Causal Discovery in High-Dimensional Point Process Networks with Hidden Nodes

Xu Wang and Ali Shojaie *

Department of Biostatistics, University of Washington, Seattle, WA 98195, USA; wangxu85@uw.edu
* Correspondence: ashojaie@uw.edu

Abstract: Thanks to technological advances leading to near-continuous time observations, emerging multivariate point process data offer new opportunities for causal discovery. However, a key obstacle in achieving this goal is that many relevant processes may not be observed in practice. Naïve estimation approaches that ignore these hidden variables can generate misleading results because of the unadjusted confounding. To plug this gap, we propose a deconfounding procedure to estimate high-dimensional point process networks with only a subset of the nodes being observed. Our method allows flexible connections between the observed and unobserved processes. It also allows the number of unobserved processes to be unknown and potentially larger than the number of observed nodes. Theoretical analyses and numerical studies highlight the advantages of the proposed method in identifying causal interactions among the observed processes.

Keywords: causal discovery; Hawkes process; high-dimensional statistics; hidden confounder

1. Introduction

Learning causal interactions from observational multivariate time series is generally impossible [1,2]. Among many challenges, two of the most important ones are that (i) the data acquisition rate may be much slower than the underlying rate of changes; and (ii) there may be unmeasured confounders [1,3]. First, due to the cost or technological constraints, the data acquisition rate may be much slower than the underlying rate of changes. In such settings, the most commonly used procedure for inferring interactions among time series, Granger causality, may both miss true interactions and identify spurious ones [4–6]. Second, the available data may only include a small fraction of potentially relevant variables, leading to unmeasured confounders. Naïve connectivity estimators that ignore these confounding effects can produce highly biased results [7]. Therefore, reliably distinguishing causal connections between pairs of observed processes from correlations induced by common inputs from unobserved confounders remains a key challenge.

Learning causal interactions between neurons is critical to understanding the neural basis of cognitive functions [8,9]. Many existing neuroscience data, such as data collected using functional magnetic resonance imaging (fMRI), have relatively low temporal resolutions, and are thus of limited utility for causal discovery [10]. This is because many important neuronal processes and interactions happen at finer time scales [11]. New technologies, such as calcium florescent imaging that generate spike train data, make it possible to collect 'live' data at high temporal resolutions [12]. The spike train data, which are multivariate point processes containing spiking times of a collection of neurons, are increasingly used to learn the latent brain connectivity networks and to glean insight into how neurons respond to external stimuli [13]. For example, Bolding and Franks [14] collected spike train data on neurons in mouse olfactory bulb region at 30 kHz under multiple laser intensity levels to study the odor identification mechanism. Despite progress in recording the activity of massive populations of neurons [15], simultaneously monitoring a complete network of spiking neurons at high temporal resolutions is still beyond the reach of the current technology. In fact, most experiments only collect data on a small fraction of neurons,

leaving many unobserved neurons [16–18]. These hidden neurons may potentially interact with the neurons inside the observed set and cannot be ignored. Nevertheless, given its high temporal resolution, spike train data provide an opportunity for causal discovery if we can account for the unmeasured confounders.

When unobserved confounders are a concern, causal effects among the observed variables can be learned using causal structural learning approaches, such as the Fast Causal Inference (FCI) algorithm and its variants [1,19]. However, these algorithms may not identify all causal edges. Specifically, instead of learning the directed acyclic graph (DAG) of causal interactions, FCI learns the maximally ancestral graph (MAG). This graph includes causal interactions between variables that are connected by directed edges, but also bi-directed edges among some other variables, leaving the corresponding causal relationships undetermined. As a result, causality discovery using these algorithms is not always satisfactory. For example, Malinsky and Spirtes [20] recently applied FCI to infer causal network of time series and found a low recall for identifying the true causal relationships. Additionally, despite recent efforts [21], causal structure learning remains computationally intensive, because the space of candidate causal graphs grows super-exponentially with the number of network nodes [22].

The Hawkes process [23] is a popular model for analyzing multivariate point process data. In this model, the probability of future events for each component can depend on the entire history of events of other components. Under straightforward conditions, the multivariate Hawkes process reveals Granger causal interactions among multivariate point processes [24]. Moreover, assuming that all relevant processes are observed in a linear Hawkes process, causal interactions among components can also be inferred [25]. The Hawkes process thus provides a flexible and interpretable framework for investigating the latent network of point processes and is widely used in neuroscience applications [26–32].

In modern applications, it is common for the number of measured components, e.g., the number of neurons, to be large compared to the observed period, e.g., the duration of neuroscience experiments. The high-dimensional nature of data in such applications poses challenges to learning the connectivity network of a multivariate Hawkes process. To address this challenge, Hansen et al. [33] and Chen et al. [34] proposed ℓ_1-regularized estimation procedures and Wang et al. [35] recently developed a high-dimensional inference procedure to characterize the uncertainty of these regularized estimators. However, due to the confounding from unobserved neurons in practice, existing estimation and inference procedures assuming complete observation from all components, may not provide reliable estimates.

Accounting for unobserved confounders in high-dimensional regression has been the subject of recent research. Two such examples are HIVE [36] and trim regression [37], which facilitate causal discovery using high-dimensional regression with unobserved confounders. However, these methods are designed for linear regression with independent observations and do not apply to the long-history temporal dependency setting of Hawkes processes. Moreover, they rely on specific assumptions on observed and unobsvered causal effects, which are not clear to hold in neuronal network settings.

In this paper, we consider learning causal interactions among high-dimensional point processes with (potentially many) hidden confounders. Considering the generalization of the above two approaches to the setting of Hawkes processes, we show that the assumption required by trim regression is more likely to hold in a stable point process network, especially when the confounders affect many observed nodes. Motivated by this finding, we propose a generalization of the trim regression, termed *hp-trim*, for causal discovery from high-dimensional point processes in the presence of (potentially many) hidden confounders. We establish a non-asymptotic convergence rate in estimating the network edges using this procedure. Unlike the previous result for independent data [37], our result considers both the temporal dependence of the Hawkes processes as well as the network sparsity. Using simulated and real data, we also show that *hp-trim* has superior

finite-sample performance compared to the corresponding generalization of HIVE for point processes and/or the naïve approach that ignores the unobserved confounders.

2. The Hawkes Processes with Unobserved Components

2.1. The Hawkes Process

Let $\{t_k\}_{k\in\mathbb{Z}}$ be a sequence of real-valued random variables, taking values in $[0, T]$, with $t_{k+1} > t_k$ and $t_1 \geq 0$ almost surely. Here, time $t = 0$ is a reference point in time, e.g., the start of an experiment, and T is the duration of the experiment. A simple point process N on $\mathbb{R}$ is defined as a family $\{N(A)\}_{A\in\mathcal{B}(\mathbb{R})}$, where $\mathcal{B}(\mathbb{R})$ denotes the Borel σ-field of the real line and $N(A) = \sum_k \mathbf{1}_{\{t_k\in A\}}$. The process N is essentially a simple counting process with isolated jumps of unit height that occur at $\{t_k\}_{k\in\mathbb{Z}}$. We write $N([t, t + dt))$ as $dN(t)$, where dt denotes an arbitrarily small increment of t.

Let $\mathbf{N}$ be a p-variate counting process $\mathbf{N} \equiv \{N_i\}_{i\in\{1,\dots,p\}}$, where, as above, N_i satisfies $N_i(A) = \sum_k \mathbf{1}_{\{t_{ik}\in A\}}$ for $A \in \mathcal{B}(\mathbb{R})$ with $\{t_{i1}, t_{i2}, \dots\}$ denoting the event times of N_i. Let $\mathcal{H}_t$ be the history of $\mathbf{N}$ prior to time t. The intensity process $\{\lambda_1(t), \dots, \lambda_p(t)\}$ is a p-variate $\mathcal{H}_t$-predictable process, defined as

$$\lambda_i(t)dt = \mathbb{P}(dN_i(t) = 1 \mid \mathcal{H}_t). \tag{1}$$

Hawkes [23] proposed a class of point process models in which past events can affect the probability of future events. The process $\mathbf{N}$ is a *linear Hawkes process* if the intensity function for each unit $i \in \{1, \dots, p\}$ takes the form

$$\lambda_i(t) = \mu_i + \sum_{j=1}^{p} (\omega_{ij} * dN_j)(t), \tag{2}$$

where

$$(\omega_{ij} * dN_j)(t) = \int_0^{t^-} \omega_{ij}(t - s)dN_j(s) = \sum_{k:t_{jk}<t} \omega_{ij}(t - t_{jk}). \tag{3}$$

Here, μ_i is the background intensity of unit i and $\omega_{ij}(\cdot) : \mathbb{R}^+ \to \mathbb{R}$ is the *transfer function*. In particular, $\omega_{ij}(t - t_{jk})$ represents the influence from the kth event of unit j on the intensity of unit i at time t.

Motivated by neuroscience applications [38,39], we consider a parametric transfer function $\omega_{ij}(\cdot)$ of the form

$$\omega_{ij}(t) = \beta_{ij}\kappa_j(t) \tag{4}$$

with a *transition kernel* $\kappa_j(\cdot) : \mathbb{R}^+ \to \mathbb{R}$ that captures the decay of the dependence on past events. This leads to $(\omega_{ij} * dN_j)(t) = \beta_{ij}x_j(t)$, where the *integrated stochastic process*

$$x_j(t) = \int_0^{t^-} \kappa_j(t - s)dN_j(s) \tag{5}$$

summarizes the entire history of unit j of the multivariate Hawkes processes. A commonly used example is the exponential transition kernel, $\kappa_j(t) = e^{-t}$ [40].

Assuming that the model holds and all relevant processes are observed, it follows from [40] that the *connectivity coefficient* β_{ij} represents the strength of the *causal* dependence of unit i's intensity on unit j's past events. A positive β_{ij} implies that past events of unit j *excite* future events of unit i and is often considered in the literature (see, e.g., [40,41]). However, we might also wish to allow for negative β_{ij} values to represent *inhibitory* effects [34,42], which are expected in neuroscience applications [43].

