Electronic Journal of Statistics
ISSN: 1935-7524

# Nearly Assumptionless Screening for the Mutually-Exciting Multivariate Hawkes Process

**Shizhe Chen**

*Department of Statistics, Columbia University, New York, NY 10027*
*e-mail:* shizhe.chen@gmail.com

**Daniela Witten*, and Ali Shojaie†**

*Department of Biostatistics and Statistics, University of Washington, Seattle, WA 98195*
*e-mail:* dwitten@u.washington.edu; ashojaie@u.washington.edu

**Abstract:** We consider the task of learning the structure of the graph underlying a mutually-exciting multivariate Hawkes process in the high-dimensional setting. We propose a simple and computationally inexpensive *edge screening* approach. Under a subset of the assumptions required for penalized estimation approaches to recover the graph, this edge screening approach has the sure screening property: with high probability, the screened edge set is a superset of the true edge set. Furthermore, the screened edge set is relatively small. We illustrate the performance of this new edge screening approach in simulation studies.

**MSC 2010 subject classifications:** Primary 60G55; secondary 62M10, 62H12.
**Keywords and phrases:** Hawkes process, screening, high-dimensionality.

## 1. Introduction

### 1.1. Overview of the Multivariate Hawkes Process

In a seminal paper, Hawkes (1971) proposed the *multivariate Hawkes process*, a multivariate point process model in which a past event may trigger the occurrence of future events. The Hawkes process and its variants have been widely applied to model recurrent events, with notable applications in modeling earthquakes (Ogata, 1988), crime rates (Mohler et al., 2011), interactions in social networks (Simma and Jordan, 2012; Perry and Wolfe, 2013; Zhou, Zha and Song, 2013a,b), financial events (Chavez-Demoulin, Davison and McNeil, 2005; Bowsher, 2007; Aït-Sahalia, Cacho-Diaz and Laeven, 2015), and spiking histories of neurons (see e.g., Brillinger, 1988; Okatan, Wilson and Brown, 2005; Paninski, Pillow and Lewi, 2007; Pillow et al., 2008).

*D.W. was supported by NIH Grant DP5OD009145, NSF CAREER Award DMS-1252624, and an Alfred P. Sloan Foundation Research Fellowship.
†A.S. was supported by NSF grant DMS-1561814 and NIH grants 1K01HL124050-01A1 and 1R01GM114029-01A1.

In this section, we provide a very brief review of the multivariate Hawkes process. A more comprehensive discussion can be found in Liniger (2009) and Zhu (2013).

Following Brémaud and Massoulié (1996), we define a simple point process $N$ on $\mathbb{R}^+$ as a family $\{N(A)\}_{A\in\mathcal{B}(\mathbb{R}^+)}$ taking integer values (including positive infinity), where $\mathcal{B}(\mathbb{R}^+)$ denotes the Borel $\sigma$-algebra of the positive half of the real line. Further let $t_1, t_2, \ldots \in \mathbb{R}^+$ be the event times of $N$. In this notation, $N(A) = \sum_i \mathbb{1}_{[t_i\in A]}$ for $A \in \mathcal{B}(\mathbb{R}^+)$. We write $N\big([t, t+\mathrm{d}t)\big)$ as $\mathrm{d}N(t)$, where $\mathrm{d}t$ denotes an arbitrary small increment of $t$. Let $\mathcal{H}_t$ be the *history* of $N$ up to time $t$. Then the $\mathcal{H}_t$-predictable *intensity* process of $N$ is defined as

$$\lambda(t)\mathrm{d}t = \mathbb{P}(\mathrm{d}N(t) = 1 \mid \mathcal{H}_t). \tag{1}$$

Now suppose that $N$ is a *marked* point process, in which each event time $t_i$ is associated with a mark $m_i \in \{1, \ldots, p\}$ (see e.g., Definition 6.4.I. in Daley and Vere-Jones, 2003). We can then view $N$ as a *multivariate* point process $\big(N_j\big)_{j=1,\ldots,p}$, of which the $j$th component process is given by $N_j(A) = \sum_i \mathbb{1}_{[t_i\in A, m_i=j]}$ for $A \in \mathcal{B}(\mathbb{R}^+)$. To simplify the notation, we let $t_{j,1}, t_{j,2}, \ldots \in \mathbb{R}^+$ denote the event times of $N_j$.

The intensity of the $j$th component process is

$$\lambda_j(t)\mathrm{d}t = \mathbb{P}(\mathrm{d}N_j(t) = 1 \mid \mathcal{H}_t).$$

In the case of the *linear* Hawkes process, this function takes the form (Brémaud and Massoulié, 1996; Hansen, Reynaud-Bouret and Rivoirard, 2015)

$$\lambda_j(t) = \mu_j + \sum_{k=1}^{p} \left( \sum_{i:t_{k,i}\le t} \omega_{j,k}(t - t_{k,i}) \right). \tag{2}$$

We refer to $\mu_j \in \mathbb{R}$ as the *background intensity*, and $\omega_{j,k}(\cdot) : \mathbb{R}^+ \mapsto \mathbb{R}$ as the *transfer function*.

For $p$ fixed, Brémaud and Massoulié (1996) established that the linear Hawkes process with intensity function (2) is stationary given the following assumption.

**Assumption 1** *Let $\mathbf{\Omega}$ be a $p\times p$ matrix whose entries are $\Omega_{j,k} = \int_0^\infty |\omega_{j,k}(\Delta)|\,\mathrm{d}\Delta$, for $j, k = 1, \ldots, p$. We assume that the spectral norm of $\mathbf{\Omega}$ is strictly less than 1, i.e., $\Gamma_{\max}(\mathbf{\Omega}) \le \gamma_\Omega < 1$, where $\gamma_\Omega$ is a generic constant.*

We now define a directed graph with node set $\{1, \ldots, p\}$ and edge set

$$\mathcal{E} \equiv \{(j, k) : \omega_{j,k} \neq 0, 1 \le j, k \le p\}, \tag{3}$$

for $\omega_{j,k}$ given in (2). Let

$$s \equiv \max_{1\le j\le p} \sum_{k=1}^{p} \mathbb{1}_{\{(j,k)\in\mathcal{E}\}} \tag{4}$$

denote the maximum in-degree of the nodes in the graph. In this paper, we propose a simple screening procedure that can be used to obtain a small superset of the edge set $\mathcal{E}$.

### 1.2. Estimation and Theory for the Hawkes Process

We first consider the low-dimensional setting, in which the dimension of the process, $p$, is fixed, and $T$, the time period during which the point process is observed, is allowed to grow. In this setting, asymptotic properties such as the central limit theorem have been established; for instance, see Bacry et al. (2013) and Zhu (2013). Consequently, estimating the edge set $\mathcal{E}$ is straightforward in low dimensions.

In high dimensions, when $p$ might be large, we can fit the Hawkes process model using a penalized estimator of the form

$$\underset{\omega_{j,k}\in\mathcal{F},1\leq j,k\leq p}{\text{minimize}} \quad \mathcal{L}\left(\omega_{j,k};\{N_j\}_{j=1}^p\right)+\lambda\sum_{j,k}\mathcal{P}\left(\omega_{j,k};\{N_j\}_{j=1}^p\right), \tag{5}$$

where $\mathcal{L}\left(\cdot\,;\{N_j\}_{j=1}^p\right)$ is a *loss function*, based on, e.g., the log-likelihood (Bacry, Gaïffas and Muzy, 2015) or least squares (Hansen, Reynaud-Bouret and Rivoirard, 2015); $\mathcal{P}\left(\cdot\,;\{N_j\}_{j=1}^p\right)$ is a *penalty function*, such as the lasso (Hansen, Reynaud-Bouret and Rivoirard, 2015); $\lambda$ is a nonnegative tuning parameter; and $\mathcal{F}$ is a suitable function class. Then, a natural estimator for $\mathcal{E}$ is $\{(j,k):\hat{\omega}_{j,k}\neq 0\}$.

Recently, Reynaud-Bouret and Schbath (2010), Bacry, Gaïffas and Muzy (2015), and Hansen, Reynaud-Bouret and Rivoirard (2015) have established that under certain assumptions, penalized estimation approaches of the form (5) are consistent in high dimensions, provided that the edge set $\mathcal{E}$ is sparse. For instance, Hansen, Reynaud-Bouret and Rivoirard (2015) establish the oracle inequality of the lasso estimator for the Hawkes process, given that certain conditions hold on the observed event times. However, to show that these conditions hold with high probability for arbitrary samples, these theoretical results require that the point process is *mutually-exciting* — that is, an event in one component process can *increase*, but cannot decrease, the probability of an event in another component process. This amounts to assuming that $\omega_{j,k}(\Delta)\geq 0$ for all $\Delta\geq 0$, for $\omega_{j,k}$ defined in (1).

When the dimension $p$ is large, penalized estimation procedures of the form (5) (Bacry, Gaïffas and Muzy, 2015; Hansen, Reynaud-Bouret and Rivoirard, 2015) become computationally expensive: they require $\mathcal{O}(Tp^2)$ operations *per iteration* in an iterative algorithm. This is problematic in contemporary applications, in which $p$ can be on the order of tens of thousands (Ahrens et al., 2013). These concerns motivate us to propose a simple and computationally-efficient edge screening procedure for estimating the true edge set $\mathcal{E}$ in high dimensions. Under very few assumptions, our proposed screening procedure is guaranteed to select a small superset of the true edge set $\mathcal{E}$.

### 1.3. Organization of Paper

The rest of this paper proceeds as follows. In Section 2, we introduce our screening procedure for estimating the edge set $\mathcal{E}$, and establish its theoretical properties. We present simulation results in support of our proposed procedure in

Section 3. Proofs of theoretical results are presented in Section 4, and the Discussion is in Section 5.

## 2. An Edge Screening Procedure

### 2.1. Approach

For $j = 1, \ldots, p$, let $\Lambda_j$ denote the *mean intensity* of the $j$th point process introduced in Section 1. That is,

$$\Lambda_j \equiv \mathbb{E}[\mathrm{d}N_j(t)]/\mathrm{d}t. \tag{6}$$

Following Equation 5 of Hawkes (1971), for any $\Delta \in \mathbb{R}$, the *(infinitesimal) cross-covariance* of the $j$th and $k$th processes is defined as

$$V_{j,k}(\Delta) \equiv \begin{cases} \mathbb{E}[\mathrm{d}N_j(t)\mathrm{d}N_k(t-\Delta)]/\{\mathrm{d}t\mathrm{d}(t-\Delta)\} - \Lambda_j\Lambda_k & j \neq k \\ \mathbb{E}[\mathrm{d}N_k(t)\mathrm{d}N_k(t-\Delta)]/\{\mathrm{d}t\mathrm{d}(t-\Delta)\} - \Lambda_k^2 - \Lambda_k\delta(\Delta) & j = k \end{cases}, \tag{7}$$

where $\delta(\cdot)$ is the Dirac delta function, which satisfies $\int_{-\infty}^{\infty} \delta(x)\mathrm{d}x = 1$ and $\delta(x) = 0$ for $x \neq 0$.

For a given value of $\Delta$, we can estimate the cross-covariance function $V_{j,k}(\Delta)$ using kernel smoothing:

$$\widehat{V}_{j,k}(\Delta) = \begin{cases} \frac{1}{Th}\iint_{[0,T]^2} K\left(\frac{(t'-t)+\Delta}{h}\right)\mathrm{d}N_j(t)\mathrm{d}N_k(t') - \frac{1}{T^2}N_j([0,T])N_k([0,T]) & j \neq k \\ \frac{1}{Th}\iint_{[0,T]^2\setminus\{t=t'\}} K\left(\frac{(t'-t)+\Delta}{h}\right)\mathrm{d}N_k(t)\mathrm{d}N_k(t') - \frac{1}{T^2}N_k^2([0,T]) & j = k \end{cases}, \tag{8}$$

where $K(\cdot)$ is a kernel function with bandwidth $h$, and $\int_0^T f(t)\mathrm{d}N_j(t)$ is the Stieltjes integral, defined as

$$\int_0^T f(t)\mathrm{d}N_j(t) \equiv \sum_{i:t_{j,i}\in[0,T]} f(t_{j,i}).$$

In this paper, we focus on kernel functions that are bounded by 1 and are defined on a bounded support, i.e., $0 \leq K(x/h) \leq 1$ for $x \in [-h, h]$, and $K(x/h) = 0$ for $x \notin [-h, h]$ (e.g., the Epanechnikov kernel).

Let $B$ denote a tuning parameter that defines the time range of interest for $V_{j,k}(\Delta)$, i.e. $\Delta \in [-B, B]$. For any $\zeta$, we define the set of screened edges as

$$\widehat{\mathcal{E}}(\zeta) \equiv \{(j,k) : \|\widehat{V}_{j,k}\|_{2,[-B,B]} > \zeta\}, \tag{9}$$

where $\|f\|_{2,[l,u]} \equiv \left\{\int_l^u f^2(t)\mathrm{d}t\right\}^{1/2}$ is the $\ell_2$-norm of a function $f$ on the interval $[l, u]$.

