Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Latent Network Structure Learning From High-Dimensional Multivariate Point Processes

Biao Cai, Jingfei Zhang & Yongtao Guan

To cite this article: Biao Cai, Jingfei Zhang & Yongtao Guan (2024) Latent Network Structure Learning From High-Dimensional Multivariate Point Processes, Journal of the American Statistical Association, 119:545, 95-108, DOI: 10.1080/01621459.2022.2102019

To link to this article: https://doi.org/10.1080/01621459.2022.2102019

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2024, VOL. 119, NO. 545, 95–108: Theory and Methods
https://doi.org/10.1080/01621459.2022.2102019

Latent Network Structure Learning From High-Dimensional Multivariate Point Processes

Biao Cai, Jingfei Zhang, and Yongtao Guan

Department of Management Science, Miami Herbert Business School, University of Miami, Coral Gables, FL

ABSTRACT

Learning the latent network structure from large scale multivariate point process data is an important task in a wide range of scientific and business applications. For instance, we might wish to estimate the neuronal functional connectivity network based on spiking times recorded from a collection of neurons. To characterize the complex processes underlying the observed data, we propose a new and flexible class of nonstationary Hawkes processes that allow both excitatory and inhibitory effects. We estimate the latent network structure using an efficient sparse least squares estimation approach. Using a thinning representation, we establish concentration inequalities for the first and second order statistics of the proposed Hawkes process. Such theoretical results enable us to establish the non-asymptotic error bound and the selection consistency of the estimated parameters. Furthermore, we describe a least squares loss based statistic for testing if the background intensity is constant in time. We demonstrate the efficacy of our proposed method through simulation studies and an application to a neuron spike train dataset. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received September 2019
Accepted July 2022

KEYWORDS
Multivariate Hawkes process; Non asymptotic error bound; Nonlinear Hawkes process; Nonstationary; Selection consistency

1. Introduction

Large-scale multivariate point process data are fast emerging in a wide range of scientific and business applications. Learning the latent network structure from such data has become an increasingly important task. For instance, one may wish to estimate the neuronal functional connectivity network based on spiking times (i.e., times when a neuron fires) recorded from a collection of neurons (Farajtabar et al. 2015), or to estimate the financial network based on trading times recorded for a collection of stocks (Linderman and Adams 2014). Both the neuron spiking times and the trading times can be viewed as realizations from multivariate point processes. To characterize the latent interactions between the different point processes, a useful class of models is the multivariate Hawkes process (Hawkes 1971). The multivariate Hawkes process is a mutually-exciting point process, in which the arrival of one event in one point process may trigger those of future events across the different processes. Because of its flexibility and interpretability, the multivariate Hawkes process has been widely used in many applications, such as social studies (Zhou, Zha, and Song 2013), criminology (Linderman and Adams 2014), finance (Bacry et al. 2013) and neuroscience (Okatan, Wilson, and Brown 2005). In the network setting, each component point process of the multivariate Hawkes process is viewed as a node, with a directed edge connecting two nodes indicating an event in the source point process increases the probability of occurrence of future events in the target point process.

Despite the popularity of the multivariate Hawkes process, there is a need of new statistical theory and methodology for its broader applications. *First*, most existing theoretical results for the Hawkes process are derived using a cluster process representation of the process. This cluster process representation by its definition depends on the mutually excitation assumption, that is, the arrival of one event increases the probability of occurrence of future events (Hawkes and Oakes 1974; Hansen, Reynaud-Bouret, and Rivoirard 2015). However, such an assumption may not be valid in certain applications. For example, it is well known that the firing activity of one neuron can inhibit the activities of other neurons (Amari 1977). A more flexible model should allow both excitatory and inhibitory effects, which renders the cluster process representation infeasible. *Second*, most existing models assume that the background intensities, that is, the baseline arrival rates of events from the different component processes, are constant in time. Under this assumption, the multivariate Hawkes process satisfies a stationary condition (e.g., Brémaud and Massoulié 1996). However, assuming constant background intensities may also be too restrictive. For example, stock trading activities tend to be much higher during market opens and closes (Engle and Russell 1998), and the associated background intensities are therefore not constant in time. A multivariate Hawkes process with constant background intensities may not fit such data well (Chen and Hall 2013). A more flexible approach instead should allow the background intensities to be time-varying. For such nonstationary models, new development

CONTACT Jingfei Zhang *ezhang@bus.miami.edu* Department of Management Science, Miami Herbert Business School, University of Miami, Coral Gables, FL, 33146.

Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

on both theory and methodology is needed, as most existing results are established assuming the underlying process to be stationary.

Some existing work have considered broadening the class of Hawkes process models. Specifically, Brémaud and Massoulié (1996), Chen et al. (2017), and Costa et al. (2018) considered a class of nonlinear Hawkes processes that allows both excitatory and inhibitory effects. A thinning process representation was used to investigate the properties of the proposed process. However, these work focused on processes with constant background intensities and the thinning representation technique depended critically on the stationarity condition. Some recent work also considered nonstationary Hawkes processes. Lewis and Mohler (2011), Chen and Hall (2013), and Roueff, Von Sachs, and Sansonnet (2016) considered Hawkes processes with time varying background intensities. However, they only considered univariate processes, and only with excitatory effects. Lemonnier and Vayatis (2014) considered a multivariate Hawkes process with time varying background intensities. However, they focused on an approximate optimization algorithm for model estimation, and did not provide any theoretical results.

In this article, we propose a flexible class of multivariate Hawkes process that admits time-varying background intensities and allows both excitatory and inhibitory effects. We show the existence of a thinning process representation of this nonstationary process. Such a result has not yet been established in the literature, and it enables our subsequent theoretical analysis. To estimate the network structure, we consider a computationally efficient penalized least squares estimation, in which both the background intensities and the transfer functions are approximated using basis functions. We establish theoretical properties of the penalized least squares estimator in the high-dimensional regime, where the dimension of the multivariate process p can grown much faster than the length of the observation window T. Specifically, we investigate the following properties in our analysis:

1. (Concentration inequalities.) We establish concentration inequalities for the first and second order statistics of the proposed Hawkes process. Such inequalities are established using a new thinning process representation result for nonlinear and nonstationary Hawkes processes.
2. (Non asymptotic error bound.) Under certain regularity conditions, we establish, in the high-dimensional regime, the non asymptotic error bound of the intensity functions estimated using basis approximations. Specifically, we verify that the design matrix satisfy a restricted eigenvalue condition and a bounded eigenvalue condition for the diagonal blocks; these bounds on eigenvalues depend on the number of basis functions.
3. (Network recovery.) We show that, under certain regularity conditions, our proposed estimation method can consistently identify the true edges in the network with probability tending to one. Moreover, we propose a consistent generalized information criterion (GIC) for regularizing parameter selection.
4. (Test for background intensity.) We propose a least squares based statistic for testing if the background intensity is constant in time. Specifically, we show that the null distribution of the test statistic is asymptotically χ^2 and the test is powerful against alternatives.

It is worth mentioning that there is another class of approaches that estimate the latent network structure from high dimensional multivariate point process data (Zhang et al. 2016; Vinci et al. 2016, 2018). These methods divide the observation window into a number of bins, and model the number of events in each bin. The network structure is estimated using methods such as correlation of event counts (Vinci et al. 2016), regularized generalized linear models (Zhang et al. 2016), or Gaussian graphical models (Vinci et al. 2018). The heuristic binning procedure may result in information loss, for example, short-term excitatory effects may be overlooked if the bins are chosen to be too wide. Choosing an appropriate binning procedure remains a challenging task.

The rest of the article is organized as follows. Section 2 introduces the proposed model, and Section 3 describes the model estimation and selection. The aforementioned theoretical results are detailed in Sections 4. Section 5 includes simulation studies. The detailed analysis of a neuron spike train dataset is presented in Section 6. A short discussion section concludes the article.

2. Model

2.1. Notation

Given a function f on $\mathcal{X} \in \mathbb{R}$, let $\|f\|_{\infty,\mathcal{X}} = \sup_{t\in\mathcal{X}} |f(t)|$ and $\|f\|_{2,\mathcal{X}} = \{\int_{t\in\mathcal{X}} f(t)^2 \mathrm{d}t\}^{1/2}$(or, respectively, $\|f\|_\infty$ and $\|f\|_2$, when there is no ambiguity). Let $f^{(k)}$ denote the kth derivative of a function f when such a derivative exists. For a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, we use $\|\mathbf{A}\|_2$, $\|\mathbf{A}\|_{\max}$ and $\|\mathbf{A}\|_\infty$ to denote its spectral norm, maximum entry-wise ℓ_1 norm and maximum row-wise ℓ_1 norm, respectively. We write $[n] = \{1, 2, \ldots, n\}$ and let $\lfloor x \rfloor$ denote the largest integer less than x. For a set $\mathcal{S}$, we use $|\mathcal{S}|$ to denote its cardinality. We write $\mathbf{1}_n$ to denote a length-n vector of 1, $\mathbf{I}_{n\times n}$ to denote a $n \times n$ identity matrix, $\mathrm{diag}\{d_1, \ldots, d_n\}$ to denote a $n \times n$ diagonal matrix with diagonal elements $d_1, \ldots, d_n$, and use $\sigma_{\min}(\cdot)$ and $\sigma_{\max}(\cdot)$ to denote the smallest and largest eigenvalues of a matrix, respectively. For two positive sequences a_n and b_n, write $a_n = \mathcal{O}(b_n)$ if there exist $c > 0$ and $N > 0$ such that $a_n < cb_n$ for all $n > N$, write $a_n \asymp b_n$ if $a_n = \mathcal{O}(b_n)$ and $b_n = \mathcal{O}(a_n)$, and $a_n = o(b_n)$ if $a_n/b_n \to 0$ as $n \to \infty$. For a sequence of random variables Y_n and a positive sequence a_n, we write $Y_n = \mathcal{O}_p(a_n)$ if for any $\epsilon > 0$, there exist $M > 0$ and $N > 0$ such that $\mathbb{P}(|Y_n/a_n| > M) < \epsilon$ for any $n > N$; we write $Y_n = o_p(a_n)$ if $\lim_{n\to\infty} \mathbb{P}(|Y_n/a_n| \geq \epsilon) = 0$ for any $\epsilon > 0$.

2.2. The Multivariate Hawkes Process Model

Consider a directed network with p nodes. For each node $j \in [p]$, we observe its event locations $\{t_{j,1}, t_{j,2}, \ldots\}$ in the time interval $[0, T]$ such that $0 < t_{j,1} < t_{j,2} < \cdots \leq T$. For node j, let the associated counting process be $N_j(t) = |\{i : t_{j,i} \leq t\}|$, $t \in [0, T]$. Write $\mathbf{N} = (N_j)_{j\in[p]}$ as the p-variate counting process. Let $\mathcal{H}_t$ denote the entire history of $\mathbf{N}$ up to time t, and write $N_j([t, t + \mathrm{d}t))$ as $\mathrm{d}N_j(t)$. The p-variate intensity function

$\boldsymbol{\lambda}(t) = (\lambda_1(t), \ldots, \lambda_p(t))^\top$ of $\mathbf{N}$ is defined as

$$\lambda_j(t)\mathrm{d}t = \mathbb{P}(\mathrm{d}N_j(t) = 1|\mathcal{H}_t), \quad j \in [p].$$

We propose a flexible class of Hawkes processes with intensity functions defined as

$$\lambda_j(t) = h\left\{\nu_j(t) + \sum_{k=1}^{p}\int_0^t \omega_{j,k}(t-u)\mathrm{d}N_k(u)\right\}, \quad j \in [p], \quad (1)$$

where $h(\cdot) : \mathbb{R} \to \mathbb{R}^+$ is a link function and it is assumed to be θ-Lipschitz (see Assumption 1), $\nu_j(\cdot) : \mathbb{R}^+ \to \mathbb{R}^+$ is the time-varying background (or baseline) intensity function of the jth process, and $\omega_{j,k}(\cdot) : \mathbb{R}^+ \to \mathbb{R}$ is the transfer function that characterizes the effect of the kth process on the jth process. Specifically,

i. $\omega_{j,k}(s) > 0$ corresponds to *excitatory* effect, that is, an event in process k increases the probability of event occurrence in process j at a time distance of s.
ii. $\omega_{j,k}(s) < 0$ corresponds to *inhibitory* effect, that is, an event in process k decreases the probability of event occurrence in process j at a time distance of s.
iii. $\omega_{j,k}(s) = 0$ corresponds to no effect, that is, an event in process k has no effect on the event occurrence in process j at a time distance of s.

The proposed model in (1) considers a time dependent background intensity function instead of the constant background intensity considered in existing multivariate Hawkes process models (Chen and Hall 2013; Hansen, Reynaud-Bouret, and Rivoirard 2015; Bacry et al. 2020; Wang, Kolar, and Shojaie 2020). Consequently, the proposed Hawkes process model is nonstationary, and its analysis requires new theoretical tools, which will be introduced in Section 4.