Denoting $\mathbf{x}(t) = (x_1(t), \dots, x_p(t))^\top \in \mathbb{R}^p$ and $\boldsymbol{\beta}_i = (\beta_{i1}, \dots, \beta_{ip})^\top \in \mathbb{R}^p$, we can write

$$\lambda_i(t) = \mu_i + \mathbf{x}^\top(t)\boldsymbol{\beta}_i. \tag{6}$$

Furthermore, let $Y_i(t) = dN_i(t)/dt$ and $\epsilon_i(t) = Y_i(t) - \lambda_i(t)$. Then the linear Hawkes process can be written compactly as

$$Y_i(t) = \mu_i + \boldsymbol{x}^\top(t)\boldsymbol{\beta}_i + \epsilon_i(t). \tag{7}$$

2.2. The Confounded Hawkes Process

Because of technology constraints, neuroscience experiments usually collect data from only a small portion of neurons. As a result, many other neurons that potentially interact with the observed neurons will be unobserved. Consider a network of $p + q$ counting processes, where we only observe the first p components. The number of unobserved neurons, q, is usually unknown and likely much greater than p. Extending (7) to include the unobserved components, we obtain the *confounded Hawkes model*,

$$Y_i(t) = \mu_i + \boldsymbol{x}^\top(t)\boldsymbol{\beta}_i + \boldsymbol{z}^\top(t)\boldsymbol{\delta}_i + \epsilon_i(t), \tag{8}$$

in which $\boldsymbol{z}(t) = (x_{p+1}(t), \ldots, x_{p+q}(t))^\top \in \mathbb{R}^q$ denotes the integrated processes of the hidden components, and $\boldsymbol{\delta}_i \in \mathbb{R}^q$ denotes the connectivity coefficients from the unobserved components to unit i.

Unless the observed and unobserved processes are independent, the naïve estimator that ignores the unobserved components will produce misleading conclusion about the causal relationship among the observed components. This is illustrated in the simple linear vector autoregressive process of Figure 1. This example includes three continuous random variables generated according to the following set of equations

$$Y_1(t) = Y_1(t-1) + Y_2(t-1) + \epsilon_1(t-1)$$
$$Y_2(t) = Y_3(t-1) + \epsilon_2(t-1)$$
$$Y_3(t) = Y_3(t-1) + \epsilon_2(t-1),$$

where ϵ_i are mean zero innovation or error terms. The Granger causal network corresponding to the above process is shown in Figure 1A. Figure 1B shows that if Y_3 is not observed, the conditional means of the observed variables Y_1 and Y_2, namely,

$$\mathbb{E}\{Y_1(t) \mid Y_1(t-1), Y_2(t-1)\} = Y_1(t-1) + Y_2(t-1)$$
$$\mathbb{E}\{Y_2(t) \mid Y_1(t-1), Y_2(t-1)\} = Y_2(t-1),$$

leads to incorrect Granger causal conclusions—in this case, a spurious autoregressive effect from the past values of Y_2. The same phenomenon occurs in Hawkes processes with unobserved components.

Throughout this paper, we assume that the confounded linear Hawkes model in (8) is *stationary*, meaning that for all units $i = 1, \ldots, p$, the spontaneous rates μ_i and strengths of transition $(\boldsymbol{\beta}_i, \boldsymbol{\delta}_i)$ are constant over the time range $[0, T]$ [44,45].
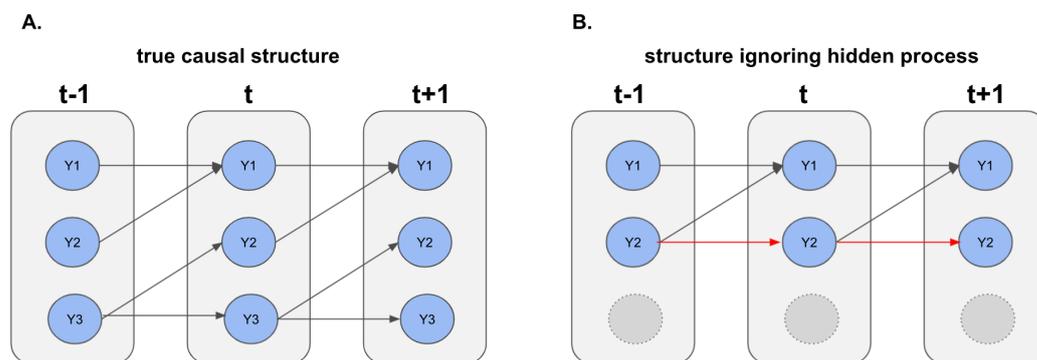
Figure 1. Illustration of the effect of hidden confounders on inferred causal interactions among the observed variables. (**A**) The true causal diagram for the complete processes. (**B**) The causal structure of the observed process when the hidden component, Y_3, is ignored, including a spurious autoregressive effect of Y_2 on its future values.

3. Estimating Causal Effects in Confounded Hawkes Processes

3.1. Extending Trim Regression to Hawkes Processes

Let $\boldsymbol{b}_i \in \mathbb{R}^p$ be the projection coefficient of $\boldsymbol{z}^\top(t)\boldsymbol{\delta}_i$ onto $\boldsymbol{x}(t)$ such that

$$\mathrm{Cov}\left(\boldsymbol{x}(t), \boldsymbol{z}^\top(t)\boldsymbol{\delta}_i - \boldsymbol{x}^\top(t)\boldsymbol{b}_i\right) = 0. \tag{9}$$

We can write the confounded linear Hawkes model in (8) in the form of the *perturbed linear model* [37]:

$$Y_i(t) = \mu_i + \boldsymbol{x}^\top(t)(\boldsymbol{\beta}_i + \boldsymbol{b}_i) + \nu_i(t), \tag{10}$$

where $\nu_i(t) = \left(\boldsymbol{z}^\top(t)\boldsymbol{\delta}_i - \boldsymbol{x}^\top(t)\boldsymbol{b}_i\right) + \epsilon_i(t)$. By the construction of $\boldsymbol{b}_i$, $\nu(t)$ is uncorrelated with the observed processes $\boldsymbol{x}(t)$ and $\boldsymbol{b}_i$ represents the bias, or the perturbation, due to the confounding from $\boldsymbol{z}^\top(t)\boldsymbol{\delta}_i$. In general, $\boldsymbol{b}_i \neq 0$ unless $\mathrm{Cov}(\boldsymbol{x}(t), \boldsymbol{z}(t)) = 0$.

The perturbed model in (10) is generally unidentifiable because we can only estimate $\boldsymbol{\beta}_i + \boldsymbol{b}_i$ from the observed data, e.g., by regressing $Y_i(t)$ on $\boldsymbol{x}(t)$. The *trim regression* [37] is a two-step deconfounding procedure to estimate $\boldsymbol{\beta}_i$ for independent and Gaussian-distributed data. The method first applies a simple spectral transformation, called trim transformation (described below), to the observed data. It then estimates $\boldsymbol{\beta}_i$, using penalized regression. When $\boldsymbol{b}_i$ is sufficiently small, the method consistently estimates $\boldsymbol{\beta}_i$. Although this condition is generally not valid for Gaussian-distributed data, previous work on Hawkes processes [34] implies that the confounding magnitude cannot be large when the underlying network is stable, particularly when the confounders affect many observed components (see the discussion following Corollary 1 in Section 4). This allows us to generalize the trim regression to learn the network of multivariate Hawkes processes.

Assume, without loss of generality, that the first p components are observed at times indexed from 1 to T. Let $X \in \mathbb{R}^{T \times p}$ be the design matrix of the observed integrated process and $Y_i = (Y_i(1), \ldots, Y_i(T))^\top \in \mathbb{R}^T$ be the vector of observed outcomes. Further, let $X = UDV^\top$ be the singular value decomposition on X, where $U \in \mathbb{R}^{T \times r}$, $D \in \mathbb{R}^{r \times r}$ and $V \in \mathbb{R}^{p \times r}$; here, $r = \min(T, p)$ is the rank of X. Denoting the non-zero diagonal entries of D by $d_1, \ldots, d_r$, the *spectral transformation* $F : \mathbb{R}^{T \times p} \to \mathbb{R}^{T \times p}$ is given by

$$F = U \begin{pmatrix} \tilde{d}_1/d_1 & 0 & \ldots & 0 \\ 0 & \tilde{d}_2/d_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \tilde{d}_r/d_r \end{pmatrix} U^\top. \tag{11}$$

Denoting by $\widetilde{D}$ a diagonal matrix with entries $\tilde{d}_1, \ldots, \tilde{d}_r$, the first step of *hp-trim* involves applying the spectral transformation to the observed data to obtain

$$\widetilde{X} = FX = U\widetilde{D}V^\top, \tag{12}$$

$$\widetilde{Y} = FY. \tag{13}$$

The spectral transformation is designed to reduce the magnitude of confounding. In particular, when $\boldsymbol{b}_i$ aligns with the top eigen-vectors of X, for an appropriate F, e.g., $\tilde{d}_k = \min(\tau, d_k)$ as used in previous work [37], the magnitude of $\widetilde{X}\boldsymbol{b}_i$ is small compared with $X\boldsymbol{b}_i$. Here, τ is a threshold parameter and the trim transformation is a special case of the spectral transformation when $\tau = \text{median}(d_1, \ldots, d_r)$. See Ćevid et al. [37] for additional details.

In the second step, we then estimate the network connectivities using the transformed data by solving the following optimization problem

$$\arg \min_{\substack{\mu_i \in \mathbb{R}, \boldsymbol{\beta}_i \in \mathbb{R}^p \\ 1 \leq i \leq p}} \sum_{i=1}^{p} \left\{ \frac{1}{T} \left\| \widetilde{Y}_i - \mu_i - \widetilde{X}\boldsymbol{\beta}_i \right\|_2^2 + \lambda \|\boldsymbol{\beta}_i\|_1 \right\}, \tag{14}$$

which is an instance of lasso regression [46] and can be solved separately for each $i \in \{1, \ldots, p\}$.

3.2. An Alternative Approach

HIdden Variable adjustment Estimation (HIVE) [36] is an alternative method for estimating coefficients of a linear model with independent and Gaussian-distributed data in the presence of latent variables. Adapted to the network of multivariate point processes, HIVE first estimates the latent column space of the unobserved connectivity matrix, $\Delta = \begin{pmatrix} \delta_1 & \ldots & \delta_p \end{pmatrix}^\top \in \mathbf{R}^{p \times q}$, with δ_i defined in (8). It then projects the outcome vector, $Y(t) = \begin{pmatrix} Y_1(t), \ldots, Y_p(t) \end{pmatrix}^\top$, onto the space orthogonal to the column space of Δ. Assuming that the column space of the observed connectivity matrix, $\Theta = \begin{pmatrix} \boldsymbol{\beta}_1 & \ldots & \boldsymbol{\beta}_p \end{pmatrix}^\top \in \mathbf{R}^{p \times p}$ is orthogonal to that of Δ, HIVE consistently estimates Θ using the transformed data. While the orthogonality assumption might be satisfied when the hidden processes are external, such as experimental perturbations in genetic studies [47], it might be too stringent in a network setting. However, when the orthogonality assumption fails, HIVE may lead to poor edge selection performance, and potentially worse than the naïve method that ignores the hidden processes. HIVE also requires the number of hidden variables to be known. Although methods in selecting the number of hidden variables have been proposed, the resulting theoretical guarantees would only be asymptotic. An over- or under-estimated number can either miss the true edges or generate false ones. Given these limitations, we outline the extension of HIVE for Hawkes processes in Appendix A and refer the interested reader to Bing et al. [36] for details.