The screened edge set $\widehat{\mathcal{E}}(\zeta)$ in (9) can be calculated quickly: $\|\widehat{V}_{j,k}\|_{2,[-B,B]}$ can be calculated in $\mathcal{O}(T)$ computations, and so $\widehat{\mathcal{E}}(\zeta)$ can be calculated in $\mathcal{O}(Tp^2)$ computations. The procedure can be easily parallellized.

There are three tuning parameters in the procedure: the bandwidth $h$ in (8), the range $B$ in (9), and the screening threshold $\zeta$ in (9). The bandwidth $h$ can be chosen by cross-validation. The range $B$ can be selected based on the problem setting. For instance, when using the multivariate Hawkes process to model a spike train data set in neuroscience, we can set $B$ to equal the maximum time gap between a spike and the spike it can possibly evoke. The choice of screening threshold $\zeta$ can be determined based on the sparsity level that we expect based on our prior knowledge. Alternatively, we may wish to use a small value of $\zeta$ in order to reduce the chance of false negative edges in $\widehat{\mathcal{E}}(\zeta)$, or a larger value due to limited computational resources in our downstream analysis.

### 2.2. Theoretical Results

We consider the asymptotics of triangular arrays (Greenshtein and Ritov, 2004), where the dimension $p$ is allowed to grow with $T$. When unrestricted, it is possible to cook up extreme networks, where, for instance, the mean intensity $\Lambda_j$ in (6) diverges to infinity. To avoid such cases, we pose the following regularity assumption.

**Assumption 2** *There exist positive constants $\Lambda_{\min}$, $\Lambda_{\max}$, and $V_{\max}$ such that $0 < \Lambda_{\min} \le \Lambda_j \le \Lambda_{\max}$ and $\max_{\Delta\in\mathbb{R}} |V_{j,k}(\Delta)| \le V_{\max}$ for all $1 \le j, k \le p$, where $\Lambda_j$ and $V_{j,k}$ are defined in (6) and (7), respectively. Furthermore, $\Lambda_{\min}$, $\Lambda_{\max}$, and $V_{\max}$ are generic constants that do not depend on $p$.*

Next, we make some standard assumptions on the transfer functions $\omega_{j,k}$ in (2).

**Assumption 3** *The following hold:*

*(a) The transfer functions are non-negative: $\omega_{j,k}(\Delta) \ge 0$ for all $\Delta \ge 0$.*

*(b) There exists a positive constant $\beta_{\min}$ such that*

$$\min_{(j,k):\omega_{j,k}\neq 0} \left( \int_0^\infty \omega_{j,k}^2(\Delta) \mathrm{d}\Delta \right) \ge \beta_{\min}^2.$$

*(c) There exist positive constants $b$, $\theta_0$, and $C$ such that, for all $1 \le j, k \le p$, and for any $\Delta_1$, $\Delta_2 \in \mathbb{R}$, $\mathrm{supp}(\omega_{j,k}) \subset (0, b]$, $\max_\Delta |\omega_{j,k}(\Delta)| \le C$, and $|\omega_{j,k}(\Delta_1) - \omega_{j,k}(\Delta_2)| \le \theta_0 |\Delta_1 - \Delta_2|$.*

Assumption 3(a) guarantees that the multivariate Hawkes process is mutually-exciting: that is, an event may trigger (but cannot inhibit) future events. This assumption is shared by the original proposal of Hawkes (1971). Furthermore, existing theory for penalized estimators for the Hawkes process requires this assumption (Bacry, Gaïffas and Muzy, 2015; Hansen, Reynaud-Bouret and Rivoirard, 2015).

Assumption 3(b) guarantees that the non-zero transfer functions are non-negligible. Such an assumption is needed in order to establish variable selection consistency (Bühlmann and van de Geer, 2011; Wainwright, 2009) for the penalized estimator (5).

Assumption 3(c) guarantees that the transfer functions are sufficiently smooth; this guarantees that the cross-covariances are smooth (see Section A.2 in Appendix), and hence can be estimated using a kernel smoother (8). Instead of Assumption 3(c), we could assume that $\omega_{j,k}$ is an exponential function (Bacry, Gaïffas and Muzy, 2015) or that it is well-approximated by a set of smooth basis functions (Hansen, Reynaud-Bouret and Rivoirard, 2015).

Recall that $s$ was defined in (4). We now state our main result.

**Theorem 1** *Suppose that the Hawkes process (2) satisfies Assumptions 1–3. Let $h = c_1 s^{-1/2} T^{-1/6}$ in (8) and $\zeta = 2c_2 s^{1/2} T^{-1/6}$ in (9) for some constants $c_1$ and $c_2$. Then, for some positive constants $c_3$ and $c_4$, with probability at least $1 - c_3 T^{7/6} s^{1/2} p^2 \exp(-c_4 T^{1/6})$,*

*(a)* $\mathcal{E} \subset \widehat{\mathcal{E}}(\zeta)$;

*(b)* $card\left(\widehat{\mathcal{E}}(\zeta)\right) = \mathcal{O}\left(card(\mathcal{E}) s^{-1} T^{1/3} \gamma_\Omega (1-\gamma_\Omega)^{-2} \Lambda_{\max}^2\right)$.

Theorem 1(a) guarantees that, with high probability, the screened edge set $\widehat{\mathcal{E}}(\zeta)$ contains the true edge set $\mathcal{E}$. Therefore, screening does not result in false negatives. This is referred to as the *sure screening property* in the literature (Fan and Lv, 2008; Fan, Samworth and Wu, 2009; Fan and Song, 2010; Fan, Feng and Song, 2011; Fan, Ma and Dai, 2014; Liu, Li and Wu, 2014; Song et al., 2014; Luo, Song and Witten, 2014). Typically, establishing the sure screening property requires assuming that the *marginal* association between a pair of nodes in $\mathcal{E}$ is sufficiently large; see e.g. Condition 3 in Fan and Lv (2008) and Condition C in Fan, Feng and Song (2011). In contrast, Theorem 1(a) requires only that the *conditional* association between a pair of nodes in $\mathcal{E}$ is sufficiently large; see Assumption 3(b).

Theorem 1(b) guarantees that $\widehat{\mathcal{E}}(\zeta)$ is a relatively small set, on the order of $\mathcal{O}(\text{card}(\mathcal{E}) s^{-1} T^{1/3})$. Suppose that $p^2 \propto s^{-1/2} \exp(c_4 T^{1/6-\epsilon})$ for some positive constant $\epsilon < 1/6$; this is the high-dimensional regime, in which the probability statement in Theorem 1 converges to one. Then the size of $\widehat{\mathcal{E}}(\zeta)$, $\mathcal{O}(\text{card}(\mathcal{E}) s^{-1} T^{1/3})$, can be much smaller than $p^2$, the total number of node pairs. We note that the rate of $T^{1/3}$ is comparable to existing results for *non-parametric* screening in the literature (see e.g., Fan, Feng and Song 2011; Fan, Ma and Dai 2014).

To summarize, Theorem 1 guarantees that *under a small subset of the assumptions required for penalized estimation methods to recover the edge set $\mathcal{E}$*, the screened edge set $\widehat{\mathcal{E}}(\zeta)$ (9) is small and contains no false negatives. We note that this is not the case for other types of models. For instance, in the case of the Gaussian graphical model, Luo, Song and Witten (2014) considered estimating the conditional dependence graph by screening the marginal covariances. In order for this procedure to have the sure screening property, one must make an assumption on the minimum marginal covariance associated with an edge in the graph, which is not required for variable selection consistency of penalized estimators (Cai, Liu and Luo, 2011; Luo, Song and Witten, 2014; Ravikumar et al., 2011; Saegusa and Shojaie, 2016).

It is important to note that Theorem 1 considers an *oracle* procedure, where

the tuning parameters depend on unknown parameters. The heuristic selection guidelines suggested at the end of Section 2.1 may not satisfy the requirements of Theorem 1. We leave the discussion of optimal tuning parameter selection criteria for future research. Also, note that the bandwidth $h \propto T^{-1/6}$ is wider than the typical bandwidth for kernel smoothing, which is $T^{-1/3}$ (Tsybakov, 2009). This is because we aim to minimize a concentration bound on $\widehat{V}_{j,k} - V_{j,k}$ (see the proof of Lemma 3 in the Appendix), rather than the usual mean integrated square error as in, e.g., Theorem 1.1 in Tsybakov (2009).

**Remark 1** *In light of Theorem 1, consider applying a constraint induced by* $\widehat{\mathcal{E}}(\zeta)$ *to* (5)*:*

$$\begin{aligned} \underset{\omega_{j,k}\in\mathcal{F}, 1\le j,k\le p}{\text{minimize}} \quad & \mathcal{L}\left(\omega_{j,k}; \{N_j\}_{j=1}^p\right) + \lambda \sum_{j,k} \mathcal{P}\left(\omega_{j,k}; \{N_j\}_{j=1}^p\right) \\ \text{subject to} \quad & \omega_{j,k} = 0 \text{ for } (j,k) \notin \widehat{\mathcal{E}}(\zeta). \end{aligned} \tag{10}$$

*Theorem 1 can be combined with existing results on consistency of penalized estimators of the Hawkes process (Bacry, Gaïffas and Muzy, 2015; Hansen, Reynaud-Bouret and Rivoirard, 2015) in order to establish that* (10) *results in consistent estimation of the transfer functions* $\omega_{j,k}$*. As a concrete example, Hansen, Reynaud-Bouret and Rivoirard (2015) considered* (10) *with* $\mathcal{L}\left(\omega_{j,k}; \{N_j\}_{j=1}^p\right)$ *taken to be the least-squares loss, and* $\mathcal{P}\left(\omega_{j,k}; \{N_j\}_{j=1}^p\right)$ *a lasso-type penalty. Our simulation experiments in Section 3 indicate that in this setting,* (10) *can actually have better small-sample performance than* (5) *when* $p$ *is very large. Furthermore, solving* (10) *can be much faster than solving* (5)*: the former requires* $\mathcal{O}(T^{4/3}s^{-1}\,card(\mathcal{E}))$ *computations* per iteration*, compared to* $\mathcal{O}(Tp^2)$ *per iteration for the latter (using e.g. coordinate descent, Friedman, Hastie and Tibshirani, 2010). In the high-dimensional regime when* $p^2 \propto s^{-1/2}\exp(c_4 T^{1/6-\epsilon})$ *for some positive constant* $\epsilon < 1/6$*, we have that* $T^{4/3}s^{-1}\,card(\mathcal{E}) \ll Tp^2$*. We note that in order to solve* (10)*, we must first compute* $\widehat{\mathcal{E}}(\zeta)$*, which requires an additional one-time computational cost of* $\mathcal{O}(Tp^2)$*.*

## 3. Simulation

### 3.1. Simulation Set-Up

In this section, we investigate the performance of our screening procedure in a simulation study with $p = 100$ point processes. Intensity functions are given by (2), with $\mu_j = 0.75$ for $j = 1, \ldots, p$, and $\omega_{j,k}(t) = 2t\exp(1 - 5t)$ for $(j,k) \in \mathcal{E}$. By definition, $\omega_{j,k} = 0$ for all $(j,k) \notin \mathcal{E}$. We consider two settings for the edge set $\mathcal{E}$, Setting A and Setting B. These settings are displayed in Figure 1.

In what follows, it will be useful to think about the (undirected) node pairs as belonging to three types. (i) We let

$$\tilde{\mathcal{E}} \equiv \{(j,k) : (j,k) \in \mathcal{E} \text{ or } (k,j) \in \mathcal{E}\}. \tag{11}$$

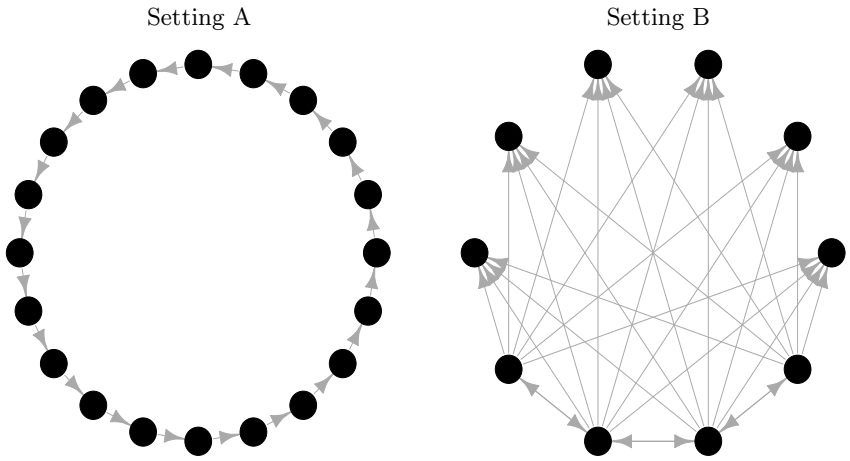


FIG 1. *Left:* In Setting A, the edge set $\mathcal{E}$ is composed of 5 connected components, each of which is a chain graph containing 20 nodes. *Right:* In Setting B, $\mathcal{E}$ is composed of 10 connected components, each of which contains 10 nodes.

