Flexible modeling and estimation for high-dimensional graphs

Shizhe Chen

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

Doctor of Philosophy

University of Washington

2016

Reading Committee:

Ali Shojaie, Chair

Daniela Witten, Chair

Mathias Drton

Program Authorized to Offer Degree:
Department of Biostatistics

University of Washington

Abstract

Flexible modeling and estimation for high-dimensional graphs

Shizhe Chen

Co-Chairs of the Supervisory Committee:
Dr. Ali Shojaie
Biostatistics and Statistics

Dr. Daniela Witten
Biostatistics and Statistics

With the wealth of large-scale data arising from biology, the Internet, and social science, there is a growing need for exploratory tools for data analysis. It is often of interest to estimate the underlying graph of the variables. This dissertation focuses on developing flexible statistical models for complex graphs motivated by scientific questions in genome science and neuroscience. We investigate three types of graphical models: mixed graphical models, systems of additive ordinary differential equations, and multivariate Hawkes processes. For each type of graphical models, we discuss the properties of the graphical model and propose efficient statistical methods for recovering the graphical structure from high-dimensional data. Furthermore, we establish statistical guarantees of the proposed procedures and conduct extensive numerical experiments to evaluate their empirical performance.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Number Page

LIST OF TABLES

ACKNOWLEDGMENTS

I would first like to thank my advisors, Daniela Witten and Ali Shojaie, who have patiently guided me through my doctoral study. I have been heavily influenced by their passion for discovery, efforts in accessible writing, and rigorousness in research. I would also like to thank Mathias Drton and Eric Shea-Brown, who never failed to give me insightful comments. Throughout the years, I have benefited from generous advice from Xiao-Hua Zhou, Scott Emerson, Jon Wellner, Lurdes Inoue, Patrick Heagerty, and other faculty in departments of biostatistics and statistics at University of Washington. I also owe my thanks to the considerate staff in the department of biostatistics for making the department a welcoming home for students.

The completion of this dissertation would not have been possible without the continuous support from my friends and my family, to whom I am always in debt. In particular, I would like to thank my cohort in the department of biostatistics, who have made the past five years a wonderful experience for a stranger in a strange land.

DEDICATION

To my family

Chapter 1

INTRODUCTION

1.1 Motivation

With the wealth of large-scale data arising from biology, the Internet, and social science, there is a growing need for exploratory tools for data analysis. It is often of interest to estimate the underlying graph of the variables. In a graph, each variable is represented as a node, and a relationship between two variables is represented as an edge, directed or undirected, between the two nodes.

As a motivating example, we consider the task of estimating gene regulatory networks from two types of gene expression data. The first type of data consists of RNA expression levels collected from multiple subjects at a single time point. The relationship of interest is conditional dependence between expression of two genes. To be specific, there is an edge between gene A and gene B if and only if the expression levels of the two genes are dependent after conditioning on the expression levels of all other genes. The second type of data consists of RNA expression levels measured from a single subject at multiple time points. This can be modeled using ordinary differential equations (ODEs). Here a directed edge from gene A to gene B means that the change in gene B's expression depends on the expression of gene A. In either case, we want to learn the network from the data.

In the classical regression setting, simple linear models can be used to explore the data. In the analysis of graphs, similar "convenient modeling" strategies are also popular due to their simplicity, but unfortunately impose strong implicit assumptions. Consider, for instance, modeling the joint distribution of p random variables, $X_1, \ldots, X_p$, over an undirected graph. If we assume a linear model for the conditional mean functions, i.e. $E[X_j|X_{-j}] = X_{-j}^{\mathrm{T}}\beta_j$, for $j = 1, \ldots, p$, then the resulting joint distribution is multivariate normal under mild conditions (Khatri and Rao, 1976). As another example, consider a system of linear ODEs $X'(t) = AX(t)$, where $X(t)$ is a $p \times 1$ vector

and A is a $p \times p$ matrix. The solution to this system is $\exp(At)X(0)$, which may not be a good approximation to a biological system of interest. In both cases, the convenient model imposes very strong assumptions, which may not be appropriate for a given dataset and scientific problem, and hence not applicable for exploratory purposes.

1.2 Graphical Models

1.2.1 Definition

A graph G is a pair (V, E) of a node set V and an edge set E. The edge set E consists of ordered or unordered pairs of nodes. Specifically, for $i, j \in V$, we use an unordered pair $\{i, j\}$ to represent an undirected edge between node i and node j, and we use an ordered pair (i, j) to represent a directed edge, also known as an arc, from node i to node j. In this dissertation, we consider a general definition of graphical models.

Definition 1. A graphical model is a probabilistic model defined on a graph. Nodes in the graph are random elements, and edges in the graph represent relationships between the nodes they connect.

In the following sections, we discuss three graphical models that are special cases of Definition 1: conditional independence graphical models (Section 1.2.2), systems of ODEs (Section 1.2.3), and multivariate Hawkes point processes (Section 1.2.4).

1.2.2 Conditional independence graphical models

The conditional independence graphical model, or simply the conditional independence graph (CIG), is a probabilistic model defined on an undirected graph. Nodes in the CIG are random variables which follow a joint distribution $P(\cdot)$, and an edge represents a conditional dependence relationship between a pair of nodes (Dawid, 1979). That is, the edge $\{X_j, X_k\} \in E$ if only if

$$P(X_k, X_j | X_{-\{k,j\}}) \neq P(X_k | X_{-\{k,j\}}) P(X_j | X_{-\{k,j\}}). \tag{1.1}$$

Many CIG models have been proposed to deal with specific scientific questions, and some have been extensively studied in the classical statistics literature. These include Gaussian graphical

models (Dempster, 1972), Ising models (Onsager, 1944), and conditional-Gaussian models (Lauritzen, 1996). In many applications, the structure of the graph – that is, the edge set E – is known. For instance, in spatial statistics, CIG models are often used to account for spatial auto-correlation, where the graphs are known from geographical information (Besag, 1974; Ripley, 2005; Cressie, 1991).

1.2.3 Graphical models of systems of ordinary differential equations

We consider a system of ODEs, whose solutions are measured with error at discrete time points. We observe $Y_1, \ldots, Y_n \in \mathbb{R}^p$, where $Y_i = X(t_i; \theta) + \epsilon_i,\ i = 1, \ldots, n$, and

$$X_j'(t; \theta) = f_j(X(t; \theta), \theta); \quad t \in [0, 1],\ j = 1, \ldots, p. \tag{1.2}$$

A directed edge from $X_k(\cdot)$ to $X_j(\cdot)$ indicates that $X_k(\cdot)$ is involved in the change of $X_j(\cdot)$, or

$$\frac{\partial f_j(X(t; \theta), \theta)}{\partial X_k(t; \theta)} \neq 0 \text{ for some } t \in [0, 1]. \tag{1.3}$$

ODEs are popular in modeling dynamics of complex systems in a number of applications. It is often the case that the form of the functions $f_j(\cdot), j = 1, \ldots, p$, are known, and hence E is known. The questions of interest are how to estimate the unknown parameters θ, and how well the system explains real world dynamics (Benson, 1979; Biegler et al., 1986; Varah, 1982; Ramsay et al., 2007; Liang and Wu, 2008; Xue et al., 2010; Brunel, 2008; Qi and Zhao, 2010; Gugushvili and Klaassen, 2012; Hall and Ma, 2014). For instance, compartmental models in epidemiology employ ODEs to model the dynamical relationships among a susceptible population (S), an infectious population (I), and a recovered population (R). This is known as the SIR model (Gibson and Renshaw, 1998; Ionides et al., 2006; M'Kendrick, 1925). In the SIR model, S, I, and R are the nodes of the graph, and the ODE assumed to take the form

$$S' = -\beta IS/N;\ I' = \beta IS/N - \gamma I;\ R' = \gamma I, \tag{1.4}$$

where $N \equiv S(t) + R(t) + I(t)$ is the total population, β describes the infectiousness of the disease, and γ the recovery rate. The SIR model can be used to study the effect of vaccination, estimate the

spread rate of a pandemic, or simulate an outbreak. In this application, the graphical structure and the parametric form of the functions f_j in the ODE system are defined by the underlying model. However, in many high-dimensional applications, such information is unavailable. Motivated by the problem of estimation of gene regulatory networks, in Chapter 3 we consider the setting where the parametric form of the functions f_j in the ODE system is not specified. In that case, estimation of graph structure, i.e. nonzero fs is of main interest.

1.2.4 Graphical models of multivariate Hawkes processes

We consider a p-variate Hawkes process first proposed by Hawkes (1971). The intensity function of the Hawkes process takes the form

$$\lambda_j(t) = \mu_j + \sum_{k=1}^{p} \big(\omega_{j,k} * dN_k\big)(t), \quad j = 1, \ldots, p, \tag{1.5}$$

where

$$\big(\omega_{j,k} * dN_k\big)(t) \equiv \int_0^\infty \omega_{j,k}(\Delta) \sum_{i:t_{k,i}\le t} \delta(t - \Delta - t_{k,i}) d\Delta = \sum_{i:t_{k,i}\le t} \omega_{j,k}(t - t_{k,i}).$$

We refer to $\mu_j \in \mathbb{R}$ as the *background intensity*, and $\omega_{j,k}(\cdot) : \mathbb{R}^+ \mapsto \mathbb{R}$ as the *transfer function*. The right-hand side of (4.4) is sometimes transformed by a *link function*, as in a generalized linear model; this leads to a non-linear Hawkes process.

We can see that the kth process affects the intensity of the jth process if $\omega_{j,k}(\Delta) \neq 0$ for some Δ. We can thus define a directed graph $G = (V, E)$ with the node set $V = \{1, \ldots, p\}$ and the edge set

$$E \equiv \{(j, k) : \exists\ \Delta \text{ such that } \omega_{j,k}(\Delta) \neq 0, 1 \le j, k \le p\}, \tag{1.6}$$

for $\omega_{j,k}$ given in (4.4). Recently, authors have connected G to the notions of directed information (Quinn et al., 2010) and Granger causality (Eichler et al., 2015).

The Hawkes process has been widely applied in modeling recurrent events, such as earthquakes (Ogata, 1988), crime rates (Mohler et al., 2011), interactions in social networks (Simma and Jordan, 2012; Perry and Wolfe, 2013), financial events (Chavez-Demoulin et al., 2005; Bowsher, 2007;

Aït-Sahalia et al., 2015), and spiking histories of neurons (see e.g., Brillinger, 1988; Okatan et al., 2005; Paninski et al., 2007; Pillow et al., 2008).

1.3 Dimensionality

As noted in the previous section, the classical literature focuses on using graphical models consisting of just a few nodes to approximate or describe real world problems. The structure of the graph is typically assumed to be known, based on scientific knowledge and experience accumulated over the years. Instead, we would like to learn the structure of the graph from the data. In other words, we would like to select the best model out of a family of candidate models. To achieve this goal, we need the family of candidate models to be flexible, the model selection procedure to be computationally efficient, and the selected graph to be reliable.

Recall that our motivating example is a gene regulatory network, which might have hundreds or thousands of nodes. The dimensionality of this problem leads to challenges, many of which are shared with supervised learning problems of the same dimension. We now digress to discuss high-dimensional regression, which is the main tool we will use in this dissertation.

The first issue is the curse of dimensionality associated with flexible modeling. Suppose that we are regressing an outcome y over p binary predictors. It is known that a fully-nonparametric model, i.e. a model with no assumptions, has at least 2^p parameters, each corresponding to one combination of values of the p binary predictors. The situation is even worse when the predictors are continuous. Such a model is not affordable even for moderate value of p. We have to introduce certain restrictions to the family of models that we consider. For instance, we could assume a linear model $\mathbb{E}y = X\beta$, an additive model $\mathbb{E}y = \sum_{i=1}^{p} f_i(X_i)$, or a regression tree model. These models offer some flexibility but also required certain assumptions. It is important to understand how robust they are in the context of variable selection.

In many contemporary setting, the parameters can outnumber the observations. This is known as the "large p small n" scenario. In gene regulatory networks, the number of genes easily exceeds the number of observations. Even if we fit a simple linear regression model with no interactions, parameters are still unidentifiable, and the estimates are thus not interpretable. To solve this issue,

penalization has been introduced to shrink some parameters towards zero (Tibshirani, 1996; Candes and Tao, 2007; Fan and Li, 2001). Among them, the ℓ_1-penalty has been particularly popular because of its parsimonious effect where some parameters are shrunken exactly to zero, which naturally leads to variable selection (Tibshirani, 1996). The theoretical properties of this penalty are well-studied (Wainwright and Jordan, 2008; Negahban et al., 2012; Lee et al., 2013), and efficient algorithms have been developed for fitting ℓ_1-penalized models (Efron et al., 2004; Friedman et al., 2010).

The last issue associated with dimension is the computational cost. Many of the techniques for learning the structure of graphical models in the classical statistics literature are quite complicated, especially for systems of ODEs. Some are already computationally challenging with moderate number of parameters. Thus, these techniques are not feasible for exploratory purposes on graphs of large scales.

1.4 Our Contributions

This dissertation will focus on the development of models for complex graphs. We will use current tools in high-dimensional regression to reconstruct graphs from high-dimensional data. The dissertation is organized as follows.

1. In Chapter 2, we consider the task of learning the structure of a CIG. We develop a general class of undirected graphical models that contains most existing parametric undirected graphical models as special cases. We discuss the existence of a joint distribution using the conditionally-specified modeling approach, and reveal implicit restrictions on the parameter space. We also propose an efficient procedure for reconstructing the graph for this type of model, and establish its variable selection consistency.

2. In Chapter 3, we study the graph reconstruction problem in a high-dimensional system of ODEs subjected to measurement error. Existing studies primarily focus on parameter estimations for ODEs of a known form. We propose a graph recovery procedure that allows for

the underlying ODE system to be additive and of unknown form. The proposed procedure is both computationally efficient and theoretically appealing, compared to existing methods.

3. In Chapter 4, we study the multivariate Hawkes process, where past events affect the future intensity of the process. We focus our discussion on the estimation of the directed graph encoded in a high-dimensional Hawkes process. We propose a neighbourhood selection procedure that recovers the true graph with high probability. Furthermore, we propose two screening methods that greatly reduce the computational cost while achieving desired statistical properties under mild conditions.

Chapter 2

MIXED GRAPHICAL MODELS

The following work is to appear in *Biometrika* (Chen et al., 2015).

2.1 Introduction

In this study, we consider the task of learning the structure of an undirected graphical model encoding pairwise conditional dependence relationships among random variables. Specifically, suppose that we have p random variables represented as nodes of the graph $G = (V, E)$, with the vertex set $V = \{1, \ldots, p\}$ and the edge set $E \subseteq V \times V$. An edge in the graph indicates a pair of random variables that are conditionally dependent given all other variables. The problem of reconstructing the graph from a set of n observations has attracted a lot of interest in recent years, especially when $p > n$ and $p(p-1)/2$ edges must be estimated from n observations.

Many authors have studied the estimation of high-dimensional undirected graphical models in the setting where the distribution of each node, conditioned on all other nodes, has the same parametric form. In particular, Gaussian graphical models have been studied extensively (Friedman et al., 2008; Meinshausen and Bühlmann, 2006; Peng et al., 2009; Ravikumar et al., 2011; Rothman et al., 2008; Wainwright and Jordan, 2008; Yuan and Lin, 2007), and have been generalized to account for non-normality and outliers (Finegold and Drton, 2011; Miyamura and Kano, 2006; Vogel and Fried, 2011; Sun and Li, 2012). Others have considered the setting in which all node-conditional distributions are Bernoulli (Höfling and Tibshirani, 2009; Lee et al., 2007; Ravikumar et al., 2010), multinomial (Jalali et al., 2011), Poisson (Allen and Liu, 2012), or any univariate distribution in the exponential family (Yang et al., 2012). An extended version of Yang et al. (2012) is available in an unpublished technical report.

In this study, we seek to estimate a graphical model in which the variables are of different types.

Here, the type of a node refers to the parametric form of its distribution, conditioned on all other nodes. For instance, the variables might include DNA nucleotides, taking binary values, and gene expression measured using RNA-sequencing, taking non-negative integer values. We could model the first set of nodes as Bernoulli, which means that each of their distributions, conditional on the other nodes, is Bernoulli; similarly, we could model the second set as Poisson. We assume that the type of each node is known a priori, and refer to this setup as a mixed graphical model.

In the low-dimensional setting, Lauritzen (1996) studied a special case of the mixed graphical model, known as the conditional Gaussian model, in which each node is either Gaussian or Bernoulli. More recent work has focused on the high-dimensional setting. Lee and Hastie (2014) proposed two algorithms for reconstructing conditional Gaussian models using a group lasso penalty. Cheng et al. (2013) modified this approach by using a weighted ℓ_1-penalty.

A related line of research considers semi-parametric or non-parametric approaches for estimating conditional dependence relationships (Fellinghauer et al., 2013; Liu et al., 2009; Voorman et al., 2014; Xue and Zou, 2012), among which Fellinghauer et al. (2013) is specifically proposed for mixed graphical models. However, despite their flexibility, these non-parametric methods are often less efficient than their parametric counterparts, if the type of each node is known.

In this study, we propose an estimator and develop theory for the parametric mixed graphical model, under a much more general setting than existing approaches (e.g. Lee and Hastie 2014). We allow the conditional distribution of each node to belong to the exponential family. Unlike Yang et al. (2012), nodes may be of different types. For instance, within a single graph, some nodes may be Bernoulli, some may be Poisson, and some may be exponential.

In parallel efforts, Yang et al. (2014) recently presented general results on strong compatibility for mixed graphical models for which the node-conditional distributions belong to the exponential family, and for which the graph contains only two types of nodes. We instead consider the setting where the graph can contain more than two types of nodes, and provide specific requirements for strong compatibility for some common distributions.

2.2 A Model for Mixed Data

2.2.1 Conditionally-Specified Models for Mixed Data

We consider the pairwise graphical model (Wainwright et al., 2007), which takes the form

$$p(x) \propto \exp\left\{\sum_{s=1}^{p} f_s(x_s) + \sum_{s=2}^{p}\sum_{t<s} f_{ts}(x_s, x_t)\right\}, \tag{2.1}$$

where $x = (x_1, ..., x_p)^{\mathrm{T}}$ and $f_{ts} = 0$ for $\{t, s\} \notin E$. Here, $f_s(x_s)$ is the node potential function, and $f_{st}(x_s, x_t)$ the edge potential function. We further simplify the pairwise interactions by assuming that $f_{st}(x_s, x_t) = \theta_{st}x_s x_t = \theta_{ts}x_s x_t$, so that we can write the parameters associated with edges in a symmetric square matrix $\Theta = (\theta_{st})_{p\times p}$ where the diagonal elements equal zero. The joint density can then be written as

$$p(x) = \exp\left\{\sum_{s=1}^{p} f_s(x_s) + \frac{1}{2}\sum_{s=1}^{p}\sum_{t\neq s} \theta_{ts}x_s x_t - A(\Theta, \alpha)\right\}, \tag{2.2}$$

where $A(\Theta, \alpha)$ is the log-partition function, a function of Θ and α. Here α is a $K \times p$ matrix of parameters involved in the node potential functions: that is, $f_s(x_s)$ involves α_s, the sth column of α. K is some known integer. For $\{s, t\} \notin E$, the edge potentials satisfy $\theta_{st} = \theta_{ts} = 0$. We define the neighbours of the sth node as $N(x_s) = \{t : \theta_{st} = \theta_{ts} \neq 0\}$.

In principle, given a parametric form for the joint density (2.2), we can estimate the conditional dependence relationships among the p variables, and hence the edges in the graph. But this approach requires the calculation of the log-partition function $A(\Theta, \alpha)$, which is often intractable. To overcome this, we instead use the framework of conditionally-specified models (Besag, 1974): we specify the distribution of each node conditional on the others, and then combine the p conditional distributions to form a single graphical model. This approach has been widely used in estimating high-dimensional graphical models where all nodes are of the same type (Allen and Liu, 2012; Meinshausen and Bühlmann, 2006; Ravikumar et al., 2010; Yang et al., 2012). However, as we will discuss in Section 2.2.2, a conditionally-specified model may not correspond to a valid joint distribution.

Define $x_{-s} = (x_1, ..., x_{s-1}, x_{s+1}, ..., x_p)^{\mathrm{T}}$. We consider conditional densities of the form

$$p(x_s \mid x_{-s}) = \exp\left\{ f_s(x_s) + \sum_{t \neq s} \theta_{ts} x_t x_s - D_s(\eta_s) \right\}, \tag{2.3}$$

where $\eta_s = \eta_s(\Theta_s, x_{-s}, \alpha_s)$ is a function of α_s, x_{-s}, and Θ_s, and Θ_s is the sth column of Θ without the diagonal element. Suppose $f_s(x_s) = \alpha_{1s} x_s + \alpha_{2s} x_s^2/2 + \sum_{k=3}^{K} \alpha_{ks} B_{ks}(x_s)$, where α_{ks} is a parameter, which could be 0, and $B_{ks}(x_s)$ is a known function for $k = 3, \ldots, K$. Under this assumption, (2.3) belongs to the exponential family.

The assumed form of $f_s(x_s)$ is quite general. We now consider some special cases of (2.3) corresponding to commonly-used distributions in the exponential family, for which $f_s(x_s)$ takes a very simple form. In the following examples, we assume that $\eta_s(\Theta_s, x_{-s}, \alpha_s) = \alpha_{1s} + \sum_{t:\, t \neq s} \theta_{ts} x_t$.

Example 1. The conditional density is Gaussian and $\alpha_{2s} = -1$:

$$p(x_s \mid x_{-s}) = \exp\left\{ -\frac{1}{2} x_s^2 + \eta_s x_s - \frac{1}{2} \eta_s^2 - \frac{1}{2} \log(2\pi) \right\}, \quad x_s \in \mathcal{R}, \tag{2.4}$$

where $f_s(x_s) = \alpha_{1s} x_s - x_s^2/2$ and $D_s(\eta_s) = \eta_s^2/2 + \log(2\pi)/2$.

Example 2. The conditional density is Bernoulli. Instead of coding x_s as $\{0, 1\}$, we code x_s as $\{-1, 1\}$. This yields the conditional density

$$p(x_s \mid x_{-s}) = \exp\left\{ \eta_s x_s - D_s(\eta_s) \right\}, \quad x_s \in \{-1, 1\}, \tag{2.5}$$

where $f_s(x_s) = \alpha_{1s} x_s$ and $D_s(\eta_s) = \log\{\exp(\eta_s) + \exp(-\eta_s)\}$.

Example 3. The conditional density is Poisson:

$$p(x_s \mid x_{-s}) = \exp\left\{ \eta_s x_s - \log(x_s!) - D_s(\eta_s) \right\}, \quad x_s \in \{0, 1, \ldots\}, \tag{2.6}$$

where $f_s(x_s) = \alpha_{1s} x_s - \log(x_s!)$ and $D_s(\eta_s) = \exp(\eta_s)$.

Example 4. The conditional density is exponential:

$$p(x_s \mid x_{-s}) = \exp\left\{ \eta_s x_s - D_s(\eta_s) \right\}, \quad x_s \in \mathcal{R}^+, \tag{2.7}$$

where $f_s(x_s) = \alpha_{1s} x_s$ and $D_s(\eta_s) = -\log(-\eta_s)$.

These four examples have been studied in the context of conditionally-specified graphical models in which all nodes are of the same type (Allen and Liu, 2012; Besag, 1974; Meinshausen and Bühlmann, 2006; Ravikumar et al., 2010; Yang et al., 2012).

In what follows, we will consider the conditionally-specified mixed graphical model, with conditional distributions given by (2.3), in which each node can be of a different type. This class of mixed graphical models is not closed under marginalization: for instance, given a graph composed of Gaussian and Bernoulli nodes, integrating out the Bernoulli nodes leads to a conditional density that is a mixture of Gaussians, which does not belong to the exponential family.

2.2.2 Compatibility of Conditionally-Specified Models

Under what circumstances does the conditionally-specified model with node-conditional distributions given in (2.3) correspond to a well-defined joint distribution? We first adapt and restate a definition from Wang and Ip (2008), which applies to any conditional density.

Definition 2. A non-negative function g is capable of generating a conditional density function $p(y \mid x)$ if

$$p(y \mid x) = \frac{g(y, x)}{\int g(y, x) dy}.$$

Two conditional densities are said to be compatible if there exists a function g that is capable of generating both conditional densities. When g is a density, the conditional densities are called strongly compatible.

The following proposition relates Definition 2 to the conditional density in (2.3). Its proof, and those of other statements in this study, are available in the Appendix.

Proposition 1. *Let $x = (x_1, ..., x_p)^{\mathrm{T}}$ be a random vector. Suppose that for each x_s, the conditional density takes the form of* (2.3)*. If $\theta_{st} = \theta_{ts}$, then the conditional densities are compatible. Furthermore, any function g that is capable of generating the conditional densities is of the form*

$$g(x) \propto \exp\left\{ \sum_{s=1}^{p} f_s(x_s) + \frac{1}{2} \sum_{s=1}^{p} \sum_{t \neq s} \theta_{ts} x_s x_t \right\}. \tag{2.8}$$

Under the conditions of Proposition 1, if we further assume that g in (2.8) is integrable, then by Definition 2, the conditional densities of the form (2.3) are strongly compatible. Proposition 1 indicates that, provided that (2.2) is a valid joint distribution, we can arrive at it via the conditional densities in (2.3). This justifies the conditionally-specified modeling approach taken in this study. Proposition 1 is closely related to Section 4·3 in Besag (1974) and Proposition 1 in Yang et al. (2012), with small modifications. More general theory is developed in Wang and Ip (2008).

We now return to the four examples (2.4)–(2.7). Lemma 1 summarizes the conditions under which a conditionally-specified model with non-degenerate conditional distributions of the form (2.4)–(2.7) leads to a valid joint distribution.

Lemma 1. *If $\theta_{st} = \theta_{ts}$, the subset of conditions with a dagger (†) in Table 2.1 is necessary and sufficient for the conditional densities in* (2.4)–(2.7) *to be compatible. Moreover, the complete set of conditions in Table 2.1 is necessary and sufficient for the conditional densities in* (2.4)–(2.7) *to be strongly compatible.*

To simplify the presentation of the conditions for the Gaussian nodes, in Table 2.1 it is assumed that J is the index set of the Gaussian nodes. Without loss of generality, we further assume that the nodes are ordered such that $J = \{1, \ldots, m\}$, and define

$$\Theta_{JJ} = \begin{pmatrix} \alpha_{21} & \theta_{12} & \cdots & \theta_{1m} \\ \theta_{21} & \alpha_{22} & \cdots & \theta_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{m1} & \theta_{m2} & \cdots & \alpha_{2m} \end{pmatrix}. \tag{2.9}$$

Table 1 reveals the set of restrictions on the parameter space that must hold in order for the conditional densities in (2.4)–(2.7) to be compatible or strongly compatible. The diagonal entries of this table were previously studied in Besag (1974). In general, strong compatibility imposes more restrictions on the parameter space than compatibility. For instance, compatibility does not place any restrictions on edges between two Poisson nodes, but for strong compatibility to hold, the edge potentials must be negative. Compatibility and strong compatibility even restrict the relationships

Table 2.1: Restrictions on the parameter space required for compatibility or strong compatibility of the conditional densities in (2.4)–(2.7)

	Gaussian	Poisson	Exponential	Bernoulli
Gaussian	$\Theta_{JJ} \prec 0$	$\theta_{ts} = 0$	$\theta_{ts} = 0^\dagger$	$\theta_{ts} \in \mathcal{R}^\dagger$
Poisson		$\theta_{ts} \leq 0$	$\theta_{ts} \leq 0^\dagger$	$\theta_{ts} \in \mathcal{R}^\dagger$
Exponential			$\theta_{ts} \leq 0^\dagger$	$\sum_{s \in I} \lvert\theta_{st}\rvert < -\alpha_{1t}^\dagger$
Bernoulli				$\theta_{ts} \in \mathcal{R}^\dagger$

The column specifies the type of the sth node, and the row specifies the type of the tth node. Conditions marked with a dagger (†) are necessary and sufficient for the conditional densities in (2.4)–(2.7) to be compatible, and the complete set of conditions is necessary and sufficient for the conditional densities to be strongly compatible. For compatibility to hold for a Gaussian node x_s, $\alpha_{2s} < 0$ is also required. Here Θ_{JJ} is as defined in (2.9), and I denotes the set of Bernoulli nodes.

that can be modeled using the conditional densities (2.4)–(2.7): for instance, no edges are possible between Gaussian and exponential nodes, or between Gaussian and Poisson nodes.

To summarize, given conditional densities of the form (2.4)–(2.7), existence of a joint density imposes substantial constraints on the parameter space, and thus limits the flexibility of the corresponding graph. However, we will see in Section 2.5 that it is possible to consistently estimate the structure of a graph even when the requirements for compatibility or strong compatibility are violated, i.e., even in the absence of a joint density.

While Table 2.1 only examines conditionally-specified models composed of the conditional densities in (2.4)–(2.7), the estimator proposed in Section 2.3 and the theory developed in Sections 2.4 and 2.5 apply to other types of conditional densities of the form (2.3).

2.3 Estimation via Neighbourhood Selection

2.3.1 Estimation

We now present a neighbourhood selection approach for recovering the structure of a mixed graphical model, by maximizing penalized conditional likelihoods node-by-node. A similar approach has been studied in the setting where all nodes in the graph are of the same type (Meinshausen and Bühlmann, 2006; Ravikumar et al., 2010; Allen and Liu, 2012; Yang et al., 2012).

Recall from Section 2.2.1 that $f_s(x_s) = \alpha_{1s}x_s + \alpha_{2s}x_s^2/2 + \sum_{k=3}^{K} \alpha_{ks}B_{ks}(x_s)$. We now simplify the problem by assuming that α_{ks} is known, and possibly zero, for $k \geq 2$. Let X denote an $n \times p$ data matrix, with the ith row given by $x^{(i)}$. From now on, we use an asterisk to denote the true parameter values. We estimate Θ_s^* and α_{1s}^*, the parameters for the sth node, as

$$\underset{\Theta_s \in \mathcal{R}^{p-1},\ \alpha_{1s} \in \mathcal{R}}{\arg\min} \quad -\ell_s(\Theta_s, \alpha_{1s}; X) + \lambda_n \|\Theta_s\|_1, \tag{2.10}$$

where $\ell_s(\Theta_s, \alpha_{1s}; X) = \sum_{i=1}^{n} \log p(x_s^{(i)} \mid x_{-s}^{(i)})/n$; recall that the conditional density $p(x_s^{(i)} \mid x_{-s}^{(i)})$ is defined in (2.3). Finally, we define the estimated neighbourhood of x_s to be $\hat{N}(x_s) = \{t : \hat{\theta}_{ts} \neq 0\}$, where $\hat{\Theta}_s$ solves (2.10), and $\hat{\theta}_{ts}$ is the element corresponding to an edge with the tth node.

In practice, to avoid a situation in which variables of different types are on different scales, we may wish to modify (2.10) in order to allow a different weight for the ℓ_1-penalty on each coefficient. We define a weight vector w equal to the empirical standard errors of the corresponding variables: $w = (\hat{\sigma}_1, ..., \hat{\sigma}_{s-1}, \hat{\sigma}_{s+1}, ..., \hat{\sigma}_p)^{\mathrm{T}}$. Then (2.10) can be replaced with

$$\underset{\Theta_s \in \mathcal{R}^{p-1},\ \alpha_{1s} \in \mathcal{R}}{\arg\min} \quad -\ell_s(\Theta_s, \alpha_{1s}; X) + \lambda_n \|\mathrm{diag}(w)\Theta_s\|_1. \tag{2.11}$$

The analysis in Sections 2.4 and 2.5 uses (2.10) for simplicity, but could be generalized to (2.11) with additional bookkeeping. Both (2.10) and (2.11) can be easily solved (see e.g. (Friedman et al., 2010)).

In the joint density (2.2), the parameter matrix Θ is symmetric, i.e., $\theta_{st} = \theta_{ts}$, but the neighbourhood selection method does not guarantee symmetric estimates: for instance, it could happen

that $\hat{\theta}_{st} = 0$ but $\hat{\theta}_{ts} \neq 0$. Our analysis in Section 2.4.2 shows that we can exploit the asymmetry in $\hat{\theta}_{st}$ and $\hat{\theta}_{ts}$ when x_s and x_t are of different types, in order to obtain more efficient edge estimates.

2.3.2 Tuning

In order to select the value of the tuning parameter λ_n in (2.10), we use the Bayesian information criterion (Zou et al., 2007; Peng et al., 2009; Voorman et al., 2014), which takes the form

$$\text{BIC}_s(\lambda_n) = -2n\ell_s(\hat{\Theta}_s, \hat{\alpha}_{1s}; X) + \log(n)\|\hat{\Theta}_s\|_0, \tag{2.12}$$

where $\|\hat{\Theta}_s\|_0$ is the number of non-zero elements in $\hat{\Theta}_s$ for a given value of λ_n. We allow a different value of λ_n for each node type. For instance, to select λ_n for the Poisson nodes, we choose the value of λ_n such that $\text{BIC}_s(\lambda_n)$, summed over the Poisson nodes, is minimized. We evaluate the performance of this approach for tuning parameter selection in Section 2.6.3.

2.4 Recovery with Strongly Compatible Conditional Distributions

2.4.1 Neighbourhood Recovery

In this subsection we show that if the conditional distributions in (2.3) are strongly compatible, as they will be under conditions discussed in Section 2.2.2, then under some additional assumptions, the true neighbourhood of each node is consistently selected using the neighbourhood selection approach proposed in Section 2.3.1. Here we rely heavily on results from Yang et al. (2012), who consider a related problem in which all nodes are of the same type.

In the following discussion, we assume that $p > n$ for simplicity. For any s, let Δ_s denote the set of indices for elements of $(\Theta_s^{\mathrm{T}}, \alpha_{1s})^{\mathrm{T}}$ that correspond to non-neighbours of the sth node, and let $Q_s^* = -\nabla^2 \ell_s(\Theta_s^*, \alpha_{1s}^*; X)$ be the negative Hessian of $\ell_s(\Theta_s, \alpha_{1s}; X)$ with respect to $(\Theta_s^{\mathrm{T}}, \alpha_{1s})^{\mathrm{T}}$, evaluated at the true values of the parameters. Below we suppress the subscript s for simplicity, and we remind the reader that all quantities are related to the conditional density of the sth node. We express Q^* in blocks:

$$Q^* = \begin{pmatrix} Q^*_{\Delta^c\Delta^c} & Q^*_{\Delta^c\Delta} \\ Q^*_{\Delta\Delta^c} & Q^*_{\Delta\Delta} \end{pmatrix}.$$

Assumption 1. There exists a positive number a such that

$$\max_{l \in \Delta} \|Q^*_{l\Delta^c}(Q^*_{\Delta^c\Delta^c})^{-1}\|_1 \leq 1 - a.$$

Assumption 1 limits the association between the neighbours and non-neighbours of the sth node: if the association is too high, then it is not possible to select the correct neighbourhood. This type of assumption is standard for variable selection consistency of ℓ_1-penalized estimators (Lee et al., 2013; Meinshausen and Bühlmann, 2006; Ravikumar et al., 2010, 2011; Wainwright, 2009; Yang et al., 2012; Zhao and Yu, 2006).

Assumption 2. There exists $\Lambda_1 > 0$ such that the smallest eigenvalue of $Q^*_{\Delta^c\Delta^c}$, $\Lambda_{\min}(Q^*_{\Delta^c\Delta^c})$, is greater than or equal to Λ_1. Also, there exists $\Lambda_2 < \infty$ such that the largest eigenvalue of $\sum_{i=1}^n x_0^{(i)}(x_0^{(i)})^{\mathrm{T}}/n$, $\Lambda_{\max}\left\{\sum_{i=1}^n x_0^{(i)}(x_0^{(i)})^{\mathrm{T}}/n\right\}$, is less than or equal to Λ_2, where $x_0 = (x_{-s}^{\mathrm{T}}, 1)^{\mathrm{T}}$.

The lower bound in Assumption 2 is needed to prevent singularity among the true neighbours, which would prevent neighbourhood recovery. The bound on the largest eigenvalue of the sample covariance matrix is needed to prevent a situation where most of the variance in the data is due to a single feature. Similar assumptions are made in Meinshausen and Bühlmann (2006); Ravikumar et al. (2010); Wainwright (2009); Yang et al. (2012); Zhao and Yu (2006).

Assumption 3. The log-partition function $D(\cdot)$ of the conditional density $p(x_s \mid x_{-s})$ is third-order differentiable, and there exist κ_2 and κ_3 such that $|D''(y)| \leq \kappa_2$ and $|D'''(y)| \leq \kappa_3$ for $y \in \{y : y \in \mathcal{D},\ |y| \leq M\delta_1 \log p\}$, where $\mathcal{D}$ is the support of $D(\cdot)$.

Remark 1. The two quantities κ_2 and κ_3 are functions of p. The quantity δ_1 is a constant to be chosen in Proposition 2. The constant M is a sufficiently large constant that plays a role in Assumption 6.

Assumption 3 controls the smoothness of the log-partition function $D(\cdot)$ for conditional densities of the form (2.3). Recall from Section 2.2.1 that the log-partition function of the node x_s is

$D(\eta_s)$, where η_s equals $\alpha_{1s} + \sum_{t \neq s} \theta_{ts} x_t$. To apply Assumption 3 to $D(\eta_s)$, we will need to bound $\sum_{t \neq s} \theta_{ts} x_t$, so that $|\eta_s| \leq M \delta_1 \log(p)$.

In order to obtain such a bound, we need another assumption.

Assumption 4. Assume that, for $t = 1, ..., p$, (i) $|E(x_t)| \leq \kappa_m$, (ii) $E(x_t^2) \leq \kappa_v$, and (iii)

$$\max_{u:|u|\leq 1} \frac{\partial^2 A}{\partial \alpha_{1t}^2}\bigg|_{\alpha_{1t}^* + u} \leq \kappa_h, \quad \max_{u:|u|\leq 1} \frac{\partial^2 A}{\partial \alpha_{2t}^2}\bigg|_{\alpha_{2t}^* + u} \leq \kappa_h.$$

Assumption 4 controls the moments of each node, as well as the local smoothness of the log-partition function A in (2.2). Given Assumption 4, the following propositions on the marginal behaviour of random variables hold; see Propositions 3 and 4 in Yang et al. (2012).

Proposition 2. *Define the event*

$$\xi_1 = \left(\max_{i \in \{1,...,n\}; t \in \{1,...,p\}} |x_t^{(i)}| < \delta_1 \log p \right).$$

Assuming $p > n$, *pr*$(\xi_1) \geq 1 - c_1 p^{-\delta_1 + 2}$, *where* $c_1 = \exp(\kappa_m + \kappa_h/2)$.

Proposition 3. *Define the event*

$$\xi_2 = \left[\max_{t \in \{1,...,p\}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (x_t^{(i)})^2 \right\} < \delta_2 \right],$$

where $\delta_2 \geq 1$. *If* $\delta_2 \leq \min(2\kappa_v/3, \kappa_h + \kappa_v)$, *and* $n \geq 8\kappa_h^2 \log p / \delta_2^2$, *then pr*$(\xi_2) \geq 1 - \exp(-c_2 \delta_2^2 n)$, *where* $c_2 = 1/(4\kappa_h^2)$.

We now present three additional assumptions that relate to the node-wise regression in (2.10).

Assumption 5. The minimum of edge potentials related to node x_s, $\min_{t \in N(x_s)} |\theta_{ts}|$, is larger than $10(d+1)^{1/2} \lambda_n / \Lambda_1$, where d is the number of neighbours of x_s.

Assumption 6. The tuning parameter λ_n is in the range

$$\left[\frac{8(2-a)}{a} \left\{ \delta_2 \kappa_2 \frac{\log(2p)}{n} \right\}^{1/2}, \min \left\{ \frac{2(2-a)}{a} \kappa_2 \delta_2 M, \frac{a \Lambda_1^2 (d+1)^{-1}}{288(2-a)\kappa_2 \Lambda_2}, \frac{\Lambda_1^2 (d+1)^{-1}}{12 \Lambda_2 \kappa_3 \delta_1 \log p} \right\} \right]. \quad (2.13)$$

Remark 2. Of the three quantities in the upper bound of λ_n, $\Lambda_1^2/\{12\Lambda_2(d+1)\kappa_3\delta_1 \log p\}$ is usually the smallest because of the $\log p$ in the denominator.

Assumption 7. The sample size n is no smaller than $8\kappa_h^2 \log p/\delta_2^2$, and also the range of feasible λ_n in Assumption 6 is non-empty, i.e.,

$$n \geq \frac{96^2(2-a)^2\Lambda_2^2}{a^2\Lambda_1^4}(d+1)^2\kappa_2\kappa_3^2\delta_1^2\delta_2 \log(2p)(\log p)^2. \tag{2.14}$$

Assumptions 5, 6, and 7 specify the minimum edge potential, the range of the tuning parameter, and the minimum sample size, required for Theorem 1 to hold, that is, for our neighbourhood selection approach (2.10) to achieve model selection consistency. Similar assumptions are made in related work (Yang et al., 2012).

Remark 3. Suppose that $n = \Omega\{(d+1)^2 \log^{3+\epsilon}(p)\}$ for $\epsilon > 0$, $\lambda_n = c\{\log(p)/n\}^{1/2}$ for some constant c, and κ_2 and κ_3 are $O(1)$. Then Assumptions 6 and 7 are satisfied asymptotically as n and p tend to infinity. Similar rates appear in Meinshausen and Bühlmann (2006); Ravikumar et al. (2010); Yang et al. (2012).

Theorem 1. *Suppose that the joint density* (2.2) *exists and Assumptions 1 – 7 hold for the* s*th node. Then with probability at least* $1 - c_1 p^{-\delta_1+2} - \exp(-c_2\delta_2^2 n) - \exp(-c_3\delta_3 n)$*, for some constants* c_1, c_2, c_3*,* $\delta_2 \leq \min(2\kappa_v/3, \kappa_h + \kappa_v)$*, and* $\delta_3 = 1/(\kappa_2\delta_2)$*, the estimator from* (2.10) *recovers the true neighbourhood of* x_s *exactly, so that* $\hat{N}(x_s) = N(x_s)$*.*

Theorem 1 shows that the probability of successful recovery converges to unity exponentially fast with the sample size n. We note that the number of neighbours d appears in Assumptions 5–7. As d increases, the minimum edge potential for each neighbour increases, the upper range for λ_n decreases, and the required sample size increases. Therefore, we need the true graph G to be sparse, $d = o(n)$, in order for Theorem 1 to be meaningful.

The quantities $\delta_2\kappa_2$ and $\delta_1\kappa_3$ appear in the upper bound of λ_n (2.13) and the minimum sample size (2.14). The fact that κ_2 and δ_2 appear together in a product implies that we can relax the restriction on δ_2 if κ_2 is small. The same applies to δ_1 and κ_3.

For certain types of nodes, Theorem 1 holds with a less stringent set of assumptions. For a Gaussian node, the second-and-higher order derivatives of $D(\cdot)$ are always bounded, i.e., $\kappa_2 = 1$ and $\kappa_3 = 0$. This has profound effects on the theory, as illustrated in Corollary 1.

Corollary 1. *Suppose that the joint density (2.2) exists and Assumptions 1–5 hold for a Gaussian node, x_s. If*

$$\lambda_n \in \left[\frac{8(2-a)}{a} \left\{ \delta_2 \frac{\log(2p)}{n} \right\}^{1/2}, \frac{2(2-a)}{a} \delta_2 M \right], \quad n \geq \frac{8\kappa_h^2 \log p}{\delta_2^2},$$

then with probability at least $1 - \exp(-c_2\delta_2^2 n) - \exp(-c_3\delta_3 n)$, for some constants c_2, c_3, $\delta_2 \leq \min(2\kappa_v/3, \kappa_h + \kappa_v)$, and $\delta_3 = 1/\delta_2$, the estimator from (2.10) *recovers the true neighbourhood of x_s exactly, so that $\hat{N}(x_s) = N(x_s)$.*

2.4.2 Combining Neighbourhoods to Estimate the Edge Set

The neighbourhood selection approach may give asymmetric estimates, in the sense that $t \in \hat{N}(x_s)$ but $s \notin \hat{N}(x_t)$. To deal with this discrepancy, two strategies for estimating a single edge set were proposed in Meinshausen and Bühlmann (2006), and adapted in other work:

$$\hat{E}_{\text{and}} = \left\{ (s,t) : s \in \hat{N}(x_t) \text{ and } t \in \hat{N}(x_s) \right\}, \quad \hat{E}_{\text{or}} = \left\{ (s,t) : s \in \hat{N}(x_t) \text{ or } t \in \hat{N}(x_s) \right\}.$$

When the sth and tth nodes are of the same type, there is no clear reason to prefer the edge estimate from $\hat{N}(x_s)$ over the one from $\hat{N}(x_t)$, and so the choice of the intersection rule, $\hat{E}_{\text{and}}$, versus the union rule, $\hat{E}_{\text{or}}$, is not crucial (Meinshausen and Bühlmann, 2006).

When the sth and tth nodes are of different types, however, the choice of neighbourhood matters. We now take a closer look at this with examples of Gaussian, Bernoulli, exponential and Poisson nodes as in (2.4)–(2.7). Quantities c_1, c_2, and c_3 in Theorem 1 are the same regardless of the node type, while the values of κ_2 and κ_3 depend on the type of node being regressed on the others in (2.10). We fix $B_1 = \kappa_3\delta_1$ for Bernoulli, Poisson and exponential nodes. For a Gaussian node, this quantity will always equal zero, since $D(\eta_s) = \eta_s^2/2 + \log(2\pi)/2$ and hence $D'''(\eta_s) = 0 = \kappa_3$. Furthermore, we fix $B_2 = 1/\delta_3 = \delta_2\kappa_2$ for all four types of

nodes. With B_1 and B_2 fixed, the minimum sample size and the feasible range of the tuning parameter for Bernoulli, Poisson and exponential nodes are exactly the same, as these quantities involve only B_1 and B_2. In particular, from Assumption 6, the range of feasible λ_n is $[8(2-a)\{\log(2p)B_2/n\}^{1/2}/a, \Lambda_1^2/\{12\Lambda_2(d+1)B_1 \log p\}]$, and from Assumption 7, the minimum sample size is $96^2(2-a)^2\Lambda_2^2(d+1)^2 B_2 B_1^2 \log(2p)(\log p)^2/(a^2\Lambda_1^4)$. These bounds are more restrictive than the corresponding bounds for Gaussian nodes in Corollary 1. We now derive a lower bound for the probability of successful neighbourhood recovery for each node type.

Example 5. If x_s is a Gaussian node, then the log-partition function is $D(\eta_s) = \eta_s^2/2 + \log(2\pi)/2$. It follows that $D^{''}(\eta_s) = 1 = \kappa_2$. Thus, $\delta_2 = B_2$. By Corollary 1, a lower bound for the probability of successful neighbourhood recovery is

$$\text{pr}\{\hat{N}(x_s) = N(x_s)\} \geq 1 - \exp(-c_2 B_2^2 n) - \exp(-c_3 n/B_2). \tag{2.15}$$

Example 6. If x_s is a Bernoulli node, then the log-partition function is $D(\eta_s) = \log\{\exp(-\eta_s) + \exp(\eta_s)\}$, so that $|D^{''}(\eta_s)| \leq 1$ and $|D^{'''}(\eta_s)| \leq 2$. Consequently, $\delta_2 = B_2$, and $\delta_1 = B_1/\kappa_3 = B_1/2$. By Theorem 1, a lower bound for the probability of successful neighbourhood recovery is

$$\text{pr}\{\hat{N}(x_s) = N(x_s)\} \geq 1 - c_1 p^{-B_1/2+2} - \exp(-c_2 B_2^2 n) - \exp(-c_3 n/B_2). \tag{2.16}$$

Example 7. If x_s is a Poisson node, then the log-partition function is $D(\eta_s) = \exp(\eta_s)$, so $D^{''}(\eta_s) = D^{'''}(\eta_s) = \exp(\eta_s)$. To bound $D^{''}(\eta_s)$ and $D^{'''}(\eta_s)$, we need to bound $\exp(\eta_s)$. Recall from Table 2.1 that strong compatibility requires that $\theta_{ts} x_t \leq 0$ when x_t is Gaussian, Poisson or exponential. Therefore, an upper bound for $\exp(\eta_s)$ is

$$\exp(\eta_s) \leq \exp\left(\alpha_{1s} + \sum_{t \in I} |\theta_{ts}|\right) \equiv b_P, \tag{2.17}$$

with I the set of Bernoulli nodes. Therefore, $\kappa_2 = \kappa_3 = b_P$, and so $\delta_2 = B_2/b_P$ and $\delta_1 = B_1/b_P$. By Theorem 1, a lower bound on the probability of successful neighbourhood recovery is

$$\text{pr}\{\hat{N}(x_s) = N(x_s)\} \geq 1 - c_1 p^{-B_1/b_P+2} - \exp(-c_2 B_2^2 n/b_P^2) - \exp(-c_3 n/B_2). \tag{2.18}$$

Example 8. If x_s is an exponential node, then the log-partition function is $D(\eta_s) = -\log(-\eta_s)$, so $D^{''}(\eta_s) = \eta_s^{-2}$ and $D^{'''}(\eta_s) = -2\eta_s^{-3}$. Furthermore,

$$\eta_s = \alpha_{1s} + \sum_{t \neq s} \theta_{ts} x_t \leq \alpha_{1s} + \sum_{t \in I} \theta_{ts} x_t \leq \alpha_{1s} + \sum_{t \in I} |\theta_{ts}| < 0, \tag{2.19}$$

with I the set of Bernoulli nodes. In (2.19), the first inequality follows from the requirement for compatibility from Table 2.1 that $\theta_{ts} x_t \leq 0$ when x_t is Gaussian, Poisson or exponential; the second inequality follows from the fact that Bernoulli nodes are coded as $+1$ and -1; and the third inequality follows from the Bernoulli-exponential entry in Table 2.1. Therefore, it follows that

$$|\eta_s| \geq \left| \alpha_{1s} + \sum_{t \in I} |\theta_{ts}| \right| \geq |\alpha_{1s}| - \sum_{t \in I} |\theta_{ts}| \equiv b_E. \tag{2.20}$$

As a result, $|D^{''}(\eta_s)|$ and $|D^{'''}(\eta_s)|$ are bounded by $\kappa_2 = b_E^{-2}$ and $\kappa_3 = 2b_E^{-3}$, respectively. For fixed B_1 and B_2, we have $\delta_2 = b_E^2 B_2$ and $\delta_1 = B_1 b_E^3/2$. By Theorem 1, a lower bound for the probability of successful neighbourhood recovery is

$$\text{pr}\{\hat{N}(x_s) = N(x_s)\} \geq 1 - c_1 p^{-b_E^3 B_1/2+2} - \exp(-c_2 b_E^4 B_2^2 n) - \exp(-c_3 n/B_2). \tag{2.21}$$

Examples 5–8 reveal that the neighbourhood of a Gaussian node is easier to recover than the neighbourhood of the other three types of nodes: the first requires a smaller minimum sample size when p is large, allows for a wider range of feasible tuning parameters, and has in general a higher probability of success. As a result, the neighbourhood of the Gaussian node should be used when estimating an edge between a Gaussian node and a non-Gaussian node.

Which neighbourhood should we use to estimate an edge between two non-Gaussian nodes? There are no clear winners: while (2.16) can be evaluated given knowledge of c_1, c_2, and c_3, (2.18) and (2.21) also require knowledge of the unknown quantities b_E and b_P, which are functions of unknown quantities Θ_s and α_{1s} in (2.17) and (2.20). One possibility is to insert a consistent estimator for these parameters (van de Geer, 2008; Bunea, 2008) in order to obtain a consistent estimator for b_P or b_E. This leads to the following lemma.

