UW Biostatistics Working Paper Series

Winter 12-20-2019

Statistical Inference for Networks of High-Dimensional Point Processes

Xu Wang

Mladen Kolar

Ali Shojaie

Statistical Inference for Networks of High-Dimensional Point Processes

Xu Wang, Mladen Kolar & Ali Shojaie

January 14, 2020

Abstract

Fueled in part by recent applications in neuroscience, high-dimensional Hawkes process have become a popular tool for modeling the network of interactions among multivariate point process data. While evaluating the uncertainty of the network estimates is critical in scientific applications, existing methodological and theoretical work have only focused on estimation. To bridge this gap, this paper proposes a high-dimensional statistical inference procedure with theoretical guarantees for multivariate Hawkes process. Key to this inference procedure is a new concentration inequality on the first- and second-order statistics for integrated stochastic processes, which summarizes the entire history of the process. We apply this concentration inequality, combining a recent result on martingale central limit theory, to give an upper bounds for the convergence rate of the test statistics. We verify our theoretical results with extensive simulation and an application to a neuron spike train data set.

Keyword: Hawkes process; high dimensional inference; hypothesis testing; confidence intervals.

1 Introduction

Multivariate point process data have become prevalent in many applications 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). These processes can be represented by a graphical model $G = (V, E)$ (Lauritzen, 1996), where each node $v \in V$ represents a component of the multivariate point process, and each *directed* edge, $(u \to v) \in E$, indicates that the history of the source node u influences the probability of future events of the target node v. Multivariate point process data provide opportunity to learn the latent connectivity structure of this network.

In his seminal paper, Hawkes (1971) proposed a class of point process models, in which the probability of future events of each component can be influenced by the entire history of past events of others. The multivariate Hawkes process model has become a popular tool for studying the connectivity structure of the network because of its flexibility and interpretability in modeling the dependence structure between different point processes. From its early application in earthquake prediction (Ogata, 1988), this model has been widely applied to learn the latent connectivity structure in many fields, including neuroscience (Chen et al., 2017), social media (Zhou et al., 2013), and finance (Linderman and Adams, 2014).

In the original model (Hawkes, 1971) and later theoretical developments (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) the Hawkes process is considered as a *mutually-exciting* process, in which an event can only excite the process. In other words, each event of each component may trigger future events of all point processes 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. Recently, Costa et al. (2018) and Chen et al. (2017) have considered a broader class of the Hawkes processes that allow for both excitatory and inhibition effects in single and multivariate Hawkes process, respectively.

In modern applications, learning the connectivity network of multivariate point processes often poses additional challenges due to high-dimensionality. This is because the number of components measured, i.e. the number of neurons, is often large compared to the observed time period, i.e. the duration of neuroscience experiments. Recent work by Hansen et al. (2015) and Chen et al. (2017) has addressed this challenge by proposing ℓ_1-regularized estimation procedures. However, existing procedures for learning networks of multivariate point processes do not provide measures of uncertainty, which are critical in scientific applications.

Recent literature on statistical inference in high dimensions (e.g., Javanmard and Montanari, 2013; Zhang and Zhang, 2014; van de Geer et al., 2014; Ning and Liu, 2017) provides the ideal starting point for developing inference procedures for multivariate point processes. However, the vast majority of existing approaches consider the setting of independent data. Therefore, these results can not be applied to time series settings. One notable exception is the recent work by Zheng and Raskutti (2018), which develops a statistical inference procedure for high-dimensional vector auto-regressive (VAR) models. However, while a significant step forward, the VAR model captures the past history for a fixed and pre-specified time lag (or order). In contrast, the Hawkes process is dependent on the *entire* past history. Therefore, developing a high-dimensional inference procedure for the multivariate Hawkes process introduces significant additional challenges. In particular, this dependence on the entire history complicates the proof of convergence of the test statistic for the multivariate Hawkes process.

In this paper, we provides the first high-dimensional inference procedure for multivariate Hawkes processes with both excitatory and inhibitory effects. To this end, we adopt the decorrelated score test framework of Ning and Liu (2017) to high-dimensional point processes. To overcome the theoretical challenges stemming from the dependence structure, we develop a new concentration inequality on the first- and second-order statistics of the multivariate Hawkes process. Importantly, unlike previous results by Costa et al. (2018) and Chen et al. (2017), our results apply to integrated stochastic processes that summarize the entire history of each component. Such processes are necessary for developing high-dimensional inference procedures. Thus, instead of using the *thinning process representation* (Brémaud and Massoulié, 1996) or a coupling construction technique (Chen et al., 2017), our proof is based on a careful investigation on the transition structure of the Hawkes process. We combine this concentration inequality with the recent martingale central limit theorem of Zheng and Raskutti (2018) to obtain an upper bound for the convergence rate of our test statistics. We also provide confidence intervals for the model parameters by extending the semi-parametric efficient confidence region of Zheng and Raskutti (2018) for VAR models to the setting of Hawkes processes.

The rest of this paper is organized as follows. Section 2 introduces the linear Hawkes process and reviews its basic properties. Our hypothesis testing and the inference procedure is outlined in Section 3. In Section 4, we present theoretical results that guarantee the weak convergence of our test statistics under null and alternative hypothesis. The construction of confidence intervals and their theoretical justification is presented in Section 5. We investigate the properties of the proposed estimator using simulations in Section 6 and illustrate its utility in neuroscience applications in Section 7. Proofs of the main theorems are given in Section 9 and we conclude with a short discussion in Section 8. Proofs of technical lemmas are given in the Appendix.

2 The Linear Hawkes Process

Let $\mathcal{B}(\mathbb{R})$ denote the Borel σ-field of the real line, and let $\{t_k\}_{k\in\mathbb{Z}}$ defined in range $[0,T]$ be a sequence of real-valued random variables such that $t_{k+1} > t_k$ and $t_1 \geq 0$. Here, time $t = 0$ is a reference point in time, e.g., the start of an experiment. For $A \in \mathcal{B}(\mathbb{R})$, we define a simple point process N on $\mathbb{R}$ as a family $\{N(A)\}_{A\in\mathcal{B}(\mathbb{R})}$ that takes on non-negative integer values such that the sequence $\{t_k\}_{k\in\mathbb{Z}}$ consists of event times of the process N, i.e., $N(A) = \sum_k \mathbf{1}_{t_k\in A}$. In this consruction, the process N is essentially a simple counting process with isolated jumps of unit height which 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 $\mathbf{N} \equiv \{N_i\}_{i\in\{1,\ldots,p\}}$, where, as above, N_i satisfies $N_i(A) = \sum_k \mathbf{1}_{t_{ik}\in A}$ for $A \in \mathcal{B}(\mathbb{R})$ and $\{t_{i1}, t_{i2}, \ldots\}$ denote the event times of N_i. Let $\mathcal{H}_t$ be the history of $\mathbf{N}$ prior to time t. The intensity process $\{\lambda_1(t), \ldots, \lambda_p(t)\}$ is a p-variate $\mathcal{H}_t$-predictable process, defined as

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

Hawkes (1971) proposed a class of point process models in which past events can affect the probability of future events. This process is called the *linear Hawkes model*, if the intensity function for unit i takes the form

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

where

$$\left(\omega_{ij} * \frac{dN_j}{dt} \right)(t) = \int_0^{\infty} \omega_{ij}(\Delta) * dN_j(t - \Delta) = \sum_{k:t_{jk} \leq t} \omega_{ij}(t - t_{jk}). \tag{3}$$

Here, μ_i is the background intensity of unit i, and $\omega_{i,j}(\cdot) : \mathbb{R}^+ \to \mathbb{R}$ is the *transfer function*, where $\omega_{i,j}(t - t_{jk})$ represents the influence from the kth event of unit i 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 $\omega_{i,j}(\cdot)$ such that

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

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

Here, the *transition kernel* $k_j(\cdot) : \mathbb{R}^+ \to \mathbb{R}$ represents the decay of the influence of a past event. A commonly used example is the exponential transition kernel, $k_j(t) = e^{-t}$, considered by Bacry et al. (2015). In this formulation, β_{ij} represents the strength of the influence of unit j's past event on the intensity of unit i. A positive β_{ij}, which implies that the past events of one unit excites future events of another, is often considered in literature (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 unit (Chen et al., 2017). For example, in neuroscience, it is well known that a spike of one neuron may inhibit the activities of other neurons (Babington, 2001).

Denoting $x(t) = (x_1(t), \ldots, x_p(t)) \in \mathbb{R}^{1 \times p}$ and $\beta_i = (\beta_{i1}, \ldots, \beta_{ip})^\top \in \mathbb{R}^{p \times 1}$, we can write

$$\lambda_i(t) = \mu_i + x(t)\beta_i, \tag{6}$$

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

$$Y_i(t) = \mu_i + x(t)\beta_i + \epsilon_i(t). \tag{7}$$

As we will discuss later, a key challenge in this 'linear model', stems from heteroscedasticity, i.e., the fact that

$$\sigma_i^2(t) \equiv Var\left(\epsilon_i(t) | \mathcal{H}_t\right) = \lambda_i(t)(1 - \lambda_i(t)) \tag{8}$$

may not necessarily be 1 and depends on $x(t)$.

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 β_{ij} are constant over the time range $[0, T]$ (Brémaud and Massoulié, 1996; Daley and Vere-Jones, 2003).

3 Testing

We consider testing a d-dimensional subset of $\{\beta_{ij}\}_{1\le j\le p}$; that is, $\beta_{iJ} = \{\beta_{ij}, j \in J \subset \{1,\ldots,p\}\}$ and $\|J\|_0 = d$:

$$H_0 : \beta_{ij} = 0, j \in J. \tag{9}$$

For ease of notation, we primarily focus on the case of 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 $d \ge 1$.

Since the variance of noise $\sigma_i^2(t)$ defined in (8) is not necessarily one, for convenience we scale the columns $x(t)$. More specifically, let $\widetilde{x}_j(t) = x_j(t)/\sigma_i(t)$ for $j = 1,\ldots,p$. Before defining the test statistics, we first define the orthogonal projection of $\widetilde{x}_j(t)$ onto $\widetilde{x}_{-j}(t)$, where $\widetilde{x}_{-j}(t) = (\widetilde{x}_1(t),\ldots,\widetilde{x}_{j-1}(t),\widetilde{x}_{j+1}(t),\ldots,\widetilde{x}_p(t))$. Let the projection coefficient $w_j^* = \begin{pmatrix} w_{j0}^* & w_{j,-j}^* \end{pmatrix} \in \mathbb{R}^p$ be such that

$$\widetilde{x}_j(t) - \left(1, \widetilde{x}_{-j}(t)\right)^\top w_j^* \perp \widetilde{x}_{-j}(t) \tag{10}$$

Denoting the orthogonal complement of $\widetilde{x}_j(t)$ after removing its projection onto $\widetilde{x}_{-j}(t)$ as

$$\widetilde{x}_j^*(t) \equiv \widetilde{x}_j(t) - \left(1, \widetilde{x}_{-j}(t)\right)^\top w_j^*, \tag{11}$$

we have

$$\mathbb{E}\left[\widetilde{x}_j^*(t)\right] = 0. \tag{12}$$

Using this construction, our de-correlated score statistic is defined as

$$S_{ij} = \frac{1}{T}\sum_{t=1}^{T} \widetilde{\epsilon}_i(t)\, \widetilde{x}_j^*(t). \tag{13}$$

Remark 3.1: The reason we construct the de-correlated score statistics based on $\widetilde{x}_j^*(t)$, rather than defining it directly based on $\widetilde{x}_j(t)$, is that we do not know the true value of the nuisance parameters, μ_i and $\beta_{i,-j}$, and use estimates of these nuisance parameters. The construction of the de-correlated score statistics helps make the error induced by the estimation of the nuisance parameter asymptotically negligible; see Ning and Liu (2017) for more details.

Remark 3.2: Zheng and Raskutti (2018) also considers a similar de-correlated score statistics but in VAR model setting under an assumption of unit variance noise over the entire time range. The difference between our score statistics and theirs is that our score takes into account the variance of noise when constructing the score statistics due to the fact that the variance of noise is not the same at each t. In addition, the variance of noise depends on the intensity value which is time varying, which makes the technical proof more challenging in our case.

Now, let

$$\Upsilon_j = Cov\left(\widetilde{\epsilon}_i(t)\,\widetilde{x}^*_j(t)\right) = \mathbb{E}\left(\left(\widetilde{x}^*_j(t)\right)^2\right), \tag{14}$$

$$V_T = \sqrt{T}\,\Upsilon_j^{-1/2} S_{ij}, \tag{15}$$

$$U_T = \|V_T\|_2^2. \tag{16}$$

Note that in the simple case testing a univariate β_{ij}, Υ_j is a scalar. When testing multiple parameters β_{iJ}, $\Upsilon_J = Cov\left(\widetilde{\epsilon}_i(t)\,\widetilde{x}^*_J(t)\right)$ is defined similarly, but now Υ_J is a $d \times d$ matrix since $\widetilde{x}^*_J(t)$ is now a d-dimension vector.

In practice, we do not know β_i and w^*_j, so we estimate them as follows.

- **Step 1**: Calculating $\widehat{\mu}_i$, $\widehat{\beta}_i$, and $\widehat{\sigma}^2_i(t)$: estimate $\widehat{\beta}_i$ using lasso regression with the original unscaled data $(Y_i(t), x(t))$. More specifically,

  $$\widehat{\mu}_i, \widehat{\beta}_i = \arg\min_{\mu_i \in \mathbb{R}, \beta \in \mathbb{R}^p} \frac{1}{T}\sum_{t=1}^{T}(Y_i(t) - \mu_i - x(t)\beta_i)^2 + \lambda\|\beta_i\|_1. \tag{17}$$

  Then,

  $$\widehat{\lambda}_i(t) = x(t)\widehat{\beta}_i \tag{18}$$

  $$\widehat{\sigma}^2_i(t) = \widehat{\lambda}_i(t)(1 - \widehat{\lambda}_i(t)). \tag{19}$$

  As shown in Lemma 1.5, $\widehat{\mu}_i, \widehat{\beta}_i, \widehat{\sigma}^2_i(t)$ are consistent for $\mu_i, \beta_i, \sigma^2_i(t)$. This follows from the prediction consistency of lasso. We also sho that the *restricted eigenvalue condition* (REC) required for the consistency of lasso (Bickel et al., 2009) is met in our case. This follows from the bounded eigenvalue of the cross-covariance matrix of the design matrix $\{x(t)\}_{1 \le t \le T}$, which is obtained using the assumptions made in the next section.

- **Step 2**: Calculating $\widehat{w}_j$ based on $\widehat{\widetilde{x}}_j(t) = \frac{x_j(t)}{\widehat{\sigma}_i(t)}$: $\widehat{w}_j$ is estimated using a lasso regression with outcome $\widehat{\widetilde{x}}_j$ and design matrix $\widehat{\widetilde{x}}_{-j}$:

  $$\widehat{w}_j = \arg\min_{w \in \mathbb{R}^p} \frac{1}{T}\sum_{t=1}^{T}\left(\widehat{\widetilde{x}}_j(t) - \begin{pmatrix}1 & \widehat{\widetilde{x}}_{-j}(t)\end{pmatrix} w\right)^2 + \lambda\|w\|_1. \tag{20}$$

  Using a similar lasso proof for estimation/prediction consistency, we can show the consistency of $\widehat{w}_j$ for w^*_j. Similar to Step 1, the REC condition for lasso is also met by the bounded eigenvalue of the cross-covariance matrix induced by the assumptions made in the next section.

- **Step 3**: Calculating $\widehat{\Upsilon}_j$: let $\widehat{\widetilde{x}}^*_j(t) = \widehat{\widetilde{x}}_j(t) - \begin{pmatrix}1 & \widehat{\widetilde{x}}_{-j}(t)\end{pmatrix}\widehat{w}_j$. Then, $\widehat{\Upsilon}_j$ is estimated by the sample covariance

  $$\widehat{\Upsilon}_j = \frac{1}{T}\sum_{t=1}^{T}\left(\widehat{\widetilde{x}}^*_j(t)\right)^2. \tag{21}$$

Note that when testing a univariate β_{ij} in (9), $\widehat{\Upsilon}_j$ is a scalar; however, when testing multiple $\beta_{iJ} = \{\beta_{ij}, j \in J\}$, $\widehat{\Upsilon}_J = \frac{1}{T}\sum_{t=1}^{T} \left(\widehat{\widetilde{x}}^*_J(t)\right)^\top \widehat{\widetilde{x}}^*_J(t)$. Our results are valid for this case as long as $d \equiv \|J\|_0 \ll p$.

- **Step 4**:

$$\widehat{\epsilon}_i(t) = Y_i(t) - \widehat{\mu}_i - x_{-j}(t)\widehat{\beta}_{i,-j} \tag{22}$$

$$\widehat{S}_{ij} = \frac{1}{T}\sum_{t=1}^{T} \frac{\widehat{\epsilon}_i(t)}{\widehat{\sigma}_i}\widehat{\widetilde{x}}^*_j(t) \tag{23}$$

$$\widehat{V}_T = \sqrt{T}\widehat{\Upsilon}_j^{-1/2}\widehat{S}_{ij} \tag{24}$$

$$\widehat{U}_T = \|\widehat{V}_T\|_2^2. \tag{25}$$

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

Remark 3.3: Although here we use the lasso regression for $\mu_i, \beta_i, \sigma_i(t)$ and w_j, we may use other estimators to obtain consistent estimates of these parameters as long as they have the same order of prediction and estimation errors as the lasso regression.

4 Theoretical Guarantees

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, and let $A^\top$ denote its transpose. Define $\Theta = \{\beta_{ij}\}_{1\le i,j\le p} \in \mathbb{R}^{p\times p}$ and $\mu = \{\mu_i\}_{1\le i\le p} \in \mathbb{R}^p$.

Assumption 1 Let Ω be a $p \times p$ matrix whose entries are $\Omega_{j,k} = \alpha\int_0^\infty |\omega_{j,k}(\Delta)|d\Delta$, for $1 \le j,k \le p$. Then, there exists a generic constant γ_Ω such that $\Lambda_{\max}(\Omega^T\Omega) \le \gamma_\Omega^2 < 1$.

This assumption is the same as Assumption 1 in Chen et al. (2017), and is a necessary requirement for a stationary Hawkes process. The constant γ_Ω does not depend on the dimension p. For any fixed p, Brémaud and Massoulié (1996) shows that the intensity process of the form (2) is stable in distribution, and thus a stationary process $\mathbf{N}$ exists given this assumption. 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 There exists a constant ρ_Ω such that

$$\max_{1\le i\le p}\left\{\sum_{l=1}^{p}\Omega_{il}, \sum_{l=1}^{p}\Omega_{li}\right\} \le \rho_\Omega < \infty.$$

Assumption 2, which was also considered in Basu and Michailidis (2015) for VAR models, requires maximum in- and out- intensity flows. This assumption helps in bounding the eigenvalues of the cross-covariance of $x(t)$. As discussed by Chen et al. (2017), this assumption prevents the intensity from concentrating to a single process. Assumption 2 can be replaced by assumptions on the structure of Ω if the magnitude of each entry of Ω is upper bounded. Note that if we assume a stronger condition in Assumption 2 with $\rho_\Omega < 1$, then Assumption 1 is satisfied by the Perron-Frobenius theorem.

Assumption 3 There exists $\lambda_{\min}$ and $\lambda_{\max}$ such that

$$0 < \lambda_{\min} \leq \lambda_i(t) \leq \lambda_{\max} < 1$$

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

Assumption 3, which is similar to Assumption 4 in Chen et al. (2017), requires that intensity values are bounded between 0 and 1. This assumption prevents degenerate processes for all units.

Assumption 4: There exists $b > a > 0$ such that the transfer kernel function satisfies

$$0 < \max_{1 \leq j \leq p} k_j(t) \leq a \exp(-bt)$$

The lower bound, $\max_{1 \leq j \leq p} k_j(t) > 0$, is needed to avoid trivial transfer functions ($k_j(t) = 0$). We also need $\max_{1 \leq j \leq p} k_j(t) \leq a \exp(-bt)$ for $b > a$ in order to have an integrable transfer function and to control the total influence from the past. Assumption 4 implies the following properties of the transfer function.

- Assumptions 3 and 4 imply bounded β_i; that is, $\exists C_\beta$, $\|\beta_i\|_\infty \leq C_\beta < \infty$.
- Assumption 4 prevents unbounded influences from past:
$$\kappa \equiv \max_{i=1,\dots,p} \sum_{t=1}^{T} k_j(t) \leq \frac{a \exp(-b)}{1 - \exp(-b)} < \infty. \tag{26}$$
- Let ω_{ij}^{*n} be n-th auto-convolution of ω_{ij}. Under Assumption 4,
$$\omega_{ij}^{*n}(t) \leq \beta_{ij} a^n \frac{t^{(n-1)}}{(n-1)!} \exp(-bt).$$
For example, for $n = 2, 3$,
$$\omega_{ij}^{*2}(t) = \int_0^T \omega_{ij}(t-s)\omega_{ij}(s)ds \leq \beta_{ij}a^2 \int_0^t a \exp(-b(t-s))a \exp(-bs)ds = \beta_{ij}a^2 t \exp(-bt);$$
$$\omega_{ij}^{*3}(t) = \int_0^T \omega_{ij}^{*2}(t-s)\omega_{ij}(s)ds \leq \int_0^t \beta_{ij}a^2(t-s)\exp(-b(t-s))a\exp(-bs)ds = \beta_{ij}a^3\frac{t^2}{2}\exp(-bt).$$
- Finally, we obtain
$$\Psi_{ij}(t) = \sum_{n=1}^{\infty} \omega_{ij}^{*n}(t) \leq \beta_{ij} a \exp(-(b-a)t); \tag{27}$$
$$\xi \equiv \max_{1 \leq i \leq p} \sum_{j=1}^{p} \sum_{t=1}^{T} |\Psi_{ij}(t)| \leq \rho C_\beta \frac{a \exp(-(b-a))}{1 - \exp(-(b-a))} < \infty. \tag{28}$$

Let $s_j = \|w_j^*\|_0$ and $s = \max_{1 \le j \le p} s_j$; $\rho_i = \|\beta_i\|_0$ and $\rho = \max_{1 \le i \le p} \rho_i$. Then, for specific connectivity matrix structures, the sparsity of w_j^* follows from the sparsity of the connectivity matrix, Θ, similar to the case of VAR models (Zheng and Raskutti, 2018). In particular, for a stationary linear Hawkes process, we show that if the connectivity matrix is block diagonal, $s \le \rho + 1$ (see Lemma S.4 in the Appendix). In general, the relationship between sparsity of w_j^* and the sparsity of Θ is not straightforward, but, the sparsity of w_j^* depends on the sign and scale of the connectivity coefficients, as well as the transition kernel.

Using the above assumptions, we next state results on weak convergence of $\widehat{U}_T$ under the null hypothesis. For brevity, we define Π_0 as the feasible set of (Θ, μ), where Assumption 1-4 are satisfied under the null hypothesis.

