Topics on Learning Network Structure from Multivariate Point Process Data

by

Muhong Gao

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Statistics)

at the

UNIVERSITY OF WISCONSIN–MADISON

2022

Date of final oral examination: 12/7/2021

The dissertation is approved by the following members of the Final Oral Committee:

Chunming Zhang, Professor, Statistics

Zhengjun Zhang, Professor, Statistics

Nicolas Garcia Trillos, Assistant Professor, Statistics

David Anderson, Professor, Mathematics

Hanbaek Lyu, Assistant Professor, Mathematics

Acknowledgments

I would like to express my gratitude to my advisor, Professor Chunming Zhang who provided substantial guidance and support in my research work. Her passion in research has influenced me in many ways which leads me to consider and pursue an academic position. Working with Professor Zhang is a very interesting and exciting process. Without her patient mentorship and encouragements, this thesis would not have been possible.

Additionally, I am grateful to my thesis committee members: Professor Zhengjun Zhang, Nicholas Garcia Trillos, David Anderson and Hanbaek Lyu for their advice and suggestions on my thesis work.

I am fortunate to join inspiring research group meeting held by my advisor. I want to thank those who did presentations and participated in the research group. In addition, I would like to also thank professors who contributed in my Ph.D. program: Yazhen Wang, Jun Zhu, Brian Yandell, Michael Newton, Po-Ling Loh, Jun Shao, Anru Zhang and Kam-Wah Tsui. Moreover, I want to thank Dr. Derek Bean for helping me in improving my teaching skills and presentations with his feedback.

I would like to acknowledge my fellow graduate students and friends in the Department of Statistics for helping me in their ways such as Yongfeng Wu, Yukun Chen, Hao Yang Teng, Yongsu Lee, Yanbo Shen, Bowen Zhang, Taiyu Ye, Muxuan Liang, Jianchang Hu, Jacob Maronge, Xiaowu Dai, and Duzhe Wang.

Last but not least, I would like to thank my parents for their unconditional love and support especially during my Ph.D. life.

The work was partially supported by the U.S. National Science Foundation grants DMS-

2013486 and DMS-1712418, and provided by the University of Wisconsin-Madison Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation.

Contents

List of Tables

List of Figures

Abstract

This thesis work introduces a new method for learning the sparse network-structured dependence among nodes from "multivariate point process" data. Learning network structure from point process data $\{\boldsymbol{T}_i\}_{i\in\mathcal{V}}$ has wide applications in neuroscience, biology and information transmission. The commonly used discrete-time modeling relies on empirical assumptions of Poisson distributions for bin counts of events in equally-spaced time bins with stationary and independent increments, which limit the scope as applied to practical applications. This thesis work develops continuous-time stochastic models of the "conditional intensity processes" $\{\lambda_i(t) : t \geq 0\}_{i\in\mathcal{V}}$ for learning the network structure, underlying an array of non-stationary "multivariate counting processes" $\{\boldsymbol{N}(t) : t \geq 0\}$ for $\{\boldsymbol{T}_i\}_{i\in\mathcal{V}}$. Furthermore, we introduce the "marked point process for intensity discontinuities", derive their "explicit forms of conditional distributions" and show the "cyclicity property of $\boldsymbol{N}(t)$ driven by recurrence time points". These theoretical properties provide new insights into establishing statistical consistency and convergence properties of the proposed penalized M-estimators for graph parameters relevant to structure recovery, under mild regularity conditions. Simulation evaluations demonstrate computational simplicity of the proposed method, and increased estimation accuracy over existing methods. Real multiple neuron spike train recordings are analyzed to infer connectivity in neuronal networks.

Chapter 1

Introduction

Literature reviews on modeling of multivariate point process, inference and estimation of network structure, and application examples are presented in this chapter.

Structured multivariate point process data has wide applications, ranging from recordings from neuron multiple spike trains, file access patterns and failure events in server farms, to queuing networks. Inference of the network structure underlying such multivariate point processes and addressing queries based on the learned structure are important issues. For example, learning the structure of cooperative activity between multiple neurons is an important task in understanding neural spike activity and identifying patterns of information transmission and storage in cortical circuits; see Brillinger and Villa (1994), Aertsen et al. (1989), Oram et al. (1999), Harris et al. (2003), Barbieri et al. (2001), and Brown et al. (2004). Analogously, learning the access patterns of files can be exploited for developing faster file access systems.

Typically, multivariate "point processes" refer to random processes of occurrences of a particular event (such as neuron spike firing) in time and geographical spaces, in the form of $\{\boldsymbol{T}_1, \ldots, \boldsymbol{T}_V\}$ recorded at V nodes, where

$$\boldsymbol{T}_i = (T_{i,1},\ \ldots,\ T_{i,\mathrm{N}_i})^\top \quad \text{with } 0 < T_{i,1} < \cdots < T_{i,\mathrm{N}_i} \leq \mathrm{T}, \quad \text{for } i \in \mathcal{V}, \tag{1.1}$$

correspond to series of time points $T_{i,\ell}$ of the ℓth event arriving at the ith node in an experiment with time length T, where the superscript $\top$ denotes transpose and $\mathcal{V} = \{1, \ldots, V\}$ is the node set. The corresponding "counting process" $N_i(t) = \sum_{\ell \geq 1} \mathrm{I}(0 \leq T_{i,\ell} \leq t)$ counts the number of events that occur up to time t at node $i \in \mathcal{V}$. An important goal is to extract the dependency structure among nodes, connected in the network, from V sequences of time series. Figure 1 illustrates the network-structured dependence (in the left panel) of multivariate point process data at 5 nodes (in the right panel).

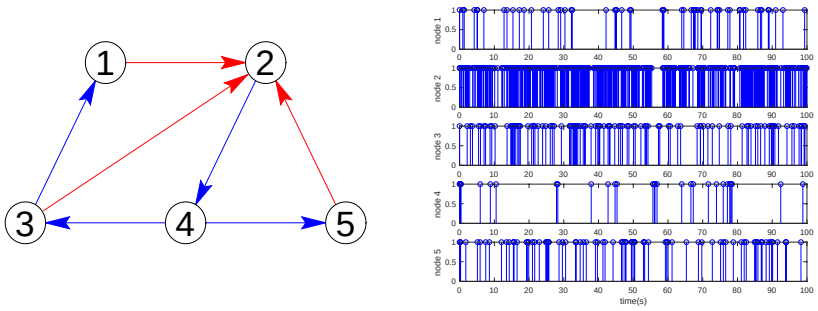

Figure 1: Each node of the network graph in the left panel corresponds to a point process in the right panel. Arrows indicate interactions (red for excitatory and blue for inhibitory effects).

Due to the stochastic nature of the point process data $\{T_{i,\ell}\}$ in (1.1) for event occurrences, there are two types of methods that are relevant to modeling multivariate point process data. (a) The discrete-time modeling approach includes the dynamic Bayesian network in Murphy (2002), and variants of generalized linear models (GLM) in Brillinger and Villa (1994), Kelly et al. (2010), Truccolo et al. (2005) and Zhang et al. (2016), which partition the time axis into equally spaced time bins, and transform the series of event arrival times into a sequence of event bin counts, empirically modelled by Poisson distributions, but meanwhile, suffer from the loss of information caused by the discrete approximation error. (b) In contrast, the continuous-time approach ideally depicts physical processes, but is substantially challenged by modeling not only the part of "intensity processes" to be time-varying, but also the part of sparsity feature underlying the network structure. Several research efforts have been made to develop specific continuous-time point process models, such as the Cox model

in Masud and Borisyuk (2011), inhomogenous Poisson process in Rajaram et al. (2005), and the Hawkes process in Chornoboy et al. (1988), Reynaud-Bouret and Schbath (2010), Xu et al. (2016). On the other hand, most existing works focus on numerical algorithms, with fewer theoretical results developed for statistical learning due to the complexity of continuous-time point processes.

To capture the dependency between point process data represented in (1.1) both qualitatively and quantitatively, we aim to develop novel network structure learning methods which integrate the utility of continuous-time and discrete-time modelings. Specifically, we build continuous-time GLM-type stochastic models (3.1) for the "conditional intensity processes" $\{\lambda_i(t)\}_{i\in\mathcal{V}}$, where each $\lambda_i(t)$ depends on short-term past events of all other nodes up to time t and embeds the magnitude and direction of interaction effects in graph parameters. A sparse network will be obtained via penalized M-estimation of parameters in the graph structure; see (5.8) and (6.6). Compared with the linear Hawkes process, our method could better capture not only the excitatory but also the inhibitory effects between nodes. Compared with the discrete-time approach, our method is partition-free, avoiding the subjective choice of the bin width.

Addressing the theoretical issues in the interface between continuous-time stochastic modeling and statistical learning procedure is a non-trivial task. Indeed, conventional tools for proving stochastic convergence and statistical consistency are not directly applicable in the context of statistical estimation from point process data, since the loss function (e.g., in (5.4)) for parameter estimation relies primarily on the non-standard dependence structure of counting processes $\{N_i(t)\}_{i\in\mathcal{V}}$, associated with point process data $\{T_{i,\ell}\}$. Hansen et al. (2015) used an approach based on the Hawkes process, which provides some insights, but requires the stationarity assumption. We will develop novel technical tools which contribute to the statistical analysis of a wide array of non-stationary multivariate point process data.

(i) Specifically, by applying a novel tool of the "marked point process $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ for intensity discontinuities", which equivalently represents the multivariate point process data $\{T_{i,\ell}\}$, we explicitly derive the probability distributions of $\breve{\boldsymbol{T}}$ and $\boldsymbol{I}$, accompanied with

a sequence of useful properties in Section 4.1.

(ii) We then prove the "cyclicity property of $\{N_i(t)\}_{i\in\mathcal{V}}$ driven by recurrence time points", and further exhibit the "asymptotic mean stationarity" in Section 4.2.

(iii) All these probabilistic results are essential for deriving the statistical properties, such as consistency of the proposed penalized M-estimation in structure learning, in Chapter 5.

(iv) Moreover, we demonstrate that the developed theoretical results are not confined to our specific model for $\{\lambda_i(t)\}_{i\in\mathcal{V}}$, but carry through to a wide array of continuous-time models arising from non-stationary point processes. See Section 4.3 and Chapter 6.

Extensive simulation studies lend support to the validity of our proposed penalization method for recovering network-structured dependency, which is also applied to a pre-frontal cortex spike train dataset to illustrate its practical utility in the analysis of real-world multivariate point process data.

The thesis is organized as follows. Chapter 2 reviews the multivariate point process, and outlines the proposed continuous-time modeling framework. Chapter 3 presents our proposed model for $\{\lambda_i(t)\}_{i\in\mathcal{V}}$, and Chapter 4 investigates related probabilistic properties of $\{N_i(t)\}_{i\in\mathcal{V}}$. Chapter 5 addresses statistical properties related to the proposed network recovery procedure. Chapter 6 introduces an extended model for structure learning. Chapter 7 illustrates simulation evaluations of the proposed method, and Chapter 8 analyzes a real spike train data. Chapter 9 gives a brief discussion. All technical details and derivations are relegated to Appendices A and B.

Chapter 2

Overview of the proposed continuous-time modeling

In this chapter, we first introduce the background of the multivariate point process with the conditional intensity process, followed by an outline of our proposed continuous-time modeling and parameter estimation procedure.

2.1. Multivariate point process in our set-up

We start with a brief review of the point process. Refer to Daley and Vere-Jones (2003) for a more comprehensive discussion of the point processes. Denote by $\mathcal{V} = \{1, \ldots, V\}$ the set of nodes. Throughout the thesis, we focus on the setting where the number V of nodes is a fixed constant. For each node $i \in \mathcal{V}$, define the univariate "point process" by $\{T_{i,\ell}\}_{\ell \geq 1}$ on the probability space $(\Omega, \mathscr{F}, \mathrm{P})$, where

$$0 < T_{i,1} < T_{i,2} < \cdots \tag{2.1}$$

denote the time-ordered sequence of event-occurrence time points at this node i. For $t \geq 0$, denote the event counts in the time interval $[0, t]$ by

$$N_i(t) = \sum_{\ell \geq 1} \mathrm{I}(0 \leq T_{i,\ell} \leq t), \tag{2.2}$$

where $\mathrm{I}(\cdot)$ denotes the indicator operator. We term $\{N_i(t)\}_{t \geq 0}$ the "counting process" of $\{T_{i,\ell}\}_{\ell \geq 1}$. More generally, we denote the event counts in any Borel set $\mathcal{T} \in \mathcal{B}([0, \infty))$ by

$$N_i(\mathcal{T}) = \sum_{\ell \geq 1} \mathrm{I}(T_{i,\ell} \in \mathcal{T}), \tag{2.3}$$

which, for $\mathcal{T} = [0, t]$, reduces to $N_i(t)$ in (2.2).

According to (2.2), a point process $\{T_{i,\ell}\}_{\ell \geq 1}$ uniquely defines a counting process $\{N_i(t)\}_{t \geq 0}$. Conversely, $\{N_i(t)\}_{t \geq 0}$ uniquely yields a point process, due to the identity $T_{i,\ell} = \inf\{t > T_{i,\ell-1} : N_i(t) > N_i(T_{i,\ell-1})\}$. Thus, the counting process $\{N_i(t)\}_{t \geq 0}$ and the point process $\{T_{i,\ell}\}_{\ell \geq 1}$ are equivalent to each other.

For the multivariate setting with V nodes, we define the vector $\boldsymbol{N}(t) = (N_1(t), \ldots, N_V(t))^\top$, and call $\{\boldsymbol{N}(t)\}_{t \geq 0}$ the "multivariate counting process", corresponding to the "multivariate point process" $\{T_{i,\ell} : \ell \geq 1\}_{i \in \mathcal{V}}$. For each $t \geq 0$, let $\mathscr{F}_t \subseteq \mathscr{F}$ be the smallest sub σ-algebra that contains all the information of the multivariate counting process in the history up to time t, formally defined as

$$\mathscr{F}_t = \sigma\big(\{N_i(s) : s \in [0, t],\ i \in \mathcal{V}\}\big). \tag{2.4}$$

From (2.4), it is seen that

$$\mathscr{F}_{t_1} \subseteq \mathscr{F}_{t_2} \subseteq \cdots, \quad \text{for any } 0 \leq t_1 \leq t_2 \leq \cdots. \tag{2.5}$$

We refer to the sequence of σ-algebras $\{\mathscr{F}_t\}_{t \geq 0}$ in (2.4), satisfying the property (2.5), as the "filtration generated by $\{\boldsymbol{N}(t)\}_{t \geq 0}$", and call $(\Omega, \mathscr{F}, \{\mathscr{F}_t\}_{t \geq 0}, \mathrm{P})$ the corresponding "filtered

probability space".

2.1.1. Total intensity process of $\boldsymbol{N}(t)$

For a single node i, the stochastic character of a counting process $N_i(t)$ is captured by the corresponding "intensity process" (also called "conditional intensity function") $\lambda_i(t)$, which measures the instantaneous rate of event occurrence at node i. In this thesis, we adopt the definition of the intensity process from Rubin (1972):

$$\lambda_i(t) = \lim_{\Delta \downarrow 0} \frac{1}{\Delta} \mathrm{P}\big(N_i(t+\Delta) = N_i(t) + 1 \mid \mathscr{F}_t\big) \tag{2.6}$$

$$= \lim_{\Delta \downarrow 0} \frac{1}{\Delta} \mathrm{P}\big(N_i(t+\Delta) \neq N_i(t) \mid \mathscr{F}_t\big), \quad \text{a.s.} \tag{2.7}$$

for $i \in \mathcal{V}$ and $t \geq 0$.

For the multivariate case with V nodes, we similarly define the "total intensity process" of $\boldsymbol{N}(t)$ by

$$\lambda^{\mathrm{sum}}(t) = \lim_{\Delta \downarrow 0} \frac{1}{\Delta} \mathrm{P}\big(\cup_{i \in \mathcal{V}} \{N_i(t+\Delta) = N_i(t) + 1\} \mid \mathscr{F}_t\big) \tag{2.8}$$

$$= \lim_{\Delta \downarrow 0} \frac{1}{\Delta} \mathrm{P}\big(\boldsymbol{N}(t+\Delta) \neq \boldsymbol{N}(t) \mid \mathscr{F}_t\big), \quad \text{a.s.}. \tag{2.9}$$

Remark 1. *For a general univariate point process, the two limits* (2.6) *and* (2.7) *may not be identical. Rubin* (1972) *introduced the notion of "regular point process" for a univariate counting process* $\{N(t)\}_{t \geq 0}$ *in which the two limits "*$\lim_{\Delta \downarrow 0} \frac{1}{\Delta} \mathrm{P}(N(t+\Delta) = N(t) + 1 \mid \mathscr{F}_t)$*" and "*$\lim_{\Delta \downarrow 0} \frac{1}{\Delta} \mathrm{P}(N(t+\Delta) \neq N(t) \mid \mathscr{F}_t)$*" are identical* a.s. *for every* $t \geq 0$*. Our work extends this definition to the multivariate counting process, and calls* $\{\boldsymbol{N}(t)\}_{t \geq 0}$ *"regular", if* $\lim_{\Delta \downarrow 0} \frac{1}{\Delta} \mathrm{P}(N_i(t+\Delta) = N_i(t) + 1 \mid \mathscr{F}_t) = \lim_{\Delta \downarrow 0} \frac{1}{\Delta} \mathrm{P}(N_i(t+\Delta) \neq N_i(t) \mid \mathscr{F}_t)$ *holds* a.s. *for every* $i \in \mathcal{V}$ *and* $t \geq 0$*. As shown in Appendix B, any multivariate "regular" counting process* $\{\boldsymbol{N}(t)\}_{t \geq 0}$ *also has identical limits* (2.8) *and* (2.9) a.s. *for every* $t \geq 0$*. Throughout the thesis, we assume that our multivariate counting process* $\{\boldsymbol{N}(t)\}_{t \geq 0}$ *is "regular". Hence, the intensity process* $\lambda_i(t)$ *at node* i *could be defined by either* (2.6) *or* (2.7)*, and the total intensity process* $\lambda^{\mathrm{sum}}(t)$ *by*

either (2.8) *or* (2.9).

2.1.2. Conditional independence for $\boldsymbol{N}(t)$

For the multivariate setting in our study, the structure of the counting process $\boldsymbol{N}(t)$ could not be fully described by solely presenting the intensity processes $\lambda_i(t)$ at individual nodes i. Besides, it needs to clarify how the increments of event counts, $N_i(t+\Delta) - N_i(t)$ and $N_j(t+\Delta) - N_j(t)$, are correlated between any pair of distinct nodes i and j. In Definition 1 below, we introduce the notion of "conditional independence", which refers to the case where $N_i(t+\Delta) - N_i(t)$ and $N_j(t+\Delta) - N_j(t)$, conditional on $\mathscr{F}_t$, are asymptotically independent for all $i \neq j$.

Definition 1 (Conditional independence)**.** *A multivariate counting process $\boldsymbol{N}(t)$ is called "conditionally independent", if for any two distinct nodes $i, j \in \mathcal{V}$, and any time $t \geq 0$,*

$$\begin{aligned}
&\lim_{\Delta \downarrow 0} \frac{1}{\Delta^2} \mathrm{P}\big(N_i(t+\Delta) = N_i(t) + 1, N_j(t+\Delta) = N_j(t) + 1 \mid \mathscr{F}_t\big) \\
&= \lim_{\Delta \downarrow 0} \frac{1}{\Delta^2} \mathrm{P}\big(N_i(t+\Delta) = N_i(t) + 1 \mid \mathscr{F}_t\big) \cdot \mathrm{P}\big(N_j(t+\Delta) = N_j(t) + 1 \mid \mathscr{F}_t\big) \\
&= \lambda_i(t)\lambda_j(t), \quad \text{a.s.}.
\end{aligned} \tag{2.10}$$

Lemma 1 shows that for a "conditionally independent" multivariate counting process $\boldsymbol{N}(t)$, the total intensity process $\lambda^{\mathrm{sum}}(t)$ in (2.8)–(2.9) equals the sum of all intensity processes $\lambda_i(t)$ over individual nodes $i \in \mathcal{V}$. In the rest of the thesis, we always assume that the multivariate counting process $\boldsymbol{N}(t)$ is "conditionally independent".

Lemma 1 (Total intensity of the multivariate counting process $\boldsymbol{N}(t)$)**.** *Assume conditions* A1 *and* A2 *in Appendix A. Assume that* $\mathrm{P}(\lambda_i(t) < \infty) = 1$ *for all $i \in \mathcal{V}$ and $t \geq 0$. If $\boldsymbol{N}(t)$ is "conditionally independent", then for any $t \geq 0$, the total intensity process $\lambda^{\mathrm{sum}}(t)$ defined in* (2.8)–(2.9) *satisfies*

$$\lambda^{\mathrm{sum}}(t) = \sum_{i=1}^{V} \lambda_i(t), \quad \text{a.s.}. \tag{2.11}$$

2.2. Outline of the proposed continuous-time modeling

The proposed continuous-time modeling procedure is outlined as follows:

(**Statistical modeling of the intensity process** $\lambda_i(t)$) Chapter 3 presents our proposed modeling (3.1) for the intensity process $\lambda_i(t)$ motivated from the GLM-type framework, with interpretations relevant to the practical network structure of neuron ensembles.

(**Stochastic framework and probabilistic properties of** $\boldsymbol{N}(t)$) Chapter 4 derives novel theoretical results for the probabilistic properties of model (3.1) for $\lambda_i(t)$, which are applicable to analyzing a wide array of multivariate point processes $\boldsymbol{N}(t)$.

(i) In Section 4.1, we introduce a new framework, the "marked point process $(\breve{\boldsymbol{T}}, \boldsymbol{I}) = (\{\breve{T}_\ell\}_{\ell\geq 1}, \{I_\ell\}_{\ell\geq 1})$" for studying discontinuity points of $\{\lambda_i(t)\}_{i\in\mathcal{V}}$ and the probabilistic features of the counting process $\boldsymbol{N}(t)$. We first derive the probability distribution of $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ in Theorem 1, together with the related probabilistic properties in Lemmas 3, 4, and 5. These results are then translated into analogous results of the counting process $\boldsymbol{N}(t)$ in Theorem 2.

(ii) In Section 4.2, a novel "cyclicity" feature of $\boldsymbol{N}(t)$ driven by "recurrence time points" $\{R_\ell\}_{\ell\geq 1}$ is given in Theorem 3, which further leads to the "asymptotic mean stationarity" in Theorem 4 for $\boldsymbol{N}(t)$.

(iii) In Section 4.3, Theorem 5 demonstrates that probabilistic results studied in Sections 4.1 and 4.2 can be extended for a wider class of intensity processes.

(**Estimation of graph parameters and learning of network structure**) We propose in Chapter 5 the penalized M-estimator $\widehat{\widetilde{\boldsymbol{\beta}}}_i$ for true values of graph parameters $\widetilde{\boldsymbol{\beta}}_i$ in modeling $\lambda_i(t)$, and verify in Section 5.3 parameter estimation consistency and network recovery consistency with Theorems 7–8 and Corollary 1, which require results of Theorems 2–4 in Chapter 4. Chapter 6 gives Theorem 9 and Corollary 2 for an extended class of $\lambda_i(t)$.

Chapter 3

Modeling $\lambda_i(t)$ by network structure

The existing work on modeling the intensity process $\lambda_i(t)$ by network structure mainly includes two types of approaches: (i) the continuous-time modeling, e.g., Chornoboy et al. (1988), Masud and Borisyuk (2011), and Hansen et al. (2015), typically based on either Hawkes or Poisson process (Definition B in Appendix A); and (ii) discrete-time modeling, e.g., Brillinger and Villa (1994), Truccolo et al. (2005), Zhang et al. (2016), and Zhao et al. (2012), usually based on the generalized linear model with the exponential link function. It is thus useful to incorporate the advantages of both modeling strategies, leading to our proposed continuous-time GLM-type modeling for $\lambda_i(t)$,

$$\lambda_i(t) = \exp\Big\{\beta_{0;i} + \sum_{j\in\mathcal{V}} \beta_{j,i}\, x_j(t)\Big\}, \quad i \in \mathcal{V}, \quad t \geq 0, \tag{3.1}$$

with interpretations of parameters $\beta_{0;i}$ and $\beta_{j,i}$ and covariates $x_j(t)$ depicted as follows:

Baseline intensity parameter $\beta_{0;i}$. Since the background intensity may vary with nodes, we include the scalar parameter $\beta_{0;i}$ into (3.1) to associate the baseline intensity parameter with each node i.

Connection strength parameter $\beta_{j,i}$. The connection strength parameter $\beta_{j,i}$ in (3.1) quantifies the magnitude and direction of a parent node j's influence on the child

node i, i.e., $\boldsymbol{j} \xrightarrow{\beta_{j,i}} \boldsymbol{i}$; particularly,

$$\begin{array}{lll} \beta_{j,i} > 0: & \text{excitatory} & \text{effect from node } j \text{ to node } i; \\ \beta_{j,i} = 0: & \text{no} & \text{effect from node } j \text{ to node } i; \\ \beta_{j,i} < 0: & \text{inhibitory} & \text{effect from node } j \text{ to node } i. \end{array}$$

In addition, we assume $\beta_{i,i} = 0$ for all $i \in \mathcal{V}$, i.e., there is no "self effect". The network graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ can be obtained from all pairs of nodes (j, i) with non-zero connection parameters $\beta_{j,i}$ in the edge set,

$$\mathcal{E} = \{(j,i) \in \mathcal{V} \times \mathcal{V} : \beta_{j,i} \neq 0,\ j \neq i\} = \mathcal{E}_+ \cup \mathcal{E}_-, \tag{3.2}$$

which distinguishes the edge set for excitatory effects

$$\mathcal{E}_+ = \{(j,i) \in \mathcal{V} \times \mathcal{V} : \beta_{j,i} > 0,\ j \neq i\}, \tag{3.3}$$

from the edge set for inhibitory effects

$$\mathcal{E}_- = \{(j,i) \in \mathcal{V} \times \mathcal{V} : \beta_{j,i} < 0,\ j \neq i\}. \tag{3.4}$$

Configuration of this graph $\mathcal{G}$ reveals the interaction effects between nodes, and learning such graph structure via statistical estimation methods is the main goal of this thesis.

Regression covariates $x_j(t)$. Regression covariates $x_j(t)$ aim to represent the effect from other nodes $j \in \mathcal{V}$ on node i, within a short period of time, earlier than t. We formulate $x_j(t)$ by

$$x_j(t) = g\big(r_{j,\phi}(t)\big), \tag{3.5}$$

via the empirical firing rate,

$$r_{j,\phi}(t) = \frac{N_j\big(\big((t-\phi) \vee 0, t\big]\big)}{\phi}, \tag{3.6}$$

during a short time interval of width $\phi \in (0, \infty)$, where $a \vee b = \max(a, b)$, and a non-linear "shape function" $g(\cdot) : [0, \infty) \to [0, \infty)$, which is continuous, non-negative, and monotonically increasing, with $g(0) = 0$. In parameter estimation and theoretical analysis, the function $g(\cdot)$ and the constant ϕ are known.

The proposed model (3.1) utilizes the GLM-type framework to associate the intensity process $\lambda_i(t)$ with both the historical data and the network structure, delivering a novel continuous-time approach for modeling multivariate point process data. As illustrated below, using two specific choices of the "shape function" $g(\cdot)$ in (3.5) (combined with (3.6)), our model (3.1) connects with two existing models.

Example 1: $g(x) = x$. Then model (3.1) is

$$\begin{aligned}\lambda_i(t) &= \exp\Big\{\beta_{0;i} + \sum_{j\in\mathcal{V}} \beta_{j,i}\, r_{j,\phi}(t)\Big\} \\ &= \exp\Big\{\beta_{0;i} + \sum_{j\in\mathcal{V}} \int_{-\infty}^{t} \frac{1}{\phi}\, \beta_{j,i}\, \mathrm{I}(0 \le t-u < \phi)\, \mathrm{d}N_j(u)\Big\}, \end{aligned} \tag{3.7}$$

a special case of the general multivariate non-linear Hawkes process in Brémaud and Massoulié (1996):

$$\lambda_i(t) = \varphi\Big(\beta_{0;i} + \sum_{j\in\mathcal{V}} \int_{-\infty}^{t} \omega_{j,i}(t-u)\, \mathrm{d}N_j(u)\Big), \tag{3.8}$$

when we set the non-linear link function $\varphi(\cdot) = \exp(\cdot)$, and the interaction function $\omega_{j,i}(u) = \beta_{j,i}\, \mathrm{I}(0 \le u < \phi)/\phi$, with $\beta_{i,i} = 0$. In a general non-linear Hawkes process (3.8), the interaction function $\omega_{j,i}(\cdot)$, not constrained to be indicator functions, allows the convolution covariate $\int_{-\infty}^{t} \omega_{j,i}(t-u)\, \mathrm{d}N_j(u)$ to be more flexible. Chapter 6 will discuss an extension of our model (3.1), and its comparison with the non-linear Hawkes process.

Example 2: $g(x) = \log(1+x)$. Then model (3.1) becomes:

$$\begin{aligned}\lambda_i(t) &= \exp\Big(\beta_{0;i} + \sum_{j\in\mathcal{V}} \beta_{j,i}\, \log\{1 + r_{j,\phi}(t)\}\Big) \\ &= \exp(\beta_{0;i}) \prod_{j\in\mathcal{V}} \{1 + r_{j,\phi}(t)\}^{\beta_{j,i}}, \end{aligned} \tag{3.9}$$

agreeing with Rajaram et al. (2005). Compared with Example 1, the "shape function" $g(x) = \log(1+x)$ in Example 2 is relatively flat, and thus moderates the steepness of the exponential link function and down-weights the influence of excessively large intensities. Thus, (3.9) is expected to better portray the dynamics of multivariate point process data in real applications.

Chapter 4

Probabilistic properties of the proposed model

The stochastic properties of a counting process primarily depend on its associated intensity process. In this chapter, we investigate the probabilistic properties of the counting process $\boldsymbol{N}(t)$ associated with the intensity processes $\{\lambda_i(t)\}_{i\in\mathcal{V}}$ in our model (3.1). These probabilistic results are essentially required for deriving our statistical properties (Theorems 6–9 and Corollaries 1–2) in Sections 5 and 6.

A distinctive feature of our intensity processes $\{\lambda_i(t)\}_{i\in\mathcal{V}}$ in (3.1) is that they are piecewise-constant functions of time t (as to be shown in Section 4.1.1). In other words, unlike many other models, our $\{\lambda_i(t)\}_{i\in\mathcal{V}}$ do not change continuously over time, yielding a countable number of discontinuity points in $(0,\infty)$ from all nodes. As we will see, the discontinuity points of $\{\lambda_i(t)\}_{i\in\mathcal{V}}$ play an important role in characterizing the stochastic features of our intensity processes. We start by investigating the set of discontinuity points of $\{\lambda_i(t)\}_{i\in\mathcal{V}}$ in Section 4.1.

4.1. Marked point process for intensity discontinuities

In this section, we conduct a step-by-step analysis based on the discontinuity points of $\{\lambda_i(t)\}_{i\in\mathcal{V}}$ in (3.1). Section 4.1.1 exhibits the piecewise-constantness of $\{\lambda_i(t)\}_{i\in\mathcal{V}}$. Section 4.1.2 defines the "marked point process $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ for intensity discontinuities", which proves (in (4.6)–(4.7)) to be equivalent for studying the point process $\{T_{i,\ell}\}_{\ell\geq 1,\, i\in\mathcal{V}}$ and the counting process $\boldsymbol{N}(t)$. Section 4.1.3 derives the probability distribution of $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ (in Theorem 1), and the related probabilistic properties (in Lemmas 3–5). By translating the results of $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ into the analogues of $\boldsymbol{N}(t)$, Section 4.1.4 presents the bounded variance and the finiteness (in Theorem 2) for our counting process $\boldsymbol{N}(t)$.

4.1.1. Piecewise-constant $\lambda_i(t)$

Recall that from (3.1), (3.5) and (3.6), the conditional intensity function at each node $i \in \mathcal{V}$ can be re-written as $\lambda_i(t) = \exp\{\beta_{0;i} + \sum_{j\in\mathcal{V}} \beta_{j,i}\, g\big(N_j(((t-\phi)\vee 0, t])/\phi\big)\}$, which is a continuous function of $\{N_j(((t-\phi)\vee 0, t])\}_{j\in\mathcal{V}}$, with

$$N_j\big(((t-\phi)\vee 0, t]\big) \;=\; N_j(t) - N_j(t-\phi), \quad t \geq 0. \tag{4.1}$$

Thus the smoothness of $\lambda_i(t)$ directly depends on that of $\{N_j(((t-\phi)\vee 0, t])\}_{j\in\mathcal{V}}$.

For each node $j \in \mathcal{V}$ and event time points $\{T_{j,\ell}\}_{\ell\geq 1}$ in (2.1), define $N_j(\{t\}) = \sum_{\ell\geq 1} \mathrm{I}(T_{j,\ell} = t)$, which is the jump size of $N_j(\cdot)$ at a single point t. Clearly, $N_j(\{t\}) \in \{0, 1\}$, and $N_j(\{t\}) = 1$ is equivalent to $t \in \{T_{j,\ell}\}_{\ell\geq 1}$. Moreover, two properties of $N_j\big(((t-\phi)\vee 0, t]\big)$ can be verified. First, $N_j(((t-\phi)\vee 0, t])$ is non-negative, right-continuous, piecewise-constant, but is not monotonically increasing in $t \in [0, \infty)$. Second, the set of discontinuity points of $N_j(((t-\phi)\vee 0, t])$ is

$$\big\{t \geq 0 : N_j(\{t\}) - N_j(\{t-\phi\}) = +1\big\} \cup \big\{t \geq 0 : N_j(\{t\}) - N_j(\{t-\phi\}) = -1\big\}, \tag{4.2}$$

where

$$N_j(\{t\}) - N_j(\{t-\phi\}) = \begin{cases} 0, & \text{if } t \notin \{T_{j,\ell}\}_{\ell\geq 1}, \text{and } t \notin \{T_{j,k}+\phi\}_{k\geq 1}, \\ +1, & \text{if } t \in \{T_{j,\ell}\}_{\ell\geq 1}, \text{and } t \notin \{T_{j,k}+\phi\}_{k\geq 1}, \\ -1, & \text{if } t \notin \{T_{j,\ell}\}_{\ell\geq 1}, \text{and } t \in \{T_{j,k}+\phi\}_{k\geq 1}, \\ 0, & \text{if } t \in \{T_{j,\ell}\}_{\ell\geq 1}, \text{and } t \in \{T_{j,k}+\phi\}_{k\geq 1}. \end{cases} \tag{4.3}$$

Following (4.3), we can rewrite the set of discontinuity points in (4.2) as

$$\Big\{t \geq 0 : t \in \{T_{j,\ell}\}_{\ell\geq 1}, \text{and } t \notin \{T_{j,k}+\phi\}_{k\geq 1}\Big\} \bigcup \Big\{t \geq 0 : t \notin \{T_{j,\ell}\}_{\ell\geq 1}, \text{and } t \in \{T_{j,k}+\phi\}_{k\geq 1}\Big\},$$

which belongs to the set

$$\{T_{j,\ell}\}_{\ell\geq 1} \cup \{T_{j,k}+\phi\}_{k\geq 1}.$$

Condition A1 in Appendix A implies $\mathrm{P}\big(\cup_{\ell\geq 1}\cup_{k\geq 1}\{T_{j,\ell} = T_{j,k}+\phi\}\big) = 0$ for all $j \in \mathcal{V}$.

Accordingly, $\{\lambda_i(t)\}_{i\in\mathcal{V}}$ are piecewise-constant functions of t, with discontinuity points listed in the sequence,

$$\{\breve{T}_1, \breve{T}_2, \ldots\} = \bigcup_{j\in\mathcal{V}} \Big\{\{T_{j,\ell}\}_{\ell\geq 1} \cup \{T_{j,k}+\phi\}_{k\geq 1}\Big\}, \quad \text{with } 0 < \breve{T}_1 < \breve{T}_2 < \cdots, \tag{4.4}$$

where $\{\breve{T}_\ell\}_{\ell\geq 1}$ are arranged in increasing order. An illustration of $N_j(t), N_j(((t-\phi)\vee 0, t])$ and $\lambda_i(t)$ is given in Figure 2.

4.1.2. Marked point process $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ for studying discontinuity points of $\{\lambda_i(t)\}_{i\in\mathcal{V}}$

To investigate $\{\lambda_i(t)\}_{i\in\mathcal{V}}$, we next introduce the notion of "marked point process $(\breve{\boldsymbol{T}}, \boldsymbol{I}) = (\{\breve{T}_\ell\}_{\ell\geq 1}, \{I_\ell\}_{\ell\geq 1})$ for intensity discontinuities" in Definition 2 below. A general marked point process $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ is a double sequence, where $\{\breve{T}_\ell\}_{\ell\geq 1}$ is a point process, and each $\breve{T}_\ell$ is

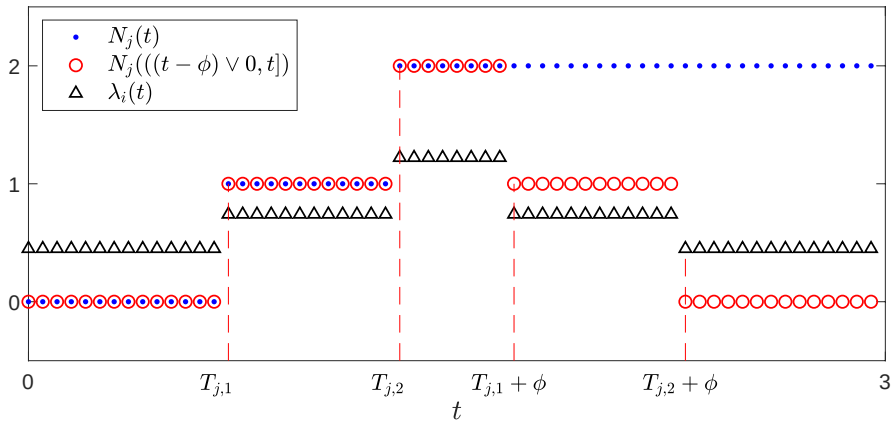

Figure 2: Illustrative plot of sample paths of stochastic processes $N_j(t)$ in (2.2), $N_j(((t-\phi)\vee 0, t])$ in (4.1), and $\lambda_i(t) = \exp\{-0.8 + 0.5 \cdot N_j(((t-\phi)\vee 0, t])\}$ in (3.1), with $\mathcal{V} = \{1, 2\}$, $i = 1$, $j = 2$, and the time-lag $\phi = 1$. Note that $N_j(t)$ and $N_j(((t-\phi)\vee 0, t])$ are overlapped in the time interval $t \in [0, 1.7)$. It is seen that $\lambda_i(t)$ is a piecewise-constant function with points of discontinuities identical to those of $N_j(((t-\phi)\vee 0, t])$.

associated with a "mark" I_ℓ, usually representing some additional features (such as labels or locations) related to the time point $\breve{T}_\ell$; see Daley and Vere-Jones (2003) and the references therein. For each $i \in \mathcal{V}$, define $\mathcal{S}_i = \{j \in \mathcal{V} \setminus i : \beta_{j,i} \neq 0\}$.

Definition 2 (Marked point process $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ for discontinuity points of $\{\lambda_i(t)\}_{i\in\mathcal{V}}$)**.** *Assume conditions* A1–A4 *in Appendix A; assume $\mathcal{S}_i \neq \varnothing$ for each $i \in \mathcal{V}$. For the strictly increasing time points $\{\breve{T}_1, \breve{T}_2, \ldots\}$ defined in* (4.4) *and integers $\ell \geq 1$, let $I_\ell \in \mathcal{V} \cup \{0\}$ be the mark corresponding to $\breve{T}_\ell$, defined by*

$$I_\ell = \begin{cases} i, & \text{if } \breve{T}_\ell \in \{T_{i,k}\}_{k\geq 1} \text{ for some node } i \in \mathcal{V}, \\ 0, & \text{if } \breve{T}_\ell \in \{T_{i,k} + \phi\}_{k\geq 1} \text{ for some node } i \in \mathcal{V}. \end{cases} \tag{4.5}$$

We call the double sequence $(\breve{\boldsymbol{T}}, \boldsymbol{I}) = (\{\breve{T}_\ell\}_{\ell\geq 1}, \{I_\ell\}_{\ell\geq 1})$ the "marked point process for intensity discontinuities".

The mark I_ℓ in (4.5) indicates the identity of $\breve{T}_\ell$: if the discontinuity point $\breve{T}_\ell$ of $\{\lambda_i(t)\}_{i\in\mathcal{V}}$ is due to an event occurrence from some node i at that time point, then I_ℓ represents the index i of that node; otherwise, we set $I_\ell = 0$. Lemma 2 below guarantees the uniqueness of

the mark I_ℓ defined in (4.5) for each $\breve{T}_\ell$.

Lemma 2 (Uniqueness of the mark I_ℓ corresponding to $\breve{T}_\ell$)**.** *Assume conditions* A1–A4 *in Appendix A; assume $\mathcal{S}_i \neq \varnothing$ for each $i \in \mathcal{V}$. Then, the mark I_ℓ in* (4.5)*, corresponding to the discontinuity point $\breve{T}_\ell$, is uniquely defined* a.s.*, i.e.,*

(i) *for any distinct $i, j \in \mathcal{V}$,* $\mathrm{P}(I_\ell = i,\, I_\ell = j) = \mathrm{P}\big(\breve{T}_\ell \in \{T_{i,k}\}_{k\geq 1}\,,\, \breve{T}_\ell \in \{T_{j,k}\}_{k\geq 1}\big) = 0$;

(ii) *for any $i \in \mathcal{V}$,* $\mathrm{P}(I_\ell = i,\, I_\ell = 0) = \mathrm{P}\big(\breve{T}_\ell \in \{T_{i,k}\}_{k\geq 1}\,,\, \breve{T}_\ell \in \{T_{j,k} + \phi\}_{j\in\mathcal{V},\, k\geq 1}\big) = 0$.