(ii) With a slight abuse of notation, we will use $\tilde{\mathcal{E}}^c \cap \text{supp}(\mathbf{V})$ to denote node pairs that are not in $\tilde{\mathcal{E}}$ with non-zero population cross-covariance, defined in (7). (iii) Continuing to slightly abuse notation, we will use $\tilde{\mathcal{E}}^c \backslash \text{supp}(\mathbf{V})$ to denote node pairs that are not in $\tilde{\mathcal{E}}$ and that have zero population cross-covariance.

Throughout the simulation, we set the bandwidth $h$ in (8) to equal $T^{-1/6}$, and the range of interest $B$ in (9) to equal 5. Thus, $h$ satisfies the requirements of Theorem 1, and $[-B, B]$ covers the majority of the mass of the transfer function $\omega_{j,k}$. However, these simulation results are not sensitive to the particular choices of $h$ or $B$.

### *3.2. Investigation of the Estimated Cross-Covariances*

In Setting A, within a single connected component, all of the node pairs that are not in $\tilde{\mathcal{E}}$ are in $\tilde{\mathcal{E}}^c \cap \text{supp}(\mathbf{V})$. However, for the most part, the population cross-covariances corresponding to node pairs in $\tilde{\mathcal{E}}^c \cap \text{supp}(\mathbf{V})$ are quite small, because they are induced by paths of length two and greater. This can be seen from the left-hand panel of Figure 2. Given the left-hand panel of Figure 2, we expect the proposed screening procedure to work very well in Setting A: for a sufficiently large value of the time period $T$, there exists a value of $\zeta$ such that, with high probability, $\widehat{\mathcal{E}}(\zeta) = \tilde{\mathcal{E}}$.

In Setting B, six nodes receive directed edges from the same set of four nodes. Therefore, we expect the pairs among these six nodes to be in the set $\tilde{\mathcal{E}}^c \cap \text{supp}(\mathbf{V})$, and to have substantial population cross-covariances. This intuition is supported by the center panel of Figure 2, which indicates that the node pairs in $\tilde{\mathcal{E}}^c \cap \text{supp}(\mathbf{V})$ have relatively large estimated cross-covariances, on the

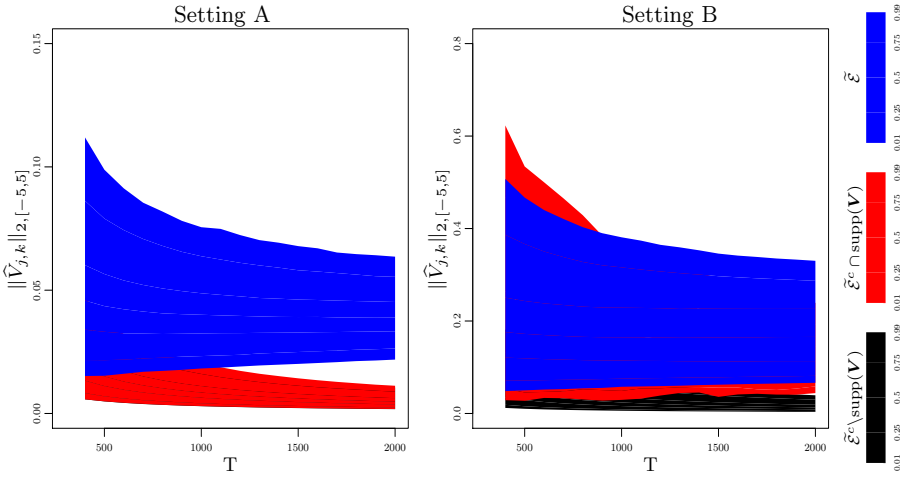


FIG 2. The quantiles of $\|\widehat{V}_{jk}\|_{2,[-5,5]}$ are displayed, for node pairs in $\tilde{\mathcal{E}}$ (11), $\tilde{\mathcal{E}}^c \cap \text{supp}(\mathbf{V})$, and $\tilde{\mathcal{E}}^c \backslash \text{supp}(\mathbf{V})$, as a function of the time period $T$. *Left:* Results for Setting A. The estimated cross-covariances of node pairs in $\tilde{\mathcal{E}}^c \backslash \text{supp}(\mathbf{V})$ and $\tilde{\mathcal{E}}^c \cap \text{supp}(\mathbf{V})$ overlap. *Center:* Results for Setting B. The estimated cross-covariances of node pairs in $\tilde{\mathcal{E}}$ and $\tilde{\mathcal{E}}^c \cap \text{supp}(\mathbf{V})$ overlap. *Right:* The color legend is displayed.

same order as the node pairs in $\tilde{\mathcal{E}}$. In light of Figure 2, we anticipate that for a sufficiently large value of the time period $T$, the screened edge set $\widehat{\mathcal{E}}(\zeta)$ will contain the edges in $\tilde{\mathcal{E}}$ as well as many of the edges in $\tilde{\mathcal{E}}^c \cap \text{supp}(\mathbf{V})$.

### 3.3. Size of Smallest Screened Edge Set

We now define $\zeta^* \equiv \max\{\zeta : \mathcal{E} \subseteq \widehat{\mathcal{E}}(\zeta)\}$, and calculate $\text{card}\big(\widehat{\mathcal{E}}(\zeta^*)\big)$. This represents the size of the smallest screened edge set that contains the true edge set.

Results, averaged over 200 simulated data sets, are shown in Figure 3.

We see that in Setting A, for sufficiently large $T$, $\text{card}\big(\widehat{\mathcal{E}}(\zeta^*)\big) = \text{card}(\tilde{\mathcal{E}})$, which implies that $\widehat{\mathcal{E}}(\zeta^*) = \tilde{\mathcal{E}}$. In other words, in Setting A, the screening procedure yields perfect recovery of the set $\tilde{\mathcal{E}}$ (11). This is in line with our intuition based on the left-hand panel of Figure 2.

In contrast, in Setting B, even when $T$ is very large, $\text{card}(\widehat{\mathcal{E}}(\zeta^*)) > \text{card}(\tilde{\mathcal{E}})$, which implies that $\widehat{\mathcal{E}}(\zeta^*) \supseteq \tilde{\mathcal{E}}$. This was expected based on the center panel of Figure 2.

### 3.4. Performance of Constrained Penalized Estimation

We now consider the performance of the estimator (10), which we obtain by calculating the screened edge set $\widehat{\mathcal{E}}(\zeta)$, and then performing a penalized regression subject to the constraint that $\omega_{jk} = 0$ for $(j,k) \notin \widehat{\mathcal{E}}(\zeta)$. Note that rather

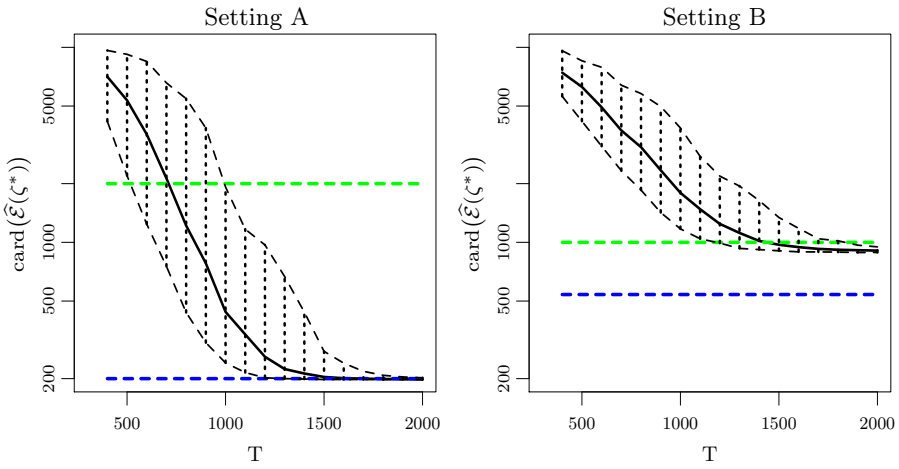


FIG 3. For each of 200 simulated data sets, we calculated card($\widehat{\mathcal{E}}(\zeta^*)$), where $\zeta^* \equiv \max\{\zeta : \mathcal{E} \subseteq \widehat{\mathcal{E}}(\zeta)\}$, as a function of the time period $T$. The curves represent the mean of card($\widehat{\mathcal{E}}(\zeta^*)$) (——); the 2.5% and 97.5% quantiles of card($\widehat{\mathcal{E}}(\zeta^*)$) (– –); card($\tilde{\mathcal{E}}$) (– –); and card(supp($\boldsymbol{V}$)) (– –). *Left:* Data generated under Setting A. *Right:* Data generated under Setting B.

than assuming a specific functional form for $\omega_{j,k}$, Hansen, Reynaud-Bouret and Rivoirard (2015) use a basis expansion to estimate $\omega_{j,k}$. Following their lead, we use a basis of step functions, of the form $\mathbb{1}_{((m-1)/2,m/2]}(t)$ for $m = 1, \ldots, 6$. Instead of applying a lasso penalty to the basis function coefficients (Hansen, Reynaud-Bouret and Rivoirard, 2015), we employ a group lasso penalty for every $1 \leq j, k \leq p$ (Yuan and Lin, 2006; Simon and Tibshirani, 2012). Thus, (10) consists of a squared error loss function and a group lasso penalty. We let

$$\widehat{\mathcal{E}}_{\mathcal{P}} \equiv \{(j,k) : \exists \Delta \text{ s.t. } \hat{\omega}_{j,k}(\Delta) \neq 0\}, \tag{12}$$

where $\hat{\omega}_{j,k}$ solves (10).

Results are shown in Figure 4. In Setting A, solving the constrained optimization problem (10) leads to substantially better performance than solving the unconstrained problem (5). The improvement is especially noticeable when $T$ is small. In Setting B, solving the constrained optimization problem (10) leads to only a slight improvement in performance relative to solving the unconstrained problem (5), since, as we have learned from Figures 2 and 3, the screened set $\widehat{\mathcal{E}}(\zeta)$ contains many edges in $\tilde{\mathcal{E}}^c \cap \text{supp}(\mathbf{V})$. In both settings, solving the constrained optimization problem leads to substantial computational improvements.

## 4. Proofs of Theoretical Results

In this section, we prove Theorem 1. In Section 4.1, we review an important property of the Hawkes process, the Wiener-Hopf integral equation. In Section 4.2,

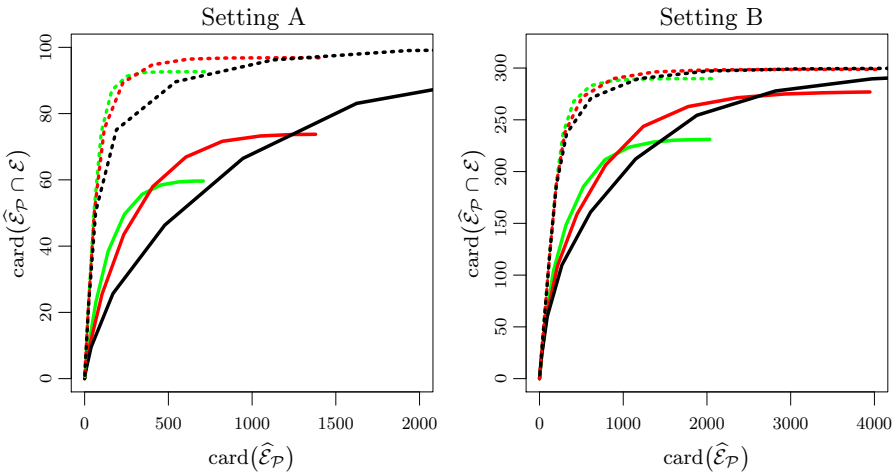


FIG 4. The constrained penalized optimization problem (10) was performed, for a range of values of the tuning parameter $\lambda$. The x-axis displays the size of the estimated edge set $\widehat{\mathcal{E}}_{\mathcal{P}}$ (12), and the $y$-axis displays the number of true positives, averaged over 200 simulated data sets. The curves represent performance when $\zeta$ is chosen to yield $\text{card}(\widehat{\mathcal{E}}(\zeta)) = 4\text{card}(\tilde{\mathcal{E}})$ ($T = 300$ [—]and $T = 600$ [$\cdots$]), and when $\zeta$ is chosen to yield $\text{card}(\widehat{\mathcal{E}}(\zeta)) = 8\text{card}(\tilde{\mathcal{E}})$ ($T = 300$ [—] and $T = 600$ [$\cdots$]). We also display performance of the unconstrained penalized optimization problem (5) ($T = 300$ [—] and $T = 600$ [$\cdots$]).

we list three technical lemmas used in the proof of Theorem 1. Theorem 1 is proved in Section 4.3. Proofs of the technical lemmas are provided in the Appendix.