Theorem 1. Suppose the linear Hawkes model defined in (2) satisfies Assumptions 1-4. Further, suppose $\widehat{\beta}_i$, $\widehat{w}_j$ and $\widehat{\Upsilon}_j$ are estimated by (17), (20) and (21). Let F_d be the cdf of χ^2-distribution with d degrees of freedom. Then, if $(\rho \vee s) \log p = o\left(\sqrt{T}\right)$ and $T > C$ for some constant C, under the null hypothesis in (9), $\widehat{U}_T$ defined in (25) satisfies

$$\sup_{(\Theta,\mu) \in \Pi_0, x \in \mathbb{R}} \left| \mathbb{P}(\widehat{U}_T \le x) - F_d(x) \right| \le \frac{C_1}{T^{1/8}} + C_2 \left(\frac{(\rho \vee s) \log p}{\sqrt{T}} \right)^{1/2} + \frac{C_3}{p^{C_4}}. \tag{29}$$

Theorem 1 shows that $\widehat{U}_T$ converges to χ_d^2 in distribution ($d = 1$ when testing univariate β_{ij}). This result is an extension of the result for the VAR model by Zheng and Raskutti (2018). Despite differences between the Hawkes process and the continuous VAR model discussed before, we obtain the same rate of convergence using the properties of the Hawkes process. Note that the difference between $\widehat{U}_T$ from $F_d(x)$ is dominated by $T^{-1/8}$ rather than $T^{-1/2}$. This difference from the standard CLT is due to the time dependence of the point process data (see Lemma 1.1).

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

$$H_a : \beta_{ij} = T^{-\phi} \Delta. \tag{30}$$

Let $\widetilde{\Delta} = \Upsilon_j^{1/2} \Delta$, where Υ_j is defined in (14) and Δ is set in (30). We define Π_a as the feasible set of (Θ, μ) such that Assumption 1-4 are satisfied under the alternative hypothesis.

Theorem 2. Suppose the linear Hawkes model defined in (2) satisfies Assumptions 1-4. Further, suppose $\widehat{\beta}_i$, $\widehat{w}_j$ and $\widehat{\Upsilon}_j$ are estimated by (17), (20) and (21). Let $F_{d, \|\widetilde{\Delta}\|_2^2}$ be the cdf of a non-central χ^2-distribution with d degrees of freedom and non-centrality parameter $\|\widetilde{\Delta}\|_2^2$. Then, if $(\rho \vee s) \log p = o\left(\sqrt{T}\right)$ and $T > C$ for some constant C, under the alternative hypothesis in (9), $\widehat{U}_T$ defined in (25) satisfies one of the following

If $\phi = \frac{1}{2}$,

$$\sup_{(\Theta,\mu) \in \Pi_a, x \in \mathbb{R}} \left| \mathbb{P}\left(\widehat{U}_T \le x\right) - F_{d, \|\widetilde{\Delta}\|_2^2}(x) \right| \le \frac{C_1}{T^{1/8}} + C_2 \left(\frac{(s \vee \rho) \log p}{\sqrt{T}} \right)^{1/2} + \frac{C_3}{p^{C_4}}; \tag{31}$$

If $0 < \phi < \frac{1}{2}$,

$$\sup_{(\Theta,\mu)\in\Pi_a, x\in\mathbb{R}} \left|\mathbb{P}\left(\widehat{U}_T \le x\right)\right| \le \frac{C_1}{T^{1/8}} + \frac{C_2}{p^{C_3}} + C_4 \exp\left\{-C_5 T^{\frac{1}{2}-\phi} + C_6\sqrt{x}\right\}; \tag{32}$$

If $\phi > \frac{1}{2}$,

$$\sup_{(\Theta,\mu)\in\Pi_a, x\in\mathbb{R}} \left|\mathbb{P}\left(\widehat{U}_T \le x\right) - F_d(x)\right| \le \frac{C_1}{T^{1/8}} + C_2\left(\frac{(s\vee\rho)\log p}{\sqrt{T}}\right)^{1/2} + \frac{C_3}{p^{C_4}} + C_3 T^{\frac{1-2\phi}{3}}. \tag{33}$$

Theorem 2 establishes the asymptotic distribution of $\widehat{U}_T$ under the alternative hypothesis. Here, depending on the scaling of β_{ij} with respect to T, i.e. 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; finally, when $\phi = 1/2$, $\widehat{U}_T$ converges to a non-central χ^2-distribution with d degree of freedom and non-centrality parameter $\|\widetilde{\Delta}\|_2^2$. This result is an extension of Theorem 3.2 in Zheng and Raskutti (2018) to the linear Hawkes model. As before, in spite of differences between the Hawkes process and the VAR model, we obtain the same rate of convergence using the properties of the Hawkes process.

5 Confidence Regions

In this section, we construct confidence intervals for β_{ij}. Similar to Ning and Liu (2017), our confidence interval is based on the one-step estimator of β_{ij} for the de-correlated score function,

$$\widehat{b}_{ij} = \widehat{\beta}_{ij} - \left(\widetilde{\Upsilon}_j\right)^{-1}\widetilde{S}_j. \tag{34}$$

Here, $\widehat{\beta}_{ij}$ is the lasso estimator in (17),

$$\widetilde{\Upsilon}_j = \frac{1}{T}\sum_{t=1}^{T} \widehat{\widetilde{x}}_j^*(t)\widehat{\widetilde{x}}_j(t), \tag{35}$$

which follows the construction of Zheng and Raskutti (2018) and is constructed slightly differently from $\widehat{\Upsilon}_j$ for theoretical convenience. Let

$$\widetilde{S}_j = \frac{1}{T}\sum_{t=1}^{T} \frac{Y_i(t) - \widehat{\mu}_i - \widehat{\beta}_i x(t)}{\widehat{\sigma}_i(t)}\widehat{\widetilde{x}}_j^*(t), \tag{36}$$

and note that $\widetilde{S}_j$ involves the entire $\widehat{\beta}_i$ instead of $\widehat{\beta}_{i,-j}$, which is different from $\widehat{S}_j$.

Next, let

$$\widehat{R}_T \equiv T(\widehat{b}_{ij} - \beta_{ij})^\top \widehat{\Upsilon}_j (\widehat{b}_{ij} - \beta_{ij}). \tag{37}$$

In the following, we show that $\widehat{R}_T$ converges weakly to $\widehat{U}_T$; therefore, we construct an asymptotically $1-\alpha$ confidence region for β_{ij} as

$$CR(\alpha) = \{\theta : T(\widehat{b}_{ij} - \theta)^\top \widehat{\Upsilon}_j (\widehat{b}_{ij} - \theta) \le \chi^2_d(1-\alpha)\}. \tag{38}$$

Theorem 3. Suppose the linear Hawkes model from (2) satisfies Assumptions 1-4. Further, suppose $\widehat{\beta}_i$, $\widehat{w}_j$ and $\widehat{\Upsilon}_j$ are estimated by (17), (20) and (21). Then, there exists constants C, c_1, c_2 such that

$$\mathbb{P}\left(\left|\widehat{R}_T - \widehat{U}_T\right| > C\sqrt{\frac{s \vee \rho \log p}{T}}\right) \le c_1 \exp(-c_2 \log p). \tag{39}$$

As mentioned before, $\widehat{\beta}_i$ can be obtained from any consistent estimator with the same order of the estimation error as lasso defined in (17).

6 Simulation Studies

In this section, we verify our theoretical results and investigate the power and convergence properties of the proposed inference procedure. We consider the linear Hawkes model with the transfer function specified in (6). For the connectivity matrix $\Theta = \{\beta_{ij}\}_{1\le i,j\le p}$, we consider three structures: chain, block and random (Figure 1). The scale of non-zero β_{ij} is set to be 0.3 and the transfer kernel function $k_{ij}(t)$ is chosen to be $\exp(-t)$. This setting satisfies our assumptions of a stable Hawkes process.

To assess the performance of our method, we use it to 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 convergence of the 95% confidence intervals for zero and non-zero coefficients. We consider graphs of $p = 50$ units and experiments lengths $T \in \{200, 1000, 2000\}$. As a benchmark, we compare the performance of our test method against an oracle procedure, which knows what coefficients are non-zero.

Figure 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.

7 Application

In this section, 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

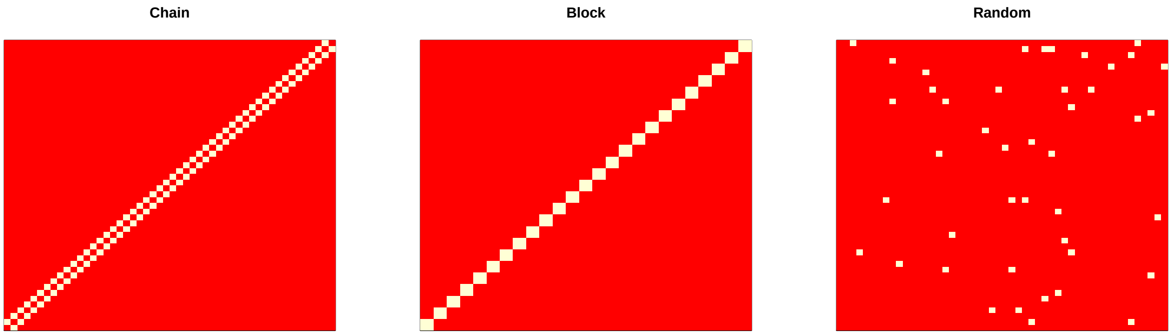

Figure 1: Connectivity matrix under chain, block and random graph structures (zero-value coefficients in red and non-zero coefficients in white).

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 (Condition 1) and 5 mW/mm^2 (Condition 2), of the subject mouse with the most neurons (29 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 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 29-neuron network.

Figure 3 illustrates the estimated connectivity coefficients in a graph representation, where each node represents a neuron and a directed edge indicates a statistically significant estimated connectivity coefficient. We see there are few common edges between the networks in the two conditions (gray edges); moreover, each condition has its own functional connectivity structures (red edge for Condition 1 and blue edge for Condition 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 3 also shows the 95% confidence interval for the estimated connectivity coefficients corresponding to the edges that are unique to each condition.

8 Discussion

In this paper, we proposed a statistical inference procedure with theoretical guarantees for high-dimensional linear Hawkes processes. To overcome the challenges from the the Hawkes process on its entire history, we develop a new concentration inequality on the first- and second-order statistics for an integrated stochastic process; these integrated processes summarize the entire history for each component. We combine this new concentration inequality with a recent martingale central limit theorem, to give an upper bounds for the conver-

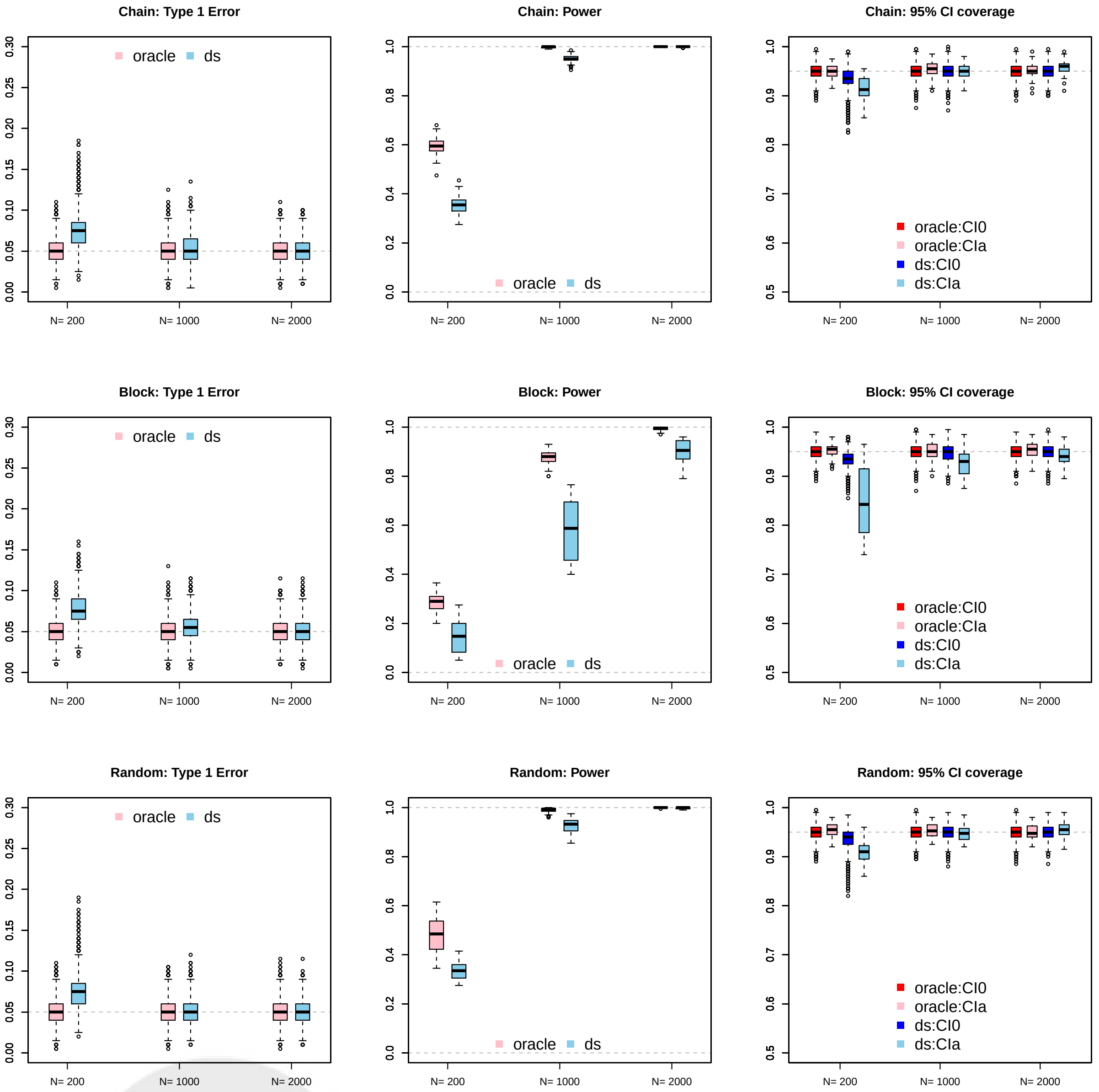

Figure 2: Simulation results under chain, block and random graph structure. CI0:95% confidence interval for zero-coefficients; CIa:95% confidence interval for non-zero coefficients. Oracle: score test under the true model with zero coefficients known; ds: de-correlated score test with nuisance coefficients.

gence rate of the test statistics. We also provide confidence intervals for the parameters as an extension of the semi-parametric efficient confidence region considered in Zheng and Raskutti (2018). Our results establish the first inferential framework for high-dimensional point processes.

In this paper, we consider a parametric transition function for the Hawkes process. Given the complex nature of the point process, one may consider nonparametric models for the

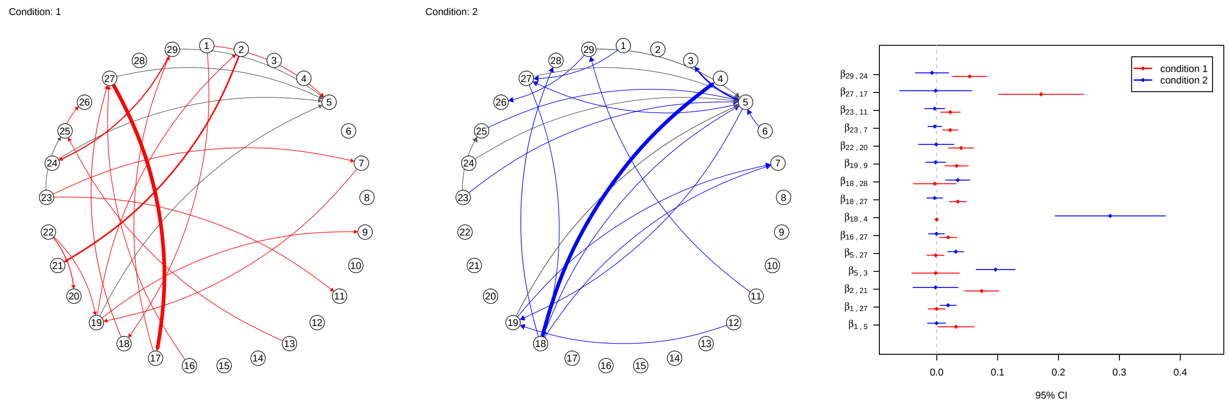

Figure 3: Estimated functional connectivities among neuronal populations from the spike train data. The gray edges are shared between two conditions of different pulse power level, the red edges are unique to condition 1 (0 mW/mm^2), and the blue edges are unique to condition 2 (5 mW/mm^2). 95% confidence intervals (CI) are shown for 15 unique edges corresponding to the edges that are unique to each condition and with largest estimated connectivity coefficients at one of the conditions.

transition functions and learn the form adaptively from the 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 statistical inference procedure for non-linear Hawkes processes.

9 Proof of Main Results

9.1 Proof of Theorem 1

First, to overcome the challenge of the unknown variance $\sigma_i^2(t)$ in $\widehat{U}_T$ as defined in (25), we introduce $\widehat{U}_T^0$ defined as

$$\widehat{U}_T^0 = \|\widehat{V}_T^0\|_2^2; \tag{40}$$

$$\widehat{V}_T^0 = \sqrt{T}\big(\widehat{\Upsilon}_j^0\big)^{-1/2}\widehat{S}_{ij}^0; \tag{41}$$

$$\widehat{\Upsilon}_j^0 = \frac{1}{T}\sum_{t=1}^{T}\big(x_j/\sigma_i^2(t) - \big(1 \quad x_{-j}/\sigma_i^2(t)\big)\,\widehat{w}_j\big)\big(x_j/\sigma_i^2(t) - \big(1 \quad x_{-j}/\sigma_i^2(t)\big)\,\widehat{w}_j\big)^\top\big(\widetilde{x}_j^*\big)^\top; \tag{42}$$

$$\widehat{S}_{ij}^0 = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{\sigma_i(t)}\big(Y_i(t) - \widehat{\mu}_i - x_{-j}(t)\widehat{\beta}_{i,-j}\big)\big(x_j(t)/\sigma_i(t) - \widehat{w}_{j0} - x_{-j}(t)/\sigma_i(t)\widehat{w}_{j,-j}\big). \tag{43}$$

As we see from the definition, the difference between $\widehat{U}_T^0$ and $\widehat{U}_T$ defined in (25) is that we replace $\widehat{\sigma}_i(t)$ by $\sigma_i(t)$.

We first prove the bound for $\widehat{U}_T^0 - U_T$ and then complete the proof using

$$\sup_{x\in\mathbb{R}} |P(\widehat{U}_T \leq x) - F_d(x)| \leq \sup_{y\in\mathbb{R}} |P(\widehat{U}_T^0 \leq y) - F_d(y)| \\ + F_d(x+\delta) - F_d(x-\delta) + P(|\widehat{U}_T - \widehat{U}_T^0| > \delta), \tag{44}$$

where we use Lemma 1.8 to bound $P\left(|\widehat{U}_T - \widehat{U}_T^0| > \delta\right)$ and use the properties of the χ^2 distribution to bound $|F_d(x+\delta) - F_d(x-\delta)| \leq C(d)\delta$ as in Zheng and Raskutti (2018).

Now we focus on bounding $\left|P\left(\widehat{U}_T^0 \leq x\right) - F_d(x)\right|$. The proof first links $P\left(\widehat{U}_T^0 \leq x\right)$ and $F_d(x)$ via $P(U_T \leq x)$ and then bounds the error among the three parts.

Note that $\forall \epsilon > 0$,

$$\begin{aligned} P\left(\widehat{U}_T^0 \leq x\right) - F_d(x) &\leq P(U_T \leq x+\epsilon) + P(\left|\widehat{U}_T^0 - U_T\right| > \epsilon) - F_d(x) \\ &\leq |P(U_T \leq x+\epsilon) - F_d(x+\epsilon)| + F_d(x+\epsilon) - F_d(x) + P(\left|\widehat{U}_T^0 - U_T\right| > \epsilon) \\ F_d(x) - P(\widehat{U}_T^0 \leq x) &= P(\widehat{U}_T^0 > x) - (1 - F_d(x)) \\ &\leq P(U_T > x-\epsilon) + P(\left|\widehat{U}_T^0 - U_T\right| > \epsilon) - (1 - F_d(x)) \\ &\leq |F_d(x-\epsilon) - P(U_T \leq x-\epsilon)| + F_d(x) - F_d(x-\epsilon) + P(\left|\widehat{U}_T^0 - U_T\right| > \epsilon). \end{aligned}$$

Then,

$$\sup |P(\widehat{U}_T^0 \leq x) - F_d(x)| \leq \underbrace{\sup_{y\in\mathbb{R}} |P(U_T \leq y) - F_d(y)|}_{A} \\ + \underbrace{F_d(x+\epsilon) - F_d(x-\epsilon)}_{B} + \underbrace{P(|\widehat{U}_T^0 - U_T| > \epsilon)}_{C} \tag{45}$$

First, we bound part A, i.e. $P(U_T \leq y) - F_d(y)$. The difference between this proof and ordinary proofs of weak convergence of a sample average is that here our data is time dependent. Therefore, instead of using an ordinary central limit theorem (CLT), we use a martingale CLT. This technique is also used for VAR models by Zheng and Raskutti (2018). The result is stated in the following lemma.

Lemma 1.1. Suppose the stationary linear Hawkes model from (6) satisfies Assumptions 1-4, and $\widehat{\beta}_i$, $\widehat{w}_j$ and $\widehat{\Upsilon}_j$ are estimated by (17), (20) and (21). Then, $\forall u \in \mathbb{R}^d$,

$$\sup_{y\in\mathbb{R}} |P(U_T + u \leq y) - F_{d,\|u\|_2^2}(y)| \leq C(\|u\|_2, d, s)T^{-1/8}, \tag{46}$$

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

The proof is based on a martingale difference sequence CLT, and extends the previous result for VAR models (Lemma 5.3 in Zheng and Raskutti (2018)) to the linear Hawkes model (2).

By Lemma 1.1, for $s, d \ll T$, we have

$$\sup_{y\in\mathbb{R}} |P(U_T \le y) - F_d(y)| \le C_2 T^{-1/8} \tag{47}$$

Next, we bound part B. Note that χ_d^2 distribution has bounded density and continuous and differentiable cdf. This implies that there exists a constant $C_3 > 0$ s.t.

$$|F_d(x+\epsilon) - F_d(x-\epsilon)| \le C_3\epsilon, \tag{48}$$

which gives a bound for part B.

Next, we check part C.

$$\begin{aligned}
|\widehat{U}_T^0 - U_T| &= \left| T\big(\widehat{S}_{ij}^0\big)^\top \big(\widehat{\Upsilon}_j^0\big)^{-1}\widehat{S}_{ij}^0 - S_{ij}^T\Upsilon_j^{-1}S_{ij}\right| \\
&\le \left| T\big(\widehat{S}_{ij}^0\big)^\top \big(\big(\widehat{\Upsilon}_j^0\big)^{-1} - \Upsilon_j^{-1}\big)\widehat{S}_{ij}^0 + T\big(\widehat{S}_{ij}^0\big)^\top \Upsilon_j^{-1}\widehat{S}_{ij} - S_{ij}^T\Upsilon_j^{-1}S_{ij}\right| \\
&\le \|\Upsilon_j^{1/2}\big(\widehat{\Upsilon}_j^0\big)^{-1}\Upsilon_j^{1/2} - I\|_\infty \|\sqrt{T}\Upsilon_j^{-1/2}\widehat{S}_{ij}^0\|_1^2 \\
&+ \|\sqrt{T}\Upsilon_j^{-1/2}(S_{ij} - \widehat{S}_{ij}^0)\|_1^2 \\
&+ 2\|V_T\|_2\|\sqrt{T}\Upsilon_j^{-1/2}(S_{ij} - \widehat{S}_{ij}^0)\|_2
\end{aligned} \tag{49}$$

Let $E = \sqrt{T}\Upsilon_j^{-1/2}(S_{ij} - \widehat{S}_{ij}^0)$, then

$$|\widehat{U}_T^0 - U_T| \le \|E\|_2^2 + 2\|V_T\|_2\|E\|_2 + \|\Upsilon_j^{1/2}\widehat{\Upsilon}_j^{-1}\Upsilon_j^{1/2} - I\|_\infty(\|V_T\|_2 + \|E\|_2)^2 \tag{50}$$

First, we bound $\|E\|_2^2$ using the following lemma.

