf{W}}}_{0,\mathbf{0}} - \widetilde{\mathbf{W}}^*\|_{\mathrm{F}} < r_{\mathrm{T}} C_\epsilon \big) \geq 1 - \epsilon.$$

Hence, we have $\|\widehat{\widetilde{\mathbf{W}}}_{0,\mathbf{0}} - \widetilde{\mathbf{W}}^*\|_{\mathrm{F}} = O_{\mathrm{P}}(r_{\mathrm{T}}) = O_{\mathrm{P}}(\sqrt{d^2/\mathrm{T}})$.

By Lemma C.8, we get $\mathrm{P}(\mathrm{Mi}(\widehat{\mathbf{W}}_{0,\mathbf{0}}, \mathbf{W}^*) = 0$ and $\mathrm{Rv}(\widehat{\mathbf{W}}_{0,\mathbf{0}}, \mathbf{W}^*) = 0) \to 1$. Furthermore, using Lemma C.11 and the fact that $\mathcal{G}(\widehat{\mathbf{W}}_{0,\mathbf{0}}) \in \mathbb{D}$, we obtain $\mathrm{P}(\mathrm{Ex}(\widehat{\mathbf{W}}_{0,\mathbf{0}}, \mathbf{W}^*) \leq d(d-1)/2 - s^*) \to 1$ as $\mathrm{T} \to \infty$. This completes the proof. ■

Proof of Lemma 2. By Lemma C.6 and the fact that $\mathcal{L}(\widetilde{\mathbf{W}}) + \mathcal{P}(\mathbf{W}; \boldsymbol{\eta})$ is convex (as verified by Lemma C.9), there exists a global minimizer $\widehat{\widetilde{\mathbf{W}}}_{\infty,\boldsymbol{\eta}}$ of (34), such that $\|\widehat{\widetilde{\mathbf{W}}}_{\infty,\boldsymbol{\eta}} - \widetilde{\mathbf{W}}^*\|_{\mathrm{F}} = O_{\mathrm{P}}(\sqrt{d^2/\mathrm{T}})$. Moreover, Lemma C.8 proves $\mathrm{P}(\mathrm{Mi}(\widehat{\mathbf{W}}_{\infty,\boldsymbol{\eta}}, \mathbf{W}^*) = 0$ and $\mathrm{Rv}(\widehat{\mathbf{W}}_{\infty,\boldsymbol{\eta}}, \mathbf{W}^*) = 0) \to 1$, and Lemma C.7 proves $\mathrm{P}(\mathrm{Ex}(\widehat{\mathbf{W}}_{\infty,\boldsymbol{\eta}}, \mathbf{W}^*) = 0) \to 1$. Combining these results, we have $\mathrm{P}(\mathrm{SHD}(\widehat{\mathbf{W}}_{\infty,\boldsymbol{\eta}}, \mathbf{W}^*) = 0) \to 1$ as $\mathrm{T} \to \infty$.

Proof of Theorem 3. Theorem 3 directly follows from Lemma 2 and Lemma C.12. ■

References

S. F. Ackley, J. Lessler, and M. M. Glymour. Dynamical Modeling as a Tool for Inferring Causation. *American Journal of Epidemiology*, 191(1):1–6, 08 2021. ISSN 0002-9262.

J. Bang-Jensen and G. Z. Gutin. *Digraphs: Theory, Algorithms and Applications*. Springer Science & Business Media, 2008.

D. P. Bertsekas. Necessary and sufficient conditions for a penalty method to be exact. *Mathematical Programming*, 9(1):87–99, 1975.

D. P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods*. Academic Press, 2014.

R. Biswas and E. Shlizerman. Statistical perspective on functional and causal neural connectomics: A comparative study. *Frontiers in Systems Neuroscience*, 16:817962, 2022.

L. Carstensen, A. Sandelin, O. Winther, and N. R. Hansen. Multivariate hawkes process models of the occurrence of regulatory elements. *BMC bioinformatics*, 11:1–19, 2010.