Lemma 2. *Suppose $\tilde{\Theta}_s$ and $\tilde{\alpha}_{1s}$ are consistent estimators of the true parameters in the conditional densities* (2.6) *and* (2.7)*. Let I be the index set of the Bernoulli nodes.*

Table 2.2: Neighbourhood to use in estimating an edge between two non-Gaussian nodes of different types

	Selection rules
Poisson & Exponential	Choose Poisson if $\tilde{b}_E^2\tilde{b}_P < 1$ and $\tilde{b}_E^3\tilde{b}_P < 2$. Choose exponential if $\tilde{b}_E^2\tilde{b}_P > 1$ and $\tilde{b}_E^3\tilde{b}_P > 2$.
Poisson & Bernoulli	Choose Poisson if $\tilde{b}_P < 1$. Choose Bernoulli if $\tilde{b}_P > 2$.
Exponential & Bernoulli	Choose exponential if $\tilde{b}_E \geq 1$. Choose Bernoulli if $\tilde{b}_E < 1$.

When the conditions in this table are not met, there is no clear preference in terms of which neighbourhood to use.

1. If x_s is a Poisson node and $\tilde{b}_P = \exp(\tilde{\alpha}_{1s} + \sum_{t\in I} |\tilde{\theta}_{ts}|)$, then

$$1 - c_1 p^{-B_1/\tilde{b}_P+2} - \exp(-c_2 B_2^2 n/\tilde{b}_P^2) - \exp(-c_3 n/B_2) \tag{2.22}$$

is a consistent estimator of a lower bound for pr$\{\hat{N}(x_s) = N(x_s)\}$.

2. If x_s is an exponential node and $\tilde{b}_E = |\tilde{\alpha}_{1s}| - \sum_{t\in I} |\tilde{\theta}_{ts}|$, then

$$1 - c_1 p^{-\tilde{b}_E^3 B_1/2+2} - \exp(-c_2 \tilde{b}_E^4 B_2^2 n) - \exp(-c_3 n/B_2) \tag{2.23}$$

is a consistent estimator of a lower bound for pr$\{\hat{N}(x_s) = N(x_s)\}$.

Therefore, by inserting consistent estimators of Θ_s and α_{1s} into (2.17) or (2.20), we can reconstruct an edge by choosing the estimate with the highest probability of correct recovery according to (2.16), (2.22), and (2.23). The rules are summarized in Table 2.2. The results in this section illustrate a worst case scenario for recovery of each neighbourhood, in that Theorem 1 provides a lower bound for the probability of successful neighbourhood recovery.

2.5 Recovery in Partially-Specified Models

In Section 2.4, we showed that the neighbourhood selection approach of Section 2.3.1 can recover the true graph when each node's conditional distribution is of the form (2.3), provided that the conditions for strong compatibility are satisfied. In this section, we consider a partially-specified model in which some of the nodes are assumed to have conditional distributions of the form (2.3), and we make no assumption on the conditional distributions of the remaining nodes. We will show that in this setting, neighbourhoods of the nodes with conditional distributions of the form (3) can still be recovered.

Here the neighbourhood of x_s is defined based upon its conditional density, (2.3), as $N^0(x_s) = \{t : \theta_{ts} \neq 0\}$. Assumption 4 in Section 2.4.1 is inappropriate since we no longer assume that all p nodes have conditional densities of the form (2.3), and consequently we are not assuming a particular form for the joint density. Therefore, we make the following assumption to replace Propositions 2 and 3.

Assumption 8. Assume that (i) $\text{pr}(\xi_1) \geq 1 - c_1 p^{-\delta_1+2}$, (ii) $\text{pr}(\xi_2) \geq 1 - \exp(-c_2\delta_2^2 n)$.

Theorem 2. *Suppose that the sth node has conditional density* (2.3)*, and that Assumptions 1 – 3 and 5 – 8 hold. Then with probability at least* $1 - c_1 p^{-\delta_1+2} - \exp(-c_2\delta_2^2 n) - \exp(-c_3\delta_3 n)$*, for some constants* c_1, c_2, c_3*, and* $\delta_3 = 1/(\kappa_2\delta_2)$*, the estimator from* (2.10) *recovers the true neighbourhood of* x_s *exactly, so that* $\hat{N}(x_s) = N^0(x_s)$.

The proof of Theorem 2 is similar to that of Theorem 1, and is thus omitted. Theorem 2 indicates that our neighbourhood selection approach can recover the neighbourhood of any node for which the conditional density is of the form (2.3), provided that Assumption 8 holds. This means that in order to recover an edge between two nodes using our neighbourhood selection approach, it suffices for one of the two nodes' conditional densities to be of the form (2.3). Consequently, we can model relationships that are far more flexible than those outlined in Table 2.1, e.g. an edge between a Poisson node and a node that takes values on the whole real line.

Although Theorem 2 allows us to go beyond some of the restrictions in Table 2.1, it is still restricted in that it only guarantees recovery of an edge between two nodes for which at least one

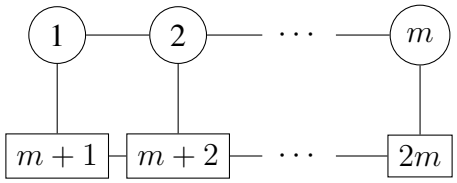

Figure 2.1: The graph used to generate the data in Sections 2.6.2–2.6.4. There are $m = p/2$ Gaussian or Poisson nodes, shown as circles, and $m = p/2$ Bernoulli nodes, shown as rectangles.

of the node-conditional densities is exactly of the form (2.3). In future work, we could generalize Theorem 2 to the case where (2.3) is simply an approximation to the true node-conditional distribution.

2.6 Numerical Studies

2.6.1 Data Generation

We consider mixed graphical models with two types of nodes, and $m = p/2$ nodes per type, for Gaussian-Bernoulli and Poisson-Bernoulli models. We order the nodes so that the Gaussian or Poisson nodes precede the Bernoulli nodes.

For both models, we construct a graph in which the jth node for $j = 1, \ldots, m$ is connected with the adjacent nodes of the same type, as well as the $(m + j)$th node of the other type, as shown in Fig. 2.1. This encodes the edge set E. For $(i, j) \in E$ and $i < j$, we generate the edge potentials θ_{ij} and θ_{ji} as

$$\theta_{ij} = \theta_{ji} = y_{ij} r_{ij},\ \text{pr}(y_{ij} = 1) = \text{pr}(y_{ij} = -1) = 0.5,\ \ r_{ij} \sim \text{Unif}(a, b). \tag{2.24}$$

We set $\theta_{ij} = \theta_{ji} = 0$ if $(i, j) \notin E$. Additional steps to ensure strong compatibility of the conditional distributions are discussed in the the Appendix. Values of a and b in (2.24), as well as the parameters of $f_s(x_s)$ in the conditional density (2.3), are specified in Sections 2.6.2–2.6.4.

To sample from the joint density $p(x)$ in (2.2) without calculating the log-partition function A, we employ a Gibbs sampler, as in Lee and Hastie (2014). Briefly, we iterate through the nodes, and

sample from each node's conditional distribution. To ensure independence, after a burn-in period of 3000 iterations, we select samples from the chain 500 iterations apart from each other.

2.6.2 Probability of Successful Neighbourhood Recovery

In Section 2.4.1 we saw that the probability of successful neighbourhood recovery for neighbourhood selection converges to unity exponentially fast with the sample size, and in Section 2.4.2 we saw that the estimates from the Gaussian nodes are superior to those from the Bernoulli nodes, in the sense that a smaller sample size is needed in order to achieve a given probability of successful recovery. We now verify those findings empirically. Here, successful neighbourhood recovery is defined to mean that the estimated and true edge sets of a graph or a sub-graph are identical.

We set $a = b = 0{\cdot}3$ in (2.24) so that Assumption 5 is satisfied, and generate one Gaussian-Bernoulli graph for each of $p = 60$, $p = 120$, and $p = 240$. We set $\alpha_{1s} = 0$ and $\alpha_{2s} = -1$ in (2.4) for Gaussian nodes, and $\alpha_{1s} = 0$ for Bernoulli nodes (2.5). For each graph, 100 independent data sets are drawn from the Gibbs sampler. We perform neighbourhood selection using the estimator from (2.11), with the tuning parameter λ_n set to be a constant c times $\{\log(p)/n\}^{1/2}$, so that it is on the scale required by Assumption 6, as illustrated in Remark 3.

In order to achieve successful neighbourhood recovery as the sample size increases, the value of c must be in a range matching the requirement of Assumption 6. We explored a range of values of c, and in Fig. 2.2 we show the probability of successful neighbourhood recovery for $c = 2.6$. For ease of viewing, we display separate empirical probability curves for the Gaussian-Gaussian, Bernoulli-Bernoulli, and Bernoulli-Gaussian subgraphs. Panels (a) and (b) are estimates obtained by regressing the Gaussian nodes onto the others, and panels (c) and (d) are the estimates from regressing the Bernoulli nodes onto the others. We see that the probability of successful recovery increases to unity once the scaled sample size exceeds the threshold required in Assumption 7 and Corollary 1. Furthermore, panels (b) and (c) agree with the conclusions of Section 2.4.2: neighbourhood recovery using the regression of a Gaussian node onto the others requires fewer samples than recovery using the regression of a Bernoulli node onto the others.

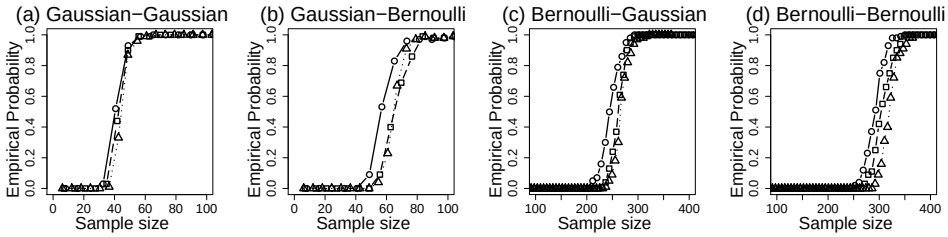

Figure 2.2: Probability of successful neighbourhood recovery, y-axis, as a function of scaled sample size $n/\{3\log(p)\}$, x-axis, for the set-up of Section 2.6.2. The curves are empirical probabilities of successful neighbourhood recovery for graphs with 60 (⊶), 120 (▫--▫), and 240 nodes (△···△), averaged over 100 independent data sets. The tuning parameter is set to be $2{\cdot}6\{\log(p)/n\}^{1/2}$. The title of each panel indicates the subgraph for which the recovery probability is displayed, and the first word in the title indicates the node type that was regressed in order to obtain the subgraph estimate. For instance, panel (b) displays probability curves for edges between Gaussian and Bernoulli nodes that are estimated from the ℓ_1-penalized linear regression of Gaussian nodes. Panel (c) displays the same quantity, estimated via an ℓ_1-penalized logistic regression of the Bernoulli nodes.

2.6.3 Comparison to Competing Approaches

In this section, we compare the proposed method to alternative approaches on a Gaussian-Bernoulli graph. We limit the number of nodes to $p = 40$ in order to facilitate comparison with the computationally intensive approach of Lee and Hastie (2014). We generate 100 random graphs with a =0·3 and b =0·6 in (2.24), and we set $\alpha_{1s} = 0$ and $\alpha_{2s} = -1$ in (2.4) for Gaussian nodes and $\alpha_{1s} = 0$ for Bernoulli nodes (2.5). Twenty independent samples of $n = 200$ observations are generated from each graph. We evaluate the performance of each approach by computing the number of correctly estimated edges as a function of the number of estimated edges in the graph. Results are averaged over 20 data sets from each of 100 random graphs, for a total of 2000 simulated data sets.

Seven approaches are compared in this study: 1) our proposal for neighbourhood selection in the mixed graphical model; 2) penalized maximum likelihood estimation in the mixed graphical model (Lee et al., 2013); 3) weighted ℓ_1-penalized regression in the mixed graphical model, as

proposed by Cheng et al. (2013); 4) graphical random forests (Fellinghauer et al., 2013); 5) neighbourhood selection in the Gaussian graphical model (Meinshausen and Bühlmann, 2006), where we use an ℓ_1-penalized linear regression to estimate the neighbourhood of all nodes; 6) the graphical lasso (Friedman et al., 2008), which treats all features as Gaussian; and 7) neighbourhood selection in the Ising model (Ravikumar et al., 2010), where we use ℓ_1-penalized logistic regression on all nodes after dichotomizing the Gaussian nodes by their means. The first four methods are designed for mixed graphical models, with Lee and Hastie (2014) and Cheng et al. (2013). specifically proposed for Gaussian-Bernoulli networks. In contrast, the last three methods ignore the presence of mixed node types. For methods based on neighbourhood selection, we use the union rule of Meinshausen and Bühlmann (2006) to reconstruct the edge set from the estimated neighbourhoods, with one exception: to estimate the Gaussian-Bernoulli edges for our proposed method, we use the estimates from the Gaussian nodes, as suggested by the theory developed in Section 2.4.2.

Due to its high computational cost, the method of Lee and Hastie (2014) is run on 250 data sets from 50 graphs rather than 2000 data sets from 100 graphs.

The left-hand panel of Fig. 2.3 displays results for Bernoulli-Bernoulli and Gaussian-Gaussian edges, and the right-hand panel displays results for edges between Gaussian and Bernoulli nodes.

The curves in Fig. 2.3 correspond to the estimated graphs as the tuning parameter for each method is varied. Recall from Section 2.3.2 that our proposal involves a tuning parameter λ_n^G for the ℓ_1-penalized linear regressions of the Gaussian nodes onto the others, and a tuning parameter λ_n^B for the ℓ_1-penalized logistic regressions of the Bernoulli nodes onto the others. The triangles in Fig. 2.3 show the average performance of our proposed method with the tuning parameters $\hat{\lambda}_n^B$ and $\hat{\lambda}_n^G$ selected using BIC summed over the Bernoulli and Gaussian nodes, respectively, as described in Section 2.3.2. This choice yields good precision (52%) and recall (95%) for edge recovery in the graph. To obtain the curves in Fig. 2.3, we set $\lambda_n^B = (\hat{\lambda}_n^B/\hat{\lambda}_n^G)\lambda_n^G$, and varied the value of λ_n^G.

In general, our proposal outperforms the competitors, which is expected since it assumes the correct model. Though the proposals of Cheng et al. (2013) and Lee and Hastie (2014) are intended for a Gaussian-Bernoulli graph, they attempt to capture more complicated relationships

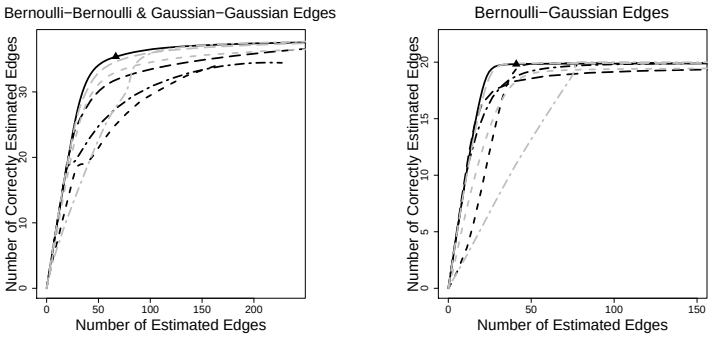

Figure 2.3: Simulation results for the Gaussian-Bernoulli graph, as described in Section 2.6.3. The number of correctly estimated edges is displayed as a function of the number of estimated edges, for a range of tuning parameter values in a graph with $p = 40$ and $n = 200$. The left panel corresponds to edges between nodes of the same type, while the right panel corresponds to the edges between Gaussian and Bernoulli nodes. The curves within each panel represent our proposal (——), Lee and Hastie (2014) (– –), Cheng et al. Cheng et al. (2013) (— –), Fellinghauer et al. (2013) (— –), neighbourhood selection in the Gaussian graphical model (— –), neighbourhood selection in the Ising model (– –), and the graphical lasso (— –). The black triangle shows the average performance of our proposed approach with the tuning parameter selected by the Bayesian information criterion (Section 2.3.2).

than in (2.2), and so they perform worse than our proposal. On the other hand, the graphical random forest of Fellinghauer et al. (2013) performs reasonably well, despite the fact that it is a nonparametric approach. Neighbourhood selection in the Gaussian graphical model performs closest to the proposed method in terms of edge selection. The Ising model suffers substantially due to dichotomization of the Gaussian variables. The graphical lasso algorithm experiences serious violations to its multivariate Gaussian assumption, leading to poor performance.

2.6.4 Application of Selection Rules for Mixed Graphical Models

In Section 2.6.3, in keeping with the results of Section 2.4.2, we always used the estimates from the Gaussian nodes in estimating an edge between a Bernoulli node and a Gaussian node. Here we consider a mixed graphical model of Poisson and Bernoulli nodes. In this case, the selection rules in Section 2.4.2 are more complex, and whether it is better to use a Poisson node or a Bernoulli node in order to estimate a Bernoulli-Poisson edge depends on the true parameter values in Table 2.2.

We generate a graph with $p = 80$ nodes as follows: $a = 0.8$ and $b = 1$ in (2.24), $\alpha_{1s} = -3$ for $s = 1, \ldots, 20$ and $\alpha_{1s} = 0$ for $s = 21, \ldots, 40$ for the Poisson nodes, and $\alpha_{1s} = 0$ for the Bernoulli nodes. This guarantees that b_P in (2.17) is smaller than 1 for the first half of the Poisson nodes, and larger than 2 for the second half, due to the structure of the graph from Fig. 2.1. In order to estimate a Bernoulli-Poisson edge, we will use the estimates from the Poisson nodes if $b_P < 1$ and the estimates from the Bernoulli nodes if $b_P > 2$, according to the selection rules in Table 2.2.

We compare the performance of our proposed approach using the selection rules in Table 2.2, with the true and estimated parameters, to our proposed approach using the union and intersection rules (Section 2.4.2), as well as the graphical random forest of Fellinghauer et al. (2013). To prevent over-shrinkage of the parameters for estimation of b_P in (2.17), we set λ_n in (2.10) to equal 0·5 times the value from the Bayesian information criterion for each node type. We present only the results for Poisson-Bernoulli edges, as the selection rules in Section 2.4.2 apply to edges between nodes of different types.

Results are shown in Fig. 2.4, averaged over 20 samples from each of 25 random graphs. The selection rules proposed in Section 2.4.2 clearly outperform the commonly-used union and intersection rules. The curve for the selection rule from Section 2.4.2 using the estimated parameter values is almost identical to the curve using the true parameter values, which indicates that in this case the quantity b_P is accurately estimated for each node. The graphical random forest slightly outperforms our proposal when few edges are estimated, but performs worse when the estimated graph includes more edges. This may indicate that as the graph becomes less sparse, the nonparametric graphical random forest approach suffers from insufficient sample size.

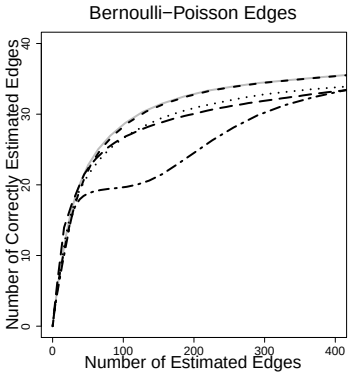

Figure 2.4: Summary of the simulation results for the Poisson-Bernoulli graph, as described in Section 2.6.4. The number of correctly estimated edges is displayed as a function of the number of estimated edges, for a range of tuning parameter values in a graph with $p = 80$ nodes from $n = 200$ observations. The curves represent the selection rule from Section 2.4.2 with the true parameters (——), the selection rule from Section 2.4.2 with estimated parameters (- -), the union rule (- —), the intersection rule (· ·), and the method from Fellinghauer et al. (2013) (— -).

2.7 Discussion

In Section 2.2.2 we saw that a stringent set of restrictions is required for compatibility or strong compatibility of the node-conditional distributions given in (2.4)–(2.7). These restrictions limit the theoretical flexibility of the conditionally-specified mixed graphical model, especially when modeling unbounded variables. It is possible that by truncating unbounded variables, we may be able to circumvent some of these restrictions. Furthermore, the model (2.2) assumes pairwise interactions in the form of $x_s x_t$, which can be seen as a second-order approximation of the true edge potentials in (2.1). We can relax this assumption by fitting non-linear edge potentials using semi-parametric penalized regressions, as in Voorman et al. (2014).

Chapter 3

GRAPHICAL ESTIMATION FOR ODE MODELS

The following work is to appear in *Journal of the American Statistical Association - Theory and Methods* (Chen et al., 2016).

3.1 Introduction

Ordinary differential equations (ODEs) have been widely used to model dynamical systems in many fields, including chemical engineering (Biegler et al., 1986), genomics (Chou and Voit, 2009), neuroscience (Izhikevich, 2007), and infectious diseases (Wu, 2005). A system of ODEs takes the form

$$X'(t;\theta) \equiv \begin{bmatrix} \frac{dX_1(t;\theta)}{dt} \\ \vdots \\ \frac{dX_p(t;\theta)}{dt} \end{bmatrix} = \begin{bmatrix} f_1(X(t;\theta),\theta) \\ \vdots \\ f_p(X(t;\theta),\theta) \end{bmatrix} \equiv f(X(t;\theta),\theta); \quad t \in [0,1], \tag{3.1}$$

where $X(t;\theta) = (X_1(t;\theta),\ldots,X_p(t;\theta))^{\mathrm{T}}$ denotes a set of variables, and the form of the functions $f = (f_1,\ldots,f_p)^{\mathrm{T}}$ may be known or unknown. In (3.1), t indexes time. Typically, there is also an initial condition of the form $X(0;\theta) = C$, where C is a p-vector. In practice, the system (3.1) is often observed on discrete time points subject to measurement errors. Let $Y_i \in \mathbb{R}^p$ be the measurement of the system at time t_i such that

$$Y_i = X(t_i;\theta^*) + \epsilon_i, \quad i = 1,\ldots,n, \tag{3.2}$$

where θ^* denotes the true set of parameter values and the random p-vector ϵ_i represents independent measurement errors. In what follows, for notational simplicity, we sometimes suppress the dependence of $X(t;\theta)$ on θ, i.e., $X(t) \equiv X(t;\theta)$ in (3.1) and $X^*(t) \equiv X(t;\theta^*)$ in (3.2).

In the context of high-dimensional time-course data arising from biology, it can be of interest to recover the structure of a system of ODEs — that is, to determine which features regulate each other. If f_j in (3.1) is a function of X_k, then we say that X_k *regulates* X_j in the sense that X_k controls the changes of X_j through its derivative X_j'. For instance, biologists might want to infer gene regulatory networks from noisy time-course gene expression data. In this case, the number of variables p exceeds the number of time points n; we refer to this as the high-dimensional setting.

In high-dimensional statistics, sparsity-inducing penalties such as the lasso (Tibshirani, 1996) and the group lasso (Yuan and Lin, 2006) have been well-studied. Such penalties have also been extensively used to recover the structure of probabilistic graphical models (e.g., Yuan and Lin, 2007; Friedman et al., 2008; Meinshausen and Bühlmann, 2010; Voorman et al., 2014). However, model selection in high-dimensional ODEs remains a relatively open problem, with the exception of some notable recent work (Lu et al., 2011; Henderson and Michailidis, 2014; Wu et al., 2014). In fact, the tasks of parameter estimation and model selection in ODEs from noisy data are very challenging, even in the classical statistical setting where $n > p$ (see e.g., Ramsay et al., 2007; Brunel, 2008; Liang and Wu, 2008; Qi and Zhao, 2010; Xue et al., 2010; Gugushvili and Klaassen, 2012; Hall and Ma, 2014; Zhang et al., 2015). Moreover, the problem of high-dimensionality is compounded if the form of the function f in (3.1) is unknown, leading to both statistical and computational issues.

In this study, we propose an efficient procedure for structure recovery of an ODE system of the form (3.1) from noisy observations of the form (3.2), in the setting where the functional form of f is unknown.

3.2 Literature Review

In this section, we review existing statistical methods for parameter estimation and/or model selection in ODEs. Most of the methods reviewed in this section are proposed for the low-dimensional setting. Even though they may not be directly applicable to the high-dimensional setting, they lay the foundation for the development of model selection procedures in high-dimensional additive ODEs.

3.2.1 Notation

Without loss of generality, assume that $0 = t_1 < t_2 < \ldots < t_n = 1$. We let Y_{ij} indicate the observation of the jth variable at the ith time point, t_i. We use $\mathcal{X}(h)$ to denote a nonparametric class of functions on $[0, 1]$ indexed by some smoothing parameter(s) h. We use $Z(\cdot)$ to represent an arbitrary function belonging to $\mathcal{X}(\cdot)$. We use $\|\cdot\|_2$ to denote the ℓ_2-norm of a vector or a matrix, and $|||f|||$ to denote the ℓ_2-norm of a function f on the interval $[0, 1]$, i.e. $|||f|||^2 \equiv \int_0^1 f^2(t)\, dt$. We use an asterisk to denote true values—for instance, θ^* denotes the true value of θ in (3.1). We use $\Lambda_{\min}(A)$ and $\Lambda_{\max}(A)$ to denote the minimum and maximum eigenvalues of a square matrix A, respectively.

3.2.2 Methods that assume a known form of f

Gold standard approach

To begin, we suppose that the function f in (3.1) takes a known form. Benson (1979) and Biegler et al. (1986) proposed to estimate the unknown parameter θ^* in (3.2) by solving the problem

$$\hat{\theta}^{\text{gold}} = \arg\min_{\theta} \sum_{i=1}^{n} \|Y_i - X(t_i; \theta)\|_2^2 \tag{3.3a}$$

$$\text{subject to} \quad X'(t; \theta) = f(X(t; \theta), \theta), \quad t \in [0, 1]. \tag{3.3b}$$

Note that $X(\cdot; \theta)$ in (3.3) is a fixed function given θ, although an analytic expression may not be available. The resulting estimator $\hat{\theta}^{\text{gold}}$ has appealing theoretical properties: for instance, when the measurement errors ϵ_i in (3.2) are Gaussian, then $\hat{\theta}^{\text{gold}}$ is the maximum likelihood estimator, and is $\sqrt{n}$-consistent. In this sense, (3.3) can thus be considered the *gold standard* approach. However, solving (3.3) is often computationally challenging.

Two-step collocation methods

In order to overcome the computational challenges associated with solving (3.3), *collocation* methods have been employed by a number of authors (Varah, 1982; Ellner et al., 2002; Ramsay et al.,

2007; Brunel, 2008; Cao and Zhao, 2008; Liang and Wu, 2008; Cao et al., 2011; Lu et al., 2011; Gugushvili and Klaassen, 2012; Brunel et al., 2014; Hall and Ma, 2014; Henderson and Michailidis, 2014; Wu et al., 2014; Dattner and Klaassen, 2015; Zhang et al., 2015).

The two-step collocation procedure first proposed by Varah (1982) involves fitting a smoothing estimate $\hat{X}(\cdot; h)$ to the observations $Y_1, \ldots, Y_n$ in (3.2) with a smoothing parameter h, and then plugging $\hat{X}(\cdot; h)$ and its derivative with respect to t into (3.1) in order to estimate θ. This amounts to solving the optimization problem

$$\hat{\theta}^{\text{TS}} = \arg\min_{\theta} \int_0^1 \left\| \hat{X}'(t; h) - f\big(\hat{X}(t; h), \theta\big) \right\|_2^2 dt, \tag{3.4a}$$

where

$$\hat{X}(\cdot; h) = \arg\min_{Z(\cdot) \in \mathcal{X}(h)} \sum_{i=1}^{n} \|Y_i - Z(t_i)\|_2^2. \tag{3.4b}$$

The two-step procedure (3.4) has a clear advantage over the gold standard approach (3.3) because the former decouples the estimation of θ and X. However, this advantage comes at a cost: due to the presence of $\hat{X}'$ in (3.4a), the properties of the estimator $\hat{\theta}^{\text{TS}}$ in (3.4) rely heavily on the smoothing estimates obtained in (3.4b), and $\sqrt{n}$-consistency has only been shown for certain values of the smoothing parameter h that are hard to choose in practice (Brunel, 2008; Liang and Wu, 2008; Gugushvili and Klaassen, 2012).

Dattner and Klaassen (2015) proposed an improvement to (3.4) for a special case of (3.1). To be more specific, they assume that $f_j(X(t), \theta)$ in (3.1) is a linear function of θ, which leads to

$$X'(t) \equiv \begin{bmatrix} \frac{dX_1(t)}{dt} \\ \vdots \\ \frac{dX_p(t)}{dt} \end{bmatrix} = \begin{bmatrix} g_1^{\mathrm{T}}(X(t))\theta \\ \vdots \\ g_p^{\mathrm{T}}(X(t))\theta \end{bmatrix} \equiv g(X(t))\theta; \quad t \in [0, 1], \tag{3.5}$$

where $g(X(t))$ is a known function of $X(t)$. Integrating both sides of (3.5) gives

$$X(t) = \left\{ \int_0^t g(X(u))\, du \right\} \theta + C, \tag{3.6}$$

where $C \equiv X(0; \theta)$. The unknown parameter θ^* is estimated by solving

$$\hat{\theta}^{\text{LM}} = \arg\min_{\theta} \int_0^1 \left\| \hat{X}(t; h) - \left\{ \int_0^t g\big(\hat{X}(u; h)\big)\, du \right\} \theta - C \right\|_2^2 dt, \tag{3.7a}$$

where

$$\hat{X}(\cdot; h) = \underset{Z(\cdot)\in\mathcal{X}(h)}{\arg\min} \sum_{i=1}^{n} \|Y_i - Z(t_i)\|_2^2. \tag{3.7b}$$

The optimization problem (3.7a) has an analytical solution, given the smoothing estimates from (3.7b). Compared with the two-step procedure (3.4), this approach requires an estimate of the integral, $\int_0^t g\big(\hat{X}(u; h)\big)\, du$ in (3.7a), rather than an estimate of the derivative, $\hat{X}'(t; h)$. This has profound effects on the asymptotic behaviour of the estimator $\hat{\theta}^{\text{LM}}$. $\sqrt{n}$-consistency of $\hat{\theta}^{\text{LM}}$ has been established under mild conditions, and it has been found that the choice of smoothing parameter h is less crucial than for other methods (Gugushvili and Klaassen, 2012).

Recently, Brunel et al. (2014) and Hall and Ma (2014) have considered alternatives to the loss function in (3.4a). Let $\mathbb{C}^1(0, 1)$ be the set of functions that are first-order differentiable on $(0, 1)$ and equal zero on the boundary points 0 and 1. Then (3.1) implies that, for any $\phi \in \mathbb{C}^1(0, 1)$,

$$\int_0^1 f(X(t), \theta)\phi(t)dt + \int_0^1 X(t)\phi'(t)dt = 0. \tag{3.8}$$

Equation (3.8) is referred to as the *variational formulation* of the ODE. A least squares loss based on (3.8) takes the form

$$\hat{\theta}^{\text{V}} = \underset{\theta}{\arg\min} \frac{1}{L} \sum_{l=1}^{L} \left\| \int_0^1 f\big(\hat{X}(t; h), \theta\big)\phi_l(t)dt + \int_0^1 \hat{X}(t; h)\phi_l'(t)dt \right\|_2^2, \tag{3.9}$$

where $\hat{X}(t; h)$ is defined in (3.4b) and $\{\phi_l, l = 1, \ldots, L\}$ is a finite set of functions in $\mathbb{C}^1(0, 1)$ (Brunel et al., 2014). In Hall and Ma (2014), the loss function is the sum of the loss functions in (3.4b) and (3.9), so that θ and the optimal bandwidth h are estimated simultaneously. It is immediately clear that the derivative $X'(\cdot; \theta)$ is not needed in (3.9), which can lead to substantial improvement compared to the two-step procedure in (3.4). A minor drawback of (3.9) is that the variational formulation (3.8) is enforced on a finite set of functions $\{\phi_l, l = 1, \ldots, L\}$ rather than on the whole class $\mathbb{C}^1(0, 1)$. Under suitable assumptions, the estimator $\hat{\theta}^{\text{V}}$ is $\sqrt{n}$-consistent (Brunel et al., 2014; Hall and Ma, 2014).

The generalized profiling method

Another collocation-based method is the generalized profiling method of Ramsay et al. (2007). Instead of the smoothing estimate $\hat{X}(\cdot; h)$ in (3.4b), the generalized profiling method uses a smoothing estimate $\check{X}(\cdot; h, \theta)$ that minimizes the weighted sum of a data-fitting loss and a model-fitting loss for any given θ. In greater detail,

$$\hat{\theta}_\lambda^{\text{GP}} = \arg\min_\theta \sum_{i=1}^n \left\| Y_i - \check{X}(t_i; h, \theta) \right\|_2^2, \tag{3.10a}$$

where

$$\check{X}(\cdot; h, \theta) = \arg\min_{Z(\cdot) \in \mathcal{X}(h)} \frac{1}{n} \sum_{i=1}^n \|Y_i - Z(t_i)\|_2^2 + \lambda \int_0^1 \|Z'(t) - f(Z(t), \theta)\|_2^2 \, dt. \tag{3.10b}$$

In Ramsay et al. (2007), the authors solve (3.10a) iteratively for a non-decreasing sequence of λ's in (3.10b). $\sqrt{n}$-consistency of the limiting estimator was later established by Qi and Zhao (2010). Zhang et al. (2015) proposed a model selection procedure by applying an *ad hoc* lasso procedure (Wang and Leng, 2007) to the estimates from (3.10).

3.2.3 Methods that do not assume the form of f

A few authors have recently considered modeling large-scale dynamical systems from biology using ODEs (Henderson and Michailidis, 2014; Wu et al., 2014), under the assumption that the right-hand side of (3.1) is additive,

$$X_j'(t) = \theta_{j0} + \sum_{k=1}^p f_{jk}(X_k(t)), \quad \theta_{j0} \in \mathbb{R}. \tag{3.11}$$

Henderson and Michailidis (2014) and Wu et al. (2014) approximate the unknown f_{jk} with a truncated basis expansion. Consider a finite basis, $\psi(x) = (\psi_1(x), \ldots, \psi_M(x))^{\mathrm{T}}$, such that

$$f_{jk}(a_k) = \psi(a_k)^{\mathrm{T}} \theta_{jk} + \delta_{jk}(a_k), \quad \theta_{jk} \in \mathbb{R}^M, \tag{3.12}$$

where $\delta_{jk}(a_k)$ denotes the residual. Using (3.12), a system of additive ODEs of the form (3.11) can be written as

$$X_j'(t) = \theta_{j0} + \sum_{k=1}^p \psi(X_k(t))^{\mathrm{T}} \theta_{jk} + \sum_{k=1}^p \delta_{jk}(X_k(t)), \quad j = 1, \ldots, p. \tag{3.13}$$

Henderson and Michailidis (2014) and Wu et al. (2014) consider the problem of estimating and selecting the non-zero elements θ_{jk} in (3.13). Roughly speaking, they propose to solve optimization problems of the form

$$\begin{aligned}\hat{\theta}_j^{\text{NP}} = \underset{\theta_{j0}\in\mathbb{R},\theta_{jk}\in\mathbb{R}^M}{\arg\min} \int_0^1 \left\| \hat{X}_j'(t;h) - \theta_{j0} - \sum_{k=1}^p \psi\big(\hat{X}_k(t;h)\big)^{\mathrm{T}}\theta_{jk} \right\|_2^2 dt \\ + \lambda_n \sum_{k=1}^p \left[\int_0^1 \{\psi\big(\hat{X}_k(t;h)\big)^{\mathrm{T}}\theta_{jk}\}^2 \, dt \right]^{1/2},\end{aligned} \tag{3.14a}$$

for $j = 1, \ldots, p$, where

$$\hat{X}(\cdot;h) = \underset{Z(\cdot)\in\mathcal{X}(h)}{\arg\min} \sum_{i=1}^n \|Y_i - Z(t_i)\|_2^2. \tag{3.14b}$$

In (3.14a), a standardized group lasso penalty forces all elements in θ_{jk} to be either zero or non-zero when λ_n is large, thereby providing variable selection.

The proposals of Henderson and Michailidis (2014) and Wu et al. (2014) are slightly more involved than (3.14): an extra ℓ_2-penalty is applied to the θ_{jk}'s in (3.14a) in Henderson and Michailidis (2014), whereas in Wu et al. (2014) (3.14a) is followed by tuning parameter selection using Bayesian information criterion (BIC), an adaptive group lasso regression, and a regular lasso. We refer the reader to Henderson and Michailidis (2014) and Wu et al. (2014) for further details.

3.3 Proposed Approach

We consider the problem of model selection in high-dimensional ODEs. As in Henderson and Michailidis (2014) and Wu et al. (2014), we assume an additive ODE model (3.11). We use a finite basis $\psi(\cdot)$ to approximate the additive components f_{jk} as in (3.12), leading to an ODE system that is linear in the unknown parameters (3.13). Following the example of Dattner and Klaassen (2015), we exploit this linearity by integrating both sides of (3.13), which yields

$$X_j(t) = X_j(0) + \theta_{j0}t + \sum_{k=1}^p \Psi_k(t)^{\mathrm{T}}\theta_{jk} + \sum_{k=1}^p \int_0^t \delta_{jk}(X_k(u))\, du, \tag{3.15}$$

where $\Psi_k(t)$ denotes the integrated basis such that

$$\Psi_k(t) = (\Psi_{k1}(t), \ldots, \Psi_{kM}(t))^{\mathrm{T}} = \int_0^t \psi(X_k(u))\, du, \ k = 1, \ldots, p, \tag{3.16}$$

and $\Psi_0(t) = t$. Our method, called *Graph Reconstruction via Additive Differential Equations* (GRADE), then solves the following problem for $j = 1, \ldots, p$:

$$\hat{\theta}_j = \underset{C_{j0}\in\mathbb{R},\theta_{j0}\in\mathbb{R},\, \theta_{j1},\ldots,\theta_{jp}\in\mathbb{R}^M}{\arg\min} \frac{1}{2n}\sum_{i=1}^{n}\left\{Y_{ij} - C_{j0} - \theta_{j0}\hat{\Psi}_0(t_i) - \sum_{k=1}^{p}\theta_{jk}^{\mathrm{T}}\hat{\Psi}_k(t_i)\right\}^2 + \lambda_{n,j}\sum_{k=1}^{p}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\theta_{jk}^{\mathrm{T}}\hat{\Psi}_k(t_i)\right\}^2\right]^{1/2}, \tag{3.17a}$$

where

$$\hat{X}(\cdot; h) = \underset{Z(\cdot)\in\mathcal{X}(h)}{\arg\min}\sum_{i=1}^{n}\|Y_i - Z(t_i)\|_2^2, \tag{3.17b}$$

and

$$\hat{\Psi}_0(t) = t;\ \hat{\Psi}_k(t) = \int_0^t \psi(\hat{X}_k(u; h))\, du,\ k = 1, \ldots, p. \tag{3.17c}$$

In (3.17a), $\lambda_{n,j}$ is a non-negative sparsity-inducing tuning parameter. We may sometimes use $\lambda_{n,j} \equiv \lambda_n$ for $j = 1, \ldots, p$ for simplicity. If the true function f_{jk}^* in (3.11) is non-zero, we say that the kth variable X_k^* is a true regulator of X_j^*. We let $S_j \equiv \{k : \|f_{jk}^*\|_2 \neq 0, k = 1, \ldots, p\}$ denote the set of true regulators. We let the estimated index set of regulators be $\hat{S}_j \equiv \{k : \left\|\hat{\theta}_{jk}\right\|_2 \neq 0, k = 1, \ldots, p\}$. We then reconstruct the network using $\hat{S}_j, j = 1, \ldots, p$.

Both (3.17a) and (3.17b) can be implemented efficiently using existing software (e.g., Loader, 2013; Meier, 2014). In our theoretical analysis in Section 3.4, we use local polynomial regression to obtain the smoothing estimate in (3.17b). We use generalized cross-validation (GCV) on the loss (3.17b) to select the smoothing tuning parameter h. We use BIC to select the number of bases M for ψ and $\hat{\Psi}$ in (3.17c), and the sparsity tuning parameter λ_n in (3.17a).

In some studies, time-course data is collected from multiple samples, or experiments. Let R denote the total number of experiments, and $Y^{(r)}$ the observations in the rth experiment. We assume that the same ODE system (3.13) applies across all experiments with the same true parameter θ_{jk}^*. We allow a different set of initial values for each experiment. Assume that each experiment

consists of measurements on the same set of time points. This leads us to modify (3.17) as follows:

$$\hat{\theta}_j = \underset{C_{j0}^{(r)}\in\mathbb{R},\theta_{j0}\in\mathbb{R},\ \theta_{j1},\ldots,\theta_{jp}\in\mathbb{R}^M}{\arg\min} \frac{1}{2Rn}\sum_{r=1}^{R}\sum_{i=1}^{n}\left\{Y_{ij}^{(r)} - C_{j0}^{(r)} - \theta_{j0}\hat{\Psi}_0(t_i) - \sum_{k=1}^{p}\theta_{jk}^{\mathrm{T}}\hat{\Psi}_k^{(r)}(t_i)\right\}^2 + \lambda_n\sum_{k=1}^{p}\left[\frac{1}{Rn}\sum_{r=1}^{R}\sum_{i=1}^{n}\left\{\theta_{jk}^{\mathrm{T}}\hat{\Psi}_k^{(r)}(t_i)\right\}^2\right]^{1/2}, \tag{3.18}$$

where

$$\hat{X}^{(r)}(\cdot;h) = \underset{Z(\cdot)\in\mathcal{X}(h)}{\arg\min}\sum_{i=1}^{n}\|Y_i^{(r)} - Z(t_i)\|_2^2,\ r=1,\ldots,R,$$

$$\hat{\Psi}_0(t) = t;\ \hat{\Psi}_k^{(r)}(t) = \int_0^t \psi\big(\hat{X}_k^{(r)}(u;h)\big)\,du,\ k=1,\ldots,p.$$

In Sections 3.4, 3.5.1, and 3.5.2, we will assume that only one experiment is available, so that our proposal takes the form (3.17). In Sections 3.5.3 and 3.6, we will apply our proposal to data from multiple experiments using (3.18).

Remark 4. To facilitate the comparison of GRADE (3.17) with other methods, we introduce an intermediate variable,

$$\tilde{X}_j(t;h,\theta) \equiv C_{j0} + \theta_{j0}t + \sum_{k=1}^{p}\theta_{jk}^{\mathrm{T}}\hat{\Psi}_k(t), \tag{3.19}$$

following from (3.15). Plugging (3.19) into the loss function in (3.17a) yields $\sum_{i=1}^{n}\big\{Y_{ij} - \tilde{X}_j(t_i;h,\theta)\big\}^2$. In the gold standard (3.3), the ODE system (3.1) is strictly satisfied due to the constraint in (3.3b). In the two-step procedure (3.4a) and (3.14a), the smoothing estimate $\hat{X}(\cdot;h)$ does not satisfy (3.1). GRADE stands in between: the initial estimate $\hat{X}(\cdot;h)$ in (3.17b) is solely based on the observations, while the intermediate estimate $\tilde{X}(\cdot;h,\theta)$ is calculated by plugging $\hat{X}(\cdot;h)$ into the additive ODE (3.13).

3.4 Theoretical Properties

In this section, we establish variable selection consistency of the GRADE estimator (3.17). Technical proofs of the statements in this section are available in Section B.1 in the supplementary material. We use s_j to denote the cardinality of S_j, and set $s = \max_j\{s_j\}$. For ease of presentation, we let $S_j^0 = \{0\}\cup S_j$, so that $\Psi_{S_j^0}(t) = \big(\Psi_0(t), \Psi_{S_j}^{\mathrm{T}}(t)\big)^{\mathrm{T}} = \big(t, \Psi_{S_j}^{\mathrm{T}}(t)\big)^{\mathrm{T}}$ is an (s_jM+1)-vector.

The proposed method (3.17) differs from the standard sparse additive model (Ravikumar et al., 2009) in that the regressors $\hat{\Psi}_k(t)$ in (3.17c) are estimated from smoothing estimates $\hat{X}(\cdot; h)$ (3.17b) instead of the true trajectories X^* in (3.2). We use local polynomial regression to compute $\hat{X}(\cdot; h)$ in (3.17b) (see e.g., Equation 1.67 of Tsybakov, 2009 for details on parameterization). To establish variable selection consistency, it is necessary to obtain a bound for the difference between $\hat{X}(\cdot; h)$ and X^*. This is addressed in Theorem 3. Using the bound in Theorem 3, we then establish variable selection consistency of the estimator in (3.17) for high-dimensional ODEs in Theorem 4.

In this study, we assume that the measurement errors in (3.2) are normally distributed. Generalizations to bounded or sub-Gaussian errors are straightforward.

Assumption 9. The measurement errors in (3.2) are independent, and $\epsilon_{ij} \sim N(0, \sigma^2), i = 1, \ldots, n, j = 1, \ldots, p$.

We also require the true trajectories X_j^* in (3.2) to be smooth.

Assumption 10. Assume that the solutions $X_j^*, 1 \leq j \leq p$, belong to a Hölder class $\Sigma(\beta_1, L_1)$, where $\beta_1 \geq 3$.

In addition, we need some regularity assumptions to hold for the smoothing estimation (3.17b). These assumptions are common and not crucial to this study, and are hence deferred to Section B.1.2 in the supplementary material (or see Section 1.6.1 in Tsybakov, 2009). We arrive at the following concentration inequality for $\left|\!\left|\!\left| \hat{X} - X^* \right|\!\right|\!\right|$.

Theorem 3. *Suppose that Assumptions 9–10 and 23–25 in the supplementary material are satisfied. Let* $\hat{X}_j$ *in* (3.17b) *be the local polynomial regression estimator of order* $\ell = \lfloor \beta_1 \rfloor$ *with bandwidth*

$$h_n \propto n^{(\alpha-1)/(2\beta_1+1)} \tag{3.20}$$

for some positive $\alpha < 1$. *Then, for each* $j = 1, \ldots, p$,

$$\left|\!\left|\!\left| \hat{X}_j - X_j^* \right|\!\right|\!\right|^2 \leq C_2 n^{\frac{2\beta_1}{2\beta_1+1}(\alpha-1)} \tag{3.21}$$

holds with probability at least $1 - 2\exp\big\{ - n^\alpha/(2C_3\sigma^2)\big\}$, *for some constants* C_2 *and* C_3.

The concentration inequality in Theorem 3 is derived using concentration bounds for Gaussian errors (Boucheron et al., 2013). Using Theorem 3, we see that the bound (3.21) holds uniformly for $j = 1, \ldots, p$ with probability at least $1 - 2p \exp\big\{ - n^{\alpha}/(2C_3\sigma^2)\big\}$. The bound in Theorem 3 thus holds uniformly for $j = 1, \ldots, p$ with probability converging to unity if $p = o\big(\exp\big\{n^{\alpha}/(2C_3\sigma^2)\big\}\big)$.

For the methods outlined in (3.14) (Henderson and Michailidis, 2014; Wu et al., 2014), variable selection consistency depends on the convergence of $\Big|\Big|\Big|\hat{X}' - (X^*)'\Big|\Big|\Big|$ and $\Big|\Big|\Big|\hat{X} - X^*\Big|\Big|\Big|$. In contrast, our method depends only on the convergence rate of $\Big|\Big|\Big|\hat{X} - X^*\Big|\Big|\Big|$. It is known that the convergence of $\Big|\Big|\Big|\hat{X}' - (X^*)'\Big|\Big|\Big|$ is slower than that of $\Big|\Big|\Big|\hat{X} - X^*\Big|\Big|\Big|$, see e.g. Gugushvili and Klaassen (2012). As a result, the rate of convergence of $\hat{\theta}_{jk}$ from (3.14) is slower than that of our proposed method (3.17).

In order to establish the main result, we need the following additional assumptions. Recall the definition of $\Psi_j(t)$ from (3.16); for convenience, we suppress the dependence of $\Psi(t)$ on t in what follows.

Assumption 11. For $j = 1, \ldots, p$, $(X_j^*)'$ is an additive function of X_k^*, $k = 1, \ldots, p$. In other words,

$$\big(X_j^*\big)'(t) = \theta_{j0}^* + \sum_{k=1}^{p} f_{jk}^*\big(X_k^*(t)\big), \quad \theta_{j0}^* \in \mathbb{R},\ j = 1, \ldots, p, \tag{3.22}$$

where $\int_0^1 f_{jk}^*\big(X_k^*(t)\big)dt = 0$ for all j, k. Furthermore, the functions f_{jk}^* $(1 \leq j, k \leq p)$ belong to a Sobolev class $W(\beta_2, L_2)$ on a finite interval with $\beta_2 \geq 3$.

Assumption 12. The eigenvalues of $\int_0^1 \Psi_{S_j^0}\Psi_{S_j^0}^{\mathrm{T}}\, dt$ are bounded from above by $C_{\max}$ and bounded from below by a positive number $C_{\min}$, and for $k \notin S_j^0$, the eigenvalues of $\int_0^1 \Psi_k\Psi_k^{\mathrm{T}}\, dt$ are bounded from below by $C_{\min}$. In other words,

$$0 < C_{\min} \leq \Lambda_{\min}\left(\int_0^1 \Psi_{S_j^0}\Psi_{S_j^0}^{\mathrm{T}}\, dt\right) \leq \Lambda_{\max}\left(\int_0^1 \Psi_{S_j^0}\Psi_{S_j^0}^{\mathrm{T}}\, dt\right) \leq C_{\max}, \tag{3.23}$$

and

$$C_{\min} \leq \Lambda_{\min}\left(\int_0^1 \Psi_k\Psi_k^{\mathrm{T}}\, dt\right), \quad \text{for} \quad k \notin S_j^0. \tag{3.24}$$

Assumption 13. Assume that

$$\max_{k \notin S_j^0} \left\| \left(\int_0^1 \Psi_k \Psi_{S_j^0}^{\mathrm{T}} \, dt \right) \left(\int_0^1 \Psi_{S_j^0} \Psi_{S_j^0}^{\mathrm{T}} \, dt \right)^{-1} \right\|_2 \leq \xi. \tag{3.25}$$

The first part of Assumption 12 ensures identifiability among the $s_j + 1$ elements in the set $\{t, X^*_{S_j}\}$, and the second part ensures that Ψ_k is non-degenerate for $k \notin S_j^0$. Assumption 13 restricts the association between the elements in the set $\{t, X^*_{S_j}\}$ and the elements in the set $X^*_{S_j^c}$. Note that in order for the parameters in an additive model such as (3.13) to be identifiable, there must be no concurvity among the variables (Buja et al., 1989). This is guaranteed by Assumptions 12 and 13, which appear often in the literature of lasso regression (Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006; Ravikumar et al., 2009; Wainwright, 2009; Lee et al., 2013). We refer the readers to Miao et al. (2011) for a detailed discussion of the identifiability of the parameters in an ODE model.

The next assumption characterizes the relationships between the quantities in Assumptions 12 and 13 and the sparsity tuning parameter λ_n in (3.17a). Similar assumptions have been made in lasso-type regression (Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006; Ravikumar et al., 2009; Wainwright, 2009; Lee et al., 2013).

Assumption 14. Assume that

$$f_{\min} > \lambda_n \frac{4\sqrt{2sC_{\max}}}{C_{\min}} \quad \text{and} \quad \xi < \frac{1}{4}\sqrt{\frac{C_{\min}}{sC_{\max}}},$$

where $f_{\min} \equiv \min_{k \in S_j} \left\{ \int_0^1 \left[f^*_{jk}(X^*_k(t)) \right]^2 dt \right\}^{1/2}$ is the minimum regulatory effect.

Furthermore, we impose some regularity conditions on the bases $\psi(\cdot)$; these are deferred to Assumption 26 in the supplementary material.

We arrive at the following theorem.

Theorem 4. *Suppose that Assumptions 9–14 and 23–26 in the supplementary material hold, and let*

$$h_n \propto n^{(\alpha-1)/(2\beta_1+1)}, \quad M \propto n^{\frac{2\beta_1(1-\alpha)}{(2\beta_1+1)(2\beta_2+1)}}, \quad \lambda_n \propto n^{-\frac{\beta_1(2\beta_2-1)(1-\alpha)}{(2\beta_1+1)(2\beta_2+1)}+2\gamma},$$

where $0 < \alpha < 1$, $0 < \gamma < H_1(\beta_1, \beta_2, \alpha)$, *and* $H_1(\beta_1, \beta_2, \alpha)$ *is a constant that depends only on* β_1, β_2 *and* α. *Then as* n *increases, the proposed procedure* (3.17) *correctly recovers the true graph, i.e.,* $\hat{S}_j = S_j$ *for all* $j = 1, \ldots, p$, *with probability converging to* 1, *if* $s = O(n^\gamma)$ *and* $pn \exp(-C_4 n^\alpha / \sigma^2) = o(1)$ *for some constant* C_4.