4. Theoretical Properties

In this section we establish the recovery of the network connectivity in the presence of hidden processes. Technical proofs for the results in this section are given in Appendix B.

We start by stating our assumptions. For a square matrix A, let $\Lambda_{\max}(A)$ and $\Lambda_{\min}(A)$ be its maximum and minimum eigenvalues, respectively.

Assumption 1. *Let $\Omega = \{\Omega_{ij}\}_{1 \leq i,j \leq p+q} \in \mathbb{R}^{(p+q) \times (p+q)}$ with entries $\Omega_{ij} = \int_0^\infty |\omega_{ij}(\Delta)| d\Delta$. There exists a constant γ_Ω such that $\Lambda_{\max}(\Omega^T \Omega) \leq \gamma_\Omega^2 < 1$.*

Assumption 1 is necessary for stationarity of a Hawkes process [34]. The constant γ_Ω does not depend on the dimension $p + q$. For any fixed dimension, Brémaud and Massoulié [44] show that given this assumption the intensity process of the form (6) is stable in distribution and, thus, a stationary process exists. Since our connectivity coefficients of interest are

ill-defined without stationarity, this assumption provides the necessary context for our estimation framework.

Assumption 2. *There exists $\lambda_{\min}$ and $\lambda_{\max}$ such that*

$$0 < \lambda_{\min} \leq \lambda_i(t) \leq \lambda_{\max} < \infty, \quad t \in [0, T]$$

for all $i = 1, \ldots, p + q$.

Assumption 2 requires that the intensity rate is strictly bounded, which prevents degenerate processes for all components of the multivariate Hawkes processes. This assumption has been considered in the previous analysis of Hawkes processes [33–35,42,48].

Assumption 3. *The transition kernel $\kappa_j(t)$ is bounded and integrable over $[0, T]$, for $1 \leq j \leq p + q$.*

Assumption 4. *There exists constants $\rho_r \in (0, 1)$ and $0 < \rho_c < \infty$ such that*

$$\max_{1 \leq i \leq p+q} \sum_{j=1}^{p+q} \Omega_{ij} \leq \rho_r \quad \text{and} \quad \max_{1 \leq j \leq p+q} \sum_{i=1}^{p+q} \Omega_{ij} \leq \rho_c.$$

Assumption 3 implies that the integrated process $x_j(t)$ in (5) is bounded. Assumption 4 requires maximum in- and out- intensity flows to be bounded, which provides a sufficient condition for bounding the eigenvalues of the cross-covariance of $x(t)$ [35]. A similar assumption is considered by Basu and Michailidis [49] in the context of VAR models. Together, Assumptions 3 and 4 imply that the model parameters are bounded, which is often required in time-series analysis [50]. Specifically, these assumptions restrict the influence of the hidden processes from being too large.

Define the set of active indices among the observed components, $S_i = \{j : \beta_{ij} \neq 0, 1 \leq j \leq p\}$, and $s_i = |S_i|$ and $s^* \equiv \max_{1 \leq i \leq p} s_i$. Let $Q = \frac{1}{T} \sum_{t=1}^{T} \begin{pmatrix} 1 \\ x(t) \end{pmatrix} \begin{pmatrix} 1 & x^\top(t) \end{pmatrix}$, and $\gamma_{\min} \equiv \Lambda_{\min}(Q)$ and $\gamma_{\max} \equiv \Lambda_{\max}(Q)$. Our first result provides a fixed sample bound on the error of estimating the connectivity coefficients.

Theorem 1. *Suppose each of the p-variate Hawkes processes with intensity function defined in (8) satisfies Assumptions 1–4. Assume $(\log p) \vee (s^*)^{1/2} = o(T^{1/5})$. Then, taking $\lambda = O(\Lambda_{\max}^2(F) T^{-2/5})$,*

$$\left\| \boldsymbol{\beta}_i - \widehat{\boldsymbol{\beta}}_i \right\|_1 \leq C_1 \Lambda_{\max}^2(F) \frac{s^*}{\gamma_{\min}^2} T^{-2/5} + C_2 \Lambda_{\max}^{-2}(F) T^{-3/5} \left\| \widetilde{X} \boldsymbol{b}_i \right\|_2^2, \quad 1 \leq i \leq p,$$

with probability at least $1 - c_1 p^2 T \exp(-c_2 T^{1/5})$, where $C_1, C_2, c_1, c_2 > 0$ depend on the model parameters and the transition kernel.

Compared to the case with independent and Gaussian-distributed data ([37], Theorem 2), we obtain a slower convergence rate because of the complex dependency of the Hawkes processes. Our rate takes into account the network sparsity among the observed components. It also does not depend on the size of unobserved components, q, which is critical in neuroscience experiments because q is often unknown and potentially very large.

The result in Theorem 1 is different from the corresponding result obtained when all processes are observed ([35], Lemma 10). More specifically, our result includes an extra error term, $\|\widetilde{X} \boldsymbol{b}_i\|_2^2$, which captures the effect of unobserved processes. Next, we show that when $\|\boldsymbol{b}_i\|_2^2$ is sufficiently small, we obtain a similar rate of convergence as the one obtained when all processes are observed.

Corollary 1. *Under the same assumptions in Theorem 1, suppose, in addition,*

$$\|\boldsymbol{b}_i\|_2^2 = O\left(\frac{s^*}{\gamma_{\min}^2 \gamma_{\max}} T^{-4/5} \Lambda_{\max}^2(F)\right),$$

$$\left\|\boldsymbol{\beta}_i - \widehat{\boldsymbol{\beta}}_i\right\|_1 = O\left(\frac{s^*}{\gamma_{\min}^2} \Lambda_{\max}^2(F) T^{-2/5}\right), \quad 1 \le i \le p,$$

with probability at least $1 - c_1 p^2 T \exp(-c_2 T^{1/5})$, *where* $c_1, c_2 > 0$ *depending on the model parameters and the transition kernel.*

The spectral transformation empirically reduces the magnitude of $\frac{1}{T}\|\widetilde{X}\boldsymbol{b}_i\|_2^2$, especially when the confounding vector, $\boldsymbol{b}_i$, stays in the sub-space spanned by top right singular vectors of X; however, this is not guaranteed to hold for arbitrary $\boldsymbol{b}_i$. Corollary 1 specifies a condition on $\boldsymbol{b}_i$ that leads to consistent estimation of $\boldsymbol{\beta}_i$, regardless of the empirical performance of the spectral transformation. While the condition does not always hold for arbitrary stochastic process, it is satisfied for a stable network of high-dimensional multivariate Hawkes processes when the confounding is dense. Specifically, by the construction of $\boldsymbol{b}_i$ in (9), Assumption 4 implies that $\|\boldsymbol{b}_i\|_1 = O(\|\boldsymbol{\delta}_i\|_1) = O(1)$. When the confounding effects are relatively dense—i.e., $\|\boldsymbol{b}_i\|_0 = O(p)$, meaning that there are large number of interactions from unobserved nodes to the observed ones—we obtain $\|\boldsymbol{b}_i\|_2^2 = O(1/p)$. Therefore, the constraint on $\|\boldsymbol{b}_i\|_2^2$ is likely satisfied under a high-dimensional network, when $p \gg T$. The high-dimensional network setting is common in modern neuroscience experiments where the number of neurons is often large compared to the duration of experiments.

Next we introduce an additional assumption to establish the edge selection consistency. To this end, we consider the *thresholded connectivity estimator*,

$$\widetilde{\beta}_{ij} = \widehat{\beta}_{ij}\mathbf{1}\left(\left|\widehat{\beta}_{ij}\right| > \tau\right), \quad 1 \le i, j \le p.$$

Thresholded estimators are used for variable selections in high-dimensional network estimation [51] as they alleviate the need for restrictive irrepresentability assumptions [52].

Assumption 5. *There exists* $\tau > 0$ *such that*

$$\min_{1 \le i,j \le p} \beta_{ij} \ge \beta_{min} > 2\tau.$$

Assumption 5 is called the β-min condition [53] and requires sufficient signal strength for the true edges in order to distinguish them from 0. Let the estimated edge set $\widehat{S} = \left\{(i,j) : \widetilde{\beta}_{ij} \ne 0, 1 \le i, j \le p\right\}$ and the true edge set $S = \left\{(i,j) : \beta_{ij} \ne 0, 1 \le i, j \le p\right\}$. The next result shows that the estimated edge set consistently recovers the true edge set.

Theorem 2. *Under the same conditions in Theorem 1, assume Assumption 5 is satisfied with* $\tau = O\left(\frac{s^*}{\gamma_{\min}^2}\Lambda_{\max}^2(F)T^{-2/5}\right)$. *Then,*

$$\mathbb{P}\left(\widehat{S} = S\right) \ge 1 - c_1 p^2 T \exp\left(-c_2 T^{1/5}\right),$$

where $c_1, c_2 > 0$ *depending on the model parameters and the transition kernel.*

Theorem 2 guarantees the recovery of causal interactions among the observed components. As before, the result is valid irrespsective of the number of unobserved components, which is important in neuroscience applications.

5. Simulation Studies

We compare our proposed method, *hp-trim*, with two alternatives, HIVE and the naïve approach that ignores the unobserved nodes. To this end, we compare the methods in terms of their abilities to identify the correct causal interactions among the observed components.