Let the directed network $\mathcal{G}(\mathcal{V}, \mathcal{E})$ summarize the relationships between the p component processes. Specifically, let $\mathcal{V} = \{1, 2, \ldots, p\}$ be the set of p nodes and $\mathcal{E}$ be the set of edges such that

$$\mathcal{E} = \{(j,k) : \omega_{j,k} \neq 0,\ j,k \in [p]\},$$

where $\omega_{j,k}$'s are the transfer functions in (1). Therefore, $(j,k) \in \mathcal{E}$ if and only if the kth process has an excitatory or inhibitory effect on the jth process.

Next, we introduce a set of regularity conditions on the background intensities and transfer functions in (1).

Assumption 1. Let $\boldsymbol{\Omega}$ be a $p \times p$ matrix with $\Omega_{jk} = \int_0^\infty |\omega_{j,k}(t)|\mathrm{d}t$, $j,k \in [p]$ and assume that $\sigma_{\max}(\boldsymbol{\Omega}^\top\boldsymbol{\Omega}) \leq \sigma_{\boldsymbol{\Omega}} < 1$. Moreover, assume that $h(\cdot)$ is a θ-Lipschitz link function with $\theta \leq 1$, and the background intensity functions are bounded, that is, $0 < \nu_j(t) \leq \nu$, $j \in [p]$, for some positive constant ν.

Assuming the Lipschitz constant θ to satisfy $\theta \leq 1$ is not restrictive. For example, if $h(\cdot)$ is K_0-Lipschitz for some $K_0 > 1$, we can reparameterize (1) by setting $\widetilde{h}(x) = h(x/K_0)$, $\widetilde{\nu}_j(t) = K_0\nu_j(t)$ and $\widetilde{\omega}_{j,k}(t) = K_0\omega_{j,k}(t)$. In this reparameterized model, $\widetilde{h}(\cdot)$ is 1-Lipschitz. The Lipschitz condition on the link function was also considered in Massoulié (1998) and Chen et al. (2017). Assumption 1 implies that $h\{\nu_j(t)\}$ is bounded as Lipschitz functions are bounded on bounded supports.

Define the mean intensity of (1) as $\bar{\lambda}_j(t) = \mathbb{E}\{\mathrm{d}N_j(t)\}/\mathrm{d}t$. Under Assumption 1, the mean intensity $\bar{\lambda}_j(t)$ is upper bounded. This can be shown in three steps. First, define a p-dimensional Hawkes process $\mathbf{N}^* = (N_j^*)_{j\in[p]}$ with intensity function

$$\lambda_j^*(t) = \nu^* + \sum_{k=1}^{p}\int_0^t |\omega_{j,k}(t-u)|\mathrm{d}N_k^*(u), \quad j \in [p], \quad (2)$$

where ν^* is a positive constant such that $\nu^* \geq h\{\nu_j(t)\}$ for any j and t, and $\omega_{j,k}$'s are as defined in (1). By Brémaud and Massoulié (1996), the point process defined in (2) satisfies a stationary condition under Assumption 1. Next, write the mean intensity of (2) as $\boldsymbol{\Lambda}^* = (\Lambda_1^*, \ldots, \Lambda_p^*)^\top$, where $\Lambda_j^* = \mathbb{E}\{\mathrm{d}N_j^*(t)\}/\mathrm{d}t$. Correspondingly, we have

$$\boldsymbol{\Lambda}^* = \boldsymbol{\nu}^* + \left\{\int_0^\infty |\boldsymbol{\omega}(t)|\mathrm{d}t\right\}\boldsymbol{\Lambda}^*, \quad (3)$$

where $\boldsymbol{\nu}^* = (\nu^*, \ldots, \nu^*)^\top \in \mathbb{R}^p$ and $\boldsymbol{\omega}(t) \in \mathbb{R}^{p\times p}$, with $\{\boldsymbol{\omega}(t)\}_{jk} = \omega_{j,k}(t)$. The mean intensity $\boldsymbol{\Lambda}^*$ in (3) can be rewritten as $\boldsymbol{\Lambda}^* = \sum_{k=0}^\infty \boldsymbol{\Omega}^k\boldsymbol{\nu}^*$, which is upper bounded given $\sigma_{\boldsymbol{\Omega}} < 1$ in Assumption 1. Finally, it can be shown that the mean intensity of (1), that is, $\bar{\lambda}_j(t)$, is upper bounded by Λ_j^* (see Lemma S6 and its proof in the supplementary materials). Consequently, $\bar{\lambda}_j(t)$ is also upper bounded under Assumption 1.

3. Estimation

From the observed event locations in $[0, T]$, our objective is to estimate the intensity functions $\lambda_j(t)$, $j \in [p]$. Furthermore, by identifying the nonzero transfer functions $\omega_{j,k}$'s in the estimated intensity functions, we can estimate the structure of the directed network $\mathcal{G}(\mathcal{V}, \mathcal{E})$. In this section, we consider the link function to be $h(x) = \max(0, x)$ in (1). It is seen that this link function is a 1-Lipschitz function. To ease notation, we write

$$\psi_j(t) = \nu_j(t) + \sum_{k=1}^{p}\int_0^t \omega_{j,k}(t-u)\mathrm{d}N_k(u), \quad (4)$$

and $\lambda_j(t) = \max\{0, \psi_j(t)\}$, $j \in [p]$.

To estimate the intensity functions, one may consider a likelihood function based approach (Ogata 1981; Chen and Hall 2013; Zhou, Zha, and Song 2013). However, minimizing the negative log-likelihood function may require a very involved and computationally intensive iterative procedure (Veen and Schoenberg 2008). To improve the estimation efficiency, we consider a least squares loss based estimation approach. That is, we consider the following loss function

$$\frac{1}{T}\sum_{j=1}^{p}\int_0^T \{\psi_j^2(t)\mathrm{d}t - 2\psi_j(t)\mathrm{d}N_j(t)\}. \quad (5)$$

The least squares loss comes from the empirical risk minimization principle (van de Geer 2000) and has been fairly commonly considered in estimating point process models (Hansen, Reynaud-Bouret, and Rivoirard 2015; Chen et al. 2017; Bacry et al. 2020). We later show that (5) can be separated into p objective functions that can be estimated individually, which significantly reduces the computation cost.

We consider a nonparametric estimation of $\nu_j(t)$ and $\omega_{j,k}(t)$, $j,k \in [p]$, using B-spline approximations. Given $[a_1, a_2] \subset \mathbb{R}$ and a set of K knots $a_1 = \zeta_0 < \zeta_1 < \cdots < \zeta_{K+1} = a_2$ such that $\max_{1\le l\le K+1} |\zeta_k - \zeta_{k-1}| = \mathcal{O}(K^{-1})$, let $\mathcal{S}_{K,l}$ be the space of polynomial splines of degree $l \ge 1$ consisting of functions satisfying: (i) restricting to each interval $[\zeta_i, \zeta_{i+1}]$, $i \in [K]$, the function is a polynomial of degree $l-1$; (ii) for $l \ge 2$ and $0 \le l' \le l-2$, the function is l' times continuously differentiable (Stone 1985). Such a space $\mathcal{S}_{K,l}$ is of dimension $m = K + l$ (Schumaker 2007) and as such, let $\{\phi_1(t), \ldots, \phi_m(t)\}$ be the normalized B-spline basis of $\mathcal{S}_{K,l}$. When $l = 1$, the basis is a set of $K+1$ step functions with jumps at knots (Stone 1985). In our procedure, we approximate the background intensity $\nu_j(t)$ with an m_0-dimensional normalized B-spline basis $\boldsymbol{\phi}_0(t) = (\phi_{0,1}(t), \ldots, \phi_{0,m_0}(t))^\top$, such that $\nu_j(t) = \boldsymbol{\beta}_{j,0}\boldsymbol{\phi}_0(t) + r_{j,0}(t)$, where $\boldsymbol{\beta}_{j,0} \in \mathbb{R}^{m_0}$ and $r_{j,0}(\cdot)$ denotes the approximation residual. Furthermore, we approximate the transfer functions $\omega_{j,k}(t)$ with an m_1-dimensional normalized B-spline basis $\boldsymbol{\phi}_1(t) = (\phi_{1,1}(t), \ldots, \phi_{1,m_1}(t))^T$, such that $\omega_{j,k}(t) = \boldsymbol{\beta}_{j,k}\boldsymbol{\phi}_1(t) + r_{j,k}(t)$, where $\boldsymbol{\beta}_{j,k} \in \mathbb{R}^{m_1}$ and $r_{j,k}(\cdot)$ denotes the approximation residual. The dimensions and degrees of the bases $\boldsymbol{\phi}_0(t)$ and $\boldsymbol{\phi}_1(t)$ are allowed to be different for more flexibility in characterizing the background intensities and transfer functions. For example, one may use cubic B-splines to approximate the background intensities and step functions to approximate the transfer functions (Hansen, Reynaud-Bouret, and Rivoirard 2015). The choices for the number and locations of knots are discussed in Section 3.1.

Write $\boldsymbol{\beta}_j = (\boldsymbol{\beta}_{j,0}, \boldsymbol{\beta}_{j,1}, \ldots, \boldsymbol{\beta}_{j,p})^\top$. We define $\boldsymbol{\alpha}_j = (\boldsymbol{\alpha}^{(j,0)}, \boldsymbol{\alpha}^{(j,1)}, \ldots, \boldsymbol{\alpha}^{(j,p)})^\top$ such that $\boldsymbol{\alpha}^{(j,0)} \in \mathbb{R}^{m_0}$ with

$$\alpha_l^{(j,0)} = \frac{1}{T}\int_0^T \phi_{0,l}(t)\mathrm{d}N_j(t), \quad l \in [m_0],$$

and $\boldsymbol{\alpha}^{(j,k)} \in \mathbb{R}^{m_1}$, $k \in [p]$ with

$$\alpha_l^{(j,k)} = \frac{1}{T}\int_0^T \int_0^t \phi_{1,l}(t-u)\mathrm{d}N_k(u)\mathrm{d}N_j(t), \quad l \in [m_1].$$

Moreover, we define $\mathbf{G} \in \mathbb{R}^{(m_0+pm_1)\times(m_0+pm_1)}$ such that

$$\mathbf{G} = \begin{pmatrix} \mathbf{G}^{(0,0)} & \mathbf{G}^{(0,1)} & \ldots & \mathbf{G}^{(0,p)} \\ \mathbf{G}^{(1,0)} & \mathbf{G}^{(1,1)} & \ldots & \mathbf{G}^{(1,p)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{G}^{(p,0)} & \mathbf{G}^{(p,1)} & \ldots & \mathbf{G}^{(p,p)} \end{pmatrix}, \tag{6}$$

where the component $\mathbf{G}^{(k_1,k_2)}$ is defined as

$$\mathbf{G}^{(k_1,k_2)} = \begin{cases} \frac{1}{T}\int_0^T \boldsymbol{\phi}_0(t)\boldsymbol{\phi}_0^\top(t)\mathrm{d}t, & \text{if } k_1 = 0, k_2 = 0, \\ \frac{1}{T}\int_0^T \boldsymbol{\phi}_0(t)\left\{\int_0^t \boldsymbol{\phi}_1^\top(t-u)\mathrm{d}N_{k_2}(u)\right\}\mathrm{d}t, & \text{if } k_1 = 0, k_2 \neq 0, \\ \frac{1}{T}\int_0^T \left\{\int_0^t \boldsymbol{\phi}_1(t-u)\mathrm{d}N_{k_1}(u)\right\}\boldsymbol{\phi}_0^\top(t)\mathrm{d}t, & \text{if } k_1 \neq 0, k_2 = 0, \\ \frac{1}{T}\int_0^T \left\{\int_0^t \boldsymbol{\phi}_1(t-u)\mathrm{d}N_{k_1}(u)\right\}\left\{\int_0^t \boldsymbol{\phi}_1^\top(t-u)\mathrm{d}N_{k_2}(u)\right\}\mathrm{d}t, & \text{if } k_1 \neq 0, k_2 \neq 0. \end{cases}$$

With the above expressions for $\boldsymbol{\beta}_j$, $\boldsymbol{\alpha}_j$ and $\mathbf{G}$, we define

$$\ell_j(\boldsymbol{\beta}_j) \triangleq -2\boldsymbol{\beta}_j^\top\boldsymbol{\alpha}_j + \boldsymbol{\beta}_j^\top\mathbf{G}\boldsymbol{\beta}_j. \tag{7}$$

Some straightforward algebra shows that the loss function in (5) can be written as $\sum_{j=1}^p \ell_j(\boldsymbol{\beta}_j)$. We note that both $\boldsymbol{\alpha}_j$ and $\mathbf{G}$ are calculated based on the observed event locations and the prespecified basis functions, that is, $\boldsymbol{\phi}_0(t)$ and $\boldsymbol{\phi}_1(t)$. Therefore, to estimate the background intensities and transfer functions, we can directly optimize (7) with respect to $\boldsymbol{\beta}_j$. Since the loss function $\sum_{j=1}^p \ell_j(\boldsymbol{\beta}_j)$ can be decomposed into p separate convex loss functions, that is, $\ell_1(\boldsymbol{\beta}_1), \ldots, \ell_p(\boldsymbol{\beta}_p)$, we can optimize each loss function separately.

