Statistical Methods for Networks of High-Dimensional Point Processes

Xu Wang

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2021

Reading Committee:

Ali Shojaie, Chair

Chongzhi Di

Noah Simon

Program Authorized to Offer Degree:
Biostatistics - Public Health

University of Washington

Abstract

Statistical Methods for Networks of High-Dimensional Point Processes

Xu Wang

Chair of the Supervisory Committee:
Professor Ali Shojaie
Department of Biostatistics

Fueled in part by recent applications in neuroscience, high-dimensional Hawkes processes are widely used for modeling the network of interactions among multivariate point processes. Despite this popularity, existing methodological and theoretical work has mainly focused on estimation for a single network, assuming all network components are observed. This dissertation aims to develop more flexible estimation and inference tools for high-dimensional Hawkes processes. In Chapter 2, we develop a new statistical inference procedure for high-dimensional Hawkes processes. The key ingredient for this inference procedure is a new concentration inequality on the first- and second-order statistics for integrated stochastic processes, which summarize the entire history of the process. Combining recent results on martingale central limit theory with the new concentration inequality, we then characterize the convergence rate of the test statistics. In Chapter 3, we propose a joint estimation procedure to combine data from multiple experiments for more efficient network estimation. Our procedure incorporates easy-to-compute weights to data-adaptively encourage similarity between the estimated networks. We also develop a powerful hierarchical multiple testing procedure for edges of all estimated networks, taking into account the hierarchical similarity structure of the multi-experiment networks guided by the proposed weights. In Chapter 4, to

cope with the challenges of confounding from unobserved components, we develop a deconfounding procedure to estimate high-dimensional point process networks with only a subset of nodes observed. Unlike existing procedures, our method allows flexible connectivity between the observed and unobserved nodes, and unknown number of hidden nodes that can be larger than the observed population. We finish the dissertation with a discussion in Chapter 5, where we outline a possible future research direction that maps brain networks to animal behavior.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I am grateful to many people for their support and guidance over my doctoral study. I would like to thank my advisor, Ali Shojaie, for his tremendous support and guidance on my dissertation work and patience in helping me build essential skills required for a statistician. I would like to thank my doctoral committee members, Chongzhi Di, Noah Simon, and Nick Steinmetz, for their invaluable support and encouragement throughout the dissertation process. I am grateful to my RA supervisors at Fred Hutchinson Cancer Research Center, Ollivier Hyrien and Chongzhi Di, for providing me opportunities to delving into exciting statistical methods and applications beyond my dissertation research. I would like to thank Mladen Kolar at University of Chicago for his insightful comments on my first project. I am extremely grateful to Lang Wu at UBC, Ana Diez Roux and Brisa Sánchez at Drexel University for teaching me how to conduct statistical and applied research and encouraging me to pursue doctoral training in biostatistics. I would like to thank the staff, faculty, and students at Department of Biostatistics for making the program a warm home. Last but not least, my biggest thanks go to my parents for their endless support and love, my wife for being my best friend forever, and my son for making me the happiest person in the world.

DEDICATION

to my family

Chapter 1

INTRODUCTION

Fueled in part by recent applications in neuroscience, high-dimensional Hawkes processes are widely used for modeling the network of interactions among multivariate point processes. Despite this popularity, existing methodological and theoretical work has mainly focused on estimation for a single network, assuming all network components are observed. In this dissertation, we develop more flexible estimation and inference tools for high-dimensional Hawkes processes. In the first part, we develop a new statistical inference procedure for high-dimensional Hawkes processes. In the second part, we propose a joint estimation to combine data from multiple experiments for more efficient network estimation. We also develop a powerful hierarchical multiple testing procedure for edges of all estimated networks. In the third part, we propose a deconfounding procedure to estimate high-dimensional point process networks with only a subset of the nodes observed. In this chapter, we outline the motivations and provide an overview of the methodological contributions for each part.

1.1 Statistical Inference for Networks of High-Dimensional Point Processes

While evaluating the uncertainty of network estimates is critical in scientific applications, existing methodological and theoretical developments [Hansen et al., 2015, Chen et al., 2019] have only focused on estimation. To bridge this gap, in this project, we develop a statistical inference procedure for high-dimensional Hawkes processes.

Tools for statistical inference in high-dimensional linear models [Zhang and Zhang, 2014, van de Geer et al., 2014, Belloni et al., 2013], graphical models [Barber and Kolar, 2018, Janková and van de Geer, 2019, Lu et al., 2018, Yu et al., 2019], and more general estimators

[Ning and Liu, 2017, Neykov et al., 2018] have been recently developed for the setting of independent data. However, existing tools cannot be directly applied in a time series setting, except for the recent work on statistical inference for high-dimensional vector auto-regressive (VAR) models [Neykov et al., 2018, Zheng and Raskutti, 2019]. While a significant step forward, the VAR model only captures dependence for a fixed and pre-specified time lag (or order). In contrast, the Hawkes process is dependent on its *entire* history. This dependence introduces significant challenges for high-dimensional inference procedures.

In this project, we provide the first high-dimensional inference procedure for multivariate Hawkes processes with both excitatory and inhibitory effects. To this end, we adopt the *de-correlated score test* framework of Ning and Liu [2017] to high-dimensional point processes. We also develop confidence intervals for model parameters by extending the semi-parametric efficient confidence region of Neykov et al. [2018], Zheng and Raskutti [2019] for VAR models to the setting of the multivariate Hawkes process. While the general steps for our inference framework are similar to those in de-correlated score test and efficient confidence regions, key challenges in adopting these tools stem from the dependence of Hawkes processes on their entire past and their continuous-time nature. In particular, to establish our inference framework, we tackle two main challenges: (i) deriving concentration inequalities for summary statistics of high-dimensional Hawkes processes; and (ii) establishing the restricted eigenvalue (RE) condition required for the estimation consistency of ℓ_1-regularized estimators [Bickel et al., 2009].

To address the above challenges, we first generalize the results by Costa et al. [2020] and Chen et al. [2019] to obtain new concentration inequalities on the first- and second-order statistics of the integrated stochastic process that summarizes the entire history of each component of the multivariate Hawkes process. These inequalities are essential for developing our high-dimensional inference procedures. For instance, together with the martingale central limit theorem (CLT) of Zheng and Raskutti [2019], they allow us to establish the convergence of our test statistics to a χ^2 distribution. They are also used to establish the

maximal inequalities needed to investigate the theoretical properties of the ℓ_1-regularized estimator. Next, to bound the eigenvalues of the covariance matrix of the integrated process, we link these eigenvalues to the spectral density of the transition matrix of the Hawkes process. By carefully examining the transition functions of the multivariate Hawkes process, we investigate structural conditions on the transition functions that are sufficient to bound the eigenvalues. These bounds allow us to establish the convergence of our test statistic, and are also used to verify the restricted eigenvalue (RE) condition [Bickel et al., 2009].

1.2 Multi-Experiment Networks of High-Dimensional Point Processes

Neuroscience experiments often consist of short stationary periods. Thus, available sample size from a single experiment may not be sufficient for reliable estimation of the Hawkes processes. To overcome this technical limitation, data are often collected over multiple experiments. For instance, Bolding and Franks [2018] repeat 10 laser stimuli to a set of neurons at 8 different intensity levels ranging from 0 to 50 mW/mm^2; this results in 80 experiments in total. Neuronal connectivity networks corresponding to different stimuli in a sequence of experiments are expected to share many edges. This shared structure motivates joint estimation of networks from multiple experiments, which could lead to more efficient estimation of common edges. However, the networks from different experiments or conditions are also expected to contain unique edges that are in fact of primary scientific interest. For example, distinct neuronal connectivities are found under laser stimulus at different intensity levels [Bolding and Franks, 2018]. Moreover, the degree of (dis)similarity between networks from two experiments depends on the similarity of neuronal responses to the corresponding stimuli, which is generally unknown. These characteristics highlight the need for joint estimation and inference of multiple networks of high-dimensional point processes while accounting for similarities between networks, a task that is not addressed by existing methods.

Joint estimation of multiple graphical models has been considered for independent and

Gaussian-distributed data [e.g. Chiquet et al., 2011, Guo et al., 2011, Danaher et al., 2014, Yajima et al., 2014, Zhu et al., 2014, Ma and Michailidis, 2016, Cai et al., 2016, Huang et al., 2018, Wang et al., 2020a]. Extensions to time- and/or space-varying networks have also been studied [e.g. Kolar et al., 2010, Qiu et al., 2016, Lin et al., 2017, Hallac et al., 2017, Yang and Peng, 2020]. However, the vast majority of existing approaches are primarily designed for learning *two* networks or implicitly assume that networks from multiple experiments are equally similar, or the network similarity is known or implied by the spatial/temporal ordering. On the other hand, the few methods designed for joint estimation of multiple networks [Peterson et al., 2015, Saegusa and Shojaie, 2016] are specific to Gaussian graphical models and either assume the network edges are independent [Peterson et al., 2015], or assume similarities in population means in different conditions reveal similarities among precision matrices [Saegusa and Shojaie, 2016]. Moreover, existing procedures do not provide inference for the estimated networks, which is critically important in many scientific applications.

Given the paucity of methods for joint analysis of point process data from multiple experiments/conditions, in this project, we propose a data-adaptive joint estimation procedure for networks of high-dimensional point processes. The proposed approach uses estimates of cross correlations in each condition to obtain a measure of similarities among neuronal connectivity networks. While cross-correlations are widely used by neuroscientists to gain insights into neuronal interaction mechanisms [de Abril et al., 2018], they do not reveal direct interactions between neurons [Tchumatchenko et al., 2011, Reid et al., 2019]. However, they can be easily computed, even in high dimensions. We also show that they provide useful information about the overall similarities among neuronal connectivity networks, especially when used to define data-driven weights for joint estimation of multiple networks. By encouraging similarity among estimated networks, such data-driven weights offer superior finite-sample performance in selecting the edges. Extending previous work under a single experiment [Chen et al., 2019], we establish a unified non-asymptotic convergence rate for edge estimation in multiple networks of generalized Hawkes processes. Our result implies a

faster convergence rate using the joint estimation approach compared to estimating networks separately under each experiment. While our method does not assume the correctness of the weights to achieve the estimation convergence, we achieve a lower estimation error when the true weights are known.

To address the need for efficient inference procedures, we develop a powerful hierarchical testing procedure for simultaneous inference on edges of all estimated networks. While statistical inference for a single high-dimensional Hawkes processes has been recently addressed [Wang et al., 2020b], implementing such an approach directly on all estimated networks would amount to a large number of tests. As a result, the testing power can diminish when adjusting for multiple comparisons, especially when the number of experiments is large. This is particularly the case in neuroscience applications, such as that in Bolding and Franks [2018] with 80 experiments. Moreover, the tests associated with network edges under each experiment have complex dependencies. Our proposed inference framework mitigates these challenges through a multiple comparison adjustment that carries out testing along the hierarchical structure of network similarities inferred from cross-correlations. By taking advantage of this hierarchical structure, our procedure greatly reduces the number of required tests, which in turn results in increased power while tightly controlling the family-wise error rate (FWER). Unlike existing hierarchical testing procedures that rely on specific assumptions on the dependency between the tests [Yekutieli, 2008, Lynch and Guo, 2016] or particular p-values calculation along the hierarchy [Bogomolov et al., 2020], our method is free of such assumptions and can incorporate p-values flexibly calculated from any valid tests. Moreover, we show that for large and sparse networks, our inference procedure is robust to misspecification of the similarity weights, which had not been previously addressed. The advantages of our estimation and inference procedures are illustrated by analyzing simulated and real data, respectively.

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

While learning causal interactions between neurons is critical to understanding the neuronal connectivity mechanisms in neuroscience, such a task is generally impossible using observational data because of the confounding from unobserved neurons. Existing estimation and inference procedures that require all interacted neurons observed thus have no guarantee of achieving reliable conclusions in practice.

The problem of causal discovery with hidden variables has been studied using causal structural learning such as Fast Causal Inference (FCI) and its variants [Spirtes et al., 2000, Glymour et al., 2019]. Unfortunately, not all causal edges can always be identified purely based on such algorithms. Specifically, while FCI identifies causal interactions between variables that are connected by directed edges, it sometimes produces bi-directed edges, leaving the corresponding causal relationship undetermined. Thus, the causality selection performance using such algorithms is not always satisfactory. For example, Malinsky and Spirtes [2018] recently applied FCI to infer causal network of time series and found a low recall in identifying the true casual relationship. Additionally, the causal structure learning often requires intensive computation, especially for large networks, because the space of candidate causal graphs grows super-exponentially with the number of network nodes.

When the variables inhibit temporal ordering as in time series data, the causal structure can be directly learned via the network skeleton using multiple regression models [Shojaie and Michailidis, 2010], which can be solved in a linear time to the number of nodes in the network. While learning causal networks in the presence of unobserved nodes remains challenging using this framework, two recently-developed methods, HIVE [Bing et al., 2020] and trim regression [Ćevid et al., 2020], shed light on solving this challenge. Specifically, HIVE generates consistent estimates on causality given that the observed and unobserved causal effects are orthogonal. While such condition might hold when the unobserved processes are external such as experimental perturbations in genetic studies [Lee et al., 2017], it might

be too stringent in a network setting like neuroscience. Trim regression applies a spectral transformation to both the outcome and the design columns, which empirically reduces the confounding magnitude. It then eliminates the confounding using penalized regression, assuming the confounding is sufficiently small. However, it is unclear whether this assumption is satisfied in a network setting where the observed and unobserved processes are complexly connected. Moreover, the long-history temporal dependency of the Hawkes processes makes the learning task more challenging. Unfortunately, both methods were developed for independent data. Also, these methods were only evaluated on their edge-estimation performance but their abilities in edge selection in a network setting are unknown.

Previous work suggests that the confounding has restricted magnitudes under a stable Hawkes process network, especially when it is dense over the observed nodes [Chen et al., 2019]. Such property motivates us to extend the trim regression to learn the causal network of multivariate point processes. Our method shows superior finite-sample performance compared to HIVE and the naive approach that neglects the unobserved processes. We also establish a non-asymptotic convergence rate in estimating the network edges. Unlike the previous result on independent data [Ćevid et al., 2020], our result considers both the temporal dependency of the Hawkes processes and the network sparsity.

Chapter 2

STATISTICAL INFERENCE FOR NETWORKS OF HIGH-DIMENSIONAL POINT PROCESSES

2.1 Introduction

Multivariate point process data have become prevalent in a number of emerging application areas. Examples include neural spike train data in neuroscience, containing times of neuron spikes of a collection of neurons [Okatan et al., 2005]; social media data, recording times when each individual in an online community takes an action [Zhou et al., 2013]; and high frequency financial data, recording times of market orders [Chavez-Demoulin and McGill, 2012]. The latent connectivity structure of these processes can be represented by a probabilistic graphical model [Lauritzen, 1996] with a graph or network $G = (V, E)$ whose nodes, $v \in V$, represent components/units in the multivariate point processes and each *directed* edge, $(u \to v) \in E$, indicates that the probability of future events of the target node v depends on the history of the source node u. Multivariate point process data can be used to learn the structure of this network.

In a seminal work, Hawkes [1971] proposed a class of multivariate point process models, where the probability of future events for a component can depend on the entire history of events of other components. Because of its flexibility and interpretability when modeling the dependence structure of component processes, the multivariate Hawkes process has become a popular tool for studying the latent network of point processes. From its early application in earthquake prediction [Ogata, 1988], the multivariate Hawkes process model has been widely used to learn the latent connectivity structure in many fields, including neuroscience [Chen et al., 2019], social media [Zhou et al., 2013], and finance [Bacry et al., 2011, Linderman and

Adams, 2014].

The Hawkes process, as introduced in Hawkes [1971] and later studied in Hawkes and Oakes [1974], Reynaud-Bouret and Roy [2007], Reynaud-Bouret and Schbath [2010], Bacry et al. [2015], Hansen et al. [2015], Etesami et al. [2016], is considered as a *mutually-exciting* process, in which an event can only excite future events. More specifically, each event in any component may trigger future events in all other components, including itself. However, in many applications, it is desired to allow for *inhibitory* effects of past events. For example, a spike in one neuron may inhibit the activities of other neurons [Babington, 2001], which means that it decreases the probability that other neurons would spike. Costa et al. [2020] and Chen et al. [2019] developed a broader class of Hawkes process models that allow for both excitatory and inhibitory effects in a single and multivariate point process data, respectively.

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 time period, e.g., the duration of neuroscience experiments. The high-dimensional nature of data in such applications poses additional challenges to learning the connectivity network of a multivariate point process. Hansen et al. [2015] and Chen et al. [2019] proposed ℓ_1-regularized estimation procedures to address this challenge. However, there are no tools to characterize the sampling distribution of these estimators and characterize their uncertainty. Such inferential tools are critical in scientific applications.

Tools for statistical inference in high-dimensional linear models [Zhang and Zhang, 2014, van de Geer et al., 2014, Belloni et al., 2013], graphical models [Barber and Kolar, 2018, Janková and van de Geer, 2019, Lu et al., 2018, Yu et al., 2019], and more general estimators [Ning and Liu, 2017, Neykov et al., 2018] are only recently developed for the setting of independent data and cannot be directly applied in a time series setting. Statistical inference for high-dimensional vector auto-regressive (VAR) models was recently studied by Neykov et al. [2018] and Zheng and Raskutti [2019]. While a significant step forward, the VAR model only captures dependence for a fixed and pre-specified time lag (or order). In contrast, the

Hawkes process is dependent on the *entire* history, which introduces significant challenges in developing inferential procedures for the high-dimensional multivariate Hawkes process. In particular, this dependence on the entire history complicates the proof of convergence of the test statistic for the multivariate Hawkes process. Moreover, unlike the time series models that are often set up in a discrete time domain [Basu and Michailidis, 2015, Zheng and Raskutti, 2019], the multivariate Hawkes process is defined in a continuous time domain and thus requires different technical tools to investigate its properties.

In this chapter, we provide the first high-dimensional inference procedure for multivariate Hawkes processes with both excitatory and inhibitory effects. To this end, we adopt the de-correlated score test framework of Ning and Liu [2017] to high-dimensional point processes. We also develop confidence intervals for model parameters by extending the semi-parametric efficient confidence region of Neykov et al. [2018], Zheng and Raskutti [2019] for VAR models to the setting of the multivariate Hawkes process. While the general steps for our inference framework are similar to those in de-correlated score test and efficient confidence regions, key challenges in adopting these tools stem from the dependence of Hawkes processes on their entire past and their continuous-time nature. In particular, to establish our inference framework, we tackle two main challenges: (i) deriving concentration inequalities for summary statistics of high-dimensional Hawkes processes; and (ii) establishing the restricted eigenvalue (RE) condition required for the estimation consistency of ℓ_1-regularized estimators [Bickel et al., 2009].

To address the above challenges, we first generalize the results by Costa et al. [2020] and Chen et al. [2019] to obtain new concentration inequalities on the first- and second-order statistics of the integrated stochastic process that summarizes the entire history of each component of the multivariate Hawkes process. These inequalities are essential for developing our high-dimensional inference procedures. For instance, together with the martingale central limit theorem (CLT) of Zheng and Raskutti [2019], they allow us to establish the convergence of our test statistics to a χ^2 distribution. They are also used to establish the

maximal inequalities needed to investigate the theoretical properties of the ℓ_1-regularized estimator. Next, to bound the eigenvalues of the covariance matrix of the integrated process, we link these eigenvalues to the spectral density of the transition matrix of the Hawkes process. By carefully examining the transition functions of the multivariate Hawkes process, we investigate structural conditions on the transition functions that are sufficient to bound the eigenvalues. These bounds allow us to establish the convergence of our test statistic, and are also used to verify the restricted eigenvalue (RE) condition [Bickel et al., 2009].

2.2 The Linear 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 $t_1 \geq 0$. 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

$$\lambda_i(t)dt = \mathbb{P}(dN_i(t) = 1 \mid \mathcal{H}_t). \tag{2.1}$$

Hawkes [1971] 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 ($i \in \{1, \dots, p\}$) takes the form

$$\lambda_i(t) = \mu_i + \sum_{j=1}^{p} (\omega_{ij} * dN_j)(t), \tag{2.2}$$

where

$$(\omega_{ij} * dN_j)(t) = \int_0^{t-} \omega_{ij}(t-s)dN_j(s) = \sum_{k:t_{jk}<t} \omega_{ij}(t-t_{jk}). \tag{2.3}$$

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 [Linderman and Adams, 2014, de Abril et al., 2018], we consider a parametric transfer function $\omega_{ij}(\cdot)$ of the form

$$\omega_{ij}(t) = \beta_{ij}\kappa_j(t) \tag{2.4}$$

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*

$$x_j(t) = \int_0^{t-} \kappa_j(t-s)dN_j(s) \tag{2.5}$$

summarizes the entire history of unit j of the multivariate Hawkes processes. A commonly used example is the exponential transition kernel, $\kappa_j(t) = e^{-t}$ [Bacry et al., 2015].

In this formulation, the *connectivity coefficient* of the underlying network, β_{ij}, represents the strength of the dependence of unit i's intensity on unit j's past events. A positive β_{ij}, which implies that past events of unit j *excite* future events of unit i, is often considered in the literature [see, e.g., Bacry et al., 2015, Etesami et al., 2016]. However, we might also wish to allow for negative β_{ij} values to represent *inhibitory* effect of one unit's past events on another [Chen et al., 2019, Costa et al., 2020], which is expected in neuroscience applications [Babington, 2001].

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

$$\lambda_i(t) = \mu_i + \boldsymbol{x}^\top(t)\boldsymbol{\beta}_i. \tag{2.6}$$

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

$$Y_i(t) = \mu_i + \boldsymbol{x}^\top(t)\boldsymbol{\beta}_i + \epsilon_i(t). \tag{2.7}$$

As we will discuss later, a key challenge in this 'linear model' stems from heteroscedasticity: the variance of $\epsilon_i(t)$ given the history of $\mathbf{N}$ up to t,

$$\sigma_i^2(t) \equiv \mathrm{Var}\left(\epsilon_i(t) \mid \mathcal{H}_t\right) = \lambda_i(t)/dt(1 - \lambda_i(t)dt), \tag{2.8}$$

may not necessarily be 1 and depends on $\boldsymbol{x}(t)$.

Throughout this paper, we assume that the linear Hawkes model described above is *stationary*, meaning that for all units $i = 1, \ldots, p$, the spontaneous rates μ_i and strengths of transition $\boldsymbol{\beta}_i$ are constant over the time range $[0, T]$ [Brémaud and Massoulié, 1996, Daley and Vere-Jones, 2003].

2.3 Testing

Let $J \subset \{1, \ldots, p\}$ be an index set of cardinality $|J| = d$ and denote $\boldsymbol{\beta}_{iJ} = \{\beta_{ij}, j \in J\}$. We consider testing a d-dimensional subset $\boldsymbol{\beta}_{iJ}$:

$$H_0 : \beta_{ij} = 0, \qquad j \in J. \tag{2.9}$$

For ease of notation, we primarily focus on the case of a single parameter; that is, we consider testing $H_0 : \beta_{ij} = 0$, which corresponds to $d = 1$. However, our inferential framework is developed for the more general case of $d \geq 1$.

In order to simplify the presentation, we scale the components of $\boldsymbol{x}(t)$ by the variance of the noise $\sigma_i^2(t)$ defined in (2.8). Denote the scaled components, $z_j(t) = x_j(t)/\sigma_i(t)$ for $j = 1, \ldots, p$. Next, we define the orthogonal projection of $z_j(t)$ onto $\boldsymbol{z}_{-j}(t)$, where $\boldsymbol{z}_{-j}(t) = (z_1(t), \ldots, z_{j-1}(t), z_{j+1}(t), \ldots, z_p(t))^\top \in \mathbb{R}^{p-1}$. Let the projection coefficient $\boldsymbol{w}_j^* = \left(w_{j0}^*, (\boldsymbol{w}_{j,-j}^*)^\top\right)^\top \in \mathbb{R}^p$ be such that

$$\mathbb{E}\left[z_j^*(t)\boldsymbol{z}_{-j}(t)\right] = 0 \qquad \text{and} \qquad \mathbb{E}\left[z_j^*(t)\right] = 0, \tag{2.10}$$

where

$$z_j^*(t) \equiv z_j(t) - \left(1, \boldsymbol{z}_{-j}^\top(t)\right) \boldsymbol{w}_j^* \tag{2.11}$$

denotes the orthogonal complement of $z_j(t)$ after removing its projection onto $\boldsymbol{z}_{-j}(t)$. Then, $z_j^*(t)$ is uncorrelated with $\boldsymbol{z}_{-j}(t)$. To be specific, such orthogonal projection is available by the projection coefficient

$$\boldsymbol{w}_j^* = \left(\mathbb{E} \begin{pmatrix} 1 \\ \boldsymbol{z}_{-j}(t) \end{pmatrix} \begin{pmatrix} 1 & \boldsymbol{z}_{-j}(t) \end{pmatrix} \right)^{-1} \mathbb{E} \begin{pmatrix} 1 \\ \boldsymbol{z}_{-j}(t) \end{pmatrix} z_j(t),$$

which only depends on the network structure and the background intensity.

In addition, let $\widetilde{\epsilon}_i(t) = \epsilon_i(t)/\sigma_i(t)$. With this notation, we define the *de-correlated score* statistic as

$$S_{ij} = \frac{1}{T} \sum_{t=1}^{T} \widetilde{\epsilon}_i(t)\, z_j^*(t). \tag{2.12}$$

The de-correlated score statistic is constructed using $z_j^*(t)$, instead of directly using $z_j(t)$, to make its sampling distribution robust to errors induced by the estimation of the unknown nuisance parameters, μ_i and $\boldsymbol{\beta}_{i,-j}$ [Ning and Liu, 2017]. In particular, we will show that the induced error is asymptotically negligible and the limiting distribution of the test statistic does not depend on model selection mistakes that occur when estimating the nuisance parameters. In practice, due to technology constraints—e.g. the precision of timing at 30Hz [Bolding and Franks, 2018]), truly continuous time processes could not be obtained. Therefore, we develop the test statistics incorporating the available data in a discrete time domain. Specifically in (2.12), we consider the spike train samples observed over T equally-spaced time indexed from 1 to T.

Neykov et al. [2018] and Zheng and Raskutti [2019] consider a similar de-correlated score statistic in the context of VAR models. Their proof strategy exploits the homoscedastic noise variance in VAR models and does not extend to the linear Hawkes process. In contrast, we need to take into account the noise variance when constructing the score statistic, as the

variance varies over time. Moreover, the noise variance depends on the intensity value, which is time varying, resulting in more challenging proof in our case.

In order to construct a test, we need to characterize the quantiles of the de-correlated score statistic. Let

$$\Upsilon_j = \mathrm{Cov}\left(z_j^*(t)\right) = \mathbb{E}\left(\left(z_j^*(t)\right)^2\right), \tag{2.13}$$

$$V_T = \sqrt{T}\,\Upsilon_j^{-1/2} S_{ij}, \tag{2.14}$$

$$U_T = \|V_T\|_2^2. \tag{2.15}$$

While Υ_j is scalar when testing a univariate β_{ij}, when testing multiple parameters $\boldsymbol{\beta}_{iJ}$, Υ_J is a $d \times d$ matrix defined as

$$\Upsilon_J = \mathrm{Cov}\left(\boldsymbol{z}_J^*(t)\right).$$

In the next section, we show that U_T converges weakly to a χ^2 distribution with degrees of freedom d, which is 1 for testing a univariate β_{ij}. The non-centrality parameter is zero under the null hypothesis and depends on the true parameters under the alternative.

In practice, $(\mu_i, \boldsymbol{\beta}_i)$ and $\boldsymbol{w}_j^*$ are not known. We next describe a procedure for estimating them.

Step 1: Calculate $\widehat{\mu}_i$, $\widehat{\boldsymbol{\beta}}_i$, and $\widehat{\sigma}_i^2(t)$. We estimate $\widehat{\mu}_i, \widehat{\boldsymbol{\beta}}_i$ using the lasso on the unscaled data $(Y_i(t), \boldsymbol{x}(t))$:

$$\widehat{\mu}_i, \widehat{\boldsymbol{\beta}}_i = \underset{\mu_i \in \mathbb{R}, \boldsymbol{\beta}_i \in \mathbb{R}^p}{\arg\min} \frac{1}{T} \sum_{t=1}^{T} \left(Y_i(t) - \mu_i - \boldsymbol{x}^\top(t)\boldsymbol{\beta}_i\right)^2 + \lambda \|\boldsymbol{\beta}_i\|_1. \tag{2.16}$$

Then,

$$\widehat{\lambda}_i(t) = \widehat{\mu}_i + \boldsymbol{x}^\top(t)\widehat{\boldsymbol{\beta}}_i \qquad \text{and} \qquad \widehat{\sigma}_i^2(t) = \widehat{\lambda}_i(t)(1 - \widehat{\lambda}_i(t)). \tag{2.17}$$

The estimation consistency for $\widehat{\mu}_i$, $\widehat{\boldsymbol{\beta}}_i$ to the corresponding true parameters is shown in Lemma A.11. The consistency of $\widehat{\sigma}^2(t)$ to $\sigma^2(t)$ follows from the prediction consistency of the lasso estimator. In our proof, we show that the *restricted eigenvalue* (RE) condition

required for the consistency of lasso [Bickel et al., 2009] is met in our case. This follows from the bounded eigenvalue of the covariance matrix of the integrated stochastic process $\boldsymbol{x}(t)$, which is obtained under the assumptions made in the following section.

The tuning parameter λ is selected via cross-validation for sequentially dependent data [Safikhani and Shojaie, 2020]. Specifically, unlike standard cross-validation for independent samples, the sequential cross-validation uses successively training sets along with validation sets that follow each of the training sets in the sequence order.

Step 2: Calculate $\widehat{\boldsymbol{w}}_j$. Let $\widehat{z}_j(t) = x_j(t)/\widehat{\sigma}_i(t)$ for $j = 1, \ldots, p$. We estimate $\widehat{\boldsymbol{w}}_j$ by regressing the outcome $\widehat{z}_j$ on the design matrix $\widehat{\boldsymbol{z}}_{-j}$ using a lasso procedure with tuning parameter selected as in Step 1:

$$\widehat{\boldsymbol{w}}_j = \underset{\boldsymbol{w}_j \in \mathbb{R}^p}{\arg\min} \frac{1}{T} \sum_{t=1}^{T} \left(\widehat{z}_j(t) - \begin{pmatrix} 1 & \widehat{\boldsymbol{z}}_{-j}^\top(t) \end{pmatrix} \boldsymbol{w}_j \right)^2 + \lambda \|\boldsymbol{w}_{j,-j}\|_1. \tag{2.18}$$

The consistency of $\widehat{\boldsymbol{w}}_j$ for $\boldsymbol{w}_j^*$ is shown in Lemma A.12. The proof is similar to the one used for Step 1.

Step 3: Calculate $\widehat{\Upsilon}_j$. Let $\widehat{z}_j^*(t) = \widehat{z}_j(t) - \begin{pmatrix} 1 & \widehat{\boldsymbol{z}}_{-j}^\top(t) \end{pmatrix} \widehat{\boldsymbol{w}}_j$. When testing a univariate β_{ij} in (2.9), Υ_j is a scalar and we estimate it by the sample covariance as

$$\widehat{\Upsilon}_j = \frac{1}{T} \sum_{t=1}^{T} \left(\widehat{z}_j^*(t) \right)^2. \tag{2.19}$$

When testing multiple $\boldsymbol{\beta}_{iJ} = \{\beta_{ij}, j \in J\}$, we estimate Υ_J as

$$\widehat{\Upsilon}_J = \frac{1}{T} \sum_{t=1}^{T} \widehat{z}_J^*(t) \left(\widehat{z}_J^*(t) \right)^\top \in \mathbb{R}^{J \times J}.$$

Our results are valid as long as $d = |J| \ll p$.

Step 4: Putting everything together, we compute the de-correlated score statistic with

estimated nuisance parameters. Let

$$\widehat{\epsilon}_i(t) = Y_i(t) - \widehat{\mu}_i - \boldsymbol{x}_{-j}^\top(t)\widehat{\boldsymbol{\beta}}_{i,-j}, \tag{2.20}$$

$$\widehat{S}_{ij} = \frac{1}{T}\sum_{t=1}^{T} \frac{\widehat{\epsilon}_i(t)}{\widehat{\sigma}_i(t)} \widehat{z}_j^*(t), \tag{2.21}$$

$$\widehat{V}_T = \sqrt{T}\widehat{\Upsilon}_j^{-1/2}\widehat{S}_{ij}, \tag{2.22}$$

$$\widehat{U}_T = \|\widehat{V}_T\|_2^2. \tag{2.23}$$

The above quantities are univariate when testing a univariate β_{ij}, but are defined as vectors and matrices when testing multivariate $\boldsymbol{\beta}_{iJ}$ of dimension d. Specifically, $\widehat{S}_{iJ} \in \mathbb{R}^d$, $\widehat{\Upsilon}_J \in \mathbb{R}^{d\times d}$ and $\widehat{V}_T \in \mathbb{R}^d$.

In the next section, we show that with high probability $\widehat{U}_T$ converges to U_T, which asymptotically follows a χ^2 distribution with d degrees of freedom under the null hypothesis. Thus, we define our test procedure for (2.9) as

$$\Phi_\alpha = \mathrm{I}\left\{\widehat{U}_T \geq \chi^2_{d,1-\alpha}\right\},$$

and reject the null hypothesis when $\Phi_\alpha = 1$.

2.4 Theoretical Guarantees

In this section, we present our main theoretical results, which characterize the limiting distribution of $\widehat{U}_T$. 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. Define $\Theta = \{\beta_{ij}\}_{1\leq i,j\leq p} \in \mathbb{R}^{p\times p}$ and $\boldsymbol{\mu} = \{\mu_i\}_{1\leq i\leq p} \in \mathbb{R}^p$.

Assumption 2.1. *Let* $\Omega = \{\Omega_{ij}\}_{1\leq i,j\leq p} \in \mathbb{R}^{p\times p}$ *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 2.1 is necessary for stationarity of a Hawkes process [Chen et al., 2019]. The constant γ_Ω does not depend on the dimension p. For any fixed p, Brémaud and Massoulié

[1996] show that given this assumption the intensity process of the form (2.2) is stable in distribution and, thus, a stationary process $\mathbf{N}$ exists. Since our connectivity coefficients of interest, Θ, are ill-defined without a stationarity, this assumption provides the necessary context for our inferential framework.

Assumption 2.2. *There exists constants* $\rho_r \in (0,1)$ *and* $0 < \rho_c < \infty$ *such that*

$$\max_{1\leq i\leq p} \sum_{j=1}^{p} \Omega_{ij} \leq \rho_r \qquad \textit{and} \qquad \max_{1\leq j\leq p} \sum_{i=1}^{p} \Omega_{ij} \leq \rho_c.$$

Assumption 2.2 requires maximum in- and out- intensity flows to be bounded, which helps in bounding the eigenvalues of the cross-covariance of $\boldsymbol{x}(t)$. A similar assumption is also considered by Basu and Michailidis [2015] in the context of VAR models. The condition of $\rho_r \in (0,1)$ prevents the intensity from concentrating to a single process [Chen et al., 2019].

Assumption 2.3. *There exists* $\lambda_{\min}$ *and* $\lambda_{\max}$ *such that*

$$0 < \lambda_{\min} \leq \lambda_i(t) \leq \lambda_{\max}$$

for all $i = 1, \ldots, p$ *and* $t \in [0, T]$.

Assumption 2.3 requires that the intensity rate is strictly bounded, which prevents degenerate processes for all units of the multivariate Hawkes process. As a consequence, $\sigma_i^2(t)$ will be bounded away from 0, and hence the construction of the de-correlated score in (2.12) is valid.

Assumption 2.4. *The transition kernel* $\kappa_j(t)$ *is positive and integrable over* $[0, T]$, *for* $1 \leq j \leq p$.

Assumption 2.4 implies that the integrated process $x_j(t)$ defined in (2.5) is bounded. Together with Assumptions 2.3, it also implies that μ_i and $\boldsymbol{\beta}_i$ are bounded for all $i = 1, \ldots, p$.

Our next two assumptions state the rate of convergence for the estimators of $(\mu_i, \boldsymbol{\beta}_i)$ and $\boldsymbol{w}_j^*$ that guarantee the weak convergence of the test statistic. We let Π_0 and Π_a denote

the feasible set of $(\boldsymbol{\mu}, \Theta)$ under the null and the alternative hypotheses, respectively, with Assumptions 2.1– 2.4 satisfied. In addition, we use $s_j = \|\boldsymbol{w}_j^*\|_0$, $s = \max_{1\le j\le p} s_j$, $\rho_i = \|\boldsymbol{\beta}_i\|_0$, and $\rho = \max_{1\le i\le p} \rho_i$ to denote the sparsity of $\boldsymbol{w}_j^*$ and $\boldsymbol{\beta}_i$.

Assumption 2.5 (Estimation error of $\boldsymbol{\beta}_i$)**.** *For $(\boldsymbol{\mu}, \Theta) \in \Pi_0 \cup \Pi_a$ and $r \in \{1, 2\}$,*

$$\left\| \begin{pmatrix} \widehat{\mu}_i \\ \widehat{\boldsymbol{\beta}}_i \end{pmatrix} - \begin{pmatrix} \mu_i \\ \boldsymbol{\beta}_i \end{pmatrix} \right\|_r \le C_1(\rho+1)^{1/r} T^{-2/5},$$

for all $1 \le j \le p$, with probability at least $1 - C_2 p^2 T \exp(-C_3 T^{1/5})$. Constants C_1, C_2, C_3 only depend on $(\boldsymbol{\mu}, \Theta)$ and the transition kernel function.

Assumption 2.6 (Estimation error of $\boldsymbol{w}_j^*$)**.** *For $(\boldsymbol{\mu}, \Theta) \in \Pi_0 \cup \Pi_a$ and $r \in \{1, 2\}$,*

$$\left\| \widehat{\boldsymbol{w}}_j - \boldsymbol{w}_j^* \right\|_r \le C_1(s+1)^{\frac{3-r}{2}} \rho T^{-2/5},$$

for all $1 \le j \le p$, with probability at least $1 - C_2 p^2 T \exp(-C_3 T^{1/5})$. Constants C_1, C_2, C_3 only depend on $(\boldsymbol{\mu}, \Theta)$ and the transition kernel function.

Assumption 2.5 and 2.6 state the rate of convergence for estimators of the nuisance components. Under a stationary linear Hawkes process that satisfies Assumptions 2.1–2.4, Lemmas 2.6 and A.11 in Appendix A.3 show that the lasso estimators in (2.16) and (2.18) satisfy Assumptions 2.5 and 2.6. However, our main results on the limiting distribution of the test statistic are valid for other high-dimensional estimators, as long as their rate of convergence satisfies the requirements in Assumptions 2.5 and 2.6.

The rate of convergence naturally depends on the sparsity of $\boldsymbol{\beta}_i$ and $\boldsymbol{w}_j^*$. In general, the relationship between the sparsity of $\boldsymbol{w}_j^*$ and the sparsity of $\boldsymbol{\beta}_i$ is not straightforward — it depends on the sign and scale of the connectivity coefficients, as well as the transition kernel. Lemma A.13 in Appendix A.3 shows that the sparsity of $\boldsymbol{w}_j^*$ is upper bounded by the sparsity of $\boldsymbol{\beta}_i$ as $s \le 2\rho + 1$ when connectivity matrix is block diagonal. The rate of convergence for an estimator of $\boldsymbol{w}_j^*$ in Assumption 2.6 depends on both the sparsity of $\boldsymbol{w}_j^*$ and $\boldsymbol{\beta}_i$. This

is because estimation of $\boldsymbol{w}_j^*$ requires an estimate of the unknown variance, which in turn depends on the estimates of $(\mu_i, \boldsymbol{\beta}_i)$.

Next, we introduce our first result, which establishes the weak convergence of $\widehat{U}_T$ under the null hypothesis.

Theorem 2.1. *Suppose the linear Hawkes process defined in (2.2) satisfies Assumptions 2.1–2.4. Furthermore assume that* $\left(\widehat{\mu}_i, \widehat{\boldsymbol{\beta}}_i\right)$ *and* $\widehat{\boldsymbol{w}}_j$ *satisfy Assumptions 2.5 and 2.6. If* $s^2\rho^2 \log p = o\left(T^{1/5}\right)$*, then, under the null hypothesis in (2.9),*

$$\sup_{(\Theta,\mu)\in\Pi_0, x\in\mathbb{R}} \left|\mathbb{P}\left(\widehat{U}_T \leq x\right) - F_d(x)\right| \leq C_1 p^2 T \exp(-C_2 T^{1/5}) + C_3 s^2 \rho^2 T^{-1/5} + C_4 T^{-1/8}, \quad (2.24)$$

where F_d *is the cdf of the* χ^2*-distribution with* d *degrees of freedom and* $C_k, k = 1, \ldots, 4$*, are constants only depending on the model parameter* $(\boldsymbol{\mu}, \Theta)$ *and the transition kernel function.*

Theorem 2.1 shows that $\widehat{U}_T$ converges to χ^2_d in distribution ($d = 1$ when testing univariate β_{ij}). The proof involves quantifying the difference between the cdf of U_T and $F_d(x)$ and establishing the convergence of $\widehat{U}_T$ to U_T. The main challenge in establishing these results stems from the dependency structure of the multivariate Hawkes process whose intensity depends on the entire history of each component. The additional complexity due to the dependency structure leads to a slower rate of convergence in quantifying the difference between $\widehat{U}_T$ and U_T than those obtained for the VAR model [Zheng and Raskutti, 2019]. Moreover, this dependence also leads to a difference between cdf of U_T and $F_d(x)$ that is dominated at least by $O\left(T^{-1/8}\right)$ (using the martingale central limit theorem in Proposition A.1) rather than $O\left(T^{-1/2}\right)$ (using the standard central limit theorem).

Next, we investigate the distribution of $\widehat{U}_T$ under the alternative hypothesis. More specifically, for $\phi > 0$, we assume

$$H_a : \beta_{ij} = T^{-\phi}\Delta. \quad (2.25)$$

Theorem 2.2. *Suppose the linear Hawkes process defined in (2.2) satisfies Assumptions 2.1–2.4. Furthermore assume that* $\left(\widehat{\mu}_i, \widehat{\boldsymbol{\beta}}_i\right)$ *and* $\widehat{\boldsymbol{w}}_j$ *satisfy Assumptions 2.5 and 2.6. Let* $F_{d,\delta}$ *be*

the cdf of a non-central χ^2-distribution with d degrees of freedom and non-centrality parameter δ. If $s^2\rho^2 \log p = o\left(T^{1/5} \wedge T^{2\phi-\frac{7}{5}}\right)$, then, under the alternative hypothesis in (2.25), we have:

- *if $\phi > \frac{1}{2}$,*

$$\sup_{(\Theta,\mu)\in\Pi_a, x\in\mathbb{R}} \left|\mathbb{P}\left(\widehat{U}_T \le x\right) - F_d(x)\right| \le C_1 p^2 T \exp(-C_2 T^{1/5}) + C_3 s^2\rho^2\left(T^{-1/5} \vee T^{\frac{7}{5}-2\phi}\right) + C_4 T^{-1/8}; \quad (2.26)$$

- *if $\phi = \frac{1}{2}$,*

$$\sup_{(\Theta,\mu)\in\Pi_a, x\in\mathbb{R}} \left|\mathbb{P}\left(\widehat{U}_T \le x\right) - F_{d,\|\widetilde{\Delta}\|_2^2}(x)\right| \le C_1 p^2 T \exp(-C_2 T^{1/5}) + C_3 s^2\rho^2 T^{-1/5} + C_4 T^{-1/8}, \quad (2.27)$$

  where $\widetilde{\Delta} = \Upsilon_j^{1/2}\Delta$ with Υ_j defined in (2.13);

- *if $\phi < \frac{1}{2}$,*

$$\sup_{(\Theta,\mu)\in\Pi_a, x\in\mathbb{R}} \left|\mathbb{P}\left(\widehat{U}_T \le x\right)\right| \le C_1 p^2 T \exp(-C_2 T^{1/5}) + C_3 T^{-1/8} + C_4 \exp(-(C_5 T^{1/2-\phi} - C_6\sqrt{x})^2); \quad (2.28)$$

here $C_k, k = 1,\dots,6$, are constants only depending on the model parameter $(\boldsymbol{\mu}, \Theta)$ and the transition kernel function.

Theorem 2.2 establishes the asymptotic distribution of $\widehat{U}_T$ under the alternative hypothesis. Depending on the scaling of β_{ij} with respect to T, which is parameterized by ϕ, the asymptotics are different. When $\phi > 1/2$, our test does not distinguish H_a from H_0, since in both cases $\widehat{U}_T$ convergences to χ^2_d. When $\phi < 1/2$, $\widehat{U}_T$ diverges to $+\infty$ in probability, resulting in trivial rejection of the null hypothesis. Finally, when $\phi = 1/2$, $\widehat{U}_T$ converges to a non-central χ^2-distribution with d degrees of freedom and non-centrality parameter $\|\widetilde{\Delta}\|_2^2$.

This result should be compared with Theorem 3.2 in Zheng and Raskutti [2019] developed for VAR models. However, since the multivariate Hawkes process is defined on a continuous time domain with intensity rate depending on the entire history, rather than a pre-specified time lag (or order), our rate of convergence is slower compared with the VAR model.

The results in Theorem 2.1 and 2.2 are established by extending the concentration inequality developed in Chen et al. [2019], which is built on a Bernstein type inequality for weakly dependent observations of the point process at different time points [Merlevède et al., 2009]. This weak dependence leads to a slower rate of convergence in the second order statistics of $\boldsymbol{x}(t)$ compared with the standard sub-Gaussian deviation bound for independent samples; see Chen et al. [2019, Theorem 4] or Merlevède et al. [2009, Theorem 1] for details. As an alternative, in Lemma A.19 in the Appendix A.4, we write the relevant statistics (i.e. the second order statistics of $\boldsymbol{x}(t)$ in our case) based on independent 'residuals', referred to as martingale compensated processes [Bacry et al., 2011]. Then, considering the point process in a discrete time domain, we use the Hansen-Wright inequality to obtain a faster rate of convergence that is comparable to the standard Gaussian deviation bounds for VAR models in Zheng and Raskutti [2019]. However, this is obtained under a more stringent requirement on the structure of the transfer function of the Hawkes processes.

2.5 Confidence Regions

We next describe a procedure for constructing a confidence intervals for β_{ij}. Similar to Ning and Liu [2017], our confidence interval is based on the one-step estimator of β_{ij}. Let $\widehat{\beta}_{ij}$ be the lasso estimator in (2.16), or any other consistent estimator with the same order of the estimation error, and let

$$\widetilde{\Upsilon}_j = \frac{1}{T}\sum_{t=1}^{T} \widehat{z}_j^*(t)\widehat{z}_j(t) \qquad \text{and} \qquad \widetilde{S}_{ij} = \frac{1}{T}\sum_{t=1}^{T} \frac{Y_i(t) - \widehat{\mu}_i - \boldsymbol{x}^\top(t)\widehat{\boldsymbol{\beta}}_i}{\widehat{\sigma}_i(t)}\widehat{z}_j^*(t). \tag{2.29}$$

Note that $\widetilde{S}_{ij}$ involves the entire $\widehat{\boldsymbol{\beta}}_i$, compared with $\widehat{S}_{ij}$ which uses $\widehat{\boldsymbol{\beta}}_{i,-j}$. We define the one-step estimator of β_{ij} as

$$\widehat{b}_{ij} = \widehat{\beta}_{ij} - \left(\widetilde{\Upsilon}_j\right)^{-1} \widetilde{S}_{ij}. \tag{2.30}$$

Finally, let

$$\widehat{R}_T = T\left(\widehat{b}_{ij} - \beta_{ij}\right)^\top \widehat{\Upsilon}_j \left(\widehat{b}_{ij} - \beta_{ij}\right). \tag{2.31}$$

Our next result shows that $\widehat{R}_T$ converges weakly to χ^2_d. Therefore, we construct an asymptotically $1-\alpha$ confidence region for β_{ij} as

$$\mathrm{CR}(\alpha) = \left\{\theta : T\left(\widehat{b}_{ij} - \theta\right)^\top \widehat{\Upsilon}_j \left(\widehat{b}_{ij} - \theta\right) \le \chi^2_d(1-\alpha)\right\}. \tag{2.32}$$

Theorem 2.3. *Suppose the linear Hawkes process defined in (2.2) satisfies Assumptions 2.1-2.4. Furthermore $\left(\widehat{\mu}_i, \widehat{\boldsymbol{\beta}}_i\right)$ and $\widehat{\boldsymbol{w}}_j$ satisfy Assumption 2.5 and 2.6. If $s^2\rho^2 \log p = o\left(T^{1/5}\right)$, then*

$$\sup_{x\in\mathbb{R}} \left|\mathbb{P}(\widehat{R}_T \le x) - F_d(x)\right| \le C_1 p^2 T \exp(-C_2 T^{1/5}) + C_3 s^2 \rho^2 T^{-1/5} + C_4 T^{-1/8},$$

where $C_k, k = 1, \ldots, 4$, are constants only depending on the model parameter $(\boldsymbol{\mu}, \Theta)$ and the transition kernel function.

2.6 Simulation

We illustrate finite sample properties of the proposed inference procedure through extensive simulations. We consider the linear Hawkes process with the transfer function specified in (2.6). For the connectivity matrix $\Theta = \{\beta_{ij}\}_{1\le i,j\le p}$, we consider three structures: chain, block and random, with $p = 50$ component processes. The chain structure contains nodes of component processes sequentially connected; the block structure contains 25 blocks with 2 component processes mutually connected within each block; the random structure is created by randomly assigning edges over all possible pairs of the component processes with a total sparsity of about 2%. Figure 2.1 illustrates the connectivity matrices under the three graph

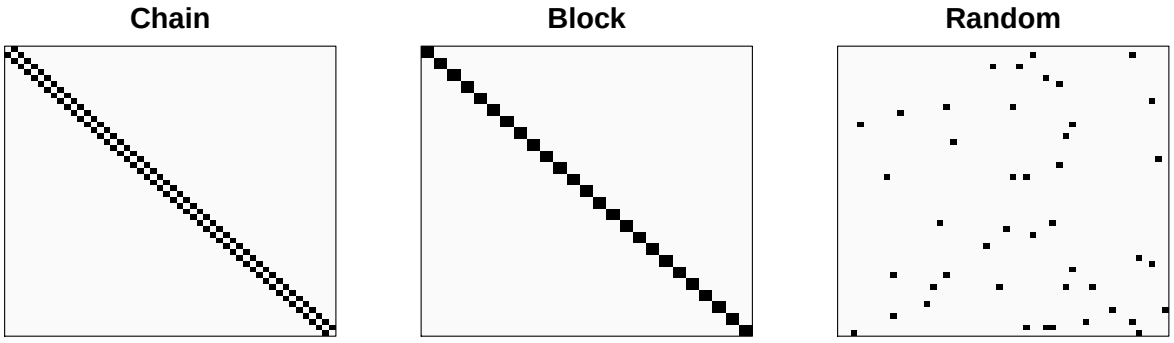


Figure 2.1: Connectivity matrices (50×50) under chain, block and random graph structures. Zero coefficients are shown in gray and non-zero coefficients are shown in black.

structures. The background intensity μ_i is set to be 0.2 and the scale of non-zero elements β_{ij} is set to be 0.3. The transfer kernel function $\kappa_j(t)$ is chosen to be $\exp(-t)$. This setting satisfies our assumptions of a stationary Hawkes process.

To assess the performance of our method, we test each of the p^2 coefficients in the connectivity matrix. We calculate the type-I error (i.e. the rejection rate among zero coefficients) and the power (i.e. the rejection rate among non-zero coefficients). We also investigate the empirical coverage of the 95% confidence intervals for zero and non-zero coefficients. We consider experiments lengths $T \in \{200, 1000, 2000\}$. As a benchmark, we compare the performance of our methods against an oracle procedure, which knows what coefficients are non-zero. Simulation is run for 200 times.

Figure 2.2 illustrates the simulation results for chain, block and random structure separately. It can be seen that as the experiment length increases, our test properly controls the type-I error rate. Moreover, the 95% confidence intervals have reasonable converge. Finally, our test also achieves power close to the oracle procedure.

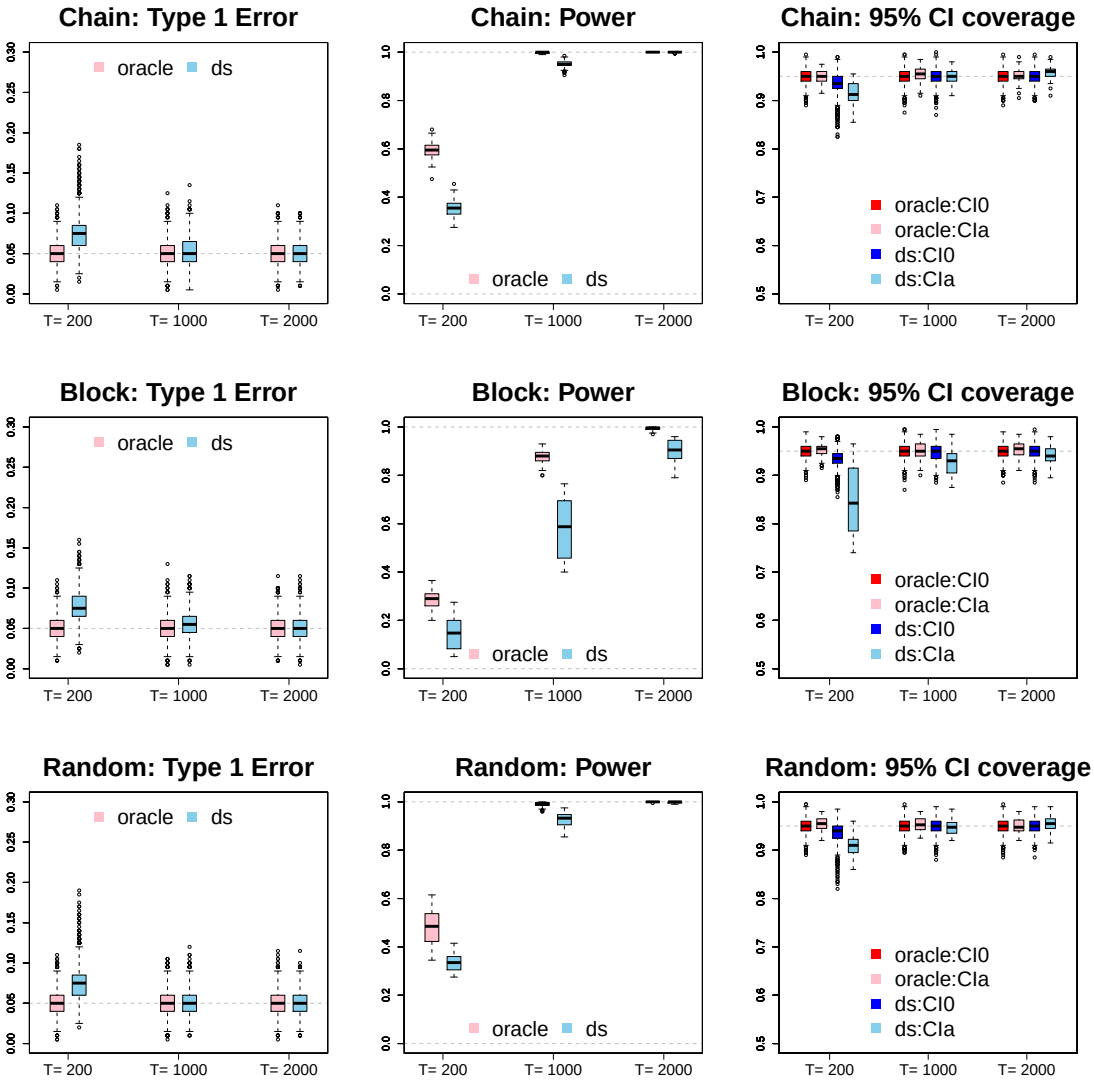


Figure 2.2: Type-I errors, powers and coverage of confidence intervals under chain, block and random graph structures. The `oracle` corresponds to the score test under the true model with known zero coefficients and `ds` corresponds to the de-correlated score test with nuisance coefficients. In the last column, CI0 and CIa correspond to the coverage of confidence intervals for zero and non-zero coefficients, respectively.