### *4.1. The Wiener-Hopf Integral Equation*

Recall that the transfer functions $\boldsymbol{\omega} = \{\omega_{j,k}\}_{1 \le j,k \le p}$ were defined in (2), the cross-covariances $\boldsymbol{V} = \{V_{j,k}\}_{1 \le j,k \le p}$ were defined in (7), and the mean intensities $\boldsymbol{\Lambda} = (\Lambda_1, \ldots, \Lambda_p)^{\mathrm{T}}$ were defined in (6). If the Hawkes process defined in (2) is stationary, then for any $\Delta \in \mathbb{R}^+$,

$$\boldsymbol{V}(\Delta) = \boldsymbol{\omega}(\Delta)\text{diag}(\boldsymbol{\Lambda}) + (\boldsymbol{\omega} * \boldsymbol{V})(\Delta), \tag{13}$$

where

$$[\boldsymbol{\omega} * \boldsymbol{V}]_{j,k}(\Delta) \equiv \sum_{l=1}^{p} [\omega_{j,l} * V_{l,k}](\Delta)$$

and

$$[\omega_{j,l} * V_{l,k}](\Delta) \equiv \int_0^\infty \omega_{j,l}(\Delta') V_{l,k}(\Delta - \Delta')\mathrm{d}\Delta'.$$

Equation (13) belongs to a class of integral equations known as the *Wiener-Hopf integral equations*.

### 4.2. Technical Lemmas

We state three lemmas used to prove Theorem 1, and provide their proofs in the Appendix. The following lemma is a direct consequence of (13) and our assumptions. Recall that $[0, b]$ is a superset of $\mathrm{supp}(\omega_{j,k})$ introduced in Assumption 3.

**Lemma 1** *Under Assumptions 1–3, for sufficiently large $B$ such that $B \geq b$, we have that $\|V_{j,k}\|_{2,[-B,B]} \geq \beta_{\min}\Lambda_{\min}$ for $(j,k) \in \mathcal{E}$.*

The next lemma shows that the cross-covariance is Lipschitz continuous given the smoothness assumption on $\omega_{j,k}$ (Assumption 3(c)). We will use this lemma in the proof of Theorem 1, in order to bound the bias of the kernel smoothing estimator (8). Recall that $s$, the maximum node in-degree, was defined in (4).

**Lemma 2** *Under Assumptions 1–3, the cross-covariance function is Lipschitz for $1 \leq j, k \leq p$. More specifically, there exists some $\theta_1 > 0$ such that $|V_{j,k}(x) - V_{j,k}(y)| \leq \theta_1 s|x - y|$ for any $x, y \in \mathbb{R}$.*

Recall that the bandwidth $h$ was defined in (8). The following concentration inequality holds on the estimated cross-covariance.

**Lemma 3** *Suppose that Assumptions 1–3 hold, and let $h = c_1 s^{-1/2} T^{-1/6}$ for some constant $c_1$. Then*

$$\mathbb{P}\left(\bigcap_{1\leq j\leq k\leq p}\left[\left\|\widehat{V}_{j,k} - V_{j,k}\right\|_{2,[-B,B]} \leq c_2 s^{1/2} T^{-1/6}\right]\right) \geq 1 - c_3 s^{1/2} T^{7/6} p^2 e^{-c_4 T^{1/6}}.$$

### 4.3. Proof of Theorem 1

**Proof.** In what follows, we will consider the event

$$\mathcal{M} \equiv \left\{\left\|\widehat{V}_{j,k} - V_{j,k}\right\|_{2,[-B,B]} \leq c_2 s^{1/2} T^{-1/6} \text{ for all } 1 \leq j, k \leq p\right\}.$$

We will first show that part (b) of Theorem 1 holds. From the Wiener-Hopf equation, (13), for each $(j, k)$, we can write

$$V_{j,k} = \omega_{j,k}\Lambda_k + \boldsymbol{\omega}_{j,\cdot} * \boldsymbol{V}_{\cdot,k}. \tag{14}$$

We thus have

$$\begin{aligned}
\|V_{j,k}\|_{2,(-\infty,\infty)} &\leq \Lambda_k \|\omega_{j,k}\|_{2,(-\infty,\infty)} + \|\boldsymbol{\omega}_{j,\cdot} * \boldsymbol{V}_{\cdot,k}\|_{2,(-\infty,\infty)} \\
&\leq \Lambda_k \|\omega_{j,k}\|_{2,(-\infty,\infty)} + \sum_{l=1}^{p} \|\omega_{j,l} * V_{l,k}\|_{2,(-\infty,\infty)} \\
&\leq \Lambda_k \|\omega_{j,k}\|_{2,(-\infty,\infty)} + \sum_{l=1}^{p} \left(\int_{-\infty}^{\infty} |\omega_{j,l}(\Delta)|\, \mathrm{d}\Delta\right) \|V_{l,k}\|_{2,(-\infty,\infty)},
\end{aligned} \tag{15}$$

where the last inequality follows from Young's inequality (see e.g., Theorem 3.9.4 in Bogachev (2007)), which takes the form

$$\|f * g\|_{r,(-\infty,\infty)} \leq \|f\|_{p,(-\infty,\infty)} \|g\|_{q,(-\infty,\infty)}, \quad \frac{1}{p} + \frac{1}{q} = \frac{1}{r} + 1, \tag{16}$$

with $\|f\|_{p,(-\infty,\infty)} \equiv \left[\int_{-\infty}^{\infty} |f(x)|^p \, \mathrm{d}x\right]^{1/p}$. Here, we let $r = q = 2$, $p = 1$, $f = \omega_{j,l}$, and $g = V_{l,k}$.

From Assumption 3(c), we know that $\omega_{j,k}$ is bounded by $C$. Therefore, by the Cauchy-Schwartz inequality,

$$\|\omega_{j,k}\|_{2,\mathbb{R}} = \left\{\int_{-\infty}^{\infty} \omega_{j,k}^2(\Delta) \, \mathrm{d}\Delta\right\}^{1/2} \leq \left\{\int_{-\infty}^{\infty} C|\omega_{j,k}(\Delta)| \, \mathrm{d}\Delta\right\}^{1/2} = C^{1/2}\Omega_{j,k}^{1/2}.$$

Using (15) and letting $\bar{V}_{j,k} \equiv \|V_{j,k}\|_{2,(-\infty,\infty)}$, we get

$$\bar{V}_{j,k} \leq C^{1/2}\Omega_{j,k}^{1/2}\Lambda_k + \boldsymbol{\Omega}_{j,\cdot} \cdot \bar{\boldsymbol{V}}_{\cdot,k}. \tag{17}$$

The $\ell_2$-norm of the vector $\bar{\boldsymbol{V}}_{\cdot,k}$ can then be bounded using the triangle inequality,

$$\|\bar{\boldsymbol{V}}_{\cdot,k}\|_2 \leq C^{1/2}\Lambda_k \left[\sum_{j=1}^{p} \Omega_{j,k}\right]^{1/2} + \|\boldsymbol{\Omega}\bar{\boldsymbol{V}}_{\cdot,k}\|_2.$$

Thus, by Assumption 1,

$$\|\bar{\boldsymbol{V}}_{\cdot,k}\|_2 \leq C^{1/2}\Lambda_k \|\boldsymbol{\Omega}_{\cdot,k}\|_1^{1/2} + \gamma_\Omega \|\bar{\boldsymbol{V}}_{\cdot,k}\|_2.$$

Rearranging the terms, and using the fact that $\gamma_\Omega < 1$, gives

$$\|\bar{\boldsymbol{V}}_{\cdot,k}\|_2 \leq C^{1/2}(1-\gamma_\Omega)^{-1}\Lambda_{\max} \|\boldsymbol{\Omega}_{\cdot,k}\|_1^{1/2}. \tag{18}$$

Hence,

$$\sum_{j,k} \bar{V}_{j,k}^2 = \sum_k \|\bar{\boldsymbol{V}}_{\cdot,k}\|_2^2 \leq C(1-\gamma_\Omega)^{-2}\Lambda_{\max}^2 \sum_{k=1}^{p} \|\boldsymbol{\Omega}_{\cdot,k}\|_1. \tag{19}$$

Now, recall that the number of non-zero elements in $\boldsymbol{\Omega}$ is $\mathrm{card}(\mathcal{E})$, and $\Omega_{j,k} \leq \gamma_\Omega$. Thus, the inequality becomes

$$\sum_{j,k} \bar{V}_{j,k}^2 \leq C(1-\gamma_\Omega)^{-2}\Lambda_{\max}^2 \mathrm{card}(\mathcal{E})\gamma_\Omega. \tag{20}$$

Hence, no more than $(C/c_2^2)\mathrm{card}(\mathcal{E})s^{-1}T^{1/3}\gamma_\Omega(1-\gamma_\Omega)^{-2}\Lambda_{\max}^2$ elements of $\bar{V}_{j,k}$ exceed $c_2 s^{1/2}T^{-1/6}$. Recalling that $\bar{V}_{j,k} = \|V_{j,k}\|_{2,(-\infty,\infty)}$, this implies that no more than

$$(C/c_2^2)\mathrm{card}(\mathcal{E})s^{-1}T^{1/3}\gamma_\Omega(1-\gamma_\Omega)^{-2}\Lambda_{\max}^2$$

elements of $\|V_{j,k}\|_{2,(-B,B)}$ exceed $c_2 s^{1/2}T^{-1/6}$.

Given the event $\mathcal{M}$, only edges in the set

$$\{(j,k) : \|V_{j,k}\|_{2,[-B,B]} \geq c_2 s^{1/2} T^{-1/6}, 1 \leq j,k \leq p\}$$

can be contained in $\widehat{\mathcal{E}}(\zeta)$ for $\zeta = 2c_2 s^{1/2} T^{-1/6}$. This implies that the size of $\widehat{\mathcal{E}}(\zeta)$ is on the order of $(C/c_2^2)\text{card}(\mathcal{E}) s^{-1} T^{1/3} \gamma_\Omega (1-\gamma_\Omega)^{-2} \Lambda_{\max}^2$.

We now proceed to prove part (a) of Theorem 1. Lemma 1 states that $\|V_{j,k}\|_{2,[-B,B]} \geq \beta_{\min}\Lambda_{\min}$ for $(j,k) \in \mathcal{E}$. If the event $\mathcal{M}$ holds, then for $T$ sufficiently large, $\|\widehat{V}_{j,k}\|_{2,[-B,B]} > 2c_2 s^{1/2} T^{-1/6} = \zeta$ for $(j,k) \in \mathcal{E}$. Therefore, $\mathcal{E} \subset \widehat{\mathcal{E}}(\zeta)$.

Finally, Theorem 1 follows from the fact that, by Lemma 3, the event $\mathcal{M}$ holds with probability at least $1 - c_3 s^{1/2} T^{7/6} p^2 \exp(-c_4 T^{1/6})$. □

## 5. Discussion

In this paper, we have proposed a very simple procedure for screening the edge set in a multivariate Hawkes process. Provided that the process is mutually-exciting, we establish that this screening procedure can lead to a very small screened edge set, without incurring any false negatives. In fact, this result holds under a subset of the conditions required to establish model selection consistency of penalized regression estimators for the Hawkes process (Wainwright, 2009; Hansen, Reynaud-Bouret and Rivoirard, 2015). Therefore, this screening should always be performed whenever estimating the graph for a mutually-exciting Hawkes process.

The proposed screening procedure boils down to just screening pairs of nodes by thresholding an estimate of their cross-covariance. In fact, this approach is commonly taken within the neuroscience literature, with a goal of estimating the *functional connectivity* among a set of $p$ neuronal spike trains (Okatan, Wilson and Brown, 2005; Pillow et al., 2008; Mishchencko, Vogelstein and Paninski, 2011; Berry et al., 2012). Therefore, this paper sheds light on the theoretical foundations for an approach that is often used in practice.

## References


Ahrens, M. B., Orger, M. B., Robson, D. N., Li, J. M. and Keller, P. J. (2013). Whole-brain functional imaging at cellular resolution using light-sheet microscopy. *Nature Methods* **10** 413–420.

Aït-Sahalia, Y., Cacho-Diaz, J. and Laeven, R. J. A. (2015). Modeling financial contagion using mutually exciting jump processes. *Journal of Financial Economics* **117** 585 - 606.

Bacry, E., Gaïffas, S. and Muzy, J.-F. (2015). A generalization error bound for sparse and low-rank multivariate Hawkes processes. *arXiv preprint arXiv:1501.00725*.

Bacry, E., Delattre, S., Hoffmann, M. and Muzy, J. F. (2013). Some limit theorems for Hawkes processes and application to financial statistics. *Stochastic Process. Appl.* **123** 2475–2499. MR3054533

Berry, T., Hamilton, F., Peixoto, N. and Sauer, T. (2012). Detecting connectivity changes in neuronal networks. *Journal of Neuroscience Methods* **209** 388 - 397.

Bogachev, V. I. (2007). *Measure Theory. Vol. I, II*. Springer-Verlag, Berlin. MR2267655

Bowsher, C. G. (2007). Modelling security market events in continuous time: Intensity based, multivariate point process models. *Journal of Econometrics* **141** 876 - 912.

Brémaud, P. and Massoulié, L. (1996). Stability of nonlinear Hawkes processes. *Ann. Probab.* **24** 1563–1588. MR1411506

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

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data. Springer Series in Statistics*. Springer, Heidelberg Methods, theory and applications. MR2807761

Cai, T., Liu, W. and Luo, X. (2011). A constrained  1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association* **106** 594–607.