Define $\mathcal{E}_j = \{k : \omega_{j,k} \neq 0, k \in [p]\}$. To estimate $\mathcal{E}_j$, we consider $\widehat{\mathcal{E}}_j = \{k : \widehat{\omega}_{j,k} \neq 0, k \in [p]\}$, where $\widehat{\omega}_{j,k}$'s are the estimated transfer functions. Note that if $\omega_{j,k} = 0$, then all coefficients associated with $\omega_{j,k}$ are zero (i.e., $\boldsymbol{\beta}_{j,k} = 0$). Thus, to encourage sparsity in the estimated network, we impose a standardized group lasso penalty on $\boldsymbol{\beta}_j$ with the coefficients in $\boldsymbol{\beta}_{j,k}$ grouped together, $k \in [p]$. Specifically, we consider the following optimization problem

$$\min_{\boldsymbol{\beta}_j \in \mathbb{R}^{m_0+pm_1}} -2\boldsymbol{\beta}_j^\top\boldsymbol{\alpha}_j + \boldsymbol{\beta}_j^\top\mathbf{G}\boldsymbol{\beta}_j + \eta_j \sum_{k=1}^p \left(\boldsymbol{\beta}_{j,k}^\top\mathbf{G}^{(k,k)}\boldsymbol{\beta}_{j,k}\right)^{1/2}. \tag{8}$$

The penalty term $\sum_{k=1}^p \left(\boldsymbol{\beta}_{j,k}^\top\mathbf{G}^{(k,k)}\boldsymbol{\beta}_{j,k}\right)^{1/2}$ is an extension of the standardized group lasso penalty (Simon et al. 2013). This optimization problem in (8) is convex and can be efficiently solved using a block coordinate descent algorithm (Simon et al. 2013). The terms $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p$, and $\mathbf{G}$ can be computed using standard numerical integration methods and such calculations can be carried out before implementing the block coordinate descent algorithm.

3.1. Tuning Parameter Selection

Our proposed estimation procedure involves a number of tuning parameters, including the numbers of B-splines (i.e., m_0 and m_1) for approximating the background and transfer functions, respectively, knots locations for the B-splines, and tuning parameter η_j's in the penalized least squares estimation in (8). Cross-validation procedures may not be appropriate for tuning parameter selections under our setting as the proposed process is nonstationary due to the time-varying background intensity in (1). As such, the data cannot be divided into training and validation sets in a straightforward manner.

Given m_0 and m_1, we let the knots be evenly distributed (Ravikumar et al. 2009; Huang, Horowitz, and Wei 2010). For m_0 and m_1, theoretical conditions in Theorem 3 can guide their empirical choices. In Section B1, we describe a heuristic

procedure for selecting m_0 and m_1; a similar procedure was considered in Kozbur (2020). In Section B2, we show that this heuristic procedure achieves good performance; additionally, we demonstrate that the estimation accuracy is not overly sensitive to the choices of m_0 and m_1. Once $\boldsymbol{\phi}_0(t)$ and $\boldsymbol{\phi}_1(t)$ are determined, we then move to select η_j's.

The tuning parameter η_j in (8) controls the sparsity of $\boldsymbol{\beta}_j$, which in turn controls the sparsity of the estimated network. To select η_j, we propose a generalized information criterion (GIC) defined as

$$\mathrm{GIC}(\eta_j) = \ell_j(\widehat{\boldsymbol{\beta}}_j) \cdot \kappa_j + (\alpha_T/T) \cdot |\widehat{\mathcal{E}}_j| \tag{9}$$

where $\ell_j(\cdot)$ is as defined in (7), $\kappa_j = T/N_j\{(0,T]\}$ is a scaling parameter, $\widehat{\boldsymbol{\beta}}_j$ is estimated from (8) with η_j, and $\alpha_T > 0$ is a parameter that scales with T and p. As $\ell_j(\widehat{\boldsymbol{\beta}}_j)$ is the squares loss and not the log-likelihood function, the GIC is not directly comparable to the likelihood based selection criteria such as the BIC or extended BIC (Schwarz 1978; Chen and Chen 2008). In Theorem 5, we show that the proposed GIC is consistent given appropriate choices of α_T, such as $\mathcal{O}((\log p)^2 \log T)$. In Section 5, we evaluate the efficacy of the proposed GIC and show it achieves satisfactory performance.

4. Theoretical Properties

In this section, we first show the existence of a thinning process representation of the proposed nonlinear and nonstationary Hawkes process. We then establish concentration inequalities for the first and second order statistics of the proposed point process. These results are useful in the subsequent analysis of the estimated intensity functions. Next, we establish the nonasymptotic error bound of the estimated intensity functions and show that our method can consistently identify the true edges in the network. Lastly, we propose a test statistic for testing if the background intensities are constant in time. We derive its asymptotic null distribution and show the test is powerful against alternatives. All proofs are collected in the supplementary materials.

4.1. Concentration Inequalities

Many existing theoretical analyses rely on the cluster process representation of the Hawkes process (Hawkes and Oakes 1974; Hansen, Reynaud-Bouret, and Rivoirard 2015; Bacry et al. 2020), which needs the transfer functions to be nonnegative. Brémaud and Massoulié (1996) employed a thinning process representation of the Hawkes process that permitted negative transfer functions, but required a stationarity condition. As such, the results in Brémaud and Massoulié (1996) are not directly applicable to our problem. Next, we first show the existence of a thinning process representation of the proposed nonlinear and nonstationary Hawkes process in (1).

Let $\overline{\mathbf{N}} = (\overline{N}_j)_{j\in[p]}$ be a p-variate homogeneous Poisson process on $\mathbb{R}^2$ with intensity 1. Let $\lambda_j^{(0)}(t) = 0$, $j \in [p]$, and $N_j^{(0)} = \varnothing$. For $n \geq 1$, construct recursively $\boldsymbol{\lambda}^{(n)}(t) = (\lambda_1^{(n)}(t), \ldots, \lambda_p^{(n)}(t))^\top$ and $\mathbf{N}^{(n)} = (N_j^{(n)})_{j\in[p]}$ as follows:

$$\begin{aligned}
\lambda_j^{(n+1)}(t) &= h\left\{\nu_j(t) + \sum_{k=1}^{p}\int_0^t \omega_{j,k}(t-u)\mathrm{d}N_k^{(n)}(u)\right\},\\
\mathrm{d}N_j^{(n+1)}(t) &= \overline{N}_j\left(\left[0, \lambda_j^{(n+1)}(t)\right] \times \mathrm{d}t\right), \quad j \in [p],
\end{aligned} \tag{10}$$

where $h(\cdot)$, ν_j and $w_{j,k}$ are as defined in (1), and $\overline{N}_j([0, \lambda_j^{(n+1)}(t)] \times \mathrm{d}t)$ denotes the number of points for $\overline{N}_j$ in the area $[0, \lambda_j^{(n+1)}(t)] \times [t, t+\mathrm{d}t]$. It follows from Lemma S2 that $\lambda_j^{(n)}(t)$ is the intensity function of the point process $N^{(n)}(t)$. Next, we show that the sequence $\{\mathbf{N}^{(n)}\}_{n=1}^{\infty}$ in (10) converges in distribution to the Hawkes process $\mathbf{N}$ with intensity function (1).

Theorem 1. Let $\boldsymbol{\lambda}(t)$ be as defined in (1) satisfying Assumption 1. Let $\{\boldsymbol{\lambda}^{(n)}(t)\}_{n=1}^{\infty}$ and $\{\mathbf{N}^{(n)}\}_{n=1}^{\infty}$ be sequences as defined in (10). Then, it holds that

(a) $\boldsymbol{\lambda}^{(n)}(t)$ converges to $\boldsymbol{\lambda}(t)$ almost surely for any t,
(b) $\{\mathbf{N}^{(n)}\}_{n=1}^{\infty}$ converges in distribution to $\mathbf{N}$ with intensity (1).

Theorem 1 shows the existence of a thinning process representation of the proposed nonstationary Hawkes Process. It provides a theoretical guarantee, analogous to Massoulié (1998), for nonstationary multivariate Hawkes processes. This new result is critical in our subsequent theoretical analysis.

Assumption 2. Assume that there exists $\lambda_{\max} > 0$ such that $\lambda_j(t) \leq \lambda_{\max}$ for any t and j and $\omega_{j,k}$, $j,k \in [p]$ are bounded functions with a bounded support $[0,b]$ for some $b > 0$.

This condition first assumes that the intensities are upper bounded. One example of such processes is when the link function $h(\cdot)$ is upper bounded by a positive constant; see also Section 7. Assumption 2 also assumes that the transfer functions $\omega_{j,k}$'s have a bounded support. The bounded support assumption has been fairly commonly considered in the analysis of multivariate Hawkes process (Hansen, Reynaud-Bouret, and Rivoirard 2015; Costa et al. 2018).

Assumption 3. There exists $\rho_{\boldsymbol{\Omega}} \in (0,1)$ such that $\sum_{k=1}^{p} \Omega_{j,k} \leq \rho_{\boldsymbol{\Omega}}$, $j \in [p]$.

This assumption requires that $\boldsymbol{\Omega}$ has bounded column sums, which prevents the intensity function from concentrating on any single process.

Recall that $\mathcal{H}_t$ denotes the history of $\mathbf{N}$ up to time t. For $\mathcal{H}_t$-predictable functions $f_1(\cdot)$ and $f_2(\cdot)$, define

$$y_k = \frac{1}{T}\int_0^T f_1(t)\mathrm{d}N_k(t),$$

$$y_{j,k} = \frac{1}{T}\int_0^T\int_0^T f_2(t-t')\mathrm{d}N_k(t')\mathrm{d}N_j(t).$$

Theorem 2. Consider a Hawkes process on $[0,T]$ with intensity as defined in (1) satisfying Assumptions 1–3. Let $f_1(t)$ be

a bounded function and $f_2(t)$ be a bounded function on a bounded support. Then, for $k \in [p]$, it holds that

$$\mathbb{P}(|y_k - \mathbb{E}y_k| \geq c_1 T^{-3/5}) \leq c_2 T \exp(-c_3 T^{1/5}), \tag{11}$$

where c_1, c_2, and c_3 are positive constants. For any $j, k \in [p]$, it holds that

$$\mathbb{P}(|y_{j,k} - \mathbb{E}y_{j,k}| \geq c_1' T^{-2/5}) \leq c_2' T \exp(-c_3' T^{1/5}), \tag{12}$$

where c_1', c_2', and c_3' are positive constants.

The proof of Theorem 2 is provided in the supplementary materials. Our proof strategy for the concentration results in Theorem 2 follows from that in Chen et al. (2017). Specifically, as in Chen et al. (2017), we first define a coupling process of $\mathbf{N}$, which is used to bound the temporal dependence of $\mathbf{N}$. Then, a Bernstein type inequality for weakly dependent sequences (Merlevède, Peligrad, and Rio 2011) is used to obtain the desired results. The main difference in the proof is that the validity of the thinning process representation in Chen et al. (2017) is ensured by Massoulié (1998), established under a stationarity condition. In the nonstationary case, the validity of the thinning process representation is established in Theorem 1.

We remark that the concentration results for stationary processes in Chen et al. (2017) may not be directly applicable to establish (11)–(12) under the nonstationary case we considered, even though $\mathbf{N}$ is dominated by a stationary process (see Lemma S6), as no existing theoretical results, to our knowledge, establish the concentration inequalities of a target process using directly the concentration inequalities of its dominating process. Theorem 2 is useful in the ensuing theoretical analysis that derives the non-asymptotic error bound of the estimated intensity functions and establishes edge selection consistency. The next corollary is a direct consequence of Theorem 2.

Corollary 1. Consider a Hawkes process on $[0, T]$ with intensity as defined in (1) satisfying Assumptions 1–3. Considering the matrix $\mathbf{G}$ defined in (6), we have

$$\mathbb{P}\Big[\bigcap_{i\neq j}\Big\{\big|\mathbf{G}_{ij} - \mathbb{E}(\mathbf{G}_{ij})\big| \leq c_4 T^{-2/5}\Big\}\Big] \geq 1 - c_5(p+1)^2 T \exp(-c_6 T^{1/5}),$$

where c_4, c_5, and c_6 are positive constants.