2.7 Application

We consider the task of learning the functional connectivity network among population of neurons, using the spike train data from Bolding and Franks [2018]. 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 in total collects spike train data on 23 mice. We consider the spike train data collected at two intensity levels, 0 mW/mm^2 and 20 mW/mm^2, in the subject mouse with the most neurons (25 neurons). 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. We apply our inference procedure separately for each intensity level, and obtain the estimated connectivity coefficients and the corresponding 95% confidence interval for the 25-neuron network.

Figure 2.3 illustrates the estimated connectivity coefficients that are specific to each laser condition in a graph representation, where each node represents a neuron and a directed edge indicates a statistically significant estimated connectivity coefficient. Compared with the control (0 mW/mm^2 laser) we see more condition-specific edges as laser is applied (at 20 mW/mm^2). This agrees with the observation by neuroscientists that the OB response is sensitive to the intensity level of the external stimuli [Bolding and Franks, 2018]. Figure 2.3 also shows the 95% confidence interval for 12 unique edges with largest estimated connectivity coefficients in one of the conditions. As expected, the confidence intervals corroborate with testing results and provide additional insight into differences in connectivity coefficients.

As discussed in Section 2.4, our inference procedure is asymptotically valid. That means with large enough samples, if the other assumptions in Section 2.4 are satisfied, the type-I

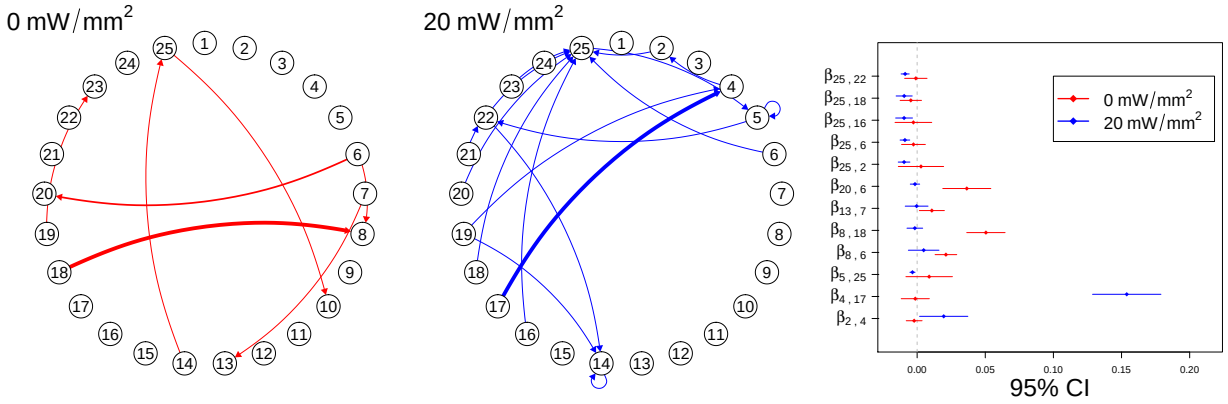


Figure 2.3: Estimated functional connectivities among neuronal populations using the spike train data from Bolding and Franks [2018]. In the condition-specific connectivity graphs, red edges are unique to 0 mW/mm^2, and blue edges are unique to 20 mW/mm^2. The last plot shows 95% confidence intervals for 12 unique edges with largest estimated connectivity coefficients in one of the conditions.

error should be controlled at the nominal level. Assessing the validity of the assumptions and estimating the type-I error in real data applications is challenging. However, we can verify the sample size requirement by estimating the type-I error rate in a 'permuted' data set where each neuron's spike train is permuted. This permutation destroys both the connections between neurons and also the temporal dependence in each neuron. As a result, the neuronal connectivity network corresponding to this data set contains no edges. Moreover, some of the other assumptions in Section 2.4 — e.g. the sparsity of $\boldsymbol{w}_j^*$ and $\boldsymbol{\beta}_i$ and the structure of the transition matrix — are trivially satisfied for this permuted data set. Thus, if the sample size is sufficient and the other assumptions are satisfied, we should not reject more than $\alpha = 0.05$ of the tests. This is indeed the case: the total rejection rate is 0.32%, suggesting that the conditions are likely satisfied.

2.8 Extensions

In practice, the transition kernel is often unknown. In addition, the background intensity can be time-varying due to the non-stationary nature of the process. In this section, we outline the steps that extend our statistical inference procedure for a flexible class of linear Hawkes process with the intensity functions defined as

$$\lambda_i(t) = \nu_i(t) + \sum_{j=1}^{p} \int_0^{t-} \omega_{ij}(t-u) dN_j(u); \quad i = 1, \dots, p, \tag{2.33}$$

where both the time-varying background intensity, $\nu_i(\cdot)$, and the transition function, $\omega_{ij}(\cdot)$ are unknown.

Following Cai et al. [2020], we approximate $\nu_i(\cdot)$ by m_0-dimensional basis $\boldsymbol{\phi}_0(t) = (\phi_{0,1}(t), \dots, \phi_{0,m_0}(t))$ such that $\nu_i(t) = \boldsymbol{\phi}_0(t)\boldsymbol{\beta}_{i,0} + r_{i,0}(t)$, where $r_{i,0}(t)$ is the approximation residual, $\boldsymbol{\beta}_{i,0} \in \mathbb{R}^{m_0}$ and $m_0 > 0$ is arbitrary. We allow the transition function $\omega_{ij}(\cdot)$ to be from a class of functions that can be approximated by m_1-dimensional basis $\boldsymbol{\phi}_1(t) = (\phi_{1,1}(t), \dots, \phi_{1,m_1}(t))$ such that $\omega_{ij}(t) = \boldsymbol{\phi}_1(t)\boldsymbol{\beta}_{i,j} + r_{i,j}(t)$, where $r_{i,j}(t)$ is the approximation residual and $\boldsymbol{\beta}_{i,j} \in \mathbb{R}^{m_1}$ and m_1 is bounded. Let $\boldsymbol{x}(t) = (\boldsymbol{x}_0(t), \dots, \boldsymbol{x}_p(t))$, where $\boldsymbol{x}_0(t) = \boldsymbol{\phi}_0(t)$ and $\boldsymbol{x}_j(t) = \int_0^{t-} \boldsymbol{\phi}_1(t-u) dN_j(u)$, $j = 1, \dots, p$. Next, we state a generalization to Assumption 2.4 that is needed to develop inference for the flexible Hawkes process (2.33).

Assumption 2.7. *For $j = 1, \dots, p$, there exists m_0, m_1, $\boldsymbol{\beta}_i = \left(\boldsymbol{\beta}_{i,0}^\top, \boldsymbol{\beta}_{i,j}^\top, \dots, \boldsymbol{\beta}_{i,p}^\top\right)^\top \in \mathbb{R}^{m_0 + m_1 p}$ such that*

$$\frac{1}{T} \int_0^T \left(\boldsymbol{x}^\top(t)\boldsymbol{\beta}_i - \lambda_i(t)\right)^2 dt = O\left(s^2 \rho^4 T^{-8/5}\right).$$

We index $\boldsymbol{\beta}_i = (\beta_{i,1}, \dots, \beta_{i,m_0+pm_1})^\top$ and $\boldsymbol{x}(t) = (x_1(t), \dots, x_{i,m_0+pm_1}(t))$. Indexing this way allows us to use our previous notations to establish the statistical inference procedure for the flexible class of linear Hawkes processes in (2.33). In particular, testing the existence of a directed edge from node j to i is equivalent to testing

$$H_0 : \left(\beta_{i,m_0+(j-1)m_1+1}, \dots, \beta_{i,m_0+jm_1}\right)^\top = \mathbf{0}. \tag{2.34}$$

Next, we show that, under the null hypothesis, our previously defined test statistic, $\widehat{U}_T$ from (2.15), converges weakly to a central χ^2 distribution with m_1 degree of freedom.

Theorem 2.4. *Suppose the flexible linear Hawkes model with its intensity function defined in (2.33) satisfies Assumptions 2.1 – 2.3 and 2.7. In addition, suppose $(\widehat{\mu}_i, \widehat{\boldsymbol{\beta}}_i)$ and $\widehat{w}_j$ satisfy Assumption 2.5 and 2.6, respectively. If $s^2\rho^2 \log p = o\left(T^{1/5}\right)$, then, under the null hypothesis in (2.34),*

$$\sup_{(\Theta,\mu)\in\Pi_0, x\in\mathbb{R}} \left|\mathbb{P}\left(\widehat{U}_T \le x\right) - F_d(x)\right| \le C_1 p^2 T \exp(-C_2 T^{1/5}) + C_3 s^2 \rho^2 T^{-1/5} + C_4 T^{-1/8}, \quad (2.35)$$

where F_d is the cdf of the χ^2-distribution with $d = m_1$ degrees of freedom and $C_k, k = 1, \ldots, 4$, are constants only depending on $\boldsymbol{\beta}_i$ and the basis function.

The proof of Theorem 2.4, given in Appendix A.1, is similar to that of Theorem 2.1 except that we need to re-examine the bound of $\widehat{S}^0_{ik} - S_{ik}$ due to our approximation of the unknown background intensity and the flexible form of transition function.

In practice, we can estimate $\boldsymbol{\beta}_i$ using a group lasso penalty in (2.16) [Yuan and Lin, 2006]. Assuming the true intensity function can be well approximated as in Assumption 2.7, a similar proof to Theorem 3 in Cai et al. [2020] shows that the estimation error, $\left\|\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i\right\|_2$, is bounded by $O_p\left(T^{-\frac{4}{5}}\right)$ with high probability, which satisfies Assumption 2.5.

Chapter 3

JOINT ESTIMATION AND INFERENCE FOR MULTI-EXPERIMENT NETWORKS OF HIGH-DIMENSIONAL POINT PROCESSES.TEX

3.1 Introduction

Multivariate point process data have become prevalent in many application areas, from finance and social networks to biology. Of prominent importance are spike train data, containing spiking times of a collection of neurons [Okatan et al., 2005]. These data, which have become more abundant thanks to the advent of calcium florescent imaging technology, are increasingly used to learn the latent brain connectivity network and glean insight into how neurons respond to external stimuli.

The Hawkes process [Hawkes, 1971] is a popular choice 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. As such, the Hawkes process offers a flexible and interpretable framework for investigating the latent network of point processes and is widely used in neuroscience applications [Brillinger, 1988, Johnson, 1996, Krumin et al., 2010, Pernice et al., 2011, Reynaud-Bouret et al., 2013, Truccolo, 2016, Lambert et al., 2018].

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. Hansen et al. [2015] and Chen et al. [2019] proposed ℓ_1-regularized estimation procedures to address this

challenge. Recently, Wang et al. [2020b] developed a high-dimensional inference procedure to characterize the sampling distribution of these estimators and their uncertainty. However, because of the complex dependence structure of the point process data, existing estimation and inference procedures require data collected over a long period to achieve satisfactory performance. Unfortunately, available data routinely consist of short stationary segments that may not satisfy these requirements. This is particularly the case in neuroscience applications, where experiments include multiple stimuli that are examined consecutively in order to investigate and contrast how neurons respond to each stimulus. For instance, Bolding and Franks [2018] apply 10 laser stimuli to a set of neurons at 8 different intensity levels ranging from 0 to 50 mW/mm^2, resulting in 80 experiments in total.

Neuronal connectivity networks corresponding to different stimuli in a sequence of experiments are expected to share many edges. This shared structure motivates joint estimation of networks from multiple experiments, which could lead to more efficient estimation of common edges. However, the networks from different experiments or conditions are also expected to contain unique edges that are in fact of primary scientific interest. For example, distinct neuronal connectivities are found under laser stimulus at different intensity levels [Bolding and Franks, 2018]. Moreover, the degree of (dis)similarity between networks from two experiments depends on the similarity of neuronal responses to the corresponding stimuli, which is generally unknown. These characteristics highlight the need for joint estimation and inference of multiple networks of high-dimensional point processes while accounting for similarities between networks, a task that is not addressed by existing methods.

Joint estimation of multiple graphical models has been considered for independent and Gaussian-distributed data [e.g. Chiquet et al., 2011, Guo et al., 2011, Danaher et al., 2014, Yajima et al., 2014, Zhu et al., 2014, Ma and Michailidis, 2016, Cai et al., 2016, Huang et al., 2018, Wang et al., 2020a]. Extensions to time- and/or space-varying networks have also been studied [e.g. Kolar et al., 2010, Qiu et al., 2016, Lin et al., 2017, Hallac et al., 2017, Yang and Peng, 2020]. However, the vast majority of existing approaches are primarily designed

for learning *two* networks or implicitly assume that networks from multiple experiments are equally similar, or the network similarity is known or implied by the spatial/temporal ordering. On the other hand, the few methods designed for joint estimation of multiple networks [Peterson et al., 2015, Saegusa and Shojaie, 2016] are specific to Gaussian graphical models and either assume the network edges are independent [Peterson et al., 2015], or assume similarities in population means in different conditions reveal similarities among precision matrices [Saegusa and Shojaie, 2016]. Moreover, existing procedures do not provide inference for the estimated networks, which is critically important in many scientific applications.

Given the paucity of methods for joint analysis of point process data from multiple experiments/conditions, in Section 3.3 we propose a data-adaptive joint estimation procedure for networks of high-dimensional point processes. The proposed approach uses estimates of cross correlations in each condition to obtain a measure of similarities among neuronal connectivity networks. While cross-correlations are widely used by neuroscientists to gain insights into neuronal interaction mechanisms [de Abril et al., 2018], they do not reveal direct interactions between neurons [Tchumatchenko et al., 2011, Reid et al., 2019]. However, they can be easily computed, even in high dimensions. We also show that they provide useful information about the overall similarities among neuronal connectivity networks, especially when used to define data-driven weights for joint estimation of multiple networks. By encouraging similarity among estimated networks, such data-driven weights offer superior finite-sample performance in selecting the edges. Extending previous work under a single experiment [Chen et al., 2019], in Section 3.4 we establish a unified non-asymptotic convergence rate for edge estimation in multiple networks of generalized Hawkes processes. Our result implies a faster convergence rate using the joint estimation approach compared to estimating networks separately under each experiment. While our method does not assume the correctness of the weights to achieve the estimation convergence, we achieve a lower estimation error when the true weights are known.

To address the need for efficient inference procedures, in Section 3.5, we develop a powerful

hierarchical testing procedure for simultaneous inference on edges of all estimated networks. While statistical inference for a single high-dimensional Hawkes processes has been recently addressed [Wang et al., 2020b], implementing such an approach directly on all estimated networks would amount to a large number of tests. As a result, the testing power can diminish when adjusting for multiple comparisons, especially when the number of experiments is large. This is particularly the case in neuroscience applications, such as that in Bolding and Franks [2018] with 80 experiments. Moreover, the tests associated with network edges under each experiment have complex dependencies. Our proposed inference framework mitigates these challenges through a multiple comparison adjustment that carries out testing along the hierarchical structure of network similarities inferred from cross-correlations. By taking advantage of this hierarchical structure, our procedure greatly reduces the number of required tests, which in turn results in increased power while tightly controlling the family-wise error rate (FWER). Unlike existing hierarchical testing procedures that rely on specific assumptions on the dependency between the tests [Yekutieli, 2008, Lynch and Guo, 2016] or particular p-values calculation along the hierarchy [Bogomolov et al., 2020], our method is free of such assumptions and can incorporate p-values flexibly calculated from any valid tests. Moreover, as we show in Section 3.5, for large and sparse networks, our inference procedure is robust to misspecification of the similarity weights, which had not been previously addressed. The advantages of our estimation and inference procedures are illustrated by analyzing simulated and real data in Sections 3.6 and 3.7, respectively.

3.2 Background: 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 $t_1 \geq 0$. 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

$$\lambda_i(t)dt = \mathbb{P}(dN_i(t) = 1 \mid \mathcal{H}_t). \tag{3.1}$$

The intensity function for the Hawkes process [Hawkes, 1971], takes the form

$$\lambda_i(t) = g_i\left(\mu_i + \sum_{j=1}^{p} (\omega_{ij} * dN_j)(t)\right), \tag{3.2}$$

where

$$(\omega_{ij} * dN_j)(t) = \int_0^{t-} \omega_{ij}(t-s)dN_j(s) = \sum_{k:t_{jk}<t} \omega_{ij}(t-t_{jk}). \tag{3.3}$$

Here, μ_i is the background intensity of unit i and $\omega_{ij}(\cdot) : \mathbb{R}^+ \mapsto \mathbb{R}$ 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. If the link function $g_i(\cdot)$ is non-linear, then $\lambda_i(\cdot)$ is the intensity of non-linear Hawkes process [Brémaud and Massoulié, 1996]. We refer to the class of Hawkes processes that allows for non-linear link functions and negative transfer functions as the *generalized Hawkes process* [Chen et al., 2019].

Motivated by applications in neuroscience [Linderman and Adams, 2014, de Abril et al., 2018], we consider a parametric transfer function $\omega_{ij}(\cdot)$ of the form

$$\omega_{ij}(t) = \beta_{ij}\kappa_j(t) \tag{3.4}$$

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*

$$x_j(t) = \int_0^{t-} \kappa_j(t-s)dN_j(s) \tag{3.5}$$

summarizes the entire history of unit j of the multivariate Hawkes processes. A commonly used example is the exponential transition kernel, $\kappa_j(t) = e^{-t}$ [Bacry et al., 2015].

In this formulation, the *connectivity coefficient* of the underlying network, β_{ij}, represents the strength of the dependence of unit i's intensity on unit j's past events. A positive β_{ij}, which implies that past events of unit j *excite* future events of unit i, is often considered in the literature [see, e.g., Bacry et al., 2015, Etesami et al., 2016]. However, we also allow for negative β_{ij} values to represent *inhibitory* effect of one unit's past events on another [Chen et al., 2019, Costa et al., 2020], which is expected in neuroscience applications [Babington, 2001].

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

$$\lambda_i(t) = g_i\left(\mu_i + \boldsymbol{x}^\top(t)\boldsymbol{\beta}_i\right). \tag{3.6}$$

3.3 Networks of Multi-Experiment Hawkes Processes

Given point process data from M experiments, let $N_i^{(m)}$ be the point process of unit i under experiment m defined on $[0, T_m]$, where $m \in \{1, \ldots, M\}$. Further, consider experiment-specific notations for the corresponding intensity function $\lambda_i^{(m)}(\cdot)$, link function $g_i^{(m)}(\cdot)$, background rate $\mu_i^{(m)}$, connectivity coefficient $\beta_{ij}^{(m)}$, transfer function $\omega_{ij}^{(m)}(\cdot)$, transition kernel function $\kappa_j^{(m)}(\cdot)$, and integrated process $x_j^{(m)}(\cdot)$. Denote the entire model parameter for unit i at experiment m as $\boldsymbol{\theta}_i^{(m)} = \left(\mu_i^{(m)}, \left(\boldsymbol{\beta}_i^{(m)}\right)^\top\right)^\top$, where $\boldsymbol{\beta}_i^{(m)} = \left(\beta_{i1}^{(m)} \quad \ldots \quad \beta_{ip}^{(m)}\right)^\top$. Let $\boldsymbol{x}^{(m)}(t) = \left(x_1^{(m)}(t) \quad \ldots \quad x_p^{(m)}(t)\right)^\top$. In addition, let $\boldsymbol{\theta}_i = \left(\left(\boldsymbol{\theta}_i^{(1)}\right)^\top \quad \ldots \quad \left(\boldsymbol{\theta}_i^{(M)}\right)^\top\right)^\top$ and $T = \sum_{m=1}^M T_m$.

Throughout the paper, we assume that the generalized Hawkes process (3.6) from each experiment is *stationary*, meaning that for all units $i \in \{1, \ldots, p\}$, the spontaneous rates $\mu_i^{(m)}$ and strengths of transition $\boldsymbol{\beta}_i^{(m)}$ are constant over the time range $[0, T_m]$ under each experiment $m \in \{1, \ldots, M\}$ [Brémaud and Massoulié, 1996]. This assumption is often satisfied in neuroscience applications, where a "white noise" period is included between

consecutive experiments—e.g., laser is turned off for one second between stimuli [Bolding and Franks, 2018].

Let $\ell(\cdot,\cdot) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ denote twice continuously differentiable loss function. To rigorously define our penalized estimator, denote the expectation of the observed outcome at $[t, t+dt)$ for unit i under experiment m as

$$f_{\boldsymbol{\theta}_i^{(m)}}\left(\boldsymbol{x}^{(m)}(t)\right) \equiv \lambda_i^{(m)}(t)dt = g_i^{(m)}\left(\mu_i^{(m)} + \left(\boldsymbol{x}^{(m)}(t)\right)^\top \boldsymbol{\beta}_i^{(m)}\right) dt.$$

Our proposed joint estimation procedure is based on solving the following optimization problem:

$$\left\{\widehat{\boldsymbol{\theta}}_i\right\}_{1\le i\le p} = \underset{\substack{\boldsymbol{\theta}_i^{(m)}\in\mathbb{R}^{p+1},\\ 1\le i\le p, 1\le m\le M}}{\arg\min} \sum_{i=1}^{p}\left\{\frac{1}{T}\sum_{m=1}^{M}\int_0^{T_m} \ell\left(dN_i^{(m)}(t), f_{\boldsymbol{\theta}_i^{(m)}}(\boldsymbol{x}^{(m)}(t))\right) + \mathcal{P}\left(\left\{\boldsymbol{\beta}_i^{(m)}\right\}_{m=1}^{M}\right)\right\}, \tag{3.7}$$

where to achieve both sparsity and similarity, we consider the penalty

$$\mathcal{P}\left(\left\{\boldsymbol{\beta}_i^{(m)}\right\}_{m=1}^{M}\right) = \underbrace{\lambda_1 \sum_{m=1}^{M}\left\|\boldsymbol{\beta}_i^{(m)}\right\|_1}_{\text{sparsity penalty}} + \underbrace{\lambda_2 \sum_{1\le m<m'\le M} w_{m,m'} \sum_{1\le i\le p}\left\|\boldsymbol{\beta}_i^{(m)} - \boldsymbol{\beta}_i^{(m')}\right\|_1}_{\text{fusion penalty}}. \tag{3.8}$$

The sparsity penalty encourages a sparse structure of the estimated networks (i.e. few non-zero $\beta_{ij}^{(m)}$, for $1 \le m \le M$). The fusion penalty [Tibshirani et al., 2005] encourages similarity in the estimated networks. The key difference of our penalty is that the extent of fusion is governed by the *data-driven weights* $w_{m,m'} \in [0,1]$ for $1 \le m, m' \le M$. A larger weight represents more similar networks in two experiments.

To construct the similarity weight between networks, we start by measuring the similarity between two matrices, $A = \{a_{ij}\}, B = \{b_{ij}\} \in \mathbf{R}^{p\times p}$, by

$$d(A,B) \equiv \sum_{1\le i,j\le p} \mathbf{1}\left(a_{ij}b_{ij} > 0\right). \tag{3.9}$$

In the context of network analysis, $d(\cdot,\cdot)$ measures the similarity of two networks according to their connectivity structures. Specifically, consider the $p\times p$ connectivity matrices, $\Theta^{(m)} =$

$\left\{\beta_{ij}^{(m)}\right\}$ and $\Theta^{(m')} = \left\{\beta_{ij}^{(m')}\right\}$, for networks under condition m and m'. The network similarity is then given by

$$d^o_{m,m'} \equiv d(\Theta^{(m)}, \Theta^{(m')}). \tag{3.10}$$

We call $d^o_{m,m'}$ the *oracle network similarity* because, in practice, the true connectivity matrices are generally unknown. Instead, we propose a surrogate measure based on cross-covariances [Hawkes, 1971]:

$$d^c_{m,m'} \equiv d(V^{(m)}, V^{(m')}), \tag{3.11}$$

where $V^{(m)} = \{V_{ij}^{(m)}\}$ and $V^{(m')} = \{V_{ij}^{(m')}\}$ are cross-covariance matrices for networks under condition m and m'. While they only represent marginal temporal associations between component processes, the cross-covariances can be easily computed, even in high-dimensional settings. For mutually-exciting networks, there exists a one-to-one mapping between the cross-covariance matrix and the connectivity matrix [see Bacry and Muzy, 2016, Theorem 1], and the set of edges represented by the non-zero cross-covariance covers the true edge set [see Chen et al., 2017, Lemma 1]. In practice, we estimate the cross-covariance matrix using its empirical estimate, $\widehat{V}_{ij}^{(m)} = \left\{\widehat{V}_{ij}^{(m)}\right\}$ [Chen et al., 2017]. Given a pre-specified threshold $\kappa > 0$, the thresholded sample cross-covariance matrix is given by

$$\widetilde{V}_{ij}^{(m)} = \widehat{V}_{ij}^{(m)} \mathbb{1}\left(\left|\widehat{V}_{ij}^{(m)}\right| > \kappa\right).$$

We then obtain the *empirical network similarity* as

$$d^e_{m,m'} \equiv d\left(\widetilde{V}_{ij}^{(m)}, \widetilde{V}_{ij}^{(m')}\right). \tag{3.12}$$

Following the consistency of the sample cross-covariance estimator to the true cross-covariance [see Chen et al., 2019, Corollary 1], the empirical network similarity is consistent to the network similarity based on the true cross-covariance, assuming that a minimum signal strength, κ, of the true cross-covariances exists. In practice, to put these sample cross-covariances in a comparable range, we transform the covariance values into z-scores using Fisher's transformation and obtain the corresponding p-values. We claim an edge in this cross-covariance

network if the p-value is below a pre-specified threshold—e.g., 0.1, where such threshold can be chosen, e.g., to achieve a certain level of sparsity in a network [Chen et al., 2017]. The *similarity weights* used in our algorithm are obtained by normalizing the empirical similarity measure by their total so that the weights are always between 0 and 1; that is,

$$w_{m,m'} \equiv \frac{d^e_{m,m'}}{\sum_{1\le k'\neq k\le M} d^e_{k,k'}} \in [0,1], \tag{3.13}$$

for $1 \le m \neq m' \le M$.

Fusion penalty has been used in previous work but with a natural ordering of the conditions defined according to location or network structure [Tang and Song, 2016, Wang et al., 2016]. Given that no ordering necessarily exists between experiments, we adopt a different computation strategy based on the smoothing proximal gradient descent algorithm [Chen et al., 2012] to solve (3.7). Implementation details are given in Appendix B.1.

Two tuning parameters are involved in (3.7) to control the sparsity and the similarity of the networks between experiments. To select the tuning parameters, we use an eBIC-type criterion [Chen and Chen, 2008, Cai et al., 2020]. Let $\widehat{\boldsymbol{\Theta}}_{\lambda_1,\lambda_2} = \left\{\widehat{\boldsymbol{\theta}}^{(m)}(\lambda_1,\lambda_2)\right\}_{m=1}^{M}$ be the model parameter estimates using pre-specified tuning parameters (λ_1,λ_2), then

$$\text{eBIC}\left(\widehat{\boldsymbol{\Theta}}_{\lambda_1,\lambda_2}\right) = \sum_{m=1}^{M}\sum_{i=1}^{p}\left\{2\ell\left(dN_i^{(m)}(t), f_{\boldsymbol{\theta}_i^{(m)}(\lambda_1,\lambda_2)}\left(\boldsymbol{x}^{(m)}(t)\right)\right) + s_i^{(m)}\log T_m + 2\gamma\log\binom{p}{s_i^{(m)}}\right\},$$

where $s_i^{(m)} = \left\|\widehat{\boldsymbol{\beta}}_i^{(m)}(\lambda_1,\lambda_2)\right\|_0$ and $\binom{p}{s_i^{(m)}}$ of the last term is the binomial coefficient.

For ease of presentation in later sections, let $d_{m,m'} = w_{m,m'}(\boldsymbol{e}_m^\top - \boldsymbol{e}_{m'}^\top)\otimes I_0 \in \mathbb{R}^{(p+1)\times(p+1)M}$ and denote $D = (d_{1,2},\ldots,d_{m,m'},\ldots,d_{M-1,M})^\top \in \mathbb{R}^{\binom{M}{2}(p+1)\times(p+1)M}$, where $\boldsymbol{e}_m$ is canonical basis in $\mathbb{R}^M$, $I_0 = \begin{pmatrix} 0 & 0 \\ 0 & I_p \end{pmatrix} \in \mathbb{R}^{(p+1)\times(p+1)}$ and $I_p \in \mathbb{R}^{p\times p}$ is the identity matrix. Then, the fusion penalty can be written in a compact form

$$\sum_{1\le m<m'\le M} w_{m,m'}\sum_{1\le i\le p}\left\|\boldsymbol{\beta}_i^{(m)} - \boldsymbol{\beta}_i^{(m')}\right\|_1 = \|D\boldsymbol{\theta}\|_1\,. \tag{3.14}$$

3.4 Asymptotics

In this section we establish consistent parameter estimation and recovery of the latent networks over multiple experiments. Proofs for the results in this section are given in Appendix B.2.

We start by stating our assumptions. Throughout, we denote the maximum and minimum eigenvalues of a square matrix A as $\Lambda_{\max}(A)$ and $\Lambda_{\min}(A)$, respectively.

Assumption 3.1. *The link functions, $g_i^{(m)}(\cdot)$, are first-order differentiable such that $|\nabla g_i^{(m)}(\cdot)| \leq \alpha_i^{(m)}$, for $1 \leq i \leq p, 1 \leq m \leq M$. Further, let $\Omega^{(m)}$ be a $p \times p$ matrix whose entries are $\Omega_{ij}^{(m)} = \alpha_i^{(m)} \int_0^\infty |\omega_{ij}^{(m)}(\Delta)| d\Delta$, for $1 \leq i, j \leq p, 1 \leq m \leq M$. Then, there exists a constant γ_Ω such that $\Lambda_{\max}((\Omega^{(m)})^\top \Omega^{(m)}) \leq \gamma_\Omega^2 < 1$, for $1 \leq m \leq M$.*

Assumption 3.1 is necessary for stationarity of a Hawkes process under a specific experiment [Chen et al., 2019]. The constant γ_Ω does not depend on the dimension p. For any fixed p, Brémaud and Massoulié [1996] show that given this assumption the intensity process of the form (3.2) is stable in distribution and, thus, a stationary process exists. Since the connectivity coefficients of interest are ill-defined without stationarity, this assumption provides the necessary context for our joint estimation framework.

Assumption 3.2. *There exists $\lambda_{\min}$ and $\lambda_{\max}$ such that*

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

for all $i = 1, \ldots, p$ and $m = 1, \ldots, M$.

Assumption 3.2 requires that the intensity rate is strictly bounded, which prevents degenerate processes for all units of the multivariate Hawkes processes. This assumption has been considered in the previous analysis of Hawkes process [Chen et al., 2019, Wang et al., 2020b].

Assumption 3.3. *The transition kernel $\kappa_i^{(m)}(t)$ is bounded and integrable over $[0, T_m]$, for $1 \le i \le p$ and $1 \le m \le M$.*

Assumption 3.3 implies that the integrated process $x_i^{(m)}(t)$ in (3.5) is bounded. Assumptions 3.2 and 3.3 imply that the model parameters are bounded, which is often required in time-series analysis [Safikhani and Shojaie, 2020].

Assumption 3.4. *There exists constants $\rho_r \in (0,1)$ and $0 < \rho_c < \infty$ such that*

$$\max_{1\le i\le p} \sum_{j=1}^{p} \Omega_{ij}^{(m)} \le \rho_r \qquad \text{and} \qquad \max_{1\le j\le p} \sum_{i=1}^{p} \Omega_{ij}^{(m)} \le \rho_c,$$

for $m = 1, \ldots, M$.

Assumption 3.4 requires maximum in- and out- intensity flows to be bounded, which helps in bounding the eigenvalues of the cross-covariance of $\boldsymbol{x}^{(m)}(t)$ [Wang et al., 2020b]. A similar assumption is also considered by Basu and Michailidis [2015] in the context of VAR models.

Define the set of active indices as $S_i^{(m)} = \{j : \beta_{ij}^{(m)} \neq 0, 1 \le j \le p\}$, and $d_i^{(m)} = |S_i^{(m)}|$, let $d^* \equiv \max_{1\le m\le M, 1\le i\le p} d_i^{(m)}$, and $S_i = \bigcup_{m=1}^{M} S_i^{(m)}$. Also denote the set of dissimilar experiment indices as $\widetilde{S}_i = \left\{(j,m) : \beta_{ij}^{(m)} \neq \beta_{ij}^{(m')}, \exists\, m' \neq m \in \{1,\ldots,M\}, \text{ for } 1 \le j \le p\right\}$ and then the dissimilarity index $r^* \equiv \max_{1\le i\le p} |\widetilde{S}_i|$. When $M = 1$, let $r^* = 0$. With a little abuse of notation, we use $m \in \widetilde{S}_i$ if $\exists\, j$ such that $(j,m) \in \widetilde{S}_i$. In addition, for $\widehat{\boldsymbol{\theta}}_i^{(m)}$ defined in (3.7), let $\Delta_i^{(m)} = \widehat{\boldsymbol{\theta}}_i^{(m)} - \boldsymbol{\theta}_i^{(m)}$ and $\Delta_i = \left(\left(\Delta_i^{(1)}\right)^\top, \ldots, \left(\Delta_i^{(M)}\right)^\top\right)^\top \in \mathbb{R}^{(p+1)M}$. With these notations, Δ_{S_i} and $\Delta_{\widetilde{S}_i}$ are vectors that collect the estimation error on those non-zero $\beta_{ij}^{(m)}$ that are non-zero, and those varying over the experiments, respectively.

Let

$$\mathcal{C} = \left\{\Delta \in R^{M(p+1)} : \frac{1}{\sqrt{M}}\|\Delta_{S_i^c}\|_1 + 2\|D_{\cdot,\widetilde{S}_i^c}\Delta_{\widetilde{S}_i^c}\|_1 \le \frac{3}{\sqrt{M}}\|\Delta_{S_i}\|_1 + 2\|D_{\cdot,\widetilde{S}_i}\Delta_{\widetilde{S}_i}\|_1\right\},$$

where $D_{\cdot,\widetilde{S}_i}$, $D_{\cdot,\widetilde{S}_i^c}$ are columns of D corresponding to index set $\widetilde{S}_i$, $\widetilde{S}_i^c$, respectively.

Next, we introduce two conditions that are required on $\ell(t;\boldsymbol{\theta}_i^{(m)}) \equiv \ell\left(dN_i^{(m)}(t), f_{\boldsymbol{\theta}_i}^{(m)}(\boldsymbol{x}^{(m)}(t))\right)$.

The first condition is known as the restricted strong convexity (RSC) [Negahban and Wainwright, 2010]. The constraint set, $\mathcal{C}$, is constructed specifically for the penalty in (3.8), which is geometrically decomposable [Lee et al., 2015]. This construction links the estimation error bound to both the sparsity and dissimilarity of the multi-experiment networks. In Corollary 3.1, we show that this condition is satisfied for linear Hawkes process or generalized Hawkes process with exponential-link under Assumption 3.2. We also show that the second condition is satisfied with common loss functions such as the least square loss and the negative-likelihood loss.

Condition 3.1. *There exist constants $\eta, c, C > 0$ such that, for $1 \le i \le p$,*

$$\mathbb{P}\left(\min_{\Delta\in\mathcal{C}} \frac{1}{T}\sum_{m=1}^{M} \Delta_i^\top \left(\int_0^{T_m} \nabla^2 \ell(t;\boldsymbol{\theta}_i^{(m)})\right) \Delta_i \ge \eta\|\Delta\|_2^2\right) \ge 1 - cp^2 \sum_{m=1}^{M} T_m \exp(-CT_m^{1/5}).$$

Condition 3.2. *There exist $c, C > 0$ such that, for $1 \le i \le p$,*

$$\mathbb{P}\left(\left\|-\frac{1}{T}\sum_{m=1}^{M}\int_0^{T_m} \nabla \ell(t;\boldsymbol{\theta}_i^{(m)})\right\|_\infty \le CT^{-2/5}\right) \ge 1 - cpM\exp(-T^{1/5}).$$

Theorem 3.1. *Assume the p-variate Hawkes processes for all M experiments—with each component process has its intensity function defined in (3.2)—satisfy Assumptions 3.1– 3.4. In addition, suppose Conditions 3.1 and 3.2 are met, and $\log p = o(\min T_m^{1/5})$ and $(d^* \vee r^*)^{1/4} = o(T^{1/5})$, where $T = \sum_{m=1}^M T_m$. Then, taking $\sqrt{M}\lambda_1 = \lambda_2 = O(T^{-2/5})$,*

$$\|\Delta_i\|_2 = O_p\left(\frac{1}{\eta}T^{-2/5}\left(3\sqrt{d^*} + 2\phi_{\widetilde{S}_i}\sqrt{r^*}\right)\right), \quad 1 \le i \le p, \tag{3.15}$$

with probability at least $1 - c_1p^2\sum_{m=1}^M T_m\exp(-c_2T_m^{1/5})$, where $\phi_{\widetilde{S}_i} = \max_{m\in\widetilde{S}_i}\sum_{m'\neq m\in\widetilde{S}_i} w_{m,m'}$, and $c_1, c_2 > 0$ depend on the model parameters and the transition kernel.

Remark 3.1. *The term $2\phi_{\widetilde{S}_i}\sqrt{r^*}$ comes from the error induced by the fusion penalty in the joint estimation. In an ideal case, when the true network skeletons are known, we can reparameterize the model so that it includes the same parameters that stay the same for*

all experiments and the experiment-specific parameters that vary across experiments. This greatly reduces the number of parameters involved, particularly when the networks are sparse and similar. Under such reparameterized model, the fusion penalty is no longer needed in the estimation step, which may thus lead to a tighter estimation error bound. However, knowing the true network skeletons is almost impossible in practice. Therefore, in this work, we focus on developing the proposed estimation procedure under the unknown network-skeleton case.

The error bound in Theorem 3.1 involves the network sparsity index d^* and the dissimilarity index r^*, which suggests a low prediction error bound associated with sparse networks that are similar between experiments. The network size, p, is allowed to grow much faster than the minimum length of the experiments, as long as $\log p = o(\min T_m^{1/5})$. The number of experiments, M, is also allowed to grow in a faster rate compared to the length of the experiments, as long as $\log \sum_{i=1}^{M} T_m = o(\min T_m^{1/5})$. Such condition is likely met in practice where the total number of experiment is usually not too large and the lengths of experiments are similar—e.g. Bolding and Franks [2018] conducted 80 experiments in total where each experiment collected data in 30kHz over 10 seconds. Compared with methods that separately estimate the network under each experiment, our proposed method achieves a faster convergence rate in the order of $T = \left(\sum_{m=1}^{M} T_m\right)^{-2/5}$ instead of $\min_{m=1,\ldots,M} T_m^{-2/5}$ in the estimation.

With normalized similarity weights,

$$\phi_{\widetilde{S}_i} = \max_{m \in \widetilde{S}_i} \sum_{m' \neq m \in \widetilde{S}_i} w_{m,m'} \leq \max_{1 \leq m \leq M} \sum_{1 \leq m' \neq m \leq M} w_{m,m'} \leq 1.$$

Thus, the estimation error is always bounded by $O_p\left(\frac{1}{\eta} T^{-2/5}\left(3\sqrt{d^*} + 2\sqrt{r^*}\right)\right)$ regardless of the choice of weights. When the weights are correctly specified—i.e., $w_{m,m'}$s are small for pairs of networks that are different, the error bound is improved. For example, consider networks under 3 conditions where the first two are exactly the same and the third is completely different. With oracle similarity weights—i.e., $w_{1,2} = 1, w_{1,3} = w_{2,3} = 0$,

$\phi_{\widetilde{S}_i} = w_{1,3} + w_{2,3} = 0$. In contrast, using uninformative weights that treat all networks equal—i.e., $w_{1,2} = w_{1,3} = w_{2,3} = \frac{1}{3}$, $\phi_{\widetilde{S}_i} = w_{1,3} + w_{2,3} = \frac{2}{3}$.

Corollary 3.1. *Assume the setting of Theorem 3.1 and in particular Assumptions 3.1– 3.4. Then, the result (3.15) holds for*

1. *linear Hawkes model estimated using the least square loss, $\ell(a,b) = (a-b)^2$;*
2. *non-linear Hawkes model, with exponential-link function, $g(\cdot) = \exp(\cdot)$, estimated using the negative log likelihood loss, $\ell(a,b) = -a\log(b) + b$.*

To establish the edge selection consistency, we next introduce an additional assumption.

Assumption 3.5. *There exists $\tau > 0$ such that*

$$\min_{\beta_{ij}^{(m)} \in \bigcup_{m=1}^{M} S_i^{(m)}} \beta_{ij}^{(m)} \geq \beta_{min} > 2\tau,$$

for $1 \leq i \leq p$.

Assumption 3.5 is known as the β-min condition [Bühlmann and van de Geer, 2011] and requires sufficient signal strength for the true edges in order to distinguish them from 0. We then construct *thresholded connectivity estimator*

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

Such thresholded estimator is often used for variable selections in high dimensions [van de Geer et al., 2011]. Let the estimated edge set $\widehat{S}^{(m)} = \left\{(i,j) : \widetilde{\beta}_{ij}^{(m)} \neq 0, 1 \leq i,j \leq p\right\}$ and the true edge set $S^{(m)} = \left\{(i,j) : \beta_{ij}^{(m)} \neq 0, 1 \leq i,j \leq p\right\}$. Next, we show that the estimated edge set consistently recovers the true edge set.

Theorem 3.2. *Under the same conditions in Theorem 3.1, suppose Assumption 3.5 is also satisfied with $\tau = O\left(\frac{1}{\eta} T^{-2/5}\left(3\sqrt{d^*} + 2\max_{1\leq i\leq p} \phi_{\widetilde{S}_i}\sqrt{r^*}\right)\right)$. Then,*

$$\mathbb{P}\left(\bigcap_{m=1}^{M}\left\{\widehat{S}^{(m)} = S^{(m)}\right\}\right) \geq 1 - c_1 p^2 \sum_{m=1}^{M} T_m \exp(-c_2 T_m^{1/5}),$$

where $\phi_{\widetilde{S}_i} = \max_{m \in \widetilde{S}_i} \sum_{m' \neq m \in \widetilde{S}_i} w_{m,m'}$, *and* $c_1, c_2 > 0$ *depending on the model parameters and the transition kernel.*

3.5 Inference

In this section, we develop a hierarchical testing procedure for edges of all networks, where the hierarchy is guided by the network similarity discussed in Section 3.3. Taking advantage of the hierarchical structure over the tests, our procedure greatly reduces the number of tests involved and thus increases the power. At the same time, it tightly controls the family-wise error rate (FWER) under arbitrary dependencies among tests. The results from this section are proved in Appendix B.4.

To guide the hierarchy of our testing procedure for M experiments, we start by building a M-level binary tree, where the left child node is always a leaf node. Here the level of a node equals to 1 + the number of the connections between the node and the root of the tree. For example, the root is on level 1 and the two nodes at the bottom are on level M. We assign a set of experiment indices to each node of the tree starting from the bottom nodes. Specifically, the right node at the bottom is assigned with an index set, $\mathcal{L}_{R,M} = \{t\}$, where t corresponds to the experiment that has the fewest edges. The left node is assigned with the index set, $\mathcal{L}_{L,M} = \{s\}$, where s, corresponds to the experiment whose underlying network is most similar to that of experiment t; i.e,

$$s = \underset{s \in \mathcal{L}_1 / \mathcal{L}_{R,M}}{\arg\max} \max_{t \in \mathcal{L}_{R,M}} d(s,t), \tag{3.16}$$

where $\mathcal{L}_1 = \{1, \ldots, M\}$ and $d(s,t)$ is the similarity between two experiments, e.g., according to their network structures. We then merge the index sets of both nodes into one set and assigned it to the right child node at the upper level—i.e., $\mathcal{L}_{R,M-1} = \mathcal{L}_{R,M} \bigcup \mathcal{L}_{L,M} = \{s,t\}$. Next, we assign the left node at level $M-1$ the index set with a single element corresponding to the experiment that is closest to $\mathcal{L}_{R,M-1}$ as defined in (3.16). We repeat this procedure until we reach the root of the binary tree where the root is always assigned with the total

index set $\mathcal{L}_1$.

For example, consider networks under 4 conditions with the number of common edges listed in the similarity matrix below:

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 100 & 40 & 10 & 5 \\ . & 80 & 30 & 10 \\ . & . & 50 & 40 \\ . & . & . & 45 \end{pmatrix} \end{array} .$$

Here, 40 edges are shared between Conditions 3 and 4, and 4 edges are shared between Conditions 1 and 4. Also, Condition 4 has the fewest edges among all conditions. The 4-level binary tree built according to the network similarity matrix above is illustrated in Figure 3.1.

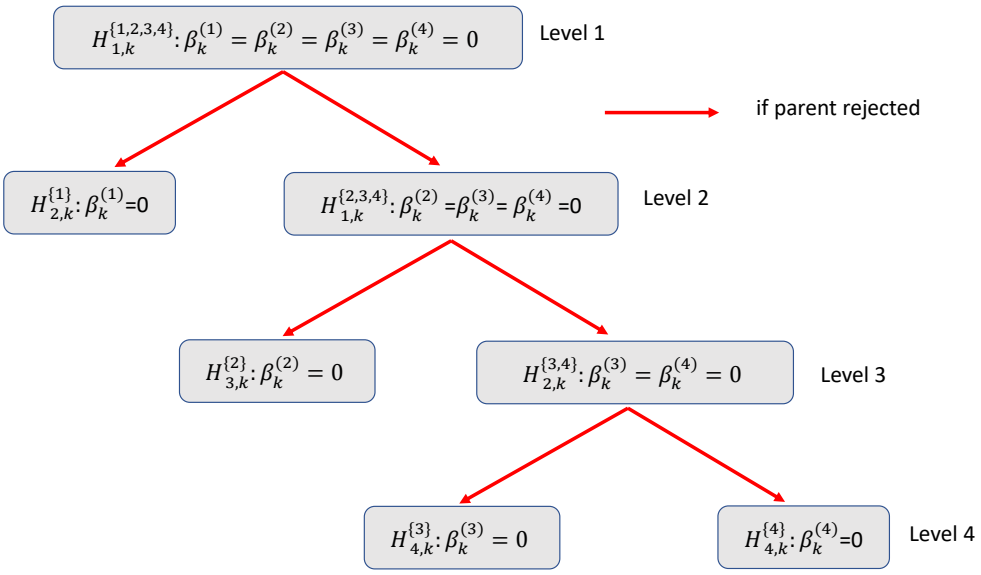


Figure 3.1: Illustration on hierarchical testing procedure

Let $\mathcal{D}$ be the binary tree created over M experiments as above. We call $\mathcal{D}$ an *oracle* or *empirical* binary tree if the tree is built using the oracle network similarity—i.e., $d(s,t) = d^o_{s,t}$ in (3.10), or the empirical network similarity—i.e., $d(s,t) = d^e_{s,t}$ in (3.12). Next we describe the hierarchical testing procedure with its hierarchy guided by the binary tree built above.

For ease of notations, we index the p^2 edge coefficients of a p-variate network at condition m as $\beta_1^{(m)}, \ldots, \beta_{p^2}^{(m)}$, for $m = 1, \ldots, M$. Besides the binary tree, $\mathcal{D}$, the algorithm takes in the critical values at each level of the tree, $\{\alpha_l\}_{l=1}^M$. Our procedure is implemented separately for each coefficient. For a specific coefficient indexed $k \in \{1, \ldots, p^2\}$, at each node of the hierarchy, we assign a test on a global hypothesis that all the edge coefficients corresponding to the experiments indexed by the index set of the node are 0–i.e., $H^{\mathcal{L}}_{l,k} : \beta_k^{(m)} = 0, m \in \mathcal{L}$ where $\mathcal{L}$ takes $\mathcal{L}_1$, $\mathcal{L}_{L,l}$, $\mathcal{L}_{R,l}$ when the node is the root of the binary tree, the left or right child node at level l, respectively. Our procedure starts testing the hypothesis at the root of the tree. If it is rejected, we move down to the next level of the tree and separately test the hypotheses assigned to each of the child nodes.

For example, consider inference on all edge coefficients under $M = 4$ conditions, where the hierarchy guided by the binary tree in Figure 3.1. Let a $p^2 \times M$ matrix I be the matrix of rejection indicators. Separately for an edge coefficient with index k, we start from the root of the tree and test $H^{\{1,2,3,4\}}_{1,k} : \beta_k^{(1)} = \beta_k^{(2)} = \beta_k^{(3)} = \beta_k^{(4)} = 0$. If $H^{\{1,2,3,4\}}_{1,k}$ is rejected, then we move down to the next level and separately test $H^{\{1\}}_{2,k} : \beta_k^{(1)} = 0$ and $H^{\{2,3,4\}}_{2,k} : \beta_k^{(2)} = \beta_k^{(3)} = \beta_k^{(4)} = 0$. If we reject $H^{\{1\}}_{2,k}$ then set $I_{k,1} = 1$; otherwise, $I_{k,1} = 0$. If we reject $H^{\{2,3,4\}}_{2,k}$ then we move down to test $H^{\{2\}}_{3,k}$ and $H^{\{3,4\}}_{3,k}$; otherwise, $I_{k,2} = I_{k,3} = I_{k,4} = 0$.

We summarize our *hierarchical testing* procedure in Algorithm 1.

Algorithm 1 Hierarchical Testing Procedure for Multi-Experiment Networks

input: $\mathcal{D}$, $\{\alpha_l\}_{l=1}^{M}$;

initialization: rejection matrix $I = \{I_{k,m} = 0\}_{1\le k\le p^2, 1\le m\le M}$;

for $k = 1, \ldots, p^2$ **do**

root: calculate p-value, $P_k^{\mathcal{L}_1}$, on $H_{1,k}^{\mathcal{L}_1} : \beta_k^{(m)} = 0, m \in \mathcal{L}_1 = \{1, \ldots, M\}$;

if $P_k^{\mathcal{L}_1} \le \alpha_1$ and $M > 1$ **then**

for $l = 2, \ldots, M$ **do**

left node: calculate p-value, $P_k^{\mathcal{L}_{L,l}}$, on $H_{l,k}^{\mathcal{L}_{L,l}} : \beta_i^{(m)} = 0, m \in \mathcal{L}_{L,l}$; if $P_k^{\mathcal{L}_{L,l}} \le \alpha_l$, set $I_{k,m} = 1$, for $m \in \mathcal{L}_{L,l}$;

right node: calculate p-value, $P_k^{\mathcal{L}_{R,l}}$, on $H_{m,k}^{\mathcal{L}_{R,l}} : \beta_i^{(m)} = 0, m \in \mathcal{L}_{R,l}$; stop the loop over l if $P_k^{\mathcal{L}_{R,l}} > \alpha_l$; otherwise, set $I_{k,m} = 1$ for $m \in \mathcal{L}_{R,l}$ if $l = M$;

end for

end if

end for

return I

Note that our procedure requires p-values calculated based on a valid test. In the context of high-dimensional point processes, recently, Wang et al. [2020b] developed a de-correlated score statistics in testing $\beta_{ij}^{(m)} = 0$. Specifically, the method calculates the *de-correlated score* statistics

$$S_{ij}^{(m)} = \frac{1}{T_m} \int_0^{T_m} \epsilon_i^{(m)}(t) \widetilde{x}_j^{(m)}(t) dt,$$

where $\epsilon_i^{(m)}(t) = \frac{dN_i^{(m)}(t)}{dt} - \lambda_i^{(m)}(t)$, and $\widetilde{x}_j^{(m)}$ is called de-correlated column that is $x_j^{(m)}(t)$ after removing its projection on the other columns $x_{-j}^{(m)}(t)$. It has been shown that $V_{ij}^{(m)} \to_d \mathcal{N}(0,1)$, where $V_{ij}^{(m)} = \sqrt{T}\left(\Upsilon_j^{(m)}\right)^{-1/2} S_{ij}^{(m)}$ and $\Upsilon_j^{(m)} = \frac{1}{T}\int_0^T \left(\widetilde{x}_j^{(m)}(t)\right)^2 dt$. Therefore, for example, to test the global hypothesis at the root, $H_{1,k}^{\mathcal{L}_1}$, we construct the test statistics

$$U_{ij} = \sum_{m\in\mathcal{L}_1} \left(V_{ij}^{(m)}\right)^2 \to_d \chi_M^2.$$

Then, the corresponding p-value is approximately $\mathbb{P}\left(\chi^2_M \geq U_{ij}\right)$. By its construction, such test is powerful when many $\beta_{ij}^{(m)} \neq 0, m \in \mathcal{L}_1$.

When the number of non-zero edges is believed not to be many, an alternative test choice is to construct U_{ij} as the maximum of $\left(V_{ij}^{(m)}\right)^2$; that is,

$$U_{ij} = \max_{m \in \mathcal{L}_1} \left(V_{ij}^{(m)}\right)^2.$$

It has been shown that as $M \to \infty$,

$$\begin{aligned} & a_M(U_{ij} - b_M) \to_d G, \\ & a_M = 1/2, \quad b_n = 2(\ln M + (d-1)\ln\ln M - \ln\Gamma(d)), \end{aligned}$$

where G follows Gumble distribution—$\mathbb{P}(G \leq x) = \exp(-\exp(-x)), x \in \mathbb{R}$ [see Page 156 Embrechts et al., 1997]. By its construction, such test requires a large M to obtain valid p-values calculated from the limiting distribution—i.e., the Gumble distribution.

Next we show that our proposed procedure controls the family-wise error rate (FWER) under arbitrary dependencies between tests.

Theorem 3.3. *The hierarchical testing procedure in Algorithm 1 with the oracle binary tree and critical value*

$$\alpha_l = \frac{\alpha}{p^2}\frac{M-l+1}{M}, \quad l = 1, \ldots, M,$$

controls the FWER for testing all p^2M hypotheses at α.

For a single experiment, our hierarchical testing procedure with the α_l chosen in Theorem 3.3 reduces to the usual Bonferroni correction. However, in multi-experiment settings, our procedure uses a less stringent critical value than that used by Bonferroni correction, as $\frac{\alpha}{p^2}\frac{M-l+1}{M} < \frac{\alpha}{Mp^2}$ for $l < M$. This makes the procedure more powerful in practice, particularly for sparse networks when most tests are carried out at shallow levels of the dendrogram. The procedure can also be applied to any subset, $\mathcal{J}$, of the p^2 edges by taking $\alpha_l = \frac{\alpha}{|\mathcal{J}|}\frac{M-l+1}{M}$ in Theorem 3.3.

Unlike existing hierarchical testing methods [e.g. Yekutieli, 2008, Lynch and Guo, 2016] that control the error rate over all the tests involved, our proposed method controls the error rate among the hypothesis associated only with the leaves of the tree, which is exactly the set of hypotheses of our interest.