As stated by Definition 2 and Lemma 2, a multivariate point process $\{T_{i,\ell}\}_{\ell\geq 1,\, i\in\mathcal{V}}$ uniquely defines a marked point process $(\breve{\boldsymbol{T}}, \boldsymbol{I})$. Conversely, $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ uniquely yields a multivariate point process $\{T_{i,\ell}\}_{\ell\geq 1,\, i\in\mathcal{V}}$ and accordingly a multivariate counting process $\{\boldsymbol{N}(t)\}_{t\geq 0}$, due to the identities,

$$T_{i,\ell} = \inf\{\breve{T}_k > T_{i,\ell-1} : I_k = i,\, k \geq 1\}, \qquad \ell \geq 1, \quad i \in \mathcal{V}, \tag{4.6}$$

$$N_i(t) = \sum_{k\geq 1} \mathrm{I}(\breve{T}_k \leq t,\, I_k = i), \qquad t \geq 0, \quad i \in \mathcal{V}, \tag{4.7}$$

where $T_{i,0} = 0$. Hence, the point process $\{T_{i,\ell}\}_{\ell\geq 1,\, i\in\mathcal{V}}$, the counting process $\{\boldsymbol{N}(t)\}_{t\geq 0}$, and the marked point process $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ can be deduced from each other.

Remark 2. *As to be shown in Theorem* 1 *below, the probability distribution of the marked point process $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ has a closed-form expression, and thus $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ is more convenient to analyze than $\{T_{i,\ell}\}_{\ell\geq 1,\, i\in\mathcal{V}}$ and $\{\boldsymbol{N}(t)\}_{t\geq 0}$.*

4.1.3. Probabilistic properties of $(\breve{\boldsymbol{T}}, \boldsymbol{I})$

For each integer $\ell \geq 1$, let $\mathcal{F}_{\breve{T}_\ell} = \big\{A \in \mathscr{F} : A \cap \{\breve{T}_\ell \leq t\} \in \mathscr{F}_t \text{ for every } t > 0\big\}$ be the stopping time σ-algebra (defined as in Fischer (2013)) with respect to $\breve{T}_\ell$, i.e., generated by the marked point process $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ up to time $\breve{T}_\ell$. For $\ell = 0$, define $\mathcal{F}_{\breve{T}_0} = \mathscr{F}_0 = \{\Omega, \varnothing\}$, $\breve{T}_0 = 0$ and $I_0 = 0$. Theorem 1 presents the probability distribution of the marked point process $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ conditional on the filtration $\{\mathcal{F}_{\breve{T}_\ell}\}_{\ell\geq 0}$ (i.e., $\mathcal{F}_{\breve{T}_0} \subseteq \mathcal{F}_{\breve{T}_1} \subseteq \cdots$). For a σ-field $\mathscr{F}$ and

a random variable X, denote by $\sigma(\mathscr{F}, X)$ the smallest σ-field that contains all the events belonging to $\mathscr{F} \cup \sigma(X)$.

Theorem 1 (Conditional distributions of $\breve{T}_{\ell+1}$ and $I_{\ell+1}$ given $\mathcal{F}_{\breve{T}_\ell}$)**.** *Assume conditions* A1–A4 *and* A2′ *in Appendix A; assume $\mathcal{S}_i \neq \varnothing$ for each $i \in \mathcal{V}$. For each integer $\ell \geq 0$, define by*

$$\mathcal{T}_\ell = \bigcup_{i \in \mathcal{V}} \big\{ t \in (\breve{T}_\ell - \phi, \breve{T}_\ell] : N_i(\{t\}) = 1 \big\} \tag{4.8}$$

the set of event-occurrence time points in the interval $(\breve{T}_\ell - \phi, \breve{T}_\ell]$. Define the $\mathcal{F}_{\breve{T}_\ell}$-measurable random variable

$$T_\ell^* = \begin{cases} \min(\mathcal{T}_\ell) + \phi, & \text{if } \mathcal{T}_\ell \neq \varnothing, \\ \infty, & \text{if } \mathcal{T}_\ell = \varnothing. \end{cases} \tag{4.9}$$

We have the following results:

(i) (*Support of $\breve{T}_{\ell+1}$*) $\mathrm{P}(\breve{T}_\ell < \breve{T}_{\ell+1} \leq T_\ell^*) = 1$.

(ii) (*Conditional distribution of $\breve{T}_{\ell+1}$*) *If $T_\ell^* < \infty$, then $\breve{T}_{\ell+1}$, conditional on $\mathcal{F}_{\breve{T}_\ell}$, has the mixed-type probability distribution, with probability mass function* (p.m.f.) *at the point T_ℓ^*,*

$$\mathrm{P}\big(\breve{T}_{\ell+1} = T_\ell^* \mid \mathcal{F}_{\breve{T}_\ell}\big) = \exp\{-\lambda^{\mathrm{sum}}(\breve{T}_\ell) \cdot (T_\ell^* - \breve{T}_\ell)\}, \tag{4.10}$$

and probability density function (p.d.f.)*,*

$$f_{\breve{T}_{\ell+1} | \mathcal{F}_{\breve{T}_\ell}}(t \mid \breve{T}_\ell) = \lambda^{\mathrm{sum}}(\breve{T}_\ell) \exp\{-\lambda^{\mathrm{sum}}(\breve{T}_\ell) \cdot (t - \breve{T}_\ell)\}, \quad \textit{for } t \in (\breve{T}_\ell, T_\ell^*), \tag{4.11}$$

where $\lambda^{\mathrm{sum}}(t) = \sum_{i=1}^V \lambda_i(t)$ denotes the total intensity process defined in (2.8)–(2.9) *and derived in* (2.11)*. If $T_\ell^* = \infty$, then* (4.10) *and* (4.11) *reduce to $(\breve{T}_{\ell+1} - \breve{T}_\ell) \mid \mathcal{F}_{\breve{T}_\ell} \sim \mathrm{Exp}\big(\lambda^{\mathrm{sum}}(\breve{T}_\ell)\big)$.*

(iii) (*Conditional distribution of $I_{\ell+1}$*) *If $T_\ell^* < \infty$, then for $i \in \mathcal{V}$,*

$$\mathrm{P}\big(I_{\ell+1} = i \mid \sigma(\mathcal{F}_{\breve{T}_\ell}, \breve{T}_{\ell+1})\big) = \begin{cases} 0, & \text{if } \breve{T}_{\ell+1} = T_\ell^*, \\ \lambda_i(\breve{T}_\ell)/\lambda^{\mathrm{sum}}(\breve{T}_\ell), & \text{if } \breve{T}_{\ell+1} \in (\breve{T}_\ell, T_\ell^*). \end{cases} \tag{4.12}$$

If $T_\ell^* = \infty$, then (4.12) *reduces to $\mathrm{P}(I_{\ell+1} = i \mid \sigma(\mathcal{F}_{\breve{T}_\ell}, \breve{T}_{\ell+1})) = \lambda_i(\breve{T}_\ell)/\lambda^{\mathrm{sum}}(\breve{T}_\ell)$, for $i \in \mathcal{V}$ and $\breve{T}_{\ell+1} \in (\breve{T}_\ell, \infty)$.*

The derivation of Theorem 1 primarily utilizes the fact that the intensity functions $\{\lambda_i(t)\}_{i\in\mathcal{V}}$ are constant in each interval $[\breve{T}_\ell, \breve{T}_{\ell+1})$. For example, if $T_\ell^* < \infty$, (4.11) indicates that conditional on $\mathcal{F}_{\breve{T}_\ell}$, the duration $\breve{T}_{\ell+1} - \breve{T}_\ell$ follows an exponential distribution with rate $\lambda^{\mathrm{sum}}(\breve{T}_\ell)$ before $\breve{T}_{\ell+1}$ reaches T_ℓ^*. Furthermore, $\breve{T}_{\ell+1} = T_\ell^*$ indicates that $\breve{T}_{\ell+1} \in \{T_{i,k} + \phi\}_{i\in\mathcal{V},k\geq 1}$, whereas $\breve{T}_{\ell+1} < T_\ell^*$ implies that the probability of the event $\{\breve{T}_{\ell+1} \in \{T_{i,k}\}_{k\geq 1}\}$, conditional on $\sigma(\mathcal{F}_{\breve{T}_\ell}, \breve{T}_{\ell+1})$, is proportional to the corresponding intensity $\lambda_i(\breve{T}_\ell)$ at node i.

In what follows, we show that Theorem 1 leads to further probabilistic results of $(\breve{\boldsymbol{T}}, \boldsymbol{I})$, as presented in Lemmas 3, 4 and 5, which are used for proving Theorem 2. For the sake of description, the following notations are required.

"Duration" τ_ℓ between two consecutive discontinuity time points $\breve{T}_\ell$:

$$\tau_\ell = \breve{T}_\ell - \breve{T}_{\ell-1}, \quad \ell \geq 1. \tag{4.13}$$

"Event counts" $M_{i,\ell}$ at node $i \in \mathcal{V}$:

$$M_{i,0} = 0, \quad M_{i,\ell} = \sum_{k=1}^{\ell} \mathrm{I}(I_k = i), \quad \ell \geq 1. \tag{4.14}$$

"Piecewise-constant intensity" at node $i \in \mathcal{V}$ within the time interval $[\breve{T}_\ell, \breve{T}_{\ell+1})$:

$$\lambda_{i,\ell} = \lambda_i(\breve{T}_\ell), \quad \ell \geq 0. \tag{4.15}$$

Lemma 3 (Expectation and variance related to $(\breve{\boldsymbol{T}}, \boldsymbol{I})$)**.** *Assume conditions* A1–A4 *and* A2′ *in Appendix A; assume $\mathcal{S}_i \neq \varnothing$ for each $i \in \mathcal{V}$. Then for each integer $k \geq 1$, we have that*

$$\begin{aligned} \mathrm{E}\big\{\mathrm{I}(I_k = i) - \lambda_{i,k-1}\,\tau_k \mid \mathcal{F}_{\breve{T}_{k-1}}\big\} &= 0, \\ \mathrm{var}\big\{\mathrm{I}(I_k = i) - \lambda_{i,k-1}\,\tau_k \mid \mathcal{F}_{\breve{T}_{k-1}}\big\} &= \mathrm{E}\big\{\mathrm{I}(I_k = i) \mid \mathcal{F}_{\breve{T}_{k-1}}\big\}, \end{aligned} \tag{4.16}$$

where $\lambda_{i,0} = \lambda_i(0)$. Furthermore, for each integer $\ell \geq 1$,

$$\mathrm{E}\Big(M_{i,\ell} - \sum_{k=1}^{\ell} \lambda_{i,k-1}\,\tau_k\Big) = 0,$$

$$\mathrm{var}\Big(M_{i,\ell} - \sum_{k=1}^{\ell} \lambda_{i,k-1}\,\tau_k\Big) = \mathrm{E}(M_{i,\ell}).$$

Lemma 4 (Martingale property related to $(\breve{\boldsymbol{T}}, \boldsymbol{I})$)**.** *Assume conditions* A1–A4 *and* A2′ *in Appendix A; assume $\mathcal{S}_i \neq \varnothing$ for each $i \in \mathcal{V}$. Then the random process*

$$\Big\{M_{i,\ell} - \sum_{k=1}^{\ell} \lambda_{i,k-1}\,\tau_k\Big\}_{\ell \geq 1}$$

is a martingale with respect to $\{\mathcal{F}_{\breve{T}_\ell}\}_{\ell \geq 1}$.

Lemma 5 (Upper bound for variance related to t-truncated $(\breve{\boldsymbol{T}}, \boldsymbol{I})$)**.** *Assume conditions* A1–A4 *and* A2′ *in Appendix A; assume $\mathcal{S}_i \neq \varnothing$ for each $i \in \mathcal{V}$. For a given deterministic time point $t \in (0, \infty)$, let*

$$L_t = \sum_{\ell=1}^{\infty} \mathrm{I}(\breve{T}_\ell \leq t) \tag{4.17}$$

count the number of discontinuity points $\{\breve{T}_\ell\}_{\ell \geq 1}$ that occur up to t. For integers $\ell \geq 1$, let

$$\tau_\ell^{[t]} = \breve{T}_\ell \wedge t - \breve{T}_{\ell-1} \wedge t \tag{4.18}$$

be the duration between t-truncated $\breve{T}_\ell$ and $\breve{T}_{\ell-1}$, where $a \wedge b = \min(a, b)$. Let $\{X_\ell\}_{\ell \geq 0}$ be a sequence of random variables, such that $X_\ell \geq 0$ is measurable with respect to $\mathcal{F}_{\breve{T}_\ell}$ for each $\ell \geq 0$, and $\sup_{\ell \geq 0} X_\ell \leq c_1$ a.s. for a constant $c_1 \in (0, \infty)$. Then, there exists a constant $c_2 \in (0, \infty)$, such that for each $i \in \mathcal{V}$,

$$\mathrm{E}\Big\{\sum_{k=1}^{L_t} X_{k-1}\,\mathrm{I}(I_k = i) - \sum_{k=1}^{L_t+1} X_{k-1}\,\lambda_{i,k-1}\,\tau_k^{[t]}\Big\} = 0, \tag{4.19}$$

and

$$\mathrm{var}\Big\{\sum_{k=1}^{L_t} X_{k-1}\,\mathrm{I}(I_k = i) - \sum_{k=1}^{L_t+1} X_{k-1}\,\lambda_{i,k-1}\,\tau_k^{[t]}\Big\}$$

$$= \mathrm{E}\Big\{ \sum_{k=1}^{L_t} X_{k-1}^2 \, \mathrm{I}(I_k = i) \Big\} \leq c_2 \, c_1^2 \, t. \tag{4.20}$$

Furthermore, for the special case of $X_k \equiv 1$*, we have*

$$\begin{aligned} \mathrm{E}\Big(M_{i,L_t} - \sum_{k=1}^{L_t+1} \lambda_{i,k-1} \, \tau_k^{[t]} \Big) &= 0, \\ \mathrm{var}\Big(M_{i,L_t} - \sum_{k=1}^{L_t+1} \lambda_{i,k-1} \, \tau_k^{[t]} \Big) &= \mathrm{E}(M_{i,L_t}) \leq c_2 \, t. \end{aligned} \tag{4.21}$$

Remark 3. *Derivations of Lemmas* 3–5 *are outlined as follows. Lemma* 3 *is obtained from direct calculations based on the probability distribution of* $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ *in Theorem* 1*. Lemma* 4 *follows from* (4.16) *in Lemma* 3*. Lemma* 5 *is a non-trivial extension of Lemma* 3*, by replacing a non-random index* ℓ *with a random index* L_t*; Lemma* 5 *aims to study the properties of the marked point process* $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ *when truncated by a fixed time point* $t \in (0, \infty)$*, which is further used for translating these results into the forms of the counting process* $\boldsymbol{N}(t)$*.*

4.1.4. Translating results of $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ into results of $\boldsymbol{N}(t)$

Using Brémaud (1981) (Theorem T8, p. 27), we can justify that for each $i \in \mathcal{V}$, our stochastic process $\{N_i(t) - \int_0^t \lambda_i(u) \, \mathrm{d}u\}_{t \geq 0}$ is a martingale with respect to $\{\mathscr{F}_t\}_{t \geq 0}$. Moreover, the equivalence verified in (4.7), between the marked point process $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ and the counting process $\boldsymbol{N}(t)$, enables us to translate the results of Lemmas 3–5 into the counterparts of $\boldsymbol{N}(t)$, and to directly obtain Theorem 2 below for some useful properties of the martingale $\{N_i(t) - \int_0^t \lambda_i(u) \, \mathrm{d}u\}_{t \geq 0}$.

Theorem 2 (Bounded variance of $N_i(t) - \int_0^t \lambda_i(u) \, \mathrm{d}u$; finiteness of $\boldsymbol{N}(t)$)**.** *Assume conditions* A1–A4 *and* A2′ *in Appendix A. Then, there exists a constant* $c_1 \in (0, \infty)$*, such that for any* $i \in \mathcal{V}$ *and any* $t \in (0, \infty)$*,*

$$\mathrm{var}\Big\{ N_i(t) - \int_0^t \lambda_i(u) \, \mathrm{d}u \Big\} = \mathrm{E}\{N_i(t)\} \leq c_1 \, t, \tag{4.22}$$

and thus the counting process $N_i(t)$ is finite a.s., *i.e.,*

$$\mathrm{P}(N_i(t) < \infty) = 1, \quad i \in \mathcal{V}. \tag{4.23}$$

Furthermore, for a random process $\{x(t)\}_{t\geq 0}$, such that $x(t)$ is $\mathscr{F}_t$-measurable, $0 \leq \inf_{t\geq 0} x(t) \leq \sup_{t\geq 0} x(t) \leq c_2$ a.s. *for a constant $c_2 \in (0,\infty)$, and $x(t)$ is constant in the interval $[\breve{T}_\ell, \breve{T}_{\ell+1})$ for each integer $\ell \geq 0$, it follows that for any $t \in (0,\infty)$,*

$$\mathrm{var}\Big[\int_0^t \big\{x(u-)\,\mathrm{d}N_i(u) - x(u)\,\lambda_i(u)\,\mathrm{d}u\big\}\Big] = \mathrm{E}\Big\{\int_0^t x^2(u)\,\lambda_i(u)\,\mathrm{d}u\Big\} \leq c_1\,c_2^2\,t, \tag{4.24}$$

where $x(u-) = \lim_{t\uparrow u} x(t)$ denotes the left limit.

By means of the marked point process $(\breve{\boldsymbol{T}}, \boldsymbol{I})$, we obtain Theorem 2, which guarantees some basic probabilistic properties of our counting process $\boldsymbol{N}(t)$. In the subsequent discussions in Sections 5 and 6, we shall show that these results are crucial for deriving the related statistical properties, as presented in Theorems 6–9 and Corollaries 1–2.

4.2. Cyclicity and asymptotic mean stationarity

A counting process $\{N(t)\}_{t\geq 0}$ is called "stationary", if for any time point $s \in [0,\infty)$, $N(t+s)-N(s) \overset{\mathrm{D}}{=} N(t)$ for each $t \geq 0$, where $X_1 \overset{\mathrm{D}}{=} X_2$ denotes random quantities X_1 and X_2 having an identical distribution; see Daley and Vere-Jones (2003) and the references therein. A stationary counting process $\{N(t)\}_{t\geq 0}$ possesses many well-known probabilistic properties, some of which are listed as follows:

(P1) Invariant distribution of the conditional intensity function: for the conditional intensity function $\lambda(t)$ defined in (2.6)–(2.7), the probability distribution of $\lambda(t)$ is invariant to any $t \in [0,\infty)$.

(P2) Constant mean intensity: for any $t \in [0,\infty)$, the mean intensity function $\mathrm{E}\{\lambda(t)\} \equiv \lambda_0$ for some constant $\lambda_0 \in (0,\infty)$.

(P3) Expectation of increments: for any $t \in (0,\infty)$ and any $s \in (0,\infty)$, $\mathrm{E}\{N(t+s)-N(s)\} = \lambda_0 \times t$. Furthermore, if $N(t)$ is ergodic, then $\lim_{t\to\infty} N(t)/t = \lambda_0$, a.s..

(P4) Finiteness of $N(t)$: for any $t \in (0, \infty)$, $\mathrm{P}(N(t) < \infty) = 1$.

These features following from the stationarity assumption significantly facilitate the theoretical analysis. Therefore, the stationarity assumption is widely imposed in the relevant literature, e.g., Hansen et al. (2015), Hawkes and Oakes (1974), Reynaud-Bouret and Schbath (2010). A multivariate counting process $\{\boldsymbol{N}(t)\}_{t\geq 0}$ is called stationary if for each dimension $i \in \mathcal{V}$, $\{N_i(t)\}_{t\geq 0}$ is stationary.

Nonetheless, Lemma 6 below shows that $\boldsymbol{N}(t)$ associated with the conditional intensity functions $\{\lambda_i(t)\}_{i\in\mathcal{V}}$ in our model (3.1) is non-stationary.

Lemma 6 (Non-stationarity of $\boldsymbol{N}(t)$)**.** *Assume conditions* A1–A4, A2′, *and* A7 *in Appendix A. Then there exists some node $i_0 \in \mathcal{V}$, such that $N_{i_0}(t)$ is non-stationary. Hence, the multivariate counting process $\boldsymbol{N}(t)$ is non-stationary.*

We justify Lemma 6 by showing that the intensity functions $\{\lambda_i(t)\}_{i\in\mathcal{V}}$ in (3.1) do not have invariant distribution for all $t \in [0, \infty)$, and thus violate property (P1) of stationarity. Without possessing properties (P1)–(P4) listed above, a non-stationary point process greatly challenges the theoretical analysis. It would be necessary to explore alternative properties for non-stationary point processes. We aim to accomplish this goal through a novel approach.

Recall that a Poisson process assumes the independent increment property, leading to the memoryless property Kass et al. (2014), i.e., for any $s \in (0, \infty)$, the time-shifted counting process $\{N(t+s) - N(s)\}_{t\geq 0}$ is independent of the history up to the time point s. Our study considers a relaxed assumption C1 for the memoryless property, i.e., the counting process $\{\boldsymbol{N}(t + R_\ell) - \boldsymbol{N}(R_\ell)\}_{t\geq 0}$, at random time points R_ℓ, is independent of a σ-field $\mathcal{F}_{R_\ell}$, with R_ℓ and $\mathcal{F}_{R_\ell}$ introduced in Definition 3.

Definition 3 (Recurrence time points R_ℓ, recurrence cycle of $\boldsymbol{N}(t)$, and $\mathcal{F}_{R_\ell}$)**.** *Let $R_0 = 0$. For each integer $\ell \geq 1$, let R_ℓ be the first time point, after $R_{\ell-1} + \phi$, such that no events occur at any node in the time interval $(R_\ell - \phi, R_\ell]$, i.e.,*

$$R_\ell = \min\{t \geq R_{\ell-1} + \phi : \boldsymbol{N}(((t - \phi) \vee 0, t]) = 0\}. \tag{4.25}$$

We call R_ℓ the ℓth "recurrence time point", and call the interval $(R_{\ell-1}, R_\ell]$ the ℓth "recurrence cycle". Denote by $\mathcal{F}_{R_\ell} = \{A \in \mathscr{F} : A \cap \{R_\ell \le t\} \in \mathscr{F}_t \text{ for every } t > 0\}$ the stopping time σ-algebra with respect to R_ℓ.

Figure 3 illustrates the recurrence time points R_ℓ. For our $\boldsymbol{N}(t)$, the existence of R_ℓ is verified by Lemma 7.

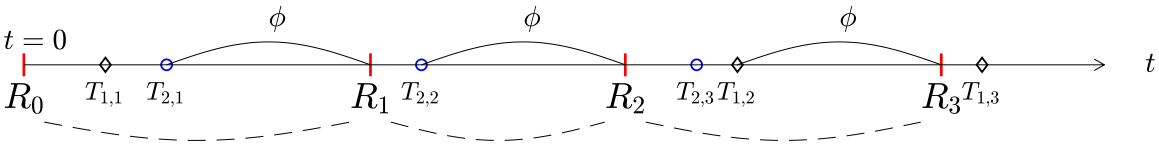

Figure 3: Illustrative plot of recurrence time points $R_0, R_1, R_2, \ldots$, and the event-occurrence time points $\{T_{i,\ell}\}_{\ell \ge 1}$ of nodes $i \in \mathcal{V}$, with the node set $\mathcal{V} = \{1, 2\}$. Condition C1 refers to that after reaching each recurrence time point R_ℓ, $\boldsymbol{N}(t)$ enters into the recurrence cycle $(R_\ell, R_{\ell+1}]$, where $\lambda_i(R_\ell) = \lambda_i(0)$, and restarts a new process $\{\boldsymbol{N}(t + R_\ell) - \boldsymbol{N}(R_\ell)\}_{t \ge 0}$, independent of $\mathcal{F}_{R_\ell}$.

Lemma 7 (Existence of R_ℓ). *Assume conditions* A1–A4 *and* A2′ *in Appendix A. For each integer $\ell \ge 1$, the recurrence time point R_ℓ in Definition* 3 *exists with probability one.*

Furthermore, the recurrence time point R_ℓ and condition C1 imply the "memoryless property" of our intensity process $\lambda_i(t)$ for $t \ge R_\ell$, briefly explained as follows. Recall that $\lambda_i(t)$ in our model (3.1) is a deterministic function of $\boldsymbol{N}(((t-\phi) \vee 0, t])$. Using $\boldsymbol{N}((R_\ell - \phi, R_\ell]) = 0$ for R_ℓ in (4.25) gives that

$$\boldsymbol{N}\big(((t-\phi) \vee 0, t]\big) = \boldsymbol{N}\big(((t-\phi) \vee R_\ell, t]\big), \quad \text{for any } t \ge R_\ell, \tag{4.26}$$

the detail of which is in Remark 4. Hence, $\lambda_i(t)$ with $t \ge R_\ell$ relies only on $\boldsymbol{N}(((t-\phi) \vee R_\ell, t])$ and thus is independent of $\mathcal{F}_{R_\ell}$.

Remark 4. *The derivation of* (4.26) *primarily utilizes the following fact: for two sets $\mathcal{T}_1$ and $\mathcal{T}_2$ of time points, if $\boldsymbol{N}(\mathcal{T}_2) = 0$, then we have $\boldsymbol{N}(\mathcal{T}_1 \setminus \mathcal{T}_2) = \boldsymbol{N}(\mathcal{T}_1)$, where $\mathcal{T}_1 \setminus \mathcal{T}_2 = \{t \ge 0 : t \in \mathcal{T}_1,\ t \notin \mathcal{T}_2\}$ denotes the set difference of $\mathcal{T}_1$ and $\mathcal{T}_2$. For any $t \ge R_\ell$, let $\mathcal{T}_1 = ((t-\phi) \vee 0, t]$ and*

$\mathcal{T}_2 = (R_\ell - \phi, R_\ell]$. *Then* $\mathcal{T}_1 \setminus \mathcal{T}_2 = ((t-\phi) \vee R_\ell, t]$, *and* R_ℓ *in* (4.25) *implies that* $\boldsymbol{N}(\mathcal{T}_2) = 0$, *and thus* $\boldsymbol{N}(\mathcal{T}_1) = \boldsymbol{N}(\mathcal{T}_1 \setminus \mathcal{T}_2) = \boldsymbol{N}(((t-\phi) \vee R_\ell, t])$.

A new "cyclicity property" of $\boldsymbol{N}(t)$ is obtained in Theorem 3.

Theorem 3 (Cyclicity of $\boldsymbol{N}(t)$ driven by R_ℓ)**.** *Assume conditions* A1–A4, A2′, A7, *and* C1 *in Appendix A. Let* $\boldsymbol{N}(t)$ *be the counting process with the intensity processes* $\{\lambda_i(t)\}_{i\in\mathcal{V}}$ *in* (3.1). *Then, for each recurrence time point* R_ℓ *in* (4.25) *with* $\ell \geq 1$,

(i) $\boldsymbol{N}(t + R_\ell) - \boldsymbol{N}(R_\ell) \overset{\mathrm{D}}{=} \boldsymbol{N}(t)$ *for each* $t \geq 0$, *with its intensity processes* $\{\lambda_i(t + R_\ell)\}_{i\in\mathcal{V}, t\geq 0}$ *independent of* $\mathcal{F}_{R_\ell}$;

(ii) $\{\boldsymbol{N}((R_{\ell-1}, R_\ell]) : \ell \geq 1\}$ *is a sequence of* i.i.d. *random vectors;*

(iii) $\{R_\ell - R_{\ell-1} : \ell \geq 1\}$ *is a sequence of* i.i.d. *random variables with the finite second moment.*

This "cyclicity" property of $\boldsymbol{N}(t)$ will be used for deriving Theorem 4 below, as well as Theorem 6 in Section 5.1.

Theorem 4 (Asymptotic mean stationarity of $\boldsymbol{N}(t)$)**.** *Assume conditions* A1–A4, A2′, A7, *and* C1 *in Appendix A. Then, there exists a constant vector* $\boldsymbol{c}_0 \in (0, \infty)^V$, *such that the counting process* $\boldsymbol{N}(t)$ *associated with the intensity processes* $\{\lambda_i(t)\}_{i\in\mathcal{V}}$ *in* (3.1) *satisfies*

$$\frac{\boldsymbol{N}(t)}{t} \overset{\mathrm{P}}{\to} \boldsymbol{c}_0, \quad \text{as } t \to \infty. \tag{4.27}$$

Theorem 4 verifies that the vector $\boldsymbol{N}(t)/t$ of average counts converges in probability to a constant vector as t approaches infinity. For a non-stationary counting process $\boldsymbol{N}(t)$, this type of property is called the "asymptotic mean stationarity"; see Nieuwenhuis (2013).

In summary, we proved in Lemma 6 that our counting process $\boldsymbol{N}(t)$ is non-stationary, violating property (P1) of stationarity. Nevertheless, we verified that $\boldsymbol{N}(t)$ possesses some desirable properties similar to stationary processes. For example, Theorem 4 is similar to the ergodicity in property (P3); Theorem 2 verifies property (P4); and part (i) of Theorem 3 indicates a feature similar to the "shift invariance" property of stationarity. Theorems 3

and 4 not only are crucial for deriving the related statistical properties (in Theorems 6–8) in Chapter 5, but also offer valuable insights into the theories for studying non-stationary point processes. By comparison with existing results, technical tools we developed are easier to interpret and utilize.

4.3. Probabilistic results extended for generalized intensity process

In the previous Sections 4.1 and 4.2, we presented a series of useful probabilistic properties of our proposed model (3.1) for $\lambda_i(t)$. A natural question arises about the reliance between these probabilistic results and the specific form of intensity process $\lambda_i(t)$ in model (3.1). In this chapter, we shall show that all those results could also be extended to point processes with a more general class of intensity processes.

Consider the following generalized model of the intensity process:

$$\lambda_i(t) = h_i\big(\boldsymbol{N}\big(((t-\phi)\vee 0, t]\big)\big), \quad i \in \mathcal{V}, \tag{4.28}$$

where $\phi \in (0,\infty)$ is the time-lag width, and the function $h_i(\cdot) : \mathbb{R}^V \to (0,\infty)$ is continuous, positive, and bounded above. Model (4.28) accounts for a wide class of intensity processes $\{\lambda_i(t)\}_{i\in\mathcal{V}}$. Clearly, model (3.1) is a special case of (4.28), when we set $h_i\big(\boldsymbol{N}(((t-\phi)\vee 0, t])\big) = \exp\{\beta_{0;i} + \sum_{j\in\mathcal{V}} \beta_{j,i}\, g\big(N_j(((t-\phi)\vee 0, t])/\phi\big)\}$. In addition, choosing $h_i(\cdot) \equiv c_i$ for a constant $c_i \in (0,\infty)$ yields $\lambda_i(t) \equiv c_i$, corresponding to a homogeneous Poisson process $N_i(t)$. With various choices of the function $h_i(\cdot)$, (4.28) is able to encode a broad type of dependency structures among nodes, and thus is more general with applications in modeling multivariate point process data. Theorem 5 shows that all the probabilistic properties derived in Sections 4.1 and 4.2 continue to hold for the extended model (4.28) for $\lambda_i(t)$.

Theorem 5 (Probabilistic properties for generalized $\lambda_i(t)$ in (4.28))**.** *Assume conditions* A1–

A3, A2′, B1, *and* C1 *in Appendix A. Then, for a counting process* $\boldsymbol{N}(t)$ *with the generalized intensity process* $\{\lambda_i(t)\}_{i\in\mathcal{V}}$ *modeled by* (4.28), *the results in Theorems* 1–4 *and Lemmas* 2–5 *in Sections* 4.1–4.2 *still hold.*

Theorem 5 demonstrates that our novel technical tools and probabilistic results, such as the "marked point process", "cyclicity" and "asymptotic mean stationarity", could be extended to a wider range of models in (4.28), i.e., not restricted to the specific form of intensity process $\lambda_i(t)$ in (3.1). This fact confirms that our probabilistic contributions provide valuable insights into the theoretical properties of a broad class of multivariate point processes.

Chapter 5

Parameter estimation via penalized M-estimation

Our primary interest is to learn the network structure from the observed data $\{\boldsymbol{T}_i\}_{i\in\mathcal{V}}$ in (1.1) of the multivariate point process in the time interval $[0, \mathrm{T}]$, where $\mathrm{T} \in (0, \infty)$ is the total time length of the experiment. Denote the true values of the conditional intensity function (3.1) as

$$\lambda_i^*(t) = \exp\left\{\widetilde{\boldsymbol{x}}_i(t)^\top \widetilde{\boldsymbol{\beta}}_i^*\right\}, \tag{5.1}$$

where $\widetilde{\boldsymbol{\beta}}_i^* = (\beta_{0;i}^*, \boldsymbol{\beta}_i^{*\top})^\top = (\beta_{0;i}^*, \beta_{1,i}^*, \ldots, \beta_{i-1,i}^*, \beta_{i+1,i}^*, \ldots, \beta_{V,i}^*)^\top \in \mathbb{R}^V$ is the vector of true parameters, and $\widetilde{\boldsymbol{x}}_i(t) = (1, \boldsymbol{x}_i(t)^\top)^\top = (1, x_1(t), \ldots, x_{i-1}(t), x_{i+1}(t), \ldots, x_V(t))^\top \in \mathbb{R}^V$ is the vector of regression covariates. The ultimate goal of our statistical learning is to estimate $\widetilde{\boldsymbol{\beta}}_i^*$ in (5.1) and recover the true network structure $\mathcal{G}^* = \{\mathcal{V}, \mathcal{E}^*\}$, where the true edge set $\mathcal{E}^* = \mathcal{E}_+^* \cup \mathcal{E}_-^*$ corresponds to $\mathcal{E} = \mathcal{E}_+ \cup \mathcal{E}_-$ in (3.2) with parameters $\beta_{j,i}$ replaced by $\beta_{j,i}^*$.

The existing parameter estimation methods include two categories: (i) moment or correlation-based approaches, e.g., Bacry and Muzy (2014), and Krumin et al. (2010); and (ii) intensity-based approaches, e.g., Chornoboy et al. (1988), Hansen et al. (2015), and Xu et al. (2016). The former approach was typically applied to the linear models of $\lambda_i(t)$, and

thus is not suitable for our non-linear model (3.1). We adopt the latter approach, where the parameter estimation is accomplished through the minimization of some loss function, appropriate for measuring the discrepancy between the true and the estimated intensity processes.

5.1. Loss function

In the existing literature, loss functions for a generic counting process $N(t)$ associated with an intensity process $\lambda(t)$ containing parameters $\boldsymbol{\beta}$ include the negative log-likelihood function in Carstensen et al. (2010), Rubin (1972), Xu et al. (2016):

$$\mathcal{L}(\boldsymbol{\beta}) = -\frac{1}{\mathrm{T}} \int_0^{\mathrm{T}} \Big[\log\{\lambda(t-)\} \, \mathrm{d}N(t) - \lambda(t) \, \mathrm{d}t \Big], \tag{5.2}$$

and the squared loss in Hansen et al. (2015), Reynaud-Bouret and Schbath (2010):

$$\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{\mathrm{T}} \int_0^{\mathrm{T}} \Big\{ \lambda^2(t) \, \mathrm{d}t - 2\lambda(t-) \, \mathrm{d}N(t) \Big\}. \tag{5.3}$$

The squared loss (5.3) is more convenient for the linear models of $\lambda(t)$, such as the linear Hawkes process in Reynaud-Bouret and Schbath (2010), whereas the negative log-likelihood function (5.2) is typically suitable for the non-linear case, such as using an exponential link function in our model (3.1). Hence, our discussion will focus on using (5.2). In our multi-dimensional setting, we choose to estimate $\widetilde{\boldsymbol{\beta}}_i^*$ at individual nodes i, and recover the network structure by aggregating estimators of $\{\widetilde{\boldsymbol{\beta}}_i^*\}_{i \in \mathcal{V}}$, with the loss function,

$$\mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i) = -\frac{1}{\mathrm{T}} \int_0^{\mathrm{T}} \Big[\widetilde{\boldsymbol{x}}_i(t-)^\top \widetilde{\boldsymbol{\beta}}_i \, \mathrm{d}N_i(t) - \exp\left\{ \widetilde{\boldsymbol{x}}_i(t)^\top \widetilde{\boldsymbol{\beta}}_i \right\} \mathrm{d}t \Big], \tag{5.4}$$

where $\widetilde{\boldsymbol{\beta}}_i = (\beta_{0;i}, \boldsymbol{\beta}_i^\top)^\top = (\beta_{0;i}, \beta_{1,i}, \ldots, \beta_{i-1,i}, \beta_{i+1,i}, \ldots, \beta_{V,i})^\top \in \mathbb{R}^V$ denotes a vector of generic parameters.

In many fields such as neuroscience, biology and finance, the number of recorded event-occurrence time points could be large, typically in the order of millions or more. This

motivates us to study the behavior of our estimation approach using true values $\widetilde{\boldsymbol{\beta}}_i^*$ for a large number $N_i(\mathrm{T})$ of event time points, or equivalently, long total time length T. Theorem 6 presents the asymptotic convergence under the true model of the gradient vector and the Hessian matrix of $\mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i^*)$ as T approaches infinity. This result will be used for deriving parameter estimation consistency (Theorems 7–8) in Section 5.3.

Theorem 6 (Asymptotic convergence related to loss function $\mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i)$ in (5.4))**.** *Assume conditions* A1–A6, A2′, *and* C1 *in Appendix A. For each $i \in \mathcal{V}$, denote by $\nabla\mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i)$ and $\nabla^2\mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i)$ the gradient vector and the Hessian matrix of $\mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i)$ in* (5.4) *respectively. Then, we have the following results:*

(i) $\nabla\mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i^*)$ *converges to* 0 *in probability at square-root rate, i.e.,*

$$\begin{aligned}\nabla\mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i^*) &= \frac{1}{\mathrm{T}}\int_0^{\mathrm{T}}\Big[\widetilde{\boldsymbol{x}}_i(t)\,\exp\big\{\widetilde{\boldsymbol{x}}_i(t)^\top\widetilde{\boldsymbol{\beta}}_i^*\big\}\,\mathrm{d}t - \widetilde{\boldsymbol{x}}_i(t-)\,\mathrm{d}N_i(t)\Big]\\ &= O_{\mathrm{P}}(\sqrt{1/\mathrm{T}}), \quad \textit{as } \mathrm{T}\to\infty. \end{aligned} \tag{5.5}$$

(ii) *There exists a constant matrix $\mathbf{C}_i$ such that*

$$\begin{aligned}\nabla^2\mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i^*) \;&=\; \frac{1}{\mathrm{T}}\int_0^{\mathrm{T}}\widetilde{\boldsymbol{x}}_i(t)\,\widetilde{\boldsymbol{x}}_i(t)^\top\exp\big\{\widetilde{\boldsymbol{x}}_i(t)^\top\widetilde{\boldsymbol{\beta}}_i^*\big\}\,\mathrm{d}t\\ &\xrightarrow{\mathrm{P}}\mathbf{C}_i, \quad \textit{as } \mathrm{T}\to\infty. \end{aligned} \tag{5.6}$$

Furthermore, $\mathbf{C}_i \succ 0$ with all entries positive.

Theorem 6 is derived from the probabilistic results of $\boldsymbol{N}(t)$ in Chapter 4. Specifically, (5.5) is obtained from the bounded variance property (4.24) of $\boldsymbol{N}(t)$ in Theorem 2, which is derived from the properties of the marked point process $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ (in Theorem 1 and Lemmas 3–5); (5.6) applies the cyclicity property of $\boldsymbol{N}(t)$ in Theorem 3 and the asymptotic mean stationarity of $\boldsymbol{N}(t)$ in Theorem 4.

Remark 5. *Conventional tools for asymptotic results, such as law of large numbers* (L.L.N) *or central limit theorems* (C.L.T)*, are not directly applicable to Theorem* 6*, due to the distinctive features of the stochastic processes $\boldsymbol{N}(t)$ and $\widetilde{\boldsymbol{x}}_i(t)$ in* (5.5) *and* (5.6)*. To be specific, the non-stationary counting*

process $\boldsymbol{N}(t)$ is closely linked with the historical events up to t (via its associated intensity processes $\{\lambda_i^*(t)\}_{i\in\mathcal{V}}$ modeled by (5.1)), yielding a complicated dependence structure of $\boldsymbol{N}(t)$ across time t. Moreover, the stochastic process $\widetilde{\boldsymbol{x}}_i(t) = (1, x_1(t), \ldots, x_{i-1}(t),\ x_{i+1}(t), \ldots, x_V(t))^\top$, with $x_j(t) = g\big(N_j(((t-\phi)\vee 0, t])/\phi\big)$ (see (3.5), (3.6) and (4.1)), relies on the special type of stochastic process $N_j(((t-\phi)\vee 0, t])$, whose probabilistic properties are not available in the existing literature. The use of Theorems 1–4 enables us to prove Theorem 6. This justifies the importance of our probabilistic results in Chapter 4.

5.2. Penalized estimation of parameters

Sparsity assumptions are imposed for the true network structure in many real applications (e.g., Xu et al. (2016); Zhang et al. (2016); Zhao et al. (2012)). To this end, penalization would serve as a useful technique for promoting a sparse network structure with the most significant interactions. Here, we employ the weighted L_1-penalty,

$$\mathcal{P}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i) = \sum_{j\in\mathcal{V}\setminus i} w_{j,i,\mathrm{T}}\,|\beta_{j,i}|, \tag{5.7}$$

where $\{w_{j,i,\mathrm{T}} : j \in \mathcal{V}\setminus i\}$ represent non-negative weights. We estimate the true parameter vector $\widetilde{\boldsymbol{\beta}}_i^*$ by the "penalized M-estimator", which minimizes the sum of the loss function (5.4) and the penalty function (5.7), i.e.,

$$\begin{aligned}\widehat{\widetilde{\boldsymbol{\beta}}}_i &= \arg\min_{\widetilde{\boldsymbol{\beta}}_i\in\mathbb{R}^V} \big\{\mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i) + \mathcal{P}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i)\big\} \\ &= \arg\min_{\widetilde{\boldsymbol{\beta}}_i\in\mathbb{R}^V} \Big\{\tfrac{1}{\mathrm{T}}\int_0^{\mathrm{T}} \Big[\exp\big\{\widetilde{\boldsymbol{x}}_i(t)^\top\widetilde{\boldsymbol{\beta}}_i\big\}\,\mathrm{d}t - \widetilde{\boldsymbol{x}}_i(t-)^\top\widetilde{\boldsymbol{\beta}}_i\,\mathrm{d}N_i(t)\Big] + \sum_{j\in\mathcal{V}\setminus i} w_{j,i,\mathrm{T}}\,|\beta_{j,i}|\Big\},\end{aligned} \tag{5.8}$$

where the vector $\widehat{\widetilde{\boldsymbol{\beta}}}_i = (\widehat{\beta}_{0;i}, \widehat{\beta}_{1,i}, \ldots, \widehat{\beta}_{i-1,i}, \widehat{\beta}_{i+1,i}, \ldots, \widehat{\beta}_{V,i})^\top$ collects $\widehat{\beta}_{0;i}$ and all $\{\widehat{\beta}_{j,i} : j \in \mathcal{V}\setminus i\}$. Accordingly, the estimated network is obtained by

$$\widehat{\mathcal{E}} = \{(j,i) \in \mathcal{V}\times\mathcal{V} : \widehat{\beta}_{j,i} \neq 0,\ j \neq i\}.$$

Furthermore, considering that the sign of an estimator indicates the type of effect, we estimate the sets of excitatory and inhibitory effects by

$$\widehat{\mathcal{E}}_{+} = \{(j,i) \in \mathcal{V} \times \mathcal{V} : \widehat{\beta}_{j,i} > 0,\, j \neq i\}, \tag{5.9}$$

$$\widehat{\mathcal{E}}_{-} = \{(j,i) \in \mathcal{V} \times \mathcal{V} : \widehat{\beta}_{j,i} < 0,\, j \neq i\}, \tag{5.10}$$

respectively.