Chavez-Demoulin, V., Davison, A. C. and McNeil, A. J. (2005). Estimating value-at-risk: a point process approach. *Quantitative Finance* **5** 227-234.

Daley, D. and Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes, volume I: Elementary Theory and Methods of Probability and its Applications*. Springer,.

Fan, J., Feng, Y. and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *J. Amer. Statist. Assoc.* **106** 544–557. MR2847969

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70** 849–911. MR2530322

Fan, J., Ma, Y. and Dai, W. (2014). Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models. *J. Amer. Statist. Assoc.* **109** 1270–1284. MR3265696

Fan, J., Samworth, R. and Wu, Y. (2009). Ultrahigh dimensional feature selection: beyond the linear model. *J. Mach. Learn. Res.* **10** 2013–2038. MR2550099

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.* **38** 3567–3604. MR2766861

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software* **33** 1.

Greenshtein, E. and Ritov, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* **10** 971–988. MR2108039

Hansen, N. R., Reynaud-Bouret, P. and Rivoirard, V. (2015). Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli* **21** 83–143. MR3322314

Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika* **58** 83–90. MR0278410

Hawkes, A. G. and Oakes, D. (1974). A cluster process representation of a self-exciting process. *J. Appl. Probability* **11** 493–503. MR0378093

Liniger, T. J. (2009). Multivariate Hawkes processes PhD thesis, Diss., Eidgenössische Technische Hochschule ETH Zürich, Nr. 18403, 2009.

Liu, J., Li, R. and Wu, R. (2014). Feature selection for varying coefficient models with ultrahigh-dimensional covariates. *J. Amer. Statist. Assoc.* **109** 266–274. MR3180562

Luo, S., Song, R. and Witten, D. (2014). Sure Screening for Gaussian Graphical Models. *arXiv preprint arXiv:1407.7819*.

Massart, P. (2007). *Concentration inequalities and model selection. Lecture Notes in Mathematics* **1896**. Springer, Berlin Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard. MR2319879

Mishchencko, Y., Vogelstein, J. T. and Paninski, L. (2011). A Bayesian approach for inferring neuronal connectivity from calcium fluorescent imaging data. *Ann. Appl. Stat.* **5** 1229–1261. MR2849773

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P. and Tita, G. E. (2011). Self-exciting point process modeling of crime. *J. Amer. Statist. Assoc.* **106** 100–108. MR2816705

Ogata, Y. (1988). Statistical Models for Earthquake Occurrences and Residual Analysis for Point Processes. *Journal of the American Statistical Association* **83** 9-27.

Okatan, M., Wilson, M. A. and Brown, E. N. (2005). Analyzing Functional Connectivity Using a Network Likelihood Model of Ensemble Neural Spiking Activity. *Neural Comput.* **17** 1927–1961.

Paninski, L., Pillow, J. and Lewi, J. (2007). Statistical models for neural encoding, decoding, and optimal stimulus design. *Progress in Brain Research* **165** 493–507.

Perry, P. O. and Wolfe, P. J. (2013). Point process modelling for directed interaction networks. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 821–849. MR3124793

Pillow, J. W., Shlens, J., Paninski, L., Sher, A., Litke, A. M., Chichilnisky, E. and Simoncelli, E. P. (2008). Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature* **454** 995–999.

Ravikumar, P., Wainwright, M. J., Raskutti, G. and Yu, B. (2011). High-dimensional covariance estimation by minimizing $\ell_1$-penalized log-determinant divergence. *Electron. J. Stat.* **5** 935–980. MR2836766

Reynaud-Bouret, P. and Roy, E. (2006). Some non asymptotic tail estimates for Hawkes processes. *Bull. Belg. Math. Soc. Simon Stevin* **13** 883–896. MR2293215

Reynaud-Bouret, P. and Schbath, S. (2010). Adaptive estimation for Hawkes processes; application to genome analysis. *Ann. Statist.* **38** 2781–2822. MR2722456

Saegusa, T. and Shojaie, A. (2016). Joint estimation of precision matrices in heterogeneous populations. *Electronic Journal of Statistics* **10** 1341–1392.

Simma, A. and Jordan, M. I. (2012). Modeling events with cascades of Poisson processes. *arXiv preprint arXiv:1203.3516*.

Simon, N. and Tibshirani, R. J. (2012). Standardization and the group lasso penalty. *Statist. Sinica* **22** 983–1001. MR2987480

Song, R., Lu, W., Ma, S. and Jeng, X. J. (2014). Censored rank independence screening for high-dimensional survival data. *Biometrika* **101** 799–814. MR3286918

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation. Springer Series in Statistics*. Springer, New York Revised and extended from the 2004 French original, Translated by Vladimir Zaiats. MR2724359 (2011g:62006)

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using $\ell_1$-constrained quadratic programming (Lasso). *Information Theory, IEEE Transactions on* **55** 2183–2202.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 49–67. MR2212574

Zhou, K., Zha, H. and Song, L. (2013a). Learning social infectivity in sparse low-rank networks using multi-dimensional Hawkes processes. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics* 641–649.

Zhou, K., Zha, H. and Song, L. (2013b). Learning triggering kernels for multi-dimensional Hawkes processes. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)* 1301–1309.

Zhu, L. (2013). Nonlinear Hawkes processes. *arXiv preprint arXiv:1304.7531*.

## Appendix A: Technical Proofs

### A.1. Proof of Lemma 1

**Proof.** First, we observe that, if $V_{j,k}$ is non-negative for all $j$ and $k$, then $\omega_{j,l} * V_{l,k}$ is non-negative for any $j, l, k$. Under Assumption 1, we know that (13) holds. We can see from (13) that

$$V_{j,k}(\Delta) \geq \omega_{j,k}(\Delta)\Lambda_k.$$

Therefore, we have

$$\|V_{j,k}(\Delta)\|_{2,[-B,B]} \geq \|\omega_{j,k}(\Delta)\|_{2,[-B,B]}\Lambda_{\min} = \|\omega_{j,k}(\Delta)\|_{2,[0,b]}\Lambda_{\min}, \tag{21}$$

where the inequality follows from Assumption 2 and the equality holds since

$$\operatorname{supp}(\omega_{j,k}) \subset (0, b] \subset [-B, B].$$

From Assumption 3(b), we have that $\|V_{j,k}(\Delta)\|_{2,[-B,B]} \geq \beta_{\min}\Lambda_{\min}$ for $(j,k) \in \mathcal{E}$.

We now show that the elements of $\boldsymbol{V}$ are non-negative, i.e., $V_{l,k}(\Delta) \geq 0$ for $1 \leq l, k \leq p$, and $\Delta \in \mathbb{R}$. Recall from the definition (7) in the main paper that

$$\begin{aligned} V_{l,k}(\Delta) \equiv& \mathbb{E}[\mathrm{d}N_l(t)\,\mathrm{d}N_k(t-\Delta)]/\{\mathrm{d}t\,\mathrm{d}(t-\Delta)\} - \Lambda_l\Lambda_k \\ =& \mathbb{E}[\lambda_l(t)\,\mathrm{d}N_k(t-\Delta)]/\{\mathrm{d}(t-\Delta)\} - \Lambda_l\Lambda_k, \end{aligned} \tag{22}$$

where the second equality follows from

$$\mathbb{E}[\,\mathrm{d}N_l(t)\,\mathrm{d}N_k(t-\Delta)] = \mathbb{E}\big[\mathbb{E}\big[\,\mathrm{d}N_l(t) \mid \mathcal{H}_t\big]\,\mathrm{d}N_k(t-\Delta)\big] = \mathbb{E}[\lambda_l(t)\,\mathrm{d}N_k(t-\Delta)]\,\mathrm{d}t. \tag{23}$$

In this proof, we use the Stieltjes integral to rewrite $\lambda_l(t)$ in (2) as

$$\lambda_l(t) = \mu_l + \sum_{k=1}^{p}\Big(\sum_{i:t_{k,i}\leq t}\omega_{l,k}(t-t_{k,i})\Big) = \mu_l + \sum_{k=1}^{p}\int_0^\infty \omega_{l,k}(\Delta)\mathrm{d}N_k(t-\Delta). \tag{24}$$

Plugging in $\lambda_l(t)$ from (24) into (22) gives

$$\begin{aligned} V_{l,k}(\Delta) =& -\Lambda_l\Lambda_k + \mathbb{E}[\mu_l\,\mathrm{d}N_k(t-\Delta)]/\{\mathrm{d}(t-\Delta)\} \\ &+ \mathbb{E}\left[\sum_{m=1}^{p}\int_0^b \omega_{l,m}(\Delta')\,\mathrm{d}N_m(t-\Delta')\,\mathrm{d}N_k(t-\Delta)\right]/\{\mathrm{d}(t-\Delta)\} \\ =& \sum_{m=1}^{p}\int_0^b \omega_{l,m}(\Delta')\mathbb{E}\big[\,\mathrm{d}N_m(t-\Delta')\,\mathrm{d}N_k(t-\Delta)\big]/\{\mathrm{d}(t-\Delta)\} \\ &+ \mathbb{E}[\mu_l\,\mathrm{d}N_k(t-\Delta)]/\{\mathrm{d}(t-\Delta)\} - \Lambda_l\mathbb{E}[\,\mathrm{d}N_k(t-\Delta)]/\{\mathrm{d}(t-\Delta)\}, \end{aligned}$$

where we use the definition $\Lambda_k = \mathbb{E}[\,\mathrm{d}N_k(t-\Delta)]/\{\mathrm{d}(t-\Delta)\}$.

Using the fact that (see e.g., Hawkes and Oakes (1974))

$$\Lambda_l = \mu_l + \sum_{m=1}^{p} \int_0^b \omega_{l,m}(\Delta')\,\mathrm{d}\Delta'\mu_m,$$

we have

$$\begin{aligned}
V_{l,k}(\Delta) &= \sum_{m=1}^{p} \int_0^b \omega_{l,m}(\Delta')\mathbb{E}\big[\,\mathrm{d}N_m(t-\Delta')\,\mathrm{d}N_k(t-\Delta)\big]/\{\mathrm{d}(t-\Delta)\} \\
&\quad + \mathbb{E}\big[\mu_l\,\mathrm{d}N_k(t-\Delta)\big]/\{\mathrm{d}(t-\Delta)\} \\
&\quad - \left\{\mu_l + \sum_{m=1}^{p} \int_0^b \omega_{l,m}(\Delta')\mu_m\,\mathrm{d}\Delta'\right\}\mathbb{E}\big[\,\mathrm{d}N_k(t-\Delta)\big]/\{\mathrm{d}(t-\Delta)\} \\
&= \sum_{m=1}^{p} \int_0^b \omega_{l,m}(\Delta')\mathbb{E}\big[\,\mathrm{d}N_m(t-\Delta')\,\mathrm{d}N_k(t-\Delta)\big]/\{\mathrm{d}(t-\Delta)\} \\
&\quad - \sum_{m=1}^{p} \int_0^b \omega_{l,m}(\Delta')\mu_m\,\mathrm{d}\Delta'\mathbb{E}\big[\,\mathrm{d}N_k(t-\Delta)\big]/\{\mathrm{d}(t-\Delta)\}.
\end{aligned}$$

Rearranging the terms gives

$$V_{l,k}(\Delta) = \sum_{m=1}^{p} \int_0^b \frac{\omega_{l,m}(\Delta')}{\mathrm{d}(t-\Delta)}\Big\{\mathbb{E}\big[\,\mathrm{d}N_m(t-\Delta')\,\mathrm{d}N_k(t-\Delta)\big] - \mathbb{E}\big[\mu_m\,\mathrm{d}\Delta'\,\mathrm{d}N_k(t-\Delta)\big]\Big\}. \tag{25}$$

Next, we will rewrite (25) by taking the conditional expectation of $\mathrm{d}N_k$ or $\mathrm{d}N_m$ as in (23). Note here that, when $\Delta' < \Delta$, we condition $\mathrm{d}N_m$ on the history up to $t-\Delta'$, i.e., $\mathcal{H}_{t-\Delta'}$. Given $\mathcal{H}_{t-\Delta'}$, $\mathrm{d}N_k(t-\Delta)$ is fixed since $t-\Delta < t-\Delta'$. When $\Delta' > \Delta$, we condition $\mathrm{d}N_k$ on the history up to $t-\Delta$. These cases are discussed separately in the following.