The result in Corollary 1 is a direct consequence of Theorem 2, once we show that the entries in $\mathbf{G}$ are first and second order statistics of the proposed Hawkes process.

4.2. Non asymptotic Error Bound

In this section, we derive the non asymptotic error bound of the estimated intensity function in the diverging p regime. @To simplify notation, we define $\mathbf{\Psi}(t) = (\mathbf{\Psi}_0^\top(t), \mathbf{\Psi}_1^\top(t), \ldots, \mathbf{\Psi}_p^\top(t))^\top$, where $\mathbf{\Psi}_0(t) = \boldsymbol{\phi}_0(t)$ and $\mathbf{\Psi}_k(t) = \int_0^t \boldsymbol{\phi}_1(t-u)\mathrm{d}N_k(u)$, $k \in [p]$. Correspondingly, it holds that $\mathbf{G} = \frac{1}{T}\int_0^T \mathbf{\Psi}(t)\mathbf{\Psi}^\top(t)\mathrm{d}t$ and $\mathbf{G}^{(k,k)} = \frac{1}{T}\int_0^T \mathbf{\Psi}_k(t)\mathbf{\Psi}_k^\top(t)\mathrm{d}t$. Let $s = \max_j |\mathcal{E}_j|$, where $\mathcal{E}_j = \{k : \omega_{j,k} \neq 0, k \in [p]\}$.

Recall the first order mean intensity function $\bar{\lambda}_k(u)$ is defined as $\bar{\lambda}_k(u) = \mathbb{E}(\mathrm{d}N_k(u))/\mathrm{d}u$, $k \in [p]$. For $k_1 \neq k_2 \in [p]$ and $k_1 = k_2 \in [p]$, $u_1 \neq u_2$, define the second order mean intensity function $\bar{\lambda}^{(2)}_{k_1,k_2}(u_1, u_2)$ as

$$\bar{\lambda}^{(2)}_{k_1,k_2}(u_1, u_2) = \mathbb{E}\{\mathrm{d}N_{k_1}(u_1)\mathrm{d}N_{k_2}(u_2)\}/(\mathrm{d}u_1\mathrm{d}u_2). \tag{13}$$

Denote the $p \times p$ covariance function as $\mathbf{C}^0(u_1, u_2)$, such that, for $k_1 \neq k_2 \in [p]$ and $k_1 = k_2 \in [p]$, $u_1 \neq u_2$, the (k_1, k_2)th entry is defined as

$$C^0_{k_1,k_2}(u_1, u_2) = \bar{\lambda}^{(2)}_{k_1,k_2}(u_1, u_2) - \bar{\lambda}_{k_1}(u_1)\bar{\lambda}_{k_2}(u_2). \tag{14}$$

When $k_1 = k_2$ and $u_1 = u_2$, it holds that $\mathbb{E}\{\mathrm{d}N_k(u)\mathrm{d}N_k(u)\} = \mathbb{E}\{\mathrm{d}N_k(u)\}$ (Hawkes 1971). Thus, the complete covariance matrix can be written as

$$\mathbf{C}(u_1, u_2) = \delta(u_1 - u_2)\bar{\mathbf{\Lambda}}(u_1) + \mathbf{C}^0(u_1, u_2),$$

where $\delta(\cdot)$ is the Dirac function, $\bar{\mathbf{\Lambda}}(u_1) = \mathrm{diag}\{\bar{\lambda}_1(u_1), \ldots, \bar{\lambda}_p(u_1)\}$ and $C^0_{k,k}(u_1, u_2)$ is continuous at $u_1 = u_2$, $k \in [p]$ (Hawkes 1971). Note that $\boldsymbol{C}(u_1, u_2)$ is in general not symmetric (Li and Zhang 2011). Specifically, $\boldsymbol{C}(u_1, u_2)$ is symmetric when $u_1 = u_2$; when $u_1 \neq u_2$, $\boldsymbol{C}(u_1, u_2)$ is the cross-covariance function, which is not necessarily symmetric, and it holds by definition that $C_{k_1,k_2}(u_1, u_2) = C_{k_2,k_1}(u_2, u_1)$.

Assumption 4. Assume there exist constants $\Lambda_{\min}, \Lambda_{\max} > 0$ such that, $\bar{\lambda}_k(t) \geq \Lambda_{\min}$ and $\bar{\lambda}^{(2)}_{k_1,k_2}(u_1, u_2) \leq \Lambda_{\max}$. Additionally, assume that $\mathbf{C}^0(u_1, u_2)$ is nonnegative definite, that is, $\int\int \mathbf{f}(u_1)^\top \mathbf{C}^0(u_1, u_2)\mathbf{f}(u_2)\mathrm{d}u_1\mathrm{d}u_2 \geq 0$ for any square-integrable functions $\mathbf{f} = (f_1, \ldots, f_p)$.

This condition assumes that the first and second order mean intensities are bounded. The lower-bounded condition on the mean intensity $\bar{\lambda}_j(t)$ can be satisfied when the inhibitory effect from the negative transfer functions is not excessive when compared to the background intensity and the excitatory effect from the positive transfer functions. The nonnegative definite assumption of $\mathbf{C}^0(u_1, u_2)$ holds true for many commonly used univariate point process models (Guan, Jalilian, and Waagepetersen 2013). In the stationary multivariate Hawkes process case, Bacry and Muzy (2016) showed that $\mathbf{C}^0(u_1, u_2)$ is directly related to the solution to an integral equation involving the transfer functions; the integral equation can be numerically solved and an estimate of $\mathbf{C}^0(u_1, u_2)$ can therefore be obtained. In our nonstationary multivariate Hawkes process setup, $\mathbf{C}^0(u_1, u_2)$ may instead be estimated through parametric bootstrap (see Section B4). Validity of the nonnegative definite assumption of $\mathbf{C}^0(u_1, u_2)$ can be subsequently assessed using an estimated $\mathbf{C}^0(u_1, u_2)$.

Assumption 5. Assume that there exist $\widetilde{\boldsymbol{\beta}}_j = (\widetilde{\boldsymbol{\beta}}_{j,0}, \widetilde{\boldsymbol{\beta}}_{j,1}, \ldots, \widetilde{\boldsymbol{\beta}}_{j,p})^\top \in \mathbb{R}^{m_0 + pm_1}$, $j \in [p]$, and a smoothness parameter $d \geq 2$ such that, for some positive constants C_1, C_2' and C_3',

$$\frac{1}{T}\int_0^T \Big\{\mathbf{\Psi}^\top(t)\widetilde{\boldsymbol{\beta}}_j - \lambda_j(t)\Big\}^2 \mathrm{d}t \leq C_1(s+1)^2 m_1^{-2d}, \tag{15}$$

with probability at least $1 - C_2' pT\exp(-C_3' T^{1/5})$, where $m_1 \asymp m_0$ and $\widetilde{\boldsymbol{\beta}}_{j,k} = \mathbf{0}$ for $k \notin \mathcal{E}_j$.

This condition assumes that the true intensity function can be well approximated by the basis functions, in that residuals from the truncated basis approximation decreases at a polynomial rate of the number of basis functions. The $d \geq 2$ is a smoothness parameter for the background intensities $\nu_j(t)$'s and transfer functions $\omega_{j,k}(t)$'s. While this parameter may differ between $\nu_j(t)$'s and $\omega_{j,k}(t)$'s, it is assumed to be the same to simplify notations in our analysis. Condition (15) can be verified when, for example, $h(x) = x$ and the approximation errors satisfy $\frac{1}{T}\|\mathbf{b}_{j,0}\boldsymbol{\phi}_0(t) - \nu_j(t)\|^2_{2,[0,T]} = \mathcal{O}(m_0^{-2d})$ and $\|\mathbf{b}_{j,k}\boldsymbol{\phi}_1(t) - \omega_{j,k}(t)\|^2_{2,[0,b]} = \mathcal{O}(m_1^{-2d})$ for some $\mathbf{b}_{j,0} \in \mathbb{R}^{m_0}$ and $\mathbf{b}_{j,k} \in \mathbb{R}^{m_1}$, $j \in [p]$, where b is as defined in Assumption 2; see a detailed proof of this statement in Section A10 and also Section 7. Such approximation errors hold for B-spline basis (Stone 1985) or trigonometric basis (Tsybakov 2008) when the target functions belong to certain function classes. For example, when $\omega_{j,k}$ is d-smooth (Chen 2007), that is, $|\omega^{(l)}_{j,k}(t) - \omega^{(l)}_{j,k}(s)| \leq c|t-s|^{d-l}$, where $l = \lfloor d \rfloor$ and c is some positive constant, there exists $\mathbf{b}_{j,k} \in \mathbb{R}^{m_1}$ for normalized B-spline basis $\boldsymbol{\phi}_1(t)$ of dimension m_1 such that $\|\mathbf{b}_{j,k}\boldsymbol{\phi}_1(t) - \omega_{j,k}(t)\|^2_{2,[0,b]} = \mathcal{O}(m_1^{-2d})$ (Stone 1985). We refer to Chen (2007) and Tsybakov (2008) for thorough reviews of basis approximations and truncation errors.

Next we establish the non-asymptotic error bound of the estimated intensity functions.

Theorem 3. Consider a Hawkes process on $[0, T]$ with intensity as defined in (1) satisfying Assumptions 1–5. For $j \in [p]$, let $\widehat{\lambda}_j(t) = \boldsymbol{\Psi}^\top(t)\widehat{\boldsymbol{\beta}}_j$, where $\widehat{\boldsymbol{\beta}}_j$ is estimated from (8). Given $\eta_j = (C_2 \log p/T)^{1/2}$, $s = o(T^{2/5})$, $\log p = \mathcal{O}(T^{1/5})$ and $sm_1 = \mathcal{O}(T^{4/5})$, we have, for $j \in [p]$,

$$\frac{1}{T}\int_0^T \left\{\widehat{\lambda}_j(t) - \lambda_j(t)\right\}^2 \mathrm{d}t \leq 32\left\{C_1(s+1)^2 m_1^{-2d} + 9s\lambda_{\max}\frac{\log p}{T}\right\}, \quad (16)$$

holds with probability at least $1 - C_3 p^{-2} - C_4 p^2 T \exp(-C_5 T^{1/5})$, where C_2, C_3, C_4, and C_5 are positive constants, and C_1 is as defined in (15).

The error bound on the right hand side of (16) consists of two terms. The first term comes from the B-spline basis approximation error (i.e., bias from approximating the nonparametric background and transfer functions using basis functions) and the second term comes from the statistical error (i.e., stochastic error in estimating the intensity functions). It is seen that when $sT/\log p = o(m_1^{2d})$, the bias term $C_1(s+1)^2 m_1^{-2d}$ would become negligible when compared to the statistical error term. When, for example, $s = \mathcal{O}(1)$, $d = 2$ and $m_1 \asymp T^{1/5}$, the error bound in (16) reduces to $\mathcal{O}(T^{-4/5} + \log p/T)$, which is comparable with the estimation error in sparse additive regressions (Raskutti, Wainwright, and Yu 2012).

Two key ingredients in the proof of Theorem 3 are establishing an upper and lower bounded eigenvalue condition for $\mathbf{G}^{(k,k)}$ (see Lemma S10) employed in the standardized group lasso penalty in (8) and a restricted eigenvalue condition for $\mathbf{G}$ under the group lasso setting (see Lemma S11). Establishing these two conditions under the proposed nonstationary process is nontrivial; it requires a delicate analysis that combines properties of the basis functions and concentration inequalities of first and second order statistics of the proposed process. Combining these two ingredients and a martingale central limit theorem for counting processes (van de Geer 1995), we are able to derive the result in Theorem 3. We note that if the basis approximation error condition in Assumption 5 is not satisfied, we may replace the first term in the error bound (16), that is, $C_1(s+1)^2 m_1^{-2d}$ with

$$R_{m_0,m_1} = \min_{\widetilde{\boldsymbol{\beta}}_j \in \mathbb{R}^{m_0+pm_1}} \frac{1}{T}\int_0^T \left\{\boldsymbol{\Psi}^\top(t)\widetilde{\boldsymbol{\beta}}_j - \lambda_j(t)\right\}^2 \mathrm{d}t$$

and Theorem 3 holds with an error bound of $32(R_{m_0,m_1} + 9s\lambda_{\max}\log p/T)$.

4.3. Network Structure Recovery

In this section, we show that, under certain regularity conditions, our proposed method can consistently identify the true edges in the network with probability tending to one. Recalling $\boldsymbol{\Psi}_k(t) = \int_0^t \boldsymbol{\phi}_1(t-u)\mathrm{d}N_k(u)$, $k \in [p]$, we introduce two assumptions.