Because the regressors $\hat{\Psi}$ are estimated, establishing variable selection consistency requires extra attention. To prove Theorem 4, we must first establish variable selection consistency of group lasso regression with errors in variables. This generalizes the recent work on errors in variables for lasso regression (Loh and Wainwright, 2012). Theorem 4 ensures that the proposed method is able to recover the true graph exactly, given sufficiently dense observations in a finite time interval if the graph is sparse. The number of variables in the system can grow exponentially fast with respect to n, which means that the result holds for the "large p, small n" scenario.

Theorem 4 does not provide us with practical guidance for selecting the bandwidth h_n for the local polynomial regression estimator $\hat{X}_j$. The next result mirrors Theorem 4 for the bandwidths selected by cross-validation or GCV, which converge to $h_n \propto n^{-1/(2\beta_1+1)}$ asymptotically (see Xia and Li, 2002; Tsybakov, 2009 for details).

Proposition 4. *Suppose that Assumptions 9–14 and 23–26 in the supplementary material hold, and let*

$$h_n \propto n^{-1/(2\beta_1+1)}, \quad M \propto n^{\frac{1}{2\beta_2+1}(\frac{2\beta_1}{2\beta_1+1}-\alpha)}, \quad \textit{and} \quad \lambda_n \propto n^{-\frac{2\beta_2-1}{4\beta_2+2}(\frac{2\beta_1}{2\beta_1+1}-\alpha)+2\gamma},$$

where $0 < \alpha < \frac{2\beta_1}{2\beta_1+1}$, $0 < \gamma < H_2(\beta_1, \beta_2, \alpha)$, *and* $H_2(\beta_1, \beta_2, \alpha)$ *is a constant that depends only on* β_1, β_2 *and* α. *Then as* n *increases, the proposed procedure* (3.17) *correctly recovers the true graph, i.e.,* $\hat{S}_j = S_j$ *for all* $j = 1, \ldots, p$, *with probability converging to* 1, *if* $s = O(n^\gamma)$ *and* $pn \exp(-C_4 n^\alpha / \sigma^2) = o(1)$ *for some constant* C_4.

We note that selecting the values of M and λ_n that yield the rate specified in Proposition 4 is challenging in practice. The rate of convergence of the sparsity tuning parameter λ_n is slower in Proposition 4 compared to Theorem 4. This results in an increase in the minimum regulatory effect $f_{\min}$ because of the relation between $f_{\min}$ and λ_n in Assumption 14.

3.5 Numerical Experiments

We study the empirical performance of our proposal in three different scenarios in the following subsections. In what follows, given a set of initial conditions and a system of ODEs, numerical solutions of the ODEs are obtained using the Euler method with step size 0.001. Observations are drawn from the solutions at an evenly-spaced time grid $\{iT/n; i = 1, \ldots, n\}$ with independent $N(0, 1)$ measurement errors, unless specified otherwise. To facilitate the comparison of GRADE with other methods, we fit the smoothing estimates $\hat{X}$ in (3.17b) using smoothing splines with bandwidth chosen by GCV. We use cubic splines with two internal knots as the basis functions in (3.17c) in Sections 3.5.1 and 3.5.3. Linear basis functions are used in Section 3.5.2. The integral $\hat{\Psi}_k(t) = \int_0^t \psi\big(\hat{X}_k(u; h)\big)\, du$ in (3.17c) is calculated numerically with step size 0.01.

3.5.1 Variable selection in additive ODEs

In this simulation, we compare GRADE with NeRDS (Henderson and Michailidis, 2014) and SA-ODE (Wu et al., 2014) described in (3.14). We consider the following system of additive ODEs, for $k = 1, \ldots, 5$:

$$\begin{cases} X'_{2k-1}(t) = \theta_{2k-1,0} + \psi(X_{2k-1}(t))^{\mathrm{T}}\theta_{2k-1,2k-1} + \psi(X_{2k}(t))^{\mathrm{T}}\theta_{2k-1,2k} \\ X'_{2k}(t) = \theta_{2k,0} + \psi(X_{2k-1}(t))^{\mathrm{T}}\theta_{2k,2k-1} + \psi(X_{2k}(t))^{\mathrm{T}}\theta_{2k,2k} \end{cases}, t \in [0, 20], \tag{3.26}$$

where $\psi(x) = (x, x^2, x^3)^{\mathrm{T}}$ is the cubic monomial basis. The parameters and initial conditions are chosen so that the solution trajectories are identifiable under an additive model (Buja et al., 1989). Detailed specification of (3.26) can be found in Section B.3 of the supplementary material.

After generating data according to (3.26) and introducing noise, we apply GRADE, NeRDS, and SA-ODE to recover the directed graph encoded in (3.26). Both NeRDS and SA-ODE are implemented using code provided by the authors. NeRDS and SA-ODE use smoothing splines to estimate $\hat{X}$ and $\hat{X}'$ in (3.14b), and cubic splines with two internal knots as the basis ψ in (3.14a). As mentioned briefly in Section 3.2, NeRDS applies an additional smoothing penalty which amounts to an ℓ_2 penalty on θ_{jk} in (3.14a), controlled by a parameter selected using GCV (Henderson and

Michailidis, 2014). We apply GRADE using the same smoothing estimates and basis functions as NeRDS and SA-ODE. To facilitate a direct comparison to NeRDS, we apply GRADE both with and without an additional ℓ_2-type penalty on the θ_{jk}'s in (3.17a). We apply all methods for a range of values of the sparsity-inducing tuning parameter (e.g., λ_n in (3.17a)), in order to yield a recovery curve of varying sparsity.

We summarize the simulation results in Figure 3.1, where the numbers of true edges selected are displayed against the total numbers of selected edges over a range of sparsity tuning parameters. We see that GRADE outperforms the other two methods, which corroborates our theoretical findings in Section 3.4 that our proposed method is more efficient than methods such as NeRDS and SA-ODE which involve derivative estimation (see e.g., comments below Theorem 1).

3.5.2 Variable selection in linear ODEs

In this simulation, we compare GRADE to two recent proposals by Brunel et al. (2014) and Hall and Ma (2014). Recall from Section 3.2.2 that Brunel et al. (2014) and Hall and Ma (2014) are proposed to estimate a few unknown parameters in an ODE system of known form. Hence, we consider a simple linear ODE system, for $k = 1, \ldots, 4$,

$$\begin{cases} X'_{2k-1}(t) = 2k\pi X_{2k}(t) \\ X'_{2k}(t) = -2k\pi X_{2k-1}(t) \end{cases}, t \in [0, 1]. \tag{3.27}$$

For each $k = 1, \ldots, 4$, we set the initial condition to be $(X_{2k-1}(0), X_{2k}(0)) = (\sin(y_k), \cos(y_k))$ where $y_k \sim N(0, 1)$. The solutions to (3.27) take the form of sine and cosine functions of frequencies ranging from 2π to 8π. The graph corresponding to (3.27) is sparse, with only eight directed edges out of 64 possible edges. We fit the model

$$X'(t) = \Theta X(t) + C, \tag{3.28}$$

where Θ is an unknown 8×8 matrix and C is an 8-vector. We apply the method in Brunel et al. (2014) using the code provided by the authors. We implement the method in Hall and Ma (2014) in R based on the authors' code in Fortran. Because the loss function in Hall and Ma (2014) is

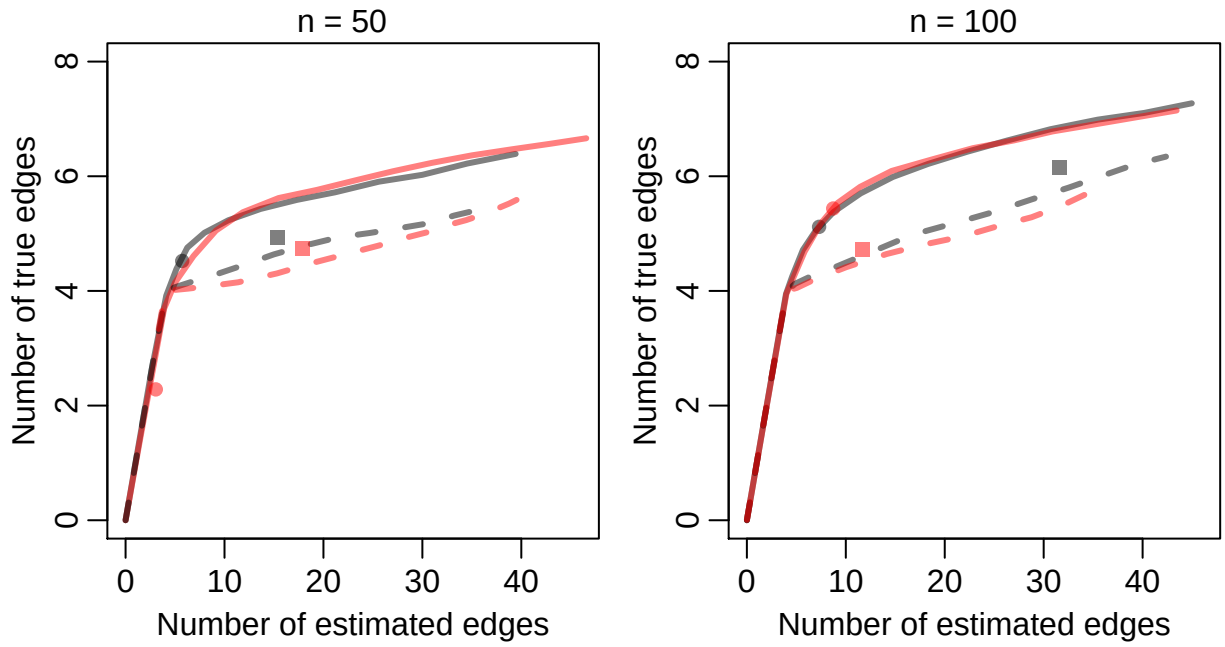

Figure 3.1: Performance of network recovery methods on the system of additive ODEs in (3.26), averaged over 400 simulations. The four curves represent SA-ODE (- -), NeRDS (- -), and GRADE without (—) and with (—) the additional smoothing penalty in (3.17a) used by NeRDS. Each point on the curves corresponds to average performance for a given sparsity tuning parameter λ_n in (3.14a) or (3.17a). The symbols indicate the sparsity tuning parameter λ_n selected using BIC (SA-ODE, ▪, and GRADE, • and •) or GCV (NeRDS, ▪).

not convex, we use five sets of random initial values and report the best performance. Since both Brunel et al. (2014) and Hall and Ma (2014) yield dense estimates for Θ in (3.28), in order to examine how well these methods recover the true graph, we threshold the estimates at a range of values in order to obtain a variable selection path. We apply GRADE using the linear basis function $\psi(x) = x$.

Results are shown in Figure 3.2. We can see that GRADE outperforms the methods in Brunel et al. (2014) and Hall and Ma (2014). This is likely due to the fact that GRADE exploits the sparsity of the true graph with a sparsity-inducing penalty. In principle, Brunel et al. (2014) and

Hall and Ma (2014) could be generalized in order to include penalties on the parameters. We leave this to future research.

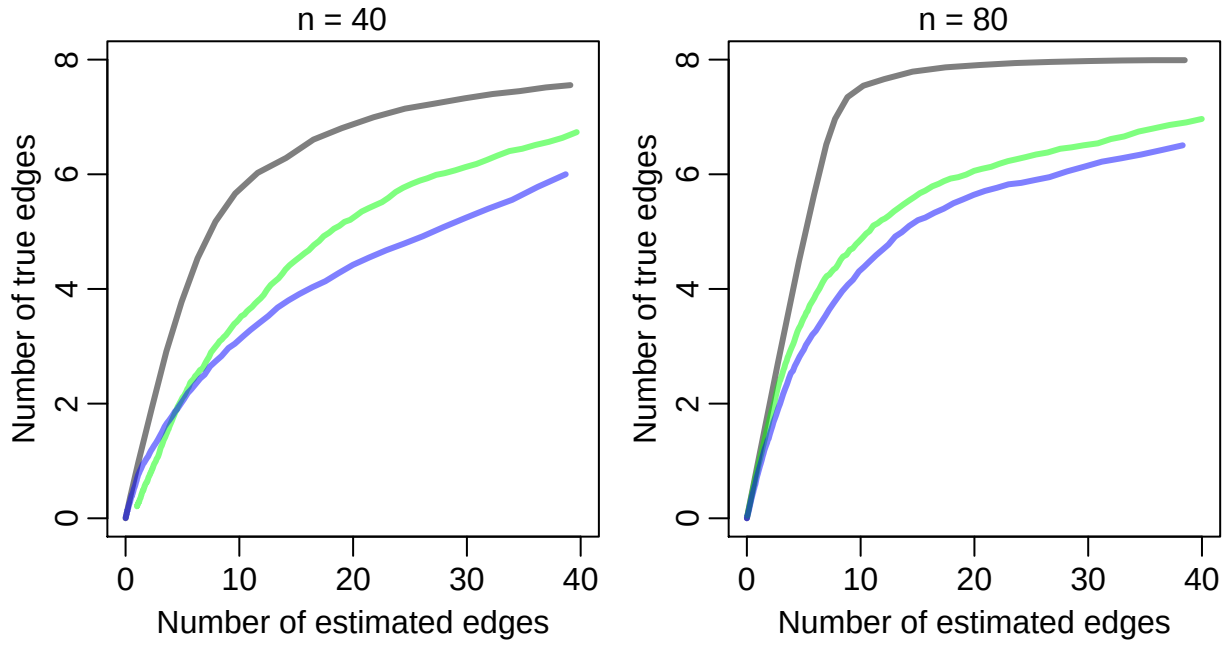

Figure 3.2: Network recovery on the system of linear ODEs (3.27), averaged over 200 simulated data sets. The three curves represent GRADE (—), Hall and Ma (2014) (—), Brunel et al. (2014) (—).

3.5.3 Robustness of GRADE to the additivity assumption

The GRADE method assumes that the true underlying model is additive (Assumption 11). However, in many systems, the additivity assumption is violated; for instance, multiplicative effects may be present in gene regulatory networks (Ma et al., 2009). In this subsection, we investigate the performance of GRADE in a setting where the true model is non-additive. We consider the

following system of ODEs, for $k = 1, \ldots, 5$,

$$\begin{cases} X'_{2k-1}(t) = f_{2k-1}\left(X_{2k-1}(t), X_{2k}(t)\right) \equiv 2X_{2k-1}(t) - vX_{2k-1}(t)X_{2k}(t) \\ X'_{2k}(t) = f_{2k}\left(X_{2k-1}(t), X_{2k}(t)\right) \equiv vX_{2k-1}(t)X_{2k}(t) - 2X_{2k}(t) \end{cases}, t \in [0, 5], \quad (3.29)$$

where v is a positive constant. For each $k = 1, \ldots, 5$, the pair of equations (3.29) is a special case of the Lotka-Volterra equations (Volterra, 1928), which represent the dynamics between predators (X_{2k}) and prey (X_{2k-1}). The parameter v defines the interaction between the two populations. For $v \neq 0$, both X'_{2k-1} and X'_{2k} are non-additive functions of X_{2k-1} and X_{2k}. We define two types of directed edges, where $\mathcal{E}_1 \equiv \{(X_j, X_j), j = 1, \ldots, 10\}$ and $\mathcal{E}_2 \equiv \{(X_{2k-1}, X_{2k}), (X_{2k}, X_{2k-1}),\ k = 1, \ldots, 5\}$ represent the self-edges and non-self-edges, respectively. Figure 3.3(a) contains an illustration of the graph and edge types for each pair of equations. In what follows, we investigate how well GRADE recovers these two types of edges as we change the parameter v, i.e., as the additivity assumption is violated.

Since measurement error is not essential to the current discussion, we generate data according to (3.29) without adding noise. To ensure that the trajectories are identifiable, we generate $R = 2$ sets of random initial values drawn from $N_{10}(0, 2I_{10})$, where I_{10} is a 10×10 identity matrix. In order to quantify the amount of signal in an edge that GRADE can detect, we introduce the quantity

$$D_{j,k}(v) = \mathbb{E}\left[R \int_0^T \left\{\frac{\partial f_j}{\partial X_k}\left(t; X(0)\right)\right\}^2 dt\right], \quad (3.30)$$

where the expectation is taken with respect to the random initial values $X(0)$ and R is the number of initial values. The measure $D_{j,k}$ in (3.30) is a loose analogy to $\left\{\int_0^1 \left[f^*_{jk}(X^*_k(t))\right]^2 dt\right\}^{1/2}$ used in Assumption 14. Note that if no edge is present from X_k to X_j, then $\partial f_j / \partial X_k \equiv 0$ and hence $D_{j,k}(v) = 0$. One immediately notes that, as R increases, the regulatory effect for a true edge increases proportionally to R, while the regulatory effect of a non-edge remains zero. For the self-edges in $\mathcal{E}_1$ and the non-self-edges in $\mathcal{E}_2$, we can define $D^{(1)}(v)$ and $D^{(2)}(v)$ as

$$D^{(1)}(v) = \min_{k=1,\ldots,10} D_{k,k}(v), \quad \text{and} \quad D^{(2)}(v) = \min_{k=1,\ldots,5} \{D_{2k-1,2k}(v), D_{2k,2k-1}(v)\}, \quad (3.31)$$

where we use the minimum because variable selection is limited by the minimum regulatory effect (see Assumption 14). With a slight abuse of definition, we refer to (3.31) as the minimum

regulatory effects in a non-additive model.

We apply GRADE using the formulation in (3.18). The sparsity parameter λ is chosen so that there are 20 directed edges in the estimated network. We record the number of estimated edges that are in $\mathcal{E}_1$ and $\mathcal{E}_2$. The edge recovery performance is shown in Figure 3.3(b). In Figure 3.3(c), we display the minimum regulatory effects defined in (3.31). Edge recovery and minimum regulatory effects show a similar trend as a function of r in (3.29). This suggests that (3.31), and thus (3.30), is a reasonable measure of the additive components of the regulatory effect of the edges. The slight deviation between the trends reflects the fact that the measure defined in (3.30) is not a direct counterpart of $\left\{\int_0^1 \left[f_{jk}^*(X_k^*(t))\right]^2 dt\right\}^{1/2}$ in a non-additive model. The edge recovery improves when a larger value of R is used, though these results are omitted due to space constraints. Our results indicate that GRADE can recover the true graph even when the additivity assumption is violated, provided that the regulatory effects (3.30) for the true edges are sufficiently large.

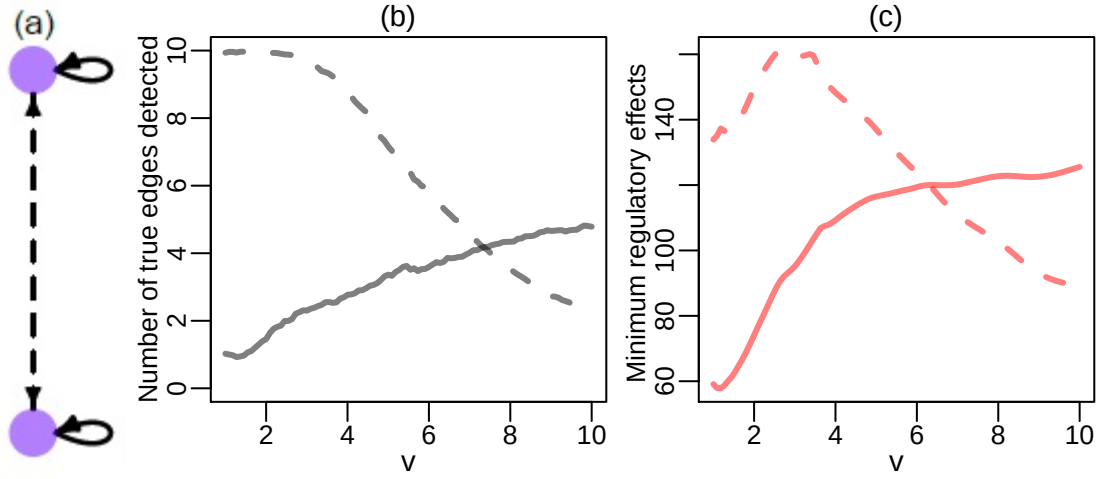

Figure 3.3: (a): The graph encoded by a pair of Lotka-Volterra equations as given in (3.29). Self-edges (—) and non-self-edges (- -) are shown. (b): Self-edge (—) and non-self-edge (- -) recovery of GRADE, averaged over 200 simulated data sets. (c): Minimum signals defined in (3.31), for self-edges, $D^{(1)}(\cdot)$ (—), and non-self-edges, $D^{(2)}(\cdot)$ (- -).

3.6 Applications

3.6.1 Application to in silico *gene expression data*

GeneNetWeaver (GNW) provides an *in silico* benchmark for assessing the performance of network recovery methods (Schaffter et al., 2011), and was used in the third DREAM challenge (Marbach et al., 2009). GNW is based upon real gene regulatory networks of yeast and *E. coli*. It extracts sub-networks from the yeast or *E. coli* gene regulatory networks, and assigns a system of ODEs to the extracted network. This system of ODEs is non-additive, and includes unobserved variables (Marbach et al., 2010). Therefore, the assumptions of GRADE are violated in the GNW data.

To mimic real-world laboratory experiments, GNW provides several data generation mechanisms. In this study, we consider data from the *perturbation* experiments. The perturbation experiments are similar to the data generating mechanisms used in Section 3.5.3, where initial conditions of the ODE system are perturbed in order to emulate the diversity of trajectories from multiple independent experiments.

We investigate ten networks from GNW that have been previously studied in Henderson and Michailidis (2014), of which five have 10 nodes and five have 100 nodes. For each network, GNW provides one set of noiseless gene expression data consisting of R perturbation experiments where the trajectories are measured at $n = 21$ evenly-spaced time points in $[0, 1]$. Here $R = 10$ for the five 10-node networks and $R = 100$ for the five 100-node networks. As in Henderson and Michailidis (2014), we add independent $N\left(0, 0.025^2\right)$ measurement errors to the data at each timepoint.

We apply NeRDS as described in Henderson and Michailidis (2014). We apply GRADE using the formulation (3.18) to handle observations from multiple experiments, with the smoothing estimates $\hat{X}$ in (3.17b) fit using smoothing splines with bandwidth chosen by GCV, and using cubic splines with two internal knots as the basis functions in (3.17c). The integral $\hat{\Psi}_k(t) = \int_0^t \psi\left(\hat{X}_k(u; h)\right) du$ in (3.17c) is calculated numerically with step size 0.01. Finally, we apply an additional ℓ_2-type penalty to the θ_{jk}'s in (3.18) in order to match the setup of NeRDS. The tuning parameter for this penalty is set to be 0.1.

Results are shown in Table 3.1. Recall that the data generating mechanism violates crucial

assumptions for both NeRDS and GRADE. We see in Table 3.1 that NeRDS outperforms GRADE in one network, while GRADE outperforms NeRDS in the other nine networks. This suggests that GRADE is a competitive exploratory tool for reconstructing gene regulatory networks.

Table 3.1: Area Under ROC Curves for NeRDS and GRADE

	$p = 10$		$p = 100$	
	NeRDS	GRADE	NeRDS	GRADE
Ecoli1	$0.450\ (0.438, 0.462)$	$\mathbf{0.545}\ (0.534, 0.557)$	$0.624\ (0.622, 0.627)$	$\mathbf{0.670}\ (0.667, 0.673)$
Ecoli2	$0.512\ (0.502, 0.523)$	$\mathbf{0.643}\ (0.634, 0.653)$	$0.637\ (0.635, 0.640)$	$\mathbf{0.653}\ (0.650, 0.656)$
Yeast1	$0.486\ (0.476, 0.495)$	$\mathbf{0.679}\ (0.666, 0.691)$	$0.610\ (0.607, 0.612)$	$\mathbf{0.636}\ (0.635, 0.638)$
Yeast2	$0.525\ (0.518, 0.532)$	$\mathbf{0.607}\ (0.600, 0.613)$	$0.568\ (0.566, 0.569)$	$\mathbf{0.584}\ (0.582, 0.585)$
Yeast3	$0.467\ (0.460, 0.474)$	$\mathbf{0.576}\ (0.566, 0.587)$	$\mathbf{0.617}\ (0.616, 0.619)$	$0.567\ (0.566, 0.568)$

The average area under the curves and 90% confidence intervals, over 100 simulated data sets. Networks and data generating mechanisms are described in Section 3.6.1. Boldface indicates the method with larger AUC.

3.6.2 Application to calcium imaging recordings

In this section, we consider the task of learning regulatory relationships among populations of neurons. We investigate the calcium imaging recording data from the Allen Brain Observatory project conducted by the Allen Institute for Brain Science[1]. Here, we investigate one of the experiments in the project. In this experiment, calcium fluorescence levels (a surrogate for neuronal activity) are recorded at 30 Hz on a region of the primary visual cortex while the subject mouse is shown forty visual stimuli. The forty visual stimuli are combinations of eight spatial orientations and five

[1]Website: ©2016 Allen Institute for Brain Science. Allen Brain Observatory [Internet]. Available from: `http://observatory.brain-map.org`.

temporal frequencies. Each stimulus lasts for two seconds and is repeated 15 times. The recorded videos are processed by the Allen Institute to identify individual neurons. In this particular experiment, there are 575 neurons. Each neuron's activity is defined as the average calcium fluorescence level of the pixels that it covers in the video.

It is known that the activities of individual neurons are noisy and sometimes misleading (Cunningham and Byron, 2014). As an alternative, neuronal populations can be studied (see e.g., Part Three of Gerstner et al., 2014). We define 25 neuronal populations by dividing the recording region into a 5×5 grid, where each population contains roughly 20 neurons. We use GRADE to capture the functional connectivity among the 25 neuronal populations. Note that functional connectivity is distinct from physical connectivity. Functional connectivity involves the relationships among neuronal populations that can be observed through neuron activities and may change across stimuli, whereas physical connectivity consists of synaptic interactions.

We estimate the functional connectivity corresponding to three different but related stimuli, consisting of frequencies of 1 Hz, 2 Hz, and 4 Hz, each at a spatial orientation of $90°$. For each stimulus, we have calcium fluorescence levels of the $p = 25$ neuronal populations for each of $R = 15$ repetitions. Since each repetition spans two seconds and the calcium fluorescence is recorded at 30 Hz, there are 60 timepoints per repetition. We apply GRADE using the formulation in (3.18) in order to reconstruct the functional connectivity under each of the three stimuli. We use smoothing splines with bandwidth h selected with GCV in order to estimate $\hat{X}$ in (3.17b), and use cubic splines with 4 internal knots as the basis functions $\psi(\cdot)$ in (3.17c). The sparsity parameter $\lambda_{j,n}$ for each nodewise regression in (3.18) is selected using BIC for each $j = 1, \ldots, 25$. For ease of visualization, we prefer a sparse network, and so we fit GRADE using tuning parameter values $\alpha(\lambda_{1,n}, \ldots, \lambda_{p,n})$, where the scalar α is selected so that each of the estimated networks contains approximately 25 edges.

Estimated functional connectivities are shown in Figure 3.4. We see that, in all three networks, the 24th neuronal population regulates many other neuronal populations, indicating that this region may contain neurons that are sensitive to this spatial orientation. Furthermore, we see that the adjacent connectivity networks in Figure 3.4 are somewhat similar to each other, whereas the

networks at 1 Hz and 4 Hz have few similarities. This agrees with the observation in neuroscience that neurons in the mouse primary visual cortex are responsive to a somewhat narrow range of temporal frequencies near their peak frequencies (see, e.g., Gao et al., 2010).

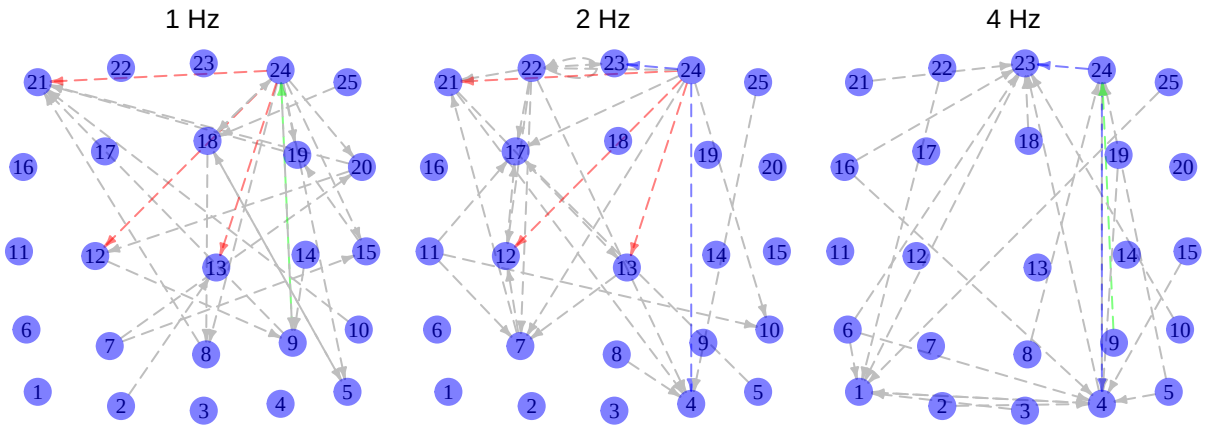

Figure 3.4: Estimated functional connectivities among neuronal populations from the calcium imaging data described in Section 3.6.2. Each node is positioned near the center of the neuronal population it represents, with jitter added for ease of display. The three red edges are shared between the estimated networks at 1 Hz and 2 Hz; the two blue edges are shared between estimated networks at 2 Hz and 4 Hz; the single green edge is shared between the estimated networks at 1 Hz and 4 Hz. For reference, given two Erdös-Rènyi graphs consisting of 25 nodes and 25 edges, the probability of having three or more shared edges is 0.07, and the probability of having two or more shared edges is 0.26.

3.7 Discussion

In this study, we propose a new approach, GRADE, for estimating a system of high-dimensional additive ODEs. GRADE involves estimation of an integral rather than a derivative. We show that estimating the integral is superior to estimating the derivatives both theoretically and empirically. We leave an extension of our work to non-additive ODEs to future research.

In this study, we have not addressed the issue of experimental design. Given a finite set of resources, one may choose to design an experiment to measure n observations on a very dense time grid, or on a coarse time grid. Alternatively, one might choose to measure n/R observations for R distinct experiments from a single ODE system (3.1), each with a different initial condition. This presents a trade-off that is especially interesting in the context of ODEs: using a dense time grid improves the quality of the smoothing estimates $\hat{X}$, as seen in Sections 3.5.1 and 3.5.2, while running multiple experiments enhance the identifiability of the true structure, as seen in Section 3.5.3. We leave a more detailed treatment of these issues to future work.

Chapter 4

GRAPHICAL ESTIMATION FOR HAWKES PROCESSES

4.1 Introduction

Hawkes (1971) proposed a class of point processes that have the so-called *mutually exciting* property: a past event may trigger the occurrence of future events. The Hawkes process and its variants are widely applied in modeling recurrent events in many fields, with notable applications in modeling earthquakes (Ogata, 1988), crime rates (Mohler et al., 2011), interactions in social networks (Simma and Jordan, 2012; Perry and Wolfe, 2013), financial events (Chavez-Demoulin et al., 2005; Bowsher, 2007; Aït-Sahalia et al., 2015), and spiking histories of neurons (see e.g., Brillinger, 1988; Okatan et al., 2005; Paninski et al., 2007; Pillow et al., 2008).

In recent years, new challenges are presented to researchers as the size of data grows dramatically. Consider, for instance, the challenges in modeling a neuronal network. The Hawkes process has been used to model neuronal networks (Brillinger, 1988; Okatan et al., 2005; Paninski et al., 2007; Pillow et al., 2008), but the statistical methods quickly become computationally unfeasible as the size of the neuronal network grows. In the meantime, new technologies make it possible to record simultaneous activities of tens of thousands of neurons (Ahrens et al., 2013). Given the large scale of the available data, any statistical method for this task must be computationally efficient. Furthermore, give the complex dynamics of the neurons, a suitable statistical method must be flexible, data-driven, and statistically efficient.

Moreover, there is currently a significant gap between the applications and statistical theory on the Hawkes process. Hawkes (1971) considered the case when an event *excites* the process, in the sense that the event may trigger more future events. Later, Hawkes and Oakes (1974) discovered a cluster process representation for the mutually-exciting Hawkes process, which has become an essential tool of many theoretical results on the Hawkes process (Reynaud-Bouret, 2003; Reynaud-

Bouret and Schbath, 2010; Hansen et al., 2015; Bacry et al., 2015). However, events might not always be excitatory in many important applications of the Hawkes process. For instance, it is well-known that a spike of one neuron may *inhibit* the activities of other neurons, meaning that there might be fewer future spikes of other neurons. In the presence of inhibitory relationships, the cluster process representation by Hawkes and Oakes (1974) is no longer available, and many existing theoretical results do not apply.

In this study, we consider the estimation of the graph encoded in a Hawkes process from possibly high-dimensional data. We defer the details to Section 4.2.2 after we introduce some necessary notations and concepts about the Hawkes process.

The main contributions of this study are three-fold:

1. We propose a flexible procedure for fitting a Hawkes process non-parametrically. Under suitable assumptions, we establish that the proposed procedure has the oracle inequality and variable selection consistency in the high-dimensional setting, in which the number of neurons grows much more quickly than the length of observed period.

2. We propose two simple screening procedures that can be applied *before* fitting the Hawkes process, in order to alleviate some of the computational burdens resulted from high dimensionality. We also show that these screening procedures will include the true edges under mild conditions.

3. We establish a new concentration inequality for the Hawkes process. To this end, we develop a new representation of the Hawkes process that applies beyond mutually-exciting processes.

4.2 Background and Literature Review

4.2.1 Notation

We use the notation $f * g(t) \equiv \int_{-\infty}^{\infty} f(\Delta) g(t-\Delta) d\Delta$ to indicate the convolution of two functions, f and g. We use $\|a\|_2$ to denote the ℓ_2-norm of a vector $a \in \mathbb{R}^p$. Furthermore, $\|f\|_{2,[l,u]} \equiv$

$\left\{ \int_l^u f^2(t)dt \right\}^{1/2}$ will denote the ℓ_2-norm of a function f on the interval $[l, u]$. We use $\Gamma_{\min}(\boldsymbol{A})$ for the minimum eigenvalue of a square matrix $\boldsymbol{A}$ and $\Gamma_{\max}(\boldsymbol{A})$ for its maximum eigenvalue.

4.2.2 Background

In this section, we provide a very brief review of point processes in general, and the Hawkes process in particular. We refer interested readers to the monograph by Daley and Vere-Jones (2003) for a comprehensive discussion of point processes.

A Brief Review of Point Processes

We define the p-variate *point process* – or the *counting process* – as $\boldsymbol{N} = (N_1, \ldots, N_p)^{\mathrm{T}}$. For $j = 1, \ldots, p$, let $\{t_{j,1}, t_{j,2}, \ldots\}$ be the event times of the jth point process N_j on $\mathbb{R}^+$. In this notation, $N_j(A) = \sum_i \mathbf{1}_{[t_{j,i} \in A]}$ for $A \in \mathcal{B}(\mathbb{R}^+)$, where $\mathcal{B}(\mathbb{R}^+)$ denotes the Borel σ-algebra of the positive half of real line. We use $\boldsymbol{N}(t)$ as a short-hand notion for $\boldsymbol{N}\big([0, t]\big)$, and write $\boldsymbol{N}\big([t, t+dt)\big)$ as $d\boldsymbol{N}(t)$, where dt denotes an arbitrary small increment of t. Let $\mathcal{H}_t$ be the *history* of $\boldsymbol{N}$ up to time t where $\mathcal{H}_t \equiv \sigma(\boldsymbol{N}(A \cap [0, t)), A \in \mathcal{B}(\mathbb{R}^+))$. The *intensity process* $\boldsymbol{\lambda}(t) = (\lambda_1(t), \ldots, \lambda_p(t))^{\mathrm{T}}$ is a p-variate $\mathcal{H}_t$-predictable process defined as

$$\lambda_j(t)dt = \mathbb{P}(dN_j(t) = 1 \mid \mathcal{H}_t),\ j = 1, \ldots, p. \tag{4.1}$$

We define the *mean intensity* $\boldsymbol{\Lambda} = (\Lambda_1, \ldots, \Lambda_p)^{\mathrm{T}} \in \mathbb{R}^p$ where, for $j = 1, \ldots, p$,

$$\Lambda_j \equiv \mathbb{E}[dN_j(t)]/dt, \tag{4.2}$$

and the *(infinitesimal) cross-covariance* $\boldsymbol{V}(\cdot) = \big(V_{j,k}(\cdot)\big)_{p \times p} : \mathbb{R} \mapsto \mathbb{R}^{p \times p}$, where, for any $\Delta \in \mathbb{R}$ and $1 \leq j, k \leq p$,

$$V_{j,k}(\Delta) \equiv \mathbb{E}[dN_j(t)dN_k(t - \Delta)]/\{dtd(t - \Delta)\} - \Lambda_j\Lambda_k - \mathbf{1}_{[j=k]}\Lambda_k\delta(\Delta) \tag{4.3}$$

following the definition in Equation 5 of Hawkes (1971). Here $\delta(\cdot)$ is the Dirac delta function, which satisfies $\delta(x) = 0$ for $x \neq 0$ and $\int_{-\infty}^{\infty} \delta(x)dx = 1$.

Note also that the mean intensity and cross-covariance are well-defined provided that $\boldsymbol{N}$ is stationary in time. We defer the discussion of stationarity in the context of the Hawkes process to Section 4.2.2.

A Brief Overview of the Hawkes Process

For the *linear* Hawkes process Hawkes (1971), the intensity function (4.1) takes the form

$$\lambda_j(t) = \mu_j + \sum_{k=1}^{p} \big(\omega_{j,k} * dN_k\big)(t), \quad j = 1, \ldots, p, \tag{4.4}$$

where

$$\big(\omega_{j,k} * dN_k\big)(t) \equiv \int_0^\infty \omega_{j,k}(\Delta) \sum_{i:t_{k,i}\le t} \delta(t - \Delta - t_{k,i}) d\Delta = \sum_{i:t_{k,i}\le t} \omega_{j,k}(t - t_{k,i}).$$

We refer to $\mu_j \in \mathbb{R}$ as the *background intensity*, and $\omega_{j,k}(\cdot) : \mathbb{R}^+ \mapsto \mathbb{R}$ as the *transfer function*. The right-hand side of (4.4) is sometimes transformed by a *link function*, as in a generalized linear model; this leads to a non-linear Hawkes process.

If the Hawkes process defined in (4.4) is stationary, we then have the following relationships between the first two moments (i.e., $\boldsymbol{\Lambda}$ and $\boldsymbol{V}$) and the background intensity $\boldsymbol{\mu} \equiv (\mu_1, \ldots, \mu_p)^{\mathrm{T}}$ and the transfer functions $\boldsymbol{\omega} \equiv (\omega_{j,k})_{p\times p}$ (see e.g., Equations 21 and 22 in Hawkes, 1971 or Theorem 1 in Bacry and Muzy, 2014):

$$\boldsymbol{\Lambda} = \boldsymbol{\mu} + \left[\int_0^\infty \boldsymbol{\omega}(\Delta) d\Delta\right] \boldsymbol{\Lambda}, \tag{4.5}$$

and

$$\boldsymbol{V}(\Delta) = \boldsymbol{\omega}(\Delta)\mathrm{diag}(\boldsymbol{\Lambda}) + (\boldsymbol{\omega} * \boldsymbol{V})(\Delta), \tag{4.6}$$

where $[\boldsymbol{\omega} * \boldsymbol{V}]_{j,k}(\Delta) \equiv \sum_{i=1}^{p} [\omega_{j,i} * V_{i,k}](\Delta)$. The second equation (4.6) belongs to a class of integral equations known as the *Wiener-Hopf integral equations*.

For any fixed p, Brémaud and Massoulié (1996) establish that the Hawkes process with intensity function (4.4) is stationary given the following assumption.

Assumption 15. Let $\boldsymbol{\Omega}$ be a $p \times p$ matrix whose entries are $\Omega_{j,k} = \int_0^\infty |\omega_{j,k}(\Delta)| \, d\Delta$, for $j, k = 1, \ldots, p$. We assume that $\Gamma_{\max}(\boldsymbol{\Omega}) \le \gamma_\Omega < 1$.

In the Gaussian graphical model, a similar assumption, proposed by Anandkumar et al. (2012), is known as *walk summability*. To see this, consider the case when $\omega_{j,k}(\Delta) \geq 0$ for all Δ, so that $\mathbf{\Omega} = \int_0^\infty \boldsymbol{\omega}(\Delta) d\Delta$. We can rewrite (4.5) as

$$\mathbf{\Lambda} = \sum_{i=0}^{\infty} \mathbf{\Omega}^i \boldsymbol{\mu}, \tag{4.7}$$

where $\mathbf{\Omega}^i$ is the ith power of the matrix $\mathbf{\Omega}$. In (4.7), $\mathbf{\Omega}^i \mu$ can be seen as the intensity induced through paths of length i. In fact, $\mathbf{\Omega}^i \boldsymbol{\mu}$ characterizes the temporal dependence of the Hawkes process $\boldsymbol{N}$ (see Appendix C.4.1 for more discussion). Assumption 15 ensures that $\|\mathbf{\Omega}^i \boldsymbol{\mu}\|_2 \leq \gamma_\Omega^i \|\boldsymbol{\mu}\|_2$; in other words, the induced intensity decreases exponentially fast as k grows.

In this study, we investigate the non-asymptotic setting, in which both p and T can be very large. Without additional constraints, $\|\mathbf{\Omega}^i \boldsymbol{\mu}\|$ can be on the order of p, which means that the temporal dependence can be arbitrarily strong. To help our discussion, we restrict our discussion to the Hawkes process that satisfies the following assumption on $\mathbf{\Omega}$.

Assumption 16. There exist a pair of positive constants (d_0, ρ_0) such that, for all p, $\mathbf{\Omega}$ satisfies $\|\mathbf{\Omega}^{d_0} \mathbf{1}\|_2 \leq \rho$.

Assumption 16 holds, for instance, when $\mathbf{\Omega}$ is the adjacency matrix of, e.g., a stochastic block model or a sparse Erdös-Rényi graph (Anandkumar et al., 2012).

We now define a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with the node set $\mathcal{V} = \{1, \ldots, p\}$ and the edge set

$$\mathcal{E} \equiv \{(j,k) : \exists\, \Delta \text{ such that } \omega_{j,k}(\Delta) \neq 0, 1 \leq j, k \leq p\}, \tag{4.8}$$

for $\omega_{j,k}$ given in (4.4). Recently, authors have connected $\mathcal{G}$ to the notions of directed information (Quinn et al., 2010) and Granger causality (Eichler et al., 2015).

In this study, we propose a new method for learning the structure of the directed graph $\mathcal{G}$: or equivalently, for estimating the edge set $\mathcal{E}$.

4.2.3 Existing Methods for Learning the Graphical Structure

It is clear from the previous section that learning the structure of the directed graph $\mathcal{G}$ reduces to identifying the non-zero transfer functions in (4.4). Existing methods for this task generally take

either an *intensity-based approach* or a *moment-based approach*.

The Intensity-Based Approach

The intensity-based approach involves fitting the model (4.4) by maximizing the likelihood (Ogata, 1981)

$$\sum_{j=1}^{p} \int_0^T \big[\log\big(\lambda_j(t)\big) dN_j(t) - \lambda_j(t) dt \big], \tag{4.9}$$

or by minimizing the squared error loss

$$\frac{1}{2T} \sum_{j=1}^{p} \int_0^T \big[\lambda_j^2(t)\, dt - 2\lambda_j(t) dN_j(t) \big]. \tag{4.10}$$

Several authors have considered learning the structure of the graph by conducting hypothesis tests, or by thresholding the estimated transfer functions (Brillinger, 1988; Okatan et al., 2005; Eldawlatly et al., 2009; Masud and Borisyuk, 2011; Berry et al., 2012; Eichler et al., 2015). Recently, a few authors have proposed to recover the graph by applying sparsity-inducing penalties to (4.9) or (4.10) (Reynaud-Bouret and Schbath, 2010; Lewis and Mohler, 2011; Simma and Jordan, 2012; Song et al., 2013; Zhou et al., 2013; Bacry et al., 2015; Hansen et al., 2015; Xu et al., 2016). Proposals have also been made to recover the graph using a Bayesian framework, with a sparsity-inducing prior (Linderman and Adams, 2014) or a random network prior (Blundell et al., 2012; Perry and Wolfe, 2013).

Under the assumption that the transfer functions $\omega_{j,k}$ in (4.4) are exponential, Simma and Jordan (2012) proposed a novel distributional expectation-maximization algorithm that can be applied to very large systems (Zhou et al., 2013; Xu et al., 2016). However, if $\omega_{j,k}$ is not exponential, this approach is computationally prohibitive when p is large, as is increasingly the case for contemporary neuronal data sets.

Analysis of the intensity-based approach follows mostly from martingale theory using the fact that $dN_j(t) - \lambda_j(t)dt$ is a martingale. Hence, standard theory for martingales applies in the classical setting, in which p is fixed (see e.g., Part II of Williams (1991)). When the number of unknown parameters is allowed to grow, oracle inequalities for penalized intensity-based estimators have

been established for the Hawkes process by Reynaud-Bouret and Schbath (2010), Hansen et al. (2015) and Bacry et al. (2015). However, these results require the transfer functions $\boldsymbol{\omega}$ to be non-negative, due to some technical constraints that we will visit in Section 4.5.

The Moment-Based Approach

Recall that the transfer functions in (4.4) , the mean intensity (4.2), and the cross-covariance functions (4.3) are connected via the set of equations (4.5) and (4.6). The moment-based approach involves plugging in estimates of the mean intensity functions and the cross-covariance functions — that is, the first and second moments of $d\boldsymbol{N}$ — into (4.5) and (4.6), and then solving for the transfer functions (Brillinger et al., 1976; Krumin et al., 2010; Bacry and Muzy, 2014; Etesami et al., 2016).

If the transfer functions $\omega_{j,k}$ have an unknown form, then solving the integral equations (4.5) and (4.6) nonparametrically is computationally intensive. Thus far, work in this area has been restricted to systems with small p (Brillinger et al., 1976; Krumin et al., 2010; Bacry and Muzy, 2014). Recently, Etesami et al. (2016) propose a moment-based estimator for large systems, under the assumption that the transfer functions are exponential functions.

In order to analyze the moment-based estimator, we need to establish the properties of estimators for the first and second moments of the Hawkes process. The event times in one realization of the Hawkes process have inherent temporal dependence by definition – except for the trivial case when $\boldsymbol{\omega} \equiv \mathbf{0}$. However, classical theory of high dimensional statistics relies heavily on the independence assumption (see e.g., Bühlmann and van de Geer (2011)). Existing work often assumes that there are multiple independent realizations of the same process (Bacry and Muzy, 2014; Etesami et al., 2016). But it is often the case that only one realization is available. In this case, it is necessary to define asymptotics in terms of the growth of the observe period (i.e., T). The discussion of this type is limited in the current literature, and is restricted to cases when all transfer functions are non-negative (Reynaud-Bouret, 2003; Reynaud-Bouret and Schbath, 2010; Hansen et al., 2015).

4.3 Graph Reconstruction via Penalized Regression

4.3.1 Estimation Procedure

We assume that the event times follow a Hawkes process whose intensities are of the form (4.4) with unknown transfer functions $\omega_{j,k}$. In this section, we propose to estimate the transfer functions non-parametrically using the framework of high-dimensional additive models (Ravikumar et al., 2009; Meier et al., 2009). A group lasso penalty (Yuan and Lin, 2006) is applied to encourage sparsity in the estimate of the edge set $\mathcal{E}$ (4.8).

As noted in Section 4.2.3, a few authors have recently used penalized regression to estimate $\mathcal{E}$, assuming that the form of the transfer functions $\omega_{j,k}$ is known (Lewis and Mohler, 2011; Simma and Jordan, 2012; Song et al., 2013; Zhou et al., 2013; Bacry et al., 2015; Xu et al., 2016) or unknown (Reynaud-Bouret and Schbath, 2010; Hansen et al., 2015). Our proposal is most related to, and is inspired by, the work of Hansen et al. (2015). Hansen et al. (2015) fit the multivariate Hawkes process with a modified group lasso penalty, and established the oracle inequality for the estimation procedure under the assumption that the transfer function is non-negative, i.e. $\omega_{j,k}(\Delta) \geq 0$ for all Δ. We relax this assumption and propose a penalty that is invariant to the signs of the transfer functions. We further establish variable selection consistency of the proposed procedure in Section 4.3.2.

We approximate the transfer functions using an M-dimensional basis, $\psi(t) = (\psi_1(t), \ldots, \psi_M(t))^{\mathrm{T}}$, so that

$$\omega_{j,k}(t) = \psi(t) \cdot \beta_{j,k} + r_{j,k}(t), \tag{4.11}$$

where $r_{j,k}(\cdot)$ denotes the residual, and where $\beta_{j,k} \in \mathbb{R}^M$. In Section 4.3.2, we will allow M to increase as the observed time period T grows.

We fit the model (4.4) for $j = 1, \ldots, p$, using the loss function (4.10). Note that (4.10) can be decomposed into p separate loss functions, one per node. Let $\Psi_k(t) \equiv \big(\psi * dN_k\big)(t)$ denote the convolution of the bases $\psi(\cdot)$ with the history of dN_k up to time t. For $j = 1, \ldots, p$, plugging

$\omega_{j,k}(t) \approx \psi(t) \cdot \beta_{j,k}$ into (4.4) and then (4.10), yields

$$-\frac{1}{T}\int_0^T \left[\mu_j + \sum_{k=1}^{p} \Psi_k(t) \cdot \beta_{j,k}\right] dN_j(t) + \frac{1}{2T}\int_0^T \left[\mu_j + \sum_{k=1}^{p} \Psi_k(t) \cdot \beta_{j,k}\right]^2 dt.$$

Our estimates $\hat{\mu}_k$, $\hat{\beta}_{j,k}$, $k = 1, \ldots, p$ solve the optimization problem

$$\begin{aligned} \underset{\mu_j \in \mathbb{R}, \beta_{j,k} \in \mathbb{R}^M}{\text{minimize}} & -\frac{1}{T}\int_0^T \left[\mu_j + \sum_{k=1}^{p} \Psi_k(t) \cdot \beta_{j,k}\right] dN_j(t) \\ & + \frac{1}{2T}\int_0^T \left[\mu_j + \sum_{k=1}^{p} \Psi_k(t) \cdot \beta_{j,k}\right]^2 dt + \eta_j \frac{1}{\sqrt{T}} \sum_{k=1}^{p} \left\|\Psi_k \cdot \beta_{j,k}\right\|_{2,[0,T]}, \end{aligned} \tag{4.12}$$

where $\|\Psi_k \cdot \beta_{j,k}\|_{2,[0,T]}$ is a standardized group lasso penalty (Simon and Tibshirani, 2012). Problem (4.12) is convex and can be solved efficiently using, e.g., a block coordinate descent algorithm (see Appendix C.1). The tuning parameter η_j controls the sparsity of $\hat{\beta}_{j,k}$, $k = 1, \ldots, p$. It can be chosen using the Bayesian information criterion, the extended Bayesian information criterion (Chen and Wu, 2008), or other using parameter selection criteria.