When $\Delta' < \Delta$, for each integral of the summation, it holds that

$$\mathbb{E}\{\,\mathrm{d}N_m(t-\Delta')\,\mathrm{d}N_k(t-\Delta)\} = \mathbb{E}\{\lambda_m(t-\Delta')\,\mathrm{d}\Delta'\,\mathrm{d}N_k(t-\Delta)\}.$$

From the definition of $\lambda_m(t)$ in (2), we know that $\lambda_m(t-\Delta') \geq \mu_m$. Hence, in (25), if $\Delta' < \Delta$, it holds that

$$\mathbb{E}\{\,\mathrm{d}N_m(t-\Delta')\,\mathrm{d}N_k(t-\Delta)\}/\{\mathrm{d}(t-\Delta)\} - \mathbb{E}\{\mu_m\,\mathrm{d}\Delta'\,\mathrm{d}N_k(t-\Delta)\}/\{\mathrm{d}(t-\Delta)\} \geq 0. \tag{26}$$

On the other hand, when $\Delta' \geq \Delta$, we have

$$\begin{aligned}
&\mathbb{E}\{\,\mathrm{d}N_m(t-\Delta')\,\mathrm{d}N_k(t-\Delta)\}/\{\mathrm{d}(t-\Delta)\} - \mathbb{E}\{\mu_m\,\mathrm{d}\Delta'\,\mathrm{d}N_k(t-\Delta)\}/\{\mathrm{d}(t-\Delta)\} \\
=&\mathbb{E}\{\,\mathrm{d}N_m(t-\Delta')\lambda_k(t-\Delta)\} - \mathbb{E}\{\mu_m\,\mathrm{d}\Delta'\lambda_k(t-\Delta)\} \\
=&\mathbb{E}\{\,\mathrm{d}N_m(t-\Delta')\lambda_k(t-\Delta)\} - \mu_m\Lambda_k\,\mathrm{d}\Delta'.
\end{aligned}$$

Expanding $\lambda_k$ and $\Lambda_k$ yields

$$\begin{aligned}
&\mathbb{E}\{\,\mathrm{d}N_m(t-\Delta')\lambda_k(t-\Delta)\} - \mu_m\Lambda_k\,\mathrm{d}\Delta' \\
=&\mathbb{E}\{\,\mathrm{d}N_m(t-\Delta')\mu_k\} + \sum_{i=1}^{p}\int_0^b \omega_{k,i}(\Delta'')\mathbb{E}\big[\,\mathrm{d}N_m(t-\Delta')\,\mathrm{d}N_i(t-\Delta-\Delta'')\big] \\
&- \mu_m\mu_k\,\mathrm{d}\Delta' - \sum_{i=1}^{p}\int_0^b \omega_{k,i}(\Delta'')\,\mathrm{d}\Delta''\mu_i\mu_m\,\mathrm{d}\Delta' \\
=&(\Lambda_m-\mu_m)\mu_k\,\mathrm{d}\Delta' \\
&+ \sum_{i=1}^{p}\int_0^b \omega_{k,i}(\Delta'')\Big\{\mathbb{E}\big[\,\mathrm{d}N_m(t-\Delta')\,\mathrm{d}N_i(t-\Delta-\Delta'')\big] - \mu_i\mu_m\,\mathrm{d}\Delta'\,\mathrm{d}\Delta''\Big\}.
\end{aligned}$$

Now, observe that $\Lambda_m \geq \mu_m$ and $\mathbb{E}\{\mathrm{d}N_i(t-\Delta-\Delta'')\,\mathrm{d}N_m(t-\Delta')\}/\{\,\mathrm{d}\Delta' d\Delta''\} \geq \mu_i\mu_m$ by the nature of the mutually-exciting process. Thus, for $\Delta' \geq \Delta$,

$$\mathbb{E}\{\,\mathrm{d}N_m(t-\Delta')\,\mathrm{d}N_k(t-\Delta)\}/\{\mathrm{d}(t-\Delta)\} - \mathbb{E}\{\mu_m\,\mathrm{d}\Delta'\,\mathrm{d}N_k(t-\Delta)\}/\{\mathrm{d}(t-\Delta)\} \geq 0. \tag{27}$$

Applying both (26) and (27) to (25) shows that $V_{l,k}(\Delta) \geq 0$.

□

### *A.2. Proof of Lemma 2*

**Proof.** For any $\Delta \geq 0$, the integral equation (13) gives

$$V_{j,k}(\Delta) = \omega_{j,k}(\Delta)\Lambda_k + (\boldsymbol{\omega}_{j,\cdot} * \boldsymbol{V}_{\cdot,k})(\Delta). \tag{28}$$

For any $x, y \geq 0$, we can write

$$\begin{aligned}
\big|V_{j,k}(x) - V_{j,k}(y)\big| =& \big|\{\omega_{j,k}(x) - \omega_{j,k}(y)\}\Lambda_k + (\boldsymbol{\omega}_{j,\cdot} * \boldsymbol{V}_{\cdot,k})(x) - (\boldsymbol{\omega}_{j,\cdot} * \boldsymbol{V}_{\cdot,k})(y)\big| \\
=& \Big|\{\omega_{j,k}(x) - \omega_{j,k}(y)\}\Lambda_k + \sum_{l=1}^{p}\big\{\omega_{j,l} * V_{l,k}(x) - \omega_{j,l} * V_{l,k}(y)\big\}\Big| \\
=& \Big|\{\omega_{j,k}(x) - \omega_{j,k}(y)\}\Lambda_k + \sum_{l\in\mathcal{E}_j}\big\{\omega_{j,l} * V_{l,k}(x) - \omega_{j,l} * V_{l,k}(y)\big\}\Big|,
\end{aligned} \tag{29}$$

where the last inequality holds since $\omega_{j,l} \equiv 0$ for $l \notin \mathcal{E}_j$. We then have

$$\big|V_{j,k}(x) - V_{j,k}(y)\big| \leq \underbrace{\big|\{\omega_{j,k}(x) - \omega_{j,k}(y)\}\Lambda_k\big|}_{\mathrm{I}} + \sum_{l\in\mathcal{E}_j}\underbrace{\big|\omega_{j,l} * V_{l,k}(x) - \omega_{j,l} * V_{l,k}(y)\big|}_{\mathrm{II}_l}. \tag{30}$$

For I, we know from Assumptions 2 and 3(c) that

$$\mathrm{I} \equiv \big|\{\omega_{j,k}(x) - \omega_{j,k}(y)\}\Lambda_k\big| \leq \theta_0\Lambda_{\max}|x-y|. \tag{31}$$

For $\mathrm{II}_l$, we can expand the convolution

$$\begin{aligned}\mathrm{II}_l &= \left| \int_0^b \omega_{j,l}(\Delta) V_{l,k}(x-\Delta)\mathrm{d}\Delta - \int_0^b \omega_{j,l}(\Delta) V_{l,k}(y-\Delta)\mathrm{d}\Delta \right| \\ &= \left| \int_{-x}^{b-x} \omega_{j,l}(\Delta'+x) V_{l,k}(-\Delta')\mathrm{d}\Delta' - \int_{-y}^{b-y} \omega_{j,l}(\Delta'+y) V_{l,k}(-\Delta')\mathrm{d}\Delta' \right|.\end{aligned}$$

Without loss of generality, we consider only the case that $x \geq y$. We can decompose the integrals into parts on the intervals $[-x,-y)$, $[-y,b-x)$, and $[b-x,b-y]$ as

$$\begin{aligned}\mathrm{II}_l \leq & \left| \int_{-y}^{b-x} \left\{ \omega_{j,l}(\Delta'+x) - \omega_{j,l}(\Delta'+y) \right\} V_{l,k}(-\Delta')\,\mathrm{d}\Delta' \right| \\ & + \left| \int_{-x}^{-y} \omega_{j,l}(\Delta'+x) V_{l,k}(-\Delta')\,\mathrm{d}\Delta' \right| + \left| \int_{b-x}^{b-y} \omega_{j,l}(\Delta'+y) V_{l,k}(-\Delta')\,\mathrm{d}\Delta' \right| \\ \leq & \int_{-y}^{b-x} \theta_0 \big|\Delta'+x-\Delta'-y\big| \big|V_{l,k}(-\Delta')\big|\,\mathrm{d}\Delta' \\ & + \int_{-x}^{-y} \big|\omega_{j,l}(\Delta'+x) V_{l,k}(-\Delta')\big|\,\mathrm{d}\Delta' + \int_{b-x}^{b-y} \big|\omega_{j,l}(\Delta'+y) V_{l,k}(-\Delta')\big|\,\mathrm{d}\Delta' \\ \leq & \int_{-y}^{b-x} \theta_0 |x-y| V_{\max}\,\mathrm{d}\Delta' + \int_{-x}^{-y} |\omega_{j,l}(\Delta'+x)| V_{\max}\,\mathrm{d}\Delta' \\ & + \int_{b-x}^{b-y} |\omega_{j,l}(\Delta'+y)| V_{\max}\,\mathrm{d}\Delta' \\ \leq & (b-x+y)\theta_0 |x-y| V_{\max} + 2C(x-y)V_{\max},\end{aligned}$$

where we use Assumption 3(c) in the second inequality, Assumptions 2 in the third inequality, and the boundedness of $\omega_{j,l}$ from Assumption 3(c) in the last inequality. Recalling that $x \geq y$, we have

$$\mathrm{II}_l \leq \big(b\theta_0 V_{\max} + 2CV_{\max}\big)|x-y|, \tag{32}$$

Finally, plugging (31) and (32) into (30) gives

$$\big|V_{j,k}(x) - V_{j,k}(y)\big| \leq \theta_0 \Lambda_{\max}|x-y| + s\big(b\theta_0 V_{\max} + 2CV_{\max}\big)|x-y| \leq s\theta_1 |x-y|, \tag{33}$$

where we set $\theta_1 \equiv \theta_0 \Lambda_{\max} + b\theta_0 V_{\max} + 2CV_{\max}$. Note that the last inequality holds as long as $s \geq 1$. (The result also holds if $s = 0$: in this case, the second term in (30) is zero for every $j$ and the bound (31) suffices.) □

### A.3. Proof of Lemma 3

Recall that the estimator of the cross-covariance (8) takes the form

$$\underbrace{\frac{1}{h}\frac{1}{T}\iint_{[0,T]^2} K\left(\frac{t-t'+\Delta}{h}\right)\mathrm{d}N_j(t')\,\mathrm{d}N_k(t)}_{\mathrm{I}_{j,k}} - \underbrace{\left[\frac{1}{T}\int_0^T \mathrm{d}N_j(t)\right]}_{\mathrm{II}_j}\underbrace{\left[\frac{1}{T}\int_0^T \mathrm{d}N_k(t)\right]}_{\mathrm{II}_k}.$$

The proof of Lemma 3 uses the following result. Lemma 4 is based on Proposition 3 of Hansen, Reynaud-Bouret and Rivoirard (2015); for completeness, we provide its proof in Section A.4.

**Lemma 4** *Suppose that Assumption 1 holds. We have*

$$\mathbb{P}\left(\bigcap_{1\le j\le k\le p}\left[|\mathrm{I}_{j,k}-\mathbb{E}\mathrm{I}_{j,k}|\ge c_6T^{-1/3}\right]\right)\le c_5p^2T\exp(-c_4T^{1/6}), \tag{34}$$

$$\mathbb{P}\left(\bigcap_{1\le j\le p}\left[|\mathrm{II}_j-\mathbb{E}\mathrm{II}_j|\ge c_6T^{-1/3+1/18}\right]\right\}\le c_5p^2T\exp(-c_4T^{1/6}), \tag{35}$$

*where $c_4$, $c_5$, and $c_6$ are constants.*

We are now ready to prove Lemma 3.

**Proof.**

First, note that

$$\begin{aligned}
&\left|\mathbb{E}\mathrm{I}_{j,k}-h[V_{j,k}(\Delta)+\Lambda_j\Lambda_k]\right|\\
&=\left|\frac{1}{T}\iint_{[0,T]^2}K\left(\frac{t-t'+\Delta}{h}\right)\mathbb{E}[\,\mathrm{d}N_j(t')\,\mathrm{d}N_k(t)]\right.\\
&\qquad\left.-\frac{1}{T}\iint_{[0,T]^2}K\left(\frac{t-t'+\Delta}{h}\right)[V_{j,k}(\Delta)+\Lambda_j\Lambda_k]\,\mathrm{d}t\,\mathrm{d}t'\right|\\
&=\left|\frac{1}{T}\iint_{[0,T]^2}K\left(\frac{t-t'+\Delta}{h}\right)\{\mathbb{E}[\,\mathrm{d}N_j(t')\,\mathrm{d}N_k(t)]-\Lambda_j\Lambda_k\,\mathrm{d}t\,\mathrm{d}t'\}\right.\\
&\qquad\left.-\frac{1}{T}\iint_{[0,T]^2}K\left(\frac{t-t'+\Delta}{h}\right)V_{j,k}(\Delta)\,\mathrm{d}t\,\mathrm{d}t'\right|\\
&=\left|\frac{1}{T}\iint_{[0,T]^2}K\left(\frac{t-t'+\Delta}{h}\right)V_{j,k}(t'-t)\,\mathrm{d}t\,\mathrm{d}t'\right.\\
&\qquad\left.-\frac{1}{T}\iint_{[0,T]^2}K\left(\frac{t-t'+\Delta}{h}\right)V_{j,k}(\Delta)\,\mathrm{d}t\,\mathrm{d}t'\right|\\
&=\left|\frac{1}{T}\iint_{[0,T]^2}K\left(\frac{t-t'+\Delta}{h}\right)[V_{j,k}(t'-t)-V_{j,k}(\Delta)]\,\mathrm{d}t\,\mathrm{d}t'\right|,
\end{aligned} \tag{36}$$

where we use the definition of $\boldsymbol{V}$ in the third equality. Using the fact that the kernel $K(x/h)$ is defined on $[-h,h]$, we can write

$$\begin{aligned}
&\left|\mathbb{E}\mathrm{I}_{j,k} - h[V_{j,k}(\Delta) + \Lambda_j\Lambda_k]\right| \\
=&\left|\frac{1}{T}\int_0^T \int_{\max(0,t-\Delta-h)}^{\min(T,t-\Delta+h)} K\left(\frac{t-t'+\Delta}{h}\right)\left[V_{j,k}(t'-t) - V_{j,k}(\Delta)\right]\mathrm{d}t\,\mathrm{d}t'\right| \\
\leq&\frac{1}{T}\int_0^T \int_{\max(0,t-\Delta-h)}^{\min(T,t-\Delta+h)} K\left(\frac{t-t'+\Delta}{h}\right)\theta_1 s|t'-t-\Delta|\,\mathrm{d}t\,\mathrm{d}t' \\
\leq&\frac{1}{T}\int_0^T \int_{\max(0,t-\Delta-h)}^{\min(T,t-\Delta+h)} K\left(\frac{t-t'+\Delta}{h}\right)\theta_1 hs\,\mathrm{d}t\,\mathrm{d}t' \\
\leq&\frac{1}{T}\int_0^T 2\theta_1 sh^2\,\mathrm{d}t \\
=&2\theta_1 sh^2,
\end{aligned} \tag{37}$$

where the first inequality follows from Lemma 2.