We consider a point process network consisting of 200 nodes with half of the nodes being observed; that is $p = q = 100$. The observed nodes are connected in blocks of five nodes, and half of the blocks are connected with the unobserved nodes (see Figure 2a). This setting exemplifies neuroscience applications, where the orthogonality assumption of HIVE is violated. As a sensitivity analysis, we also consider a second setting similar to the first, in which we remove the connections of the blocks that are not connected with the unobserved nodes This setting, shown in Figure 3a, satisfies HIVE's orthogonality assumption.
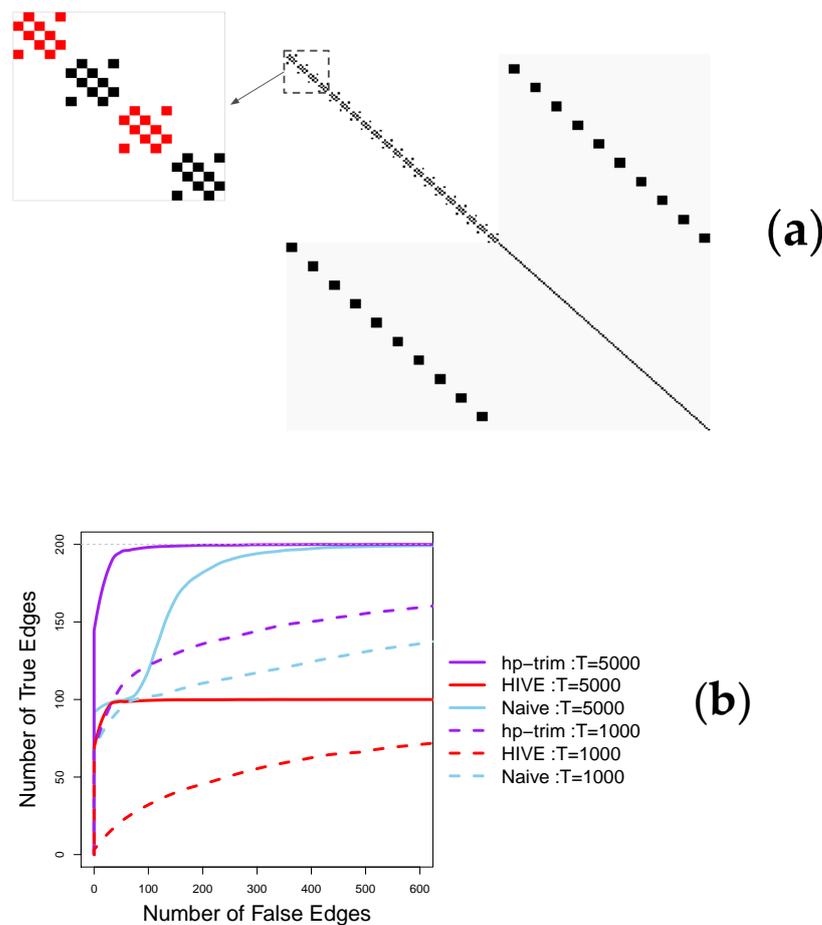

Figure 2. Edge selection performance of the proposed *hp-trim* approach compared with estimators based on HIVE (run with the known (oracle) number of latent features) and the naïve approach. Here, $p = q = 100$. (**a**) Visualization of the connectivity matrix, with unobserved connecitivies colored in gray and entries corresponding to edges shown in black. This setting violates the orthogonality condition of HIVE because of the connections between the observed and the hidden nodes (represented by the non-zero coefficients colored in red). (**b**) Average number of true positive and false positive edges detected using each method over 100 simulation runs.

To generate point process data, we consider $\beta_{ij} = 0.12$ and $\delta_{ij} = 0.10$ in the setting of Figure 2a, and $\beta_{ij} = 0.2$ and $\delta_{ij} = 0.18$ in the setting of Figure 3b. The background intensity, μ_i, is set to 0.05 in both settings. The transfer kernel function is chosen to be $\exp(-t)$. These settings satisfy the assumptions of stationary Hawkes processes. In both settings, we set the length of the time series to $T \in \{1000, 5000\}$.
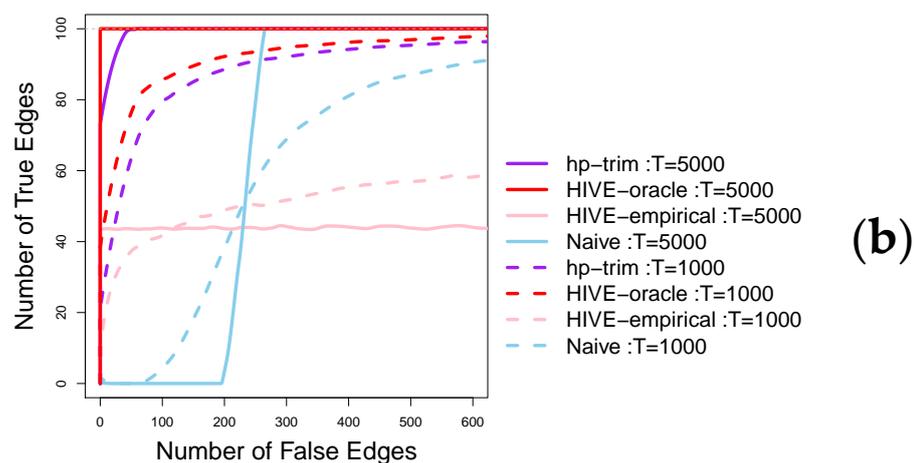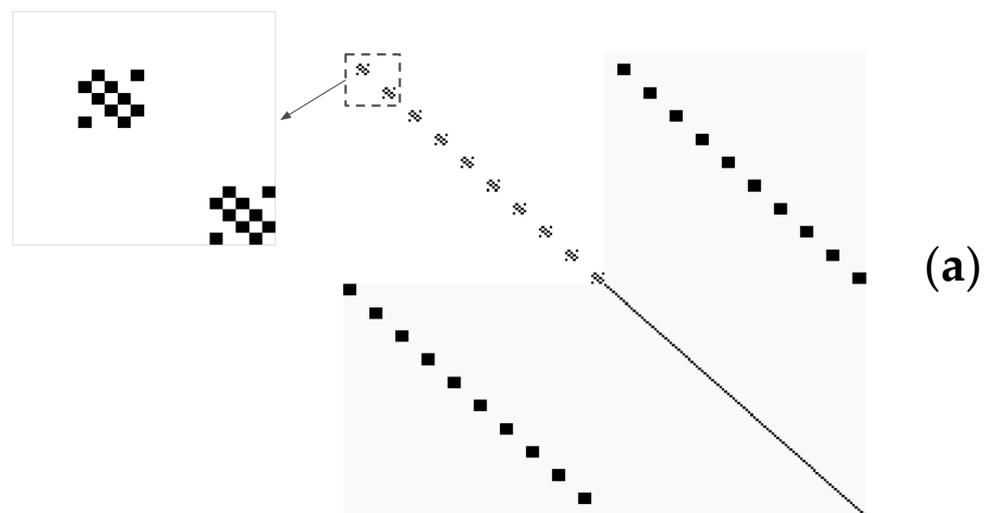
Figure 3. Edge selection performance of the proposed *hp-trim* approach compared with estimators based on HIVE and the naïve approach. Here, $p = q = 100$. (**a**) Visualization of the connectivity matrix, with unobserved connecitivies colored in gray and entries corresponding to edges shown in black. This setting satisfies the orthogonality condition of HIVE, which is run both with and without assuming known number of latent features. These two versions are denoted HIVE-oracle and HIVE-empirical, respectively. In HIVE-empirical the number of latent factors is estimated based on the estimate with highest frequency over the 100 simulation runs (estimated $\hat{q} = 79$). (**b**) Average number of true positive and false positive edges detected using each method over 100 simulation runs.

The results in Figure 2b shown that *hp-trim* offers superior performance for both small and large sample sizes in the first setting. For example, with large sample size, $T = 5000$, *hp-trim* is able to detect almost all 200 true edges at the expense of about 50 falsely detected edges; this is almost twice as large as the number of true edges detected by HIVE and the naïve method, which only detect half of the true edges at the same level of falsely detected edges. The naïve method eventually detects all true edges but at much bigger cost of about 400 falsely detected edges. In this case, HIVE performs poorly and detects at most half of the true edges, no matter the tolerance level of the number of falsely detected edges. The poor performance of HIVE is because its stringent orthogonality condition is violated in this simulation setting. When the orthogonality condition is satisfied (Figure 3a), HIVE shows the best performance. Specifically, with large sample size, $T = 5000$, HIVE detects all true edges almost without identifying any falsely detected edges (the red solid line in

Figure 3b). However, this advantage requires knowledge of the correct number of latent features. When the number of latent features is unknown and estimated from data, HIVE's performance deteriorates, especially with an insufficient sample size. For example, HIVE with empirically estimated number of latent features only detect about 40 true edges (out of a total of 100) at the expense of 100 falsely detected edges (pink lines in Figure 3b). In contrast, *hp-trim*'s performance with both moderate and large sample sizes is close to the oracle version of HIVE (HIVE-oracle). Specifically, with a large sample size, $T = 5000$, *hp-trim* captures all 100 true edges at the expense of 50 falsely detected edges, again than twice as many true edges as HIVE-empirical.

Although our main focus is on the edge selection relevant for causal discovery, in Appendix C we also examine the estimation performance of our algorithm on the connectivity coefficients associated with the observed processes. Not surprisingly, the results indicate that *hp-trim* can also offer advantages in estimating the parameters, especially in settings where it offers improved edge selection.

6. Analysis of Mouse Spike Train Data

We consider the task of learning causal interactions among the observed population of neurons, using the spike train data from Bolding and Franks [14]. In this experiment, spike times are recorded at 30 kHz on a region of the mice olfactory bulb (OB), while a laser pulse is applied directly on the OB cells of the subject mouse. The laser pulse has been applied at increasing intensities from 0 to 50 (mW/mm^2). The laser pulse at each intensity level lasts 10 seconds and is repeated 10 times on the same set of neuron cells of the subject mouse.

The experiment consists of spike train data multiple mice and we consider data from the subject mouse with the most detected neurons (25) under laser (20 mW/mm^2) and no laser conditions. In particular, we use the spike train data from one laser pulse at each intensity level. Since one laser pulse spans 10 seconds and the spike train data is recorded at 30 kHz, there are 300,000 time points per experimental replicate.

The population of observed neurons is a small subset of all the neurons in mouse's brain. Therefore, to discover causal interactions among the $p = 25$ observed neurons, we apply our estimation procedure, *hp-trim*, along with HIVE and naïve approaches, separately for each intensity level, and obtain the estimated connectivity coefficients for the observed neurons. For ease of comparison, the tuning parameters for both methods are chosen to have about 30 estimated edges; moreover, for HIVE, q is estimated following the procedure in Bing et al. [36], which is based on the maximum decrease in eigenvalue of the covariance matrix of the errors, $\widetilde{E}(t)$ in (A1).