Using (4.12), the estimator of the true edge set $\mathcal{E}$ in (4.8) is

$$\widehat{\mathcal{E}} \equiv \cup_{j=1}^{p} \big\{(j,k) : \left\|\Psi_k \cdot \hat{\beta}_{j,k}\right\|_{2,[0,T]} \neq 0, k = 1, \ldots, p\big\}. \tag{4.13}$$

Remark 5. In (4.12), $T^{-1/2}\left\|\Psi_k \cdot \beta_{j,k}\right\|_{2,[0,T]} \equiv T^{-1/2}\{\int_0^T [\Psi_k(t)\beta_{j,k}]^2 dt\}^{1/2}$ is a standardized group lasso penalty. Hansen et al. (2015) considered a variant of this penalty, which takes the form

$$\left\{\frac{1}{T}\int_0^T \left[\Psi_k(t) \cdot \beta_{j,k}\right]^2 dN_j(t)\right\}^{1/2}. \tag{4.14}$$

In (4.14), $\Psi_k(t) \cdot \beta_{j,k}$ contributes to the penalty only when there is an event in $dN_j(t)$. We show in Section 4.6.1 through numerical experiments that this penalty leads to inferior edge recovery when negative transfer functions are present.

Remark 6. A closer look at the loss function (4.12) reveals the connection between the squared error loss (4.10) and equations (4.5) and (4.6). Let $\widetilde{\Lambda}_j \equiv N_j(T)/T$ and

$$\widetilde{V}_{l,k}(\Delta) \equiv \frac{1}{T}\sum_i \sum_{i'} \delta(t_{li} - t_{ki'} - \Delta) - \widetilde{\Lambda}_k \widetilde{\Lambda}_l - \widetilde{\Lambda}_k \delta(\Delta) \mathbf{1}_{[k=l]}. \tag{4.15}$$

Here we can see $\widetilde{\Lambda}_j$ and $\widetilde{V}_{l,k}(\Delta)$ as empirical versions of Λ_j and $V_{l,k}(\Delta)$ in (4.2) and (4.3). Ignoring the penalty term in (4.12), the first order condition of (4.12) with respect to $\beta_{j,k}$ $(j \neq k)$ is

$$\begin{aligned} 0 = &\left\{\widetilde{\Lambda}_k \int_0^\infty \psi(\Delta) d\Delta\right\} \left\{\widetilde{\Lambda}_j - \mu_j - \sum_{l=1}^p \widetilde{\Lambda}_l \int_0^\infty \psi(\Delta) \cdot \hat{\beta}_{j,l} d\Delta\right\} \\ &+ \int_0^\infty \psi(\Delta) \left\{\widetilde{V}_{j,k}(\Delta) - \psi(\Delta) \cdot \hat{\beta}_{j,k} \widetilde{\Lambda}_k - \sum_{l=1}^p \left[\left(\psi \cdot \hat{\beta}_{j,l}\right) * \widetilde{V}_{l,k}\right](\Delta)\right\} d\Delta. \end{aligned} \tag{4.16}$$

A detailed derivation of (4.16) is provided in Appendix C.2.1. Recall that $\psi \cdot \hat{\beta}_{j,k}$ is an estimator for $\omega_{j,k}$. The first term on the right-hand side of (4.16) then corresponds to (4.5), and the second term corresponds to (4.6). The only difference is that $\mathbf{\Lambda}$ and $\boldsymbol{V}$ in (4.2) and (4.3) are replaced by $\widetilde{\mathbf{\Lambda}}$ and $\widetilde{\boldsymbol{V}}$ in (4.15). Equation (4.16) thus reveals the close connection between the intensity-based approach and the moment-based approach.

4.3.2 Theory

We now study the theoretical properties of the estimator (4.12) and the estimated edge set $\widehat{\mathcal{E}}$ in (4.13).

Let $\hat{\lambda}_j(t) \equiv \mu_j + \sum_{k=1}^p \Psi_k(t) \cdot \hat{\beta}_{j,k}$ be the estimated intensity function. The first result is the so-called oracle inequality on $\hat{\lambda}_j$. We note that a similar oracle inequality have been shown in Hansen et al. (2015), in the special case where all transfer functions are assumed to be non-negative, i.e., $\omega_{j,k}(t) \geq 0$ for all t.

In what follows, we use $\mathcal{E}_j \equiv \{k : (j,k) \in \mathcal{E}\}$ to denote nodes who are connected to the jth node in the true graph $\mathcal{G}$. Let s denote the maximum node degree of the true graph $\mathcal{G}$, i.e. $s = \max_j \text{card}(\mathcal{E}_j)$.

To establish the oracle inequality, we need the following assumptions.

Assumption 17. (Compatibility) There exist positive constants ξ_1 and c_1 such that, for any vectors $a_1, \ldots, a_p \in \mathbb{R}^M$ that satisfy the inequality

$$\sum_{k \notin \mathcal{E}_j} \left\{\mathbb{E} \int_0^T \left[\Psi_k(t) \cdot a_k\right]^2 dt\right\}^{1/2} \leq 2 \sum_{k \in \mathcal{E}_j} \left\{\mathbb{E} \int_0^T \left[\Psi_k(t) \cdot a_k\right]^2 dt\right\}^{1/2}, \tag{4.17}$$

we have

$$\xi_1^{1/2} \sum_{k \in \mathcal{E}_j} \left\{ \mathbb{E} \int_0^T \left[\Psi_k(t) \cdot a_k \right]^2 dt \right\}^{1/2} \leq s^{1/2} \left\{ \mathbb{E} \int_0^T \left[\sum_{k=1}^p \Psi_k(t) \cdot a_k \right]^2 dt \right\}^{1/2}. \tag{4.18}$$

Assumption 17 is essential for establishing oracle inequalities for lasso-type estimators (van de Geer and Bhlmann, 2009). Instead of Assumption 17, Hansen et al. (2015) assume that $\omega_{j,k} \geq 0$ for all j and k, which implies Assumption 17 (see Proposition 4 in Hansen et al. (2015)).

The next assumption guarantees the non-degeneracy of the basis functions.

Assumption 18. For $j = 1, \ldots, p$, the basis functions are non-degenerate:

$$0 < \gamma_{\min} \leq \Gamma_{\min} \left(\frac{1}{T} \mathbb{E} \int_0^T \Psi_k(t) \Psi_k^{\mathrm{T}}(t) \, dt \right) \leq \Gamma_{\max} \left(\frac{1}{T} \mathbb{E} \int_0^T \Psi_k(t) \Psi_k^{\mathrm{T}}(t) \, dt \right) \leq \gamma_{\max}. \tag{4.19}$$

Next, we require that the transfer functions are smooth, and that the residuals from the truncated basis approximation (4.11) decrease polynomially fast as M, the number of basis functions, grows. The second statement of Assumption 19 is known to hold for certain basis functions, such as trigonometric bases (see, e.g., Tsybakov, 2009).

Assumption 19. For $k \in \mathcal{E}_j$, $\omega_{j,k}$ belongs to a Sobolev class $W(\theta_1, L_1)$ on the interval $[0, b]$ for some integer $\theta_1 \geq 2$, i.e., $\omega_{j,k}^{(\theta_1 - 1)}$ is absolutely continuous and $\|\omega_{j,k}^{(\theta_1)}\|_{2,[0,b]} \leq L_1$, where $\omega_{j,k}^{(l)}$ denotes the lth derivative of $\omega_{j,k}$. Furthermore, for $k \in \mathcal{E}_j$, there exists $\tilde{\beta}_{j,k} \in \mathbb{R}^M$ such that

$$\frac{1}{T} \int_0^T \left\{ \Psi_k(t) \cdot \tilde{\beta}_{j,k} - \big(\omega_{j,k} * dN_k \big)(t) \right\}^2 dt \leq Q T^{-\frac{2\theta_1}{2\theta_1 + 1}}, \tag{4.20}$$

where $M = \lfloor q T^{1/(2\theta_1 + 1)} \rfloor$ for some global constants q and Q.

We arrive at the following theorem.

Theorem 5. *For a Hawkes process on $[0, T]$ whose intensities follow* (4.4)*, suppose that Assumptions 15–19 hold. Furthermore, assume that there exists a $\lambda_{\max}$ such that $\lambda_j(t) \leq \lambda_{\max}$ for $t \in [0, T]$. Given $M = \lfloor q T^{1/(2\theta_1 + 1)} \rfloor$ and $\eta_j = (8 \lambda_{\max} \xi_1 \log(p)/T)^{1/2}$, the oracle inequality*

$$\frac{1}{T} \int_0^T \left[\hat{\lambda}_j(t) - \lambda_j(t) \right]^2 dt \leq 2^4 s^2 Q T^{-\frac{2\theta_1}{2\theta_1 + 1}} + 2^6 s \lambda_{\max} \frac{\log(p)}{T} \tag{4.21}$$

holds for all $j = 1, \ldots, p$ with probability at least $1 - c_2 p^2 T \exp(-c_3 T^{1/5}) - c_4 p^{-1}$, where Q, c_2, c_3, and c_4 are positive constants.

In order to establish variable selection consistency of the estimated edge set $\widehat{\mathcal{E}}$ (4.13), we need two additional assumptions.

Assumption 20. (Irrepresentability) There exists a constant $0 < \xi_2 < 1$, such that for all $j = 1, \ldots, p$,

$$\max_{k \notin \mathcal{E}_j} \left\| \left(\mathbb{E} \int_0^T \Psi_k(t) \Psi_{\mathcal{E}_j}^{\mathrm{T}}(t)\, dt \right) \left(\mathbb{E} \int_0^T \Psi_{\mathcal{E}_j}(t) \Psi_{\mathcal{E}_j}^{\mathrm{T}}(t)\, dt \right)^{-1} \right\|_2 \leq \xi_2, \tag{4.22}$$

where $\Psi_{\mathcal{E}_j}(t) \in \mathbb{R}^{M \mathrm{card}(\mathcal{E}_j)}$ is composed by concatenating vectors in $\big\{ \Psi_k(t) : k \in \mathcal{E}_j \big\}$.

Assumption 21. (Beta-min) There exits $\beta_{\min} > 0$ such that for $(j,k) \in \mathcal{E}$,

$$\left\{ \frac{1}{T} \int_0^T \big[\big(\omega_{j,k} * dN_k \big)(t) \big]^2 \, dt \right\}^{1/2} \geq \beta_{\min}.$$

Moreover,

$$\sqrt{\frac{2}{\xi_1}} q^{1/2} T^{\frac{-1}{4\theta_1+2}} \left(\frac{T^{\frac{-\theta_1}{2\theta_1+1}} s Q^{1/2}}{4\sqrt{\lambda_{\max} \log(p)/T}} + 1 \right) + 1 \leq \frac{1}{2\xi_2} \sqrt{\frac{\gamma_{\min}}{s \gamma_{\max}}}.$$

Assumption 20 is almost necessary to establish variable selection consistency of lasso estimators (van de Geer and Bhlmann, 2009); however, it can be relaxed if we instead use an adaptive lasso or a thresholded lasso estimator (van de Geer et al., 2011). The first statement of Assumption 21 describes the signal strength that is needed in order for (4.12) to be able to identify the true edges, and the second statement is a technical requirement on the parameters in Assumptions 17 – 20.

We arrive at the following theorem:

Theorem 6. *For a Hawkes process on* $[0,T]$ *whose intensities follow* (4.4)*, suppose that Assumptions 15–21 hold. Assume also that there exists a* $\lambda_{\max}$ *such that* $\lambda_j(t) \leq \lambda_{\max}$ *for* $t \in [0,T]$*. Furthermore, assume that* $s = O(\log(p))$*. For* $\eta_j = (8\lambda_{\max}\xi_1 \log(p)/T)^{1/2}$ *and* $M = \lfloor q T^{1/(2\theta_1+1)} \rfloor$*, if* $p^2 T \exp(-c_3 T^{1/5}) = o(1)$*, we have that*

$$\textit{Prob}(\widehat{\mathcal{E}} = \mathcal{E}) \geq 1 - o(1),$$

where Q*,* c_2*,* c_3*, and* c_4 *are the same constants as in Theorem 5.*

Theorem 6 guarantees that the proposed estimator (4.12) recovers the true graph with probability converging to unity, even if the dimension p grows much faster than the observed time period T.

4.4 Reducing Computational Cost via Screening

The computation complexity of solving (4.12) for each $j = 1, \ldots, p$ is roughly $O(Tp^2)$ per iteration using the coordinate descent algorithm outlined in Appendix C.1. In order to reduce the computational cost when p is large, we propose two screening methods that can be quickly applied to reduce the number of potential edges to consider in the penalized regression (4.12). The screening methods are based on the cross-covariances (4.3) that can be estimated at low computational cost. Furthermore, we show that both screening methods satisfy the *sure screening property* introduced by Fan and Lv (2008); that is, all true edges are in the selected set with probability tending to unity. In particular, theoretical guarantees of the screening method in Section 4.4.1 requires only a subset of assumptions for Theorem 6. In contrast, the screening method in Section 4.4.2 requires an additional assumption but might lead to greater reduction in computational cost.

4.4.1 Screening the Connected Components

A connected component is a set of nodes that are connected by paths in an undirected graph. Recall that $\mathcal{G}$ is a directed graph. We drop the directions of edges and define the corresponding undirected (skeleton) graph $\mathcal{G}^u$ as $\mathcal{G}^u = (\mathcal{V}, \mathcal{E}^u)$, where

$$\mathcal{E}^u \equiv \{(j,k) : \exists \Delta \in \mathbb{R}^+ \text{ such that } \omega_{j,k}(\Delta) \neq 0 \text{ or } \omega_{k,j}(\Delta) \neq 0, 1 \leq j,k \leq p\}. \tag{4.23}$$

One can see that $\mathcal{E}^u$ is a superset of $\mathcal{E}$ in (4.8). Let $\left\{\mathcal{C}_l\right\}_{l=1}^{L}$ denote the set of connected components of the graph $\mathcal{G}^u$; we will refer to these as the true connected components in what follows. If the true connected components were known, we only need to estimate $\omega_{j,k}$ in (4.12) for nodes j and k that are in the same connected component, because otherwise we know that $\omega_{j,k} = 0$. A reduction in computational cost in solving (4.12) is guaranteed as long as $L > 1$.

Of course, the true connected components are not known in practice. To estimate the connected components, we use the fact that the cross-covariance between two independent point processes is the zero function. As a result, the ℓ_2-norm of the cross-covariance between two nodes is zero if they are not in the same connected component. In contrast, nodes in the same connected component may have non-zero cross-covariances. Based on these observations, we screen for the connected component using the following strategy.

1) First, for any $j, k \in \mathcal{V}$, we use kernel smoothing to estimate the cross-covariance $V_{j,k}(\Delta)$ in (4.3) for $\Delta \in [-B, B]$ as

$$\widehat{V}_{j,k}(\Delta) = \frac{1}{Th} \iint_{[0,T]^2} K\big(\frac{\Delta - \{t' - t\}}{h}\big)\, dN_j(t) dN_k(t') - \frac{1}{T} N_j(T) \frac{1}{T} N_k(T), \tag{4.24}$$

where $K(\cdot)$ is a kernel function with bandwidth h. Here B is a tuning parameter that defines the time range of interest. The bandwidth h can be chosen by cross-validation or simply set to be $T^{-1/5}$. When $j = k$, we set $\hat{V}_{k,k}(0) = 0$.

2) For some threshold level ζ, we define an adjacency matrix A as $A_{j,k} = 1$ if $\|\widehat{V}_{j,k}\|_{2,[-B,B]} > \zeta$ and $A_{j,k} = 0$ otherwise.

3) We compute the connected components $\big\{\widehat{\mathcal{C}}_l(\zeta)\big\}_{l=1}^{\hat{L}}$ of the adjacency matrix A.

The following theorem establishes that, under mild conditions, the aforementioned procedure will result in exact recovery of the true connected components.

Theorem 7. *Suppose Assumptions 15, 16, 17, 19, and 21 hold. Assume also that there exists* $\lambda_{\max} > 0$ *so that* $\lambda_j(t) \leq \lambda_{\max}$ *for all* $j = 1, \ldots, p$ *and* $t \in [0, T]$. *Furthermore, assume that* $s = O(\log(p))$. *If* $p^2 T^{6/5} \exp\big(-c_6 T^{1/5}\big) = o(1)$, *then* $\hat{L} = L$ *and* $\big\{\widehat{\mathcal{C}}_l(\zeta)\big\}_{l=1}^{\hat{L}} = \big\{\mathcal{C}_l\big\}_{l=1}^{L}$ *for* $\zeta = c_5 T^{-1/5}$, *with probability converging to unity as* $T \to \infty$.

Theorem 7 relies on a concentration inequality for $\widehat{V}_{j,k} - V_{j,k}$, which we defer to Section 4.5 (see Corollary 2). Once the deviation of $\widehat{V}_{j,k}$ from $V_{j,k}$ is bounded, Theorem 7 follows from the facts that (a) $V_{j,k}(\cdot) \equiv 0$ for j, k not in the same connected component, and (b) a node and its

neighbours belong to the same estimated connected component. The latter statement follows from the compatibility condition and the beta-min condition (Assumptions 17 and 21) together with the Wiener-Hopf integral equation (4.6). The detailed proof of Theorem 7 can be found in Appendix C.3.

Note that Assumptions 17 and 21 are required in Theorem 6 for successful recovery of the true graph by fitting the (conditional) intensity λ_j (4.12). Theorem 7 shows that these assumptions also imply that the connected components (though not the individual edges) can be recovered from simply estimating and thresholding the cross-covariances.

In light of Theorem 3, rather than solving (4.12), we instead solve (4.12) subject to the constraint

$$\beta_{j,k} = 0, \text{ for } (j,k) \notin \widehat{\mathcal{E}}^{cc}(\zeta), \tag{4.25}$$

where the selected edges are defined as

$$\widehat{\mathcal{E}}^{cc}(\zeta) \equiv \{(j,k) : j,k \in \widehat{\mathcal{C}}_l(\zeta), l = 1, \ldots, \hat{L}\}. \tag{4.26}$$

Combining Theorems 6 and 7, it follows that solving (4.12) subject to (4.25) will lead to recovery of the true edge set $\mathcal{E}$ (4.8) with probability converging to unity.

It is immediately clear that it is faster to solve the constrained optimization problem than to solve (4.12) without the constraint (4.25) as long as $\hat{L} > 1$, which implies $L > 1$ under Theorem 7. This is because the computational complexity of solving (4.12) subject to (4.25) is roughly $O\big(T \max_l \text{card}(\widehat{\mathcal{C}}_l)^2\big)$ per iteration, where $\max_l \widehat{\mathcal{C}}_l \leq p-1$ if $\hat{L} > 1$. The reduction in computation increases if the size of the largest estimated connected component grows slower than p grows. The cost of estimating $\widehat{\mathcal{E}}^{cc}(\zeta)$ is negligible compared to the iterative algorithm for solving (4.12): the computational complexity of estimating the cross-covariances is $O(Tp^2)$, and the computational complexity of estimating the connected components is $O(p^2)$ (Tarjan, 1972), of which both are one-time operations. As we will see in the numerical experiments in Section 4.6.3, the constrained optimization problem may also have lower sample complexity than the unconstrained optimization problem (4.12).

4.4.2 Screening the Edges

We now consider an alternative screening method in which we directly screen the edge set $\mathcal{E}$, rather than estimating the connected components. This alternative approach is preferable when $L = 1$; however, it requires additional assumptions. For any given ζ, we define the set of selected edges as

$$\widehat{\mathcal{E}}^{ss}(\zeta) \equiv \{(j,k) : \|\widehat{V}_{j,k}\|_{2,[-B,B]} > \zeta\}, \tag{4.27}$$

where $\widehat{V}_{j,k}$ is defined in (4.24).

To show that (4.27) has the sure screening property (i.e., $\mathcal{E} \subset \widehat{\mathcal{E}}^{ss}(\zeta)$), an additional assumption is needed.

Assumption 22. (Minimum marginal signal) There exists a $c_7 > 0$ so that, for some $\kappa \in [0, 1/5)$, $\min_{(j,k)\in\mathcal{E}} \|V_{j,k}\|_{2,[-B,B]} \geq c_7 T^{-\kappa}$.

Assumption 22 appears throughout the sure screening literature (see, among others, Fan and Lv, 2008; Fan et al., 2009; Fan and Song, 2010; Fan et al., 2011, 2014; Liu et al., 2014; Song et al., 2014; Luo et al., 2014). If a true edge corresponds to a very small cross-covariance, then Assumption 22 will fail; in this case, the edge may not be included in $\widehat{\mathcal{E}}^{ss}(\zeta)$. Assumption (22) holds, for instance, under Assumption 21 in the setting of Hawkes processes with non-negative transfer functions considered in, e.g., Hansen et al. (2015).

The following theorem characterizes the properties of $\widehat{\mathcal{E}}^{ss}(\zeta)$: it is a (relatively) small set, and it satisfies the sure screening property.

Theorem 8. *Suppose Assumptions 15, 16, and 19 hold. Further suppose that there exists a positive constant* $\lambda_{\max}$ *so that* $\lambda_{\max} \geq \lambda_j(t)$ *for* $j = 1, \ldots, p$ *and* $t \in [0, T]$. *Let* $\zeta = 2c_5 T^{-1/5}$. *If* $p^2 T^{6/5} \exp\left(-c_6 T^{1/5}\right) = o(1)$, *then card*$\left(\widehat{\mathcal{E}}^{ss}(\zeta)\right) = O\left(spT^{5/2}\gamma_\Omega^2(1-\gamma_\Omega)^{-2}(1+\lambda_{\max})^2\right)$ *with probability converging to unity as* $T \to \infty$. *Furthermore, if Assumption 22 holds, then* $\mathcal{E} \subset \widehat{\mathcal{E}}^{ss}(\zeta)$ *with probability converging to unity as* $T \to \infty$.

In light of Theorem 8, rather than solving (4.12), we can instead solve (4.12) subject to the constraint

$$\beta_{j,k} = 0, \text{ for } (j,k) \notin \widehat{\mathcal{E}}^{ss}(\zeta). \tag{4.28}$$

For any given threshold ζ, $\widehat{\mathcal{E}}^{ss}(\zeta)$ is always a subset of $\widehat{\mathcal{E}}^{cc}(\zeta)$. Therefore, the computational complexity of solving (4.12) subject to (4.28) is no larger than that of solving (4.12) subject to (4.25).

4.5 A Concentration Inequality for Hawkes Processes

In this section, we address the remaining technical challenges from Sections 4.3 and 4.4. To prove Theorems 5 and 6, we need to show that the sample version of the population assumptions Assumptions 17 – 21 holds with high probability. To prove Theorems 7 and 8, we need a concentration bound on how far $\widehat{V}_{j,k}$ deviates from $V_{j,k}$. All these quantities are related to second-order statistics of the Hawkes process, which take the form

$$\bar{y}_{j,k} \equiv \frac{1}{T}\int_0^T \int_0^T f(t-t')dN_j(t')dN_k(t). \tag{4.29}$$

For instance, $f(x) = K((x-\Delta)/h)/h$ in (4.24).

It is immediately clear that $\bar{y}_{j,k}$ is defined on sequences of temporal dependence, i.e., dN_j and dN_k. As a result, the deviation of $\bar{y}_{j,k}$ from its mean $\mathbb{E}\bar{y}_{j,k}$ cannot be bounded using classic concentration inequalities on independent random variables. To apply existing concentration inequalities for dependent sequences (see, among others, Merlevède et al., 2011), we need to quantify the temporal dependence of the Hawkes process, which is a non-trivial task.

Previous theoretical analysis of Hawkes processes assume that the transfer functions are non-negative, i.e. $\omega_{j,k} \geq 0$ (Reynaud-Bouret, 2003; Reynaud-Bouret and Schbath, 2010; Hansen et al., 2015). This assumption is infeasible in many real-world applications. In particular, it is known that inhibitory connections are abundant in neuronal circuits and are crucial for maintaining balance in the nervous system. Nevertheless, the non-negativity assumption is essential in the existing literature: the proof techniques rely heavily on the Poisson branching representation of the Hawkes process (Hawkes and Oakes, 1974), which is only available for non-negative transfer functions.

To relax the non-negativity assumption, we establish a concentration inequality for second-order statistics of the general Hawkes process using a novel proof technique. The following theorem summarizes our main results.

Theorem 9. *Suppose that Assumptions 15, 16, and 19 hold. Further suppose that there exists* $\lambda_{\max} > 0$ *so that* $\lambda_j(t) \leq \lambda_{\max}$ *for all* $j = 1, \ldots, p$ *and* $t \in [0, T]$. *Let* $f(\cdot)$ *be a bounded function on a bounded support, i.e., supp*$(f) = [b_1, b_2]$ *and* $\|f\|_\infty \equiv \max_x |f(x)| \leq C_f$. *Then, for each pair of* (j, k)

$$\mathbb{P}\left(|\bar{y}_{j,k} - \mathbb{E}\bar{y}_{j,k}| \geq c_5 T^{-2/5}\right) \leq c_7 T \exp(-c_6 T^{1/5}). \tag{4.30}$$

The detailed proof of Theorem 9 is given in Appendix C.4. We provide a sketch here. The core of the new proof is an iterative construction procedure based on the Poisson embedding theorem pioneered by Brémaud and Massoulié (1996), who study the stationary condition of the Hawkes process. The iterative construction procedure is then used to derive an upper bound of the τ-dependence coefficient using the result in (Dedecker and Prieur, 2004). Roughly speaking, the τ-dependence coefficient of a given time gap u (i.e., $\tau(u)$) measures the dependence between the history $\mathcal{H}_z$ and the point process $d\boldsymbol{N}(z + u)$ for any z (Merlevède et al., 2011). In particular, we show that, given the conditions in Theorem 9, $\tau(u)$ decreases exponentially fast as a function of u. This finding allows us to use the Bernstein inequality for dependence sequences in Merlevède et al. (2011), which gives the desired results.

As a direct result, we have the following corollary on $\|\widehat{V}_{j,k}\|$. The proof of Corollary 2 is also given in Appendix C.4.

Corollary 2. *Suppose that Assumptions 15, 16, and 19 hold. Further suppose that there exists* $\lambda_{\max} > 0$ *so that* $\lambda_j(t) \leq \lambda_{\max}$ *for all* $j = 1, \ldots, p$ *and* $t \in [0, T]$. *Then, for all* $1 \leq j, k \leq p$, *if* $p^2 T^{6/5} \exp\left(-c_6 T^{1/5}\right) = o(1)$,

$$\mathbb{P}\big(\big\|\widehat{V}_{j,k} - V_{j,k}\big\|_{2,[-B,B]} \leq c_5 T^{-1/5}\big) \geq 1 - o(1),$$

where c_5 *and* c_6 *are constants that depend only on* s, γ_Ω, ρ, *and* d_0.

4.6 Simulation Studies

This section is organized as follows. In Section 4.6.1, we study the choice of the penalty function in the penalized regression presented in Section 4.3. In Section 4.6.2, we examine the numerical

performances of the two screening methods proposed in Section 4.4. In Section 4.6.3, we combine screening and penalized regression in order to reconstruct the graph, and compare the performance to that of competing methods from the literature.

In the simulations, the intensities are of the form (4.4) with $\omega_{j,k}(t) = 5a_{j,k}t\exp(1-5t)$. We set the magnitudes $a_{j,k}$ to be 0.4 for a positive edge, -0.2 for a negative edge, and 0 in the absence of an edge. The signs of the edges are specified in the following sections, as are the baseline intensities $\mu_1, \ldots, \mu_p$. Given the form of the intensity function, we use the thinning method by Ogata (1988) to draw samples from the Hawkes process.

4.6.1 A Comparison of Penalties in Penalized Regression

We consider a directed graph of 21 nodes with the edge set $\mathcal{E} \equiv \{(1,k) : k = 2, \ldots, 11\}$. Specifically, we study two data-generating scenarios.

Scenario 1: The graph is shown in Figure 4.1(a). We set $\mu_1 = 0.75$, $\mu_j = 0.25$ for $j = 2, \ldots, 11$, and $\mu_j = 0.5$ for $j = 12, \ldots, 21$.

Scenario 2: The graph is shown in Figure 4.1(b). We set $\mu_1 = 2$ and $\mu_j = 0.5$ for $j = 2, \ldots, 21$.

We generate 200 simulated data sets on $[0, 900]$ under each scenario.

We compare the performances of three methods: (i) the proposal of Hansen et al. (2015); (ii) our proposal (4.12); and (iii) a modification of our proposal that replaces the standardized group lasso penalty in (4.12) with an unstandardized group lasso penalty of the form $\|\beta_{j,k}\|_2$. Thus, the three methods differ only in terms of the penalties on the parameters $\beta_{j,k}$. The unstandardized group lasso penalizes the ℓ_2-norm of the parameters $\beta_{j,k}$. The penalty of Hansen et al. (2015) penalizes the cumulative transfer functions $\Psi_{j,k}(t) \cdot \beta_{j,k}$ only when there is an event in dN_j, as mentioned in Remark 5.

Each method is applied using step function bases, $\psi_m = 1_{[(m-1)/M, m/M]}(x)$, as considered in Hansen et al. (2015), in order to recover the neighborhood of the first node. The edge recovery

performances over 200 simulated data sets are shown in Figures 4.1(c)–(d).

In Scenario 1, the value of the background intensity, μ_k, is much higher for the non-neighbours (i.e., $\{k : \omega_{1,k} = 0\}$) than for the true neighbours (i.e., $\{k : \omega_{1,k} \neq 0\}$) of the first node. The unstandardized group lasso penalty tends to assign non-zero edge estimates to nodes with higher intensities, which are the non-neighbours in this scenario. Thus, it performs poorly relative to the other two methods, as can be seen in Figure 4.1(c).

As seen in Remark 5, the penalty of Hansen et al. (2015) is not invariant to sign flips of the transfer function $\omega_{j,k}$: the penalty is higher for a positive transfer function than for its negative copy. Due to this lack of invariance, the proposal of Hansen et al. (2015) performs poorly in terms of edge recovery performance in Scenario 2, as we can see from Figure 4.1(d).

4.6.2 A Comparison of Screening Methods

We now examine the empirical performance of the two screening methods proposed in Section 4.4 in two scenarios.

> *Scenario 3:* There are 5 identical connected components, each of which is a circle of 20 nodes, as shown in Figure 4.2(a). We set $\mu_j = 0.75$ for all $j = 1, \ldots, 100$.
>
> *Scenario 4:* There are 20 identical connected components, each of which is a dense graph of 5 nodes, as shown in Figure 4.2(b). We set $\mu_{1+5(k-1)} = 1$ and $\mu_{j+5(k-1)} = 0.8$ for $j = 2, \ldots, 5, k = 1, \ldots, 20$.

In Scenario 3, all edges are positive, so that Assumption 22 holds on the Hakwes process associated with this graph. In contrast, in Scenario 4, there are both positive and negative edges, designed so that certain cross-covariances are diluted due to simultaneous presence of an inhibitory path and an excitatory path. Therefore, Assumption 22 is violated.

We generate 200 simulated data sets under each scenario on $[0, T]$, for a range of T from 300 to 2000. We use a kernel smoother with smoothing bandwidth $h = T^{-1/5}$, in order to obtain $\widehat{V}_{j,k}$, an estimate of the cross-covariance $V_{j,k}$.

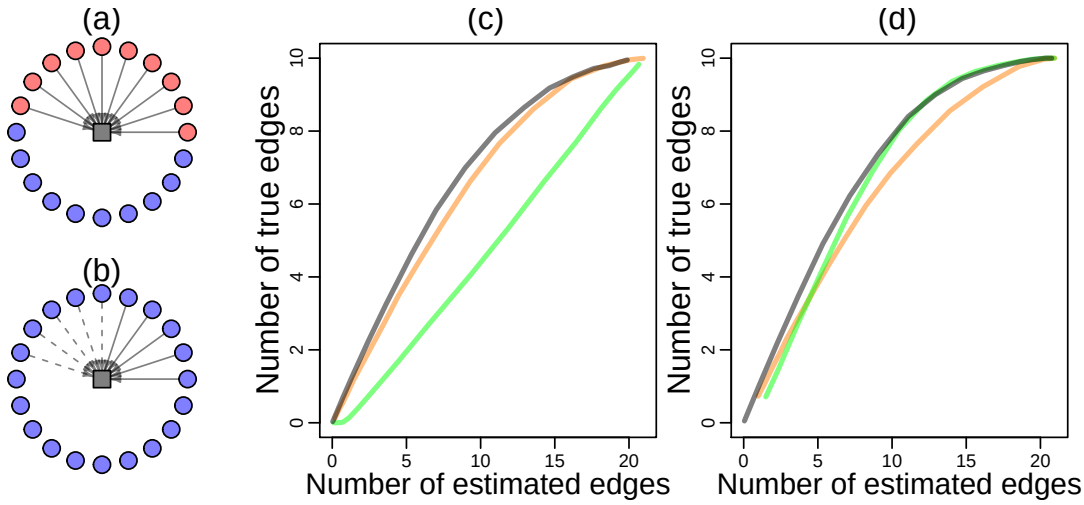

Figure 4.1: Edge recovery performance of (4.12) with different choices of penalties. *(a):* the directed graph corresponding to Scenario 1; *(b):* the directed graph corresponding to Scenario 2; *(c):* edge recovery performance under Scenario 1; and *(d):* edge recovery performance under Scenario 2. In Panels (a) and (b), ■ represents Node 1, ● represents nodes with baseline intensity $\mu_j = 0.5$, ● represents nodes with baseline intensity $\mu_j = 0.25$, → represents edges with $a_{j,k} = 0.4$, and ⇢ represents edges with $a_{j,k} = -0.2$. In Panels (c) and (d), each point represents the recovery of the neighborhood of Node 1, for a given value of η_1, averaged over 200 simulated data sets. The three curves represent the performance of our proposal (—), our proposal with unstandardized group lasso (—), and the proposal in Hansen et al. (2015) (—).

We consider the marginal covariances between pairs of nodes from the two sets $\mathcal{E}$, the true edge set in (4.8), and $\mathcal{I}$, where $\mathcal{I} \equiv \{(j,k) : j \in \mathcal{C}_l, k \in \mathcal{C}_m, m \neq l\}$ and $\mathcal{C}_l$, $l = 1, \ldots L$ are the true connected components introduced in Section 4.4.

Figure 4.2(c) indicates that in Scenario 3, for T sufficiently large, the empirical quantiles of $\|\widehat{V}_{j,k}\|_{2,[-5,5]}$ for node pairs in $\mathcal{I}$ do not overlap the empirical quantiles of $\|\widehat{V}_{j,k}\|_{2,[-5,5]}$ for node pairs in $\mathcal{E}$. This suggests that the screening procedure described in Section 4.4.2 will perform well in Scenario 3. However, Figure 4.2(d) indicates that in Scenario 4, there is overlap between the

empirical quantiles of $\|\widehat{V}_{j,k}\|_{2,[-5,5]}$ for node pairs in $\mathcal{I}$ and $\mathcal{E}$, regardless of the value of T. This suggests that the screening procedure described in Section 4.4.2 will perform poorly in Scenario 4.

Finally, we apply the screening procedures described in Sections 4.4.1 and 4.4.2 to data generated under Scenarios 3 and 4. Results are shown in Figure 4.3. As expected, the screened set $\widehat{\mathcal{E}}^{ss}(\zeta)$ (4.27) performs well in Scenario 3, while $\widehat{\mathcal{E}}^{cc}(\zeta)$ (4.26) performs well in Scenario 4.

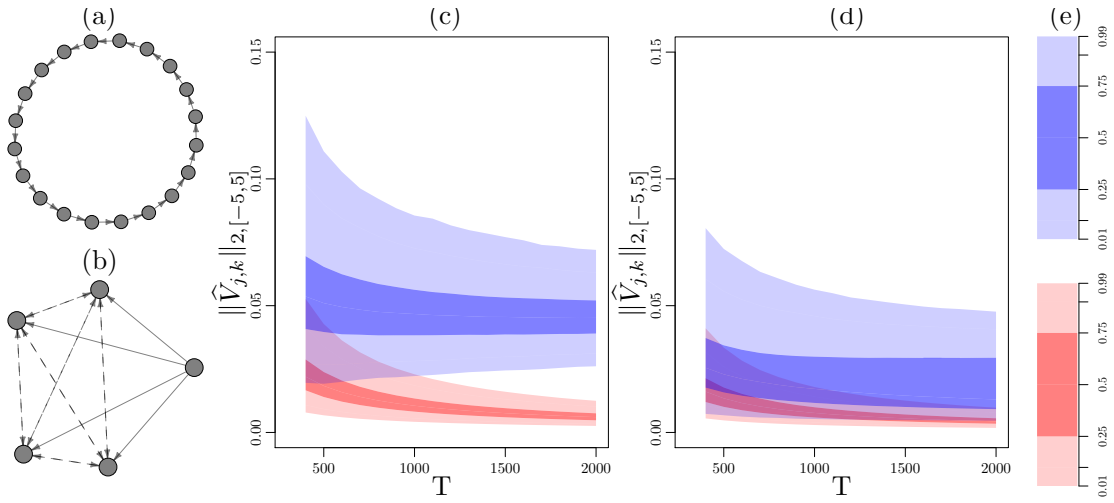

Figure 4.2: *(a):* A single connected component under Scenario 3; *(b):* a single connected component under Scenario 4; *(c):* empirical quantiles of $\|\widehat{V}_{j,k}\|_{2,[-5,5]}$ for node pairs in $\mathcal{I}$ (shown in red) and $\mathcal{E}$ (shown in blue) as a function of time T, for Scenario 3; (d) empirical quantiles of $\|\widehat{V}_{j,k}\|_{2,[-5,5]}$ for node pairs in $\mathcal{I}$ (shown in red) and $\mathcal{E}$ (shown in blue) as a function of time T, for Scenario 4; and *(e):* the heat maps used to display the empirical quantiles of $\|\widehat{V}_{j,k}\|_{2,[-5,5]}$ in $\mathcal{I}$ (red) and $\mathcal{E}$ (blue). In (a) and (b), we use $\longrightarrow$ to represent the positive edges ($\omega_{j,k} > 0$) and $\dashrightarrow$ to represent the negative edges ($\omega_{j,k} < 0$).

4.6.3 Graph Reconstruction

Finally, we consider reconstructing the graph. We generate data as follows:

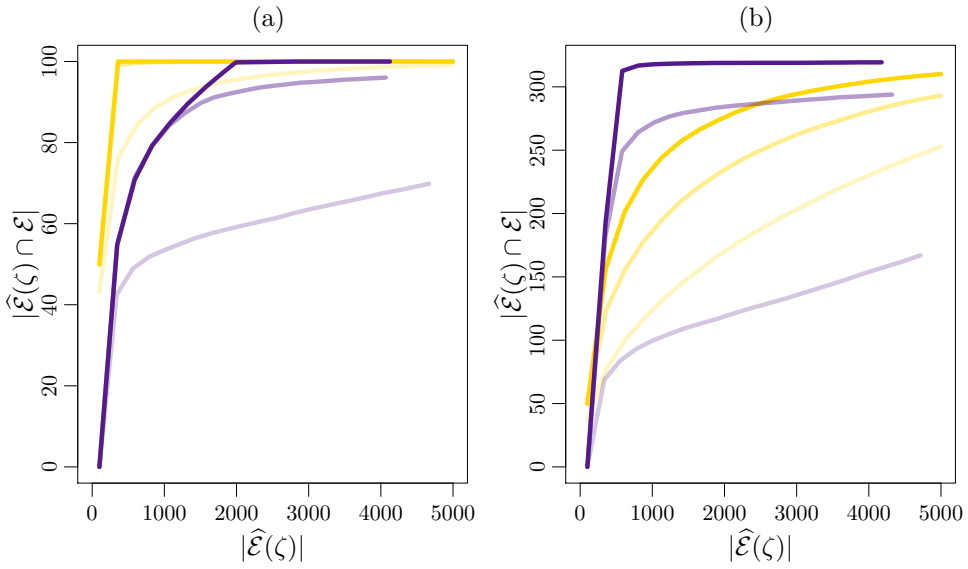

Figure 4.3: The performances of two screening methods $\widehat{\mathcal{E}}^{cc}(\zeta)$ and $\widehat{\mathcal{E}}^{ss}(\zeta)$, as a function of the tuning parameter ζ, for *(a):* Scenario 3, which has $\text{card}(\mathcal{E}) = 100$; and *(b):* Scenario 4, which has $\text{card}(\mathcal{E}) = 320$. In both panels, the x-axes are the number of edges selected by the screening methods ($|\widehat{\mathcal{E}}(\zeta)| \equiv \text{card}(\widehat{\mathcal{E}}(\zeta))$) and the y-axes are the number of true edges among the selected edges ($|\widehat{\mathcal{E}}(\zeta) \cap \mathcal{E}| \equiv \text{card}(\widehat{\mathcal{E}}(\zeta) \cap \mathcal{E})$), where $\widehat{\mathcal{E}}(\zeta)$ can be $\widehat{\mathcal{E}}^{cc}(\zeta)$ or $\widehat{\mathcal{E}}^{ss}(\zeta)$. The curves represent the performances of $\widehat{\mathcal{E}}^{cc}(\zeta)$ with $T = 500$ (—), 1000 (—), and 1500 (—); as well as the performances of $\widehat{\mathcal{E}}^{ss}(\zeta)$ with $T = 500$ (—), 1000 (—), and 1500 (—). Each point displayed represents the results for a single value of ζ, averaged over 200 simulated data sets.

> *Scenario 5:* 500 nodes are grouped into 50 connected components. Each connected component is as displayed in Figure 4.4(a). We set $\mu_j = 0.8$ for all $j = 1, \ldots, 500$.

This scenario is designed so that some of the edges within a connected component have small cross-covariance.

In order to reconstruct the graph, we take a two-step approach: (i) We perform screening of edges (4.27) or of connected components (4.26), with the threshold ζ selected so that the size of the screened edge set (not including self-loops) is either 2625, 5750, or 12000. (ii) We perform

neighborhood selection (4.12) on the subset of selected edges, for a range of values of the sparsity tuning parameter $\eta = \eta_1 = \ldots = \eta_p$.

Results averaged over 200 simulated data sets are shown in Figures 4.4(b)–(c). Screening the connected components (4.26) with an appropriate choice of threshold leads to improved statistical efficiency, as well as faster computation. In this particular simulation, the best performance is achieved by applying penalized regression on the screened set $\widehat{\mathcal{E}}^{cc}(\zeta)$, where ζ is chosen so that $\text{card}\big(\widehat{\mathcal{E}}^{cc}(\zeta)\big) = 5,750$. In this simulation setting, screening the edges (4.27) yields poor results, because some edges are hard to detect using merely the cross-covariance, and hence the screened edge set $\widehat{\mathcal{E}}^{ss}(\zeta)$ contains many false positives.

We omit the results of edge recovery without screening (i.e. $\zeta = 0$) due to its high computational cost. For the same reason, we omit the results from the method by Hansen et al. (2015). We expect Hansen et al. (2015) to perform poorly in this setting, due to the presence of negative edges (see discussion in Section 4.6.1, and Figure 4.1).

4.7 Discussion

There are some limitations of this study. First, we have not investigated the issue of tuning parameter selection for the sparsity tuning parameter η_j ($j = 1, \ldots, p$) or the threshold level ζ for the screening methods. Second, it might be possible to relax Assumption 15 and 16. For instance, when the true graph is a tree, the Hawkes process might still be stationary even if Assumption 15 is violated. Third, the concentration inequality in Theorem 9 might not be sharp. However, these questions are beyond the scope of this study, and we leave them for future research.

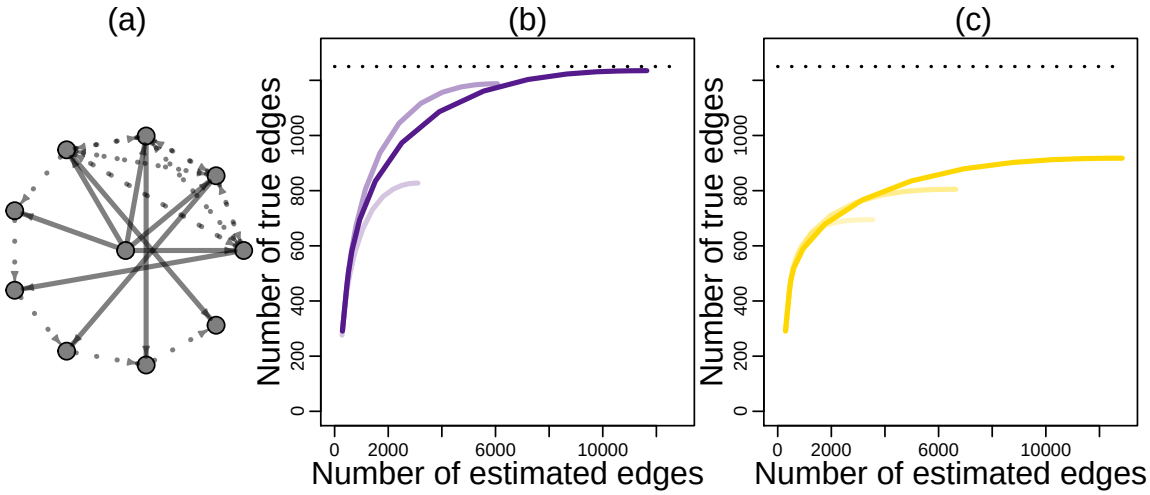

Figure 4.4: Edge recovery results for Scenario 5, with $T = 1500$. *(a)*: One connected component of the graph in Scenario 5. In total, the graph contains 50 connected components, and 1250 edges. *(b):* Results from screening using $\widehat{\mathcal{E}}^{cc}(\zeta)$ in (4.26) before performing neighborhood selection, with ζ chosen to yield 2625 (—), 5750 (—), and 12000 (—) non-self-loop edges. *(c):* Results from screening using $\widehat{\mathcal{E}}^{ss}(\zeta)$ in (4.27) before performing neighborhood selection, with ζ chosen to yield 2625 (—), 5750 (—), and 12000 (—) non-self-loop edges. In (a), we use $\longrightarrow$ to represent the positive edges ($\omega_{j,k} > 0$) and $\dashrightarrow$ to represent the negative edges ($\omega_{j,k} < 0$). Each point on the curves in Panels (b) and (c) represents the results, averaged over the simulated data sets, for a given value of the sparsity tuning parameter $\eta = \eta_1 = \ldots = \eta_p$ in the neighborhood selection problem (4.12). In Panels (b) and (c), the number of estimated edges on the x-axis cannot exceed the size of the screened edge set, which is either 2625, 5750, or 12000 plus the 500 self-loops. The dotted lines ($\cdots$) in Panels (b) and (c) indicate the total number of true edges (1250).

BIBLIOGRAPHY

Ahrens, M. B., M. B. Orger, D. N. Robson, J. M. Li, and P. J. Keller (2013). Whole-brain functional imaging at cellular resolution using light-sheet microscopy. *Nature methods 10*(5), 413–420.

Aït-Sahalia, Y., J. Cacho-Diaz, and R. J. Laeven (2015). Modeling financial contagion using mutually exciting jump processes. *Journal of Financial Economics 117*(3), 585 – 606.

Allen, G. I. and Z. Liu (2012). A log-linear graphical model for inferring genetic networks from high-throughput sequencing data. In *IEEE International Conference on Bioinformatics and Biomedicine*, pp. 1–6.

Anandkumar, A., V. Y. F. Tan, F. Huang, and A. S. Willsky (2012, August). High-dimensional gaussian graphical model selection: Walk summability and local separation criterion. *J. Mach. Learn. Res. 13*(1), 2293–2337.

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

Bacry, E. and J.-F. Muzy (2014). Second order statistics characterization of hawkes processes and non-parametric estimation. *arXiv preprint arXiv:1401.0903*.

Benson, M. (1979). Parameter fitting in dynamic models. *Ecological Modelling 6*(2), 97 – 115.

Berry, T., F. Hamilton, N. Peixoto, and T. Sauer (2012). Detecting connectivity changes in neuronal networks. *Journal of Neuroscience Methods 209*(2), 388 – 397.

Besag, J. E. (1974). Spatial interaction and the statistical analysis of lattice systems (with Discussion). *Journal of the Royal Statistical Society. Series B (Methodological) 36*(2), 192–236.

Biegler, L. T., J. J. Damiano, and G. E. Blau (1986). Nonlinear parameter estimation: A case study comparison. *AIChE Journal 32*(1), 29–45.

Blundell, C., J. Beck, and K. A. Heller (2012). Modelling reciprocating relationships with hawkes processes. In *Advances in Neural Information Processing Systems*, pp. 2600–2608.

Boucheron, S., G. Lugosi, and P. Massart (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press.

Bowsher, C. G. (2007). Modelling security market events in continuous time: Intensity based, multivariate point process models. *Journal of Econometrics 141*(2), 876 – 912.

Breheny, P. and J. Huang (2015). Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and Computing 25*, 173–187.

Brémaud, P. and L. Massoulié (1996, 07). Stability of nonlinear hawkes processes. *Ann. Probab. 24*(3), 1563–1588.

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

Brillinger, D. R., H. L. Bryant Jr, and J. P. Segundo (1976). Identification of synaptic interactions. *Biological cybernetics 22*(4), 213–228.

Brunel, N. J.-B. (2008). Parameter estimation of odes via nonparametric estimators. *Electronic Journal of Statistics 2*, 1242–1267.

Brunel, N. J.-B., Q. Clairon, and F. dAlch Buc (2014). Parametric estimation of ordinary differential equations with orthogonality conditions. *Journal of the American Statistical Association 109*(505), 173–185.

Bühlmann, P. and S. van de Geer (2011). *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg. Methods, theory and applications.

Buja, A., T. J. Hastie, and R. J. Tibshirani (1989). Linear smoothers and additive models. *Ann. Statist. 17*(2), 453–555.

Bunea, F. (2008). Honest variable selection in linear and logistic regression models via ℓ_1 and ℓ_1+ ℓ_2 penalization. *Electronic Journal of Statistics 2*, 1153–1194.

Candes, E. and T. Tao (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 2313–2351.

Cao, J., L. Wang, and J. Xu (2011). Robust estimation for ordinary differential equation models. *Biometrics 67*(4), 1305–1313.

Cao, J. and H. Zhao (2008). Estimating dynamic models for gene regulation networks. *Bioinformatics 24*(14), 1619–1624.

Chavez-Demoulin, V., A. C. Davison, and A. J. McNeil (2005). Estimating value-at-risk: a point process approach. *Quantitative Finance 5*(2), 227–234.

Chen, J. and H. Wu (2008). Efficient local estimation for time-varying coefficients in deterministic dynamic models with applications to hiv-1 dynamics. *Journal of the American Statistical Association 103*(481), 369–384.

Chen, S., A. Shojaie, and D. M. Witten (2016+). Network reconstruction from high-dimensional ordinary differential equations. *Journal of the American Statistical Association (to appear)*.

Chen, S., D. M. Witten, and A. Shojaie (2015). Selection and estimation for mixed graphical models. *Biometrika 102*(1), 47–64.

Cheng, J., E. Levina, and J. Zhu (2013). High-dimensional mixed graphical models. *arXiv preprint arXiv:1304.2810*.

Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics 23*(4), 493–507.

Chou, I.-C. and E. O. Voit (2009). Recent developments in parameter estimation and structure identification of biochemical and genomic systems. *Mathematical Biosciences 219*(2), 57 – 83.

Cressie, N. (1991). Statistics for spatial data. *Wiley series in probability and mathematical statistics*.

Cunningham, J. P. and M. Y. Byron (2014). Dimensionality reduction for large-scale neural recordings. *Nature neuroscience 17*(11), 1500–1509.

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

Dattner, I. and C. A. J. Klaassen (2015). Optimal rate of direct estimators in systems of ordinary differential equations linear in functions of the parameters. *Electron. J. Stat. 9*(2), 1939–1973.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–31.

Dedecker, J. and C. Prieur (2004). Coupling for τ-dependent sequences and applications. *Journal of Theoretical Probability 17*(4), 861–885.

Dempster, A. P. (1972). Covariance selection. *Biometrics*, 157–175.

Efron, B., T. Hastie, I. Johnstone, R. Tibshirani, et al. (2004). Least angle regression. *The Annals of statistics 32*(2), 407–499.

Eichler, M., R. Dahlhaus, and J. Dueck (2015). Graphical modeling for multivariate Hawkes process with nonparametric link. *preprint*.

Eldawlatly, S., R. Jin, and K. G. Oweiss (2009). Identifying functional connectivity in large-scale neural ensemble recordings: a multiscale data mining approach. *Neural computation 21*(2), 450–477.