Lemma 1.2. Suppose the stationary linear Hawkes model from (6) satisfies Assumptions 1-4, and let $\Upsilon_x = Cov(x(t))$ and Υ_j be defined in (14). Then,

$$0 < C_1(\beta) \le \Lambda_{\min}\big(\Upsilon_x\big) \le \Lambda_{\max}\big(\Upsilon_x\big) \le C_2(\beta) < \infty \tag{51}$$

$$c_1\Lambda_{\min}\big(\Upsilon_x\big) \le \Lambda_{\min}\big(\Upsilon_j\big) \le \Lambda_{\max}\big(\Upsilon_j\big) \le c_2\Lambda_{\max}\big(\Upsilon_x\big). \tag{52}$$

The proof uses a result on eigenvalues of the cross-covariance matrix of the outcome under a covariance-stationary process given by Prop. 2.3 in Basu and Michailidis (2015) and a modified result linking the spectral density of the cross-covariance matrix and the covariance matrix of the outcome given by (Bacry et al., 2011) or Theorem 3 in (Etesami et al., 2016).

By Lemma 1.2, there exists constant C s.t.

$$\|E\|_2 \le C\sqrt{T}\left\|\widehat{S}_{ij}^0 - S_{ij}\right\|_2. \tag{53}$$

Next, we bound $\left\|\widehat{S}^0_{ij} - S_{ij}\right\|_2$.

$$\begin{aligned}\widehat{S}^0_{ij} - S_{ij} &= (\widehat{w}_j - w^*_j)^\top \frac{1}{T}\sum_{t=1}^{T} \widetilde{\epsilon}_i(t) \begin{pmatrix}1\\ \widetilde{x}_{-j}(t)\end{pmatrix} \\ &+ \frac{1}{T}\sum_{t=1}^{T} \widetilde{x}^*_j(t) \begin{pmatrix}1 & \widetilde{x}_{-j}(t)\end{pmatrix} \left(\begin{pmatrix}\hat{\mu}_i\\ \widehat{\beta}_{i,-j}\end{pmatrix} - \begin{pmatrix}\mu_i\\ \beta_{i,-j}\end{pmatrix}\right) \\ &- (\widehat{w}_j - w^*_j)^\top \left(\frac{1}{T}\sum_{t=1}^{T} \begin{pmatrix}1\\ \widetilde{x}^\top_{-j}(t)\end{pmatrix}\begin{pmatrix}1 & \widetilde{x}^\top_{-j}(t)\end{pmatrix}\right)\left(\begin{pmatrix}\hat{\mu}_i\\ \widehat{\beta}_{i,-j}\end{pmatrix} - \begin{pmatrix}\mu_i\\ \beta_{i,-j}\end{pmatrix}\right). \end{aligned} \tag{54}$$

Next, we introduce two lemmas to bound the terms $\frac{1}{T}\sum_{t=0}^{T-1} \widetilde{x}^*_j(t) \begin{pmatrix}1\\ \widetilde{x}_{-j}(t)\end{pmatrix}$ and $\frac{1}{T}\sum_{t=1}^{T} \widetilde{\epsilon}_i(t) \begin{pmatrix}1\\ \widetilde{x}_{-j}(t)\end{pmatrix}$.

Lemma 1.3. Under Assumption 1-4 and the linear Hawkes model (2), for $T > C\log p$, for some constant $C > 0$,

$$P\left(\|\frac{1}{T}\sum_{t=0}^{T-1}(\widetilde{x}_j(t) - w^*_{j0} - \widetilde{x}_{-j}(t)w^*_j) \begin{pmatrix}1\\ \widetilde{x}_{-j}(t)\end{pmatrix}\|_\infty > C\sqrt{\frac{\log p}{T}}\right) \leq C_1\exp(-C_2\log p) \tag{55}$$

The proof essentially depends on deviation bounds for the first-order and quadratic forms of $\widetilde{x}(t)$, described in Lemma S.2 and Lemma S.3.

Lemma 1.4. Under Assumption 1-4 and the linear Hawkes model (2) ,

$$P\left(\|\frac{1}{T}\sum_{t=1}^{T}\widetilde{\epsilon}_i(t) \begin{pmatrix}1\\ \widetilde{x}_{-j}(t)\end{pmatrix}\|_\infty > C\sqrt{\frac{s\log p}{T}}\right) \leq C_1\exp(-C_2\log p). \tag{56}$$

The proof is a direct application of the martingale inequality given by van de Geer (1995) as $x_j(t), \sigma_i(t)$ are bounded under Assumption 3 and 4.

The next lemma helps us bound the estimation error of $\widehat{\beta}$ and $\widehat{w}$.

Lemma 1.5. Let $\theta_i = (\mu_i, \beta_i)$, $\rho_i = \|\beta_i\|_0$ and define $H = \frac{1}{T}\sum_{t=1}^{T} \begin{pmatrix}1\\ x(t)\end{pmatrix}\begin{pmatrix}1 & x(t)\end{pmatrix}$. Suppose the linear Hawkes model (2) satisfies Assumption 1-4 and $\hat{\theta}_i$ is given by (17). Then, for $\lambda \asymp \sqrt{\frac{\log p}{T}}$, when $T \geq C(\rho_i + 1)\log p$,

$$\begin{aligned} \|\hat{\theta}_i - \theta_i\|_2 &\leq C\sqrt{(\rho_i+1)\log p/T} \\ (\hat{\theta}_i - \theta_i)^\top H(\hat{\theta}_i - \theta_i) &\leq C(\rho_i+1)\log p/T \\ \|\hat{\theta}_i - \theta_i\|_1 &\leq C(\rho_i+1)\sqrt{\log p/T}. \end{aligned}$$

with probability at least $1 - c_1\exp(-c_2\log p)$.

The proof of this lemma follows typical proofs of estimation and prediction consistency of lasso. However, instead of assuming the restricted eigenvalue condition (REC) condition required for the consistency of lasso, we show that for a model satisfying our assumptions the REC condition is met. This leads to bounded eigenvalues of the covariance matrix as shown in Lemma 1.2. The introduction of ρ_i+1, instead of is ρ_i, is because by Assumption 3 $\mu_i > 0$.

Lemma 1.6. Let $s_j = \|w_j^*\|_0$ and let $H = \frac{1}{T}\sum_{t=1}^T \begin{pmatrix} 1 \\ \widehat{\widetilde{x}}_{-j}(t) \end{pmatrix} \begin{pmatrix} 1 & \widehat{\widetilde{x}}_{-j}(t) \end{pmatrix}$. Suppose the linear Hawkes model (2) satisfies Assumptions 1-4, and $\widehat{w}$ is given by (20). Then, for $\lambda \asymp \sqrt{\frac{s_i \log p}{T}}$,

$$\begin{aligned} \|\widehat{w}_j - w_j^*\|_2 &\leq C\sqrt{(\rho_i+1) \vee s_i \log p/T} \\ (\widehat{w}_j - w_j^*)^\top H(\widehat{w}_j - w_j^*) &\leq C(\rho_i+1) \vee s_i \log p/T \\ \|\widehat{w}_j - w_j^*\|_1 &\leq C\sqrt{(\rho_i+1) \vee s_i \log p/T}, \end{aligned}$$

with probability at least $1 - c_1 \exp(-c_2 \log p)$.

The proof is similar to that of Lemma 1.5. Since $\widehat{w}_j$ is based on a $\widehat{\widetilde{x}}(t)$ instead of $\widetilde{x}(t)$, the proof in addition needs to bound the difference $\widehat{\sigma}_i^2(t) - \sigma_i^2(t)$ as $O(\sqrt{\frac{(\rho_i+1)\log p}{T}})$ based on the estimation consistency by Lemma 1.5. That is why in the error bound we have both $\rho_i + 1$ and s_j.

By Assumptions 2 and 3, $x_j(t)$ is bounded for every $j = 1, \ldots, p$ and $t = 0, \ldots, T$. Therefore, by Lemmas 1.3-1.6,

$$\|\widehat{S}_{ij}^0 - S_{ij}\|_2 \leq C\frac{(s_i \vee \rho_i)\log p}{T} \tag{57}$$

Combining the above, $\|E\|_2^2 \leq C\frac{(s_i \vee \rho_i)\log p}{\sqrt{T}}$, with probability at least $1 - c_1\exp(-c_2 \log p)$.

Next, we obtain a bound for $\|V_T\|_2$. We can directly use the intermediate result in the proof in Theorem 3.1 of Zheng and Raskutti (2018) because the result only requires that Lemma 1.1 holds for V_T. Thus,

$$P(\|V_T\|_2 > y) \leq CT^{-1/8} + Cy^{-2} \tag{58}$$

To reach the final conclusion, we introduce the following lemma.

Lemma 1.7. Suppose the stationary linear Hawkes model defined in (6) satisfies Assumption 1-4 and $\widehat{\beta}$, $\widehat{w}$, and $\widehat{\Upsilon}_j$ are given by (17), (20) and (14). Then,

$$\|\Upsilon_j^{1/2}\widehat{\Upsilon}_j^{-1}\Upsilon_j^{1/2} - I\|_\infty \leq C\sqrt{\rho \log p/T}$$

and

$$\|\Upsilon_j^{1/2}(\widehat{\Upsilon}_j^0)^{-1}\Upsilon_j^{1/2} - I\|_\infty \leq C\sqrt{\rho \log p/T}$$

with probability at least $1 - c_1 \exp(-c_2 \log p)$, where

$$\widehat{\Upsilon}_j^0 = \frac{1}{T}\sum_{t=1}^T (x_j/\sigma_i^2(t) - \begin{pmatrix}1 & x_{-j}/\sigma_i^2(t)\end{pmatrix}\widehat{w}_j)(x_j/\sigma_i^2(t) - \begin{pmatrix}1 & x_{-j}/\sigma_i^2(t)\end{pmatrix}\widehat{w}_j)^\top (\widetilde{x}_j^*)^\top.$$

Due to the difficulty involving $\widehat{\sigma}_i^2(t)$ in the estimator, we first bound the difference between $\widehat{\Upsilon}_j^0$ (which replace $\widehat{\sigma}_i^2(t)$ by $\sigma_i^2(t)$ in $\widehat{\Upsilon}_j$) from Υ_j using the 1st-order and the 2nd-order (i.e the quadratic form) deviation bounds on $x(t)$ by Lemma S.3. Then, we bound the difference between $\widehat{\Upsilon}_j^0$ and $\widehat{\Upsilon}_j$ based on the consistency of $\widehat{\sigma}_i^2(t)$ to $\sigma_i^2(t)$.

Plugging in $y = \left(\frac{s_i \vee \rho_i \log p}{\sqrt{T}}\right)^{-1/4}$ into (58), we have

$$|\widehat{U}_T^0 - U_T| \leq C\left(\frac{s_i \vee \rho_i \log p}{\sqrt{T}}\right)^{1/2} \tag{59}$$

with probability at least $1 - c_1 \exp(-c_2 \log p) - c_3 T^{-1/8} - c_4\left(\frac{s_i \vee \rho_i \log p}{\sqrt{T}}\right)^{1/2}$.

Further, taking $\epsilon = C\left(\frac{s_i \vee \rho_i \log p}{\sqrt{T}}\right)^{1/2}$, we get

$$\sup_{x\in\mathbb{R}} |P(\widehat{U}_T^0 \leq x) - F_d(x)| \leq \frac{C_1}{T^{1/8}} + c_2\big(\frac{(s_i \vee \rho_i)\log p}{\sqrt{T}}\big)^{1/2} + \frac{C_3}{p^{C_4}}. \tag{60}$$

The next lemma helps us bound $|\widehat{U}_T - \widehat{U}_T^0|$.

Lemma 1.8. Suppose the linear Hawkes model defined in (2) satisfies Assumption 1 - 4. Then,

$$\left|\widehat{U}_T - \widehat{U}_T^0\right| \leq C\left(\frac{\rho \log p}{\sqrt{T}}\right)^{1/2}$$

with probability at least $1 - c_1 \exp(-c_2 \log p) - c_3 T^{-1/8} - c_4\left(\frac{(s\vee\rho)\log p}{\sqrt{T}}\right)^{1/2}$.

To bound $\left|\widehat{U}_T - \widehat{U}_T^0\right|$, it is enough to bound $\widehat{\Upsilon}_j - \widehat{\Upsilon}_j^0$ which can be bounded using the result in Lemma 1.7, and to bound $\widehat{S}_{ij} - \widehat{S}_{ij}^0$ where $\widehat{S}_{ij}^0$ defined in (43) is $\widehat{S}_{ij}$ with $\widehat{\sigma}_i^2(t)$ by $\sigma_i^2(t)$. Thus, in the main proof of Lemma 1.8, we show $\left\|\widehat{S}_{ij} - \widehat{S}_{ij}^0\right\|_2^2$ is bounded by $C\frac{\rho_i \log p}{T}$ with probability at least $1 - c_1 \exp(-c_2 \log p)$. To finish the proof, we use the weak convergence result for $\widehat{U}_T^0$ to χ_d^2 which is proofed in the above (60).

Then, following Lemma 1.8, $\delta = \left(\frac{\rho \log p}{\sqrt{T}}\right)^{1/2}$,

$$\begin{aligned}
&\sup_{x\in\mathbb{R}} |P(\widehat{U}_T \leq x) - F_d(x)| \\
&\leq \sup_{y\in\mathbb{R}} |P(\widehat{U}_T^0 \leq y) - F_d(y)| + F_d(x+\delta) - F_d(x-\delta) + P(|\widehat{U}_T - \widehat{U}_T^0| > \delta) \\
&\leq \frac{c_1}{T^{1/8}} + c_2\big(\frac{(s\vee\rho)\log p}{\sqrt{T}}\big)^{1/2} + \frac{c_3}{p^{c_4}} + \frac{c_5}{T^{1/8}} + c_6\big(\frac{\rho\log p}{\sqrt{T}}\big)^{1/2} + \frac{c_7}{p^{c_8}} \\
&\leq \frac{C_1}{T^{1/8}} + C_2\big(\frac{(s\vee\rho)\log p}{\sqrt{T}}\big)^{1/2} + \frac{C_3}{p^{C_4}}.
\end{aligned}$$

9.2 Proof of Theorem 2

As before, we start by bounding $\widehat{U}_T^0 - U_T$, where $\widehat{U}_T^0$ defined in (40).

First, consider $\phi = 1/2$. In this case, $\forall \epsilon > 0$,

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

Note that by Lemma 1.2 on the bounded eigenvalue of the covariance matrix Υ_j, $\|\widetilde{\Delta}\|_2^2 \le \Lambda_{\max}\Big(\Upsilon_j\Big)\|\Delta\|_2^2$. Then, for part A, applying Lemma 1.1 on the martingale CLT, we get

$$\sup_{y\in\mathbb{R}}\left|P(\|V_T - \widetilde{\Delta}\|_2^2 \le y) - F_{d,\|\widetilde{\Delta}\|_2^2}(y)\right| \le C\|\Delta\|_2 T^{-1/8}$$

For part B, we use an intermediate result from Zheng and Raskutti (2018),

$$: F_{d,\|\widetilde{\Delta}\|_2^2}(x+\epsilon) - F_{d,\|\widetilde{\Delta}\|_2^2}(x-\epsilon). \le C(d)\epsilon$$

The derivation of the result is purely based on the properties of non-central χ^2 distribution, thus is applicable in our case; see the proof of Theorem 3.2 in Zheng and Raskutti (2018).

Next, we bound part C. Let $E = \sqrt{T}\big(\Upsilon_j\big)^{-1/2}\widehat{S}_{ij} - V_T + \widetilde{\Delta}$. Then,

$$\left|\widehat{U}_T^0 - \left\|V_T - \widetilde{\Delta}\right\|_2^2\right| \le \|E\|_2^2 + \left\|V_T - \widetilde{\Delta}\right\|_2 \|E\|_2$$
$$+ \left\|\Big(\Upsilon_j\Big)^{1/2}\Big(\widehat{\Upsilon}_j^0\Big)^{-1}\Big(\Upsilon_j\Big)^{1/2} - I\right\|_\infty \Big(\left\|V_T - \widetilde{\Delta}\right\|_2 + \|E\|_2\Big)^2$$

Let $\widetilde{S}_{ij} = \frac{1}{T}\sum_{t=1}^T(\widetilde{Y}_i(t) - \mu_i - \widetilde{x}_{-j}(t)\beta_{i,-j})\widetilde{x}_j^*$, and define W_j^* such that $\widetilde{x}_j^*(t) = \begin{pmatrix}1 & \widetilde{x}(t)\end{pmatrix} W_j^*$. Let $\Upsilon = \mathbb{E}\left(\begin{pmatrix}1 & \widetilde{x}(t)\end{pmatrix}\begin{pmatrix}1 \\ \widetilde{x}(t)\end{pmatrix}\right)$. Then,

$$\Upsilon_j = Cov(\widetilde{x}_j^*(t)) = Cov\left(\begin{pmatrix}1 & \widetilde{x}(t)\end{pmatrix} W_j^*\right) = (W_j^*)^\top \Upsilon W_j^*.$$

Next,

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

Therefore, taking $\beta_{ij} = \sqrt{T}\Delta$, then by Lemma 1.2 of bounded eigenvalue Υ_j and Lemma S.1 of bounded w_j^* and S.3 of deviation bound for the 1st-order and the quadratic form of $\tilde{x}(t)$,

$$\begin{aligned}
\|E\|_2 &\le \left\|\sqrt{T}(\Upsilon_j)^{-1/2}(\widehat{S}^0_{ij} - \widetilde{S}_{ij})\right\|_2 + \left\|(\Upsilon_j)^{-1/2} W_j^* \left(\frac{1}{T}\sum_{t=1}^{T} \begin{pmatrix}1\\ \widetilde{x}(t)\end{pmatrix}\begin{pmatrix}1 & \widetilde{x}(t)\end{pmatrix} - \Upsilon_{\cdot,j}\right)\Delta\right\|_2 \\
&\le \sqrt{T}C\left\|\widehat{S}^0_{ij} - \widetilde{S}_{ij}\right\|_2 + C\sqrt{d}\|W_j^*\|_1 \left\|\frac{1}{T}\sum_{t=1}^{T} \begin{pmatrix}1\\ \widetilde{x}(t)\end{pmatrix}\begin{pmatrix}1 & \widetilde{x}(t)\end{pmatrix} - \Upsilon\right\|_\infty \\
&\le C\sqrt{T}\|\widehat{S}^0_{ij} - \widetilde{S}_{ij}\|_2 + C\sqrt{\frac{s\log p}{T}}.
\end{aligned}$$

Recall that d is the dimension of β_{ij}. When testing a univariate β_{ij} (9), $d = 1$. If $d > 1$, then $W_j^* \in \mathbb{R}^{d\times(p+1)}$. Now,

$$\begin{aligned}
\widehat{S}^0_{ij} - \widetilde{S}_{ij} =& (\widehat{w}_j - w_j^*)^\top \frac{1}{T}\sum_{t=1}^{T}\begin{pmatrix}1\\ \widetilde{x}_{-j}(t)\end{pmatrix}\left(\widetilde{\epsilon}_i(t) + T^{-1/2}\Delta^\top\widetilde{x}_j(t)\right) \\
&+ \left(\begin{pmatrix}\hat{\mu}_i & \widehat{\beta}_{i,-j}\end{pmatrix} - \begin{pmatrix}\mu_i & \beta_{i,-j}\end{pmatrix}\right)\frac{1}{T}\sum_{t=1}^{T}\begin{pmatrix}1\\ \widetilde{x}_{-j}(t)\end{pmatrix}\widetilde{x}_j^*(t) \\
&- \left(\begin{pmatrix}\hat{\mu}_i & \widehat{\beta}_{i,-j}\end{pmatrix} - \begin{pmatrix}\mu_i & \beta_{i,-j}\end{pmatrix}\right)\frac{1}{T}\sum_{t=1}^{T}\begin{pmatrix}1\\ \widetilde{x}_{-j}(t)\end{pmatrix}\begin{pmatrix}1 & \widetilde{x}_{-j}(t)\end{pmatrix}(\widehat{w}_j - w_j^*) \\
\equiv& I + II + III.
\end{aligned}$$

Then, by Lemma 1.4 and Lemma 1.6 on the lasso estimation consistency of $\widehat{w}_j$ together with Assumption 4 of bounded transition kernel function, we get

$$I \le C\sqrt{\frac{(s\vee\rho)\log p}{T}}\left(\sqrt{\frac{\log p}{T}} + T^{-1/2}\Delta\right) \le \frac{(s\vee\rho)\log p}{T}.$$

Also, by Lemma 1.3 and Lemma 1.5 of the lasso estimation consistency on $\widehat{\mu}_i, \widehat{\beta}_i$, $II \le C\frac{\log p}{T}$.

Finally, by the lasso prediction consistency by Lemma 1.5 and Lemma 1.6,

$$III \le C\sqrt{\frac{(s\vee\rho)\log p}{T}}\sqrt{\frac{\rho\log p}{T}}.$$

Combining the above, with probability at least $1 - c_1\exp(-c_2\log p)$, we have

$$\|\widehat{S}^0_{ij} - \widetilde{S}_{ij}\|_2^2 \le \frac{(s\vee\rho)\log p}{T}.$$

Thus, $\|E\|_2 \le C\frac{(s\vee\rho)\log p}{\sqrt{T}}$.

Next, by Lemma 1.1 and the tail bound for χ^2 distribution (Lemma 1 in Laurent and Massart (2000)),

$$P\left(\left\|V_T - \widetilde{\Delta}\right\|_2 > y\right) \le C_1T^{-1/8} + C_2y^{-2}. \tag{61}$$

Then, taking $y = \left(\frac{(s\vee\rho)\log p}{\sqrt{T}}\right)^{-1/4}$,

$$\left|\widehat{U}_T^0 - \left\|V_T - \widetilde{\Delta}\right\|_2^2\right| \le C\left(\frac{(s\vee\rho)\log p}{\sqrt{T}}\right)^{1/2},$$

with probability at least $1 - c_1\exp(-c_2\log p) - c_3T^{-1/8} - c_4\left(\frac{(s\vee\rho)\log p}{\sqrt{T}}\right)^{1/2}$.