5.3. Asymptotic results for structure learning

In this section, we focus on the asymptotic properties of the penalized M-estimator $\widehat{\widetilde{\boldsymbol{\beta}}}_i$ in (5.8) with T approaching infinity. To derive the estimation consistency, we provide the following types of conditions for the weights $w_{j,i,\mathrm{T}}$ in (5.7):

$$\max_{j \in \mathcal{S}_i^*} w_{j,i,\mathrm{T}} = O_{\mathrm{P}}(\sqrt{1/\mathrm{T}}); \tag{5.11}$$

$$\max_{j \in \mathcal{S}_i^*} w_{j,i,\mathrm{T}} = o_{\mathrm{P}}(\sqrt{1/\mathrm{T}}); \tag{5.12}$$

$$\min_{j \in \mathcal{V} \setminus \{\mathcal{S}_i^* \cup i\}} \sqrt{\mathrm{T}}\, w_{j,i,\mathrm{T}} \xrightarrow{\mathrm{P}} \infty, \quad \text{as } \mathrm{T} \to \infty, \tag{5.13}$$

with $\mathcal{S}_i^* = \{j \in \mathcal{V} \setminus i : \beta_{j,i}^* \neq 0\}$ collecting nodes having true non-zero effect on node i. An example of weights $\{w_{j,i,\mathrm{T}}\}$ which satisfy (5.11)–(5.13) is that of the adaptive lasso penalty in Zou (2006), in which $w_{j,i,\mathrm{T}} = \eta_{\mathrm{T}} |\breve{\beta}_{j,i}|^{\gamma}$, with $\eta_{\mathrm{T}} = O(1/\mathrm{T}^a)$ for $1/2 < a < 3/2$, $\gamma = -2$, and $\breve{\widetilde{\boldsymbol{\beta}}}_i = (\breve{\beta}_{0;i}, \breve{\beta}_{1,i}, \ldots, \breve{\beta}_{i-1,i}, \breve{\beta}_{i+1,i}, \ldots, \breve{\beta}_{V,i})^{\top}$ denoting the minimizer of $\mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i)$.

Theorem 7 guarantees the existence of a $\sqrt{1/\mathrm{T}}$-consistent estimator $\widehat{\widetilde{\boldsymbol{\beta}}}_i$ in (5.8).

Theorem 7 (Existence of a consistent penalized M-estimator)**.** *Assume conditions* A1–A6, A2′, *and* C1 *in Appendix A. Assume that the weights* $w_{j,i,\mathrm{T}}$ *satisfy* (5.11)*. Then, there exists a local minimizer* $\widehat{\widetilde{\boldsymbol{\beta}}}_i$ *in* (5.8) *such that* $\|\widehat{\widetilde{\boldsymbol{\beta}}}_i - \widetilde{\boldsymbol{\beta}}_i^*\| = O_{\mathrm{P}}(\sqrt{1/\mathrm{T}})$*, as* $\mathrm{T} \to \infty$*.*

Following Theorem 7, the sparsistency of the penalized M-estimator is given in Theorem 8 below. Before stating it, we introduce some notations. We partition the true parameter

vector as $\widetilde{\boldsymbol{\beta}}_i^* = (\beta_{0;i}^*, \boldsymbol{\beta}_i^{*\top})^\top = (\beta_{0;i}^*, \boldsymbol{\beta}_i^{*(\mathrm{I})\top}, \boldsymbol{\beta}_i^{*(\mathrm{II})\top})^\top = (\widetilde{\boldsymbol{\beta}}_i^{*(\mathrm{I})\top}, \boldsymbol{\beta}_i^{*(\mathrm{II})\top})^\top$, where $\boldsymbol{\beta}_i^{*(\mathrm{II})} = 0$, and $\boldsymbol{\beta}_i^{*(\mathrm{I})}$ collects all the non-zero components in $\boldsymbol{\beta}_i^*$. Similarly, for the estimator $\widehat{\widetilde{\boldsymbol{\beta}}}_i$, we adopt the partition $\widehat{\widetilde{\boldsymbol{\beta}}}_i = (\widehat{\beta}_{0;i}, \widehat{\boldsymbol{\beta}}_i^{(\mathrm{I})\top}, \widehat{\boldsymbol{\beta}}_i^{(\mathrm{II})\top})^\top = (\widehat{\widetilde{\boldsymbol{\beta}}}_i^{(\mathrm{I})\top}, \widehat{\boldsymbol{\beta}}_i^{(\mathrm{II})\top})^\top$, with index sets "I" and "II" corresponding to those of $\boldsymbol{\beta}_i^{*(\mathrm{I})}$ and $\boldsymbol{\beta}_i^{*(\mathrm{II})}$ respectively.

Theorem 8 (Sparsistency of penalized M-estimator)**.** *Assume conditions* A1–A6, A2′, *and* C1 *in Appendix A. Assume that the weights $w_{j,i,\mathrm{T}}$ satisfy* (5.12)–(5.13)*. Then, any $\sqrt{1/\mathrm{T}}$-consistent local minimizer $\widehat{\widetilde{\boldsymbol{\beta}}}_i = (\widehat{\widetilde{\boldsymbol{\beta}}}_i^{(\mathrm{I})\top}, \widehat{\boldsymbol{\beta}}_i^{(\mathrm{II})\top})^\top$ in* (5.8) *satisfies*

$$\mathrm{P}\big(\widehat{\boldsymbol{\beta}}_i^{(\mathrm{II})} = 0\big) \to 1, \quad \textit{as } \mathrm{T} \to \infty. \tag{5.14}$$

The sparsistency in (5.14) immediately yields the network recovery consistency in Corollary 1.

Corollary 1 (Network recovery consistency)**.** *Assume the same conditions as in Theorem* 8*. Then, the network structure estimators $\widehat{\mathcal{E}}_+$ in* (5.9) *and $\widehat{\mathcal{E}}_-$ in* (5.10)*, based on $\widehat{\widetilde{\boldsymbol{\beta}}}_i$ in* (5.8)*, are consistent to the true edges $\mathcal{E}_+^*$ and $\mathcal{E}_-^*$, i.e.,*

$$\mathrm{P}\big(\widehat{\mathcal{E}}_+ = \mathcal{E}_+^*, \quad \widehat{\mathcal{E}}_- = \mathcal{E}_-^*\big) \to 1, \quad \textit{as } \mathrm{T} \to \infty.$$

Corollary 1 shows that our method could consistently recover the true network structure as the total time length T grows. This provides strong theoretical support for the utility of our proposed statistical learning procedure.

Remark 6. *The proofs of Theorems* 7 *and* 8 *rely on the asymptotic convergence of $\nabla \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i^*)$ and $\nabla^2 \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i^*)$ in Theorem* 6*, which is derived from the probabilistic results of $\boldsymbol{N}(t)$ in Theorems* 1–4*.*

Chapter 6

Extension of structure learning

The non-linear Hawkes process (3.8) with

$$\lambda_i(t) = \varphi\Big(\beta_{0;i} + \sum_{j\in\mathcal{V}} \int_{-\infty}^{t} \omega_{j,i}(t-u)\,\mathrm{d}N_j(u)\Big),$$

is a well-known continuous-time model for structured multivariate point processes. The distinctive feature of (3.8) is the use of the interaction function $\omega_{j,i}(\cdot)$, which offers more flexibility, but in the meantime, substantially challenges the model estimation. Typically, $\omega_{j,i}(\cdot)$ could not be directly estimated, but can be approximated by some linear combination of basis functions (e.g., Hansen et al. (2015), Reynaud-Bouret and Schbath (2010)). Thus it would be desirable to have some models that can better balance the trade-off between model flexibility and estimation efficiency.

Here, we achieve this task by proposing the following extended model,

$$\lambda_i(t) = \exp\Big\{\beta_{0;i} + \sum_{j\in\mathcal{V}} \sum_{k=1}^{K} \beta_{j,i,k}\, x_{j,k}(t)\Big\}, \quad i \in \mathcal{V}, \quad t \geq 0, \tag{6.1}$$

where

$$x_{j,k}(t) = g\big(r_{j,\phi}(t-(k-1)\,\phi)\big)$$

$$= \begin{cases} 0, & \text{if } 0 \le t \le (k-1)\,\phi, \\ g\big(N_j\big(((t-k\,\phi)\vee 0,\, t-(k-1)\phi]\big)/\phi\big), & \text{if } t > (k-1)\phi, \end{cases} \tag{6.2}$$

is the kth covariate of node j, with the shape function $g(\cdot)$ identical to that in (3.5). The intervals $\{(t-k\,\phi, t-(k-1)\phi] : k = 1, \ldots, K\}$ form an equally-spaced partition of the "lag-window" $(t-K\phi, t]$. The connection strength parameter $\beta_{j,i,k}$ corresponds to the covariate $x_{j,k}(t)$, and represents the strength of the effect imposed on node i from the historical events of node j in the kth sub interval $(t-k\,\phi, t-(k-1)\phi]$ of the lag-window.

Clearly, model (6.1) includes model (3.1) as a special case, with $K = 1$, and the self-effect parameters $\{\beta_{i,i,k}\}_{i\in\mathcal{V};k=1,\ldots,K}$ excluded from analysis. On the other hand, with $g(x) = x$ in (6.2), model (6.1) reduces to a special case of the nonlinear Hawkes process (3.8) with the link function $\varphi(\cdot) = \exp(\cdot)$, and the interaction function

$$\omega_{j,i}(t) = \sum_{k=1}^{K} \frac{\beta_{j,i,k}}{\phi}\, \mathrm{I}((k-1)\phi \le t < k\,\phi). \tag{6.3}$$

With sufficiently large K and small ϕ, those piecewise-constant $\omega_{j,i}(t)$ in (6.3) could well approximate an arbitrary continuous interaction function $\omega_{j,i}(t)$ in (3.8). Thus, models (6.1) and (3.8) share similar flexibility in modeling the mutual effects between nodes.

We now show that parameters in model (6.1) could be estimated via the approach similar to that in Chapter 5. Denote the true values of the conditional intensity function (6.1) as

$$\lambda_i^*(t) = \exp\big\{\widetilde{\boldsymbol{x}}_i(t)^\top \widetilde{\boldsymbol{\beta}}_i^*\big\}, \tag{6.4}$$

where $\widetilde{\boldsymbol{\beta}}_i^* = \big(\beta_{0;i}^*, (\beta_{1,i,1}^*, \ldots, \beta_{1,i,K}^*), \ldots, (\beta_{V,i,1}^*, \ldots, \beta_{V,i,K}^*)\big)^\top \in \mathbb{R}^{VK+1}$ denotes the vector of true parameters, and $\widetilde{\boldsymbol{x}}_i(t) = \big(1, (x_{1,1}(t), \ldots, x_{1,K}(t)), \ldots, (x_{V,1}(t), \ldots, x_{V,K}(t))\big)^\top \in$

$\mathbb{R}^{VK+1}$ denotes the vector of covariates in (6.2). Denote the true interaction function by

$$\omega^*_{j,i}(t) = \sum_{k=1}^{K} \frac{\beta^*_{j,i,k}}{\phi} \mathrm{I}((k-1)\phi \leq t < k\,\phi). \tag{6.5}$$

We propose the penalized M-estimator $\widehat{\widetilde{\boldsymbol{\beta}}}_i$,

$$\begin{aligned}\widehat{\widetilde{\boldsymbol{\beta}}}_i &= \arg\min_{\widetilde{\boldsymbol{\beta}}_i \in \mathbb{R}^{VK+1}} \left\{ \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i) + \mathcal{P}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i) \right\} \\ &= \arg\min_{\widetilde{\boldsymbol{\beta}}_i \in \mathbb{R}^{VK+1}} \left\{ \frac{1}{\mathrm{T}} \int_0^{\mathrm{T}} \left[\exp\left\{ \widetilde{\boldsymbol{x}}_i(t)^\top \widetilde{\boldsymbol{\beta}}_i \right\} \mathrm{d}t - \widetilde{\boldsymbol{x}}_i(t-)^\top \widetilde{\boldsymbol{\beta}}_i \, \mathrm{d}N_i(t) \right] + \mathcal{P}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i) \right\}, \end{aligned} \tag{6.6}$$

and estimate the interaction function (6.3) by

$$\widehat{\omega}_{j,i}(t) = \sum_{k=1}^{K} \frac{\widehat{\beta}_{j,i,k}}{\phi} \mathrm{I}((k-1)\phi \leq t < k\,\phi). \tag{6.7}$$

Analogous to Theorem 7, we present the estimation consistency for the parameters and interaction function in Theorem 9 below.

Theorem 9 (Consistency of penalized M-estimator in the extended model (6.1))**.** *Assume conditions* A1–A4, A2′, *and* C1 *in Appendix A. Let* $\{\boldsymbol{N}(t)\}_{t\geq 0}$ *be a multivariate counting process with the conditional intensity function* (6.4)*, with* $K \geq 1$*. For each* $i \in \mathcal{V}$*, let* $\mathcal{S}^*_i = \{(j,k) \in \mathcal{V} \times \{1,\ldots,K\} : \beta^*_{j,i,k} \neq 0\}$*. Assume that* $\mathcal{P}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i) = \sum_{j\in\mathcal{V}} \sum_{k=1}^{K} w_{j,i,k,\mathrm{T}} \, |\beta_{j,i,k}|$ *is the weighted* L_1*-penalty satisfying* $\max_{(j,k)\in\mathcal{S}^*_i} w_{j,i,k,\mathrm{T}} = o_{\mathrm{P}}(\sqrt{1/\mathrm{T}})$*, and* $\min_{(j,k)\in\mathcal{V}\times\{1,\ldots,K\}\setminus\mathcal{S}^*_i} \sqrt{\mathrm{T}}\, w_{j,i,k,\mathrm{T}} \xrightarrow{\mathrm{P}} \infty$ *as* $\mathrm{T} \to \infty$*. Then, there exists a local minimizer* $\widehat{\widetilde{\boldsymbol{\beta}}}_i$ *in* (6.6) *such that* $\|\widehat{\widetilde{\boldsymbol{\beta}}}_i - \widetilde{\boldsymbol{\beta}}^*_i\| = O_{\mathrm{P}}(\sqrt{1/\mathrm{T}})$*. Furthermore, the estimator* $\widehat{\omega}_{j,i}(t)$ *in* (6.7) *is* $\sqrt{1/\mathrm{T}}$*-consistent for the interaction function* $\omega^*_{j,i}(t)$ *in* (6.5)*, i.e.,*

$$\sup_{t\geq 0} |\widehat{\omega}_{j,i}(t) - \omega^*_{j,i}(t)| = O_{\mathrm{P}}(\sqrt{1/\mathrm{T}}).$$

Define the true edge set by edges corresponding to true non-zero interaction functions

$\omega^*_{j,i}(t)$ in (6.5), i.e.,

$$\mathcal{E}^* = \Big\{(j,i) \in \mathcal{V} \times \mathcal{V} : \int_0^\infty |\omega^*_{j,i}(t)|\,\mathrm{d}t \neq 0,\ j \neq i\Big\}, \tag{6.8}$$

and define the estimated edge set by

$$\widehat{\mathcal{E}} = \Big\{(j,i) \in \mathcal{V} \times \mathcal{V} : \int_0^\infty |\widehat{\omega}_{j,i}(t)|\,\mathrm{d}t \neq 0,\ j \neq i\Big\}. \tag{6.9}$$

Similar to Theorem 8 and Corollary 1, we obtain the network recovery consistency in Corollary 2 below.

Corollary 2 (Network recovery consistency for the extended model (6.1))**.** *Assume the same conditions as in Theorem 9. Then,*

$$\mathrm{P}\big(\widehat{\mathcal{E}} = \mathcal{E}^*\big) \to 1, \quad \textit{as } \mathrm{T} \to \infty.$$

Remark 7. *The proofs of Theorem 9 and Corollary 2 rely on the same probabilistic framework proposed in Chapter 4, with slight modifications in accordance with the extended model (6.1). Specifically, with Definitions 2 and 3 slightly modified, the probabilistic results in Theorems 1–4 also hold for model (6.1). Following these results and using the same proofs as in Theorems 6–8, we further derive the estimation consistency for model (6.1) in Theorem 9 and Corollary 2.*

Theorem 9 and Corollary 2 indicate that after extending model (3.1) to model (6.1), the similar theoretical properties still hold. Besides the non-linear Hawkes process, we may apply the similar extension strategy and connect our model (3.1) with a number of other useful models of continuous-time point processes.

Chapter 7

Simulation study

In this chapter, we perform numerical experiments to illustrate the practical utility of our continuous-time modeling approach.

7.1. Choice of network

In this simulation study, we consider three simulated networks with various degrees of complexity, which are illustrated in Figure 4. Network 1 is a simple network of 10 nodes,

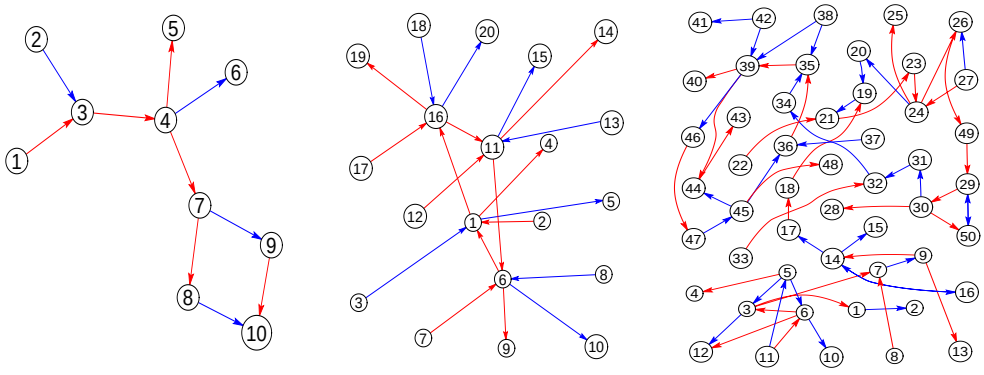

Figure 4: (**Networks** 1**,** 2 **and** 3) The left panel is for Network 1, a simple network with 10 nodes. The middle panel is for Network 2, a medium-complex network with 20 nodes. The right panel is for Network 3, a complex network with 50 nodes. Red arrows indicate excitatory effects and blue arrows indicate inhibitory effects.

with 6 excitatory and 4 inhibitory effects. Motivated from the applications in neuroscience,

Network 1 is a manually designed directed acyclic graph (DAG), aiming to better mimic the information flow from sensory neurons to motor neurons, which is conventionally recognized as an acyclic network structure by neurology scientists. Network 2, cited from Zhao et al. (2012), is a medium-complex network of 20 nodes, with 12 excitatory and 8 inhibitory effects, intending to mimic the potential "hub" and "leaves" structures in neuron ensembles. The four nodes 1, 6, 11 and 16 in Network 2 are hub nodes with degree 6, and others are leaves with degree 1. Network 3 is a complex network of 50 nodes, with 30 excitatory and 30 inhibitory effects, mainly based on the same design motivation as Network 1. The data were generated via simulation algorithms (e.g., Ogata (1981)) using conditional intensity function (3.1), with covariates

$$x_i(t) = g\big(r_{i,\phi}(t)\big),\ i \in \mathcal{V},\ t \in [0, \mathrm{T}], \quad \text{where } g(x) = \log(1 + x \wedge 10). \tag{7.1}$$

7.2. Methods for comparison

We compare methods listed below:

(i) Continuous-time modeling (our proposed method). Method (i) estimates the parameters by the penalized M-estimator in (5.8), with the penalty (5.7) in the following two scenarios.

   a) L_1-penalty in Tibshirani (1996), $\mathcal{P}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i) = \eta \sum_{j \in \mathcal{V} \setminus i} |\beta_{j,i}|$. The tuning parameter η is selected by minimizing the Bayesian Information Criterion (BIC) function in Nishii (1984).

   b) Adaptive lasso penalty in Zou (2006), $\mathcal{P}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i) = \sum_{j \in \mathcal{V} \setminus i} \eta_{\mathrm{T}} |\breve{\beta}_{j,i}|^{\gamma} \cdot |\beta_{j,i}|$, where $\breve{\beta}_{j,i}$ is the M-estimator of $\beta^*_{j,i}$. We set $\gamma = -2$, and use BIC for data-driven selection of η_{T}.

(ii) Discrete-time approximation modeling. Method (ii) takes the discrete-time approximation, which partitions the entire time interval $[0, \mathrm{T}]$ into n equally-spaced time bins $\{(t_{k-1}, t_k] : k = 1, \ldots, n\}$, each of length T/n, followed by transforming the observed point process $\{\boldsymbol{T}_i\}_{i \in \mathcal{V}}$ into sequences of bin counts $\{N_{i,k}\}_{i \in \mathcal{V}; k = 1, \ldots, n}$. Interaction parameters are estimated by the penalized M-estimation similar to (5.8), except that a Poisson distribution with rate $\lambda_i(t_{k-1})\, \mathrm{T}/n$ is assumed for $N_{i,k}$ at node i.

(iii) Discrete-time modeling with groups of connection parameters in Zhao et al. (2012). Method (iii) is similar to method (ii), except that the effect from node j to i is modeled by a group of parameters $\{\beta_{j,i,q} : q = 1, \ldots, Q\}$, instead of a single parameter $\beta_{j,i}$, with $Q = [\phi/(\mathrm{T}/n)]$.

(iv) "SIE-GLM" method in Zhang et al. (2016). Method (iv) is an extension of method (iii), by considering the structural information in the parameter space and employing the sparse group lasso (SGL) penalty for parameter estimation.

(v) Bayesian method in Rajaram et al. (2005). Method (v) is a continuous-time modeling approach using the same loss function as that in our proposed method (i). This method searches all subsets of components in $\widetilde{\boldsymbol{\beta}}_i$, and selects the best subset which has the maximum Bayesian posterior density. To make the comparison fair, method (v) assumes the uniform prior (i.e., no prior information) for $\widetilde{\boldsymbol{\beta}}_i$.

The coordinate descent algorithm in Friedman et al. (2007) is used for solving (5.8). To facilitate further discussion, we categorize all methods in Table 1.

Table 1: Description of the methods in simulation.

abbreviation of method		description
"Discrete_L1"	bin=0.5	method (ii) with L_1-penalty, and bin width $= 0.5$.
	bin=0.25	method (ii) with L_1-penalty, and bin width $= 0.25$.
	bin=0.1	method (ii) with L_1-penalty, and bin width $= 0.1$.
"Continuous_L1"		method (i) with L_1-penalty.
"Discrete_AL"	bin=0.5	method (ii) with adaptive lasso penalty, and bin width $= 0.5$.
	bin=0.25	method (ii) with adaptive lasso penalty, and bin width $= 0.25$.
	bin=0.1	method (ii) with adaptive lasso penalty and bin width $= 0.1$.
"Continuous_AL"		method (i) with adaptive lasso penalty.
"Zhao_2012"	bin=0.5	method (iii) with bin width $= 0.5$.
	bin=0.25	method (iii) with bin width $= 0.25$.
	bin=0.1	method (iii) with bin width $= 0.1$.
"SIE-GLM"	bin=0.5	method (iv) with bin width $= 0.5$.
	bin=0.25	method (iv) with bin width $= 0.25$.
	bin=0.1	method (iv) with bin width $= 0.1$.
"Raj_2005"	parent=3	method (v) with maximum parent number = 3.
	parent=2	method (v) with maximum parent number = 2.

7.3. Simulation result

The total time length $\mathrm{T} \in \{500,\ 1000,\ 2000\}$ is considered. For each $i \in \mathcal{V}$, the true baseline intensity parameter is $\beta^*_{0;i} = -0.8$, with the background intensity $\exp(-0.8) \approx 0.45$. The

true time-lag width is set at $\phi = 1$. For the true connection strength parameters $\{\beta^*_{j,i} : i, j \in \mathcal{V},\ j \neq i\}$, we set $\beta^*_{j,i} = \beta$ if there is excitatory effect from node j to node i; $\beta^*_{j,i} = -\beta$ if there is inhibitory effect from j to i; and $\beta^*_{j,i} = 0$ if there is no effect, where $\beta \in \{0.4,\ 0.5\}$ reflects the magnitude of connection strength.

The performance of each method is assessed by criterion measures "Corret_All" (for correctly detected number of excitatory and inhibitory effects), "Detected_A" (for correctly detected number of excitatory effects), "Detected_B" (for correctly detected number of inhibitory effects), and "Correct_NC" (for correctly detected number of non-effects). As a comparison, "Corret_All", "Detected_A" and "Detected_B" reflect the level of "sensitivity" (defined as the percentage of correctly identified effects, measuring how sensitive is each method for detecting excitatory or inhibitory effect), and "Correct_NC" indicates the level of "specificity" (defined as the percentage of correctly identified non-effects, representing the ability of specification for the non-existence of an effect).

7.3.1. Complex network

In this section, we compare the performance of all methods under Network 3; see Figure 4 (right panel).

We first compare in Table 2 the performance of each method under varying connection strength $\beta \in \{0.4,\ 0.5\}$. Most methods can successfully detect the sparse structure of the network, i.e., most true non-effects are correctly identified and a fairly good level of specificity is achieved. In contrast, the performance in sensitivity is relatively worse than specificity. All methods with $\beta = 0.5$ have better results in sensitivity than with $\beta = 0.4$. This is reasonable, since larger connection strength parameter results in stronger interaction between nodes, which makes the detection easier. In settings of both strength parameters, continuous-time methods ("Continuous_L1", "Continuous_AL") outperform the discrete-time approximation methods ("Discrete_L1", "Discrete_AL", bin = $0.5,\ 0.25,\ 0.1$) in terms of sensitivity. A remarkable finding is that, for methods "Discrete_L1" and "Discrete_AL", the smaller bin width yields the better results, but does not surpass the corresponding

continuous-time methods "Continuous_L1" and "Continuous_AL". This fact suggests that the continuous-time modeling could be recognized as a limiting case of discrete-time modeling when the bin width approaches zero, and thus provides the most accurate results. For penalties in methods (i) and (ii), using adaptive lasso penalty consistently has better results than using L_1-penalty when the same loss function is used. Methods (iii) ("Zhao_2012", bin $= 0.5, 0.25$) and (iv) ("SIE-GLM", bin = $0.5, 0.25$) have relatively reduced sensitivity performance than other methods, with "Correct_All" less than 38 out of 60. For method (v) ("Raj_2005", parent = 2), in order to reduce the computation cost of searching all possible subsets of parents for each node, we only search those subsets with size no greater than 2. Since the true network is sparse with degree no greater than 2 for each node, this setting of maximum number of parents is most favourable for method (v). Even so, method (v) only has good performance in terms of sensitivity, while has significantly worse behaviour than any other methods in terms of specificity. In summary, our proposed continuous-time method with adaptive lasso penalty ("Continuous_AL") has the best overall performance across $\beta \in \{0.4, 0.5\}$.

Table 2: Simulation of Network 3 with varying connection strength β. We set time length $T = 2000$. Results are averaged over 100 replications, with standard errors indicated in parentheses.

		Correct_All		Detected_A		Detected_B		Correct_NC	
strength $\beta =$		0.4	0.5	0.4	0.5	0.4	0.5	0.4	0.5
Discrete_L1	bin=0.5	21.13 (0.41)	42.55 (0.33)	13.68 (0.25)	24.92 (0.17)	7.45 (0.25)	17.63 (0.26)	2385.91 (0.25)	2379.91 (0.33)
	bin=0.25	36.20 (0.45)	53.39 (0.25)	21.28 (0.25)	28.83 (0.10)	14.92 (0.30)	24.56 (0.21)	2382.15 (0.32)	2376.76 (0.42)
	bin=0.1	44.04 (0.41)	56.93 (0.17)	24.76 (0.18)	29.67 (0.05)	19.28 (0.29)	27.26 (0.15)	2379.40 (0.38)	2375.46 (0.38)
Continuous_L1		51.86 (0.33)	58.44 (0.14)	27.33 (0.15)	29.87 (0.03)	24.53 (0.26)	28.57 (0.12)	2368.25 (0.60)	2367.50 (0.43)
Discrete_AL	bin=0.5	40.02 (0.39)	54.05 (0.25)	22.69 (0.22)	28.86 (0.11)	17.32 (0.29)	25.19 (0.20)	2380.52 (0.39)	2379.92 (0.36)
	bin=0.25	50.60 (0.28)	58.55 (0.11)	27.05 (0.17)	29.83 (0.03)	23.55 (0.21)	28.72 (0.09)	2380.34 (0.39)	2381.44 (0.34)
	bin=0.1	54.81 (0.22)	59.44 (0.07)	28.51 (0.12)	29.97 (0.01)	26.30 (0.18)	29.47 (0.06)	2380.06 (0.36)	2382.73 (0.30)
Continuous_AL		56.80 (0.21)	59.72 (0.05)	29.08 (0.11)	30.00 (0.00)	27.72 (0.17)	29.72 (0.05)	2380.73 (0.34)	2383.44 (0.25)
Zhao_2012	bin=0.5	10.08 (0.29)	26.42 (0.37)	6.36 (0.20)	16.46 (0.24)	3.72 (0.16)	9.96 (0.19)	2388.66 (0.13)	2385.15 (0.29)
	bin=0.25	3.62 (0.17)	9.41 (0.26)	3.28 (0.16)	8.08 (0.22)	0.34 (0.06)	1.33 (0.11)	2389.66 (0.05)	2388.86 (0.14)
SIE-GLM	bin=0.5	19.88 (0.40)	39.28 (0.33)	14.00 (0.26)	24.54 (0.18)	5.88 (0.22)	14.74 (0.25)	2387.55 (0.18)	2384.09 (0.28)
	bin=0.25	17.43 (0.35)	37.24 (0.31)	13.29 (0.23)	24.28 (0.20)	4.13 (0.20)	12.96 (0.22)	2388.23 (0.13)	2384.84 (0.25)
Raj_2005	parent=2	57.34 (0.08)	57.91 (0.02)	29.37 (0.07)	29.45 (0.06)	27.97 (0.09)	28.46 (0.06)	2347.34 (0.08)	2347.91 (0.02)
TRUE		60		30		30		2390	

We next compare the results using different values of the total time length T. We set $T \in \{1000, 2000\}$, with the result presented in Table 3. It is clearly seen that $T = 2000$ has better performance than $T = 1000$ for all methods. This is as expected, since larger datasets

provide more information and yield more accurate estimation. This finding also agrees with our theoretical result of network recovery consistency in Corollary 1 of Section 5.3, i.e., the detected network will become closer to the true network with growing time length T. Under each setting of T, the pattern of results is similar to that in Table 2. "Continuous_AL" continues to have the best overall performance.

Table 3: Simulation of Network 3 with varying time length T. The connection strength β is 0.5. Results are averaged over 100 replications, with standard errors indicated in parentheses.

		Correct_All		Detected_A		Detected_B		Correct_NC	
time length T =		1000	2000	1000	2000	1000	2000	1000	2000
Discrete_L1	bin=0.5	16.75 (0.39)	42.55 (0.33)	11.74 (0.26)	24.92 (0.17)	5.01 (0.22)	17.63 (0.26)	2385.86 (0.23)	2379.91 (0.33)
	bin=0.25	28.25 (0.40)	53.39 (0.25)	18.20 (0.24)	28.83 (0.10)	10.05 (0.25)	24.56 (0.21)	2383.25 (0.30)	2376.76 (0.42)
	bin=0.1	35.75 (0.41)	56.93 (0.17)	21.91 (0.24)	29.67 (0.05)	13.84 (0.26)	27.26 (0.15)	2379.78 (0.34)	2375.46 (0.38)
Continuous_L1		48.77 (0.32)	58.44 (0.14)	27.83 (0.12)	29.87 (0.03)	20.94 (0.28)	28.57 (0.12)	2359.83 (0.81)	2367.50 (0.43)
Discrete_AL	bin=0.5	33.18 (0.45)	54.05 (0.25)	19.69 (0.26)	28.86 (0.11)	13.49 (0.30)	25.19 (0.20)	2376.71 (0.42)	2379.92 (0.36)
	bin=0.25	44.57 (0.34)	58.55 (0.11)	25.04 (0.19)	29.83 (0.03)	19.53 (0.26)	28.72 (0.09)	2377.21 (0.42)	2381.44 (0.34)
	bin=0.1	49.88 (0.30)	59.44 (0.07)	27.24 (0.15)	29.97 (0.01)	22.64 (0.25)	29.47 (0.06)	2376.30 (0.41)	2382.73 (0.30)
Continuous_AL		52.61 (0.26)	59.72 (0.05)	28.06 (0.12)	30.00 (0.00)	24.55 (0.21)	29.72 (0.05)	2375.96 (0.42)	2383.44 (0.25)
Zhao_2012	bin=0.5	6.15 (0.25)	26.42 (0.37)	4.33 (0.20)	16.46 (0.24)	1.82 (0.13)	9.96 (0.19)	2388.78 (0.12)	2385.15 (0.29)
	bin=0.25	2.84 (0.17)	9.41 (0.26)	2.60 (0.16)	8.08 (0.22)	0.24 (0.04)	1.33 (0.11)	2389.59 (0.08)	2388.86 (0.14)
SIE-GLM	bin=0.5	15.07 (0.36)	39.28 (0.33)	11.77 (0.27)	24.54 (0.18)	3.30 (0.18)	14.74 (0.25)	2387.59 (0.18)	2384.09 (0.28)
	bin=0.25	12.24 (0.36)	37.24 (0.31)	10.30 (0.29)	24.28 (0.20)	1.94 (0.15)	12.96 (0.22)	2388.46 (0.12)	2384.84 (0.25)
Raj_2005	parent=2	55.44 (0.16)	57.91 (0.02)	28.96 (0.09)	29.45 (0.06)	26.48 (0.13)	28.46 (0.06)	2345.44 (0.16)	2347.91 (0.02)
TRUE		60		30		30		2390	

In practice, the true time-lag width ϕ in (3.6) is unknown, and thus in real applications, we need to first estimate ϕ via either prior knowledge or data-driven methods. In the next numerical experiment, we will investigate how robust our method is against the misspecified time-lag width ϕ.

We fix the true lag width to be $\phi = 1$, and select $\phi_a \in \{0.5, 1, 1.5\}$ as candidates of the "assumed time-lag width" used in the estimation procedure. Note that only $\phi_a = 1$ coincides with the true parameter, whereas $\phi_a \in \{0.5,\ 1.5\}$ are two misspecified time-lag widths. The result is provided in Table 4. As expected, $\phi_a = \phi$ has the best performance. The misspecified $\phi_a \in \{0.5,\ 1.5\}$ does not impact specificity very much, but reduces sensitivity for most of the listed methods. Among all the listed methods, the continuous-time methods ("Continuous_L1", "Continuous_AL") still have best overall performance. In particular, the sensitivity of "Continuous_L1" decreases for less than 15% under both of the two misspecified time-lag widths, which lends support to the robustness of our method against this

misspecification to some extent.

7.3.2. Simple and medium-complex networks

For Networks 1 and 2, we conduct the same simulation evaluation as for Network 3. The results of the two networks for comparing strength, time length and time-lag width are omitted. The pattern of results is similar to that in Network 3. Among all methods, "Continuous_AL" continues to have the best overall performance in each setting. This finding reflects that our conclusions are consistent across varying types of true networks.

In summary, all these simulation results verified the superiority of our proposed continuous-time method over all other irrespective of the complexity level of the true network structure.

Table 4: Simulation of Network 3 with varying assumed time-lag width ϕ_a. The connection strength β is 0.5 and the time length $T = 2000$. True time-lag width $\phi = 1$. Results are averaged over 100 replications, with standard errors indicated in parentheses.

		Correct_All			Detected_A			Detected_B			Correct_NC		
assumed time-lag width =		0.5	1	1.5	0.5	1	1.5	0.5	1	1.5	0.5	1	1.5
Discrete_L1	bin=0.5	37.11 (0.36)	42.55 (0.33)	23.27 (0.45)	22.58 (0.19)	24.92 (0.17)	15.42 (0.27)	14.54 (0.25)	17.63 (0.26)	7.85 (0.27)	2382.25 (0.31)	2379.91 (0.33)	2384.80 (0.25)
	bin=0.25	37.11 (0.42)	53.39 (0.25)	37.15 (0.43)	22.70 (0.25)	28.83 (0.10)	22.39 (0.25)	14.41 (0.25)	24.56 (0.21)	14.76 (0.27)	2381.75 (0.30)	2376.76 (0.42)	2381.04 (0.32)
	bin=0.1	36.77 (0.42)	56.93 (0.17)	44.88 (0.38)	22.26 (0.21)	29.67 (0.05)	25.88 (0.19)	14.51 (0.29)	27.26 (0.15)	19.00 (0.28)	2382.06 (0.29)	2375.46 (0.38)	2379.42 (0.34)
Continuous_L1		43.90 (0.32)	58.44 (0.14)	54.26 (0.26)	25.70 (0.17)	29.87 (0.03)	28.06 (0.14)	18.20 (0.25)	28.57 (0.12)	26.20 (0.19)	2362.46 (0.60)	2367.50 (0.43)	2369.14 (0.60)
Discrete_AL	bin=0.5	51.81 (0.25)	54.05 (0.25)	40.63 (0.36)	27.69 (0.13)	28.86 (0.11)	23.83 (0.21)	24.12 (0.19)	25.19 (0.20)	16.80 (0.27)	2380.06 (0.42)	2379.92 (0.36)	2380.05 (0.36)
	bin=0.25	51.52 (0.29)	58.55 (0.11)	50.86 (0.28)	27.60 (0.15)	29.83 (0.03)	27.85 (0.11)	23.92 (0.24)	28.72 (0.09)	23.01 (0.23)	2379.61 (0.34)	2381.44 (0.34)	2380.11 (0.40)
	bin=0.1	51.44 (0.28)	59.44 (0.07)	54.72 (0.23)	27.56 (0.13)	29.97 (0.01)	29.10 (0.08)	23.88 (0.22)	29.47 (0.06)	25.62 (0.20)	2380.23 (0.34)	2382.73 (0.30)	2380.28 (0.35)
Continuous_AL		51.05 (0.28)	59.72 (0.05)	56.48 (0.17)	27.58 (0.15)	30.00 (0.00)	29.43 (0.07)	23.47 (0.21)	29.72 (0.05)	27.05 (0.16)	2380.23 (0.36)	2383.44 (0.25)	2381.29 (0.34)
Zhao_2012	bin=0.5	34.61 (0.35)	26.42 (0.37)	23.51 (0.37)	20.62 (0.20)	16.46 (0.24)	14.86 (0.25)	14.00 (0.23)	9.96 (0.19)	8.65 (0.18)	2383.59 (0.33)	2385.15 (0.29)	2386.07 (0.21)
	bin=0.25	10.59 (0.31)	9.41 (0.26)	8.53 (0.26)	8.76 (0.25)	8.08 (0.22)	7.41 (0.23)	1.83 (0.13)	1.33 (0.11)	1.12 (0.11)	2388.87 (0.12)	2388.86 (0.14)	2388.94 (0.12)
SIE-GLM	bin=0.5	36.04 (0.32)	39.28 (0.33)	35.00 (0.32)	22.71 (0.18)	24.54 (0.18)	22.95 (0.20)	13.33 (0.22)	14.74 (0.25)	12.05 (0.23)	2385.73 (0.21)	2384.09 (0.28)	2384.92 (0.25)
	bin=0.25	25.84 (0.39)	37.24 (0.31)	29.83 (0.35)	18.60 (0.25)	24.28 (0.20)	21.09 (0.22)	7.24 (0.22)	12.96 (0.22)	8.74 (0.21)	2387.31 (0.17)	2384.84 (0.25)	2386.44 (0.21)
Raj_2005	parent=2	55.34 (0.14)	57.91 (0.02)	57.16 (0.08)	28.99 (0.09)	29.45 (0.06)	29.44 (0.06)	26.35 (0.15)	28.46 (0.06)	27.72 (0.09)	2345.34 (0.14)	2347.91 (0.02)	2347.16 (0.08)
TRUE			60			30			30			2390	

Chapter 8

Real data analysis

In this chapter, we apply our method to real-world multivariate point process data. We analyze the prefrontal cortex spike train dataset "pfc-6" on CRCNS at `https://crcns.org/data-sets/pfc/pfc-6/about-pfc-6`. This dataset consists of neuronal ensemble recordings from the medial prefrontal cortex (mostly prelimbic cortex) of freely moving rats using tetrodes. The data were collected while the animals were learning a behavioral contingency task, as well as during sleep before and after the task. This dataset contains a total of 90 sessions, each corresponding to an experiment. We choose the session folder "181020" in our real data experiment. The spike train data of 55 neurons in an experiment period of 6500 seconds are recorded in the file "181020_SpikeData.dat", containing 1309619 spikes from all 55 neurons.

We apply our continuous-time modeling method with the adaptive lasso penalty to this dataset; see method (i) (b). The same BIC tuning parameter selection procedure is adopted as in the simulation study. The estimated network structure is displayed in Figure 5 (left panel). There are a total of 568 connections identified, among which 338 are excitatory effects and 230 are inhibitory effects. Several findings are remarkable. For example, some pairs of neurons $\{6, 7\}$, $\{24, 34\}$, $\{38, 42\}$, $\{25, 27\}$, $\{21, 23\}$ demonstrate strong mutual excitatory effects, which implies that very close functional connectivity and similarity within these pairs may exist. We see that neuron 13 imposed 44 excitatory effects on other neurons, which

is significantly more than any other neuron, whereas it did not impose any inhibitory effect. This suggests that neuron 13 is potentially a hub neuron which may play an important role in triggering the activities of the entire neuron ensemble.