S. Chen, A. Shojaie, E. Shea-Brown, and D. Witten. The multivariate hawkes process in high dimensions: beyond mutual excitation. *arXiv preprint arXiv:1707.04928*, 2017.

S. Chen, Y. Xie, and S. Yang. Causal graph and social network analysis for the spread of covid-19 from self-reported indicator data. In *2021 55th Asilomar Conference on Signals, Systems, and Computers*, pages 1302–1306. IEEE, 2021.

D. M. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3(Nov):507–554, 2002.

D. J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*. Springer, 2003.

E. Dobryakova, O. Boukrina, and G. R. Wylie. Investigation of information flow during a novel working memory task in individuals with traumatic brain injury. *Brain Connectivity*, 5(7):433–441, 2015.

J. Fan and H. Peng. Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics*, 32(3):928–961, 2004.

F. Fu and Q. Zhou. Learning sparse causal gaussian networks with experimental intervention: regularization and coordinate descent. *Journal of the American Statistical Association*, 108(501):288–300, 2013.

S. Fujisawa, A. Amarasingham, M. T. Harrison, and G. Buzsáki. Behavior-dependent short-term assembly dynamics in the medial prefrontal cortex. *Nature Neuroscience*, 11 (7):823–833, 2008.

M. Gao, C. Zhang, and J. Zhou. Learning network-structured dependence from non-stationary multivariate point process data. *IEEE Transactions on Information Theory*, 70(8):5935–5968, 2024. doi: 10.1109/TIT.2024.3396778.

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

D. Heckerman, D. Geiger, and D. M. Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243, 1995.

C.-j. Hsieh, I. Dhillon, P. Ravikumar, and M. Sustik. Sparse inverse covariance matrix estimation using quadratic approximation. *Advances in Neural Information Processing Systems*, 24, 2011.

R. E. Kass, U. T. Eden, E. N. Brown, et al. *Analysis of neural data*, volume 491. Springer, 2014.

P.-L. Loh and P. Bühlmann. High-dimensional learning of linear causal networks via inverse covariance estimation. *Journal of Machine Learning Research*, 15(1):3065–3105, 2014.

D. Luo, H. Xu, Y. Zhen, X. Ning, H. Zha, X. Yang, and W. Zhang. Multi-task multi-dimensional hawkes processes for modeling event sequences. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

P. Nandy, A. Hauser, and M. H. Maathuis. High-dimensional consistency in score-based and hybrid structure learning. *Annals of Statistics*, 46(6A):3151–3183, 2018.

A. V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy. Dynamic bayesian networks for audio-visual speech recognition. *EURASIP Journal on Advances in Signal Processing*, 2002(11):1–15, 2002.

R. Nishii. Asymptotic properties of criteria for selection of variables in multiple regression. *Annals of Statistics*, 12(2):758–765, 1984.

J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, 1999.

R. Pamfil, N. Sriwattanaworachai, S. Desai, P. Pilgerstorfer, K. Georgatzis, P. Beaumont, and B. Aragam. Dynotears: Structure learning from time-series data. In *International Conference on Artificial Intelligence and Statistics*, pages 1595–1605. PMLR, 2020.

J. Pearl. *Causality*. Cambridge University Press, 2009.

P. O. Perry and P. J. Wolfe. Point process modelling for directed interaction networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(5):821–849, 2013.

S. Rajaram, T. Graepel, and R. Herbrich. Poisson-networks: A model for structured poisson processes. In R. G. Cowell and Z. Ghahramani, editors, *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, volume R5 of *Proceedings of Machine Learning Research*, pages 277–284. PMLR, 06–08 Jan 2005.

P. Reynaud-Bouret and S. Schbath. Adaptive estimation for hawkes processes; application to genome analysis. *Annals of Statistics*, 38(5):2781–2822, 2010.