Theorem 3.3 assumes an oracle binary tree while in practice, the data-driven similarity in Section 3.3 can be used to create an empirical binary tree. The next result shows that even when the oracle binary tree or its proxy are not available, our procedure is robust to misspecification of the binary tree for multiple large and sparse networks.

Theorem 3.4. *The hierarchical testing procedure in Algorithm 1 using a binary tree built based on arbitrary network similarity and critical values*

$$\alpha_l = \frac{\alpha}{p^2}\frac{M-l+1}{M}, \quad l = 1, \ldots, M,$$

controls FWER at $\alpha\left(1 + \frac{d^* M(M-1)}{2p}\right)$.

Theorem 3.4 implies that under $d^* M(M-1) = o(p)$, FWER is controlled at $\alpha\,(1 + o(1))$ with arbitrary hierarchy guided by the binary tree used in our hierarchical testing procedure. Such condition is met when the underlying network is sparse—i.e, $d^* = o(p)$ and the number of experiment M is $o(\sqrt{p})$. Seen from our technical proof in Appendix B.4, such robustness property comes from the unique construction of the binary tree where the left child node is always a leaf.

3.6 Simulations

We first investigate the edge selection performance of the proposed joint estimation procedure. We consider networks of $p = 100$ linear Hawkes processes under $M = 3$ experiments. Network 1 consists of 20 circles, Network 3 consists of 20 stars, and Network 2 is a mix of 18 circles and 2 stars (see Figure 3.2). The edge coefficients of circles and stars are set to be 0.3 and 0.6, respectively. The background intensity is set to 0.2 for all nodes in all experiments.

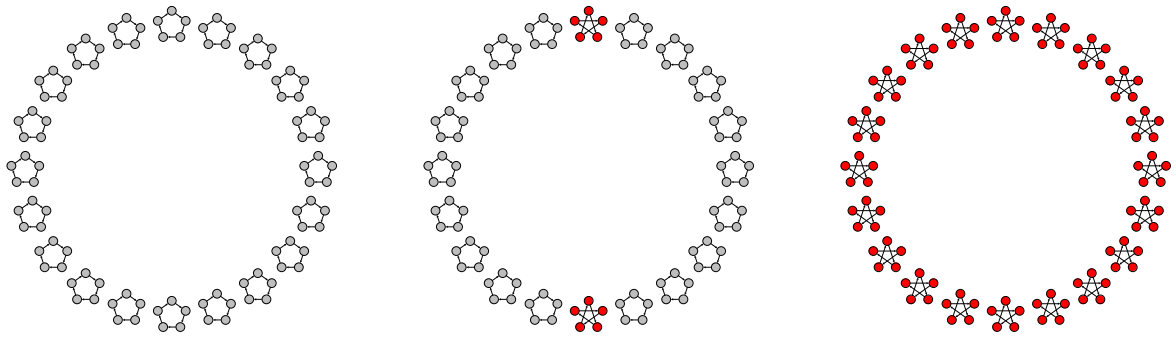
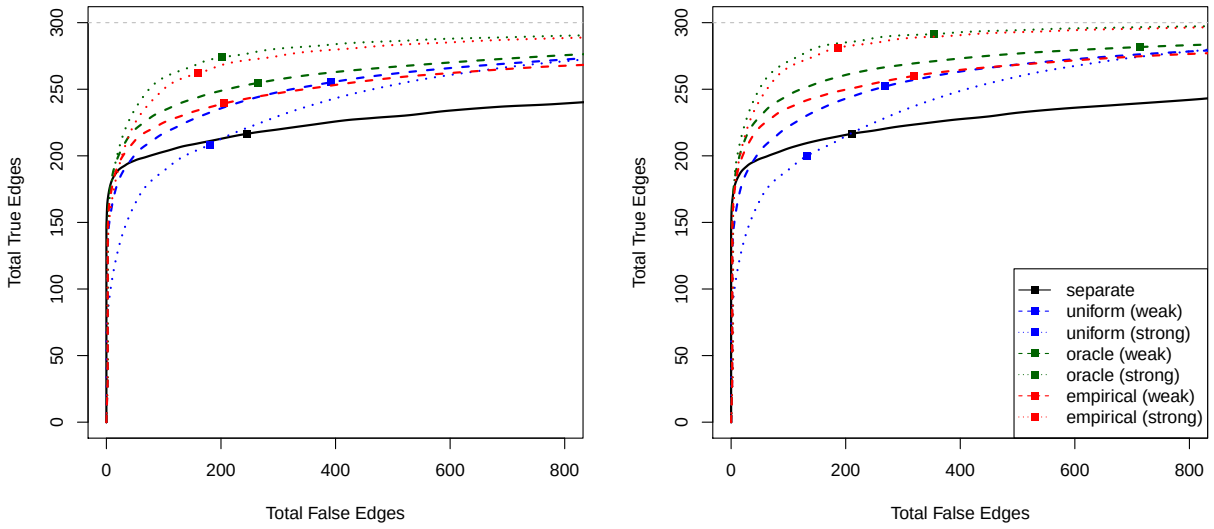


Figure 3.2: Networks of $p = 100$ processes under $M = 3$ experiments. Network 1 (left) consists of 20 circles, Network 3 (right) consists of 20 stars, and Network 2 (middle) is a mix of 18 circles and 2 stars.

The transfer kernel function is chosen to be $\exp(-t)$, for all nodes in all experiments. This setting satisfies our assumptions of a stationary Hawkes process under each experiment. The time periods, T_m, are 200, 500, 300 for $m = 1, 2, 3$, respectively.

We consider three weight choices: informative weights based on the true networks (oracle) and the cross-correlation (empirical), and uniform weights that treat all networks equal. We consider weak and strong fusion penalties and compare them to the method that separately estimates each network.

Simulation results are summarized in Figure 3.3. It can be seen that our proposed empirical weights perform very similar to the oracle weights and both versions of informative weights greatly improve the edge selection performance compared with the uniform weights. Moreover, while the benefits are more pronounced with larger fusion penalty, the improvements from informative weights are robust to the choice of tuning parameter for the fusion penalty. Finally, while the advantages of the informative weights are clear, even uniform weights can perform better than method that estimates each network separately; however, with uniform weights, the performance of the method is sensitive to the choice of the tuning



(a) (b)

Figure 3.3: Edge selection performance of the proposed joint estimation method in a simulation study focused on inferring edges in 3 networks of linear Hawkes processes. The plots show average number of true positive and false positive edges, over 100 simulation runs, for the joint estimation method with different choices of weights, compared to separate estimation of each network. Weight strategies include oracle, empirical and uniform weights. Solid squares (■) correspond the choice of tuning parameter using eBIC. (a): Network 2 shares 90% edges with Network 1 and 10% with Network 3 as in Figure 3.2. (b): Network 2 is the same as Network 1.

parameter for the fusion penalty. The benefit of our estimation procedure depends on the similarity between networks. For example, when we alter Network 2 to be the same as Network 1, we observe greater advantages of our method compared to estimating each network separately or using the uniform weights (Figure 3.3(b)). Also, our procedure gets improved performance with increased number of experiments where the additional experiments share

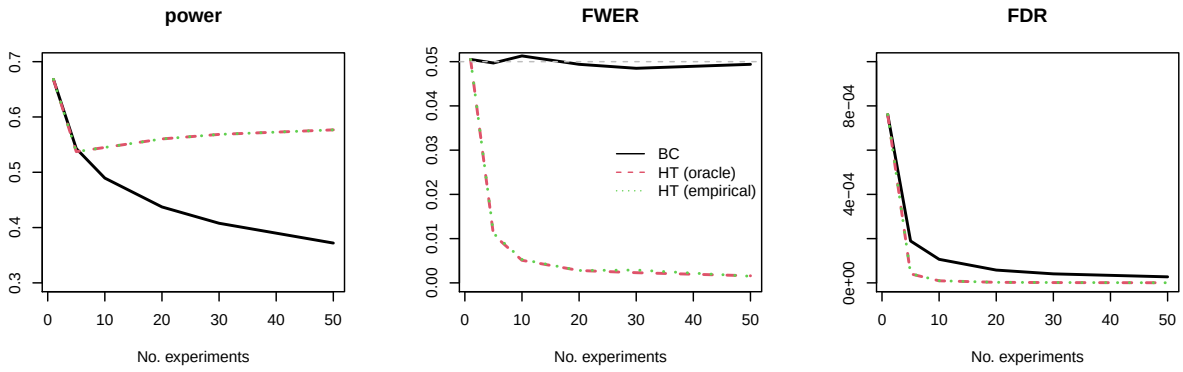


Figure 3.4: Power, FWER and false discovery rate (FDR) for Bonferroni correction (left) and the proposed hierarchical testing procedure using oracle (middle) and the empirical (right) binary trees over 1000 simulation runs. The FWER is controlled at $\alpha = 0.05$ (gray dashline).

similarity to the existed ones (Figure B.5 in Appendix B.5).

Next, to evaluate the performance of the hierarchical testing procedure, we consider $M = 1, 5, 10, 20, 30, 50$ experiments where the network of the first $M - 1$ is Network 1 described above and the Mth network is Network 3. We compare our proposed procedure with Bonferroni correction in terms of power, control of FWER and false discovery rate (FDR). We run our procedure using the oracle and empirical binary trees. As in Figure 3.3, we observe similar performances using both types of weights. While controlling FWER, our proposed procedure greatly improves the power compared with the non-hierarchical method, particularly when the number of experiments–i.e. M–is large (see Figure 3.4). Finally, while our method still controls the FWER with misspecified similarity between networks, the power is lowered when the hierarchy is poorly constructed (see Figure B.2 in Appendix B.5).

3.7 Application

We consider the task of learning the functional connectivity network among population of neurons, using the spike train data from Bolding and Franks [2018]. 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 is applied at increasing intensities at 8 levels 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.

While a total of 80 laser stimuli were applied on neurons of multiple mice, for illustration purposes, we consider the spike train data collected at three stimuli at 0, 10 and 20 mW/mm^2 in a single mouse with the most neurons detected in OB ($p = 25$ neurons). Since one laser pulse spans 10 seconds and the spike train data is recorded at 30 kHz, there are 300,000 time points per stimulus. We apply our joint estimation procedure using data under the three stimuli and evaluate the uncertainty of estimates using the hierarchical testing procedure.

Figure 3.5 illustrates the estimated connectivity coefficients that are specific to each laser condition in a graph representation, where each node represents a neuron and a directed edge indicates a non-zero estimated connectivity coefficient. More edges are observed when laser is applied (32 under 10 mW/mm^2 and 39 under 20 mW/mm^2 v.s. 27 under no laser). Both positive and negative edges are found in all conditions, which empirically echoes the neuroscience hypothesis that both excitatory and inhibitory synapses facilitate maintaining stimulus specificity across odorant concentrations [Bolding and Franks, 2018]. Additionally, we find more common edges in the two laser conditions. Specifically, there are 17 edges (in blue) uniquely shared in the laser conditions compared to 4 edges (in red) shared in all conditions. We find this difference significantly from 0 (p-value $< 1e-5$) based on a permutation test that compares this difference—i.e., 17—to the distribution of the number of edges uniquely shared in the laser condition (by randomly generating networks with the same node-degree distributions at each condition). This finding agrees with the observation by neuroscientists that the OB response is sensitive to the intensity level of the external stimuli [Bolding and Franks, 2018].

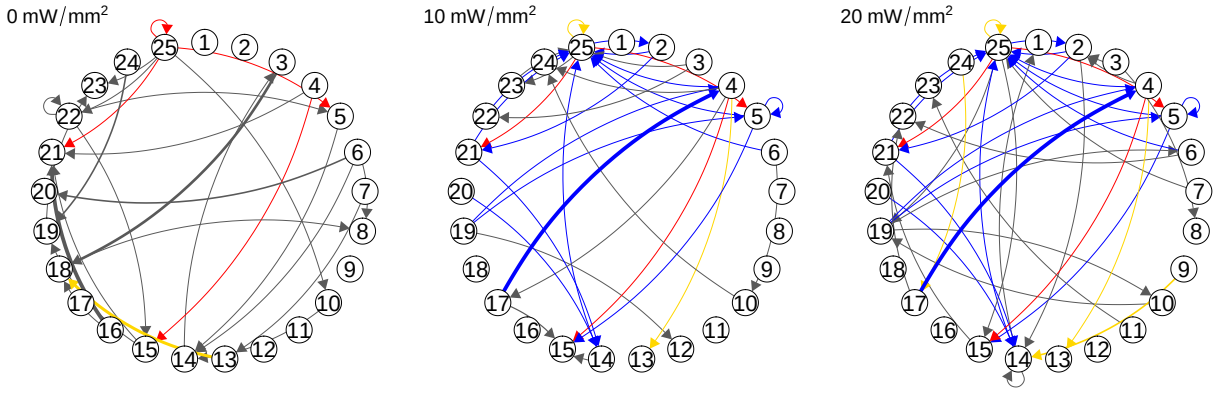


Figure 3.5: Estimated functional connectivities among neuronal populations using the spike train data from Bolding and Franks [2018]. Common edges across all experiments are in red. Edges shared only under laser conditions are in blue. Statistically significant edges controlling FWER are in gold ($\alpha = 0.05$). Edges that are unique to each condition are in gray.

3.8 Discussion

In this paper, we developed a joint estimation procedure for networks of high-dimensional Hawkes processes under multiple experiments. The optimization problem corresponding to our proposed estimation procedure is solved using a smoothing proximal gradient descent algorithm [Chen and Chen, 2008]. Although the algorithm works well for linear models, it empirically shows slow and unstable behavior with non-linear Hawkes models. Since non-linear link functions are often used when analyzing spike train data [Paninski et al., 2007, Pillow et al., 2008], developing more computationally-efficient and stable algorithms for the non-linear models would be a potential direction of future research.

Our proposed hierarchical testing procedure improves the testing power by taking advantage of the multi-experiment structure, while controlling the family-wise error rate. Given large-scale networks, a testing procedure that instead controls the FDR [e.g. Benjamini and

Hochberg, 1995] may offer additional power. Bogomolov et al. [2020] recently proposed a multiple testing adjustment procedure that controls the FDR by using the tree structure of the tests. While improving the power, the method requires a bottom-up p-value calculation, where the upper-level p-values needs to be a specific combination of those from the lower levels. Such procedure becomes compute-intensive, particularly when the number of tests is large, since all the hypotheses on the leaves need to be tested at beginning. It is thus desirable to develop a procedure that allows p-values flexibly calculated on each node of the tree. For example, a procedure that allows a downward p-value calculation avoids intensive computation in calculating p-values over all the leaves, which becomes particularly important in sparse networks, when most leaf-hypotheses are null. Moreover, the existing literature [e.g. Li and Barber, 2019, Bogomolov et al., 2020] that control FDR for structured tests often require the structure to be correctly identified. Given that such structural information is not always available or may not be accurately estimated, developing FDR controlling procedures that are robust to the structure misspecification would be another direction of future research.

Theorem 3.4 shows that, for large and sparse networks, our proposed hierarchical testing procedure is robust to potential misspecification of the hierarchical structure defined based on the proposed similarity in (3.11). Nonetheless, consistent estimation of similarities between networks may still be of interest. In particular, such an estimate could, for instance, facilitate the development of hierarchical FDR controlling procedures discussed above. A key requirement for developing consistent estimates of similarities between connectivity networks based on cross-covariances is to develop a measures of similarity that is *order-preserving*; that is, the order of the similarity based on cross-covariances is the same as that given by the oracle similarity based on the (unknown) connectivity network. One such measure of similarity can be defined based on the connected components of the networks. A *connected component* is a set of nodes that are connected by paths in an undirected graph. In the setting of our problem, the edges of this undirected graph are given by $\mathcal{E}^u = \{(i,j) : |\beta_{ij}| \neq 0, 1 \leq i,j \leq p\}$.

Let $\{\mathcal{C}_l^{(m)}\}_{l=1}^{L^{(m)}}$ and $\{\mathcal{C}_l^{(m')}\}_{l=1}^{L^{(m')}}$ denote the connected components of two p-variate networks in conditions m and m'. The *connected-component similarity* can be defined as

$$d^{cc}\left(\{\mathcal{C}_l^{(m)}\}_{l=1}^{L^{(m)}}, \{\mathcal{C}_l^{(m')}\}_{l=1}^{L^{(m')}}\right) = \sum_{l,l'} r_{l,l'} \left\{\log(r_{l,l'}/p_l) + \log(r_{l,l'}/q_{l'})\right\}, \tag{3.17}$$

where $p_l = \left\|\mathcal{C}_l^{(m)}\right\|/p$, $q_{l'} = \left\|\mathcal{C}_l^{(m')}\right\|/p$ and $r_{l,l'} = \left\|\mathcal{C}_l^{(m)} \cap \mathcal{C}_{l'}^{(m')}\right\|/p$. This measure, which is also known as *variation of information*, is often used to compare the similarity between two clusterings [Meila, 2003].

While the true connected components are unknown in practice, they can be consistently estimated. In particular, we can obtain estimates $\{\widehat{\mathcal{C}}_l^{(m)}\}_{l=1}^{\hat{L}^{(m)}}$ from undirected graphs corresponding to the nonzero values of thresholded empirical cross-covariances (with thresholding at κ) as $\widehat{\mathcal{E}}^u(\kappa) = \left\{(i,j) : \left|\widehat{V}_{ij}\right| > \kappa, 1 \leq i, j \leq p\right\}$. Chen [2016] has shown that the connected components of the true network can be consistently identified using the empirical cross-covariances—i.e., $\mathbb{P}\left(\{\mathcal{C}_l^{(m)}\}_{l=1}^{L} = \{\widehat{\mathcal{C}}_l^{(m)}\}_{l=1}^{\hat{L}^{(m)}}\right) \to 1$ as $T^{(m)} \to \infty$ with $\kappa = o\left(\left(T^{(m)}\right)^{-1/5}\right)$. Thus, as a natural estimator of $d^{cc}\left(\{\mathcal{C}_l^{(m)}\}_{l=1}^{L^{(m)}}, \{\mathcal{C}_l^{(m')}\}_{l=1}^{L^{(m')}}\right)$, $d^{cc}\left(\{\widehat{\mathcal{C}}_l^{(m)}\}_{l=1}^{\hat{L}^{(m)}}, \{\widehat{\mathcal{C}}_l^{(m')}\}_{l=1}^{\hat{L}^{(m')}}\right)$, based on the empirical cross-covariances, consistently represent the similarity in the connected components of the true networks, thus it is order-preserving. However, the effectiveness of a similarity measure based on connected-components depends on the structure of the underlying networks. For instance, while the networks of circles and stars in Figure 3.2 are quite different, their connected-component structures are identical. As a result, similarity weights and dedrograms defined based on connected component woudl not be informative in this case (see Figure B.4 in Appendix B.5). Developing more effective order-preserving network similarity measures is thus an important area of future research.

Chapter 4

CAUSAL DISCOVERY IN HIGH-DIMENSIONAL POINT PROCESS NETWORKS WITH HIDDEN NODES

4.1 Introduction

Learning causality over multivariate time series is generally impossible using observational data for many reasons, including that the data acquisition rate may be much slower than the underlying rate of changes and there may be unmeasured confounding causes [Glymour et al., 2019, Reid et al., 2019]. For example, neuroscience imaging technologies, like functional magnetic resonance imaging, have relatively low temporal resolutions. However, many important neural processes and interactions happen at finer time scales [Zhou et al., 2014]. It is well-established that the most common procedure, Granger causality, based on the slower-rate data may both miss true interactions and add spurious ones [Breitung and Swanson, 2002, Silvestrini and Veredas, 2008, Tank et al., 2019]. Additionally, despite the swift progress in recording activity in massive populations of neurons [Berényi et al., 2014], simultaneously monitoring a complete network of spiking neurons at high temporal resolutions is still beyond the reach of current technology. Nevertheless, there exists many unobserved neurons interacting with the neurons inside the observed population [Trong and Rieke, 2008, Tchumatchenko et al., 2011, Huang, 2015]. Therefore, reliably distinguishing causal connections between pairs of observed neurons from correlations induced by common inputs from unobserved neurons makes the causal learning task challenging. Naive connectivity estimators that neglect these common input effects can produce highly biased results [Soudry et al., 2014].

Learning causal interactions between neurons is critical to understanding the neuronal

connectivity mechanisms in neuroscience. The spike train data is a type of multivariate point process data that contains spiking times of a collection of neurons [Okatan et al., 2005]. Such data have become more abundant at high temporal resolutions thanks to the availability of calcium florescent imaging technology. They are increasingly used to learn the latent brain connectivity networks and glean insight into how neurons respond to external stimuli. For example, Bolding and Franks [2018] collected spike train data at 30kHz at multiple laser intensity levels in studying mouse brain activity.

The Hawkes process [Hawkes, 1971] is a popular choice 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. As such, the Hawkes process offers a flexible and interpretable framework for investigating the latent network of point processes and is widely used in neuroscience applications [Brillinger, 1988, Johnson, 1996, Krumin et al., 2010, Pernice et al., 2011, Reynaud-Bouret et al., 2013, Truccolo, 2016, Lambert et al., 2018]. Bacry and Muzy [2016] has shown that causal interactions can be identified from multivariate point process data, assuming complete data generated from linear Hawkes processes.

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. Hansen et al. [2015] and Chen et al. [2019] proposed ℓ_1-regularized estimation procedures to address this challenge. Wang et al. [2020b] recently developed a high-dimensional inference procedure to characterize the sampling distribution of these estimators and their uncertainty. However, because of the confounding from unobserved neurons in practice, existing estimation and inference procedures, which depend on complete observation from all components, have no guarantee of achieving reliable conclusions.

The problem of causal discovery with hidden variables has been studied using causal

structural learning such as Fast Causal Inference (FCI) and its variants [Spirtes et al., 2000, Glymour et al., 2019]. Unfortunately, not all causal edges can always be identified purely based on such algorithms. Specifically, while FCI identifies causal interactions between variables that are connected by directed edges, it sometimes produces bi-directed edges, leaving the corresponding causal relationship undetermined. Thus, the causality selection performance using such algorithms is not always satisfactory. For example, Malinsky and Spirtes [2018] recently applied FCI to infer causal network of time series and found a low recall in identifying the true casual relationship. Additionally, the causal structure learning often requires intensive computation, especially for large networks, because the space of candidate causal graphs grows super-exponentially with the number of network nodes.

When the variables inhibit temporal ordering as in time series data, the causal structure can be directly learned via the network skeleton using multiple regression models [Shojaie and Michailidis, 2010], which can be solved in a linear time to the number of nodes in the network. While learning causal networks in the presence of unobserved nodes remains challenging using this framework, two recently-developed methods, HIVE and trim regression, shed light on solving this challenge. Specifically, HIVE [Bing et al., 2020] estimates the latent space of the unobserved causal effects and then projects the outcome onto the corresponding orthogonal space. It generates consistent estimates on causality given that the observed and unobserved causal effects are orthogonal. While such condition might hold when the unobserved processes are external such as experimental perturbations in genetic studies [Lee et al., 2017], it might be too stringent in a network setting like neuroscience. Trim regression [Ćevid et al., 2020] applies a spectral transformation to both the outcome and the design columns, which empirically reduces the confounding magnitude. It then eliminates the confounding using penalized regression, assuming the confounding is sufficiently small. It is unclear whether this assumption is satisfied in a network setting where the observed and unobserved processes are complexly connected. Moreover, the long-history temporal dependency of the Hawkes processes makes the learning task more challenging. Unfortunately,

both methods were developed for independent data. Also, these methods were only evaluated on their edge-estimation performance but their abilities in edge selection in a network setting are unknown.

Previous work suggests that the confounding has restricted magnitudes under a stable Hawkes process network, especially when it is dense over the observed nodes [Chen et al., 2019]. Such property motivates us extending the trim regression to learn the causal network of multivariate point processes. Our method shows superior finite-sample performance compared to HIVE and the naive approach that neglects the unobserved processes. Moreover, we establish a non-asymptotic convergence rate in estimating the network edges. Unlike the previous result on independent data [Ćevid et al., 2020], our result considers both the temporal dependency of the Hawkes processes and the network sparsity.

4.2 The Multivariate Hawkes Processes with Unobserved Components

4.2.1 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

$$\lambda_i(t)dt = \mathbb{P}(dN_i(t) = 1 \mid \mathcal{H}_t). \tag{4.1}$$

Hawkes [1971] 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 $i \in \{1, \ldots, p\}$ takes the form

$$\lambda_i(t) = \mu_i + \sum_{j=1}^{p} \left(\omega_{ij} * dN_j\right)(t), \tag{4.2}$$

where

$$\left(\omega_{ij} * dN_j\right)(t) = \int_0^{t-} \omega_{ij}(t-s)dN_j(s) = \sum_{k:t_{jk}<t} \omega_{ij}(t - t_{jk}). \tag{4.3}$$

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 [Linderman and Adams, 2014, de Abril et al., 2018], we consider a parametric transfer function $\omega_{ij}(\cdot)$ of the form

$$\omega_{ij}(t) = \beta_{ij}\kappa_j(t) \tag{4.4}$$

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 $\left(\omega_{ij} * dN_j\right)(t) = \beta_{ij}x_j(t)$, where the *integrated stochastic process*

$$x_j(t) = \int_0^{t-} \kappa_j(t-s)dN_j(s) \tag{4.5}$$

summarizes the entire history of unit j of the multivariate Hawkes processes. A commonly used example is the exponential transition kernel, $\kappa_j(t) = e^{-t}$ [Bacry et al., 2015].

In this formulation, the *connectivity coefficient* of the underlying network, β_{ij}, represents the strength of the *causal* dependence of unit i's intensity on unit j's past events. A positive β_{ij}, which implies that past events of unit j *excite* future events of unit i, is often considered

in the literature [see, e.g., Bacry et al., 2015, Etesami et al., 2016]. However, we might also wish to allow for negative β_{ij} values to represent *inhibitory* effect of one unit's past events on another [Chen et al., 2019, Costa et al., 2020], which is expected in neuroscience applications [Babington, 2001].

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

$$\lambda_i(t) = \mu_i + \boldsymbol{x}^\top(t)\boldsymbol{\beta}_i. \tag{4.6}$$

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

$$Y_i(t) = \mu_i + \boldsymbol{x}^\top(t)\boldsymbol{\beta}_i + \epsilon_i(t). \tag{4.7}$$

4.2.2 The Confounded Hawkes Model

Because of technology constraints, neuroscience experiments usually collect data from only a small portion of neurons while leaving many neurons, interacting with the neurons inside the collected population, 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 (4.7) to including the unobserved components, we obtain the *confounded Hawkes model* as

$$Y_i(t) = \mu_i + \boldsymbol{x}^\top(t)\boldsymbol{\beta}_i + \boldsymbol{z}^\top(t)\boldsymbol{\delta}_i + \epsilon_i(t). \tag{4.8}$$

where $\boldsymbol{z}(t) = (x_{p+1}(t), \ldots, x_{p+q}(t))^\top \in \mathbb{R}^q$ 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, naive estimator that neglects the unobserved components will produce misleading conclusion on the causal relationship among the observed components (e.g., see Figure 4.1).

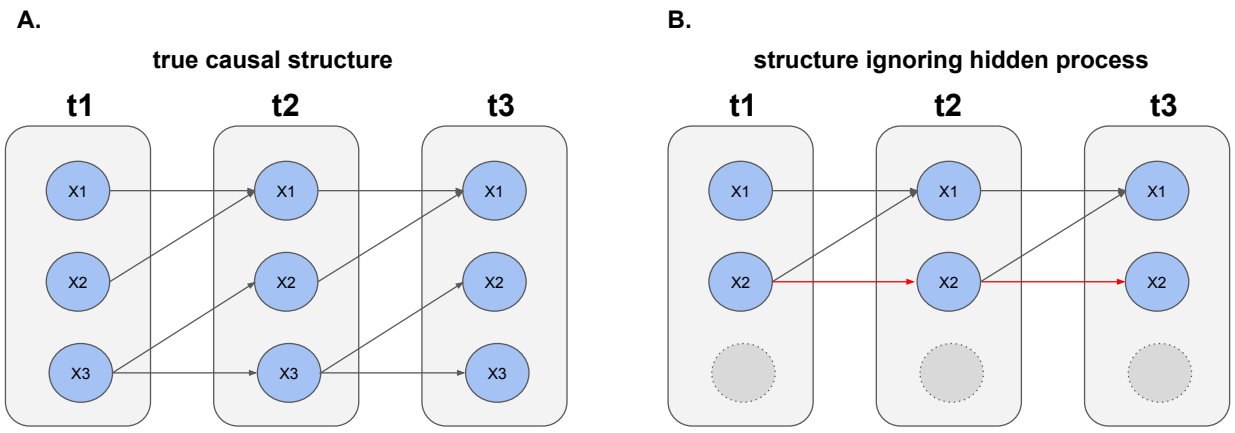


Figure 4.1: Graphical depiction of how hidden process confounds temporal causal interactions. A) The true causal diagram for the complete processes. B) The estimated causal structure of the observed process when the hidden process is ignored.

Throughout this paper, we assume that the confounded linear Hawkes model in (4.8) 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 $[0, T]$ [Brémaud and Massoulié, 1996, Daley and Vere-Jones, 2003].

4.3 Estimation

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

$$\operatorname{Cov}\left(\boldsymbol{x}(t), \boldsymbol{z}^\top(t)\boldsymbol{\delta}_i - \boldsymbol{x}^\top(t)\boldsymbol{b}_i\right) = 0. \tag{4.9}$$

Then, we write the confounded linear Hawkes model in (4.8) in the form of the *perturbed linear model* [Ćevid et al., 2020]:

$$Y_i(t) = \mu_i + \boldsymbol{x}^\top(t)\left(\boldsymbol{\beta}_i + \boldsymbol{b}_i\right) + \nu_i(t), \tag{4.10}$$

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 $\boldsymbol{x}(t)$. $\boldsymbol{b}_i$ 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 purterbed model in (4.10) 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 $\boldsymbol{x}(t)$. Recently, Ćevid et al. [2020] developed a two-step deconfounding procedure to estimate $\boldsymbol{\beta}_i$ for independent and Gaussian-distributed data. First, it applies a simple spectral transformation, called trim transformation (described later), to the observed data. It then estimates $\boldsymbol{\beta}_i$, using penalized regression. We call this procedure *trim regression* for short. Assuming $\boldsymbol{b}_i$ is sufficiently small, the method consistently estimate $\boldsymbol{\beta}_i$. While such assumption is generally not held for Gaussian-distributed data, previous work on Hawkes processes [Chen et al., 2019] implies that the confounding magnitude cannot be large when the underlying network is stable, particularly when the confounding are dense over the observed processes. This allows us to use the trim regression to learn the network of multivariate Hawkes processes.

Consider observations only available from the first p processes at time indexed from 1 to T. Let $X \in \mathbb{R}^{T\times p}$ be the design matrix of the observed integrated process and $Y_i = (Y_i(1), \ldots, Y_i(T))^\top \in \mathbb{R}^T$. 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 $V \in \mathbb{R}^{p\times r}$. $r = \min(T, p)$ is the rank of X. Let non-zero $d_1, \ldots, d_r$ be the diagonal items of D. The *spectral transformation* $F : \mathbb{R}^{T\times p} \to \mathbb{R}^{T\times p}$ is defined as

$$F = U \begin{pmatrix} \tilde{d}_1/d_1 & 0 & \ldots & 0 \\ 0 & \tilde{d}_2/d_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \tilde{d}_r/d_r \end{pmatrix} U^\top. \tag{4.11}$$

Let $\widetilde{D}$ be a diagonal matrix with diagonal elements $\tilde{d}_1, \ldots, \tilde{d}_r$. Applying the spectral trans-

formation to the observed data,

$$\widetilde{X} = FX = U\widetilde{D}V^\top, \tag{4.12}$$

$$\widetilde{Y} = FY. \tag{4.13}$$

Then, we estimate the network connectivities using the transformed data by solving the following optimization problem:

$$\underset{\substack{\mu_i\in\mathbb{R},\boldsymbol{\beta}_i\in\mathbb{R}^p\\ 1\le i\le p}}{\arg\min} \sum_{i=1}^{p} \left\{ \frac{1}{T} \left\| \widetilde{Y}_i - \mu_i - \widetilde{X}\boldsymbol{\beta}_i \right\|_2^2 + \lambda \left\| \boldsymbol{\beta}_i \right\|_1 \right\}. \tag{4.14}$$

The optimization problem above can be separately solved for each $i \in \{1, \ldots, p\}$, using Lasso [Tibshirani, 1996].

Spectral transformation is designed to reduce the magnitude of the confounding. Especially when $\boldsymbol{b}_i$ aligns with the top eigen-vectors of X, choosing appropriate F, e.g., $\tilde{d}_k = \min(\tau, d_k)$, the magnitude of $\widetilde{X}\boldsymbol{b}_i$ is small, compared to $X\boldsymbol{b}_i$. τ is the threshold parameter. Trim transformation is a special case of the spectral transformation when $\tau = median\,(d_1, \ldots, d_r)$.

HIVE [Bing et al., 2020] is another method initially developed to learn linear models with latent variables for independent and Gaussian-distributed data. Adapted to the network of multivariate point processes, it first estimates the latent column space of the unobserved connectivity matrix, $\Delta = \begin{pmatrix} \boldsymbol{\delta}_1^\top \\ \ldots \\ \boldsymbol{\delta}_p^\top \end{pmatrix} \in \mathbf{R}^{p\times q}$, and then projects the outcome vector, $Y(t) = (Y_1(t), \ldots, Y_p(t))^\top$, onto the corresponding orthogonal space. Assuming that the column space of the observed connectivity matrix, $\Theta = \begin{pmatrix} \boldsymbol{\beta}_1^\top \\ \ldots \\ \boldsymbol{\beta}_p^\top \end{pmatrix} \in \mathbf{R}^{p\times p}$, 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 [Lee et al., 2017], it might be too stringent in a network

setting. When the orthogonality assumption fails, it leads poor edge selection performance, sometimes worse than a naive method neglecting the hidden processes. HIVE also requires the number of hidden variables known. Although methods in selecting the number of hidden factors have been proposed, the theoretical guarantee is only asymptotic. An over- or under-estimated number can either miss the true edges or generate false ones. We outline HIVE in Appendix C.1 and refer the interested reader to Bing et al. [2020] for details.

4.4 Theory

In this section we present our main results that guarantee the recovery of the network connectivity when hidden processes present. 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 4.1. *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 4.1 is necessary for stationarity of a Hawkes process [Chen et al., 2019]. The constant γ_Ω does not depend on the dimension $p+q$. For any fixed dimension, Brémaud and Massoulié [1996] show that given this assumption the intensity process of the form (4.6) 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 4.2. *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 4.2 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 [Chen et al., 2019, Wang et al., 2020b].

Assumption 4.3. *The transition kernel $\kappa_j(t)$ is bounded and integrable over $[0, T]$, for $1 \leq j \leq p+q$.*

Assumption 4.3 implies that the integrated process $x_j(t)$ in (4.5) is bounded.

Assumption 4.4. *There exists constants $\rho_r \in (0,1)$ and $0 < \rho_c < \infty$ such that*

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

Assumption 4.4 requires maximum in- and out- intensity flows to be bounded, which provides a sufficient condition in bounding the eigenvalues of the cross-covariance of $\boldsymbol{x}(t)$ [Wang et al., 2020b]. A similar assumption is considered by Basu and Michailidis [2015] in the context of VAR models. Assumptions 4.3 and 4.4 imply that the model parameters are bounded, which is often required in time-series analysis [Safikhani and Shojaie, 2020]. Specifically, the assumptions restrict the influence from the hidden processes from being 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 $Q = \frac{1}{T}\sum_{t=1}^{T} \begin{pmatrix} 1 \\ \boldsymbol{x}(t) \end{pmatrix} \begin{pmatrix} 1 & \boldsymbol{x}^\top(t) \end{pmatrix}$, and $\gamma_{\min} \equiv \Lambda_{\min}(Q)$ and $\gamma_{\max} \equiv \Lambda_{\max}(Q)$.

Theorem 4.1. *Suppose each of the p-variate Hawkes processes has its intensity function defined in (4.8), satisfying Assumptions 4.1– 4.4. Assume $\log p \vee (s^*)^{1/2} = o(T^{1/5})$. Then, taking $\lambda = O(\Lambda^2_{\max}(F)\, T^{-2/5})$,*

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

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 [Ćevid et al., 2020, 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.

Our result is different from that when all processes are observed [Wang et al., 2020b, Lemma 10] by including an extra error term—$\|\widetilde{X}\boldsymbol{b}_i\|_2^2$—from the unobserved process. Next, we show that when $\|\boldsymbol{b}_i\|_2^2$ is sufficiently small, we obtain a similar convergence rate as that given under the case when all processes are observed.

Corollary 4.1. *Under the same assumptions in Theorem 4.1, in addition, suppose* $\|\boldsymbol{b}_i\|_2^2 = O\left(\frac{s^*}{\gamma_{\min}^2\gamma_{\max}}T^{-4/5}\Lambda_{\max}^2(F)\right)$,

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

with probability at least $1 - c_1p^2T\exp(-c_2T^{1/5})$, *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, such result is not guaranteed to hold for arbitrary $\boldsymbol{b}_i$. Corollary 4.1 specifies the condition on $\boldsymbol{b}_i$ that leads to consistent estimates on $\boldsymbol{\beta}_i$ regardless of the empirical performance of the spectral transformation. While the condition does not always hold for arbitrary stochastic process, it is satisfied under a stable network of high-dimensional multivariate Hawkes processes when the confounding is dense. Specifically, by the construction of $\boldsymbol{b}_i$ in (4.9), Assumption 4.4 implies $\|\boldsymbol{b}_i\|_1 = O\left(\|\boldsymbol{\delta}_i\|_1\right) = O(1)$. When the confounding are relatively dense over the observed components—i.e., $\|\boldsymbol{b}_i\|_0 = O(p)$, $\|\boldsymbol{b}_i\|_2^2 = O(1/p)$. Therefore, the magnitude requirement on $\|\boldsymbol{b}_i\|_2^2$ is likely satisfied under a high-dimensional network—i.e., $p \gg T$. The high-dimensional network setting is common in modern neuro-

science experiments where the number of neurons is often large compared to the duration of experiments.

Next we introduce one additional assumption to establish the edge selection consistency.

Assumption 4.5. *There exists $\tau > 0$ such that*

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

Assumption 4.5 is called the β-min condition [Buhlmann, 2013] that requires sufficient signal strength of the true edges in order to distinguish them from 0, given the amount of information observed. We then construct *thresholded connectivity estimator*

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

Such thresholded estimator is often used for variable selections in high dimensions [van de Geer et al., 2011]. Let the estimated edge set $\widehat{S} = \left\{(i,j) : \widetilde{\beta}_{ij} \neq 0, 1 \le i, j \le p\right\}$ and the true edge set $S = \{(i,j) : \beta_{ij} \neq 0, 1 \le i, j \le p\}$. Next, we show that the estimated edge set consistently recovers the true edge set.

Theorem 4.2. *Under the same conditions in Theorem 4.1 and assume Assumption 4.5 is satisfied with $\tau = O\left(\frac{s^*}{\gamma_{\min}^2}\Lambda_{\max}^2(F)T^{-2/5}\right)$. Then,*

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

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

Theorem 4.2 guarantees the recovery of causal interactions among the observed components using our proposed procedure.

We add the proofs of both theorems and the corollary in Appendix C.2.

4.5 Simulation

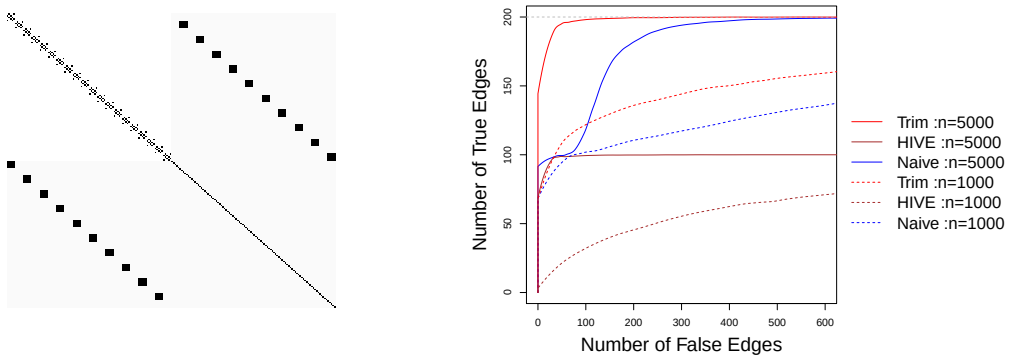
We compare our proposed method–trim regression to two alternatives, HIVE and the (naive) approach that neglects the unobserved nodes, on their abilities in identifying the correct connections in the observed network.

We consider a network of 200 nodes with half of the nodes observed. The observed nodes are connected in blocks of five nodes, and half of the blocks are connected with the unobserved nodes (see Figure 4.2a). This setting violates the orthogonal condition of HIVE. As a sensitivity analysis, we consider another setting similar to the first but removing the connections of the blocks that are not connected with the unobserved nodes. This setting satisfies the orthogonality condition (see Figure 4.3a). β_{ij} and δ_{ij} are 0.12 and 0.10 in the setting of Figure 4.2a, and 0.2 and 0.18 in the setting of Figure 4.3b. The background intensity, μ_i, is 0.05 in both settings. The transfer kernel function is chosen to be $\exp(-t)$. We consider the time period, $T \in \{1000, 5000\}$. These settings satisfy the assumptions of stationary Hawkes processes.

Our proposed method shows superior performance in both small and large sample sizes in the first setting (Figure 4.2b). HIVE performs poorly, worse than the naive approach, because the orthogonal assumption is violated. When the orthogonality condition is satisfied (Figure 4.3b), HIVE shows the best performance. However, such advantage depends on knowing the correct number of latent factors. HIVE generates poor results when the number of latent factors is correctly estimated, especially with an insufficient sample size. In contrast, trim regression gives a satisfying performance close to the oracle have, with both moderate and large sample sizes.

4.6 Application

We consider the task of learning the functional connectivity network among the observed population of neurons, using the spike train data from Bolding and Franks [2018]. In this



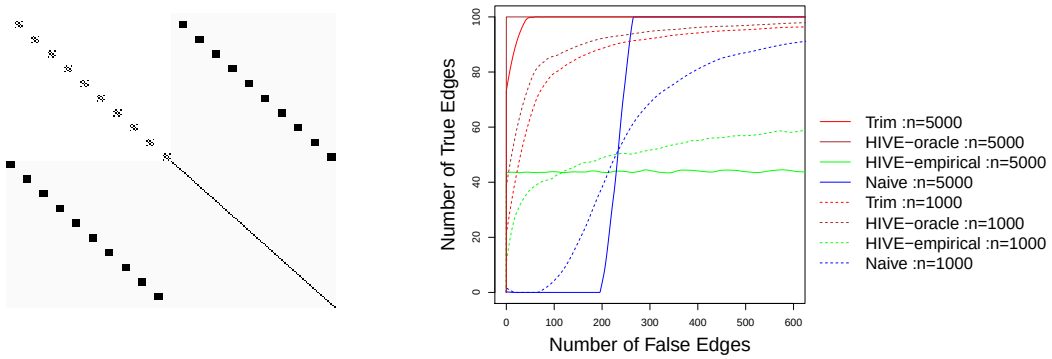
(a) (b)

Figure 4.2: Edge selection using trim regression compared to HIVE and the naive approach for a 200-node network with half of the nodes observed over 100 simulation runs. The connectivity matrix in (a) (unobserved connecitivies colored in gray) violates the orthogonality condition of HIVE. HIVE method is run with the oracle number of latent factors.

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 in total collects spike train data on 23 mice. We consider the spike train data collected in the subject mouse with the most neurons (25 neurons) 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. We apply our estimation procedure separately for each intensity level, and obtain the estimated connectivity coefficients for the observed neurons.

Figure 4.4 illustrates the estimated connectivity coefficients specific to each laser condi-



(a) (b)

Figure 4.3: Edge selection using trim regression compared to HIVE and the naive approach for a 200-node network with half of the nodes observed over 100 simulation runs. The connectivity matrix in (a) (unobserved connecitivies colored in gray) satisfies the orthogonality condition of HIVE. The green line visualizes the performance of HIVE using the estimated number of latent factors with the highest frequency over the 100 simulation runs (estimated N=79 vs. the truth N=50).

tion in a graph representation, where 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 [Bolding and Franks, 2018]. Compared to our proposed method, the naive approach generates a more similar network than HIVE under both laser and no-laser conditions.

As discussed in Section 4.4, our inference procedure is asymptotically valid. That means with large enough samples, if the other assumptions in Section 4.4 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 verify 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 subset of neurons. Such estimation stability is indeed the case in our data set under the laser condition (Figure C.1 in Appendix C.4), suggesting that the conditions are likely satisfied under when laser stimuli are applied.

4.7 Discussion

We proposed a causal-estimation procedure with theoretical guarantees for the high-dimensional network of multivariate Hawkes processes in the presence of hidden nodes. Our estimates are obtained by extending trim regression [Ćevid et al., 2020] to the point process data. Our procedure requires less stringent conditions than existing methods like HIVE, especially for a stable point process network with dense confounding over the observed nodes. Empirically, our procedure shows superior edge-selection performance compared with HIVE and a naive method that neglects the unobserved nodes.

We considered a parametric transition function for the Hawkes process. Given the complex nature of a point process, one may consider modeling the transition function non-

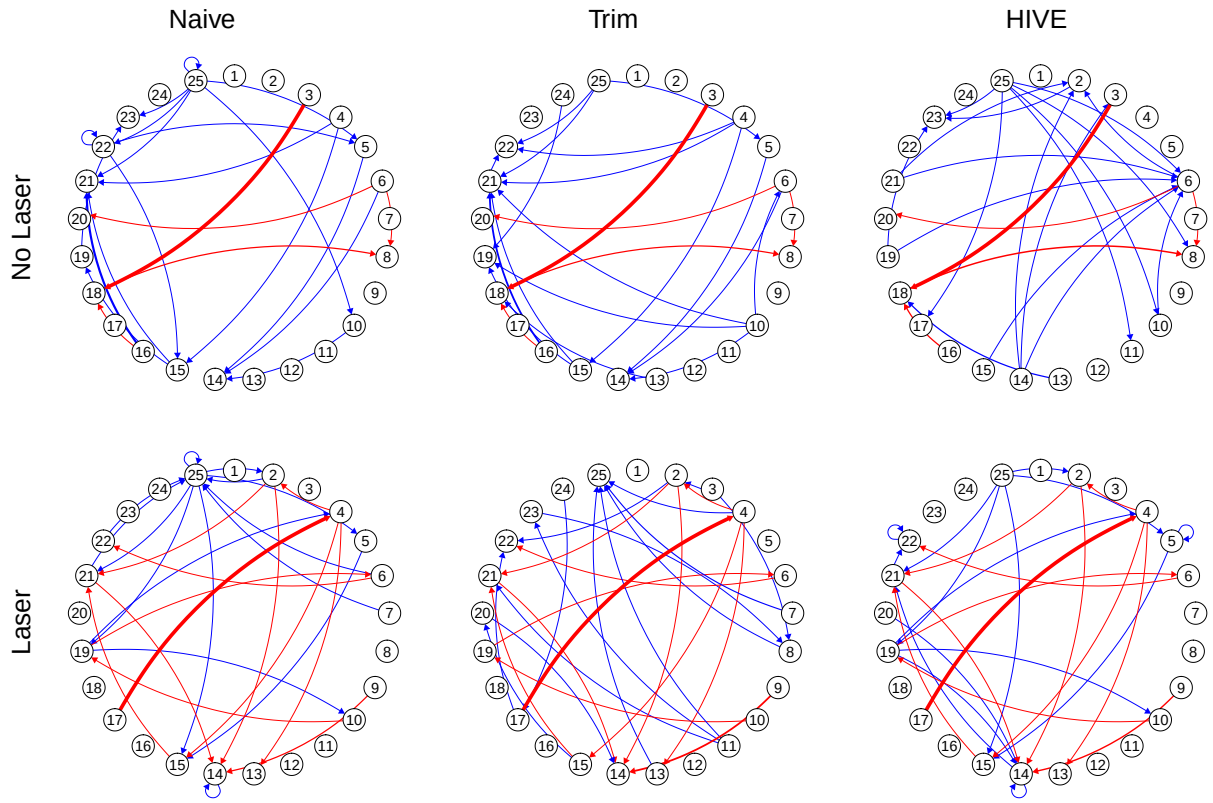


Figure 4.4: Estimated functional connectivities among neuronal populations using the spike train data under laser and no-laser conditions from Bolding and Franks [2018]. Common edges estimated by the three methods are in red and the other edges are in blue. Thicker edges indicate estimated connectivity coefficients of larger magnitudes.

parametrically and learn the form adaptively from data. In addition, since non-linear link functions are often used when analyzing spike train data [Paninski et al., 2007, Pillow et al., 2008], it would also be of interest to develop casual-estimation procedure for non-linear Hawkes processes.

Chapter 5

DISCUSSION

5.1 Summary

In this dissertation, we developed flexible estimation and inference tools for high-dimensional Hawkes processes.

In Chapter 2, we developed a new statistical inference procedure for high-dimensional Hawkes processes. While evaluating the uncertainty of the network estimates is critical in scientific applications, existing methodological and theoretical work has primarily addressed estimation. Our proposed statistical inference procedure thus bridges this gap. The key ingredient for our inference procedure is a new concentration inequality on the first- and second-order statistics for integrated stochastic processes, which summarize the entire history of the process. Combining recent results on martingale central limit theory with the new concentration inequality, we then characterize the convergence rate of the test statistics.

In Chapter 3, we developed a joint estimation procedure to combine data from multiple experiments for more efficient network estimation. Modern high-dimensional point process data often involve observations from multiple conditions and/or experiments. Networks of interactions corresponding to these conditions are expected to share many edges, but also exhibit unique, condition-specific ones. However, the degree of similarity among the networks from different conditions is generally unknown. Unlike existing approaches for multivariate point processes, our proposed joint estimation procedure takes these structures into account by incorporating easy-to-compute weights in order to data-adaptively encourage similarity between the estimated networks. We also propose a powerful hierarchical multiple testing

procedure for edges of all estimated networks, which takes into account the hierarchical similarity structure of the multi-experiment networks guided by the proposed weights. Compared to conventional multiple testing procedures, our proposed procedure greatly reduces the number of tests and results in improved power, while tightly controlling the family-wise error rate. Unlike existing procedures, our method is also free of dependency assumptions on the tests, flexible to calculate p-values along the hierarchy, and robust to misspecification of the hierarchical structure.

In Chapter 4, to cope with the challenges of confounding from unobserved components, we develop a deconfounding procedure to estimate high-dimensional point process networks with only a subset of nodes observed. Unlike existing procedures, our method allows flexible connectivity between the observed and unobserved nodes. It also allows the number of unobserved nodes unknown, which can be much larger than the observed population. With the confounding shrunk by trim transformation and further eliminated using penalized regression, our proposed procedure shows a superior edge-selection performance.

5.2 Future Research

Mapping neuronal connectivity to animal behavior is critical to understanding how the brain controls the body. To achieve this goal, neuroscientists collect data both on neural activity and animal behavior at the same time. For example, Stringer et al. [2019] records the activity of 10000 neurons in the visual cortex of awake mice using two-photon calcium imaging, and simultaneously monitors the facial movements using an infrared camera, where the facial movements were found to be predictable to individual neuron activity. However, how neuronal connectivity structure evolves according to the facial movement remains unknown.

In the previous chapters, we assume the underlying point process is stationary where the neuronal connectivity is stable under a fixed experimental condition, e.g., a specified intensity of laser stimulus. However, in a "free" environment where external factors keep changing, e.g., when a subject mouse watches a video, the fixed connectivity assumption

becomes invalid. Specifically, the connectivity is time-varying and possibly depends on the external (time-varying) stimulus or the mouse's behavior in response to the stimulus. Thus, an interesting question is whether a neuronal connectivity structure can be reconstructed given the external stimulus or animal behavior. While our work in this thesis focuses on learning the neural network itself, our method can be extended to provide insights into this question. We outline our approach as follows.

Consider a p-variate network, we introduce *non-stationary* Hawkes process

$$\lambda_i(t) = \mu_i + \boldsymbol{x}^\top(t)\boldsymbol{\beta}_i(t), \quad i \in \{1, \dots, p\}, \tag{5.1}$$

where μ_i is the spontaneous rate, and $\boldsymbol{x}(t)$ is the integrated process defined in (A.50), and the connectivity coefficients, $\boldsymbol{\beta}_i(t) = (\beta_{i1}(t), \dots, \beta_{ip}(t)) \in \mathbb{R}^p$, vary according to external processes, $\boldsymbol{s}_i(t) = (s_{i1}(t), \dots, s_{iK}(t)) \in \mathbb{R}^K$:

$$\beta_{ij}(t) = \beta_{ij0} + \sum_{k=1}^{K} \beta_{ijk} s_{ik}(t). \tag{5.2}$$

In the context of behavior-dependent brain networks, $\boldsymbol{s}_i(t) = (s_{i1}(t), \dots, s_{iK}(t)) \in \mathbb{R}^K$ represent features of behavior at time t. Specifically, when the behavior of interest can be categorized into total countable types, e.g., running, walking, sniffing and sleeping, $\boldsymbol{s}_i(t) \in \mathbb{R} \in \{1, \dots, C\}$. When the behavior of interest is complicated like facial expression recorded in a movie, $\boldsymbol{s}_i(t)$ can be chosen as the projection scores of the movie frame matrix on the "eigenFace" vectors at time t [Stringer et al., 2019].

Similar as before, $\boldsymbol{\beta}_{ij}(t) \neq 0$ for $t \in [0, T]$ indicates the strength of the dependence/connectivity of unit i's intensity on unit j's past events. Additionally, when there exists $\beta_{ijk} \neq 0$ for some $k \in \{1, \dots, K\}$, then such connectivity between neurons is *behavior-dependent* (when $\boldsymbol{s}_i(t)$ represents behavior features at time t).

Denote the unknown parameters in (5.1) as $\boldsymbol{\theta}_i = (\mu_i, \boldsymbol{\beta}_{i1}, \dots, \boldsymbol{\beta}_{ip}) \in \mathbb{R}^{p\times(K+1)+1}$ for $1 \leq i \leq p$, where $\boldsymbol{\beta}_{ij} = (\beta_{ij0}, \beta_{ij1}, \dots, \beta_{ijK})$. Let $\ell(\cdot, \cdot) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ denote twice continuously differentiable loss function. To rigorously define our penalized estimator, denote

the expectation of the observed outcome at $[t, t+dt)$ for unit i under experiment m as $f_{\boldsymbol{\theta}_i}(\boldsymbol{x}(t)) \equiv \lambda_i(t)dt$. We then estimate these unknown parameters by solving the following optimization problem:

$$\underset{\substack{\boldsymbol{\theta}_i \in \mathbb{R}^{p\times(K+1)+1} \\ 1\le i\le p}}{\arg\min} \sum_{i=1}^{p} \left\{ \frac{1}{T} \int_0^T \ell\left(dN_i(t), f_{\boldsymbol{\theta}_i}(\boldsymbol{x}(t))\right) + \mathcal{P}\left(\{\boldsymbol{\beta}_{ij}\}_{j=1}^{p}\right) \right\}, \tag{5.3}$$

where to achieve sparsity, we use *sparse group-lasso* penalty [Simon et al., 2013]

$$\mathcal{P}\left(\{\boldsymbol{\beta}_{ij}\}_{j=1}^{p}\right) = \underbrace{\lambda_1 \sum_{j=1}^{p} \|\boldsymbol{\beta}_{ij}\|_2}_{\text{network sparsity}} + \underbrace{\lambda_2 \sum_{j=1}^{p} \|\boldsymbol{\beta}_{ij}\|_1}_{\text{behavior-dependency sparsity}} . \tag{5.4}$$

The sparse group-lasso penality encourages a (groupwise) sparse structure of the estimated networks (i.e. few non-constant-zero $\beta_{ij}(t)$, for $1 \le t \le T$) via the tuning parameter λ_1, and a (within-group) sparse structure of the behavior dependency (i.e. few non-zero β_{ijk}, for $1 \le k \le K$) via the tuning parameter λ_2. The optimization problem can be solved using blockwise descent algorithm [Simon et al., 2013].