Figure 4 shows the estimated connectivity coefficients specific to each laser condition in a graph representation. In this representation, each node represents a neuron, and a directed edge indicates a non-zero estimated connectivity coefficient. We see different network connectivity structures when laser stimulus is applied, which agrees with the observation by neuroscientists that the OB response is sensitive to the external stimuli [14].

Compared to our proposed method, the naïve approach generates a more similar network than HIVE under both laser and no-laser conditions, which is likely an indication that the naïve estimate is incorrect in this application.

As discussed in Section 4, our inference procedure is asymptotically valid. In other words, with large enough sample size, if the other assumptions in Section 4 are satisfied, the estimated edges should represent the true edges. Assessing the validity of the assumptions and selecting the true edges in real data applications is challenging. However, we can assess the sample size requirement and the validity of assumptions by estimating the edges over a subset of neurons as if the other removed neurons are unobserved. If the sample size is sufficient and the other assumptions are satisfied, we should obtain similar connectivities among the observed subset of neurons, even when some neurons are hidden. Figure 5 shows the result of such a stability analysis for the laser condition using *hp-trim*. Comparing the connectivities in this graph with those in Figure 4 indicates that the estimated edges

using the subsets of neurons are all consistent with those estimated using all neurons. Thus, the assumptions are likely satisfied in this application.
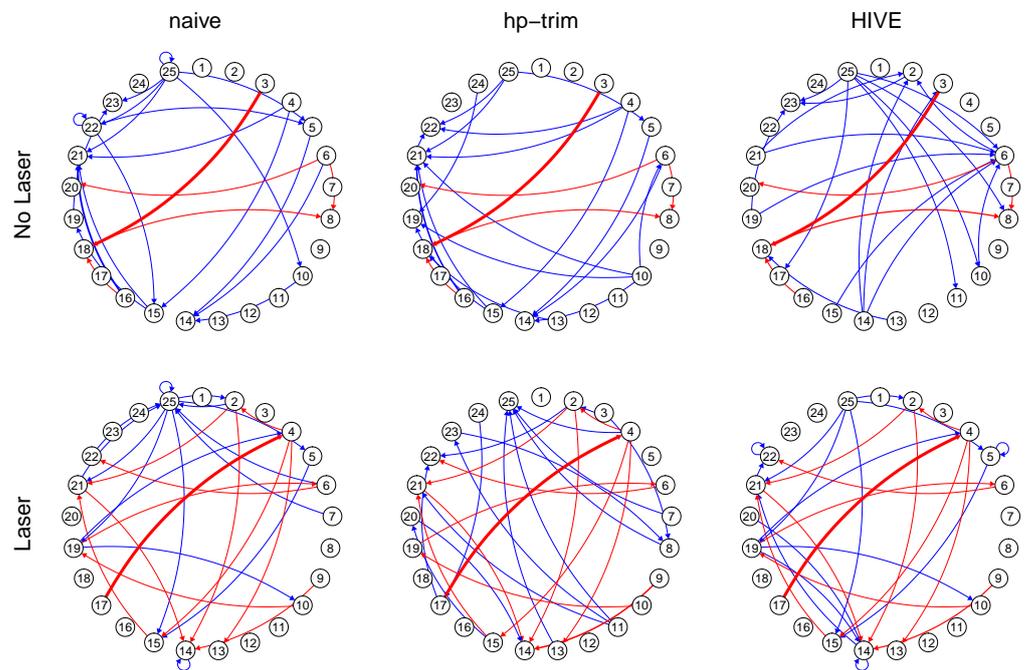

Figure 4. Estimated functional connectivities among neurons using mouse spike train data from laser and no-laser conditions [14]. Common edges estimated by the three methods are in red and the method-specific edges are in blue. Thicker edges indicate estimated connectivity coefficients of larger magnitudes.
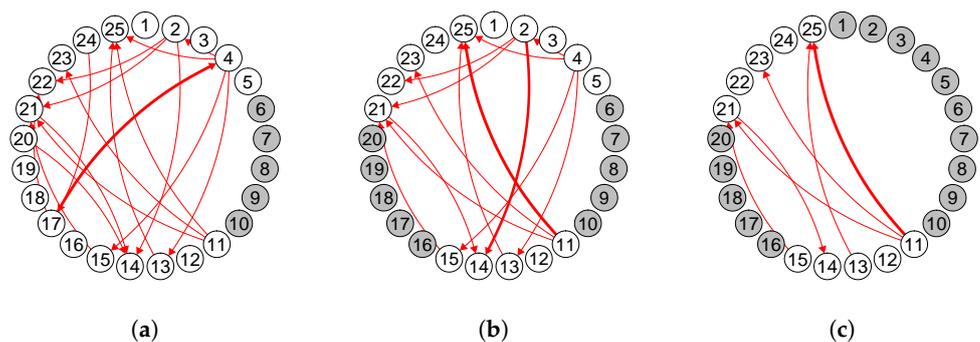

Figure 5. Estimated functional connectivities using *hp-trim* among multiple subset of neurons. Here, data is the same as that used in Figure 4 under the laser condition, except that 5, 10 and 15 neurons (shown in gray) are considered hidden. Thicker edges indicate estimated connectivity coefficients of larger magnitudes. All estimated edges using the subsets of neurons are also found in the estimated network using all neurons (**a**–**c**).

7. Conclusions and Future Work

We proposed a causal-estimation procedure with theoretical guarantees for high-dimensional network of multivariate Hawkes processes in the presence of hidden confounders. Our method extends the trim regression [37] to the setting of point process data. The choice of trim regression as the starting point was motivated by the fact that its assumptions are less stringent than conditions required for the alternative HIVE procedure, especially for a stable point process network with dense confounding effects. Empirically,

our procedure, *hp-trim*, shows superior performance in identifying edges in the causal network compared with HIVE and a naïve method that ignores the unobserved nodes.

Causal discovery from observational time series is a challenging problem and the success of our method is not without limitations. First, the theoretical guarantees for *hp-trim* require the magnitude of the hidden confounding to be bounded. As we discussed in the paper, this condition is likely met for a stable network of high-dimensional multivariate Hawkes processes when the confounding is dense. Nonetheless a careful examination of this condition is required when applying the method in other settings. When certain structure exists between the observed and hidden network connectivities, more structure-specific methods, such as HIVE, may be able to better utilize the structural property of the network for improved performance in identifying the causal effects. Second, our estimates assume a linear Hawkes process with a particular parametric form of the transition function. We also assume the underlying Hawkes process is stationary, where certain structural requirements of the process (specified as assumptions in Section 4) must be satisfied. The proposed method is guaranteed to identify causal effects only if these modeling assumptions are valid. When the modeling assumptions are violated, the estimated effects may not be causal. In other words, the method is primarily designed to generate causal hypotheses— or facilitate *causal discovery*—and the results should be interpreted with caution. Extending the proposed approach to model the transition function nonparametrically, learning its form adaptively from data and capturing time-varying processes would be important future research directions. Finally, given that non-linear link functions are often used when analyzing spike train data [54,55], it would also be of interest to develop causal-estimation procedure for non-linear Hawkes processes.

Appendix A. Additional Details on HIVE

We introduce additional notations before illustrating the method.

Let $Y(t) = \left(Y_1(t), \ldots, Y_p(t)\right)^\top$, $X(t) = \left(x_1(t), \ldots, x_p(t)\right)^\top$, $Z(t) = \left(z_1(t), \ldots, z_q(t)\right)^\top$ and $E(t) = \left(\epsilon_1(t), \ldots, \epsilon_p(t)\right)^\top$. Then, we rewrite (8) simultaneously for all components:

$$Y(t) = \boldsymbol{\mu} + \Theta X(t) + \Delta Z(t) + E(t), \tag{A1}$$

where $\Theta = \begin{pmatrix} \boldsymbol{\beta}_1^\top \\ \ldots \\ \boldsymbol{\beta}_p^\top \end{pmatrix} \in \mathbf{R}^{p \times p}$ and $\Delta = \begin{pmatrix} \boldsymbol{\delta}_1^\top \\ \ldots \\ \boldsymbol{\delta}_p^\top \end{pmatrix} \in \mathbf{R}^{p \times q}$ are connectivity matrix between the

observed and unobserved components, respectively. $\boldsymbol{\mu} = \left(\mu_1, \ldots, \mu_p\right)^\top \in \mathbf{R}^p$ is the vector of spontaneous rate.

To illustrate the confounding induced by the hidden process, we project $Z(t)$ onto the space spanned by $X(t)$ as

$$Z(t) = \boldsymbol{\nu} + A X(t) + W(t), \tag{A2}$$

where A is the projection matrix, representing the cross-sectional correlation between Z and X. Then, (A1) becomes

$$Y(t) = \widetilde{\boldsymbol{\mu}} + \widetilde{\Theta}X(t) + \widetilde{E}(t), \tag{A3}$$

where

$$\widetilde{\boldsymbol{\mu}} = \boldsymbol{\mu} + \Delta\boldsymbol{\nu},$$
$$\widetilde{\Theta} = \Theta + \Delta A,$$
$$\widetilde{E}(t) = E(t) + \Delta W(t).$$

From the above, it is easy to see that the correlations between the observed and unobserved processes determine the strength the confounding. Specifically, unless $A = 0$—i.e., when the observed and unobserved processes are independent, directly regressing $Y(t)$ on $X(t)$ produces biased estimates on Θ. Under the condition that $\Theta \perp \Delta$—i.e., the column space of Θ is orthogonal to the column space of Δ, HIVE gets around this issue by finding a projection matrix, $P_{\Delta^\perp}$, that projects Δ onto its orthogonal space—i.e., $P_{\Delta^\perp}\Delta = 0$. Moreover, because of the orthogonality assumption, $P_{\Delta^\perp}\Theta = \Theta$. Therefore, when multiplying both sides in (A1) by $P_{\Delta^\perp}$, the unobserved term disappears. Specifically, letting $\widetilde{Y}(t) = P_{\Delta^\perp}Y(t)$, (A1) becomes

$$\widetilde{Y}(t) = P_{\Delta^\perp}\boldsymbol{\mu} + \Theta X(t) + P_{\Delta^\perp}E(t). \tag{A4}$$

Consequently, regressing $\widetilde{Y}(t)$ on $X(t)$ produces unbiased estimates on Θ (using penalized regression with ℓ_1-penalty on Θ under the high-dimensional setting when p is allowed to grow with the sample size T). In order to obtain $P_{\Delta^\perp}$, HIVE first calculates $\widetilde{E}(t)$ in (A3) and then implement *heteroPCA* algorithm [56] to estimate the latent column space of Δ thus to obtain P_Δ. Then, the method obtains the corresponding orthogonal project as $P_{\Delta^\perp} = I - P_\Delta$. We refer the interested readers to Bing et al. [36] for details about the method.