Ellner, S. P., Y. Seifu, and R. H. Smith (2002). Fitting population dynamic models to time-series data by gradient matching. *Ecology 83*(8), 2256–2270.

Etesami, J., N. Kiyavash, K. Zhang, and K. Singhal (2016). Learning network of multivariate hawkes processes: A time series approach. *arXiv preprint arXiv:1603.04319*.

Fan, J., Y. Feng, and R. Song (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *J. Amer. Statist. Assoc. 106*(494), 544–557.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*(456), 1348–1360.

Fan, J. and R. Li (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of the American Statistical Association 99*(467), 710–723.

Fan, J. and J. Lv (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(5), 849–911.

Fan, J., Y. Ma, and W. Dai (2014). Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models. *Journal of the American Statistical Association 109*(507), 1270–1284.

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

Fan, J. and R. Song (2010, 12). Sure independence screening in generalized linear models with np-dimensionality. *Ann. Statist. 38*(6), 3567–3604.

Fellinghauer, B., P. Bühlmann, M. Ryffel, M. von Rhein, and J. D. Reinhardt (2013). Stable graphical model estimation with random forests for discrete, continuous, and mixed variables. *Computational Statistics & Data Analysis 64*, 132–142.

Finegold, M. and M. Drton (2011, 06). Robust graphical modeling of gene networks using classical and alternative t-distributions. *The Annals of Applied Statistics 5*(2A), 1057–1080.

Friedman, J. H., T. J. Hastie, and R. J. Tibshirani (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics 9*(3), 432–441.

Friedman, J. H., T. J. Hastie, and R. J. Tibshirani (2010, 2). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*(1), 1–22.

Gao, E., G. C. DeAngelis, and A. Burkhalter (2010). Parallel input channels to mouse primary visual cortex. *The Journal of neuroscience 30*(17), 5912–5926.

Gershgorin, S. (1931). über die abgrenzung der eigenwerte einer matrix. *Bulletin de l'Académie des Sciences de l'URSS. Classe des sciences mathématiques et na* (6), 749–754.

Gerstner, W., W. M. Kistler, R. Naud, and L. Paninski (2014). *Neuronal dynamics: From single neurons to networks and models of cognition*. Cambridge University Press.

Gibson, G. J. and E. Renshaw (1998). Estimating parameters in stochastic compartmental models using markov chain methods. *Mathematical Medicine and Biology 15*(1), 19–40.

Gugushvili, S. and C. A. Klaassen (2012, 08). $\sqrt{n}$-consistent parameter estimation for systems of ordinary differential equations: bypassing numerical integration via smoothing. *Bernoulli 18*(3), 1061–1098.

Hall, P. and Y. Ma (2014). Quick and easy one-step parameter estimation in differential equations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76*(4), 735–748.

Hansen, N. R., P. Reynaud-Bouret, and V. Rivoirard (2015, 02). Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli 21*(1), 83–143.

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

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

Henderson, J. and G. Michailidis (2014, 04). Network reconstruction using nonparametric additive ode models. *PLoS ONE 9*(4), e94003.

Höfling, H. and R. J. Tibshirani (2009, jun). Estimation of sparse binary pairwise Markov networks using pseudo-likelihoods. *The Journal of Machine Learning Research 10*, 883–906.

Ionides, E., C. Bretó, and A. King (2006). Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences 103*(49), 18438–18443.

Izhikevich, E. M. (2007). *Dynamical systems in neuroscience*. MIT press.

Jalali, A., P. K. Ravikumar, V. Vasuki, and S. Sanghavi (2011). On learning discrete graphical models using group-sparse regularization. In *International Conference on Artificial Intelligence and Statistics*, pp. 378–387.

Khatri, C. and C. Rao (1976). Characterizations of multivariate normality. i. through independence of some statistics. *Journal of Multivariate Analysis 6*(1), 81 – 94.

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).

Lauritzen, S. (1996). *Graphical models*, Volume 17. Oxford University Press, U.S.A.

Lee, J. D. and T. J. Hastie (2014). Learning the structure of mixed graphical models. *Journal of Computational and Graphical Statistics, in press*.

Lee, J. D., Y. Sun, and J. E. Taylor (2013). On model selection consistency of penalized M-estimators: a geometric theory. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26*, pp. 342–350. Curran Associates, Inc.

Lee, S.-I., V. Ganapathi, and D. Koller (2007). Efficient structure learning of markov networks using ℓ_1-regularization. In B. Schölkopf, J. Platt, and T. Hoffman (Eds.), *Advances in Neural Information Processing Systems 19*, pp. 817–824. MIT Press.

Lewis, E. and G. Mohler (2011). A nonparametric em algorithm for multiscale hawkes processes. *preprint*, 1–16.

Liang, H. and H. Wu (2008). Parameter estimation for differential equation models using a framework of measurement error in regression models. *Journal of the American Statistical Association 103*(484), 1570–1583.

Linderman, S. W. and R. P. Adams (2014). Discovering latent network structure in point process data. *arXiv preprint arXiv:1402.0914*.

Liu, H., J. D. Lafferty, and L. A. Wasserman (2009, December). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *The Journal of Machine Learning Research 10*, 2295–2328.

Liu, J., R. Li, and R. Wu (2014). Feature selection for varying coefficient models with ultrahigh-dimensional covariates. *Journal of the American Statistical Association 109*(505), 266–274.

Loader, C. (2013). *locfit: Local Regression, Likelihood and Density Estimation.* R package version 1.5-9.1.

Loh, P.-L. and M. J. Wainwright (2012, 06). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics 40*(3), 1637–1664.

Lu, T., H. Liang, H. Li, and H. Wu (2011). High-dimensional odes coupled with mixed-effects modeling techniques for dynamic gene regulatory network identification. *Journal of the American Statistical Association 106*(496), 1242–1258.

Luo, S., R. Song, and D. Witten (2014). Sure screening for gaussian graphical models. *arXiv preprint arXiv:1407.7819*.

Ma, W., A. Trusina, H. El-Samad, W. A. Lim, and C. Tang (2009). Defining network topologies that can achieve biochemical adaptation. *Cell 138*(4), 760–773.

Ma, Y. and R. Li (2010). Variable selection in measurement error models. *Bernoulli 16*(1), 274–300.

Marbach, D., R. J. Prill, T. Schaffter, C. Mattiussi, D. Floreano, and G. Stolovitzky (2010). Revealing strengths and weaknesses of methods for gene network inference. *Proceedings of the National Academy of Sciences 107*(14), 6286–6291.

Marbach, D., T. Schaffter, C. Mattiussi, and D. Floreano (2009). Generating realistic in silico gene networks for performance assessment of reverse engineering methods. *Journal of Computational Biology 16*(2), 229–239.

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.

Meier, L. (2014). *grplasso: Fitting user specified models with group lasso penalty*. R package version 0.4-4.

Meier, L., S. Van De Geer, and P. Bühlmann (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(1), 53–71.

Meier, L., S. van de Geer, and P. Bühlmann (2009). High-dimensional additive modeling. *Ann. Statist. 37*(6B), 3779–3821.

Meinshausen, N. and P. Bühlmann (2006, 06). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics 34*(3), 1436–1462.

Meinshausen, N. and P. Bühlmann (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72*(4), 417–473.

Merlevède, F., M. Peligrad, and E. Rio (2011). A bernstein type inequality and moderate deviations for weakly dependent sequences. *Probability Theory and Related Fields 151*(3-4), 435–474.

Miao, H., X. Xia, A. Perelson, and H. Wu (2011). On identifiability of nonlinear ODE models and applications in viral dynamics. *SIAM Rev. 53*(1), 3–39.

Miyamura, M. and Y. Kano (2006). Robust Gaussian graphical modeling. *Journal of Multivariate Analysis 97*(7), 1525–1550.

M'Kendrick, A. (1925). Applications of mathematics to medical problems. *Proceedings of the Edinburgh Mathematical Society 44*, 98–130.

Mohler, G. O., M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita (2011). Self-exciting point process modeling of crime. *Journal of the American Statistical Association 106*(493), 100–108.

Negahban, S. N., P. K. Ravikumar, M. J. Wainwright, B. Yu, et al. (2012, 11). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science 27*(4), 538–557.

Ogata, Y. (1981, Jan). On lewis' simulation method for point processes. *Information Theory, IEEE Transactions on 27*(1), 23–31.

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

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

Onsager, L. (1944, Feb). Crystal statistics. i. a two-dimensional model with an order-disorder transition. *Phys. Rev. 65*, 117–149.

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

Peng, J., P. Wang, N. Zhou, and J. Zhu (2009). Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association 104*(486), 735–746.

Perry, P. O. and P. J. Wolfe (2013). Point process modelling for directed interaction networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75*(5), 821–849.

Pillow, J. W., J. Shlens, L. Paninski, A. Sher, A. M. Litke, E. Chichilnisky, and E. P. Simoncelli (2008). Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature 454*(7207), 995–999.

Qi, X. and H. Zhao (2010, 02). Asymptotic efficiency and finite-sample properties of the generalized profiling estimation of parameters in ordinary differential equations. *The Annals of Statistics 38*(1), 435–481.

Quinn, C. J., T. P. Coleman, N. Kiyavash, and N. G. Hatsopoulos (2010). Estimating the directed information to infer causal relationships in ensemble neural spike train recordings. *Journal of Computational Neuroscience 30*(1), 17–44.

Ramsay, J. O., G. Hooker, D. Campbell, and J. Cao (2007). Parameter estimation for differential equations: a generalized smoothing approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69*(5), 741–796.

Ravikumar, P. K., J. Lafferty, H. Liu, and L. Wasserman (2009). Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(5), 1009–1030.

Ravikumar, P. K. and J. D. Lafferty (2004). Variational Chernoff bounds for graphical models. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI '04, pp. 462–469. Arlington, Virginia, United States: AUAI Press.

Ravikumar, P. K., M. J. Wainwright, and J. D. Lafferty (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics 38*(3), 1287–1319.

Ravikumar, P. K., M. J. Wainwright, G. Raskutti, and B. Yu (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics 5*, 935–980.

Reynaud-Bouret, P. (2003). Adaptive estimation of the intensity of inhomogeneous Poisson processes via concentration inequalities. *Probab. Theory Related Fields 126*(1), 103–153.

Reynaud-Bouret, P. and S. Schbath (2010, 10). Adaptive estimation for hawkes processes; application to genome analysis. *Ann. Statist. 38*(5), 2781–2822.

Ripley, B. D. (2005). *Spatial statistics*, Volume 575. John Wiley & Sons.

Rosenbaum, M. and A. B. Tsybakov (2010, 10). Sparse recovery under matrix uncertainty. *The Annals of Statistics 38*(5), 2620–2651.

Roth, V. and B. Fischer (2008). The group-lasso for generalized linear models: Uniqueness of solutions and efficient algorithms. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, New York, NY, USA, pp. 848–855. ACM.

Rothman, A. J., P. J. Bickel, E. Levina, and J. Zhu (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics 2*, 494–515.

Schaffter, T., D. Marbach, and D. Floreano (2011). Genenetweaver: in silico benchmark generation and performance profiling of network inference methods. *27*(16), 2263–2270.

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

Simon, N. and R. J. Tibshirani (2012). Standardization and the group lasso penalty. *Statist. Sinica 22*(3), 983–1001.

Song, D., H. Wang, C. Y. Tu, V. Z. Marmarelis, R. E. Hampson, S. A. Deadwyler, and T. W. Berger (2013). Identification of sparse neural functional connectivity using penalized likelihood estimation and basis functions. *Journal of Computational Neuroscience 35*(3), 335–357.

Song, R., W. Lu, S. Ma, and X. J. Jeng (2014). Censored rank independence screening for high-dimensional survival data. *Biometrika 101*(4), 799–814.

Sun, H. and H. Li (2012). Robust Gaussian graphical modeling via ℓ_1 penalization. *Biometrics 68*(4), 1197–1206.

Tarjan, R. (1972). Depth-first search and linear graph algorithms. *SIAM J. Comput. 1*(2), 146–160.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 267–288.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. New York, NY: Springer Science+ Business Media, LLC.

van de Geer, S. (1995, 10). Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes. *Ann. Statist. 23*(5), 1779–1801.

van de Geer, S., P. Bühlmann, S. Zhou, et al. (2011). The adaptive and the thresholded lasso for potentially misspecified models (and a lower bound for the lasso). *Electronic Journal of Statistics 5*, 688–749.

van de Geer, S. A. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics 36*(2), 614–645.

van de Geer, S. A. and P. Bhlmann (2009). On the conditions used to prove oracle results for the lasso. *Electron. J. Statist. 3*, 1360–1392.

Varah, J. (1982). A spline least squares method for numerical parameter estimation in differential equations. *SIAM Journal on Scientific and Statistical Computing 3*(1), 28–46.

Vogel, D. and R. Fried (2011). Elliptical graphical modelling. *Biometrika 98*(4), 935–951.

Volterra, V. (1928). Variations and fluctuations of the number of individuals in animal species living together. *J. Cons. Int. Explor. Mer 3*(1), 3–51.

Voorman, A. L., A. Shojaie, and D. M. Witten (2014). Graph estimation with joint additive models. *Biometrika 101*(1), 85–101.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory 55*(5), 2183–2202.

Wainwright, M. J. and M. I. Jordan (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning 1*(1-2), 1–305.

Wainwright, M. J., J. D. Lafferty, and P. K. Ravikumar (2007). High-dimensional graphical model selection using ℓ_1-regularized logistic regression. In B. Schölkopf, J. Platt, and T. Hoffman (Eds.), *Advances in Neural Information Processing Systems 19*, pp. 1465–1472. MIT Press.

Wang, H. and C. Leng (2007). Unified LASSO estimation by least squares approximation. *J. Amer. Statist. Assoc. 102*(479), 1039–1048.

Wang, Y. J. and E. H. Ip (2008). Conditionally specified continuous distributions. *Biometrika 95*(3), 735–746.

Weyl, H. (1912). Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung). *Math. Ann. 71*(4), 441–479.

Williams, D. (1991). *Probability with martingales*. Cambridge university press.

Wu, H. (2005). Statistical methods for hiv dynamic studies in aids clinical trials. *Statistical Methods in Medical Research 14*(2), 171–192.

Wu, H., T. Lu, H. Xue, and H. Liang (2014). Sparse additive ordinary differential equations for dynamic gene regulatory network modeling. *Journal of the American Statistical Association 109*(506), 700–716.

Xia, Y. and W. Li (2002). Asymptotic behavior of bandwidth selected by the cross-validation method for local polynomial fitting. *Journal of Multivariate Analysis 83*(2), 265 – 287.

Xu, H., M. Farajtabar, and H. Zha (2016). Learning granger causality for hawkes processes. *arXiv preprint arXiv:1602.04511*.

Xue, H., H. Miao, and H. Wu (2010, 08). Sieve estimation of constant and time-varying coefficients in nonlinear ordinary differential equation models by considering both numerical error and measurement error. *The Annals of Statistics 38*(4), 2351–2387.

Xue, L. and H. Zou (2012, 10). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *The Annals of Statistics 40*(5), 2541–2571.

Yang, E., G. I. Allen, Z. Liu, and P. K. Ravikumar (2012). Graphical models via generalized linear models. In P. Bartlett, F. Pereira, C. Burges, L. Bottou, and K. Weinberger (Eds.), *Advances in Neural Information Processing Systems 25*, pp. 1367–1375.

Yang, E., Y. Baker, P. K. Ravikumar, G. I. Allen, and Z. Liu (2014). Mixed graphical models via exponential families. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, pp. 1042–1050.

Yang, Y. and H. Zou (2015). A fast unified algorithm for solving group-lasso penalize learning problems. *Statistics and Computing 25*(6), 1129–1141.

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

Yuan, M. and Y. Lin (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika 94*(1), 19–35.

Zhang, X., J. Cao, and R. J. Carroll (2015). On the selection of ordinary differential equation models with application to predator-prey dynamical models. *Biometrics 71*(1), 131–138.

Zhao, P. and B. Yu (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research 7*, 2541–2563.

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

Zou, H., T. J. Hastie, and R. J. Tibshirani (2007). On the "degrees of freedom" of the lasso. *The Annals of Statistics 35*(5), 2173–2192.

Appendix A

APPENDIX FOR CHAPTER 2

A.1 A Proof for Proposition 1

Proof. First of all, it is easy to see that if $\theta_{st} = \theta_{ts}$, then any function g such that

$$g(x) \propto \exp\left\{\sum_{s=1}^{p} f_s(x_s) + \frac{1}{2}\sum_{s=1}^{p}\sum_{t\neq s}\theta_{ts}x_s x_t\right\} \tag{A.1}$$

is capable of generating the conditional densities in (2.3) of the main paper as long as the function g is integrable with respect to x_s for $s = 1, \ldots, p$. The function g can be decomposed as

$$g(x) \propto \exp\left\{f_s(x_s) + \frac{1}{2}\sum_{t:t\neq s}(\theta_{ts}+\theta_{st})x_s x_t\right\}\exp\left\{\sum_{t\neq s} f_t(x_t) + \frac{1}{2}\sum_{t:t\neq s,\ j:j\neq s,\ j\neq t}\theta_{tj}x_j x_t\right\},$$

so the integrability of the conditional density $p(x_s \mid x_{-s})$ guarantees the integrability of g with respect to x_s. Therefore, the conditional densities of the form in (2.3) in Chapter 2 are compatible if $\theta_{ts} = \theta_{st}$.

We now prove that any function h that is capable of generating the conditional density in (2.3) of Chapter 2 is in the form (A.1). The following proof is essentially the same as that in Besag (1974). Suppose h is a function that is capable of generating the conditional densities. Define $P(x) = \log\{h(x)/h(0)\}$, where 0 can be replaced by any interior point in the sample space.

By definition, $P(0) = \log\{h(0)/h(0)\} = 0$. Therefore, P can be written in the general form

$$P(x) = \sum_{s=1}^{p} x_s G_s(x_s) + \sum_{t\neq s}\frac{G_{ts}(x_t, x_s)}{2}x_t x_s + \sum_{t\neq s, t\neq j, j\neq s}\frac{G_{tsj}(x_t, x_s, x_j)}{6}x_t x_s x_j + \cdots,$$

where we write the function P as the sum of interactions of different orders. Note that the factor of $1/2$ is due to $G_{st}(x_s, x_t) = G_{ts}(x_t, x_s)$; similar factors apply for higher-order interactions. Recalling that we assume h is capable of generating the conditional density $p(x_s \mid x_{-s})$, from

Definition 2 in Chapter 2 we know that

$$P(x) - P(x_s^0) = \log\left\{\frac{h(x)/\int h(x)dx_s}{h(x_s^0)/\int h(x)dx_s}\right\} = \log\left\{\frac{p(x_s \mid x_{-s})}{p(0 \mid x_{-s})}\right\},$$

where $x_s^0 = (x_1, \ldots, x_{s-1}, 0, x_{s+1}, \ldots, x_p)^{\mathrm{T}}$ and $p(x_s \mid x_{-s})$ is the conditional density in (2.3) of the main paper. It follows that

$$\log\left\{\frac{p(x_s \mid x_{-s})}{p(0 \mid x_{-s})}\right\} = P(x) - P(x_s^0) = x_s\left(G_s(x_s) + \sum_{t:t\neq s} x_t G_{ts}(x_t, x_s) + \cdots\right). \tag{A.2}$$

Letting $x_t = 0$ for $t \neq s$ in (A.2) and using the form of the conditional densities in (2.3) in Chapter 2, we have

$$x_s G_s(x_s) = f_s(x_s) - f_s(0). \tag{A.3}$$

Here we set $f_s(0) = 0$ since $f_s(0)$ is a constant. For the second-order interaction G_{ts}, we let $x_j = 0$ for $j \neq t, j \neq s$ in (A.2):

$$x_s G_s(x_s) + x_s x_t G_{ts}(x_t, x_s) = \theta_{st} x_t x_s + f_s(x_s).$$

Similarly, applying the previous argument on $P(x) - P(x_t^0)$, we have

$$x_t G_t(x_t) + x_s x_t G_{st}(x_s, x_t) = \theta_{ts} x_t x_s + f_t(x_t).$$

Therefore, if $\theta_{st} = \theta_{ts}$, then by (A.3),

$$G_{st}(x_s, x_t) = G_{ts}(x_t, x_s) = \theta_{st}.$$

It is easy to show that, by setting $x_k = 0$ $(k \neq s, k \neq t, k \neq j)$ in (A.2), the third-order interactions in $P(x)$ are zero. Similarly, we can show that fourth-and-higher-order interactions are zero. Hence, we arrive at the following formula for P:

$$P(x) = \sum_{s=1}^{p} f_s(x_s) + \frac{1}{2}\sum_{s=1}^{p}\sum_{t\neq s} \theta_{ts} x_s x_t.$$

Furthermore, $P(x) = \log\{h(x)/h(0)\}$, so the function h takes the form

$$h(x) \propto \exp\{P(x)\} = \exp\left\{\sum_{s=1}^{p} f_s(x_s) + \frac{1}{2}\sum_{s=1}^{p}\sum_{t\neq s} \theta_{ts} x_s x_t\right\},$$

which is the same as (A.1). □

A.2 A Proof for Lemma 1

Proof. We first prove the claim about compatibility.

It is easy to verify that the conditional densities are integrable given the restrictions with daggers in Table 2.1 in Chapter 2. Therefore, these restrictions are sufficient for compatibility.

We now show that the restrictions with daggers in Table 2.1 in Chapter 2 are necessary, by investigating each of the distributions in Equations 2.4 to 2.7 of Chapter 2. Note that we have limited our discussion to the case where all conditional densities are non-degenerate. Recall that we refer to the type of distribution of x_s given the others as the node type of x_s.

Suppose that x_s is exponential, as in (2.7) of Chapter 2. By definition of the exponential distribution, it must be that $\eta_s = \alpha_{1s} + \sum_{t \neq s} \theta_{ts} x_t < 0$. This leads to the following restrictions on θ_{ts}: 1) When x_t is Poisson or exponential, it must be that $\theta_{ts} \leq 0$ since x_t is unbounded in $\mathcal{R}^+$. 2) When x_t is Gaussian, then it must be that $\theta_{ts} = 0$ since x_t is unbounded on the real line. 3) Let I denote the indices of the Bernoulli variables. Then it must be that $\sum_{t \in I} |\theta_{ts}| < -\alpha_{1s}$ so that $\eta_s < 0$ for any combination of $\{x_t\}_{t \in I}$.

Suppose that x_s is Gaussian, as in (2.4) in Chapter 2. Then α_{2s} has to be negative for the conditional density to be well-defined.

Suppose that x_s is Bernoulli or Poisson, as in Equations 2.5 or 2.6 in Chapter 2. We can see that there are no restrictions on η_s, and thus no restrictions on θ_{ts} or α_{1s}.

Hence, the conditions with daggers in Table 2.1 in Chapter 2 are necessary for the conditional densities in Equations 2.4 to 2.7 in Chapter 2 to be compatible.

We now show the statement about strong compatibility.

We first prove the necessity of the conditions in Table 2.1 in Chapter 2. Recall from Definition 2 in Chapter 2 that in order for strong compatibility to hold, compatibility must hold, and any function g that satisfies (2.8) in Chapter 2 must be integrable. Therefore, we derive the necessary conditions for g to be integrable.

For Gaussian nodes that are indexed by J, recall that Θ_{JJ} is defined as in (2.9) of Chapter 2. Then, from properties of the multivariate Gaussian distribution, Θ_{JJ} must be negative definite if

the joint density exists and is non-degenerate.

Let x_1 be a Poisson node, and x_2 an exponential node. Consider the ratio

$$G(x_1, x_2) = \frac{g(x_1, x_2, 0, ..., 0)}{g(0, 0, 0, ..., 0)} = \exp\{-\log(x_1!) + \alpha_{11}x_1 + \theta_{12}x_1x_2 + \alpha_{12}x_2\},$$

where g is the function in (2.8) of Chapter 2. It is not hard to see that integrability of $G(x_1, x_2)$ is a necessary condition for integrability of the joint density. Summing over x_1 yields

$$\sum_{i=0}^{\infty} G(i, x_2) = \exp\{\alpha_{12}x_2 + \exp(\alpha_{11} + \theta_{12}x_2)\}.$$

Therefore, if $\sum_{i=0}^{\infty} G(i, x_2)$ is integrable with respect to the exponential node x_2, it must be the case that $\theta_{12} = \theta_{21} \leq 0$. Following a similar argument, the edge potential $\theta_{12} = \theta_{21}$ has to be non-positive when x_2 is Poisson, and zero when x_2 is Gaussian.

A similar argument to the one just described can be applied to the exponential nodes. Such an argument reveals that conditions on the edge potentials of the exponential nodes that are necessary for g to be a density are those stated in Table 2.1 of Chapter 2.

For Bernoulli nodes, no restrictions on the edge potentials are necessary in order for g to be a density.

Therefore, the conditions listed in Table 2.1 in Chapter 2 are necessary for the conditional densities in Equations 2.4 to 2.7 in Chapter 2 to be strongly compatible.

We now show that the conditions listed in Table 2.1 in Chapter 2 are sufficient for the conditional densities to be strongly compatible. We can restrict the discussion by conditioning on the Bernoulli nodes, since integrating over Bernoulli variables yields a mixture of finite components. Table 2.1 in Chapter 2 guarantees that the Gaussian nodes are isolated from the Poisson and exponential nodes, as the corresponding edge potentials are zero. From Table 2.1 in Chapter 2, the distribution of Gaussian nodes is integrable, as Θ_{JJ} in (2.9) of Chapter 2 is negative definite. Now we consider the Poisson and exponential nodes. For these,

$$\exp\left\{\sum_{s=1}^{p} f_s(x_s) + \frac{1}{2}\sum_{s=1}^{p}\sum_{t\neq s}\theta_{st}x_sx_t\right\} \leq \exp\left\{\sum_{s=1}^{p} f_s(x_s)\right\}$$

since $\theta_{st}x_sx_t \leq 0$. So the joint density is dominated by the density of a model with no interactions, which is integrable since α_{1t} for an exponential node x_t is non-positive; this follows from the fact that $0 \leq \sum_{s\in I} |\theta_{st}| < -\alpha_{1t}$, as stated in Table 2.1 of Chapter 2. Therefore, the conditions listed in Table 2.1 in Chapter 2 are also sufficient for the conditional densities in Equations 2.4 to 2.7 in Chapter 2 to be strongly compatible.

□

A.3 A Proof for Theorem 1

Proof. Our proof is similar to that of Theorem 1 in Yang et al. (2012), and is based on the primal-dual witness method Wainwright (2009). The primal-dual witness method studies the property of ℓ_1-penalized estimators by investigating the sub-gradient condition of an oracle estimator. We assume that readers are familiar with the primal-dual witness method; for reference, see Ravikumar et al. (2011) and Yang et al. (2012). Without loss of generality, we assume $s = p$ to avoid cumbersome notation. For other values of s, a similar proof holds with more complicated notation. Below we denote Θ_p as θ, η_p as η, and ℓ_p as ℓ for simplicity. We also denote the neighbours of x_p, $N(x_p)$, as N.

The sub-gradient condition for (2.10) in Chapter 2 with respect to $(\theta^{\mathrm{T}}, \alpha_{1p})^{\mathrm{T}}$ is

$$-\nabla\ell(\theta, \alpha_{1p}; X) + \lambda_n Z = 0; \quad Z_t = \text{sgn}(\theta_t) \quad \text{for } t < p; \quad Z_p = 0, \tag{A.4}$$

where

$$\text{sgn}(x) = \begin{cases} x/|x|, & x \neq 0, \\ \gamma \in [-1, 1], & x = 0. \end{cases}$$

We construct the oracle estimator $(\hat{\theta}_N^{\mathrm{T}}, \hat{\theta}_\Delta^{\mathrm{T}}, \hat{\alpha}_{1p})^{\mathrm{T}}$ as follows: first, let $\hat{\theta}_\Delta = 0$ where Δ indicates the set of non-neighbours; second, obtain $\hat{\theta}_N, \hat{\alpha}_{1p}$ by solving (2.10) in Chapter 2 with an additional restriction that $\hat{\theta}_\Delta = 0$; third, set $\hat{Z}_t = \text{sgn}(\hat{\theta}_t)$ for $t \in N$ and $\hat{Z}_p = 0$; last, estimate $\hat{Z}_\Delta$ from (A.4) by plugging in $\hat{\theta}, \hat{\alpha}_{1p}$ and $\hat{Z}_{\Delta^c}$. To complete the proof, we verify that $(\hat{\theta}_N^{\mathrm{T}}, \hat{\theta}_\Delta^{\mathrm{T}}, \hat{\alpha}_{1p})^{\mathrm{T}}$ and $\hat{Z} = (\hat{Z}_N^{\mathrm{T}}, \hat{Z}_\Delta^{\mathrm{T}}, 0)^{\mathrm{T}}$ is a primal-dual pair of (2.10) in Chapter 2 and recovers the true neighbourhood exactly.

Applying the mean value theorem on each element of $\nabla \ell(\hat{\theta}, \hat{\alpha}_{1p}; X)$ in the subgradient condition (A.4) gives

$$Q^* \begin{pmatrix} \hat{\theta} - \theta^* \\ \hat{\alpha}_{1p} - \alpha^*_{1p} \end{pmatrix} = -\lambda_n \hat{Z} + W^n + R^n, \tag{A.5}$$

where $W^n = \nabla \ell(\theta^*, \alpha^*_{1p}; X)$ is the sample score function evaluated at the true parameter $(\theta^{*\mathrm{T}}, \alpha^*_{1p})^{\mathrm{T}}$. Recall that $Q^* = -\nabla^2 \ell(\theta^*, \alpha^*_{1p}; X)$ is the negative Hessian of $\ell(\theta, \alpha_{1p}; X)$ with respect to $(\theta^{\mathrm{T}}, \alpha_{1p})^{\mathrm{T}}$, evaluated at the true values of the parameters. In (A.5), R^n is the residual term from the mean value theorem, whose kth term is

$$R^n_k = [\nabla^2 \ell(\bar{\theta}^k, \bar{\alpha}^k_{1p}; X) - \nabla^2 \ell(\theta^*, \alpha^*_{1p}; X)]^{\mathrm{T}}_k \begin{pmatrix} \hat{\theta} - \theta^* \\ \hat{\alpha}_{1p} - \alpha^*_{1p} \end{pmatrix}, \tag{A.6}$$

where $\bar{\theta}^k$ denotes an intermediate point between θ^* and $\hat{\theta}$, $\bar{\alpha}^k_{1p}$ denotes an intermediate point between α^*_{1p} and $\hat{\alpha}_{1p}$, and $[\cdot]^{\mathrm{T}}_k$ denotes the kth row of a matrix.

By construction, $\hat{\theta}_\Delta = 0$. Thus, (A.5) can be rearranged as

$$\lambda_n \hat{Z}_\Delta = (W^n_\Delta + R^n_\Delta) - Q^*_{\Delta\Delta^c}(Q^*_{\Delta^c\Delta^c})^{-1}(W^n_{\Delta^c} + R^n_{\Delta^c} - \lambda_n \hat{Z}_{\Delta^c}). \tag{A.7}$$

We obtain an estimator $\hat{Z}_\Delta$ by plugging $\hat{\theta}, \hat{\alpha}_{1p}$ and $\hat{Z}_{\Delta^c}$ into (A.7). To complete the proof, we need to verify strict dual feasibility,

$$\|\hat{Z}_\Delta\|_\infty < 1, \tag{A.8}$$

and sign consistency,

$$\mathrm{sgn}(\hat{\theta}_t) = \mathrm{sgn}(\theta^*_t) \quad \text{for any } t \in N. \tag{A.9}$$

In (A.7), $\max_{l \in \Delta} \|Q^*_{l\Delta^c}(Q^*_{\Delta^c\Delta^c})^{-1}\|_1 \leq 1 - a$ by Assumption 1 in Chapter 2. The following lemmas characterize useful concentration inequalities regarding W^n, R^n, and $\hat{\theta}_N - \theta^*_N$. Proofs of Lemmas 3 and 4 are given in Sections A.4 and A.5, respectively.

Lemma 3. *Suppose that*

$$\frac{8(2-a)}{a}\{\delta_2 \kappa_2 \log(2p)/n\}^{1/2} \leq \lambda_n \leq \frac{2(2-a)}{a}\delta_2 \kappa_2 M,$$

where δ_2 is defined in Proposition 3, and a and κ_2 are defined in Assumptions 1 and 3 of Chapter 2, respectively. Then,

$$pr\left(\|W^n\|_\infty > \frac{a\lambda_n}{8-4a}\middle|\,\xi_2,\xi_1\right) \le \exp(-c_3\delta_3 n),$$

where $\delta_3 = 1/(\kappa_2\delta_2)$ and c_3 is some positive constant.

Lemma 4. *Suppose that ξ_1 and $\|W^n\|_\infty \le a\lambda_n/(8-4a)$ hold and*

$$\lambda_n \le \min\left\{\frac{a\Lambda_1^2(d+1)^{-1}}{288(2-a)\kappa_2\Lambda_2}, \frac{\Lambda_1^2(d+1)^{-1}}{12\Lambda_2\kappa_3\delta_1\log p}\right\},$$

where δ_1 is defined in Proposition 2, and a and κ_3 are defined in Assumptions 1 and 3 of Chapter 2, respectively. Then with probability 1,

$$\|\hat{\theta}_N - \theta_N^*\|_2 < \frac{10}{\Lambda_1}(d+1)^{1/2}\lambda_n, \quad \|R^n\|_\infty \le \frac{a\lambda_n}{8-4a}.$$

We now continue with the proof of Theorem 1 in Chapter 2. Given Assumption 6, the conditions regarding λ_n are met for Lemmas 3 and 4.

We now assume that ξ_1, ξ_2 and the event $\|W^n\|_\infty \le a\lambda_n/(8-4a)$ are true so that the conditions for the two lemmas are satisfied. We derive the lower bound for the probability of these events at the end of the proof.

First, applying Lemma 4 and Assumption 1 to (A.7) yields

$$\begin{aligned}\|\hat{Z}_\Delta\|_\infty \le & \max_{l\in\Delta}\|Q^*_{l\Delta^c}(Q^*_{\Delta^c\Delta^c})^{-1}\|_1\left(\|W^n_{\Delta^c}\|_\infty + \|R^n_{\Delta^c}\|_\infty + \lambda_n\|\hat{Z}_{\Delta^c}\|_\infty\right)/\lambda_n + \\ & \left(\|W^n_\Delta\|_\infty + \|R^n_\Delta\|_\infty\right)/\lambda_n \\ \le & (1-a) + (2-a)\left\{\frac{a}{4(2-a)} + \frac{a}{4(2-a)}\right\} < 1.\end{aligned} \tag{A.10}$$

Next, applying Lemma 4 and a norm inequality to $\|\hat{\theta}_N - \theta_N^*\|_\infty$ gives

$$\|\hat{\theta}_N - \theta_N^*\|_\infty \le \|\hat{\theta}_N - \theta_N^*\|_2 < \frac{10}{\Lambda_1}(d+1)^{1/2}\lambda_n \le \min_t|\theta_t|, \tag{A.11}$$

since $\min_t|\theta_t| \ge 10(d+1)^{1/2}\lambda_n/\Lambda_1$ by Assumption 5. The strict inequality in (A.11) ensures that the sign of the estimator is consistent with the sign of the true value for all edges.

Equations A.10 and A.11 are sufficient to establish the result, i.e., $\hat{N} = N$. Let A be the event $\|W^n\|_\infty \leq a\lambda_n/(8-4a)$. Recall that we have assumed events A, ξ_1, and ξ_2 to be true in order to prove (A.10) and (A.11). We now derive the lower bound for the probability of $A \cap \xi_1 \cap \xi_2$.

Using the fact that

$$\text{pr}\{(A \cap \xi_1 \cap \xi_2)^c\} \leq \text{pr}(A^c \mid \xi_1 \cap \xi_2) + \text{pr}\{(\xi_1 \cap \xi_2)^c\} \leq \text{pr}(A^c \mid \xi_1, \xi_2) + \text{pr}(\xi_1^c) + \text{pr}(\xi_2^c),$$

we know the probability of $A \cap \xi_1 \cap \xi_2$ satisfies

$$\text{pr}\left\{\left(\|W^n\|_\infty \leq \frac{a}{2-a}\frac{\lambda_n}{4}\right) \cap \xi_2 \cap \xi_1\right\} \geq 1 - c_1 p^{-\delta_1+2} - \exp(-c_2\delta_2^2 n) - \exp(-c_3\delta_3 n),$$

where c_1, c_2, and c_3 are constants from Proposition 2, Proposition 3, and Lemma 3. Thus, the event $A \cap \xi_1 \cap \xi_2$ happens with high probability when the sample size n is large. This completes the proof. □

A.4 A Proof for Lemma 3

Proof. Recall that $\eta^{(i)} = \alpha_{1p} + \sum_{t<p} \theta_t x_t^{(i)}$ and that we have assumed that α_{kp} is known for $k \geq 2$. We can rewrite the conditional density in (2.3) of Chapter 2 as

$$p(x_p \mid x_{-p}) \propto \exp\{\eta x_p - D(\eta)\}.$$

For any $t < p$,

$$W_t^n = \frac{\partial \ell}{\partial \theta_t} = \sum_{i=1}^n \frac{\partial \ell}{\partial \eta^{(i)}}\frac{\partial \eta^{(i)}}{\partial \theta_t} = \frac{1}{n}\sum_{i=1}^n \{x_p^{(i)} - D'(\eta^{(i)})\}x_t^{(i)}. \tag{A.12}$$

Recall that M is a large constant introduced in Assumption 3 in Chapter 2. Suppose that M is sufficiently large that $|\alpha_{1p}^*| + \sum_{k<p}|\theta_k^*| < M/2$. For every v such that $0 < v < M/2$,

$$\begin{aligned}
E\left(\exp\left[vx_t^{(i)}\left\{x_p^{(i)} - D'(\eta^{(i)})\right\}\right]\Big| X_{-p}\right) &= E\left\{\exp\left(vx_t^{(i)}x_p^{(i)}\right)\Big| X_{-p}\right\}\exp\left\{-vx_t^{(i)}D'(\eta^{(i)})\right\} \\
&= \exp\left\{D(\eta^{(i)} + vx_t^{(i)}) - D(\eta^{(i)})\right\}\exp\left\{-vx_t^{(i)}D'(\eta^{(i)})\right\} \\
&= \exp\left\{vx_t^{(i)}D'(\eta^{(i)}) + (vx_t^{(i)})^2\frac{D''(\tilde{\eta})}{2}\right\}\exp\left\{-vx_t^{(i)}D'(\eta^{(i)})\right\} \\
&= \exp\left\{(vx_t^{(i)})^2\frac{D''(\tilde{\eta})}{2}\right\}, \quad \tilde{\eta} \in [\eta^{(i)}, \eta^{(i)} + vx_t^{(i)}],
\end{aligned} \tag{A.13}$$

where the second equality was derived using the properties of the moment generating function of the exponential family, and the third equality follows from a second-order Taylor expansion. Since $\tilde{\eta} \in [\eta^{(i)}, \eta^{(i)} + vx_t^{(i)}]$, the event ξ_1 implies that

$$|\tilde{\eta}| \leq |\alpha_{1p}^*| + \sum_{k<p} |x_k^{(i)}\theta_k^*| + |vx_t^{(i)}| \leq |\alpha_{1p}^*| + (\sum_{k<p} |\theta_k^*| + |v|)\max_{t,i}|x_t^{(i)}| \leq M\delta_1 \log p. \tag{A.14}$$

Therefore, the condition of Assumption 3 is satisfied, and thus $|D^{''}(\tilde{\eta})| \leq \kappa_2$. Recalling that $\{x^{(i)}\}_{i=1}^n$ are independent samples, it follows from (A.12) and (A.13) that

$$\begin{aligned} E\left\{\exp(vnW_t^n) \mid \xi_2, \xi_1\right\} =& E\left[E\left\{\exp(vnW_t^n) \mid X_{-p}, \xi_2, \xi_1\right\} \mid \xi_2, \xi_1\right] \\ \leq& E\left[\exp\left\{v^2\frac{\kappa_2}{2}\sum_{i=1}^n (x_t^{(i)})^2\right\}\middle| \xi_2, \xi_1\right] \\ \leq& \exp(nv^2\kappa_2\delta_2/2), \end{aligned} \tag{A.15}$$

where we use the event ξ_2 in the last inequality. Similarly,

$$E\left\{\exp(-nvW_t^n) \mid \xi_2, \xi_1\right\} \leq \exp(nv^2\kappa_2\delta_2/2). \tag{A.16}$$

Furthermore, one can see from a similar argument as in (A.12) and (A.13) that

$$\begin{aligned} E\left\{\exp(nvW_p^n) \mid \xi_1\right\} =& E\left\{\exp\left(vn\frac{\partial \ell}{\partial \alpha_{1p}}\right)\middle| \xi_1\right\} \\ =& \prod_{i=1}^n E\left(\exp[v\{x_p^{(i)} - D^{'}(\eta^{(i)})\}]\middle| \xi_1\right) \leq \exp(n\kappa_2 v^2/2). \end{aligned}$$

We focus on the discussion of (A.15) and (A.16) since $\delta_2 \geq 1$. For some δ to be specified, we let $v = \delta/(\kappa_2\delta_2)$ and apply the Chernoff bound Chernoff (1952); Ravikumar and Lafferty (2004) with (A.15) and (A.16) to get

$$\text{pr}(|W_t^n| > \delta \mid \xi_2, \xi_1) \leq \frac{E\{\exp(vnW_t^n) \mid \xi_2, \xi_1\}}{\exp(vn\delta)} + \frac{E\{\exp(-vnW_t^n) \mid \xi_2, \xi_1\}}{\exp(vn\delta)} \leq 2\exp\left(-n\frac{\delta^2}{2\kappa_2\delta_2}\right).$$

Letting $\delta = a\lambda_n/(8-4a)$ and using the Bonferroni inequality, we get

$$\begin{aligned} \text{pr}\left(\|W^n\|_\infty > \frac{a}{2-a}\frac{\lambda_n}{4}\middle| \xi_2, \xi_1\right) \leq& 2\exp\left\{-n\frac{a^2\lambda_n^2}{32(2-a)^2\kappa_2\delta_2} + \log(p)\right\} \\ \leq& \exp\left\{-\frac{a^2\lambda_n^2}{64(2-a)^2\kappa_2\delta_2}n\right\} = \exp(-c_3\delta_3 n), \end{aligned} \tag{A.17}$$

where $\delta_3 = 1/(\kappa_2\delta_2)$ and $c_3 = a^2\lambda_n^2/\{64(2-a)^2\}$. In (A.17), we made use of the assumption that $\lambda_n \geq 8(2-a)\{\kappa_2\delta_2 \log(2p)/n\}^{1/2}/a$, and we also require that $\lambda_n \leq 2(2-a)\kappa_2\delta_2 M/a$ since $v = a\lambda_n/\{(8-4a)\kappa_2\delta_2\} \leq M/2$. □

A.5 A Proof for Lemma 4

Proof. We first prove that $\|\hat{\theta}_N - \theta_N^*\|_2 < 10(d+1)^{1/2}\lambda_n/\Lambda_1$.

Following the method in Fan and Li (2004) and Ravikumar et al. (2010), we construct a function $F(u)$ as

$$F(u) = -\ell(\theta^* + u_{-p}, \alpha_{1p}^* + u_p; X) + \ell(\theta^*, \alpha_{1p}^*; X) + \lambda_n\|\theta^* + u_{-p}\|_1 - \lambda_n\|\theta^*\|_1, \qquad \text{(A.18)}$$

where u is a p-dimensional vector and $u_\Delta = 0$. $F(u)$ has some nice properties: (i) $F(0) = 0$ by definition; (ii) $F(u)$ is convex in u given the form of (2.3) in Chapter 2; and (iii) by the construction of the oracle estimator $\hat{\theta}$, $F(u)$ is minimized by $\hat{u}$ with $\hat{u}_{-p} = \hat{\theta} - \theta^*$ and $\hat{u}_p = \hat{\alpha}_{1p} - \alpha_{1p}^*$.

We claim that if there exists a constant B such that $F(u) > 0$ for any u such that $\|u\|_2 = B$ and $u_\Delta = 0$, then $\|\hat{u}\|_2 \leq B$. To show this, suppose that $\|\hat{u}\|_2 > B$ for such a constant. Let $t = B/\|\hat{u}\|_2$. Then, $t < 1$, and the convexity of $F(u)$ gives

$$F(t\hat{u}) \leq (1-t)F(0) + tF(\hat{u}) \leq 0.$$

Thus, $\|t\hat{u}\|_2 = B$ and $(t\hat{u})_\Delta = t\hat{u}_\Delta = 0$, but $F(t\hat{u}) \leq 0$, which is a contradiction.

Applying a Taylor expansion to the first term of $F(u)$ gives

$$\begin{aligned} F(u) =& -\nabla\ell(\theta^*, \alpha_{1p}^*; X)^{\mathrm{T}}u - u^{\mathrm{T}}\nabla^2\ell(\theta^* + vu_{-p}, \alpha_{1p}^* + vu_p; X)u/2 + \lambda_n(\|\theta^* + u_{-p}\|_1 - \|\theta^*\|_1) \\ =& \mathrm{I} + \mathrm{II}/2 + \mathrm{III}, \end{aligned}$$

for some $v \in [0,1]$. Recall that $u_\Delta = 0$ as defined in (A.18). The gradient and Hessian are with respect to the vector $(\theta^T, \alpha_{1p})^{\mathrm{T}}$.

We now proceed to find a B such that for $\|u\|_2 = B$ and $u_\Delta = 0$, the function $F(u)$ is always greater than 0. First, given that $\|W^n\|_\infty \leq a\lambda_n/(8-4a)$ and $a < 1$ assumed in Assumption 1,

$$|\mathrm{I}| = |(W^n)^{\mathrm{T}}u| \leq \|W^n\|_\infty\|u\|_1 \leq \frac{a}{2-a}\frac{\lambda_n}{4}(d+1)^{1/2}B \leq \frac{\lambda_n}{4}(d+1)^{1/2}B.$$

Next, by the triangle inequality and the Cauchy-Schwarz inequality,

$$\text{III} \geq -\lambda_n \|u_{-p}\|_1 \geq -\lambda_n d^{1/2} \|u_{-p}\|_2 \geq -\lambda_n (d+1)^{1/2} B.$$

To bound II, we note that

$$-\nabla^2 \ell(\theta^* + v u_{-p}, \alpha_{1p}^* + v u_p; X) = \frac{1}{n} \sum_{i=1}^n x_0^{(i)} (x_0^{(i)})^{\mathrm{T}} D^{''}(\eta_r^{(i)}),$$

where $x_0 = (x_{-p}^{\mathrm{T}}, 1)^{\mathrm{T}}$ as in Assumption 2, and $\eta_r^{(i)} = \alpha_{1p}^* + v u_p + \sum_{t<p} (\theta_t^* + v u_t) x_t^{(i)}$. Applying a Taylor expansion on each $D^{''}(\eta_r^{(i)})$ at $\eta^{(i)} = \alpha_{1p}^* + \sum_{t<p} \theta_t^* x_t^{(i)}$, we get

$$\begin{aligned}
-\nabla^2 \ell(\theta^* + v u_{-p}, \alpha_{1p}^* + v u_p; X) =& \frac{1}{n} \sum_{i=1}^n x_0^{(i)} (x_0^{(i)})^{\mathrm{T}} D^{''}(\eta^{(i)}) + \frac{1}{n} \sum_{i=1}^n x_0^{(i)} (x_0^{(i)})^{\mathrm{T}} D^{'''}(\tilde{\eta}^{(i)}) \left(v u^{\mathrm{T}} x_0^{(i)} \right) \\
=& Q^* + \frac{1}{n} \sum_{i=1}^n x_0^{(i)} (x_0^{(i)})^{\mathrm{T}} D^{'''}(\tilde{\eta}^{(i)}) \left(v u^{\mathrm{T}} x_0^{(i)} \right),
\end{aligned}$$

where $\tilde{\eta}^{(i)} \in [\eta^{(i)}, \eta_r^{(i)}]$. Using the argument on $\tilde{\eta}$ in (A.14) and the fact that $v \leq 1$ and $\|u\|_2 = B$, we can see that $\tilde{\eta}^{(i)}$ is in the range required for Assumption 3 to hold given ξ_1. Therefore, applying Assumption 3 we can write

$$\begin{aligned}
\text{II} \geq & \min_{u: \|u\|_2 = B, u_\Delta = 0} \{ -u^T \nabla^2 \ell(\theta^* + v u_{-p}, \alpha_{1p}^* + v u_p; X) u \} \\
\geq & B^2 \Lambda_{\min}(Q^*_{\Delta^c \Delta^c}) - \max_{v \in [0,1]} \max_{u: \|u\|_2 = B, u_\Delta = 0} u^{\mathrm{T}} \left\{ \frac{1}{n} \sum_{i=1}^n D^{'''}(\tilde{\eta}^{(i)}) (v u^{\mathrm{T}} x_0^{(i)}) x_0^{(i)} (x_0^{(i)})^{\mathrm{T}} \right\} u \\
\geq & \Lambda_1 B^2 - \max_{u: \|u\|_2 = B, u_\Delta = 0} \left\{ \max_{i, v \in [0,1]} (v u^{\mathrm{T}} x_0^{(i)}) \max_{\tilde{\eta}^{(i)}} D^{'''}(\tilde{\eta}^{(i)}) \frac{1}{n} \sum_{i=1}^n (u^{\mathrm{T}} x_0^{(i)})^2 \right\} \\
\geq & \Lambda_1 B^2 - \kappa_3 \max_{i, u: \|u\|_2 = B, u_\Delta = 0, v \in [0,1]} (v u^{\mathrm{T}} x_0^{(i)}) \max_{u: \|u\|_2 = B, u_\Delta = 0} \left\{ \frac{1}{n} \sum_{i=1}^n (u^{\mathrm{T}} x_0^{(i)})^2 \right\}.
\end{aligned}$$

By inspection, the maximum of $u^{\mathrm{T}} x_0^{(i)}$ is non-negative. Thus, the maximum of $v u^T x_0^{(i)}$ is achieved at $v = 1$. Then, using ξ_1 and Assumption 2,

$$\begin{aligned}
\text{II} \geq & \Lambda_1 B^2 - \kappa_3 B (d+1)^{1/2} \delta_1 \log(p) B^2 \Lambda_{\max} \left\{ \frac{1}{n} \sum_{i=1}^n x_0^{(i)} (x_0^{(i)})^{\mathrm{T}} \right\} \\
\geq & \Lambda_1 B^2 - \kappa_3 B^3 (d+1)^{1/2} \delta_1 \log(p) \Lambda_2.
\end{aligned}$$

Thus, if our choice of B satisfies

$$\Lambda_1 - \delta_1 \log(p) B \kappa_3 (d+1)^{1/2} \Lambda_2 \geq \frac{\Lambda_1}{2}, \tag{A.19}$$

then the lower bound of $F(u)$ is

$$F(u) \geq -\frac{\lambda_n}{4}(d+1)^{1/2}B + \frac{\Lambda_1}{4}B^2 - \lambda_n (d+1)^{1/2} B.$$

So, $F(u) > 0$ for any $B > 5(d+1)^{1/2}\lambda_n/\Lambda_1$. We can hence let

$$B = 6(d+1)^{1/2}\lambda_n/\Lambda_1 \tag{A.20}$$

to get

$$\|\hat{\theta}_N - \theta_N^*\|_2 \leq \|\hat{u}\|_2 \leq B = \frac{6}{\Lambda_1}(d+1)^{1/2}\lambda_n. \tag{A.21}$$

And thus, $\|\hat{\theta}_N - \theta_N^*\|_2 < 10\lambda_n(d+1)^{1/2}/\Lambda_1$. It is easy to show that (A.20) satisfies (A.19) provided that

$$\lambda_n \leq \frac{\Lambda_1^2 (d+1)^{-1}}{12 \Lambda_2 \kappa_3 \delta_1 \log p}.$$

To find the bound for R^n defined in (A.6), we first recall that $(\bar{\theta}^{\mathrm{T}}, \bar{\alpha})^{\mathrm{T}}$ is an intermediate point between $(\theta^{*\mathrm{T}}, \alpha_{1p}^*)^{\mathrm{T}}$ and $(\hat{\theta}^{\mathrm{T}}, \hat{\alpha}_{1p})^{\mathrm{T}}$. We denote $\bar{\eta}^{(i)} = \bar{\alpha}_{1p} + \sum_{t<p} \bar{\theta}_t x_t^{(i)}$, and observe that $|\bar{\eta}^{(i)}| \leq M\delta_1 \log p$ for $i = 1, \ldots, n$ using the argument of (A.14), which implies that Assumption 3 is applicable. Thus,

$$\begin{aligned}
\Lambda_{\max}\{\nabla^2 \ell(\bar{\theta}, \bar{\alpha}_{1p}; X) - \nabla^2 \ell(\theta^*, \alpha_{1p}^*; X)\} &= \max_{\|u\|_2 = 1} u^{\mathrm{T}} \{\nabla^2 \ell(\bar{\theta}, \bar{\alpha}_{1p}; X) - \nabla^2 \ell(\theta^*, \alpha_{1p}^*; X)\} u \\
&= \max_{\|u\|_2 = 1} u^{\mathrm{T}} \left[\frac{1}{n} \sum_{i=1}^n \left\{ D''(\bar{\eta}^{(i)}) - D''(\eta^*) \right\} x_0^{(i)} (x_0^{(i)})^{\mathrm{T}} \right] u.
\end{aligned}$$

By Assumption 3, $|D''(\bar{\eta}^{(i)}) - D''(\eta^*)| \leq 2\kappa_2$, and so

$$\begin{aligned}
\Lambda_{\max}\{\nabla^2 \ell(\bar{\theta}, \bar{\alpha}_{1p}; X) - \nabla^2 \ell(\theta^*, \alpha_{1p}^*; X)\} &= \max_{\|u\|_2 = 1} u^{\mathrm{T}} \left[\frac{1}{n} \sum_{i=1}^n \left\{ D''(\bar{\eta}^{(i)}) - D''(\eta^*) \right\} x_0^{(i)} (x_0^{(i)})^{\mathrm{T}} \right] u \\
&\leq 2\kappa_2 \max_{\|u\|_2 = 1} u^{\mathrm{T}} \left\{ \frac{1}{n} \sum_{i=1}^n x_0^{(i)} (x_0^{(i)})^{\mathrm{T}} \right\} u \leq 2\kappa_2 \Lambda_2,
\end{aligned}$$

using Assumption 2 at the last inequality. Hence, we arrive at

$$\begin{aligned}
\|R^n\|_\infty \leq & \|R^n\|_2^2 = \left\| \{\nabla^2 \ell(\bar{\theta}, \bar{\alpha}_{1p}; X) - \nabla^2 \ell(\theta^*, \alpha_{1p}^*; X)\}^{\mathrm{T}} \begin{pmatrix} \hat{\theta} - \theta^* \\ \hat{\alpha}_{1p} - \alpha_{1p}^* \end{pmatrix} \right\|_2^2 \\
\leq & \Lambda_{\max}\{\nabla^2 \ell(\bar{\theta}, \bar{\alpha}_{1p}; X) - \nabla^2 \ell(\theta^*, \alpha_{1p}^*; X)\} \left\| \begin{pmatrix} \hat{\theta} - \theta^* \\ \hat{\alpha}_{1p} - \alpha_{1p}^* \end{pmatrix} \right\|_2^2 \\
= & \Lambda_{\max}\{\nabla^2 \ell(\bar{\theta}, \bar{\alpha}_{1p}; X) - \nabla^2 \ell(\theta^*, \alpha_{1p}^*; X)\} \|\hat{u}\|_2^2 \\
\leq & \frac{72\kappa_2 \Lambda_2}{\Lambda_1^2}(d+1)\lambda_n^2,
\end{aligned}$$

where the last inequality follows from (A.21). So $\|R^n\|_\infty \leq a\lambda_n/(8-4a)$ if

$$\lambda_n \leq \frac{a}{2-a} \frac{\Lambda_1^2}{288(d+1)\kappa_2 \Lambda_2}, \tag{A.22}$$

which holds by assumption. □

A.6 A Proof for Corollary 1

Proof. The proof is essentially the same as the proof in Section A.3. We first show that a modified version of Lemma 4 holds with fewer conditions.