To derive the theoretical guarantee on both edge- and network-behavior-dependency selection using the penalized estimator above, we need to verify the restrict eigen-value (RE) condition [Negahban and Wainwright, 2010] for the new design matrix,

$$\frac{1}{T} \int_0^T \begin{pmatrix} 1 \\ \widetilde{\boldsymbol{x}}(t) \end{pmatrix} \begin{pmatrix} 1 & \widetilde{\boldsymbol{x}}(t) \end{pmatrix} dt,$$

where $\widetilde{\boldsymbol{x}}(t) = (\widetilde{\boldsymbol{x}}_1(t), \ldots, \widetilde{\boldsymbol{x}}_p(t)) \in \mathbb{R}^{p\times(K+1)}$ and $\widetilde{\boldsymbol{x}}_i(t) = (x_i(t), s_{i1}(t)x_i(t), \ldots, s_{iK}(t)x_i(t))$. When such condition is met, the behavior-dependency indicators, $\{(i,j,k) : \beta_{ijk} \ne 0, 1 \le i, j \le p, 0 \le k \le K\}$, can be consistently selected following a similar argument as we show in Theorem 3.1 and Theorem 3.2 in Chapter 3.

Evaluating the uncertainty of the connectivity coefficients is also critical. To be specific, we are interested in

- confidence band for the time-varying connectivity coefficient $\beta_{ij}(t)$; i.e.,

$$H_0 : \beta_{ij}(t) = 0, \quad t \in [0, T]; \tag{5.5}$$

- test whether any of the behavior-dependency coefficients is non-zero; i.e.,

$$H_0 : \beta_{ijk} = 0, \quad 1 \leq k \leq K. \tag{5.6}$$

Testing the hypotheses in (5.5) and (5.6) requires the knowledge of the distribution of $(\beta_{ij0}, \ldots, \beta_{ijK})$ under the null. While the inference framework in Chapter 2 can be applied here, the main challenge is to valid the bounded eigen-value condition of the new design matrix —i.e. $\mathrm{Var}(\widetilde{\boldsymbol{x}}(t))$, which may require additional regularity conditions on both the point process and the behavior feature process.

BIBLIOGRAPHY

Kevin A. Bolding and Kevin M. Franks. Recurrent cortical circuits implement concentration-invariant odor coding. Science, 361(6407), 2018.

Niels Richard Hansen, Patricia Reynaud-Bouret, and Vincent Rivoirard. Lasso and probabilistic inequalities for multivariate point processes. Bernoulli, 21(1):83–143, 2015.

Shizhe Chen, Ali Shojaie, Eric Shea-Brown, and Daniela Witten. The multivariate hawkes process in high dimensions: Beyond mutual excitation. arXiv 1707.04928, 2019.

Cun-Hui Zhang and Stephanie S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76(1):217–242, 2014.

Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. Ann. Statist., 42(3): 1166–1202, 2014.

Alexandre Belloni, Victor Chernozhukov, and Christian B. Hansen. Inference on treatment effects after selection amongst high-dimensional controls. Rev. Econ. Stud., 81(2):608–650, Nov 2013.

Rina Foygel Barber and Mladen Kolar. Rocket: Robust confidence intervals via kendall's tau for transelliptical graphical models. Ann. Statist., 46(6B):3422–3450, 2018.

Jana Janková and Sara van de Geer. Inference in high-dimensional graphical models. In Handbook of graphical models, Chapman & Hall/CRC Handb. Mod. Stat. Methods, pages 325–349. CRC Press, Boca Raton, FL, 2019.

Junwei Lu, Mladen Kolar, and Han Liu. Post-regularization inference for time-varying nonparanormal graphical models. J. Mach. Learn. Res., 18(203):1–78, 2018.

Ming Yu, Varun Gupta, and Mladen Kolar. Simultaneous inference for pairwise graphical models with generalized score matching. arXiv 1905.06261, 2019.

Yang Ning and Han Liu. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. Ann. Statist., 45(1):158–195, 2017.

Matey Neykov, Yang Ning, Jun S. Liu, and Han Liu. A unified theory of confidence regions and testing for high-dimensional estimating equations. Statist. Sci., 33(3):427–443, 2018.

Lili Zheng and Garvesh Raskutti. Testing for high-dimensional network parameters in auto-regressive models. Electronic Journal of Statistics, 13(2):4977–5043, 2019.

Peter J Bickel, Ya'acov Ritov, Alexandre B Tsybakov, et al. Simultaneous analysis of lasso and dantzig selector. Ann. Statist., 37(4):1705–1732, 2009.

Manon Costa, Carl Graham, Laurence Marsalle, and Viet Chi Tran. Renewal in hawkes processes with self-excitation and inhibition. Advances in Applied Probability, 52(3): 879–915, 2020.

Julien Chiquet, Yves Grandvalet, and Christophe Ambroise. Inferring multiple graphical structures. Statistics and Computing, 21(4):537–553, Oct 2011.

Jian Guo, Elizaveta Levina, George Michailidis, and Ji Zhu. Joint estimation of multiple graphical models. Biometrika, 98(1):1–15, 02 2011.

Patrick Danaher, Pei Wang, and Daniela M. Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. Journal of the Royal Statistical Society. Series B, Statistical methodology, 76(2):373–397, Mar 2014.

Masanao Yajima, Donatello Telesca, Yuan Ji, and Peter Müller. Detecting differential patterns of interaction in molecular pathways. Biostatistics, 16(2):240–251, 12 2014.

Yunzhang Zhu, Xiaotong Shen, and Wei Pan. Structural pursuit over multiple undirected graphs. Journal of the American Statistical Association, 109(508):1683–1696, 2014.

Jing Ma and George Michailidis. Joint structural estimation of multiple graphical models. Journal of Machine Learning Research, 17(166):1–48, 2016.

T. Tony Cai, Hongzhe Li, Weidong Liu, and Jichun Xie. Joint estimation of multiple high-dimensional precision matrices. Statistica Sinica, 26(2):445–464, 2016.

F. Huang, S. Chen, and S. Huang. Joint estimation of multiple conditional gaussian graphical models. IEEE Transactions on Neural Networks and Learning Systems, 29(7):3034–3046, 2018.

Yuhao Wang, Santiago Segarra, and Caroline Uhler. High-dimensional joint estimation of multiple directed gaussian graphical models. arXiv 1804.00778, 2020a.

Mladen Kolar, Le Song, Amr Ahmed, and Eric P. Xing. Estimating time-varying networks. Ann. Appl. Stat., 4(1):94–123, 03 2010.

Huitong Qiu, Fang Han, Han Liu, and Brian Caffo. Joint estimation of multiple graphical models from high dimensional time series. Journal of the Royal Statistical Society. Series B (Statistical Methodology), 78(2):487–504, 2016.

Zhixiang Lin, Tao Wang, Can Yang, and Hongyu Zhao. On joint estimation of gaussian graphical models for spatial and temporal data. Biometrics, 73(3):769–779, 2017.

David Hallac, Youngsuk Park, Stephen Boyd, and Jure Leskovec. Network inference via the time-varying graphical lasso. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '17, page 205–213, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450348874.

Jilei Yang and Jie Peng. Estimating time-varying graphical models. Journal of Computational and Graphical Statistics, 29(1):191–202, 2020.

Christine Peterson, Francesco C Stingo, and Marina Vannucci. Bayesian inference of multiple gaussian graphical models. Journal of the American Statistical Association, 110(509):159–174, 2015.

Takumi Saegusa and Ali Shojaie. Joint estimation of precision matrices in heterogeneous populations. Electron. J. Statist., 10(1):1341–1392, 2016.

Ildefons Magrans de Abril, Junichiro Yoshimoto, and Kenji Doya. Connectivity inference from neural recording data: Challenges, mathematical bases and research directions. Neural Networks, 102:120–137, 2018.

Tatjana Tchumatchenko, Theo Geisel, Maxim Volgushev, and Fred Wolf. Spike correlations – what can they tell about synchrony? Frontiers in Neuroscience, 5:68, 2011. ISSN 1662-453X.

Andrew T. Reid, Drew B. Headley, Ravi D. Mill, Ruben Sanchez-Romero, Lucina Q. Uddin, Daniele Marinazzo, Daniel J. Lurie, Pedro A. Valdés-Sosa, Stephen José Hanson, Bharat B. Biswal, Vince Calhoun, Russell A. Poldrack, and Michael W. Cole. Advancing functional connectivity research from association to causation. Nature Neuroscience, 22(11):1751–1760, Nov 2019. ISSN 1546-1726.

Xu Wang, Mladen Kolar, and Ali Shojaie. Statistical inference for networks of high-dimensional point processes, 2020b.

Daniel Yekutieli. Hierarchical false discovery rate–controlling methodology. Journal of the American Statistical Association, 103(481):309–316, 2008.

Gavin Lynch and Wenge Guo. On procedures controlling the fdr for testing hierarchically ordered hypotheses. arXiv:1612.04467, 2016.

W Bogomolov, Christine B Peterson, Yoav Benjamini, and Chiara Sabatti. Hypotheses on a tree: new error rates and testing strategies. Biometrika, Oct. 2020.

P. Spirtes, C. Glymour, and R. Scheines. Causation, Prediction, and Search. MIT press, 2nd edition, 2000.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. Frontiers in Genetics, 10:524, 2019. ISSN 1664-8021.

Daniel Malinsky and Peter Spirtes. Causal structure learning from multivariate time series in settings with unmeasured confounding. In Proceedings of 2018 ACM SIGKDD Workshop on Causal Disocvery, volume 92 of Proceedings of Machine Learning Research, pages 23–47, London, UK, 20 Aug 2018. PMLR.

Ali Shojaie and George Michailidis. Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. Biometrika, 97(3):519–538, 07 2010. ISSN 0006-3444.

Xin Bing, Yang Ning, and Yaosheng Xu. Adaptive estimation of multivariate regression with hidden variables. arXiv: 2003.13844, 2020.

Domagoj Ćevid, Peter Bühlmann, and Nicolai Meinshausen. Spectral deconfounding via perturbed sparse linear models. Journal of Machine Learning Research, 21(232):1–41, 2020.

Seunggeun Lee, Wei Sun, Fred A. Wright, and Fei Zou. An improved and explicit surrogate variable analysis procedure by coefficient adjustment. Biometrika, 104(2):303–316, 04 2017.

Murat Okatan, Matthew A. Wilson, and Emery N. Brown. Analyzing functional connectivity using a network likelihood model of ensemble neural spiking activity. Neural Computation, 17(9):1927–1961, 2005.

Ke Zhou, Hongyuan Zha, and Le Song. Learning social infectivity in sparse low-rank networks using multi-dimensional Hawkes processes. In AISTATS, 2013.

V. Chavez-Demoulin and J.A. McGill. High-frequency financial data modeling using Hawkes processes. Journal of Banking and Finance, 36(12):3415 – 3426, 2012.

Steffen L Lauritzen. Graphical models, volume 17. Clarendon Press, 1996.

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

Yosihiko Ogata. Statistical models for earthquake occurrences and residual analysis for point processes. J. Am. Statist. Ass., 83(401):9–27, 1988.

Emmanuel Bacry, Khalil Dayri, and J.F. Muzy. Non-parametric kernel estimation for symmetric Hawkes processes. application to high frequency financial data. The European Physical Journal B, 85, 2011.

Scott Linderman and Ryan Adams. Discovering latent network structure in point process data. 31st International Conference on Machine Learning, ICML 2014, 4, 02 2014.

Alan G Hawkes and David Oakes. A cluster process representation of a self-exciting process. J Appl. Probab., 11(3):493–503, 1974.

Patricia Reynaud-Bouret and Emmanuel Roy. Some non asymptotic tail estimates for Hawkes processes. Bull. Belg. Math. Soc. Simon Stevin, 13(5):883–896, 2007.

Patricia Reynaud-Bouret and Sophie Schbath. Adaptive estimation for Hawkes processes; application to genome analysis. Ann. Statist., 38(5):2781–2822, 2010.

Emmanuel Bacry, Iacopo Mastromatteo, and J.F. Muzy. Hawkes processes in finance. Market Microstructure and Liquidity, 01, 02 2015.

Jalal Etesami, Negar Kiyavash, Kun Zhang, and Kushagra Singhal. Learning network of multivariate hawkes processes: A time series approach. arXiv, abs/1603.04319, 2016.

Peter Babington. Neuroscience (Second ed.). Sunderland, MA: Sinauer Associates, 2 edition, 2001.

Sumanta Basu and George Michailidis. Regularized estimation in sparse high-dimensional time series models. Ann. Statist., 43(4):1535–1567, 2015.

Pierre Brémaud and Laurent Massoulié. Stability of nonlinear Hawkes processes. Ann. Probab., 24(3):1563–1588, 1996.

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

Abolfazl Safikhani and Ali Shojaie. Joint structural break detection and parameter estimation in high-dimensional nonstationary var models. Journal of the American Statistical Association, 0(0):1–14, 2020.

Florence Merlevède, Magda Peligrad, and Emmanuel Rio. A bernstein type inequality and moderate deviations for weakly dependent sequences. Probability Theory and Related Fields, 151, 03 2009.

Biao Cai, Jingfei Zhang, and Yongtao Guan. Latent network structure learning from high dimensional multivariate point processes. arXiv: 2004.03569, 2020.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68(1):49–67, 2006.

D. R. Brillinger. Maximum likelihood analysis of spike trains of interacting nerve cells. Biological Cybernetics, 59(3):189–200, Aug 1988.

Don H. Johnson. Point process models of single-neuron discharges. Journal of Computational Neuroscience, 3(4):275–299, Dec 1996.

Michael Krumin, Inna Reutsky, and Shy Shoham. Correlation-based analysis and generation of multiple spike trains using hawkes models with an exogenous input. Frontiers in computational neuroscience, 4:147–147, Nov 2010.

Volker Pernice, Benjamin Staude, Stefano Cardanobile, and Stefan Rotter. How structure determines correlations in neuronal networks. PLoS computational biology, 7(5):e1002059–e1002059, May 2011. ISSN 1553-7358.

P. Reynaud-Bouret, V. Rivoirard, and C. Tuleau-Malot. Inference of functional connectivity in neurosciences via hawkes processes. In 2013 IEEE Global Conference on Signal and Information Processing, pages 317–320, 2013.

Wilson Truccolo. From point process observations to collective neural dynamics: Nonlinear hawkes process glms, low-dimensional dynamics and coarse graining. Journal of Physiology-Paris, 110(4, Part A):336 – 347, 2016.

Régis C. Lambert, Christine Tuleau-Malot, Thomas Bessaih, Vincent Rivoirard, Yann Bouret, Nathalie Leresche, and Patricia Reynaud-Bouret. Reconstructing the functional connectivity of multiple spike trains using hawkes models. Journal of Neuroscience Methods, 297:9 – 21, 2018.

Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(1):91–108, 2005.

E. Bacry and J. Muzy. First- and second-order statistics characterization of hawkes processes and non-parametric estimation. IEEE Transactions on Information Theory, 62(4):2184–2202, 2016.

Shizhe Chen, Daniela Witten, and Ali Shojaie. Nearly assumptionless screening for the mutually-exciting multivariate Hawkes process. Electronic Journal of Statistics, 11(1): 1207 – 1234, 2017.

Lu Tang and Peter X.K. Song. Fused lasso approach in regression coefficients clustering – learning parameter heterogeneity in data integration. Journal of Machine Learning Research, 17(113):1–23, 2016.

Fei Wang, Lu Wang, and Peter Song. Fused lasso with the adaptation of parameter ordering in merging multiple studies with repeated measurements. Biometrics, 72, 02 2016.

Xi Chen, Qihang Lin, Seyoung Kim, Jaime G. Carbonell, and Eric P. Xing. Smoothing proximal gradient method for general structured sparse regression. The Annals of Applied Statistics, 6(2):719–752, 2012.

Jiahua Chen and Zehua Chen. Extended Bayesian information criteria for model selection with large model spaces. Biometrika, 95(3):759–771, 09 2008.

Sahand Negahban and Martin Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. Computing Research Repository - CORR, 13, 09 2010.

Jason D. Lee, Yuekai Sun, and Jonathan E. Taylor. On model selection consistency of regularized m-estimators. Electron. J. Statist., 9(1):608–642, 2015.

Peter Bühlmann and Sara van de Geer. Statistics for high-dimensional data: methods, theory and applications. Springer Science & Business Media, 2011.

Sara van de Geer, Peter Bühlmann, and Shuheng Zhou. The adaptive and the thresholded Lasso for potentially misspecified models (and a lower bound for the Lasso). Electronic Journal of Statistics, 5:688 – 749, 2011.

Paul Embrechts, Claudia Klüppelberg, and Thomas Mikosch. Modelling Extremal Events for Insurance and Finance. Springer-Verlag Berlin Heidelberg, 1 edition, 1997.

Liam Paninski, Jonathan Pillow, and Jeremy Lewi. Statistical models for neural encoding, decoding, and optimal stimulus design. In Computational Neuroscience: Theoretical

Insights into Brain Function, volume 165 of Progress in Brain Research, pages 493 – 507. Elsevier, 2007.

Jonathan Pillow, Jonathon Shlens, Liam Paninski, Alexander Sher, Alan Litke, E.J. Chichilnisky, and Eero Simoncelli. Spatio-temporal correlations and visual signaling in a complete neuronal population. Nature, 454:995–9, 2008.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological), 57(1):289–300, 1995.

Ang Li and Rina Foygel Barber. Multiple testing with the structure-adaptive benjamini–hochberg algorithm. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 81(1):45–74, 2019.

Marina Meila. Comparing Clusterings by the Variation of Information, volume 2777. Lecture Notes in Computer Science, 2003.

Shizhe Chen. Flexible modeling and estimation for high-dimensional graphs. PhD thesis, University of Washington, 2016.

D. Zhou, Y. Zhang, Yanyang Xiao, and D. Cai. Analysis of sampling artifacts on the granger causality analysis for topology extraction of neuronal dynamics. Frontiers in Computational Neuroscience, 8, 2014.

Jörg Breitung and Norman R. Swanson. Temporal aggregation and spurious instantaneous causality in multiple time series models. Journal of Time Series Analysis, 23(6):651–665, 2002.

Andrea Silvestrini and David Veredas. Temporal aggregation of univariate and multivariate time series models: A survey. Journal of Economic Surveys, 22(3):458–497, 2008.

A Tank, E B Fox, and A Shojaie. Identifiability and estimation of structural vector autoregressive models for subsampled and mixed-frequency time series. Biometrika, 106(2): 433–452, 04 2019. ISSN 0006-3444.

Antal Berényi, Zoltán Somogyvári, Anett J. Nagy, Lisa Roux, John D. Long, Shigeyoshi Fujisawa, Eran Stark, Anthony Leonardo, Timothy D. Harris, and György Buzsáki. Large-scale, high-density (up to 512 channels) recording of local circuits in behaving animals. Journal of Neurophysiology, 111(5):1132–1149, 2014.

Philipp Trong and Fred Rieke. Origin of correlated activity between parasol retinal ganglion cells. Nature neuroscience, 11:1343–51, 10 2008.

Haiping Huang. Effects of hidden nodes on network structure inference. Journal of Physics A: Mathematical and Theoretical, 48(35):355002, Aug. 2015.

Daniel Soudry, Suraj Keshri, Patrick Stinson, Min hwan Oh, Garud Iyengar, and Liam Paninski. A shotgun sampling solution for the common input problem in neural connectivity inference. arXiv: 1309.3724, 2014.

Robert Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 58(1):267–288, 1996.

Peter Buhlmann. Statistical significance in high-dimensional linear models. Bernoulli, 19(4): 1212–1242, 09 2013.

Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Charu Bai Reddy, Matteo Carandini, and Kenneth D. Harris. Spontaneous behaviors drive multidimensional, brainwide activity. Science, 364(6437), 2019.

Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. Journal of Computational and Graphical Statistics, 22(2):231–245, 2013.

B Laurent and P Massart. Adaptive estimation of a quadratic functional by model selection. Ann. Statist., 28(5):1302–1338, 2000.

Ion Grama and E. Haeusler. An asymptotic expansion for probabilities of moderate deviations for multivariate martingales. J Theor. Probab., 19:1–44, 2006.

L Paninski and J Cunningham. Neural data science: accelerating the experiment-analysis-theory cycle in large-scale neuroscience. Current opinion in neurobiology, 50:232–241, 2018.

Sara van de Geer. Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes. Ann. Statist., 23(5):1779–1801, 1995.

Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. J. Mach. Learn. Res.,, 15, 2013.

P. Reynaud-Bouret, V. Rivoirard, and Christine Tuleau-Malot. Inference of functional connectivity in neurosciences via hawkes processes. 2013 IEEE Global Conference on Signal and Information Processing, pages 317–320, 2013.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. arXiv:1011.3027, 2010.

A. Skripnikov and G. Michailidis. Joint estimation of multiple network granger causal models. Econometrics and Statistics, 10, 08 2018.

Xiaotong Shen and Hsin-Cheng Huang. Grouping pursuit through a regularization solution surface. Journal of the American Statistical Association, 105:727–739, 06 2010.

Jerome Friedman, Trevor Hastie, Holger Höfling, and Robert Tibshirani. Pathwise coordinate optimization. Ann. Appl. Stat., 1(2):302–332, 12 2007.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sciences, 2:183–202, 2009.

Amir Beck. First-Order Methods in Optimization. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2017.

Michael Chichignoud, Johannes Lederer, and Martin J. Wainwright. A practical scheme and fast algorithm to tune the lasso with optimality guarantees. Journal of Machine Learning Research, 17(229):1–20, 2016.

Marlene Cohen and Adam Kohn. Measuring and interpreting neuronal correlations. Nat Neurosci, 14:811–819, 2011.

Ali Shojaie. Differential network analysis: A statistical perspective. arXiv:2003.04235, 2020.

Anqi Fu, Balasubramanian Narasimhan, and Stephen Boyd. CVXR: An R package for disciplined convex optimization. Journal of Statistical Software, 94(14):1–34, 2020.

Yinchu Zhu and Jelena Bradic. Linear hypothesis testing in dense high-dimensional linear models. Journal of the American Statistical Association, 113(524):1583–1600, 2018.

Hongteng Xu, Mehrdad Farajtabar, and Hongyuan Zha. Learning granger causality for hawkes processes. arXiv: 1602.04511, 2016.

Michael Eichler, Rainer Dahlhaus, and Johannes Dueck. Graphical modeling for multivariate hawkes processes with nonparametric link functions. Journal of Time Series Analysis, 38 (2):225–242, 2017.

Michael Eichler. Causal inference with multiple time series: Principles and problems. Philosophical transactions. Series A, Mathematical, physical, and engineering sciences, 371:20110613, 07 2013.

Clive Granger. Investigating causal relations by econometric models and cross-spectral methods. Econometrica, 37(3):424–38, 1969.

Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvrinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. Journal of Machine Learning Research, 7(72): 2003–2030, 2006.

Filippo Moauro and Giovanni Savio. Temporal disaggregation using multivariate structural time series models. The Econometrics Journal, 8(2):214–234, 07 2005.

Daniel Stram and William Wei. A methodological note on the disaggregation of time series totals. Journal of Time Series Analysis, 7:293 – 302, 05 2008.

J. C. G. Boot, W. Feibes, and J. H. C. Lisman. Further methods of derivation of quarterly figures from annual data. Journal of the Royal Statistical Society: Series C (Applied Statistics), 16(1):65–75, 1967.

Beatrice Miccoli, Carolina Mora Lopez, Erkuden Goikoetxea, Jan Putzeys, Makrina Sekeri, Olga Krylychkina, Shuo-Wen Chang, Andrea Firrincieli, Alexandru Andrei, Veerle Reumers, and Dries Braeken. High-density electrical recording and impedance imaging with a multi-modal cmos multi-electrode array chip. Frontiers in Neuroscience, 13:641, 2019.

Anru Zhang, T. Tony Cai, and Yihong Wu. Heteroskedastic pca: Algorithm, optimality, and applications. arXiv: 1810.08316, 2019.

G. W. Stewart and Jiguang Sun. Matrix Perturbation Theory. ACADEMIC PressINC, 1990.

Peter Bühlmann and Sara van de Geer. Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer Publishing Company, Incorporated, 1st edition, 2011. ISBN 3642201911.

Massimiliano Zanin and David Papo. Efficient neural codes can lead to spurious synchronization. Frontiers in Computational Neuroscience, 7:125, 2013.

David Marc Anton Mehler and Konrad Paul Kording. The lure of misleading causal statements in functional connectivity research. arXiv: 1812.03363, 2020.

Aaron Alexander-Bloch, Nitin Gogtay, David Meunier, Rasmus Birn, Liv Clasen, Francois Lalonde, Rhoshel Lenroot, Jay Giedd, and Edward Bullmore. Disrupted modularity and local connectivity of brain functional networks in childhood-onset schizophrenia. Frontiers in systems neuroscience, 4:147, 10 2010.

Shinji Nishimoto, An T. Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L. Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. Current biology : CB, 21(19):1641–1646, Oct 2011.

Kean Ming Tan, Daniela Witten, and Ali Shojaie. The cluster graphical lasso for improved estimation of gaussian graphical models. Computational Statistics & Data Analysis, 85: 23–36, 2015.

Appendix A

APPENDIX FOR CHAPTER 2

A.1 Proof Outline

Before presenting the formal proof of Theorems 2.1– 2.3, we outline the key technical steps in this section. We also briefly discuss key steps in the proof of lemmas used to prove the theorems, as well as auxiliary lemmas, which are presented in Appendix A.3. Recall that S_{ij} is the de-correlated score defined in (2.12), Υ_j, defined in (2.13), is the covariance of $z_j^*(t)$, where $z_j^*(t)$ is the scaled version of the design column $x_j(t)$ after removing its projection onto the other columns. Here $x_j(t)$, defined in (2.5), is an integrated stochastic process that summarizes the past events of the jth feature of the multivariate Hawkes process. For theoretical convenience, when calculating S_{ij}, we scale $x_j(t)$ by the standard deviation of the ith process at time t, $\sigma_i(t)$ (see details in Section 2.3).

Theorem 2.1: This theorem establishes the convergence of the test statistics $\widehat{U}_T$ (2.23) to a χ^2-distribution under the null hypothesis. While the result is comparable to that in Neykov et al. [2018] and Zheng and Raskutti [2019], in our case $\widehat{U}_T$ is a function of the estimate of the time-varying variance of the point process, $\widehat{\sigma}^2(t)$. Proving the convergence in this case is different and requires additional care. To this end, we (i) show the convergence in probability of $\widehat{U}_T$ to the test statistic $\widehat{U}_T^0$, which is defined similar to $\widehat{U}_T$ but with $\widehat{\sigma}^2(t)$ replaced with the true $\sigma^2(t)$; (ii) show that $\widehat{U}_T^0$ converges in probability to U_T; and (iii) establish that U_T weakly converges to a χ^2-distribution. Next, we provide some details on each of these steps.

To show the weak convergence of U_T to a χ^2-distribution under the null hypothesis, we adopt the recently developed martingale central limit theorem (CLT) [Grama and Haeusler,

2006, Zheng and Raskutti, 2019], which is given as a special case of Proposition A.1 (for $\Delta = 0$).

To show the convergence of $\widehat{U}_T^0$ to U_T, we expand the difference in terms of differences between $\widehat{S}_{ij}^0$ and S_{ij}, and $\widehat{\Upsilon}_j^0$ and Υ_j, and bound each term. Here, $\widehat{S}_{ij}^0$ and $\widehat{\Upsilon}_j^0$ are estimates of S_{ij} and Υ_j but with the true $\sigma^2(t)$. By the construction of the decorrelated score [Neykov et al., 2018], bounding the difference between $\widehat{S}_{ij}^0$ and S_{ij} is equivalent to evaluating the estimation error of the lasso estimators of $(\mu_i, \boldsymbol{\beta}_i)$ and $\boldsymbol{w}_j^*$. Bounds on these estimation errors are given in Lemmas A.11 and A.12, respectively, in both ℓ_1- and ℓ_2-norms. To establish these bounds, we show that the restricted eigenvalue condition [Bickel et al., 2009] is satisfied with high probability in our case, by using the concentration bounds for the first and second order statistics of $\boldsymbol{x}(t)$ (shown in Lemma A.17). We also show that the eigenvalue of the covariance matrix of $\boldsymbol{x}(t)$ are bounded under Assumptions 2.1–2.4 (shown in Proposition A.2). To bound the difference between $\widehat{S}_{ij}^0$ and S_{ij}, we need to bound the ℓ_∞ norms of (i) the first and the second order statistics of $z_j^*(t)$, and (ii) the average of the scaled version of the residual $\epsilon_i(t)$ and its inner product with the scaled integrated feature, $z_j^*(t)$. The bounds for (i) are presented in Lemma A.18 using the concentration bound for the first and second order statistics of $\boldsymbol{x}(t)$ (shown in Lemma A.17). The bounds for (ii) are given in Lemma A.18 by a direct application of a martingale inequality [van de Geer, 1995] using the fact that $x_j(t), \sigma_i(t)$ are bounded under Assumptions 2.3 and 2.4. The proof for the convergence of $\widehat{\Upsilon}_j^0$ to Υ_j, which is shown in Lemma A.4, is similar to that for the convergence of $\widehat{S}_{ij}^0$ to S_{ij}; however, we only need the estimation error bound for $\boldsymbol{w}_j$ as the construction of Υ_j only involves the integrated process (or the design columns) $\boldsymbol{x}(t)$.

To complete the proof of Theorem 2.1, we establish the convergence of $\widehat{U}_T$ to $\widehat{U}_T^0$ in Lemma A.5. The proof of the lemma first expands the difference between $\widehat{U}_T$ and $\widehat{U}_T^0$ in terms of differences between (i) $\widehat{S}_{ij}^0$ to $\widehat{S}_{ij}$ (2.21), and (ii) $\widehat{\Upsilon}_j^0$ and $\widehat{\Upsilon}_j$ (2.19). (ii) can be bounded using Lemma A.4. To quantify the bound of (i), we first replace $\sigma^2(t)$ by $\mu_i + \boldsymbol{x}^\top(t)\boldsymbol{\beta}_i$ in $\widehat{S}_{ij}^0$ and then bound the term by carefully applying the bound for convergence of $(\widehat{\mu}_i, \widehat{\boldsymbol{\beta}}_i)$ to

$(\mu_i, \boldsymbol{\beta}_i)$ by Assumption 2.5 which is also proved to be satisfied in Lemma A.11.

Theorem 2.2: This theorem characterizes the behavior of the test statistics $\widehat{U}_T = \|\widehat{V}_T\|_2^2$ (2.23) under the alternative hypothesis. More specifically, it shows that for different signal strengths characterized by parameter ϕ, the test statistic behaves differently around the cutoff of $\phi = 1/2$. This is because (i) $\widehat{U}_T$ converges to $U_T = \|V_T\|_2^2$ defined in (2.15) in probability (see the following paragraphs for the proof outline), and (ii) under the alternative, we have $\mathbb{E}(V_T) = O(T^{1/2-\phi}\Delta)$. If the alternative signal is too small with $\phi > 1/2$, the expectation of V_T converges to 0 as T increases. In this case, we can show that $\widehat{U}_T$ converges weakly to a central χ^2 distribution, and is hence indistinguishable from null. In contrast, if the signal strength is too large with $\phi < 1/2$, then V_T diverges as T goes to ∞. Finally, when the alternative signal is moderate with $\phi = 1/2$, the expectation of V_T is a constant, $\widetilde{\Delta} = \Upsilon_j^{1/2}\Delta$. In this case, we apply Proposition A.1 to show the weak convergence of $\widehat{U}_T$ to a non-central χ^2 distribution with non-centrality parameter $\|\widetilde{\Delta}\|_2^2$.

For the case of $\phi > 1/2$, we use a similar strategy as that for the proof of Theorem 2.1 under the null hypothesis. More specifically, we split the proof into two parts by bounding $\widehat{U}_T - \widehat{U}_T^0$ and $\widehat{U}_T^0 - U_T$. U_T is shown to follow a central χ^2 distribution in Proposition A.1. The bound of $\widehat{U}_T - \widehat{U}_T^0$ is given in Lemma A.5. Similar to in the proof of Theorem 2.1, the key in bounding $\widehat{U}_T^0 - U_T$ under the alternative hypothesis is also to bound $\sqrt{T}\left\|\left(\Upsilon_j^0\right)^{-1/2}(\widehat{S}_{ij}^0 - S_{ij})\right\|_2$. The difference from the proof of Theorem 2.1 is an extra term involving $T^{1/2-\phi}\Delta$ in the bound of $\|\widehat{S}_{ij}^0 - S_{ij}\|_2$ because under the alternative hypothesis, $\beta_{ij} = T^{-\phi}\Delta$. Such difference leads to an extra term of order $O\left(s^2\rho^2\left(T^{\frac{7}{5}-2\phi}\right)\right)$ in bounding $\widehat{U}_T^0 - U_T$. As a result, we reach a modified rate of weak convergence of $\widehat{U}_T$ in this case compared to the rate in Theorem 2.1.

For the case of $\phi = 1/2$, our proof uses a strategy similar to the proof of Theorem 2.1. Specifically, we split the proof by bounding $\widehat{U}_T^0$ to $\widehat{U}_T$, and $\widehat{U}_T^0$ to $V_T + \widetilde{\Delta}$. The proof of the first part is the same as that for Theorem 2.1 and is given in Lemma A.5. Although

the second part involves non-zero $\widetilde{\Delta}$, the proof strategy is similar and involves explicitly expressing the difference between $\widehat{U}_T^0$ and $V_T + \widetilde{\Delta}$ in terms of the difference between $\widehat{S}_{ij}^0$ and S_{ij}, the difference between $\widehat{\Upsilon}_j^0$ and $\widehat{\Upsilon}_j$, and an extra item involving Δ. The first two items are bounded using the same strategy as in Theorem 2.1. The term involving Δ is also bounded since the leading term involves the ℓ_∞ bound of the first and the second order statistics of $\boldsymbol{z}(t)$, which are bounded by Lemma A.18. We complete the proof by applying Proposition A.1 to show that $V_T + \widetilde{\Delta}$ converges weakly to $\chi^2_{d,\|\widetilde{\Delta}\|_2^2}$.

For the case of $\phi < 1/2$, the test statistic diverges in probability as T goes to ∞. Therefore, we derive a lower bound for $\widehat{U}_T$ which requires quantifying the lower bound of $\widehat{S}_{ij} - S_{ij}$ and the upper bound of $\|V_T\|_2$. Due to the estimate on the unknown variance involved in $\widehat{S}_{ij}$, we split the difference in the first part into (i) $\widehat{S}_{ij}^0 - S_{ij}$ and (ii) $\widehat{S}_{ij} - \widehat{S}_{ij}^0$. We show that the lower bound of part (i) is in order of $T^{1/2-\phi}$ under the setting of the alternative of $\phi < 1/2$. To bound part (ii), similar as before, we expand the term into components involving differences between the estimates of $(u_i, \boldsymbol{\beta}_i)$ and $\boldsymbol{w}_j^*$, whose error bounds are assumed in Assumptions 2.5 and 2.6, and parts with the first and second order statistics of $\boldsymbol{x}(t)$ and $\boldsymbol{z}(t)$ given in Lemmas A.17 and A.18. To quantify the upper bound of $\|V_T\|_2$, we apply a result on the tail bound of the central χ^2 distribution since $\|V_T\|_2^2$ follows a central χ^2-distribution by Proposition A.1; this bound is also used to prove Theorem 3.2 of Zheng and Raskutti [2019].

Theorem 2.3: Recall from (2.30) that $\widehat{b}$ is the one-step debiased lasso estimates of $\boldsymbol{\beta}_i$. This theorem shows that $\widehat{R} = T\left(\widehat{b}_{ij} - \beta_{ij}\right)^\top \widehat{\Upsilon}_j \left(\widehat{b}_{ij} - \beta_{ij}\right)$ converges weakly to a central χ^2 distribution. This allows us to construct the optimal confidence regions in (2.32), since the asymptotic variance of $\widehat{R}$ is close to the inverse of the partial information, $\Upsilon_j = \mathrm{Cov}\left(z_j^*(t)\right)$ (2.13).

For this proof, we introduce a new quantity $\check{S}_{ij}$:

$$\begin{aligned}\check{S}_{ij} &= \frac{1}{T}\sum_{t=1}^{T}\frac{Y_i(t) - x_j(t)\beta_{i,j} - \widehat{\mu}_i - \boldsymbol{x}_{-j}^{\top}(t)\widehat{\boldsymbol{\beta}}_{i,-j}}{\widehat{\sigma}_i(t)}\widehat{z}_j^*(t) \\ &= \frac{1}{T}\sum_{t=1}^{T}\frac{\epsilon_i(t) + (\mu_i - \widehat{\mu}_i) + \boldsymbol{x}_{-j}^{\top}(t)(\boldsymbol{\beta}_{i,-j} - \widehat{\boldsymbol{\beta}}_{i,-j})}{\widehat{\sigma}_i(t)}\widehat{z}_j^*(t),\end{aligned}$$

which is equivalent to the $\widehat{S}_{ij}$ defined in (2.21) under the null, and similar to that under the alternative, except that here $\Delta = 0$ or $\phi = \infty$. Letting $\check{U}_T = \check{S}_{ij}^{\top}\widehat{\Upsilon}_j^{-1}\check{S}_{ij}$, we then show that $\check{U}_T$ weakly converges to a χ^2 distribution under both null and alternative hypotheses. For the null, we follow Theorem 2.1, while for the alternative, we repeat the steps in Theorem 2.2 (case $\phi > 1/2$) but replace $\Delta = 0$ or $\phi = \infty$ throughout.

Since by the construction of $\widehat{R}_T$, we have $\widehat{R}_T = \left(\check{S}_{ij}\right)^{\top}\left(\widetilde{\Upsilon}_j\right)^{-1}\widehat{\Upsilon}_j\left(\widetilde{\Upsilon}_j\right)^{-1}\check{S}_{ij}$, what is left is to bound the difference between $\widehat{R}_T$ and $\check{U}_T$. This requires to bound $\widetilde{\Upsilon}_j - \widehat{\Upsilon}_j$ where $\widetilde{\Upsilon}_j$ and $\widehat{\Upsilon}_j$ are defined in (2.29) and (2.19), respectively. To obtain such a bound with the unknown variance, we define $\widetilde{\Upsilon}_j^0$ and $\widehat{\Upsilon}_j^0$ similar to $\widetilde{\Upsilon}_j$ and $\widehat{\Upsilon}_j$, but with $\widehat{\sigma}^2$ replace by the true $\sigma^2(t)$. We then separately bound (i) $\widetilde{\Upsilon}_j - \widetilde{\Upsilon}_j^0$, (ii) $\widetilde{\Upsilon}_j^0 - \widehat{\Upsilon}_j^0$, and (iii) $\widehat{\Upsilon}_j^0 - \widehat{\Upsilon}_j$: (i) is bounded according to the consistency of $\widehat{\sigma}^2(t)$ for the true $\sigma^2(t)$ following a similar strategy used in Lemma A.18; (ii) is bounded by carefully evaluating the lasso estimation error of $\boldsymbol{w}_j^*$; (iii) is bounded by Lemma A.18.

Key steps in proof of lemmas: Proofs of supporting lemmas crucially rely on properties of the integrated stochastic process $\boldsymbol{x}(t) = \begin{pmatrix} x_1(t) & \cdots & x_p(t)\end{pmatrix}$, which summarizes all the past event history of the Hawks process. More specifically, the main theorems rely on the bounded eigenvalue of $\Upsilon_x = \mathrm{Cov}(\boldsymbol{x}(t))$ (Proposition A.2) and the concentration bounds on the first and second order statistics of scaled $\boldsymbol{x}(t)$ (i.e. $\boldsymbol{z}(t)$) (Lemmas A.17 and A.18). These results are used to show the condition of the martingale CLT (Proposition A.1), the restrict eigenvalue (RE) condition used in the estimation consistency of lasso (Lemma A.11 and A.12), and the convergence rate of $\widehat{\Upsilon}_j$ (Lemma A.4).

A key challenge in establishing results for the integrated process stems from the complicated (and non-Markovian) dependence structure of the Hawkes process. In particular, each column of $\boldsymbol{x}(t)$, $x_j(t)$, is a stochastic process with non-trivial serial dependence due to the integration over the past history. To show that the eigenvalues of $\Upsilon_x = \mathrm{Cov}(\boldsymbol{x}(t))$ are bounded, we show that the eigenvalues of the cross-variance for a stationary stochastic process can be bounded by its spectral density in the Hawkes process. The proof follows a similar strategy as Basu and Michailidis [2015] in the VAR models, but is specialized for the Hawkes process in a continuous time domain with an integrable transition kernel. Next, we establish a relationship between the spectral densities of the cross-covariance and the transition matrix of the Hawkes process by generalizing Theorem 1 in Bacry et al. [2011] or Theorem 3 in Etesami et al. [2016] (where they assume a non-negative transfer function) to real-value transfer functions. At last, we utilize the martingale inequality in van de Geer [1995] and the concentration inequality on weak dependent samples in Chen et al. [2019] to establish the concentration bounds on the first and second order statistics of $\boldsymbol{x}(t)$ in Lemma A.17. The final concentration bound for $\boldsymbol{z}(t)$ in Lemma A.18 directly follows from the deviation bounds for $\boldsymbol{x}(t)$ but requires a lengthy derivation, due to the scaling factor of the time-varying variance of the point process, $\sigma^2(t)$, in $\boldsymbol{z}(t)$.

A.2 Proof of Main Results

For ease of notations, we prove the result for testing an univariate β_{ij}; i.e., $d = 1$. Our proof can be extended to $d > 1$ by replacing the scalars by vectors or matrices in the corresponding norms when needed. In the following, we use C_k, c_k with some subscript k to represent constants that only depend on the model parameter $(\boldsymbol{\mu}, \Theta)$ and the transition kernel function.

Proof of Theorem 2.1

Our proof strategy is to relate $\widehat{U}_T$ in (2.23) to U_T in (2.15), and show that U_T converges weakly to χ^2_d. Directly comparing $\widehat{U}_T$ with U_T is difficult due to the unknown time-varying variance $\sigma_i^2(t)$ involved. Therefore, we start by introducing $\widehat{U}_T^0$ and then show that $\widehat{U}_T \approx \widehat{U}_T^0$. Let

$$\widehat{S}_{ij}^0 = \frac{1}{T}\sum_{t=1}^{T} \frac{1}{\sigma_i(t)} \left(Y_i(t) - \widehat{\mu}_i - \boldsymbol{x}_{-j}^\top(t)\widehat{\boldsymbol{\beta}}_{i,-j}\right)\left(x_j(t)/\sigma_i(t) - \widehat{w}_{j0} - \boldsymbol{x}_{-j}^\top(t)/\sigma_i(t)\widehat{\boldsymbol{w}}_{j,-j}\right), \tag{A.1}$$

$$\widehat{\Upsilon}_j^0 = \frac{1}{T}\sum_{t=1}^{T} \left(x_j(t)/\sigma_i(t) - \widehat{w}_{j0} - \boldsymbol{x}_{-j}^\top(t)/\sigma_i(t)\widehat{\boldsymbol{w}}_{j,-j}\right)^2, \tag{A.2}$$

$$\widehat{V}_T^0 = \sqrt{T}\left(\widehat{\Upsilon}_j^0\right)^{-1/2}\widehat{S}_{ij}^0, \tag{A.3}$$

$$\widehat{U}_T^0 = \|\widehat{V}_T^0\|_2^2. \tag{A.4}$$

The difference between $\widehat{U}_T$ and $\widehat{U}_T^0$ is that we replace $\widehat{\sigma}_i(t)$ by $\sigma_i(t)$.

For any $\delta > 0$, we have

$$\begin{aligned}\mathbb{P}\left(\widehat{U}_T \le x\right) - F_d(x) &\le \mathbb{P}(U_T \le x+\delta) + \mathbb{P}\left(\left|\widehat{U}_T - \widehat{U}_T^0\right| > \delta\right) - F_d(x) \\ &\le \left|\mathbb{P}(U_T \le x+\delta) - F_d(x+\delta)\right| + F_d(x+\delta) - F_d(x) + \mathbb{P}\left(\left|\widehat{U}_T - \widehat{U}_T^0\right| > \delta\right),\end{aligned}$$

and

$$F_d(x) - \mathbb{P}\left(\widehat{U}_T \le x\right) \le \mathbb{P}(\widehat{U}_T^0 > x - \delta) + \mathbb{P}\left(\left|\widehat{U}_T - \widehat{U}_T^0\right| > \delta\right) - (1 - F_d(x))$$
$$\le |F_d(x-\delta) - \mathbb{P}(U_T \le x - \delta)| + F_d(x) - F_d(x-\delta) + \mathbb{P}\left(\left|\widehat{U}_T - \widehat{U}_T^0\right| > \delta\right).$$

Combining the two inequalities gives

$$\left|\mathbb{P}\left(\widehat{U}_T \le x\right) - F_d(x)\right| \le \sup_{y\in\mathbb{R}}\left|\mathbb{P}\left(\widehat{U}_T^0 \le y\right) - F_d(y)\right|$$
$$+ F_d(x+\delta) - F_d(x-\delta) + \mathbb{P}\left(\left|\widehat{U}_T - \widehat{U}_T^0\right| > \delta\right). \quad \text{(A.5)}$$

Next, we bound

$$\left|\mathbb{P}\left(\widehat{U}_T^0 \le x\right) - F_d(x)\right|$$

by showing $\widehat{U}_T^0 \approx U_T \approx \chi_d^2$. Following a similar deduction as (A.5), for any $\epsilon > 0$, we have

$$\left|\mathbb{P}(\widehat{U}_T^0 \le x) - F_d(x)\right| \le \underbrace{\sup_{y\in\mathbb{R}}|\mathbb{P}(U_T \le y) - F_d(y)|}_{A}$$
$$+ \underbrace{F_d(x+\epsilon) - F_d(x-\epsilon)}_{B} + \underbrace{\mathbb{P}(|\widehat{U}_T^0 - U_T| > \epsilon)}_{C}. \quad \text{(A.6)}$$

Direct application of Proposition A.1 shows that $\mathrm{A} = O\left(T^{-1/8}\right)$. Using the fact that a χ_d^2 random variable has bounded density gives $|\mathrm{B}| = O(\epsilon)$. Thus, we control the term C in the rest of the proof. Notice that

$$\begin{aligned}
\left|\widehat{U}_T^0 - U_T\right| &= \left|T\left(\widehat{S}_{ij}^0\right)^\top \left(\widehat{\Upsilon}_j^0\right)^{-1} \widehat{S}_{ij}^0 - S_{ij}^T \Upsilon_j^{-1} S_{ij}\right| \\
&\le \left|T\left(\widehat{S}_{ij}^0\right)^\top \left(\left(\widehat{\Upsilon}_j^0\right)^{-1} - \Upsilon_j^{-1}\right)\widehat{S}_{ij}^0 + T\left(\widehat{S}_{ij}^0\right)^\top \Upsilon_j^{-1}\widehat{S}_{ij} - S_{ij}^T\Upsilon_j^{-1}S_{ij}\right| \\
&\le \left\|\Upsilon_j^{1/2}\left(\widehat{\Upsilon}_j^0\right)^{-1}\Upsilon_j^{1/2} - I\right\|_\infty \left\|\sqrt{T}\Upsilon_j^{-1/2}\widehat{S}_{ij}^0\right\|_1^2 \\
&\quad + \|\sqrt{T}\Upsilon_j^{-1/2}(S_{ij} - \widehat{S}_{ij}^0)\|_1^2 \\
&\quad + 2\|V_T\|_2 \left\|\sqrt{T}\Upsilon_j^{-1/2}(S_{ij} - \widehat{S}_{ij}^0)\right\|_2. \quad \text{(A.7)}
\end{aligned}$$

Let $E = \sqrt{T}\Upsilon_j^{-1/2}(S_{ij} - \widehat{S}_{ij}^0)$, where Υ_j is defined in (2.13). Then

$$|\widehat{U}_T^0 - U_T| \leq \|E\|_2^2 + 2\|V_T\|_2\|E\|_2 + \left\|\Upsilon_j^{1/2}\left(\widehat{\Upsilon}_j^0\right)^{-1}\Upsilon_j^{1/2} - I\right\|_\infty (\|V_T\|_2 + \|E\|_2)^2 . \quad \text{(A.8)}$$

Next, we provide probabilistic bounds for $\|E\|_2$, $\|V_T\|_2$, and $\left\|\Upsilon_j^{1/2}\left(\widehat{\Upsilon}_j^0\right)^{-1}\Upsilon_j^{1/2} - I\right\|_\infty$ in order to bound $|\widehat{U}_T^0 - U_T|$. First, from Proposition A.2 and Lemma A.3, we get $\|E\|_2^2 = O_p\left(s^2\rho^4 T^{-3/5}\right)$, with probability at least $1 - c_1p^2T\exp(-c_2T^{1/5})$. Second, Lemma A.9 and Proposition A.1 lead $\mathbb{P}(\|V_T\|_2 > T^{1/10}) = O\left(T^{-1/8}\right)$. Third, Lemma A.4 gives

$$\left\|\Upsilon_j^{1/2}\left(\widehat{\Upsilon}_j^0\right)^{-1}\Upsilon_j^{1/2} - I\right\|_\infty = O_p(s^2\rho^2T^{-2/5}),$$

with probability at least $1 - c_3p^2T\exp(-c_4T^{1/5})$. Therefore,

$$\mathbb{P}\left(|\widehat{U}_T^0 - U_T| > c_5s^2\rho^2T^{-1/5}\right) \leq c_6p^2T\exp(-c_7T^{1/5}) + c_8T^{-1/8}. \quad \text{(A.9)}$$

Turning back to (A.6) with the bounds of A, B, and C gives us

$$\sup_{x\in\mathbb{R}}|\mathbb{P}(\widehat{U}_T^0 \leq x) - F_d(x)| \leq c_6p^2T\exp(-c_7T^{1/5}) + c_8T^{-1/8} + c_9s^2\rho^2T^{-1/5}. \quad \text{(A.10)}$$

By the bounded density of χ^2 distribution

$$|F_d(x+\delta) - F_d(x-\delta)| \leq C(d)\delta.$$

Finally, Lemma A.5 gives us

$$\mathbb{P}\left(|\widehat{U}_T - \widehat{U}_T^0| > c_{10}\rho T^{-1/5}\right) \leq c_{11}p^2T\exp(-c_{12}T^{1/5}) + c_{13}T^{-1/8} + c_{14}s^2\rho^2T^{-1/5}.$$

Combining the last three displays gives us a bound on (A.5), which completes the proof.

Proof of Theorem 2.2

We study the three cases separately.

Case: $\phi > 1/2$. The proof for this case is similar to the proof of Theorem 2.1. However, due to $\beta_{ij} \neq 0$ under the alternative hypothesis, we have a modified rate of convergence depending

on the scale of ϕ. A new bound is needed for $\left|\widehat{U}_T^0 - U_T\right|$, which is obtained modifying the bound for the difference between $\widehat{S}_{ij}^0$ and S_{ij}. To be specific,

$$\begin{aligned}
\widehat{S}_{ij}^0 - S_{ij} = &(\widehat{\boldsymbol{w}}_j - \boldsymbol{w}_j^*)^\top \frac{1}{T}\sum_{t=0}^{T-1} \widetilde{\epsilon}_i(t)\boldsymbol{z}_{-j}^\top(t) \\
&+ \frac{1}{T}\sum_{t=0}^{T-1} z_j^* \begin{pmatrix} 1 & \boldsymbol{z}_{-j}^\top(t) \end{pmatrix} \left(\begin{pmatrix} \widehat{\mu}_i \\ \widehat{\boldsymbol{\beta}}_{i,-j} \end{pmatrix} - \begin{pmatrix} \mu_i \\ \boldsymbol{\beta}_{i,-j} \end{pmatrix} \right) \\
&- (\widehat{\boldsymbol{w}}_j - \boldsymbol{w}_j^*)^\top \left(\frac{1}{T}\sum_{t=0}^{T-1} \begin{pmatrix} 1 \\ \boldsymbol{z}_{-j}(t) \end{pmatrix} \begin{pmatrix} 1/\sigma_i(t) & \boldsymbol{z}_{-j}^\top(t) \end{pmatrix} \right) \left(\begin{pmatrix} \widehat{\mu}_i \\ \widehat{\boldsymbol{\beta}}_{i,-j} \end{pmatrix} - \begin{pmatrix} \mu_i \\ \boldsymbol{\beta}_{i,-j} \end{pmatrix} \right) \\
&- T^{-\phi}\frac{1}{T}\sum_{t=1}^{T} \left(z_j(t) - \begin{pmatrix} 1 & \boldsymbol{z}^\top(t) \end{pmatrix} \widehat{\boldsymbol{w}}_j \right) z_j(t)\Delta. \qquad \text{(A.11)}
\end{aligned}$$

The first three terms are bounded by $O_p(s\rho^2 T^{-4/5})$, as shown in the proof of Lemma A.3 (see (A.33)). The last term shows up because under the alternative hypothesis setting, $\beta_{ij} = T^{-\phi}\Delta \neq 0$. Now we examine the fourth item in (A.11). By Lemma A.8, $\|z_j(t)\|_\infty = O(1)$. Then,

$$\left\| \frac{1}{T}\sum_{t=1}^{T} \left(z_j(t) - \begin{pmatrix} 1 & \boldsymbol{z}^\top(t) \end{pmatrix} \widehat{\boldsymbol{w}}_j \right) z_j(t)\Delta \right\|_2 \leq |\Delta| \|z_j(t)\|_\infty^2 = O(1).$$

Using the same notation as the proof in Theorem 2.1, let $E = \sqrt{T}\left(\Upsilon_j\right)^{-1/2}\left(\widehat{S}_{ij}^0 - S_{ij}\right)$. Then, with the bounds in each part of (A.11),

$$|E| = O_p\left(s\rho^2 T^{-3/10} + T^{1/2-\phi}\right), \qquad \text{(A.12)}$$

with probability at least $1 - c_1 p^2 T \exp(-c_2 T^{1/5})$. Here we see that $\phi > 1/2$ is important to bound the term; otherwise, it is not guaranteed that the term is bounded as T increases.

By repeating the same steps in Theorem 2.1 with the bound of $|E|$, we show

$$\left|\widehat{U}_T^0 - U_T\right| = O_p\left(s^2\rho^2\left(T^{-1/5} \vee T^{\frac{7}{5}-2\phi}\right)\right),$$

with probability at least

$$1 - c_3 p^2 T \exp(-c_4 T^{1/5}) - c_5 s^2 \rho^2 \left(T^{-1/5} \vee T^{\frac{7}{5}-2\phi}\right)$$

and

$$\sup_{x\in\mathbb{R}}|\mathbb{P}(\widehat{U}_T^0\le x)-F_d(x)|\le c_3p^2T\exp(-c_4T^{1/5})+c_5s^2\rho^2\left(T^{-1/5}\vee T^{\frac{7}{5}-2\phi}\right)+c_6T^{-1/8}.$$

At last, using (A.5) and Lemma A.5 to bound $|\widehat{U}_T-\widehat{U}_T^0|$, we reach the conclusion.