Appendix B. Proof of Main Results

Since our focus is on the estimation error for $\boldsymbol{\beta}_i$, we consider the perturbation model in (10) in the following.

Let $\boldsymbol{\theta}_i = \begin{pmatrix} \mu_i & \boldsymbol{\beta}_i \end{pmatrix}^\top$ be the true model parameter and $\widehat{\boldsymbol{\theta}}_i = \begin{pmatrix} \widehat{\mu}_i & \widehat{\boldsymbol{\beta}}_i \end{pmatrix}^\top$ be the optimizer for (14). Recall that the set of active indices, $S_i = \{j : \beta_{ij} \neq 0, 1 \leq j \leq p\}$, and $s_i = |S_i|$ and $s^* \equiv \max_{1 \leq i \leq p} s_i$. Because optimization problem (14) can be solved separately for each component process, in the follows we focus on the estimation consistency for one component process. For ease of notation, we drop the subscript i; that is, we use $x(t)$ for $x_i(t)$, $\boldsymbol{\theta}$ for $\boldsymbol{\theta}_i$, $dN(t)$ for $dN_i(t)$, $\lambda(t)$ for $\lambda_i(t)$, $\boldsymbol{b}$ for $\boldsymbol{b}_i$, S for S_i and $\widetilde{S}$ for $\widetilde{S}_i$.

Next, we state two lemmas that will be used in the proof of main results.

Lemma A1 (van de Geer [57]). *Suppose there exists $\lambda_{\max}$ such that $\lambda(t) \leq \lambda_{\max}$ where $\lambda(t)$ is the intensity function of Hawkes process defined in (2). Let $H(t)$ be a bounded function that is $\mathcal{H}_t$-predictable. Then, for any $\epsilon > 0$,*

$$\frac{1}{T}\int_0^T H(t)\left\{\lambda(t)dt - dN(t)\right\} \leq 4\left\{\frac{\lambda_{\max}}{2T}\int_0^T H^2(t)dt\right\}^{1/2}\epsilon^{1/2},$$

with probability at least $1 - C\exp(-\epsilon T)$, for some constant C.

Lemma A2 (Wang et al. [35]). *Suppose the Hawkes process defined in* (2) *satisfies Assumptions* 1–4. *Let* $Q = \frac{1}{T} \int_0^T \begin{pmatrix} 1 \\ x(t) \end{pmatrix} (1 \quad x^\top(t)) dt$, *where* $x(t)$ *is defined in* (5). *Then, there exists* $\gamma_{\max} \geq \gamma_{\min} > 0$ *such that*

$$\gamma_{\max} \geq \Lambda_{\max}(Q) \geq \Lambda_{\min}(Q) \geq \gamma_{\min} > 0,$$

with probability at least $1 - c_1 p^2 T \exp(-c_2 T^{1/5})$, *where constants* c_1, c_2 *depending on the model parameters and the transition kernel.*

Proof of Theorem 1**.** While the skeleton of the proof follows from (Ćevid et al. [37], Theorem 2), the following two conditions are needed because of the Hawkes process data's unique dependency structure. □

Condition 1. *There exist constants* $\gamma_{\min}, c, C > 0$ *such that*

$$\mathbb{P}\left(\min_{\Delta \in \mathcal{C}(L,S)} \frac{1}{T} \left\| \widetilde{X}\Delta \right\|_2^2 \geq \gamma_{\min} \|\Delta\|_2^2 \right) \geq 1 - cp^2 T \exp(-CT^{1/5}),$$

where $\mathcal{C}(L,S) = \{ \boldsymbol{\alpha} : \|\boldsymbol{\alpha}_{S^c}\|_1 \leq L\|\boldsymbol{\alpha}_S\|_1 \}$.

Condition 1 is referred as the restrict strong convexity (RSC) [58]. Lemma A2 by Wang et al. [35] has shown Condition 1 holds when $\widetilde{X} = X$ under Assumptions 1–4. Since the min eigenvalue of $\widetilde{X}$ stays the same with our choice of F, Condition 1 holds for $\widetilde{X} = FX$.

Condition 2. *There exist* $c, C > 0$ *such that*

$$\mathbb{P}\left(\frac{1}{T} \left\| \widetilde{X}\nu \right\|_\infty \leq C\Lambda_{\max}^2(F) T^{-2/5} \right) \geq 1 - cp \exp(-T^{1/5}),$$

where ν *is defined in* (10).

Condition 2 holds as a result of Lemma A1 by van de Geer [57].
Under the two conditions, we achieve the conclusion as follows.
Because $\widehat{\theta}$ is the optimizer for (14),

$$\frac{1}{T}\|\widetilde{Y} - \widetilde{X}\widehat{\theta}\|_2^2 + \lambda\|\widehat{\beta}\|_1 \leq \frac{1}{T}\|\widetilde{Y} - \widetilde{X}\theta\|_2^2 + \lambda\|\beta\|_1$$

$$\frac{1}{T}\left\| \widetilde{X}\left(\widehat{\theta} - \theta - b\right) \right\|_2^2 + \lambda\|\widehat{\beta}\|_1 \leq \frac{2}{T}\int_{t=0}^T \nu(t)\widetilde{X}(t)\left(\widehat{\theta} - \theta\right) + \frac{1}{T}\|\widetilde{X}b\|_2^2 + \lambda\|\beta\|_1$$

Under Condition 2,

$$\frac{2}{T}\int_{t=0}^T \nu(t)\widetilde{X}(t)\left(\widehat{\theta} - \theta\right) \leq \frac{2}{T}\left\| \int_{t=0}^T \nu(t)\widetilde{X}(t) \right\|_\infty \left\| \widehat{\theta} - \theta \right\|_1 \leq \psi\left\| \widehat{\theta} - \theta \right\|_1,$$

with probability at least $1 - c_1 p \exp(-T^{1/5})$, where $\psi = C_1 \Lambda_{\max}^2(F) T^{-2/5}$.
Letting $\boldsymbol{\theta}_S = \begin{pmatrix} u & \boldsymbol{\beta}_S \end{pmatrix}^\top$ and $\boldsymbol{\theta}_{S^c} = \begin{pmatrix} u & \boldsymbol{\beta}_{S^c} \end{pmatrix}^\top$,

$$\frac{1}{T}\left\| \widetilde{X}\left(\widehat{\theta} - \theta - b\right) \right\|_2^2 + \lambda\|\widehat{\beta}\|_1 \leq \psi\left\| \widehat{\theta} - \theta \right\|_1 + \frac{1}{T}\|\widetilde{X}b\|_2^2 + \lambda\|\beta\|_1$$

$$\frac{1}{T}\left\| \widetilde{X}\left(\widehat{\theta} - \theta - b\right) \right\|_2^2 + (\lambda - \psi)\|\widehat{\theta}_{S^c} - \theta_{S^c}\|_1 \leq (\lambda + \psi)\left\| \widehat{\theta}_S - \theta_S \right\|_1 + \frac{1}{T}\|\widetilde{X}b\|_2^2$$

Next, we discuss in two conditions: i) $\frac{1}{T}\|\widetilde{X}b\|_2^2 \leq \lambda\left\| \widehat{\theta}_S - \theta_S \right\|_1$ and ii) $\frac{1}{T}\|\widetilde{X}b\|_2^2 \geq \lambda\left\| \widehat{\theta}_S - \theta_S \right\|_1$.

First, when $\frac{1}{T}\|\widetilde{X}b\|_2^2 \le \lambda\left\|\widehat{\theta}_S - \theta_S\right\|_1$,

$$\frac{1}{T}\left\|\widetilde{X}\left(\widehat{\theta} - \theta - b\right)\right\|_2^2 + (\lambda - \psi)\|\widehat{\theta}_{S^c} - \theta_{S^c}\|_1 \le (2\lambda + \psi)\left\|\widehat{\theta}_S - \theta_S\right\|_1.$$

The above implies

$$(\lambda - \psi)\|\widehat{\theta}_{S^c} - \theta_{S^c}\|_1 \le (2\lambda + \psi)\left\|\widehat{\theta}_S - \theta_S\right\|_1,$$

which means $\widehat{\alpha}_{S^c} - \alpha_{S^c} \in \mathcal{C}(L, S) = \{\alpha : \|\alpha_{S^c}\|_1 \le L\|\alpha_S\|_1\}$ for $L = \frac{2\lambda + \psi}{\lambda - \psi}$.

Taking $\lambda = 2\psi$,

$$
\begin{aligned}
&\frac{1}{T}\left\|\widetilde{X}\left(\widehat{\theta} - \theta - b\right)\right\|_2^2 + (\lambda - \psi)\left\|\widehat{\theta} - \theta\right\|_1\\
&\le 3\lambda\sqrt{s^*}\|\widehat{\theta}_S - \theta_S\|_2\\
&\le 3\lambda\sqrt{s^*}\frac{1}{\gamma_{\min}\sqrt{T}}\left\|\widetilde{X}\left(\widehat{\theta} - \theta\right)\right\|_2\\
&\le 3\lambda\sqrt{s^*}\frac{1}{\gamma_{\min}\sqrt{T}}\left\{\left\|\widetilde{X}\left(\widehat{\theta} - \theta - b\right)\right\|_2 + \left\|\widetilde{X}b\right\|_2\right\}\\
&\le 3\lambda\sqrt{s^*}\frac{1}{\gamma_{\min}\sqrt{T}}\left\|\widetilde{X}\left(\widehat{\theta} - \theta - b\right)\right\|_2 + 3\lambda\sqrt{s^*}\frac{1}{\gamma_{\min}\sqrt{T}}\left\|\widetilde{X}b\right\|_2\\
&\le \frac{9}{2}\lambda^2 s^*\frac{1}{\gamma_{\min}^2} + \frac{1}{T}\left\|\widetilde{X}\left(\widehat{\theta} - \theta - b\right)\right\|_2^2 + \frac{1}{T}\left\|\widetilde{X}b\right\|_2^2,
\end{aligned}
$$

where the second inequality is by Condition 1 and the last step is by using $xy \le \frac{1}{4}x^2 + y^2$ twice. Therefore, we get

$$(\lambda - \psi)\left\|\widehat{\theta} - \theta\right\|_1 \le \frac{9}{2}\lambda^2 s^*\frac{1}{\gamma_{\min}^2} + \frac{1}{T}\left\|\widetilde{X}b\right\|_2^2.$$