Recall that $\mathrm{II}_j \equiv T^{-1}N_j(T)$ and $\mathrm{II}_k \equiv T^{-1}N_k(T)$. Applying Lemma 4 and (37), we have, with probability at least $1 - 2c_5p^2T\exp(-c_4T^{1/6})$,

$$\begin{aligned}
\left|\widehat{V}_{j,k}(\Delta) - V_{j,k}(\Delta)\right| \leq&\frac{1}{h}|\mathrm{I}_{j,k} - \mathbb{E}\mathrm{I}_{j,k}| + \frac{1}{h}|\mathbb{E}\mathrm{I}_{j,k} - \mathbb{E}[\mathrm{d}N_j(t-\Delta)\mathrm{d}N_k(t)]/(\mathrm{d}t\mathrm{d}\Delta)| \\
&+\left|\frac{1}{T^2}(N_j(T) - T\Lambda_j)N_k(T)\right| + \left|\Lambda_j\frac{1}{T}N_k(T) - \Lambda_j\Lambda_k\right| \\
\leq&c_6T^{-1/3}h^{-1} + 2\theta_1 hs + \\
&(\|\Lambda_{\max} + c_6T^{-1/3+1/18})c_6T^{-1/3+1/18} + \Lambda_{\max}c_6T^{-1/3+1/18}.
\end{aligned} \tag{38}$$

Letting $h = c_1 s^{-1/2}T^{-1/6}$, (38) can be written as

$$\left|\widehat{V}_{j,k}(\Delta) - V_{j,k}(\Delta)\right| \leq c_2' s^{1/2}T^{-1/6}. \tag{39}$$

Lastly, we need a uniform bound on $\widehat{V}_{j,k} - V_{j,k}$ on the region $[-B,B]$. We first see that the above probability statement holds for a grid of $\lceil s^{1/2}T^{1/6}\rceil$ points on $[-B,B]$, denoted as $\{\Delta_i\}_{i=1}^{\lceil s^{1/2}T^{1/6}\rceil}$. The gap between adjacent points on this grid is bounded by $2Bs^{-1/2}T^{-1/6}$. Furthermore, for any $\Delta \in [-B,B]$, we can find a point on the grid $\Delta_i$ such that $|\Delta - \Delta_i| \leq 2B/\lceil s^{1/2}T^{1/6}\rceil \leq 2Bs^{-1/2}T^{-1/6}$. From basic calculus we get that, for all $\Delta \in [-B,B]$,

$$\begin{aligned}
&\left|\widehat{V}_{j,k}(\Delta) - V_{j,k}(\Delta)\right| \\
=&\left|\widehat{V}_{j,k}(\Delta) - \widehat{V}_{j,k}(\Delta_i) + \widehat{V}_{j,k}(\Delta_i) - V_{j,k}(\Delta_i) + V_{j,k}(\Delta_i) - V_{j,k}(\Delta)\right| \\
\leq&2Bs^{-1/2}T^{-1/6} + c_2' s^{1/2}T^{-1/6} + \theta_1 s s^{-1/2}T^{-1/6} \\
\leq&c_2 s^{1/2}T^{-1/6},
\end{aligned} \tag{40}$$

for some constant $c_2$.

Therefore, with probability at least $1 - c_3 s^{1/2} p^2 T^{7/6} \exp(-c_4 T^{1/6})$,

$$\left\| \widehat{V}_{j,k} - V_{j,k} \right\|_{2,[-B,B]} \leq c_2 s^{1/2} T^{-1/6}. \tag{41}$$

□

### A.4. Proof of Lemma 4

Lemma 4 follows directly from the proof of Proposition 3 in Hansen, Reynaud-Bouret and Rivoirard (2015). The only difference is that we want a polynomial bound on the deviation, while Hansen, Reynaud-Bouret and Rivoirard (2015) consider a logarithmic bound. For completeness, we state the proof of Lemma 4 below, but note that the proof is almost identical to the proof of Proposition 3 in Hansen, Reynaud-Bouret and Rivoirard (2015). We refer the interested readers to the original proof in Section 7.4.3 of Hansen, Reynaud-Bouret and Rivoirard (2015) for more details.

Throughout this section, we assume that $\boldsymbol{N} \equiv (N_1, \ldots, N_p)^{\mathrm{T}}$ is defined on the full real line. We first state some notation that is only used in this section.

1. Following Hansen, Reynaud-Bouret and Rivoirard (2015), we use $C^{(i)}_{a_1,a_2,\ldots}$ to denote a constant that depends only on $a_1, a_2, \ldots$; and we use the superscript $i$ to indicate that this is the $i$th constant appearing in the proof.
2. Without loss of generality, we assume that $\mathrm{supp}(\omega_{j,k}) \subset (0, 1]$, as in Hansen, Reynaud-Bouret and Rivoirard (2015).
3. As in Hansen, Reynaud-Bouret and Rivoirard (2015), we introduce a function $Z(\boldsymbol{N})$ such that $Z(\boldsymbol{N})$ depends only on $\{\mathrm{d}\boldsymbol{N}(t'), t' \in [-A, 0)\}$, and there exist two non-negative constants $\eta$ and $d$ such that

$$|Z(\boldsymbol{N})| \leq d \left\{ 1 + \left( \sum_{l=1}^{p} N_l([-A, 0)] \right)^{\eta} \right\}. \tag{42}$$

4. We also introduce the (time) shift operator $S_t$ so that $Z \circ S_t(\boldsymbol{N})$ depends only on $\{\mathrm{d}\boldsymbol{N}(t'), t' \in [-A + t, t)\}$, in the same way as $Z(\boldsymbol{N})$ depends on the points of $\boldsymbol{N}$ in $[-A, 0)$.

We are now ready to prove the lemma. When proving the bound (34), we only discuss the case when $j \neq k$. The proof for the case when $j = k$ follows from the same argument and is thus omitted.

**Proof.**

In this proof, we will consider a probability bound for $\left[ Z \circ S_t(\boldsymbol{N}) - \mathbb{E}(Z) \right] \mathrm{d}t \geq u$, where, for some $\kappa \in (0, 1)$ to be specified later,

$$u = c_6 T^{(1-\kappa)(1-\eta)+\kappa}. \tag{43}$$

Note that, by applying the bound to $-Z(\cdot)$, we can obtain a bound for $\big|Z \circ S_t(\boldsymbol{N}) - \mathbb{E}(Z)\big|$. To complete the proof, we will verify the statements (34) and (35) by considering some specific choices of $Z(\cdot)$.

For any positive integer $k$ such that $x \equiv T/(2k) > A$, we have

$$\begin{aligned}
&\mathbb{P}\bigg(\int_0^T \big[Z \circ S_t(\boldsymbol{N}) - \mathbb{E}(Z)\big]\,\mathrm{d}t \geq u\bigg)\\
=&\mathbb{P}\bigg(\sum_{q=0}^{k-1}\int_{2qx}^{2qx+x} \big[Z \circ S_t(\boldsymbol{N}) - \mathbb{E}(Z)\big]\,\mathrm{d}t + \int_{2qx+x}^{2qx+2x} \big[Z \circ S_t(\boldsymbol{N}) - \mathbb{E}(Z)\big]\,\mathrm{d}t \geq u\bigg)\\
\leq&2\mathbb{P}\bigg(\sum_{q=0}^{k-1}\int_{2qx}^{2qx+x} \big[Z \circ S_t(\boldsymbol{N}) - \mathbb{E}(Z)\big]\,\mathrm{d}t \geq \frac{u}{2}\bigg),
\end{aligned}$$

where the inequality follows from the stationarity of $\boldsymbol{N}$. As in Reynaud-Bouret and Roy (2006), let $\{\tilde{M}_q^x\}_{q=1}^{\infty}$ be a sequence of independent Hawkes processes, each of which is stationary with intensities $\boldsymbol{\lambda}(t) \equiv (\lambda_1(t), \ldots, \lambda_p(t))^{\mathrm{T}}$. See Section 3 of Reynaud-Bouret and Roy (2006) for more details on the construction of $\{\tilde{M}_q^x\}_{q=1}^{\infty}$. For each $q$, let $M_q^x$ be the truncated process associated with $\tilde{M}_q^x$, where truncation means that we only consider the points in $[2qx - A, 2qx + x]$. Now, if we set

$$F_q = \int_{2qx}^{2qx+x} \big[Z \circ S_t\big(M_q^x\big) - \mathbb{E}(Z)\big]\,\mathrm{d}t, \tag{44}$$

then

$$\mathbb{P}\bigg(\int_0^T \big[Z \circ S_t(\boldsymbol{N}) - \mathbb{E}(Z)\big]\,\mathrm{d}t \geq u\bigg) \leq 2\mathbb{P}\bigg(\sum_{q=0}^{k-1} F_q \geq \frac{u}{2}\bigg) + 2\sum_{q=0}^{k-1}\mathbb{P}\bigg(T_{e,q} > \frac{T}{2k} - A\bigg), \tag{45}$$

where $T_{e,q}$ is the time to extinction of the process $M_q^x$. The extinction time $T_{e,q}$ is introduced in Sections 2.2 and 3 in Reynaud-Bouret and Roy (2006). Roughly speaking, it is the last time when there is an event for the Hawkes process with intensity $\boldsymbol{\lambda}(t)$ of the form (2), with background intensity $\boldsymbol{\mu} \equiv (\mu_1, \ldots, \mu_p)^{\mathrm{T}}$ set to $\mathbf{0}$ for $t \geq 0$. Since $T_{e,q}$ is identically distributed for all $q$, we can focus on one $T_{e,q}$. Denoting by $a_l$ the ancestral points with marks $l$ and by $H_{a_l}^l$ the length of the corresponding cluster whose origin is $a_l$, we have:

$$T_{e,q} = \max_{l \in \{1,\ldots,p\}} \max_{a_l} \big\{a_l + H_{a_l}^l\big\}. \tag{46}$$

Then by the exact argument on page 48 of Hansen, Reynaud-Bouret and Rivoirard (2015), we have

$$\mathbb{P}(T_{e,q} \leq a) \geq 1 - \sum_{l=1}^{p} \mu^{(l)}\mathsf{c}_l/\vartheta_l \exp(-\vartheta_l a). \tag{47}$$

Thus, there exists a constant $C_A^{(1)}$ depending on $A$ such that if we take $k =$

$\lfloor C_A^{(1)} T^\kappa \rfloor$, for some $\kappa \in (0,1)$ to be specified later, then

$$\sum_{q=0}^{k-1} \mathbb{P}\left( T_{e,q} > \frac{T}{2k} - A \right) \leq T^\kappa p \exp(-c_4 T^{1-\kappa}), \tag{48}$$

where $c_4$ is a constant. Note that $x = T/2k \approx T^{1-\kappa}$ is larger than $A$ for $T$ large enough (depending on $A$).