Case: $\phi=1/2$. The proof strategy here is to quantify the difference between the cdf of $\widehat{U}_T^0$ and χ^2 distribution. For any $\epsilon>0$, we have

$$\begin{aligned}\left|\mathbb{P}(\widehat{U}_T^0\le x)-F_{d,\|\widetilde{\Delta}\|_2^2}(x)\right|\le&\underbrace{\sup_{y\in\mathbb{R}}\left|\mathbb{P}(\|V_T+\widetilde{\Delta}\|_2^2\le y)-F_{d,\|\widetilde{\Delta}\|_2^2}(y)\right|}_{A}\\&+\underbrace{F_{d,\|\widetilde{\Delta}\|_2^2}(x+\epsilon)-F_{d,\|\widetilde{\Delta}\|_2^2}(x-\epsilon)}_{B}+\underbrace{\mathbb{P}\left(\left|\widehat{U}_T^0-\|V_T+\widetilde{\Delta}\|_2^2\right|>\epsilon\right)}_{C},\end{aligned}$$

where $\widetilde{\Delta}=\Upsilon_j^{1/2}\Delta$, and $\widehat{U}_T^0$, V_T are defined in (A.4) and (2.14), respectively.

Lemma A.1 gives us $A=O\left(T^{-1/8}\right)$. By the bounded density of non-central χ^2 distribution, we have $B=O(\epsilon)$. Therefore, in the rest of the proof, we bound part C.

Let $E=\widehat{V}_T^0-V_T-\widetilde{\Delta}$, where $\widehat{V}_T^0=\sqrt{T}\left(\Upsilon_j\right)^{-1/2}\widehat{S}_{ij}^0$ is defined in (A.3) and $\widehat{U}_T^0=\|\widehat{V}_T^0\|_2^2$. Then,

$$\begin{aligned}\left|\widehat{U}_T^0-\left\|V_T+\widetilde{\Delta}\right\|_2^2\right|\le&\|E\|_2^2+\left\|V_T+\widetilde{\Delta}\right\|_2\|E\|_2\\&+\left\|\left(\Upsilon_j\right)^{1/2}\left(\widehat{\Upsilon}_j^0\right)^{-1}\left(\Upsilon_j\right)^{1/2}-I\right\|_\infty\left(\left\|V_T+\widetilde{\Delta}\right\|_2+\|E\|_2\right)^2.\end{aligned}\tag{A.13}$$

We bound $\left\|\left(\Upsilon_j\right)^{1/2}\left(\widehat{\Upsilon}_j^0\right)^{-1}\left(\Upsilon_j\right)^{1/2}-I\right\|_\infty$ using Lemma A.4, as we did in Theorem 2.1. However, the bounds for E and $V_T+\widetilde{\Delta}$ need to be modified.

Let $\boldsymbol{w}_j$ be such that $z_j^*(t)=\begin{pmatrix}1 & \boldsymbol{z}^\top(t)\end{pmatrix}\boldsymbol{w}_j$, where

$$\boldsymbol{w}_j=\left(w_{j0}^*,\{w_{jl}^*\mathbf{1}(l\ne j)+\mathbf{1}(l=j)\}_{1\le l\le p}\right)^\top\in\mathbb{R}^{p+1}.\tag{A.14}$$

Then

$$\Upsilon_j=\mathrm{Cov}(z_j^*(t))=\mathrm{Cov}\left(\begin{pmatrix}1 & \boldsymbol{z}^\top(t)\end{pmatrix}\boldsymbol{w}_j\right)=\boldsymbol{w}_j^\top\Upsilon\boldsymbol{w}_j,$$

where $\Upsilon = \mathbb{E}\left(\begin{pmatrix}1\\ \boldsymbol{z}(t)\end{pmatrix}\begin{pmatrix}1 & \boldsymbol{z}^\top(t)\end{pmatrix}\right)$ and $\|\boldsymbol{w}_j\|_1 = \|\boldsymbol{w}_j^*\|_1 + 1$. Recall that $S_{ij} = \frac{1}{T}\sum_{t=1}^T \widetilde{\epsilon}_i(t)\, z_j^*(t)$ as defined in (2.12). In addition, define $\widetilde{S}_{ij} = S_{ij} + \frac{1}{T}\sum_{t=1}^T z_j(t)\beta_{ij} z_j^*$. Thus, with $\beta_{ij} = T^{-1/2}\Delta$,

$$\begin{aligned}
V_T + \widetilde{\Delta} &= \sqrt{T}\,(\Upsilon_j)^{-1/2}\, S_{ij} + \Upsilon_j^{1/2}\Delta \\
&= \sqrt{T}\,(\Upsilon_j)^{-1/2}\,(S_{ij} + \Upsilon_j\beta_{ij}) \\
&= \sqrt{T}\,(\Upsilon_j)^{-1/2}\left(S_{ij} + z_j(t)\beta_{ij}z_j^* - z_j(t)\beta_{ij}z_j^* + \Upsilon_j\beta_{ij}\right) \\
&= \sqrt{T}\,(\Upsilon_j)^{-1/2}\left(\widetilde{S}_{ij} - \boldsymbol{w}_j^\top\left(\frac{1}{T}\sum_{t=1}^T\begin{pmatrix}1\\ \boldsymbol{z}(t)\end{pmatrix} z_j(t) - \Upsilon_{\cdot,j}\right)\beta_{ij}\right).
\end{aligned}$$

Then, by Lemma A.2 of the bounded eigenvalue of Υ_j,

$$\begin{aligned}
\|E\|_2 &= \|\widehat{V}_T^0 - V_T - \widetilde{\Delta}\|_2 \\
&= \left\|\sqrt{T}\,(\Upsilon_j)^{-1/2}\,\widehat{S}_{ij}^0 - \sqrt{T}\,(\Upsilon_j)^{-1/2}\left(\widetilde{S}_{ij} - \boldsymbol{w}_j^\top\left(\frac{1}{T}\sum_{t=1}^T\begin{pmatrix}1\\ \boldsymbol{z}(t)\end{pmatrix} z_j(t) - \Upsilon_{\cdot,j}\right)\beta_{ij}\right)\right\|_2 \\
&\le \left\|\sqrt{T}\,(\Upsilon_j)^{-1/2}\left(\widehat{S}_{ij}^0 - \widetilde{S}_{ij}\right)\right\|_2 \\
&+ \left\|(\Upsilon_j)^{-1/2}\,\boldsymbol{w}_j\left(\frac{1}{T}\sum_{t=1}^T\begin{pmatrix}1\\ \boldsymbol{z}(t)\end{pmatrix}\begin{pmatrix}1 & \boldsymbol{z}^\top(t)\end{pmatrix} - \Upsilon_{\cdot,j}\right)\sqrt{T}\beta_{ij}\right\|_2 \\
&= O\left(\sqrt{T}\left\|\widehat{S}_{ij}^0 - \widetilde{S}_{ij}\right\|_2\right) \\
&+ O\left(\left|\sqrt{T}\beta_{ij}\right|\|\boldsymbol{w}_j\|_1\left\|\frac{1}{T}\sum_{t=1}^T\begin{pmatrix}1\\ \boldsymbol{z}(t)\end{pmatrix}\begin{pmatrix}1 & \boldsymbol{z}^\top(t)\end{pmatrix} - \Upsilon\right\|_\infty\right).
\end{aligned} \tag{A.15}$$

First, we examine the second item on RHS. Using Lemma A.14, $\|\boldsymbol{w}_j\|_1 = O(\sqrt{s})$, and using Lemma A.18 we get

$$\left\|\frac{1}{T}\sum_{t=1}^T\begin{pmatrix}1\\ \boldsymbol{z}(t)\end{pmatrix}\begin{pmatrix}1 & \boldsymbol{z}^\top(t)\end{pmatrix} - \Upsilon\right\|_\infty = O_p(\rho T^{-2/5}),$$

with probability at least $1 - c_1 p^2 T \exp(-c_2 T^{1/5})$. Then, with $\beta_{ij} = T^{-1/2}\Delta$,

$$\left|\sqrt{T}\beta_{ij}\right| \|\boldsymbol{w}_j\|_1 \left\| \frac{1}{T}\sum_{t=1}^{T} \begin{pmatrix} 1 \\ \boldsymbol{z}(t) \end{pmatrix} \begin{pmatrix} 1 & \boldsymbol{z}^\top(t) \end{pmatrix} - \Upsilon \right\|_\infty = O_p\left(\sqrt{s}\rho T^{-2/5}\right).$$

Next, we quantify $\sqrt{T}\left\|\widehat{S}_{ij}^0 - \widetilde{S}_{ij}\right\|_2$. Expanding the difference, we have

$$\begin{aligned}
\widehat{S}_{ij}^0 - \widetilde{S}_{ij} =& (\widehat{\boldsymbol{w}}_j - \boldsymbol{w}_j^*)^\top \frac{1}{T}\sum_{t=1}^{T} \begin{pmatrix} 1 \\ \boldsymbol{z}_{-j}(t) \end{pmatrix} \widetilde{\epsilon}_i(t) \\
&+ \left(\begin{pmatrix} \widehat{\mu}_i & \widehat{\boldsymbol{\beta}}_{i,-j}^\top \end{pmatrix} - \begin{pmatrix} \mu_i & \boldsymbol{\beta}_{i,-j}^\top \end{pmatrix} \right) \frac{1}{T}\sum_{t=1}^{T} \begin{pmatrix} 1/\sigma_i(t) \\ \boldsymbol{z}_{-j}(t) \end{pmatrix} z_j^*(t) \\
&- \left(\begin{pmatrix} \widehat{\mu}_i & \widehat{\boldsymbol{\beta}}_{i,-j}^\top \end{pmatrix} - \begin{pmatrix} \mu_i & \boldsymbol{\beta}_{i,-j}^\top \end{pmatrix} \right) \frac{1}{T}\sum_{t=1}^{T} \begin{pmatrix} 1/\sigma_i(t) \\ \boldsymbol{z}_{-j}(t) \end{pmatrix} \begin{pmatrix} 1 & \boldsymbol{z}_{-j}^\top(t) \end{pmatrix} (\widehat{\boldsymbol{w}}_j - \boldsymbol{w}_j^*) \\
&+ (\widehat{\boldsymbol{w}}_j - \boldsymbol{w}_j^*)^\top \frac{1}{T}\sum_{t=1}^{T} \begin{pmatrix} 1 \\ \boldsymbol{z}_{-j}(t) \end{pmatrix} z_j(t)\beta_{ij}. \qquad \text{(A.16)}
\end{aligned}$$

According to the proof of Lemma A.3, the first three terms above are bounded as $O_p\left(s\rho^2 T^{-4/5}\right)$.

With $\|z_j(t)\|_\infty = O(1)$ by Lemma A.8, and $\|\widehat{\boldsymbol{w}}_j - \boldsymbol{w}_j^*\|_1 = O_p(s\rho T^{-2/5})$ in Assumption 2.6, the last term is bounded as

$$(\widehat{\boldsymbol{w}}_j - \boldsymbol{w}_j^*)^\top \frac{1}{T}\sum_{t=1}^{T} \begin{pmatrix} 1 \\ \boldsymbol{z}_{-j}(t) \end{pmatrix} z_j(t)\beta_{ij} = O\left(\|\widehat{\boldsymbol{w}}_j - \boldsymbol{w}_j^*\|_1 \cdot |\beta_{ij}|\right) = O_p\left(s\rho\Delta T^{-9/10}\right),$$

with probability at least $1 - c_3 p^2 T \exp(-c_4 T^{1/5})$, where $\beta_{ij} = T^{-1/2}\Delta$ as set in the alternative. Thus,

$$\left|\widehat{S}_{ij}^0 - \widetilde{S}_{ij}\right| = O_p\left(s\rho^2 T^{-4/5}\right) + O_p\left(s\rho\Delta T^{-9/10}\right) = O_p\left(s\rho^2 T^{-4/5}\right).$$

As a result,

$$\|E\|_2 = O_p\left(s\rho^2 T^{-3/10}\right) + O_p\left(\sqrt{s}\rho T^{-2/5}\Delta\right) = O_p\left(s\rho^2 T^{-3/10}\right),$$

with probability at least $1 - c_5 p^2 T \exp(-c_6 T^{1/5})$.

Next, we quantify the bound of $V_T + \widetilde{\Delta}$. Setting $y = T^{1/10}$ in (A.45) of Lemma A.9 and $\|V_T\|_2^2$ weakly converges to χ^2 in Proposition A.1, we get

$$P\left(\left\|V_T + \widetilde{\Delta}\right\|_2 > T^{1/10}\right) = O(T^{-1/8}). \tag{A.17}$$

At last, we take the bounds of $\left\|V_T + \widetilde{\Delta}\right\|_2$ and $\|E\|_2$ back to (A.13), and

$$\left|\widehat{U}_T^0 - \left\|V_T + \widetilde{\Delta}\right\|_2^2\right| = O_p\left(s^2\rho^2 T^{-1/5}\right),$$

with probability at least $1 - c_7 p^2 T \exp(-c_8 T^{1/5}) - c_9 T^{-1/8}$. Finally, using decomposition in (A.6), we get

$$\sup_{x\in\mathbb{R}} |\mathbb{P}(\widehat{U}_T^0 \le x) - F_{d,\|\widetilde{\Delta}\|_2^2}(x)| \le c_7 p^2 T \exp(-c_8 T^{1/5}) + c_{10} T^{-1/8} + c_{11} s^2 \rho^2 T^{-1/5}. \tag{A.18}$$

The conclusion follows from the decomposition in (A.5) and Lemma A.5 that quantify the bound of $|\widehat{U}_T - \widehat{U}_T^0|$.

Case: $\phi < 1/2$. Recall that S_{ij} is defined in (2.12) and V_T is defined in (2.14). First, notice that

$$\begin{aligned}
\widehat{U}_T &= T\widehat{S}_{ij}\left(\widehat{\Upsilon_j}\right)^{-1}\widehat{S}_{ij} \\
&= T\widehat{S}_{ij}\left(\Upsilon_j^{-1} + \left(\widehat{\Upsilon_j}\right)^{-1} - \Upsilon_j^{-1}\right)\widehat{S}_{ij} \\
&\ge T\|(\Upsilon_j)^{-1/2}\widehat{S}_{ij}\|_2^2\left(1 - d\left\|\Upsilon_j^{1/2}\left(\widehat{\Upsilon_j}\right)^{-1}\Upsilon_j^{1/2} - I\right\|_\infty\right) \\
&\ge c_1 T\|(\Upsilon_j)^{-1/2}\widehat{S}_{ij}\|_2^2 \\
&= c_1\left(\|\sqrt{T}\,(\Upsilon_j)^{-1/2}(\widehat{S}_{ij} - S_{ij})\|_2 - \|V_T\|_2\right)^2,
\end{aligned} \tag{A.19}$$

where the second inequality follows from Lemma A.4 as $\left\|\Upsilon_j^{1/2}\left(\widehat{\Upsilon_j^0}\right)^{-1}\Upsilon_j^{1/2} - I\right\|_\infty$ converges to 0 when $s^2\rho^2\log p = o(T^{1/5})$.

Next, we bound $\widehat{S}_{ij} - S_{ij}$ by separately bounding $\widehat{S}_{ij}^0 - S_{ij}$ and $\widehat{S}_{ij} - \widehat{S}_{ij}^0$, where $\widehat{S}_{ij}^0$ is defined in (A.1).

First,

$$\widehat{S}_{ij}^{0} - S_{ij} = \widehat{S}_{ij}^{0} - \widetilde{S}_{ij} + \frac{1}{T}\sum_{t=1}^{T} z_j^*(t) z_j^\top(t)\beta_{ij}. \tag{A.20}$$

A similar deduction as (A.16), but with $\beta_{ij} = T^{-\phi}\Delta$ leads to $\widehat{S}_{ij}^{0} - \widetilde{S}_{ij} = O_p(s\rho T^{-\frac{2}{5}-\phi})$. Thus, in the follows, we give a lower bound for $\frac{1}{T}\sum_{t=1}^{T} z_j^*(t) z_j^\top(t)\beta_{ij}$.

By the construction of $z_j^*(t)$ and the projection coefficients w_j^* defined in (2.10), $\Upsilon_j = \mathrm{Cov}\left(z_j^*(t)\right) = \mathbb{E}\left(z_j^*(t)\left(\boldsymbol{z}_j - \boldsymbol{w}_{-j}^*\begin{pmatrix}1\\ \boldsymbol{z}_{-j}(t)\end{pmatrix}\right)\right) = \mathbb{E}\left(z_j^*(t) z_j(t)\right)$. Then,

$$\begin{aligned}\left|\frac{1}{T}\sum_{t=1}^{T} z_j^*(t) z_j(t) - \Upsilon_j\right| &= \left|\frac{1}{T}\sum_{t=1}^{T} z_j^*(t) z_j(t) - \mathbb{E}\left(z_j^*(t) z_j(t)\right)\right| \\ &\le \left(\|\boldsymbol{w}_j^*\|_1 + 1\right)\left\|\frac{1}{T}\sum_{t=1}^{T}\begin{pmatrix}1\\ \boldsymbol{z}(t)\end{pmatrix}\begin{pmatrix}1 & \boldsymbol{z}^\top(t)\end{pmatrix} - \Upsilon\right\|_\infty . \end{aligned} \tag{A.21}$$

By Lemma A.18, $\left\|\frac{1}{T}\sum_{t=1}^{T}\begin{pmatrix}1\\ \boldsymbol{z}(t)\end{pmatrix}\begin{pmatrix}1 & \boldsymbol{z}^\top(t)\end{pmatrix} - \Upsilon\right\|_\infty = O_p(\rho T^{-2/5})$. In addition, $\|\boldsymbol{w}_j^*\|_1 = O(\sqrt{s})$ by Lemma A.14. Therefore,

$$\left|\frac{1}{T}\sum_{t=1}^{T}\widehat{z}_j^*(t) z_j(t) - \Upsilon_j\right| = O_p((\sqrt{s}+1)\rho T^{-2/5}),$$

with probability at least $1 - c_2 p^2 T\exp(-c_3 T^{1/5})$. Then, by $\beta_{ij} = T^{-\phi}\Delta$, $s^2\rho^2\log p = o(T^{1/5})$ and $\Lambda_{\min}(\Upsilon_j) > 0$ in Proposition A.2,

$$\frac{1}{T}\sum_{t=1}^{T}\widehat{z}_j^*(t) z_j^\top(t)\beta_{ij} \ge \left(\Upsilon_j - c_4 T^{-2/5}(\sqrt{s}+1)\rho\right)|\beta_{ij}| \ge c_5 T^{-\phi},$$

with probability at least $1 - c_2 p^2 T\exp(-c_3 T^{1/5})$. Taking the bounds above back to (A.20), under $s^2\rho^2\log p = o(T^{1/5})$, we have

$$T\left\|(\Upsilon_j)^{-1/2}\left(\widehat{S}_{ij}^{0} - S_{ij}\right)\right\|_2^2 \ge O_p(T^{1-2\phi}). \tag{A.22}$$

with probability at least $1 - c_6 p^2 T\exp(-c_7 T^{1/5})$.

Next, we bound $\widehat{S}_{ij} - \widehat{S}^0_{ij}$. Without repeating the technical details, the bound of $\widehat{S}_{ij} - \widehat{S}^0_{ij}$ can be derived as follows: under $\phi < 1/2$, we add an addition term of $x_j(t)\beta_{ij}$ to $\epsilon_i(t)$ because $\beta_{ij} \neq 0$ under the alternative; then the bound for A_1 and B_1 in the proof of Lemma A.5 are dominated by $O_p(\rho T^{-\frac{2}{5}-\phi} \vee T^{-4/5})$ instead of $O_p(\rho T^{-\frac{4}{5}})$ under the alternative, which implies $\|\widehat{S}_{ij} - \widehat{S}^0_{ij}\|_2 = O_p\left(\rho T^{-\frac{2}{5}-\phi} \vee T^{-4/5}\right)$. Combining the results above,

$$\begin{aligned} T\left\|(\Upsilon_j)^{-\frac{1}{2}}\left(\widehat{S}_{ij} - S_{ij}\right)\right\|_2^2 &\geq T\left\|(\Upsilon_j)^{-\frac{1}{2}}\left(\widehat{S}_{ij} - \widehat{S}^0_{ij} + \widehat{S}^0_{ij} - S_{ij}\right)\right\|_2^2 \\ &\geq T\left\|(\Upsilon_j)^{-\frac{1}{2}}\left(\widehat{S}^0_{ij} - S_{ij}\right)\right\|_2^2 - T\left\|(\Upsilon_j)^{-\frac{1}{2}}\left(\widehat{S}_{ij} - \widehat{S}^0_{ij}\right)\right\|_2^2 \\ &= O_p(T^{1-2\phi}), \end{aligned} \tag{A.23}$$

with probability at least $1 - c_{10}p^2T\exp(-c_{11}T^{1/5})$.

Note that $\|V_T\|_2^2$ weakly converges to χ^2 by Proposition A.1. In addition, taking $y = c_{12}T^{1/2-\phi} - c_1^{-1}\sqrt{x}$ in Lemma A.9,

$$\mathbb{P}(\|V_T\|_2 \geq c_{12}T^{1/2-\phi} - c_1^{-1}\sqrt{x}) \leq c_{13}T^{-1/8} + c_{14}\exp(-(c_{12}T^{1/2-\phi} - c_1^{-1}\sqrt{x})^2). \tag{A.24}$$

Taking (A.23) and (A.24) back to (A.19), we reach the conclusion:

$$\begin{aligned} &\mathbb{P}\left(\widehat{U}_T \geq x\right) \\ &\quad\geq \mathbb{P}\left(\left\{\|\sqrt{T}\,(\Upsilon_j)^{-1/2}\,(\widehat{S}_{ij} - S_{ij})\|_2 \geq c_{12}T^{1/2-\phi}\right\} \cap \left\{\|V_T\|_2 \leq c_{12}T^{1/2-\phi} - c_1^{-1}\sqrt{x}\right\}\right) \\ &\qquad\geq 1 - c_{10}p^2T\exp(-c_{11}T^{1/5}) - c_{13}T^{-1/8} - c_{14}\exp\left(-\left(c_{12}T^{1/2-\phi} - c_1^{-1}\sqrt{x}\right)^2\right). \end{aligned}$$

Proof of Theorem 2.3

Denote

$$\check{S}_{ij} = \frac{1}{T}\sum_{t=1}^{T}\frac{Y_i(t) - x_j(t)\beta_{i,j} - \widehat{\mu}_i - \boldsymbol{x}_{-j}^\top(t)\widehat{\boldsymbol{\beta}}_{i,-j}}{\widehat{\sigma}_i(t)}\widehat{z}_j^*(t).$$

We have the following decomposition.

$$\widetilde{S}_{ij} = \check{S}_{ij} + \frac{1}{T}\sum_{t=1}^{T}\widehat{z}_j^*(t)\widehat{z}_j(t)(\widehat{\beta}_{ij} - \beta_{ij}) = \check{S}_{ij} + \widetilde{\Upsilon}_j\left(\widehat{\beta}_{ij} - \beta_{ij}\right).$$

By the decomposition above, we have

$$\widehat{b}_{ij} - \beta_{ij} = -\left(\widetilde{\Upsilon}_j\right)^{-1} \check{S}_{ij},$$

and

$$\widehat{R}_T = \left(\check{S}_{ij}\right)^\top \left(\widetilde{\Upsilon}_j\right)^{-1} \widehat{\Upsilon}_j \left(\widetilde{\Upsilon}_j\right)^{-1} \check{S}_{ij}.$$

Notice that

$$\check{S}_{ij} = \frac{1}{T}\sum_{t=1}^{T} \frac{\epsilon_i(t) + (\mu_i - \widehat{\mu}_i) + \boldsymbol{x}_{-j}^\top(t)(\boldsymbol{\beta}_{i,-j} - \widehat{\boldsymbol{\beta}}_{i,-j})}{\widehat{\sigma}_i(t)} \widehat{z}_j^*(t),$$

which is equivalent to $\widehat{S}_{ij}$ defined in (2.21) under the null hypothesis. Let

$$\check{U}_T = \check{S}_{ij}^\top \widehat{\Upsilon}_j^{-1} \check{S}_{ij}.$$

$\check{U}_T$ weakly converges to χ^2 distribution following the proof of Theorem 2.1:

$$\sup_{x\in\mathbb{R}} \left|\mathbb{P}(\check{U}_T \le x) - F_d(x)\right| \le c_1 p^2 T \exp(-c_2 T^{1/5}) + c_3 s^2 \rho^2 T^{-1/5} + c_4 T^{-1/8}. \tag{A.25}$$

Next, we bound

$$\widehat{R}_T - \check{U}_T = \check{S}_{ij}^\top \left(\left(\widetilde{\Upsilon}_j\right)^{-1} \widehat{\Upsilon}_j \left(\widetilde{\Upsilon}_j\right)^{-1} - \left(\widehat{\Upsilon}_j\right)^{-1}\right) \check{S}_{ij}. \tag{A.26}$$

Using Assumptions 2.3-2.6 and the consistency of estimators in Assumption 2.5 and 2.6, it is easy to see that $\check{S}_{ij} = O_p(1)$. Therefore, it is enough to quantify $(\widetilde{\Upsilon}_j)^{-1}\widehat{\Upsilon}_j(\widetilde{\Upsilon}_j)^{-1} - (\widehat{\Upsilon}_j)^{-1}$ in order to quantify $\widehat{R}_T - \check{U}_T$. Let $E = \widetilde{\Upsilon}_j - \widehat{\Upsilon}_j$. Then,

$$\widetilde{\Upsilon}_j \left(\widehat{\Upsilon}_j\right)^{-1} \widetilde{\Upsilon}_j = \left(\widehat{\Upsilon}_j + E\right)\left(\widehat{\Upsilon}_j\right)^{-1}\left(\widehat{\Upsilon}_j + E\right) = \widehat{\Upsilon}_j + E + E + E\left(\widehat{\Upsilon}_j\right)^{-1} E,$$

which leads to

$$\widetilde{\Upsilon}_j \left(\widehat{\Upsilon}_j\right)^{-1} \widetilde{\Upsilon}_j - \widehat{\Upsilon}_j = E + E + E\Upsilon_j^{-1}E + E\left(\left(\widehat{\Upsilon}_j\right)^{-1} - \Upsilon_j^{-1}\right)E. \tag{A.27}$$

Proposition A.2 implies that $E\Upsilon_j^{-1}E \le O(E^2)$. Thus, the first three items are bounded by $O(|E| \vee E^2)$. By Proposition A.2 $\Upsilon_j^{-1} = O(1)$, and Theorem 2.5 in Stewart and Sun [1990] gives us

$$\left\|\left(\widehat{\Upsilon}_j\right)^{-1} - \Upsilon_j^{-1}\right\|_2 \le \frac{\|\Upsilon_j^{-1}\|_2 \|\widehat{\Upsilon}_j - \Upsilon_j\|_2}{1 - \|\Upsilon_j^{-1}\|_2\|\widehat{\Upsilon}_j - \Upsilon_j\|_2} = O(\|\widehat{\Upsilon}_j - \Upsilon_j\|_2).$$

Then, taking this result back to (A.27), we get

$$\widetilde{\Upsilon}_j \left(\widehat{\Upsilon}_j\right)^{-1} \widetilde{\Upsilon}_j - \widehat{\Upsilon}_j = O(|E|) + O(E^2) + O(E^2) O\left(\left\|\widehat{\Upsilon}_j - \Upsilon_j\right\|_2\right).$$

By Lemma A.4, $\|\widehat{\Upsilon}_j - \Upsilon_j\|_2 = O_p\left(s^2\rho^2 T^{-2/5}\right)$. Therefore, to bound $\widetilde{\Upsilon}_j \left(\widehat{\Upsilon}_j\right)^{-1} \widetilde{\Upsilon}_j - \widehat{\Upsilon}_j$, it is sufficient to quantify the bound of E. We first write $E = \widetilde{\Upsilon}_j - \widehat{\Upsilon}_j = \widetilde{\Upsilon}_j - \widetilde{\Upsilon}_j^0 + \widetilde{\Upsilon}_j^0 - \widehat{\Upsilon}_j^0 + \widehat{\Upsilon}_j^0 - \widehat{\Upsilon}_j$ and then quantify the bound for each difference.

We start with the bound of $E^0 \equiv \widetilde{\Upsilon}_j^0 - \widehat{\Upsilon}_j^0$. Notice that

$$\begin{aligned}
\|E^0\|_\infty &\leq \left\| \frac{1}{T} \sum_{t=1}^{T} \left(z_j(t) - \begin{pmatrix} 1 & \boldsymbol{z}_{-j}^\top(t) \end{pmatrix} \widehat{\boldsymbol{w}}_j \right) \begin{pmatrix} 1 & \boldsymbol{z}_{-j}^\top(t) \end{pmatrix} \widehat{\boldsymbol{w}}_j \right\|_\infty \\
&\leq \left\| \frac{1}{T} \sum_{t=1}^{T} \left(z_j(t) - \begin{pmatrix} 1 & \boldsymbol{z}_{-j}^\top(t) \end{pmatrix} \boldsymbol{w}_j^* \right) \boldsymbol{z}_{-j}(t) \right\|_\infty (\|\boldsymbol{w}_j^*\|_1 + \|\boldsymbol{w}_j^* - \widehat{\boldsymbol{w}}_j\|_1) \\
&+ \left| (\widehat{\boldsymbol{w}}_j - \boldsymbol{w}_j^*)^\top \frac{1}{T} \sum_{t=1}^{T} \left(\begin{pmatrix} 1 \\ \boldsymbol{z}_{-j}(t) \end{pmatrix} \begin{pmatrix} 1 & \boldsymbol{z}_{-j}^\top(t) \end{pmatrix} \right) (\widehat{\boldsymbol{w}}_j - \boldsymbol{w}_j^*) \right| \\
&+ \left\| \frac{1}{T} \sum_{t=1}^{T} \begin{pmatrix} 1 \\ \boldsymbol{z}_{-j}(t) \end{pmatrix} \begin{pmatrix} 1 & \boldsymbol{z}_{-j}^\top(t) \end{pmatrix} \boldsymbol{w}_j^* \right\|_\infty \|\widehat{\boldsymbol{w}}_j - \boldsymbol{w}_j^*\|_1.
\end{aligned}$$

Using Lemma A.6 (to bound $\left\| \frac{1}{T} \sum_{t=1}^{T} \left(z_j(t) - \begin{pmatrix} 1 & \boldsymbol{z}_{-j}^\top(t) \end{pmatrix} \boldsymbol{w}_j^* \right) \boldsymbol{z}_{-j}(t) \right\|_\infty$), Lemma A.14 of bounded norm of $\boldsymbol{w}_j^*$, and the estimation consistency of $\widehat{\boldsymbol{w}}_j$ in Assumption 2.6, the first item on the RHS is $O_p(\sqrt{s}\rho T^{-2/5})$. Using Assumption 2.6 again and Lemma A.8, the second item on RHS is $O_p(s\rho T^{-2/5})$. by Lemma A.14 and A.8,

$$\left\| \frac{1}{T} \sum_{t=1}^{T} \begin{pmatrix} 1 \\ \boldsymbol{z}_{-j}(t) \end{pmatrix} \begin{pmatrix} 1 & \boldsymbol{z}_{-j}^\top(t) \end{pmatrix} \boldsymbol{w}_j^* \right\|_\infty \leq \|\boldsymbol{w}_j^*\|_1 \left\| \begin{pmatrix} 1 \\ \boldsymbol{z}_{-j}(t) \end{pmatrix} \begin{pmatrix} 1 & \boldsymbol{z}_{-j}^\top(t) \end{pmatrix} \right\|_\infty = O(\sqrt{s}).$$

Then, under Assumption 2.6, the last term on the RHS is bounded by $O_p(s^{3/2}\rho T^{-2/5})$. Therefore,

$$E^0 = \widetilde{\Upsilon}_j^0 - \widehat{\Upsilon}_j^0 = O_p(s^{3/2}\rho T^{-2/5}),$$

with probability at least $1 - c_5 p^2 T \exp(-c_6 T^{1/5})$. By Lemma A.4, $\widehat{\Upsilon}_j^0 - \widehat{\Upsilon}_j$ is bounded by $O_p(\rho T^{-2/5})$ and $\widetilde{\Upsilon}_j^0 - \widetilde{\Upsilon}_j = O_p(\rho T^{-2/5})$. Combining these results, we get $|E| =$

$O_p(s^{3/2}\rho T^{-2/5})$. Next, turning back to (A.27), $\widetilde{\Upsilon}_j \left(\widehat{\Upsilon}_j\right)^{-1} \widetilde{\Upsilon}_j - \widehat{\Upsilon}_j = O_p(s^{3/2}\rho T^{-2/5})$. Then, referring to (A.26) at beginning,

$$\begin{aligned}\widehat{R}_T - \check{U}_T &= \check{S}_{ij}^\top \left(\left(\widetilde{\Upsilon}_j\right)^{-1} \widehat{\Upsilon}_j \left(\widetilde{\Upsilon}_j\right)^{-1} - \left(\widehat{\Upsilon}_j\right)^{-1} \right) \check{S}_{ij} \\ &= O_p(1) O_p(s^{3/2}\rho T^{-2/5}) O_p(1) = O_p(s^{3/2}\rho T^{-2/5}),\end{aligned}$$

with probability at least $1 - c_6 p^2 T \exp(-c_7 T^{1/5})$. At the end, taking $\delta = s^{3/2}\rho T^{-2/5}$ and using (A.25),

$$\begin{aligned}&\sup_{x\in\mathbb{R}} |\mathbb{P}(\widehat{R}_T \le x) - F_d(x)| \\ &\le \sup_{y\in\mathbb{R}} |\mathbb{P}(\check{U}_T \le y) - F_d(y)| + \sup_{x\in\mathbb{R}} (F_d(x+\delta) - F_d(x-\delta)) + \mathbb{P}(|\widehat{R}_T - \check{U}_T| > \delta) \\ &\le c_8 p^2 T \exp(-c_9 T^{1/5}) + c_{10} s^2 \rho^2 T^{-1/5} + c_{11} T^{-1/8}.\end{aligned}$$

Proof of Theorem 2.4

To simplify the presentation, we prove the theorem for $d = 1$ when testing $H_0 : \beta_{ik} = 0$ for some $k \in \{1, \ldots, m_0 + pm_1\}$. Because of the complexity due to the approximation of the unknown time-varying background intensity and that the transition function is flexible in (2.33), we need to re-examine the bound of $\widehat{S}^0_{ik} - S_{ik}$ to account for the extra approximation error. In Lemma A.10, we show that under Assumption 2.7, $\widehat{S}^0_{ik} - S_{ik}$ has the same error bound as that given by Lemma A.3 that is used to show Theorem 2.1. Then, the rest of the proof follows from an argument similar to that in Theorem 2.1 and is hence omitted.

A.3 Auxiliary Results

The following result shows that $\mathbb{P}\left(U_T \leq y\right) \approx F_d(y)$ uniformly in y. The proof is based on a martingale central limit theorem. Recall that $s_j = \|\boldsymbol{w}_j^*\|_0$ and $s = \max_{1\leq j\leq p} s_j$; $\rho_i = \|\boldsymbol{\beta}_i\|_0$ and $\rho = \max_{1\leq i\leq p} \rho_i$.

Proposition A.1. *Suppose the linear Hawkes model with its intensity function defined in (2.6) is stationary and satisfies Assumptions 2.1 – 2.4. Then* $\forall u \in \mathbb{R}^d$,

$$\sup_{y\in\mathbb{R}} \left|\mathbb{P}(\|V_T + u\|_2^2 \leq y) - F_{d,\|u\|_2^2}(y)\right| \leq C(\|u\|_2, d)T^{-1/8}, \tag{A.28}$$

where $C(\|u\|_2, d)$ *is a constant that is non-decreasing w.r.t.* $\|u\|_2$ *and* d.

Proof. The main part of the proof is based on the result on the martingale difference sequence in Lemma A.2. To reach the conclusion, we verify the conditions required to apply the result in the setting of the multivariate Hawkes process.

Let

$$\xi_{T,t} = -\frac{1}{\sqrt{T}} \left(\Upsilon_j\right)^{-1/2} \frac{\epsilon_i(t)}{\sigma_i(t)} z_j^*(t),$$

where $\sigma_i(t)$, $z_j^*(t)$ are defined in (2.8) and (2.11), respectively.

Recall that $\mathcal{H}_{T,t}$ is information filtration of the past. Then $(\xi_{T,t}, \mathcal{H}_{T,t})$ is a martingale difference sequence. Then, $V_T = \sum_{t=1}^T \xi_{T,t}$, where V_T is defined in (2.14). Following the same notation as Lemma A.2, denote

$$L_\delta^{n,d} = \sum_{t=1}^T \mathbb{E}\|\xi_{Tt}\|_2^{2+2\delta}, \tag{A.29}$$

$$N_\delta^{T,d} = \mathbb{E}\left\|\left(\Upsilon_j\right)^{-1/2} \left(\frac{1}{T}\sum_{t=1}^T z_j^*(t)\left(z_j^*(t)\right)^\top - \Upsilon_j\right)\left(\Upsilon_j\right)^{-1/2}\right\|_{tr}^{1+\delta}. \tag{A.30}$$

To apply Lemma A.2, we need to evaluate the bound of $R_\delta^{n,d} \equiv L_\delta^{n,d} + N_\delta^{T,d}$.

We start with the bound of $L_\delta^{n,d}$. Notice that

$$L_\delta^{n,d} = \sum_{t=1}^{T} \mathbb{E}\|\xi_{Tt}\|_2^{2+2\delta} \leq \Lambda_{\min}^{-1/2}(\Upsilon_j)\, T^{-(1+\delta)} \sum_{i=1}^{T} \mathbb{E}\left\| \frac{1}{\sigma_i(t)} \epsilon_i(t) z_j^*(t) \right\|_2^{2+2\delta}.$$

Proposition A.2 implies $\Lambda_{\min}^{-1/2}(\Upsilon_j) = O(1)$. Lemma A.8 shows that $0 < \sigma_i^2(t)$ and $x_j(t)$, $\epsilon_i(t)$ and $z_j(t)$ are bounded. Therefore,

$$\mathbb{E}\left\| \frac{1}{\sigma_i(t)} \epsilon_i(t) z_j^*(t) \right\|_2^{2+2\delta} \leq O(\mathbb{E}\left\| z_j^*(t) \right\|_2^{2+2\delta}).$$

Recall that $z_j^*(t)$ is $z_j(t)$ after removing its projection onto $\boldsymbol{z}_{-j}$. Then, $\mathbb{E}\left\|z_j^*(t)\right\|_2^2 \leq \mathbb{E}\left\|z_j(t)\right\|_2^2 = O(1)$. In addition, since $x^{1+\delta}$ is convex under $\delta \in [0, 1/2]$ for $x \geq 0$,

$$\mathbb{E}\left(\left\|z_j^*(t)\right\|_2^2\right)^{1+\delta} \leq \left(\mathbb{E}\left\|z_j^*(t)\right\|_2^2\right)^{1+\delta} \leq \left(\mathbb{E}\left\|z_j(t)\right\|_2^2\right)^{1+\delta} = O(1).$$

Thus, $L_\delta^{n,d} = O\left(T^{-\delta}\right)$.

Next, we quantify the bound of $N_\delta^{n,d}$. Notice that

$$\sum_{t=0}^{T-1} \mathbb{E}\left(\xi_{T,t}\xi_{T,t}^\top \mid \mathcal{H}_t\right) - I = (\Upsilon_j)^{-1/2} \left(\frac{1}{T} \sum_{t=1}^{T} z_j^*(t) \left(z_j^*(t)\right)^\top - \Upsilon_j \right) (\Upsilon_j)^{-1/2}.$$

By Proposition A.2, the rank of

$$(\Upsilon_j)^{-1/2} \left(\frac{1}{T} \sum_{t=1}^{T} z_j^*(t) \left(z_j^*(t)\right)^\top - \Upsilon_j \right) (\Upsilon_j)^{-1/2}$$

is at most d (where $d = 1$ in the case of testing an univariate β_{ij}). Since $\|B\|_{tr} \leq d\|B\|_2$ and $\|B_d\|_2 \leq d\|B_d\|_\infty$ for $B \in \mathbb{R}^{d\times d}$, we have

$$\begin{aligned}
N_\delta^{T,d} &= \mathbb{E}\left\| (\Upsilon_j)^{-1/2} \left(\frac{1}{T} \sum_{t=1}^{T} z_j^*(t) \left(z_j^*(t)\right)^\top - \Upsilon_j \right) (\Upsilon_j)^{-1/2} \right\|_{tr}^{1+\delta} \\
&\leq \mathbb{E}\left(d \left\| (\Upsilon_j)^{-1/2} \left(\frac{1}{T} \sum_{t=1}^{T} z_j^*(t) \left(z_j^*(t)\right)^\top - \Upsilon_j \right) (\Upsilon_j)^{-1/2} \right\|_2 \right)^{1+\delta} \\
&\leq \Lambda_{\min}(\Upsilon_j)^{-1}\, \mathbb{E}\left(d^2 \left\| \frac{1}{T} \sum_{t=1}^{T} z_j^*(t) \left(z_j^*(t)\right)^\top - \Upsilon_j \right\|_\infty \right)^{1+\delta},
\end{aligned}$$

where the last step follows from Proposition A.2.

Now,

$$\left\|\frac{1}{T}\sum_{t=1}^{T} z_j^*(t)\left(z_j^*(t)\right)^\top - \Upsilon_j\right\|_\infty \leq$$
$$\left(1+\|\boldsymbol{w}_{j,-j}^*\|_1^2\right)\left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{z}(t)\boldsymbol{z}^\top(t) - \mathbb{E}\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{z}^\top(t)\boldsymbol{z}^\top(t)\right)\right\|_\infty + \|w_{j0}\|_1^2\left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{z}(t) - \mathbb{E}\left(\boldsymbol{z}(t)\right)\right\|_\infty,$$

where bounds on

$$\left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{z}^\top(t)\boldsymbol{z}(t) - \mathbb{E}\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{z}^\top(t)\boldsymbol{z}(t)\right)\right\|_\infty \quad \text{and} \quad \left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{z}(t) - \mathbb{E}\left(\boldsymbol{z}(t)\right)\right\|_\infty$$

are given in Lemma A.18. In addition, the ℓ_1-norm of $\boldsymbol{w}_j^*$ is bounded in Lemma A.14. Therefore, assuming $s^2\rho^2\log p = o(T^{1/5})$,

$$\begin{aligned}
N_\delta^{T,d} &\leq \int_0^\infty \mathbb{P}\left(\left(d^2\left\|\frac{1}{T}\sum_{t=1}^{T} z_j^*(t)\left(z_j^*(t)\right)^\top - \Upsilon_j\right\|_\infty\right)^{1+\delta} > r\right) dr \\
&= \int_0^\infty \mathbb{P}\left(d^2\left\|\frac{1}{T}\sum_{t=1}^{T} z_j^*(t)\left(z_j^*(t)\right)^\top - \Upsilon_j\right\|_\infty > r^{1/(1+\delta)}\right) dr \\
&\leq \int_0^\infty C_1\exp\left(-C_2\min\left\{\left(\frac{T}{s\rho}r^{1/(1+\delta)}\right)^{1/3}, \frac{T}{\rho}r^{1/(1+\delta)}\right\}\right) dr \\
&\leq C(\delta)T^{-1-\delta}(s\rho)^{1+\delta},
\end{aligned}$$

where the last step is based on the integral of the gamma function.

Then,

$$R_\delta^{n,d} = L_\delta^{n,d} + N_\delta^{n,d} \leq C(\delta)T^{-\delta} + C'(\delta)T^{-1-\delta}(s\rho)^{1+\delta},$$

where the first term dominates under $s^2\rho^2\log p = o(T^{1/5})$.

Then taking $\delta = \frac{1}{2}$, Therefore, by Lemma A.2, we have for any $x \geq 0$, $u \in \mathbb{R}^d$, and $\delta \in [0, 1/2]$,

$$|\mathbb{P}(\widehat{U}_T + u \leq x) - F_{d,\|u\|_2^2}(x)| \leq C(\|u\|_2^2, d)T^{-1/8}, \tag{A.31}$$

which completes the proof. □

Proposition A.2. *Suppose the linear Hawkes model with its intensity function defined in (2.6) is stationary and satisfies Assumptions 2.1 – 2.4. Let* $\Upsilon_x = Cov(\boldsymbol{x}(t))$. *Then,*

$$0 < C_1 \leq \Lambda_{\min}(\Upsilon_x) \leq \Lambda_{\max}(\Upsilon_x) \leq C_2 < \infty,$$
$$C_3\Lambda_{\min}(\Upsilon_x) \leq \Lambda_{\min}(\Upsilon_j) \leq \Lambda_{\max}(\Upsilon_j) \leq C_4\Lambda_{\max}(\Upsilon_x),$$

where constants $C_k, k = 1, \ldots, 4$, *only depend on* $(\Theta, \boldsymbol{\mu})$ *and the transition kernel function.*

Proof. Recall that $x_j(t) = \int_0^{t-} \kappa_j(t-s)dN_j(s)$ for $j = 1, \ldots, p$. Let

$$K_t = \begin{pmatrix} \kappa_1(t) & & \\ & \ddots & \\ & & \kappa_p(t) \end{pmatrix}$$

and $\boldsymbol{dN}_s = (dN_1(s), \ldots, dN_p(s))^\top$. Then, $\boldsymbol{x}(t) = (x_1(t), \ldots, x_p(t))^\top = \int_0^{t-} K_{t-s}\boldsymbol{dN}_s$. We consider bounding eigenvalues of

$$\begin{aligned}
\Upsilon_x &= \mathrm{Cov}(\boldsymbol{x}(t)) \\
&= \int_0^{t^-}\int_0^{t^-} K_{t-s}\mathbb{E}\left(\boldsymbol{dN}_s - \Lambda ds\right)\left(\boldsymbol{dN}_r^\top - \Lambda^\top dr\right)K_{t-r} \\
&= \int_0^{t^-}\int_0^{t^-} K_{t-s}\Gamma(s-r)K_{t-r}drds,
\end{aligned}$$

where $\Gamma(l) = \mathbb{E}\left(\boldsymbol{dN}_t - \Lambda dt\right)\left(\boldsymbol{dN}_{t+l}^\top - \Lambda^\top dt\right)/(dt)^2 \in \mathbb{R}^{p\times p}$.

Let $f_\Gamma(\theta) = \int_{-\infty}^{\infty}\Gamma(l)e^{-i\theta l}dl$. Thus, $\Gamma(l) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f_\Gamma(\theta)e^{il\theta}d\theta$. In addition, let $G_{t,s}(\theta) = \int_0^{t-} K(t-s)e^{-is\theta}ds$. Then,

$$\begin{aligned}
\Upsilon_x &= \int_0^{t^-}\int_0^{t^-} K_{t-s}e^{i\theta s}\Gamma(s-r)e^{i\theta(r-s)}K_{t-r}e^{-i\theta r}drds \\
&= \int_0^{t^-}\int_0^{t^-} K_{t-s}e^{i\theta s}\left(\frac{1}{2\pi}\int_{-\pi}^{\pi} f_\Gamma(\theta)e^{i(s-r)}d\theta\right)e^{i\theta(r-s)}K_{t-r}e^{-i\theta r}drds \\
&= \int_0^{t^-}\int_0^{t^-} K_{t-s}e^{i\theta s}\left(\frac{1}{2\pi}\int_{-\pi}^{\pi} f_\Gamma(\theta)d\theta\right)K_{t-r}e^{-i\theta r}drds \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi} G_{t,s}^*(\theta)f_\Gamma(\theta)G_{t,r}(\theta)d\theta.
\end{aligned}$$

Since $f_\Gamma(\theta)$ is Hermitian and $G_t^*(s)f_\Gamma(\theta)G_t(r)$ is real,

$$\mathfrak{m}(f_\Gamma)G_{t,s}^*(\theta)G_{t,r}(\theta) \leq G_{t,s}^*(\theta)f_\Gamma(\theta)G_{t,r}(\theta) \leq \mathfrak{M}(f_\Gamma)G_{t,s}^*(\theta)G_{t,r}(\theta),$$

where

$$\mathfrak{M}(f_\Gamma) = ess \sup_{\theta\in[-\pi,\pi]} \sqrt{\Lambda_{\max}\left(f_\Gamma(\theta)f_\Gamma(\theta)^*\right)},$$
$$\mathfrak{m}(f_\Gamma) = ess \inf_{\theta\in[-\pi,\pi]} \sqrt{\Lambda_{\min}\left(f_\Gamma(\theta)f_\Gamma(\theta)^*\right)}.$$

In addition, notice that

$$\begin{aligned}\frac{1}{2\pi}\int_{-\pi}^{\pi} G_{t,s}^*(\theta)G_{t,r}(\theta)d\theta &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\int_0^{t^-}\int_0^{t^-} K_{t-s}K_{t-r}e^{i\theta(s-r)}dsdrd\theta \\ &= \int_0^{t^-}\int_{-t^-}^{t^-} K_{t-r-l}\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i\theta l}d\theta K_{t-r}dldr \\ &= \int_0^{t^-} K_{t-r}^2 dr.\end{aligned}$$

Letting $Q(t) = \int_0^{t^-} K_{t-r}^2 dr$,

$$\mathfrak{m}(f_\Gamma)\Lambda_{\min}(Q(t)) \leq \Lambda_{\min}\left(\Upsilon_x\right) \leq \Lambda_{\max}\left(\Upsilon_x\right) \leq \mathfrak{M}(f_\Gamma)\Lambda_{\max}(Q(t)),$$

Next, we introduce the result linking $f_\Gamma(\cdot)$ to the transition matrix Ω.

Lemma A.1. *For a stationary multivariate Hawkes process,*

$$f_\Gamma(\theta) = \left(I - f_\omega(\theta)\right)^{-1} diag(\Lambda)\left(I - f_\omega^*(\theta)\right)^{-1},$$

where $\Lambda = \begin{pmatrix}\mathbb{E}\lambda_1(t) & \dots & \mathbb{E}\lambda_p(t)\end{pmatrix}^\top \in \mathbb{R}^p$ *and* $f_\omega(\theta)$ *is the Fourier transformation on the matrix of the transition function* $\omega(t) = (\omega_{ij}(t))_{1\leq i,j\leq p}$.

A similar result as Lemma A.1 has been shown by Bacry et al. [2011, Theorem 1] or Etesami et al. [2016, Theorem 3], but they assume a non-negative transfer function. Lemma A.1 extends the class of linear Hawkes processes by including non-mutually exciting structure. The proof is directly established based on the proof in Bacry et al. [2011, Theorem 1] or

Etesami et al. [2016, Theorem 3] but we re-write a real function into its positive part and its negative part, and then using the distributive property of the convolution operation.

As a result, one set of sufficient conditions to bound the spectral radius of Γ is to assume:

- $0 < \lambda_{\min} \leq \min_{1\leq i\leq p} \lambda_i(t) \leq \max_{1\leq i\leq p} \lambda_i(t) \leq \lambda_{\max} < \infty$ as stated in Assumption 2.3;
- $\Lambda_{\max}(\Omega) < 1$ which leads $\mathfrak{m}(I - f^*_\omega(\theta)) > 0$ assumed in Assumption 2.1;
- Bounded row and column sum of Ω as assume in Assumption 2.2; that is,

$$\mathfrak{M}(I - f^*_\omega(\theta)) \leq 1 + \left(\max_{1\leq i\leq p} \sum_{j=1}^{p} \Omega_{ij} + \max_{1\leq j\leq p} \sum_{i=1}^{p} \Omega_{ij} \right) /2 < \infty.$$

Then, with the assumptions above,

$$\frac{2\min_{1\leq i\leq p}(\mathbb{E}\lambda_i(t))}{(\mathfrak{M}(I - f^*_\omega(\theta)))^2} \leq \mathfrak{m}(f_\Gamma) \leq \mathfrak{M}(f_\Gamma) \leq \frac{2\max_{1\leq i\leq p}(\mathbb{E}\lambda_i(t))}{(\mathfrak{m}(I - f^*_\omega(\theta)))^2}.$$

Therefore, with bounded $Q(t)$ by an integrable and non-trivial transition kernel function $\kappa_j(t)$ in Assumption 2.4, we reach the conclusion. Since both Λ and ω are constants depending only on $\Theta = (\beta_{ij})_{1\leq i,j\leq p}$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^\top$, we have constants C_1, C_2 only depending on the model parameter and the transition kernel function such that

$$0 < C_1(\Theta, \mu) \leq \Lambda_{\min}(\Upsilon_x) \leq \Lambda_{\max}(\Upsilon_x) \leq C_2(\Theta, \mu) < \infty.$$

Since $\lambda_i(t)$ is bounded, there exists c_1, c_2 such that $0 < c_1 \leq \sigma_i^2(t) = \lambda_i(t)(1 - \lambda_i(t)) \leq c_2 \leq \infty$. Let $\Upsilon = \mathrm{Cov}(\boldsymbol{x}(t)/\sigma_i(t))$,

$$c_2^{-1}\Upsilon_x \leq \Upsilon \leq c_1^{-1}\Upsilon_x. \tag{A.32}$$

Notice that

$$\mathrm{Cov}(z_j^*(t))^{-1} = \left(\Upsilon^{-1}\right)_{jj},$$

which means that $\mathrm{Cov}(z_j^*(t))^{-1}$ is a principal submatrix of Υ^{-1}. Then, by the Cauchy's interlace theorem for eigenvalues of Hermitian matrices,

$$\Lambda_{\min}\left(\mathrm{Cov}(z_j^*(t))^{-1}\right) \geq \Lambda_{\min}\left(\Upsilon^{-1}\right),$$
$$\Lambda_{\max}\left(\mathrm{Cov}(z_j^*(t))^{-1}\right) \leq \Lambda_{\max}\left(\Upsilon^{-1}\right).$$

Therefore,

$$c_1\Lambda_{\min}\left(\Upsilon_x^{-1}\right) \leq \Lambda_{\min}\left(\Upsilon_j^{-1}\right) \leq \Lambda_{\max}\left(\Upsilon_j^{-1}\right) \leq c_2\Lambda_{\max}\left(\Upsilon_x^{-1}\right),$$

which completes the proof. □

Lemma A.2 (Lemma C.1 Zheng and Raskutti [2019]). *Let $(\xi_{n,i}, \mathcal{H}_{n,i})_{0\leq i\leq n}$ be a martingale difference sequence taking values in $\mathbb{R}^d$. Let*

$$X_n^k = \sum_{i=1}^{k} \xi_{ni} \quad \text{and} \quad \langle X^n\rangle_k = \sum_{i=1}^{k} a_{ni} \equiv \sum_{i=1}^{k} \mathbb{E}\left(\xi_{ni}\xi_{ni}^\top \mid \mathcal{H}_{n,i-1}\right).$$

Define $R_\delta^{n,d} = L_\delta^{n,d} + N_\delta^{n,d}$, where

$$L_\delta^{n,d} = \sum_{i=1}^{n} \mathbb{E}\|\xi_{ni}\|_2^{2+2\delta} \quad \text{and} \quad N_\delta^{n,d} = \sum_{i=1}^{n} \mathbb{E}\|\langle X^n\rangle_n - I\|_{tr}^{1+\delta}.$$

If $R_\delta^{n,d} \leq 1$, then for any $u \in \mathbb{R}^d$, $r \geq 0$, and $0 < \delta \leq 1/2$, we have

$$P\left(\|X_n^n + u\|_2 \leq r\right) - P\left(\|Z + u\|_2 \leq r\right) \leq C(\|u\|_2, d, \delta)\left(R_\delta^{n,d}\right)^{\frac{1}{3+2\delta}},$$

where $Z_{d\times 1} \sim N(0, I)$ and $C(\|u\|_2, d, \delta)$ is a non-decreasing as $\|u\|_2$ increases.