When $\frac{1}{T}\|\widetilde{X}b\|_2^2 \ge \lambda\left\|\widehat{\theta}_S - \theta_S\right\|_1$,

$$\frac{1}{T}\left\|\widetilde{X}\left(\widehat{\theta} - \theta - b\right)\right\|_2^2 + (\lambda - \psi)\left\|\widehat{\theta} - \theta\right\|_1 \le \frac{3}{T}\left\|\widetilde{X}b\right\|_2^2.$$

Combining the two cases, we always have

$$(\lambda - \psi)\left\|\widehat{\theta} - \theta\right\|_1 \le \frac{9}{2}\lambda^2 s^*\frac{1}{\gamma_{\min}^2} + \frac{3}{T}\left\|\widetilde{X}b\right\|_2^2.$$

Thus, taking $\lambda = 2\psi = O(\Lambda_{\max}^2(F)T^{-2/5})$ and dividing both sides by $\frac{1}{2}\lambda$, we achieve the conclusion that

$$\left\|\widehat{\theta} - \theta\right\|_1 \le C_1\Lambda_{\max}^2(F)\frac{s^*}{\gamma_{\min}^2}T^{-2/5} + C_2 T^{-3/5}\Lambda_{\max}^{-2}(F)\left\|\widetilde{X}b\right\|_2^2.$$

Proof of Corollary 1. Notice that

$$\frac{1}{T}\|\widetilde{X}b\|_2^2 \le \Lambda_{\max}^2(F)\frac{1}{T}\|Xb\|_2^2 \le \Lambda_{\max}^2(F)\gamma_{\max}\|b\|_2^2,$$

with probability at least $1 - c_1 p^2 T \exp(-c_2 T^{1/5})$, where the second inequality is by Lemma A2. Then, Corollary 1 is a direct result from Theorem 1 by plugging in $\|b\|_2^2$. □

Proof of Theorem 2. Recall $S = \{\beta_{ij} : \beta_{ij} \ne 0, 1 \le i, j \le p\}$ and $S_C = \{\beta_{ij} : \beta_{ij} = 0, 1 \le i, j \le p\}$. To establish selection consistency, we need two parts. First, we show

that our estimates on the true zero and non-zero coefficients can be separated with high probability; that is, there exists some constant $\Delta > 0$ such that for $\beta_S \in S$ and $\beta_{S_C} \in S_C$, $|\widehat{\beta}_S - \widehat{\beta}_{S_C}| \geq \Delta$ with high probability. By the β-min condition specified in Assumption 5, we have $\beta_{ij} \in S \geq 2\tau$. Theorem 1 shows that for $1 \leq i, j \leq p$, $|\widehat{\beta}_{ij} - \beta_{ij}| \leq \tau$ with probability at least $1 - c_1 p^2 T \exp(-c_2 T^{1/5})$. Then, for any $\beta_S \in S$ and $\beta_{S_C} \in S_C$,

$$
\begin{aligned}
|\widehat{\beta}_S - \widehat{\beta}_{S_C}| &= |\widehat{\beta}_S - \beta_S - (\widehat{\beta}_{S_C} - \beta_{S_C}) + \beta_S - \beta_{S_C}| \\
&\geq |\beta_S - \beta_{S_C}| - |\widehat{\beta}_S - \beta_S| - |\widehat{\beta}_{S_C} - \beta_{S_C}| \\
&\geq \beta_{min} - 2\tau.
\end{aligned}
$$

This means the estimates on zero and non-zero coefficients can be separated with high probability.

Next, we show there exists a post-selection threshold that allows to correctly identify S and S_C based on the estimates. In fact, the post-selection estimator is

$$
\widetilde{\beta} = \widehat{\beta}\mathbf{1}(|\widehat{\beta}| > \tau).
$$

By Theorem 1, we have $|\widehat{\beta}_{S_C}| \leq \tau$, with probability $1 - c_1 p^2 T \exp(-c_2 T^{1/5})$. Then,

$$
\widetilde{\beta}_{S_C} = \widehat{\beta}_{S_C}\mathbf{1}(\widehat{\beta}_{S_C} > \tau_S) = 0,
$$

which means $\widetilde{\beta}$ selects β_{S_C} into S_C with high probability. In addition, since $|\widehat{\beta}_S - \beta_S| \leq \tau$,

$$
|\widehat{\beta}_S| \geq |\beta_S| - \tau \geq \beta_{min} - \tau > \tau > 0.
$$

Therefore,

$$
\widetilde{\beta}_S = \widehat{\beta}_S\mathbf{1}(|\widehat{\beta}_S| > \tau) = \widehat{\beta}_S \neq 0,
$$

which means $\widetilde{\beta}_S$ selects β_S into S with high probability.

Combining the two sides, the post-selection estimator $\widetilde{\beta}$ identifies S and S_C with high probability. $\square$

Appendix C. Parameter Estimation Performance

In this section we examine estimation performance of our algorithm on the connectivity coefficients associated with the observed processes. To this end, we compare the optimal root-mean squared error (RMSE) of the various methods (*hp-trim*, HIVE and Naïve) over all connectivity coefficients for the observed processes. Here, the optimal RMSE is the minimum RMSE for each estimation method over the range of tuning parameters in each simulation run.

We find that in the case when *hp-trim* performs the best in terms of edge selection (i.e., under the setting by Figure 2a), the method also gives the lowest RMSE (see Unorthogonal ($T = 5000$ and $T = 1000$) in Figure A1). In contrast, when the orthogonality condition is met for HIVE (i.e., under the setting by Figure 3a), HIVE-oracle gives the best RMSE (see Orthogonal ($T = 5000$ and $T = 1000$) in Figure A1). However, HIVE-oracle is not available in practice, and even when the orthogonality assumption is satisfied, the empirical version of HIVE (HIVE-empirical) performs worse than *hp-trim*.
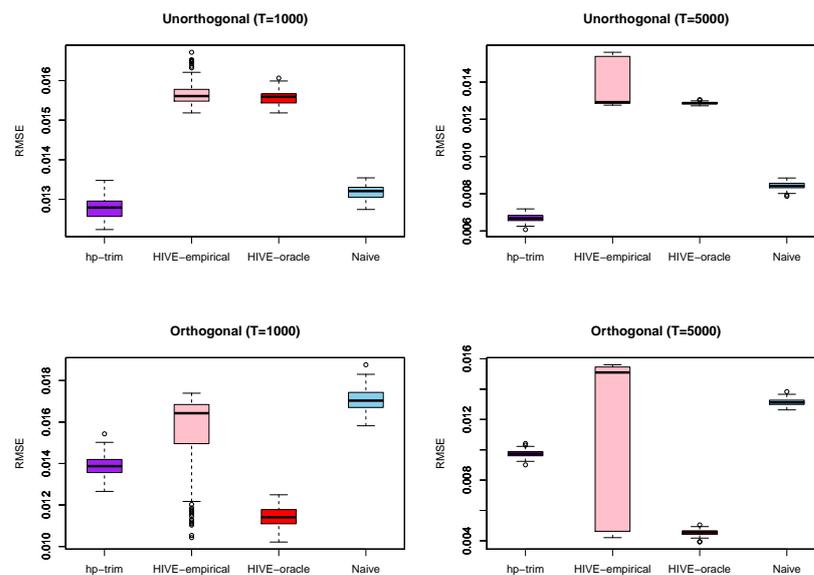
Figure A1. Boxplot of optimal RMSE over all connectivity coefficients for *hp-trim*, HIVE and Naïve. Unorthogonal ($T = 5000$ and $T = 1000$) conditions refer to the setting in Figure 2a in the main text; Orthogonal ($T = 5000$ and $T = 1000$) conditions refer to the setting in Figure 3a in the main text. RMSE over all connectivity coefficients is calculated as $\sqrt{\frac{1}{p^2} \sum_{1 \le i,j, \le p} (\hat{\beta}_{ij}^{(k)} - \beta_{ij})^2}$, where $\hat{\beta}_{ij}^{(k)}$ is the estimate of the true parameter value, β_{ij}, from the kth simulation run ($k = 1, \ldots, 100$) and $p = 100$ observed processes are considered as in the simulation study in the main text.

References

1. Glymour, C.; Zhang, K.; Spirtes, P. Review of causal discovery methods based on graphical models. *Front. Genet.* **2019**, *10*, 524. [CrossRef] [PubMed]
2. Shojaie, A.; Fox, E.B. Granger causality: A review and recent advances. *arXiv* **2021**, arXiv:2105.02675.
3. Reid, A.T.; Headley, D.B.; Mill, R.D.; Sanchez-Romero, R.; Uddin, L.Q.; Marinazzo, D.; Lurie, D.J.; Valdés-Sosa, P.A.; Hanson, S.J.; Biswal, B.B.; et al. Advancing functional connectivity research from association to causation. *Nat. Neurosci.* **2019**, *22*, 1751–1760. [CrossRef]
4. Breitung, J.; Swanson, N.R. Temporal aggregation and spurious instantaneous causality in multiple time series models. *J. Time Ser. Anal.* **2002**, *23*, 651–665. [CrossRef]
5. Silvestrini, A.; Veredas, D. Temporal aggregation of univariate and multivaraite time series models: A survey. *J. Econ. Surv.* **2008**, *22*, 458–497. [CrossRef]
6. Tank, A.; Fox, E.B.; Shojaie, A. Identifiability and estimation of structural vector autoregressive models for subsampled and mixed-frequency time series. *Biometrika* **2019**, *106*, 433–452. [CrossRef]
7. Soudry, D.; Keshri, S.; Stinson, P.; hwan Oh, M.; Iyengar, G.; Paninski, L. A shotgun sampling solution for the common input problem in neural connectivity inference. *arXiv* **2014**, arXiv:1309.3724.
8. Yang, Y.; Qiao, S.; Sani, O.G.; Sedillo, J.I.; Ferrentino, B.; Pesaran, B.; Shanechi, M.M. Modelling and prediction of the dynamic responses of large-scale brain networks during direct electrical stimulation. *Nat. Biomed. Eng.* **2021**, *5*, 324–345. [CrossRef] [PubMed]
9. Bloch, J.; Greaves-Tunnell, A.; Shea-Brown, E.; Harchaoui, Z.; Shojaie, A.; Yazdan-Shahmorad, A. Cortical network structure mediates response to stimulation: An optogenetic study in non-human primates. *bioRxiv* **2021**. [CrossRef]
10. Lin, F.H.; Ahveninen, J.; Raij, T.; Witzel, T.; Chu, Y.H.; Jääskeläinen, I.P.; Tsai, K.W.K.; Kuo, W.J.; Belliveau, J.W. Increasing fMRI sampling rate improves Granger causality estimates. *PLoS ONE* **2014**, *9*, e100319. [CrossRef] [PubMed]
11. Zhou, D.; Zhang, Y.; Xiao, Y.; Cai, D. Analysis of sampling artifacts on the Granger causality analysis for topology extraction of neuronal dynamics. *Front. Comput. Neurosci.* **2014**, *8*, 75. [CrossRef] [PubMed]
12. Prevedel, R.; Yoon, Y.G.; Hoffmann, M.; Pak, N.; Wetzstein, G.; Kato, S.; Schrödel, T.; Raskar, R.; Zimmer, M.; Boyden, E.S.; et al. Simultaneous whole-animal 3D imaging of neuronal activity using light-field microscopy. *Nat. Methods* **2014**, *11*, 727–730. [CrossRef] [PubMed]
13. Okatan, M.; Wilson, M.A.; Brown, E.N. Analyzing functional connectivity using a network likelihood model of ensemble neural spiking activity. *Neural Comput.* **2005**, *17*, 1927–1961. [CrossRef] [PubMed]