Lemma 5. *Suppose that $p(x_p|x_{-p})$ follows a Gaussian distribution as in* (2.4) *of Chapter 2, and $\|W^n\|_\infty \leq a\lambda_n/(8-4a)$. Then*

$$\|\hat{\theta}_N - \theta_N^*\|_2 < \frac{10}{\Lambda_1}(d+1)^{1/2}\lambda_n, \quad \|R^n\|_\infty = 0.$$

Proof. To prove this lemma, we go through the argument in Section A.5. But for II we note that

$$\begin{aligned}
\mathrm{II} \geq & \min_{u:\|u\|_2 = B, u_\Delta = 0} \{-u^T \nabla^2 \ell(\theta^* + v u_{-p}, \alpha_{1p}^* + v u_p; X) u\} \\
\geq & B^2 \Lambda_{\min}(-Q_{\Delta^c\Delta^c}^*) - \max_{v\in[0,1]} \max_{u:\|u\|_2 = B, u_\Delta = 0} u^{\mathrm{T}} \left\{ \frac{1}{n} \sum_{i=1}^n D'''(\tilde{\eta}^{(i)})(v u^{\mathrm{T}} x_0^{(i)}) x_0^{(i)} (x_0^{(i)})^{\mathrm{T}} \right\} u \\
\geq & \Lambda_1 B^2 - 0,
\end{aligned}$$

since $D'''(\tilde{\eta}^{(i)}) = 0$ for a Gaussian distribution. Therefore,

$$F(u) \geq -\frac{\lambda_n}{4}(d+1)^{1/2}B + \frac{1}{2}\Lambda_1 B^2 - \lambda_n(d+1)^{1/2}B.$$

So, $F(u) > 0$ for $B > 5\lambda_n(d+1)^{1/2}/(2\Lambda_1)$. We can hence let $B = 5(d+1)^{1/2}\lambda_n/\Lambda_1$ to get

$$\|\hat{\theta}_N - \theta_N^*\|_2 \leq \|\hat{u}\|_2 \leq B = \frac{5}{\Lambda_1}(d+1)^{1/2}\lambda_n.$$

Thus, $\|\hat{\theta}_N - \theta_N^*\|_2 < 10\lambda_n(d+1)^{1/2}/\Lambda_1$. And $\|R^n\|_\infty = 0$ trivially as $D''(\bar{\eta}^{(i)}) - D''(\eta^*) = 0$ for a Gaussian distribution. □

With Lemma 5, we can then verify (A.10) and (A.11) as in Section A.3. Finally, we drop the requirement of ξ_1 in the condition of Lemma 5, so the probability of $\hat{N} = N$ is

$$\text{pr}\left\{\left(\|W^n\|_\infty \leq \frac{a}{2-a}\frac{\lambda_n}{4}\right) \cap \xi_2\right\} \geq 1 - \exp(-c_2\delta_2^2 n) - \exp(-c_3\delta_3 n),$$

where c_2 and c_3 are constants from Proposition 3 in Chapter 2 and Lemma 3. □

A.7 Additional Details of Data-Generation Procedure

Here we provide additional details of the data-generation procedure described in Section 2.6.1 of Chapter 2. In particular, we describe the approach used to guarantee that the conditions listed in Table 2.1 of Chapter 2 for strong compatibility of the conditional distributions are satisfied.

Recall from Table 2.1 of Chapter 2 that in order for strong compatibility to hold, the matrix Θ_{JJ} in (2.9) in Chapter 2 that contains the edge potentials between the Gaussian nodes must be negative definite. If Θ_{JJ} generated as described in Section 2.6.1 of Chapter 2 is not negative definite, then we define a matrix T_{JJ} as

$$T_{JJ} = -\Theta_{JJ} + \{\Lambda_{\min}(\Theta_{JJ}) - 0.1\}\, I,$$

where $\Lambda_{\min}(\Theta_{JJ})$ denotes the minimum eigenvalue of Θ_{JJ}. Thus, T_{JJ} is guaranteed to be negative definite, as all its eigenvalues are no larger than -0.1. We then standardize T_{JJ} so that its diagonal elements equal -1,

$$\tilde{T}_{JJ} = \text{diag}(|T_{11}|^{-1/2}, \ldots, |T_{mm}|^{-1/2})\; T_{JJ}\; \text{diag}(|T_{11}|^{-1/2}, \ldots, |T_{mm}|^{-1/2}).$$

Finally, we replace Θ_{JJ} with $\tilde{T}_{JJ}$.

Table 2.1 of Chapter 2 also indicates that for strong compatibility to hold, the edge potential between two Poisson nodes must be negative. Therefore, after generating edge potentials as described in Section 2.6.1 of Chapter 2, we replace θ_{st} with $-|\theta_{st}|$ where x_s and x_t are Poisson nodes.

Appendix B

APPENDIX FOR CHAPTER 3

B.1 Proofs

B.1.1 Outline

In this section, we prove Theorems 3 and 4 from Section 3.4 in Chapter 3. The remaining subsections are organized as follows. In Section B.1.2, we list the additional assumptions for Theorem 3 in Chapter 3 and give the proof of Theorem 3 in Chapter 3. In Section B.1.3, we prove a theorem on variable selection consistency for group lasso regression with errors in variables, which itself is of independent interest. In Section B.1.4, we introduce Assumption 26 on the bases $\psi(\cdot)$, and several technical lemmas that are useful in proving Theorem 4 in Chapter 3. In Section B.1.5, we finish the proof of Theorem 4 in Chapter 3. And in Section B.1.6, we prove Proposition 4 in Chapter 3. The proofs of the technical lemmas presented in Section B.1.4 are provided in Section B.2.

B.1.2 Proof of Theorem 3

In this section, we follow closely the notation in Section 1.6 of Tsybakov (2009). We first present some necessary notation and assumptions. Denote the local polynomial estimator of degree ℓ as

$$\hat{X}(t;h) = \sum_{i=1}^{n} Y_i W_{ni}(t;h), \tag{B.1}$$

where

$$W_{ni}(t;h) = \frac{1}{nh} U^{\mathrm{T}}(0) B_{nt}^{-1} U\left(\frac{t_i - t}{h}\right) K\left(\frac{t_i - t}{h}\right), \tag{B.2}$$

$$B_{nt} = \frac{1}{nh}\sum_{i=1}^{n} U\left(\frac{t_i - t}{h}\right) U^{\mathrm{T}}\left(\frac{t_i - t}{h}\right) K\left(\frac{t_i - t}{h}\right),$$

$$U(u) = \left(1, u, u^2/2!, \ldots, u^{\ell}/\ell!\right)^{\mathrm{T}},$$

and $K(\cdot)$ is a kernel function. In (B.2), $W_{ni}(t;h)$ is the weight for observation Y_i in (B.1), which satisfies

$$\sum_{i=1}^{n} W_{ni}(t;h) = 1. \tag{B.3}$$

See e.g., Proposition 1.12 in Tsybakov (2009), for a rigorous proof of (B.3). We introduce the following assumptions on the kernel function $K(\cdot)$ and the time points $t_1, \ldots, t_n$. These assumptions are common in the study of local polynomial estimators (see e.g. Tsybakov, 2009).

Assumption 23. There exists a real number $\lambda_0 > 0$ and a positive integer n_0 such that the smallest eigenvalue $\Lambda_{\min}(B_{nt})$ of B_{nt} satisfies

$$\Lambda_{\min}(B_{nt}) \geq \lambda_0$$

for all $n \geq n_0$ and any $t \in [0,1]$.

Assumption 24. The time points $t_1, \ldots, t_n$ are evenly-spaced on the interval $[0,1]$.

Assumption 25. The kernel K has compact support belonging to $[-1,1]$, and there exists a number $K_{\max} < \infty$ such that $|K(u)| \leq K_{\max}$, $\forall u \in \mathbb{R}$.

These assumptions lead to the following lemma (Lemma 1.3 in Tsybakov, 2009).

Lemma 6. *Under Assumptions 23–25, for all $n \geq n_0$, $h \geq 1/(2n)$, and $t \in [0,1]$, the weights W_{ni} in* (B.2) *satisfy:*

i. $\sup_{i,t} |W_{ni}(t;h)| \leq C_3/nh$*;*

ii. $\sum_{i=1}^{n} |W_{ni}(t;h)| \leq C_3$*,*

where the constant C_3 depends only on λ_0 and $K_{\max}$.

Recall that we also assume the unknown true solutions X_j^*, $j = 1, \ldots, p$, belong to a Hölder class in Assumption 10 in Chapter 3. We state the definition here for completeness.

Definition 3. Let T be an interval in $\mathbb{R}$ and let β_1 and L_1 be two positive numbers. The Hölder class $\Sigma(\beta_1, L_1)$ on T is defined as the set of $\ell = \lfloor \beta_1 \rfloor$ times differentiable functions $f : T \to \mathbb{R}$ whose ℓth order derivative $f^{(\ell)}(\cdot)$ satisfies

$$|f^{(\ell)}(x) - f^{(\ell)}(x')| \leq L_1 |x - x'|^{\beta_1 - \ell}, \quad \forall x, x' \in T.$$

We are now ready to prove Theorem 3 of Chapter 3.

Proof.

$$\begin{aligned}
\left\| \hat{X}_j - X_j^* \right\|^2 &= \int_0^1 \{\hat{X}_j(u;h) - X_j^*(u)\}^2 \, du = \int_0^1 \left\{ \sum_{i=1}^n Y_{ij} W_{ni}(u;h) - X_j^*(u) \right\}^2 du \\
&= \int_0^1 \left[\sum_{i=1}^n \{X_j^*(t_i) + \epsilon_{ji}\} W_{ni}(u;h) - X_j^*(u) \right]^2 du.
\end{aligned}$$

Using the property (B.3) of the weights W_{ni} and the fact that $(a+b)^2 \leq 2a^2 + 2b^2$,

$$\begin{aligned}
\left\| \hat{X}_j - X_j^* \right\|^2 &\leq 2 \int_0^1 \left[\sum_{i=1}^n \{X_j^*(t_i) - X_j^*(u)\} W_{ni}(u;h) \right]^2 du \\
&\quad + 2 \int_0^1 \left\{ \sum_{i=1}^n \epsilon_{ji} W_{ni}(u;h) \right\}^2 du \qquad \text{(B.4)} \\
&\equiv 2 \int_0^1 \text{bias}^2(u) \, du + 2 \int_0^1 g^2(\epsilon_j/\sigma, u, h) \, du,
\end{aligned}$$

where

$$\text{bias}(u) = \sum_{i=1}^n \{X_j^*(t_i) - X_j^*(t)\} W_{ni}(u;h), \qquad \text{(B.5)}$$

$$g(a,u,h) = \sigma \sum_{i=1}^n a_i W_{ni}(u;h), \quad \epsilon_j = (\epsilon_{1j}, \ldots, \epsilon_{nj})^{\mathrm{T}}, \qquad \text{(B.6)}$$

and where σ is defined in Assumption 9 in Chapter 3.

In what follows, for convenience, we denote the ℓth derivative of $X_j^*(t)$ as $X_j^{(\ell)}$. Under Assumption 10 in Chapter 3 and Assumptions 23–25, it follows from Proposition 1.13 in Tsybakov

(2009) that $|\text{bias}(u)| \leq q_1 h^{\beta_1}$, where $q_1 = C_3 L_1/\ell!$. Therefore,

$$\int_0^1 \text{bias}^2(u)\, du \leq q_1^2 h^{2\beta_1}. \tag{B.7}$$

Next, we bound $g(\epsilon_j/\sigma, t, h)$ in (B.6) using Theorem 5.6 in Boucheron et al. (2013). The theorem states that if $Z = (Z_1, \ldots, Z_n)$ is a vector of n independent standard normal random variables and f is an L-Lipschitz function, then for all $v > 0$,

$$\Pr\{f(Z) - \mathbb{E}f(Z) \geq v\} \leq \exp\{-v^2/(2L^2)\}.$$

Applying the theorem to $f(z)$ and $-f(z)$, we get

$$\Pr\{|f(Z) - \mathbb{E}f(Z)| \geq v\} \leq 2\exp\{-v^2/(2L^2)\}.$$

We now show that $g(x, t, h)$ is an L_3-Lipschitz function with $L_3 = \sigma C_3 (nh)^{-0.5}$:

$$\begin{aligned}
|g(a, u, h) - g(b, u, h)| =& \sigma \left| \sum_{i=1}^n (a_i - b_i) W_{ni}(u; h) \right| \\
\leq& \sigma \left\{ \sum_{i=1}^n W_{ni}^2(u; h) \right\}^{\frac{1}{2}} \|a - b\|_2 \\
\leq& \sigma \left\{ \sup_{i,u} |W_{ni}(u; h)| \sum_{i=1}^n |W_{ni}(u, h)| \right\}^{\frac{1}{2}} \|a - b\|_2 \\
\leq& \sigma C_3 \sqrt{\frac{1}{nh}} \|a - b\|_2,
\end{aligned}$$

where the last inequality follows from Lemma 6. Hence, from Theorem 5.6 in Boucheron et al. (2013), we have

$$\Pr\{|g(\epsilon_j/\sigma, u, h) - \mathbb{E}g(\epsilon_j/\sigma, u, h)| \geq v\} \leq 2\exp\{-v^2/(2L_3^2)\}.$$

Letting $v = n^{\alpha/2 - 0.5} h^{-0.5}$ and noting that $\mathbb{E}[g(\epsilon_j/\sigma, u, h)] = 0$, we have

$$\Pr\{|g(\epsilon_j/\sigma, u, h)| \geq n^{\alpha/2-0.5} h^{-0.5}\} \leq 2\exp\{-n^\alpha/(2\sigma^2 C_3^2)\}. \tag{B.8}$$

Combining (B.4), (B.7), and (B.8), we have

$$\begin{aligned}
\left\| \hat{X}_j - X_j^* \right\|^2 &\leq 2 \int_0^1 \text{bias}^2(u)\, du + 2 \int_0^1 g^2(\epsilon_j/\sigma, u, h)\, du \\
&\leq 2q_1^2 h^{2\beta_1} + 2n^{\alpha-1} h^{-1},
\end{aligned} \tag{B.9}$$

with probability at least $1 - 2\exp\{-n^\alpha/(2\sigma^2 C_3^2)\}$.

Minimizing the right-hand side of (B.9) with respect to h, we find that the minimizer h_n satisfies

$$2\beta_1 q_1^2 h_n^{2\beta_1+1} = n^{\alpha-1}.$$

Thus, for $h_n \propto n^{(\alpha-1)/(2\beta_1+1)}$, the error bound is

$$\left|\left|\left| \hat{X}_j - X_j^* \right|\right|\right|^2 \leq C_2 n^{\frac{2\beta_1}{2\beta_1+1}(\alpha-1)},$$

for some global constant C_2. □

B.1.3 Variable selection consistency of group lasso in error-in-variable models

We first review some notation that is heavily used in this section. In (3.17c) of Chapter 3, we made use of the notation

$$\hat{\Psi}_0(t) = t;\ \hat{\Psi}_k(t) = \int_0^t \psi(\hat{X}_k(u;h))\, du,\ k = 1, \ldots, p.$$

Therefore, $\hat{\Psi}_k(t)$ is an M-vector for $k = 1, \ldots, p$ and a scalar for $k = 0$. We sometimes use sets, e.g. S_j and S_j^0, as the subscripts. In this case, $\hat{\Psi}_{S_j}(t)$ is an Ms_j-vector, which is composed of $\hat{\Psi}_k$ for $k \in S_j$. Furthermore, $\hat{\Psi}_{S_j^0} = (\hat{\Psi}_0(t), \hat{\Psi}_{S_j}^{\mathrm{T}}(t))^{\mathrm{T}}$ is an $(Ms_j + 1)$-vector. Without subscripts, $\hat{\Psi}(t) \equiv (\hat{\Psi}_0(t), \hat{\Psi}_1^{\mathrm{T}}(t), \ldots, \hat{\Psi}_p^{\mathrm{T}}(t))^{\mathrm{T}}$ is of dimension $Mp + 1$. We will also apply subscripts to the quantities θ_j^*, $\hat{\theta}_j$, $\hat{g}$, and R. For instance, $\hat{\theta}_{jk} = (\theta_{jk1}, \ldots, \theta_{jkM})^{\mathrm{T}}$ for $k = 1, \ldots, p$, and $\hat{\theta}_j = (\hat{\theta}_{j0}, \hat{\theta}_{j1}^{\mathrm{T}}, \ldots, \hat{\theta}_{jp}^{\mathrm{T}})^{\mathrm{T}}$. The products of these vectors are defined as usual, e.g., $\hat{\theta}_{jS_j^0}^{\mathrm{T}} \hat{\Psi}_{S_j^0}(t)$ is a scalar, and $\hat{\Psi}_{S_j^0}(t)\hat{\Psi}_{S_j^0}^{\mathrm{T}}(t)$ is an $(Ms_j + 1) \times (Ms_j + 1)$ matrix.

The optimization problem (3.17a) in Chapter 3 is a standardized group lasso problem (Simon and Tibshirani, 2012). Because the regressors $\hat{\Psi}_1, \ldots, \hat{\Psi}_p$ are estimated, establishing variable selection consistency requires extra attention. For ease of discussion, we re-state the optimization problem (3.17a),

$$\hat{\theta}_j = \underset{C_0 \in \mathbb{R}, \theta_{j0} \in \mathbb{R},\ \theta_{jk} \in \mathbb{R}^M}{\arg\min} \frac{1}{2n} \sum_{i=1}^n \left\{ Y_{ij} - C_0 - \theta_{j0} \hat{\Psi}_0(t_i) - \sum_{k=1}^p \theta_{jk}^{\mathrm{T}} \hat{\Psi}_k(t_i) \right\}^2 + \lambda_{n,j} \sum_{k=1}^p \left[\frac{1}{n} \sum_{i=1}^n \{\theta_{jk}^{\mathrm{T}} \hat{\Psi}_k(t_i)\}^2 \right]^{1/2},$$

where

$$\hat{X}(\cdot;h) = \underset{Z(\cdot)\in\mathcal{X}(h)}{\arg\min} \sum_{i=1}^{n} \|Y_i - Z(t_i)\|_2^2,$$

$$\hat{\Psi}_0(t) = t;\ \hat{\Psi}_k(t) = \int_0^t \psi(\hat{X}_k(u;h))\,du,\ k = 1,\ldots,p.$$

In what follows, for simplicity we assume that $X_j^*(0) = 0$, and that $\lambda_{n,1} = \cdots = \lambda_{n,p} \equiv \lambda_n$. For any $1 \le j,k \le p$, let $\theta_{jk}^* \in \mathbb{R}^M$ be the coefficients of the true functions f_{jk}^* on the bases $\psi(\cdot)$, i.e.,

$$f_{jk}^*(a) = \psi(a)^{\mathrm{T}}\theta_{jk}^* + \delta_{jk}(a), \tag{B.10}$$

where f_{jk}^* is introduced in Assumption 11 in Chapter 3. Here we establish variable selection consistency for group lasso regression with errors in variables. We extend the recent work of Loh and Wainwright (2012) for lasso regression; related results can be found in Ma and Li (2010) and Rosenbaum and Tsybakov (2010). In order for variable selection consistency to hold, we need four conditions. In Section B.1.5, we will show that these conditions hold with high probability given Assumptions 9–14 in Chapter 3 and Assumptions 23–26.

Condition 1. Suppose that

$$\begin{aligned}
&0 < \frac{1}{2}C_{\min} \le \Lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^{n} \hat{\Psi}_{S_j^0}(t_i)\hat{\Psi}_{S_j^0}^{\mathrm{T}}(t_i)\right),\\
&\Lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^{n} \hat{\Psi}_{S_j^0}(t_i)\hat{\Psi}_{S_j^0}^{\mathrm{T}}(t_i)\right) \le 2C_{\max},\\
&0 < \frac{1}{2}C_{\min} \le \Lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^{n} \hat{\Psi}_{k}(t_i)\hat{\Psi}_{k}^{\mathrm{T}}(t_i)\right), \quad k \notin S_j^0,
\end{aligned}$$

where $C_{\min}$ and $C_{\max}$ are introduced in Assumption 12 in Chapter 3.

Condition 2. Assume that

$$\max_{k\notin S_j^0} \left\| \left(\frac{1}{n}\sum_{i=1}^{n} \hat{\Psi}_{k}(t_i)\hat{\Psi}_{S_j^0}^{\mathrm{T}}(t_i)\right)\left(\frac{1}{n}\sum_{i=1}^{n} \hat{\Psi}_{S_j^0}(t_i)\hat{\Psi}_{S_j^0}^{\mathrm{T}}(t_i)\right)^{-1} \right\|_2 \le 2\xi,$$

where ξ is introduced in Assumption 13.

The next condition was first proposed in Loh and Wainwright (2012) as the deviation condition. Specifically, (B.11) is a special case of Equation 3.1 in Loh and Wainwright (2012). Recall that the true parameters θ^*_{j0} and θ^*_{jk} are introduced in Assumption 11 of Chapter 3 and (B.10), respectively.

Condition 3. For $j = 1, \ldots, p$, let $\Delta \equiv \max_{j=1,\ldots,p} \left|\!\left|\!\left| \hat{X}_j - X^*_j \right|\!\right|\!\right|$. Assume that

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \hat{\Psi}_k(t_i) Y_{ij} - \frac{1}{n} \sum_{i=1}^{n} \hat{\Psi}_k(t_i) \hat{\Psi}^{\mathrm{T}}_{S^0_j}(t_i) \theta^*_{jS^0_j} \right\|_2 \leq \eta, \ k = 0, \ldots, p \tag{B.11}$$

where $\eta = M^{1/2} \left\{ sM^{-\beta_2} Q^{1/2} B + BD \|\theta^*_S\|_1 \Delta + n^{\alpha/2 - 1/2} \right\}$.

Note that the global constant Q in Condition 3 also appears in Assumption 26 in Section B.1.4.

Condition 4 places constraints on the quantities involved in the proof of Theorem 10. In the proof of Theorem 4 in Chapter 3, we will show that Condition 4 holds with high probability.

Condition 4. The following inequalities hold:

$$\frac{2\sqrt{s+1}}{C_{\min}} \eta + \lambda_n \frac{\sqrt{8sC_{\max}}}{C_{\min}} \leq \frac{2}{3} \theta_{\min},$$

$$\frac{2\xi\sqrt{s+1} + 1}{\lambda_n} \eta + 2\xi\sqrt{s}\sqrt{2C_{\max}} < \sqrt{C_{\min}/2},$$

where $\theta_{\min} \equiv \min_{k \in S^0_j} \|\theta^*_{jk}\|_2$, and $\xi, \eta, C_{\min}$, and $C_{\max}$ are introduced in Assumptions 12–14 of Chapter 3.

We arrive at the following theorem.

Theorem 10. *Suppose that Conditions 1–4 are met. Then the estimator $\hat{\theta}_j$ from* (3.17a) *has the correct support, i.e. $\hat{S}_j = S_j$ for all $j = 1, \ldots, p$.*

Proof. We establish variable selection consistency using the primal-dual witness method (Wainwright, 2009). For simplicity, we drop the subscript j in what follows: for instance, we drop the subscript j in Y_{ij} and $\hat{\theta}_j$ in (3.17a), and in the estimated neighbourhood $\hat{S}_j$.

A vector $\hat{\theta}$ solves the optimization problem (3.17a) in Chapter 3 if it satisfies the Karush-Kuhn-Tucker (KKT) condition, which is

$$\frac{1}{n} \sum_{i=1}^{n} \hat{\Psi}_k(t_i) \left\{ \hat{\Psi}^{\mathrm{T}}(t_i) \hat{\theta} - Y_i \right\} + \lambda_n \hat{g}_k = 0, \quad k = 1, \ldots, p, \tag{B.12}$$

with

$$\begin{aligned} \hat{g}_k &= \frac{\sum_{i=1}^n \hat{\Psi}_k(t_i)\hat{\Psi}_k^{\mathrm{T}}(t_i)\hat{\theta}_k/n}{\sqrt{\hat{\theta}_k^{\mathrm{T}}\sum_{i=1}^n \hat{\Psi}_k(t_i)\hat{\Psi}_k^{\mathrm{T}}(t_i)\hat{\theta}_k/n}} \quad \text{if } \hat{\theta}_k \neq 0, \\ \hat{g}_k^{\mathrm{T}}\left(\frac{1}{n}\sum_{i=1}^n \hat{\Psi}_k(t_i)\hat{\Psi}_k^{\mathrm{T}}(t_i)\right)^{-1} \hat{g}_k &< 1 \quad \text{if } \hat{\theta}_k = 0. \end{aligned} \tag{B.13}$$

The KKT condition for $\hat{\theta}_0$ is

$$\frac{1}{n}\sum_{i=1}^n \hat{\Psi}_0(t_i)\left\{\hat{\Psi}^{\mathrm{T}}(t_i)\hat{\theta} - Y_i\right\} = 0. \tag{B.14}$$

Note that, in the previous equations, we drop the parameter C_0 that appears in (3.17a) of Chapter 3 to avoid cumbersome bookkeeping.

We will construct an oracle estimator $\hat{\theta}$ and will verify that it satisfies the KKT conditions (B.12), (B.13), and (B.14), which means that it solves the optimization problem (3.17a) in Chapter 3.

We construct an oracle primal-dual pair $(\hat{\theta}, \hat{g})$ as follows:

1. Set $\hat{\theta}_k = 0$ for $k \notin S^0$.

2. Let

$$\hat{\theta}_{S^0} = \underset{\theta_{S^0}\in\mathbb{R}^{sM+1}}{\arg\min} \frac{1}{2n}\sum_{i=1}^n \left\{Y_i - \theta_{S^0}^{\mathrm{T}}\hat{\Psi}_{S^0}(t_i)\right\}^2 + \lambda_n \sum_{k\in S}\left[\frac{1}{n}\sum_{i=1}^n \{\theta_{jk}^{\mathrm{T}}\hat{\Psi}_k(t_i)\}^2\right]^{1/2}. \tag{B.15}$$

3. Define $\hat{g}_{S^0} = (0, \hat{g}_S^{\mathrm{T}})^{\mathrm{T}}$ as in (B.13).

4. Solve $\hat{g}_k$ from the sub-gradient condition (B.12) for $k \notin S^0$.

We will verify the support recovery consistency

$$\max_{k\in S} \|\hat{\theta}_k - \theta_k^*\|_2 \leq \frac{2}{3}\theta_{\min} \tag{B.16}$$

and strict dual feasibility

$$\max_{k\notin S^0} \hat{g}_k^{\mathrm{T}}\left(\frac{1}{n}\sum_{i=1}^n \hat{\Psi}_k(t_i)\hat{\Psi}_k^{\mathrm{T}}(t_i)\right)^{-1} \hat{g}_k < 1. \tag{B.17}$$

(B.16) implies that the oracle estimator $\hat{\theta}$ recovers the support of θ^* exactly, and (B.17) implies that $\hat{\theta}$ solves (3.17a).

Further, if the optimal solution to (3.17a) is unique, then the oracle estimator is the unique estimator. If the optimal solution is not unique, then from Theorem 2 in Roth and Fischer (2008), the null set of any optimal solution should contain S^c, and thus any optimal solution satisfies the construction of the oracle estimator. Therefore, the statement of Theorem 10 holds for any optimal solution for (3.17a).

We now establish (B.16). The subgradient condition for the constrained problem (B.15) is

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\Psi}_{S^0}(t_i)\big\{\hat{\Psi}_{S^0}^{\mathrm{T}}(t_i)\hat{\theta}_{S^0}-Y_i\big\}+\lambda_n\hat{g}_{S^0}=0. \tag{B.18}$$

Adding and subtracting $\frac{1}{n}\sum_{i=1}^{n}\hat{\Psi}_{S^0}(t_i)\hat{\Psi}_{S^0}^{\mathrm{T}}(t_i)\theta^*_{S^0}$, we get

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\Big\{\hat{\Psi}_{S^0}(t_i)\hat{\Psi}_{S^0}^{\mathrm{T}}(t_i)\hat{\theta}_{S^0}-\hat{\Psi}_{S^0}(t_i)\hat{\Psi}_{S^0}^{\mathrm{T}}(t_i)\theta^*_{S^0}\Big\}+\\&\frac{1}{n}\sum_{i=1}^{n}\Big\{\hat{\Psi}_{S^0}(t_i)\hat{\Psi}_{S^0}^{\mathrm{T}}(t_i)\theta^*_{S^0}-\hat{\Psi}_{S^0}(t_i)Y_i\Big\}+\lambda_n\hat{g}_{S^0}=0.\end{aligned}$$

Rearranging the terms and letting

$$R_{S^0}\equiv\frac{1}{n}\sum_{i=1}^{n}\hat{\Psi}_{S^0}(t_i)\hat{\Psi}_{S^0}^{\mathrm{T}}(t_i)\theta^*_{S^0}-\frac{1}{n}\sum_{i=1}^{n}\hat{\Psi}_{S^0}(t_i)Y_i, \tag{B.19}$$

we get

$$\hat{\theta}_{S^0}-\theta^*_{S^0}=-\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\Psi}_{S^0}(t_i)\hat{\Psi}_{S^0}^{\mathrm{T}}(t_i)\right)^{-1}\left(R_{S^0}+\lambda_n\hat{g}_{S^0}\right). \tag{B.20}$$

By the definition of R_{S^0} in (B.19), for each $k\in S$, we have that

$$R_k=\frac{1}{n}\sum_{i=1}^{n}\hat{\Psi}_{k}(t_i)\hat{\Psi}_{S^0}^{\mathrm{T}}(t_i)\theta^*_{S^0}-\frac{1}{n}\sum_{i=1}^{n}\hat{\Psi}_{k}(t_i)Y_i, \tag{B.21}$$

and $R_0=\frac{1}{n}\sum_{i=1}^{n}t_i\{\hat{\Psi}_{S^0}^{\mathrm{T}}(t_i)\theta^*_{S^0}-Y_i\}$. By Condition 3, we know that $\|R_k\|_2\le\eta$ for $k\in S^0$. Hence,

$$\|R_{S^0}\|_2\le\eta\sqrt{s+1}. \tag{B.22}$$

By Condition 1, we have that

$$\Lambda_{\max}\left\{\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\Psi}_{S^0}(t_i)\hat{\Psi}_{S^0}^{\mathrm{T}}(t_i)\right)^{-1}\right\}\leq\frac{2}{C_{\min}}. \tag{B.23}$$

From (B.13) and the fact that the largest eigenvalue of a submatrix is no greater than the largest eigenvalue of the matrix,

$$\frac{1}{2C_{\max}}\|\hat{g}_k\|_2^2\leq\hat{g}_k^{\mathrm{T}}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\Psi}_k(t_i)\hat{\Psi}_k^{\mathrm{T}}(t_i)\right)^{-1}\hat{g}_k=1,\quad k\in S.$$

Furthermore, $\hat{g}_0=0$ by construction. Hence,

$$\|\hat{g}_{S^0}\|_2=\left\{\|\hat{g}_0\|_2^2+\|\hat{g}_S\|_2^2\right\}^{1/2}\leq\sqrt{2sC_{\max}}. \tag{B.24}$$

Therefore, combining (B.20), (B.22), (B.23), and (B.24), it follows that

$$\max_{k\in S}\|\hat{\theta}_k-\theta_k^*\|_2\leq\|\hat{\theta}_{S^0}-\theta_{S^0}^*\|_2\leq\frac{2\eta\sqrt{s+1}}{C_{\min}}+\lambda_n\frac{\sqrt{8sC_{\max}}}{C_{\min}}\leq\frac{2}{3}\theta_{\min},$$

where the last inequality follows from Condition 4.

Next, we verify strict feasibility (B.17). For $k\notin S^0$, from (B.12),

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\Psi}_k(t_i)\big(\hat{\Psi}_{S^0}^{\mathrm{T}}(t_i)\hat{\theta}_{S^0}-Y_i\big)+\lambda_n\hat{g}_k=0.$$

Adding and subtracting $\frac{1}{n}\sum_{i=1}^{n}\hat{\Psi}_k(t_i)\hat{\Psi}_{S^0}^{\mathrm{T}}(t_i)\theta_{S^0}^*$ yields

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\left\{\hat{\Psi}_k(t_i)\hat{\Psi}_{S^0}^{\mathrm{T}}(t_i)\hat{\theta}_{S^0}-\hat{\Psi}_k(t_i)\hat{\Psi}_{S^0}^{\mathrm{T}}(t_i)\theta_{S^0}^*\right\}+\\&\frac{1}{n}\sum_{i=1}^{n}\left\{\hat{\Psi}_k(t_i)\hat{\Psi}_{S^0}^{\mathrm{T}}(t_i)\theta_{S^0}^*-\hat{\Psi}_k(t_i)Y_i\right\}+\lambda_n\hat{g}_k=0.\end{aligned}$$

Rearranging the terms and plugging in (B.20) and (B.21), we get

$$\lambda_n\hat{g}_k=\frac{1}{n}\sum_{i=1}^{n}\hat{\Psi}_k(t_i)\hat{\Psi}_{S^0}^{\mathrm{T}}(t_i)\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\Psi}_{S^0}(t_i)\hat{\Psi}_{S^0}^{\mathrm{T}}(t_i)\right)^{-1}(R_{S^0}+\lambda_n\hat{g}_{S^0})-R_k.$$

By Condition 2, we know that

$$\max_{k\notin S^0}\left\|\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\Psi}_k(t_i)\hat{\Psi}_{S^0}^{\mathrm{T}}(t_i)\right)\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\Psi}_{S^0}(t_i)\hat{\Psi}_{S^0}^{\mathrm{T}}(t_i)\right)^{-1}\right\|_2\leq 2\xi.$$

Recall from Condition 3 that $\|R_k\|_2 \leq \eta$ for $1 \leq k \leq p$. Using (B.22) and (B.24), we have that

$$\|\hat{g}_k\|_2 \leq \frac{2\xi\sqrt{s+1}+1}{\lambda_n}\eta + 2\xi\sqrt{s}\sqrt{2C_{\max}}, \quad k \notin S^0.$$

By Condition 4, $\|\hat{g}_k\|_2 < \sqrt{C_{\min}/2}$, and thus, applying Condition 1,

$$\hat{g}_k^{\mathrm{T}} \left(\frac{1}{n} \sum_{i=1}^{n} \hat{\Psi}_k(t_i) \hat{\Psi}_k^{\mathrm{T}}(t_i) \right)^{-1} \hat{g}_k \leq \frac{2\|\hat{g}_k\|_2^2}{C_{\min}} < 1, \quad k \notin S^0.$$

Therefore, we have established (B.17). □

B.1.4 Assumption 26 and technical lemmas

Theorem 10 characterizes the samples on which the GRADE estimator is able to reconstruct the true network. We must now establish that with high probability, the observations satisfy Conditions 1–4. In Section B.1.5, Lemmas 8–10, stated below, will be used to show that Conditions 1–4, needed for Theorem 10, hold with high probability. Lemma 7 is used to prove Lemmas 8–10. Lemmas 7 – 10 are proven in Appendix B.2.

First, we state the regularity condition on the bases ψ mentioned in Section 3.4 in Chapter 3.

Assumption 26. The basis functions are orthonormal, i.e., $\int_0^1 \psi_{jk}\big(X_k^*(u)\big)\psi_{jk}^{\mathrm{T}}\big(X_k^*(u)\big)du = I_M$, where I_M is an $M \times M$ identity matrix. The basis functions are bounded and have bounded first order derivative, i.e. $|\psi_m(x)| \leq B,\ |\psi_m'(x)| \leq D, m = 1, \ldots, M$. Further, under Assumption 11 in Chapter 3, for any j, k,

$$\int_0^1 \delta_{jk}^2(u)du = \int_0^1 \left\{ f_{jk}^*(X_k^*(u)) - \psi^{\mathrm{T}}(X_k^*(u))\theta_{jk}^* \right\}^2 du \leq Q(M+1)^{-2\beta_2}, \tag{B.25}$$

where θ_{jk}^* is defined in (B.10) and Q is a global constant.

Remark 7. Assumption 26 holds, for instance, when $\psi(\cdot)$ is the set of trigonometric basis functions (see, e.g., Section 1.7.3 in Tsybakov (2009)).

We next state the technical lemmas used in the proof of Theorem 4 in Chapter 3.

Lemma 7. *Suppose that Assumption 11 in Chapter 3 and Assumption 26 hold, and* $\psi(t) = (\psi_0(t), \psi_1(t), \ldots, \psi_M(t))^{\mathrm{T}}$ *is of degree* M*. Then,*

$$\left| \|\theta^*_{jk}\|_2 - \left\{ \int_0^1 \left[f^*_{jk}(X^*_k(u)) \right]^2 du \right\}^{1/2} \right| \leq \sqrt{Q} M^{-\beta_2}. \tag{B.26}$$

$$\left| \left| \left| X^*_j - \Psi^{\mathrm{T}}_{S^0} \theta^*_{S^0} \right| \right| \right| \leq s\sqrt{QM^{-2\beta_2}}, \tag{B.27}$$

and

$$\frac{1}{n} \sum_{i=1}^{n} \{ X^*_j(t_i) - \Psi^{\mathrm{T}}_{S^0}(t_i) \theta^*_{S^0} \}^2 \leq s^2 Q M^{-2\beta_2} + o\left(n^{-2}\right), \tag{B.28}$$

where θ^*_{jk} *is defined in* (B.10) *and* Q *is a constant in Assumption 26.*

Lemma 8. *Suppose that Assumptions 11 and 12 in Chapter 3 and Assumption 26 hold. Let* $\Delta \equiv \max_{j=1,\ldots,p} \left| \left| \left| \hat{X}_j - X^*_j \right| \right| \right|$*. The following bounds on the eigenvalues of* $\sum_{i=1}^n \hat{\Psi}_{S^0} \hat{\Psi}^{\mathrm{T}}_{S^0} / n$ *hold:*

$$\begin{aligned}
&\Lambda_{\min} \left(\frac{1}{n} \sum_{i=1}^{n} \hat{\Psi}_{S^0}(t_i) \hat{\Psi}^{\mathrm{T}}_{S^0}(t_i) \right) \geq C_{\min} - \left(2BD\Delta + \frac{BD + B^2}{6n^2} \right) (Ms + 1), \\
&\Lambda_{\max} \left(\frac{1}{n} \sum_{i=1}^{n} \hat{\Psi}_{S^0}(t_i) \hat{\Psi}^{\mathrm{T}}_{S^0}(t_i) \right) \leq C_{\max} + \left(2BD\Delta + \frac{BD + B^2}{6n^2} \right) (Ms + 1), \\
\textit{and} \quad &\Lambda_{\min} \left(\frac{1}{n} \sum_{i=1}^{n} \hat{\Psi}_k(t_i) \hat{\Psi}^{\mathrm{T}}_k(t_i) \right) \geq C_{\min} - \left(2BD\Delta + \frac{BD + B^2}{6n^2} \right) M, \quad k \notin S^0_j.
\end{aligned} \tag{B.29}$$

Lemma 9. *Suppose that Assumptions 11 and 13 in Chapter 3 and Assumption 26 hold. Let* $\Delta \equiv \max_{j=1,\ldots,p} \left| \left| \left| \hat{X}_j - X^*_j \right| \right| \right|$*. Then,*

$$\begin{aligned}
&\left\| \left(\frac{1}{n} \sum_{i=1}^{n} \hat{\Psi}_k(t_i) \hat{\Psi}^{\mathrm{T}}_{S^0}(t_i) \right) \left(\frac{1}{n} \sum_{i=1}^{n} \hat{\Psi}_{S^0}(t_i) \hat{\Psi}^{\mathrm{T}}_{S^0}(t_i) \right)^{-1} \right\|_2 \leq \\
&\xi + \left\{ c_1 \hat{C}^{-2}_{\min} M (Ms + 1)^3 \Delta^2 \right\}^{1/2} + \left\{ c_2 M (Ms + 1) \Delta^2 \right\}^{1/2} + \left\{ c_3 M (Ms + 1)^3 / 6n^2 \right\}^{1/2},
\end{aligned} \tag{B.30}$$

where $\hat{C}_{\min} \equiv C_{\min} - \left(2BD\Delta + \frac{BD + B^2}{6n^2} \right) (Ms + 1)$*, and* c_1, c_2, c_3 *are constants.*

Lemma 10. *Suppose Assumptions 9, 10, and 11 in Chapter 3 and Assumption 26 hold. Let* $\Delta \equiv \max_{j=1,\ldots,p} \left| \left| \left| \hat{X}_j - X^*_j \right| \right| \right|$*. For each* $k = 0, \ldots, p$*,*

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \hat{\Psi}_k(t_i) Y_{ij} - \frac{1}{n} \sum_{i=1}^{n} \hat{\Psi}_k(t_i) \hat{\Psi}^{\mathrm{T}}_{S^0}(t_i) \theta^*_{S^0} \right\|_2 \leq \eta, \tag{B.31}$$

where

$$\eta \equiv M^{1/2}\left\{sM^{-\beta_2}Q^{1/2}B + BD\|\theta_S^*\|_1\Delta + n^{\alpha/2-1/2}\right\}$$

with probability at least $1 - 2M\exp\{-n^\alpha/(2B^2\sigma^2)\}$.

B.1.5 Proof of Theorem 4

Proof. Notice that Theorem 10 offers the desired result of Theorem 4 in Chapter 3. We now verify that Conditions 1–4 hold with high probability given the assumptions for Theorem 4 of Chapter 3. This completes the proof of Theorem 4 of Chapter 3.

First of all, Lemma 10 tells us that Condition 3 holds with probability at least $1-2pM\exp^{-n^\alpha/(2B^2\sigma^2)}$. This probability converges to unity as p and n grow, because $M \propto n^{\frac{2}{2\beta_2+1}\frac{\beta_1}{2\beta_1+1}(1-\alpha)} = o(n)$ and $pn\exp(-C_4n^\alpha/\sigma^2) = o(1)$ as required in Theorem 4 of Chapter 3, where $C_4 \equiv \min\{1/(2B^2), 1/(2C_3^2)\}$. Thus, Condition 3 holds with high probability.

Next, we verify that Condition 4 holds with high probability. Given Assumptions 9–10 and 23–25, we know from Theorem 3 in Chapter 3 that

$$\max_j \left\|\left\|\hat{X}_j - X_j^*\right\|\right\| \equiv \Delta = O\left(n^{\frac{\beta_1}{2\beta_1+1}(\alpha-1)}\right), \tag{B.32}$$

with probability at least $1 - 2p\exp\{-n^\alpha/(2C_3\sigma^2)\}$. Recall that in Theorem 4 of Chapter 3 we require that $s = O(n^\gamma)$ and $M \propto n^{\frac{2}{2\beta_2+1}\frac{\beta_1}{2\beta_1+1}(1-\alpha)}$. Furthermore, $\|\theta_k^*\|_1 < \sqrt{M}\|\theta_k^*\|_2$, and $\|\theta_k^*\|_2$ is bounded by a constant due to the fact that f_{jk}^* is bounded and (B.26). Combining these with (B.32), we know that the three terms of η in Condition 3 satisfy

$$sM^{-\beta_2+1/2}Q^{1/2}B = O\left(n^{-\frac{2\beta_2-1}{2\beta_2+1}\frac{\beta_1}{2\beta_1+1}(1-\alpha)+\gamma}\right),$$

$$M^{1/2}BD\|\theta_S^*\|_1\Delta = O\left(n^{-\frac{2\beta_2-1}{2\beta_2+1}\frac{\beta_1}{2\beta_1+1}(1-\alpha)+\gamma}\right),$$

and

$$M^{1/2}n^{\alpha/2-1/2} = O\left(n^{(\frac{1}{2\beta_2+1}\frac{\beta_1}{2\beta_1+1}-\frac{1}{2})(1-\alpha)}\right).$$

These lead to

$$\left\|\frac{1}{n}\sum_{i=1}^n \hat{\Psi}_k(t_i)Y_{ij} - \frac{1}{n}\sum_{i=1}^n \hat{\Psi}_k(t_i)\hat{\Psi}_{S^0}^{\mathrm{T}}(t_i)\theta_{S^0}^*\right\|_2 \leq \eta = O\left(n^{-\frac{2\beta_2-1}{2\beta_2+1}\frac{\beta_1}{2\beta_1+1}(1-\alpha)+\gamma}\right) \tag{B.33}$$

with probability at least $1 - 2pM\exp\{-n^\alpha/(2B^2\sigma^2)\}$ for all $k = 0, \ldots, p$, from Lemma 10.

In Theorem 4 of Chapter 3, we require that $\lambda_n \propto n^{-\frac{\beta_1}{2\beta_1+1}\frac{2\beta_2-1}{2\beta_2+1}(1-\alpha)+2\gamma}$. Given (B.33) and $s = O(n^\gamma)$, we know that $\sqrt{s}\eta = o(\lambda_n)$. Furthermore, define

$$H_1(\beta_1, \beta_2, \alpha) \equiv \min\left\{\frac{\beta_1}{2\beta_1+1}\frac{2\beta_2-1}{4\beta_2+2}(1-\alpha), \frac{2}{3}\frac{\beta_1}{2\beta_1+1}\frac{2\beta_2-3}{2\beta_2+1}(1-\alpha)\right\}. \tag{B.34}$$

Then,

$$-\frac{\beta_1}{2\beta_1+1}\frac{2\beta_2-1}{2\beta_2+1}(1-\alpha) + 2\gamma \leq -2H_1(\beta_1, \beta_2, \alpha) + 2\gamma.$$

Thus, $\lambda_n = o(1)$ for $\gamma < H_1(\beta_1, \beta_2, \alpha)$. Further notice that $M^{-\beta_2} \propto n^{-\frac{2\beta_2}{2\beta_2+1}\frac{\beta_1}{2\beta_1+1}(1-\alpha)} = o(1)$, which implies that $\theta_{\min} \geq 3f_{\min}/4$ for sufficiently large n from (B.26) in Lemma 7. As a result, the two inequalities in Condition 4 become

$$o(\lambda_n) + \lambda_n\frac{\sqrt{8sC_{\max}}}{C_{\min}} \leq \frac{f_{\min}}{2},$$

$$o(1) + 2\xi\sqrt{s}\sqrt{2C_{\max}} < \sqrt{C_{\min}/2},$$

which hold for sufficiently large n under Assumption 14 of Chapter 3.

Note that the probability that (B.32) and (B.33) both hold is at least $1-2pM\exp\{-n^\alpha/(2B^2\sigma^2)\}-2p\exp\{-n^\alpha/(2C_3^2\sigma^2)\}$. Letting $C_4 = \min\{1/(2B^2), 1/(2C_3^2)\}$, we know from Theorem 4 that $pn\exp(-C_4n^\alpha/\sigma^2) = o(1)$. Combining this with $M \propto n^{\frac{2}{2\beta_2+1}\frac{\beta_1}{2\beta_1+1}(1-\alpha)} = o(n)$, we know that $1-2pM\exp\{-n^\alpha/(2B^2\sigma^2)\}-2p\exp\{-n^\alpha/(2C_3^2\sigma^2)\}$ converges to 1 as p, s, and n grow. Therefore, Condition 4 holds with high probability.

Finally, we establish that Conditions 1 and 2 hold with high probability. Note that the dominant terms not involving $C_{\min}, C_{\max}$ or ξ in the bounds in (B.29) in Lemma 8 and (B.30) in Lemma 9 involve $sM\Delta$ and $s^{3/2}M^2\Delta$, respectively. Given (B.32), one can check that

$$sM\Delta \propto n^{\frac{\beta_1}{2\beta_1+1}\frac{2\beta_2-1}{2\beta_2+1}(1-\alpha)+\gamma} = o(1), \text{ and} \tag{B.35}$$

$$s^{3/2}M^2\Delta \propto n^{\frac{\beta_1}{2\beta_1+1}\frac{2\beta_2-3}{2\beta_2+1}(1-\alpha)+\frac{3}{2}\gamma} = o(1), \tag{B.36}$$

where we have used the fact that $\beta_2 \geq 3$ in Assumption 11 in Chapter 3 as well as the fact that $\gamma < H_1(\beta_1, \beta_2, \alpha)$ from the statement of Theorem 4 in Chapter 3. Since (B.32) and (B.33) hold

with high probability, combining the inequalities in Lemmas 8 and 9 with (B.35) and (B.36), we see that Conditions 1 and 2 hold with high probability given Assumptions 11, 12 and 13 in Chapter 3.