Then, following Lemma 1.8 (and the discussion for $\phi = 1/2$), and taking $\delta = \left(\frac{\rho\log p}{\sqrt{T}}\right)^{1/2}$,

$$\begin{aligned}
\sup_{x\in\mathbb{R}}|P(\widehat{U}_T \le x) - F_d(x)| \le& \sup_{y\in\mathbb{R}}|P(U_T^0 \le y) - F_d(y)| \\
&+ F_d(x+\delta) - F_d(x-\delta) + P(|\widehat{U}_T - \widehat{U}_T^0| > \delta) \\
&\le c_1\exp(-c_2\log p) + c_3T^{-1/8} + c_4\left(\frac{(s\vee\rho)\log p}{\sqrt{T}}\right)^{1/2} \\
&+ c_5\left(\frac{\rho\log p}{\sqrt{T}}\right)^{1/2} + c_6\exp(-c_7\log p) + c_8T^{-1/8} + c_9\left(\frac{(s\vee\rho)\log p}{\sqrt{T}}\right)^{1/2} \\
&\le C_1\exp(-C_2\log p) + C_3T^{-1/8} + C_4\left(\frac{(s\vee\rho)\log p}{\sqrt{T}}\right)^{1/2}.
\end{aligned}$$

Next, we discuss $0 < \phi < 1/2$. By Lemma 1.7, $\left\|\Upsilon_j^{1/2}\left(\widehat{\Upsilon}_j^{0}\right)^{-1}\Upsilon_j^{1/2} - I\right\|_\infty$ converges to 0 w.r.t T, then for $T > c$ for some constant c, with probability at least $1 - c_1\exp(-c_2\log p)$,

$$\begin{aligned}
\widehat{U}_T &= T\widehat{S}_{ij}^0\left(\widehat{\Upsilon}_j^{0}\right)^{-1}\widehat{S}_{ij}^0 \\
&\ge T\|\left(\widehat{\Upsilon}_j^{0}\right)^{-1/2}\widehat{S}_{ij}^0\|_2^2\left(1 - d\left\|\Upsilon_j^{1/2}\left(\widehat{\Upsilon}_j^{0}\right)^{-1}\Upsilon_j^{1/2} - I\right\|_\infty\right) \\
&\ge CT\|\left(\widehat{\Upsilon}_j^{0}\right)^{-1/2}\widehat{S}_{ij}^0\|_2^2 \\
&\ge C\left(T\|\left(\widehat{\Upsilon}_j^{0}\right)^{-1/2}(\widehat{S}_{ij}^0 - S_{ij})\|_2 - \|V_T\|_2\right)^2.
\end{aligned}$$

But,

$$\begin{aligned}
\widehat{S}_{ij}^0 - S_{ij} =&(\widehat{w}_j - w_j^*)^\top\frac{1}{T}\sum_{t=1}^{T}\left(1 \quad \widetilde{x}_{-j}(t)\right)\widetilde{\epsilon}_i(t) \\
&-\frac{1}{T}\sum_{t=1}^{T}\left(\widetilde{x}_j(t) - \left(1 \quad \widetilde{x}_{-j}(t)\right)\widehat{w}_j\right)\widetilde{x}_j^\top(t)\beta_{ij} \\
&+\frac{1}{T}\sum_{t=1}^{T}\left(\widetilde{x}_j(t) - \widehat{w}_j^\top\begin{pmatrix}1\\ \widetilde{x}_{-j}(t)\end{pmatrix}\right)\,)\,\left(1 \quad \widetilde{x}_{-j}(t)\right)\,(\begin{pmatrix}\widehat{\mu}_i\\ \widehat{\beta}_{i,-j}\end{pmatrix} - \begin{pmatrix}\mu_i\\ \beta_{i,-j}\end{pmatrix}) \\
\equiv&E_1 + E_2 + E_3.
\end{aligned}$$

By Lemma 1.4 and Lemma 1.6, $\|E_1\|_2 \le \frac{(s\vee\rho)\log p}{T}$, with probability at least $1-c_1\exp(-c_2\log p)$.

Also, by Lemma 1.5, 1.6 and 2.1, we can show

$$
\begin{aligned}
\|E_3\|_2 &= \|\frac{1}{T}\sum_{t=1}^{T}\left(\widetilde{x}_j(t) - \widehat{w}_j^\top\begin{pmatrix}1\\ \widetilde{x}_{-j}(t)\end{pmatrix}\right)\left(1 \quad \widetilde{x}_{-j}(t)\right)(\begin{pmatrix}\widehat{\mu}_i\\ \widehat{\beta}_{i,-j}\end{pmatrix} - \begin{pmatrix}\mu_i\\ \beta_{i,-j}\end{pmatrix})\|_2 \\
&\le \left\|(\widehat{w}_j - w_j^*)^\top\frac{1}{T}\sum_{t=1}^{T}\begin{pmatrix}1\\ \widetilde{x}_{-j}(t)\end{pmatrix}\left(1 \quad \widetilde{x}_{-j}(t)\right)(\begin{pmatrix}\widehat{\mu}_i\\ \widehat{\beta}_{i,-j}\end{pmatrix} - \begin{pmatrix}\mu_i\\ \beta_{i,-j}\end{pmatrix})\right\|_2 \\
&+ \sqrt{d}\left\|\frac{1}{T}\sum_{t=1}^{T}\left(\widetilde{x}_j(t) - \left(1 \quad \widetilde{x}_{-j}(t)\right)w_j^* + \left(1 \quad \widetilde{x}_{-j}(t)\right)(w_j^* - \widehat{w}_j)\right)\begin{pmatrix}1\\ \widetilde{x}_{-j}(t)\end{pmatrix}\right\|_\infty \left\|\begin{pmatrix}\widehat{\mu}_i\\ \widehat{\beta}_{i,-j}\end{pmatrix} - \begin{pmatrix}\mu_i\\ \beta_{i,-j}\end{pmatrix}\right\|_1 \\
&\le \left\|(\widehat{w}_j - w_j^*)^\top\frac{1}{T}\sum_{t=1}^{T}\begin{pmatrix}1\\ \widetilde{x}_{-j}(t)\end{pmatrix}\left(1 \quad \widetilde{x}_{-j}(t)\right)(\begin{pmatrix}\widehat{\mu}_i\\ \widehat{\beta}_{i,-j}\end{pmatrix} - \begin{pmatrix}\mu_i\\ \beta_{i,-j}\end{pmatrix})\right\|_2 \\
&+ \sqrt{d}\left\|\frac{1}{T}\sum_{t=1}^{T}\widetilde{x}_j^*(t)\begin{pmatrix}1\\ \widetilde{x}_{-j}(t)\end{pmatrix}\right\|_\infty \left\|\begin{pmatrix}\widehat{\mu}_i\\ \widehat{\beta}_{i,-j}\end{pmatrix} - \begin{pmatrix}\mu_i\\ \beta_{i,-j}\end{pmatrix}\right\|_1 \\
&+ \sqrt{d}\|w_j^* - \widehat{w}_j\|_1\left\|\frac{1}{T}\sum_{t=1}^{T}\begin{pmatrix}1\\ \widetilde{x}_{-j}(t)\end{pmatrix}\left(1 \quad \widetilde{x}_{-j}(t)\right)\right\|_\infty \left\|\begin{pmatrix}\widehat{\mu}_i\\ \widehat{\beta}_{i,-j}\end{pmatrix} - \begin{pmatrix}\mu_i\\ \beta_{i,-j}\end{pmatrix}\right\|_1 \\
&\le C_1\sqrt{\frac{(s\vee\rho)\log p}{T}}\sqrt{\frac{\rho\log p}{T}} + C_2\sqrt{\frac{\rho\log p}{T}}\sqrt{\frac{\rho\log p}{T}} + C_2\sqrt{\frac{(s\vee\rho)\log p}{T}}\sqrt{\frac{\rho\log p}{T}} \\
&\le C\frac{(s\vee\rho)\log p}{T},
\end{aligned}
$$

with probability at least $1-c_1\exp(-c_2\log p)$.

To bound E_2, we can write, by Lemmas 1.6, S.1 and S.3,

$$
\begin{aligned}
&\|\frac{1}{T}\sum_{t=1}^{T}\left(\widetilde{x}_j(t) - \left(1 \quad \widetilde{x}(t)\right)\widehat{w}_j\right)\widetilde{x}_j(t) - \Upsilon_j\|_\infty \\
&\le \|W_j^*\|_1\left\|\frac{1}{T}\sum_{t=1}^{T}\begin{pmatrix}1\\ \widetilde{x}(t)\end{pmatrix}\left(1 \quad \widetilde{x}(t)\right) - \Upsilon\right\|_\infty \qquad (62)\\
&+ (1 + \|\widehat{w}_j - w_j^*\|_1)\left\|\frac{1}{T}\sum_{t=1}^{T}\begin{pmatrix}1\\ \widetilde{x}(t)\end{pmatrix}\left(1 \quad \widetilde{x}(t)\right)\right\|_\infty \\
&\le C\sqrt{\frac{(s\vee\rho)\log p}{T}} \qquad (63)
\end{aligned}
$$

with probability at least $1-c_1\exp(-c_2\log p)$.

Then, by Lemma 1.2 and $\beta_{ij} = T^{-\phi}\Delta$,

$$
\|E_2\|_2 \ge T^{-\phi}\|\Upsilon_j\Delta_j\|_2 - C\sqrt{\frac{(s\vee\rho)\log p}{T}}T^{-\phi} \ge CT^{-\phi}.
$$

Hence,

$$T\left\|\left(\widehat{\Upsilon}_j^{0}\right)^{-1/2}\left(\widehat{S}_{ij}^0 - S_{ij}\right)\right\|_2^2 \geq \left(C_1T^{1/2-\phi} - C_2\frac{(s\vee\rho)\log p}{\sqrt{T}}\right)^2 \geq CT^{1-2\phi}, \tag{64}$$

with probability at least $1 - c_1\exp(-c_2\log p)$.

In addition, by Lemma 1.1 and taking the intermediate results in the proof of Theorem 3.2 in Zheng and Raskutti (2018) about the tail bound of χ^2 distribution,

$$P(\|V_T\|_2 \geq c_1T^{1/2-\phi} - c_2\sqrt{x}) \leq cT^{-1/8} + C\exp(-(c_1T^{1/2-\phi} - c_2\sqrt{x})^2).$$

Therefore,

$$P(\widehat{U}_T^0 \leq x) \leq c_1\exp(-c_2\log p) + c_3T^{-1/8} + c_4\exp(-(c_5T^{1/2-\phi} - c_5\sqrt{x})^2).$$

Then, following Lemma 1.8 (see the discussion for $0 < \phi < 1/2$), we have

$$P(\widehat{U}_T^0 \leq x) \leq c_1\exp(-c_2\log p) + c_3T^{-1/8} + c_4\exp(-(c_5T^{1/2-\phi} - c_5\sqrt{x})^2)$$

Lastly, we discuss the case of $\phi > 1/2$. As before, one key part to bound $\left|\widehat{U}_T^0 - U_T\right|$ is $\widehat{S}_{ij}^0 - S_{ij}$.

$$\begin{aligned}
\widehat{S}_{ij}^0 - S_{ij} &= (\widehat{w}_j - w_j^*)^\top \frac{1}{T}\sum_{t=0}^{T-1}\widetilde{\epsilon}_i(t)\widetilde{x}_{-j}^\top(t) \\
&\quad + \frac{1}{T}\sum_{t=0}^{T-1}\widetilde{x}_j^*\begin{pmatrix}1 & \widetilde{x}_{-j}\end{pmatrix}\left(\begin{pmatrix}\widehat{\mu}_i \\ \widehat{\beta}_{i,-j}\end{pmatrix} - \begin{pmatrix}\mu_i \\ \beta_{i,-j}\end{pmatrix}\right) \\
&\quad - (\widehat{w}_j - w_j^*)^\top\left(\frac{1}{T}\sum_{t=0}^{T-1}\begin{pmatrix}1 \\ \widetilde{x}_{-j}\end{pmatrix}\begin{pmatrix}1 & \widetilde{x}_{-j}\end{pmatrix}\right)\left(\begin{pmatrix}\widehat{\mu}_i \\ \widehat{\beta}_{i,-j}\end{pmatrix} - \begin{pmatrix}\mu_i \\ \beta_{i,-j}\end{pmatrix}\right) \\
&\quad - T^{-(1+\phi)}\sum_{t=1}^{T}(\widetilde{x}_j(t) - \begin{pmatrix}1 & \widetilde{x}(t)\end{pmatrix}\widehat{w}_j)\widetilde{x}_j(t)\Delta
\end{aligned}$$

By (63) and Lemma 1.2,

$$\left\|\frac{1}{T}\sum_{t=1}^{T}(\widetilde{x}_j(t) - \begin{pmatrix}1 & \widetilde{x}(t)\end{pmatrix}\widehat{w}_j)\widetilde{x}_j(t)\Delta\right\|_2 \leq \|\Upsilon_j\Delta\|_2 + Cs\sqrt{\frac{\log p}{T}} \leq C$$

For the first three items, we use Lemmas 1.3-1.6 and Assumption 4, then

$$\|\sqrt{T}(\Upsilon_j)^{-1/2}(\widehat{S}_{ij}^0 - S_{ij})\|_2 \leq C_1\frac{(s\vee\rho)\log p}{\sqrt{T}} + C_2T^{1/2-\phi},$$

with probability at least $1 - c_1\exp(-c_2\log p)$.

Then, combining (61) with $y = \left(\frac{(s\vee\rho)\log p}{\sqrt{T}}\right)^{-1/4} \wedge T^{(2\phi-1)/6}$, then following (16),

$$\left|\widehat{U}_T^0 - U_T\right| \leq C_1\left(\frac{(s\vee\rho)\log p}{\sqrt{T}}\right)^{1/2} + C_2T^{(2\phi-1)/3},$$

with probability at least

$$1 - c_1\exp(-c_2\log p) - c_3T^{-1/8}\left(\frac{(s\vee\rho)\log p}{\sqrt{T}}\right)^{1/2} - C_5T^{(2\phi-1)/3},$$

if $(s\vee\rho)\log p = o(\sqrt{T})$ and $T > C$ for some constant C. Therefore, taking $\epsilon = C_1\left(\frac{(s\vee\rho)\log p}{\sqrt{T}}\right)^{1/2} + C_2T^{(1-2\phi)/3}$,

$$\sup|P(\widehat{U}_T^0 \leq x) - F_d(x)| \leq \frac{C_1}{T^{1/8}} + C_2\big(\frac{(s\vee\rho)\log p}{\sqrt{T}}\big)^{1/2} + \frac{C_3}{p^{C_4}} + C_3T^{\frac{1-2\phi}{3}}$$

Then, following Lemma 1.8 (Remark for $\phi > 1/2$), taking $\delta = \left(\frac{\rho\log p}{\sqrt{T}}\right)^{1/2}$,

$$\begin{aligned}\sup|P(\widehat{U}_T \leq x) - F_d(x)| &\leq \sup_{y\in\mathbb{R}}|P(U_T^0 \leq y) - F_d(y)| \\ &+ F_d(x+\delta) - F_d(x-\delta) + P(|\widehat{U}_T - \widehat{U}_T^0| > \delta) \\ &\leq \frac{c_1}{T^{1/8}} + c_2\big(\frac{(s\vee\rho)\log p}{\sqrt{T}}\big)^{1/2} + \frac{c_3}{p^{c_4}} + c_3T^{\frac{1-2\phi}{3}} \\ &+ c_4\left(\frac{\rho\log p}{\sqrt{T}}\right)^{1/2} + \frac{c_5}{T^{1/8}} + c_6\big(\frac{(s\vee\rho)\log p}{\sqrt{T}}\big)^{1/2} + \frac{c_7}{p^{C_4}} + c_8T^{\frac{1-2\phi}{3}} \\ &\leq \frac{C_1}{T^{1/8}} + C_2\big(\frac{(s\vee\rho)\log p}{\sqrt{T}}\big)^{1/2} + \frac{C_3}{p^{C_4}} + C_5T^{\frac{1-2\phi}{3}}.\end{aligned}$$

9.3 Proof of Theorem 3

The proof extends the proof of Theorem 3.4 in Zheng and Raskutti (2018) for the VAR model to the case of linear Hawkes model.

First, note that

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

where $\widehat{S}_{ij}$ is defined previously (23).

Then,

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

Thus,

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

Next,

$$\widehat{R}_T - \widehat{U}_T = (\widehat{S}_{ij})^\top \Big(\Big(\widetilde{\Upsilon}_j\Big)^{-1} \widehat{\Upsilon}_j \Big(\widetilde{\Upsilon}_j\Big)^{-1} - \Big(\widehat{\Upsilon}_j\Big)^{-1}\Big) \widehat{S}_{ij}$$

According to the proof in Lemma 1.8, we know $P\bigg(\|\widehat{S}_{ij} - S_{ij}\|_2 > \frac{(s_i \vee \rho_i)\log p}{T}\bigg) \leq c_1 \exp(-c_2 \log p)$. In addition, S_{ij} is bounded based on Assumption 3 and 4 and Lemma S.1 which leads that $\widetilde{x}_j^*(t)$ is bounded for all t. Thus, $\widehat{S}_{ij}$ is bounded with high probability. Therefore, we focus on bounding $\Big(\widetilde{\Upsilon}_j\Big)^{-1} \widehat{\Upsilon}_j \Big(\widetilde{\Upsilon}_j\Big)^{-1} - \Big(\widehat{\Upsilon}_j\Big)^{-1}$.

Let $E \equiv \widetilde{\Upsilon}_j - \widehat{\Upsilon}_j$. We write E into $E = \widetilde{\Upsilon}_j - \widetilde{\Upsilon}_j^0 + \widetilde{\Upsilon}_j^0 - \widehat{\Upsilon}_j^0 + \widehat{\Upsilon}_j^0 - \widehat{\Upsilon}_j$.

First, we bound $E^0 \equiv \widetilde{\Upsilon}_j^0 - \widehat{\Upsilon}_j^0$. Notice that

$$\begin{aligned}
\|E^0\|_\infty \leq & \left\| \frac{1}{T} \sum_{t=1}^T \Big(\widetilde{x}_j(t) - \begin{pmatrix} 1 & \widetilde{x}_{-j}(t) \end{pmatrix} \widehat{w}_j \Big) \begin{pmatrix} 1 & \widetilde{x}_{-j}(t) \end{pmatrix} \widehat{w}_j \right\|_\infty \\
\leq & \left\| \frac{1}{T} \sum_{t=1}^T \Big(\widetilde{x}_j(t) - \begin{pmatrix} 1 & \widetilde{x}_{-j}(t) \end{pmatrix} w_j^* \Big) \widetilde{x}_{-j}(t) \right\|_\infty \big(\|w_j^*\|_1 + \|w_j^* - \widehat{w}_j\|_1 \big) \\
& + \max \left| (\widehat{w}_j - w_j^*)^\top \frac{1}{T} \sum_{t=1}^T \bigg(\begin{pmatrix} 1 \\ \widetilde{x}_{-j}(t) \end{pmatrix} \begin{pmatrix} 1 & \widetilde{x}_{-j}(t) \end{pmatrix} \bigg) (\widehat{w}_j - w_j^*) \right| \\
& + \left\| \frac{1}{T} \sum_{t=1}^T \begin{pmatrix} 1 \\ \widetilde{x}_{-j}(t) \end{pmatrix} \begin{pmatrix} 1 & \widetilde{x}_{-j}(t) \end{pmatrix} w_j^* \right\|_\infty \|\widehat{w}_j - w_j^*\|_1.
\end{aligned}$$

With probability at least $1 - c_1 \exp(-c_2 \log p)$, the first item on RHS is bounded by $C\sqrt{\frac{\rho \log p}{T}}$ by Lemma 1.3 and 1.6; the second item is bounded by $C\frac{s \vee \rho \log p}{T}$ based on Lemma 1.6 and Assumption 4. For the third item, by Lemmas 1.2, S.1 and S.3,

$$\begin{aligned}
\left\| \frac{1}{T} \sum_{t=1}^T \begin{pmatrix} 1 \\ \widetilde{x}_{-j}(t) \end{pmatrix} \begin{pmatrix} 1 & \widetilde{x}_{-j}(t) \end{pmatrix} w_j^* \right\|_\infty \leq & \|\Upsilon_{-j,-j} w_j^*\|_\infty \\
& + \left\| \frac{1}{T} \sum_{t=1}^T \begin{pmatrix} 1 \\ \widetilde{x}_{-j}(t) \end{pmatrix} \begin{pmatrix} 1 & \widetilde{x}_{-j}(t) \end{pmatrix} - \Upsilon_{-j,-j} \right\|_\infty \|w_j^*\|_1 \\
\leq & \Lambda_{\max}\Big(\Upsilon\Big) \max(w_j^*) + c\sqrt{\frac{\rho \log p}{T}} \leq C
\end{aligned}$$

Then, by Lemma 1.6 and combining the first two items in E^0, $\|E^0\|_\infty \leq C\sqrt{\frac{s \vee \rho \log p}{T}}$, with probability at least $1 - c_1 \exp(-c_2 \log p)$.

By Lemma 1.7, we know that

$$P\left(\|\widehat{\Upsilon}_j^0 - \widehat{\Upsilon}_j\|_2 > \sqrt{\rho_i \frac{\log p}{T}}\right) \leq c_1 \exp(-c_2 \log p).$$

Following a similar proof as we did to bound $\widehat{\Upsilon}_j^0 - \widehat{\Upsilon}_j$ in Lemma 1.7 based on the consistency of $\widehat{\sigma}_i(t)$ to $\sigma_i(t)$, we can show that

$$P\left(\|\widetilde{\Upsilon}_j^0 - \widetilde{\Upsilon}_j\|_2 > \sqrt{\rho_i \frac{\log p}{T}}\right) \leq c_1 \exp(-c_2 \log p).$$

Therefore, $\|E\|_\infty \leq C\sqrt{\frac{s \vee \rho \log p}{T}}$, with probability at least $1 - c_1 \exp(-c_2 \log p)$.

Next, we have

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

and

$$\left(\widetilde{\Upsilon}_j\right)^{-1} \widehat{\Upsilon}_j \left(\widetilde{\Upsilon}_j\right)^{-1} - \left(\widehat{\Upsilon}_j\right)^{-1} \leq E + E^\top + E\big(\widehat{\Upsilon}_j\big)^{-1} E^\top.$$

Based on Lemma 1.2 and 1.7, there exists $C > 0$ such that with probability at least $1 - c_1 \exp(-c_2 \log p)$,

$$\Lambda_{\min}\big(\widehat{\Upsilon}_j\big) \geq \Lambda_{\min}\big(\Upsilon_j\big) - d\|\widehat{\Upsilon}_j - \Upsilon_j\|_\infty \geq C$$

which implies $\Lambda_{\max}\big(\widehat{\Upsilon}_j\big)^{-1} < C^{-1}$ and $\|\mathbb{E}\left(\widehat{\Upsilon}_j\right)^{-1} E^\top\|_\infty \leq Cd\|E\|_\infty$ with probability at least $1 - c_1 \exp(-c_2 \log p)$.

Then,

$$\left\|\left(\widetilde{\Upsilon}_j\right)\widehat{\Upsilon}_j^{-1}\left(\widetilde{\Upsilon}_j\right) - \left(\widehat{\Upsilon}_j\right)\right\|_2 \leq d\left\|\left(\widetilde{\Upsilon}_j\right)\widehat{\Upsilon}_j^{-1}\left(\widetilde{\Upsilon}_j\right) - \left(\widehat{\Upsilon}_j\right)\right\|_\infty \leq d\|E\|_\infty$$

Let $P = \left(\widetilde{\Upsilon}_j\right)\widehat{\Upsilon}_j^{-1}\left(\widetilde{\Upsilon}_j\right) - \left(\widehat{\Upsilon}_j\right)$ and $Q = \widehat{\Upsilon}_j$, using the invertible matrix inequality result given by Lemma A.2 in (Zheng and Raskutti, 2018) that

$$\|\big(P + Q\big)^{-1} - Q^{-1}\|_2 \leq \frac{\|Q^{-1}\|_2\|P\|_2}{1 - \|Q^{-1}\|_2\|P\|_2}$$

we show

$$\begin{aligned} \|\left(\widetilde{\Upsilon}_j\right)^{-1} \widehat{\Upsilon}_j \left(\widetilde{\Upsilon}_j\right)^{-1} - \left(\widehat{\Upsilon}_j\right)^{-1}\|_\infty &\leq \|\left(\widetilde{\Upsilon}_j\right)^{-1} \widehat{\Upsilon}_j \left(\widetilde{\Upsilon}_j\right)^{-1} - \left(\widehat{\Upsilon}_j\right)^{-1}\|_2 \\ &\leq C\left\|\left(\widetilde{\Upsilon}_j\right)\widehat{\Upsilon}_j^{-1}\left(\widetilde{\Upsilon}_j\right) - \left(\widehat{\Upsilon}_j\right)\right\|_2 \\ &\leq Cd\|E\|_\infty \leq C\sqrt{\frac{s \vee \rho \log p}{T}} \end{aligned}$$

with probability at least $1 - c_1 \exp(-c_2 \log p)$.