Lemma A.3. *Suppose the linear Hawkes model with its intensity function defined in (2.6) is stationary and satisfies Assumptions 2.1 – 2.4. In addition, $(\widehat{\mu}_i, \widehat{\boldsymbol{\beta}}_i)$ and $\widehat{w}_j$ satisfy Assumption 2.5 and 2.6, respectively. Given S_{ij} in (2.12) and $\widehat{S}_{ij}^0$ in (A.1),*

$$\mathbb{P}\left(\|\widehat{S}_{ij}^0 - S_{ij}\|_2 \leq C_1 s\rho^2 T^{-4/5}\right) \geq 1 - C_2 p^2 T\exp(-C_3 T^{1/5}),$$

where $C_k, k = 1, \ldots, 3$ are constants only depending on the model parameter $(\boldsymbol{\mu}, \Theta)$ and the transition kernel function.

Proof. We start with the following decomposition

$$\widehat{S}_{ij}^0 - S_{ij} = \underbrace{(\widehat{\boldsymbol{w}}_j - \boldsymbol{w}_j^*)^\top \frac{1}{T}\sum_{t=1}^{T} \widetilde{\epsilon}_i(t) \begin{pmatrix} 1 \\ \boldsymbol{z}_{-j}(t) \end{pmatrix}}_{A}$$
$$+ \underbrace{\frac{1}{T}\sum_{t=1}^{T} z_j^*(t) \begin{pmatrix} 1/\sigma_i(t) & \boldsymbol{z}_{-j}^\top(t) \end{pmatrix} \left(\begin{pmatrix} \widehat{\mu}_i \\ \widehat{\boldsymbol{\beta}}_{i,-j} \end{pmatrix} - \begin{pmatrix} \mu_i \\ \boldsymbol{\beta}_{i,-j} \end{pmatrix} \right)}_{B}$$
$$\underbrace{- (\widehat{\boldsymbol{w}}_j - \boldsymbol{w}_j^*)^\top \left(\frac{1}{T}\sum_{t=1}^{T} \begin{pmatrix} 1 \\ \boldsymbol{z}_{-j}(t) \end{pmatrix} \begin{pmatrix} 1/\sigma_i(t) & \boldsymbol{z}_{-j}^\top(t) \end{pmatrix} \right) \left(\begin{pmatrix} \widehat{\mu}_i \\ \widehat{\boldsymbol{\beta}}_{i,-j} \end{pmatrix} - \begin{pmatrix} \mu_i \\ \boldsymbol{\beta}_{i,-j} \end{pmatrix} \right)}_{C}. \quad \text{(A.33)}$$

In the follows, we bound A, B, C. Lemma A.7 and Assumption 2.6 give

$$A \le \left\|\widehat{\boldsymbol{w}}_j - \boldsymbol{w}_j^*\right\|_1 \left\| \frac{1}{T}\sum_{t=1}^{T} \widetilde{\epsilon}_i(t) \begin{pmatrix} 1 \\ \boldsymbol{z}_{-j}(t) \end{pmatrix} \right\|_\infty = O_p(s\rho T^{-4/5}).$$

Lemma A.6 and Assumption 2.5 give

$$B \le \left\| \frac{1}{T}\sum_{t=1}^{T} z_j^*(t) \begin{pmatrix} 1/\sigma_i(t) & \boldsymbol{z}_{-j}^\top(t) \end{pmatrix} \right\|_\infty \left\| \begin{pmatrix} \widehat{\mu}_i \\ \widehat{\boldsymbol{\beta}}_{i,-j} \end{pmatrix} - \begin{pmatrix} \mu_i \\ \boldsymbol{\beta}_{i,-j} \end{pmatrix} \right\|_1 = O_p(s^{1/2}\rho^2 T^{-4/5}).$$

Combining Assumption 2.5 and 2.6 and Lemma A.8 gives

$$C \le \left\|\widehat{\boldsymbol{w}}_j - \boldsymbol{w}_j^*\right\|_1 \left\| \frac{1}{T}\sum_{t=1}^{T} \begin{pmatrix} 1 \\ \boldsymbol{z}_{-j}(t) \end{pmatrix} \begin{pmatrix} 1/\sigma_i(t) & \boldsymbol{z}_{-j}^\top(t) \end{pmatrix} \right\|_\infty \left\| \begin{pmatrix} \widehat{\mu}_i \\ \widehat{\boldsymbol{\beta}}_{i,-j} \end{pmatrix} - \begin{pmatrix} \mu_i \\ \boldsymbol{\beta}_{i,-j} \end{pmatrix} \right\|_1$$
$$= O_p(s\rho^2 T^{-4/5}).$$

Therefore,

$$\left|\widehat{S}_{ij}^0 - S_{ij}\right| = O_p\left(s\rho^2 T^{-4/5}\right), \quad \text{(A.34)}$$

which probability at least $1 - c_1 p^2 T \exp(-c_2 T^{1/5})$, where c_1, c_2 are constants only depending on the model parameter $(\boldsymbol{\mu}, \Theta)$ and the transition kernel function.

□

Lemma A.4. *Suppose the linear Hawkes model with its intensity function defined in (2.6) is stationary and satisfies Assumptions 2.1 – 2.4. In addition, $(\widehat{\mu}_i, \widehat{\boldsymbol{\beta}}_i)$ and $\widehat{w}_j$ satisfy Assumption 2.5 and 2.6, respectively. Then,*

$$\begin{aligned}
\left\|\widehat{\Upsilon}_j - \widehat{\Upsilon}_j^0\right\|_\infty &= O_p(\rho T^{-2/5}), \\
\left\|\widetilde{\Upsilon}_j - \widetilde{\Upsilon}_j^0\right\|_\infty &= O_p(\rho T^{-2/5}), \\
\left\|\widehat{\Upsilon}_j^0 - \Upsilon_j\right\|_\infty &= O_p(s^2\rho^2 T^{-2/5}),
\end{aligned}$$

and

$$\begin{aligned}
\left\|\Upsilon_j^{1/2}\widehat{\Upsilon}_j^{-1}\Upsilon_j^{1/2} - I\right\|_\infty &= O_p(s^2\rho^2 T^{-2/5}), \\
\left\|\Upsilon_j^{1/2}\left(\widehat{\Upsilon}_j^0\right)^{-1}\Upsilon_j^{1/2} - I\right\|_\infty &= O_p(s^2\rho^2 T^{-2/5}), \\
\left\|\left(\widehat{\Upsilon}_j^0\right)^{1/2}\left(\widehat{\Upsilon}_j\right)^{-1}\left(\widehat{\Upsilon}_j^0\right)^{1/2} - I\right\|_\infty &= O_p(\rho T^{-2/5}),
\end{aligned}$$

with probability at least $1 - C_1 p^2 T \exp(-C_2 T^{1/5})$, where C_1, C_2 are constants only depending on the model parameter $(\boldsymbol{\mu}, \Theta)$ and the transition kernel function, and $\widehat{\Upsilon}_j^0$ and $\widetilde{\Upsilon}_j$ are defined in (A.2), and (2.29), separately.

Proof. First, applying Talyor expansion on $\Upsilon_j^{1/2}\widehat{\Upsilon}_j^{-1}\Upsilon_j^{1/2} - I$ (treating $\widehat{\Upsilon}_j^{-1}$ as variable) at Υ_j leads to

$$\Upsilon_j^{1/2}\widehat{\Upsilon}_j^{-1}\Upsilon_j^{1/2} - I = \Upsilon_j^{1/2}\Upsilon_j^{-1}\Upsilon_j^{1/2} - I + \Upsilon_j^{-1}(\widehat{\Upsilon}_j - \Upsilon_j) + o(\widehat{\Upsilon}_j - \Upsilon_j). \tag{A.35}$$

Then, by Proposition A.2 that $\Lambda_{\min}(\Upsilon_j) > 0$,

$$\left\|\Upsilon_j^{1/2}\widehat{\Upsilon}_j^{-1}\Upsilon_j^{1/2} - I\right\|_\infty = O\left(\left\|\Upsilon_j^{-1}(\widehat{\Upsilon}_j - \Upsilon_j\right\|_\infty\right) = O\left(\left\|\widehat{\Upsilon}_j - \Upsilon_j\right\|_\infty\right). \tag{A.36}$$

In the following, we focus on quantifying the bound of $\|\widehat{\Upsilon}_j - \Upsilon_j\|_\infty$.

Recall

$$\widehat{\Upsilon}_j^0 = \frac{1}{T}\sum_{t=1}^{T}\left(z_j(t) - \begin{pmatrix}1 & \boldsymbol{z}_{-j}^\top(t)\end{pmatrix}\widehat{\boldsymbol{w}}_j\right)^2$$

defined in (A.2). The difference between $\widehat{\Upsilon}_j^0$ and $\widehat{\Upsilon}_j$ is that we replace $\widehat{\sigma}_i^2(t)$ by the true value $\sigma_i^2(t)$. We bound the two parts $\widehat{\Upsilon}_j^0 - \Upsilon_j$ and $\widehat{\Upsilon}_j - \widehat{\Upsilon}_j^0$ separately.

First, we have

$$\begin{aligned}
\widehat{\Upsilon}_j^0 - \Upsilon_j =& \left(\frac{1}{T}\sum_{t=1}^{T}\left(z_j^*(t)\right)^2 - \Upsilon_j\right) \\
&+(\widehat{\boldsymbol{w}}_j - \boldsymbol{w}_j^*)^\top \left(\frac{2}{T}\sum_{t=1}^{T}\begin{pmatrix}1\\ \boldsymbol{z}_{-j}(t)\end{pmatrix} z_j^*(t)\right) \\
&+(\widehat{\boldsymbol{w}}_j - \boldsymbol{w}_j^*)^\top \left(\frac{1}{T}\sum_{t=1}^{T}\begin{pmatrix}1\\ \boldsymbol{z}_{-j}(t)\end{pmatrix}\begin{pmatrix}1 & \boldsymbol{z}_{-j}^\top(t)\end{pmatrix}\right)(\widehat{\boldsymbol{w}}_j - \boldsymbol{w}_j^*) \\
\equiv& E_1 + 2E_2 + E_3.
\end{aligned}$$

By Lemma A.8 and A.14, Assumption 2.6,

$$\begin{aligned}
\|E_2\|_1 &\le \left\|\widehat{\boldsymbol{w}}_j - \boldsymbol{w}_j^*\right\|_1 \left\|\frac{2}{T}\sum_{t=1}^{T}\begin{pmatrix}1\\ \boldsymbol{z}_{-j}(t)\end{pmatrix} z_j^*(t)\right\|_\infty \\
&= \left\|\widehat{\boldsymbol{w}}_j - \boldsymbol{w}_j^*\right\|_1 \left\|\frac{2}{T}\sum_{t=1}^{T}\begin{pmatrix}1\\ \boldsymbol{z}_{-j}(t)\end{pmatrix}\begin{pmatrix}1 & \boldsymbol{z}(t)\end{pmatrix}\right\|_\infty \left(\|\boldsymbol{w}_j^*\|_1 + 1\right) \\
&= O_p(s^{3/2}\rho T^{-2/5}).
\end{aligned}$$

Using Lemma A.8 and Assumption 2.6,

$$\|E_3\|_1 \le \left\|\widehat{\boldsymbol{w}}_j - \boldsymbol{w}_j^*\right\|_1 \left\|\frac{1}{T}\sum_{t=1}^{T}\begin{pmatrix}1\\ \boldsymbol{z}_{-j}(t)\end{pmatrix}\begin{pmatrix}1 & \boldsymbol{z}_{-j}^\top(t)\end{pmatrix}\right\|_\infty \left\|\widehat{\boldsymbol{w}}_j - \boldsymbol{w}_j^*\right\|_1 = O_p(s^2\rho^2 T^{-4/5}).$$

Finally, we bound E_1. Recall that $\boldsymbol{w}_j$ is defined such that $z_j^*(t) = \begin{pmatrix}1 & \boldsymbol{z}^\top(t)\end{pmatrix}\boldsymbol{w}_j$, where $\boldsymbol{w}_j = (w_{j0}^*, \{w_{jl}^*\mathbf{1}(l \ne j) + \mathbf{1}(l = j)\}_{1\le l\le p})^\top \in \mathbb{R}^{p+1}$ and $\|\boldsymbol{w}_j\|_1 = \|\boldsymbol{w}_j^*\|_1 + 1$. Then,

$$\left(z_j^*(t)\right)^2 = \boldsymbol{w}_j^\top \begin{pmatrix}1 & \boldsymbol{z}^\top(t)\end{pmatrix}\begin{pmatrix}1\\ \boldsymbol{z}(t)\end{pmatrix}\boldsymbol{w}_j.$$

Using Lemma A.14 that $\|\boldsymbol{w}_j^*\|_1^2 = O(s)$ and Lemma A.18,

$$|E_1| \le \|\boldsymbol{w}_j\|_1^2 \left\| \frac{1}{T}\sum_{t=1}^{T} \begin{pmatrix} 1 \\ \boldsymbol{z}(t) \end{pmatrix} \begin{pmatrix} 1 & \boldsymbol{z}^\top(t) \end{pmatrix} - \mathbb{E} \begin{pmatrix} 1 \\ \boldsymbol{z}(t) \end{pmatrix} \begin{pmatrix} 1 & \boldsymbol{z}^\top(t) \end{pmatrix} \right\|_\infty$$
$$= O_p\left(s\rho T^{-2/5}\right).$$

Therefore, combining the bounds of E_1, E_2, E_3,

$$|\widehat{\Upsilon}_j^0 - \Upsilon_j| = O_p\left(s^2\rho^2 T^{-2/5}\right), \tag{A.37}$$

with probability at least $1 - C_1 p^2 T \exp(-C_2 T^{1/5})$.

Next, we bound $\widehat{\Upsilon}_j^0 - \widehat{\Upsilon}_j$. Letting $D_i = \frac{\widehat{\sigma}_i(t)}{\sigma_i(t)}$, we have

$$\widehat{\Upsilon}_j^0 - \widehat{\Upsilon}_j = \frac{1}{T}\sum_{t=1}^{T}(D_i^2 - 1)(\widehat{z}_j(t) - \widehat{\boldsymbol{z}}_{-j}^\top(t)\widehat{\boldsymbol{w}}_{j,-j})^2 - 2(D_i - 1)\widehat{w}_{j0}\left(\widehat{z}_j(t) - \widehat{\boldsymbol{z}}_{-j}^\top(t)\widehat{\boldsymbol{w}}_{j,-j}\right).$$

Recall that $\sigma_i^2(t) = \lambda_i(t)(1-\lambda_i(t))$ and $\lambda_i(t) = \mu_i + \boldsymbol{x}^\top(t)\boldsymbol{\beta}_i$. By Lemma A.8, $\sigma_i^2(t) = O(1)$. By Lemma A.8 that $\boldsymbol{x}(t) = O(1)$, and $\left\| \begin{pmatrix} \mu_i \\ \boldsymbol{\beta}_i \end{pmatrix} - \begin{pmatrix} \widehat{\mu}_i \\ \widehat{\boldsymbol{\beta}}_i \end{pmatrix} \right\|_1 = O_p\left(\rho T^{-2/5}\right)$ in Assumption 2.5,

$$|\lambda_i(t) - \widehat{\lambda}_i(t)| = \left| \begin{pmatrix} 1 & \boldsymbol{x}(t) \end{pmatrix} \left(\begin{pmatrix} \mu_i \\ \boldsymbol{\beta}_i \end{pmatrix} - \begin{pmatrix} \widehat{\mu}_i \\ \widehat{\boldsymbol{\beta}}_i \end{pmatrix} \right) \right|$$
$$\le \left\| \begin{pmatrix} 1 & \boldsymbol{x}(t) \end{pmatrix} \right\|_\infty \left\| \begin{pmatrix} \mu_i \\ \boldsymbol{\beta}_i \end{pmatrix} - \begin{pmatrix} \widehat{\mu}_i \\ \widehat{\boldsymbol{\beta}}_i \end{pmatrix} \right\|_1 = O_p\left(\rho T^{-2/5}\right).$$

Thus,

$$
\begin{aligned}
|D_i^2 - 1| &= \left| \frac{\widehat{\sigma}_i^2(t) - \sigma_i^2(t)}{\sigma_i^2(t)} \right| \\
&= \left| \frac{\widehat{\lambda}_i(t) - \widehat{\lambda}_i^2(t) - \lambda_i(t) + \lambda_i^2(t)}{\sigma_i^2(t)} \right| \\
&= \left| \frac{\left(\lambda_i(t) - \widehat{\lambda}_i(t)\right)\left(1 - 2\lambda_i(t) + \lambda_i(t) - \widehat{\lambda}_i(t)\right)}{\sigma_i^2(t)} \right| \\
&\leq \frac{1}{\sigma_i^2(t)} \left|\lambda_i(t) - \widehat{\lambda}_i(t)\right| \left(|1 - 2\lambda_i(t)| + \left|\lambda_i(t) - \widehat{\lambda}_i(t)\right| \right) \\
&= O\left(\lambda_i(t) - \widehat{\lambda}_i(t)\right) = O_p\left(\rho T^{-2/5}\right).
\end{aligned}
$$

Using a similar deduction, we get $|D_i - 1| = O_p\left(\rho T^{-2/5}\right)$. In addition, the estimation error bounds of $\widehat{\boldsymbol{w}}_j$ and $(\widehat{\mu}_i, \widehat{\boldsymbol{\beta}}_i)$ in Assumptions 2.3 and 2.4 imply that $\widehat{z}_j(t) - \widehat{\boldsymbol{z}}_{-j}^\top(t)\widehat{\boldsymbol{w}}_{j,-j}$ and $\widehat{w}_{j0}\left(\widehat{z}_j(t) - \widehat{\boldsymbol{z}}_{-j}^\top(t)\widehat{\boldsymbol{w}}_{j,-j}\right)$ are bounded in probability. Combining these results,

$$
\left| \widehat{\Upsilon}_j^0 - \widehat{\Upsilon}_j \right| = O_p\left(\rho T^{-2/5}\right), \tag{A.38}
$$

with probability at least $1 - C_3 p^2 T \exp(-C_4 T^{1/5})$. Therefore, using (A.36),

$$
\begin{aligned}
\|\Upsilon_j^{1/2} \widehat{\Upsilon}_j^{-1} \Upsilon_j^{1/2} - I\|_\infty &= O\left(\|\widehat{\Upsilon}_j - \Upsilon_j\|_\infty\right) \\
&= O\left(\|\widehat{\Upsilon}_j - \widehat{\Upsilon}_j^0\|_\infty\right) + O\left(\|\widehat{\Upsilon}_j^0 - \Upsilon_j\|_\infty\right) \\
&= O_p\left(s^2 \rho^2 T^{-2/5}\right),
\end{aligned}
$$

with probability at least $1 - C_5 p^2 T \exp(-C_6 T^{1/5})$

Following a similar deduction and separately using the bounds of $\|\widehat{\Upsilon}_j - \widehat{\Upsilon}_j^0\|_\infty$ and $\|\widehat{\Upsilon}_j^0 - \Upsilon_j\|_\infty$ derived in the above, we get the bounds for $\left\| \left(\widehat{\Upsilon}_j^0\right)^{1/2} \left(\widehat{\Upsilon}_j\right)^{-1} \left(\widehat{\Upsilon}_j^0\right)^{1/2} - I \right\|_\infty = O_p(\rho T^{-2/5})$ and $\left\| \Upsilon_j^{1/2} \left(\widehat{\Upsilon}_j^0\right)^{-1} \Upsilon_j^{1/2} - I \right\|_\infty = O_p\left(s^2 \rho^2 T^{-2/5}\right)$, respectively.

In addition, following a similar deduction as (A.38), we can show the same probabilistic bound holds for $\left|\widetilde{\Upsilon}_j^0 - \widetilde{\Upsilon}_j\right|$ (except for a change in constants), where $\widetilde{\Upsilon}_j$ is defined in (2.29) and $\widetilde{\Upsilon}_j^0$ is the same as $\widetilde{\Upsilon}_j$ except it involves the true $\sigma_i(t)$ in its construction. □

Lemma A.5. *Suppose the stationary linear Hawkes model defined in (2.2) satisfies Assumption 2.1-2.4. In addition, $(\widehat{\mu}_i, \widehat{\boldsymbol{\beta}}_i)$ and $\widehat{w}_j$ satisfy Assumption 2.5 and 2.6, respectively. Under the null hypothesis and the alternative with $\phi \geq 1/2$,*

$$\mathbb{P}\left(|\widehat{U}_T - \widehat{U}_T^0| > C_1 \rho T^{-1/5}\right) \leq C_2 p^2 T \exp(-C_3 T^{1/5}) + C_4 T^{-1/8} + C_5 s^2 \rho^2 T^{-1/5}, \tag{A.39}$$

where $C_k, k = 1, \ldots, 5$ are constants only depending on the model parameter $(\boldsymbol{\mu}, \Theta)$ and the transition kernel function. $\widehat{U}_T$ and $\widehat{U}_T^0$ are defined in (2.23) and (A.4), respectively.

Proof. We first give a proof under the null hypothesis when $\beta_{ij} = 0$. We extend the proof for the alternative hypothesis setting at the end.

Similar to the proof of Theorem 2.1, we quantify $\widehat{U}_T^0 - U_T$ (see equation (A.8)). To study the difference between $\widehat{U}_T^0$ and $\widehat{U}_T$ as shown in (A.7), it is sufficient to bound $\widehat{S}_{ij} - \widehat{S}_{ij}^0$ and $\widehat{\Upsilon}_j - \widehat{\Upsilon}_j^0$, where $\widehat{S}_{ij}^0$, $\widehat{\Upsilon}_j^0$ are defined in (A.1) and (A.4).

The bound for $\widehat{\Upsilon}_j - \widehat{\Upsilon}_j^0$ is given by Lemma A.4. We focus on quantifying the difference

between $\widehat{S}_{ij}$ and $\widehat{S}^0_{ij}$. We have

$$
\begin{aligned}
&\widehat{S}_{ij} - \widehat{S}^0_{ij} \\
&= \frac{1}{T}\sum_{t=1}^{T} \frac{1}{\widehat{\sigma}_i(t)} \left(Y_i(t) - \widehat{\mu}_i - \boldsymbol{x}_{-j}^\top(t)\widehat{\boldsymbol{\beta}}_{i,-j}\right) \left(x_j(t)/\widehat{\sigma}_i(t) - \widehat{w}_{j0} - \boldsymbol{x}_{-j}^\top(t)/\widehat{\sigma}_i(t)\widehat{\boldsymbol{w}}_{j,-j}\right) \\
&- \frac{1}{T}\sum_{t=1}^{T} \frac{1}{\sigma_i(t)} \left(Y_i(t) - \widehat{\mu}_i - \boldsymbol{x}_{-j}^\top(t)\widehat{\boldsymbol{\beta}}_{i,-j}\right) \left(x_j(t)/\sigma_i(t) - \widehat{w}_{j0} - \boldsymbol{x}_{-j}^\top(t)/\sigma_i(t)\widehat{\boldsymbol{w}}_{j,-j}\right) \\
&= \frac{1}{T}\sum_{t=1}^{T} \frac{\sigma_i^2(t)}{\widehat{\sigma}_i^2(t)}\frac{1}{\sigma_i(t)} \left(Y_i(t) - \widehat{\mu}_i - \boldsymbol{x}_{-j}^\top(t)\widehat{\boldsymbol{\beta}}_{i,-j}\right) \left(x_j(t)/\sigma_i(t) - \widehat{w}_{j0}\widehat{\sigma}_i(t)/\sigma_i(t) - \boldsymbol{x}_{-j}^\top(t)/\sigma_i(t)\widehat{\boldsymbol{w}}_{j,-j}\right) \\
&- \frac{1}{T}\sum_{t=1}^{T} \frac{1}{\sigma_i(t)} \left(Y_i(t) - \widehat{\mu}_i - \boldsymbol{x}_{-j}^\top(t)\widehat{\boldsymbol{\beta}}_{i,-j}\right) \left(x_j(t)/\sigma_i(t) - \widehat{w}_{j0} - \boldsymbol{x}_{-j}^\top(t)/\sigma_i(t)\widehat{\boldsymbol{w}}_{j,-j}\right) \\
&= \frac{1}{T}\sum_{t=1}^{T} \left(\frac{\sigma_i^2(t)}{\widehat{\sigma}_i^2(t)} - 1\right)\frac{1}{\sigma_i(t)} \left(Y_i(t) - \widehat{\mu}_i - \boldsymbol{x}_{-j}^\top(t)\widehat{\boldsymbol{\beta}}_{i,-j}\right) \left(x_j(t)/\sigma_i(t) - \widehat{w}_{j0} - \boldsymbol{x}_{-j}^\top(t)/\sigma_i(t)\widehat{\boldsymbol{w}}_{j,-j}\right) \\
&+ \frac{1}{T}\sum_{t=1}^{T} \frac{\sigma_i^2(t)}{\widehat{\sigma}_i^2(t)}\frac{1}{\sigma_i(t)} \left(Y_i(t) - \widehat{\mu}_i - \boldsymbol{x}_{-j}^\top(t)\widehat{\boldsymbol{\beta}}_{i,-j}\right) \widehat{w}_{j0} \left(1 - \frac{\widehat{\sigma}_i(t)}{\sigma_i(t)}\right) \\
&= \frac{1}{T}\sum_{t=1}^{T} \left(\frac{\sigma_i^2(t)}{\widehat{\sigma}_i^2(t)} - 1\right)\frac{1}{\sigma_i(t)} \left(Y_i(t) - \widehat{\mu}_i - \boldsymbol{x}_{-j}^\top(t)\widehat{\boldsymbol{\beta}}_{i,-j}\right) \left(x_j(t)/\sigma_i(t) - \widehat{w}_{j0} - \boldsymbol{x}_{-j}^\top(t)/\sigma_i(t)\widehat{\boldsymbol{w}}_{j,-j}\right) \\
&+ \frac{1}{T}\sum_{t=1}^{T} \frac{\sigma_i^2(t)}{\widehat{\sigma}_i^2(t)}\frac{1}{\sigma_i(t)} \left(Y_i(t) - \widehat{\mu}_i - \boldsymbol{x}_{-j}^\top(t)\widehat{\boldsymbol{\beta}}_{i,-j}\right) \widehat{w}_{j0}\frac{\sigma_i^2(t) - \widehat{\sigma}_i^2(t)}{\sigma_i(t)\left(\sigma_i(t) + \widehat{\sigma}_i(t)\right)} \\
&\equiv A + B.
\end{aligned}
$$

For ease of notation, let

$$
\boldsymbol{\eta} = \begin{pmatrix} \mu_i \\ \boldsymbol{\beta}_i \end{pmatrix}, \quad \widehat{\boldsymbol{\eta}} = \begin{pmatrix} \widehat{\mu}_i \\ \widehat{\boldsymbol{\beta}}_i \end{pmatrix}, \quad \text{and} \quad \boldsymbol{x}_t = \begin{pmatrix} 1 & \boldsymbol{x}^\top(t) \end{pmatrix}^\top.
$$

Then,

$$
\sigma_i^2(t) - \widehat{\sigma}_i^2(t) = \boldsymbol{x}_t^\top\boldsymbol{\eta}(1 - \boldsymbol{x}_t^\top\boldsymbol{\eta}) - \boldsymbol{x}_t^\top\widehat{\boldsymbol{\eta}}(1 - \boldsymbol{x}_t^\top\widehat{\boldsymbol{\eta}}) = (\boldsymbol{\eta} - \widehat{\boldsymbol{\eta}})^\top\boldsymbol{x}_t\left(1 - \boldsymbol{x}_t^\top(\widehat{\boldsymbol{\eta}} + \boldsymbol{\eta})\right). \quad \text{(A.40)}
$$

Since $\|\boldsymbol{x}_t\|_\infty$ is bounded by Assumption 2.3, we have

$$
\|\sigma_i^2(t) - \widehat{\sigma}_i^2(t)\|_\infty \leq \|\boldsymbol{\eta} - \widehat{\boldsymbol{\eta}}\|_1\|\boldsymbol{x}_t\|_\infty\|1 - \boldsymbol{x}_t^\top(\widehat{\boldsymbol{\eta}} + \boldsymbol{\eta}))\|_\infty = O\left(\|\boldsymbol{\eta} - \widehat{\boldsymbol{\eta}}\|_1\right) = O_p(\rho T^{-2/5}),
$$

where the bound of $\|\boldsymbol{\eta} - \widehat{\boldsymbol{\eta}}\|_1$ is given in Assumption 2.5.

Let

$$\boldsymbol{\eta}_{-j} = \begin{pmatrix} \mu_i \\ \boldsymbol{\beta}_{i,-j} \end{pmatrix} \quad \text{and} \quad \boldsymbol{x}_t^{-j} = \begin{pmatrix} 1 \\ \boldsymbol{x}_{-j}(t) \end{pmatrix}.$$

Under the null hypothesis that $\beta_{ij} = 0$,

$$\begin{aligned} Y_i(t) - \widehat{\mu}_i - \boldsymbol{x}_{-j}^\top(t)\widehat{\boldsymbol{\eta}}_{i,-j} &= \epsilon_i(t) + (\mu_i - \widehat{\mu}_i) + \boldsymbol{x}_{-j}^\top(t)(\boldsymbol{\eta}_{i,-j} - \widehat{\boldsymbol{\eta}}_{i,-j}) \\ &= \epsilon_i(t) + \left(\boldsymbol{x}_t^{-j}\right)^\top (\boldsymbol{\eta}_{-j} - \widehat{\boldsymbol{\eta}}_{-j}). \end{aligned}$$

Let $C_i(t) = \left(1 - \boldsymbol{x}_t^\top(\widehat{\boldsymbol{\eta}} + \boldsymbol{\eta})\right) \frac{1}{\widehat{\sigma}_i^2(t)\sigma_i(t)} \left(x_j(t)/\sigma_i(t) - \widehat{w}_{j0} - \boldsymbol{x}_{-j}^\top(t)/\sigma_i(t)\widehat{\boldsymbol{w}}_{j,-j}\right)$. Then, we write part A as follows:

$$\begin{aligned} A &= \frac{1}{T}\sum_{t=1}^{T} \left(\sigma_i^2(t) - \widehat{\sigma}_i^2(t)\right) \epsilon_i(t) \frac{1}{\widehat{\sigma}_i^2(t)\sigma_i(t)} \left(x_j(t)/\sigma_i(t) - \widehat{w}_{j0} - \boldsymbol{x}_{-j}^\top(t)/\sigma_i(t)\widehat{\boldsymbol{w}}_{j,-j}\right) \\ &+ \frac{1}{T}\sum_{t=1}^{T} \left(\sigma_i^2(t) - \widehat{\sigma}_i^2(t)\right) \boldsymbol{x}_t^{-j}(\boldsymbol{\eta}_{-j} - \widehat{\boldsymbol{\eta}}_{-j}) \frac{1}{\widehat{\sigma}_i^2(t)\sigma_i(t)} \left(x_j(t)/\sigma_i(t) - \widehat{w}_{j0} - \boldsymbol{x}_{-j}^\top(t)/\sigma_i(t)\widehat{\boldsymbol{w}}_{j,-j}\right) \\ &= (\boldsymbol{\eta} - \widehat{\boldsymbol{\eta}})^\top \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t^\top \epsilon_i(t) C_i(t) + (\boldsymbol{\eta} - \widehat{\boldsymbol{\eta}})^\top \left(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t^\top C_i(t) \boldsymbol{x}_t^{-j}\right) (\boldsymbol{\eta}_{-j} - \widehat{\boldsymbol{\eta}}_{-j}) \\ &\equiv A_1 + A_2. \end{aligned}$$

By the estimation consistency of $\widehat{w}_j$ on w_j^* and $(\widehat{\mu}_i, \widehat{\boldsymbol{\beta}}_i)$ on $(\mu_i, \boldsymbol{\beta}_i)$ in Assumptions 2.5 and 2.6 as well as the bounded $\sigma^2(t)$ and $x_j(t)$ implied by Assumptions 2.3 and 2.4, $C_i(t)$ is bounded in probability. Thus, applying Lemma A.7 and under Assumption 2.5,

$$|A_1| \leq \|\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}\|_1 \left\| \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t^\top \epsilon_i(t) C_i(t) \right\|_\infty \leq O_p\left(\rho T^{-4/5}\right),$$

with probability at least $1 - C_1 p \exp(-C_2 T^{1/5})$. Similarly,

$$|A_2| \leq |C_i(t)| \|x_j(t)\|_\infty^2 \|\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}\|_2^2 \leq O_p\left(\rho T^{-4/5}\right).$$

Next, let $C_i'(t) = \left(1 - \boldsymbol{x}_t^\top(\widehat{\boldsymbol{\eta}} + \boldsymbol{\eta})\right) \frac{\widehat{w}_{j0}}{\widehat{\sigma}_i^2(t)(\sigma_i(t)+\widehat{\sigma}_i(t))}$. Then,

$$\begin{aligned} B &= \frac{1}{T}\sum_{t=1}^{T}\left(\epsilon_i(t) + \boldsymbol{x}_t^\top(\boldsymbol{\eta} - \widehat{\boldsymbol{\eta}})\right)\widehat{w}_{j0}\frac{\sigma_i^2(t) - \widehat{\sigma}_i^2(t)}{\widehat{\sigma}_i^2(t)\left(\sigma_i(t) + \widehat{\sigma}_i(t)\right)} \\ &= (\boldsymbol{\eta} - \widehat{\boldsymbol{\eta}})^\top \frac{1}{T}\sum_{t=1}^{T}\epsilon_i(t)\boldsymbol{x}_t C'(t) \\ &+ (\boldsymbol{\eta} - \widehat{\boldsymbol{\eta}})^\top \left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t C'(t)(\boldsymbol{x}_t^{-j})^\top\right)(\boldsymbol{\eta}_{-j} - \widehat{\boldsymbol{\eta}}_{-j}) \\ &\equiv B_1 + B_2. \end{aligned}$$

By the estimation consistency of $\widehat{w}_j$ on w_j^* and $(\widehat{\mu}_i, \widehat{\boldsymbol{\beta}}_i)$ on $(\mu_i, \boldsymbol{\beta}_i)$ in Assumptions 2.5 and 2.6 as well as Lemma A.8 that $\sigma^2(t)$ and $x_j(t)$ are bounded, we get $C_i'(t) = O_p(1)$. In addition, applying Lemma A.16,

$$|B_1| \le \|\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}\|_1 \left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t^\top \epsilon_i(t) C_i'(t)\right\|_\infty = O_p\left(\rho T^{-4/5}\right),$$

with probability at least $1 - C_3 p\exp(-C_4 T^{1/5})$. Similarly,

$$|B_2| \le |C_i'(t)|\|\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}\|_1 \|\boldsymbol{x}_t\boldsymbol{x}_t^\top\|_\infty \|\widehat{\boldsymbol{\eta}}_{-j} - \boldsymbol{\eta}_{-j}\|_1 \le O_p\left(\rho^2 T^{-4/5}\right).$$

Then, $|B| = O_p\left(\rho T^{-4/5}\right)$.

Therefore, taking the bound of A and B back to (A.3),

$$\left|\widehat{S}_{ij} - \widehat{S}_{ij}^0\right| = O_p\left(\rho T^{-4/5}\right), \tag{A.41}$$

with probability at least $1 - C_5\exp(-C_6 T^{1/5})$.

We follow a similar deduction as (A.8) and let $E = \sqrt{T}\left(\widehat{\Upsilon}_j^0\right)^{-1/2}(\widehat{S}_{ij} - \widehat{S}_{ij}^0)$. Then,

$$|\widehat{U}_T^0 - \widehat{U}_T| \le \|E\|_2^2 + 2\|\widehat{V}_T\|_2\|E\|_2 + \left\|\left(\widehat{\Upsilon}_j^0\right)^{1/2}\left(\widehat{\Upsilon}_j\right)^{-1}\left(\widehat{\Upsilon}_j^0\right)^{1/2} - I\right\|_\infty \left(\|\widehat{V}_T\|_2 + \|E\|_2\right)^2. \tag{A.42}$$

Using the consistency of $\widehat{\Upsilon}_j$ to Υ_j in Lemma A.4 and (A.41), $\|E\|_2^2 = O_p\left(\rho^2 T^{-3/5}\right)$. Lemma A.4 gives $\left\|\left(\widehat{\Upsilon}_j^0\right)^{1/2}\left(\widehat{\Upsilon}_j\right)^{-1}\left(\widehat{\Upsilon}_j^0\right)^{1/2} - I\right\|_\infty = O_p(\rho T^{-2/5})$. Using an intermediate

result shown in Theorem 2.1 that $\|V_T^0\|_2^2$ weakly converges to χ_d^2 in (A.10) and then applying Lemma A.9, we get

$$\mathbb{P}(\|\widehat{V}_T^0\|_2 > T^{1/10}) \leq C_7\rho^2 T\exp(-C_8T^{1/5}) + C_9s^2\rho^2T^{-1/5} + C_{10}T^{-1/8}.$$

Therefore,

$$\mathbb{P}\left(|\widehat{U}_T - \widehat{U}_T^0| > C_{11}\rho T^{-1/5}\right) \leq C_{12}p^2T\exp(-C_{13}T^{1/5}) + C_{14}s^2\rho^2T^{-1/5} + C_{15}T^{-1/8}, \quad \text{(A.43)}$$

which establishes the result under the null hypothesis.

Next, we examine the difference between $\widehat{U}_T$ and $\widehat{U}_T^0$ under the alternative hypothesis setting, ($\beta_{ij} = \Delta T^{-\phi}$). Here we modify A_1 and B_1, where we replace $\epsilon_i(t)$ by $\epsilon_i(t) + x_j(t)\beta_{ij}$ when bounding $\|\widehat{S}_{ij} - \widehat{S}_{ij}^0\|_2$.

Notice that for $\phi \geq \frac{1}{2}$, the order of the upper bounds of A_1, B_1 remain unchanged. Using the estimation error of $(\mu_i, \boldsymbol{\beta}_i)$ given in Assumption 2.5,

$$\begin{aligned}|A_1| &\leq \|\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}\|_1 \left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t^\top\epsilon_i(t)C_i(t)\right\|_\infty + \|\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}\|_1\|x_j(t)\|_\infty^2|\beta_{ij}| \\ &= O_p\left(\rho T^{-4/5}\right) + O_p\left(T^{-2/5-\phi}\Delta\right) = O_p\left(\rho T^{-4/5}\right),\end{aligned}$$

with probability at least $1 - C_1p^2T\exp(-C_2T^{1/5})$. Similarly, we show part B_1 is bounded in probability with the same order of bound as before. Therefore, under the alternative hypothesis with $\phi \geq \frac{1}{2}$, $\left|\widehat{U}_T - \widehat{U}_T^0\right|$ is bounded by the same order of upper bound with a similar probability (except for a change in constants) as that under the null.

□

Lemma A.6. *Suppose the linear Hawkes model with its intensity function defined in (2.6) is stationary and satisfies Assumptions 2.1 – 2.4. Then,*

$$P\left(\left\|\frac{1}{T}\sum_{t=1}^{T}z_j^*(t)\begin{pmatrix}1/\sigma_i(t) & \boldsymbol{z}_{-j}^\top(t)\end{pmatrix}\right\|_\infty > C_1\left(\sqrt{s}+1\right)\rho T^{-2/5}\right) \leq C_2p^2T\exp(-C_3T^{1/5}),$$

where C_1, C_2, C_3 are constants only depending on the model parameter $(\boldsymbol{\mu}, \Theta)$ and the transition kernel function.

Proof. We separately bound

$$\frac{1}{T}\sum_{t=1}^{T} z_j^*(t)z_k(t), \text{ for } k \neq j, \quad \text{and} \quad \frac{1}{T}\sum_{t=1}^{T}\frac{1}{\sigma_i(t)}z_j^*(t).$$

Notice that $x_k(t) = (\boldsymbol{x}(t))^\top e_k$, where e_k is the kth canonical basis vector. Let $\boldsymbol{w}_j$ be such that $z_j^*(t) = \begin{pmatrix}1 & \boldsymbol{z}^\top(t)\end{pmatrix}\boldsymbol{w}_j$, where $\boldsymbol{w}_j = \left(w_{j0}^*, \{w_{jl}^*\mathbf{1}(l \neq j) + \mathbf{1}(l = j)\}_{1\leq l\leq p}\right)^\top \in \mathbb{R}^{p+1}$ and $\|\boldsymbol{w}_j\|_1 = \|\boldsymbol{w}_j^*\|_1 + 1$. Then,

$$\frac{1}{T}\sum_{t=1}^{T} z_j^*(t)z_k(t) = \boldsymbol{w}_j^\top\left(\frac{1}{T}\sum_{t=1}^{T}\begin{pmatrix}1\\ \boldsymbol{z}(t)\end{pmatrix}\boldsymbol{z}^\top(t)\right)e_k.$$

Note that $\mathbb{E}\left(\frac{1}{T}\sum_{t=1}^{T} z_j^*(t)z_k(t)\right) = 0, k \neq j$, by the construction of $z_j^*(t)$ in (2.11). Then,

$$\begin{aligned}
\left\|\frac{1}{T}\sum_{t=1}^{T} z_j^*(t)z_k(t)\right\|_\infty &= \left\|\frac{1}{T}\sum_{t=1}^{T} z_j^*(t)z_k(t) - \mathbb{E}\left(\frac{1}{T}\sum_{t=1}^{T} z_j^*(t)z_k(t)\right)\right\|_\infty \\
&\leq \|\boldsymbol{w}_j\|_1\left\|\frac{1}{T}\sum_{t=1}^{T}\begin{pmatrix}1\\ \boldsymbol{z}(t)\end{pmatrix}\boldsymbol{z}^\top(t) - \mathbb{E}\left(\begin{pmatrix}1\\ \boldsymbol{z}(t)\end{pmatrix}\boldsymbol{z}^\top(t)\right)\right\|_\infty \|e_k\|_1 \\
&\leq \left(\|\boldsymbol{w}_j^*\|_1 + 1\right)\left\|\frac{1}{T}\sum_{t=1}^{T}\begin{pmatrix}1\\ \boldsymbol{z}(t)\end{pmatrix}\boldsymbol{z}^\top(t) - \mathbb{E}\left(\begin{pmatrix}1\\ \boldsymbol{z}(t)\end{pmatrix}\boldsymbol{z}^\top(t)\right)\right\|_\infty \|e_k\|_1 \\
&\leq O_p\left((\sqrt{s_j}+1)\rho T^{-2/5}\right),
\end{aligned}$$

with probability at least $1 - C_1p^2T\exp(-C_2T^{1/5})$, where the last inequality is by Lemma A.14 and Lemma A.18 (taking $\epsilon = T^{-2/5}$).

For the second part, by the construction of $z_j^*(t)$ that $\mathbb{E}z_j^*(t) = 0$ in (2.11),

$$\begin{aligned}
\frac{1}{\sigma_i(t)}z_j^*(t) &= \frac{1}{\sigma_i(t)}\left(z_j^*(t) - \mathbb{E}z_j^*(t)\right) \\
&= \frac{1}{\sigma_i(t)}\left(\boldsymbol{w}_j^*\right)^\top\left(\boldsymbol{z}(t) - \mathbb{E}\boldsymbol{z}(t)\right).
\end{aligned}$$

Then,

$$\left\|\frac{1}{T}\sum_{t=1}^{T}\frac{1}{\sigma_i(t)}z_j^*(t)\right\|_\infty \leq \|\boldsymbol{w}_j^*\|_1\left\|\frac{1}{T}\sum_{t=1}^{T} 1/\sigma_i(t)\left(\boldsymbol{z}(t) - \mathbb{E}\boldsymbol{z}(t)\right)\right\|_\infty.$$

To derive the bound for $\left\| \frac{1}{T} \sum_{t=1}^{T} 1/\sigma_i(t) \left(\boldsymbol{z}(t) - \mathbb{E}\boldsymbol{z}(t) \right) \right\|_\infty$, we repeat the steps of (A.48) and (A.49) in Lemma A.17, but with $f_1(s) = \frac{1}{\sigma_i^2(t)} \int_s^T k_j(t-s)dt$. With $\sigma_i(t) = O(1)$ by Lemma A.8,

$$\left\| \frac{1}{T} \sum_{0}^{T} \boldsymbol{z}(t)/\sigma_i(t) - \mathbb{E}\boldsymbol{z}(t)/\sigma_i(t) \right\|_\infty = O_p(T^{-2/5}).$$

Combining the above,

$$\left\| \frac{1}{T} \sum_{t=1}^{T} \frac{1}{\sigma_i(t)} z_j^*(t) \right\|_\infty = O_p\left(\sqrt{s_j} \rho T^{-2/5} \right),$$

with probability at least $1 - C_3 p^2 T \exp(-C_4 T^{1/5})$.

□

Lemma A.7. *Suppose the linear Hawkes model with its intensity function defined in (2.6) is stationary and satisfies Assumptions 2.1 – 2.4. Then*

$$P\left(\left\| \frac{1}{T} \sum_{t=1}^{T} \widetilde{\epsilon}_i(t) \begin{pmatrix} 1 \\ \boldsymbol{z}_{-j}(t) \end{pmatrix} \right\|_\infty > C_1 T^{-2/5} \right) \leq C_2 p \exp(-T^{1/5}), \tag{A.44}$$

where C_1, C_2 are constants only depending on the model parameter $(\boldsymbol{\mu}, \Theta)$ and the transition kernel function.

Proof. The concentration inequality for the linear Hawkes process has been discussed Chen et al. [2019]. The result is a direct application of the martingale inequality by van de Geer [1995, Theorem 3.1] stated in Lemma A.16.

Using Lemma A.16, we reach the conclusion by taking $\epsilon = T^{-4/5}$, and separately set $H(t) = \frac{1}{\sigma_i(t)} z_k(t)$, $H(t) = \frac{1}{\sigma_i(t)}$, for $k \neq j$, where $H(t)$ is bounded by Assumption 2.3 and 2.4 of a bounded intensity function and an integrable transition kernel.

□

Lemma A.8. *Under Assumptions 2.3 and 2.4, for* $1 \le i, j \le p$ *and* $\forall t \in [0, T]$,

$$x_j(t) = O(1), \quad 0 < \sigma_i^2(t) = O(1), \quad z_j(t) = O(1), \quad \epsilon_i(t) = O(1),$$
$$\left\| \frac{1}{T} \sum_{t=1}^{T} \begin{pmatrix} 1 \\ \boldsymbol{x}(t) \end{pmatrix} \begin{pmatrix} 1 & \boldsymbol{x}^\top(t) \end{pmatrix} \right\|_\infty = O(1), \quad \left\| \frac{1}{T} \sum_{t=1}^{T} \begin{pmatrix} 1 \\ \boldsymbol{z}(t) \end{pmatrix} \begin{pmatrix} 1 & \boldsymbol{z}^\top(t) \end{pmatrix} \right\|_\infty = O(1),$$

where $x_j(t), \sigma_i^2(t)$, $z_j(t)$ *and* $\epsilon_i(t)$ *are defined in Section 2.2, and* $\boldsymbol{x}(t) = (x_1(t), \ldots, x_p(t))^\top$ *and* $\boldsymbol{z}(t) = (z_1(t), \ldots, z_p(t))^\top$.

Proof. The result follows directly from Assumptions 2.3 and 2.4. □

Lemma A.9. *Let* $\|V_T\|_2^2 \sim \chi_d^2$, *then for a constant* Δ,

$$P\left(\|V_T + \Delta\|_2 > y\right) \le y^{-2}. \tag{A.45}$$

Proof. The conclusion is by the tail bound on χ^2 distribution given in Zheng and Raskutti [2019](see Theorem 3.2). □

Lemma A.10. *Suppose the flexible linear Hawkes model with its intensity function defined in* (2.33) *satisfies Assumptions 2.1 – 2.3, & 2.7. In addition,* $\widehat{\boldsymbol{\beta}}_i$ *and* $\widehat{w}_j$ *satisfy Assumption 2.5 and 2.6, respectively. Given* S_{ik} *in* (2.12) *and* $\widehat{S}_{ik}^0$ *in* (A.1),

$$\mathbb{P}\left(\left\| \widehat{S}_{ik}^0 - S_{ik} \right\|_2 \le C_1 s \rho^2 T^{-4/5} \right) \ge 1 - C_2 p^2 T \exp(-C_3 T^{1/5}),$$

where C_1, C_2, C_3 *are constants only depending on* $\boldsymbol{\beta}_i$ *and the basis function.*

Proof. Let

$$\widetilde{S}_{ik} = \frac{1}{T} \sum_{t=1}^{T} \frac{1}{\sigma_i(t)} \left(Y_i(t) - \boldsymbol{x}^\top(t) \boldsymbol{\beta}_i \right) z_k^*(t). \tag{A.46}$$

When the intensity function can be perfectly described by the basis (i.e. $\lambda_i(t) = \boldsymbol{x}^\top(t)\boldsymbol{\beta}_i$), $\widetilde{S}_{ik} = S_{ik}$ and our proof then follows that in Lemma A.3. Otherwise, we split the difference $\widehat{S}_{ik}^0 - S_{ik}$ into $\widetilde{S}_{ik} - S_{ik}$ and $\widehat{S}_{ik}^0 - \widetilde{S}_{ik}$. Note that the bound for the latter term can be quantified following the same argument as that in Lemma A.3. Now we check the first term.

With $Y_i(t) = \lambda_i(t) + \epsilon_i(t)$,

$$\begin{aligned}\widetilde{S}_{ik} - S_{ik} &= \frac{1}{T}\sum_{t=1}^{T}\frac{1}{\sigma_i(t)}\left(Y_i(t) - \boldsymbol{x}^\top(t)\boldsymbol{\beta}_i\right)z_k^*(t) - \frac{1}{T}\sum_{t=1}^{T}\widetilde{\epsilon}_i(t)z_k^*(t) \\ &= \frac{1}{T}\sum_{t=1}^{T}\frac{1}{\sigma_i(t)}\left(Y_i(t) - \boldsymbol{x}^\top(t)\boldsymbol{\beta}_i - \epsilon_i\right)z_k^*(t) \\ &= \frac{1}{T}\sum_{t=1}^{T}\left(\lambda_i(t) - \boldsymbol{x}^\top(t)\boldsymbol{\beta}_i\right)\frac{1}{\sigma_i^2(t)}x_k^*(t).\end{aligned}$$

By Cauchy–Schwarz inequality,

$$\left\|\widetilde{S}_{ik} - S_{ik}\right\|_2^2 \le \frac{1}{T}\sum_{t=1}^{T}\left(\lambda_i(t) - \boldsymbol{x}^\top(t)\boldsymbol{\beta}_i\right)^2\frac{1}{T}\sum_{t=1}^{T}\frac{1}{\sigma_i^4(t)}\left(x_k^*(t)\right)^2.$$

By Lemma A.8, $\sigma_i^2(t) = O(1)$. Then, by the construction of $x_k^*(t)$,

$$\frac{1}{T}\sum_{t=1}^{T}\left(x_k^*(t)\right)^2 \le \frac{1}{T}\sum_{t=1}^{T}\left(x_k(t)\right)^2 = O(1),$$

where the last equality follows from the fact that $x_k(t) = \int_0^{t-}\phi_k(t)dN_j(t)$ and $\phi_k(t)$ is an integrable basis function.

Then, under Assumption 2.7,

$$\left\|\widetilde{S}_{ik} - S_{ik}\right\|_2 = O\left(s\rho^2 T^{-4/5}\right).$$

The result follows from combining the bounds of $\widetilde{S}_{ik} - S_{ik}$ and $\widehat{S}_{ik}^0 - \widetilde{S}_{ik}$. □

Lemma A.11. *Let* $H = \frac{1}{T}\sum_{t=1}^{T}\begin{pmatrix}1\\ \boldsymbol{x}(t)\end{pmatrix}\begin{pmatrix}1 & \boldsymbol{x}^\top(t)\end{pmatrix}$. *Suppose the linear Hawkes model with its intensity function defined in (2.6) is stationary and satisfies Assumptions 2.1 – 2.4. In addition,* $(\widehat{\mu}_i, \widehat{\boldsymbol{\beta}}_i)$ *are given in (2.16). Then, taking* $\lambda = O\left(T^{-2/5}\right)$ *and assuming*

$\sqrt{\rho} \vee \log p = o\left(T^{1/5}\right)$, $\forall i = 1, \ldots, p$,

$$\left\| \begin{pmatrix} \widehat{\mu}_i \\ \widehat{\boldsymbol{\beta}}_i \end{pmatrix} - \begin{pmatrix} \mu_i \\ \boldsymbol{\beta}_i \end{pmatrix} \right\|_2 \leq C_1 \sqrt{\rho + 1} T^{-2/5}$$

$$\left(\begin{pmatrix} \widehat{\mu}_i \\ \widehat{\boldsymbol{\beta}}_i \end{pmatrix} - \begin{pmatrix} \mu_i \\ \boldsymbol{\beta}_i \end{pmatrix} \right)^\top H \left(\begin{pmatrix} \widehat{\mu}_i \\ \widehat{\boldsymbol{\beta}}_i \end{pmatrix} - \begin{pmatrix} \mu_i \\ \boldsymbol{\beta}_i \end{pmatrix} \right) \leq C_1 \sqrt{\rho + 1} T^{-2/5}$$

$$\left\| \begin{pmatrix} \widehat{\mu}_i \\ \widehat{\boldsymbol{\beta}}_i \end{pmatrix} - \begin{pmatrix} \mu_i \\ \boldsymbol{\beta}_i \end{pmatrix} \right\|_1 \leq C_1 (\rho + 1) T^{-2/5},$$

with probability at least $1 - C_2 p^2 T \exp(-C_3 T^{1/5})$, *where* C_1, C_2, C_3 *are constants only depending on the model parameter* $(\boldsymbol{\mu}, \Theta)$ *and the transition kernel function, and* $\rho_i = \|\boldsymbol{\beta}_i\|_0$ *and* $\rho = \max_{1 \leq i \leq p} \rho_i$.

Proof. The proof follows a typical framework for the analysis of lasso-type estimators [Bühlmann and van de Geer, 2011, Negahban and Wainwright, 2010]. The crucial difficulty in the proof is showing that the key conditions are satisfied for the linear Hawkes process. To be specific, we start with the basic inequality by the construction of the lasso estimator in (2.16). We then bound the prediction error of the lasso regression using the results of Lemma A.7. In addition, we show that the restricted eigen-value (RE) condition is satisfied with high probability in the setting with the multivariate Hawkes process, where the proof essentially uses Proposition A.2 and Lemma A.17. In what follows $C_k, k = 1, \ldots, 5$ are constants only depending on the model parameter $(\boldsymbol{\mu}, \Theta)$ and the transition kernel function.