In summary, we have shown that Conditions 1–4 hold with high probability. Applying Theorem 10 establishes that the GRADE estimator $\hat{S}_j$ in (3.17) in Chapter 3 recovers the true support S_j^*. □

B.1.6 Proof of Proposition 4

In Proposition 4, the choice of bandwidth h_n is different from that in Theorems 3 and 4 of Chapter 3. In order to prove Proposition 4 of Chapter 3, we establish the following concentration inequality for $\left|\left|\left|\hat{X}_j - X_j^*\right|\right|\right|$, where the bandwidth is chosen as specified in Proposition 4 of Chapter 3.

Proposition 5. *Suppose that Assumptions 9–10 in Chapter 3 and 23–25 hold. Let $\hat{X}_j$ be the local polynomial regression estimator of order $\ell = \lfloor \beta_1 \rfloor$ with bandwidth*

$$h_n \propto n^{-1/(2\beta_1+1)}.$$

There exists a constant $C_2 < \infty$ such that for each $j = 1, \ldots, p$,

$$\left|\left|\left|\hat{X}_j - X_j^*\right|\right|\right|^2 \leq C_2 n^{\alpha - \frac{2\beta_1}{2\beta_1+1}}$$

holds with probability at least $1 - 2\exp\{-n^\alpha/(2\sigma^2 C_3^2)\}$.

The proof of Proposition 5 is similar to that for Theorem 3 in Chapter 3 by plugging in $h_n \propto n^{-1/(2\beta_1+1)}$ in (B.9).

Given Proposition 5, the proof of Proposition 4 in Chapter 3 follows from a similar argument as in the proof of Theorem 4 in Chapter 3, and is thus omitted here. The constant $H_2(\beta_1, \beta_2, \alpha)$ is defined as

$$H_2(\beta_1, \beta_2, \alpha) \equiv \min\left\{\frac{\beta_1}{2\beta_1+1}\frac{2\beta_2-1}{2\beta_2+1} - \alpha, \frac{1}{3}\frac{\beta_1}{2\beta_1+1}\frac{2\beta_2-3}{2\beta_2+1} - \alpha\right\}.$$

B.2 Proofs of Technical Lemmas

B.2.1 Proof of Lemma 7

In this section, in the interest of clarity, we bring back the subscript j in $\theta_j^*, \theta_{jk}^*, \theta_{jS^0}^*$ and f_{jk}^*.

Proof. Recall that in Assumption 26, (B.25) says that

$$\int_0^1 \delta_{jk}^2(u)du = \int_0^1 \left\{f_{jk}^*(X_k^*(u)) - \psi^{\mathrm{T}}(X_k^*(u))\theta_{jk}^*\right\}^2 du \leq Q(M+1)^{-2\beta_2}.$$

It follows from the triangle inequality that

$$\left|\left\{\int_0^1 \left[\psi^{\mathrm{T}}(X_k^*(u))\theta_{jk}^*\right]^2 du\right\}^{1/2} - \left\{\int_0^1 \left[f_{jk}^*(X_k^*(u))\right]^2 du\right\}^{1/2}\right| \leq \sqrt{Q}M^{-\beta_2}.$$

The orthogonality of ψ in Assumption 26 then leads to (B.26), i.e.,

$$\left|\|\theta_{jk}^*\|_2 - \left\{\int_0^1 \left[f_{jk}^*(X_k^*(u))\right]^2 du\right\}^{1/2}\right| \leq \sqrt{Q}M^{-\beta_2}.$$

From (B.25), we can also see that

$$\begin{aligned}\left|\int_0^t \delta_{jk}(u)\,du\right| \leq \left\{\int_0^t \delta_{jk}^2(u)\,du\right\}^{1/2} \left\{\int_0^t 1^2\,du\right\}^{1/2} \leq \left\{\int_0^1 \delta_{jk}^2(u)\,du\right\}^{1/2} \\ \leq \sqrt{Q(M+1)^{-2\beta_2}} \leq \sqrt{QM^{-2\beta_2}},\end{aligned}$$

where we use the fact that $t \in [0,1]$.

Recall from (3.15) in Chapter 3 and (B.10) that

$$X_j^*(t) = \theta_{j0}^* t + \sum_{k=1}^p \Psi_k^{\mathrm{T}}(t)\theta_{jk}^* + \sum_{k=1}^p \int_0^t \delta_{jk}(u)\,du,$$

where we let $X_j^*(0) = 0$ for ease of discussion. We know that both θ_{jk}^* and δ_{jk} are zero for $k \notin S$. Thus, the errors that result from the use of truncated bases are bounded by

$$\begin{aligned}\left|\left|\left|X_j^* - \Psi_{S_j^0}^{\mathrm{T}}\theta_{jS_j^0}^*\right|\right|\right| = \left|\left|\left|X_j^* - \theta_{j0}^* t - \sum_{k\in S_j} \Psi_k^{\mathrm{T}}\theta_{jk}^*\right|\right|\right| = \left[\int_0^1 \left\{\sum_{k\in S_j}\int_0^t \delta_{jk}(u)du\right\}^2 dt\right]^{\frac{1}{2}} \\ \leq \left[\int_0^1 \left\{s\sqrt{QM^{-2\beta_2}}\right\}^2 dt\right]^{\frac{1}{2}} \leq s\sqrt{QM^{-2\beta_2}}.\end{aligned}$$

The error bound in (B.27) is on the whole trajectories, whereas we only observe discrete measurements of the trajectories in reality. The bound in (B.28) addresses this case and is proved below.

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\{X_j^*(t_i)-\Psi_{S_j^0}^{\mathrm{T}}(t_i)\theta_{jS_j^0}^*\}^2 &= \frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{k\in S_j}\int_0^{t_i}\delta_{jk}(u)du\right\}^2\\ &\le \int_0^1\left\{\sum_{k\in S_j}\int_0^{t}\delta_{jk}(u)du\right\}^2 dt + o\left(\frac{1}{n^2}\right)\\ &\le s^2QM^{-2\beta_2}+o\left(n^{-2}\right),\end{aligned}$$

where the last inequality follows from (B.27) and the second to last inequality follows from the trapezoidal rule on a uniform grid.

□

B.2.2 Proof of Lemma 8

We first review some known results on matrix norms and eigenvalues. For an $m\times n$ matrix A,

$$\|A\|_2=\sup_{x\in\mathbb{R}^n}\frac{\|Ax\|_2}{\|x\|_2}=\sup_{\|x\|_2=1}\left\{\sum_{i=1}^{m}\left(\sum_{j=1}^{n}a_{ij}x_j\right)^2\right\}^{\frac{1}{2}}\le\left(\sum_{i=1}^{m}\sum_{j=1}^{n}a_{ij}^2\right)^{\frac{1}{2}}\equiv\|A\|_F, \tag{B.37}$$

where $\|\cdot\|_F$ is the Frobenius norm. We remind the reader that for a symmetric matrix A that is not positive semi-definite, $\Lambda_{\max}(A)\le\|A\|_2$. The following two inequalities are useful in the proofs. Let A and $\hat{A}$ be two $n\times n$ symmetric matrices.

1. Weyl's inequality (Weyl, 1912) states that

$$\Lambda_{\min}(A)-\Lambda_{\max}(\hat{A}-A)\le\Lambda_{\min}(\hat{A}),\ \text{and}\ \Lambda_{\max}(\hat{A})\le\Lambda_{\max}(A)+\Lambda_{\max}(\hat{A}-A),$$

   which leads to

$$\Lambda_{\min}(A)-\|\hat{A}-A\|_2\le\Lambda_{\min}(\hat{A}),\ \text{and}\ \Lambda_{\max}(\hat{A})\le\Lambda_{\max}(A)+\|\hat{A}-A\|_2. \tag{B.38}$$

2. The Gershgorin circle theorem (Gershgorin, 1931) states that

$$\|\hat{A} - A\|_2 \leq \max_i \sum_{j=1}^{n} |(\hat{A} - A)_{ij}| \leq n\|\hat{A} - A\|_\infty, \tag{B.39}$$

where the norm $\|\cdot\|_\infty$ is defined as $\|A\|_\infty = \max_{i,j} |A_{ij}|$.

We are now ready to prove Lemma 8.

Proof. Let $A \equiv \int_0^1 \Psi_{S^0}(t)\Psi_{S^0}^T(t)\,dt$, $A_n \equiv \frac{1}{n}\sum_{i=1}^n \Psi_{S^0}(t_i)\Psi_{S^0}^T(t_i)$, $\hat{A}_n \equiv \frac{1}{n}\sum_{i=1}^n \hat{\Psi}_{S^0}(t_i)\hat{\Psi}_{S^0}^T(t_i)$, which are $(Ms+1)\times(Ms+1)$ matrices. Then,

$$\begin{aligned}\Lambda_{\min}(\hat{A}_n) \geq& \Lambda_{\min}(A) - \|\hat{A}_n - A\|_2 \\ \geq& \Lambda_{\min}(A) - \|A_n - A\|_2 - \|\hat{A}_n - A_n\|_2,\end{aligned} \tag{B.40}$$

where the first inequality follows from (B.38) and the second follows from the triangle inequality.

Furthermore,

$$\begin{aligned}\|\hat{A}_n - A_n\|_2 \leq& (Ms+1)\|\hat{A}_n - A_n\|_\infty \\ \leq& (Ms+1)\left\|\frac{1}{n}\sum_{i=1}^n \left\{\hat{\Psi}_{S^0}(t_i)\hat{\Psi}_{S^0}^{\mathrm{T}}(t_i) - \Psi_{S^0}(t_i)\Psi_{S^0}^{\mathrm{T}}(t_i)\right\}\right\|_\infty \\ \leq& \frac{Ms+1}{n}\left\|\sum_{i=1}^n \hat{\Psi}_{S^0}(t_i)\left\{\hat{\Psi}_{S^0}^{\mathrm{T}}(t_i) - \Psi_{S^0}^{\mathrm{T}}(t_i)\right\}\right\|_\infty + \\ & \frac{Ms+1}{n}\left\|\sum_{i=1}^n \Psi_{S^0}(t_i)\left\{\hat{\Psi}_{S^0}^{\mathrm{T}}(t_i) - \Psi_{S^0}^{\mathrm{T}}(t_i)\right\}\right\|_\infty \\ \leq& \frac{Ms+1}{n}\left\|\sum_{i=1}^n \hat{\Psi}_{S^0}(t_i)D\Delta\right\|_\infty + \frac{Ms+1}{n}\left\|\sum_{i=1}^n \Psi_{S^0}(t_i)D\Delta\right\|_\infty \\ \leq& \frac{2Ms+2}{n}\|nBD\Delta\|_\infty = 2(Ms+1)BD\Delta,\end{aligned} \tag{B.41}$$

where the first inequality follows from (B.39), the last inequality follows from the bounds in Assumption 26, and the second to last inequality follows from the following inequality: for $k \in S^0$

and $m = 1, \ldots, M$,

$$\begin{aligned}
|\hat{\Psi}_{km}(t_i) - \Psi_{km}(t_i)| &= \left| \int_0^{t_i} \psi_m(\hat{X}_k(u))\, du - \int_0^{t_i} \psi_m(X_k^*(u))\, du \right| \\
&= \left| \int_0^{t_i} \{\psi_m(\hat{X}_k(u)) - \psi_m(X_k^*(u))\}\, du \right| \\
&\leq \left| \int_0^{t_i} |D\{\hat{X}_k(u) - X_k^*(u)\}|\, du \right| \\
&\leq \left\{ \int_0^{t_i} D^2\, du \right\}^{1/2} \left\{ \int_0^{t_i} (\hat{X}_k(u) - X_k^*(u))^2\, du \right\}^{1/2} \\
&\leq D \left\| \hat{X}_k - X_k^* \right\| \leq D\Delta.
\end{aligned} \tag{B.42}$$

Here the first inequality follows from the mean-value theorem and the bounds in Assumption 26.

Now, from (B.39),

$$\|A_n - A\|_2 \leq (Ms+1)\|A_n - A\|_\infty \leq (Ms+1)\frac{BD + B^2}{6n^2}, \tag{B.43}$$

where for each element of the matrix $A_n - A = \frac{1}{n}\sum_{i=1}^n \Psi_{S^0}(t_i)\Psi_{S^0}^{\mathrm{T}}(t_i) - \int_0^1 \Psi_{S^0}(t)\Psi_{S^0}^{\mathrm{T}}(t)\, dt$,

$$\begin{aligned}
&\left| \frac{1}{n}\sum_{i=1}^n \Psi_{km_1}(t_i)\Psi_{lm_2}(t_i) - \int_0^1 \Psi_{km_1}(t)\Psi_{lm_2}(t)\, dt \right| \\
\leq &\frac{\left| \{\Psi_{km_1}(u)\Psi_{lm_2}(u)\}'' \right|}{12n^2} \leq \frac{\left| 2\Psi'_{km_1}(u)\Psi'_{lm_2}(u) + \Psi''_{km_1}(u)\Psi_{lm_2}(u) + \Psi'_{km_1}(u)\Psi''_{lm_2}(u) \right|}{12n^2} \\
\leq &\frac{2B^2 + BD + BD}{12n^2} = \frac{BD + B^2}{6n^2},
\end{aligned}$$

where derivatives are taken with respect to t. By the trapezoid rule on a uniform grid, the first inequality holds for some $u \in [0, 1]$. The second inequality makes use of the bounds in Assumption 26, which imply that

$$|\Psi'_{km}(t)| = \left| \left(\int_0^t \psi_{km}(s)\, ds \right)' \right| = |\psi_{km}(t)| \leq B$$

and

$$|\Psi''_{km}(t)| = \left| \left(\int_0^t \psi_{km}(s)\, ds \right)'' \right| = |\psi'_{km}(t)| \leq D.$$

In summary, combining (B.40), (B.41), and (B.43),

$$\begin{aligned}\Lambda_{\min}(\hat{A}_n) \geq & \Lambda_{\min}(A) - \left(2BD\Delta + \frac{BD+B^2}{6n^2}\right)(Ms+1) \\ \geq & C_{\min} - \left(2BD\Delta + \frac{BD+B^2}{6n^2}\right)(Ms+1).\end{aligned}$$

The upper bound for $\Lambda_{\max}(\hat{A}_n)$ and the lower bound for $\Lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\Psi}_k(t_i)\hat{\Psi}_k^{\mathrm{T}}(t_i)\right)$ can be established in a similar manner. □

B.2.3 Proof of Lemma 9

Proof. Define A, A_n, and $\hat{A}_n$ as in the proof for Lemma 8. We let $F = \int_0^1 \Psi_k \Psi_{S^0}^{\mathrm{T}}\, dt$, $F_n = \sum_{i=1}^{n}\Psi_k(t_i)\Psi_{S^0}^{\mathrm{T}}(t_i)/n$, and $\hat{F}_n = \sum_{i=1}^{n}\hat{\Psi}_k(t_i)\hat{\Psi}_{S^0}^{\mathrm{T}}(t_i)/n$. F, F_n, and $\hat{F}_n$ are $M \times (Ms+1)$ matrices. We let $\hat{C}_{\min}$ denote the lower bound of $\Lambda_{\min}(\hat{A}_n)$ established in Lemma 8, i.e.,

$$\hat{C}_{\min} \equiv C_{\min} - \left(2BD\Delta + \frac{BD+B^2}{6n^2}\right)(Ms+1).$$

To prove the result, we need to bound $\|\hat{F}_n\hat{A}_n^{-1}\|_2$. Note that

$$\begin{aligned}\|\hat{F}_n\hat{A}_n^{-1}\|_2 \leq & \|\hat{F}_n\hat{A}_n^{-1} - \hat{F}_nA_n^{-1} + \hat{F}_nA_n^{-1} - F_nA_n^{-1} + F_nA_n^{-1}\|_2 \\ \leq & \|\hat{F}_n(\hat{A}_n^{-1} - A_n^{-1})\|_2 + \|(\hat{F}_n - F_n)A_n^{-1}\|_2 + \|F_nA_n^{-1}\|_2 \\ \equiv & \|\mathrm{I}\|_2 + \|\mathrm{II}\|_2 + \|\mathrm{III}\|_2.\end{aligned}$$

Using sub-multiplicity of the ℓ_2-norm of matrices,

$$\|\mathrm{I}\|_2^2 \leq \|\hat{F}_n\|_2^2\|\hat{A}_n^{-1} - A_n^{-1}\|_2^2.$$

Applying (B.37) to $\hat{F}_n$, we get

$$\|\mathrm{I}\|_2^2 \leq M(Ms+1)\left(\max_{i,j}\hat{F}_{n,ij}^2\right)\|\hat{A}_n^{-1} - A_n^{-1}\|_2^2.$$

Recalling that $\hat{F}_n = \sum_{i=1}^{n}\hat{\Psi}_k(t_i)\hat{\Psi}_{S^0}^{\mathrm{T}}(t_i)/n$ and that $|\hat{\Psi}_{km}(t_i)| \leq B$,

$$\|\mathrm{I}\|_2^2 \leq M(Ms+1)(\sum_{i=1}^{n}B^2/n)^2\|\hat{A}_n^{-1} - A_n^{-1}\|_2^2.$$

Note that $\hat{A}_n^{-1} - A_n^{-1} = \hat{A}_n^{-1}(A_n - \hat{A}_n)A_n^{-1}$. Thus,

$$\begin{aligned}
\|\mathrm{I}\|_2^2 \leq & M(Ms+1)B^4\|\hat{A}_n^{-1}\|_2^2\|\hat{A}_n - A_n\|_2^2\|A_n^{-1}\|_2^2 \\
\leq & M(Ms+1)B^4\hat{C}_{\min}^{-2}\|\hat{A}_n - A_n\|_2^2 C_{\min}^{-2} \\
\leq & M(Ms+1)B^4\{2(Ms+1)DB\Delta\}^2\hat{C}_{\min}^{-2}C_{\min}^{-2}, \\
\equiv & c_1\hat{C}_{\min}^{-2}M(Ms+1)^3\Delta^2,
\end{aligned}$$

where the last two inequalities follow from the proof of Lemma 8.

Next, note that

$$\begin{aligned}
\|\mathrm{II}\|_2^2 = & \|(\hat{F}_n - F_n)A_n^{-1}\|_2^2 \\
\leq & C_{\min}^{-2}\left\|\frac{1}{n}\sum_{i=1}^n \hat{\Psi}_k(t_i)\hat{\Psi}_{S^0}^{\mathrm{T}}(t_i) - \frac{1}{n}\sum_{i=1}^n \Psi_k(t_i)\Psi_{S^0}^{\mathrm{T}}(t_i)\right\|_2^2 \\
\leq & C_{\min}^{-2}\left\|\frac{1}{n}\sum_{i=1}^n \hat{\Psi}_k(t_i)\left\{\hat{\Psi}_{S^0}^{\mathrm{T}}(t_i) - \Psi_{S^0}^{\mathrm{T}}(t_i)\right\}\right\|_2^2 + \\
& C_{\min}^{-2}\left\|\frac{1}{n}\sum_{i=1}^n \left\{\hat{\Psi}_k(t_i) - \Psi_k(t_i)\right\}\Psi_{S^0}^{\mathrm{T}}(t_i)\right\|_2^2 \\
\leq & 2C_{\min}^{-2}B^2D^2\Delta^2M(Ms+1) \\
\equiv & c_2M(Ms+1)\Delta^2,
\end{aligned}$$

where the first inequality follows from sub-multiplicity of norms of matrices, and the last from (B.37), (B.42), and the bounds in Assumption 26.

Finally,

$$
\begin{aligned}
\|\mathrm{III}\|_2 =& \|F_n A_n^{-1}\|_2 = \|F_n(A_n^{-1} - A^{-1}) + (F_n - F)A^{-1} + FA^{-1}\|_2 \\
\leq& \xi + \|F_n\|_2 \|A_n^{-1} - A^{-1}\|_2 + \|(F_n - F)A^{-1}\|_2 \\
\leq& \xi + \left\{M(Ms+1)B^4 \|A_n^{-1} - A^{-1}\|_2^2\right\}^{1/2} + \{\|F_n - F\|_2^2 C_{\min}^{-2}\}^{1/2} \\
\leq& \xi + \left\{M(Ms+1)B^4 \|A_n^{-1}\|_2^2 \|A_n - A\|_2^2 \|A^{-1}\|_2^2\right\}^{1/2} + \{\|F_n - F\|_2^2 C_{\min}^{-2}\}^{1/2} \\
\leq& \xi + \left\{M(Ms+1)B^4 \hat{C}_{\min}^{-2} C_{\min}^{-2} \|A_n - A\|_2^2\right\}^{1/2} + \{\|F_n - F\|_2^2 C_{\min}^{-2}\}^{1/2} \\
\leq& \xi + \left\{M(Ms+1)B^4 \hat{C}_{\min}^{-2} C_{\min}^{-2} (Ms+1)^2 \frac{BD + B^2}{6n^2}\right\}^{1/2} + \\
& \left\{M(Ms+1) C_{\min}^{-2} \frac{BD + B^2}{6n^2}\right\}^{1/2} \\
\leq& \xi + \left\{c_3 M(Ms+1)^3/6n^2\right\}^{1/2},
\end{aligned}
$$

where the first inequality follows from Assumption 13 in Chapter 3 and the second to last inequality follows from (B.43).

In summary,

$$
\begin{aligned}
&\left\| \frac{1}{n} \sum_{i=1}^{n} \hat{\Psi}_k(t_i) \hat{\Psi}_{S^0}^{\mathrm{T}}(t_i) \left(\frac{1}{n} \sum_{i=1}^{n} \hat{\Psi}_{S^0}(t_i) \hat{\Psi}_{S^0}^{\mathrm{T}}(t_i) \right)^{-1} \right\|_2 \leq \\
&\xi + \left\{c_1 M(Ms+1)^3 \Delta^2\right\}^{1/2} + \left\{c_2 M(Ms+1)\Delta^2\right\}^{1/2} + \left\{c_3 M(Ms+1)^3/6n^2\right\}^{1/2}.
\end{aligned}
$$

where c_1, c_2, c_3 are constants. □

B.2.4 Proof of Lemma 10

Proof. For $k = 1, \ldots, p$,

$$
\begin{aligned}
&\left\| \frac{1}{n} \sum_{i=1}^{n} \hat{\Psi}_k(t_i) Y_{ij} - \frac{1}{n} \sum_{i=1}^{n} \hat{\Psi}_k(t_i) \hat{\Psi}_{S^0}^{\mathrm{T}}(t_i) \theta_{S^0}^* \right\|_2 \\
=&\frac{1}{n} \left\| \sum_{i=1}^{n} \hat{\Psi}_k(t_i) X_j^*(t_i) + \sum_{i=1}^{n} \hat{\Psi}_k(t_i) \epsilon_{ji} - \sum_{i=1}^{n} \hat{\Psi}_k(t_i) \Psi_{S^0}^{\mathrm{T}}(t_i) \theta_{S^0}^* + \right. \\
&\left. \sum_{i=1}^{n} \hat{\Psi}_k(t_i) \Psi_{S^0}^{\mathrm{T}}(t_i) \theta_{S^0}^* - \sum_{i=1}^{n} \hat{\Psi}_k(t_i) \hat{\Psi}_{S^0}^{\mathrm{T}}(t_i) \theta_{S^0}^* \right\|_2 \\
\leq &\left\| \frac{1}{n} \sum_{i=1}^{n} \hat{\Psi}_k(t_i) \{ X_j^*(t_i) - \Psi_{S^0}^{\mathrm{T}}(t_i) \theta_{S^0}^* \} \right\|_2 + \left\| \frac{1}{n} \sum_{i=1}^{n} \hat{\Psi}_k(t_i) \{ \Psi_{S^0}^{\mathrm{T}}(t_i) - \hat{\Psi}_{S^0}^{\mathrm{T}}(t_i) \} \theta_{S^0}^* \right\|_2 + \\
&\left\| \frac{1}{n} \sum_{i=1}^{n} \hat{\Psi}_k(t_i) \epsilon_{ji} \right\|_2 \\
\equiv &\|\mathrm{I}\|_2 + \|\mathrm{II}\|_2 + \|\mathrm{III}\|_2.
\end{aligned}
$$

First, applying the Cauchy-Schwarz inequality to $\|\mathrm{I}\|_2^2$,

$$
\|\mathrm{I}\|_2^2 \leq \sum_{m=1}^{M} \left[\frac{1}{n^2} \sum_{i=1}^{n} \hat{\Psi}_{km}^2(t_i) \sum_{i=1}^{n} \left\{ X_j^*(t_i) - \Psi_{S^0}^{\mathrm{T}}(t_i) \theta_{S^0}^* \right\}^2 \right].
$$

From the bounds in Assumption 26 and (B.28) ,

$$
\|\mathrm{I}\|_2^2 \leq M \left\{ \frac{1}{n^2} \left(nB^2 \right) \left(s^2 n Q M^{-2\beta_2} \right) \right\} = s^2 M^{-2\beta_2 + 1} Q B^2.
$$

Next, note that $\hat{\Psi}_0(t_i) - \Psi_0(t_i) = t_i - t_i = 0$, we have $\left\{ \Psi_{S^0}^{\mathrm{T}}(t_i) - \hat{\Psi}_{S^0}^{\mathrm{T}}(t_i) \right\} \theta_{S^0}^* = \left\{ \Psi_S(t_i) - \hat{\Psi}_S(t_i) \right\}^{\mathrm{T}} \theta_S^*$. Thus applying the Cauchy-Schwarz inequality to $\|\mathrm{II}\|_2^2$,

$$
\|\mathrm{II}\|_2^2 \leq \sum_{m=1}^{M} \left(\frac{1}{n^2} \sum_{i=1}^{n} \hat{\Psi}_{km}^2(t_i) \sum_{i=1}^{n} \left[\left\{ \Psi_S(t_i) - \hat{\Psi}_S(t_i) \right\}^{\mathrm{T}} \theta_S^* \right]^2 \right).
$$

Applying the norm inequality $a^{\mathrm{T}} b \leq \|a\|_\infty \|b\|_1$ to $\left\{ \Psi_S(t_i) - \hat{\Psi}_S(t_i) \right\}^{\mathrm{T}} \theta_S^*$ and using the inequality (B.42) as well as the bounds in Assumption 26, we get

$$
\|\mathrm{II}\|_2^2 \leq M \left\{ \frac{1}{n^2} n B^2 \sum_{i=1}^{n} \|\theta_S^*\|_1^2 D^2 \Delta^2 \right\} \leq M B^2 D^2 \|\theta_S^*\|_1^2 \Delta^2.
$$

Finally, III $= \frac{1}{n}\sum_{i=1}^{n}\hat{\Psi}_k(t_i)\epsilon_{ji}$ is an M-vector. For each $m = 1, \ldots, M$, we let $g(\epsilon_j/\sigma) = \sum_{i=1}^{n}\hat{\Psi}_{km}(t_i)\epsilon_{ji}/n$. Then, for $a, b \in \mathbb{R}^p$,

$$\begin{aligned}|g(a) - g(b)| =& \left|\sigma\sum_{i=1}^{n}\hat{\Psi}_{km}(t_i)(a_i - b_i)/n\right| \\ \leq& \frac{\sigma}{n}\left\{\sum_{i=1}^{n}\hat{\Psi}_{km}^2(t_i)\right\}^{0.5}\|a-b\|_2 \leq \frac{\sigma}{n}\sqrt{nB^2}\|a-b\|_2.\end{aligned}$$

This shows that $g(\cdot)$ is an L_3-Lipshitz function with $L_3 = \sigma B/\sqrt{n}$. Note that $\mathbb{E}g(\epsilon_j/\sigma) = 0$. Thus, by Theorem 5.6 in Boucheron et al. (2013) presented in Section B.1.2, we have

$$\Pr(|g(\epsilon_j/\sigma)| \geq v) \leq 2\exp\{-v^2 n/(2B^2\sigma^2)\}.$$

Letting $v = n^{\alpha/2-0.5}$, $\|\text{III}\|_2^2 \leq n^{\alpha-1}M$ holds with probability at least $1-2M\exp\{-n^\alpha/(2B^2\sigma^2)\}$.

Combining all of the pieces, we find that

$$\|\text{I}\|_2 + \|\text{II}\|_2 + \|\text{III}\|_3 \leq \eta \equiv M^{1/2}\left\{sM^{-\beta_1}Q^{1/2}B + BD\|\theta_S^*\|_1\Delta + n^{\frac{\alpha}{2}-\frac{1}{2}}\right\}$$

with probability at least $1 - 2M\exp\{-n^\alpha/(2B^2\sigma^2)\}$.

For $k = 0$,

$$\begin{aligned}&\left\|\frac{1}{n}\sum_{i=1}^{n}\hat{\Psi}_0(t_i)\left\{Y_{ij} - \hat{\Psi}_{S^0}^{\mathrm{T}}(t_i)\theta_{S^0}^*\right\}\right\|_2 = \left\|\frac{1}{n}\sum_{i=1}^{n}t_i\left\{Y_{ij} - \hat{\Psi}_{S^0}^{\mathrm{T}}(t_i)\theta_{S^0}^*\right\}\right\|_2 \\ \leq& \left\|\frac{1}{n}\sum_{i=1}^{n}t_i\{X_j^*(t_i) - \Psi_{S^0}^{\mathrm{T}}(t_i)\theta_{S^0}^*\}\right\|_2 + \left\|\frac{1}{n}\sum_{i=1}^{n}t_i\{\Psi_{S^0}^{\mathrm{T}}(t_i) - \hat{\Psi}_{S^0}^{\mathrm{T}}(t_i)\}\theta_{S^0}^*\right\|_2 + \\ &\left\|\frac{1}{n}\sum_{i=1}^{n}t_i\epsilon_{ji}\right\|_2.\end{aligned}$$

Recall that $t \in [0,1]$ and, without loss of generality, let $B \geq 1$. Thus, we can see from the same argument that $\left\|\frac{1}{n}\sum_{i=1}^{n}t_i\left\{Y_{ij} - \hat{\Psi}_{S^0}^{\mathrm{T}}(t_i)\theta_{S^0}^*\right\}\right\|_2 \leq \eta$ holds with the same probability.

□

B.3 Details About Data Generation

In this section, we provide details about the parameters used for generating data in Section 3.5.1 of Chapter 3 (see Equation 3.26). Three pairs of variables, $(X_1, X_2), (X_3, X_4), (X_5, X_6)$, are solutions of (3.26) in Chapter 3 with the following parameters and initial values:

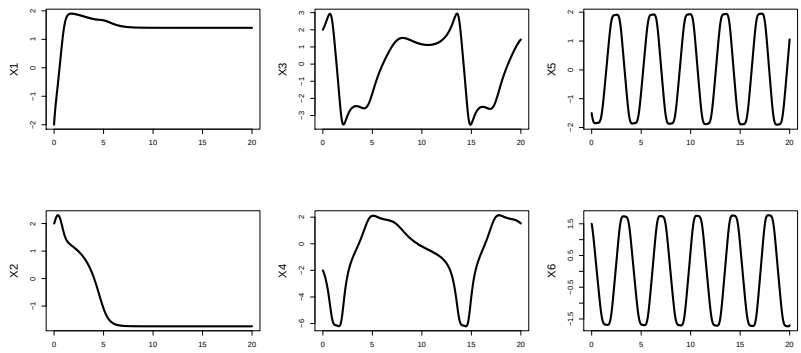

Figure B.1: The curves $X_1, \ldots, X_6$ on $[0, 20]$ described in Section 3.5.1 of Chapter 3 and Section B.3 of the supplementary material.

1. (X_1, X_2) are generated according to (3.26) from Chapter 3 with $\theta_{1,0} = 0$, $\theta_{1,1} = (1.2, 0.3, -0.6)^{\mathrm{T}}$, $\theta_{1,2} = (0.1, 0.2, 0.2)^{\mathrm{T}}$, $\theta_{2,0} = 0.4$, $\theta_{2,1} = (-2, 0, 0.4)^{\mathrm{T}}$, $\theta_{2,2} = (0.5, 0.2, -0.3)^{\mathrm{T}}$, and initial values $X_1(0) = -2, X_2(0) = 2$.

2. (X_3, X_4) are generated according to (3.26) from Chapter 3 with $\theta_{3,0} = -0.2, \theta_{3,3} = (0, 0, 0)^{\mathrm{T}}, \theta_{3,4} = (-0.3, 0.4, 0.1)^{\mathrm{T}}, \theta_{4,0} = -0.2, \theta_{4,3} = (0.2, -0.1, -0.2)^{\mathrm{T}}, \theta_{4,4} = (0, 0, 0)^{\mathrm{T}}$, and initial values $X_3(0) = 2, X_4(0) = -2$.

3. (X_5, X_6) are generated according to (3.26) from Chapter 3 with $\theta_{5,0} = 0.05$, $\theta_{5,5} = (0, 0, 0)^{\mathrm{T}}$, $\theta_{5,6} = (0.1, 0, -0.8)^{\mathrm{T}}$, $\theta_{6,0} = -0.05$, $\theta_{6,5} = (0, 0, 0.5)^{\mathrm{T}}$, $\theta_{6,6} = (0, 0, 0)^{\mathrm{T}}$, and initial values $X_5(0) = -1.5, X_6(0) = 1.5$.

Solution trajectories of $X_1, \ldots, X_6$ are shown in Figure B.1. For $X_7, \ldots, X_{10}$, we drew the initial values $X_j(0), j = 7, \ldots, 10$, and the $\theta_{j,0}, j = 7, \ldots, 10$, from a normal distribution. All other parameters were set to zero, so that $X_7, \ldots, X_{10}$ represent "noise" variables. The directed graph of $X_1, \ldots, X_{10}$ is showing in Figure B.2.

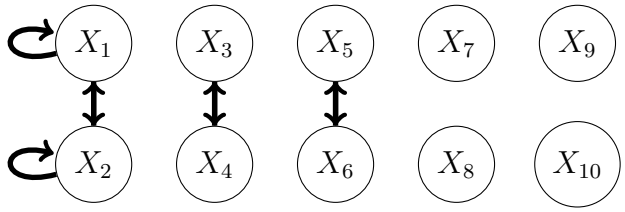

Figure B.2: The network of $\{X_1, \ldots, X_{10}\}$. A directed edge $j \rightarrow k$ indicates that the jth node regulates the kth node.

Appendix C

APPENDIX FOR CHAPTER 4

C.1 Algorithm

C.1.1 Overview

Recall that in Chapter 4, the loss function (4.12) takes the form

$$-\frac{1}{T}\int_0^T \left[\mu_j + \sum_{k=1}^{p} \Psi_k(t)\cdot\beta_{j,k}\right] dN_j(t) + \frac{1}{2T}\int_0^T \left[\mu_j + \sum_{k=1}^{p} \Psi_k(t)\cdot\beta_{j,k}\right]^2 dt$$
$$+ \eta_j \sum_{k=1}^{p} \left\|\Psi_k(t)\cdot\beta_{j,k}\right\|_{2,[0,T]},$$

where $\Psi_k(t) \equiv \big(\psi * dN_k\big)(t)$ for $k = 1, \ldots, p$. Recall also that

$$\frac{1}{T}\int_0^T \left[\mu_j + \sum_{k=1}^{p} \Psi_k(t)\cdot\beta_{j,k}\right] dN_j(t) = \frac{1}{T}\sum_{i=1}^{n_j} \left[\mu_j + \sum_{k=1}^{p} \Psi_k(t_{j,i})\cdot\beta_{j,k}\right],$$

where $n_j \equiv N_j(T)$ is the number of evens of N_j in $[0, T]$. This loosely resembles the loss function for linear regression with a standardized group lasso penalty (see e.g., Simon and Tibshirani (2012)), in which $dN_j(t)$ is the response, and $\Big(\Psi_1(t) \quad \ldots \quad \Psi_K(t)\Big)$ is the design matrix. However, in order to minimize (4.12) using standard software for solving the standardized group lasso, such as in `grpreg` (Breheny and Huang, 2015), `grplasso` (Meier et al., 2008), or `gglasso` (Yang and Zou, 2015), we need the response and the design matrix to be defined at discrete time points rather than in continuous time. There are two obvious options for how we might do this.

1. We could discretize the response $dN_j(t)$ and the design matrix $\Big(\Psi_1(t) \quad \ldots \quad \Psi_K(t)\Big)$ on an evenly-spaced grid. We would need to use a very fine grid, in order to ensure that there is no loss of information in discretizing the response. However, using such a fine grid would be computationally burdensome.

2. We could discretize $dN_j(t)$ and $\begin{pmatrix}\Psi_1(t) & \dots & \Psi_K(t)\end{pmatrix}$ on a grid that is based on the observed event times. In this case, the grid would be irregular and stochastic, which might cause trouble assessing the precision in estimation of the design matrix.

To avoid the need to discretize the integrals in (4.12), we directly solve (4.12) using a block coordinate descent algorithm detailed in the following two sections.

C.1.2 Change of Variables

Although we can choose $\psi(t) = \begin{pmatrix}\psi_1(t) & \dots & \psi_M(t)\end{pmatrix}^T$ to be an orthogonal set of basis functions, the design matrix $\Psi_k(t) \equiv \big(\psi * dN_k\big)(t)$ in (4.12) is not necessarily orthogonal. Therefore, we define the symmetric matrix square root

$$U_k \equiv \left(T^{-1}\int_0^T \Psi_k(t)\Psi_k^{\mathrm{T}}(t)dt\right)^{1/2},$$

and let $\theta_{j,k} = U_k\beta_{j,k}$ and thus $\beta_{j,k} = U_k^{-1}\theta_{j,k}$. We can then rewrite (4.12) as

$$\begin{aligned}\underset{\mu_j\in\mathbb{R},\theta_{j,k}\in\mathbb{R}^M}{\arg\min} &-\frac{1}{T}\int_0^T\left[\mu_j+\sum_{k=1}^p\Psi_k^{\mathrm{T}}(t)U_k^{-1}\theta_{j,k}\right]dN_j(t)+\frac{1}{2T}\int_0^T\left[\mu_j+\sum_{k=1}^p\Psi_k^{\mathrm{T}}(t)U_k^{-1}\theta_{j,k}\right]^2dt\\ &+\eta_j\sum_{k=1}^p\big\|\theta_{j,k}\big\|_2,\end{aligned} \tag{C.1}$$

where we have used the fact that

$$\begin{aligned}\big\|\Psi_k\cdot\beta_{j,k}\big\|_{2,[0,T]}^2 =&\beta_{j,k}^{\mathrm{T}}\left[\frac{1}{T}\int_0^T\Psi_k(t)\Psi_k^{\mathrm{T}}(t)\,dt\right]\beta_{j,k}\\ =&\theta_{j,k}^{\mathrm{T}}U_k^{-1}\left[\frac{1}{T}\int_0^T\Psi_k(t)\Psi_k^{\mathrm{T}}(t)\,dt\right]U_k^{-1}\theta_{j,k}=\big\|\theta_{j,k}\big\|_2^2.\end{aligned}$$

Suppose that $\{\hat{\mu}_j,\ \hat{\theta}_{j,k}, k = 1,\dots,p\}$ is the solution to (C.1). It then follows that $\{\hat{\mu}_j, \hat{\beta}_{j,k} = U_k^{-1}\hat{\theta}_{j,k}, k = 1,\dots,p\}$ is a solution to (4.12).

C.1.3 Block Coordinate Descent Algorithm

We now derive a coordinate descent algorithm to solve (C.1). In the algorithm, we will update one parameter at a time while keeping the rest fixed.

To begin, consider updating μ_j while keeping $\theta_{j,1}, \ldots, \theta_{j,p}$ fixed. After removing terms that do not involve μ_j, (C.1) can be written as

$$\arg\min_{\mu_j \in \mathbb{R}} -\frac{n_j}{T}\mu_j + \frac{1}{2}\mu_j^2 + \frac{1}{T}\mu_j \sum_{l=1}^{p} \left[\int_0^T \Psi_l(t)\, dt\right]^{\mathrm{T}} U_l^{-1}\theta_{j,l}.$$

The solution to this problem is

$$\mu_j^{\text{new}} = \frac{1}{T}n_j - \frac{1}{T}\sum_{l=1}^{p} \left[\int_0^T \Psi_l(t)\, dt\right]^{\mathrm{T}} U_l^{-1}\theta_{j,l}. \tag{C.2}$$

Consider an update for $\theta_{j,k}$ while keeping $\mu_j, \theta_{j,l}, l \neq k$ fixed. After removing terms that do not involve $\theta_{j,k}$, (C.1) can be written as

$$\arg\min_{\theta_{j,k} \in \mathbb{R}^M} \frac{1}{T}\theta_{j,k}^{\mathrm{T}} U_k^{-1} \left\{\mu_j \int_0^T \Psi_k(t)\, dt + \sum_{l \neq k}\left[\int_0^T \Psi_k(t)\Psi_l^{\mathrm{T}}(t)dt\right] U_l^{-1}\theta_{j,l} - \sum_{i=1}^{n_j}\Psi_k(t_{j,i})\right\}$$
$$+ \frac{1}{2}\left\|\theta_{j,k}\right\|_2^2 + \eta_j\left\|\theta_{j,k}\right\|_2.$$

The solution to this problem is

$$\theta_{j,k}^{\text{new}} = \begin{cases} r_k\big(\|r_k\|_2 - \eta_j\big)/\|r_k\|_2 & \text{if } \|r_k\|_2 \geq \eta_j \\ 0 & \text{if } \|r_k\|_2 < \eta_j \end{cases},$$

where

$$r_k = \frac{1}{T}U_k^{-1}\left\{\sum_{i=1}^{n_j}\Psi_k(t_{j,i}) - \mu_j\int_0^T \Psi_k(t)\, dt - \sum_{l \neq k}\left[\int_0^T \Psi_k(t)\Psi_l^{\mathrm{T}}(t)dt\right] U_l^{-1}\theta_{j,l}\right\}. \tag{C.3}$$

The algorithm for solving (C.1) is provided in Algorithm 1. In other to carry out Algorithm 1, we need to compute the integrals in (C.2) and (C.3). These integrals can be pre-computed before beginning the algorithm, using any numerical integration methods that reach a desired precision. In our numerical experiments, we compute the integrals by summing over an evenly-spaced grid in which the distance between two adjacent points is 0.01.

C.2 Proofs of Results in Section 4.3

In this section, we present the derivations behind Remark 6 in Chapter 4, and prove Theorems 5 and 6 in Chapter 4.

Algorithm 1 Coordinate descent algorithm for solving (4.12) in Chapter 4

Input: the point process $\boldsymbol{N}(\cdot)$, $\eta_j \in \mathbb{R}^+$, $\tau \in \mathbb{R}^+$

Initialization: $\theta_{j,k}^{\text{new}} = 0 \in \mathbb{R}^M$, $k = 1, \ldots, p$, $\mu_j^{\text{new}} \in \mathbb{R}$, $\delta = 10$

while $\delta < \tau$ **do**

$\mu_j^{\text{old}} \leftarrow \mu_j^{\text{new}}$, $\theta_{j,k}^{\text{old}} \leftarrow \theta_{j,k}^{\text{new}}$, $k = 1, \ldots, p$

$$\mu_j^{\text{new}} \leftarrow n_j/T - \tfrac{1}{T}\textstyle\sum_{l=1}^{p} \left(\int_0^T \Psi_l(t)dt\right)^{\mathrm{T}} U_l^{-1}\theta_{j,l}^{\text{new}}$$

For $k = 1, \ldots, p$, **do**

$$r_k \leftarrow T^{-1}U_k^{-1}\left\{\sum_{i=1}^{n_j}\Psi_k(t_{j,i}) - \mu_j^{\text{new}}\int_0^T \Psi_k(t)\,dt - \sum_{l\neq k}\left[\int_0^T \Psi_k(t)\Psi_l^{\mathrm{T}}(t)dt\right]U_l^{-1}\theta_{j,l}^{\text{new}}\right\}$$

$$\theta_{j,k}^{\text{new}} \leftarrow \begin{cases} r_k\big(\|r_k\|_2 - \eta_j\big)/\|r_k\|_2 & \|r_k\|_2 \geq \eta_j \\ 0 & \|r_k\|_2 < \eta_j \end{cases}$$

$$\delta \leftarrow \big[(\mu_j^{\text{new}} - \mu_j^{\text{old}})^2 + \textstyle\sum_{k=1}^{p}\|\theta_{j,k}^{\text{new}} - \theta_{j,k}^{\text{old}}\|_2^2\big] / \big[\big(\mu_j^{\text{new}}\big)^2 + \textstyle\sum_{k=1}^{p}\|\theta_{j,k}^{\text{new}}\|_2^2 + 1\big]$$

Output: $\hat{\mu}_j = \mu_j^{\text{new}}$, $\hat{\beta}_{j,k} = U_k^{-1}\theta_{j,k}^{\text{new}}$ for $k = 1, \ldots, p$

The derivation of Equation 4.16 in Remark 6 is provided in Appendix C.2.1. It follows from a careful examination of Stieljies integrals and Dirac delta functions.

The proofs of Theorems 5 and 6 are shown in Appendices C.2.2–C.2.4. The proofs for the oracle inequality and variable selection consistency in high-dimensional regression have become quite standard following Wainwright (2009) and van de Geer and Bhlmann (2009). We here provide a brief sketch of the proof, with details to follow in the corresponding sections. First, we will establish that required conditions on the observed event times hold with high probability under the population assumptions in Section 4.3 in Chapter 4 (Appendix C.2.2). Then, we show that Theorem 5 holds with high probability given the compatibility condition (Assumption 17) and other regularity assumptions (Appendix C.2.3). Finally, with the additional irrepresentability (Assump-

tion 20) and Beta-min (Assumption 21) conditions, we can show that, with high probability, the penalized regression recovers the true graph as claimed in Theorem 6 (Appendix C.2.4).

C.2.1 Derivation of Remark 6

The first-order condition of (4.12) (without penalty) with respect to $\beta_{j,k}$ $(j \neq k)$ takes the form

$$\underbrace{-\frac{1}{T}\int_0^T \Psi_k(t)dN_j(t)}_{\text{I}} + \underbrace{\frac{1}{T}\int_0^T \mu_j \Psi_k(t)dt}_{\text{II}} + \underbrace{\frac{1}{T}\int_0^T \left[\sum_{l=1}^{p} \Psi_l(t)\cdot\beta_{j,l}\right]\Psi_k(t)dt}_{\text{III}} = 0, \tag{C.4}$$

where $\Psi_l(t) = (\psi * dN_l)(t) = \int_0^\infty \psi(\Delta)dN_l(t-\Delta)$.

We then have the following set of equations for the three terms in (C.4).

$$\begin{aligned}
\text{I} =& -\frac{1}{T}\int_0^T\int_0^\infty \psi(\Delta)dN_k(t-\Delta)dN_j(t) = -\frac{1}{T}\sum_i \int_0^\infty \psi(\Delta)dN_k(t_{ji}-\Delta) \\
=& -\frac{1}{T}\sum_i \int_0^\infty \psi(\Delta)\sum_{i'}\delta(t_{ki'}-t_{ji}+\Delta)d\Delta = -\int_0^\infty \psi(\Delta)\left[\frac{1}{T}\sum_i\sum_{i'}\delta(t_{ki'}-t_{ji}+\Delta)\right]d\Delta \\
=& -\int_0^\infty \psi(\Delta)\big[\widetilde{V}_{jk}(\Delta)+\widetilde{\Lambda}_k\widetilde{\Lambda}_j\big]d\Delta \\
=& -\underbrace{\int_0^\infty \psi(\Delta)\widetilde{V}_{jk}(\Delta)d\Delta}_{\text{I(i)}} - \underbrace{\int_0^\infty \psi(\Delta)\widetilde{\Lambda}_k\widetilde{\Lambda}_j d\Delta}_{\text{I(ii)}}.
\end{aligned}$$

where we use the fact that $dN_k(t_{ji}-\Delta) = \sum_{i'}\delta(t_{ki'}-t_{ji}+\Delta)d\Delta$ in the third equality, and we use the definitions of $\widetilde{V}_{j,k}$ and $\widetilde{\Lambda}_k$ from Chapter 4 in the second-to-last equality.

$$\begin{aligned}
\text{II} =& \frac{1}{T}\int_0^T \mu_j \int_0^\infty \psi(\Delta)dN_k(t-\Delta)dt = \int_0^\infty \mu_j\psi(\Delta)\left\{\frac{1}{T}\int_0^T dN_k(t-\Delta)dt\right\} \\
=& \mu_j\widetilde{\Lambda}_k\int_0^\infty \psi(\Delta)d\Delta,
\end{aligned}$$

$$
\begin{aligned}
\text{III} =& \frac{1}{T}\int_0^T \left\{\sum_{l=1}^{p}\int_0^\infty \psi(\Delta')\cdot\beta_{j,l}\sum_i \delta(t-\Delta'-t_{li})d\Delta'\right\}\left\{\int_0^\infty \psi(\Delta)\sum_{i'}\delta(t-\Delta-t_{ki'})d\Delta\right\}dt \\
=& \sum_{l=1}^{p}\int_0^\infty\int_0^\infty \psi(\Delta')\cdot\beta_{j,l}\psi(\Delta)\left\{\frac{1}{T}\int_0^T\sum_i\delta(t-\Delta'-t_{li})\sum_{i'}\delta(t-\Delta-t_{ki'})dt\right\}d\Delta d\Delta' \\
=& \sum_{l=1}^{p}\int_0^\infty\int_0^\infty \psi(\Delta')\cdot\beta_{j,l}\psi(\Delta)\left\{\frac{1}{T}\int_0^T\sum_i\sum_{i'}\delta(t-\Delta'-t_{li})\delta(t-\Delta-t_{ki'})dt\right\}d\Delta d\Delta' \\
=& \sum_{l=1}^{p}\int_0^\infty\int_0^\infty \psi(\Delta')\cdot\beta_{j,l}\psi(\Delta)\left\{\frac{1}{T}\sum_i\sum_{i'}\delta(\Delta-\Delta'+t_{ki'}-t_{li})\right\}d\Delta d\Delta' \\
=& \sum_{l=1}^{p}\int_0^\infty\int_0^\infty \psi(\Delta')\cdot\beta_{j,l}\psi(\Delta)\left\{\widetilde{V}_{k,l}(\Delta-\Delta')+\widetilde{\Lambda}_l\widetilde{\Lambda}_k+\widetilde{\Lambda}_k\delta(\Delta-\Delta')\mathbf{1}_{[l=k]}\right\}d\Delta d\Delta' \\
=& \sum_{l=1}^{p}\int_0^\infty\int_0^\infty \psi(\Delta')\cdot\beta_{j,l}\psi(\Delta)\widetilde{V}_{k,l}(\Delta-\Delta')d\Delta d\Delta' + \sum_{l=1}^{p}\int_0^\infty\int_0^\infty \psi(\Delta')\cdot\beta_{j,l}\psi(\Delta)\widetilde{\Lambda}_l\widetilde{\Lambda}_k d\Delta d\Delta' \\
&+ \int_0^\infty\int_0^\infty \psi(\Delta')\cdot\beta_{j,k}\psi(\Delta)\widetilde{\Lambda}_k\delta(\Delta-\Delta')d\Delta d\Delta' \\
=& \underbrace{\sum_{l=1}^{p}\int_0^\infty \psi(\Delta)\big[(\psi\cdot\beta_{j,l})*\widetilde{V}_{k,l}\big](\Delta)d\Delta}_{\text{III(i)}} + \underbrace{\sum_{l=1}^{p}\widetilde{\Lambda}_k\widetilde{\Lambda}_l\int_0^\infty\psi(\Delta)\left[\int_0^\infty\psi(\Delta')\cdot\beta_{j,l}d\Delta'\right]d\Delta}_{\text{III(ii)}} \\
&+ \underbrace{\widetilde{\Lambda}_k\int_0^\infty\big(\psi(\Delta)\cdot\beta_{j,k}\big)\psi(\Delta)d\Delta}_{\text{III(iii)}},
\end{aligned}
$$

where III(iii) follows from the fact that $\int_0^\infty \psi(\Delta')\delta(\Delta-\Delta')d\Delta' = \psi(\Delta)$ since $\delta(\cdot)$ is the Dirac delta function.