References

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

E. Bacry, K. Dayri, and J. Muzy. Non-parametric kernel estimation for symmetric hawkes processes. application to high frequency financial data. *The European Physical Journal B*, 85, 2011.

E. Bacry, I. Mastromatteo, and J. Muzy. Hawkes processes in finance. *Market Microstructure and Liquidity*, 01, 2015.

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

P. J. Bickel, Y. Ritov, A. B. Tsybakov, et al. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

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

P. Brémaud and L. Massoulié. Stability of nonlinear hawkes processes. *The Annals of Probability*, 24(3):1563–1588, 1996.

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

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

M. Costa, C. Graham, L. Marsalle, and V. C. Tran. Renewal in hawkes processes with self-excitation and inhibition. 2018.

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.

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

J. Etesami, N. Kiyavash, K. Zhang, and K. Singhal. Learning network of multivariate hawkes processes: A time series approach, 2016.

I. Grama and E. Haeusler. An asymptotic expansion for probabilities of moderate deviations for multivariate martingales. *Journal of Theoretical Probability*, 19:1–44, 2006.

N. R. Hansen, P. Reynaud-Bouret, and V. Rivoirard. Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli*, 21(1):83–143, 2015.

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

A. G. Hawkes and D. Oakes. A cluster process representation of a self-exciting process. *Journal of Applied Probability*, 11(3):493–503, 1974.

A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15, 2013.

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

S. L. Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

S. W. Linderman and R. P. Adams. Discovering latent network structure in point process data, 2014.

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

Y. Ogata. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association*, 83(401):9–27, 1988.

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

L. Paninski, J. Pillow, and J. 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.

J. Pillow, J. Shlens, L. Paninski, A. Sher, A. Litke, E. Chichilnisky, and E. Simoncelli. Spatio-temporal correlations and visual signaling in a complete neuronal population. *Nature*, 454: 995–9, 2008.

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

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

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

S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, 42(3):1166–1202, 2014.

C.-H. Zhang and S. 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.

L. Zheng and G. Raskutti. Testing for high-dimensional network parameters in auto-regressive models, 2018.

K. Zhou, H. Zha, and L. Song. Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. In *AISTATS*, 2013.

Appendix: Proof of Technical Lemmas

To proof Theorem 1-3, we introduce technical lemmas 1.1-1.8 which essentially depends on Lemma S.1-S.3. We give proof of all these lemmas. We also use results from previous work in Lemma D.1-D.4. We refer audience to the original source for details.

Lemma 1.1-1.8

Proof of Lemma 1.1

The main part of the proof is based on the result on the martingale difference sequence. Our proof extends the previous result in VAR model (Lemma 5.3 in (Zheng and Raskutti, 2018)) to the linear Hawkes model (2).

Let

$$\xi_{T,t} = -\frac{1}{\sqrt{T}}\big(\Upsilon_j\big)^{-1/2}\frac{\epsilon_i(t)}{\sigma_i(t)}\widetilde{x}_j^*(t) \tag{65}$$

where $\sigma_i(t)$ defined in (8) and $\widetilde{x}_j^*(t)$ defined in (11).

As defined previously, $\mathcal{H}_{T,t}$ is information filtration of the past. Then $(\xi_{T,t}, \mathcal{H}_{T,t})$ is a martingale difference sequence, and $V_T = \sum_{t=0}^{T-1} \xi_{T,t}$. To complete the proof, we use Lemma D.1 as follows, which is a technical Lemma from (Zheng and Raskutti, 2018) which is a modified version of Lemma 4 by (Grama and Haeusler, 2006). Although the original application of this Lemma is on VAR model, the Lemma requires a martingale difference sequences as an object to consider thus can be applied to $\xi_{T,t}$.

Lemma D.1 Let $(\xi_{n,i}, \mathcal{H}_{n,i})_{0\le i\le n}$ be a martingale difference sequence taking values in $\mathbb{R}^d$. Let $X_n^k = \sum_{i=1}^k \xi_{ni}$, and $\langle X^n\rangle_k = \sum_{i=1}^k a_{ni} \equiv \sum_{i=1}^k E\big(\xi_{ni}\xi_{ni}^\top|\mathcal{H}_{n,i-1}\big)$. Define $R_\delta^{n,d} = L_\delta^{n,d} + N_\delta^{n,d}$,

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

Then $\forall u \in \mathbb{R}^d, r \ge 0, 0 < \delta \le 1/2$, when $R_\delta^{n,d} \le 1$,

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

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

By Lemma D.1, to complete the proof, we need to check the bound for $R_\delta^{n,d} = L_\delta^{n,d} + N_\delta^{n,d}$.

First, by Lemma 1.2, both $\Lambda_{\max}\big(\Upsilon_j^{-1}\big)$ and $\Lambda_{\max}\big(\Upsilon_j^{-1}\big)$ are bounded. Second, by the bounded intensity by Assumption 3 (i.e. $0 < \lambda_{\min} \le \lambda_i(t) \le \lambda_{\max} < 1$) and $Y_i(t) \in \{0, 1\}$,

$\epsilon_i(t) = Y_i(t) - \lambda_i(t)$ and $\sigma_i(t)$ are bounded. Third, by Lemma S.1, $\|w_j^*\|_\infty \le \|w_j^*\|_2 \le C$. At last, $x_j(t) = x_j(t) \le \int_0^T k_j(t)dt$ is bounded by Assumption 4. Therefore,

$$\|\widetilde{x}_j^*(t)\|_2^2 = \|\widetilde{x}_j\|_2^2 + \|\big(1 \quad \widetilde{x}_{-j}\big)\, w_j^*\|_2^2 \le (1 + ds\|w_j^*\|_\infty^2) \max_{1\le j\le p} |x_j(t)/\sigma_i(t)|^2 \le Cds,$$

which implies

$$L_\delta^{n,d} = \sum_{t=1}^T E\|\xi_{Tt}\|_2^{2+2\delta} \le CT^{-(1+\delta)} \sum_{i=1}^T E\|\frac{1}{\sigma_i(t)}\epsilon_i(t)\widetilde{x}_j^*(t)\|_2^{2+2\delta} \le C(d,\delta)T^{-\delta}(ds)^{1+1\delta}$$

Remark: Note that we have an easier proof comparing to Zheng et al (2018) because in our case according to assumptions to the Hawkes model, $x(t)$ and $\epsilon_i(t)$ are bounded deterministically, while in Zheng et al (2018), they consider sub-gaussian error so they need to proof the bound $L_\delta^{n,d}$ based on $x(t)$ and $\epsilon_i(t)$ with high probability. d here is dimension of β_{ij} to test. In our case (9), $d = 1$ but the proof is applicable for any fixed low dimension $d > 1$.

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

$$\sum_{t=0}^{T-1} E\big(\xi_{T,t}\xi_{T,t}^\top|\mathcal{H}_t\big) - I = \big(\Upsilon_j\big)^{-1/2}\big(\frac{1}{T}\sum_{t=1}^T \widetilde{x}_j^*(t)\big(\widetilde{x}_j^*(t)\big)^\top - \Upsilon_j\big)\big(\Upsilon_j\big)^{-1/2}$$

By Lemma 1.2, the rank of $\big(\Upsilon_j\big)^{-1/2}\Bigg(\frac{1}{T}\sum_{t=1}^T \widetilde{x}_j^*(t)\big(\widetilde{x}_j^*(t)\big)^\top - \Upsilon_j\Bigg)\big(\Upsilon_j\big)^{-1/2}$ is at most d. By matrix norm inequality as introduced in A.3 (Zheng and Raskutti, 2018) that for $B \in \mathbb{R}^{d\times d}$

$$\|B\|_{tr} \le d\|B\|_2,$$

then

$$\begin{aligned} N_\delta^{T,d} &= E\left\|\big(\Upsilon_j\big)^{-1/2}\big(\frac{1}{T}\sum_{t=1}^T \widetilde{x}_j^*(t)\big(\widetilde{x}_j^*(t)\big)^\top - \Upsilon_j\big)\big(\Upsilon_j\big)^{-1/2}\right\|_{tr}^{1+\delta} \\ &\le E\Bigg(d\left\|\big(\Upsilon_j\big)^{-1/2}\big(\frac{1}{T}\sum_{t=1}^T \widetilde{x}_j^*(t)\big(\widetilde{x}_j^*(t)\big)^\top - \Upsilon_j\big)\big(\Upsilon_j\big)^{-1/2}\right\|_2\Bigg)^{1+\delta} \\ &\le CE\big(d^2\left\|\big(\frac{1}{T}\sum_{t=1}^T \widetilde{x}_j^*(t)\big(\widetilde{x}_j^*(t)\big)^\top - \Upsilon_j\big)\right\|_\infty\big)^{1+\delta} \end{aligned}$$

where the last step is based on matrix norm inequality

$$\|B_d\|_2 \le d\|B_d\|_\infty$$

and Lemma 1.2.

We introduce the follow Lemma to bound

$$\|\frac{1}{T}\sum_{t=1}^{T}\widetilde{x}_j^*(t)\big(\widetilde{x}_j^*(t)\big)^\top - \Upsilon_j\|_\infty \le (1+\|w_{j,-j}^*\|_1^2)\|\frac{1}{T}\sum_{t=1}^{T}\widetilde{x}(t)\big(\widetilde{x}(t)\big)^\top - E\bigg(\frac{1}{T}\sum_{t=1}^{T}\big(\widetilde{x}(t)\big)^\top\widetilde{x}^\top(t)\bigg)\|_\infty$$
$$+\|w_{j0}\|_1^2\|\frac{1}{T}\sum_{t=1}^{T}\widetilde{x}(t) - E\big(\widetilde{x(t)}\big)\|_\infty$$

Lemma S.3 Under Assumption 1-4 and the stationary linear Hawkes model defined in (6), we have

$$P\bigg(\|\frac{1}{T}\sum_{t=1}^{T}\big(\widetilde{x}(t)\big)^\top\widetilde{x}(t) - E\frac{1}{T}\sum_{t=1}^{T}\big(\widetilde{x}(t)\big)^\top\widetilde{x}(t)\|_\infty > \delta\bigg) \le c_1\exp\bigg(-c_2\min\big\{\sqrt{\frac{T}{s}}\delta, \frac{T}{s}\delta^2\big\}\bigg) \tag{66}$$

and

$$P\bigg(\|\frac{1}{T}\sum_{t=1}^{T}\widetilde{x}(t) - E\bigg(\widetilde{x}(t)\bigg)\|_\infty > C\delta\bigg) \le\le c_1\exp(-c_2T\delta^2).$$

Remark: The proof is essentially based on the deviation bound of the 1st-order and quadratic form show in (81) and (80) in Lemma S.2 and apply Taylor expansion to reach the conclusion.

By Lemma S.3,

$$\begin{aligned}
N_\delta^{T,d} &\le \int_0^\infty P\bigg(\big(d^2\|\frac{1}{T}\sum_{t=1}^{T}\big(\widetilde{x}_j^*(t)\big)^\top\widetilde{x}_j^*(t) - \Upsilon_j\|_\infty\big)^{1+\delta} > r\bigg)dr \\
&= \int_0^\infty P\bigg(d^2\|\frac{1}{T}\sum_{t=1}^{T}\big(\widetilde{x}_j^*(t)\big)^\top\widetilde{x}_j^*(t) - \Upsilon_j\|_\infty > r^{1/(1+\delta)}\bigg)dr \\
&\le \int_0^\infty c_1\exp\bigg(-c_2\min\big(\frac{T}{s}r^{2/(1+\delta)/d^4}, \sqrt{\frac{T}{s}}r^{1/(1+\delta)/d^2}\big)\bigg)dr \\
&= C(\delta)\max\{T^{-(1+\delta)/2}s^{(1+\delta)/2}d^{2(1+\delta)}, T^{-(1+\delta)}s^{(1+\delta)}d^{4(1+\delta)}\}
\end{aligned}$$

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

Assume $s = o(T)$ and $d = o(T)$ and combine the two parts,

$$R_\delta^{n,d} = L_\delta^{n,d} + N_\delta^{n,d} \le C(\delta, s, d)(T^{-(1+\delta)/2} + T^{-\delta})$$

Here $C(\delta, s, d)$ is increasing with s which implies bounding $R_\delta^{n,d}$ becomes more difficulty when the sparsity s and/or the dimension of parameter to test d becomes larger.

Therefore, by Lemma D.1, we have $\forall x \geq 0, u \in \mathbb{R}^d$, and $\delta \in [0, 1/2]$, when $T > C(\delta)$,

$$|P(\widehat{U}_T + u \leq x) - F_{d,\|u\|_2^2}(x)| \leq C(\|u\|_2^2, d, \delta)\left(R_\delta^{T,d}\right)^{\frac{1}{3+2\delta}} \tag{67}$$

The best rate is achieved when taking $\delta = 1/2$, when $T > C$,

$$|P(\widehat{U}_T + u \leq x) - F_{d,\|u\|_2^2}(x)| \leq C(\|u\|_2^2, s, d)T^{-1/8} \tag{68}$$

■

Proof of Lemma 1.2

For consider bounding the eigenvalue of $\Upsilon_x = Cov(x(t))$.

Without loss of generality, consider a discrete time scenario with unit time window ($dt = 1$). Then, $x_j(t) = \sum_{s=1}^{t-1} k_j(t-s)Y_j(s)$ and $x_j(1) = 0$. Let

$$K_{t-1} = \begin{pmatrix} k_1(1)e_1 & k_1(2)e_1 & k_1(3)e_1 & \dots & \dots & k_1(t-1)e_1 \\ k_2(1)e_2 & k_2(2)e_2 & k_2(3)e_2 & \dots & \dots & k_2(t-1)e_2 \\ . & . & . & . & . & . \\ k_p(1)e_p & k_p(2)e_p & k_p(3)e_p & \dots & \dots & k_p(t-1)e_p \end{pmatrix} \in \mathbb{R}^{p \times (t-1)p},$$

where $e_j \in \mathbb{R}^p$ with 1 on the j-th entry and 0 all other entires.

Let $Y_t = (Y_1(t), \dots, Y_p(t))^\top \in \mathbb{R}^p$. Then, $x(t) = K_{t-1}Y_{t-1}$, where $Y_t = (Y(t), \dots, Y(1))^\top$.

$$\Upsilon_x = Cov(x(t)) = K_{t-1}Cov(Y_{t-1})K_{t-1}^\top$$

Therefore, to bound the eigenvalue of Υ_x, we need

- Condition 1: $0 < \min_j k_j(1) \leq \max_j \sum_{t=1}^{\infty} k_j(t) \leq C < \infty$
- Condition 2: $0 < \Lambda_{\min} Cov(Y) \leq \Lambda_{\max} Cov(Y) < 1$

The first condition is met by assuming an integrable and non-trivial transfer kernel function by Assumption 4. We show the second condition is met by a specific structure of the transfer function.

We first introduce a result (Prop. 2.3,(Basu and Michailidis, 2015)) linking the eigen-value of cross-covariance matrix with its spectral radius. Although the original application of this result is for the VAR model, this result is valid for a cross-covariance stationary process.

Let $Y = (Y(T), \dots, Y(1)) \in \mathbb{R}^{Tp}$, $\Sigma = Cov(Y) \in \mathbb{R}^{Tp \times Tp}$ and $\Gamma(l) = Cov(Y(t), Y(t+l))$. Define the spectral density function, which is the Fourier transformation on the cross-covariance

$\Gamma(l)$, $f_\Gamma(\theta) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\Gamma(l)\exp(-i\theta l)dl$.

Lemma D.2 Under the linear Hawkes process with cross-covariance stationary,

$$2\pi m(f_\Gamma) \leq \Lambda_{\min}\Sigma \leq \Lambda_{\max}\Sigma \leq 2\pi M(f_\Gamma)$$

Here

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

Next, we introduce the result linking the spectral density Γ to the spectral density of the matrix of transition function Ω.

Lemma D.3 The Fourier transform of the normalized covariance matrix of a stationary multivariate Hawkes process with window size z is given by

$$f_\Gamma(w) = 2\Big(I - f_\Omega(w)\Big)^{-1} diag(\Lambda)\Big(I - f_\Omega^*(w)\Big)^{-1}$$

The result is a direct extension of Theorem 1 in (Bacry et al., 2011) or Theorem 3 in (Etesami et al., 2016) from assuming non-negative transfer function to a general real function. This result extends the class of linear Hawkes process including non-mutually exciting structure. The proof is directly established by re-writing a real function into its positive part and its negative part, and then complete the proof by the distributive property of the convolution operation.

By Lemma D.2 and Lemma D.3,

$$\frac{2\min(\Lambda)}{M\Big(I - f_\Omega^*(w)\Big)} \leq \Lambda(f_\Gamma) \leq \frac{2\max(\Lambda)}{m\Big(I - f_\Omega^*(w)\Big)}$$

Therefore, one sufficient condition to bound the spectral radius of f_Γ is to assume

- Bounded intensity function as assumed in Assumption 3;
- $\Lambda_{\max}\int_0^\infty |\Omega(t)|dt < 1$ which leads $m\Big(I - f_\Omega^*(w)\Big) > 0$ as assumed in Assumption 1;
- Bounded row and column sum of Ω as assume in Assumption 2; that is,

$$M\Big(I - f_\Omega^*(w)\Big) \leq 1 + \Big(\max_{1\leq i\leq p}\sum_{j=1}^{p}\Omega_{ij} + \max_{1\leq j\leq p}\sum_{i=1}^{p}\Omega_{ij}\Big)/2 < \infty$$

Since for a fixed time range T, both Λ and Ω are constants only containing β, we have

$$0 < C_1(\beta) \le \Lambda_{\min}\big(\Upsilon_x\big) \le \Lambda_{\max}\big(\Upsilon_x\big) \le C_2(\beta) < \infty$$

By Assumption 3 of bounded intensity such that $0 < \lambda_{\max} \le \lambda_i(t) \le \lambda_{\min} < 1$, $0 < c_1 \equiv \min\{\lambda_{\max}(1-\lambda_{\max}), \lambda_{\min}(1-\lambda_{\min})\} \le \sigma_i^2(t) = \lambda_i(t)(1-\lambda_i(t)) \le 1/4 \equiv c_2$. Let $\Upsilon = Cov\Bigg(x(t)/\sigma_i(t)\Bigg)$,

$$c_2^{-1}\Upsilon_x \le \Upsilon \le c_1^{-1}\Upsilon_x. \tag{69}$$

Notice that

$$Cov(\widetilde{x}^*_j(t))^{-1} = \Bigg(\Upsilon^{-1}\Bigg)_{jj}$$

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

$$\begin{gathered}\Lambda_{\min}\left(Cov(\widetilde{x}^*_j(t))^{-1}\right) \ge \Lambda_{\min}\left(\Upsilon^{-1}\right)\\ \Lambda_{\max}\left(Cov(\widetilde{x}^*_j(t))^{-1}\right) \le \Lambda_{\max}\left(\Upsilon^{-1}\right)\end{gathered}$$

Therefore,

$$c_1\Lambda_{\min}\Bigg(\Upsilon_x^{-1}\Bigg) \le \Lambda_{\min}\Bigg(\Upsilon_j^{-1}\Bigg) \le \Lambda_{\max}\Bigg(\Upsilon_j^{-1}\Bigg) \le c_2\Lambda_{\max}\Bigg(\Upsilon_x^{-1}\Bigg)$$

■

Remark: In the above, we specify the condition to upper bound $M\Bigg(I - f^*_\Omega(w)\Bigg)$ in order to be comparable with the results in VAR model shown in (Basu and Michailidis, 2015). Other condition could also be available to bound the spectral radius of f_Γ. For example, we could consider a bounded transition function value; i.e., $\max_{1\le i,j\le p}\Omega_{ij} \le C_\Omega < \infty$, and consider sparsity of the transition matrix Ω such that $\rho = \sum_{i=1}^p \rho_i$, as defined previously $\rho_i = \|\beta_i\|_0$, where ρ does not goes up with p. Then $M\Bigg(I - f^*_\Omega(w)\Bigg) \le 1 + \rho C_\Omega$.

Proof of Lemma 1.3

Lemma 1.3 is similar to Lemma 5.6 in (Zheng and Raskutti, 2018). One of the key part of the proof is the construction of w^* such that $\mathbb{E}\left(\widetilde{x}^*_j(t)\widetilde{x}_k(t)\right) = 0$ for any $k \ne j$.

WLOG, consider $j = 1$ for convenience of notation. Notice that $x_k^\top(t) = x(t)e_k$, $e_k = (\underbrace{0,\dots,0}_{k-1 \text{ 0's}}, 1, \underbrace{0,\dots,0}_{p-k-1 \text{ 0's}})^\top \in \mathbb{R}^p$. Denote W_j^* such that $\widetilde{x}_j^*(t) = \begin{pmatrix}1 & \widetilde{x}(t)\end{pmatrix} W_j^*$. Then,

$$\frac{1}{T}\sum_{t=1}^{T} \widetilde{x}_j^*(t)\widetilde{x}_k^\top(t) = W_j^*\left(\frac{1}{T}\sum_{t=1}^{T}\begin{pmatrix}1\\ \widetilde{x}(t)\end{pmatrix}\widetilde{x}(t)\right)e_k$$

Also, note that $\mathbb{E}\big(\frac{1}{T}\sum_{t=1}^{T}\widetilde{x}_j^*(t)\widetilde{x}_k^\top(t)\big) = 0, k \neq j$ by the definition of w_j^*.

Then, by Lemma S.3 of the 1st and 2nd order deviation bound on $\widetilde{x}(t)$ and Lemma S.1 of bounded w_j^*, we have

$$\begin{aligned}
\|\frac{1}{T}\sum_{t=1}^{T}\widetilde{x}_j^*(t)\widetilde{x}_k^\top(t)\|_\infty &= \|\frac{1}{T}\sum_{t=1}^{T}\widetilde{x}_j^*(t)\widetilde{x}_k^\top(t) - \mathbb{E}\big(\frac{1}{T}\sum_{t=1}^{T}\widetilde{x}_j^*(t)\widetilde{x}_k^\top(t)\big)\|_\infty \\
&\le \|W_j^*\|_1 \left\|\frac{1}{T}\sum_{t=1}^{T}\begin{pmatrix}1\\ \widetilde{x}(t)\end{pmatrix}\widetilde{x}(t)\right) - \mathbb{E}\left(\frac{1}{T}\sum_{t=1}^{T}\begin{pmatrix}1\\ \widetilde{x}(t)\end{pmatrix}\widetilde{x}(t)\right)\right\|_\infty \|e_k\|_1 \\
&\le (\|w_j^*\|_1 + 1)\left\|\frac{1}{T}\sum_{t=1}^{T}\begin{pmatrix}1\\ \widetilde{x}(t)\end{pmatrix}\widetilde{x}(t)\right) - \mathbb{E}\left(\frac{1}{T}\sum_{t=1}^{T}\begin{pmatrix}1\\ \widetilde{x}(t)\end{pmatrix}\widetilde{x}(t)\right)\right\|_\infty \|e_k\|_1 \\
&\le C\sqrt{\frac{s\log p}{T}}
\end{aligned}$$

with probability at least $1 - c_1\exp(-c_2\log p)$.