Assumption 6. For all $j \in [p]$, we assume that

$$\max_{k \notin \mathcal{E}_j} \left\| \left\{\mathbb{E}\int_0^T \boldsymbol{\Phi}_k(t)\boldsymbol{\Phi}^\top_{\mathcal{E}_j}(t)\mathrm{d}t\right\}\left\{\mathbb{E}\int_0^T \boldsymbol{\Phi}_{\mathcal{E}_j}(t)\boldsymbol{\Phi}^\top_{\mathcal{E}_j}(t)\mathrm{d}t\right\}^{-1}\right\|_2 \leq \frac{\gamma_{\min}}{6\sqrt{s}\gamma_{\max}},$$

where $\boldsymbol{\Phi}_k(t) = \int_0^t \boldsymbol{\phi}_1(t-s)\left\{\mathrm{d}N_k(s) - \bar{\lambda}_k(s)\mathrm{d}s\right\}$, $\boldsymbol{\Psi}_{\mathcal{E}_j}(t) \in \mathbb{R}^{m_1|\mathcal{E}_j|}$ is the concatenation of vectors $\{\boldsymbol{\Psi}_k(t) : k \in \mathcal{E}_j\}$, and $\gamma_{\min}$ and $\gamma_{\max}$ are constants as defined in Lemma S10.

This is the irrepresentable condition (Zhao and Yu 2006) under our setting and it is a condition on covariances between the component processes. Considering the jth component process, this condition stipulates that the $\boldsymbol{\Phi}_k(t)$ from non neighbors of j (i.e., $k \notin \mathcal{E}_j$) has small covariances with $\boldsymbol{\Phi}_k(t)$ from neighbors of j (i.e., $k \in \mathcal{E}_j$). A trivial sufficient condition is if the covariance function $C^0_{k_1,k_2}(u_1,u_2) = 0$ for $k_1 \notin \mathcal{E}_j$ and $k_2 \in \mathcal{E}_j$. More generally, Assumption 6 is satisfied if the covariance $|C^0_{k_1,k_2}(u_1,u_2)| \leq c_0/s$ for $k_1 \notin \mathcal{E}_j$, $k_2 \in \mathcal{E}_j$ and some constant $c_0 > 0$; see a detailed proof in Section A11. The condition can be further relaxed if the adaptive lasso (Zou 2006; Huang, Horowitz, and Wei 2010) penalty term is considered and we plan to investigate this extension in our future work. The next condition is a minimal signal condition.

Assumption 7. There exists a constant $\beta_{\min} > 0$ such that $\|\widetilde{\boldsymbol{\beta}}_{j,k}\|_2 \geq \beta_{\min}$ for $k \in \mathcal{E}_j$, where $\widetilde{\boldsymbol{\beta}}_j$ is as defined in Assumption 5.

Note that this condition is not placed on $\widetilde{\boldsymbol{\beta}}_{j,0}$ since $\boldsymbol{\beta}_{j,0}$ is not included in the penalty term. This minimal signal condition can be relaxed; see discussions in Section 7.

Theorem 4. Consider a Hawkes process on $[0, T]$ with intensity as defined in (1) satisfying Assumptions 1–7. Assume that $\eta_j =$

$(C_2 \log p/T)^{1/2}$, $s^2T/\log p = \mathcal{O}(m_1^{2d})$, $s = \mathcal{O}(T^{1/5})$, $\log p = \mathcal{O}(T^{1/5})$, $s^2 m_1 = o(T^{4/5})$. It holds that, for $j \in [p]$,

$$\widehat{\mathcal{E}}_j = \mathcal{E}_j,$$

with probability at least $1 - 2C_3 p^{-2} - 3C_4 p^2 T \exp(-C_5 T^{1/5})$, where C_2, C_3, C_4, and C_5 are as in Theorem 3.

This result establishes selection consistency. The condition $s^2T/\log p = \mathcal{O}(m_1^{2d})$ is needed in selection to ensure the bias term $\mathcal{O}(s^2 m_1^{-2d})$ does not dominate the group-wise estimation error $\mathcal{O}(\log p/T)$. The conditions in Theorem 4 are satisfied when, for example, $s = \mathcal{O}(1)$, $d = 2$, $\log p \asymp T^{1/5}$ and $m_1 \asymp T^{1/5}$. We note that selection consistency was also studied in Chen et al. (2017), under a stationary Hawkes process setting. In comparison, our result is established under the more flexible nonstationary setting. Moreover, our result significantly relaxes a restrictive condition in Chen et al. (2017). Specifically, Assumption 7 (second equation) in Chen et al. (2017), after some simplification, would require T to be upper bounded. This result on the selection consistency has an important implication in practice, as it ensures that our method can correctly identify the true edges in the latent network.

Next, we investigate the selection consistency of the proposed GIC in (9). We use $\widehat{\mathcal{E}}_j^{\eta_j}$ to denote the estimated $\mathcal{E}_j$ with tuning parameter η_j. Let $\eta_{\max}$ and $\eta_{\min}$ be, respectively, the upper and lower limits of the tuning parameter η_j, where $\eta_{\max}$ can be easily chosen such that $\widehat{\mathcal{E}}_j^{\eta_{\max}}$ is empty and $\eta_{\min}$ can be chosen such that $\widehat{\mathcal{E}}_j^{\eta_{\min}}$ is sparse, and the corresponding model size $s_0 = |\widehat{\mathcal{E}}_j^{\eta_{\min}}|$ satisfies conditions in Theorem 5. We partition the interval $[\eta_{\min}, \eta_{\max}]$ into two subsets

$$\Gamma_- = \left\{\eta_j \in [\eta_{\min}, \eta_{\max}] : \widehat{\mathcal{E}}_j^{\eta_j} \not\supset \mathcal{E}_j\right\},$$
$$\Gamma_+ = \left\{\eta_j \in [\eta_{\min}, \eta_{\max}] : \widehat{\mathcal{E}}_j^{\eta_j} \supset \mathcal{E}_j \text{ and } \widehat{\mathcal{E}}_j^{\eta_j} \neq \mathcal{E}_j\right\},$$

corresponding to η_j's that result in under-fitted and over-fitted models, respectively. The next result states that the proposed GIC is consistent in model selection.

Theorem 5. Consider a Hawkes process on $[0, T]$ with intensity as defined in (1) satisfying Assumptions 1–5 and 7 and that $s^2T/\log p = \mathcal{O}(m_1^{2d})$, $s = \mathcal{O}(T^{1/5})$, $\log p = \mathcal{O}(T^{1/5})$ and $m_1 = \mathcal{O}(T^{2/5})$. Assume that there exists $\eta_j^* \in [\eta_{\min}, \eta_{\max}]$ such that $\widehat{\mathcal{E}}_j^{\eta_j^*} = \mathcal{E}_j$. Consider the GIC function defined in (9). When $s_0 = |\widehat{\mathcal{E}}_j^{\eta_{\min}}| = o(T^{2/5})$, $s = o(s_0)$, $sm_1\alpha_T/T = o(1)$ and $s \log p/\alpha_T = o(1)$, it holds that

$$\mathbb{P}\left(\inf_{\eta_j \in \Gamma_- \cup \Gamma_+} \mathrm{GIC}(\eta_j) - \mathrm{GIC}(\eta_j^*) > 0\right) \to 1.$$

The assumption that there exists $\eta_j^* \in [\eta_{\min}, \eta_{\max}]$ such that $\widehat{\mathcal{E}}_j^{\eta_j^*} = \mathcal{E}_j$ is satisfied, for example, by the result in Theorem 4, which would additionally require Assumption 6. This is true by noting $|\widehat{\mathcal{E}}_j^{\eta_{\max}}| = 0$, $|\widehat{\mathcal{E}}_j^{\eta_{\min}}| = s_0$, $s = o(s_0)$ and the size of the selected model decreases as η_j increases (Zhang, Li, and Tsai 2010). The main challenge in establishing Theorem 5 is the large number of candidate models in the over-fitted case, which increases combinatorially fast with p. To overcome this challenge, we introduce a proxy criterion on a support of size $s_0 < p$ (Zhang, Li, and Tsai 2010); see proof details in Section A7. The two conditions on α_T specify a range that ensures consistency. Specifically, $s \log p/\alpha_T = o(1)$ suggests that α_T should diverge adequately fast such that the true model is not dominated by over-fitted models. On the other hand, $sm_1\alpha_T/T = o(1)$ restricts the rate of divergence for α_T based on the size of the true model s and observation window length T. If we take, for example, $\alpha_T \asymp (\log p)^2 \log T$, both conditions on α_T are met. While other choices of α_T can also satisfy both conditions, we have chosen a uniform choice $\alpha_T \asymp (\log p)^2 \log T$ in our empirical investigations. Moreover, we take $s_0 \asymp T^{2/5}/\log\log T$ and choose $\eta_{\max}$ such that that $|\widehat{\mathcal{E}}_j^{\eta_{\max}}| = 0$ and $\eta_{\min}$ such that $|\widehat{\mathcal{E}}_j^{\eta_{\min}}| = s_0$.

4.4. Test of Background Intensity

In this section, we consider the problem of testing if the background intensities of the proposed Hawkes process are constant in time. If the background intensities $\nu_j(t)$'s in (1) are constant, under Assumption 1, a stationary process whose intensity follows (1) exists (Brémaud and Massoulié 1996).

In our model, the background intensity $\nu_j(t)$ for the jth process is represented as $\boldsymbol{\phi}_0(t)\boldsymbol{\beta}_{j,0}$. Without loss of generality, let the first term in the basis $\boldsymbol{\phi}_0(t)$, that is, $\phi_{01}(t)$, be the constant term. Testing if $\nu_j(t)$ is constant in time can then be formulated as testing the following hypotheses:

$$\mathrm{H}_0 : \mathbf{A}\boldsymbol{\beta}_{j,0} = \mathbf{0} \qquad \text{versus} \qquad \mathrm{H}_1 : \mathbf{A}\boldsymbol{\beta}_{j,0} \neq \mathbf{0} \tag{17}$$

where $\mathbf{A} = \begin{bmatrix} 0 & \mathbf{0}_{m_0-1} \\ \mathbf{0}_{m_0-1}^\top & \mathbf{I}_{(m_0-1)\times(m_0-1)} \end{bmatrix} \in \mathbb{R}^{m_0 \times m_0}$. The test in (17) can detect any fixed departure in $\nu_j(t)$ from a constant provided that m_0 is sufficiently large (Fan, Zhang, and Zhang 2001). Recall that $\widehat{\boldsymbol{\beta}}_j$ is obtained from

$$\widehat{\boldsymbol{\beta}}_j = \arg\min_{\boldsymbol{\beta}_j \in \mathbb{R}^{m_0+m_1 p}} \left\{\ell_j(\boldsymbol{\beta}_j) + \eta_j \sum_{k=1}^{p} \left(\boldsymbol{\beta}_{j,k}^\top \mathbf{G}^{(k,k)} \boldsymbol{\beta}_{j,k}\right)^{1/2}\right\}, \tag{18}$$

where $\ell_j(\boldsymbol{\beta}_j)$ is defined as in (7). Next, letting $\mathrm{supp}_1(\boldsymbol{\beta}_j) = \{k \in [p] : \boldsymbol{\beta}_{j,k} \neq \mathbf{0}\}$, and define the refitted estimator $\widehat{\boldsymbol{\beta}}_j^1$ and the restricted estimator $\widehat{\boldsymbol{\beta}}_j^{H_0}$ under H_0 as

$$\widehat{\boldsymbol{\beta}}_j^1 = \arg \min_{\substack{\mathbf{b}_j \in \mathbb{R}^{m_0+m_1 p} \\ \mathrm{supp}_1(\mathbf{b}_j) = \mathrm{supp}_1(\widehat{\boldsymbol{\beta}}_j)}} \ell_j(\mathbf{b}_j),$$
$$\widehat{\boldsymbol{\beta}}_j^{H_0} = \arg \min_{\substack{\mathbf{b}_j \in \mathbb{R}^{m_0+m_1 p} : \mathbf{A}\mathbf{b}_{j,0} = \mathbf{0} \\ \mathrm{supp}_1(\mathbf{b}_j) = \mathrm{supp}_1(\widehat{\boldsymbol{\beta}}_j)}} \ell_j(\mathbf{b}_j). \tag{19}$$

Note that the number of basis functions used in (18) and (19) may not be the same; see discussion after Theorem 6. Finally, we define the test statistic as

$$S_j = T\left\{\ell_j(\widehat{\boldsymbol{\beta}}_j^{H_0}) - \ell_j(\widehat{\boldsymbol{\beta}}_j^1)\right\},$$

where $\ell_j(\boldsymbol{\beta}_j)$ is defined as in (7). The following theorem states the asymptotic null distribution of the test statistic.