R. W. Robinson. Counting unlabeled acyclic digraphs. In *Combinatorial mathematics V*, pages 28–43. Springer, 1977.

F. Roux, G. Parish, R. Chelvarajah, D. T. Rollings, V. Sawlani, H. Hamer, S. Gollwitzer, G. Kreiselmeyer, M. J. ter Wal, L. Kolibius, B. P. Staresina, M. Wimber, M. W. Self, and S. Hanslmayr. Oscillations support short latency co-firing of neurons during human episodic memory formation. *eLife*, 11:e78109, nov 2022. ISSN 2050-084X.

I. Rubin. Regular point processes and their detection. *IEEE Transactions on Information Theory*, 18(5):547–557, 1972.

M. Scanagatta, C. P. de Campos, G. Corani, and M. Zaffalon. Learning bayesian networks with thousands of variables. *Advances in Neural Information Processing Systems*, 28, 2015.

S. Shalev-Shwartz and A. Tewari. Stochastic methods for l 1 regularized loss minimization. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 929–936, 2009.

C. Shi and L. Li. Testing mediation effects using logic of boolean matrices. *Journal of the American Statistical Association*, 117(540):2014–2027, 2022.

S. Shimizu, P. O. Hoyer, A. Hyvärinen, A. Kerminen, and M. Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

P. L. Spirtes. Directed cyclic graphical representations of feedback models. *arXiv preprint arXiv:1302.4982*, 2013.

X. Tang and L. Li. Multivariate temporal point process regression. *Journal of the American Statistical Association*, pages 1–16, 2021.

W. Truccolo, U. T. Eden, M. R. Fellows, J. P. Donoghue, and E. N. Brown. A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *Journal of Neurophysiology*, 93(2):1074–1089, 2005.

S. Van de Geer and P. Bühlmann. l_0-penalized maximum likelihood for sparse directed acyclic graphs. *Annals of Statistics*, 41(2):536–567, 2013.

S. J. Wright. Accelerated block-coordinate relaxation for regularized optimization. *SIAM Journal on Optimization*, 22(1):159–186, 2012.

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

X. Yang, Y. Guo, and Y. Liu. Bayesian-inference-based recommendation in online social networks. *IEEE Transactions on Parallel and Distributed Systems*, 24(4):642–651, 2012.

G.-X. Yuan, C.-H. Ho, and C.-J. Lin. An improved glmnet for l1-regularized logistic regression. *Journal of Machine Learning Research*, 13(64):1999–2030, 2012.

C. Zhang, Y. Jiang, and Y. Chai. Penalized bregman divergence for large-dimensional regression and classification. *Biometrika*, 97(3):551–566, 2010.

C. Zhang, Y. Chai, X. Guo, M. Gao, D. Devilbiss, and Z. Zhang. Statistical learning of neuronal functional connectivity. *Technometrics*, 58(3):350–359, 2016.

G. Zhang, B. Cai, A. Zhang, Z. Tu, L. Xiao, J. M. Stephen, T. W. Wilson, V. D. Calhoun, and Y.-P. Wang. Detecting abnormal connectivity in schizophrenia via a joint directed acyclic graph estimation model. *NeuroImage*, 260:119451, 2022a. ISSN 1053-8119.

G. Zhang, B. Cai, A. Zhang, Z. Tu, L. Xiao, J. M. Stephen, T. W. Wilson, V. D. Calhoun, and Y.-P. Wang. Detecting abnormal connectivity in schizophrenia via a joint directed acyclic graph estimation model. *NeuroImage*, 260:119451, 2022b. ISSN 1053-8119.

M. Zhao, A. Batista, J. P. Cunningham, C. Chestek, Z. Rivera-Alvidrez, R. Kalmar, S. Ryu, K. Shenoy, and S. Iyengar. An l1-regularized logistic model for detecting short-term neuronal interactions. *Journal of Computational Neuroscience*, 32(3):479–497, 2012.

X. Zheng, B. Aragam, P. K. Ravikumar, and E. P. Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, 31, 2018.

K. Zhong, I. E.-H. Yen, I. S. Dhillon, and P. K. Ravikumar. Proximal quasi-newton for computationally intensive l1-regularized m-estimators. *Advances in Neural Information Processing Systems*, 27, 2014.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.