At the end, we take a union bound over all $k \neq j$ to reach the conclusion. ■

Proof of Lemma 1.4

The deviation bound for linear Hawkes process has been discussed in previous literature (Chen et al., 2017). We first introduce a Lemma given in (Chen et al., 2017) which is a direct result based on the martingale inequality from Theorem 3.1 in (van de Geer, 1995).

Lemma D.4 Under the linear Hawkes model, let $H(t)$ be a bounded function that is $\mathcal{H}_t$-predictable. Then, for any $\epsilon > 0$, the inequality

$$\frac{1}{T}\int_0^T H(t)\Big\{\lambda_i(t)dt - dN_i(t)\Big\} \le 4\Big\{\frac{\lambda_{\max}}{2T}\int_0^T H^2(t)dt\Big\}^{1/2}\epsilon^{1/2}$$

holds with probability at least $1 - c\exp(-\epsilon T)$ for some constant c, for any $i = 1,\dots,p$.

By Assumption 2, there exists constant C such that $\|x(t)\|_2^2 \le \Big(\sum_{j=1}^{p}\int_0^T |k_j(t)|dt\Big)^2 < C$.

By $\{\epsilon_i(t), \mathcal{H}_t\}$ is a martingale sequence and Lemma D.4 taking $\epsilon = \frac{\log p}{Td}$ and union bound over low dimension d,

$$P\Big(\|\frac{1}{T}\sum_{t=1}^{T}\epsilon_i(t)x_j(t)\|_\infty > C\sqrt{\frac{\log p}{T}}\Big) \leq c_1\exp(-c_2\log p) \tag{70}$$

Next consider $\|\frac{1}{T}\sum_{t=1}^{T}\widetilde{\epsilon}_i(t)\widetilde{x}^*_j(t)\|_\infty$.

Let $H(t) = \frac{1}{\sigma_i(t)}\widetilde{x}^*_j(t)$. Notice that $\|\widetilde{x}^*_j(t)\|_2^2 \leq \|\widetilde{x}_j(t)\|_2^2 \leq C$ by the choice of w^* and Assumption 4 of bounded transition kernel function. Then, according Assumption 3 of bounded intensity function which lead bounded $\sigma_i^2(t)$,

$$\frac{1}{T}\sum_{1}^{T}H^2(t) \leq C$$

Therefore, applying Lemma D.4,

$$P\Big(\|\frac{1}{T}\sum_{t=1}^{T}\widetilde{\epsilon}_i(t)\widetilde{x}^*_j(t)\|_\infty > C\sqrt{\frac{\log p}{T}}\Big) \leq C_1\exp(-C_2\log p) \tag{71}$$

Another result by Lemma D.4, taking $H(t) = 1$,

$$P\Big(\|\frac{1}{T}\sum_{t=1}^{T}\widetilde{\epsilon}_i(t)\|_\infty > C\sqrt{\frac{\log p}{T}}\Big) \leq C_1\exp(-C_2\log p) \tag{72}$$

■

Proof of Lemma 1.5

The proof is typical for estimation consistency in lasso estimates. We start with the basic inequality induced by the construction of the lasso estimator in (17). We then bound the prediction error of the lasso regression using the results of Lemma 1.4. Next, as one of the key part, we show that the restricted eigen-value condition (REC) is satisfied with high probability in the linear Hawkes process under our setting based on what we show about the bounded eigenvalue of Υ_x in Lemma 1.2.

By the construction of the lasso estimator,

$$\widehat{\mu}_i, \widehat{\beta}_i = \arg\min_{\mu\in\mathbb{R},\beta\in\mathbb{R}^p}\frac{1}{T}\sum_{t=1}^{T}(Y_i(t) - \mu - x(t)\beta)^2 + \lambda\|\beta\|_1,$$

we have

$$\frac{1}{T}\sum_{t=1}^{T}(Y_i(t)-\widehat{\mu}_i-x(t)\widehat{\beta}_i)^2+\lambda\|\widehat{\beta}_i\|_1 \le \frac{1}{T}\sum_{t=1}^{T}(Y_i(t)-\mu_i-x(t)\beta_i)^2+\lambda\|\beta_i\|_1$$

Let $H=\frac{1}{T}\sum_{t=1}^{T}\begin{pmatrix}1\\(x(t))^\top\end{pmatrix}\begin{pmatrix}1 & (x(t))^\top\end{pmatrix}$ and $u=\widehat{\mu}_i-\mu_i$, $v=\widehat{\beta}_i-\beta_i$. Define $s=\{j:\beta_{ij}\neq 0\}$ and $s^c=\{j:\beta_{ij}=0\}$. Re-organize the above,

$$\begin{aligned}
\begin{pmatrix}u & v^\top\end{pmatrix}H\begin{pmatrix}u\\v\end{pmatrix} &\le 2\left\|\frac{1}{T}\sum_{t=1}^{T}\epsilon_i(t)\begin{pmatrix}1\\(x(t))^\top\end{pmatrix}\right\|_\infty\left\|\begin{pmatrix}u\\v\end{pmatrix}\right\|_1+\lambda\|v_s\|_1-\lambda\|v_{s^c}\|_1\\
&\le 2\left\|\frac{1}{T}\sum_{t=1}^{T}\epsilon_i(t)\begin{pmatrix}1\\(x(t))^\top\end{pmatrix}\right\|_\infty\left\|\begin{pmatrix}u\\v\end{pmatrix}\right\|_1+\lambda\|v_s\|_1-\lambda\|v_{s^c}\|_1\\
&\le 2\left\|\frac{1}{T}\sum_{t=1}^{T}\epsilon_i(t)\begin{pmatrix}1\\(x(t))^\top\end{pmatrix}\right\|_\infty\left\|\begin{pmatrix}u\\v\end{pmatrix}\right\|_1+\lambda\|v_s\|_1-\lambda\|v_{s^c}\|_1
\end{aligned}$$

By Lemma 1.4,

$$P\left(\left\|\frac{1}{T}\sum_{t=1}^{T}\epsilon_i(t)\begin{pmatrix}1\\(x(t))^\top\end{pmatrix}\right\|_\infty\le C\sqrt{\log p/T}\right)\ge 1-c_1\exp(-c_2\log p)$$

Taking $\lambda=4C\sqrt{\log p/T}$ leads that

$$0\le\begin{pmatrix}u & v^\top\end{pmatrix}H\begin{pmatrix}u\\v\end{pmatrix}\le\frac{3\lambda}{2}\|v_s\|_1-\frac{\lambda}{2}\|v_{s^c}\|_1+\frac{1}{2}\lambda\|u\|_1\le\frac{3\lambda}{2}\|\begin{pmatrix}u & v_s\end{pmatrix}\|_1-\frac{\lambda}{2}\|v_{s^c}\|_1$$

Let $\theta=\begin{pmatrix}u\\v\end{pmatrix}\in\mathbb{R}^{p+1}$. Define $\theta_s=(u,v_s)$ and $\theta_{s^c}=(v_{s^c})$. Then,

$$\begin{aligned}
0\le\theta^T H\theta\le\frac{3\lambda}{2}\|\theta_s\|_1-\frac{\lambda}{2}\|\theta_{s^c}\|_1\\
\|\theta_{s^c}\|_1\le 3\|\theta_s\|_1
\end{aligned}$$

Next we introduce Lemma A.1 from (Zheng and Raskutti, 2018) (with only change in notation) to bound the minimum eigen-value of H with high probability.

Lemma A.1: Assume Assumption 1-4 are satisfied and a stationary linear Hawkes model satisfying (6). For any set $J\subset\{1,\ldots,p\}$, H satisfies the following REC

$$\inf\{\theta^\top H\theta:\theta\in\mathcal{C}(J,\kappa),\|\theta\|_2\le 1\}\ge C_1\ge 0$$

with probability at least $1-2\exp(-cT)$, when $|J|\log p\le C_2T$. Here $\mathcal{C}(J,\kappa)=\{\theta:\|\theta_{J^c}\|_1\le\kappa\|\theta_J\|_1\}$, constant C depends on β_i. c and C_2 depend on κ and β_i.

Here we give a sketch of proof. Let $\check{\Gamma} = \mathbb{E}\left(\begin{pmatrix}\mathbf{1}^T \\ x\end{pmatrix}\begin{pmatrix}\mathbf{1}^T & x\end{pmatrix}\right)$. The proof can be split into two parts. First, by Lemma D.6 in (Zheng and Raskutti, 2018) and Lemma S.2 (the concentration bound for the 1st-moment and quadratic form of $x(t)$), we show $\left|\theta^\top\left(H - \check{\Gamma}\right)\theta\right|$ are bounded with high probability for any $\theta \in \mathcal{C}(J,\kappa)$. Since

$$\begin{aligned}
&\inf\{\theta^\top H\theta : \theta \in \mathcal{C}(J,\kappa), \|\theta\|_2 \leq 1\} \\
&\geq \Lambda_{\min}\check{\Gamma} - \sup\left\{\left|\theta^\top\left(H - \check{\Gamma}\right)\theta\right| : \theta \in \mathcal{C}(J,\kappa), \|\theta\|_2 \leq 1\right\} \\
&\geq \frac{1}{2}\Lambda_{\min}\check{\Gamma},
\end{aligned}$$

with probability at least $1 - c_1\exp(-c_2 T)$, the second part is to show $\Lambda_{\min}\left(\check{\Gamma}\right)$ is bounded from 0. By Lemma 1.2, the bounded minimum eigenvalue of Γ_x, we have $\Lambda_{\min}\mathbb{E}\left(x^\top(t)x(t)\right) \geq \Lambda_{\min}\left(\Gamma_x\right) > 0$, which implies that the p-unit multivariate process $\{x_j(t)\}_{1\leq j\leq p}$ are not linearly correlated. In addition, by Assumption 3, the process $x_j(t)$ is not a trivial process of constants, we conclude the minimum eigenvalue of $\check{\Gamma}$ is strictly positive (otherwise, we met contradiction to the result of Lemma 1.2 or Assumption 3).

Therefore, by Lemma A.1,

$$\begin{aligned}
C_1\|\begin{pmatrix}u & v^\top\end{pmatrix}\|_2^2 &\leq \begin{pmatrix}u & v^\top\end{pmatrix} H \begin{pmatrix}u \\ v\end{pmatrix} \\
&\leq \frac{3\lambda}{2}\|\begin{pmatrix}u & v_s^\top\end{pmatrix}\|_1 \\
&\leq \frac{3}{2}4C\sqrt{\frac{\log p}{T}}\sqrt{\rho_i + 1}\|\begin{pmatrix}u & v^\top\end{pmatrix}\|_2 \\
&\leq 6C\sqrt{\frac{\log p}{T}}\sqrt{\rho_i + 1}\|\begin{pmatrix}u & v^\top\end{pmatrix}\|_2
\end{aligned}$$

Then, combining all constants into one term,

$$\begin{aligned}
\|\theta\|_2 &\leq C\sqrt{(\rho_i + 1)\log p/T} \\
\theta^\top H\theta &\leq C(\rho_i + 1)\log p/T \\
\|\theta\|_1 &\leq 4\|\theta_s\|_1 \leq 4\sqrt{\rho_i + 1}\|\theta_s\|_2 \leq C(\rho_i + 1)\sqrt{\log p/T}
\end{aligned}$$

■

Proof of Lemma 1.6

The proof is almost the same as Lemma 1.5 except that 1) we work with the scaled data $(\widetilde{Y}_i(t), \widetilde{x}(t))$ and 2) we use Lemma 1.3 instead of Lemma 1.4 to bound estimation error on $\widehat{w}_j$. ■

Proof of Lemma 1.7

Notice that

$$\begin{aligned}\|\Upsilon_j^{1/2}\widehat{\Upsilon}_j^{-1}\Upsilon_j^{1/2}-I\|_\infty &\le \|\Upsilon_j^{1/2}\widehat{\Upsilon}_j^{-1}\Upsilon_j^{1/2}-I\|_2\\ &\le (\Lambda_{\min}(\Upsilon_j))^{-1}\|\widehat{\Upsilon}_j-\Upsilon_j\|_2\\ &\le dC\|\widehat{\Upsilon}_j-\Upsilon_j\|_\infty\end{aligned}$$

Next, we bound $\|\widehat{\Upsilon}_j-\Upsilon_j\|_\infty$.

Let

$$\widehat{\Upsilon}_j^0=\frac{1}{T}\sum_{t=1}^{T}(x_j/\sigma_i^2(t)-\begin{pmatrix}1 & x_{-j}/\sigma_i^2(t)\end{pmatrix}\widehat{w}_j)(x_j/\sigma_i^2(t)-\begin{pmatrix}1 & x_{-j}/\sigma_i^2(t)\end{pmatrix}\widehat{w}_j)^\top\left(\widetilde{x}_j^*\right)^\top$$

The difference between $\widehat{\Upsilon}_j^0$ and $\widehat{\Upsilon}_j$ is that we replace $\widehat{\sigma}_i^2(t)$ by the true value $\sigma_i^2(t)$. Thus, to bound the difference between $\widehat{\Upsilon}_j-\Upsilon_j$, we bound the two parts $\widehat{\Upsilon}_j^0-\Upsilon_j$ and $\widehat{\Upsilon}_j-\widehat{\Upsilon}_j^0$ separately.

Then, first consider,

$$\begin{aligned}\widehat{\Upsilon}_j^0-\Upsilon_j=&\left(\frac{1}{T}\sum_{t=1}^{T}\widetilde{x}_j^*\left(\widetilde{x}_j^*\right)^\top-\Upsilon_j\right)\\ &+(\widehat{w}_j-w_j^*)^\top\left(\frac{2}{T}\sum_{t=1}^{T}\begin{pmatrix}1\\ \widetilde{x}_{-j}(t)\end{pmatrix}\widetilde{x}_j^*(t)\right)\\ &+(\widehat{w}_j-w_j^*)^\top\left(\frac{1}{T}\sum_{t=1}^{T}\begin{pmatrix}1\\ \widetilde{x}_{-j}(t)\end{pmatrix}\begin{pmatrix}1 & \widetilde{x}_{-j}(t)\end{pmatrix}\right)(\widehat{w}_j-w_j^*)\\ =&E_1+2E_2+E_3\end{aligned}$$

By Assumption 3 and 4 and Lemma S.1, we have $x_j(t)$, $\sigma_i^2(t)$ bounded for all t. In addition, by Lemma S.1, $\widetilde{x}_j^*(t)\le\frac{1}{\sigma_i(t)}\|w_j^*\|_1\|x_j(t)\|_\infty\le C$. Then, by Lemma 1.5 of the lasso estimation consistency,

$$\|E_2\|_2\le\|C(\widehat{w}_j-w_j^*)\|_1\le C\sqrt{\rho_i\vee s_j\frac{\log p}{T}}$$

$$\|E_3\|_2\le C\|\widehat{w}_j-w_j^*\|_1^2\le\rho_i\vee s_j\frac{\log p}{T}$$

with probability at least $1-c_1\exp(-c_2\log p)$.

Next, we look at E_1. WLOG, consider $j=1$,

$$\begin{aligned}\widetilde{x}_j^*(t)\left(\widetilde{x}_j^*(t)\right)^\top&=\left(\widetilde{x}_j(t)-w_{j0}^*-w_{j,-j}^*\widetilde{x}_{-j}(t)\right)\left(\widetilde{x}_j(t)-w_{j0}^*-w_{j,-j}^*\widetilde{x}_{-j}(t)\right)^\top\\ &=\left(\widetilde{x}_j(t)-w_{j,-j}^*\widetilde{x}_{-j}(t)\right)\left(\widetilde{x}_j(t)-w_{j,-j}^*\widetilde{x}_{-j}(t)\right)^\top+\left(w_{j0}^*\right)^2-2w_{j0}^*\left(\widetilde{x}_j(t)-w_{j,-j}^*\widetilde{x}_{-j}(t)\right)\\ &=\begin{pmatrix}1 & -(w_{j,-j}^*)^\top\end{pmatrix}\frac{1}{\sigma_i^2(t)}x(t)x^\top(t)\begin{pmatrix}1\\ -w_{j,-j}^*\end{pmatrix}+\left(w_{j0}^*\right)^2-2w_{j0}^*\frac{1}{\sigma_i(t)}x(t)\begin{pmatrix}1\\ -w_{j,-j}^*\end{pmatrix}\end{aligned}$$

Then,

$$\|E_1\|_\infty = \|\begin{pmatrix}1 & -(w^*_{j,-j})^\top\end{pmatrix}\|_2^2 \| \frac{1}{T}\sum_{t=1}^T \frac{1}{\sigma_i^2(t)} x(t)x^\top(t) - \mathbb{E}\Big(\frac{1}{\sigma_i^2(t)}x(t)x^\top(t)\Big)\|_\infty$$

$$+ 2\|w^*_{j0}\|_2 \| \begin{pmatrix}1 \\ -w^*_{j,-j}\end{pmatrix}\|_2 \| \frac{1}{T}\sum_{t=1}^T \frac{1}{\sigma_i(t)} x(t) - \mathbb{E}\Big(\frac{1}{\sigma_i(t)}x(t)\Big)\|_\infty$$

Therefore, by the 1st and 2nd deviation bound on $x(t)$ shown Lemma S.3 (81) and (80) together with Lemma S.1,

$$\|E_1\|_\infty \le \sqrt{\frac{\log p}{T}}$$

with probability at least $1 - c_1\exp(-c_2\log p)$.

Combining all together,

$$\|\widehat{\Upsilon}_j^0 - \Upsilon_j\|_\infty \le C\sqrt{\frac{\rho_i \vee s_j \log p}{T}}$$

with probability at least $1 - c_1\exp(-c_2\log p)$.

Next, we bound $\widehat{\Upsilon}_j^0 - \widehat{\Upsilon}_j$. Let $D_i = \frac{\widehat{\sigma}_i(t)}{\sigma_i(t)}$. Then,

$$\widehat{\Upsilon}_j^0 - \widehat{\Upsilon}_j = \frac{1}{T}\sum_{t=1}^T (D_i^2 - 1)(\widehat{\widetilde{x}}_j(t) - \widehat{\widetilde{x}}_{-j}(t)\widehat{w}_{j,-j})^2 - 2(D_i - 1)\widehat{w}_{j0}\big(\widehat{\widetilde{x}}_j(t) - \widehat{\widetilde{x}}_{-j}(t)\widehat{w}_{j,-j}\big)$$

$$\le \|D_i^2 - 1\|_\infty \frac{1}{T}\sum_{t=1}^T (\widehat{\widetilde{x}}_j(t) - \widehat{\widetilde{x}}_{-j}(t)\widehat{w}_{j,-j})^2 + 2\|D_i - 1\|_\infty \frac{1}{T}\sum_{t=1}^T \Big|\widehat{w}_{j0}\big(\widehat{\widetilde{x}}_j(t) - \widehat{\widetilde{x}}_{-j}(t)\widehat{w}_{j,-j}\big)\Big|$$

By Assumption 3 of bounded intensity function, the prediction consistency in Lemma 1.5,

$$\begin{aligned}\|D_i^2 - 1\|_\infty &= \|\frac{\widehat{\sigma}_i^2(t) - \sigma_i^2(t)}{\sigma_i^2(t)}\|_\infty \\ &\le C\|\widehat{\lambda}_i(t) - \lambda_i(t)\|_\infty \\ &\le C\sqrt{\rho_i \log p / T}\end{aligned}$$

with probability at least $1 - c_1\exp(-c_2\log p)$.

By Lemma 1.6 together with Lemma S.1 and Assumption 4,

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^T (\widehat{\widetilde{x}}_j(t) - \widehat{\widetilde{x}}_{-j}(t)\widehat{w}_{j,-j})^2 &\le \frac{1}{T}\sum_{t=1}^T \left(\widehat{\widetilde{x}}_j^*(t) + \widehat{w}_{j0}\right)^2 \\ &= \frac{1}{T}\|\widehat{\widetilde{x}}_j^* + \widehat{w}_{j0}\|_2^2 \\ &\le \frac{1}{T}\|\widetilde{x}_j^*\|_2^2 + \frac{1}{T}\|\widetilde{x}_j^* - \widehat{\widetilde{x}}_j^*\|_2^2 + \|w^*_{j0}\|_2^2 + \|w^*_{j0} - \widehat{w}_{j0}\|_2^2 \\ &\le C_1 + C_2\frac{\rho_i \vee s_j \log p}{T} + C_3 + C_4\frac{\rho_i \vee s_j \log p}{T} \le C'\end{aligned}$$

with probability at least $1 - c_1 \exp(-c_2 \log p)$ when $T \gg (\rho_i \vee s_j) \log p$.

Following a similar derivation based on the estimation consistency of lasso for w_j^*, we can show that

$$\frac{1}{T}\sum_{t=1}^{T}\left|\widehat{w}_{j0}\big(\widehat{\widetilde{x}}_j(t) - \widehat{\widetilde{x}}_{-j}(t)\widehat{w}_{j,-j}\big)\right|$$

is bounded with probability at least $1 - c_1 \exp(-c_2 \log p)$.

Then,

$$\widehat{\Upsilon}_j^0 - \widehat{\Upsilon}_j \le C\sqrt{\rho_i \vee s_j \log p/T}$$

with probability at least $1 - c_1 \exp(-c_2 \log p)$ when $T \gg \rho_i \log p$.

Therefore, combining all the above,

$$\|\Upsilon_j^{1/2}\widehat{\Upsilon}_j^{-1}\Upsilon_j^{1/2} - I\|_\infty \le C\sqrt{\rho_i \vee s_j \log p/T}$$

and

$$\|\Upsilon_j^{1/2}\big(\widehat{\Upsilon}_j^0\big)^{-1}\Upsilon_j^{1/2} - I\|_\infty \le C\sqrt{\rho_i \vee s_j \log p/T}$$

with probability at least $1 - c_1 \exp(-c_2 \log p)$. ■

Proof of Lemma 1.8

We start with the proof under null hypothesis, that is to consider $\beta_{ij} = 0$. We actually get the same the result under alternative hypothesis and we discuss this at the end.

Due to the difficulty involving the unknown variance $\sigma_i(t)$ in $\widehat{U}_T$ as defined in (25), we introduce $\widehat{U}_T^0$ which is defined as

$$\widehat{U}_T^0 = \|\widehat{V}_T^0\|_2^2 \tag{73}$$

$$\widehat{V}_T^0 = \sqrt{T}\widehat{\Upsilon}_j^{-1/2}\widehat{S}_{ij}^0 \tag{74}$$

$$\widehat{S}_{ij}^0 = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{\sigma_i(t)}\big(Y_i(t) - \widehat{\mu}_i - x_{-j}(t)\widehat{\beta}_{i,-j}\big)\big(x_j(t)/\sigma_i(t) - \widehat{w}_{j0} - x_{-j}(t)/\sigma_i(t)\widehat{w}_{j,-j}\big) \tag{75}$$

We see that the difference between $\widehat{S}_{ij}^0$ and $\widehat{S}_{ij}$ defined in (23) is that we place $\widehat{\sigma}_i(t)$ by $\sigma_i(t)$.

Based on the technical details in the proof of Theorem 1, we can find that to bound

$\left|\widehat{U}_T - \widehat{U}_T^0\right|$, it is enough to bound $\widehat{S}_{ij} - \widehat{S}_{ij}^0$ and $\widehat{\Upsilon}_j - \widehat{\Upsilon}_j^0$.

By the proof of Lemma 1.7, we have $\left|\widehat{\Upsilon}_j - \widehat{\Upsilon}_j^0\right| \leq \sqrt{\rho_i \log p/T}$ with probability at least $1 - c_1 \exp(-c_2 \log p)$. Therefore, in the follows, we show $\left\|\widehat{S}_{ij} - \widehat{S}_{ij}^0\right\|_2^2$ is bounded by $C\frac{\rho_i \log p}{T}$ with probability at least $1 - c_1 \exp(-c_2 \log p)$.