Theorem 6. Assume that all conditions in Theorem 4 are satisfied and the m_1 used in (19) satisfy $s^3T = o(m_1^{2d+1})$. Under H_0, we have that

$$S_j/\bar{\lambda}_j \xrightarrow{\mathcal{D}} \chi^2_{m_0-1},$$

where $\bar{\lambda}_j = \mathbb{E}\{dN_j(t)\}/dt$ is a constant under H_0.

Theorem 6 is derived for a two-step procedure, where the first step involves a regularized estimation in (18) and requires appropriate regularization to achieve selection consistency and the second step involves refitting based on the selected set of edges from step 1. The selection consistency in step 1 is ensured by Theorems 4–5, and the tuning parameter η_j used to calculate $\boldsymbol{\beta}_j$ is selected following the proposed GIC in (9). In step 2, the condition $s^3T = o(m_1^{2d+1})$ is needed such that the bias from approximating the background and transfer functions using basis functions is asymptotically negligible relative to variance of the test statistic. Such a condition is usually referred to as under-smoothing (Chen 2007) and is common in nonparametric regression testing problems. For instance, if $m_1 \asymp T^{1/5}$ is used in (18) (see discussion of Theorem 4), the under-smoothing condition $s^3T = o(m_1^{2d+1})$ will be satisfied if we multiply m_1 by, for example, a factor of $T^{1/20}$. We do not assume m_0 is fixed when estimating $\widehat{\beta}_j^1$, that is, m_0 may increase with T. As m_0 increases, we may alternatively write the limiting distribution as $(S_j/\bar{\lambda}_j - m_0 + 1)/(2m_0 - 2)^{1/2} \to^d \mathcal{N}(0,1)$, which does not depend on m_0. In practice, given the m_0 and m_1 used in network structure estimation (see Section B1), we multiply both of them by an under-smoothing factor (e.g., $T^{1/20}$), which results in the under-smoothed m_0 and m_1 used in the testing procedure.

Based on the result in Theorem 6, we would reject the null $H_0 : \mathbf{A}\boldsymbol{\beta}_{j,0} = \mathbf{0}$ if $S_j/\bar{\lambda}_j \geq z_{1-\alpha}$, where z_α is the αth quantile of $\chi^2_{m_0-1}$ and $\bar{\lambda}_j$ is estimated with $\bar{\lambda}_j = \frac{1}{T}\int_0^T \widehat{\lambda}_j(t)dt$. As the limiting distribution in Theorem 6 is derived under perfect model recovery, ensured by Assumptions 6–7, cautions should be exercised when performing this test, as the testing procedure may be invalid when perfect model recovery does not hold. We plan to relax this assumption on perfect model recovery in our future research; see discussions in Section 7.

Next, we discuss the asymptotic power of our proposed test. The following theorem provides a lower bound on the growth rate of the test statistic under alternatives.

Proposition 1. Assume that all conditions in Theorem 4 are satisfied. For any alternative H_1 such that $\|\mathbf{A}\widetilde{\boldsymbol{\beta}}_{j,0}\|_2/(s^2 m_1 \log p/T)^{1/2} \to \infty$, we have

$$\mathbb{P}(S_j > M_1 s \log p) \to 1,$$

for some constant $M_1 > 0$.

This result shows that the growth rate of S_j under the alternative is at least $s \log p$, while p diverges. The asymptotic null distribution in Theorem 6 and the growth rate under the alternative together suggest that the null and the alternative hypotheses are well separated, and our proposed test is asymptotically powerful against alternatives. Moreover, the test is locally powerful, that is, $\|\mathbf{A}\widetilde{\boldsymbol{\beta}}_{j,0}\|_2$ is allowed to tend to 0 as long as it decreases no faster than the rate of $(s^2 m_1 \log p/T)^{1/2}$.

Compared with the inferential result in Chen and Hall (2013) for the maximum likelihood estimator of univariate Hawkes process, Theorem 6 is derived for least squares based estimation. Hence, the specific form of the martingale, the characterization of its large jumps and the asymptotic variance are different when applying the martingale central limit theorem. Moreover, as S_j calculates the difference between two least squares losses, a careful treatment was needed to derive its asymptotic expansion; see Step 1 in the proof. Finally, Chen and Hall (2013) consider parametric forms of the background intensity and transfer function while we consider a nonparametric estimation using B-splines. Consequently, our analysis is challenged by a bias term of $\int_0^T \boldsymbol{\Psi}_{\bar{\mathcal{E}}_j}(t)\left\{\widetilde{\lambda}_j(t) - \lambda_j(t)\right\}dt$ in the test statistic, which is bounded by combining properties of the B-spline basis and results in Lemmas S2 and S11.

5. Simulation Studies

In this section, we carry out simulation studies to investigate the finite sample performance of our proposed method, and to compare with existing solutions. We consider three simulation settings. In Simulation 1, we investigate the estimation accuracy of our proposed method; in Simulation 2, we investigate the network edge selection accuracy; in Simulation 3, we evaluate the size and power of our proposed test of hypothesis. Due to space limitations, results from Simulation 3 is relegated to Section B3 of the supplement. We refer to our proposed method for nonstationary Hawkes processes as `NStaHawkes`. In Simulations 1 and 2, we compare our method with Chen et al. (2017), referred to as `StaHawkes`, which was proposed for stationary Hawkes processes. We also compare with a binning based approach in Zhang et al. (2016) on selection accuracy, referred to as `BinGLM`.

In all simulations, we use the criterion in (9) with $\alpha_T = (\log p)^2 \log T/2$ to select the tuning parameter for `NStaHawkes`. The tuning parameters in `StaHawkes` and `BinGLM` are selected using the BIC-type functions recommended, respectively, in Chen et al. (2017) and Zhang et al. (2016).

Simulation 1

In this simulation, we consider two different network settings. The first setting considers the network in Figure 1(a), where

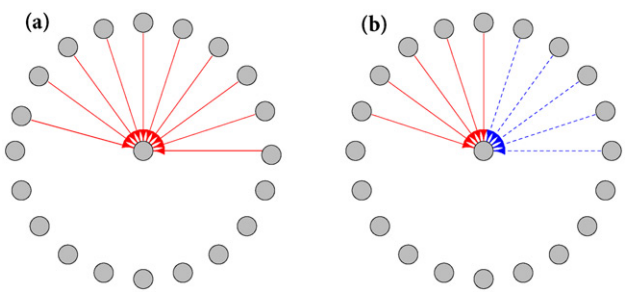

Figure 1. Directed network structures in Simulation 1 with (a) considered in Setting 1.1 and (b) considered in Setting 1.2. Red (solid) edges represent excitatory effects and blue (dashed) edges represent inhibitory effects.

all transfer functions are positive, corresponding to excitatory effects. The second setting considers the network in Figure 1(b), with both positive and negative transfer functions, corresponding to excitatory and inhibitory effects, respectively. Let the network edge set be $\mathcal{E} = \{(k,1), k = 2,\ldots,11\}$. The background intensity functions and transfer functions for each setting are as follows:

Setting 1.1:

$$\begin{aligned}
&\nu_1(t) = 60 + 50 \times \sin(2\pi t/T),\\
&\nu_j(t) = \alpha_j + \alpha_j \times \sin(2\pi t/T), \quad j = 2,\ldots,21,\\
&\omega_{1,k} = 20000(x + 0.001)\exp(1 - 500x), \quad k = 2,\ldots,11,
\end{aligned}$$

Setting 1.2:

$$\begin{aligned}
&\nu_1(t) = 60 + 50 \times \sin(2\pi t/T),\\
&\nu_j(t) = \alpha_j + \alpha_j \times \sin(2\pi t/T), \quad j = 2,\ldots,21,\\
&\omega_{1,k} = 20000(x + 0.001)\exp(1 - 500x), \quad k = 2,\ldots,6,\\
&\omega_{1,k} = -15000(x + 0.001)\exp(1 - 500x), \quad k = 7,\ldots,11,
\end{aligned}$$

where α_j is generated from $N(30, 5^2)$. We let the supports of all transfer functions be $[0, 0.01]$, and simulate events in $[0, T]$ with the intensity function (1) under Settings 1.1–1.2. To estimate the background intensities and transfer functions, we use cubic B-splines with equally spaced knots. To select the numbers of B-splines m_0 and m_1, we first perform selection using the proposed procedure in Section B1 over 20 data replications. The numbers of B-splines are then fixed at the respective averages of the 20 selected values for m_0 and m_1. It is worth noting that the estimation and selection accuracy are not overly sensitive to the number of B-splines used in the estimation (see additional results in Section B2). We have also considered larger ranges for the transfer functions and the results are very similar. We thus focus on the current setting when reporting our simulation results.

To evaluate the estimation accuracy, we report the mean squared errors. For the background intensity, it is calculated as $\mathrm{MSE}(\nu) = \frac{1}{p}\sum_{j\in[p]}\mathrm{MSE}(\nu_j)$, where

$$\mathrm{MSE}(\nu_j) = \left\{\frac{1}{T}\int_0^T (\widehat{\nu}_j(t) - \nu_j(t))^2 \mathrm{d}t\right\}^{1/2}$$

and $\widehat{\nu}_j(t)$ is the estimate of $\nu_j(t)$. For the transfer functions, it is calculated as $\mathrm{MSE}(\omega) = \frac{1}{p}\sum_{j\in[p]}\mathrm{MSE}(\omega_{j,\cdot})$, where

$$\mathrm{MSE}(\omega_{j,\cdot}) = \left\{\sum_{k=1}^{p}\int_0^b (\widehat{\omega}_{j,k}(t) - \omega_{j,k}(t))^2 \mathrm{d}t\right\}^{1/2}$$

and $\widehat{\omega}_{j,k}(t)$ is the estimate of $\omega_{j,k}(t)$. To evaluate the selection accuracy, we report the false positive rate (FNR), the false positive rate (FPR) and the F_1 score (Forman 2003; Ho, Parikh, and Xing 2012), calculated as 2TP/(2TP+FP+FN), where TP is the true positive count, FP is the false positive count, and FN is the false negative count. The highest F_1 score is 1 indicating perfect selection. For both `NStaHawkes` and `StaHawkes` estimators, we report the estimation and selection accuracy. For the `BinGLM` estimator, we only report the selection accuracy, as this method cannot be used to estimate the intensity functions. Table 1 reports the average criteria from the three methods, with standard errors in the parentheses, over 100 data replications. The proposed method `NStaHawkes` is seen to achieve the best performance, both in terms of the estimation accuracy and edge selection accuracy, and this holds true for different observation window length T. Moreover, it is seen that the estimation error of `NStaHawkes` decreases as T increases. Such an observation agrees with our theoretical result in Theorem 3.

Simulation 2

In this simulation, we evaluate the edge selection accuracy of our proposed method. We consider two types of networks. The first type of networks are assumed to follow an Erdos–Renyi network model with edge probability p_e (Erdős and Renyi 1969). In an Erdos–Renyi network model, edges are generated independently from a Bernoulli distribution with probability p_e. The second type of networks follow a scale-free network

Table 1. Comparison of the three methods with varying observation window length T in Simulation 1.

Setting 1											
T	Method	MSE(ν)		MSE(ω)		FNR		FPR		F_1 score	
10	`NStaHawkes`	7.921	(0.071)	0.332	(0.003)	0.008	(0.003)	0.011	(0.001)	0.816	(0.007)
	`StaHawkes`	25.576	(0.073)	1.363	(0.026)	0.000	(0.000)	0.305	(0.008)	0.139	(0.003)
	`BinGLM`	–		–		0.003	(0.002)	0.058	(0.001)	0.443	(0.002)
20	`NStaHawkes`	7.389	(0.038)	0.279	(0.003)	0.002	(0.001)	0.005	(0.000)	0.908	(0.006)
	`StaHawkes`	24.574	(0.034)	0.899	(0.010)	0.000	(0.000)	0.321	(0.005)	0.129	(0.002)
	`BinGLM`	–		–		0.000	(0.000)	0.058	(0.001)	0.447	(0.002)
Setting 2											
T	Method	MSE(ν)		MSE(ω)		FNR		FPR		F_1 score	
10	`NStaHawkes`	6.111	(0.038)	0.279	(0.003)	0.113	(0.012)	0.005	(0.000)	0.835	(0.007)
	`StaHawkes`	25.549	(0.083)	1.297	(0.028)	0.042	(0.012)	0.305	(0.009)	0.135	(0.004)
	`BinGLM`	–		–		0.342	(0.019)	0.041	(0.001)	0.376	(0.007)
20	`NStaHawkes`	5.545	(0.026)	0.252	(0.002)	0.094	(0.011)	0.001	(0.000)	0.924	(0.006)
	`StaHawkes`	24.648	(0.047)	0.938	(0.014)	0.414	(0.010)	0.316	(0.007)	0.080	(0.002)
	`BinGLM`	–		–		0.359	(0.018)	0.094	(0.004)	0.236	(0.007)

NOTE: `NStaHawkes` refers to the proposed method, `StaHawkes` refers to Chen et al. (2017) and `BinGLM` refers to Zhang et al. (2016). Standard errors are shown in parentheses.