By the construction of the lasso estimator in (2.16), we have

$$\frac{1}{T} \sum_{t=1}^{T} \left(Y_i(t) - \widehat{\mu}_i - \boldsymbol{x}(t) \widehat{\boldsymbol{\beta}}_i \right)^2 + \lambda \|\widehat{\boldsymbol{\beta}}_i\|_1 \leq \frac{1}{T} \sum_{t=1}^{T} \left(Y_i(t) - \mu_i - \boldsymbol{x}^\top(t) \boldsymbol{\beta}_i \right)^2 + \lambda \|\boldsymbol{\beta}_i\|_1. \quad \text{(A.47)}$$

Let $u = \widehat{\mu}_i - \mu_i$ and $\boldsymbol{v} = \widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i$. Define $S = \{j : \beta_{ij} \neq 0\}$ and $S^c = \{j : \beta_{ij} = 0\}$. It follows from (A.47) that

$$\begin{pmatrix} u & \boldsymbol{v}^\top \end{pmatrix} H \begin{pmatrix} u \\ \boldsymbol{v} \end{pmatrix} \leq 2 \left\| \frac{1}{T} \sum_{t=1}^{T} \epsilon_i(t) \begin{pmatrix} 1 \\ \boldsymbol{x}(t) \end{pmatrix} \right\|_\infty \left\| \begin{pmatrix} u \\ \boldsymbol{v} \end{pmatrix} \right\|_1 + \lambda \|\boldsymbol{v}_S\|_1 - \lambda \|\boldsymbol{v}_{S^c}\|_1.$$

Taking $\lambda = 4 \left\| \frac{1}{T} \sum_{t=1}^{T} \epsilon_i(t) \begin{pmatrix} 1 \\ \boldsymbol{x}(t) \end{pmatrix} \right\|_\infty$, we have

$$0 \leq \begin{pmatrix} u & \boldsymbol{v}^\top \end{pmatrix} H \begin{pmatrix} u \\ \boldsymbol{v} \end{pmatrix} \leq \frac{3\lambda}{2}\|\boldsymbol{v}_S\|_1 - \frac{\lambda}{2}\|\boldsymbol{v}_{S^c}\|_1 + \frac{1}{2}\lambda\|u\|_1 \leq \frac{3\lambda}{2}\|\begin{pmatrix} u & \boldsymbol{v}_S \end{pmatrix}\|_1 - \frac{\lambda}{2}\|\boldsymbol{v}_{S^c}\|_1.$$

Let $\boldsymbol{\eta} = \begin{pmatrix} u \\ \boldsymbol{v} \end{pmatrix} \in \mathbb{R}^{p+1}$. Define $\boldsymbol{\eta}_s = (u, \boldsymbol{v}_S)$ and $\boldsymbol{\eta}_{s^c} = (\boldsymbol{v}_{S^c})$. Then,

$$0 \leq \boldsymbol{\eta}^T H \boldsymbol{\eta} \leq \frac{3\lambda}{2}\|\boldsymbol{\eta}_s\|_1 - \frac{\lambda}{2}\|\boldsymbol{\eta}_{s^c}\|_1 \quad \text{and} \quad \|\boldsymbol{\eta}_{s^c}\|_1 \leq 3\|\boldsymbol{\eta}_s\|_1.$$

Let $\mathcal{C}(J, \kappa) = \{\boldsymbol{\eta} : \|\boldsymbol{\eta}_{J^c}\|_1 \leq \kappa\|\boldsymbol{\eta}_J\|_1\}$. Thus, $\boldsymbol{\eta} = \mathcal{C}(S, 3)$. Further, denote $\check{\Gamma} = \mathbb{E}H$. Then,

$$\begin{aligned}
&\inf\{\boldsymbol{\eta}^\top H \boldsymbol{\eta} : \boldsymbol{\eta} \in \mathcal{C}(J, \kappa), \|\boldsymbol{\eta}\|_2 \leq 1\} \\
&\geq \inf\{\boldsymbol{\eta}^\top \check{\Gamma} \boldsymbol{\eta} : \boldsymbol{\eta} \in \mathcal{C}(J, \kappa), \|\boldsymbol{\eta}\|_2 \leq 1\} - \sup\left\{\left|\boldsymbol{\eta}^\top \left(H - \check{\Gamma}\right) \boldsymbol{\eta}\right| : \boldsymbol{\eta} \in \mathcal{C}(J, \kappa), \|\boldsymbol{\eta}\|_2 \leq 1\right\} \\
&\geq \inf\{\boldsymbol{\eta}^\top \check{\Gamma} \boldsymbol{\eta} : \|\boldsymbol{\eta}\|_2 \leq 1\} - \sup_{\boldsymbol{\eta} \in \mathcal{C}(J, \kappa), \|\boldsymbol{\eta}\|_2 \leq 1} \left\{\left|\boldsymbol{\eta}^\top \left(H - \check{\Gamma}\right) \boldsymbol{\eta}\right|\right\} \\
&\geq \Lambda_{\min}\left(\check{\Gamma}\right) - \sup_{\boldsymbol{\eta} \in \mathcal{C}(J, \kappa), \|\boldsymbol{\eta}\|_2 \leq 1} \left\{\left|\boldsymbol{\eta}^\top \left(H - \check{\Gamma}\right) \boldsymbol{\eta}\right|\right\}.
\end{aligned}$$

Proposition A.2 shows that $\Lambda_{\min}\left(\mathrm{Cov}\left(\boldsymbol{x}(t)\right)\right) > 0$, which implies that the p-unit multivariate process $\{x_j(t)\}_{1 \leq j \leq p}$ are not linearly dependent. In addition, by Assumption 2.4, the process $x_j(t)$ is not a trivial process of constants. We conclude the minimum eigenvalue of $\check{\Gamma}$ is strictly positive. Otherwise, there exists $\eta \in \mathbb{R}^{p+1}$ and $\|\eta\|_1 > 0$, s.t., $\eta^\top \begin{pmatrix} \mathbf{1} \\ \boldsymbol{x}(t) \end{pmatrix} = 0$. This implies either $\{x_j(t)\}_{j=1}^p$ are linearly dependent or $x_j(t)$ is a trivial process of constants, which leads to a contradiction.

Using Lemma A.17, with $\epsilon = -2/5$,

$$\begin{aligned}
&\left\{\left|\boldsymbol{\eta}^\top \left(H - \check{\Gamma}\right) \boldsymbol{\eta}\right| : \|\boldsymbol{\eta}\|_2 \leq 1, \boldsymbol{\eta} \in \mathcal{C}(J, \kappa)\right\} \\
&\leq \|\boldsymbol{\eta}\|_1^2 \left\|H - \check{\Gamma}\right\|_{\max} \\
&\leq 4\|\boldsymbol{\eta}_s\|_1^2 \left\|H - \check{\Gamma}\right\|_{\max} \\
&\leq 4(\rho_i + 1)\|\boldsymbol{\eta}_s\|_2^2 \left\|H - \check{\Gamma}\right\|_{\max} = O_p\left((\rho_i + 1)T^{-2/5}\right),
\end{aligned}$$

with probability at least $1 - C_1 p^2 T \exp(-C_2 T^{1/5})$. Thus, assuming $\sqrt{\rho_i} \vee \log p = o(T^{1/5})$, there exists a constant C_3 such that when $T > C_3$,

$$\sup_{\|\boldsymbol{\eta}\|_2 \leq 1} \left\{ \left| \boldsymbol{\eta}^\top \left(H - \check{\Gamma} \right) \boldsymbol{\eta} \right| \right\} \geq \frac{1}{2} \Lambda_{\min} \left(\check{\Gamma} \right).$$

Thus,

$$\inf\{\boldsymbol{\eta}^\top H \boldsymbol{\eta} : \boldsymbol{\eta} \in \mathcal{C}(J, \kappa), \|\boldsymbol{\eta}\|_2 \leq 1\} \geq \Lambda_{\min} \left(\check{\Gamma} \right) - \frac{1}{2} \Lambda_{\min} \left(\check{\Gamma} \right) \geq \frac{1}{2} \Lambda_{\min} \left(\check{\Gamma} \right),$$

with probability at least $1 - C_1 p^2 T \exp(-C_2 T^{1/5})$. Then, with the same probability,

$$\begin{aligned} \frac{1}{2} \Lambda_{\min} \left(\check{\Gamma} \right) \left\| \begin{pmatrix} u & \boldsymbol{v}^\top \end{pmatrix} \right\|_2^2 &\leq \begin{pmatrix} u & \boldsymbol{v}^\top \end{pmatrix} H \begin{pmatrix} u \\ \boldsymbol{v} \end{pmatrix} \\ &\leq \frac{3\lambda}{2} \left\| \begin{pmatrix} u & \boldsymbol{v}_S^\top \end{pmatrix} \right\|_1 \leq \frac{3}{2} \lambda \sqrt{\rho_i + 1} \left\| \begin{pmatrix} u & \boldsymbol{v}^\top \end{pmatrix} \right\|_2. \end{aligned}$$

At last, using Lemma A.7 to bound $\lambda = 4 \left\| \frac{1}{T} \sum_{t=1}^T \epsilon_i(t) \begin{pmatrix} 1 \\ \boldsymbol{x}(t) \end{pmatrix} \right\|_\infty$, we have

$$\begin{aligned} \|\boldsymbol{\eta}\|_2 &\leq C_4 \sqrt{\rho_i + 1} T^{-2/5}, \\ \boldsymbol{\eta}^\top H \boldsymbol{\eta} &\leq C_4 (\rho_i + 1) T^{-4/5}, \\ \|\boldsymbol{\eta}\|_1 \leq 4 \|\boldsymbol{\eta}_s\|_1 &\leq 4 \sqrt{\rho_i + 1} \|\boldsymbol{\eta}_s\|_2 \leq C_4 (\rho_i + 1) T^{-2/5}, \end{aligned}$$

with probability at least $1 - C_5 p^2 T \exp(-C_6 T^{1/5})$.

□

Lemma A.12. *Let* $\widehat{H} = \frac{1}{T} \sum_{t=1}^T \begin{pmatrix} 1 \\ \widehat{\boldsymbol{z}}(t) \end{pmatrix} \begin{pmatrix} 1 & \widehat{\boldsymbol{z}}^\top(t) \end{pmatrix}$. *Suppose the linear Hawkes model with its intensity function defined in (2.6) is stationary and satisfies Assumptions 2.1 – 2.4. In addition,* $\widehat{\boldsymbol{w}}_j$ *is given in (2.18). Then, taking* $\lambda = O\left(\sqrt{s+1} \rho T^{-2/5} \right)$ *and assuming* $\sqrt{(s+1)\rho} \vee \log p = o(T^{1/5})$,

$$\begin{aligned} \|\widehat{\boldsymbol{w}}_j - \boldsymbol{w}_j\|_2 &\leq C_1 \sqrt{s+1} \rho T^{-2/5} \\ (\widehat{\boldsymbol{w}}_j - \boldsymbol{w}_j)^\top \widehat{H} (\widehat{\boldsymbol{w}}_j - \boldsymbol{w}_j) &\leq C_1 \sqrt{s+1} \rho T^{-2/5} \\ \|\widehat{\boldsymbol{w}}_j - \boldsymbol{w}_j\|_1 &\leq C_1 (s+1) \rho T^{-2/5}, \end{aligned}$$

with probability at least $1-C_2p^2T\exp(-C_3T^{1/5})$*, where* C_1, C_2, C_3 *are constants only depending on the model parameter* $(\boldsymbol{\mu}, \Theta)$ *and the transition kernel function, and* $s_j = \|\boldsymbol{w}_j^*\|_0$ *and* $s = \max_{1\le j\le p} s_j$*;* $\rho_i = \|\boldsymbol{\beta}_i\|_0$ *and* $\rho = \max_{1\le i\le p}\rho_i$*.*

Proof. The proof is similar to the steps in Lemma A.11 except that we use Lemma A.6 to bound $\left\|\frac{1}{T}\sum_{t=1}^{T}\widehat{z}_j^*(t)\begin{pmatrix}1 & \widehat{\boldsymbol{z}}_{-j}^\top(t)\end{pmatrix}\right\|_\infty$, and use Lemma A.18 to verify the RE condition for $\widehat{H} = \frac{1}{T}\sum_{t=1}^{T}\begin{pmatrix}1\\ \widehat{\boldsymbol{z}}(t)\end{pmatrix}\begin{pmatrix}1 & \widehat{\boldsymbol{z}}^\top(t)\end{pmatrix}$.

□

Lemma A.13. $s_j = \|\boldsymbol{w}_j^*\|_0$ *and* $s = \max_{1\le j\le p} s_j$*;* $\rho_i = \|\boldsymbol{\beta}_i\|_0$ *and* $\rho = \max_{1\le i\le p}\rho_i$*. Suppose the linear Hawkes model with its intensity function defined in* (2.6) *is stationary and satisfies Assumptions 2.1 – 2.4. For the connectivity matrix of block structure,* $s \le 2\rho + 1$*.*

Proof. By the choice of $\boldsymbol{w}_j^*$ in (2.10),

$$\mathrm{Cov}\left(z_j(t) - \boldsymbol{z}_{-j}^\top(t)\boldsymbol{w}_{j,-j}^*, \boldsymbol{z}_{-j}(t)\right) = 0$$

This leads to

$$\begin{aligned}\boldsymbol{w}_{j,-j}^* &= \mathrm{Cov}\left(z_j(t), \boldsymbol{z}_{-j}(t)\right)\left(\mathrm{Cov}\left(\boldsymbol{z}_{-j}(t), \boldsymbol{z}_{-j}(t)\right)\right)^{-1}\\ &= \mathrm{Cov}\left(z_j(t), \boldsymbol{z}_{-j}(t)\right)\left((\Upsilon)_{-j,-j}\right)^{-1},\end{aligned}$$

where $\Upsilon_{-j,-j} = \mathrm{Var}\left(\boldsymbol{z}_{-j}\right)$. Therefore,

$$\|\boldsymbol{w}_{j,-j}^*\|_0 \le \|\{k : j\ne k, \mathrm{Cov}\left(z_j(t), z_k(t)\right)\ne 0\}\|_0.$$

Recall that $z_j(t) = x_j(t)/\sigma_i(t)$ and $\sigma_i(t) = \lambda_i(t)(1-\lambda_i(t))$,where $\lambda_i(t) = \mu_i + \boldsymbol{x}^\top(t)\boldsymbol{\beta}_i$. Notice that by the sparsity of $\boldsymbol{\beta}_i$ tat $\|\boldsymbol{\beta}_i\|_0 \le \rho$, the number of $x_j(t)$'s that $\lambda_i(t)$ depends on is at most ρ. Let $S = \{j : \boldsymbol{\beta}_{i,j}\ne 0\}$ and $S^c = \{j : \boldsymbol{\beta}_{i,j} = 0\}$. Thus,

$$\begin{aligned}\|\boldsymbol{w}_{j,-j}^*\|_0 &\le \|\{k : k\ne j \quad \text{and} \quad k\in S^c, \mathrm{Cov}\left(z_j(t), z_k(t)\right)\ne 0\}\|_0 + \rho\\ &\le \|\{k : k\ne j \quad \text{and} \quad k\in S^c, \mathrm{Cov}\left(x_j(t), x_k(t)\right)\ne 0\}\|_0 + \rho.\end{aligned}$$

Recalling that $x_j(t) = \int_0^t k_j(t-s)dN_j(s)$,

$$\begin{aligned}\mathrm{Cov}\left(x_j(t), x_k(t)\right) &= \mathrm{Cov}\left(\int_0^t k_j(t-s)dN_j(s), \int_0^t k_j(t-s)dN_j(s)\right) \\ &= \int_0^t \int_0^t k_j(t-s)k_j(t-r)\mathrm{Cov}\left(dN_j(s), dN_j(r)\right).\end{aligned}$$

Therefore, with a positive transition kernel in Assumption 2.4,

$$\begin{aligned}&\left\|\{k : k \neq j \quad \text{and} \quad k \in S^c, \mathrm{Cov}\left(x_j(t), x_k(t)\right) \neq 0\}\right\|_0 \\ &\qquad \leq \left\|\{k : k \neq j \quad \text{and} \quad k \in S^c, \mathrm{Cov}\left(dN_j(s), dN_j(r)\right) \neq 0\}\right\|_0.\end{aligned}$$

Consider a connectivity matrix of block structure, for each j, all units that the unit j depends on must stay in one of the blocks on the connectivity matrix. Therefore, the possible number of units it depends on is at most ρ; that is,

$$\left\|\{k : j \neq k, \mathrm{Cov}\left(dN_j(s), dN_j(r)\right) \neq 0\}\right\|_0 \leq \rho,$$

which implies

$$s = \max_{1\leq j\leq p} \|\boldsymbol{w}_j^*\|_0 \leq 1 + \max_{1\leq j\leq p} \|\boldsymbol{w}_{j,-j}^*\|_0 \leq 2\rho + 1.$$

□

Lemma A.14. *Suppose the linear Hawkes model with its intensity function defined in (2.6) is stationary and satisfies Assumptions 2.1 – 2.4. Then,*

$$\|\boldsymbol{w}_j^*\|_2^2 \leq C \quad \textit{and} \quad \|\boldsymbol{w}_j^*\|_1 \leq \sqrt{s_j C},$$

where C is a constant only depending on the model parameter $(\boldsymbol{\mu}, \Theta)$ and the transition kernel function.

Proof. Recall that $\Upsilon = \mathrm{Cov}\left(\boldsymbol{z}(t)\right)$ whose eigenvalue is bounded by $\Lambda_{\max}\left(\Upsilon_x\right)$ and $\Lambda_{\min}\left(\Upsilon_x\right)$ since $\sigma_i^2(t)$ is bounded implied by Assumptions 2.3 and 2.4. Let $\Upsilon_{-j,-j} = \mathrm{Var}\left(\boldsymbol{z}_{-j}\right)$. Then,

$\Lambda_{\max}(\Upsilon_{-j,-j}) \leq \Lambda_{\max}(\Upsilon)$. In addition, let $\Upsilon_{j,-j} = \mathrm{Cov}(z_j, \boldsymbol{z}_{-j})$. Then, $\|\Upsilon_{j,-j}\|_2^2 \leq \|\Upsilon_{j,\cdot}\|_2^2 = (\Upsilon^2)_{j,j} \leq \Lambda_{\max}(\Upsilon^2) \leq \Lambda_{\max}^2(\Upsilon)$. Thus,

$$\begin{aligned}\|\boldsymbol{w}_j^*\|_2^2 &= 1 + \|\boldsymbol{w}_{j,-j}^*\|_2^2 \\ &\leq 1 + \left(\Lambda_{\max}\left(\Upsilon_{-j,-j}^{-1}\right)\right)^2 \|\Upsilon_{j,-j}\|_2^2 \\ &\leq 1 + \left(\Lambda_{\min}(\Upsilon)\right)^{-2} \Lambda_{\max}^2(\Upsilon) \leq C,\end{aligned}$$

where the last inequality is by Proposition A.2 for some constant C only depend on $(\Theta, \boldsymbol{\mu})$ and the transition kernel function. Then,

$$\|\boldsymbol{w}_j^*\|_1 \leq \sqrt{s_j}\|\boldsymbol{w}_j^*\|_2 \leq \sqrt{s_j C}.$$

□

Lemma A.15 (Theorem 3, Chen et al. [2019])**.** *Suppose the linear Hawkes model with its intensity function defined in (2.6) is stationary and satisfies Assumptions 2.1 – 2.3. For $\mathcal{H}_t$-predictable function $f(\cdot)$ that is bounded, let*

$$y_{ij} = \frac{1}{T}\int_0^T \int_0^T f(t,t')dN_i(t)dN_j(t').$$

Then there exists constants $C_1, C_2 > 0$ such that

$$\mathbb{P}\left(|y_{ij} - \mathbb{E}y_{ij}| > \epsilon\right) \leq C_1 T \exp\left(-C_2(\epsilon T)^{1/3}\right).$$

Proof. The conditions required by Chen et al. [2019] are satisfied by our Assumptions 2.1 and 2.3, where the Assumption 2 in Chen et al. [2019] is satisfied with $b_0 = 0$ by the Assumption 2.2 in our case. Thus, taking $r \to \infty$, Lemma A.15 is a direct result following the proof of Theorem 3 in Chen et al. [2019]. □

Lemma A.16 (van de Geer [1995])**.** *Suppose the linear Hawkes model with its intensity function defined in (2.6) is stationary and there exists $\lambda_{\max}$ such that $\lambda_i(t) \leq \lambda_{\max}$ for all $1 \leq i \leq p$. Let $H(t)$ be a bounded function that is $\mathcal{H}_t$-predictable. Then, for any $\epsilon > 0$,*

$$\frac{1}{T}\int_0^T H(t)\Big\{\lambda_i(t)dt - dN_i(t)\Big\} \leq 4\Big\{\frac{\lambda_{\max}}{2T}\int_0^T H^2(t)dt\Big\}^{1/2}\epsilon^{1/2},$$

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

Lemma A.17. *Suppose the linear Hawkes model with its intensity function defined in (2.6) is stationary and satisfies Assumptions 2.1 – 2.4. Then, for $\delta > 0$ and $1 \le i, j \le p$,*

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T} x_j(t) - \mathbb{E}x_j(t)\right| > C_1\epsilon^{1/2}\right) \le C_2\exp(-\epsilon T)$$

and

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T} x_i(t)x_j(t) - \mathbb{E}x_i(t)x_j(t)\right| > \epsilon\right) \le C_3 T\exp(-C_4(\epsilon T)^{1/3}),$$

where $C_k, k = 1, \ldots, 4$ are constants only depending on the model parameter $(\boldsymbol{\mu}, \Theta)$ and the transition kernel function.

Proof. The concentration bound for the first order statistics of $x_j(t), 1 \le j \le p$ is a direct application of Lemma A.16. Notice that

$$\begin{aligned}\frac{1}{T}\int_0^T (x_j(t) - \mathbb{E}x_j(t))\,dt &= \frac{1}{T}\int_0^T \int_0^t \kappa_j(t-s)(dN_j(s) - \lambda_j(s)ds)dt \\ &= \frac{1}{T}\int_0^T \int_s^T \kappa_j(t-s)dt(dN_j(s) - \lambda_j(s)ds), \end{aligned} \tag{A.48}$$

where the last equality is by Fubini's theorem. Let $f_1(s) = \int_s^T \kappa_j(t-s)dt$. Since the transition kernel is integrable, $f_1(s)$ is bounded. In addition, $\lambda_j(t)$ is bounded by $\lambda_{\max}$ by Assumption 2.3. Thus, by Lemma A.16,

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T} x_j(t) - \mathbb{E}x_j(t)\right| > C_1\epsilon^{1/2}\right) \le C_2\exp(-\epsilon T). \tag{A.49}$$

Next, we consider the second order statistics of $x_j(t)$. Let $f_2(s,r) = \int_{\max\{s,r\}}^T \kappa_i(t-s)\kappa_j(t-r)dt$. Then,

$$\begin{aligned}\frac{1}{T}\int_0^T x_i(t)x_j(t)dt &= \frac{1}{T}\int_0^T dt \int_0^t \int_0^t \kappa_i(t-s)\kappa_j(t-r)dN_i(s)dN_j(r) \\ &= \frac{1}{T}\int_0^T dN_i(s)\int_s^T \int_0^t \kappa_i(t-s)\kappa_j(t-r)dN_j(r)dt \\ &= \frac{1}{T}\int_0^T dN_i(s)\int_0^T dN_j(r)f_2(s,r),\end{aligned}$$

where the second and third equalities are based on Fubini's theorem. By Assumption 2.4, $\kappa_i(t)$ is integrable. Therefore, $|f_2(s,r)| \leq \left(\int_0^T \kappa_i(t)dt\right)^2 < \infty$ is bounded. Applying Lemma A.15,

$$\mathbb{P}\left(\left|\frac{1}{T}\int_0^T x_i(t)x_j(t) - \mathbb{E}x_i(t)x_j(t)dt\right| > \epsilon\right) \leq C_3 T \exp(-C_4(\epsilon T)^{1/3}).$$

Since the observations are in discrete time, we replace the integral above by the numerical integration to reach the conclusion. □

Lemma A.18. *Suppose the linear Hawkes model with its intensity function defined in (2.6) is stationary and satisfies Assumptions 2.1 – 2.4. Then,*

$$\mathbb{P}\left(\left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{z}(t) - \mathbb{E}\boldsymbol{z}(t)\right\| > C_1\epsilon^{1/2}\right) \leq C_2 p \exp(-\epsilon T)$$

and

$$\mathbb{P}\left(\left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{z}^\top(t)\boldsymbol{z}(t) - \mathbb{E}\boldsymbol{z}^\top(t)\boldsymbol{z}(t)\right\|_\infty > \rho\epsilon\right) \leq C_3 p^2 T \exp(-C_4(\epsilon T)^{1/3}).$$

Proof. The proof is essentially based on the concentration inequality of the first and second order statistics of $\boldsymbol{x}(t)$ established in Lemma A.17. The difference is that this lemma focuses on the scaled design column, $\boldsymbol{z}(t) = \boldsymbol{x}(t)/\sigma_i(t)$ when testing β_{ij}. We denote $\sigma_i(t)$ as $\sigma(t)$ for short from now on. The extra term, $\sigma(t)$, makes the technical proof challenging. Fortunately, since $\sigma^2(t)$ depends on $\boldsymbol{x}(t)$ and $\boldsymbol{x}(t)\boldsymbol{x}^\top(t)$, we can expand the term $\boldsymbol{z}(t) = \boldsymbol{x}(t)/\sigma(t)$ using Taylor expansion in the combinations of $\boldsymbol{x}(t)$ and $\boldsymbol{x}(t)\boldsymbol{x}^\top(t)$, and then apply Lemma A.17 to establish a similar concentration inequality for $\boldsymbol{z}(t)$. In general, for any function $h(\boldsymbol{x}(t), \boldsymbol{\eta})$ as a function of $\boldsymbol{x}(t) \in \mathbb{R}^p$ and $\boldsymbol{\eta} \in \mathbb{R}^{p+1}$ that is second-order differentiable with bounded derivatives w.r.t $\boldsymbol{x}(t)$, the term $\frac{1}{T}\sum_{t=1}^{T} h(\boldsymbol{x}(t), \boldsymbol{\eta}) - \mathbb{E}h(\boldsymbol{x}(t), \boldsymbol{\eta})$ can be bounded using the concentration inequality of the first- and second order statistics of $\boldsymbol{x}(t)$ in Lemma A.17.

First, let h_{jk} be a function of $\boldsymbol{x}(t)$ and $\boldsymbol{\eta} = (\mu, \boldsymbol{\beta}^\top)^\top$,

$$h_{jk}\left(\boldsymbol{x}(t), \boldsymbol{\eta} = (\mu, \boldsymbol{\beta}^\top)^\top\right) \equiv \frac{1}{\sigma^2(t)}x_j(t)x_k(t) = \frac{1}{\sigma^2(t)}\boldsymbol{x}^\top(t)I_{jk}\boldsymbol{x}(t).$$

Here $I_{jk} = \boldsymbol{e}_j \boldsymbol{e}_k^\top$ and $\boldsymbol{e}_i$ are canonical basis vector, while, with a little abuse of notation, $\sigma^2(t)$ is a function of $\boldsymbol{\eta} = (\mu, \boldsymbol{\beta}^\top)^\top$ and $\boldsymbol{x}(t)$, where $\sigma^2(t) = \lambda(t)(1-\lambda(t))$ and $\lambda(t) = \mu + \boldsymbol{x}^\top(t)\boldsymbol{\beta}$.

The derivative of h_{jk} w.r.t $\boldsymbol{x}(t)$ is:

$$h'_{jk}\left(\mathbb{E}\left(\boldsymbol{x}(t)\right), \boldsymbol{\eta}\right) = \sigma^{-4}(t)(1 - 2(\mu + (\mathbb{E}\boldsymbol{x}(t))^\top\boldsymbol{\beta}))\boldsymbol{\beta}\mathbb{E}\left(\boldsymbol{x}^\top(t)\right) I_{jk}\mathbb{E}\left(\boldsymbol{x}(t)\right) + \sigma^{-2}(t)\mathbb{E}\left(\boldsymbol{x}(t)\right)(I_{jk} + I_{jk}^T).$$

Let $\rho = \|\boldsymbol{\beta}\|_0$. Since $\lambda(t)$ and $\boldsymbol{x}(t)$ are bounded, as implied by Assumption 2.3 and 2.4,

$$\left\| \frac{1}{T}\sum_{t=1}^{T} h'_{jk}(\mathbb{E}\left(\boldsymbol{x}(t)\right), \boldsymbol{\eta})\left(\boldsymbol{x}(t) - \mathbb{E}\left(\boldsymbol{x}(t)\right)\right)^\top \right\|_\infty = O\left(\rho \left\| \frac{1}{T}\sum_{t=1}^{T} \left(\boldsymbol{x}(t) - \mathbb{E}\left(\boldsymbol{x}(t)\right)\right)^\top \right\|_\infty\right).$$

The second derivative of h_{jk} w.r.t $\boldsymbol{x}(t)$ is:

$$\begin{aligned}
&h^{(2)}_{jk}(\mathbb{E}\left(\boldsymbol{x}(t)\right), \boldsymbol{\eta}) \\
&= -2\sigma^{-6}(t)(1 - 2(\mu + (\mathbb{E}\boldsymbol{x}(t))^\top\boldsymbol{\beta}))\boldsymbol{\beta}\left((1 - 2(\mu + (\mathbb{E}\boldsymbol{x}(t))^\top\boldsymbol{\beta}))\boldsymbol{\beta}\right)^\top \mathbb{E}\left(\boldsymbol{x}^\top(t)\right) I_{jk}\mathbb{E}\left(\boldsymbol{x}(t)\right) \\
&+ \sigma^{-4}(t)(-2\boldsymbol{\beta}\boldsymbol{\beta}^\top\mathbb{E}\left(\boldsymbol{x}^\top(t)\right) I_{jk}\mathbb{E}\left(\boldsymbol{x}(t)\right) \\
&+ (1 - 2(\mu + (\mathbb{E}\boldsymbol{x}(t))^\top\boldsymbol{\beta}))\boldsymbol{\beta}(\boldsymbol{e}_j^\top + \boldsymbol{e}_k^\top)\mathbb{E}\left(\boldsymbol{x}(t)\right)) + \sigma^{-2}(t)(I_{jk} + I_{jk}^T).
\end{aligned}$$

Thus, with bounded $x_j(t)$ and $\lambda_i(t)$,

$$\|h^{(2)}_{jk}(\mathbb{E}\left(\boldsymbol{x}(t)\right), \boldsymbol{\eta})\|_1 = O\left(\rho\|\boldsymbol{\beta}\|_\infty^2\right) + O\left(\|\boldsymbol{\beta}\|_\infty\right),$$

and

$$\|h^{(2)}_{jk}(\mathbb{E}\left(\boldsymbol{x}(t)\right), \boldsymbol{\eta})\mathbb{E}\left(\boldsymbol{x}(t)\right)\|_1 = O\left(\rho\|\boldsymbol{\beta}\|_\infty\right).$$

In addition, under a stationary Hawkes process, $\mathbb{E}\left(\boldsymbol{x}(t)\right)$ is a constant only depending on the model parameters $(\boldsymbol{\mu}, \Theta)$. Thus,

$$\left| \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}^\top(t)h^{(2)}_{jk}(\mathbb{E}\left(\boldsymbol{x}(t)\right), \boldsymbol{\eta})\boldsymbol{x}(t) - \mathbb{E}\left(\boldsymbol{x}^\top(t)h^{(2)}_{jk}(\mathbb{E}\left(\boldsymbol{x}(t)\right), \boldsymbol{\eta})\boldsymbol{x}(t)\right) \right| = O\left(\rho \left\| \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}^\top(t)\boldsymbol{x}(t) - \mathbb{E}\left(\boldsymbol{x}(t)^\top\boldsymbol{x}(t)\right) \right\|_\infty\right),$$

and

$$\begin{aligned}
&\left\|\frac{1}{T}\sum_{t=1}^{T}\left(\boldsymbol{x}(t)-\mathbb{E}\left(\boldsymbol{x}(t)\right)\right)^{\top}h_{jk}^{(2)}(\mathbb{E}\left(\boldsymbol{x}(t)\right),\boldsymbol{\eta})\mathbb{E}\left(\boldsymbol{x}(t)\right)\right\|_{\infty}\\
&\qquad\leq\left\|h_{jk}^{(2)}(\mathbb{E}\left(\boldsymbol{x}(t)\right),\boldsymbol{\eta})\mathbb{E}\left(\boldsymbol{x}(t)\right)\right\|_{1}\left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}(t)-\mathbb{E}\left(\boldsymbol{x}(t)\right)\right\|_{\infty}\\
&\qquad\leq O\left(\rho\left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}(t)-\mathbb{E}\left(\boldsymbol{x}(t)\right)\right\|_{\infty}\right).
\end{aligned}$$

The Taylor expansion of $h_{jk}(\boldsymbol{x}(t),\boldsymbol{\eta})$ around $\mathbb{E}\left(\boldsymbol{x}(t)\right)$ is

$$\begin{aligned}
h_{jk}(\boldsymbol{x}(t),\boldsymbol{\eta})=&h_{jk}(\mathbb{E}\left(\boldsymbol{x}(t)\right),\boldsymbol{\eta})+h_{jk}'\left(\mathbb{E}\left(\boldsymbol{x}(t)\right),\boldsymbol{\eta}\right)^{\top}\left(\boldsymbol{x}(t)-\mathbb{E}\left(\boldsymbol{x}(t)\right)\right)\\
&+\frac{1}{2}\left(\boldsymbol{x}(t)-\mathbb{E}\boldsymbol{x}(t)\right)^{\top}h_{jk}^{(2)}\left(\mathbb{E}\left(\boldsymbol{x}(t)\right),\boldsymbol{\eta}\right)\left(\boldsymbol{x}(t)-\mathbb{E}\left(\boldsymbol{x}(t)\right)\right)\\
&+o\left(\left(\boldsymbol{x}(t)-\mathbb{E}\left(\boldsymbol{x}(t)\right)\right)^{\top}\left(\boldsymbol{x}(t)-\mathbb{E}\left(\boldsymbol{x}(t)\right)\right)\right).
\end{aligned}$$

Then,

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}h_{jk}(\boldsymbol{x}(t),\boldsymbol{\eta})-\mathbb{E}\left(h(\boldsymbol{x}(t),\boldsymbol{\eta})\right)\\
=&\frac{1}{T}\sum_{t=1}^{T}\left(\boldsymbol{x}(t)-\mathbb{E}\left(\boldsymbol{x}(t)\right)\right)^{\top}h_{jk}'(\mathbb{E}\left(\boldsymbol{x}(t)\right),\boldsymbol{\eta})\\
&+\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}^{\top}(t)h_{jk}^{(2)}(\mathbb{E}\left(\boldsymbol{x}(t)\right),\boldsymbol{\eta})\boldsymbol{x}(t)-\mathbb{E}\left(\boldsymbol{x}^{\top}(t)h_{jk}^{(2)}(\mathbb{E}\left(\boldsymbol{x}(t)\right),\boldsymbol{\eta})\boldsymbol{x}(t)\right)\\
&-2\frac{1}{T}\sum_{t=1}^{T}\left(\boldsymbol{x}(t)-\mathbb{E}\left(\boldsymbol{x}(t)\right)\right)^{\top}h_{jk}^{(2)}(\mathbb{E}\left(\boldsymbol{x}(t)\right),\boldsymbol{\eta})\mathbb{E}\left(\boldsymbol{x}(t)\right)\\
&+o\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}^{\top}(t)\boldsymbol{x}(t)-\mathbb{E}\left(\boldsymbol{x}^{\top}(t)\boldsymbol{x}(t)\right)-2\frac{1}{T}\sum_{t=1}^{T}\left(\boldsymbol{x}(t)-\mathbb{E}\left(\boldsymbol{x}(t)\right)\right)^{\top}\mathbb{E}\left(\boldsymbol{x}(t)\right)\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\left\|\frac{1}{T}\sum_{t=1}^{T}h_{jk}(\boldsymbol{x}(t),\boldsymbol{\eta})-\mathbb{E}\left(h_{jk}(\boldsymbol{x}(t),\boldsymbol{\eta})\right)\right\|_{\infty}\leq&O\left(\rho\left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}^{\top}(t)\boldsymbol{x}(t)-\mathbb{E}\left(\boldsymbol{x}(t)^{\top}\boldsymbol{x}(t)\right)\right\|_{\infty}\right)\\
&+O\left(\rho\left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}(t)-\mathbb{E}\left(\boldsymbol{x}(t)\right)\right\|_{\infty}\right).
\end{aligned}$$

Using Lemma A.17 and taking a union bound,

$$\mathbb{P}\left(\left|\frac{1}{T}\int_0^T \boldsymbol{z}(t)\boldsymbol{z}^\top(t) - \mathbb{E}\boldsymbol{z}(t)\boldsymbol{z}^\top(t)dt\right| > \rho\epsilon\right) \leq C_1 p^2 T \exp(-C_2(\epsilon T)^{1/3}).$$

To derive the bound for $\left\|\frac{1}{T}\sum_{t=1}^T \boldsymbol{z}(t) - \mathbb{E}\boldsymbol{z}(t)\right\|$, we repeat the steps of (A.48) and (A.49) in Lemma A.17, but with $f_1(s) = \frac{1}{\sigma_i(t)}\int_s^T \kappa_j(t-s)dt$. By $\sigma_i(t) = O(1)$ in Lemma A.8, we reach the conclusion.

□

A.4 An Alternative Concentration Inequality for Discrete Time Domain

Following the discussion at the end of Section 2.4, here we give an alternative concentration inequality for the first- and second-order integrated processed in discrete time domain. We first state an alternative assumption on the transfer kernel function.

Assumption A.1. *There exists $b > a > 0$ such that the transfer kernel function satisfies*

$$0 < \max_{1\le j\le p} \kappa_j(t) \le a\exp(-bt).$$

Compared to Assumption 2.4, we require a more stringent structure on the transition kernel function. Such structure allows us to generate an improved convergence rate of the second order statistics of $\boldsymbol{x}(t)$ as stated in Lemma A.19.

Proposition A.3. *Suppose Assumption 2.3 and Assumption A.1 hold. Then*

1. *there exists C_β, such that $\|\boldsymbol{\beta}_i\|_\infty \le C_\beta < \infty$, $i = 1, \dots, p$;*
2. *$\max_{1\le j\le p} \sum_{t=1}^{T} \kappa_j(t) < \infty$;*

Furthermore, let ω_{ij}^{*n} be n-th auto-convolution of ω_{ij}. Then

$$\omega_{ij}^{*n}(t) \le \beta_{ij} a^n \frac{t^{(n-1)}}{(n-1)!}\exp(-bt),$$
$$\Psi_{ij}(t) \equiv \sum_{n=1}^{\infty} \omega_{ij}^{*n}(t) \le \beta_{ij} a\exp(-(b-a)t),$$

and

$$\xi \equiv \max_{1\le i\le p} \sum_{j=1}^{p}\sum_{t=1}^{T} |\Psi_{ij}(t)| \le \rho C_\beta \frac{a\exp(-(b-a))}{1-\exp(-(b-a))} < \infty.$$

Proof. First, Assumption A.1 implies that $0 < x_j(t) < \infty$, $j = 1, \dots, p$. Thus, if $\exists \beta_{ij} = \pm\infty$, then $\lambda_i(t) = \mu_i + \boldsymbol{x}^\top(t)\boldsymbol{\beta}_i$ goes to ∞, which contradicts with Assumption 2.3 that $\lambda_i(t)$ is bounded.

By direct algebra, we have

$$\max_{1\leq j\leq p}\sum_{t=1}^{T}\kappa_j(t)\leq\frac{a\exp(-b)}{1-\exp(-b)}<\infty, \tag{A.50}$$

which proves the second point.

For the n-th auto-convolution, we have for $n=2$,

$$\omega_{ij}^{*2}(t)=\int_0^T\omega_{ij}(t-s)\omega_{ij}(s)ds \leq\beta_{ij}a^2\int_0^t a\exp(-b(t-s))a\exp(-bs)ds=\beta_{ij}a^2t\exp(-bt);$$

and, for $n=3$,

$$\omega_{ij}^{*3}(t)=\int_0^T\omega_{ij}^{*2}(t-s)\omega_{ij}(s)ds \leq\int_0^t\beta_{ij}a^2(t-s)\exp(-b(t-s))a\exp(-bs)ds=\beta_{ij}a^3\frac{t^2}{2}\exp(-bt).$$

Suppose the result holds for n, then for $n+1$:

$$\omega_{ij}^{*(n+1)}(t)=\int_0^T\omega_{ij}^{*n}(t-s)\omega_{ij}(s)ds \leq\int_0^t\beta_{ij}a^n\frac{t^{(n-1)}}{(n-1)!}\exp(-b(t-s))a\exp(-bs)ds=\beta_{ij}a^{n+1}\frac{t^n}{n!}\exp(-bt),$$

which implies that the result holds by induction. Then, by direct algebra, we have

$$\sum_{n=1}^{\infty}\omega_{ij}^{*n}(t)\leq\beta_{ij}a\exp(-(b-a)t)$$

and

$$\max_{1\leq i\leq p}\sum_{j=1}^{p}\sum_{t=1}^{T}|\Psi_{ij}(t)|\leq\rho C_\beta\frac{a\exp(-(b-a))}{1-\exp(-(b-a))}<\infty.$$

□

Lemma A.19. *Suppose the linear Hawkes process with its intensity function defined in (2.6) is a stationary process defined in a discrete time domain that satisfies Assumptions 2.1–2.3 and Assumption A.1. Then, for $\delta > 0$ and $1 \le i, j \le p$,*

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T} x_i(t)x_j(t)^\top - \mathbb{E}\left(\frac{1}{T}\sum_{t=1}^{T} x_i(t)x_j(t)^\top\right)\right| > \delta\right) \le C_1 \exp\left(-C_2 \min\left\{\sqrt{\frac{T}{\rho}}\delta, \frac{T}{\rho}\delta^2\right\}\right), \quad \text{(A.51)}$$

and

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T} x_j(t) - \mathbb{E}x_j(t)\right| > \delta\right) \le C_3 \exp(-C_4 T\delta^2), \quad \text{(A.52)}$$

where $\rho = \max\|\boldsymbol{\beta}_i\|_0$ and where $C_k, k = 1, \ldots, 4$ are constants only depending on the model parameter $(\boldsymbol{\mu}, \Theta)$ and the transition kernel function.

Proof. Consider the linear Hawkes process defined on a discrete time domain in $t = 1, \ldots, T$. As defined in (2.5), $x_j(t) = \sum_{s=1}^{t-1} \kappa_j(t-s)Y_j(s), t > 1$, and set $x_j(1) = 0$. Let

$$K_j = \begin{pmatrix} 0 & \kappa_j(1) & \kappa_j(2) & \kappa_j(3) & \ldots & \kappa_j(T-1) \\ 0 & 0 & \kappa_j(1) & \kappa_j(2) & \ldots & \kappa_j(T-2) \\ 0 & 0 & 0 & \kappa_j(1) & \ldots & \kappa_j(T-3) \\ 0 & . & . & . & \ldots & . \\ 0 & 0 & 0 & 0 & 0 & \kappa_j(1) \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \in \mathbb{R}^{T\times T},$$

where $\kappa_j(\cdot)$ is the transition kernel function, and let $Y_{jt} = (Y_j(t), \ldots, Y_j(1))^\top \in \mathbb{R}^t$. Then,

$$\boldsymbol{x}_j \equiv \begin{pmatrix} x_j(T) \\ \ldots \\ x_j(1) \end{pmatrix} = K_j Y_{jT}.$$

Recall that $\lambda_j(t)$, defined in (2.6), is the intensity function of unit j at time t; $\epsilon_j(t) = Y_j(t) - \lambda_j(t)$. Let $\boldsymbol{\epsilon}(s) = (\epsilon_1(s), \ldots, \epsilon_p(s))^\top \in \mathbb{R}^p$. Under a stationary Hawkes process [Bacry

et al., 2011, Proposition 1]:

$$Y_j(t) = \Lambda_j + \Psi_j * \epsilon_j(t) = \Lambda_j + \sum_{s=1}^{t} \Psi_j(t-s)\boldsymbol{\epsilon}(s).$$

Here $\Lambda_j = \mathbb{E}\lambda_j(t)$; $\Psi_j(t) = (\Psi_{j1}(t), \ldots, \Psi_{jp}(t))$ and $\Psi_{jl}(t) = \sum_{n=1}^{\infty} \omega_{jl}^{*n}(t)$, where ω_{jl}^{*n} is n-th auto-convolution of the transfer function ω_{jl} defined in (2.3).

Let

$$\Xi_j = \begin{pmatrix} \Psi_j(1) & \Psi_j(2) & \Psi_j(3) & \ldots & \Psi_j(T) \\ 0 & \Psi_j(1) & \Psi_j(2) & \ldots & \Psi_j(T-1) \\ 0 & 0 & \Psi_j(1) & \ldots & \Psi_j(T-2) \\ . & . & . & \ldots & . \\ 0 & 0 & 0 & 0 & \Psi_j(1) \end{pmatrix} \in \mathbb{R}^{T \times Tp},$$

Further, let $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}(T), \ldots, \boldsymbol{\epsilon}(1))^\top \in \mathbb{R}^{Tp}$, then

$$Y_{jT} = \Lambda_j + \Xi_j \boldsymbol{\epsilon}.$$

Thus,

$$\boldsymbol{x}_j = K_j(\Lambda_j + \Xi_j \boldsymbol{\epsilon}) = K_j\Lambda_j + K_j\Xi_j\boldsymbol{\epsilon},$$
$$\boldsymbol{x}_j - \mathbb{E}\boldsymbol{x}_j = K_j\Xi_j\boldsymbol{\epsilon},$$

which leads to

$$\boldsymbol{x}_i^\top \boldsymbol{x}_j - \mathbb{E}\left(\boldsymbol{x}_i^\top \boldsymbol{x}_j\right) = \Lambda_i^\top K_i^\top K_j \Xi_j \boldsymbol{\epsilon} + \Lambda_j^\top K_j^\top K_i \Xi_i \boldsymbol{\epsilon} + \boldsymbol{\epsilon}^\top \Xi_i^\top K_i^\top K_j \Xi_j \boldsymbol{\epsilon}.$$

We bound each term in the display above. First, notice that

$$\|\Lambda_i^\top K_i^\top K_j \Xi_j\|_2^2 \leq \|\Lambda_i\|_2^2 \Lambda_{\max}\left(K_i^\top K_j \Xi_j \Xi_j^\top K_j^\top K_i\right).$$

According to Assumption 2.3, $\|\Lambda_i\|_2^2 \leq \lambda_{\max} T$. In addition,

$$\begin{aligned} \Lambda_{\max}\left(K_i^\top K_j \Xi_j \Xi_j^\top K_j^\top K_i\right) &\leq \Lambda_{\max}\left(K_i\right)^2 \Lambda_{\max}\left(K_j\right)^2 \Lambda_{\max}\left(\Xi_j\right)^2 \\ &\leq \left(\sum_{t=1}^{T} \kappa_i(t)\right)^2 \left(\sum_{t=1}^{T} \kappa_j(t)\right)^2 \left(\sum_{k=1}^{p}\sum_{t=1}^{T} \Psi_{ik}(t)\right)^2, \end{aligned}$$

which is bounded by Assumption A.1 and its implication in Proposition A.3, where the second inequality is based on Perron–Frobenius theorem. Therefore, applying the sub-Gaussian deviation bound [Vershynin, 2010, Prop 5.10],

$$\mathbb{P}\left(\left\|\Lambda_i^\top K_i^\top K_j \Xi_j \boldsymbol{\epsilon}\right\|_2 > T\delta\right) \leq c_1 \exp(-c_2 T\delta^2).$$

Similarly,

$$\mathbb{P}\left(\left\|\Lambda_j^\top K_j^\top K_i \Xi_i \boldsymbol{\epsilon}\right\|_2 > T\delta\right) \leq c_3 \exp(-c_4 T\delta^2).$$

Next, we bound $\boldsymbol{\epsilon}^\top \Xi_i^\top K_i^\top K_j \Xi_j \boldsymbol{\epsilon}$ by examining the structure of K_i and Ξ_i. By Assumption A.1 and its implication in Proposition A.3, $\kappa_j(t) \leq a\exp(-bt)$ and $\Psi_{ij}(t) \leq C_1 \exp(-ct)$, where $c = b - a$ and $C_1 = aC_\beta$. Note that $\|\boldsymbol{\beta}_i\|_1 = \rho_i \leq \rho = \max_{1\leq i\leq p} \rho_i$ implies that there are at most ρ items of $\Psi_{il}(t) \neq 0$. So instead of considering the entire p-variate process, we only consider at most 2ρ units such that $\Psi_{il}(t) \neq 0$ and $\Psi_{jl}(t) \neq 0$. Let

$$\widetilde{K} = a \begin{pmatrix} 0 & \exp(-bt) & \exp(-2bt) & \exp(-3bt) & \dots & \exp(-b(T-1)) \\ 0 & 0 & \exp(-bt) & \exp(-2bt) & \dots & \exp(-b(T-1)) \\ 0 & 0 & 0 & \exp(-bt) & \dots & \exp(-b(T-3)) \\ 0 & . & . & . & \dots & . \\ 0 & 0 & 0 & 0 & 0 & \exp(-bt) \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \in \mathbb{R}^{T\times T},$$

and

$$\widetilde{\Xi} = C_1 \begin{pmatrix} \mathbf{0}_{2\rho}^\top & \exp(-ct)\mathbf{1}_{2\rho}^\top & \exp(-2ct)\mathbf{1}_{2\rho}^\top & \exp(-3ct)\mathbf{1}_{2\rho}^\top & \dots & \exp(-c(T-1))\mathbf{1}_{2\rho}^\top \\ \mathbf{0}_{2\rho}^\top & \mathbf{0}_{2\rho}^\top & \exp(-ct)\mathbf{1}_{2\rho}^\top & \exp(-2ct)\mathbf{1}_{2\rho}^\top & \dots & \exp(-c(T-2))\mathbf{1}_{2\rho}^\top \\ \mathbf{0}_{2\rho}^\top & \mathbf{0}_{2\rho}^\top & \mathbf{0}_{2\rho}^\top & \exp(-ct)\mathbf{1}_{2\rho}^\top & \dots & \exp(-c(T-3))\mathbf{1}_{2\rho}^\top \\ \mathbf{0}_{2\rho}^\top & . & . & . & \dots & . \\ \mathbf{0}_{2\rho}^\top & \mathbf{0}_{2\rho}^\top & \mathbf{0}_{2\rho}^\top & \mathbf{0}_{2\rho}^\top & \mathbf{0}_{2\rho}^\top & \exp(-ct)\mathbf{1}_{2\rho}^\top \\ \mathbf{0}_{2\rho}^\top & \mathbf{0}_{2\rho}^\top & \mathbf{0}_{2\rho}^\top & \mathbf{0}_{2\rho}^\top & \mathbf{0}_{2\rho}^\top & \mathbf{0}_{2\rho}^\top \end{pmatrix} \in \mathbb{R}^{T\times 2\rho T},$$

where $\mathbf{a}_{2\rho}^\top = \underbrace{(a, a, \ldots, a)}_{2\rho}$ and $a \in \{0, 1\}$. Then,

$$\left\|\Xi_i^\top K_i^\top K_j \Xi_j\right\|_2^2 \leq \left\|\widetilde{\Xi}^\top \widetilde{K}^\top \widetilde{K} \widetilde{\Xi}\right\|_2^2.$$

Let $\Theta = \widetilde{K}\widetilde{\Xi}$, then we calculate

$$\Theta_{1,k(s-1)+1} = aC_1 \sum_{s=1}^{k} \exp(-bs)\exp(-c(k+1-s)) \leq C_2 \exp(-ak),$$

where $C_2 = aC_1$. Therefore, $\left\|\widetilde{K}\widetilde{\Xi}\right\|_2^2 \leq \left\|\widetilde{\Theta}\right\|_2^2$ where

$$\widetilde{\Theta} = C_2 \begin{pmatrix} \mathbf{0}_{2\rho}^\top & \exp(-a)\mathbf{1}_{2\rho}^\top & \exp(-2a)\mathbf{1}_{2\rho}^\top & \exp(-3a)\mathbf{1}_{2\rho}^\top & \ldots & \exp(-a(T-1))\mathbf{1}_{2\rho}^\top \\ \mathbf{0}_{2\rho}^\top & \mathbf{0}_{2\rho}^\top & \exp(-a)\mathbf{1}_{2\rho}^\top & \exp(-2a)\mathbf{1}_{2\rho}^\top & \ldots & \exp(-a(T-2))\mathbf{1}_{2\rho}^\top \\ \mathbf{0}_{2\rho}^\top & \mathbf{0}_{2\rho}^\top & \mathbf{0}_{2\rho}^\top & \exp(-a)\mathbf{1}_{2\rho}^\top & \ldots & \exp(-a(T-3))\mathbf{1}_{2\rho}^\top \\ \mathbf{0}_{2\rho}^\top & . & . & . & \ldots & . \\ \mathbf{0}_{2\rho}^\top & \mathbf{0}_{2\rho}^\top & \mathbf{0}_{2\rho}^\top & \mathbf{0}_{2\rho}^\top & \mathbf{0}_{2\rho}^\top & \exp(-a)\mathbf{1}_{2\rho}^\top \\ \mathbf{0}_{2\rho}^\top & \mathbf{0}_{2\rho}^\top & \mathbf{0}_{2\rho}^\top & \mathbf{0}_{2\rho}^\top & \mathbf{0}_{2\rho}^\top & \mathbf{0}_{2\rho}^\top \end{pmatrix} \in \mathbb{R}^{T \times 2\rho T}.$$

Next, we check $M = \widetilde{\Theta}^\top \widetilde{\Theta}$. Due to the structure of $\widetilde{\Theta}$, we get

$$M = C_2^2 \begin{pmatrix} \mathbf{0}_{2\rho}^\top & \mathbf{0}_{2\rho}^\top & \mathbf{0}_{2\rho}^\top & \mathbf{0}_{2\rho}^\top & \ldots & \mathbf{0}_{2\rho}^\top \\ \mathbf{0}_{2\rho}^\top & m_1\mathbf{1}_{2\rho}^\top & m_1\exp(-a)\mathbf{1}_{2\rho}^\top & m_1\exp(-2a)\mathbf{1}_{2\rho}^\top & \ldots & m_1\exp(-a(T-2))\mathbf{1}_{2\rho}^\top \\ . & . & m_2\mathbf{1}_{2\rho}^\top & m_2\exp(-a)\mathbf{1}_{2\rho}^\top & \ldots & m_2\exp(-a(T-3))\mathbf{1}_{2\rho}^\top \\ . & . & . & m_3\mathbf{1}_{2\rho}^\top & \ldots & m_3\exp(-a(T-4))\mathbf{1}_{2\rho}^\top \\ . & . & . & . & \ldots & . \\ . & . & . & . & . & m_{T-1}\mathbf{1}_{2\rho}^\top \\ . & . & . & . & . & \mathbf{0}_{2\rho}^\top \end{pmatrix} \in \mathbb{R}^{2\rho T \times 2\rho T},$$

where $m_t = \sum_{l=1}^{t} \exp(-al) \leq C_3 \exp(-a)$ for $t = 1, \ldots, T-1$. We only write out the upper triangle part of M in the above since $M = M^T$. Therefore,

$$\|M\|_2^2 \leq 2 \sum_{i=1}^{T} 2\rho \sum_{k=1}^{T-1} m_i \exp(-ka) \leq O(\rho T).$$

Since $\{\epsilon_i(t)\}_{1\leq i\leq p; t=1,\ldots,T}$ are mutually independent centered at 0 with bounded variance according to Assumption 2.3, applying Hanson-Wright inequality we get

$$\mathbb{P}\left(\left|\boldsymbol{\epsilon}^\top \Xi_i^\top K_i^\top K_j \Xi_j \boldsymbol{\epsilon} - \mathbb{E}\left(\boldsymbol{\epsilon}^\top \Xi_i^\top K_i^\top K_j \Xi_j \boldsymbol{\epsilon}\right)\right| > T\delta\right) \leq c_5 \exp(-c_6 \min\{T\delta/\|M\|_2, T^2\delta^2/\|M\|_2^2\}).$$

Therefore,

$$\begin{aligned}\mathbb{P}\left(\left|\boldsymbol{x}_i^\top \boldsymbol{x}_j - \mathbb{E}\boldsymbol{x}_i^\top \boldsymbol{x}_j\right| > T\delta\right) &\leq \mathbb{P}\left(\left|\Lambda_i^\top K_i^\top K_j \Xi_j \boldsymbol{\epsilon}\right| > T\delta\right) \\ &+ \mathbb{P}\left(\left|\Lambda_j^\top K_j^\top K_i \Xi_i \boldsymbol{\epsilon}\right| > T\delta\right) \\ &+ \mathbb{P}\left(\left|\boldsymbol{\epsilon}^\top \Xi_i^\top K_i^\top K_j \Xi_j \boldsymbol{\epsilon}\right| > T\delta\right) \\ &\leq c_7 \exp(-c_8 \min\left\{\sqrt{\frac{T}{s}}\delta, \frac{T}{s}\delta^2\right\}).\end{aligned}$$

The above gives the concentration inequality of the second order statistics of $\boldsymbol{x}(t)$.