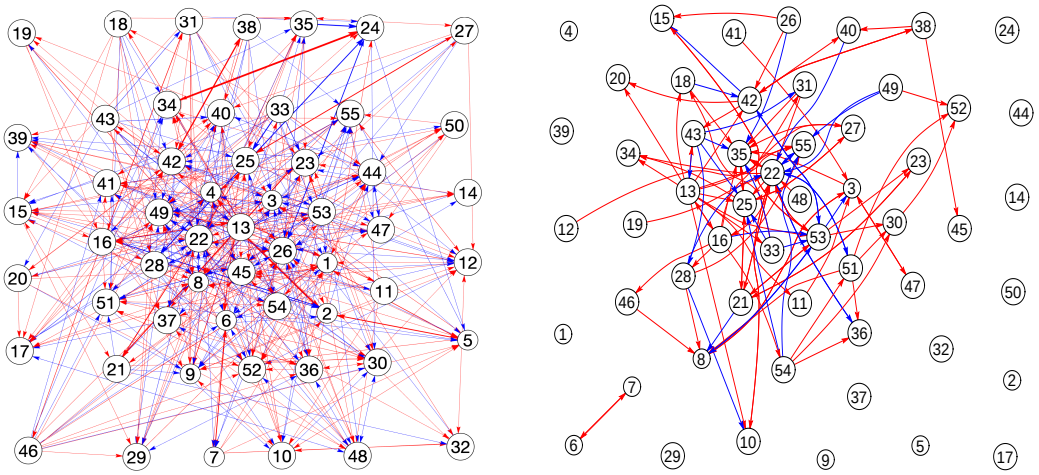

Figure 5: The left panel is the estimated network using continuous-time modeling method with adaptive lasso and BIC criterion. The right panel is the estimated network using continuous-time modeling method with adaptive lasso and GIC criterion. Red arrows indicate excitatory effects and blue arrows indicate inhibitory effects. Thicker arrows represent stronger interactions.

To draw a comparison with BIC, we also adopt the "generalized information criterion" (GIC) in Nishii (1984) for tuning parameter selection and obtain a sparser estimated network in Figure 5 (right panel). We set $a_{\mathrm{T}} = V \log(\mathrm{T})$ in the GIC function. This method may detect more significant effects, while also produces a number of isolated neurons which are disconnected with others. In this sense, GIC could not incorporate all the potential interactions, as compared with BIC in our experiment.

Chapter 9

Discussion

Motivated from inferring neural connectivity from the ensemble neural spike train data, this thesis aims to learn the "network-structured dependence" underlying non-stationary multivariate point process data. A novel continuous-time stochastic modeling of the intensity processes is developed. The related theoretical framework is formulated and probabilistic properties are analyzed to provide new insights into understanding statistical properties of the proposed "penalized M-estimator" for graph parameters relevant to mining the causal relation among nodes. In particular, we develop novel technical tools, such as the marked point process with related conditional distributions, the recurrence time points and cyclicity property, which are useful for analyzing probabilistic properties of a wide array of continuous-time models for point processes.

In discrete-time modeling approaches, a smaller bin width yields less information loss resulting from the discretization procedure. Our continuous-time model could be viewed as a limit case of the corresponding discrete-time model when the bin width approaches zero, thus minimizing the information loss and enhancing the estimation accuracy. This point of view accords with simulation studies in Chapter 7, where our continuous-time modeling (method (i)) consistently outperforms the corresponding discrete-time approximation (method (ii)) in various scenarios.

Our proposed framework is not confined to the learning of interaction effects among

nodes; other factors such as autoregressive effects, experimental units, and other extrinsic conditions, could also be incorporated into model (3.1). Model (6.1) guides a direction for such an extension, and other desirable extensions could be similarly made according to specific goals. A thorough discussion for this line of work is beyond the scope of this thesis, and could be desirable for future research.

Appendix A

Notations, conditions and definitions

A.1. Notations in the proof

For an event A in the sample space Ω, the event $\overline{A}$ denotes the complement of A. For two events A and B, the event $A \setminus B$ denotes $A \cap \overline{B}$. For $s \neq t$, the event $\{\boldsymbol{N}(s) \neq \boldsymbol{N}(t)\}$ denotes $\cup_{i \in \mathcal{V}} \{N_i(s) \neq N_i(t)\}$. For an event A, write $\sigma(\mathscr{F}, A) = \sigma(\mathscr{F}, \mathrm{I}(A))$. A positive definite matrix C is denoted by $C \succ 0$.

A.2. Conditions

The conditions are not the weakest possible, but facilitate the derivations.

A1. The number of nodes $V \in \mathbb{Z}_+$ is fixed. In the multivariate point process, event occurrence time points $\{T_{i,\ell}\}_{i \in \mathcal{V}, \ell \geq 1}$ are not mutually linearly dependent with probability one, satisfying $0 < T_{i,1} < T_{i,2} < \cdots$ for each $i \in \mathcal{V}$.

A2. The multivariate counting process $\boldsymbol{N}(t)$ is "regular".

A2′. There exists a random variable $Z > 0$ with $\mathrm{E}(Z) < \infty$, such that for any $\Delta > 0$ and $t \geq 0$, $\mathrm{P}(\boldsymbol{N}(t + \Delta) \neq \boldsymbol{N}(t) \mid \mathscr{F}_t)/\Delta \leq Z$, a.s..

A3. The multivariate counting process $\boldsymbol{N}(t)$ is "conditionally independent".

A4. In (3.5), the "shape function" $g(\cdot) : [0, \infty) \to [0, \infty)$ is continuous, non-negative, monotonically increasing, and bounded above, with $g(0) = 0$. (From (3.1) and

(3.5), this condition implies that there exist constants $c_1, c_2, c_3 \in (0, \infty)$, such that $\sup_{i\in\mathcal{V},\, t\geq 0} |x_i(t)| \leq c_1$, and $c_2 \leq \inf_{i\in\mathcal{V},\, t\geq 0} \lambda_i(t) \leq \sup_{i\in\mathcal{V},\, t\geq 0} \lambda_i(t) \leq c_3$ hold a.s..)

A5. For all $i \in \mathcal{V}$, the true self-effect parameter $\beta^*_{i,i} = 0$.

A6. The true edge set $\mathcal{E}^* \neq \varnothing$.

A7. The edge set $\mathcal{E}$ in (3.2) satisfies $\mathcal{E} \neq \varnothing$.

B1. For each $i \in \mathcal{V}$, the function $h_i(\cdot) : \mathbb{R}^V \to (0, \infty)$ in (4.28) is continuous and positive, with $\sup_{\boldsymbol{x}\in\mathbb{R}^V} h_i(\boldsymbol{x}) \leq c$ for some constant $c \in (0, \infty)$.

C1. For deterministic $t \geq 0$ and integers $\ell \geq 1$, $\boldsymbol{N}(t + R_\ell) - \boldsymbol{N}(R_\ell)$ is independent of $\mathcal{F}_{R_\ell}$.

A.3. Two versions of definitions of a Poisson process

Definition A (cited from Chapter 19.2 of Kass et al. (2014)). A counting process $\{N(t)\}_{t\geq 0}$ is called a Poisson process with a deterministic intensity function $\lambda(t)$, if the following two properties hold:

Poisson distribution: $N(\mathcal{T})$ follows the Poisson distribution with rate $\int_{s\in\mathcal{T}} \lambda(s)\, \mathrm{d}s$ for any $\mathcal{T} \in \mathcal{B}([0, \infty))$.

Independent increments (or memoryless) property: $N(\mathcal{T}_1), \ldots, N(\mathcal{T}_m)$ are mutually independent for non-overlapping time intervals $\mathcal{T}_1, \ldots, \mathcal{T}_m \in \mathcal{B}([0, \infty))$.

Furthermore, a Poisson process $\{N(t)\}_{t\geq 0}$ is called homogeneous if $\lambda(t)$ is a constant, and is called inhomogeneous otherwise.

Definition B (cited from Rubin (1972)). Let $\{N(t)\}_{t\geq 0}$ be a counting process equipped with the filtered probability space $(\Omega, \mathscr{F}, \{\mathscr{F}_t\}_{t\geq 0}, \mathrm{P})$. For $t \geq 0$, define the conditional intensity function (similar to (2.6)–(2.7)):

$$\begin{aligned} \lambda(t \mid \mathscr{F}_t) &= \lim_{\Delta\downarrow 0} \frac{1}{\Delta} \big[1 - \mathrm{P}(N(t + \Delta) = N(t) \mid \mathscr{F}_t)\big] \\ &= \lim_{\Delta\downarrow 0} \frac{1}{\Delta} \mathrm{P}(N(t + \Delta) = N(t) + 1 \mid \mathscr{F}_t). \end{aligned} \tag{A.1}$$

$\{N(t)\}_{t\geq 0}$ is called a Poisson process if $\lambda(t \mid \mathscr{F}_t) \equiv \lambda(t)$ is a deterministic function.

(Remark: Chapter 19.2 of the book Kass et al. (2014) first defined the Poisson process using Definition A; then Chapter 19.3.4 claimed that this Poisson process also satisfies the property in Definition B. Thus, Definitions A and B refer to the same type of point process, and they are two equivalent definitions of a Poisson process.)

Appendix B

Proofs of main results

B.1. Proof of the statement in Remark 1

We aim to prove the statement: any "multivariate regular point process" also has identical limits (2.8) and (2.9).

From the definition of a multivariate regular point process $\boldsymbol{N}(t)$, we have $\lim_{\Delta\downarrow 0}\frac{1}{\Delta}\mathrm{P}(N_i(t+\Delta)=N_i(t)+1 \mid \mathscr{F}_t)=\lim_{\Delta\downarrow 0}\frac{1}{\Delta}\mathrm{P}(N_i(t+\Delta)\neq N_i(t) \mid \mathscr{F}_t)$, a.s., for any $i\in\mathcal{V}$ and $t\geq 0$. Since $\{N_i(t+\Delta)=N_i(t)+1\}\subseteq\{N_i(t+\Delta)\neq N_i(t)\}$, we further get

$$\lim_{\Delta\downarrow 0}\frac{1}{\Delta}\mathrm{P}\big(\{N_i(t+\Delta)\neq N_i(t)\}\setminus\{N_i(t+\Delta)=N_i(t)+1\}\,\big|\,\mathscr{F}_t\big)=0,\quad \text{a.s.}. \tag{B.1}$$

Thus,

$$\begin{aligned}
0&\leq\lim_{\Delta\downarrow 0}\frac{1}{\Delta}\Big[\mathrm{P}\big(\boldsymbol{N}(t+\Delta)\neq\boldsymbol{N}(t)\mid\mathscr{F}_t\big)-\mathrm{P}\Big(\bigcup_{i\in\mathcal{V}}\{N_i(t+\Delta)=N_i(t)+1\}\,\Big|\,\mathscr{F}_t\Big)\Big]\\
&=\lim_{\Delta\downarrow 0}\frac{1}{\Delta}\mathrm{P}\Big(\Big[\bigcup_{i\in\mathcal{V}}\{N_i(t+\Delta)\neq N_i(t)\}\Big]\setminus\Big[\bigcup_{i\in\mathcal{V}}\{N_i(t+\Delta)=N_i(t)+1\}\Big]\,\Big|\,\mathscr{F}_t\Big)\\
&\leq\lim_{\Delta\downarrow 0}\frac{1}{\Delta}\mathrm{P}\Big(\bigcup_{i\in\mathcal{V}}\Big[\{N_i(t+\Delta)\neq N_i(t)\}\setminus\{N_i(t+\Delta)=N_i(t)+1\}\Big]\,\Big|\,\mathscr{F}_t\Big)\\
&\leq\sum_{i\in\mathcal{V}}\lim_{\Delta\downarrow 0}\frac{1}{\Delta}\mathrm{P}\Big(\{N_i(t+\Delta)\neq N_i(t)\}\setminus\{N_i(t+\Delta)=N_i(t)+1\}\,\Big|\,\mathscr{F}_t\Big)
\end{aligned}$$

$$= 0, \quad \text{a.s.},$$

where the last equality is from (B.1). Hence, we obtain $\lim_{\Delta \downarrow 0} \frac{1}{\Delta} \mathrm{P}\big(\cup_{i \in \mathcal{V}} \{N_i(t+\Delta) = N_i(t) + 1\} \mid \mathscr{F}_t \big) = \lim_{\Delta \downarrow 0} \frac{1}{\Delta} \mathrm{P}(\boldsymbol{N}(t+\Delta) \neq \boldsymbol{N}(t) \mid \mathscr{F}_t)$, a.s.. This completes the proof. ■

B.2. Proof of Lemma 1

Define the events $A_{i,\Delta} = \{N_i(t+\Delta) = N_i(t) + 1\}$. Then

$$\mathrm{P}\Big(\bigcup_{i \in \mathcal{V}} \{N_i(t+\Delta) = N_i(t) + 1\} \,\Big|\, \mathscr{F}_t \Big) = \mathrm{P}\Big(\bigcup_{i \in \mathcal{V}} A_{i,\Delta} \,\Big|\, \mathscr{F}_t \Big). \tag{B.2}$$

By the inclusion-exclusion formula, we have that

$$\begin{aligned} &\mathrm{P}\Big(\bigcup_{i \in \mathcal{V}} A_{i,\Delta} \,\Big|\, \mathscr{F}_t \Big) \\ &= \sum_{k=1}^{V} (-1)^{k+1} \sum_{\{i_1,\ldots,i_k\} \subseteq \mathcal{V}} \mathrm{P}\Big(\bigcap_{j \in \{i_1,\ldots,i_k\}} A_{j,\Delta} \,\Big|\, \mathscr{F}_t \Big) \\ &= \sum_{i=1}^{V} \mathrm{P}(A_{i,\Delta} \mid \mathscr{F}_t) - \sum_{k=2}^{V} (-1)^{k} \sum_{\{i_1,\ldots,i_k\} \subseteq \mathcal{V}} \mathrm{P}\Big(\bigcap_{j \in \{i_1,\ldots,i_k\}} A_{j,\Delta} \,\Big|\, \mathscr{F}_t \Big). \end{aligned} \tag{B.3}$$

For mutually distinct $\{i_1, \ldots, i_k\}$ with $k \geq 2$, the conditional independence condition (2.10) implies that

$$\begin{aligned} \mathrm{P}\Big(\bigcap_{j \in \{i_1,\ldots,i_k\}} A_{j,\Delta} \,\Big|\, \mathscr{F}_t \Big) &\leq \mathrm{P}(A_{i_1,\Delta} \cap A_{i_2,\Delta} \mid \mathscr{F}_t) \\ &= \Delta^2 \{\lambda_{i_1}(t)\, \lambda_{i_2}(t) + o(1)\}, \quad \text{a.s.} \end{aligned} \tag{B.4}$$

as $\Delta \downarrow 0$. Plugging (B.4) into (B.3), we obtain

$$\frac{1}{\Delta} \mathrm{P}\Big(\bigcup_{i \in \mathcal{V}} A_{i,\Delta} \,\Big|\, \mathscr{F}_t \Big)$$

$$= \frac{1}{\Delta} \sum_{i=1}^{V} \mathrm{P}(A_{i,\Delta} \mid \mathscr{F}_t) + O(\Delta) = \sum_{i=1}^{V} \lambda_i(t) + o(1), \quad \text{a.s.}$$

as $\Delta \downarrow 0$. It follows that $\lambda^{\mathrm{sum}}(t) = \lim_{\Delta \downarrow 0} \frac{1}{\Delta} \mathrm{P}(\cup_{i \in \mathcal{V}} A_{i,\Delta} \mid \mathscr{F}_t) = \sum_{i=1}^{V} \lambda_i(t)$. This completes the proof. ■

B.3. Proof of Lemma 2

We first prove part (i). From condition A1, for any distinct nodes $i, j \in \mathcal{V}$, and any integers $k \geq 1$ and $r \geq 1$, we have $\mathrm{P}(T_{i,k} = T_{j,r}) = 0$, which implies

$$\begin{aligned} &\mathrm{P}\big(\breve{T}_\ell \in \{T_{i,k}\}_{k \geq 1}\,,\, \breve{T}_\ell \in \{T_{j,k}\}_{k \geq 1}\big) \\ &\leq \mathrm{P}\big(\cup_{k \geq 1} \cup_{r \geq 1} \{T_{i,k} = T_{j,r}\}\big) \,=\, 0. \end{aligned} \tag{B.5}$$

Next we prove part (ii). Using the similar proof of (B.5), for any $i \in \mathcal{V}$, we have

$$\begin{aligned} &\mathrm{P}\big(\breve{T}_\ell \in \{T_{i,k}\}_{k \geq 1}\,,\, \breve{T}_\ell \in \{T_{j,k} + \phi\}_{k \geq 1}\big) \\ &\leq \mathrm{P}\big(\cup_{k \geq 1} \cup_{r \geq 1} \{T_{i,k} = T_{j,r} + \phi\}\big) \,=\, 0 \end{aligned} \tag{B.6}$$

for any $j \in \mathcal{V}$, and thus

$$\begin{aligned} &\mathrm{P}\big(\breve{T}_\ell \in \{T_{i,k}\}_{k \geq 1}\,,\, \breve{T}_\ell \in \{T_{j,k} + \phi\}_{j \in \mathcal{V},\, k \geq 1}\big) \\ &\leq \sum_{j \in \mathcal{V}} \mathrm{P}\big(\breve{T}_\ell \in \{T_{i,k}\}_{k \geq 1}\,,\, \breve{T}_\ell \in \{T_{j,k} + \phi\}_{k \geq 1}\big) = 0. \end{aligned} \tag{B.7}$$

The proof is completed. ■

B.4. Proof of Theorem 1

Before proving Theorem 1, we first show Lemmas B.1 and B.2 following Definition 4.

Definition 4 ("$\mathrm{E}(X \parallel \mathscr{F}, A)$")**.** *Let $(\Omega, \mathscr{G}, \mathrm{P})$ be a probability space. Let $\mathscr{F} \subseteq \mathscr{G}$ be a sub σ-field, and $A \in \mathscr{G}$ be an event such that $A \notin \mathscr{F}$ and $\mathrm{P}(A \mid \mathscr{F}) > 0$ almost surely. For any random variable X in $(\Omega, \mathscr{G}, \mathrm{P})$, define*

$$\mathrm{E}(X \parallel \mathscr{F}, A) = \frac{\mathrm{E}\{X\,\mathrm{I}(A) \mid \mathscr{F}\}}{\mathrm{P}(A \mid \mathscr{F})}. \tag{B.8}$$

(*Remark: when $\mathscr{F} = \{\Omega, \varnothing\}$, $\mathrm{E}(X \parallel \mathscr{F}, A)$ in* (B.8) *reduces to $\mathrm{E}(X \mid A) = \mathrm{E}\{X\,\mathrm{I}(A)\}/\mathrm{P}(A)$; when X is independent of $\mathscr{F}$ and A, $\mathrm{E}(X \parallel \mathscr{F}, A)$ in* (B.8) *reduces to $\mathrm{E}(X)$.*)

Lemma B.1 (Conditional probability $\mathrm{P}(\boldsymbol{N}(t) = \boldsymbol{N}(s) \mid \mathscr{F}_s)$)**.** *Assume conditions* A1–A4 *and* A2′ *in Appendix A. Then for $t \geq s \geq 0$,*

$$\mathrm{P}\big(\boldsymbol{N}(t) = \boldsymbol{N}(s) \mid \mathscr{F}_s\big) = \exp\Big\{ -\int_s^t \lambda^{\mathrm{sum}}(u; \mathscr{F}_s)\,\mathrm{d}u \Big\}, \tag{B.9}$$

with

$$\lambda^{\mathrm{sum}}(t; \mathscr{F}_s) = \mathrm{E}\{\lambda^{\mathrm{sum}}(t) \parallel \mathscr{F}_s, \boldsymbol{N}(t) = \boldsymbol{N}(s)\}, \tag{B.10}$$

where $\lambda^{\mathrm{sum}}(t)$ is the total intensity in (2.8)–(2.9)*.* (*Remark: $\lambda^{\mathrm{sum}}(t; \mathscr{F}_s)$ in* (B.10) *is motivated by the definition "$\lambda_{N(t)}(t, \mathscr{B}_s)$" in Lemma* 1 *of Rubin* (*1972*)*, and* (B.9) *is motivated by Corollary* 1 *in Rubin* (*1972*).)

Proof: For $t \geq s \geq 0$, and $\Delta > 0$, note that

$$\{\boldsymbol{N}(t + \Delta) = \boldsymbol{N}(s)\} = \{\boldsymbol{N}((s, t + \Delta]) = 0\} \subseteq \{\boldsymbol{N}((s, t]) = 0\} = \{\boldsymbol{N}(t) = \boldsymbol{N}(s)\},$$

which implies that

$$\begin{aligned}
&\mathrm{P}\big(\boldsymbol{N}(t) = \boldsymbol{N}(s) \mid \mathscr{F}_s\big) - \mathrm{P}\big(\boldsymbol{N}(t + \Delta) = \boldsymbol{N}(s) \mid \mathscr{F}_s\big) \\
&= \mathrm{P}\big(\boldsymbol{N}(t) = \boldsymbol{N}(s),\, \boldsymbol{N}(t + \Delta) \neq \boldsymbol{N}(s) \mid \mathscr{F}_s\big) \\
&= \mathrm{P}\big(\boldsymbol{N}(t + \Delta) \neq \boldsymbol{N}(t),\, \boldsymbol{N}(t) = \boldsymbol{N}(s) \mid \mathscr{F}_s\big).
\end{aligned}$$

Combining this with the fact that $\mathscr{F}_s \subseteq \mathscr{F}_t$ (from (2.5)), and (2.9), (B.8), (B.10), we obtain

$$
\begin{aligned}
&\frac{\partial \mathrm{P}\big(\boldsymbol{N}(t) = \boldsymbol{N}(s) \mid \mathscr{F}_s\big)}{\partial t} \\
&= \lim_{\Delta \downarrow 0} \frac{1}{\Delta} \big\{ \mathrm{P}\big(\boldsymbol{N}(t+\Delta) = \boldsymbol{N}(s) \mid \mathscr{F}_s\big) - \mathrm{P}\big(\boldsymbol{N}(t) = \boldsymbol{N}(s) \mid \mathscr{F}_s\big) \big\} \\
&= -\lim_{\Delta \downarrow 0} \frac{1}{\Delta} \mathrm{P}\big(\boldsymbol{N}(t+\Delta) \neq \boldsymbol{N}(t),\, \boldsymbol{N}(t) = \boldsymbol{N}(s) \mid \mathscr{F}_s\big) \\
&= -\lim_{\Delta \downarrow 0} \frac{1}{\Delta} \mathrm{E}\Big[\mathrm{E}\big\{ \mathrm{I}\big(\boldsymbol{N}(t+\Delta) \neq \boldsymbol{N}(t),\, \boldsymbol{N}(t) = \boldsymbol{N}(s)\big) \mid \mathscr{F}_t \big\} \mid \mathscr{F}_s \Big] \\
&= -\lim_{\Delta \downarrow 0} \frac{1}{\Delta} \mathrm{E}\big\{ \mathrm{P}\big(\boldsymbol{N}(t+\Delta) \neq \boldsymbol{N}(t) \mid \mathscr{F}_t\big) \cdot \mathrm{I}\big(\boldsymbol{N}(t) = \boldsymbol{N}(s)\big) \mid \mathscr{F}_s \big\} \\
&= -\mathrm{E}\Big\{ \lim_{\Delta \downarrow 0} \frac{1}{\Delta} \mathrm{P}\big(\boldsymbol{N}(t+\Delta) \neq \boldsymbol{N}(t) \mid \mathscr{F}_t\big) \cdot \mathrm{I}\big(\boldsymbol{N}(t) = \boldsymbol{N}(s)\big) \mid \mathscr{F}_s \Big\} \qquad \text{(B.11)} \\
&= -\mathrm{E}\big\{ \lambda^{\mathrm{sum}}(t) \cdot \mathrm{I}\big(\boldsymbol{N}(t) = \boldsymbol{N}(s)\big) \mid \mathscr{F}_s \big\} \\
&= -\mathrm{E}\big\{ \lambda^{\mathrm{sum}}(t) \parallel \mathscr{F}_s, \boldsymbol{N}(t) = \boldsymbol{N}(s) \big\} \cdot \mathrm{P}\big(\boldsymbol{N}(t) = \boldsymbol{N}(s) \mid \mathscr{F}_s\big) \\
&= -\lambda^{\mathrm{sum}}(t; \mathscr{F}_s) \cdot \mathrm{P}\big(\boldsymbol{N}(t) = \boldsymbol{N}(s) \mid \mathscr{F}_s\big),
\end{aligned}
$$

where the interchange of limit and expectation in (B.11) follows by the dominated convergence theorem and condition A2′. Solving the above differential equation completes the proof of (B.9). ■

Lemma B.2 ($\mathrm{E}\{\lambda_i(S) \parallel \mathscr{F}_s, \boldsymbol{N}(S) = \boldsymbol{N}(s)\}$)**.** *Assume conditions* A1–A4 *in Appendix A. Define* $\mathcal{T}_s = \cup_{j \in \mathcal{V}} \{t \in (s-\phi, s] : N_j(\{t\}) = 1\}$*, and*

$$
T_s^{\mathrm{o}} = \begin{cases} \min(\mathcal{T}_s) + \phi, & \text{if } \mathcal{T}_s \neq \varnothing, \\ \infty, & \text{if } \mathcal{T}_s = \varnothing, \end{cases} \qquad \text{(B.12)}
$$

for a fixed time point $s \in [0, \infty)$*. Then for any random variable* S *satisfying* $s \leq S < T_s^{\mathrm{o}}$*, we have*

$$
\mathrm{E}\big\{ \lambda_i(S) \parallel \mathscr{F}_s, \boldsymbol{N}(S) = \boldsymbol{N}(s) \big\} = \lambda_i\big(s\big), \text{ for all } i \in \mathcal{V}. \qquad \text{(B.13)}
$$

Proof: For $S = s$, (B.13) obviously holds. It suffices to prove (B.13) for $s < S < T_s^{\mathrm{o}}$.

First, we show the following statement:

$$\begin{aligned} &\text{for } s < S < T_s^{\mathrm{o}},\ \boldsymbol{N}(S) = \boldsymbol{N}(s) \text{ implies } \cup_{j\in\mathcal{V}} \big\{t \in (s, S] : N_j(\{t\}) = 1\big\} = \varnothing \\ &\qquad\qquad \text{and } \cup_{j\in\mathcal{V}} \big\{t \in (s, S] : N_j(\{t-\phi\}) = 1\big\} = \varnothing. \end{aligned} \tag{B.14}$$

Let $B_1 = \cup_{j\in\mathcal{V}}\big\{t \in (s, S] : N_j(\{t\}) = 1\big\}$ and $B_2 = \cup_{j\in\mathcal{V}}\big\{t \in (s, S] : N_j(\{t-\phi\}) = 1\big\}$. Note that $\boldsymbol{N}(S) = \boldsymbol{N}(s)$ directly implies $B_1 = \varnothing$. To prove (B.14), it suffices to show $B_2 = \varnothing$, whose proof is given according to whether $\mathcal{T}_s \neq \varnothing$ or not.

If $\mathcal{T}_s \neq \varnothing$, then (B.12) yields that $s - \phi < \min(\mathcal{T}_s)$, implying $\cup_{j\in\mathcal{V}}\big\{t \in (s, \min(\mathcal{T}_s) + \phi) : N_j(\{t-\phi\}) = 1\big\} - \phi = \cup_{j\in\mathcal{V}}\big\{t \in (s-\phi, \min(\mathcal{T}_s)) : N_j(\{t\}) = 1\big\} = \varnothing$. This, together with $s < S < \min(\mathcal{T}_s) + \phi$, gives $B_2 = \varnothing$.

If $\mathcal{T}_s = \varnothing$, we obtain

$$\begin{aligned} B_2 &\subseteq \big\{ \cup_{j\in\mathcal{V}} \{t \in (s, s+\phi] : N_j(\{t-\phi\}) = 1\}\big\} \\ &\quad \bigcup \big\{ \cup_{j\in\mathcal{V}} \{t \in ((s+\phi)\wedge S, S] : N_j(\{t-\phi\}) = 1\}\big\} \\ &= \{\mathcal{T}_s + \phi\} \bigcup \big\{ \cup_{j\in\mathcal{V}} \{t \in ((s+\phi)\wedge S, S] : N_j(\{t-\phi\}) = 1\}\big\} \\ &\subseteq \{\mathcal{T}_s + \phi\} \cup \{B_1 + \phi\} \\ &= \varnothing, \end{aligned}$$

where $\mathcal{T}_s + \phi = \{t + \phi : t \in \mathcal{T}_s\}$.

Combining the above two cases, we verified (B.14).

Next, we prove Lemma B.2. If $\boldsymbol{N}(S) = \boldsymbol{N}(s)$, then (B.14) indicates that the intensity functions $\{\lambda_i(t)\}_{i\in\mathcal{V}}$ have no discontinuity points in the interval $(s, S]$. Since $\{\lambda_i(t)\}_{i\in\mathcal{V}}$ are piecewise-constant functions, it follows that if $\boldsymbol{N}(S) = \boldsymbol{N}(s)$, then $\lambda_i(S) = \lambda_i(s)$, i.e.,

$$\lambda_i(S) \cdot \mathrm{I}(A) = \lambda_i(s) \cdot \mathrm{I}(A), \quad i \in \mathcal{V},$$

where A denotes the event $\{\boldsymbol{N}(S) = \boldsymbol{N}(s)\}$. Combining this with (B.8), we obtain

$$\mathrm{E}\big\{\lambda_i(S) \parallel \mathscr{F}_s,\, A\big\} \;=\; \frac{\mathrm{E}\{\lambda_i(S)\cdot \mathrm{I}(A) \mid \mathscr{F}_s\}}{\mathrm{P}(A \mid \mathscr{F}_s)} \;=\; \frac{\mathrm{E}\{\lambda_i(s)\cdot \mathrm{I}(A) \mid \mathscr{F}_s\}}{\mathrm{P}(A \mid \mathscr{F}_s)} \;=\; \lambda_i(s).$$

This completes the proof. ■

Now we prove Theorem 1. We first show part (i). By the definition in (4.4), $\breve{T}_\ell < \breve{T}_{\ell+1}$ holds for any integer $\ell \geq 1$. It suffices to show $\breve{T}_{\ell+1} \leq T^*_\ell$. If $\mathcal{T}_\ell = \varnothing$, then $T^*_\ell = \infty$ in (4.9) completes the proof. If $\mathcal{T}_\ell \neq \varnothing$, then any $t_\ell \in \mathcal{T}_\ell$ in (4.8) indicates that $t_\ell = T_{i,k}$ for some integers $i \in \mathcal{V}$ and $k \geq 1$, and $\breve{T}_\ell < t_\ell + \phi \equiv T_{i,k} + \phi \leq \breve{T}_\ell + \phi$. Also, $t_\ell + \phi \equiv T_{i,k} + \phi \in \{\breve{T}_1, \breve{T}_2, \ldots\}$. This combined with $\breve{T}_\ell < \breve{T}_{\ell+1}$ implies $\breve{T}_{\ell+1} \leq t_\ell + \phi$. Thus $\breve{T}_{\ell+1} \leq \min\{t_\ell : t_\ell \in \mathcal{T}_\ell\} + \phi = \min(\mathcal{T}_\ell) + \phi = T^*_\ell$.

Before proving parts (ii) and (iii), preparations (a), (b) (c) and (d) are made below.

(a) Since $\mathcal{F}_{\breve{T}_\ell} = \sigma\{(\breve{T}_0, I_0), \ldots, (\breve{T}_\ell, I_\ell)\}$, it suffices to show that (4.10)–(4.12) hold conditional on each realization

$$\Big\{\{(\breve{T}_0, I_0), \ldots, (\breve{T}_\ell, I_\ell)\} = \{(\breve{t}_0, i_0), \ldots, (\breve{t}_\ell, i_\ell)\}\Big\} = \text{“}\bullet\text{”} \in \mathcal{F}_{\breve{T}_\ell}. \tag{B.15}$$

Note that the realization “$\bullet$” in (B.15) is known from the history up to time $\breve{t}_\ell$. Following the notation “$\mathscr{F}_t$” in (2.4), we also have “$\bullet$” $\in \mathscr{F}_{\breve{t}_\ell}$, and thus for a random variable $X : \Omega \to \mathbb{R}$, which is measurable with respect to either $\mathcal{F}_{\breve{T}_\ell}$ or $\mathscr{F}_{\breve{t}_\ell}$, denote by “$X(\bullet)$” the value of X at the realization “$\bullet$”. For example, we can write $\breve{T}_\ell(\bullet) = \breve{t}_\ell$ and $I_\ell(\bullet) = i_\ell$.

(b) Comparing the random variables T^*_ℓ in (4.9) and $T^{\mathrm{o}}_{\breve{t}_\ell}$ in (B.12) (with $s = \breve{t}_\ell$), we observe that they have the same value t^*_ℓ at the realization “$\bullet$”, i.e.,

$$t^*_\ell = T^*_\ell(\bullet) = T^{\mathrm{o}}_{\breve{t}_\ell}(\bullet), \;\text{ and }\quad \breve{t}_\ell < t^*_\ell.$$

(c) Also, we verify the following equation for $t \in (\breve{t}_\ell, t^*_\ell)$:

$$\mathrm{P}\big(\boldsymbol{N}(t) = \boldsymbol{N}(\breve{t}_\ell) \mid \mathscr{F}_{\breve{t}_\ell}\big)(\bullet) = \exp\{-\lambda^{\mathrm{sum}}(\breve{t}_\ell)(\bullet)\cdot(t - \breve{t}_\ell)\}. \tag{B.16}$$

Using (B.9) (with $s = \breve{t}_\ell$) yields that

$$\mathrm{P}\big(\boldsymbol{N}(t) = \boldsymbol{N}(\breve{t}_\ell) \mid \mathscr{F}_{\breve{t}_\ell}\big) = \exp\Big\{ - \int_{\breve{t}_\ell}^{t} \lambda^{\mathrm{sum}}\big(u; \mathscr{F}_{\breve{t}_\ell}\big)\, \mathrm{d}u \Big\}, \text{ for } t > \breve{t}_\ell. \tag{B.17}$$

Since both sides of (B.17) are $\mathscr{F}_{\breve{t}_\ell}$-measurable random variables, it follows that

$$\mathrm{P}\big(\boldsymbol{N}(t) = \boldsymbol{N}(\breve{t}_\ell) \mid \mathscr{F}_{\breve{t}_\ell}\big)(\bullet) = \exp\Big\{ - \int_{\breve{t}_\ell}^{t} \lambda^{\mathrm{sum}}\big(u; \mathscr{F}_{\breve{t}_\ell}\big)(\bullet)\, \mathrm{d}u \Big\}, \text{ for } t > \breve{t}_\ell. \tag{B.18}$$

For $t \in (\breve{t}_\ell, t_\ell^*)$, define a random variable S to satisfy $S(\bullet) = t$ and $\breve{t}_\ell \le S < T^{\mathrm{o}}_{\breve{t}_\ell}$ (e.g., if $T_\ell^* < \infty$, then let $S = \eta\, \breve{t}_\ell + (1-\eta) T^{\mathrm{o}}_{\breve{t}_\ell}$, with $\eta = (t_\ell^* - t)/(t_\ell^* - \breve{t}_\ell) \in (0,1)$; if $T_\ell^* = \infty$, then let $S = \breve{t}_\ell \cdot \mathrm{I}(T^{\mathrm{o}}_{\breve{t}_\ell} < \infty) + t \cdot \mathrm{I}(T^{\mathrm{o}}_{\breve{t}_\ell} = \infty)$, where $S(\bullet) = t$ holds due to $T^{\mathrm{o}}_{\breve{t}_\ell}(\bullet) = t_\ell^*$, and $\breve{t}_\ell \le S < T^{\mathrm{o}}_{\breve{t}_\ell}$ is valid due to $\breve{t}_\ell < T^{\mathrm{o}}_{\breve{t}_\ell}$). Applying Lemma B.2 (with $s = \breve{t}_\ell$), we have

$$\mathrm{E}\big\{ \lambda_i(S) \parallel \mathscr{F}_{\breve{t}_\ell}, \boldsymbol{N}(S) = \boldsymbol{N}(\breve{t}_\ell) \big\} = \lambda_i\big(\breve{t}_\ell\big), \text{ for all } i \in \mathcal{V}. \tag{B.19}$$

Similarly to (B.10), for $i \in \mathcal{V}$ and $t \ge s \ge 0$, define

$$\lambda_i(t; \mathscr{F}_s) = \mathrm{E}\{ \lambda_i(t) \parallel \mathscr{F}_s, \boldsymbol{N}(t) = \boldsymbol{N}(s) \}. \tag{B.20}$$

For $t \in (\breve{t}_\ell, t_\ell^*)$, using the definition (B.20) (with $s = \breve{t}_\ell$), the fact $S(\bullet) = t$ and (B.19), we obtain

$$\begin{aligned} \lambda_i\big(t; \mathscr{F}_{\breve{t}_\ell}\big)(\bullet) &= \mathrm{E}\big\{ \lambda_i(t) \parallel \mathscr{F}_{\breve{t}_\ell}, \boldsymbol{N}(t) = \boldsymbol{N}(\breve{t}_\ell) \big\}(\bullet) \\ &= \mathrm{E}\big\{ \lambda_i(S) \parallel \mathscr{F}_{\breve{t}_\ell}, \boldsymbol{N}(t) = \boldsymbol{N}(\breve{t}_\ell) \big\}(\bullet) \;=\; \lambda_i\big(\breve{t}_\ell\big)(\bullet), \text{ for all } i \in \mathcal{V}. \end{aligned} \tag{B.21}$$

Summing over $i \in \mathcal{V}$ on both sides of (B.21) gives that

$$\lambda^{\mathrm{sum}}\big(t; \mathscr{F}_{\breve{t}_\ell}\big)(\bullet) = \lambda^{\mathrm{sum}}\big(\breve{t}_\ell\big)(\bullet). \tag{B.22}$$

By (B.22), we simplify (B.18) to be

$$\begin{aligned}\mathrm{P}\big(\boldsymbol{N}(t)=\boldsymbol{N}(\breve{t}_\ell)\mid\mathscr{F}_{\breve{t}_\ell}\big)(\bullet)&=\exp\Big\{-\int_{\breve{t}_\ell}^{t}\lambda^{\mathrm{sum}}\big(u\big)(\bullet)\,\mathrm{d}u\Big\}\\&=\exp\{-\lambda^{\mathrm{sum}}(\breve{t}_\ell)(\bullet)\cdot(t-\breve{t}_\ell)\},\end{aligned}$$

which proves (B.16).

(d) We show the following statement:

$$\text{for } t\in(\breve{T}_\ell,T_\ell^*),\ \{\boldsymbol{N}(t)=\boldsymbol{N}(\breve{T}_\ell)\}\text{ is equivalent to }\{\breve{T}_{\ell+1}>t\}. \tag{B.23}$$

The proof for the sufficiency part is similar to that of (B.14). For the necessity part, if $\boldsymbol{N}(t)\neq\boldsymbol{N}(\breve{T}_\ell)$, then there exists $T_{i,k}\in(\breve{T}_\ell,t]$ for some integers $i\in\mathcal{V}$ and $k\geq 1$. Also, $\breve{T}_\ell$ and $\breve{T}_{\ell+1}$ are two consecutive discontinuity points in the set (4.4), which implies that $\cup_{i\in\mathcal{V}}\cup_{k\geq1}\{T_{i,k}:\breve{T}_\ell<T_{i,k}<\breve{T}_{\ell+1}\}=\varnothing$. Combining this with the fact that $T_{i,k}\in(\breve{T}_\ell,t]$, we obtain $\breve{T}_\ell<\breve{T}_{\ell+1}\leq T_{i,k}\leq t$, and thus $\breve{T}_{\ell+1}\leq t$.

We next prove parts (ii) and (iii) of Theorem 1. **Proof of part** (ii). For $T_\ell^*<\infty$, using the similar proof as that of (B.23), we have that $\boldsymbol{N}((\breve{T}_\ell,T_\ell^*))=0$ is equivalent to $\breve{T}_{\ell+1}\geq T_\ell^*$. Also, the result $\breve{T}_{\ell+1}\in(\breve{T}_\ell,T_\ell^*]$ in part (i) implies that $\breve{T}_{\ell+1}\geq T_\ell^*$ is equivalent to $\breve{T}_{\ell+1}=T_\ell^*$. It follows that

$$\mathrm{P}\big(\breve{T}_{\ell+1}=T_\ell^*\,\big|\,\mathcal{F}_{\breve{T}_\ell}\big)=\mathrm{P}\big(\breve{T}_{\ell+1}\geq T_\ell^*\,\big|\,\mathcal{F}_{\breve{T}_\ell}\big)=\mathrm{P}\big(\boldsymbol{N}\big((\breve{T}_\ell,\,T_\ell^*)\big)=0\mid\mathcal{F}_{\breve{T}_\ell}\big). \tag{B.24}$$

Evaluating both sides of (B.24) at the realization"$\bullet$" and using $t_\ell^*=T_\ell^*(\bullet)$, we obtain

$$\begin{aligned}\mathrm{P}\big(\breve{T}_{\ell+1}=T_\ell^*\mid\mathcal{F}_{\breve{T}_\ell}\big)(\bullet)&=\mathrm{P}\big(\boldsymbol{N}\big((\breve{T}_\ell,\,T_\ell^*)\big)=0\mid\mathcal{F}_{\breve{T}_\ell}\big)(\bullet)\\&=\mathrm{P}\big(\boldsymbol{N}\big((\breve{t}_\ell,\,t_\ell^*)\big)=0\mid\mathscr{F}_{\breve{t}_\ell}\big)(\bullet)\\&=\lim_{t\uparrow t_\ell^*}\mathrm{P}\big(\boldsymbol{N}\big((\breve{t}_\ell,\,t]\big)=0\mid\mathscr{F}_{\breve{t}_\ell}\big)(\bullet)\\&=\lim_{t\uparrow t_\ell^*}\mathrm{P}\big(\boldsymbol{N}(t)=\boldsymbol{N}(\breve{t}_\ell)\mid\mathscr{F}_{\breve{t}_\ell}\big)(\bullet)\end{aligned}$$

$$= \lim_{t \uparrow t_\ell^*} \exp\{-\lambda^{\mathrm{sum}}(\breve{t}_\ell)(\bullet) \cdot (t - \breve{t}_\ell)\} \tag{B.25}$$

$$= \exp\{-\lambda^{\mathrm{sum}}(\breve{t}_\ell)(\bullet) \cdot (t_\ell^* - \breve{t}_\ell)\}$$

$$= \exp\big\{ - \lambda^{\mathrm{sum}}(\breve{T}_\ell) \cdot (T_\ell^* - \breve{T}_\ell)\big\}(\bullet),$$

where (B.25) is derived from (B.16) with $t \in (\breve{t}_\ell, t_\ell^*)$. This proves (4.10). By using $\breve{t}_\ell = \breve{T}_\ell(\bullet)$, (B.23) and (B.16), for $t \in (\breve{T}_\ell, T_\ell^*)$, we have

$$\begin{aligned} f_{\breve{T}_{\ell+1}|\mathcal{F}_{\breve{T}_\ell}}(t)(\bullet) &= \frac{\partial \, \mathrm{P}\big(\breve{T}_{\ell+1} \le t \mid \mathcal{F}_{\breve{T}_\ell}\big)(\bullet)}{\partial \, t} \\ &= \frac{-\partial \, \mathrm{P}(\boldsymbol{N}(t) = \boldsymbol{N}(\breve{t}_\ell) \mid \mathscr{F}_{\breve{t}_\ell})(\bullet)}{\partial \, t} \\ &= \frac{-\partial \exp\{-\lambda^{\mathrm{sum}}\big(\breve{t}_\ell\big)(\bullet) \cdot (t - \breve{t}_\ell)\}}{\partial \, t} \\ &= \lambda^{\mathrm{sum}}\big(\breve{t}_\ell\big)(\bullet) \cdot \exp\{-\lambda^{\mathrm{sum}}\big(\breve{t}_\ell\big)(\bullet) \cdot (t - \breve{t}_\ell)\} \\ &= \lambda^{\mathrm{sum}}\big(\breve{T}_\ell\big)(\bullet) \cdot \exp\{-\lambda^{\mathrm{sum}}\big(\breve{T}_\ell\big) \cdot (t - \breve{T}_\ell)\}(\bullet), \end{aligned} \tag{B.26}$$

which verifies (4.11). For $T_\ell^* = \infty$, following the same proof as that of (B.26), we have that $(\breve{T}_{\ell+1} - \breve{T}_\ell) \mid \mathcal{F}_{\breve{T}_\ell} \sim \mathrm{Exp}\big(\lambda^{\mathrm{sum}}(\breve{T}_\ell)\big)$.