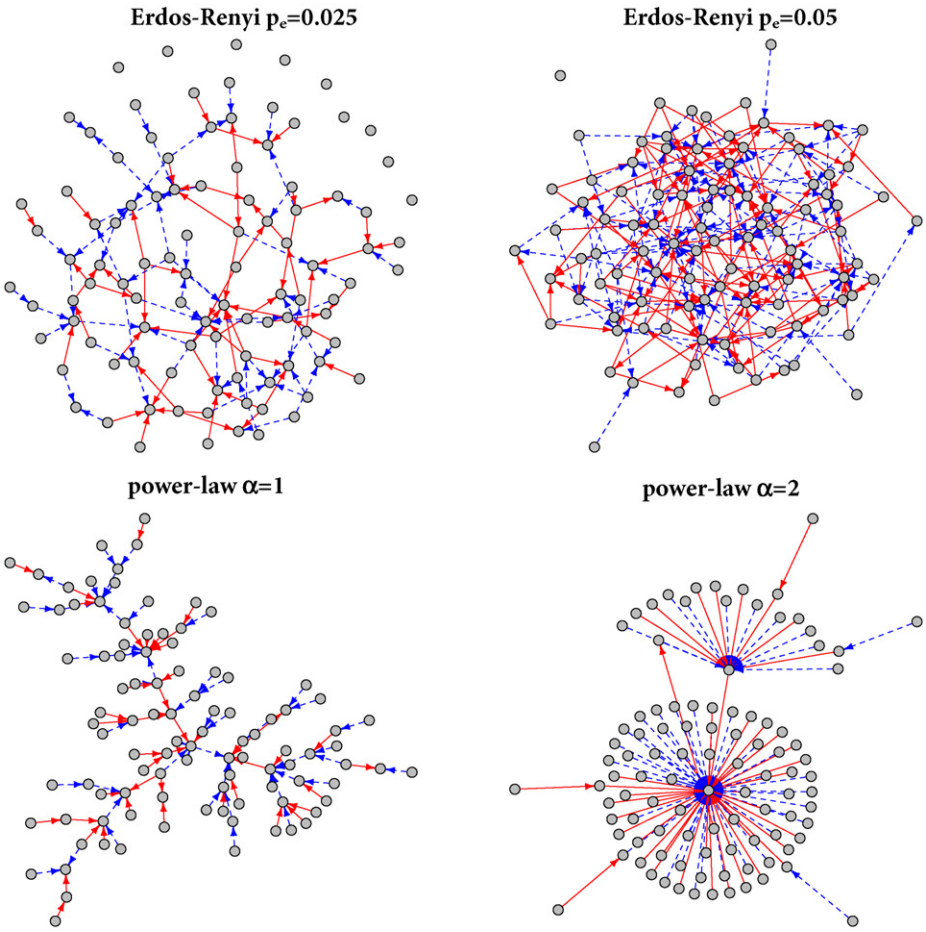

Figure 2. Directed network structures in Simulation 2. Red (solid) edges represent excitatory effects and blue (dashed) edges represent inhibitory effects.

model, with the degrees of nodes generated from a power-law distribution with parameter α; such networks have a skewed degree distributions and a larger α indicates a higher degree heterogeneity (Clauset, Shalizi, and Newman 2009). We set $p = 100$, $p_e = 0.025, 0.05$ and $\alpha = 1, 2$. The generated networks are shown in Figure 2. Based on the generated networks, we simulate data using the following setting.

Setting 2: $\nu_j(t) = \alpha_j + \alpha_j \times \sin(2\pi ft/T), j = 1, \ldots, 100,$

where α_j is generated independently from $N(100, 5^2)$ for each node. The transfer functions are the same as in Setting 1.2. We simulate events in $[0, T]$ with intensity function (1) with $f = 5$ and $T = 20$. To estimate the background intensities, we use cubic B-splines with equally spaced knots and to estimate the transfer functions, we use step functions with equally spaced knots, as considered in Hansen, Reynaud-Bouret, and Rivoirard (2015) and Chen et al. (2017). The numbers of basis functions m_0 and m_1 are selected following the same procedure as in Simulation 1. Table 2 compares the false negative rate, false positive rate and F_1 score of the three methods over 100 data replications. It is seen that `NStaHawkes` achieves the best edge selection accuracy, in terms of F_1 scores, across all settings; `StaHawkes` shows a large false positive rate and this is likely due to the biased estimation of the background intensity functions; `BinGLM` shows a large false negative rate and this is possibly due to the loss of information in the binning approach. We have also considered $p = 200$, where `NStaHawkes` continues to achieve a satisfactory edge selection accuracy (see Section B3).

6. Application to Neurophysiological Data

In this section, we apply our proposed method to a neuron spike train dataset and estimate the functional connectivity network of neurons in the rat prefrontal cortex. The data were obtained from adult male Sprague-Dawley rats performing a T-maze based delayed-alternation task of working memory (Devilbiss and Waterhouse 2004). In the experiment, the animal was trained to navigate down the T-maze and choose one of two arms (opposite to the one previously visited) for food rewards. In each trial, the animal was released after being placed in a start box for a fixed length of delay. On a correct trial (i.e., the arm with food was chosen), the animal was rewarded and returned to the start box. On an incorrect trial (i.e., the arm without food was chosen), the animal was returned to the start box without being rewarded. In the study, the animal remained in a training period until it reached 90%–100% accuracy on 40 trials. After the training period, a recording session was performed. The spike train recording consisted of 73 neurons in an experiment of 40 trials. Each trials took about 36 sec and the total recording had 1434.22 sec. See Zhang et al. (2016) for more information about data collection and processing.

Table 2. Comparison of the three methods with varying network parameters in Simulation 2.

	Erdős–Renyi network					
	$p_e = 0.025$			$p_e = 0.05$		
	FNR	FPR	F_1 score	FNR	FPR	F_1 score
NStaHawkes	0.003 (0.001)	0.002 (0.000)	0.935 (0.002)	0.005 (0.000)	0.005 (0.000)	0.909 (0.001)
StaHawkes	0.000 (0.000)	0.859 (0.001)	0.027 (0.000)	0.000 (0.000)	0.882 (0.001)	0.053 (0.000)
BinGLM	0.394 (0.003)	0.133 (0.001)	0.097 (0.001)	0.369 (0.002)	0.171 (0.001)	0.148 (0.000)
	power-law network					
	$\alpha = 1$			$\alpha = 2$		
	FNR	FPR	F_1 score	FNR	FPR	F_1 score
NStaHawkes	0.005 (0.001)	0.002 (0.000)	0.928 (0.002)	0.003 (0.001)	0.003 (0.000)	0.870 (0.001)
StaHawkes	0.000 (0.000)	0.848 (0.001)	0.023 (0.000)	0.002 (0.000)	0.824 (0.001)	0.024 (0.000)
BinGLM	0.379 (0.003)	0.121 (0.000)	0.090 (0.000)	0.787 (0.004)	0.060 (0.000)	0.059 (0.001)

NOTE: NStaHawkes refers to the proposed method, StaHawkes refers to Chen et al. (2017) and BinGLM refers to Zhang et al. (2016). Standard errors are shown in parentheses.

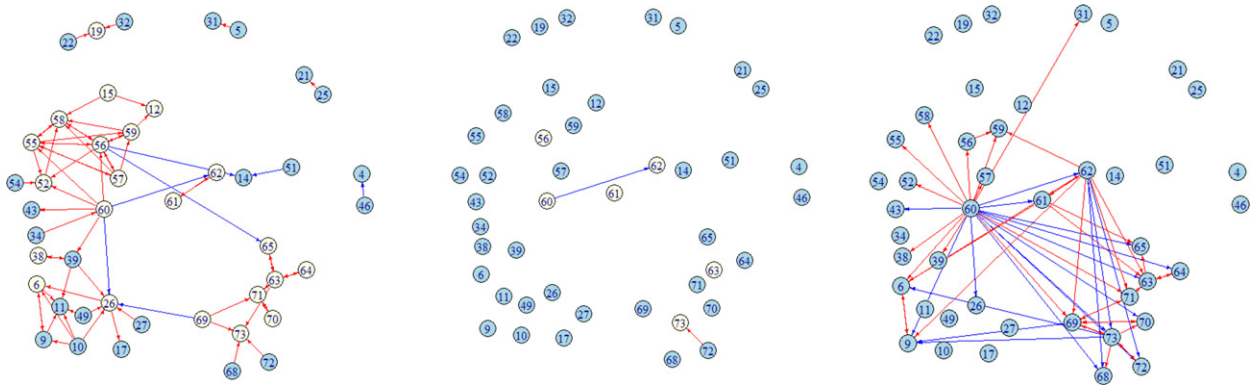

Figure 3. Estimated neuronal networks using NStaHawkes (left), StaHawkes (middle) and BinGLM (right). The red and blue arrows represent excitatory and inhibitory effects, respectively. Light colored nodes represent neurons that have self-exciting effects.

We applied our proposed method to this dataset. To estimate the background intensities, we used cubic B-splines with equally spaced knots and to estimate the transfer functions, we used step functions with equally spaced knots, as considered in Hansen, Reynaud-Bouret, and Rivoirard (2015) and Chen et al. (2017). The numbers of basis functions m_0 and m_1 were selected using the proposed procedure in Section B1. The GIC in (9) was used to select the tuning parameters with $\alpha_T = (\log p)^2 \log T/2$. The range of the transfer functions were set to $[0, 2]$. We have also considered a larger range, and the results remain very similar. First, we performed the proposed test of hypothesis for each neuron to assess the if the background intensity is constant in time. Based on the p-values from the tests, 38 neurons had time-varying background intensity functions (significance level was set to 0.05). Next, we move to estimate the neuronal connectivity network using the proposed NStaHawkes. When estimating the network structure, we also considered StaHawkes and BinGLM. The tuning parameters in StaHawkes and BinGLM were selected using their recommended BIC functions, respectively. Figure 3 shows the estimated neuronal networks from the three different methods. We can see all three estimated networks are sparse, with both excitatory and inhibitory relationships. However, their structures are quite different. The network estimated from our method is highly clustered and has a power-law degree distribution, which are two unique features of real world networks (Barabási and Albert 1999). Also interestingly, about 70% of the identified edges in our estimated network are within the right prefrontal cortex, which agrees with existing findings that the right prefrontal cortex is highly related to the episodic memory retrieval (Henson, Shallice, and Dolan 1999). The biological significance of the identified edges requires further investigation.

Compared to our estimated network, StaHawkes identified a very sparse network. This difference is likely due to the bias in estimating the background intensity function from their method. BinGLM also identified a very different network structure. This network has two hub (or densely connected) nodes, namely, neurons 60 and 62, and a small clustering coefficient. We find that neurons 60 and 62 are the two most frequently fired neurons in the ensemble. Specifically, neurons 60 and 62 have

14,433 and 8191 firing events, respectively, while other neurons have on average 501 firing events during the experiment. The regularized generalized linear model framework in `BinGLM` penalizes the frequently and infrequently firing neurons equally when encouraging sparsity. This can potentially lead to over selection for the frequently firing neurons, and under selection for the infrequently firing neurons.

7. Discussion

As summarized below, Theorems 3–6 each require a set of rate conditions on s, p and m_1.

Theorem 3	$s = o(T^{2/5})$, $\log p = \mathcal{O}(T^{1/5})$, $sm_1 = \mathcal{O}(T^{4/5})$
Theorem 4	$s = \mathcal{O}(T^{1/5})$, $\log p = \mathcal{O}(T^{1/5})$, $s^2 m_1 = o(T^{4/5})$, $s^2 m_1^{-2d} = \mathcal{O}(\log p/T)$
Theorem 5	$s = \mathcal{O}(T^{1/5})$, $\log p = \mathcal{O}(T^{1/5})$, $m_1 = \mathcal{O}(T^{2/5})$, $s^2 m_1^{-2d} = \mathcal{O}(\log p/T)$
Theorem 6	same as Theorem 4, $s^3 T = o(m_1^{2d+1})$

In Theorem 4, the requirement $s^2 m_1 = o(T^{4/5})$ is needed to characterize the estimation error of $\boldsymbol{\beta}_j$ in $\ell_{2,\infty}$ norm, that is, $\max_{k\in\mathcal{E}_j} \|\widehat{\boldsymbol{\beta}}_{j,k} - \widetilde{\boldsymbol{\beta}}_{j,k}\|_2^2$. In Theorem 5, the requirement $m_1 = \mathcal{O}(T^{2/5})$ is needed to ensure GIC selection consistency in the under-fitted case. In Theorems 4–5, the additional requirement $s^2 m_1^{-2d} = \mathcal{O}(\log p/T)$ is to ensure that the B-spline approximation error does not dominate the estimation error. In Theorem 6, the condition $s^3 T = o(m_1^{2d+1})$ is required such that the approximation error is negligible relative to the variance of the test statistic. For example, when the smoothness parameter $d = 2$ and $s = \mathcal{O}(1)$, Theorems 3–5 are satisfied when $\log p \asymp T^{1/5}$ and $m_1 \asymp T^{1/5}$, and Theorem 6 is satisfied when $m_1 \asymp T^{1/4}$. We remark that the minimal signal condition in Assumption 7 can be relaxed with more stringent assumptions on s or $\log p$ in Theorems 4–5 to control the estimation error and with modified conditions on α_T in Theorem 5 to control the penalty strength in GIC.