14. Bolding, K.A.; Franks, K.M. Recurrent cortical circuits implement concentration-invariant odor coding. *Science* **2018**, *361*, 6407. [CrossRef] [PubMed]

15. Berényi, A.; Somogyvári, Z.; Nagy, A.J.; Roux, L.; Long, J.D.; Fujisawa, S.; Stark, E.; Leonardo, A.; Harris, T.D.; Buzsáki, G. Large-scale, high-density (up to 512 channels) recording of local circuits in behaving animals. *J. Neurophysiol.* **2014**, *111*, 1132–1149. [CrossRef] [PubMed]

16. Trong, P.K.; Rieke, F. Origin of correlated activity between parasol retinal ganglion cells. *Nat. Neurosci.* **2008**, *11*, 1343–1351. [CrossRef]

17. Tchumatchenko, T.; Geisel, T.; Volgushev, M.; Wolf, F. Spike correlations—What can they tell about synchrony? *Front. Neurosci.* **2011**, *5*, 68. [CrossRef]

18. Huang, H. Effects of hidden nodes on network structure inference. *J. Phys. A Math. Theor.* **2015**, *48*, 355002. [CrossRef]

19. Spirtes, P.; Glymour, C.; Scheines, R. *Causation, Prediction, and Search*, 2nd ed.; MIT Press: Cambridge, MA, USA, 2000.

20. Malinsky, D.; Spirtes, P. Causal structure learning from multivariate time series in settings with unmeasured confounding. In Proceedings of 2018 ACM SIGKDD Workshop on Causal Disocvery, London, UK, 20 August 2018; Le, T.D., Zhang, K., Kıcıman, E., Hyvärinen, A., Liu, L., Eds.; PMLR: London, UK, 2018; Volume 92, pp. 23–47.

21. Chen, W.; Drton, M.; Shojaie, A. Causal structural learning via local graphs. *arXiv* **2021**, arXiv:2107.03597.

22. Shojaie, A.; Michailidis, G. Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika* **2010**, *97*, 519–538. [CrossRef]

23. Hawkes, A.G. Spectra of some self-exciting and mutually exciting point processes. *Biometrika* **1971**, *58*, 83–90. [CrossRef]

24. Eichler, M.; Dahlhaus, R.; Dueck, J. Graphical modeling for multivariate Hawkes processes with nonparametric link functions. *J. Time Ser. Anal.* **2017**, *38*, 225–242. [CrossRef]

25. Bacry, E.; Muzy, J. First- and second-order statistics characterization of Hawkes processes and non-parametric estimation. *IEEE Trans. Inf. Theory* **2016**, *62*, 2184–2202. [CrossRef]

26. Brillinger, D.R. Maximum likelihood analysis of spike trains of interacting nerve cells. *Biol. Cybern.* **1988**, *59*, 189–200. [CrossRef] [PubMed]

27. Johnson, D.H. Point process models of single-neuron discharges. *J. Comput. Neurosci.* **1996**, *3*, 275–299. [CrossRef]

28. Krumin, M.; Reutsky, I.; Shoham, S. Correlation-based analysis and generation of multiple spike trains using Hawkes models with an exogenous input. *Front. Comput. Neurosci.* **2010**, *4*, 147. [CrossRef] [PubMed]

29. Pernice, V.; Staude, B.; Cardanobile, S.; Rotter, S. How structure determines correlations in neuronal networks. *PLoS Comput. Biol.* **2011**, *7*, e1002059. [CrossRef] [PubMed]

30. Reynaud-Bouret, P.; Rivoirard, V.; Tuleau-Malot, C. Inference of functional connectivity in neurosciences via Hawkes processes. In Proceedings of the 2013 IEEE Global Conference on Signal and Information Processing, Austin, TX, USA, 3–5 December 2013; pp. 317–320.

31. Truccolo, W. From point process observations to collective neural dynamics: Nonlinear Hawkes process GLMs, low-dimensional dynamics and coarse graining. *J. Physiol.-Paris* **2016**, *110*, 336–347. [CrossRef]

32. Lambert, R.C.; Tuleau-Malot, C.; Bessaih, T.; Rivoirard, V.; Bouret, Y.; Leresche, N.; Reynaud-Bouret, P. Reconstructing the functional connectivity of multiple spike trains using Hawkes models. *J. Neurosci. Methods* **2018**, *297*, 9–21. [CrossRef]

33. Hansen, N.R.; Reynaud-Bouret, P.; Rivoirard, V. Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli* **2015**, *21*, 83–143. [CrossRef]

34. Chen, S.; Shojaie, A.; Shea-Brown, E.; Witten, D. The multivariate Hawkes process in high dimensions: Beyond mutual excitation. *arXiv* **2019**, arXiv:1707.04928.

35. Wang, X.; Kolar, M.; Shojaie, A. Statistical inference for networks of high-dimensional point processes. *arXiv* **2020**, arXiv:2007.07448.

36. Bing, X.; Ning, Y.; Xu, Y. Adaptive estimation of multivariate regression with hidden variables. *arXiv* **2020**, arXiv:2003.13844.

37. Ćevid, D.; Bühlmann, P.; Meinshausen, N. Spectral deconfounding via perturbed sparse linear models. *arXiv* **2020**, arXiv:1811.05352.

38. Linderman, S.; Adams, R. Discovering latent network structure in point process data. In *International Conference on Machine Learning*; PMLR: Beijing, China, 2014; Volume 32.

39. De Abril, I.M.; Yoshimoto, J.; Doya, K. Connectivity inference from neural recording data: Challenges, mathematical bases and research directions. *Neural Netw.* **2018**, *102*, 120–137.

40. Bacry, E.; Mastromatteo, I.; Muzy, J. Hawkes processes in finance. *Mark. Microstruct. Liq.* **2015**, *1*, 1550005. [CrossRef]

41. Etesami, J.; Kiyavash, N.; Zhang, K.; Singhal, K. Learning network of multivariate Hawkes processes: A time series approach. *arXiv* **2016**, arXiv:1603.04319.

42. Costa, M.; Graham, C.; Marsalle, L.; Tran, V.C. Renewal in Hawkes processes with self-excitation and inhibition. *arXiv* **2018**, arXiv:1801.04645.

43. Babington, P. *Neuroscience*, 2nd ed.; Sinauer Associates: Sunderland, MA, USA, 2001.

44. Brémaud, P.; Massoulié, L. Stability of nonlinear Hawkes processes. *Ann. Probab.* **1996**, *24*, 1563–1588. [CrossRef]

45. Daley, D.J.; Vere-Jones, D. *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*; Probability and its Applications; Springer: New York, NY, USA, 2003.

46. Tibshirani, R. Regression shrinkage and selection via the Lasso. *J. R. Stat. So. Ser. B (Methodol.)* **1996**, *58*, 267–288. [CrossRef]

47. Lee, S.; Sun, W.; Wright, F.A.; Zou, F. An improved and explicit surrogate variable analysis procedure by coefficient adjustment. *Biometrika* **2017**, *104*, 303–316. [CrossRef]

48. Cai, B.; Zhang, J.; Guan, Y. Latent network structure learning from high dimensional multivariate point processes. *arXiv* **2020**, arXiv:2004.03569.

49. Basu, S.; Michailidis, G. Regularized estimation in sparse high-dimensional time series models. *Ann. Stat.* **2015**, *43*, 1535–1567. [CrossRef]

50. Safikhani, A.; Shojaie, A. Joint structural break detection and parameter estimation in high-dimensional nonstationary VAR models. *J. Am. Stat. Assoc.* **2020**, 1–14. [CrossRef]

51. Shojaie, A.; Basu, S.; Michailidis, G. Adaptive thresholding for reconstructing regulatory networks from time-course gene expression data. *Stat. Biosci.* **2012**, *4*, 66–83. [CrossRef]

52. Van de Geer, S.; Bühlmann, P.; Zhou, S. The adaptive and the thresholded Lasso for potentially misspecified models (and a lower bound for the Lasso). *Electron. J. Stat.* **2011**, *5*, 688 – 749. [CrossRef]

53. Buhlmann, P. Statistical significance in high-dimensional linear models. *Bernoulli* **2013**, *19*, 1212–1242. [CrossRef]

54. Paninski, L.; Pillow, J.; Lewi, J. Statistical models for neural encoding, decoding, and optimal stimulus design. In *Computational Neuroscience: Theoretical Insights into Brain Function*; Elsevier: Amsterdam, The Netherlands, 2007; Volume 165, pp. 493–507.

55. Pillow, J.; Shlens, J.; Paninski, L.; Sher, A.; Litke, A.; Chichilnisky, E.; Simoncelli, E. Spatio-temporal correlations and visual signaling in a complete neuronal population. *Nature* **2008**, *454*, 995–999. [CrossRef]

56. Zhang, A.; Cai, T.T.; Wu, Y. Heteroskedastic PCA: Algorithm, optimality, and applications. *arXiv* **2019**, arXiv:1810.08316.

57. van de Geer, S. Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes. *Ann. Stat.* **1995**, *23*, 1779–1801. [CrossRef]

58. Negahban, S.; Wainwright, M. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *J. Mach. Learn. Res.* **2010**, *13*, 1665–1697.