Proof of part (iii)**.** For $T_\ell^* < \infty$, if $\breve{T}_{\ell+1} = T_\ell^*$, then (4.8) and (4.9) indicate that $\mathcal{T}_\ell \neq \varnothing$ and $\breve{T}_{\ell+1} - \phi = T_\ell^* - \phi \in \mathcal{T}_\ell$, i.e., $\breve{T}_{\ell+1} = T_{i,k} + \phi$ for some integers $i \in \mathcal{V}$ and $k \ge 1$. This combined with (4.5) gives that $I_{\ell+1} = 0$, and thus $\mathrm{P}(I_{\ell+1} = i \mid \sigma(\mathcal{F}_{\breve{T}_\ell}, \breve{T}_{\ell+1})) = 0$ for $i \in \mathcal{V}$. If $\breve{T}_{\ell+1} \in (\breve{T}_\ell, T_\ell^*)$, then for $i \in \mathcal{V}$ and $t \in (\breve{T}_\ell, T_\ell^*)$, using $\breve{t}_\ell = \breve{T}_\ell(\bullet)$, $\mathscr{F}_{\breve{t}_\ell} \subseteq \mathscr{F}_t$, (2.7), (B.20), (B.21), and (B.16), we have

$$\begin{aligned} &\frac{\partial \, \mathrm{P}\big(\breve{T}_{\ell+1} \le t, \, I_{\ell+1} = i \mid \mathcal{F}_{\breve{T}_\ell}\big)(\bullet)}{\partial \, t} \\ &= \lim_{\Delta \downarrow 0} \frac{1}{\Delta} \Big\{ \mathrm{P}\big(\breve{T}_{\ell+1} \le t + \Delta, \, I_{\ell+1} = i \mid \mathcal{F}_{\breve{T}_\ell}\big)(\bullet) \\ &\quad - \mathrm{P}\big(\breve{T}_{\ell+1} \le t, \, I_{\ell+1} = i \mid \mathcal{F}_{\breve{T}_\ell}\big)(\bullet) \Big\} \\ &= \lim_{\Delta \downarrow 0} \frac{1}{\Delta} \mathrm{P}\big(t < \breve{T}_{\ell+1} \le t + \Delta, \, I_{\ell+1} = i \mid \mathscr{F}_{\breve{t}_\ell}\big)(\bullet) \end{aligned}$$

$$
\begin{aligned}
&= \lim_{\Delta \downarrow 0} \frac{1}{\Delta} \mathrm{P}\big(N_i(t+\Delta) \neq N_i(t), \boldsymbol{N}(t) = \boldsymbol{N}(\breve{t}_\ell) \mid \mathscr{F}_{\breve{t}_\ell}\big)(\bullet) \\
&= \lim_{\Delta \downarrow 0} \frac{1}{\Delta} \mathrm{E}\big\{\mathrm{P}(N_i(t+\Delta) \neq N_i(t) \mid \mathscr{F}_t) \cdot \mathrm{I}(\boldsymbol{N}(t) = \boldsymbol{N}(\breve{t}_\ell)) \mid \mathscr{F}_{\breve{t}_\ell}\big\}(\bullet) \\
&= \mathrm{E}\Big\{ \lim_{\Delta \downarrow 0} \frac{1}{\Delta} \mathrm{P}(N_i(t+\Delta) \neq N_i(t) \mid \mathscr{F}_t) \cdot \mathrm{I}(\boldsymbol{N}(t) = \boldsymbol{N}(\breve{t}_\ell)) \mid \mathscr{F}_{\breve{t}_\ell} \Big\}(\bullet) \qquad \text{(B.27)} \\
&= \mathrm{E}\big\{\lambda_i(t) \cdot \mathrm{I}(\boldsymbol{N}(t) = \boldsymbol{N}(\breve{t}_\ell)) \mid \mathscr{F}_{\breve{t}_\ell}\big\}(\bullet) \\
&= \mathrm{E}\big\{\lambda_i(t) \parallel \mathscr{F}_{\breve{t}_\ell}, \boldsymbol{N}(t) = \boldsymbol{N}(\breve{t}_\ell)\big\}(\bullet) \cdot \mathrm{P}\big(\boldsymbol{N}(t) = \boldsymbol{N}(\breve{t}_\ell) \mid \mathscr{F}_{\breve{t}_\ell}\big)(\bullet) \\
&= \lambda_i\big(t; \mathscr{F}_{\breve{t}_\ell}\big)(\bullet) \cdot \mathrm{P}\big(\boldsymbol{N}(t) = \boldsymbol{N}(\breve{t}_\ell) \mid \mathscr{F}_{\breve{t}_\ell}\big)(\bullet) \\
&= \lambda_i\big(\breve{t}_\ell\big)(\bullet) \cdot \exp\{-\lambda^{\mathrm{sum}}\big(\breve{t}_\ell\big)(\bullet) \cdot (t - \breve{t}_\ell)\} \\
&= \lambda_i\big(\breve{T}_\ell\big)(\bullet) \cdot \exp\{-\lambda^{\mathrm{sum}}\big(\breve{T}_\ell\big) \cdot (t - \breve{T}_\ell)\}(\bullet),
\end{aligned}
$$

where the interchange of limit and expectation in (B.27) follows by the dominated convergence theorem and condition A2′. This implies that

$$
\frac{\partial\, \mathrm{P}\big(\breve{T}_{\ell+1} \leq t,\, I_{\ell+1} = i \mid \mathcal{F}_{\breve{T}_\ell}\big)}{\partial\, t} \;=\; \lambda_i\big(\breve{T}_\ell\big) \exp\{-\lambda^{\mathrm{sum}}\big(\breve{T}_\ell\big) \cdot (t - \breve{T}_\ell)\}. \qquad \text{(B.28)}
$$

Let $t \in (\breve{t}_\ell, t_\ell^*)$, where $t_\ell^* = T_\ell^*(\bullet)$ is defined in preparation (b). Analogously as (B.15), define the realization

$$
\begin{aligned}
\text{“}\circ\text{”} &= \text{“}\bullet\text{”} \cap \{\breve{T}_{\ell+1} = t\} \\
&= \Big\{\{(\breve{T}_0, I_0), \ldots, (\breve{T}_\ell, I_\ell), \breve{T}_{\ell+1}\} = \{(\breve{t}_0, i_0), \ldots, (\breve{t}_\ell, i_\ell), t\}\Big\} \;\in\; \sigma(\mathcal{F}_{\breve{T}_\ell}, \breve{T}_{\ell+1}).
\end{aligned}
$$

Combining the fact $\breve{T}_{\ell+1}(\circ) = t$, (B.28), and (B.26), for $i \in \mathcal{V}$, we have

$$
\begin{aligned}
&\mathrm{P}\big(I_{\ell+1} = i \mid \sigma(\mathcal{F}_{\breve{T}_\ell}, \breve{T}_{\ell+1})\big)(\circ) \\
&= \mathrm{P}\big(I_{\ell+1} = i \mid \sigma(\mathcal{F}_{\breve{T}_\ell}, \{\breve{T}_{\ell+1} = t\})\big)(\circ) \\
&= \frac{\mathrm{P}(I_{\ell+1} = i,\, \breve{T}_{\ell+1} = t \mid \mathcal{F}_{\breve{T}_\ell})(\circ)}{\mathrm{P}(\breve{T}_{\ell+1} = t \mid \mathcal{F}_{\breve{T}_\ell})(\circ)} \\
&= \frac{\partial\, \mathrm{P}\big(\breve{T}_{\ell+1} \leq t,\, I_{\ell+1} = i \mid \mathcal{F}_{\breve{T}_\ell}\big)}{\partial\, t}(\circ) \Big/ \frac{\partial\, \mathrm{P}\big(\breve{T}_{\ell+1} \leq t \mid \mathcal{F}_{\breve{T}_\ell}\big)}{\partial\, t}(\circ)
\end{aligned}
$$

$$= \frac{\lambda_i(\breve{T}_\ell)(\circ) \cdot \exp\{-\lambda^{\text{sum}}(\breve{T}_\ell) \cdot (t - \breve{T}_\ell)\}(\circ)}{\lambda^{\text{sum}}(\breve{T}_\ell)(\circ) \cdot \exp\{-\lambda^{\text{sum}}(\breve{T}_\ell) \cdot (t - \breve{T}_\ell)\}(\circ)}$$
$$= \frac{\lambda_i(\breve{T}_\ell)(\circ)}{\lambda^{\text{sum}}(\breve{T}_\ell)(\circ)}, \tag{B.29}$$

which proves (4.12). For $T_\ell^* = \infty$, following the same proof as that of (B.29), we have $\mathrm{P}(I_{\ell+1} = i \mid \sigma(\mathcal{F}_{\breve{T}_\ell}, \breve{T}_{\ell+1})) = \lambda_i(\breve{T}_\ell)/\lambda^{\text{sum}}(\breve{T}_\ell)$, for $i \in \mathcal{V}$ and $\breve{T}_{\ell+1} \in (\breve{T}_\ell, \infty)$. The proof is completed. ■

B.5. Proofs of Lemmas 3–5, and Theorem 2

The proofs are divided into three parts as follows:

- In Part 1, Definition 5 introduces a class of marked point processes, called "Exponential Marked Point Process (EMPP)", which generalizes the "marked point process $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ for intensity discontinuities" in Definition 2, and facilitates derivations. Lemmas B.3–B.5 will present the probabilistic properties of the EMPP.
- In Part 2, Definition 6 introduces the notion of "t-truncated EMPP". Lemmas B.6–B.8 will present the probabilistic properties of the "t-truncated EMPP".
- In Part 3, by applying the results in Parts 1 and 2 to the "marked point process $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ for intensity discontinuities", we provide the proofs of Lemmas 3–5 and Theorem 2.

B.5.1. Part 1: EMPP and its probabilistic properties

Definition 5 (Exponential Marked Point Process (EMPP) $(\boldsymbol{T}, \boldsymbol{I})$)**.** *Let the node set $\mathcal{V} = \{1, 2, \ldots, V\}$ and let the mark set be $\mathcal{V} \cup \{0\}$. Let $T_0 = 0, I_0 = 0$ and $\mathcal{F}_{T_0} = \{\Omega, \varnothing\}$. Let $\{\mathcal{F}_{T_\ell}\}_{\ell \geq 0}$ be the filtration generated by a marked point process $(\boldsymbol{T}, \boldsymbol{I}) = (\{T_\ell\}_{\ell \geq 0}, \{I_\ell\}_{\ell \geq 0}) \in ([0, \infty), \mathcal{V} \cup \{0\})$, where $0 < T_1 < T_2 < \cdots$. We call $(\boldsymbol{T}, \boldsymbol{I})$ an "Exponential Marked Point Process (EMPP)", if for each integer $\ell \geq 0$, there exist $\mathcal{F}_{T_\ell}$-measurable random variables $\Delta_\ell \in [0, \infty]$ and $\{\lambda_{i,\ell}\}_{i \in \mathcal{V}} \in (0, \infty)$, such that the distributions of $T_{\ell+1}$ and $I_{\ell+1}$ meet the following conditions:*

(i) (*Support of $T_{\ell+1}$*) $\mathrm{P}(T_\ell < T_{\ell+1} \leq T_\ell + \Delta_\ell) = 1$.

(ii) (*Conditional distribution of $T_{\ell+1}$*) *If $\Delta_\ell < \infty$, then $T_{\ell+1}$ conditional on $\mathcal{F}_{T_\ell}$ has a mixed-type distribution, with the* p.m.f.

$$\mathrm{P}(T_{\ell+1} = T_\ell + \Delta_\ell \mid \mathcal{F}_{T_\ell}) \;=\; \exp(-\lambda_\ell^{\mathrm{sum}} \cdot \Delta_\ell) \tag{B.30}$$

at $T_\ell + \Delta_\ell$, and the p.d.f.

$$f_{T_{\ell+1}|\mathcal{F}_{T_\ell}}(x \mid \mathcal{F}_{T_\ell}) \;=\; \lambda_\ell^{\mathrm{sum}} \, \exp\{-\lambda_\ell^{\mathrm{sum}} \cdot (x - T_\ell)\}, \tag{B.31}$$

for $x \in (T_\ell, T_\ell + \Delta_\ell)$, where $\lambda_\ell^{\mathrm{sum}} = \sum_{i=1}^{V} \lambda_{i,\ell}$. If $\Delta_\ell = \infty$, then $(T_{\ell+1} - T_\ell) \mid \mathcal{F}_{T_\ell} \sim \mathrm{Exp}(\lambda_\ell^{\mathrm{sum}})$.

(iii) (*Conditional distribution of $I_{\ell+1}$*) *If $\Delta_\ell < \infty$, then $I_{\ell+1}$ has the conditional distribution: for $i \in \mathcal{V}$,*

$$\mathrm{P}(I_{\ell+1} = i \mid \sigma(\mathcal{F}_{T_\ell}, T_{\ell+1})) = \begin{cases} 0, & \text{if } T_{\ell+1} - T_\ell = \Delta_\ell, \\ \lambda_{i,\ell}/\lambda_\ell^{\mathrm{sum}}, & \text{if } 0 < T_{\ell+1} - T_\ell < \Delta_\ell. \end{cases} \tag{B.32}$$

If $\Delta_\ell = \infty$, then (B.32) *reduces to $\mathrm{P}(I_{\ell+1} = i \mid \sigma(\mathcal{F}_{T_\ell}, T_{\ell+1})) = \lambda_{i,\ell}/\lambda_\ell^{\mathrm{sum}}$, for $i \in \mathcal{V}$ and $T_{\ell+1} \in (T_\ell, \infty)$.*

Remark 8. *The "Exponential Marked Point Process* (EMPP) *$(\boldsymbol{T}, \boldsymbol{I})$" in Definition* 5 *generalizes the class of "marked point process $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ for intensity discontinuities" in Definition* 2 *which follow the distribution in Theorem* 1*, according to*

$$(\boldsymbol{T}, \boldsymbol{I}) = (\breve{\boldsymbol{T}}, \boldsymbol{I}),\ \mathcal{F}_{T_\ell} = \mathcal{F}_{\breve{T}_\ell},\ \lambda_{i,\ell} = \lambda_i(\breve{T}_\ell) = \exp\big\{\beta_{0;i} + \sum_{j \in \mathcal{V}} \beta_{j,i}\, x_j(\breve{T}_\ell)\big\},\ \Delta_\ell = T_\ell^* - \breve{T}_\ell.$$

To prove Lemmas 3–5*, we will first show probabilistic properties of "*EMPP *$(\boldsymbol{T}, \boldsymbol{I})$" which then apply to our "$(\breve{\boldsymbol{T}}, \boldsymbol{I})$ for intensity discontinuities".*

For ease of exposition, we introduce some notations similar to (4.13) and (4.14).

Duration τ_ℓ between two consecutive time points:

$$\tau_\ell = T_\ell - T_{\ell-1}, \quad \ell \geq 1. \tag{B.33}$$

Event counts $M_{i,\ell}$ at node $i \in \mathcal{V}$:

$$M_{i,0} = 0, \quad M_{i,\ell} = \sum_{k=1}^{\ell} \mathrm{I}(I_k = i), \quad \ell \geq 1. \tag{B.34}$$

Lemma B.3 presents the conditional expectation and variance of τ_k and $\mathrm{I}(I_k = i)$.

Lemma B.3 (Conditional expectation and variance related to an EMPP $(\boldsymbol{T}, \boldsymbol{I})$)**.** *Consider an* EMPP $(\boldsymbol{T}, \boldsymbol{I})$ *in Definition* 5*. For integers* $k \geq 1$ *and* $i \in \mathcal{V}$*, we have*

$$\mathrm{var}\big\{\mathrm{I}(I_k = i) - \lambda_{i,k-1}\tau_k \mid \mathcal{F}_{T_{k-1}}\big\} = \mathrm{E}\big\{\mathrm{I}(I_k = i) \mid \mathcal{F}_{T_{k-1}}\big\} = \lambda_{i,k-1}\mathrm{E}(\tau_k \mid \mathcal{F}_{T_{k-1}}). \quad \text{(B.35)}$$

Proof: For $\Delta_{k-1} < \infty$, the conditional distribution of τ_k is given by (B.30) and (B.31), and the conditional distribution of $\mathrm{I}(I_k = i)$ is given by (B.32). Direct calculations yield the following (B.36)–(B.38):

$$\begin{aligned}
&\mathrm{E}(\lambda_{i,k-1}\,\tau_k \mid \mathcal{F}_{T_{k-1}}) = \lambda_{i,k-1}\,\mathrm{E}(\tau_k \mid \mathcal{F}_{T_{k-1}}) \\
&= \lambda_{i,k-1}\Big\{\Delta_{k-1}\mathrm{P}(\tau_k = \Delta_{k-1} \mid \mathcal{F}_{T_{k-1}}) + \int_0^{\Delta_{k-1}} x\, f_{\tau_k|\mathcal{F}_{T_{k-1}}}(x \mid \mathcal{F}_{T_{k-1}})\,\mathrm{d}x\Big\} \\
&= \lambda_{i,k-1}\cdot\Big(\Delta_{k-1}\, e^{-\lambda_{k-1}^{\mathrm{sum}}\cdot\Delta_{k-1}} + \int_0^{\Delta_{k-1}} x\,\lambda_{k-1}^{\mathrm{sum}}\cdot e^{-\lambda_{k-1}^{\mathrm{sum}}\cdot x}\,\mathrm{d}x\Big) \\
&= \frac{\lambda_{i,k-1}}{\lambda_{k-1}^{\mathrm{sum}}}\cdot(1 - e^{-\lambda_{k-1}^{\mathrm{sum}}\cdot\Delta_{k-1}}) \\
&= \frac{\lambda_{i,k-1}}{\lambda_{k-1}^{\mathrm{sum}}}\,\mathrm{P}(\tau_k \neq \Delta_{k-1} \mid \mathcal{F}_{T_{k-1}}) \\
&= \mathrm{E}\big[\mathrm{E}\{\mathrm{I}(I_k = i) \mid \sigma(\mathcal{F}_{T_{k-1}}, T_k)\} \mid \mathcal{F}_{T_{k-1}}\big] \\
&= \mathrm{E}\big\{\mathrm{I}(I_k = i) \mid \mathcal{F}_{T_{k-1}}\big\},
\end{aligned} \quad \text{(B.36)}$$

together with

$$\begin{aligned}
&\mathrm{E}\big\{\lambda_{i,k-1}\,\tau_k\,\mathrm{I}(I_k = i) \mid \mathcal{F}_{T_{k-1}}\big\} \\
&= \lambda_{i,k-1}\int_0^{\Delta_{k-1}} \frac{\lambda_{i,k-1}}{\lambda_{k-1}^{\mathrm{sum}}}\, x\, f_{\tau_k|\mathcal{F}_{T_{k-1}}}(x \mid \mathcal{F}_{T_{k-1}})\,\mathrm{d}x \\
&= \lambda_{i,k-1}\int_0^{\Delta_{k-1}} \frac{\lambda_{i,k-1}}{\lambda_{k-1}^{\mathrm{sum}}}\, x\,\lambda_{k-1}^{\mathrm{sum}}\cdot e^{-\lambda_{k-1}^{\mathrm{sum}}\cdot x}\,\mathrm{d}x \\
&= \frac{\lambda_{i,k-1}^2}{(\lambda_{k-1}^{\mathrm{sum}})^2}\cdot\big\{1 - (1 + \lambda_{k-1}^{\mathrm{sum}}\cdot\Delta_{k-1})\, e^{-\lambda_{k-1}^{\mathrm{sum}}\cdot\Delta_{k-1}}\big\},
\end{aligned} \quad \text{(B.37)}$$

and

$$
\begin{aligned}
&\mathrm{E}(\lambda_{i,k-1}^2\,\tau_k^2 \mid \mathcal{F}_{T_{k-1}}) \\
&= \lambda_{i,k-1}^2 \cdot \Big\{\Delta_{k-1}^2\, e^{-\lambda_{k-1}^{\mathrm{sum}}\cdot\Delta_{k-1}} + \int_0^{\Delta_{k-1}} x^2\, \lambda_{k-1}^{\mathrm{sum}} \cdot e^{-\lambda_{k-1}^{\mathrm{sum}}\cdot x}\,\mathrm{d}x\Big\} \\
&= \lambda_{i,k-1}^2\,\Delta_{k-1}^2\, e^{-\lambda_{k-1}^{\mathrm{sum}}\cdot\Delta_{k-1}} \\
&\quad + \frac{\lambda_{i,k-1}^2}{(\lambda_{k-1}^{\mathrm{sum}})^2} \cdot \big[2 - \big\{(\lambda_{k-1}^{\mathrm{sum}})^2\Delta_{k-1}^2 + 2\,\lambda_{k-1}^{\mathrm{sum}}\cdot\Delta_{k-1} + 2\big\}\, e^{-\lambda_{k-1}^{\mathrm{sum}}\cdot\Delta_{k-1}}\big] \\
&= \frac{\lambda_{i,k-1}^2}{(\lambda_{k-1}^{\mathrm{sum}})^2} \cdot \{2 - (2\,\lambda_{k-1}^{\mathrm{sum}}\cdot\Delta_{k-1} + 2)\, e^{-\lambda_{k-1}^{\mathrm{sum}}\cdot\Delta_{k-1}}\}. \qquad \text{(B.38)}
\end{aligned}
$$

Combining (B.37) and (B.38), we have

$$
\mathrm{E}\big\{\lambda_{i,k-1}^2\,\tau_k^2 - 2\,\lambda_{i,k-1}\,\tau_k\,\mathrm{I}(I_k = i) \mid \mathcal{F}_{T_{k-1}}\big\} = 0. \qquad \text{(B.39)}
$$

For $\Delta_{k-1} = \infty$, the conditional distributions of τ_k and $\mathrm{I}(I_k = i)$ are given in parts (ii) and (iii) of Definition 5. Using this with the similar calculations as in (B.36)–(B.39), we verify that (B.36) and (B.39) also hold for $\Delta_{k-1} = \infty$. By (B.36) and (B.39), we have

$$
\begin{aligned}
&\mathrm{var}\big\{\mathrm{I}(I_k = i) - \lambda_{i,k-1}\,\tau_k \mid \mathcal{F}_{T_{k-1}}\big\} \\
&= \mathrm{E}\Big[\big\{\mathrm{I}(I_k = i) - \lambda_{i,k-1}\tau_k\big\}^2 \mid \mathcal{F}_{T_{k-1}}\Big] \\
&= \mathrm{E}\big\{\mathrm{I}(I_k = i)^2 - 2\lambda_{i,k-1}\,\tau_k\,\mathrm{I}(I_k = i) + \lambda_{i,k-1}^2\,\tau_k^2 \mid \mathcal{F}_{T_{k-1}}\big\} \\
&= \mathrm{E}\big\{\mathrm{I}(I_k = i) \mid \mathcal{F}_{T_{k-1}}\big\}. \qquad \text{(B.40)}
\end{aligned}
$$

Combining (B.36) and (B.40) completes the proof of (B.35). ■

Lemma B.4 follows from Lemma B.3.

Lemma B.4 (Martingale property for EMPP)**.** *In an* EMPP $(\boldsymbol{T}, \boldsymbol{I})$*, for each* $i \in \mathcal{V}$*, the random process*

$$
\Big\{M_{i,\ell} - \sum_{k=1}^{\ell} \lambda_{i,k-1}\,\tau_k\Big\}_{\ell \geq 1}
$$

is a martingale with respect to $\{\mathcal{F}_{T_\ell}\}_{\ell\geq 1}$.

Proof: By (B.35), for each integer $k \geq 1$, we have

$$\mathrm{E}\big\{\mathrm{I}(I_k = i) - \lambda_{i,k-1}\,\tau_k \mid \mathcal{F}_{T_{k-1}}\big\} \;=\; 0. \tag{B.41}$$

This completes the proof. ■

Lemma B.5 derives the variance of the martingale.

Lemma B.5 (Variance of the martingale $\{M_{i,\ell} - \sum_{k=1}^{\ell} \lambda_{i,k-1}\,\tau_k\}_{\ell\geq 1}$)**.** *In an* EMPP $(\boldsymbol{T}, \boldsymbol{I})$*, for integers* $i \in \mathcal{V}$ *and* $\ell \geq 1$*, we have*

$$\mathrm{var}\Big(M_{i,\ell} - \sum_{k=1}^{\ell} \lambda_{i,k-1}\,\tau_k\Big) \;=\; \mathrm{E}(M_{i,\ell}).$$

Proof: Using (B.41), for any indices r and k such that $1 \leq r < k$, we have

$$\begin{aligned}
&\mathrm{E}\Big[\big\{\mathrm{I}(I_r = i) - \lambda_{i,r-1}\,\tau_r\big\}\,\big\{\mathrm{I}(I_k = i) - \lambda_{i,k-1}\,\tau_k\big\}\Big] \\
&= \mathrm{E}\Big(\mathrm{E}\big[\{\mathrm{I}(I_r = i) - \lambda_{i,r-1}\,\tau_r\}\,\{\mathrm{I}(I_k = i) - \lambda_{i,k-1}\,\tau_k\}\,\big|\,\mathcal{F}_{T_{k-1}}\big]\Big) \\
&= \mathrm{E}\Big[\big\{\mathrm{I}(I_r = i) - \lambda_{i,r-1}\,\tau_r\big\}\,\mathrm{E}\big\{\mathrm{I}(I_k = i) - \lambda_{i,k-1}\,\tau_k\,\big|\,\mathcal{F}_{T_{k-1}}\big\}\Big] \\
&= 0.
\end{aligned} \tag{B.42}$$

For any index $\ell \geq 1$, uses of (B.42), (B.35), and (B.40) give that

$$\begin{aligned}
&\mathrm{E}\Big\{\Big(M_{i,\ell} - \sum_{k=1}^{\ell} \lambda_{i,k-1}\,\tau_k\Big)^2\Big\} \\
&= \mathrm{E}\Big(\Big[\sum_{k=1}^{\ell} \big\{\mathrm{I}(I_k = i) - \lambda_{i,k-1}\,\tau_k\big\}\Big]^2\Big) \\
&= \sum_{k=1}^{\ell} \mathrm{E}\Big[\big\{\mathrm{I}(I_k = i) - \lambda_{i,k-1}\,\tau_k\big\}^2\Big] \\
&\quad + \sum_{1\leq k\neq r\leq \ell} \sum \mathrm{E}\Big[\big\{\mathrm{I}(I_r = i) - \lambda_{i,r-1}\,\tau_r\big\}\,\big\{\mathrm{I}(I_k = i) - \lambda_{i,k-1}\,\tau_k\big\}\Big]
\end{aligned}$$

$$= \sum_{k=1}^{\ell} \mathrm{E}\Big[\big\{\mathrm{I}(I_k = i) - \lambda_{i,k-1}\,\tau_k\big\}^2\Big]$$
$$= \sum_{k=1}^{\ell} \mathrm{E}\big\{\mathrm{I}(I_k = i)\big\} = \mathrm{E}(M_{i,\ell}).$$

This completes the proof. ■

B.5.2. Part 2: t-truncated EMPP and its probabilistic properties

We next derive the probabilistic results of EMPP $(\boldsymbol{T}, \boldsymbol{I})$ when the point process $\boldsymbol{T} = \{T_0, T_1, \ldots\}$ reaches a pre-specified time point $t \in (0, \infty)$. Definition 6 introduces the notion of "t-truncated EMPP".

Definition 6 (t-truncated EMPP)**.** *Consider an* EMPP $(\boldsymbol{T}, \boldsymbol{I}) = (\{T_\ell\}_{\ell\geq 0}, \{I_\ell\}_{\ell\geq 0})$. *Let* $t \in (0, \infty)$ *be a given deterministic time point. Define the marked point process* $(\boldsymbol{T}^{[t]}, \boldsymbol{I}^{[t]}) = (\{T_\ell^{[t]}\}_{\ell\geq 0}, \{I_\ell^{[t]}\}_{\ell\geq 0})$*, with*

$$T_\ell^{[t]} = T_\ell \wedge t, \;\; I_\ell^{[t]} = I_\ell\,\mathrm{I}(T_\ell \leq t). \tag{B.43}$$

We call $(\boldsymbol{T}^{[t]}, \boldsymbol{I}^{[t]})$ *the "t-truncated* EMPP *from* $(\boldsymbol{T}, \boldsymbol{I})$*".*

Lemma B.6 states that any double sequence $(\boldsymbol{T}^{[t]}, \boldsymbol{I}^{[t]})$ defined in (B.43) is an EMPP.

Lemma B.6. *Let* $(\boldsymbol{T}, \boldsymbol{I}) = (\{T_\ell\}_{\ell\geq 0}, \{I_\ell\}_{\ell\geq 0})$ *be an* EMPP *defined in Definition* 5 *associated with* $\{\lambda_{i,\ell}\}_{i\in\mathcal{V}}$ *and* Δ_ℓ *in* (B.30)–(B.32)*. For* $t \in (0, \infty)$*, let* $(\boldsymbol{T}^{[t]}, \boldsymbol{I}^{[t]})$ *be the corresponding* t*-truncated* EMPP *as in Definition* 6 *Then the probability distributions of* $\boldsymbol{T}^{[t]}$ *and* $\boldsymbol{I}^{[t]}$*) meet the conditions* (i) *and* (ii) *in Definition* 5 *associated with* $\{\lambda_{i,\ell}\}_{i\in\mathcal{V}}$ *and* $\Delta_{\ell,t}$ *(instead of* Δ_ℓ*), where*

$$\Delta_{\ell,t} \;=\; \Delta_\ell \wedge (t - T_\ell^{[t]}), \tag{B.44}$$

and thus $(\boldsymbol{T}^{[t]}, \boldsymbol{I}^{[t]})$ *is an* EMPP*.*

Proof: To prove Lemma B.6, it suffices to show that for each integer $\ell \geq 0$, $(T_{\ell+1}^{[t]}, I_{\ell+1}^{[t]})$ follows the conditional distribution in (B.30)–(B.32) with T_ℓ and Δ_ℓ replaced by $T_\ell^{[t]}$ and $\Delta_{\ell,t}$. We proceed by cases of $T_\ell^{[t]}$.

Case (i) $T_\ell^{[t]} < t - \Delta_\ell$.

From (B.43) and (B.44), we observe that $T_\ell^{[t]} = T_\ell$ and $\Delta_{\ell,t} = \Delta_\ell$. Also, from (B.30) and (B.31), we know that $T_{\ell+1} \leq T_\ell + \Delta_\ell \leq t$. Thus, $(T_{\ell+1}^{[t]},\ I_{\ell+1}^{[t]}) = (T_{\ell+1} \wedge t,\ I_{\ell+1}\mathrm{I}(T_{\ell+1} \leq t)) = (T_{\ell+1},\ I_{\ell+1})$, follows the conditional distribution in (B.30)–(B.32).

Case (ii) $t - \Delta_\ell \leq T_\ell^{[t]} < t$.

By (B.43) and (B.44), we have $T_\ell^{[t]} = T_\ell$ and $\Delta_{\ell,t} = t - T_\ell$. Using (B.30) and (B.31), we obtain

$$\mathrm{P}((T_{\ell+1} \wedge t) = t \mid \mathcal{F}_{T_\ell}) \,=\, \exp\{-\lambda_\ell^{\mathrm{sum}} \cdot (t - T_\ell)\},$$

and

$$f_{(T_{\ell+1}\wedge t)|\mathcal{F}_{T_\ell}}(x \mid \mathcal{F}_{T_\ell}) \,=\, \lambda_\ell^{\mathrm{sum}} \exp\{-\lambda_\ell^{\mathrm{sum}} \cdot (x - T_\ell)\}, \quad \text{for } x \in (T_\ell, t).$$

This indicates that $T_{\ell+1}^{[t]} = T_{\ell+1} \wedge t$ follows the distribution in (B.30)–(B.31) with T_ℓ and Δ_ℓ replaced by $T_\ell^{[t]} = T_\ell$ and $\Delta_{\ell,t} = t - T_\ell$. Also, by checking (B.32), we have that $I_{\ell+1}^{[t]} = I_{\ell+1}\mathrm{I}(T_{\ell+1} \leq t)$ conditional on $T_{\ell+1}^{[t]}$, follows the distribution in (B.32).

Case (iii) $T_\ell^{[t]} = t$.

By (B.44), we know $\Delta_{\ell,t} = 0$. Then $(T_{\ell+1}^{[t]}, I_{\ell+1}^{[t]}) = (t, 0)$ follows the distribution in (B.30)–(B.32) with T_ℓ and Δ_ℓ replaced by $T_\ell^{[t]} = t$ and $\Delta_{\ell,t} = 0$.

Summarizing the above three cases completes the proof. ■

Similar to (B.33) and (B.34), we use notations $\tau_\ell^{[t]}$ and $M_{i,\ell}^{[t]}$ for the duration and event counts respectively of the t-truncated EMPP $(\boldsymbol{T}^{[t]}, \boldsymbol{I}^{[t]})$, i.e.,

$$\tau_\ell^{[t]} = T_\ell^{[t]} - T_{\ell-1}^{[t]}, \quad \ell \geq 1, \tag{B.45}$$

$$M_{i,0}^{[t]} = 0,\ M_{i,\ell}^{[t]} = \sum_{k=1}^{\ell} \mathrm{I}(I_k^{[t]} = i), \quad \ell \geq 1. \tag{B.46}$$

Define $M_{i,\infty}^{[t]} = \lim_{\ell\to\infty} M_{i,\ell}^{[t]}$. Let $L_t = \sum_{\ell=1}^{\infty} \mathrm{I}(T_\ell \le t)$. From (B.43),

$$M_{i,\infty}^{[t]} = \sum_{k=1}^{\infty} \mathrm{I}(I_k^{[t]} = i) = \sum_{k=1}^{L_t} \mathrm{I}(I_k = i) = M_{i,L_t} \tag{B.47}$$

is the total event counts of node i in the time interval $[0, t]$. Lemma B.7 shows an upper bound for $\mathrm{E}(M_{i,\infty}^{[t]})$.

Lemma B.7 (Upper bound for $\mathrm{E}(M_{i,\infty}^{[t]})$)**.** *Let* $(\boldsymbol{T}, \boldsymbol{I})$ *be an* EMPP*, and* $(\boldsymbol{T}^{[t]}, \boldsymbol{I}^{[t]})$ *be the corresponding t-truncated* EMPP *from* $(\boldsymbol{T}, \boldsymbol{I})$ *as in Definition* 6*. Assume that* $\sup_{i\in\mathcal{V},\ell\ge 0} \lambda_{i,\ell} \le c$ *for a constant* $c \in (0, \infty)$*. Then*

$$\mathrm{E}(M_{i,\infty}^{[t]}) \le c\, t. \tag{B.48}$$

Proof: Lemma B.6 verified that $(\boldsymbol{T}^{[t]}, \boldsymbol{I}^{[t]})$ is an EMPP. Applying Lemma B.4 to $(\boldsymbol{T}^{[t]}, \boldsymbol{I}^{[t]})$ gives that for each integer $\ell \ge 1$,

$$\mathrm{E}(M_{i,\ell}^{[t]}) \;=\; \mathrm{E}\Big(\sum_{k=1}^{\ell} \lambda_{i,k-1}\tau_k^{[t]}\Big).$$

Note that $M_{i,\ell}^{[t]}$ and $\sum_{k=1}^{\ell} \lambda_{i,k-1}\tau_k^{[t]}$ are monotonically increasing in ℓ. By the monotone convergence theorem, we obtain

$$\begin{aligned} &\mathrm{E}(M_{i,\infty}^{[t]}) = \lim_{\ell\to\infty} \mathrm{E}(M_{i,\ell}^{[t]}) \\ &= \lim_{\ell\to\infty} \mathrm{E}\Big(\sum_{k=1}^{\ell} \lambda_{i,k-1}\,\tau_k^{[t]}\Big) = \mathrm{E}\Big(\sum_{k=1}^{\infty} \lambda_{i,k-1}\,\tau_k^{[t]}\Big) \\ &\le c\,\mathrm{E}\Big(\sum_{k=1}^{\infty} \tau_k^{[t]}\Big) \le c\, t. \end{aligned} \tag{B.49}$$

This completes the proof. ■

Lemma B.8 (Upper bound for $\mathrm{var}\{\sum_{k=1}^{\infty} X_{k-1}\,\mathrm{I}(I_k^{[t]} = i) - \sum_{k=1}^{\infty} X_{k-1}\,\lambda_{i,k-1}\,\tau_k^{[t]}\}$)**.** *Let* $(\boldsymbol{T}, \boldsymbol{I})$ *be an* EMPP*, and* $(\boldsymbol{T}^{[t]}, \boldsymbol{I}^{[t]})$ *be the corresponding t-truncated* EMPP *from* $(\boldsymbol{T}, \boldsymbol{I})$ *as in Definition*

6. *Assume that* $\sup_{i\in\mathcal{V},\ell\geq 0}\lambda_{i,\ell} \leq c$ *for a constant* $c \in (0,\infty)$. *Let* $\{X_\ell\}_{\ell\geq 0}$ *be a sequence of random variables, such that* $X_\ell \geq 0$ *is measurable with respect to* $\mathcal{F}_{T_\ell}$ *for each integer* $\ell \geq 0$, *and* $\sup_{\ell\geq 0} X_\ell \leq c_2$ a.s. *for a constant* $c_2 \in (0,\infty)$. *Then*

$$\mathrm{E}\Big\{\sum_{k=1}^{\infty} X_{k-1}\,\mathrm{I}(I_k^{[t]}=i) - \sum_{k=1}^{\infty} X_{k-1}\,\lambda_{i,k-1}\,\tau_k^{[t]}\Big\} = 0, \tag{B.50}$$

and

$$\begin{aligned}&\mathrm{var}\Big\{\sum_{k=1}^{\infty} X_{k-1}\,\mathrm{I}(I_k^{[t]}=i) - \sum_{k=1}^{\infty} X_{k-1}\,\lambda_{i,k-1}\,\tau_k^{[t]}\Big\}\\ &\qquad = \mathrm{E}\Big\{\sum_{k=1}^{\infty} X_{k-1}^2\,\mathrm{I}(I_k^{[t]}=i)\Big\} \leq c\,c_2^2\,t.\end{aligned} \tag{B.51}$$

Proof: An argument similar to (B.49) gives that

$$\begin{aligned}&\mathrm{E}\Big\{\sum_{k=1}^{\infty} X_{k-1}\,\mathrm{I}(I_k^{[t]}=i)\Big\} = \lim_{\ell\to\infty}\mathrm{E}\Big\{\sum_{k=1}^{\ell} X_{k-1}\,\mathrm{I}(I_k^{[t]}=i)\Big\}\\ &= \lim_{\ell\to\infty}\mathrm{E}\Big(\sum_{k=1}^{\ell} X_{k-1}\,\lambda_{i,k-1}\,\tau_k^{[t]}\Big) = \mathrm{E}\Big(\sum_{k=1}^{\infty} X_{k-1}\,\lambda_{i,k-1}\,\tau_k^{[t]}\Big),\end{aligned}$$

which proves (B.50). Using the similar proof as that of Lemma B.5, we have

$$\mathrm{E}\Big\{\Big(\sum_{k=1}^{\ell} X_{k-1}\,\mathrm{I}(I_k^{[t]}=i) - \sum_{k=1}^{\ell} X_{k-1}\,\lambda_{i,k-1}\,\tau_k^{[t]}\Big)^2\Big\} = \mathrm{E}\Big\{\sum_{k=1}^{\ell} X_{k-1}^2\,\mathrm{I}(I_k^{[t]}=i)\Big\}. \tag{B.52}$$

Note that $\lim_{\ell\to\infty}\sum_{k=1}^{\ell} X_{k-1}\,\mathrm{I}(I_k^{[t]}=i) = \sum_{k=1}^{\infty} X_{k-1}\,\mathrm{I}(I_k^{[t]}=i)$, a.s., and

$$\lim_{\ell\to\infty}\sum_{k=1}^{\ell} X_{k-1}\,\lambda_{i,k-1}\tau_k^{[t]} = \sum_{k=1}^{\infty} X_{k-1}\,\lambda_{i,k-1}\,\tau_k^{[t]} \leq \sum_{k=1}^{\infty} c\,c_2\,\tau_k^{[t]} = c\,c_2\,t < \infty, \quad \text{a.s..}$$

It follows that

$$\sum_{k=1}^{\ell} X_{k-1}\,\mathrm{I}(I_k^{[t]}=i) - \sum_{k=1}^{\ell} X_{k-1}\,\lambda_{i,k-1}\,\tau_k^{[t]}$$

$$\stackrel{\text{a.s.}}{\to} \sum_{k=1}^{\infty} X_{k-1}\,\mathrm{I}(I_k^{[t]} = i) - \sum_{k=1}^{\infty} X_{k-1}\,\lambda_{i,k-1}\,\tau_k^{[t]}, \quad \text{as } \ell \to \infty. \tag{B.53}$$

Also, for each integer $\ell \geq 1$, we have

$$\begin{aligned}
&\Big\{\sum_{k=1}^{\ell} X_{k-1}\,\mathrm{I}(I_k^{[t]} = i) - \sum_{k=1}^{\ell} X_{k-1}\,\lambda_{i,k-1}\,\tau_k^{[t]}\Big\}^2 \\
&\quad \leq \Big\{\sum_{k=1}^{\ell} X_{k-1}\,\mathrm{I}(I_k^{[t]} = i)\Big\}^2 + \Big(\sum_{k=1}^{\ell} X_{k-1}\,\lambda_{i,k-1}\,\tau_k^{[t]}\Big)^2 \\
&\quad \leq (c_2\,M_{i,\infty}^{[t]})^2 + (c\,c_2\,t)^2.
\end{aligned} \tag{B.54}$$

By (B.52), (B.53), (B.54) and the dominated convergence theorem, we have

$$\begin{aligned}
&\mathrm{E}\Big\{\Big(\sum_{k=1}^{\infty} X_{k-1}\,\mathrm{I}(I_k^{[t]} = i) - \sum_{k=1}^{\infty} X_{k-1}\,\lambda_{i,k-1}\,\tau_k^{[t]}\Big)^2\Big\} \\
&= \lim_{\ell\to\infty} \mathrm{E}\Big\{\Big(\sum_{k=1}^{\ell} X_{k-1}\,\mathrm{I}(I_k^{[t]} = i) - \sum_{k=1}^{\ell} X_{k-1}\,\lambda_{i,k-1}\,\tau_k^{[t]}\Big)^2\Big\} \\
&= \lim_{\ell\to\infty} \mathrm{E}\Big\{\sum_{k=1}^{\ell} X_{k-1}^2\,\mathrm{I}(I_k^{[t]} = i)\Big\} \\
&= \mathrm{E}\Big\{\sum_{k=1}^{\infty} X_{k-1}^2\,\mathrm{I}(I_k^{[t]} = i)\Big\} \\
&\leq c_2^2\,\mathrm{E}(M_{i,\infty}^{[t]}) \;\leq\; c\,c_2^2\,t,
\end{aligned}$$

where the last inequality is from (B.48). Thus, (B.51) is proved. ■

B.5.3. Part 3: Proofs of Lemmas 3–5 and Theorem 2

Remark 8 verified that the "marked point process $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ for intensity discontinuities" is an EMPP. Let $(\breve{\boldsymbol{T}}^{[t]}, \boldsymbol{I}^{[t]})$ be the "t-truncated EMPP" from $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ as in Definition 6, i.e., for each integer $\ell \geq 1$, $(\breve{T}_\ell^{[t]}, I_\ell^{[t]}) = \big(\breve{T}_\ell \wedge t\,,\; I_\ell\,\mathrm{I}(\breve{T}_\ell \leq t)\big)$. Recall that $M_{i,\ell}^{[t]}$ defined in (B.46) is the event counts corresponding to $(\breve{\boldsymbol{T}}^{[t]}, \boldsymbol{I}^{[t]})$, and $M_{i,\infty}^{[t]} = \lim_{\ell\to\infty} M_{i,\ell}^{[t]}$. Recall L_t defined in

(4.17). Using (4.7), (4.14) and (B.47), $N_i(t)$ has the equivalent expressions:

$$N_i(t) \;=\; M_{i,L_t} \;=\; \sum_{k=1}^{L_t} \mathrm{I}(I_k = i) \;=\; M_{i,\infty}^{[t]}. \tag{B.55}$$

Following (B.45), let $\tau_\ell^{[t]} = \breve{T}_\ell^{[t]} - \breve{T}_{\ell-1}^{[t]}$ be the duration between two consecutive time points $\breve{T}_{\ell-1}^{[t]}$ and $\breve{T}_\ell^{[t]}$. It is easy to check that this $\tau_\ell^{[t]}$ is identical to that defined in (4.18). Using (4.15) and (4.18), the integral $\int_0^t \lambda_i(u)\,\mathrm{d}u$ has the following equivalent expressions:

$$\int_0^t \lambda_i(u)\,\mathrm{d}u \;=\; \sum_{k=1}^{L_t+1} \lambda_{i,k-1}\,\tau_k^{[t]} \;=\; \sum_{k=1}^{\infty} \lambda_{i,k-1}\,\tau_k^{[t]}. \tag{B.56}$$

After clarifying the facts above, we next prove Lemmas 3–5 and Theorem 2.