We conclude the article with a brief discussion on some potential future directions. We hypothesize that the bounded intensity condition in Assumption 2 can be relaxed at the cost of an additional $\log T$ term in the estimation error and Assumption 5 can be verified under nonlinear link functions. The limiting distribution result in Theorem 6 requires conditions for establishing selection consistency (i.e., irrepresentable condition in Assumption 6 and minimal signal condition in Assumption 7). To derive a valid inferential procedure that does not reply on such conditions, we could consider the decorrelated score testing procedure in Neykov et al. (2018); Wang, Kolar, and Shojaie (2020) or the double selection procedure in Bach et al. (2020). We plan to investigate these directions in our future work. In our work, we investigated empirically the reliance of the testing procedure on model selection accuracy. In Simulation 3.1, we showed that the size of the proposed test is well controlled; for this setting, the average false negative rate is 0, false positive rate is 0.015. In Simulation 3.2, we showed that the proposed test is powerful against alternatives; for this setting and $\rho = 0.25$, the average false negative rate is 0, false positive rate is 0.016. These results suggest that the proposed testing procedure may not be overly sensitive to errors in model selection.

Supplementary Materials

The supplementary materials contain proofs to all theoretical results, additional simulation results and computational details.

Acknowledgments

The authors are very grateful for the constructive comments from the editor, associate editor, and two anonymous reviewers.

Funding

Zhang's research is supported by NSF DMS-2015190 and Guan's research is supported by NSF DMS-1810591.

References

Amari, S.-i. (1977), "Dynamics of Pattern Formation in Lateral-Inhibition Type Neural Fields," *Biological Cybernetics*, 27, 77–87. [95]

Bach, P., Klaassen, S., Kueck, J., and Spindler, M. (2020), "Uniform Inference in High-Dimensional Generalized Additive Models," arXiv preprint arXiv:2004.01623. [107]

Bacry, E., Bompaire, M., Gaïffas, S., and Muzy, J.-F. (2020), "Sparse and Low-Rank Multivariate Hawkes Processes," *Journal of Machine Learning Research*, 21, 1–32. [97,99]

Bacry, E., Delattre, S., Hoffmann, M., and Muzy, J.-F. (2013), "Modelling Microstructure Noise with Mutually Exciting Point Processes," *Quantitative Finance*, 13, 65–77. [95]

Bacry, E., and Muzy, J.-F. (2016), "First- and Second-order Statistics Characterization of Hawkes Processes and Non-parametric Estimation," *IEEE Transactions on Information Theory*, 62, 2184–2202. [100]

Barabási, A.-L., and Albert, R. (1999), "Emergence of Scaling in Random Networks," *Science*, 286, 509–512. [106]

Brémaud, P., and Massoulié, L. (1996), "Stability of Nonlinear Hawkes Processes," *The Annals of Probability*, 24, 1563–1588. [95,96,97,99,102]

Chen, F., and Hall, P. (2013), "Inference for a Nonstationary Self-Exciting Point Process with an Application in Ultra-High Frequency Financial Data Modeling," *Journal of Applied Probability*, 50, 1006–1024. [95,96,97,103]

Chen, J., and Chen, Z. (2008), "Extended Bayesian Information Criteria for Model Selection with Large Model Spaces," *Biometrika*, 95, 759–771. [99]

Chen, S., Shojaie, A., Shea-Brown, E., and Witten, D. (2017), "The Multivariate Hawkes Process in High Dimensions: Beyond Mutual Excitation," arXiv preprint arXiv:1707.04928. [96,97,100,102,103,104,105,106]

Chen, X. (2007), "Large Sample Sieve Estimation of Semi-nonparametric Models," *Handbook of Econometrics*, 6, 5549–5632. [101,103]

Clauset, A., Shalizi, C. R., and Newman, M. E. (2009), "Power-Law Distributions in Empirical Data," *SIAM Review*, 51, 661–703. [105]

Costa, M., Graham, C., Marsalle, L., and Tran, V. C. (2018), "Renewal in Hawkes Processes with Self-Excitation and Inhibition," arXiv preprint arXiv:1801.04645. [96,99]

Devilbiss, D. M., and Waterhouse, B. D. (2004), "The Effects of Tonic Locus Ceruleus Output on Sensory-Evoked Responses of Ventral Posterior Medial Thalamic and Barrel Field Cortical Neurons in the Awake Rat," *Journal of Neuroscience*, 24, 10773–10785. [105]

Engle, R. F., and Russell, J. R. (1998), "Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data," *Econometrica*, 66, 1127–1162. [95]

Erdős, P., and Renyi, A. (1969), "On Random Graphs. I," *Publicationes Mathematicae*, 6, 290–297. [104]

Fan, J., Zhang, C., and Zhang, J. (2001), "Generalized Likelihood Ratio Statistics and Wilks Phenomenon," *Annals of Statistics*, 29, 153–193. [102]

Farajtabar, M., Wang, Y., Rodriguez, M. G., Li, S., Zha, H., and Song, L. (2015), "Coevolve: A Joint Point Process Model for Information Diffusion and Network Co-evolution," in *Advances in Neural Information Processing Systems*, pp. 1954–1962. [95]

Forman, G. (2003), "An Extensive Empirical Study of Feature Selection Metrics for Text Classification," *Journal of machine learning Research*, 3, 1289–1305. [104]

Guan, Y., Jalilian, A., and Waagepetersen, R. (2013), "Decomposition of Variance for Spatial Cox Processes," *Scandinavian Journal of Statistics*, 40, 119–137. [100]

Hansen, N. R., Reynaud-Bouret, P., and Rivoirard, V. (2015), "Lasso and Probabilistic Inequalities for Multivariate Point Processes," *Bernoulli*, 21, 83–143. [95,97,98,99,105,106]

Hawkes, A. G. (1971), "Spectra of Some Self-Exciting and Mutually Exciting Point Processes," *Biometrika*, 58, 83–90. [95,100]

Hawkes, A. G., and Oakes, D. (1974), "A Cluster Process Representation of a Self-Exciting Process," *Journal of Applied Probability*, 11, 493–503. [95,99]

Henson, R., Shallice, T., and Dolan, R. J. (1999), "Right Prefrontal Cortex and Episodic Memory Retrieval: A Functional MRI Test of the Monitoring Hypothesis," *Brain*, 122, 1367–1381. [106]

Ho, Q., Parikh, A. P., and Xing, E. P. (2012), "A Multiscale Community Blockmodel for Network Exploration," *Journal of the American Statistical Association*, 15, 916–934. [104]

Huang, J., Horowitz, J. L., and Wei, F. (2010), "Variable Selection in Nonparametric Additive Models," *Annals of Statistics*, 38, 2282–2313. [98,101]

Kozbur, D. (2020), "Inference in Additively Separable Models with a High-Dimensional Set of Conditioning Variables," *Journal of Business & Economic Statistics*, 39, 984–1000. [99]

Lemonnier, R., and Vayatis, N. (2014), "Nonparametric Markovian Learning of Triggering Kernels for Mutually Exciting and Mutually Inhibiting Multivariate Hawkes Processes," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, pp. 161–176. [96]

Lewis, E., and Mohler, G. (2011), "A Nonparametric EM Algorithm for Multiscale Hawkes Processes," *Journal of Nonparametric Statistics*, 1, 1–20. [96]

Li, B., and Zhang, H. (2011), "An Approach to Modeling Asymmetric Multivariate Spatial Covariance Structures," *Journal of Multivariate Analysis*, 102, 1445–1453. [100]

Linderman, S., and Adams, R. (2014), "Discovering Latent Network Structure in Point Process Data," in *International Conference on Machine Learning*, pp. 1413–1421. [95]

Massoulié, L. (1998), "Stability Results for a General Class of Interacting Point Processes Dynamics, and Applications," *Stochastic Processes and their Applications*, 75, 1–30. [97,99,100]

Merlevède, F., Peligrad, M., and Rio, E. (2011), "A Bernstein Type Inequality and Moderate Deviations for Weakly Dependent Sequences," *Probability Theory and Related Fields*, 151, 435–474. [100]

Neykov, M., Ning, Y., Liu, J. S., and Liu, H. (2018), "A Unified Theory of Confidence Regions and Testing for High-Dimensional Estimating Equations," *Statistical Science*, 33, 427–443. [107]

Ogata, Y. (1981), "On Lewis' Simulation Method for Point Processes," *IEEE Transactions on Information Theory*, 27, 23–31. [97]

Okatan, M., Wilson, M. A., and Brown, E. N. (2005), "Analyzing Functional Connectivity using a Network Likelihood Model of Ensemble Neural Spiking Activity," *Neural Computation*, 17, 1927–1961. [95]

Raskutti, G., Wainwright, M. J., and Yu, B. (2012), "Minimax-Optimal Rates for Sparse Additive Models Over Kernel Classes via Convex Programming," *The Journal of Machine Learning Research*, 13, 389–427. [101]

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009), "Sparse Additive Models," *Journal of the Royal Statistical Society*, Series B, 71, 1009–1030. [98]

Roueff, F., Von Sachs, R., and Sansonnet, L. (2016), "Locally Stationary Hawkes Processes," *Stochastic Processes and their Applications*, 126, 1710–1743. [96]

Schumaker, L. (2007), *Spline Functions: Basic Theory*, Cambridge: Cambridge University Press. [98]

Schwarz, G. (1978), "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461–464. [99]

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013), "A Sparse-Group Lasso," *Journal of Computational and Graphical Statistics*, 22, 231–245. [98]

Stone, C. J. (1985), "Additive Regression and other Nonparametric Models," *Annals of Statistics*, 13, 689–705. [98,101]

Tsybakov, A. B. (2008), *Introduction to Nonparametric Estimation*, New York: Springer. [101]

van de Geer, S. (1995), "Exponential Inequalities for Martingales, with Application to Maximum Likelihood Estimation for Counting Processes," *Annals of Statistics*, 23, 1779–1801. [101]

——— (2000), *Empirical Processes in M-estimation* (Vol. 6), Cambridge: Cambridge University Press. [97]

Veen, A., and Schoenberg, F. P. (2008), "Estimation of Space–Time Branching Process Models in Seismology using an EM–type Algorithm," *Journal of the American Statistical Association*, 103, 614–624. [97]

Vinci, G., Ventura, V., Smith, M. A., and Kass, R. E. (2016), "Separating Spike Count Correlation from Firing Rate Correlation," *Neural Computation*, 28, 849–881. [96]

——— (2018), "Adjusted Regularization in Latent Graphical Models: Application to Multiple-Neuron Spike Count Data," *The Annals of Applied Statistics*, 12, 1068–1095. [96]

Wang, X., Kolar, M., and Shojaie, A. (2020), "Statistical Inference for Networks of High-Dimensional Point Processes," arXiv preprint arXiv:2007.07448. [97,107]

Zhang, C., Chai, Y., Guo, X., Gao, M., Devilbiss, D., and Zhang, Z. (2016), "Statistical Learning of Neuronal Functional Connectivity," *Technometrics*, 58, 350–359. [96,103,104,105,106]

Zhang, Y., Li, R., and Tsai, C.-L. (2010), "Regularization Parameter Selections via Generalized Information Criterion," *Journal of the American Statistical Association*, 105, 312–323. [102]

Zhao, P., and Yu, B. (2006), "On Model Selection Consistency of Lasso," *Journal of Machine Learning Research*, 7, 2541–2563. [101]

Zhou, K., Zha, H., and Song, L. (2013), "Learning Social Infectivity in Sparse Low-Rank Networks Using Multi-Dimensional Hawkes Processes," in *Artificial Intelligence and Statistics*, pp. 641–649. [95,97]

Zou, H. (2006), "The Adaptive Lasso and its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429. [101]