First, we see the follows:

$$
\begin{aligned}
\widehat{S}_{ij} - \widehat{S}_{ij}^0 &= \frac{1}{T}\sum_{t=1}^{T} \frac{1}{\widehat{\sigma}_i(t)}\big(Y_i(t) - \widehat{\mu}_i - x_{-j}(t)\widehat{\beta}_{i,-j}\big)\big(x_j(t)/\widehat{\sigma}_i(t) - \widehat{w}_{j0} - x_{-j}(t)/\widehat{\sigma}_i(t)\widehat{w}_{j,-j}\big) \\
&- \frac{1}{T}\sum_{t=1}^{T} \frac{1}{\sigma_i(t)}\big(Y_i(t) - \widehat{\mu}_i - x_{-j}(t)\widehat{\beta}_{i,-j}\big)\big(x_j(t)/\sigma_i(t) - \widehat{w}_{j0} - x_{-j}(t)/\sigma_i(t)\widehat{w}_{j,-j}\big) \\
&= \frac{1}{T}\sum_{t=1}^{T} \frac{\sigma_i^2(t)}{\widehat{\sigma}_i^2(t)}\frac{1}{\sigma_i(t)}\big(Y_i(t) - \widehat{\mu}_i - x_{-j}(t)\widehat{\beta}_{i,-j}\big)\big(x_j(t)/\sigma_i(t) - \widehat{w}_{j0}\widehat{\sigma}_i(t)/\sigma_i(t) - x_{-j}(t)/\sigma_i(t)\widehat{w}_{j,-j}\big) \\
&- \frac{1}{T}\sum_{t=1}^{T} \frac{1}{\sigma_i(t)}\big(Y_i(t) - \widehat{\mu}_i - x_{-j}(t)\widehat{\beta}_{i,-j}\big)\big(x_j(t)/\sigma_i(t) - \widehat{w}_{j0} - x_{-j}(t)/\sigma_i(t)\widehat{w}_{j,-j}\big) \\
&= \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)}\big(Y_i(t) - \widehat{\mu}_i - x_{-j}(t)\widehat{\beta}_{i,-j}\big)\big(x_j(t)/\sigma_i(t) - \widehat{w}_{j0} - x_{-j}(t)/\sigma_i(t)\widehat{w}_{j,-j}\big) \\
&+ \frac{1}{T}\sum_{t=1}^{T} \frac{\sigma_i^2(t)}{\widehat{\sigma}_i^2(t)}\frac{1}{\sigma_i(t)}\big(Y_i(t) - \widehat{\mu}_i - x_{-j}(t)\widehat{\beta}_{i,-j}\big)\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)}\big(Y_i(t) - \widehat{\mu}_i - x_{-j}(t)\widehat{\beta}_{i,-j}\big)\big(x_j(t)/\sigma_i(t) - \widehat{w}_{j0} - x_{-j}(t)/\sigma_i(t)\widehat{w}_{j,-j}\big) \\
&+ \frac{1}{T}\sum_{t=1}^{T} \frac{\sigma_i^2(t)}{\widehat{\sigma}_i^2(t)}\frac{1}{\sigma_i(t)}\big(Y_i(t) - \widehat{\mu}_i - x_{-j}(t)\widehat{\beta}_{i,-j}\big)\widehat{w}_{j0}\frac{\sigma_i^2(t) - \widehat{\sigma}_i^2(t)}{\sigma_i(t)\big(\sigma_i(t) + \widehat{\sigma}_i(t)\big)} \\
&\equiv A + B
\end{aligned}
$$

For ease of notation, we take $\beta = \begin{pmatrix}\mu_i \\ \beta_i\end{pmatrix}$ and $x_t = \begin{pmatrix}1 & x(t)\end{pmatrix}$ for short. Then,

$$
\sigma_i^2(t) - \widehat{\sigma}_i^2(t) = x_t\beta(1 - x_t\beta) - x_t\widehat{\beta}(1 - x_t\widehat{\beta}) = (\beta - \widehat{\beta})^\top x_t^\top\big(1 - x_t(\widehat{\beta} + \beta)\big)
$$

Note that $1 - x_t(\widehat{\beta} + \beta)$ is bounded with probability at least $1 - c_1 \exp(-c_2 \log p)$ according to Lemma 1.4 on lasso prediction consistency and Assumption 3 of bounded intensity function. In addition, by Assumption 4 of bounded transition kernel function and Lemma 1.4, with probability at least $1 - c_1 \exp(-c_2 \log p)$,

$$
\|\sigma_i^2(t) - \widehat{\sigma}_i^2(t)\|_\infty \leq \|\beta - \widehat{\beta}\|_1 \|x_t\|_\infty \|1 - x_t(\widehat{\beta} + \beta))\|_\infty \leq C\rho_i\sqrt{\log p/T}
$$

which implies that $\widehat{\sigma}_i^2(t)$ is bounded for all t with probability at least $1 - c_1 \exp(-c_2 \log p)$.

Again for ease of notation, define $\beta_{-j} = \begin{pmatrix} \mu_i \\ \beta_{i,-j} \end{pmatrix}$ and $x_t^{-j} = \begin{pmatrix} 1 & x_{-j}(t) \end{pmatrix}$. Then,

$$\begin{aligned} Y_i(t) - \widehat{\mu}_i - x_{-j}(t)\widehat{\beta}_{i,-j} &= \epsilon_i(t) + (\mu_i - \widehat{\mu}_i) + x_{-j}(t)(\beta_{i,j} - \widehat{\beta}_{i,-j}) \\ &= \epsilon_i(t) + x_t^{-j}(\beta_{-j} - \widehat{\beta}_{-j}) \end{aligned}$$

Then, we write part A as follows:

$$\begin{aligned} A &= \frac{1}{T}\sum_{t=1}^{T} \big(\sigma_i^2(t) - \widehat{\sigma}_i^2(t)\big)\epsilon_i(t)\frac{1}{\widehat{\sigma}_i^2(t)\sigma_i(t)}\big(x_j(t)/\sigma_i(t) - \widehat{w}_{j0} - x_{-j}(t)/\sigma_i(t)\widehat{w}_{j,-j}\big) \\ &+ \frac{1}{T}\sum_{t=1}^{T} \big(\sigma_i^2(t) - \widehat{\sigma}_i^2(t)\big)x_t^{-j}(\beta_{-j} - \widehat{\beta}_{-j})\frac{1}{\widehat{\sigma}_i^2(t)\sigma_i(t)}\big(x_j(t)/\sigma_i(t) - \widehat{w}_{j0} - x_{-j}(t)/\sigma_i(t)\widehat{w}_{j,-j}\big) \\ &= (\beta - \widehat{\beta})^\top \frac{1}{T}\sum_{t=1}^{T} x_t^\top \epsilon_i(t)\big(1 - x_t(\widehat{\beta} + \beta)\big)\frac{1}{\widehat{\sigma}_i^2(t)\sigma_i(t)}\big(x_j(t)/\sigma_i(t) - \widehat{w}_{j0} - x_{-j}(t)/\sigma_i(t)\widehat{w}_{j,-j}\big) \\ &+ (\beta - \widehat{\beta})^\top \left(\frac{1}{T}\sum_{t=1}^{T} x_t^\top \big(1 - x_t(\widehat{\beta} + \beta)\big)\frac{1}{\widehat{\sigma}_i^2(t)\sigma_i(t)}\big(x_j(t)/\sigma_i(t) - \widehat{w}_{j0} - x_{-j}(t)/\sigma_i(t)\widehat{w}_{j,-j}\big)x_t^{-j}\right)(\beta_{-j} - \widehat{\beta}_{-j}) \\ &\equiv A_1 + A_2 \end{aligned}$$

Let $C_i(t) = \big(1 - x_t(\widehat{\beta} + \beta)\big)\frac{1}{\widehat{\sigma}_i^2(t)\sigma_i(t)}\big(x_j(t)/\sigma_i(t) - \widehat{w}_{j0} - x_{-j}(t)/\sigma_i(t)\widehat{w}_{j,-j}\big)$. By Assumption 3 of bounded intensity function, Assumption 4 of bounded transition kernel function and Lemma S.1, we can show that with probability at least $1 - c_1 \exp(-c_2 \log p)$, $C_i(t) \leq C$. Then, by Lemma 1.5 of estimation consistency on β_i,

$$\|A_1\|_\infty \leq \|\widehat{\beta} - \beta\|_1 \|\frac{1}{T}\sum_{t=1}^{T} x_t^\top \epsilon_i(t) C_i(t)\|_\infty \leq C\frac{\rho_i \log p}{T}$$

probability at least $1 - c_1 \exp(-c_2 \log p)$. The last inequality is by Lemma D.4 (van de Geer, 1995) (described in Lemma 1.4) where we take $H(t) = x_t^\top C_i(t)$.

By the same reason for bounding A_1,

$$\|A_2\|_\infty \leq C_i(t)\|x_j(t)\|_\infty^2 \|\widehat{\beta} - \beta\|_2^2 \leq C\frac{\rho_i \log p}{T}$$

Next, we bound part B in a similar fashion.

$$\begin{aligned}
B &= \frac{1}{T}\sum_{t=1}^{T}\big(\epsilon_i(t) + (\beta - \widehat{\beta})x_t^\top\big)\widehat{w}_{j0}\frac{\sigma_i^2(t) - \widehat{\sigma}_i^2(t)}{\widehat{\sigma}_i^2(t)\big(\sigma_i(t) + \widehat{\sigma}_i(t)\big)} \\
&= (\beta - \widehat{\beta})^\top \frac{1}{T}\sum_{t=1}^{T}\epsilon_i(t)x_t^\top\big(1 - x_t(\widehat{\beta} + \beta)\big)\frac{\widehat{w}_{j0}}{\widehat{\sigma}_i^2(t)\big(\sigma_i(t) + \widehat{\sigma}_i(t)\big)} \\
&+ (\beta - \widehat{\beta})^\top\left(\frac{1}{T}\sum_{t=1}^{T}x_t^\top\big(1 - x_t(\widehat{\beta} + \beta)\big)\frac{\widehat{w}_{j0}}{\widehat{\sigma}_i^2(t)\big(\sigma_i(t) + \widehat{\sigma}_i(t)\big)}x_t^{-j}\right)(\beta_{-j} - \widehat{\beta}_{-j}) \\
&\equiv B_1 + B_2
\end{aligned}$$

Similar as bounding part A_1, by the consistency of $\hat{\sigma}_(t)$ to $\sigma_i(t)$ discussed before, Assumption 3 and 4 and Lemma S.1, we can show that $\big(1 - x_t(\widehat{\beta} + \beta)\big)\frac{\widehat{w}_{j0}}{\widehat{\sigma}_i^2(t)\big(\sigma_i(t) + \widehat{\sigma}_i(t)\big)}$ is bounded with probability at least $1 - c_1\exp(-c_2\log p)$. Then, applying Lemma D.4 (van de Geer, 1995),

$$\|B_1\|_2 \leq C\frac{\rho_i\log p}{T}$$

With the same reason bounding B_1 plus Assumption 4 of bounded transition kernel function,

$$\|B_2\|_2 \leq C\frac{\rho_i\log p}{T}$$

Therefore,

$$\|\widehat{S}_{ij} - \widehat{S}_{ij}^0\|_2 \leq C\rho_i\frac{\log p}{T}, \tag{76}$$

with probability at least $1 - c_1\exp(-c_2\log p)$.

Finally, we repeat what we have done in Theorem 1 to bound $\left|\widehat{U}_T^0 - U_T\right|$ and take advantage of the weakly convergence result of $\widehat{U}_T^0$ to χ_d^2, we reach the conclusion that

$$\left|\widehat{U}_T - \widehat{U}_T^0\right| \leq C\left(\frac{\rho_i\log p}{\sqrt{T}}\right)^{1/2}$$

with probability at least $1 - c_1\exp(-c_2\log p) - c_3T^{-1/8} - c_4\left(\frac{\rho_i\log p}{\sqrt{T}}\right)^{1/2}$.

Remark: Although we proof the result under the null hypothesis in the above, for $\phi = -\frac{1}{2}$ and $\phi > -\frac{1}{2}$, we can show that $\widehat{U}_T$ has the same weakly convergence results as $\widehat{U}_T^0$ shown in Theorem 2 under the alternative hypothesis using the same proof steps above but with the weakly convergence results of $\widehat{U}_T^0$ to $\chi_{d,\|\Delta\|_2}^2$ and χ_d^2 respectively for each case.

For $0 < \phi < \frac{1}{2}$, we can repeat the proof in Theorem 2 at (64), we write an extra step as follows, using the result we just show in (76),

$$T\left\|\left(\widehat{\Upsilon}_j\right)^{-\frac{1}{2}}\left(\widehat{S}_{ij} - S_{ij}\right)\right\|_2^2 \geq T\left\|\left(\widehat{\Upsilon}_j^0\right)^{-\frac{1}{2}}\left(\widehat{S}_{ij} - \widehat{S}_{ij}^0 + \widehat{S}_{ij}^0 - S_{ij}\right)\right\|_2^2 \tag{77}$$

$$\geq T\left\|\left(\widehat{\Upsilon}_j^0\right)^{-1/2}\left(C_1 T^{1/2-\phi} - C_2\frac{\rho_i \log p}{\sqrt{T}} - C_3\frac{(s_j \vee \rho_i)\log p}{\sqrt{T}}\right)\right\|_2^2 \tag{78}$$

$$\geq CT^{1-2\phi}. \tag{79}$$

Then, we continue with everything else the same in the proof of Theorem 2 for $0 < \phi < 1/2$ to reach the same result for $P\left(\widehat{U}_T \leq x\right)$ as that for $P\left(\widehat{U}_T^0 \leq x\right)$.

Lemmas S.1 - S.3

Lemma S.1 Let $s_j = \|w_j^*\|_0$ and $s = \max_{j=1,\ldots,p} s_j$. Under Assumption 1-4 and the linear Hawkes model (2), there exist some constant C, s.t.,

$$\|w_j^*\|_2^2 \leq C$$
$$\|w_j^*\|_1 \leq C\sqrt{s_j}$$

Proof of Lemma S.1: By Lemma 1.2 of bounded eigenvalue of Υ_x, we have

$$\|w_j^*\|_2^2 = 1 + \|w_{j,-j}^*\|_2^2 \leq 1 + \Lambda_{\max}(\Upsilon_x^{-1})_{-j,-j}^2 \|(\Upsilon_x)_{j,j}\|_2^2 \leq C$$

Then,

$$\|w_j^*\|_1 \leq \sqrt{s_j}\|w_j^*\|_2 \leq C\sqrt{s_j}$$

■

Lemma S.2: Assume Assumption 1 - 4 are satisfied. Consider the linear Hawkes model follows (2). Then, $\forall \delta > 0$ and $i, j \in \{1, \ldots, p\}$,

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T} x_i(t)x_j(t)^\top - \mathbb{E}\left(\frac{1}{T}\sum_{t=1}^{T} x_i(t)x_j(t)^\top\right)\right| > \delta\right) \leq C_1 \exp(-C_2 \min\{\sqrt{\frac{T}{\rho}}\delta, \frac{T}{\rho}\delta^2\}), \tag{80}$$

where $\rho = \max\|\beta_i\|_0$ and C_1, C_2 are constants.

In addition,

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T} x_j(t) - \mathbb{E}x_j(t)\right| > \delta\right) \leq c_1 \exp(-c_2 T\delta^2). \tag{81}$$

Proof of Lemma S.2:
The key to proof Lemma S.2 is first to write the quadratic form of $x^\top(t)x(t)$ and $x(t)$ into independent noise term $\epsilon_i(t)$, then we bound the deviation based on Hansen-Wright inequality to bound the deviation from the corresponding expectation. Another key technical issue is to bound the l_2-norm of the transformation matrix that links the independent noise part and $x(t)$. This step is made possible by the technical assumptions on the structure of transition functions and also the bounded mean intensity specified in Assumption 3 and 4.

Let $Y_{jt} = (Y_j(t), \dots, Y_j(1))^\top \in \mathbb{R}^\top$. Then $x_j(t) = \sum_{s=1}^{t-1} k_j(t-s)Y_j(s), t > 1$, $x_j(1) = 0$.

Let

$$K_j = \begin{pmatrix} 0 & k_j(1) & k_j(2) & k_j(3) & \dots & \dots & k_j(T-1) \\ 0 & 0 & k_j(1) & k_j(2) & \dots & \dots & k_j(T-2) \\ 0 & 0 & 0 & k_j(1) & \dots & \dots & k_j(T-3) \\ 0 & . & . & . & \dots & \dots & . \\ 0 & 0 & 0 & 0 & 0 & 0 & k_j(1) \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \in \mathbb{R}^{T\times T}$$

Then,

$$x_j = K_j Y_{jT}$$

In addition, by Proposition 1 in (Bacry et al., 2011),

$$Y_j(t) = \Lambda_j + \Psi_j * \epsilon_j(t) = \Lambda_j + \sum_{s=1}^{t-1} \Psi_j(t-s)\epsilon_j(s)$$

Here $\epsilon(s) = (\epsilon_1(s), \dots, \epsilon_p(s))^\top \in \mathbb{R}^p$ and $\Psi_j(t) = (\Psi_{j1}(t), \dots, \Psi_{jp}(t))$, $\Psi_{jl}(t) = \sum_{n=1}^{\infty} \omega_{jl}^{*n}(t)$, where ω_{jl}^{*n} is n-th auto-convolution of ω_{jl}.

Let

$$\Xi_j = \begin{pmatrix} \Psi_j(1) & \Psi_j(2) & \Psi_j(3) & \dots & \dots & \Psi_j(T) \\ 0 & \Psi_j(1) & \Psi_j(2) & \dots & \dots & \Psi_j(T-1) \\ 0 & 0 & \Psi_j(1) & \dots & \dots & \Psi_j(T-2) \\ . & . & . & \dots & \dots & . \\ 0 & 0 & 0 & 0 & 0 & \Psi_j(1) \end{pmatrix} \in \mathbb{R}^{T\times Tp}$$

Let $\epsilon = (\epsilon(T), \dots, \epsilon(1))^\top \in \mathbb{R}^{Tp}$, then

$$Y_{jT} = \Lambda_j + \Xi_j\epsilon$$

Then,

$$x_j = K_j(\Lambda_j + \Xi_j\epsilon) = K_j\Lambda_j + K_j\Xi_j\epsilon$$
$$x_j - Ex_j = K_j\Xi_j\epsilon$$

Then,

$$x_i^\top x_j - Ex_i^\top x_j = \Lambda_i^\top K_i^\top K_j \Xi_j \epsilon + \Lambda_j^\top K_j^\top K_i \Xi_i \epsilon + \epsilon^\top \Xi_i^\top K_i^\top K_j \Xi_j \epsilon$$

Next, we bound each of the items on RHS in the above.

First, notice that

$$\|\Lambda_i^\top K_i^\top K_j \Xi_j\|_2^2 \leq \|\Lambda_i\|_2^2 \Lambda_{\max}\left(K_i^\top K_j \Xi_j \Xi_j^\top K_j^\top K_i\right)$$

where the second inequality is based on Perron–Frobenius theorem and the last inequality is based on Assumption 2 and 4.

By Assumption 3, $\|\Lambda_i\|_2^2 \leq \lambda_{\max} T$. Also,

$$\begin{aligned}\Lambda_{\max}\left(K_i^\top K_j \Xi_j \Xi_j^\top K_j^\top K_i\right) &\leq \Lambda_{\max}\left(K_i\right)^2 \Lambda_{\max}\left(K_j\right)^2 \Lambda_{\max}\left(\Xi_j\right)^2 \\ &\leq \left(\sum_{t=1}^{T} k_i(t)\right)^2 \left(\sum_{t=1}^{T} k_j(t)\right)^2 \left(\sum_{k=1}^{p}\sum_{t=1}^{T} \Psi_{ik}(t)\right)^2 \leq C\end{aligned}$$

where the second inequality is based on Perron–Frobenius theorem and the last inequality is based on Assumption 2 and 4.

Therefore, by the inequality result on sub-gaussian deviation bound (Vershynin 2010, Prop 5.10) and $E\epsilon = 0$, we have

$$P(\left\|\Lambda_i^\top K_i^\top K_j \Xi_j \epsilon\right\|_2 > T\delta) \leq c_1 \exp(-\frac{T^2\delta^2}{\lambda_{\max} TC}) = c_1 \exp(-c_2 T\delta^2)$$

Similarly,

$$P(\left\|\Lambda_j^\top K_j^\top K_i \Xi_i \epsilon\right\|_2 > T\delta) \leq c_1 \exp(-\frac{T^2\delta^2}{\lambda_{\max} TC}) = c_3 \exp(-c_4 T\delta^2)$$

Next, we bound $\epsilon^\top \Xi_i^\top K_i^\top K_j \Xi_j \epsilon$ based on the special structure of K_i and Ξ_i.

By Assumption 4, $k_j(t) \leq a\exp(-bt)$ and $\Psi_{ij} \leq C\exp(-ct)$, where $c = b - a$. By the sparse signal assumption, for each i, there at most s items of $\Psi_{il}(t) \neq 0$. So instead of considering the entire p-unit system, we only consider at most $2s$ units such that $\Psi_{il}(t) \neq 0$ and $\Psi_{jl}(t) \neq 0$. Then, let

$$\widetilde{K} = C_1 \begin{pmatrix} 0 & \exp(-bt) & \exp(-2bt) & \exp(-3bt) & \dots & \dots & \exp(-b(T-1)) \\ 0 & 0 & \exp(-bt) & \exp(-2bt) & \dots & \dots & \exp(-b(T-1)) \\ 0 & 0 & 0 & \exp(-bt) & \dots & \dots & \exp(-b(T-3)) \\ 0 & . & . & . & \dots & \dots & . \\ 0 & 0 & 0 & 0 & 0 & 0 & \exp(-bt) \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \in \mathbb{R}^{T\times T},$$

and

$$\widetilde{\Xi} = C_1 \begin{pmatrix} \mathbf{0}_{2s}^\top & \exp(-ct)\mathbf{1}_{2s}^\top & \exp(-2ct)\mathbf{1}_{2s}^\top & \exp(-3ct)\mathbf{1}_{2s}^\top & \ldots & \ldots & \exp(-c(T-1))\mathbf{1}_{2s}^\top \\ \mathbf{0}_{2s}^\top & \mathbf{0}_{2s}^\top & \exp(-ct)\mathbf{1}_{2s}^\top & \exp(-2ct)\mathbf{1}_{2s}^\top & \ldots & \ldots & \exp(-c(T-2))\mathbf{1}_{2s}^\top \\ \mathbf{0}_{2s}^\top & \mathbf{0}_{2s}^\top & \mathbf{0}_{2s}^\top & \exp(-ct)\mathbf{1}_{2s}^\top & \ldots & \ldots & \exp(-c(T-3))\mathbf{1}_{2s}^\top \\ \mathbf{0}_{2s}^\top & . & . & . & \ldots & \ldots & . \\ \mathbf{0}_{2s}^\top & \mathbf{0}_{2s}^\top & \mathbf{0}_{2s}^\top & \mathbf{0}_{2s}^\top & \mathbf{0}_{2s}^\top & \mathbf{0}_{2s}^\top & \exp(-ct)\mathbf{1}_{2s}^\top \\ \mathbf{0}_{2s}^\top & \mathbf{0}_{2s}^\top & \mathbf{0}_{2s}^\top & \mathbf{0}_{2s}^\top & \mathbf{0}_{2s}^\top & \mathbf{0}_{2s}^\top & \mathbf{0}_{2s}^\top \end{pmatrix} \in \mathbf{R}^{T\times 2sT},$$

where $\mathbf{a}_{2s}^\top = \underbrace{(a, a, \ldots, a)}_{2s}$ and $a \in \{0, 1\}$.