B.5.4. Proof of Lemma 3

Lemma 3 is directly obtained by Lemmas B.3 and B.5. ■

B.5.5. Proof of Lemma 4

Lemma 4 is directly obtained by Lemma B.4. ■

B.5.6. Proof of Lemma 5

From (B.43), (B.46) and (B.56), we have $\sum_{k=1}^{L_t} X_{k-1}\,\mathrm{I}(I_k = i) = \sum_{k=1}^{\infty} X_{k-1}\,\mathrm{I}(I_k^{[t]} = i)$, and $\sum_{k=1}^{L_t+1} X_{k-1}\,\lambda_{i,k-1}\,\tau_k^{[t]} = \sum_{k=1}^{\infty} X_{k-1}\,\lambda_{i,k-1}\,\tau_k^{[t]}$. Then all the results in Lemma 5 could be directly obtained by Lemma B.8. ■

B.5.7. Proof of Theorem 2

Recall that $N_i(t) = M_{i,L_t}$ in (B.55) and $\int_0^t \lambda_i(u)\,\mathrm{d}u = \sum_{k=1}^{L_t+1} \lambda_{i,k-1}\,\tau_k^{[t]}$ in (B.56). For each integer $\ell \geq 0$, let $X_\ell = x(\breve{T}_\ell)$. We further have $\int_0^t x(u-)\,\mathrm{d}N_i(u) = \sum_{k=1}^{L_t} X_{k-1}\,\mathrm{I}(I_k = i)$ and $\int_0^t x(u)\,\lambda_i(u)\,\mathrm{d}u = \sum_{k=1}^{L_t+1} X_{k-1}\,\lambda_{i,k-1}\,\tau_k^{[t]}$. Then (4.22) and (4.24) in Theorem 2 are directly

implied by (4.21) and (4.20) respectively in Lemma 5. The finiteness of $\boldsymbol{N}(t)$ in (4.23) is directly implied by (4.22). ■

B.6. Proof of Lemma 6

Before proving Lemma 6, we first present Lemma B.9.

Lemma B.9 (Probabilistic inequalities for $\lambda_i(t)$ in our (3.1))**.** *Assume conditions* A1–A4 *and* A2′ *in Appendix A. Let* $\{\lambda_i(t)\}_{i\in\mathcal{V}}$ *be the conditional intensity functions defined in* (3.1)*. Then for any distinct* $i, j \in \mathcal{V}$*, there exist constants* $c_0, c_1, c_2, c_3 \in (0,\infty)$*, such that for any* $t \in (0,\phi)$*, the following assertions hold:*

$$\mathrm{P}(\lambda_i(t) = \exp(\beta_{0;i})) \geq \exp(-c_0\, t), \tag{B.57}$$

$$\mathrm{P}\big(\lambda_i(t) = \exp\{\beta_{0;i} + \beta_{j,i}\cdot g(1/\phi)\}\big) \geq c_1\cdot\exp(-c_2\, t)\cdot\{1-\exp(-c_3\, t)\}. \tag{B.58}$$

Proof: For any $t \in (0,\phi)$, (3.1), (3.5) and (3.6) indicate that "$\{\boldsymbol{N}(t)=0\} \subseteq \{\lambda_i(t) = \exp(\beta_{0;i})\}$" and "$\{N_j(t)=1 \text{ and } N_k(t)=0 \text{ for all } k\in\mathcal{V}\diagdown j\} \subseteq \{\lambda_i(t) = \exp(\beta_{0;i}+\beta_{j,i}\cdot g(1/\phi))\}$". To verify (B.57) and (B.58), it suffices to show that for any $t\in(0,\phi)$,

$$\mathrm{P}(\boldsymbol{N}(t)=0) \;=\; \exp(-c_0\, t), \tag{B.59}$$

$$\begin{aligned}\mathrm{P}(N_j(t)=1,\ &\text{and } N_k(t)=0 \text{ for all } k\in\mathcal{V}\diagdown j)\\ &\geq\; c_1\cdot\exp(-c_2\, t)\cdot\{1-\exp(-c_3\, t)\}.\end{aligned} \tag{B.60}$$

Recall that in Theorem 1, we have the following facts: $\breve{T}_0 = 0$, $\mathcal{F}_{\breve{T}_0} = \mathscr{F}_0 = \{\Omega,\varnothing\}$, $\mathcal{T}_0 = \varnothing$ and $T_0^* = \infty$. By part (ii) of Theorem 1 and the fact that $T_0^*=\infty$, we have that $\breve{T}_1$ is a continuous random variable with the p.d.f.:

$$f_{\breve{T}_1}(x) = \lambda^{\mathrm{sum}}(0)\exp\{-\lambda^{\mathrm{sum}}(0)\cdot x\},\quad x\in(0,\infty). \tag{B.61}$$

On the other hand, it is easy to verify from (4.4) that the first discontinuity point $\breve{T}_1 =$

$\min\{T_{j,\ell} : j \in \mathcal{V}, \ell \geq 1\}$ is exactly the first event-occurrence time point, i.e.,

$$\boldsymbol{N}(t) = 0 \quad \text{if and only if} \quad 0 \leq t < \breve{T}_1. \tag{B.62}$$

Combining (B.61) and (B.62), we have

$$\mathrm{P}\big(\boldsymbol{N}(t) = 0\big) = \exp\{-\lambda^{\mathrm{sum}}(0) \cdot t\}.$$

This proves (B.59) with a constant $c_0 = \lambda^{\mathrm{sum}}(0) = \sum_{i=1}^{V} \exp(\beta_{0;i})$.

Similarly, parts (ii) and (iii) of Theorem 3 yield that

$$\begin{aligned}
&\mathrm{P}\big(N_j(t) = 1, \text{ and } N_k(t) = 0 \text{ for all } k \in \mathcal{V} \diagdown j\big) \\
&= \mathrm{P}\big(\breve{T}_1 \leq t,\, I_1 = j,\, \breve{T}_2 > t\big) \\
&= \int_0^t f_{\breve{T}_1}(x) \cdot \mathrm{P}\big(I_1 = j \mid \breve{T}_1 = x\big) \cdot \mathrm{P}\big(\breve{T}_2 > t \mid \breve{T}_1 = x,\, I_1 = j\big)\, \mathrm{d}x \\
&= \int_0^t \lambda^{\mathrm{sum}}(0) \exp\{-\lambda^{\mathrm{sum}}(0) \cdot x\} \cdot \frac{\lambda_j(0)}{\lambda^{\mathrm{sum}}(0)} \\
&\qquad\qquad \times \Big[1 - \int_x^t \lambda^{\mathrm{sum}}(x) \exp\{-\lambda^{\mathrm{sum}}(x) \cdot (u - x)\}\, \mathrm{d}u\Big]\, \mathrm{d}x \\
&= \int_0^t \lambda_1 \exp(-\lambda_2 \cdot x) \cdot \exp\{-\lambda_3 \cdot (t - x)\}\, \mathrm{d}x \\
&= \begin{cases} \lambda_1 \cdot \{\exp(-\lambda_3 \cdot t) - \exp(-\lambda_2 \cdot t)\}/(\lambda_2 - \lambda_3), & \text{if } \lambda_2 \neq \lambda_3, \\ \lambda_1 \cdot t \cdot \exp(-\lambda_2 \cdot t), & \text{if } \lambda_2 = \lambda_3, \end{cases}
\end{aligned} \tag{B.63}$$

where $\lambda_1 = \lambda_j(0) = \exp(\beta_{0;j})$, $\lambda_2 = \lambda^{\mathrm{sum}}(0) = \sum_{i=1}^{V} \exp(\beta_{0;i})$, and $\lambda^{\mathrm{sum}}(x)$ reduces to $\lambda_3 = \sum_{i=1}^{V} \exp\{\beta_{0;i} + \beta_{j,i}\, g(1/\phi)\}$.

For $\lambda_2 \neq \lambda_3$, if $\lambda_2 - \lambda_3 = \delta > 0$, then (B.63) gives that

$$\mathrm{P}\big(N_j(t) = 1, \text{ and } N_k(t) = 0 \text{ for all } k \in \mathcal{V} \diagdown j\big) = \lambda_1/\delta \cdot \exp(-\lambda_3 \cdot t)\{1 - \exp(-\delta\, t)\}.$$

Thus, (B.60) holds with $c_1 = \lambda_1/\delta$, $c_2 = \lambda_3$ and $c_3 = \delta$. Due to the symmetry between λ_2

and λ_3 in (B.63), the similar argument holds when $\lambda_2 - \lambda_3 < 0$.

For $\lambda_2 = \lambda_3$, (B.63) yields that

$$\begin{aligned}
&\mathrm{P}\big(N_j(t) = 1, \text{ and } N_k(t) = 0 \text{ for all } k \in \mathcal{V} \diagdown j\big) \\
&= \lambda_1 \cdot t \cdot \exp(-\lambda_2 \cdot t) \geq \lambda_1 \cdot \exp(-\lambda_2 \cdot t) \cdot \{1 - \exp(-t)\}.
\end{aligned}$$

Hence, (B.60) holds with $c_1 = \lambda_1$, $c_2 = \lambda_2$ and $c_3 = 1$. This completes the proof. ■

Next we prove Lemma 6 using a proof by contradiction. For any $(j_0, i_0) \in \mathcal{E}$ with $\beta_{j_0,i_0} \neq 0$, if $\{N_{i_0}(t)\}_{t\geq 0}$ is stationary, then property (P1) implies that the conditional intensity function $\lambda_{i_0}(t)$ has the same distribution for all $t \in (0, \phi)$. Thus, for the two possible values $\exp(\beta_{0;i_0})$ and $\exp\{\beta_{0;i_0} + \beta_{j_0,i_0} \cdot g(1/\phi)\}$ of $\lambda_{i_0}(t)$, there exist some constants $c_4, c_5 \in [0, 1]$, such that $\mathrm{P}(\lambda_{i_0}(t) = \exp(\beta_{0;i_0})) \equiv c_4$ and $\mathrm{P}\big(\lambda_{i_0}(t) = \exp\{\beta_{0;i_0} + \beta_{j_0,i_0} \cdot g(1/\phi)\}\big) \equiv c_5$ hold for any $t \in (0, \phi)$. Combining this with (B.57) and (B.58), for any $t \in (0, \phi)$, we have

$$\begin{aligned}
&\mathrm{P}\big(\lambda_{i_0}(t) = \exp(\beta_{0;i_0}), \text{ or } \lambda_{i_0}(t) = \exp\{\beta_{0;i_0} + \beta_{j_0,i_0} \cdot g(1/\phi)\}\big) \\
&\equiv c_4 + c_5 \\
&\geq \sup_{t\in(0,\phi)} \big\{ \exp(-c_0\, t)\big\} + \sup_{t\in(0,\phi)} \big\{c_1 \cdot \exp(-c_2\, t) \cdot \{1 - \exp(-c_3\, t)\}\big\} \\
&\geq 1 + c_1 \cdot \exp(-c_2 \cdot \phi/2) \cdot \{1 - \exp(-c_3 \cdot \phi/2)\} > 1,
\end{aligned}$$

which obviously contradicts. This completes the proof. ■

B.7. Proof of Lemma 7

From (4.25), for each integer $\ell \geq 1$, we have $R_\ell = \min(\mathcal{U}_\ell)$, where $\mathcal{U}_\ell = \{t \geq R_{\ell-1} + \phi : \boldsymbol{N}((t - \phi, t]) = 0\}$. Thus, R_ℓ exists if and only if the following two conditions hold:

(i) $\mathcal{U}_\ell \neq \varnothing$. (Since $\mathcal{U}_\ell$ is bounded below, this indicates $\inf(\mathcal{U}_\ell)$ exists.)

(ii) $\inf(\mathcal{U}_\ell) \in \mathcal{U}_\ell$. (This indicates $\min(\mathcal{U}_\ell) = \inf(\mathcal{U}_\ell)$.)

We start by proving the existence of R_1. We first prove that condition (i) holds with probability one for $\mathcal{U}_1$. Note that $\mathcal{U}_1 \neq \varnothing$ if and only if there exists $t \geq \phi$ such that $\boldsymbol{N}((t-\phi, t]) = 0$. It suffices to show that

$$\mathrm{P}\Big(\bigcup_{t \geq \phi} \big\{ \boldsymbol{N}((t-\phi, t]) = 0 \big\} \Big) = 1. \tag{B.64}$$

By condition A4, there exists some constant $c \in (0, \infty)$ such that $\lambda^{\mathrm{sum}}(t) \leq c$. This together with (B.10) gives that $\lambda^{\mathrm{sum}}(t; \mathscr{F}_s) \leq c$. Then by (B.9), for $t > s \geq 0$, we have

$$\begin{aligned} \mathrm{P}\big(\boldsymbol{N}(t) = \boldsymbol{N}(s) \mid \mathscr{F}_s\big) &= \exp\Big\{ - \int_s^t \lambda^{\mathrm{sum}}(u; \mathscr{F}_s)\, \mathrm{d}u \Big\} \\ &\geq \exp\{-c \cdot (t-s)\}. \end{aligned} \tag{B.65}$$

For each integer $k \geq 1$, plugging $s = (k-1)\,\phi$ and $t = k\,\phi$ into (B.65), we obtain

$$\mathrm{P}(A_k^* \mid \mathscr{F}_{(k-1)\,\phi}) \;\geq\; \exp(-c\,\phi), \tag{B.66}$$

where the event

$$A_k^* = \Big\{ \boldsymbol{N}\big(((k-1)\,\phi, k\,\phi]\big) = 0 \Big\} \;=\; \Big\{ \boldsymbol{N}\big((k-1)\,\phi\big) = \boldsymbol{N}(k\,\phi) \Big\}.$$

Letting $k = 1$ in (B.66) yields that

$$\mathrm{P}(A_1^*) \;\geq\; \exp(-c\,\phi). \tag{B.67}$$

Also, for integers $k \geq 2$, by (B.66) and the fact that $\{\bigcap_{m=1}^{k-1} \overline{A_m^*}\} \in \mathscr{F}_{(k-1)\,\phi}$, we have

$$\mathrm{P}\Big(A_k^* \,\Big|\, \bigcap_{m=1}^{k-1} \overline{A_m^*} \Big) \;\geq\; \exp(-c\,\phi).$$

Combining this with (B.67), for any integer $\ell \geq 2$, we have

$$\begin{aligned}
\mathrm{P}\Big(\bigcap_{k=1}^{\ell} \overline{A_k^*} \Big) &= \mathrm{P}(\overline{A_1^*}) \cdot \prod_{k=2}^{\ell} \mathrm{P}\Big(\overline{A_k^*} \,\Big|\, \bigcap_{m=1}^{k-1} \overline{A_m^*} \Big) \\
&= \{1 - \mathrm{P}(A_1^*)\} \cdot \prod_{k=2}^{\ell} \Big\{ 1 - \mathrm{P}\Big(A_k^* \,\Big|\, \bigcap_{m=1}^{k-1} \overline{A_m^*} \Big) \Big\} \\
&\leq \big\{ 1 - \exp(-c\,\phi) \big\}^{\ell}, \qquad \text{(B.68)}
\end{aligned}$$

which gives that

$$\mathrm{P}\Big(\bigcup_{k=1}^{\ell} A_k^* \Big) \;=\; 1 - \mathrm{P}\Big(\bigcap_{k=1}^{\ell} \overline{A_k^*} \Big) \;\geq\; 1 - \{1 - \exp(-c\,\phi)\}^{\ell}. \qquad \text{(B.69)}$$

Letting $\ell \to \infty$ in (B.69) yields that $\mathrm{P}\big(\bigcup_{k=1}^{\infty} A_k^* \big) = 1$. It follows that

$$\begin{aligned}
\mathrm{P}\Big(\bigcup_{t \geq \phi} \big\{ \boldsymbol{N}((t-\phi, t]) = 0 \big\} \Big) &\geq \mathrm{P}\Big(\bigcup_{k \geq 1} \big\{ \boldsymbol{N}(((k-1)\,\phi, k\,\phi]) = 0 \big\} \Big) \\
&= \mathrm{P}\Big(\bigcup_{k=1}^{\infty} A_k^* \Big) = 1, \qquad \text{(B.70)}
\end{aligned}$$

which proves (B.64).

We then prove that condition (ii) holds for $\mathcal{U}_1$ using a proof by contradiction. Let $R_1^* = \inf(\mathcal{U}_1)$. If $R_1^* \notin \mathcal{U}_1$, then there exists a sequence of time points $\{u_k\}_{k \geq 1} \in \mathcal{U}_1$ such that $u_1 > u_2 > \cdots$ and $\lim_{k \to \infty} u_k = R_1^*$. Note that $u_k \in \mathcal{U}_1$ implies that $\boldsymbol{N}(u_k) - \boldsymbol{N}(u_k - \phi) = \boldsymbol{N}((u_k - \phi, u_k]) = 0$. Using this and the fact that $\boldsymbol{N}(t)$ is right-continuous in $t \geq 0$, we have

$$\boldsymbol{N}((R_1^* - \phi, R_1^*]) \;=\; \boldsymbol{N}(R_1^*) - \boldsymbol{N}(R_1^* - \phi) \;=\; \lim_{k \to \infty} \{ \boldsymbol{N}(u_k) - \boldsymbol{N}(u_k - \phi) \} \;=\; 0,$$

which contradicts with $R_1^* \notin \mathcal{U}_1$.

Next, for each integer $\ell \geq 2$, we prove that R_ℓ exists with probability one. Following the same proof of condition (ii) for the case of $\mathcal{U}_1$, we can verify that condition (ii) also holds for $\mathcal{U}_\ell$ with integers $\ell \geq 2$. Now we prove condition (i) holds with probability one for $\mathcal{U}_\ell$

with $\ell \geq 2$. Similar to (B.64), it suffices to show that

$$\mathrm{P}\Big(\bigcup_{t \geq R_{\ell-1}+\phi} \big\{ \boldsymbol{N}((t-\phi, t]) = 0 \big\} \Big) = 1. \tag{B.71}$$

For each integer $k \geq 1$ and real number $r > 0$, define the event

$$A^*_{k,r} = \Big\{ \boldsymbol{N}\big(((k-1)\phi + r, k\phi + r]\big) = 0 \Big\} = \Big\{ \boldsymbol{N}((k-1)\phi + r) = \boldsymbol{N}(k\phi + r) \Big\}.$$

Using the same proof as that of (B.66)–(B.70), we obtain $\mathrm{P}\big(\bigcup_{k=1}^{\infty} A^*_{k,r} \big) = 1$, and

$$\begin{aligned} \mathrm{P}\Big(\bigcup_{t \geq r+\phi} \big\{ \boldsymbol{N}((t-\phi, t]) = 0 \big\} \Big) &\geq \mathrm{P}\Big(\bigcup_{k \geq 1} \big\{ \boldsymbol{N}\big(((k-1)\phi + r, k\phi + r]\big) = 0 \big\} \Big) \\ &= \mathrm{P}\Big(\bigcup_{k=1}^{\infty} A^*_{k,r} \Big) = 1. \end{aligned}$$

It follows that for each realization $R_{\ell-1} = r$, we have

$$\mathrm{P}\Big(\bigcup_{t \geq R_{\ell-1}+\phi} \big\{ \boldsymbol{N}((t-\phi, t]) = 0 \big\} \,\Big|\, R_{\ell-1} = r \Big) = 1,$$

which proves (B.71). ■

B.8. Proof of Theorem 3

B.8.1. Proof of part (i)

Recall the random processes $N_i(t)$ in (2.2), $\lambda_i(t)$ in (2.6), and $r_{i,\phi}(t)$ in (3.6). For $i \in \mathcal{V}$ and $t \geq 0$, define the following "time-shifted-by-R_ℓ" random processes:

$$\mathrm{vec}N_i(t) = N_i(t + R_\ell), \quad \mathrm{vec}\lambda_i(t) = \lambda_i(t + R_\ell), \quad \mathrm{vec}r_{i,\phi}(t) = r_{i,\phi}(t + R_\ell). \tag{B.72}$$

Let $\mathrm{vec}\boldsymbol{N}(t) = (\mathrm{vec}N_1(t), \ldots, \mathrm{vec}N_V(t))^\top$, $\mathrm{vec}\boldsymbol{\lambda}(t) = (\mathrm{vec}\lambda_1(t), \ldots, \mathrm{vec}\lambda_V(t))^\top$ and $\mathrm{vec}\boldsymbol{r}_\phi(t) = (\mathrm{vec}r_{1,\phi}(t), \ldots, \mathrm{vec}r_{V,\phi}(t))^\top$ be the vectors of these random processes. Also, for $0 \leq s < t$,

let $\text{vec}N_i((s,t]) = \text{vec}N_i(t) - \text{vec}N_i(s)$. We have the three facts below:

Fact (a) : $\text{vec}\lambda_i(t)$ is the conditional intensity function of $\text{vec}N_i(t)$. This is because

$$\begin{aligned}\text{vec}\lambda_i(t) &= \lambda_i(t+R_\ell)\\ &= \lim_{\Delta\downarrow 0}\frac{1}{\Delta}\mathrm{P}(N_i(t+R_\ell+\Delta) = N_i(t+R_\ell)+1 \mid \mathscr{F}_{t+R_\ell})\\ &= \lim_{\Delta\downarrow 0}\frac{1}{\Delta}\mathrm{P}(\text{vec}N_i(t+\Delta) = \text{vec}N_i(t)+1 \mid \text{vec}\mathscr{F}_t), \quad t\geq 0,\end{aligned}$$

agreeing with the definition in (2.6) of an intensity function, where $\text{vec}\mathscr{F}_t = \mathscr{F}_{t+R_\ell} = \{A\in\mathscr{F} : A\cap\{t+R_\ell\leq u\}\in\mathscr{F}_u \text{ for every } u>t\}$ denotes the time-shifted σ-field.

Fact (b) : $\text{vec}\lambda_i(t)$, $\text{vec}r_{i,\phi}(t)$, and $\text{vec}N_i(t)$ follow the same model as in (3.1), (3.5) and (3.6). This could be seen from the following identities:

$$\begin{aligned}\text{vec}\lambda_i(t) &= \exp\Big\{\beta_{0;i} + \sum_{j\in\mathcal{V}}\beta_{j,i}\, g\big(r_{j,\phi}(t+R_\ell)\big)\Big\}\\ &= \exp\Big\{\beta_{0;i} + \sum_{j\in\mathcal{V}}\beta_{j,i}\, g\big(\text{vec}r_{j,\phi}(t)\big)\Big\}, \quad t\geq 0,\end{aligned} \tag{B.73}$$

which is identical to (3.1) and (3.5), and

$$\begin{aligned}\text{vec}r_{i,\phi}(t) &= N_i\big(((t+R_\ell-\phi)\vee 0\,,\, t+R_\ell]\big)/\phi\\ &= N_i\big(((t+R_\ell-\phi)\vee R_\ell\,,\, t+R_\ell]\big)/\phi\\ &= \text{vec}N_i\big(((t-\phi)\vee 0\,,\, t]\big)/\phi, \qquad t\geq 0,\end{aligned} \tag{B.74}$$

which is identical to (3.6). Here, (B.74) is implied by (4.26).

Fact (c) : The following mappings are deterministic, with $\mathrm{M1} = \mathrm{M1}'$, $\mathrm{M2} = \mathrm{M2}'$, $\mathrm{M3} = \mathrm{M3}'$, $\mathrm{M4} = \mathrm{M4}'$, and $\mathrm{M5} = \mathrm{M5}'$.

(M1) from $\{\boldsymbol{N}(t+R_\ell) - \boldsymbol{N}(R_\ell)\}_{t\geq 0}$ to $\{\boldsymbol{N}(t+R_\ell) - \boldsymbol{N}((t-\phi)\vee 0 + R_\ell)\}_{t\geq 0}$.
(M2) from $\{\boldsymbol{N}(t+R_\ell) - \boldsymbol{N}(R_\ell)\}_{t\geq 0}$ to $\{\text{vec}\boldsymbol{r}_\phi(t)\}_{t\geq 0}$.
(M3) from $\{\boldsymbol{N}(t+R_\ell) - \boldsymbol{N}(R_\ell)\}_{t\geq 0}$ to $\{\text{vec}\boldsymbol{\lambda}(t)\}_{t\geq 0}$.
(M4) from $\{\boldsymbol{N}(t+R_\ell) - \boldsymbol{N}(R_\ell)\}_{t\geq 0}$ to $R_{\ell+1} - R_\ell$.
(M5) from $\{\boldsymbol{N}(t+R_\ell) - \boldsymbol{N}(R_\ell)\}_{t\geq 0}$ to $\boldsymbol{N}((R_\ell, R_{\ell+1}])$.

and

(M1′) from $\{\boldsymbol{N}(t)\}_{t\geq 0}$ to $\{\boldsymbol{N}(t) - \boldsymbol{N}((t-\phi)\vee 0)\}_{t\geq 0}$.
(M2′) from $\{\boldsymbol{N}(t)\}_{t\geq 0}$ to $\{\boldsymbol{r}_\phi(t)\}_{t\geq 0}$.
(M3′) from $\{\boldsymbol{N}(t)\}_{t\geq 0}$ to $\{\boldsymbol{\lambda}(t)\}_{t\geq 0}$.
(M4′) from $\{\boldsymbol{N}(t)\}_{t\geq 0}$ to R_1.

$(\mathrm{M5}')$ from $\{\boldsymbol{N}(t)\}_{t\geq 0}$ to $\boldsymbol{N}(R_1)$.

To show this, note that for any $t \geq 0$,

$$\begin{aligned}&\boldsymbol{N}(t+R_\ell) - \boldsymbol{N}((t-\phi)\vee 0 + R_\ell)\\ &= \{\boldsymbol{N}(t+R_\ell) - \boldsymbol{N}(R_\ell)\} - \{\boldsymbol{N}((t-\phi)\vee 0 + R_\ell) - \boldsymbol{N}(R_\ell)\}\end{aligned}$$

implies that the mapping $\mathrm{M1}$ is deterministic, and $\mathrm{M1} = \mathrm{M1}'$. This combined with (B.74) and (3.6) gives that the mappings $\mathrm{M2} = \mathrm{M2}'$ are deterministic. By using (B.73), (3.1), (3.5) and the fact that the mappings $\mathrm{M2} = \mathrm{M2}'$ are deterministic, we have that the mappings $\mathrm{M3} = \mathrm{M3}'$ are deterministic. From (4.25), we observe that

$$\begin{aligned}R_1 &= \min\{t \geq \phi : \boldsymbol{N}(t) - \boldsymbol{N}((t-\phi)\vee 0) = 0\},\\ R_{\ell+1} - R_\ell &= \min\{t \geq \phi : \boldsymbol{N}(t+R_\ell) - \boldsymbol{N}((t-\phi)\vee 0 + R_\ell) = 0\}.\end{aligned}$$

This together with the fact that $\mathrm{M1} = \mathrm{M1}'$ are deterministic yields that the mappings $\mathrm{M4} = \mathrm{M4}'$ are deterministic. From

$$\boldsymbol{N}((R_\ell, R_{\ell+1}]) = \boldsymbol{N}((R_{\ell+1} - R_\ell) + R_\ell) - \boldsymbol{N}(R_\ell),$$

we have that the mapping from $(\{\boldsymbol{N}(t+R_\ell) - \boldsymbol{N}(R_\ell)\}_{t\geq 0},\ R_{\ell+1} - R_\ell)$ to $\boldsymbol{N}((R_\ell, R_{\ell+1}])$ is deterministic, and identical to that from $(\{\boldsymbol{N}(t)\}_{t\geq 0},\ R_1)$ to $\boldsymbol{N}(R_1)$. This together with $\mathrm{M4} = \mathrm{M4}'$ yields that $\mathrm{M5} = \mathrm{M5}'$ are deterministic.

Fact (a) verifies that vec$\boldsymbol{\lambda}(t)$ is the vector of intensity processes of vec$\boldsymbol{N}(t)$. From Fact (c), we have that the deterministic mappings $\mathrm{M3} = \mathrm{M3}'$. Combining these yields that $\big(\boldsymbol{N}(t+R_\ell) - \boldsymbol{N}(R_\ell), \mathrm{vec}\boldsymbol{\lambda}(t)\big) \overset{\mathrm{D}}{=} \big(\boldsymbol{N}(t), \boldsymbol{\lambda}(t)\big)$ for any $t \geq 0$, which proves $\boldsymbol{N}(t+R_\ell) - \boldsymbol{N}(R_\ell) \overset{\mathrm{D}}{=} \boldsymbol{N}(t)$. On the other hand, using condition $\mathrm{C1}$ and the fact that the mapping $\mathrm{M3}$ is deterministic, we have that vec$\boldsymbol{\lambda}(t)$ is independent of $\mathcal{F}_{R_\ell}$. This completes the proof of part (i). ■

B.8.2. Proof of part (ii)

Using the facts that the mappings $\mathrm{M5} = \mathrm{M5}'$ are deterministic and that $\boldsymbol{N}(t+R_\ell) - \boldsymbol{N}(R_\ell) \overset{\mathrm{D}}{=} \boldsymbol{N}(t)$ for any $t \geq 0$ from Theorem 3 (i), we obtain $\boldsymbol{N}((R_\ell, R_{\ell+1}]) \overset{\mathrm{D}}{=} \boldsymbol{N}(R_1)$. Thus, $\{\boldsymbol{N}((R_\ell, R_{\ell+1}])\}_{\ell\geq 0}$ is a sequence of identically distributed random vectors. On the other hand, combining condition $\mathrm{C1}$ and the fact that the mapping $\mathrm{M5}$ is deterministic, we have that $\boldsymbol{N}((R_\ell, R_{\ell+1}])$ is independent of $\mathcal{F}_{R_\ell}$. Also, $\boldsymbol{N}((R_k, R_{k+1}])$ is $\mathcal{F}_{R_\ell}$-measurable for any

$0 \le k \le \ell - 1$. It follows that $\boldsymbol{N}((R_\ell, R_{\ell+1}])$ is independent of $\{\boldsymbol{N}((R_k, R_{k+1}])\}_{0 \le k \le \ell-1}$. Hence, $\{\boldsymbol{N}((R_\ell, R_{\ell+1}])\}_{\ell \ge 0}$ is a sequence of independent random vectors. The proof is completed. ■

B.8.3. Proof of part (iii)

Before proving part (iii), we first show Lemmas B.10 and B.11.

Lemma B.10. *Assume conditions* A1–A4 *and* A2′ *in Appendix A. Let* R_1 *be the first "recurrence time point" defined in* (4.25) *with* $\ell = 1$. *Then* $\mathrm{E}(R_1^2) < \infty$.

Proof: Let the random variable $L^* = \min\{k \ge 1 : \mathrm{I}(A_k^*) = 1\}$, where the event $A_k^* = \big\{\boldsymbol{N}\big(((k-1)\,\phi, k\,\phi]\big) = 0\big\}$ is defined in (B.66). By (B.68), for any integer $\ell > 2$, we have

$$\mathrm{P}(L^* \ge \ell) = \mathrm{P}\Big(\bigcap_{k=1}^{\ell-1} \overline{A_k^*}\Big) \le \big\{1 - \exp(-c\,\phi)\big\}^{\ell-1}.$$

Thus, the tail probability of L^* is bounded by that of the geometric distribution with the constant success probability $\exp(-c\,\phi)$. Since geometric distribution has finite first and second moments, we have $\mathrm{E}\{(L^*)^2\} < \infty$. Also, $\mathrm{I}(A_{L^*}^*) = 1$ implies that $\boldsymbol{N}(((L^*-1)\,\phi, L^*\,\phi]) = 0$, which combined with (4.25) gives that $R_1 \le L^*\,\phi$. Hence, we obtain

$$\mathrm{E}(R_1^2) \le \mathrm{E}\{(L^*\,\phi)^2\} = \phi^2\,\mathrm{E}\{(L^*)^2\} < \infty.$$

This completes the proof. ■

Lemma B.11. *Assume conditions* A1–A4, A2′, *and* C1 *in Appendix A. Let* $\boldsymbol{\lambda}(t) = (\lambda_1(t), \ldots, \lambda_V(t))^\top$ *be the vector of the intensity processes in* (3.1), *and* $\boldsymbol{r}_\phi(t) = (r_{1,\phi}(t), \ldots, r_{V,\phi}(t))^\top$ *be the vector of the empirical firing rates defined in* (3.6). *Let* $h(\cdot) : \mathbb{R}^{2V} \to \mathbb{R}$ *be a continuous function. For each integer* $\ell \ge 1$, *define the random variable*

$$\mathrm{S}_\ell = \int_{R_{\ell-1}}^{R_\ell} h(\boldsymbol{\lambda}(t), \boldsymbol{r}_\phi(t))\,\mathrm{d}t. \tag{B.75}$$

Then $\{\mathrm{S}_\ell\}_{\ell\geq 1}$ *is a sequence of* i.i.d. *random variables.*

Proof: Following the notations in (B.72), for $\ell \geq 1$, we have

$$\mathrm{S}_{\ell+1} = \int_{R_\ell}^{R_{\ell+1}} h(\boldsymbol{\lambda}(t), \boldsymbol{r}_\phi(t))\,\mathrm{d}t = \int_0^{R_{\ell+1}-R_\ell} h(\mathrm{vec}\boldsymbol{\lambda}(t), \mathrm{vec}\boldsymbol{r}_\phi(t))\,\mathrm{d}t. \tag{B.76}$$

By comparing (B.76) with $\mathrm{S}_1 = \int_0^{R_1} h(\boldsymbol{\lambda}(t), \boldsymbol{r}_\phi(t))\,\mathrm{d}t$, and using the fact that the mappings $\mathrm{M2} = \mathrm{M2}'$, $\mathrm{M3} = \mathrm{M3}'$ and $\mathrm{M4} = \mathrm{M4}'$ are deterministic, we have that the mappings

$$(\mathrm{M6}) \text{ from } \{\boldsymbol{N}(t+R_\ell) - \boldsymbol{N}(R_\ell)\}_{t\geq 0} \text{ to } \mathrm{S}_{\ell+1},$$
$$(\mathrm{M6}') \text{ from } \{\boldsymbol{N}(t)\}_{t\geq 0} \text{ to } \mathrm{S}_1,$$

are both deterministic, with $\mathrm{M6} = \mathrm{M6}'$. Combining this with the fact that $\boldsymbol{N}(t+R_\ell) - \boldsymbol{N}(R_\ell) \overset{\mathrm{D}}{=} \boldsymbol{N}(t)$ for any $t \geq 0$ from Theorem 3 (i), we obtain $\mathrm{S}_{\ell+1} \overset{\mathrm{D}}{=} \mathrm{S}_1$. Thus, $\{\mathrm{S}_\ell\}_{\ell\geq 1}$ is a sequence of identically distributed random variables. On the other hand, using condition C1 and the fact that the mapping M6 is deterministic, we have that $\mathrm{S}_{\ell+1}$ is independent of $\mathcal{F}_{R_\ell}$. Also, S_k is $\mathcal{F}_{R_\ell}$-measurable for $1 \leq k \leq \ell$. Hence, $\{\mathrm{S}_\ell\}_{\ell\geq 1}$ is a sequence of independent variables. The proof is completed. ■

Now we prove part (iii) of Theorem 3. Note that $D_\ell = R_\ell - R_{\ell-1}$ is a special case of S_ℓ in (B.75) with $h(\cdot) \equiv 1$. Applying Lemma B.11, we conclude that $\{D_\ell\}_{\ell\geq 1}$ is a sequence of i.i.d. random variables. Furthermore, Lemma B.10 proved that $D_1 = R_1$ has the finite second moment. This completes the proof. ■

B.9. Proof of Theorem 4

Before proving Theorem 4, we first show Lemma B.12.

Lemma B.12 (Asymptotic convergence of $\Lambda_i(t) = \int_0^t \lambda_i(u)\,\mathrm{d}u$)**.** *Assume conditions* A1–A4, A2′, *and* C1 *in Appendix A. For each* $i \in \mathcal{V}$*, consider the random process* $\Lambda_i(t) = \int_0^t \lambda_i(u)\,\mathrm{d}u$ *for*

$t > 0$. *Then there exists a constant* $c_i \in (0, \infty)$, *such that*

$$\frac{\Lambda_i(t)}{t} \xrightarrow{\mathrm{P}} c_i, \quad \textit{as } t \to \infty. \tag{B.77}$$

Proof: Denote the increment of $\Lambda_i(t)$ in the ℓth "recurrence cycle" $(R_{\ell-1}, R_\ell]$ by

$$\mathrm{S}_{i,\ell} \;=\; \Lambda_i(R_\ell) - \Lambda_i(R_{\ell-1}) \;=\; \int_{R_{\ell-1}}^{R_\ell} \lambda_i(t)\,\mathrm{d}t.$$

Applying $h(\boldsymbol{\lambda}(t), \boldsymbol{r}_\phi(t)) = \lambda_i(t)$ to (B.75) in Lemma B.11 indicates that $\{\mathrm{S}_{i,\ell}\}_{\ell \geq 1}$ is a sequence of i.i.d. random variables.