Now we derive the deviation bound for the first order statistics of $\boldsymbol{x}(t)$. Recall that $\boldsymbol{x}_j - \mathbb{E}\boldsymbol{x}_j = K_j \Xi_j \boldsymbol{\epsilon}$ that we show earlier. In addition,

$$\|K_j \Xi_j\|_2^2 \leq \left\|\widetilde{K}\widetilde{\Xi}\right\|_2^2 \leq \left\|\widetilde{\Theta}\right\|_2^2 = O\left(T\exp(-2a)\right).$$

Then, applying the sub-Gaussian deviation bound stated in Vershynin [2010, Prop. 5.10],

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T} x_j(t) - \mathbb{E}x_j(t)\right| > \delta\right) \leq c_9 \exp(-c_{10} T\delta^2).$$

□

Appendix B

APPENDIX FOR CHAPTER 3

B.1 The Smoothing Gradient Descent Algorithm

Our estimator in (3.7) is solved using the smoothing proximal gradient descent algorithm [Chen et al., 2012]. The algorithm replaces the non-smooth fusion penalty by a smoothing approximation thus makes the original problem easier to solve using the fast iterative shrinkage thresholding algorithm [Beck and Teboulle, 2009]. In the follows, we specify the algorithm to the cases of linear and non-linear Hawkes processes.

Let $I_0 = \begin{pmatrix} 0 & 0 \\ 0 & I_p \end{pmatrix} \in \mathbb{R}^{(p+1)\times(p+1)}$ and I_p is the identity matrix. Then the sparse penalty is written as

$$\lambda_1 \sum_{m=1}^{M} \|\boldsymbol{\beta}_i^{(m)}\|_1 = \|\Lambda \boldsymbol{\theta}_i\|_1,$$

where $\Lambda = \lambda_1 I_M \otimes I_0$ and $I_M \in \mathbb{R}^{M\times M}$ is an identity matrix.

Let $d_{m,m'} = w_{m,m'}(\boldsymbol{e}_m^\top - \boldsymbol{e}_{m'}^\top) \otimes I_0 \in \mathbb{R}^{(p+1)\times(p+1)M}$, where $\boldsymbol{e}_m$ is canonical basis in $\mathbb{R}^M$. Let $D = \begin{pmatrix} d_{1,2} \\ \cdots \\ d_{m,m'} \\ \cdots \\ d_{M-1,M} \end{pmatrix} \in \mathbb{R}^{\binom{M}{2}(p+1)\times M(p+1)}$. Define $C = \lambda_2 D$. Then the fusion penalty becomes

$$\lambda_2 \sum_{1\le m<m'\le M} w_{m,m'} \left\| \boldsymbol{\beta}_i^{(m)} - \boldsymbol{\beta}_i^{(m')} \right\|_1 = \|C\boldsymbol{\theta}\|_1 . \tag{B.1}$$

Thus, the penalty in (3.8) can be written as

$$\mathcal{P}(\boldsymbol{\theta}_i) = \|\Lambda\boldsymbol{\theta}_i\|_1 + \|C\boldsymbol{\theta}_i\|_1. \tag{B.2}$$

Written using its dual norm,

$$\|C\boldsymbol{\theta}_i\|_1 = \max_{\|\boldsymbol{\alpha}_i\|_\infty \leq 1} \boldsymbol{\alpha}_i^T C\boldsymbol{\theta}_i. \tag{B.3}$$

Next, let $f_u(\boldsymbol{\theta}_i)$ be a smoothing approximation function such that

$$f_u(\boldsymbol{\theta}_i) = \|C\boldsymbol{\theta}_i\|_1 - u\frac{1}{2}\|\boldsymbol{\alpha}_i\|_2^2 = \max_{\|\boldsymbol{\alpha}_i\|_\infty \leq 1} \boldsymbol{\alpha}_i^T C\boldsymbol{\theta}_i - u\frac{1}{2}\|\boldsymbol{\alpha}_i\|_2^2, \tag{B.4}$$

where $u > 0$ is a smoothness parameter that controls the level of approximation to the fusion penalty. For example, if we take $u = \frac{4\epsilon}{M(M-1)}$, then the approximation error is up-to ϵ. Chen et al. [2012] shows that $f_u(\boldsymbol{\theta}_i)$ is convex and continuous differentiable where

$$\nabla f_u(\boldsymbol{\theta}_i) = C^T\boldsymbol{\alpha}_i^*. \tag{B.5}$$

Here $\boldsymbol{\alpha}_i^* = S(\frac{C\boldsymbol{\theta}_i}{u})$ and $S(\boldsymbol{z})$ is a coordinate-wise projection operator that projects each entry of $\boldsymbol{z}$ to the ℓ_∞-ball — i.e.,

$$S(z) = \begin{cases} -1 & z < 1 \\ z & z \in [-1, 1] \\ 1 & z > 1 \end{cases}.$$

Substituting the fusion penalty with (B.4), solving (3.7) becomes to solve

$$\widehat{\boldsymbol{\theta}}_i = \arg\min_{\boldsymbol{\theta}_i \in \mathbb{R}^{(p+1)M}} \{h(\boldsymbol{\theta}_i) + \|\Lambda_i\boldsymbol{\theta}_i\|_1\} \quad i = 1, \ldots, p, \tag{B.6}$$

where

$$h(\boldsymbol{\theta}_i) = \frac{1}{T}\sum_{m=1}^{M}\int_0^{T_m} \ell\left(dN_i^{(m)}(t), f_{\boldsymbol{\theta}_i^{(m)}}(\boldsymbol{x}^{(m)}(t))\right) + f_u(\boldsymbol{\theta}_i),$$

and $T = \sum_{m=1}^{M} T_m$. (B.6) can be solved using the fast iterative shrinkage thresholding (FISTA) algorithm [Beck and Teboulle, 2009]. The FISTA algorithm is a first-order optimization method [Beck, 2017] that requires evaluating the first derivative of $h(\boldsymbol{\theta}_i)$, and

a Lipschitz constant L calculated as a upper bound on the spectral radius of the second derivative of $h(\boldsymbol{\theta}_i)$.

Next, we specify the first derivative of $h(\boldsymbol{\theta}_i)$ and the Lipschitz constant L for the linear Hawkes model with least square loss -i.e. $\ell(a,b) = (a-b)^2$.

Let $\gamma_{ij}^{(m)} = \int_0^{T_m} x_j^{(m)}(t) dN_i^{(m)}(t)$, $\gamma_{i0}^{(m)} = \int_0^{T_m} dN_i^{(m)}(t)$, and $\boldsymbol{\gamma}_i^{(m)} = \left(\gamma_{i0}^{(m)}, \gamma_{i1}^{(m)}, \ldots, \gamma_{ip}^{(m)}\right)^\top \in \mathbb{R}^{p+1}$. Denote $\mathrm{Q}^{(m)} = \int_0^{T_m} \begin{pmatrix} 1 \\ \boldsymbol{x}^{(m)}(t) \end{pmatrix} \begin{pmatrix} 1 & \left(\boldsymbol{x}^{(m)}(t)\right)^\top \end{pmatrix} dt$.

Let $\mathrm{Q} = \begin{pmatrix} \mathrm{Q}^{(1)} & & & \\ & \mathrm{Q}^{(2)} & & \\ & & \ldots & \\ & & & \mathrm{Q}^{(M)} \end{pmatrix}$, and $\boldsymbol{\gamma} = \left(\left(\boldsymbol{\gamma}^{(1)}\right)^\top, \ldots, \left(\boldsymbol{\gamma}^{(m)}\right)^\top\right)^\top$, where $\mathrm{Q}^{(m)}$ and $\boldsymbol{\gamma}_i^{(m)}$ can be pre-calculated given data and the pre-specified transfer kernel function. With these notations,

$$\sum_{m=1}^{M} \int_0^{T_m} \ell\left(dN_i^{(m)}(t), f_{\boldsymbol{\theta}_i^{(m)}}(\boldsymbol{x}^{(m)}(t))\right) = \boldsymbol{\theta}_i^\top \mathrm{Q} \boldsymbol{\theta}_i - 2\boldsymbol{\theta}_i^\top \boldsymbol{\gamma}_i + \sum_{m=1}^{M} \int_0^{T_m} \left(dN_i^{(m)}(t)\right)^2, \quad \text{(B.7)}$$

which leads to

$$\nabla h(\boldsymbol{\theta}_i) = \frac{1}{T}\mathrm{Q}\boldsymbol{\theta}_i - \frac{1}{T}\boldsymbol{\gamma} + C^T \boldsymbol{\alpha}^*. \quad \text{(B.8)}$$

In addition, $\nabla^2 h(\boldsymbol{\theta}_i) \preccurlyeq \frac{1}{T}\mathrm{Q} + \frac{C^T C}{u}$, which leads us to choose $L = \Lambda_{\max}\{\frac{1}{T}\mathrm{Q} + \frac{C^T C}{u}\}$. Notice that both $\nabla h(\boldsymbol{\theta}_i)$ and L can be pre-calculated and stored in memory when implementing the FISTA algorithm. This feature makes the computation scalable to large size data collected over a long time range. The edge selection performance using this algorithm for the linear Hawkes process is illustrated in Figure 3.3 in Section 3.6.

Next, we specific $\nabla h(\boldsymbol{\theta}_i)$ and L for non-linear Hawkes model with the exponential-link function, $g_i(\cdot) = \exp(\cdot)$ and the negative log likelihood loss -i.e. $\ell(a,b) = -a\log(b) + b$.

With some algebra,

$$\nabla h(\boldsymbol{\theta}_i) = \frac{1}{T}\widetilde{Q}_i - \frac{1}{T}\boldsymbol{\gamma} + C^T \boldsymbol{\alpha}^*, \quad \text{(B.9)}$$

where

$$\widetilde{Q}_i = \begin{pmatrix} \widetilde{Q}_i^{(1)} & & & \\ & \widetilde{Q}_i^{(2)} & & \\ & & \cdots & \\ & & & \widetilde{Q}_i^{(M)} \end{pmatrix},$$

$$\widetilde{Q}_i^{(m)} = \int_0^{T_m} \boldsymbol{x}^{(m)}(t) \exp\left(\begin{pmatrix} 1 & \left(\boldsymbol{x}^{(m)}(t)\right)^\top \end{pmatrix} \boldsymbol{\theta}_i^{(m)}\right) dt, \quad m \in \{1, \ldots, M\}.$$

Unlike the linear model, $\nabla h(\boldsymbol{\theta}_i)$ depends on the unknown parameter. Therefore, we need to evaluate $\nabla h(\boldsymbol{\theta})$ at each step in the FISTA algorithm which slows down the algorithm given observations over long time periods.

Notice that $\nabla^2 h(\boldsymbol{\theta}_i) \preccurlyeq \max_{1\leq m\leq M} \left\{\exp\left(\begin{pmatrix} 1 & \left(\boldsymbol{x}^{(m)}(t)\right)^\top \end{pmatrix} \boldsymbol{\theta}_i^{(m)}\right)\right\} \frac{1}{T}\mathrm{Q} + \frac{C^T C}{u}$, which leads to a choice of L. However, we find that this choice on L leads slow converged or diverged results when $\frac{C^T C}{u}$ is large—e.g. with large λ_2 and small u. To mitigate this issue, we use a general convex programming solver as an alternative–e.g. `CVXR` in `R` [Fu et al., 2020], when the algorithm meets convergence problem. The edge selection performance using the algorithm for the non-linear Hawkes process is illustrated in Figure B.1.

We summarize the computing steps above in Algorithm 2. Notice that when the number of experiments, M, is large, u becomes very small (proportional to $O(1/M^2)$), which leads L very large. Then, the algorithm convergence becomes slow because the step-size, $1/L$, becomes tiny.

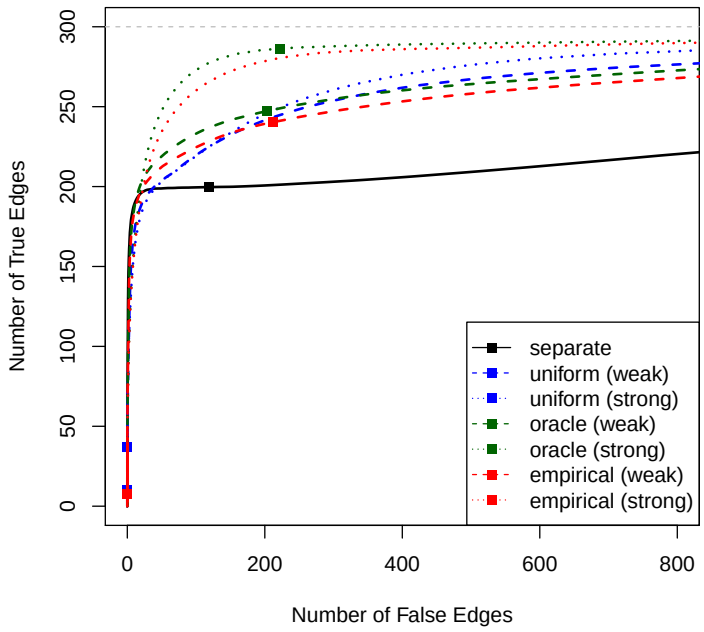


Figure B.1: Edge selection performance of the proposed joint estimation method in a simulation study focused on inferring edges in 3 networks of generalized Hawkes processes with exponential link function. The plots show average number of true positive and false positive edges, over 100 simulation runs, for the joint estimation method with different choices of weights, compared to separate estimation of each network. Weight strategies include oracle, empirical and uniform weights. Solid squares (■) correspond the choice of tuning parameter using eBIC. (a): Network 2 shares 90% edges with Network 1 and 10% with Network 3 as in Figure 3.2. (b): Network 2 is the same as Network 1.

Algorithm 2 Smoothing Proximal Gradient Descent for Generalized Hawkes Process

for $i = 1, \ldots, p$ **do**

Input: $\{N_i^{(m)}\}_{m=1}^M$, C, Λ, $\boldsymbol{\theta}_i^0$, L, desired accuracy ϵ

Initialization: set $u = \frac{4\epsilon}{(M-1)M}$, $\delta^0 = 1$, $\boldsymbol{w}^0 = \boldsymbol{\theta}_i^0$

repeat

1: Compute $\nabla h(\boldsymbol{w}^t)$;

2: Solve the proximal operator associated with ℓ_1-norm penalty:

$$\begin{aligned}\boldsymbol{\theta}_i^{t+1} &= \arg\min_{\boldsymbol{\theta}_i} \left(h(\boldsymbol{w}^t) + (\boldsymbol{\theta}_i - \boldsymbol{w}^t)^\top \nabla h(\boldsymbol{w}^t) + \frac{L}{2}\|\boldsymbol{\theta}_i - \boldsymbol{w}^t\|_2^2 + \lambda_1 \|\boldsymbol{\theta}_i\|_1 \right) \\ &= \arg\min_{\boldsymbol{\theta}_i} \frac{1}{2}\|\boldsymbol{\theta}_i - \boldsymbol{v}\|_2^2 + \frac{\lambda_1}{2}\|\boldsymbol{\theta}_i\|_1 \\ &= sign(\boldsymbol{v}) \max\left(0, |\boldsymbol{v}| - \frac{diag(\Lambda)}{L}\right),\end{aligned}$$

where $\boldsymbol{v} = \boldsymbol{w}^t - \frac{1}{L}\nabla h(\boldsymbol{w}^t)$.

3: $\delta^{t+1} = \frac{2}{t+3}$;

4: $\boldsymbol{w}^{t+1} = \boldsymbol{\theta}_i^{t+1} + \frac{1-\delta^t}{\delta^t}\delta^{t+1}(\boldsymbol{\theta}_i^{t+1} - \boldsymbol{\theta}_i^t)$

until convergence of $\boldsymbol{\theta}_i^t$;

end for

B.2 Proof of Main Theorems

For ease of presentation, we stack data from M experiments and re-label the time index over $[0, T]$ where $T = \sum_{m=1}^{M} T_m$. In particular, we denote $N_i(t) = N_i^{(m)}(t - \sum_{l=0}^{m-1} T_l)$ if $t \in \left(\sum_{l=0}^{m-1} T_l, \sum_{l=0}^{m} T_l\right]$ for $1 \le m \le M$, where $T_0 = 0$. Then, denote

$$\boldsymbol{x}(t) = \begin{pmatrix} \mathbf{1}(T_0 < t \le T_1) \\ \boldsymbol{x}^{(1)}(t)\mathbf{1}(T_0 < t \le T_1) \\ \dots \\ \mathbf{1}(\sum_{m=0}^{M-1} T_m < t \le T) \\ \boldsymbol{x}^{(M)}(t)\mathbf{1}(\sum_{m=0}^{M-1} T_m < t \le T) \end{pmatrix}, \quad \boldsymbol{\theta}_i = \begin{pmatrix} \boldsymbol{\theta}_i^{(1)} \\ \dots \\ \boldsymbol{\theta}_i^{(M)} \end{pmatrix}, \quad \lambda_i(t) = g_i(\boldsymbol{x}^\top(t)\boldsymbol{\theta}_i).$$

With these notations,

$$\sum_{m=1}^{M} \int_0^{T_m} \ell\left(dN_i^{(m)}(t), f_{\boldsymbol{\theta}_i^{(m)}}(\boldsymbol{x}^{(m)}(t))\right) = \int_0^T \ell\left(dN_i(t), f_{\boldsymbol{\theta}_i}(\boldsymbol{x}(t))\right).$$

In addition, denote $\ell(t; \boldsymbol{\theta}_i) = \ell\left(dN_i(t), f_{\boldsymbol{\theta}_i}(\boldsymbol{x}(t))\right)$.

Because optimization problem (3.7) can be solved separately for each component process, in the follows we illustrate the estimation consistency using the estimator (3.7) for one component process. Therefore, for ease of notation, we drop the subscript i; that is, we use $\boldsymbol{x}(t)$ for $\boldsymbol{x}_i(t)$, $\boldsymbol{\theta}$ for $\boldsymbol{\theta}_i$, $dN(t)$ for $dN_i(t)$, $g(\cdot)$ for $g_i(\cdot)$ and $\lambda(t)$ for $\lambda_i(t)$, S for S_i, and $\widetilde{S}$ for $\widetilde{S}_i$.

Proof of Theorem 3.1 :

Let $\Delta = \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}$. We linearize $\ell(t; \boldsymbol{\theta})$ w.r.t $\boldsymbol{\theta}$ using Taylor expansion:

$$\ell(t; \widehat{\boldsymbol{\theta}}) - \ell(t; \boldsymbol{\theta}) = \left(\nabla \ell(t; \boldsymbol{\theta})\right)^\top \Delta + \frac{1}{2}\Delta^\top \nabla^2 \ell(t; \boldsymbol{\theta})\Delta + o(\|\Delta\|_2^2).$$

Let $R(\boldsymbol{\theta}) = \|\Lambda\boldsymbol{\theta}\|_1 + \|C\boldsymbol{\theta}\|_1$, where Λ and C are defined in Appendix B.1. Taking $\widehat{\boldsymbol{\theta}}$ given

by (3.7),

$$\frac{1}{T}\int_0^T \ell(t;\widehat{\boldsymbol{\theta}}) - \frac{1}{T}\int_0^T \ell(t;\boldsymbol{\theta}) \leq R(\boldsymbol{\theta}) - R(\widehat{\boldsymbol{\theta}}).$$

Combining the above,

$$0 \leq \Delta^\top \left(\frac{1}{2T}\int_0^T \nabla^2 \ell(t;\boldsymbol{\theta})\right)\Delta \leq -\frac{1}{T}\int_0^T \left(\nabla \ell(t;\boldsymbol{\theta})\right)^\top \Delta + R(\boldsymbol{\theta}) - R(\widehat{\boldsymbol{\theta}}), \tag{B.10}$$

where

$$R(\boldsymbol{\theta}) - R(\widehat{\boldsymbol{\theta}}) \leq \lambda_1\|\Delta_S\|_1 - \lambda_1\|\Delta_{S^c}\|_1 + \lambda_2\left\|D_{\cdot,\widetilde{S}}\Delta_{\widetilde{S}}\right\|_1 - \lambda_2\left\|D_{\cdot,\widetilde{S}^c}\Delta_{\widetilde{S}^c}\right\|_1.$$

In addition,

$$-\frac{1}{T}\int_0^T \left(\nabla \ell(t;\boldsymbol{\theta})\right)^\top \Delta \leq \left\|-\frac{1}{T}\int_0^T \nabla \ell(t;\boldsymbol{\theta})\right\|_\infty \|\Delta\|_1.$$

Taking $\widetilde{\lambda} = \sqrt{M}\lambda_1 = \lambda_2 = 2\left\|-\frac{1}{T}\int_0^T \nabla \ell(t;\boldsymbol{\theta})\right\|_\infty$, we get

$$\frac{1}{\sqrt{M}}\|\Delta_{S^c}\|_1 + 2\|D_{\cdot,\widetilde{S}^c}\Delta_{\widetilde{S}^c}\|_1 \leq \frac{3}{\sqrt{M}}\|\Delta_S\|_1 + 2\|D_{\cdot,\widetilde{S}}\Delta_{\widetilde{S}}\|_1.$$

Let

$$\mathcal{C} = \left\{\Delta \in R^{M(p+1)} : \frac{1}{\sqrt{M}}\|\Delta_{S^c}\|_1 + 2\|D_{\cdot,\widetilde{S}^c}\Delta_{\widetilde{S}^c}\|_1 \leq \frac{3}{\sqrt{M}}\|\Delta_S\|_1 + 2\|D_{\cdot,\widetilde{S}}\Delta_{\widetilde{S}}\|_1\right\}.$$

By Condition 3.1, $\forall \Delta \in \mathcal{C}$, there exists $\eta, c, C > 0$ such that

$$\min_{\Delta\in\mathcal{C}} \Delta^\top \left(\frac{1}{T}\int_0^T \nabla^2 \ell(t;\boldsymbol{\theta})\right)\Delta \geq \eta\|\Delta\|_2^2, \tag{B.11}$$

with probability at least $1 - cp^2\sum_{m=1}^M T_m \exp(-CT_m^{1/5})$.

Moreover, letting $\Delta^{(m)} = \widehat{\boldsymbol{\theta}}^{(m)} - \boldsymbol{\theta}^{(m)}$ and $\Delta_j^{(m)} = \widehat{\theta}_j^{(m)} - \theta_j^{(m)}$,

$$\begin{aligned}
\|D_{\cdot,\widetilde{S}}\Delta_{\widetilde{S}}\|_1 &= \sum_{m\neq m'\in\widetilde{S}} w_{m,m'}\left\|\Delta_j^{(m)} - \Delta_j^{(m')}\right\|_1 \\
&\leq \sum_{m\neq m'\in\widetilde{S}} w_{m,m'}\left(\left\|\Delta_j^{(m)}\|_1 + \|\Delta_j^{(m')}\right\|_1\right) \\
&= \sum_{m\in\widetilde{S}}\sum_{\substack{m'\in\widetilde{S}\\ m'\neq m}} w_{m,m'}\left\|\Delta_j^{(m)}\right\|_1 \\
&\leq \phi_{\widetilde{S}}\left\|\Delta_{\widetilde{S}}\right\|_1,
\end{aligned} \tag{B.12}$$

where $\phi_{\widetilde{S}} = \max_{m\in\widetilde{S}} \sum_{m'\neq m\in\widetilde{S}} w_{m,m'}$. Here, with a little abuse of notation, $m \in \widetilde{S}$ means there exists j such that $(j,m) \in \widetilde{S}$. Because the weights are normalized—i.e., $\sum_{1\leq m\neq m'\leq M} w_{m,m'} = 1$, $\phi_{\widetilde{S}} \leq \max_{1\leq m\leq M} \sum_{1\leq m'\neq m\leq M} w_{m,m'} \leq 1$.

Next, plugging (B.11) and (B.12) into (B.10),

$$\begin{aligned}
\eta\|\Delta\|_2^2 &\leq 3\frac{\widetilde{\lambda}}{\sqrt{M}}\|\Delta_S\|_1 - \frac{\widetilde{\lambda}}{\sqrt{M}}\|\Delta_{S^c}\|_1 + 2\widetilde{\lambda}\|D_{\cdot,\widetilde{S}}\Delta_{\widetilde{S}}\|_1 - 2\widetilde{\lambda}\|D_{\cdot,\widetilde{S}^c}\Delta_{\widetilde{S}^c}\|_1 \\
&\leq 3\frac{\widetilde{\lambda}\sqrt{d^*M}}{\sqrt{M}}\|\Delta_S\|_2 + 2\widetilde{\lambda}\phi_{\widetilde{S}}\sqrt{r^*}\|\Delta_{\widetilde{S}}\|_2 \\
&\leq \widetilde{\lambda}\left(3\sqrt{d^*} + 2\phi_{\widetilde{S}}\sqrt{r^*}\right)\|\Delta\|_2, \qquad \text{(B.13)}
\end{aligned}$$

where the second inequality is by $\|\Delta_S\|_1 \leq \sqrt{|S|}\|\Delta_S\|_2$ and $|S| \leq d^*M$, and $\|\Delta_{\widetilde{S}}\|_1 \leq \sqrt{|\widetilde{S}|}\|\Delta_{\widetilde{S}}\|_2$ and $|\widetilde{S}| \leq r^*$.

At the end, we reach the conclusion by plugging $\widetilde{\lambda} = 2\left\|-\frac{1}{T}\int_0^T \nabla\ell(t;\boldsymbol{\theta})\right\|_\infty$ in (B.13), and by Condition 3.2, with probability at least $1 - c'pM\exp(-T^{1/5})$,

$$\left\|-\frac{1}{T}\int_0^T \nabla\ell(t;\boldsymbol{\theta})\right\|_\infty \leq C'T^{-2/5},$$

where c', C' are positive constants.

Proof of Corollary 3.1 : We first verify the conditions for the linear Hawkes model with least square loss -i.e. $\ell(a,b) = (a-b)^2$. In this case, we have

$$\nabla\ell(t;\boldsymbol{\theta}) = 2\left(dN(t) - \lambda(t)dt\right)\boldsymbol{x}(t).$$

By Lemma B.1 and taking the union bound over all entries of $\boldsymbol{x}(t)$,

$$\left\|\frac{1}{T}\int_0^T \nabla\ell(t;\boldsymbol{\theta})\right\|_\infty = \left\|\frac{1}{T}\int_0^T 2\left(dN(t) - \lambda(t)dt\right)\boldsymbol{x}(t)\right\|_\infty \leq CT^{-2/5},$$

with probability at least $1 - CpM\exp(-T^{1/5})$. Thus, Condition 3.2 is satisfied.

In addition,

$$\frac{1}{T}\int_0^T \nabla^2\ell(t;\boldsymbol{\theta}) = \frac{1}{T}\int_0^T \boldsymbol{x}(t)\boldsymbol{x}^\top(t)dt.$$

Thus, Condition 3.1 is satisfied after applying Lemma B.2,

Next, we verify the conditions for the non-linear Hawkes process with the exponential-link function, $g(\cdot) = \exp(\cdot)$ and estimated using the negative log likelihood loss -i.e. $\ell(a, b) = -a \log(b) + b$. In this case, we have

$$\nabla \ell(t; \boldsymbol{\theta}) = (dN(t) - \lambda(t)dt)\, \boldsymbol{x}(t).$$

Similar as the linear case, Condition 3.2 is satisfied using Lemma B.1.

Under Assumption 3.2, there exists $\lambda_{\min}$ such that $\lambda^{(m)}(t) \geq \lambda_{\min} > 0$

$$\frac{1}{T}\int_0^T \nabla^2 \ell(t; \boldsymbol{\theta}) = \frac{1}{T}\int_0^T \lambda(t)\boldsymbol{x}(t)\boldsymbol{x}^\top(t)dt \geq \lambda_{\min}\frac{1}{T}\int_0^T \boldsymbol{x}(t)\boldsymbol{x}^\top(t)dt.$$

Thus, Condition 3.1 is satisfied following Lemma B.2.

Proof of Theorem 3.2: Recall $S^{(m)} = \{\beta_{ij}^{(m)} : \beta_{ij}^{(m)} \neq 0, 1 \leq i, j \leq p\}$ and $S_C^{(m)} = \{\beta_{ij}^{(m)} : \beta_{ij}^{(m)} = 0, 1 \leq i, j \leq p\}$, $m \in \{1, \ldots, M\}$. 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 $\beta_{S^{(m)}} \in S^{(m)}$ and $\beta_{S_C^{(m)}} \in S_C^{(m)}$, $|\widehat{\beta}_{S^{(m)}} - \widehat{\beta}_{S_C^{(m)}}| \geq \Delta$ with high probability. By the β-min condition specified in Assumption 3.5, we have $\beta_{ij}^{(m)} \in S^{(m)} \geq 2\tau$. Theorem 3.1 shows that for $m = 1, \ldots, M$ and $1 \leq i, j \leq p$, $|\widehat{\beta}_{ij}^{(m)} - \beta_{ij}^{(m)}| \leq \tau$ with probability at least $1 - c_1 p^2 M^2 T \exp(-c_2 T^{1/5})$. Then, for any $\beta_{S^{(m)}} \in S^{(m)}$ and $\beta_{S_C^{(m)}} \in S_C^{(m)}$,

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

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 $S^{(m)}$ and $S_C^{(m)}$ based on the estimates. In fact, the post-selection estimator is

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

By Theorem 3.1, we have $|\widehat{\beta}_{S_C^{(m)}}| \leq \tau$, with probability $1 - c_1 p^2 M^2 T \exp(-c_2 T^{1/5})$. Then,

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

which means $\widetilde{\beta}$ selects $\beta_{S_C^{(m)}}$ into $S_C^{(m)}$ with high probability. In addition, since $|\widehat{\beta}_{S^{(m)}} - \beta_{S^{(m)}}| \leq \tau$,

$$|\widehat{\beta}_{S^{(m)}}| \geq |\beta_{S^{(m)}}| - \tau \geq \beta_{min} - \tau > \tau > 0.$$

Therefore,

$$\widetilde{\beta}_{S^{(m)}} = \widehat{\beta}_{S^{(m)}}\mathbf{1}(|\widehat{\beta}_{S^{(m)}}| > \tau) = \widehat{\beta}_{S^{(m)}} \neq 0,$$

which means $\widetilde{\beta}_{S^{(m)}}$ selects $\beta_{S^{(m)}}$ into $S^{(m)}$ with high probability.

Combining the two sides, the post-selection estimator $\widetilde{\beta}$ identifies $S^{(m)}$ and $S_C^{(m)}$ with high probability, for all $m = 1, \ldots, M$.

B.3 Supporting Lemmas

Lemma B.1 (van de Geer [1995]). *Suppose there exists* $\lambda_{\max}$ *such that* $\lambda(t) \leq \lambda_{\max}$ *where* $\lambda(t)$ *is the intensity function of Hawkes process defined in* (3.1). *Let* $H(t)$ *be a bounded function that is* $\mathcal{H}_t$*-predictable. Then, for any* $\epsilon > 0$,

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

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

Lemma B.2 (Wang et al. [2020b]). *Suppose the Hawkes process defined in* (3.6) *satisfies Assumptions 3.1– 3.4. Let* $Q = \frac{1}{T}\int_0^T \begin{pmatrix} 1 \\ \boldsymbol{x}(t) \end{pmatrix}\begin{pmatrix} 1 & \boldsymbol{x}^\top(t) \end{pmatrix} dt$, *where* $\boldsymbol{x}(t)$ *is defined in* (3.5). *Then, there exists* $\gamma > 0$ *such that*

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

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.*

B.4 Proof of FWER Control

Proof of Theorem 3.3:

We call the binary tree or its sub-tree is a null-tree if the true edge coefficients to be tested on its leaves are all 0. Notice that for any null edge—i.e. the corresponding edge coefficient is 0, there must be a null tree with this edge tested on its leaf (e.g. in a trivial case, the null tree is the leaf node itself with the null edge to be tested). The level of a null tree, l, is the level of the null tree's root - i.e., the connections between the root of the null tree and the root of the binary tree plus 1. It is easy to see that by the binary tree construction, $1 \leq l \leq M$ and the number of leaves for a null-tree at level l is $M - l + 1$. For example, when $\beta_k^{(m)} = 0$ for all $m = 1, \ldots, M$, the binary tree itself is a null tree of level 1. Another example is that when $\beta_k^{(m)} = 0$ only for $m = M$ and the binary tree allocates the Mth network to its deepest level—i.e., $\mathcal{L}_{R,M} = \{M\}$, the null tree for $\beta_k^{(M)}$ is the right child node at the bottom level M.

Recall that we index the connectivity/edge coefficients from 1 to p^2. Denote $\mathcal{H}_{l,k}$ as a collection of null hypothesis in testing $H_k^{(m)} : \beta_k^{(m)} = 0$, $m \in \{1, \ldots, M\}$, where their corresponding null-trees at level l. Let $\mathcal{H}_l = \{\mathcal{H}_{l,k} : \mathcal{H}_{l,k} \neq \emptyset, 1 \leq k \leq p^2\}$, and $n_l = |\mathcal{H}_l|$.

Following Step 2 in the hierarchical testing procedure, the root of a level-l null tree is rejected with probability not higher than α_l. In addition, our proposed hierarchical testing procedure requires a lower level test to be implemented only when the test of its parent node is rejected. Therefore,

$$\mathbb{P}\left(\bigcup_{H_k^{(m)} \in \mathcal{H}_{l,k}} H_k^{(m)} \text{ is rejected} \right) \leq \mathbb{P}\left(\text{ the root of the null-tree is rejected} \right) \leq \alpha_l.$$

Then, the FWER is controlled as follows.

$$\begin{aligned} FWER &= \mathbb{P}\left(\bigcup_{l=1}^{M}\bigcup_{\mathcal{H}_{l,k}\in\mathcal{H}_l}\bigcup_{H_k^{(m)}\in\mathcal{H}_{l,k}} H_k^{(m)} \text{ is rejected}\right) \\ &\leq \sum_{l=1}^{M} n_l \max_{\mathcal{H}_{l,k}\in\mathcal{H}_l} \mathbb{P}\left(\bigcup_{H_k^{(m)}\in\mathcal{H}_{l,k}} H_k^{(m)} \text{ is rejected}\right) \\ &\leq \sum_{l=1}^{M} n_l \alpha_l, \end{aligned}$$

where the first inequality is by Boole's inequality.

Let π_0 be the total number of null edges, which is no greater than the total edges p^2M. Let C_l be the number of leaves under a null-tree of level l where $C_l \leq M - l + 1$. Thus,

$$\sum_{l=1}^{M} n_l C_l = \pi_0 \leq p^2 M.$$

Then, taking $\alpha_l = \frac{\alpha}{p^2}\frac{C_l}{M}$,

$$FWER \leq \sum_{l=1}^{M} n_l \frac{\alpha}{p^2}\frac{C_l}{M} = \frac{\pi_0}{p^2 M}\alpha \leq \alpha.$$

Particularly, using an binary tree built on the hierarchical clustering with oracle similarity distance between networks and the null-edges are always put to lower-level of the tree, $C_l = M - l + 1$ and $\alpha_l = \frac{\alpha}{p^2}\frac{M-l+1}{M}$.

Proof of Theorem 3.4:

Next, we show that given large and sparse networks under a large size of experiments, the hierarchical testing procedure still controls the FWER without the knowledge of the oracle binary tree.

For edge indexed k that is not null at least at 1 experiment, we have

$$\sum_{l=2}^{M} r_l^{(k)}(M-l+1) \leq \frac{M(M-1)}{2},$$

where $r_l^{(k)}$ is the number of level-l null-tree under the binary tree for edge i. Since there is at most 1 null-tree of level l, $r_l^{(k)} \leq 1$. The maximum is achieved when there are $M-1$ null edge and 1 non-null edge which is allocated on the right corner —i.e. the deepest level of the tree.

Let $\frac{n_1}{p^2} \leq 1$ indicate the proportion of null trees—i.e. the true edge coefficients to be test on all leaves are null.

Taking $\alpha_l = \frac{\alpha}{p^2}\frac{M-l+1}{M}$,

$$\begin{aligned} FWER &= n_1\alpha_1 + \sum_{l=2}^{M} n_l\alpha_l \\ &= \frac{\alpha}{p^2}n_1 + \frac{\alpha}{p^2 M}\sum_{l=2}^{M} n_l(M-l+1) \\ &\leq \frac{\alpha}{p^2}n_1 + \frac{\alpha}{p^2 M}\frac{M(M-1)}{2}(p^2-n_1) \end{aligned}$$

Recall that $d^* \equiv \max_{1\leq m\leq M, 1\leq i\leq p} d_i^{(m)}$, where $d_i^{(m)} = |S_i^{(m)}|$ and $S_i^{(m)} = \{j : \beta_{ij}^{(m)} \neq 0, 1 \leq j \leq p\}$ as defined in Section 3.4.

Thus, $n_1 + d^* pM \geq p^2$. Then,

$$p^2 - n_1 \leq d^* pM,$$

which leads

$$FWER \leq \frac{\alpha}{p^2}n_1 + \frac{\alpha}{p^2}\frac{d^* pM(M-1)}{2}.$$

Notice that $n_1 \leq p^2$, then

$$FWER \leq \alpha\left(1 + \frac{d^* M(M-1)}{2p}\right).$$

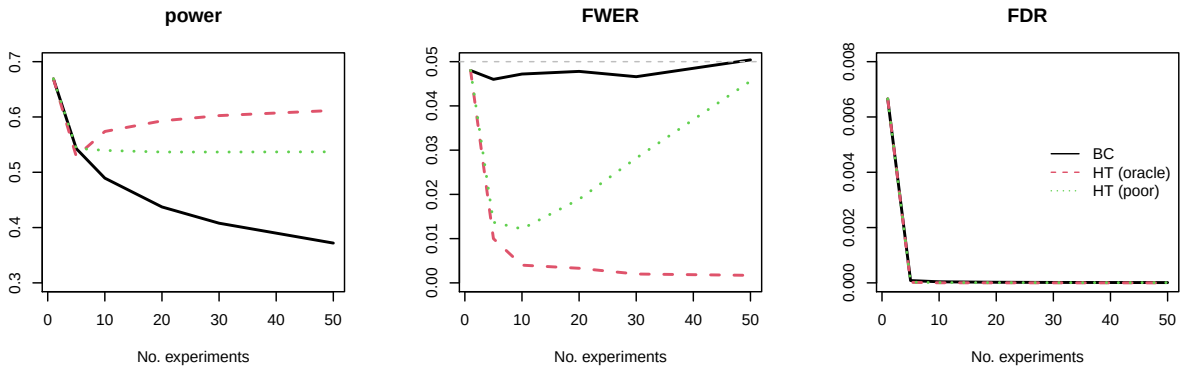


Figure B.2: Power, FWER and false discovery rate (FDR) between Bonferroni correction (BC) and the hierarchical testing procedure using oracle and poorly constructed binary trees (poor). The FWER is controlled at $\alpha = 0.05$ (gray dashline).

B.5 Additional Results

B.5.1 Testing

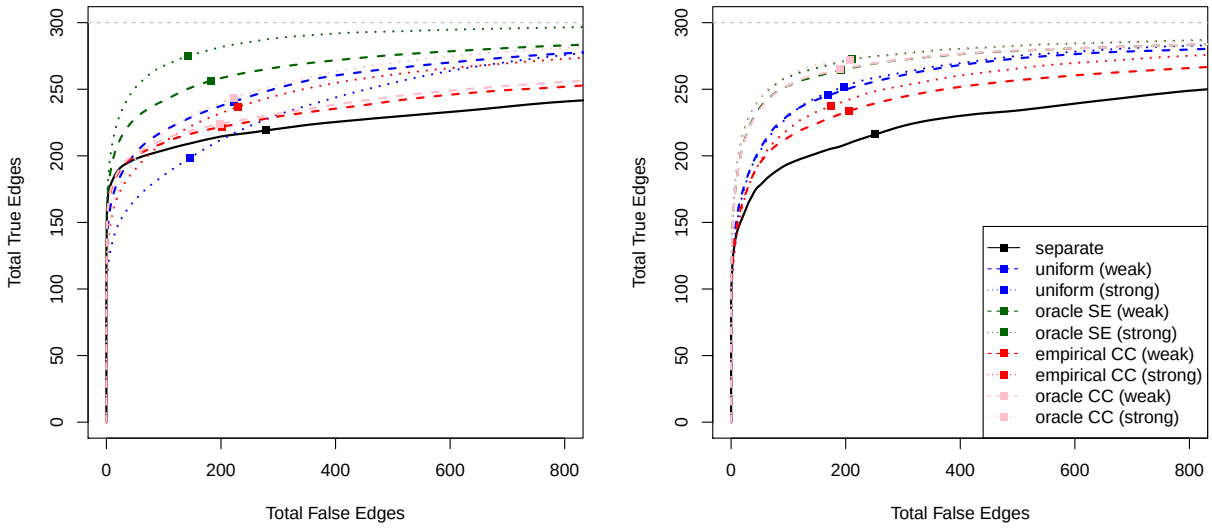
To illustrate how a poorly constructed hierarchy—i.e., the binary tree—shapes the power of the hierarchical testing procedure, we consider networks of $p = 100$ nodes under $M = 1, 5, 10, 20, 30, 50$ experiments. Half of networks are set to be inactive with 0.1% edges. The other half are active with 5% edges. No common edges are shared between the two types of networks. Note that we require the oracle binary tree put all inactive networks to deeper levels of the tree. Thus, as an extreme case, we consider a poorly constructed binary tree where active networks are put to deeper levels of the tree.

As expected, our procedure with the poorly constructed hierarchy generates lower power than that given using the oracle binary tree. However, it still tightly controls FWER and is more powerful than Bonferroni correction (see Figure B.2).

B.5.2 Connected-Component Similarity

We present additional simulation results on the effectiveness using the weight based on the connectivity-component similarity. We consider networks under three conditions, where the first two conditions are networks of circles as Figure 3.2 (left) and the third is the network of stars as Figure 3.2 (right). Then, we run another setting where the connectivity matrix of the star-network is flipped with the connected blocks from bottom left to right (Figure B.3). The difference between the two settings is that in the first one, even the networks are totally different, the connected components are the same across the three networks, while in the second setting, the third network is different from the first two in both edges and the connected components. It is not surprising to see the weight based on the connected-component similarity is not informative in the first simulation setting but is effective (as the oracle weight) in the second setting (Figure B.4). Notice that, in the second simulation setting, the empirical connected-component weight is not effective. Such poor performance comes from the nature of constructing the connected components which are easily influenced by noise, particularly using limited sample size in practice.

Figure B.3: Connectivity matrices of networks of $p = 100$ processes for network of circles (left) same as Figure 3.2 (left), stars (middle) Figure 3.2 (right) and flipped network of stars. The networks of circles and stars are totally different but their connected components are exactly the same.

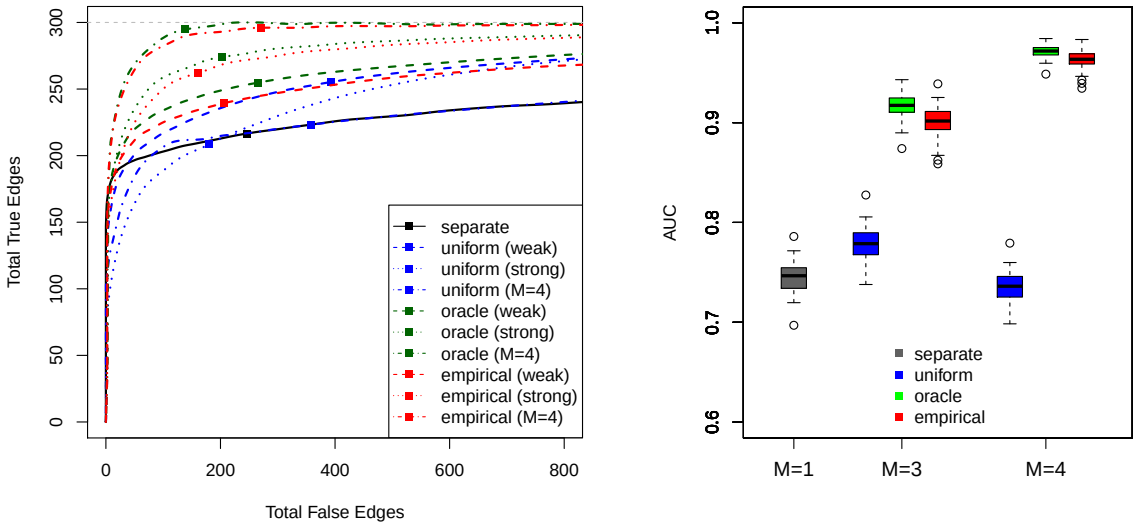


(a) (b)

Figure B.4: Edge selection performance of the proposed joint estimation method in a simulation study focused on inferring edges in 3 networks of linear Hawkes processes. The plots show average number of true positive and false positive edges, over 100 simulation runs, for the joint estimation method with different choices of weights, compared to separate estimation of each network.Weight strategies include oracle weight based on the number of shared edges (SE), connected components (CC), and empirical connect component weight, in addition to estimating each network separately. ■ indicates the optimal selection of edges using eBIC. (a): the first two conditions are networks of circles as Figure 3.2 (left) and the third is the star-network as Figure 3.2 (right). (b): same as (a) except the connectivity matrix of the third network is flipped with connected blocks from bottom left to right.

B.5.3 Increased Number of Experiments

We present additional simulation results on the benefit using more experiments where the additional experiments are similar as the existed ones. We consider networks under four conditions, where the first three conditions are exactly the same as in Figure 3.2 and the fourth network has the same structure as Network 1. The experiment periods are set the same as before for the first three condition and the fourth one is 500. We found that the inclusion of the extra condition improved the edge selection performance among the first three networks (Figure B.5). This finding is expected from our theoretical investigation where the error bound becomes tighter when the total experiment period gets larger and more similar conditions are involved.



(a) (b)

Figure B.5: Edge selection performance of the proposed joint estimation method in a simulation study evaluating the benefit of including extra experiments in the estimation procedure. The main result focused on inferring edges using 3 networks of linear Hawkes processes as in Figure 3.2. The performance is compared with the result using 4 networks (indicated as $M = 4$) with strong fusion penalty where the first three is the same as before and the extra one has the same structure as the first network. The plots in (a) show average number of true positive and false positive edges, over 100 simulation runs, for the joint estimation method with different choices of weights, compared to separate estimation of each network. Weight strategies include oracle, empirical and uniform weights. Solid squares (■) correspond the choice of tuning parameter using eBIC. The boxplots in (b) show the distribution of the area under the curve (AUC) values corresponding to the edge selection performance in (a) over 100 simulation runs for separate estimation ($M = 1$), and the joint estimation with different choices of weights using $M = 3$ and $M = 4$ experiments. The total numbers of true and false edges are normalized between 0 and 1 when calculating the AUCs.

Appendix C

APPENDIX FOR CHAPTER 4

C.1 HIVE

We introduce additional notations before illustrating the method.

Let $Y(t) = (Y_1(t), \ldots, Y_p(t))^\top$, $X(t) = (x_1(t), \ldots, x_p(t))^\top$, $Z(t) = (z_1(t), \ldots, z_q(t))^\top$ and $E(t) = (\epsilon_1(t), \ldots, \epsilon_p(t))^\top$. Then, we rewrite (4.8) simultaneously for all components:

$$Y(t) = \boldsymbol{\mu} + \Theta X(t) + \Delta Z(t) + E(t), \tag{C.1}$$

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. $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^\top \in \mathbf{R}^p$ 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

$$Z(t) = \boldsymbol{\nu} + AX(t) + W(t), \tag{C.2}$$

where A is the projection matrix, representing the cross-sectional correlation between Z and X. Then, (C.1) becomes

$$Y(t) = \widetilde{\boldsymbol{\mu}} + \widetilde{\Theta} X(t) + \widetilde{E}(t), \tag{C.3}$$

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 (C.1) by $P_{\Delta^\perp}$, the unobserved term disappears. Specifically, letting $\widetilde{Y}(t) = P_{\Delta^\perp}Y(t)$, (C.1) becomes

$$\widetilde{Y}(t) = P_{\Delta^\perp}\boldsymbol{\mu} + \Theta X(t) + P_{\Delta^\perp}E(t). \tag{C.4}$$

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 (C.3) and then implement *heteroPCA* algorithm [Zhang et al., 2019] 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. [2020] for details about the method.

C.2 Proof of Main Theorems

Since our focus is on the estimation error for $\boldsymbol{\beta}_i$, we consider the perturbation model in (4.10) 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 (4.14). 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 (4.14) 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 $\boldsymbol{x}(t)$ for $\boldsymbol{x}_i(t)$, $\boldsymbol{\theta}$ 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 $\widetilde{S}_i$.

Proof of Theorem 4.1 : While the skeleton of the proof follows from Ćevid et al. [2020, Theorem 2], the following two conditions are needed because of the Hawkes process data's unique dependency structure.

Condition C.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 C.1 is referred as the restrict strong convexity (RSC) [Negahban and Wainwright, 2010]. Lemma C.2 by Wang et al. [2020b] has shown Condition C.1 holds when $\widetilde{X} = X$ under Assumption 4.1- 4.4. Since the min eigenvalue of $\widetilde{X}$ stays the same with our choice of F, Condition C.1 holds for $\widetilde{X} = FX$.

Condition C.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* (4.10).

Condition C.2 holds as a result of Lemma C.1 by van de Geer [1995].

Under the two conditions, we achieve the conclusion as follows.

Because $\widehat{\boldsymbol{\theta}}$ is the optimizer for (4.14),

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

Under Condition C.2,

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

with probability at least $1-c_1p\exp(-T^{1/5})$, 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 $\boldsymbol{\theta}_{S^c}=\begin{pmatrix}u & \boldsymbol{\beta}_{S^c}\end{pmatrix}^\top$,

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

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

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

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

The above implies

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

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

Taking $\lambda = 2\psi$,

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

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

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

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

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

Combining the two cases, we always have

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

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

$$\left\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right\|_1 \leq C_1\Lambda_{\max}^2(F)\frac{s^*}{\gamma_{\min}^2}T^{-2/5}+C_2T^{-3/5}\Lambda_{\max}^{-2}(F)\left\|\widetilde{X}\boldsymbol{b}\right\|_2^2.$$

Proof of Corollary 4.1 : Notice that

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

with probability at least $1 - c_1 p^2 T \exp(-c_2 T^{1/5})$, where the second inequality is by Lemma C.2.

Then, Corollary 4.1 is a direct result from Theorem 4.1 by plugging in $\|b\|_2^2$.

Proof of Theorem 4.2: Recall $S = \{\beta_{ij} : \beta_{ij} \neq 0, 1 \leq i, j \leq p\}$ and $S_C = \{\beta_{ij} : \beta_{ij} = 0, 1 \leq i, j \leq p\}$. 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 4.5, we have $\beta_{ij} \in S \geq 2\tau$. Theorem 4.1 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 4.1, 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.

C.3 Supporting Lemmas

Lemma C.1 (van de Geer [1995]). *Suppose there exists* $\lambda_{\max}$ *such that* $\lambda(t) \leq \lambda_{\max}$ *where* $\lambda(t)$ *is the intensity function of Hawkes process defined in* (4.2). *Let* $H(t)$ *be a bounded function that is* $\mathcal{H}_t$*-predictable. Then, for any* $\epsilon > 0$,

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

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

Lemma C.2 (Wang et al. [2020b]). *Suppose the Hawkes process defined in* (4.2) *satisfies Assumptions 4.1– 4.4. Let* $Q = \frac{1}{T}\int_0^T \begin{pmatrix} 1 \\ \boldsymbol{x}(t) \end{pmatrix}\begin{pmatrix} 1 & \boldsymbol{x}^\top(t) \end{pmatrix} dt$, *where* $\boldsymbol{x}(t)$ *is defined in* (4.5). *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.*

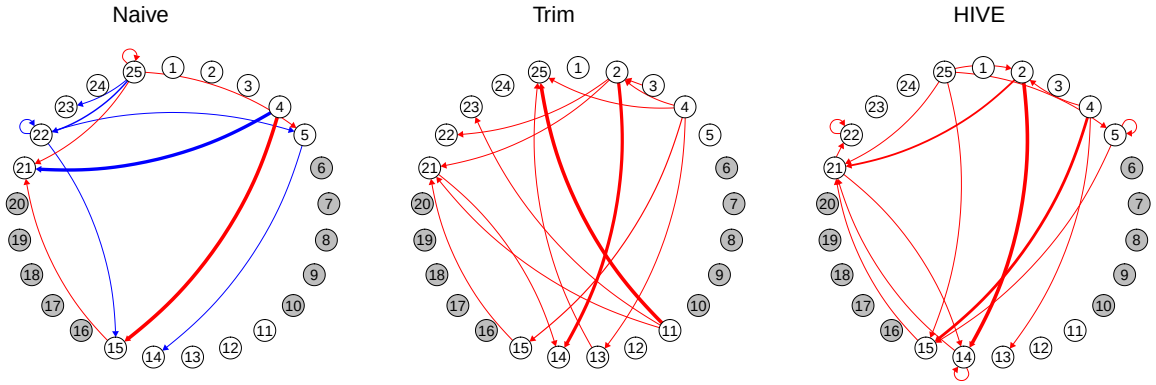


Figure C.1: Estimated functional connectivities among a subset of neuronal populations (removed neurons in gray) using the spike train data under laser from Bolding and Franks [2018]. Edges estimated by the three methods that are the same as those estimated using all neurons are in red and the other edges are in blue. Thicker edges indicate estimated connectivity coefficients of larger magnitudes.

C.4 Additional Data Analysis

We estimate the functional connectivities among a subset of neurons (15 out of 25) using the spike train data under 20 mW/mm^2 from Bolding and Franks [2018], separately using our method—trim regression, HIVE and the naive approach. The estimated connectivities among the subset of neurons by our method only using the subset neurons is the same as that using all neurons (see Figure C.1). Such estimation stability is also observed using the HIVE method but not in that given by the Naive method in this data set.