Then,

$$\left\|\Xi_i^\top K_i^\top K_j \Xi_j\right\|_2^2 \le \left\|\widetilde{\Xi}^\top \widetilde{K}^\top \widetilde{K}\widetilde{\Xi}\right\|_2^2$$

Let $\Theta = \widetilde{K}\widetilde{\Xi}$, then we calculate

$$\Theta_{1,k(s-1)+1} = \sum_{s=1}^{k} \exp(-as)\exp(-c(k+1-s)) \le C_2 \exp(-ak)$$

Therefore, $\left\|\widetilde{K}\widetilde{\Xi}\right\|_2^2 \le \left\|\widetilde{\Theta}\right\|_2^2$ where

$$\widetilde{\Theta} = C_2 \begin{pmatrix} \mathbf{0}_{2s}^\top & \exp(-a)\mathbf{1}_{2s}^\top & \exp(-2a)\mathbf{1}_{2s}^\top & \exp(-3a)\mathbf{1}_{2s}^\top & \ldots & \ldots & \exp(-a(T-1))\mathbf{1}_{2s}^\top \\ \mathbf{0}_{2s}^\top & \mathbf{0}_{2s}^\top & \exp(-a)\mathbf{1}_{2s}^\top & \exp(-2a)\mathbf{1}_{2s}^\top & \ldots & \ldots & \exp(-a(T-2))\mathbf{1}_{2s}^\top \\ \mathbf{0}_{2s}^\top & \mathbf{0}_{2s}^\top & \mathbf{0}_{2s}^\top & \exp(-a)\mathbf{1}_{2s}^\top & \ldots & \ldots & \exp(-a(T-3))\mathbf{1}_{2s}^\top \\ \mathbf{0}_{2s}^\top & . & . & . & \ldots & \ldots & . \\ \mathbf{0}_{2s}^\top & \mathbf{0}_{2s}^\top & \mathbf{0}_{2s}^\top & \mathbf{0}_{2s}^\top & \mathbf{0}_{2s}^\top & \mathbf{0}_{2s}^\top & \exp(-a)\mathbf{1}_{2s}^\top \\ \mathbf{0}_{2s}^\top & \mathbf{0}_{2s}^\top & \mathbf{0}_{2s}^\top & \mathbf{0}_{2s}^\top & \mathbf{0}_{2s}^\top & \mathbf{0}_{2s}^\top & \mathbf{0}_{2s}^\top \end{pmatrix} \in \mathbf{R}^{T\times 2sT}.$$

Next, we check $M = \widetilde{\Theta}^\top\widetilde{\Theta}$. Due to the structure of $\widetilde{\Theta}$, we get

$$M = C_2^2 \begin{pmatrix} \mathbf{0}_{2s}^\top & \mathbf{0}_{2s}^\top & \mathbf{0}_{2s}^\top & \mathbf{0}_{2s}^\top & \ldots & \ldots & \mathbf{0}_{2s}^\top \\ \mathbf{0}_{2s}^\top & m_1\mathbf{1}_{2s}^\top & m_1\exp(-a)\mathbf{1}_{2s}^\top & m_1\exp(-2a)\mathbf{1}_{2s}^\top & \ldots & \ldots & m_1\exp(-a(T-2))\mathbf{1}_{2s}^\top \\ . & . & m_2\mathbf{1}_{2s}^\top & m_2\exp(-a)\mathbf{1}_{2s}^\top & \ldots & \ldots & m_2\exp(-a(T-3))\mathbf{1}_{2s}^\top \\ . & . & . & m_3\mathbf{1}_{2s}^\top & \ldots & \ldots & m_3\exp(-a(T-4))\mathbf{1}_{2s}^\top \\ . & . & . & . & \ldots & \ldots & . \\ . & . & . & . & . & . & m_{T-1}\mathbf{1}_{2s}^\top \\ . & . & . & . & . & . & \mathbf{0}_{2s}^\top \end{pmatrix} \in \mathbf{R}^{2sT\times 2sT},$$

where $m_t = \sum_{l=1}^{\top} \exp(-al) \le C_3 \exp(-a)$ for $t = 1, \ldots, T-1$. We only write out the upper triangle part of M in the above since $M = M^T$.

Therefore,

$$\|M\|_2^2 \le 2\sum_{i=1}^{\top} 2s \sum_{k=1}^{T-1} m_i \exp(-ka) \le C_4 sT$$

Since $\{\epsilon_i(t)\}_{1\le i\le p; t=1,\dots,T}$ are mutually independent centered at 0 with bounded variance of each $\epsilon_i(t)$ according to the linear Hawkes model (2), we apply Hanson-Wright inequality and get

$$P\left(\left|\epsilon^\top \Xi_i^\top K_i^\top K_j \Xi_j \epsilon - E\left(\epsilon^\top \Xi_i^\top K_i^\top K_j \Xi_j \epsilon\right)\right| > T\delta\right) \le c_5 \exp(-c_6 \min\{T\delta/\|M\|_2, T^2\delta^2/\|M\|_2^2\})$$
$$\le c_5 \exp(-c_6 \min\{T\delta/\sqrt{C_4 sT}, T^2\delta^2/(C_4 sT)\})$$

Therefore, combining the deviation bound for $\Lambda_i^\top K_i^\top K_j \Xi_j \epsilon$, we get

$$P\left(\left|x_i^\top x_j - Ex_i^\top x_j\right| > T\delta\right) \le P\left(\left|\Lambda_i^\top K_i^\top K_j \Xi_j \epsilon\right| > T\delta\right) \tag{82}$$
$$+ P\left(\left|\Lambda_j^\top K_j^\top K_i \Xi_i \epsilon\right| > T\delta\right) \tag{83}$$
$$+ P\left(\left|\epsilon^\top \Xi_i^\top K_i^\top K_j \Xi_j \epsilon\right| > T\delta\right) \tag{84}$$
$$\le C_1 \exp(-C_2 \min\left\{\sqrt{\frac{T}{s}}\delta, \frac{T}{s}\delta^2\right\}) \tag{85}$$

The above gives a 2nd-order deviation bound. Now we deviate the 1st order deviation bound. Since in the above we have $\|K_j\Xi_j\|_2^2 \le \left\|\widetilde{K}\widetilde{\Xi}\right\|_2^2 \le \left\|\widetilde{\Theta}\right\|_2^2 \le CT\exp(-2a)$, by the inequality result on sub-gaussian deviation bound (Vershyim 2010, Prop 5.10) and $E\epsilon = 0$,

$$P(\left|\frac{1}{T}\sum_{t=1}^{T} x_j(t) - Ex_j(t)\right| > \delta) \le c_1 \exp(-\frac{T^2\delta^2}{TC}) = c_1\exp(-c_2 T\delta^2) \tag{86}$$

■

Corollary S.2: Assume Assumption 1 - 4 are satisfied. Consider the linear Hawkes model follows (2). $\forall u, v \in \mathbf{R}^r$, $x(t) \in \mathbb{R}^r$,

$$P\left(\left|\frac{1}{T}\sum_{t=1}^{T} x_i(t) u^T v x_j(t)^\top - E\left(\frac{1}{T}\sum_{t=1}^{T} x_i(t) u^T v x_j(t)^\top\right)\right| > \delta\right)$$
$$\le C_1 r^2 \exp(-C_2 \min\left\{\sqrt{\frac{T}{s}}\delta\frac{1}{\|u\|_1\|v\|_1}, \frac{T}{s}\delta^2\frac{1}{\|u\|_1^2\|v\|_1^2}\right\}\Bigg).$$

Proof of Corollary S.2:

Notice that

$$\begin{aligned}
&\left|\frac{1}{T}\sum_{t=1}^{T} x_i(t)u^T v x_j(t)^\top - \Big(\frac{1}{T}\sum_{t=1}^{T} x_i(t)u^T v x_j(t)^\top\Big)\right| \\
&= \left|u^T\Big(\frac{1}{T}\sum_{t=1}^{T} x_i^\top(t)x_j(t) - E\Big(\frac{1}{T}\sum_{t=1}^{T} x_i^\top(t)x_j(t)\Big)\Big)v\right| \\
&\leq \|u\|_1\|v\|_1\left\|\frac{1}{T}\sum_{t=1}^{T} x_i^\top(t)x_j(t) - E\Big(\frac{1}{T}\sum_{t=1}^{T} x_i^\top(t)x_j(t)\Big)\right\|_\infty
\end{aligned}$$

Applying Lemma S.2 and taking an union bound, we have

$$\begin{aligned}
&P\bigg(\left\|\frac{1}{T}\sum_{t=1}^{T} x_i^\top(t)x_j(t) - E\Big(\frac{1}{T}\sum_{t=1}^{T} x_i^\top(t)x_j(t)\Big)\right\|_\infty > \delta\bigg) \\
&= P\bigg(\bigcup_{1\leq i,j\leq l}\left\|\frac{1}{T}\sum_{t=1}^{T} x_i^\top(t)x_j(t) - E\Big(\frac{1}{T}\sum_{t=1}^{T} x_i^\top(t)x_j(t)\Big)\right\|_\infty > \delta\bigg) \\
&\leq r^2 C_1 \exp(-C_2 \min\Big\{\sqrt{\frac{T}{s}}\delta, \frac{T}{s}\delta^2\Big\}\Big)
\end{aligned}$$

Therefore, we complete the proof by replacing δ by $\frac{\delta}{\|u\|_1\|v\|_1}$. ■

Proof of Lemma S.3:

Lemma S.3 plays an important role in proof the other technical lemmas. By Lemma S.2 and Corollary S.2, we get the deviation bound for quadratic form of the design columns $x(t)$; however, different from previous work e.g. (Zheng and Raskutti, 2018), in our case, the variance of error term $\sigma^2(t)$ is a function of the mean structure, therefore, we need to deal with a more complex case bounding the deviation of $\frac{1}{T}\sum_{t=1}^{T}\frac{1}{\sigma^2(t)}x^\top(t)x(t)$ from its mean. To achieve this, we use Taylor expansion to expand the term on its 2nd order. Then apply the results on the 1st order and 2nd order (the quadratic form) deviation bound of $x(t)$ to derive the deviation bound for $\frac{1}{T}\sum_{t=1}^{T}\frac{1}{\sigma^2(t)}x^\top(t)x(t)$.

Note that under stationary condition, $\mathbb{E}\big(x(t)\big) = \int_0^t k(s)d\Lambda$, where Λ is mean intensity vector for $Y_1,\ldots,Y_p$. By Assumption 3 and 4, $E\big(x(t)\big)$ is bounded.

First, consider

$$h_{jk}(x(t),\theta = (\mu,\beta)) \equiv \frac{1}{\sigma^2(t)}x_j(t)x_k(t) = \frac{1}{\sigma^2(t)}x(t)I_{jk}x^\top(t)$$

where $I_{jk} = e_j e_K^\top$ and e_i are vector containing all 0's except 1 at position i.

Next, we check derivative of h:

$$\begin{aligned} h'_{jk}(\mathbb{E}\big(x(t)\big),\theta) &= \sigma^{-4}(t)(1-2(\mu+\mathbb{E}\big(x(t)\big)\beta))\beta^\top\mathbb{E}\big(x(t)\big)I_{jk}\mathbb{E}\big(x^\top(t)\big)+\sigma^{-4}(t)\mathbb{E}\big(x(t)\big)(I_{jk}+I_{jk}^T) \\ &= C_1\beta^\top + c_2\mathbb{E}\big(x_j(t)\big)(e_j^T+e_k^T) \\ &= C_1\beta^\top + C_2(e_j^T+e_k^T) \end{aligned}$$

where the last step is because $\mathbb{E}\big(x(t)\big)$ and θ can be regarded as constants independent with t under stationary stochastic process and bounded according to Assumption 3 an 4.

Then,

$$\begin{aligned} \|\frac{1}{T}\sum_{t=1}^{T} h'_{jk}(\mathbb{E}\big(x(t)\big),\theta)\big(x(t)-\mathbb{E}\big(x(t)\big)\big)^\top\|_\infty &= C\|C_1\beta^\top + C_2(e_j^T+e_k^T)\|_1\|\frac{1}{T}\sum_{t=1}^{T} x(t)-\mathbb{E}\big(x(t)\big)\|_\infty \\ &\leq C(C_1\rho\max\{\beta\}+2C_2)\delta \end{aligned}$$

with probability at least $1-c_1\exp(-c_2T\delta^2)$ due to the consistency of $\frac{1}{T}\sum_{t=1}^{T}x(t)$ to $\mathbb{E}\big(x(t)\big)$ proofed in Lemma S.2 (81). For example, $\delta=\sqrt{\frac{\log p}{T}}$

Next, check $h^{(2)}(\mathbb{E}\big(x(t)\big),\theta)$. Similar as above, due to the constancy of $\frac{1}{T}\sum_{t=1}^{T}x(t)$ to $\mathbb{E}\big(x(t)\big)$ and $\theta=(\beta,\mu)$,

$$\begin{aligned} h^{(2)}_{jk}(\mathbb{E}\big(x(t)\big),\theta) &= c_1\beta\beta^\top + c_2(c_3\beta\beta^\top + c_4\beta(e_j^T+e_k^T)) - 2c_5\beta(e_j^T+e_k^T)+2c_6I_{jk} \\ &= C_1\beta\beta^\top + C_2\beta(e_j^T+e_k^T)+C_3(I_{jk}+I_{kj}) \end{aligned}$$

Thus, according to the implication of Assumption 4 that $\max\beta$ is bounded and $T\gg\rho^2$,

$$\|h^{(2)}_{jk}(\mathbb{E}\big(x(t)\big),\theta)\|_1 \leq C_1\rho^2\max\{\beta\}^2 + C_2 2\max\{\beta\}+2C_3 \leq C$$

Then, by the quadratic form deviation bound proofed in Lemma S.2 (80),

$$\begin{aligned} &\|\frac{1}{T}\sum_{t=1}^{T} x(t)h^{(2)}_{jk}(\mathbb{E}\big(x(t)\big),\theta)x^\top(t) - \mathbb{E}\bigg(x(t)h^{(2)}_{jk}(\mathbb{E}\big(x(t)\big),\theta)x^\top(t)\bigg)\|_\infty \\ &\leq \|\frac{1}{T}\sum_{t=1}^{T} x^\top(t)x(t) - \mathbb{E}\bigg(x(t)^\top x(t)\bigg)\|_\infty \|h^{(2)}(\mathbb{E}\big(x(t)\big),\theta)\|_1 \\ &\leq C\delta \end{aligned}$$

with probability at least $1-c_1\exp(-c_2\min\{\sqrt{\frac{T}{s}}\delta,\frac{T}{s}\delta^2\})$.

In addition, by Assumption 3 and Assumption 4 and its implication on bounded β,μ,

$$\begin{aligned} &\|h^{(2)}_{jk}(\mathbb{E}\big(x(t)\big),\theta)\mathbb{E}\big(x^\top(t)\big)\|_1 \\ &\leq \|C_1\beta\beta^\top\mathbb{E}\big(x^\top(t)\big) + C_2\beta(e_j^T+e_k^T)\mathbb{E}\big(x^\top(t)\big)+C_3(I_{jk}+I_{kj})\mathbb{E}\big(x^\top(t)\big)\|_1 \\ &\leq C_1\big(\max|\lambda(t)|+|\mu|\big)\|\beta\|_1 + 2C_2\max\mathbb{E}\big(x(t)\big)+2C_3\max\mathbb{E}\big(x(t)\big) \leq C_1'\rho\max\beta + C_2' \end{aligned}$$

Thus,

$$\begin{aligned}
&\|\frac{1}{T}\sum_{t=1}^{T}\Big(x(t)-\mathbb{E}\big(x(t)\big)\Big)^{\top}h_{jk}^{(2)}(\mathbb{E}\big(x(t)\big),\theta)\mathbb{E}\big(x^{\top}(t)\big)\|_{\infty}\\
&\leq C\|h_{jk}^{(2)}(\mathbb{E}\big(x(t)\big),\theta)\mathbb{E}\big(x^{\top}(t)\big)\|_{1}\|\frac{1}{T}\sum_{t=1}^{T}x(t)-\mathbb{E}\big(x(t)\big)\|_{\infty}\\
&\leq \big(C_1'\rho\max\beta+C_2'\big)\delta
\end{aligned}$$

with probability at least $1-c_1\exp(-c_2\min\{\sqrt{\frac{T}{s}}\delta,\frac{T}{s}\delta^2\})$.

Now, we expand h around $\mathbb{E}\big(x(t)\big)$, we get

$$\begin{aligned}
h_{jk}(x(t),\theta)=&\,h_{jk}(\mathbb{E}\big(x(t)\big),\theta)+h_{jk}'(\mathbb{E}\big(x(t)\big),\theta)\big(x(t)-\mathbb{E}\big(x(t)\big)\big)\\
&+\frac{1}{2}\big(x(t)-\mathbb{E}\big(x(t)\big)\big)h_{jk}^{(2)}(\mathbb{E}\big(x(t)\big),\theta)\big(x(t)-\mathbb{E}\big(x(t)\big)\big)^{\top}\\
&+o(\big(x(t)-\mathbb{E}\big(x(t)\big)\big)\big(x(t)-\mathbb{E}\big(x(t)\big)\big)^{\top})
\end{aligned}$$

Then,

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}h_{jk}(x(t),\theta)-E\Big(h(x(t),\theta)\Big)\\
=&\frac{1}{T}\sum_{t=1}^{T}\big(x(t)-\mathbb{E}\big(x(t)\big)\big)h_{jk}'(\mathbb{E}\big(x(t)\big),\theta)\\
+&\frac{1}{T}\sum_{t=1}^{T}x(t)h_{jk}^{(2)}(\mathbb{E}\big(x(t)\big),\theta)x^{\top}(t)-\mathbb{E}\Big(x(t)h_{jk}^{(2)}(\mathbb{E}\big(x(t)\big),\theta)x^{\top}(t)\Big)\\
-&2\frac{1}{T}\sum_{t=1}^{T}\Big(x(t)-\mathbb{E}\big(x(t)\big)\Big)^{\top}h_{jk}^{(2)}(\mathbb{E}\big(x(t)\big),\theta)\mathbb{E}\big(x^{\top}(t)\big)\\
+&o\Big(\frac{1}{T}\sum_{t=1}^{T}x(t)x^{\top}(t)-\mathbb{E}\big(x(t)x^{\top}(t)\big)-2\frac{1}{T}\sum_{t=1}^{T}\big(x(t)-\mathbb{E}\big(x(t)\big)\big)^{\top}\mathbb{E}\big(x(t)\big)\Big)
\end{aligned}$$

Therefore, combining the deviation bound for each of the item on the RHS, we reach the conclusion that

$$\|\frac{1}{T}\sum_{t=1}^{T}h_{jk}(x(t),\theta)-E\Big(h_{jk}(x(t),\theta)\Big)\|_{\infty}\leq C(\rho,\max\{\beta\})\delta$$

with probability at least $1-c_1\exp(-c_2\min\{\sqrt{\frac{T}{s}}\delta,\frac{T}{s}\delta^2\})$.

Following similar steps above and based on the 1st-order consistency of $x(t)$ to $\mathbb{E}x(t)$ shown in (81), we can also proof

$$\|\frac{1}{T}\sum_{t=1}^{T}\frac{1}{\sigma_i(t)}x_j(t) - E\left(\frac{1}{\sigma_i(t)}x_j(t)\right)\|_\infty \leq C(\rho, \max\{\beta\})\delta$$

with probability at least $1 - c_1\exp(-c_2T\delta^2)$.

Finally, to a conclusion that involves entire $x(t)$, we take a union bound over all p^2 (j,k) pairs for the quadratic form or p variable for the 1st-order deviation bound. The final form of probability actually does not change if we assume $\log p \asymp o(\sqrt{T})$ ■

Let $s_j = \|w_j^*\|_0$ and $s = \max_{1\leq j\leq p} s_j$; $\rho_i = \|\beta_i\|_0$ and $\rho = \max_{1\leq i\leq p} \rho_i$.

Lemma S.4 Assume Assumption 1-4 are satisfied and a stationary linear Hawkes model satisfying (6). For the connectivity matrix of block structure, $s \leq \rho + 1$.

Proof of Lemma S.4:

By the choice of w_j^* in (10),

$$Cov\left(\widetilde{x}_j(t) - \widetilde{x}_{-j}(t)w_{j,-j}^*, \widetilde{x}_{-j}(t)\right) = 0$$

Under stationary condition of the linear Hawkes process and by Assumption 1, as indicated in (Chen et al., 2017), the mean intensity, $\Lambda = (\lambda_1, \ldots, \lambda_p)^\top$, can be written as

$$\Lambda = \sum_{i=1}^{\infty} \Omega^i \mu,$$

where Ω^i is the ith power of the transition matrix Ω. This implies that under stationary condition, $\sigma_j^2(t) = \sigma_j^2 = \lambda_j(1 - \lambda_j)$. Therefore, the choice of w_j^* in (10) implies

$$Cov\left(x_j(t) - x_{-j}(t)w_{j,-j}^*, x_{-j}(t)\right) = 0$$

Then,

$$\begin{aligned} w_{j,-j}^* &= Cov\left(x_j(t), x_{-j}(t)\right)\left(Cov\left(x_{-j}(t), x_{-j}(t)\right)\right)^{-1} \\ &= Cov\left(x_j(t), x_{-j}(t)\right)\left((\Upsilon_x)_{-j,-j}\right)^{-1} \end{aligned}$$

By Lemma 1.2 of the bounded eigenvalue of Υ_x,

$$\|w_{j,-j}^*\|_0 \leq \|\{k : j \neq k, Cov\left(x_j(t), x_k(t)\right) \neq 0\}\|_0$$

Without loss of generality, consider a discrete time scenario with unit time window ($dt = 1$). Then, $x_j(t) = \sum_{s=1}^{t-1} k_j(t-s)Y_j(s)$ and $x_j(1) = 0$.

$$\begin{aligned} Cov\left(x_j(t), x_k(t)\right) &= Cov\left(\sum_{s=1}^{t-1} k_j(t-s)Y_j(s), \sum_{s=1}^{t-1} k_j(t-s)Y_j(s)\right) \\ &= \sum_{s=1}^{t-1}\sum_{s'=1}^{t-1} k_j(t-s)k_j(t-s')Cov\left(Y_j(s), Y_j(s')\right) \end{aligned}$$

Therefore,

$$\|\{k : j \neq k, Cov\left(x_j(t), x_k(t)\right) \neq 0\}\|_0 \leq \|\{k : j \neq k, Cov\left(Y_j(t), Y_k(t)\right) \neq 0\}\|_0.$$

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,

$$\|\{k : j \neq k, Cov\left(Y_j(t), Y_k(t)\right) \neq 0\}\|_0 \leq \rho,$$

which implies

$$s = \max_{1 \leq j \leq p} \|w_j^*\|_0 \leq 1 + \max_{1 \leq j \leq p} \|w_{j,-j}^*\|_0 \leq \rho + 1$$

■

Remark: The lemma above makes use of the block structure of a connectivity matrix where units are correlated within clusters. Therefore, to analyze the sparsity of w_j^*, we only need to check the size of cluster that unit j is in. Due to the sparsity assumption on the connectivity matrix, the largest size of cluster is upto ρ. For a general structure connectivity matrix, the order between sparsity of w_j^* and the sparsity of β is not straightforward. The sparsity of w_j^* depends on the sign and scale of the connectivity coefficients and also the transition kernel.