By condition A4, there exists a constant $c \in (0, \infty)$, such that

$$\mathrm{S}_{i,\ell} \;=\; \int_{R_{\ell-1}}^{R_\ell} \lambda_i(t)\,\mathrm{d}t \;\leq\; \int_{R_{\ell-1}}^{R_\ell} c\,\mathrm{d}t \;=\; c\,D_\ell, \tag{B.78}$$

where $D_\ell = R_\ell - R_{\ell-1}$. Combining Lemma B.10 and (B.78) implies that the second moments of D_ℓ and $\mathrm{S}_{i,\ell}$ are finite. Applying the strong law of large numbers, we obtain

$$\frac{1}{\ell}\sum_{k=1}^{\ell} \mathrm{S}_{i,k} \xrightarrow{\text{a.s.}} \mathrm{E}(\mathrm{S}_{i,1}), \quad \frac{1}{\ell}\sum_{k=1}^{\ell} D_k \xrightarrow{\text{a.s.}} \mathrm{E}(D_1), \quad \text{as } \ell \to \infty.$$

Thus, for arbitrarily small $\delta > 0$ and $\epsilon > 0$, we could find sufficiently large C_1, such that

$$\mathrm{P}\Big(\sup_{\ell > C_1} \max\Big\{ \Big| \frac{1}{\ell}\sum_{k=1}^{\ell} \mathrm{S}_{i,k} - \mathrm{E}(\mathrm{S}_{i,1}) \Big|, \Big| \frac{1}{\ell}\sum_{k=1}^{\ell} D_k - \mathrm{E}(D_1) \Big| \Big\} > \epsilon \Big) < \delta. \tag{B.79}$$

On the other hand, for any time point $t > 0$, let $L_t = \sup\{\ell \geq 0 : R_\ell \leq t\}$ be the number of recurrence time points up to t. We have

$$\sum_{k=1}^{L_t} \mathrm{S}_{i,k} \leq \Lambda_i(t) \leq \sum_{k=1}^{L_t+1} \mathrm{S}_{i,k}, \quad \sum_{k=1}^{L_t} D_k \leq t \leq \sum_{k=1}^{L_t+1} D_k,$$

which directly yields that

$$\frac{\sum_{k=1}^{L_t} \mathrm{S}_{i,k}}{\sum_{k=1}^{L_t+1} D_k} \leq \frac{\Lambda_i(t)}{t} \leq \frac{\sum_{k=1}^{L_t+1} \mathrm{S}_{i,k}}{\sum_{k=1}^{L_t} D_k}. \tag{B.80}$$

Since $\mathrm{E}(D_1^2) < \infty$, it is easy to show that $L_t \xrightarrow{\mathrm{P}} \infty$ as $t \to \infty$. Thus for arbitrarily small $\delta_2 > 0$ and large $C_2 > C_1$, there exists $t_0 > 0$, such that for $\forall t > t_0$, $\mathrm{P}(L_t > C_2) > 1 - \delta_2$. Combining (B.79) and (B.80), the following (B.81) holds with probability at least $1 - \delta - \delta_2$ for $t > t_0$:

$$\frac{C_2 \left\{\mathrm{E}(\mathrm{S}_{i,1}) - \epsilon\right\}}{(C_2+1) \left\{\mathrm{E}(D_1) + \epsilon\right\}} \leq \frac{\Lambda_i(t)}{t} \leq \frac{(C_2+1) \left\{\mathrm{E}(\mathrm{S}_{i,1}) + \epsilon\right\}}{C_2 \left\{\mathrm{E}(D_1) - \epsilon\right\}}. \tag{B.81}$$

Since ϵ, δ, and δ_2 are arbitrarily small and C_2 is arbitrarily large, (B.81) implies that

$$\frac{\Lambda_i(t)}{t} \xrightarrow{\mathrm{P}} \frac{\mathrm{E}(\mathrm{S}_{i,1})}{\mathrm{E}(D_1)}, \quad \text{as } t \to \infty.$$

We complete the proof with $c_i = \mathrm{E}(\mathrm{S}_{i,1})/\mathrm{E}(D_1)$ in (B.77). ■

Now we prove Theorem 4. By Theorem 2, for each $i \in \mathcal{V}$ and $t \in (0, \infty)$, we have

$$\mathrm{var}\Big\{\frac{N_i(t) - \Lambda_i(t)}{t}\Big\} \leq \frac{c_1}{t},$$

which together with Lemma 5 implies that

$$\frac{N_i(t) - \Lambda_i(t)}{t} \xrightarrow{\mathrm{P}} 0, \quad \text{as } t \to \infty. \tag{B.82}$$

By Lemma B.12 and (B.82), we obtain

$$\frac{\boldsymbol{N}(t)}{t} \xrightarrow{\mathrm{P}} \boldsymbol{c}_0, \quad \text{as } t \to \infty,$$

where $\boldsymbol{c}_0 = (\mathrm{E}(\mathrm{S}_{1,1}), \ldots, \mathrm{E}(\mathrm{S}_{V,1}))^\top / \mathrm{E}(D_1)$. This completes the proof. ■

B.10. Proof of Theorem 5

Note that all the previous proofs in Theorems 1–4 and Lemmas 2–5 rely only on the fact that (i) $\lambda_i(t)$ is a positive continuous function of $\boldsymbol{N}\big(((t-\phi)\vee 0, t]\big)$, and (ii) "$\sup_{i\in\mathcal{V},\, t\geq 0} \lambda_i(t) \leq c_3$" as in condition A4. Clearly, (i) does not depend on the specific form of the function, and $\sup_{\boldsymbol{x}\in\mathbb{R}^V} h_i(\boldsymbol{x}) \leq c$ in condition B1 verifies the assumption (ii). Thus extending model (3.1) to (4.28) does not alter the proofs of Theorems 1–4 and Lemmas 2–5. This completes the proof of Theorem 5. ■

B.11. Proof of Theorem 6

Before proving Theorem 6, we first show Lemmas B.13, B.14 and B.15.

Lemma B.13. *Assume conditions* A1–A4, A2′, *and* C1 *in Appendix A. Let* $f(\cdot) : \mathbb{R}^V \to [0,\infty)$ *be a non-negative continuous function bounded above by* $c_0 \in (0,\infty)$. *For* $t \geq 0$, *let* $Y(t) = f(\boldsymbol{r}_\phi(t))$, *where* $\boldsymbol{r}_\phi(t)$ *is defined in Lemma* $B.11$. *Let* R_1 *be the first recurrence time point defined in* (4.25) *with* $\ell = 1$. *Assume that* $\mathrm{E}\{\int_0^{R_1} Y(t)\,\mathrm{d}t\} > 0$. *Then there exists a constant* $c_i \in (0,\infty)$, *such that*

$$\frac{\int_0^t \lambda_i(u)\, Y(u)\,\mathrm{d}u}{t} \xrightarrow{\mathrm{P}} c_i, \quad \text{as } t \to \infty.$$

Proof: For each integer $\ell \geq 1$, define

$$\mathrm{S}^*_{i,\ell} = \int_{R_{\ell-1}}^{R_\ell} \lambda_i(t)\, Y(t)\,\mathrm{d}t.$$

Applying Lemma B.11 with $h(\boldsymbol{\lambda}(t), \boldsymbol{r}_\phi(t)) = \lambda_i(t) f(\boldsymbol{r}_\phi(t)) = \lambda_i(t)\, Y(t)$, we have that $\{\mathrm{S}^*_{i,\ell}\}_{\ell\geq 1}$ is a sequence of i.i.d. random variables. By condition A4, there exist constants $c_2, c_3 \in (0,\infty)$ such that $c_2 \leq \lambda_i(t) \leq c_3$ for any $t \in [0,\infty)$. We obtain the following moment inequalities:

$$\begin{aligned} \mathrm{E}(\mathrm{S}^*_{i,1}) &= \mathrm{E}\Big\{ \int_0^{R_1} \lambda_i(t)\, Y(t)\,\mathrm{d}t \Big\} \;\geq\; c_2\, \mathrm{E}\Big\{ \int_0^{R_1} Y(t)\,\mathrm{d}t \Big\} \;>\; 0, \qquad \text{(B.83)} \\ \mathrm{E}\big\{(\mathrm{S}^*_{i,1})^2\big\} &= \mathrm{E}\Big[\Big\{ \int_0^{R_1} \lambda_i(t)\, Y(t)\,\mathrm{d}t \Big\}^2 \Big] \;\leq\; c_3^2\, c_0^2\, \mathrm{E}(R_1^2) \;<\; \infty. \end{aligned}$$

Applying the same proof as Lemma B.12 with $\mathrm{S}_{i,\ell} = \mathrm{S}^*_{i,\ell}$, one can show that

$$\frac{\int_0^t \lambda_i(u)\, Y(u)\, \mathrm{d}u}{t} \xrightarrow{\mathrm{P}} \frac{\mathrm{E}\{\int_0^{R_1} \lambda_i(u)\, Y(u)\, \mathrm{d}u\}}{\mathrm{E}(R_1)} = \frac{\mathrm{E}(\mathrm{S}^*_{i,1})}{\mathrm{E}(R_1)} > 0, \quad \text{as } t \to \infty.$$

This completes the proof. ■

Lemma B.14. *Assume conditions* A1–A4 *and* A2′ *in Appendix A. For $i \in \mathcal{V}$, let $x_i(t) = g\big(r_{i,\phi}(t)\big)$ be the covariate defined in* (3.5)*, and $x_0(t) \equiv 1$. Then for any $i, j \in \mathcal{V} \cup \{0\}$ (not necessarily distinct), we have*

$$\mathrm{E}\Big\{ \int_0^{R_1} x_i(u)\, \mathrm{d}u \Big\} > 0, \tag{B.84}$$

$$\mathrm{E}\Big\{ \int_0^{R_1} x_i(u)\, x_j(u)\, \mathrm{d}u \Big\} > 0. \tag{B.85}$$

Proof: If $i = 0$, then (B.84) obviously holds; if either i or j is zero, then (B.85) reduces to (B.84). Thus, to prove Lemma B.14, it suffices to verify (B.84) and (B.85) for the case of $i, j \in \mathcal{V}$.

By the similar proof to that below (B.62), for any $t > 0$, we have

$$\mathrm{P}(N_i(t) \geq 1) \;=\; 1 - \mathrm{P}(N_i(t) = 0) \;=\; 1 - \exp\{-\lambda_i(0) \cdot t\} \;>\; 0. \tag{B.86}$$

The conditional independence condition (2.10) implies that

$$\mathrm{P}(N_i(t)\, N_j(t) \geq 1) \;=\; \lambda_i(0)\, \lambda_j(0)\, t^2 + o(t^2),$$

as $t \to 0$, and thus there exists $t_0 \in (0, \phi)$, such that

$$\mathrm{P}(N_i(t_0)\, N_j(t_0) \geq 1) \;>\; 0. \tag{B.87}$$

For $t \in (0, \phi)$, we observe that $r_{i,\phi}(t) = N_i\big(((t - \phi) \vee 0, t]\big)/\phi = N_i(t)/\phi$. Hence, $r_{i,\phi}(t)$ is increasing in $t \in (0, \phi)$, which implies that $x_i(t) = g\big(r_{i,\phi}(t)\big)$ is also increasing in $t \in (0, \phi)$.

Together with (B.86), (B.87) and the fact that $R_1 \geq \phi > t_0$, we obtain

$$\begin{aligned}
\mathrm{E}\Big\{ \int_0^{R_1} x_i(u)\,\mathrm{d}u \Big\} &\geq \mathrm{E}\Big\{ \int_{t_0}^{\phi} x_i(u)\,\mathrm{d}u \Big\} \\
&\geq \mathrm{E}\{x_i(t_0)\cdot(\phi - t_0)\} \\
&\geq \mathrm{P}(N_i(t_0)\geq 1)\, g(1/\phi)\cdot(\phi - t_0) \\
&> 0, \\
\mathrm{E}\Big\{ \int_0^{R_1} x_i^2(u)\,\mathrm{d}u \Big\} &\geq \mathrm{E}\Big\{ \int_{t_0}^{\phi} x_i^2(u)\,\mathrm{d}u \Big\} \\
&\geq \mathrm{E}\{x_i^2(t_0)\cdot(\phi - t_0)\} \\
&\geq \mathrm{P}(N_i(t_0)\geq 1)\, g^2(1/\phi)\cdot(\phi - t_0) \\
&> 0, \\
\mathrm{E}\Big\{ \int_0^{R_1} x_i(u)\,x_j(u)\,\mathrm{d}u \Big\} &\geq \mathrm{E}\Big\{ \int_{t_0}^{\phi} x_i(u)\,x_j(u)\,\mathrm{d}u \Big\} \\
&\geq \mathrm{E}\{x_i(t_0)\,x_j(t_0)\cdot(\phi - t_0)\} \\
&\geq \mathrm{P}(N_i(t_0)\,N_j(t_0)\geq 1)\, g^2(1/\phi)\cdot(\phi - t_0) \\
&> 0.
\end{aligned}$$

These complete the proof. ■

Lemma B.15. *Assume conditions* A1–A4 *and* A2′ *in Appendix A. Let* $\widetilde{\boldsymbol{x}}(t) = (1, x_1(t), x_2(t), \ldots, x_V(t))^\top$ *be the vector of the covariates defined in* (3.5)*. Then for any* $\widetilde{\boldsymbol{u}} \in \mathbb{R}^{V+1}$ *with* $\|\widetilde{\boldsymbol{u}}\| > 0$*,*

$$\mathrm{E}\Big[\int_0^{R_1} \big\{ \widetilde{\boldsymbol{x}}(t)^\top \widetilde{\boldsymbol{u}} \big\}^2 \,\mathrm{d}t \Big] > 0.$$

Proof: Let $\widetilde{\boldsymbol{u}} = (u_0, u_1, \ldots, u_V)^\top$. We proceed by cases of u_0.

Case (i) $u_0 \neq 0$.
Consider $t_0 \in (0, \phi)$. By (B.65), we have

$$\mathrm{P}(\boldsymbol{N}(t_0) = 0) \geq \exp\{-c\cdot(t_0 - 0)\} > 0. \tag{B.88}$$

Note that $\boldsymbol{N}(t_0) = 0$ implies that $\widetilde{\boldsymbol{x}}(t)^\top \widetilde{\boldsymbol{u}} = u_0$ for $t \in [0, t_0]$. Combining (B.88) and the fact that $R_1 \geq \phi > t_0$, we obtain

$$\begin{aligned}\mathrm{E}\Big[\int_0^{R_1} \big\{\widetilde{\boldsymbol{x}}(t)^\top \widetilde{\boldsymbol{u}}\big\}^2 \,\mathrm{d}t\Big] &\geq \mathrm{E}\Big[\int_0^{t_0} \big\{\widetilde{\boldsymbol{x}}(t)^\top \widetilde{\boldsymbol{u}}\big\}^2 \cdot \mathrm{I}(\boldsymbol{N}(t_0) = 0) \,\mathrm{d}t\Big] \\ &= t_0\, u_0^2\, \mathrm{P}(\boldsymbol{N}(t_0) = 0) > 0.\end{aligned}$$

Case (ii) $u_0 = 0$.

Since $\|\widetilde{\boldsymbol{u}}\| > 0$ and $u_0 = 0$, there exists some $i \in \mathcal{V}$ such that $u_i \neq 0$. We have

$$\begin{aligned}&\mathrm{P}(N_j(t) = 0 \text{ for all } j \in \mathcal{V} \setminus i, \quad N_i(t) = 1) \\ &\geq \mathrm{P}(N_i(t) = 1) - \sum_{j \in \mathcal{V}\setminus i} \mathrm{P}(N_i(t) = 1, N_j(t) \neq 0) \\ &\geq \mathrm{P}(N_i(t) = 1) - \sum_{j \in \mathcal{V}\setminus i} \mathrm{P}(N_i(t) = 1, N_j(t) = 1) - \sum_{j \in \mathcal{V}\setminus i} \mathrm{P}(N_j(t) > 1) \\ &= \lambda_i(0)\, t + o(t) - \sum_{j \in \mathcal{V}\setminus i} \{\lambda_i(0)\, \lambda_j(0)\, t^2 + o(t^2)\} - o(t) \qquad \text{(B.89)} \\ &= \lambda_i(0)\, t + o(t),\end{aligned}$$

as $t \to 0$, where (B.89) is derived from (2.6), (2.7), and (2.10). Hence, there exists $t_0 \in (0, \phi)$, such that

$$\mathrm{P}(N_j(t_0) = 0 \text{ for all } j \in \mathcal{V} \setminus i,\ N_i(t_0) = 1) > 0. \qquad \text{(B.90)}$$

Let $t_1 \in (t_0, \phi)$. From (B.65) and (B.90), we have

$$\begin{aligned}&\mathrm{P}(N_j(t_1) = 0 \text{ for all } j \in \mathcal{V} \setminus i,\ N_i(t_0) = N_i(t_1) = 1) \\ &= \mathrm{P}(\boldsymbol{N}(t_1) = \boldsymbol{N}(t_0),\ N_j(t_0) = 0 \text{ for all } j \in \mathcal{V} \setminus i,\ N_i(t_0) = 1) \\ &= \mathrm{P}(\boldsymbol{N}(t_1) = \boldsymbol{N}(t_0) \mid N_j(t_0) = 0 \text{ for all } j \in \mathcal{V} \setminus i,\ N_i(t_0) = 1) \\ &\quad \times \mathrm{P}(N_j(t_0) = 0 \text{ for all } j \in \mathcal{V} \setminus i,\ N_i(t_0) = 1) \\ &\geq \exp\{-c \cdot (t_1 - t_0)\} \cdot \mathrm{P}(N_j(t_0) = 0 \text{ for all } j \in \mathcal{V} \setminus i,\ N_i(t_0) = 1) \\ &> 0. \qquad \text{(B.91)}\end{aligned}$$

Note that the event $\{N_j(t_1) = 0 \text{ for all } j \in \mathcal{V} \setminus i, \text{ and } N_i(t_0) = N_i(t_1) = 1\}$ implies that $\boldsymbol{x}(t)^\top \widetilde{\boldsymbol{u}} = x_i(t)\, u_i = g(1/\phi)\, u_i$ for $t \in (t_0, t_1)$. Together with (B.91) and the fact that $R_1 \geq \phi > t_1$, we obtain

$$\begin{aligned}&\mathrm{E}\Big[\int_0^{R_1} \{\widetilde{\boldsymbol{x}}(t)^\top \widetilde{\boldsymbol{u}}\}^2 \,\mathrm{d}t\Big] \\ &\geq \mathrm{E}\Big[\int_{t_0}^{t_1} \{\widetilde{\boldsymbol{x}}(t)^\top \widetilde{\boldsymbol{u}}\}^2 \cdot \mathrm{I}\big(N_j(t_1) = 0 \text{ for all } j \in \mathcal{V} \setminus i,\ N_i(t_0) = N_i(t_1) = 1\big) \,\mathrm{d}t\Big] \\ &\geq (t_1 - t_0)\, g^2(1/\phi)\, u_i^2 \cdot \mathrm{P}\big(N_j(t_1) = 0 \text{ for all } j \in \mathcal{V} \setminus i,\ N_i(t_0) = N_i(t_1) = 1\big) \\ &> 0.\end{aligned}$$

Combining the results of cases (i) and (ii) completes the proof. ■

Now we prove Theorem 6. Recall (5.4),

$$\mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i) = \frac{1}{\mathrm{T}} \int_0^{\mathrm{T}} \Big[\exp\big\{\widetilde{\boldsymbol{x}}_i(t)^\top \widetilde{\boldsymbol{\beta}}_i\big\}\, \mathrm{d}t - \{\widetilde{\boldsymbol{x}}_i(t-)^\top \widetilde{\boldsymbol{\beta}}_i\}\, \mathrm{d}N_i(t) \Big].$$

With some algebra, we obtain the gradient vector and Hessian matrix of $\mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i)$:

$$\nabla \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i) = \frac{1}{\mathrm{T}} \int_0^{\mathrm{T}} \Big[\widetilde{\boldsymbol{x}}_i(t)\, \exp\big\{\widetilde{\boldsymbol{x}}_i(t)^\top \widetilde{\boldsymbol{\beta}}_i\big\}\, \mathrm{d}t - \widetilde{\boldsymbol{x}}_i(t-)\, \mathrm{d}N_i(t) \Big], \tag{B.92}$$

$$\nabla^2 \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i) = \frac{1}{\mathrm{T}} \int_0^{\mathrm{T}} \widetilde{\boldsymbol{x}}_i(t)\, \widetilde{\boldsymbol{x}}_i(t)^\top \exp\big\{\widetilde{\boldsymbol{x}}_i(t)^\top \widetilde{\boldsymbol{\beta}}_i\big\}\, \mathrm{d}t. \tag{B.93}$$

Let $\widetilde{x}_{i,j}(t)$ denote the jth component of $\widetilde{\boldsymbol{x}}_i(t)$, i.e.,

$$\widetilde{x}_{i,j}(t) = \begin{cases} 1, & \text{if } j = 1, \\ x_{j-1}(t), & \text{if } 1 < j \le i, \\ x_j(t), & \text{if } i < j \le V. \end{cases} \tag{B.94}$$

For each $j \in \mathcal{V}$, applying (4.19) in Lemma 5 gives that

$$\mathrm{E}\Big(\frac{1}{\mathrm{T}} \int_0^{\mathrm{T}} \Big[\widetilde{x}_{i,j}(t)\, \exp\big\{\widetilde{\boldsymbol{x}}_i(t)^\top \widetilde{\boldsymbol{\beta}}_i^*\big\}\, \mathrm{d}t - \widetilde{x}_{i,j}(t-)\, \mathrm{d}N_i(t) \Big] \Big) = 0. \tag{B.95}$$

Also, using (4.24) in Theorem 2, we have

$$\mathrm{var}\Big(\frac{1}{\mathrm{T}} \int_0^{\mathrm{T}} \Big[\widetilde{x}_{i,j}(t)\, \exp\big\{\widetilde{\boldsymbol{x}}_i(t)^\top \widetilde{\boldsymbol{\beta}}_i^*\big\}\, \mathrm{d}t - \widetilde{x}_{i,j}(t-)\, \mathrm{d}N_i(t) \Big] \Big) \le c_1/\mathrm{T}, \tag{B.96}$$

where $c_1 \in (0, \infty)$ is some constant. By Chebyshev's inequality, (B.95) and (B.96), we have

$$\frac{1}{\mathrm{T}} \int_0^{\mathrm{T}} \Big[\widetilde{\boldsymbol{x}}_i(t) \exp\big\{\widetilde{\boldsymbol{x}}_i(t)^\top \widetilde{\boldsymbol{\beta}}_i^*\big\}\, \mathrm{d}t - \widetilde{\boldsymbol{x}}_i(t-)\, \mathrm{d}N_i(t) \Big] = O_{\mathrm{P}}(\sqrt{1/\mathrm{T}}). \tag{B.97}$$

This verifies (5.5) in part (i).

Next we prove part (ii). Write $\mathbf{H}_{i,\mathrm{T}} = \nabla^2 \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i^*)$. For any $j, k \in \mathcal{V}$, from (B.93) and

(B.94), the (j,k)th entry of $\mathbf{H}_{i,\mathrm{T}}$ is

$$\mathbf{H}_{i,\mathrm{T}}(j,k) = \frac{1}{\mathrm{T}} \int_0^{\mathrm{T}} \widetilde{x}_{i,j}(t)\, \widetilde{x}_{i,k}(t)\, \exp\left\{\widetilde{\boldsymbol{x}}_i(t)^\top \widetilde{\boldsymbol{\beta}}_i^*\right\} \mathrm{d}t.$$

Let $Y_1(t) = \widetilde{x}_{i,j}(t)\, \widetilde{x}_{i,k}(t)$. Lemma B.14 proved that $\mathrm{E}\{\int_0^{R_1} Y_1(t)\, \mathrm{d}t\} > 0$. This enables us to apply Lemma B.13 with $Y(t) = Y_1(t)$, which yields that

$$\mathbf{H}_{i,\mathrm{T}}(j,k) = \frac{1}{\mathrm{T}} \int_0^{\mathrm{T}} \widetilde{x}_{i,j}(t)\, \widetilde{x}_{i,k}(t)\, \lambda_i^*(t)\, \mathrm{d}t \xrightarrow{\mathrm{P}} c_{j,k}, \quad \text{as } \mathrm{T} \to \infty,$$

where $c_{j,k} \in (0,\infty)$ is some constant. Denote by $\mathbf{C}_i = (c_{j,k})_{V\times V}$ the matrix consisting of the entries $\{c_{j,k} : j \in \mathcal{V}, k \in \mathcal{V}\}$. It follows that all entries in $\mathbf{C}_i$ are positive, and

$$\nabla^2 \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i^*) = \mathbf{H}_{i,\mathrm{T}} \xrightarrow{\mathrm{P}} \mathbf{C}_i, \quad \text{as } \mathrm{T} \to \infty. \tag{B.98}$$

This proves the asymptotic convergence in (5.6).

We next show $\mathbf{C}_i \succ 0$ using a proof by contradiction. If $\mathbf{C}_i$ is not positive definite, then there exists a vector $\widetilde{\boldsymbol{u}}$ with $\|\widetilde{\boldsymbol{u}}\| > 0$, such that $\widetilde{\boldsymbol{u}}^\top \mathbf{C}_i \widetilde{\boldsymbol{u}} \le 0$. Then (B.98) implies that

$$\widetilde{\boldsymbol{u}}^\top \mathbf{H}_{i,\mathrm{T}} \widetilde{\boldsymbol{u}} \xrightarrow{\mathrm{P}} \widetilde{\boldsymbol{u}}^\top \mathbf{C}_i \widetilde{\boldsymbol{u}} \le 0, \quad \text{as } \mathrm{T} \to \infty. \tag{B.99}$$

Let $Y_2(t) = \{\widetilde{\boldsymbol{x}}_i(t)^\top \widetilde{\boldsymbol{u}}\}^2$. Lemma B.15 verifies that $\mathrm{E}\{\int_0^{R_1} Y_2(t)\, \mathrm{d}t\} > 0$. Form Lemma B.13, there exists a constant $c_i \in (0,\infty)$ such that

$$\widetilde{\boldsymbol{u}}^\top \mathbf{H}_{i,\mathrm{T}} \widetilde{\boldsymbol{u}} = \frac{1}{\mathrm{T}} \int_0^{\mathrm{T}} \{\widetilde{\boldsymbol{x}}_i(t)^\top \widetilde{\boldsymbol{u}}\}^2\, \lambda_i^*(t)\, \mathrm{d}t \xrightarrow{\mathrm{P}} c_i > 0, \quad \text{as } \mathrm{T} \to \infty, \tag{B.100}$$

which contradicts (B.99). The proof is completed. ■

B.12. Proof of Theorem 7

Before proving Theorem 7, we first show Lemma B.16.

Lemma B.16 (Consistency of M-estimator)**.** *Assume conditions* A1–A6, A2′, *and* C1 *in Appendix A. As* $\mathrm{T} \to \infty$*, there exists a local minimizer* $\widehat{\widetilde{\boldsymbol{\beta}}}_i$ *of the loss function* $\mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i)$ *in* (5.4) *such that* $\|\widehat{\widetilde{\boldsymbol{\beta}}}_i - \widetilde{\boldsymbol{\beta}}_i^*\| = O_{\mathrm{P}}(\sqrt{1/\mathrm{T}})$.

Proof: Let $r_{\mathrm{T}} = \sqrt{1/\mathrm{T}}$ and $\widetilde{\boldsymbol{u}} \in \mathbb{R}^V$. Following the arguments of Theorem 1 in Zhang et al. (2010), it suffices to show that for any given $\epsilon > 0$, there is a sufficiently large constant $C_\epsilon \in (0, \infty)$ such that, for sufficiently large T the following holds:

$$\mathrm{P}\Big(\inf_{\|\widetilde{\boldsymbol{u}}\| = C_\epsilon} \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i^* + r_{\mathrm{T}}\, \widetilde{\boldsymbol{u}}) > \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i^*) \Big) \geq 1 - \epsilon.$$

Let $\widetilde{\boldsymbol{\beta}}_i = r_{\mathrm{T}}\, \widetilde{\boldsymbol{u}} + \widetilde{\boldsymbol{\beta}}_i^*$ and $\|\widetilde{\boldsymbol{u}}\| = C_\epsilon$. By Taylor's expansion of $\mathcal{L}_{i,\mathrm{T}}(\cdot)$ at $\widetilde{\boldsymbol{\beta}}_i^*$, we get

$$\mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i) - \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i^*) \equiv I_{1,1} + I_{1,2} + I_{1,3}, \tag{B.101}$$

with

$$\begin{aligned}
I_{1,1} &= \frac{1}{\mathrm{T}} \int_0^{\mathrm{T}} \Big[\widetilde{\boldsymbol{x}}_i(t)^\top r_{\mathrm{T}}\, \widetilde{\boldsymbol{u}} \exp\big\{ \widetilde{\boldsymbol{x}}_i(t)^\top \widetilde{\boldsymbol{\beta}}_i^* \big\}\, \mathrm{d}t - \widetilde{\boldsymbol{x}}_i(t-)^\top r_{\mathrm{T}}\, \widetilde{\boldsymbol{u}}\, \mathrm{d}N_i(t) \Big], \\
I_{1,2} &= \frac{1}{2\mathrm{T}} \int_0^{\mathrm{T}} \{ \widetilde{\boldsymbol{x}}_i(t)^\top r_{\mathrm{T}}\, \widetilde{\boldsymbol{u}} \}^2 \exp\big\{ \widetilde{\boldsymbol{x}}_i(t)^\top \widetilde{\boldsymbol{\beta}}_i^* \big\}\, \mathrm{d}t, \\
I_{1,3} &= \frac{1}{6\mathrm{T}} \int_0^{\mathrm{T}} \{ \widetilde{\boldsymbol{x}}_i(t)^\top r_{\mathrm{T}}\, \widetilde{\boldsymbol{u}} \}^3 \exp\big\{ \widetilde{\boldsymbol{x}}_i(t)^\top \widetilde{\boldsymbol{\beta}}_i^{**} \big\}\, \mathrm{d}t,
\end{aligned} \tag{B.102}$$

where $\widetilde{\boldsymbol{\beta}}_i^{**}$ lies between $\widetilde{\boldsymbol{\beta}}_i^*$ and $\widetilde{\boldsymbol{\beta}}_i$.

For the term $I_{1,1}$, using (5.5) in Theorem 6 gives that

$$\begin{aligned}
|I_{1,1}| &\leq \Big\| \frac{1}{\mathrm{T}} \int_0^{\mathrm{T}} \Big[\widetilde{\boldsymbol{x}}_i(t) \exp\big\{ \widetilde{\boldsymbol{x}}_i(t)^\top \widetilde{\boldsymbol{\beta}}_i^* \big\}\, \mathrm{d}t - \widetilde{\boldsymbol{x}}_i(t-)\, \mathrm{d}N_i(t) \Big] \Big\| \|r_{\mathrm{T}}\, \widetilde{\boldsymbol{u}}\| \\
&= O_{\mathrm{P}}(\sqrt{1/\mathrm{T}})\, r_{\mathrm{T}}\, \|\widetilde{\boldsymbol{u}}\|.
\end{aligned} \tag{B.103}$$

For the term $I_{1,2}$, (5.6) in Theorem 6 implies that

$$I_{1,2} = r_{\mathrm{T}}^2\, \widetilde{\boldsymbol{u}}^\top \nabla^2 \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i^*) \widetilde{\boldsymbol{u}} \;=\; r_{\mathrm{T}}^2\, \widetilde{\boldsymbol{u}}^\top \{ \mathbf{C}_i + o(1) \} \widetilde{\boldsymbol{u}}. \tag{B.104}$$

For the term $I_{1,3}$, condition A4 implies that each component of $\widetilde{\boldsymbol{x}}_i(t)$ is bounded above by some positive constant. Hence, we have

$$|I_{1,3}| \leq c\, r_{\mathrm{T}}^3 \|\widetilde{\boldsymbol{u}}\|^3, \tag{B.105}$$

for some constant $c \in (0, \infty)$.

Combining (B.103), (B.104) and (B.105), we obtain

$$\begin{aligned}
&\mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i) - \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i^*) \\
&= I_{1,1} + I_{1,2} + I_{1,3} \\
&= O_{\mathrm{P}}(\sqrt{1/\mathrm{T}})\, r_{\mathrm{T}} \|\widetilde{\boldsymbol{u}}\| + r_{\mathrm{T}}^2 \{\widetilde{\boldsymbol{u}}^\top \mathbf{C}_i \widetilde{\boldsymbol{u}} + o_{\mathrm{P}}(1) \|\widetilde{\boldsymbol{u}}\|^2\} + c\, r_{\mathrm{T}}^3 \|\widetilde{\boldsymbol{u}}\|^3 \\
&= \frac{1}{\mathrm{T}} \{O_{\mathrm{P}}(1) \|\widetilde{\boldsymbol{u}}\| + \widetilde{\boldsymbol{u}}^\top \mathbf{C}_i \widetilde{\boldsymbol{u}} + o_{\mathrm{P}}(1) \|\widetilde{\boldsymbol{u}}\|^2 + o_{\mathrm{P}}(1) \|\widetilde{\boldsymbol{u}}\|^3\}.
\end{aligned} \tag{B.106}$$

By (B.106), we can choose some large C_ϵ, such that all terms in brackets in (B.106) are dominated by the term $\widetilde{\boldsymbol{u}}^\top \mathbf{C}_i \widetilde{\boldsymbol{u}}$, which is positive by the fact that $\mathbf{C}_i \succ 0$ from Theorem 6. This completes the proof. ■

Now we prove Theorem 7. Let $r_{\mathrm{T}} = \sqrt{1/\mathrm{T}}$ and $\widetilde{\boldsymbol{u}} = (u_0, u_1, \ldots, u_{i-1}, u_{i+1}, \ldots, u_V)^\top \in \mathbb{R}^V$. Denote by $\ell_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i)$ the objective function in (5.8), i.e.,

$$\ell_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i) \;=\; \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i) + \sum_{j \in \mathcal{V} \setminus i} w_{j,i,\mathrm{T}} |\beta_{j,i}|. \tag{B.107}$$

Similar to the proof of Lemma B.16, it suffices to show that for any given $\epsilon > 0$, there is a sufficiently large constant $C_\epsilon \in (0, \infty)$ such that, for large T,

$$\mathrm{P}\Big(\inf_{\|\widetilde{\boldsymbol{u}}\| = C_\epsilon} \ell_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i^* + r_{\mathrm{T}}\, \widetilde{\boldsymbol{u}}) > \ell_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i^*) \Big) \;\geq\; 1 - \epsilon.$$

From (B.107), we have

$$\ell_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i^* + r_{\mathrm{T}}\, \widetilde{\boldsymbol{u}}) - \ell_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i^*)$$

$$= \{\mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i^* + r_{\mathrm{T}}\,\widetilde{\boldsymbol{u}}) - \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i^*)\} + \sum_{j \in \mathcal{V} \setminus i} w_{j,i,\mathrm{T}} \cdot (|\beta_{j,i}^* + r_{\mathrm{T}} u_j| - |\beta_{j,i}^*|)$$

$$\geq \{\mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i^* + r_{\mathrm{T}}\,\widetilde{\boldsymbol{u}}) - \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i^*)\} + \sum_{j \in \mathcal{S}_i^*} w_{j,i,\mathrm{T}} \cdot (|\beta_{j,i}^* + r_{\mathrm{T}} u_j| - |\beta_{j,i}^*|)$$

$$\equiv I_1 + I_2.$$

For the term I_1, (B.106) yields that

$$I_1 = \frac{1}{\mathrm{T}}\{O_{\mathrm{P}}(1)\|\widetilde{\boldsymbol{u}}\| + \widetilde{\boldsymbol{u}}^\top \mathbf{C}_i \widetilde{\boldsymbol{u}} + o_{\mathrm{P}}(1)\|\widetilde{\boldsymbol{u}}\|^2 + o_{\mathrm{P}}(1)\|\widetilde{\boldsymbol{u}}\|^3\}.$$

For the term I_2, by triangle inequality and condition (5.11), we have

$$|I_2| \;\leq\; \sum_{j \in \mathcal{S}_i^*} w_{j,i,\mathrm{T}}\, r_{\mathrm{T}}\, |u_j| \;\leq\; r_{\mathrm{T}}\, \|\widetilde{\boldsymbol{u}}\|_1 \max_{j \in \mathcal{S}_i^*} w_{j,i,\mathrm{T}} \;=\; O_{\mathrm{P}}(1/\mathrm{T})\, \|\widetilde{\boldsymbol{u}}\|_1,$$

which is dominated by $\widetilde{\boldsymbol{u}}^\top \mathbf{C}_i \widetilde{\boldsymbol{u}}/\mathrm{T}$ for a sufficiently large C_ϵ. Hence, we conclude that I_2 is dominated by I_1. The remaining proof is the same as that of Lemma B.16. ■

B.13. Proof of Theorem 8

For a $\sqrt{1/\mathrm{T}}$-consistent estimator $\widehat{\widetilde{\boldsymbol{\beta}}}_i$ of $\widetilde{\boldsymbol{\beta}}_i^*$, for any $\epsilon > 0$, there exists constant C_ϵ, such that for sufficiently large T,

$$\mathrm{P}(\|\widehat{\widetilde{\boldsymbol{\beta}}}_i - \widetilde{\boldsymbol{\beta}}_i^*\| \leq r_{\mathrm{T}} C_\epsilon) \;>\; 1 - \epsilon. \tag{B.108}$$

Let $r_{\mathrm{T}} = \sqrt{1/\mathrm{T}}$. Recall condition A4 implies that $\widetilde{\boldsymbol{x}}_i(t)$ is bounded above. This together with (B.93) yields that there exists a constant $c \in (0, \infty)$, such that

$$\sup_{\widetilde{\boldsymbol{\beta}}_i : \|\widetilde{\boldsymbol{\beta}}_i - \widetilde{\boldsymbol{\beta}}_i^*\| \leq r_{\mathrm{T}} C_\epsilon} \Big| \frac{\partial^2 \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i)}{\partial \beta_{j,i} \partial \beta_{0;i}} \Big| \leq c, \;\; \text{for any } j \in \mathcal{V} \setminus i,$$

$$\sup_{\widetilde{\boldsymbol{\beta}}_i : \|\widetilde{\boldsymbol{\beta}}_i - \widetilde{\boldsymbol{\beta}}_i^*\| \leq r_{\mathrm{T}} C_\epsilon} \Big| \frac{\partial^2 \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i)}{\partial \beta_{j,i} \partial \beta_{k,i}} \Big| \leq c, \;\; \text{for any } j, k \in \mathcal{V} \setminus i.$$

Combining this with Taylor's expansion, (B.92), and (B.97), for $j \in \mathcal{V} \setminus i$, we have

$$\begin{aligned}
&\sup_{\widetilde{\boldsymbol{\beta}}_i : \|\widetilde{\boldsymbol{\beta}}_i - \widetilde{\boldsymbol{\beta}}_i^*\| \le r_{\mathrm{T}} C_\epsilon} \Big| \frac{\partial \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i)}{\partial \beta_{j,i}} \Big| \\
&\le \Big| \frac{\partial \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i^*)}{\partial \beta_{j,i}} \Big| \\
&\quad + \sup_{\widetilde{\boldsymbol{\beta}}_i : \|\widetilde{\boldsymbol{\beta}}_i - \widetilde{\boldsymbol{\beta}}_i^*\| \le r_{\mathrm{T}} C_\epsilon} \Big[\Big\{ \Big| \frac{\partial^2 \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i)}{\partial \beta_{j,i} \partial \beta_{0;i}} \Big| + \sum_{k \in \mathcal{V} \setminus i} \Big| \frac{\partial^2 \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i)}{\partial \beta_{j,i} \partial \beta_{k,i}} \Big| \Big\} \cdot \|\widetilde{\boldsymbol{\beta}}_i - \widetilde{\boldsymbol{\beta}}_i^*\| \Big] \\
&\le \Big| \frac{\partial \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i^*)}{\partial \beta_{j,i}} \Big| + V \cdot c\, r_{\mathrm{T}}\, C_\epsilon \\
&= O_{\mathrm{P}}(\sqrt{1/\mathrm{T}}) + O(r_{\mathrm{T}}) = O_{\mathrm{P}}(\sqrt{1/\mathrm{T}}).
\end{aligned} \tag{B.109}$$

Consider $\widetilde{\boldsymbol{\beta}}_i$ in the ball $\{\widetilde{\boldsymbol{\beta}}_i : \|\widetilde{\boldsymbol{\beta}}_i - \widetilde{\boldsymbol{\beta}}_i^*\| \le r_{\mathrm{T}}\, C_\epsilon\}$. For $j \in \mathcal{V} \setminus \{\mathcal{S}_i^* \cup i\}$, if $\beta_{j,i} > 0$, then (5.13) and (B.109) yield that

$$\begin{aligned}
\frac{\partial \ell_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i)}{\partial \beta_{j,i}} &= \frac{\partial \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i)}{\partial \beta_{j,i}} + w_{j,i,\mathrm{T}} \operatorname{sign}(\beta_{j,i}) \\
&\ge - \sup_{\widetilde{\boldsymbol{\beta}}_i : \|\widetilde{\boldsymbol{\beta}}_i - \widetilde{\boldsymbol{\beta}}_i^*\| \le r_{\mathrm{T}} C_\epsilon} \Big| \frac{\partial \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i)}{\partial \beta_{j,i}} \Big| + \min_{j \in \mathcal{V} \setminus \{\mathcal{S}_i^* \cup i\}} w_{j,i,\mathrm{T}} \\
&> 0,
\end{aligned} \tag{B.110}$$

with probability tending to 1 as $\mathrm{T} \to \infty$. Similarly, if $\beta_{j,i} < 0$, we have

$$\begin{aligned}
\frac{\partial \ell_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i)}{\partial \beta_{j,i}} &= \frac{\partial \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i)}{\partial \beta_{j,i}} + w_{j,i,\mathrm{T}} \operatorname{sign}(\beta_{j,i}) \\
&\le \sup_{\widetilde{\boldsymbol{\beta}}_i : \|\widetilde{\boldsymbol{\beta}}_i - \widetilde{\boldsymbol{\beta}}_i^*\| \le r_{\mathrm{T}} C_\epsilon} \Big| \frac{\partial \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i)}{\partial \beta_{j,i}} \Big| - \min_{j \in \mathcal{V} \setminus \{\mathcal{S}_i^* \cup i\}} w_{j,i,\mathrm{T}} \\
&< 0,
\end{aligned} \tag{B.111}$$

with probability tending to 1 as $\mathrm{T} \to \infty$.