Now, note that the event $\mathcal{T} \equiv \{T_{e,q} \leq T/2k - A, \text{ for all } q = 0, \ldots, k\}$ only depends on the process $\boldsymbol{N}$. We will first find a probability bound for the first term in (45). In other words, we will show that, given the event $\mathcal{T}$,

$$\mathbb{P}\left( \int_0^T \left[ Z \circ S_t(\boldsymbol{N}) - \mathbb{E}(Z) \right] \mathrm{d}t \geq u \right) \leq c_5 T \exp(-c_4 T^{1-\kappa}). \tag{49}$$

Let

$$B = \mathbb{P}\left( \sum_{q=0}^{k-1} F_q \geq \frac{u}{2} \right).$$

Consider the measurable events

$$\Omega_q = \left\{ \sup_t \left\{ M_q^x|_{[t-A,t)} \right\} \leq \tilde{\mathcal{N}} \right\},$$

where $\tilde{\mathcal{N}}$ is a constant that will be defined later and $M_q^x|_{[t-A,t)}$ represents the number of points of $M_q^x$ lying in $[t-A, t)$. Let $\Omega = \bigcap_{0 \leq q \leq k-1} \Omega_q$. Then

$$B \leq \mathbb{P}\left( \sum_{q=0}^{k-1} F_q \geq u/2 \text{ and } \Omega \right) + \mathbb{P}\left(\Omega^c\right).$$

We have $\mathbb{P}(\Omega^c) \leq \sum_q \mathbb{P}(\Omega_q^c)$, where each $\mathbb{P}(\Omega_q^c)$ can be easily controlled. Indeed, it is sufficient to split $[2qx - A, 2qx + x]$ into intervals of size $A$ (there are about $C_A^{(2)} T^{1-\kappa}$ of these) and require the number of points in each sub-interval to be smaller than $\tilde{\mathcal{N}}/2$. By stationarity, we then obtain

$$\mathbb{P}\left(\Omega_q^c\right) \leq C_A^{(2)} T^{1-\kappa} \mathbb{P}(N_{[-A,0)} > \tilde{\mathcal{N}}/2).$$

Using Proposition 2 in Hansen, Reynaud-Bouret and Rivoirard (2015) with $u = \lceil \tilde{\mathcal{N}}/2 \rceil + 1/2$, we obtain:

$$\mathbb{P}\left(\Omega_q^c\right) \leq C_A^{(2)} T^{1-\kappa} \exp(-C_A^{(3)}, \tilde{\mathcal{N}})$$

and, thus,

$$\mathbb{P}\left(\Omega^c\right) \leq C_A^{(4)} T \exp(-C_A^{(3)} \tilde{\mathcal{N}}).$$

Note that this control holds for any positive choice of $\tilde{\mathcal{N}}$. Thus, for any $\tilde{\mathcal{N}} > 0$,

$$\mathbb{P}\left(\exists t \in [0,T] \text{ such that } M_q^x|_{[t-A,t)} > \tilde{\mathcal{N}}\right) \leq C_A^{(2)} T^{1-\kappa} \exp(-C_A^{(3)} \tilde{\mathcal{N}}). \tag{50}$$

Hence by taking $\tilde{\mathcal{N}} = C_A^{(5)} T^{1-\kappa}$, for $C_A^{(5)}$ large enough, the right-hand side of (50) is smaller than $C_A^{(2)} T^{1-\kappa} \exp(-c_4 T^{1-\kappa})$.

It remains to obtain the rate of $D \equiv \mathbb{P}(\sum_q F_q \geq u/2 \text{ and } \Omega)$. For any positive constant $\epsilon$ that will be chosen later, we have:

$$\begin{aligned} D &\leq \mathrm{e}^{-\epsilon u/2} \mathbb{E}\left( \mathrm{e}^{\epsilon \sum_q F_q} \prod_q \mathbb{1}_{\Omega_q} \right) \\ &\leq \mathrm{e}^{-\epsilon u/2} \prod_q \mathbb{E}\big(\mathrm{e}^{\epsilon F_q} \mathbb{1}_{\Omega_q}\big), \end{aligned} \tag{51}$$

since the variables $\{M_q^x\}_q$ are independent. But,

$$\mathbb{E}\big(\mathrm{e}^{\epsilon F_q} \mathbb{1}_{\Omega_q}\big) = 1 + \epsilon \mathbb{E}(F_q \mathbb{1}_{\Omega_q}) + \sum_{j \geq 2} \frac{\epsilon^j}{j!} \mathbb{E}\big(F_q^j \mathbb{1}_{\Omega_q}\big)$$

and $\mathbb{E}(F_q \mathbb{1}_{\Omega_q}) = \mathbb{E}(F_q) - \mathbb{E}(F_q \mathbb{1}_{\Omega_q^c}) = -\mathbb{E}(F_q \mathbb{1}_{\Omega_q^c})$.

Next note that if for any integer $l$,

$$l\tilde{\mathcal{N}} < \sup_t M_q^x|_{[t-A,t)} \leq (l+1)\tilde{\mathcal{N}},$$

then

$$|F_q| \leq xd\big[(l+1)^\eta \tilde{\mathcal{N}}^\eta + 1\big] + x\mathbb{E}(Z).$$

Hence, cutting $\Omega_q^c$ into slices of the type $\{l\tilde{\mathcal{N}} < \sup_t M_q^x|_{[t-A,t)} \leq (l+1)\tilde{\mathcal{N}}\}$ and using (50) with $\tilde{\mathcal{N}} = C_A^{(5)} T^{1-\kappa}$ for a large enough $C_A^{(5)}$, we obtain

$$\begin{aligned} &\big|\mathbb{E}(F_q \mathbb{1}_{\Omega_q})\big| = \big|\mathbb{E}(F_q \mathbb{1}_{\Omega_q^c})\big| \\ \leq\ & \sum_{l=1}^{+\infty} x\big(d\big[(l+1)^\eta \tilde{\mathcal{N}}^\eta + 1\big] + \big|\mathbb{E}(Z)\big|\big) \\ & \times \mathbb{P}\big(\exists t \in [0,T] \text{ such that } M_q^x|_{[t-A,t)} > \ell\tilde{\mathcal{N}}\big) \\ \leq\ & C_A^{(2)} \sum_{l=1}^{+\infty} x\big(d\big[(l+1)^\eta \tilde{\mathcal{N}}^\eta + 1\big] + \big|\mathbb{E}(Z)\big|\big) T^{1-\kappa} \exp(-c_4 l\tilde{\mathcal{N}}) \\ \leq\ & C_A^{(6)} \sum_{l=1}^{+\infty} x\big(d\tilde{\mathcal{N}}^\eta + \big|\mathbb{E}(Z)\big|\big) T^{1-\kappa} 2^{l\eta} \exp\{-c_4 l\tilde{\mathcal{N}}\} \\ \leq\ & C_A^{(7)} T^{2-2\kappa} d\tilde{\mathcal{N}}^\eta \frac{\exp(-c_4\tilde{\mathcal{N}})}{1 - 2^\eta \exp(-c_4\tilde{\mathcal{N}})}, \end{aligned}$$

where in the last inequality, we have used the fact that $|\mathbb{E}(Z)| \leq d\mathbb{E}[\boldsymbol{N}_{[-A,0)}^\eta]$ by (42). Plugging $\tilde{\mathcal{N}} = C_A^{(5)} T^{1-\kappa}$ into the above equation gives

$$\big|\mathbb{E}(F_q \mathbb{1}_{\Omega_q})\big| \leq z_1 \equiv C_A^{(8)} dT^{2-2\kappa} T^{(1-\kappa)\eta} \exp(-c_4 T^{1-\kappa}).$$

In the same way, following Hansen, Reynaud-Bouret and Rivoirard (2015), we can write

$$\mathbb{E}\big(F_q^j \mathbb{1}_{\Omega_q}\big) \le \mathbb{E}\big(F_q^2 \mathbb{1}_{\Omega_q}\big) z_b^{j-2}, \tag{52}$$

where $z_b \equiv xd[\tilde{\mathcal{N}}^\eta + 1] + x\mathbb{E}(Z) = C_{\eta,A}^{(9)} dT^{(1-\kappa)(1+\eta)}$. Then, by stationarity,

$$\begin{aligned}
\mathbb{E}\big(F_q^2 \mathbb{1}_{\Omega_q}\big) &\le x\mathbb{E}\left[\int_{2qx}^{2qx+x} \big[Z \circ S_{t'}\big(M_q^x\big) - \mathbb{E}(Z)\big]^2 \cap_{\tau\in\mathbb{R}} \mathbb{1}_{M_q^x|_{[\tau-A,\tau)} \le \tilde{\mathcal{N}}\}} \,\mathrm{d}t'\right] \\
&\le x\mathbb{E}\left[\int_{2qx}^{2qx+x} \big[Z \circ S_{t'}\big(M_q^x\big) - \mathbb{E}(Z)\big]^2 \mathbb{1}_{\{M_q^x|_{[t'-A,t')} \le \tilde{\mathcal{N}}\}} \,\mathrm{d}t'\right] \\
&\le x^2\mathbb{E}\big(\big[Z(\boldsymbol{N}) - \mathbb{E}(Z)\big]^2 \mathbb{1}_{N_{[-A,0)} \le \tilde{\mathcal{N}}}\big) \\
&\le z_v \equiv C_{\eta,A}^{(10)} T^{2-2\kappa}\sigma^2,
\end{aligned}$$

where $\sigma^2 \equiv \mathbb{E}\big[Z(\boldsymbol{N}) - \mathbb{E}(Z)\big]$. Going back to (51), by (52), we have

$$\begin{aligned}
D &\le \exp\left[-\frac{\epsilon u}{2} + k\log\left(1 + \epsilon z_1 + \sum_{j\ge 2} z_v z_b^{j-2}\frac{\epsilon^j}{j!}\right)\right] \\
&\le \exp\left[-\epsilon\left(\frac{u}{2} - kz_1\right) + k\sum_{j\ge 2} z_v z_b^{j-2}\frac{\epsilon^j}{j!}\right],
\end{aligned}$$

using the fact that $\log(1+u) \le u$. Since

$$kz_1 = C_\eta^{(10)} dT^\kappa T^{(2+\eta)(1-\kappa)} \exp(-c_4 T^{1-\kappa}),$$

one can choose $c_6$ in the definition (43) of $u$ (not depending on $d$) such that $u/2 - kz_1 \ge \sqrt{2kz_v z} + \frac{1}{3}z_b z$ for some $z = c_4 T^{\kappa - 2\eta(1-\kappa)}$. Hence,

$$D \le \exp\left[-\epsilon\left(\sqrt{2kz_v z} + \frac{1}{3}z_b z\right) + k\sum_{j\ge 2} z_v z_b^{j-2}\frac{\epsilon^j}{j!}\right].$$

One can choose $\epsilon$ (as in the proof of the Bernstein inequality in Massart (2007), page 25) to obtain a bound on the right-hand side in the form of $\mathrm{e}^{-z}$. We can then choose $c_4$ large enough, and only depending on $\eta$ and $A$, to guarantee that $D \le \mathrm{e}^{-z} \le c_5 \exp(-c_4 T^{1-\kappa})$.

In summary, we have shown that, given the event $\mathcal{T}$,

$$\mathbb{P}\left(\int_0^T \big[Z \circ S_t(\boldsymbol{N}) - \mathbb{E}(Z)\big]\,\mathrm{d}t \ge u\right) \le c_5 \exp(-c_4 T^{1-\kappa}) + C_A^{(4)} T \exp(-c_4 T^{1-\kappa}).$$

With a slight abuse of notation, letting $c_5 = \max(c_5, C_A^{(4)})$ gives (49).

To complete the proof, we apply the concentration inequality (49) with some specific choices of $Z(\cdot)$.

For each pair $(j,k)$, let

$$Z \circ S_t(\boldsymbol{N}) \equiv \int_{t-h}^{t+h} K\left(\frac{t' - t + \Delta}{h}\right) \mathrm{d}N_j(t')\,\mathrm{d}N_k(t)/\mathrm{d}t.$$

We can check that $d = 1$ and $\eta = 2$ satisfy (42). Then with $\kappa = 5/6$ in (49), we get, given the event $\mathcal{T}$,

$$\mathbb{P}\left(|\mathrm{I}_{j,k} - \mathbb{E}\mathrm{I}_{j,k}| \geq c_6 T^{-1/3}\right) \leq c_5 T \exp(-c_4 T^{1/6}).$$

Applying a union bound for all pairs $(j, k)$, we have, given the event $\mathcal{T}$,

$$\mathbb{P}\left(\bigcap_{1\leq j\leq k\leq p}\left[|\mathrm{I}_{j,k} - \mathbb{E}\mathrm{I}_{j,k}| \geq c_6 T^{-1/3}\right]\right) \leq c_5 T p^2 \exp(-c_4 T^{1/6}). \tag{53}$$

Recall from the concentration inequality (48) that the event $\mathcal{T}$ holds with probability at least $1 - pT^{1/6}\exp(-c_4 T^{1/6})$. Thus, given that $pT^{1/6}\exp(-c_4 T^{1/6})$ is dominated by the right-hand side of (53), it holds unconditionally that

$$\mathbb{P}\left(\bigcap_{1\leq j\leq k\leq p}\left[|\mathrm{I}_{j,k} - \mathbb{E}\mathrm{I}_{j,k}| \geq c_6 T^{-1/3}\right]\right\} \leq c_5 T p^2 \exp(-c_4 T^{1/6}),$$

which is the statement on $\mathrm{I}_{j,k}$ in (34).

The statement on $\mathrm{II}_l$, $l = j, k$, in (35) can be shown in a similar manner by taking $Z \circ S_t(\boldsymbol{N}) \equiv \mathrm{d}N_j(t)/\mathrm{d}t$, with $\eta = 1$, and $\kappa = 13/18$. □