By (B.110) and (B.111), the following argument holds with probability tending to 1 as $\mathrm{T} \to \infty$: for all $\widetilde{\boldsymbol{\beta}}_i$ with $\|\widetilde{\boldsymbol{\beta}}_i - \widetilde{\boldsymbol{\beta}}_i^*\| \le r_{\mathrm{T}} C_\epsilon$ and all $j \in \mathcal{V} \setminus \{\mathcal{S}_i^* \cup i\}$, $\partial \ell_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i)/\partial \beta_{j,i}$ has the

same sign as $\beta_{j,i}$. Together with (B.108) and the first order condition of $\widehat{\widetilde{\boldsymbol{\beta}}}_i$, it follows that

$$\mathrm{P}\big(\widehat{\beta}_{j,i} = 0, \text{ for all } j \in \mathcal{V} \setminus \{\mathcal{S}_i^* \cup i\}\big) \geq 1 - 2\epsilon \tag{B.112}$$

holds for sufficiently large T. Since ϵ is arbitrary, letting $\epsilon \to 0$ in (B.112) yields that

$$\mathrm{P}\big(\widehat{\beta}_{j,i} = 0, \text{ for all } j \in \mathcal{V} \setminus \{\mathcal{S}_i^* \cup i\}\big) \to 1, \quad \text{as } \mathrm{T} \to \infty. \tag{B.113}$$

Note that the vector $\widehat{\boldsymbol{\beta}}_i^{(\mathrm{II})}$ collects all the components in $\widehat{\widetilde{\boldsymbol{\beta}}}_i$ whose indices belong to the set $\mathcal{V} \setminus \{\mathcal{S}_i^* \cup i\}$. Hence, (B.113) implies that

$$\mathrm{P}(\widehat{\boldsymbol{\beta}}_i^{(\mathrm{II})} = 0) \to 1, \quad \text{as } \mathrm{T} \to \infty.$$

This completes the proof. ■

B.14. Proof of Corollary 1

Recall that the true edge set $\mathcal{E}^* \neq \varnothing$ is assumed in condition A6. To prove Corollary 1, it suffices to show that for each pair of distinct nodes $(j, i) \in \mathcal{V} \times \mathcal{V}$,

$$\begin{cases} \mathrm{P}((j,i) \in \widehat{\mathcal{E}}_+) \to 1, & \text{if } (j,i) \in \mathcal{E}_+^*, \\ \mathrm{P}((j,i) \in \widehat{\mathcal{E}}_-) \to 1, & \text{if } (j,i) \in \mathcal{E}_-^*, \text{ as } \mathrm{T} \to \infty. \\ \mathrm{P}((j,i) \notin \widehat{\mathcal{E}}) \to 1, & \text{if } (j,i) \notin \mathcal{E}^*, \end{cases}$$

If $(j, i) \in \mathcal{E}_+^*$, then (3.3) implies that $\beta_{j,i}^* > 0$. By Theorem 7, we have $\widehat{\beta}_{j,i} \xrightarrow{\mathrm{P}} \beta_{j,i}^* > 0$. Thus, $\mathrm{P}((j,i) \in \widehat{\mathcal{E}}_+) = \mathrm{P}(\widehat{\beta}_{j,i} > 0) \to 1$, as $\mathrm{T} \to \infty$.

If $(j, i) \in \mathcal{E}_-^*$, then similarly as above, we have $\mathrm{P}((j,i) \in \widehat{\mathcal{E}}_-) = \mathrm{P}(\widehat{\beta}_{j,i} < 0) \to 1$, as $\mathrm{T} \to \infty$.

If $(j, i) \notin \mathcal{E}^*$, then (3.2) implies that $\beta_{j,i}^* = 0$. By (5.14) in Theorem 8, we obtain

$\mathrm{P}((j,i) \notin \widehat{\mathcal{E}}) = \mathrm{P}(\widehat{\beta}_{j,i} = 0) \to 1$, as $\mathrm{T} \to \infty$. This completes the proof. ■

B.15. Proof of Theorem 9

The proof of Theorem 9 is similar to that of Theorem 7, with slight modifications in accordance with the extended model (6.1). We omit the lengthy details and outline the key steps of the proof as follows.

(**a**) In accordance with model (6.1), the set of discontinuity points (4.4) in Definition 2 is modified as

$$\{\breve{T}_1, \breve{T}_2, \ldots\} = \bigcup_{k\in\{0,1,\ldots,K\}} \{T_{j,\ell} + k\,\phi\}_{j\in\mathcal{V},\ell\geq 1}.$$

In addition, the definition of the mark I_ℓ in (4.5) is modified as

$$I_\ell = \begin{cases} i, & \text{if } \breve{T}_\ell \in \{T_{i,r}\}_{r\geq 1} \text{ for some node } i \in \mathcal{V}, \\ 0, & \text{if } \breve{T}_\ell \in \{T_{i,r} + k\phi\}_{r\geq 1,\, k\in\{1,\ldots,K\}} \text{ for some node } i \in \mathcal{V}. \end{cases} \tag{B.114}$$

Further, the set $\mathcal{T}_\ell$ in (4.8) of Theorem 1 is modified as

$$\mathcal{T}_\ell = \bigcup_{j\in\mathcal{V};\, k\in\{1,\ldots,K\}} \big\{t \in (\breve{T}_\ell - \phi, \breve{T}_\ell] : N_{j,k}(\{t\}) = 1\big\}. \tag{B.115}$$

After the above modifications of definitions, we apply the same proof techniques used in Section 4.1, and verify that Lemmas 3–5, Theorems 1 and 2 hold for model (6.1).

(**b**) In Definition 3, the ℓth recurrence time point R_ℓ in (4.25) is modified as

$$R_\ell = \min\{t \geq R_{\ell-1} + K\phi : \boldsymbol{N}((t - K\phi, t]) = 0\}. \tag{B.116}$$

Following the same proof, Theorems 3 and 4 hold for model (6.1).

(**c**) After verifying Theorems 2–4 for model (6.1), we are able to prove that Theorem 6 holds for model (6.1). By applying the same proof as that of Theorem 7, there exists a local minimizer $\widehat{\widetilde{\boldsymbol{\beta}}}_i$ of (6.6) such that $\|\widehat{\widetilde{\boldsymbol{\beta}}}_i - \widetilde{\boldsymbol{\beta}}_i^*\| = O_{\mathrm{P}}(\sqrt{1/\mathrm{T}})$, which implies that

$$\begin{aligned} \sup_{t\geq 0} |\widehat{\omega}_{j,i}(t) - \omega_{j,i}^*(t)| &= \sup_{1\leq k\leq K} |\widehat{\beta}_{j,i,k} - \beta_{j,i,k}^*| \\ &\leq \|\widehat{\widetilde{\boldsymbol{\beta}}}_i - \widetilde{\boldsymbol{\beta}}_i^*\| = O_{\mathrm{P}}(\sqrt{1/\mathrm{T}}). \end{aligned}$$

This completes the proof. ■

B.16. Proof of Corollary 2

Recall that the true edge set $\mathcal{E}^* \neq \varnothing$ is assumed in condition A6. To prove Corollary 2, it suffices to show that for each pair of distinct nodes $(j,i) \in \mathcal{V} \times \mathcal{V}$, as $\mathrm{T} \to \infty$, we have

$$\begin{cases} \mathrm{P}((j,i) \in \widehat{\mathcal{E}}) \to 1, & \text{if } (j,i) \in \mathcal{E}^*, \\ \mathrm{P}((j,i) \notin \widehat{\mathcal{E}}) \to 1, & \text{if } (j,i) \notin \mathcal{E}^*. \end{cases}$$

If $(j,i) \in \mathcal{E}^*$, then there exists $k \in \{1,\ldots,K\}$ such that $\beta^*_{j,i,k} \neq 0$. By Theorem 9, we have $\widehat{\beta}_{j,i,k} \xrightarrow{\mathrm{P}} \beta^*_{j,i,k}$. Thus,

$$\begin{aligned} \int_0^\infty |\widehat{\omega}_{j,i}(t)|\,\mathrm{d}t &\geq \int_{(k-1)\phi}^{k\,\phi} |\widehat{\omega}_{j,i}(t)|\,\mathrm{d}t \\ &= |\widehat{\beta}_{j,i,k}| = |\beta^*_{j,i,k}| + o_{\mathrm{P}}(1). \end{aligned}$$

This implies that $\mathrm{P}((j,i) \in \widehat{\mathcal{E}}) \to 1$ as $\mathrm{T} \to \infty$.

If $(j,i) \notin \mathcal{E}^*$, then (6.8) implies that $\beta^*_{j,i,k} = 0$ for any $k = 1,\ldots,K$. Following the same proof as that of Theorem 8, we have $\mathrm{P}(\widehat{\beta}_{j,i,k} = 0) \to 1$ as $\mathrm{T} \to \infty$, for any $k \in \{1,\ldots,K\}$. Thus,

$$\mathrm{P}\Big(\int_0^\infty |\widehat{\omega}_{j,i}(t)|\,\mathrm{d}t \;=\; 0\Big) = \mathrm{P}\Big(\sum_{k=1}^{K} |\widehat{\beta}_{j,i,k}| = 0\Big) \to 1, \quad \text{as } \mathrm{T} \to \infty.$$

This implies that $\mathrm{P}((j,i) \notin \widehat{\mathcal{E}}) \to 1$ as $\mathrm{T} \to \infty$. The proof is completed. ■

Appendix C

Supplementary material for simulation

Details for methods in simulation

Selection of tuning parameter η in method (i)

The tuning parameter η is selected by minimizing the Bayesian Information Criterion (BIC) function in Nishii (1984)

$$\mathrm{BIC}_i(\widehat{\widetilde{\boldsymbol{\beta}}}_{i,\eta}) = 2\mathcal{L}_{i,\mathrm{T}}(\widehat{\widetilde{\boldsymbol{\beta}}}_{i,\eta}) + \mathrm{df}(\widehat{\widetilde{\boldsymbol{\beta}}}_{i,\eta})\ \log(\mathrm{T})/\mathrm{T}, \tag{C.1}$$

where $\widehat{\widetilde{\boldsymbol{\beta}}}_{i,\eta}$ denotes the penalized M-estimator with tuning parameter η, and $\mathrm{df}(\widehat{\widetilde{\boldsymbol{\beta}}}_{i,\eta})$ is the number of non-zero components of $\widehat{\widetilde{\boldsymbol{\beta}}}_{i,\eta}$ excluding baseline term $\widehat{\beta}_{0;i,\eta}$. The η that minimizes (C.1) is found by searching the grid points $\{\eta_{\max}\, h^k : k = 0, 1, \ldots, 11\}$, where $\eta_{\max} = \sup\{\eta : \mathrm{df}(\widehat{\widetilde{\boldsymbol{\beta}}}_{i,\eta}) > 0\}$, and the step size h is chosen to be 0.7.

Detailed descriptions for method (ii)

In method (ii), the entire time interval $[0, \mathrm{T}]$ is partitioned into the equally-spaced time bins $\{(t_{k-1}, t_k] : k = 1, \ldots, n\}$, each of length $\delta = \mathrm{T}/n$, and the observed point process $\{\boldsymbol{T}_i\}_{i\in\mathcal{V}}$

is transformed into sequences of bin counts $\{N_{i,k}\}_{i\in\mathcal{V};k=1,\ldots,n}$. The event count $N_{i,k}$ at node i in the kth bin approximately follows a Poisson distribution with rate $\lambda_i(t_{k-1})\,\delta$, where

$$\lambda_i(t_{k-1}) = \exp\Big\{\beta_{0;i} + \sum_{j\in\mathcal{V}} \beta_{j,i}\, \log\Big(1 + \Big\{\frac{N_j((t_{k-1}-\phi, t_{k-1}])}{\phi} \wedge 10\Big\}\Big)\Big\} \tag{C.2}$$

corresponds to the true conditional intensity function $\lambda_i(t)$ in (3.1) with covariate $x_i(t)$ in (7.1), at time $t = t_{k-1}$. The loss function $\mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i)$ in (5.4) is approximated by the negative log-likelihood (i.e., the joint p.m.f.) function $\mathcal{L}^D_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i)$ of $\{N_{i,k}\}_{k=1,\ldots,n}$, i.e.,

$$\begin{aligned}
\mathcal{L}^D_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i) &= -\frac{1}{\mathrm{T}} \log\Big(\prod_{k=1}^{n} \frac{\{\lambda_i(t_{k-1})\delta\}^{N_{i,k}}}{N_{i,k}!}\, \exp\{-\lambda_i(t_{k-1})\delta\}\Big) \\
&\equiv -\frac{1}{\mathrm{T}} \log\Big(\prod_{k=1}^{n} \lambda_i(t_{k-1})^{N_{i,k}}\, \exp\{-\lambda_i(t_{k-1})\delta\}\Big) + c \\
&= -\frac{1}{\mathrm{T}} \sum_{k=1}^{n} \Big[N_{i,k}\, \log\{\lambda_i(t_{k-1})\} - \lambda_i(t_{k-1})\,\delta\Big],
\end{aligned} \tag{C.3}$$

where the constant c is ignored since it does not influence the estimation result. We estimate parameters by minimizing $\mathcal{L}^D_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}) + \mathcal{P}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}})$, and adopt the same two scenarios (a) and (b) of penalties in method (i) for a fair comparison.

Detailed descriptions for method (iii)

Method (iii) adopts the same discretization procedure as in method (ii). The event count $N_{i,k}$ at node i in the kth bin is assumed to follow a Poisson distribution with rate $\lambda_i(t_{k-1})\,\delta$, where

$$\lambda_i(t_{k-1}) = \exp\Big\{\beta_{0;i} + \sum_{j\in\mathcal{V}} \sum_{q=1}^{Q} \beta_{j,i,q} N_{j,k-q}\Big\}.$$

Here, the effect from node j to node i is modeled by a group of parameters $\{\beta_{j,i,q} : q = 1,\ldots,Q\}$, with $Q = [\phi/\delta]$. Method (iii) employs the same loss function (C.3) as in method (ii) for parameter estimation.

Simulation results for Network 1

Table 5: Simulation of Network 1 with varying connection strength β. We set time length $T = 500$.

		Correct_All		Detected_A		Detected_B		Correct_NC	
strength $\beta =$		0.4	0.5	0.4	0.5	0.4	0.5	0.4	0.5
Discrete_L1	bin=0.5	1.11 (0.11)	2.39 (0.15)	0.81 (0.09)	1.81 (0.12)	0.30 (0.05)	0.57 (0.07)	79.70 (0.07)	79.48 (0.08)
	bin=0.25	1.77 (0.11)	3.65 (0.16)	1.32 (0.09)	2.70 (0.11)	0.45 (0.06)	0.95 (0.08)	79.61 (0.06)	79.17 (0.10)
	bin=0.1	2.39 (0.14)	4.63 (0.17)	1.71 (0.10)	3.32 (0.11)	0.68 (0.07)	1.31 (0.09)	79.48 (0.07)	78.91 (0.14)
Continuous_L1		4.92 (0.16)	6.56 (0.14)	3.05 (0.10)	4.24 (0.09)	1.87 (0.10)	2.31 (0.09)	76.25 (0.23)	76.58 (0.21)
Discrete_AL	bin=0.5	2.14 (0.13)	3.92 (0.15)	1.51 (0.10)	2.75 (0.12)	0.63 (0.07)	1.17 (0.08)	79.17 (0.09)	78.98 (0.10)
	bin=0.25	3.24 (0.15)	5.34 (0.16)	2.22 (0.11)	3.65 (0.11)	1.02 (0.08)	1.69 (0.09)	79.05 (0.10)	78.91 (0.12)
	bin=0.1	3.86 (0.15)	6.51 (0.15)	2.62 (0.11)	4.34 (0.10)	1.24 (0.08)	2.17 (0.10)	78.92 (0.10)	78.80 (0.12)
Continuous_AL		5.19 (0.14)	7.44 (0.14)	3.32 (0.10)	4.80 (0.09)	1.78 (0.08)	2.64 (0.08)	78.31 (0.13)	78.67 (0.11)
Zhao_2012	bin=0.5	0.23 (0.04)	0.83 (0.08)	0.18 (0.04)	0.67 (0.07)	0.05 (0.02)	0.16 (0.03)	79.93 (0.02)	79.83 (0.04)
	bin=0.25	0.22 (0.04)	0.63 (0.06)	0.21 (0.04)	0.60 (0.06)	0.01 (0.00)	0.03 (0.01)	79.98 (0.01)	79.95 (0.01)
	bin=0.1	0.08 (0.02)	0.20 (0.04)	0.08 (0.02)	0.20 (0.04)	0.00 (0.00)	0.00 (0.00)	79.94 (0.02)	79.95 (0.01)
SIE-GLM	bin=0.5	0.80 (0.08)	2.00 (0.13)	0.70 (0.08)	1.63 (0.11)	0.10 (0.03)	0.38 (0.05)	79.83 (0.03)	79.69 (0.06)
	bin=0.25	0.67 (0.08)	1.81 (0.11)	0.56 (0.07)	1.61 (0.10)	0.10 (0.03)	0.20 (0.04)	79.84 (0.04)	79.68 (0.06)
	bin=0.1	0.22 (0.04)	0.82 (0.08)	0.20 (0.04)	0.80 (0.08)	0.02 (0.01)	0.02 (0.01)	79.86 (0.07)	79.94 (0.02)
Raj_2005	parent=3	9.02 (0.08)	9.68 (0.04)	5.49 (0.06)	5.85 (0.03)	3.54 (0.06)	3.83 (0.04)	59.56 (0.12)	60.15 (0.08)
TRUE		10		6		4		80	

Table 6: Simulation of Network 1 with varying time length T. The connection strength β is 0.5.

		Correct_All		Detected_A		Detected_B		Correct_NC	
time length T =		500	1000	500	1000	500	1000	500	1000
Discrete_L1	bin=0.5	2.39 (0.15)	5.70 (0.16)	1.81 (0.12)	3.89 (0.11)	0.57 (0.07)	1.81 (0.08)	79.48 (0.08)	78.94 (0.11)
	bin=0.25	3.65 (0.16)	7.46 (0.14)	2.70 (0.11)	4.88 (0.09)	0.95 (0.08)	2.58 (0.09)	79.17 (0.10)	78.48 (0.14)
	bin=0.1	4.63 (0.17)	8.40 (0.12)	3.32 (0.11)	5.37 (0.07)	1.31 (0.09)	3.03 (0.08)	78.91 (0.14)	78.47 (0.13)
Continuous_L1		6.56 (0.14)	9.44 (0.07)	4.24 (0.09)	5.80 (0.04)	2.31 (0.09)	3.65 (0.05)	76.58 (0.21)	76.48 (0.21)
Discrete_AL	bin=0.5	3.92 (0.15)	7.34 (0.14)	2.75 (0.12)	4.76 (0.10)	1.17 (0.08)	2.58 (0.09)	78.98 (0.10)	78.95 (0.11)
	bin=0.25	5.34 (0.16)	8.69 (0.11)	3.65 (0.11)	5.51 (0.06)	1.69 (0.09)	3.18 (0.08)	78.91 (0.12)	79.11 (0.10)
	bin=0.1	6.51 (0.15)	9.25 (0.08)	4.34 (0.10)	5.76 (0.04)	2.17 (0.10)	3.49 (0.07)	78.80 (0.12)	79.18 (0.09)
Continuous_AL		7.44 (0.14)	9.52 (0.06)	4.80 (0.09)	5.90 (0.03)	2.64 (0.08)	3.62 (0.06)	78.67 (0.11)	79.03 (0.10)
Zhao_2012	bin=0.5	0.83 (0.08)	2.84 (0.16)	0.67 (0.07)	1.92 (0.12)	0.16 (0.03)	0.92 (0.07)	79.83 (0.04)	79.52 (0.07)
	bin=0.25	0.63 (0.06)	1.19 (0.09)	0.60 (0.06)	1.06 (0.08)	0.03 (0.01)	0.13 (0.03)	79.95 (0.01)	79.91 (0.03)
	bin=0.1	0.20 (0.04)	0.32 (0.05)	0.20 (0.04)	0.32 (0.05)	0.00 (0.00)	0.00 (0.00)	79.95 (0.01)	79.95 (0.02)
SIE-GLM	bin=0.5	2.00 (0.13)	5.12 (0.17)	1.63 (0.11)	3.67 (0.11)	0.38 (0.05)	1.45 (0.09)	79.69 (0.06)	79.20 (0.11)
	bin=0.25	1.81 (0.11)	4.63 (0.15)	1.61 (0.10)	3.51 (0.10)	0.20 (0.04)	1.12 (0.09)	79.68 (0.06)	79.39 (0.08)
	bin=0.1	0.82 (0.08)	2.50 (0.13)	0.80 (0.08)	2.27 (0.11)	0.02 (0.01)	0.22 (0.04)	79.94 (0.02)	79.68 (0.10)
Raj_2005	parent=3	9.68 (0.04)	9.97 (0.01)	5.85 (0.03)	5.99 (0.00)	3.83 (0.04)	3.98 (0.01)	60.15 (0.08)	63.78 (0.20)
TRUE		10		6		4		80	

Table 7: Simulation of Network 1 with varying assumed time-lag width ϕ_a. The connection strength $\beta = 0.5$ and the total time length $T = 500$. The true time-lag width $\phi = 1$. Results are averaged over 100 replications, with standard errors indicated in parentheses.

		Correct_All			Detected_A			Detected_B			Correct_NC		
assumed time-lag width =		0.5	1	1.5	0.5	1	1.5	0.5	1	1.5	0.5	1	1.5
Discrete_L1	bin=0.5	2.22 (0.14)	2.39 (0.15)	1.18 (0.10)	1.76 (0.11)	1.81 (0.12)	0.94 (0.08)	0.46 (0.06)	0.57 (0.07)	0.24 (0.04)	79.56 (0.07)	79.48 (0.08)	79.69 (0.08)
	bin=0.25	2.11 (0.12)	3.65 (0.16)	1.96 (0.12)	1.64 (0.10)	2.70 (0.11)	1.48 (0.09)	0.47 (0.05)	0.95 (0.08)	0.48 (0.06)	79.63 (0.06)	79.17 (0.10)	79.45 (0.08)
	bin=0.1	2.13 (0.13)	4.63 (0.17)	2.48 (0.14)	1.57 (0.11)	3.32 (0.11)	1.85 (0.11)	0.56 (0.07)	1.31 (0.09)	0.63 (0.06)	79.52 (0.08)	78.91 (0.14)	79.39 (0.09)
Continuous_L1		4.23 (0.15)	6.56 (0.14)	5.66 (0.16)	2.88 (0.11)	4.24 (0.09)	3.54 (0.13)	1.35 (0.09)	2.31 (0.09)	2.12 (0.10)	75.50 (0.23)	76.58 (0.21)	76.27 (0.22)
Discrete_AL	bin=0.5	3.55 (0.16)	3.92 (0.15)	2.39 (0.13)	2.50 (0.11)	2.75 (0.12)	1.71 (0.11)	1.05 (0.08)	1.17 (0.08)	0.59 (0.06)	79.19 (0.09)	78.98 (0.10)	79.11 (0.11)
	bin=0.25	3.49 (0.16)	5.34 (0.16)	3.49 (0.15)	2.40 (0.12)	3.65 (0.11)	2.44 (0.11)	1.09 (0.09)	1.69 (0.09)	1.05 (0.09)	79.06 (0.09)	78.91 (0.12)	79.05 (0.10)
	bin=0.1	3.44 (0.15)	6.51 (0.15)	4.34 (0.15)	2.31 (0.11)	4.34 (0.10)	3.01 (0.12)	1.12 (0.08)	2.17 (0.10)	1.34 (0.09)	79.00 (0.11)	78.80 (0.12)	78.87 (0.10)
Continuous_AL		3.71 (0.15)	7.44 (0.14)	5.50 (0.14)	2.35 (0.12)	4.80 (0.09)	3.63 (0.11)	1.36 (0.07)	2.64 (0.08)	1.87 (0.10)	78.88 (0.11)	78.67 (0.11)	78.33 (0.14)
Zhao_2012	bin=0.5	1.50 (0.12)	0.83 (0.08)	0.64 (0.08)	1.12 (0.10)	0.67 (0.07)	0.53 (0.07)	0.38 (0.05)	0.16 (0.03)	0.11 (0.03)	79.70 (0.06)	79.83 (0.04)	79.84 (0.04)
	bin=0.25	0.66 (0.07)	0.63 (0.06)	0.55 (0.06)	0.61 (0.07)	0.60 (0.06)	0.52 (0.06)	0.05 (0.02)	0.03 (0.01)	0.03 (0.01)	79.94 (0.02)	79.95 (0.01)	79.94 (0.02)
	bin=0.1	0.23 (0.04)	0.20 (0.04)	0.19 (0.04)	0.23 (0.04)	0.20 (0.04)	0.19 (0.04)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	79.92 (0.03)	79.95 (0.01)	79.95 (0.01)
SIE-GLM	bin=0.5	1.89 (0.13)	2.00 (0.13)	1.67 (0.11)	1.51 (0.11)	1.63 (0.11)	1.41 (0.10)	0.38 (0.05)	0.38 (0.05)	0.26 (0.04)	79.78 (0.04)	79.69 (0.06)	79.59 (0.07)
	bin=0.25	1.22 (0.10)	1.81 (0.11)	1.22 (0.09)	1.09 (0.09)	1.61 (0.10)	1.13 (0.09)	0.13 (0.03)	0.20 (0.04)	0.08 (0.02)	79.84 (0.04)	79.68 (0.06)	79.79 (0.04)
	bin=0.1	0.53 (0.06)	0.82 (0.08)	0.59 (0.07)	0.53 (0.06)	0.80 (0.08)	0.53 (0.06)	0.00 (0.00)	0.02 (0.01)	0.06 (0.02)	79.89 (0.03)	79.94 (0.02)	79.30 (0.21)
Raj_2005	parent=3	8.40 (0.11)	9.68 (0.04)	9.11 (0.08)	5.20 (0.07)	5.85 (0.03)	5.62 (0.04)	3.20 (0.07)	3.83 (0.04)	3.49 (0.07)	59.41 (0.15)	60.15 (0.08)	59.17 (0.08)
TRUE			10			6			4			80	

Simulation results for Network 2

Table 8: Simulation of Network 2 with varying connection strength β. We set time length $T = 1000$.

		Correct_All		Detected_A		Detected_B		Correct_NC	
strength $\beta =$		0.4	0.5	0.4	0.5	0.4	0.5	0.4	0.5
Discrete_L1	bin=0.5	4.97 (0.23)	11.71 (0.25)	3.53 (0.17)	8.13 (0.18)	1.44 (0.10)	3.58 (0.12)	359.06 (0.10)	357.29 (0.17)
	bin=0.25	8.97 (0.32)	15.90 (0.17)	6.16 (0.23)	10.62 (0.12)	2.81 (0.13)	5.28 (0.10)	357.91 (0.19)	355.76 (0.22)
	bin=0.1	11.69 (0.31)	17.62 (0.14)	7.99 (0.21)	11.38 (0.09)	3.70 (0.14)	6.24 (0.10)	357.24 (0.21)	355.47 (0.25)
Continuous_L1		16.09 (0.23)	19.32 (0.09)	10.07 (0.15)	11.71 (0.05)	6.02 (0.11)	7.61 (0.06)	348.16 (0.49)	346.97 (0.47)
Discrete_AL	bin=0.5	9.39 (0.26)	15.37 (0.16)	6.31 (0.20)	10.17 (0.12)	3.08 (0.12)	5.20 (0.11)	357.41 (0.20)	356.91 (0.17)
	bin=0.25	13.40 (0.27)	17.95 (0.13)	8.79 (0.18)	11.37 (0.09)	4.70 (0.14)	6.58 (0.11)	357.12 (0.22)	356.95 (0.18)
	bin=0.1	15.03 (0.23)	18.80 (0.08)	9.60 (0.16)	11.69 (0.05)	5.43 (0.12)	7.11 (0.08)	357.30 (0.19)	356.94 (0.22)
Continuous_AL		16.23 (0.17)	19.28 (0.07)	10.36 (0.12)	11.82 (0.03)	5.88 (0.11)	7.46 (0.06)	357.13 (0.19)	357.17 (0.19)
Zhao_2012	bin=0.5	1.90 (0.17)	5.53 (0.29)	1.33 (0.12)	4.01 (0.20)	0.56 (0.07)	1.52 (0.11)	359.65 (0.06)	359.28 (0.10)
	bin=0.25	0.63 (0.08)	1.75 (0.14)	0.57 (0.07)	1.55 (0.13)	0.05 (0.02)	0.20 (0.04)	359.91 (0.02)	359.81 (0.05)
	bin=0.1	0.17 (0.04)	0.50 (0.06)	0.17 (0.04)	0.48 (0.06)	0.00 (0.00)	0.02 (0.01)	359.90 (0.03)	359.87 (0.03)
SIE-GLM	bin=0.5	4.50 (0.22)	10.35 (0.24)	3.46 (0.17)	7.65 (0.17)	1.04 (0.09)	2.70 (0.13)	359.34 (0.09)	358.23 (0.16)
	bin=0.25	3.90 (0.21)	10.35 (0.23)	3.32 (0.18)	7.83 (0.17)	0.57 (0.06)	2.52 (0.12)	359.54 (0.07)	358.44 (0.14)
	bin=0.1	1.47 (0.13)	5.29 (0.25)	1.42 (0.12)	4.70 (0.20)	0.05 (0.02)	0.59 (0.07)	359.87 (0.03)	359.54 (0.07)
Raj_2005	parent=3	19.21 (0.09)	19.80 (0.04)	11.66 (0.05)	11.90 (0.03)	7.55 (0.06)	7.91 (0.02)	319.23 (0.09)	319.81 (0.04)
TRUE		20		12		8		360	

Table 9: Simulation of Network 2 with varying time length T. The connection strength β is 0.5.

		Correct_All		Detected_A		Detected_B		Correct_NC	
time length T =		500	1000	500	1000	500	1000	500	1000
Discrete_L1	bin=0.5	4.54 (0.21)	11.71 (0.25)	3.34 (0.16)	8.13 (0.18)	1.20 (0.09)	3.58 (0.12)	358.99 (0.10)	357.29 (0.17)
	bin=0.25	7.87 (0.30)	15.90 (0.17)	5.71 (0.21)	10.62 (0.12)	2.16 (0.12)	5.28 (0.10)	358.20 (0.17)	355.76 (0.22)
	bin=0.1	9.64 (0.32)	17.62 (0.14)	6.76 (0.23)	11.38 (0.09)	2.88 (0.13)	6.24 (0.10)	357.42 (0.18)	355.47 (0.25)
Continuous_L1		14.14 (0.25)	19.32 (0.09)	9.28 (0.17)	11.71 (0.05)	4.84 (0.13)	7.61 (0.06)	347.04 (0.33)	346.97 (0.47)
Discrete_AL	bin=0.5	8.25 (0.26)	15.37 (0.16)	5.55 (0.19)	10.17 (0.12)	2.70 (0.13)	5.20 (0.11)	356.36 (0.20)	356.91 (0.17)
	bin=0.25	12.01 (0.25)	17.95 (0.13)	8.00 (0.18)	11.37 (0.09)	4.01 (0.12)	6.58 (0.11)	355.93 (0.21)	356.95 (0.18)
	bin=0.1	14.00 (0.21)	18.80 (0.08)	9.10 (0.17)	11.69 (0.05)	4.90 (0.12)	7.11 (0.08)	355.98 (0.21)	356.94 (0.22)
Continuous_AL		14.92 (0.20)	19.28 (0.07)	9.61 (0.14)	11.82 (0.03)	5.30 (0.12)	7.46 (0.06)	355.92 (0.23)	357.17 (0.19)
Zhao_2012	bin=0.5	1.37 (0.13)	5.53 (0.29)	1.03 (0.10)	4.01 (0.20)	0.34 (0.06)	1.52 (0.11)	359.64 (0.07)	359.28 (0.10)
	bin=0.25	0.88 (0.08)	1.75 (0.14)	0.83 (0.08)	1.55 (0.13)	0.05 (0.02)	0.20 (0.04)	359.72 (0.05)	359.81 (0.05)
	bin=0.1	0.34 (0.05)	0.50 (0.06)	0.34 (0.05)	0.48 (0.06)	0.00 (0.00)	0.02 (0.01)	359.71 (0.06)	359.87 (0.03)
SIE-GLM	bin=0.5	3.77 (0.21)	10.35 (0.24)	3.04 (0.17)	7.65 (0.17)	0.73 (0.08)	2.70 (0.13)	359.20 (0.09)	358.23 (0.16)
	bin=0.25	3.06 (0.18)	10.35 (0.23)	2.67 (0.16)	7.83 (0.17)	0.39 (0.05)	2.52 (0.12)	359.46 (0.08)	358.44 (0.14)
	bin=0.1	1.27 (0.11)	5.29 (0.25)	1.25 (0.10)	4.70 (0.20)	0.02 (0.01)	0.59 (0.07)	359.66 (0.05)	359.54 (0.07)
Raj_2005	parent=3	18.18 (0.11)	19.80 (0.04)	11.21 (0.08)	11.90 (0.03)	6.97 (0.09)	7.91 (0.02)	318.18 (0.11)	319.81 (0.04)
TRUE		20		12		8		360	

Table 10: Simulation of Network 2 with varying assumed time-lag width ϕ_a. The connection strength β is 0.5 and the time length $\mathrm{T} = 1000$. The true time-lag width $\phi = 1$. Results are averaged over 100 replications, with standard errors indicated in parentheses.

		Correct_All			Detected_A			Detected_B			Correct_NC		
assumed time-lag width =		0.5	1	1.5	0.5	1	1.5	0.5	1	1.5	0.5	1	1.5
Discrete_L1	bin=0.5	10.27 (0.32)	11.71 (0.25)	5.84 (0.24)	7.16 (0.21)	8.13 (0.18)	4.34 (0.17)	3.11 (0.14)	3.58 (0.12)	1.50 (0.10)	357.84 (0.18)	357.29 (0.17)	358.75 (0.13)
	bin=0.25	10.31 (0.28)	15.90 (0.17)	10.25 (0.29)	7.36 (0.19)	10.62 (0.12)	7.26 (0.20)	2.95 (0.11)	5.28 (0.10)	2.99 (0.13)	357.94 (0.15)	355.76 (0.22)	357.48 (0.18)
	bin=0.1	10.41 (0.27)	17.62 (0.14)	13.30 (0.26)	7.45 (0.19)	11.38 (0.09)	9.14 (0.17)	2.96 (0.12)	6.24 (0.10)	4.16 (0.13)	357.74 (0.16)	355.47 (0.25)	356.88 (0.19)
Continuous_L1		13.35 (0.27)	19.32 (0.09)	16.94 (0.17)	8.63 (0.16)	11.71 (0.05)	10.96 (0.11)	4.72 (0.15)	7.61 (0.06)	5.98 (0.10)	349.74 (0.36)	346.97 (0.47)	346.47 (0.42)
Discrete_AL	bin=0.5	14.03 (0.22)	15.37 (0.16)	10.24 (0.24)	9.25 (0.14)	10.17 (0.12)	7.05 (0.17)	4.78 (0.12)	5.20 (0.11)	3.19 (0.12)	357.23 (0.16)	356.91 (0.17)	357.15 (0.18)
	bin=0.25	14.28 (0.21)	17.95 (0.13)	14.15 (0.21)	9.36 (0.16)	11.37 (0.09)	9.52 (0.14)	4.92 (0.12)	6.58 (0.11)	4.63 (0.12)	356.72 (0.20)	356.95 (0.18)	356.86 (0.20)
	bin=0.1	14.22 (0.20)	18.80 (0.08)	16.32 (0.16)	9.34 (0.14)	11.69 (0.05)	10.72 (0.11)	4.88 (0.12)	7.11 (0.08)	5.61 (0.11)	356.71 (0.19)	356.94 (0.22)	356.35 (0.21)
Continuous_AL		14.48 (0.19)	19.28 (0.07)	17.03 (0.15)	9.55 (0.13)	11.82 (0.03)	11.10 (0.10)	4.93 (0.12)	7.46 (0.06)	5.93 (0.10)	357.10 (0.17)	357.17 (0.19)	356.78 (0.20)
Zhao_2012	bin=0.5	8.34 (0.33)	5.53 (0.29)	4.71 (0.25)	5.70 (0.22)	4.01 (0.20)	3.50 (0.18)	2.64 (0.13)	1.52 (0.11)	1.21 (0.10)	358.54 (0.13)	359.28 (0.10)	359.36 (0.10)
	bin=0.25	1.81 (0.14)	1.75 (0.14)	1.59 (0.14)	1.63 (0.13)	1.55 (0.13)	1.42 (0.12)	0.18 (0.04)	0.20 (0.04)	0.17 (0.04)	359.79 (0.04)	359.81 (0.05)	359.86 (0.04)
	bin=0.1	0.37 (0.06)	0.50 (0.06)	0.41 (0.06)	0.36 (0.06)	0.48 (0.06)	0.39 (0.05)	0.01 (0.00)	0.02 (0.01)	0.02 (0.01)	359.93 (0.02)	359.87 (0.03)	359.89 (0.03)
SIE-GLM	bin=0.5	9.24 (0.29)	10.35 (0.24)	8.99 (0.25)	6.86 (0.21)	7.65 (0.17)	6.74 (0.19)	2.38 (0.13)	2.70 (0.13)	2.25 (0.11)	358.99 (0.11)	358.23 (0.16)	358.41 (0.15)
	bin=0.25	6.38 (0.27)	10.35 (0.23)	8.11 (0.24)	5.11 (0.21)	7.83 (0.17)	6.49 (0.18)	1.27 (0.10)	2.52 (0.12)	1.63 (0.11)	359.20 (0.10)	358.44 (0.14)	358.93 (0.11)
	bin=0.1	2.60 (0.17)	5.29 (0.25)	3.08 (0.19)	2.45 (0.16)	4.70 (0.20)	2.88 (0.17)	0.15 (0.03)	0.59 (0.07)	0.20 (0.04)	359.75 (0.06)	359.54 (0.07)	359.68 (0.16)
Raj_2005	parent=3	18.12 (0.12)	19.80 (0.04)	19.10 (0.08)	11.14 (0.08)	11.90 (0.03)	11.73 (0.05)	6.98 (0.09)	7.91 (0.02)	7.37 (0.06)	318.16 (0.13)	319.81 (0.04)	319.10 (0.08)
TRUE			20			12			8			360	

Bibliography

Aertsen, A., G. Gerstein, M. Habib, and G. Palm (1989). Dynamics of neuronal firing correlation: modulation of" effective connectivity". *Journal of neurophysiology 61*(5), 900–917.

Bacry, E. and J.-F. Muzy (2014). Second order statistics characterization of hawkes processes and non-parametric estimation. preprint. *arXiv 1401*.

Barbieri, R., M. C. Quirk, L. M. Frank, M. A. Wilson, and E. N. Brown (2001). Construction and analysis of non-poisson stimulus-response models of neural spiking activity. *Journal of neuroscience methods 105*(1), 25–37.

Brémaud, P. (1981). *Point processes and queues: martingale dynamics*, Volume 50. Springer.

Brillinger, D. and A. Villa (1994, 01). Examples of the investigation of neural information processing by point process analysis.

Brown, E. N., R. E. Kass, and P. P. Mitra (2004). Multiple neural spike train data analysis: state-of-the-art and future challenges. *Nature Neuroscience 7*(5), 456–461.

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

Carstensen, L., A. Sandelin, O. Winther, and N. R. Hansen (2010). Multivariate hawkes process models of the occurrence of regulatory elements. *BMC Bioinformatics 11*(1), 456.

Chornoboy, E. S., L. P. Schramm, and A. F. Karr (1988). Maximum likelihood identification of neural point process systems. *Biol Cybern 59*(4-5), 265–275.

Daley, D. J. and D. Vere-Jones (2003). *An introduction to the theory of point processes: volume I: elementary theory and methods*. Springer.

Fischer, T. (2013). On simple representations of stopping times and stopping time sigma-algebras. *Statistics & Probability Letters 83*(1), 345–349.

Friedman, J., T. Hastie, H. Höfling, and R. Tibshirani (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics 1*(2), 302 – 332.

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

Harris, K. D., J. Csicsvari, H. Hirase, G. Dragoi, and G. Buzsáki (2003). Organization of cell assemblies in the hippocampus. *Nature 424*(6948), 552–556.

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

Kass, R. E., U. T. Eden, and E. N. Brown (2014). *Analysis of neural data*, Volume 491. Springer.

Kelly, R. C., R. E. Kass, M. A. Smith, and T. S. Lee (2010). Accounting for network effects in neuronal responses using l1 regularized point process models. *Advances in neural information processing systems 23*(2), 1099.

Krumin, M., I. Reutsky, and S. Shoham (2010). Correlation-based analysis and generation of multiple spike trains using hawkes models with an exogenous input. *Frontiers in computational neuroscience 4*, 147.

Masud, M. S. and R. Borisyuk (2011). Statistical technique for analysing functional connectivity of multiple spike trains. *Journal of Neuroscience Methods 196*(1), 201–219.

Murphy, K. P. (2002). Dynamic bayesian networks: Representation, inference and learning, dissertation. *PhD thesis, UC Berkley, Dept. Comp. Sci*.

Nieuwenhuis, G. (2013). Asymptotic mean stationarity and absolute continuity of point process distributions. *Bernoulli 19*(5A), 1612–1636.

Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics*, 758–765.

Ogata, Y. (1981). On lewis' simulation method for point processes. *IEEE transactions on information theory 27*(1), 23–31.

Oram, M., M. Wiener, R. Lestienne, and B. Richmond (1999). Stochastic nature of precisely timed spike patterns in visual system neuronal responses. *Journal of neurophysiology 81*(6), 3021–3033.

Rajaram, S., T. Graepel, and R. Herbrich (2005). Poisson-networks: A model for structured point processes. In *Proc. 10th Intl. Workshop on AI and Stat*.

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

Rubin, I. (1972). Regular point processes and their detection. *IEEE Transactions on Information Theory 18*(5), 547–557.

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

Truccolo, W., U. T. Eden, M. R. Fellows, J. P. Donoghue, and E. N. Brown (2005). A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *Journal of neurophysiology 93*(2), 1074–1089.

Xu, H., M. Farajtabar, and H. Zha (2016). Learning granger causality for hawkes processes. In *International Conference on Machine Learning*, pp. 1717–1726. PMLR.

Zhang, C., Y. Chai, X. Guo, M. Gao, D. Devilbiss, and Z. Zhang (2016). Statistical learning of neuronal functional connectivity. *Technometrics 58*(3), 350–359.

Zhang, C., Y. Jiang, and Y. Chai (2010). Penalized bregman divergence for large-dimensional regression and classification. *Biometrika 97*(3), 551–566.

Zhao, M., A. Batista, J. P. Cunningham, C. Chestek, Z. Rivera-Alvidrez, R. Kalmar, S. Ryu, K. Shenoy, and S. Iyengar (2012). An l 1-regularized logistic model for detecting short-term neuronal interactions. *Journal of computational neuroscience 32*(3), 479–497.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association 101*(476), 1418–1429.