Then, II+III(ii)-I(ii) gives

$$
\left\{\widetilde{\Lambda}_k\int_0^\infty\psi(\Delta)d\Delta\right\}\left\{-\widetilde{\Lambda}_j+\mu_j+\sum_{l=1}^{p}\left[\widetilde{\Lambda}_l\int_0^\infty\psi(\Delta')\cdot\beta_{j,l}d\Delta'\right]\right\}. \tag{C.5}
$$

And, III(i)+III(iii)-I(i) gives

$$
\int_0^\infty\psi(\Delta)\left\{-\widetilde{V}_{j,k}(\Delta)+\sum_{l=1}^{p}\big[(\psi\cdot\beta_{j,l})*\widetilde{V}_{k,l}\big](\Delta)+\psi(\Delta)\cdot\beta_{j,k}\widetilde{\Lambda}_k\right\}d\Delta. \tag{C.6}
$$

Combining (C.5) and (C.6) yields the statement in Remark 6.

C.2.2 Technical lemmas

In this section, we state five technical lemmas that are useful in proving Theorems 5 and 6 in Chapter 4. The first four lemmas connect the assumptions on expectations (Assumptions 17–21) with the observed samples. These lemmas can be obtained from Theorem 9 in Section 4.5 in Chapter 4, and their proofs are similar to the proof of Corollary 2 and are thus omitted. The last lemma characterizes the deviation of the point process $d\boldsymbol{N}(t)$ from its incremental intensity $\lambda_j(t)dt$ using the fact that $dN_j(t) - \lambda_j(t)dt$ is a martingale. This lemma directly follows from known inequalities of martingales (see, among others, Theorem 3.1 in van de Geer (1995) or Theorem 3 in Hansen et al. (2015)).

Lemma 11. *Assume that Assumptions 15 – 17 hold. For any M-vector $a_1, \ldots, a_p \in \mathbb{R}^M$ satisfying*

$$\sum_{k \notin \mathcal{E}_j} \left\{ \int_0^T \left[\Psi_k(t) \cdot a_k \right]^2 dt \right\}^{1/2} \leq \sum_{k \in \mathcal{E}_j} \left\{ \int_0^T \left[\Psi_k(t) \cdot a_k \right]^2 dt \right\}^{1/2}, \tag{C.7}$$

we have, for each $j = 1, \ldots, p$ and sufficiently large T,

$$\frac{\xi_1}{2} \left\{ \sum_{k \in \mathcal{E}_j} \left\{ \int_0^T \left[\Psi_k(t) \cdot a_k \right]^2 dt \right\}^{1/2} \right\}^2 \leq s \int_0^T \left[\sum_{k=1}^p \Psi_k(t) \cdot a_k \right]^2 dt, \tag{C.8}$$

with probability at least $1 - c_2' p^2 T \exp(-c_3 T^{1/5})$.

Lemma 12. *Assume that Assumptions 15, 16, and 18 hold. Then, for each $j = 1, \ldots, p$,*

$$0 < \frac{\gamma_{\min}}{2} \leq \Gamma_{\min} \left(\frac{1}{T} \int_0^T \Psi_k(t) \Psi_k^{\mathrm{T}}(t)\, dt \right) \leq \Gamma_{\max} \left(\frac{1}{T} \int_0^T \Psi_k(t) \Psi_k^{\mathrm{T}}(t)\, dt \right) \leq 2\gamma_{\max}, \tag{C.9}$$

with probability at least $1 - c_2' T \exp(-c_3 T^{1/5})$.

Lemma 13. *Assume that Assumptions 15, 16, and 20 hold. Then, for each $j = 1, \ldots, p$,*

$$\max_{k \notin \mathcal{E}_j} \left\| \left(\int_0^T \Psi_k(t) \Psi_{\mathcal{E}_j}^{\mathrm{T}}(t)\, dt \right) \left(\int_0^T \Psi_{\mathcal{E}_j}(t) \Psi_{\mathcal{E}_j}^{\mathrm{T}}(t)\, dt \right)^{-1} \right\|_2 \leq 2\xi_2, \tag{C.10}$$

with probability at least $1 - c_2' p^2 T \exp(-c_3 T^{1/5})$, where $\Psi_{\mathcal{E}_j}$ is introduced in Assumption 20.

Lemma 14. *Assume that Assumptions 15, 16, and 21 hold. Then, for* $(j,k) \in \mathcal{E}$ *and* $\beta_{\min}$ *as defined in Assumption 21 in Chapter 4,*

$$\min_{(j,k)\in\mathcal{E}} \left\{ \frac{1}{T} \int_0^T \left[\big(\omega_{j,k} * dN_k\big)(t) \right]^2 dt \right\}^{1/2} \geq \frac{1}{2}\beta_{\min},$$

with probability at least $1 - c_2' p^2 T \exp(-c_3 T^{1/5})$.

Lemma 15. *Suppose that there exists* $\lambda_{\max}$ *such that* $\lambda_j(t) \leq \lambda_{\max}$ *for all* t *and for* $1 \leq j \leq p$. *Let* $H(t)$ *be a bounded function that is* $\mathcal{H}_t$*-predictable. Then, for any* $\epsilon > 0$, *the inequality*

$$\frac{1}{T} \int_0^T H(t)[\lambda_j(t)dt - dN_j(t)]dt \leq 4 \left\{ \frac{\lambda_{\max}}{2T} \int_0^T H^2(t)dt \right\}^{1/2} \epsilon^{1/2} \tag{C.11}$$

holds with probability at least $1 - c_4' \exp(-\epsilon T)$.

C.2.3 Proof of Theorem 5

Proof. In this proof, we establish the oracle inequality in Theorem 5 for learning the intensity of a given node j using the penalized regression (4.12). Recall that, for $k \in \mathcal{E}_j$, $\tilde{\beta}_{j,k} \in \mathbb{R}^M$ was introduced in Assumption 19 of Chapter 4. Let $\tilde{\beta}_{j,k} = 0$ for $k \notin \mathcal{E}_j$. Further define $\tilde{\lambda}_j(t) \equiv \sum_{k=1}^p \Psi_k(t) \cdot \tilde{\beta}_{j,k} + \mu_j$. From Assumption 19 and the Cauchy-Schwartz inequality, we know that

$$\frac{1}{T} \int_0^T \left[\tilde{\lambda}_j(t) - \lambda_j(t)\right]^2 dt \leq s^2 Q T^{\frac{-2\theta_1}{2\theta_1+1}}. \tag{C.12}$$

Since j is fixed throughout the proof, in the rest of this proof we will drop the subscript j to avoid cumbersome bookkeeping. For instance, we use β_k to denote $\beta_{j,k}$, $\tilde{\beta}_k$ to denote $\tilde{\beta}_{j,k}$, and λ to denote λ_j.

Let $\mathcal{M}$ denote the event that the statements of Lemmas 11 – 14 being true, given Assumptions 15–21. We know from the lemmas that $\mathcal{M}$ holds with probability at least $1 - c_2 p^2 T \exp(-c_3 T^{1/5})$, where $c_2 \equiv 4c_2'$. In what follows, we assume that $\mathcal{M}$ holds.

Since $\hat{\beta}$ is the minimizer of (4.12), the following basic inequality holds for $\hat{\beta}$ and $\tilde{\beta}$,

$$\begin{aligned} &- \int_0^T \hat{\lambda}(t)dN(t) + \frac{1}{2}\int_0^T \hat{\lambda}^2(t)dt + \eta\sqrt{T}\sum_k \left\{ \int_0^T \left[\Psi_k(t)\cdot\hat{\beta}_k\right]^2 dt \right\}^{1/2} \\ \leq &- \int_0^T \tilde{\lambda}(t)dN(t) + \frac{1}{2}\int_0^T \tilde{\lambda}^2(t)dt + \eta\sqrt{T}\sum_k \left\{ \int_0^T \left[\Psi_k(t)\cdot\tilde{\beta}_k\right]^2 dt \right\}^{1/2}. \end{aligned} \tag{C.13}$$

Adding $1/2\int_0^T \lambda^2(t)dt - \int_0^T \hat{\lambda}(t)\lambda(t)dt$ to both sides of (C.13), and adding and subtracting $\int_0^T \tilde{\lambda}(t)\lambda(t)dt$ on the right-hand side of the inequality, we get

$$\begin{aligned}
&\frac{1}{2}\int_0^T \left[\hat{\lambda}(t)-\lambda(t)\right]^2 dt \\
\leq &\frac{1}{2}\int_0^T \left[\tilde{\lambda}(t)-\lambda(t)\right]^2 dt + \int_0^T \left[\sum_{k=1}^{p} \Psi_k(t)\cdot(\hat{\beta}_k-\tilde{\beta}_k)\right][\lambda(t)dt - dN(t)] \\
&+\eta\sqrt{T}\sum_k \left\{\left\{\int_0^T \left[\Psi_k(t)\cdot\tilde{\beta}_k\right]^2 dt\right\}^{1/2} - \left\{\int_0^T \left[\Psi_k(t)\cdot\hat{\beta}_k\right]^2 dt\right\}^{1/2}\right\}.
\end{aligned} \tag{C.14}$$

The term on the left-hand side of (C.14) is the quantity of interest. On the right-hand side, the first term is the approximation error by the truncated basis expansion, the second term involves the stochastic error, and the third term is due to the penalty. We will use Assumption 19, Lemma 15, and Lemma 11 to bound these three terms.

To proceed, we consider the following two situations:

Case 1. It holds that

$$\begin{aligned}
\frac{1}{2}\int_0^T \left[\hat{\lambda}(t)-\lambda(t)\right]^2 dt \geq &\frac{1}{2}\int_0^T \left[\tilde{\lambda}(t)-\lambda(t)\right]^2 dt + \\
&\int_0^T \left[\sum_{k=1}^{p} \Psi_k(t)\cdot(\hat{\beta}_k-\tilde{\beta}_k)\right][\lambda(t)dt - dN(t)].
\end{aligned} \tag{C.15}$$

Then from (C.14) and (C.15), we see that

$$\sum_{k\notin\mathcal{E}}\left\{\int_0^T \left[\Psi_k(t)\cdot(\hat{\beta}_k-\tilde{\beta}_k)\right]^2 dt\right\}^{1/2} \leq \sum_{k\in\mathcal{E}}\left\{\int_0^T \left[\Psi_k(t)\cdot(\tilde{\beta}_k-\hat{\beta}_k)\right]^2 dt\right\}^{1/2}, \tag{C.16}$$

where we use the fact that $\tilde{\beta}_k = 0$ for $k \notin \mathcal{E}$ and triangle inequality for the right-hand side of the inequality. Given the event $\mathcal{M}$, applying Assumption 17 and Lemma 11 gives that

$$\sum_{k\in\mathcal{E}}\left\{\int_0^T \left[\Psi_k(t)\cdot(\hat{\beta}_k-\tilde{\beta}_k)\right]^2 dt\right\}^{1/2} \leq \sqrt{\frac{2s}{\xi_1}}\left\{\int_0^T \left[\sum_{k=1}^{p} \Psi_k(t)\cdot(\hat{\beta}_k-\tilde{\beta}_k)\right]^2 dt\right\}^{1/2}. \tag{C.17}$$

Plugging (C.17) into (C.14) gives

$$\begin{aligned}
&\frac{1}{2}\int_0^T \left[\hat{\lambda}(t)-\lambda(t)\right]^2 dt \\
\leq &\frac{1}{2}\int_0^T \left[\tilde{\lambda}(t)-\lambda(t)\right]^2 dt + \int_0^T \left[\sum_{k=1}^p \Psi_k(t)\cdot(\hat{\beta}_k-\tilde{\beta}_k)\right][\lambda(t)dt-dN(t)] \\
&+\eta\sqrt{T}\sqrt{\frac{2s}{\xi_1}}\left\{\int_0^T\left[\sum_{k=1}^p \Psi_k(t)\cdot(\hat{\beta}_k-\tilde{\beta}_k)\right]^2 dt\right\}^{1/2}.
\end{aligned} \tag{C.18}$$

Case 2. The inequality (C.15) does not hold. Then (C.18) follows directly.

In summary, (C.18) holds given the event $\mathcal{M}$.

Consider the second term in (C.18). Applying Lemma 15 with $\epsilon = 2\log(p)/T$ and $H(t) \equiv \sum_{k=1}^p \Psi_k(t)\cdot(\hat{\beta}_k-\tilde{\beta}_k)$, we see that, with probability at least $1-c_4' p^{-2}$,

$$\begin{aligned}
&\int_0^T \left[\sum_{k=1}^p \Psi_k(t)\cdot(\hat{\beta}_k-\tilde{\beta}_k)\right][\lambda(t)dt-dN(t)] \\
\leq &4\left\{\frac{\lambda_{\max}}{2T}\int_0^T\left[\sum_{k=1}^p \Psi_k(t)\cdot(\hat{\beta}_k-\tilde{\beta}_k)\right]^2 dt\right\}^{1/2}\left\{\frac{2\log(p)}{T}\right\}^{1/2}.
\end{aligned} \tag{C.19}$$

Plugging (C.19) into (C.18) gives

$$\begin{aligned}
&\frac{1}{2}\int_0^T \left[\hat{\lambda}(t)-\lambda(t)\right]^2 dt \\
\leq &\frac{1}{2}\int_0^T \left[\tilde{\lambda}(t)-\lambda(t)\right]^2 dt + 4\sqrt{T}\sqrt{\lambda_{\max}\frac{\log(p)}{T}}\left\{\int_0^T\left[\sum_{k=1}^p \Psi_k(t)\cdot(\hat{\beta}_k-\tilde{\beta}_k)\right]^2 dt\right\}^{1/2} \\
&+\eta\sqrt{T}\sqrt{\frac{2s}{\xi_1}}\left\{\int_0^T\left[\sum_{k=1}^p \Psi_k(t)\cdot(\hat{\beta}_k-\tilde{\beta}_k)\right]^2 dt\right\}^{1/2} \\
\leq &\frac{1}{2}\int_0^T \left[\tilde{\lambda}(t)-\lambda(t)\right]^2 dt + \sqrt{T}\left[\sqrt{\frac{2s}{\xi_1}}\eta + 4\sqrt{\lambda_{\max}\frac{\log(p)}{T}}\right]\left\{\int_0^T\left[\sum_{k=1}^p \Psi_k(t)\cdot(\hat{\beta}_k-\tilde{\beta}_k)\right]^2 dt\right\}^{1/2},
\end{aligned} \tag{C.20}$$

which holds with probability at least $1 - c_4' p^{-2}$ given the event $\mathcal{M}$. Let the tuning parameters η be chosen such that $\eta = (8\lambda_{\max}\xi_1)^{1/2}\sqrt{\log(p)/T}$. Then

$$\begin{aligned}
&\frac{1}{2}\int_0^T \left[\hat{\lambda}(t) - \lambda(t)\right]^2 dt \\
\leq &\frac{1}{2}\int_0^T \left[\tilde{\lambda}(t) - \lambda(t)\right]^2 dt + 8(s\lambda_{\max})^{1/2}\sqrt{\frac{\log(p)}{T}}\sqrt{2T}\left\{\frac{1}{2}\int_0^T \left[\hat{\lambda}(t) - \tilde{\lambda}(t)\right]^2 dt\right\}^{1/2},
\end{aligned} \tag{C.21}$$

Note that (C.21) contains square distance between three intensities: the true intensity λ, the approximated intensity $\tilde{\lambda}$, and the estimated intensity $\hat{\lambda}$. In the remaining of the proof, we discuss the relationship between these three square distances.

We consider the following three cases

Case I. $\frac{1}{2}\int_0^T \left[\hat{\lambda}(t) - \lambda(t)\right]^2 dt \geq \frac{1}{2}\int_0^T \left[\tilde{\lambda}(t) - \lambda(t)\right]^2 dt$ and $\frac{1}{2}\int_0^T \left[\hat{\lambda}(t) - \lambda(t)\right]^2 dt \geq \frac{1}{2}\int_0^T \left[\hat{\lambda}(t) - \tilde{\lambda}(t)\right]^2 dt$.

Case II. $\frac{1}{2}\int_0^T \left[\hat{\lambda}(t) - \lambda(t)\right]^2 dt < \frac{1}{2}\int_0^T \left[\tilde{\lambda}(t) - \lambda(t)\right]^2 dt$.

Case III. $\frac{1}{2}\int_0^T \left[\hat{\lambda}(t) - \lambda(t)\right]^2 dt < \frac{1}{2}\int_0^T \left(\hat{\lambda}(t) - \tilde{\lambda}(t)\right)^2 dt$.

Note that, although the three cases are not mutually exclusive, they cover all possible situations. In Case I, we have that

$$\begin{aligned}
&\frac{1}{2}\int_0^T \left[\hat{\lambda}(t) - \lambda(t)\right]^2 dt \\
\leq &\left\{\frac{1}{2}\int_0^T \left[\tilde{\lambda}(t) - \lambda(t)\right]^2 dt\right\}^{1/2}\left\{\frac{1}{2}\int_0^T \left[\hat{\lambda}(t) - \lambda(t)\right]^2 dt\right\}^{1/2} \\
&+ 8(s\lambda_{\max})^{1/2}\sqrt{\frac{\log(p)}{T}}\sqrt{2T}\left\{\frac{1}{2}\int_0^T \left[\hat{\lambda}(t) - \lambda(t)\right]^2 dt\right\}^{1/2}.
\end{aligned} \tag{C.22}$$

Using (C.12), we have

$$\frac{1}{T}\int_0^T \left[\hat{\lambda}(t) - \lambda(t)\right]^2 dt \leq 2s^2 Q T^{-\frac{2\theta_1}{2\theta_1+1}} + 2^6 s\lambda_{\max}\frac{\log(p)}{T}. \tag{C.23}$$

In Case II, it follows directly from (C.12) that

$$\frac{1}{T}\int_0^T \left[\hat{\lambda}(t) - \lambda(t)\right]^2 dt \leq s^2 Q T^{\frac{-2\theta_1}{2\theta_1+1}},$$

which implies (C.23).

In Case III, we add $1/2\int_0^T \tilde{\lambda}^2(t)dt - \int_0^T \hat{\lambda}(t)\tilde{\lambda}(t)dt$ on both sides of the basic inequality (C.13), and rearrange the right-hand side (add and subtract $\int_0^T \hat{\lambda}(t)[\tilde{\lambda}(t) - \hat{\lambda}(t)]dt$) to get

$$\begin{aligned}
&\frac{1}{2}\int_0^T \left[\hat{\lambda}(t) - \tilde{\lambda}(t)\right]^2 dt \\
&\leq \int_0^T \left[\tilde{\lambda}(t) - \hat{\lambda}(t)\right]\left[\lambda(t)dt - dN(t)\right] - \int_0^T \left[\tilde{\lambda}(t) - \lambda(t)\right]\left[\tilde{\lambda}(t) - \hat{\lambda}(t)\right]dt \\
&\quad + \eta\sqrt{T}\sum_k \left\{\left\{\int_0^T \left[\Psi_k(t)\cdot\tilde{\beta}_k\right]^2 dt\right\}^{1/2} - \left\{\int_0^T \left[\Psi_k(t)\cdot\hat{\beta}_k\right]^2 dt\right\}^{1/2}\right\}.
\end{aligned} \tag{C.24}$$

Plugging the inequality in Case III to the left-hand side of (C.24), we get

$$\begin{aligned}
&\frac{1}{2}\int_0^T \left[\hat{\lambda}(t) - \lambda(t)\right]^2 dt \\
&\leq \int_0^T \left[\tilde{\lambda}(t) - \lambda(t)\right]\left[\tilde{\lambda}(t) - \hat{\lambda}(t)\right]dt + \int_0^T \left[\tilde{\lambda}(t) - \hat{\lambda}(t)\right]\left[\lambda(t)dt - dN(t)\right] \\
&\quad + \eta\sqrt{T}\sum_k \left\{\left\{\int_0^T \left[\Psi_k(t)\cdot\tilde{\beta}_k\right]^2 dt\right\}^{1/2} - \left\{\int_0^T \left[\Psi_k(t)\cdot\hat{\beta}_k\right]^2 dt\right\}^{1/2}\right\}.
\end{aligned} \tag{C.25}$$

Expanding the left-hand side gives

$$\begin{aligned}
&\frac{1}{2}\int_0^T \left[\hat{\lambda}(t) - \tilde{\lambda}(t)\right]^2 dt + \frac{1}{2}\int_0^T \left[\tilde{\lambda}(t) - \lambda(t)\right]^2 dt + \\
&\int_0^T \left[\hat{\lambda}(t) - \tilde{\lambda}(t)\right]\left[\tilde{\lambda}(t) - \lambda(t)\right] dt \\
&\leq \int_0^T \left[\tilde{\lambda}(t) - \lambda(t)\right]\left[\tilde{\lambda}(t) - \hat{\lambda}(t)\right]dt + \int_0^T \left[\tilde{\lambda}(t) - \hat{\lambda}(t)\right)\left]\left[\lambda(t)dt - dN(t)\right] \\
&\quad + \eta\sqrt{T}\sum_k \left\{\left\{\int_0^T \left[\Psi_k(t)\cdot\tilde{\beta}_k\right]^2 dt\right\}^{1/2} - \left\{\int_0^T \left[\Psi_k(t)\cdot\hat{\beta}_k\right]^2 dt\right\}^{1/2}\right\}.
\end{aligned} \tag{C.26}$$

Since the second term on the left-hand side is non-negative, the inequality reduces to

$$\begin{aligned}
&\frac{1}{2}\int_0^T \left[\hat{\lambda}(t) - \tilde{\lambda}(t)\right]^2 dt \\
&\leq \int_0^T \left[\tilde{\lambda}(t) - \hat{\lambda}(t)\right]\left[\lambda(t)dt - dN(t)\right] + 2\int_0^T \left[\tilde{\lambda}(t) - \lambda(t)\right]\left[\tilde{\lambda}(t) - \hat{\lambda}(t)\right]dt \\
&\quad + \eta\sqrt{T}\sum_k \left\{\left\{\int_0^T \left[\Psi_k(t)\cdot\tilde{\beta}_k\right]^2 dt\right\}^{1/2} - \left\{\int_0^T \left[\Psi_k(t)\cdot\hat{\beta}_k\right]^2 dt\right\}^{1/2}\right\}.
\end{aligned} \tag{C.27}$$

Using the same argument as the one for obtaining (C.18), we have

$$\begin{aligned}
&\frac{1}{2}\int_0^T \left[\hat{\lambda}(t)-\tilde{\lambda}(t)\right]^2 dt \\
&\leq \int_0^T \left[\tilde{\lambda}(t)-\hat{\lambda}(t)\right]\left[\lambda(t)dt - dN(t)\right] + 2\int_0^T \left[\tilde{\lambda}(t)-\lambda(t)\right]\left[\tilde{\lambda}(t)-\hat{\lambda}(t)\right]dt \\
&\quad + 8(\lambda_{\max})^{-1/2}\sqrt{\frac{\log(p)}{T}}\sqrt{2T}\left\{\frac{1}{2}\int_0^T \left[\hat{\lambda}(t)-\tilde{\lambda}(t)\right]^2 dt\right\}^{1/2},
\end{aligned} \tag{C.28}$$

where we plug in the tuning parameter $\eta = (8\lambda_{\max}\xi_1)^{1/2}\sqrt{\log(p)/T}$. Applying Lemma 15 with $H(t) \equiv \sum_{k=1}^p \Psi_k(t)\cdot(\hat{\beta}_k - \tilde{\beta}_k)$ and $\epsilon = 2\log(p)/T$ to the first term on the right-hand side, and using Cauchy-Schwartz inequality and (C.12) on the second term, we have

$$\begin{aligned}
&\frac{1}{2T}\int_0^T \left[\hat{\lambda}(t)-\tilde{\lambda}(t)\right]^2 dt \\
&\leq \left[2^{3/2}sQ^{1/2}T^{\frac{-\theta_1}{2\theta_1+1}} + 8(s\lambda_{\max})^{-1/2}\sqrt{\frac{\log(p)}{T}}\sqrt{2}\right]\left\{\frac{1}{2T}\int_0^T \left[\hat{\lambda}(t)-\tilde{\lambda}(t)\right]^2 dt\right\}^{1/2},
\end{aligned} \tag{C.29}$$

which holds with probability at least $1 - c_4' p^{-2}$ given the event $\mathcal{M}$. Recall that the condition in Case III implies that left-hand side of (C.29) is larger than $T^{-1}\int \left[\hat{\lambda}(t)-\lambda(t)\right]^2 dt$. Therefore, given the event $\mathcal{M}$, we know that, with probability at least $1 - 2c_4' p^{-2}$,

$$\frac{1}{2T}\int_0^T \left[\hat{\lambda}(t)-\lambda(t)\right]^2 dt \leq 2^4 s^2 Q T^{-\frac{2\theta_1}{2\theta_1+1}} + 2^6 s\lambda_{\max}\frac{\log(p)}{T}. \tag{C.30}$$

Combining results from three cases, (C.23) holds with probability at least $1 - 2c_4' p^{-2}$ for each $j = 1, \ldots, p$ given the event $\mathcal{M}$. Then, using a Bonferroni correction, we know that (C.23) holds for all j with probability at least $1 - c_4 p^{-1}$ given the event $\mathcal{M}$. Recalling that the event $\mathcal{M}$ holds with probability at least $1 - c_2 p^2 T\exp(-c_3 T^{1/5})$, we know that (C.23) holds for all j with probability at least $1 - c_4 p^{-1} - c_2 p^2 T\exp(-c_3 T^{1/5})$. □

C.2.4 Proof of Theorem 6

We establish variable selection consistency using the primal-dual witness method (Wainwright, 2009). For simplicity, we assume that $\mu_j = 0$ is known in this proof; otherwise, we have one extra parameter to keep track of while the main conclusion will not be affected. As in the proof

of Theorem 5, we drop the subscript j in what follows. We also assume that the event $\mathcal{M}$ holds. Recall that $\mathcal{M}$ is introduced in the proof of Theorem 5 to denote the event that the statements of Lemmas 11 – 14 being true, given Assumptions 15–21. We know from the lemmas that $\mathcal{M}$ holds with probability at least $1 - c_2 p^2 T \exp(-c_3 T^{1/5})$, where $c_2 \equiv 4c_2'$.

A vector $\hat{\beta}$ solves the optimization problem (4.12) in Chapter 4 if it satisfies the Karush-Kuhn-Tucker (KKT) condition, i.e.,

$$-\frac{1}{T}\int_0^T \Psi_k(t)dN(t) + \frac{1}{T}\int_0^T \Psi_k(t)\left[\sum_l \Psi_l^{\mathrm{T}}(t)\hat{\beta}_l\right]dt + \eta\hat{g}_k = 0, \tag{C.31}$$

where

$$\hat{g}_k = \frac{\int_0^T \Psi_k(t)\Psi_k^{\mathrm{T}}(t)dt\hat{\beta}_k}{\left(\hat{\beta}_k^{\mathrm{T}}\int_0^T \Psi_k(t)\Psi_k^{\mathrm{T}}(t)dt\hat{\beta}_k\right)^{1/2}} \quad \text{if } \hat{\beta}_k \neq 0, \tag{C.32}$$

$$\hat{g}_k^{\mathrm{T}}\left(\int_0^T \Psi_k(t)\Psi_k^{\mathrm{T}}(t)dt\right)^{-1}\hat{g}_k < 1 \quad \text{if } \hat{\beta}_k = 0. \tag{C.33}$$

We will construct an oracle estimator $\{\hat{\beta}_k\}_{k=1}^p$ and will verify that it satisfies the KKT conditions (C.31), (C.32) and (C.33), which means that it solves the optimization problem (4.12) in Chapter 4. If the optimal solution to (4.12) is unique, then the oracle estimator is the unique estimator. If the optimal solution is not unique, then by Theorem 2 in Roth and Fischer (2008), the null set of any optimal solution should contain $\mathcal{E}^c$, and thus any optimal solution satisfies the construction of the oracle estimator. Therefore, the statement of Theorem 6 holds for any optimal solution of (4.12).

We construct an oracle primal-dual pair $\{\hat{\beta}_k, \hat{g}_k\}_{k=1}^p$ as follows:

1. Let

$$\begin{aligned}\hat{\beta}_{\mathcal{E}} = \underset{\beta_k\in\mathbb{R}^M, k\in\mathcal{E}}{\arg\min} & -\frac{1}{T}\int_0^T\left[\sum_{k\in\mathcal{E}}\Psi_k(t)\cdot\beta_k\right]dN(t) \\ & + \frac{1}{2T}\int_0^T\left[\sum_{k\in\mathcal{E}}\Psi_k(t)\cdot\beta_k\right]^2 dt + \eta\sum_{k=1}^p\left\|\Psi_k\cdot\beta_k\right\|_{2,[0,T]},\end{aligned} \tag{C.34}$$

   and set $\hat{\beta}_k = 0$ for $k \notin \mathcal{E}$.

2. Define $\hat{g}_k$ for $k \in \mathcal{E}$ as in (C.32).

3. Solve $\hat{g}_k$ from the sub-gradient condition (C.31) for $k \notin \mathcal{E}$.

We will then verify the support recovery consistency

$$\max_{k\in\mathcal{E}} \left\{ \int_0^T \left[\Psi_k(t) \cdot \hat{\beta}_k - \int_0^t \omega_{j,k}(t-s) dN_k(s) \right]^2 dt \right\}^{1/2} \leq \frac{1}{4}\beta_{\min}, \tag{C.35}$$

which, by the triangle inequality and Assumption 21, implies that $\int_0^T [\Psi_k(t) \cdot \hat{\beta}_k]^2 dt \neq 0$ for $k \in \mathcal{E}$. And, we also verify strict dual feasibility (C.33) for $k \notin \mathcal{E}$. This will complete our argument that the oracle estimator recovers the support exactly.

First, we show the support recovery consistency (C.35). This follows from applying the result of Theorem 5 on the constrained problem (C.34), which yields a bound on $T^{-1}\int_0^T \left[\hat{\lambda}(t) - \lambda(t)\right]^2 dt$. Combining the bound on $T^{-1}\int_0^T \left[\hat{\lambda}(t) - \lambda(t)\right]^2 dt$ with Assumption 19 and the inequality

$$\left[\hat{\lambda}(t) - \tilde{\lambda}(t)\right]^2 \leq 2\left[\hat{\lambda}(t) - \lambda(t)\right]^2 + 2\left[\tilde{\lambda}(t) - \lambda(t)\right]^2,$$

we arrive at a bound for $\hat{\lambda}(t) - \tilde{\lambda}(t) = \sum_{k=1}^p \left[\Psi_k(t) \cdot \left(\hat{\beta}_k - \tilde{\beta}_k\right)\right]$

$$\frac{1}{T}\int_0^T \left[\sum_{k=1}^p \Psi_k(t) \cdot \left(\hat{\beta}_k - \tilde{\beta}_k\right) \right]^2 dt \leq 2^4 s^2 Q T^{-\frac{2\theta_1}{2\theta_1+1}} + 2^6 s \lambda_{\max} \frac{\log(p)}{T}. \tag{C.36}$$

which holds with probability at least $1 - c_4 p^{-2}$ given the event $\mathcal{M}$. Since $\hat{\beta}_k$ and $\tilde{\beta}_k$ are zero for $k \notin \mathcal{E}$, by Assumption 17 and Lemma 11,

$$\frac{\xi_1}{2}\left\{ \sum_{k\in\mathcal{E}} \left\{ \frac{1}{T}\int_0^T \left[\Psi_k(t) \cdot (\hat{\beta}_k - \tilde{\beta}_k)\right]^2 dt \right\}^{1/2} \right\}^2 \leq 2^4 s^2 Q T^{-\frac{2\theta_1}{2\theta_1+1}} + 2^6 s \lambda_{\max} \frac{\log(p)}{T}. \tag{C.37}$$

It follows that

$$\begin{aligned}
&\max_{k\in\mathcal{E}} \left\{ \int_0^T \left[\Psi_k(t) \cdot \hat{\beta}_k - \int_0^t \omega_{j,k}(t-s) dN_k(s)\right]^2 dt \right\}^{1/2} \\
=&\max_{k\in\mathcal{E}} \left\{ \int_0^T \left[\Psi_k(t) \cdot \hat{\beta}_k - \Psi_k(t) \cdot \tilde{\beta}_k + \Psi_k(t) \cdot \tilde{\beta}_k - \int_0^t \omega_{j,k}(t-s) dN_k(s)\right]^2 dt \right\}^{1/2} \\
\leq& \left\{ 2^6 \frac{s^2}{\xi_1} Q T^{-\frac{2\theta_1}{2\theta_1+1}} + 2^8 s \lambda_{\max} \frac{\log(p)}{T} + 2 Q T^{-\frac{2\theta_1}{2\theta_1+1}} \right\}^{1/2}.
\end{aligned}$$

where we use Assumption 19 in the last inequality. Since $s = O\big(\log(p)\big)$ and $p^2 T \exp(-c_3 T^{1/5}) = o(1)$, we know that $s\log(p)/T = o(1)$ and $s^2 T^{-\frac{2\theta_1}{2\theta_1+1}} = o(1)$ (recall that $\theta_1 \geq 2$ from Assumption 19 in Chapter 4). For sufficiently large T, we know that

$$\max_{k\in\mathcal{E}} \left\{ \int_0^T \big[\Psi_k(t)\cdot\hat{\beta}_k - \int_0^t \omega_{j,k}(t-s)dN_k(s)\big]^2 dt \right\}^{1/2} \leq \frac{1}{4}\beta_{\min}. \tag{C.38}$$

Next, we verify strict dual feasibility (C.33) for $k \notin \mathcal{E}$. From (C.31) for the constrained problem, we have

$$\frac{1}{T}\int_0^T \Psi_{\mathcal{E}}(t)\Psi_{\mathcal{E}}^{\mathrm{T}}(t)\big(\hat{\beta}_{\mathcal{E}} - \tilde{\beta}_{\mathcal{E}}\big)\,dt + \frac{1}{T}\int_0^T \Psi_{\mathcal{E}}(t)\left\{\Psi_{\mathcal{E}}^{\mathrm{T}}(t)\tilde{\beta}_{\mathcal{E}}dt - dN(t)\right\} + \eta\hat{g}_{\mathcal{E}} = 0,$$

where $\Psi_{\mathcal{E}}(t)$ is introduced in Assumption 20 and $\hat{g}_{\mathcal{E}}$ is the joint vector of $\hat{g}_k$ for $k \in \mathcal{E}$. Rearranging the terms gives

$$\hat{\beta}_{\mathcal{E}} - \tilde{\beta}_{\mathcal{E}} = \left\{\frac{1}{T}\int_0^T \Psi_{\mathcal{E}}(t)\Psi_{\mathcal{E}}^{\mathrm{T}}(t)\,dt\right\}^{-1}\big[R_{\mathcal{E}} + \eta\hat{g}_{\mathcal{E}}\big], \tag{C.39}$$

where

$$R_{\mathcal{E}} \equiv \frac{1}{T}\int_0^T \Psi_{\mathcal{E}}(t)\big[\tilde{\lambda}(t) - \lambda(t)\big]dt + \frac{1}{T}\int_0^T \Psi_{\mathcal{E}}(t)[\lambda(t)dt - dN(t)]. \tag{C.40}$$

Now for $k \notin \mathcal{E}$, we calculate $\hat{g}_k$ from the subgradient condition (C.31)

$$\frac{1}{T}\int_0^T \Psi_k(t)\Psi_{\mathcal{E}}^{\mathrm{T}}(t)\big[\hat{\beta}_{\mathcal{E}} - \tilde{\beta}_{\mathcal{E}}\big]\,dt + \frac{1}{T}\int_0^T \Psi_k(t)\left\{\Psi_{\mathcal{E}}^{\mathrm{T}}(t)\tilde{\beta}_{\mathcal{E}}dt - dN(t)\right\} + \eta\hat{g}_k = 0.$$

Rearranging the terms and plugging in (C.39) gives

$$\eta\hat{g}_k = \left\{\frac{1}{T}\int_0^T \Psi_k(t)\Psi_{\mathcal{E}}^{\mathrm{T}}(t)dt\right\}\left\{\frac{1}{T}\int_0^T \Psi_{\mathcal{E}}(t)\Psi_{\mathcal{E}}^{\mathrm{T}}(t)\,dt\right\}^{-1}[R_{\mathcal{E}} + \eta\hat{g}_{\mathcal{E}}] + R_k, \tag{C.41}$$

where

$$R_k \equiv \frac{1}{T}\int_0^T \Psi_k(t)\big[\tilde{\lambda}(t) - \lambda(t)\big]dt + \frac{1}{T}\int_0^T \Psi_k(t)[\lambda(t)dt - dN(t)]..$$

Applying Assumption 19 and Lemma 15 with $H(t) = \Psi_{\mathcal{E}}(t)$, we arrive at the following bound for each entry of $R_{\mathcal{E}}$: for $l \in \mathcal{E}$ and $m = 1, \ldots, M$,

$$|R_{lm}| \leq \left\{\frac{1}{2T}\int_0^T \Psi_{lm}^2(t)\,dt\right\}^{1/2}\left\{\sqrt{s^2 Q T^{-\frac{2\theta_1}{2\theta_1+1}}} + 4\sqrt{\lambda_{\max}\frac{\log(p)}{T}}\right\}, \tag{C.42}$$

which holds with probability at least $1 - c_4 p^{-2}$ given the event $\mathcal{M}$. And thus,

$$\begin{aligned}\|R_{\mathcal{E}}\|_2^2 =& \sum_{l\in\mathcal{E}} \|R_l\|_2^2 = \sum_{l\in\mathcal{E}}\sum_{m=1}^{M} R_{lm}^2 \\ \leq& \left\{\sqrt{s^2 Q T^{-\frac{2\theta_1}{2\theta_1+1}}} + 4\sqrt{\lambda_{\max}\frac{\log(p)}{T}}\right\}^2 \sum_{k\in\mathcal{E}}\sum_{m=1}^{M}\left\{\frac{1}{T}\int_0^T \Psi_{lm}^2(t)\,dt\right\} \\ \leq& 2sM\gamma_{max}\left\{\sqrt{s^2 Q T^{-\frac{2\theta_1}{2\theta_1+1}}} + 4\sqrt{\lambda_{\max}\frac{\log(p)}{T}}\right\}^2.\end{aligned}$$

This implies that

$$\|R_{\mathcal{E}}\|_2^2 \approx 2sqT^{1/(2\theta_1+1)}\gamma_{max}\left\{\sqrt{s^2 Q T^{-\frac{2\theta_1}{2\theta_1+1}}} + 4\sqrt{\lambda_{\max}\frac{\log(p)}{T}}\right\}^2. \tag{C.43}$$

On the other hand,

$$\|\hat{g}_{\mathcal{E}}\|_2 \leq \sqrt{2s\gamma_{\max}}, \tag{C.44}$$

since

$$\frac{1}{2\gamma_{\max}}\|\hat{g}_k\|_2^2 \leq \hat{g}_k^{\mathrm{T}}\left(\frac{1}{T}\int_0^T \Psi_k(t)\Psi_k^{\mathrm{T}}(t)dt\right)^{-1}\hat{g}_k = 1, \quad k\in\mathcal{E}.$$

Finally, by Assumption 20 and Lemma 13, we know that, given that the event $\mathcal{M}$ holds,

$$\max_{k\notin\mathcal{E}}\left\|\left(\frac{1}{T}\int_0^T \Psi_k(t)\Psi_{\mathcal{E}}^{\mathrm{T}}(t)\,dt\right)\left(\frac{1}{T}\int_0^T \Psi_{\mathcal{E}}(t)\Psi_{\mathcal{E}}^{\mathrm{T}}(t)\,dt\right)^{-1}\right\|_2 \leq \frac{\xi_2}{2}.$$

Using (C.41), (C.43), and (C.44), we thus have

$$\|\hat{g}_k\|_2 \leq \frac{1}{\eta}\sqrt{2q\gamma_{max}}\left\{\frac{\xi_2}{2}\sqrt{s^2 T^{1/(2\theta_1+1)}}+1\right\}\left\{\sqrt{sQT^{-\frac{2\theta_1}{2\theta_1+1}}} + 4\sqrt{\lambda_{\max}\frac{\log(p)}{T}}\right\} + \frac{\xi_2}{2}\sqrt{2s\gamma_{\max}}.$$

Suppose that T is sufficiently large so that $\xi_2 s^{1/2} T^{\frac{-1}{4\theta_1+2}} \geq 2$. Rearranging the terms and recalling that $\eta = (8\lambda_{\max}\xi_1\log(p)/T)^{1/2}$, we get

$$\|\hat{g}_k\|_2 \leq \xi_2\sqrt{\frac{s\gamma_{max}}{2}}\sqrt{\frac{2}{\xi_1}}q^{1/2}T^{\frac{-1}{4\theta_1+2}}\left(\frac{T^{\frac{-\theta_1}{2\theta_1+1}}sQ^{1/2}}{4\sqrt{\lambda_{\max}\log(p)/T}}+1\right) + \xi_2\sqrt{\frac{s\gamma_{max}}{2}}.$$

Now, by Assumption 21, $\|\hat{g}_k\|_2 < \sqrt{\gamma_{\min}/2}$, and thus, applying Assumption 18 and Lemma 12 gives

$$\hat{g}_k^{\mathrm{T}}\left(\frac{1}{T}\int_0^T \Psi_k(t)\Psi_k^{\mathrm{T}}(t)dt\right)^{-1}\hat{g}_k \leq \frac{2\|\hat{g}_k\|_2^2}{\gamma_{\min}} < 1 \quad k\notin\mathcal{E}. \tag{C.45}$$

Therefore, we have established (C.33).

Given support recovery (C.38) and strict dual feasibility (C.45), results in Theorem 6 follows from the fact that, for all $j = 1, \ldots, p$, the above results hold jointly with probability at least $1 - c_4 p^{-1} - c_2 p^2 T \exp(-c_3 T^{1/5})$ as p and T grow given the condition in Theorem 5.

C.3 Proofs of Results in Section 4.4

In this section, we prove Theorems 7 and 8 in Chapter 4. As mentioned in Chapter 4, the building block of these proofs is a concentration inequality of $\left\|\widehat{V}_{j,k} - V_{j,k}\right\|_{2,[-B,B]}$ in Corollary 2 in Chapter 4. Therefore, we can focus our discussion given the event $\mathcal{M} \equiv \left\{\left\|\widehat{V}_{j,k} - V_{j,k}\right\|_{2,[-B,B]} \leq c_5 T^{-1/5}\right\}$, which holds with probability converging to unity given $p^2 T^{6/5} \exp\left(-c_6 T^{1/5}\right) = o(1)$ from Corollary 2.

Sketch of proofs:

Theorem 7: On one hand, we show that all neighbours of a node are in the same connected component with the node itself if we threshold $\left\|V_{j,k}\right\|_{2,[-B,B]}$ by $2\zeta = 2c_5 T^{-1/5}$. This is a consequence of the compatibility conditions and Beta-min condition, which we will examine in details in Section C.3.1. Under the event $\mathcal{M}$, we know that thresholding $\left\|\widehat{V}_{j,k}\right\|_{2,[-B,B]}$ by ζ gives supersets of the true connected components. On the other hand, we use the fact that $\|V_{j,k}\|_{2,[-B,B]} = 0$ for j and k not in the same connected component. Given event $\mathcal{M}$, we know that thresholding $\left\|\widehat{V}_{j,k}\right\|_{2,[-B,B]}$ by ζ gives susets of the true connected components. Combining these two statements, we know that the screened connected components agree with the true connected components.

Theorem 8: We use the Wiener-Hopf integral equation (Equation 4.6 in Chapter 4) to bound $\sum_{j,k} \|V_{j,k}\|^2_{2,[-B,B]}$. Under the event $\mathcal{M}$, however, only the edges that satisfy $\|V_{j,k}\|_{2,[-B,B]} \geq \zeta/2$ can be selected when thresholding $\left\|\widehat{V}_{j,k}\right\|_{2,[-B,B]}$ by ζ, where $\zeta = 2c_5 T^{-1/5}$. Hence, the number of edges in the set $\left\{(j,k) : \left\|\widehat{V}_{j,k}\right\|_{2,[-B,B]} > \zeta\right\}$ are inversely proportional to ζ^2. Then, we use Assumption 22 to show that all true edges are included since $\|V_{j,k}\|_{2,[-B,B]} >$

$\zeta = 2c_5 T^{-1/5}$ for sufficiently large T for the true edges. Combing these statements yields Theorem 8.

C.3.1 Implications of Assumptions 17 and Assumptions 21

Before proving Theorem 7, we need to rewrite the compatibility assumption and the Beta-min assumptions (Assumptions 17 and 21) in terms of $\boldsymbol{\omega}$ and $\boldsymbol{V}$.

Suppose now we are studying the neighbours of Node $s+1$. Let the oracle basis be chosen that $\psi = \boldsymbol{\omega}_{s+1,\cdot}$ and let $a \in \mathbb{R}^p$ be defined such that $a_k = 1$ for $k \in \mathcal{E}^*$ and $a_k = 0$ for $k \notin \mathcal{E}^*$, where $\mathcal{E}^*$ is an arbitrary subset of $\mathcal{E}_{s+1}$. We can see that the vector a satisfies the requirement in Assumption 17. It follows that

$$\xi_1 \left\{ \sum_{k \in \mathcal{E}^*} \left\{ \frac{1}{T} \int_0^T \mathbb{E} \left[\big(\omega_{s+1,k} * dN_k \big)(t) \right]^2 dt \right\}^{1/2} \right\}^2 \leq s \left\{ \frac{1}{T} \int_0^T \mathbb{E} \left[\sum_{k \in \mathcal{E}^*} \big(\omega_{s+1,k} * dN_k \big)(t) \right]^2 dt \right\}.$$

On the left-hand side, we have

$$\begin{aligned}
&\xi_1 \sum_{k \in \mathcal{E}^*} \left\{ \frac{1}{T} \int_0^T \mathbb{E} \int_0^\infty \omega_{s+1,k}(\Delta) dN_k(t-\Delta) d\Delta \int_0^\infty \omega_{s+1,k}(\Delta') dN_k(t-\Delta') d\Delta' dt \right\}^{1/2} \\
=&\xi_1 \sum_{k \in \mathcal{E}^*} \left\{ \frac{1}{T} \int_0^T \mathbb{E} \int_0^\infty \int_0^\infty \omega_{s+1,k}(\Delta) dN_k(t-\Delta) \omega_{s+1,k}(\Delta') dN_k(t-\Delta') d\Delta' d\Delta dt \right\}^{1/2} \\
=&\xi_1 \sum_{k \in \mathcal{E}^*} \left\{ \int_0^\infty \int_0^\infty \omega_{s+1,k}(\Delta) \mathbb{E} \left\{ \frac{1}{T} \int_0^T dN_k(t-\Delta) dN_k(t-\Delta') dt \right\} \omega_{s+1,k}(\Delta') d\Delta' d\Delta \right\}^{1/2} \\
=&\xi_1 \sum_{k \in \mathcal{E}^*} \left\{ \int_0^\infty \int_0^\infty \omega_{s+1,k}(\Delta) \big\{ V_{k,k}(\Delta - \Delta') + \Lambda_k \delta(\Delta - \Delta') + \Lambda_k^2 \big\} \omega_{s+1,k}(\Delta') d\Delta' d\Delta \right\}^{1/2},
\end{aligned}$$

where the last equality follows from the definition of cross-covariance. On the right-hand side, we

have

$$
\begin{aligned}
&\left\{\frac{1}{T}\int_0^T \mathbb{E}\left[\sum_{k\in\mathcal{E}^*}\big(\omega_{s+1,k} * dN_k\big)(t)\right]^2 dt\right\} \\
=&\left\{\frac{1}{T}\int_0^T \mathbb{E}\left[\int_0^\infty \boldsymbol{\omega}_{s+1,\mathcal{E}^*}^{\mathrm{T}}(\Delta) d\boldsymbol{N}(t-\Delta)d\Delta\right]\left[\int_0^\infty \boldsymbol{\omega}_{s+1,\mathcal{E}^*}^{\mathrm{T}}(\Delta') d\boldsymbol{N}(t-\Delta')d\Delta'\right] dt\right\} \\
=&\left\{\frac{1}{T}\int_0^T \mathbb{E}\left[\int_0^\infty \boldsymbol{\omega}_{s+1,\mathcal{E}^*}^{\mathrm{T}}(\Delta) d\boldsymbol{N}(t-\Delta)d\Delta\right]\left[\int_0^\infty \boldsymbol{\omega}_{s+1,\mathcal{E}^*}^{\mathrm{T}}(\Delta') d\boldsymbol{N}(t-\Delta')d\Delta'\right] dt\right\} \\
=&\int_0^\infty\int_0^\infty \boldsymbol{\omega}_{s+1,\mathcal{E}^*}^{\mathrm{T}}(\Delta)\mathbb{E}\left\{\frac{1}{T}\int_0^T d\boldsymbol{N}_{\mathcal{E}^*}(t-\Delta)d\boldsymbol{N}_{\mathcal{E}^*}^{\mathrm{T}}(t-\Delta')dt\right\}\boldsymbol{\omega}_{s+1,\mathcal{E}^*}(\Delta')d\Delta d\Delta' \\
=&\int_0^\infty\int_0^\infty \boldsymbol{\omega}_{s+1,\mathcal{E}^*}^{\mathrm{T}}(\Delta)\big\{\boldsymbol{V}_{\mathcal{E}^*,\mathcal{E}^*}(\Delta-\Delta') + \mathrm{diag}\big(\boldsymbol{\Lambda}_{\mathcal{E}^*}\big)\delta(\Delta-\Delta') + \boldsymbol{\Lambda}_{\mathcal{E}^*}\boldsymbol{\Lambda}_{\mathcal{E}^*}^{\mathrm{T}}\big\}\boldsymbol{\omega}_{s+1,\mathcal{E}^*}(\Delta')d\Delta d\Delta',
\end{aligned}
$$

where the last equality follows from the definition of cross-covariance. In summary, the compatibility assumption gives that

$$
\begin{aligned}
&\xi_1\left\{\sum_{k\in\mathcal{E}^*}\left\{\int_0^\infty\int_0^\infty \omega_{s+1,k}(\Delta)\big\{V_{k,k}(\Delta-\Delta')+\Lambda_k\delta(\Delta-\Delta')+\Lambda_k^2\big\}\omega_{s+1,k}(\Delta')d\Delta' d\Delta\right\}^{1/2}\right\}^2 \\
\le& s\int_0^\infty\int_0^\infty \boldsymbol{\omega}_{s+1,\mathcal{E}^*}^{\mathrm{T}}(\Delta)\big\{\boldsymbol{V}_{\mathcal{E}^*,\mathcal{E}^*}(\Delta-\Delta') + \mathrm{diag}\big(\boldsymbol{\Lambda}_{\mathcal{E}^*}\big)\delta(\Delta-\Delta') + \boldsymbol{\Lambda}_{\mathcal{E}^*}\boldsymbol{\Lambda}_{\mathcal{E}^*}^{\mathrm{T}}\big\}\boldsymbol{\omega}_{s+1,\mathcal{E}^*}(\Delta')d\Delta d\Delta',
\end{aligned}
\tag{C.46}
$$

which holds for any $\mathcal{E}^* \subset \mathcal{E}_{s+1}$.

The *Beta-min* assumption (Assumption 21 in Chapter 4) places a lower bound on every entry in the left-hand side of (C.46). To see this, consider the following equation, for $k \in \mathcal{E}_{s+1}$,

$$
\begin{aligned}
&\frac{1}{T}\int_0^T \mathbb{E}\left[\int_0^\infty \omega_{s+1,k}(\Delta) dN_k(t-\Delta)\right]\left[\int_0^\infty \omega_{s+1,k}(\Delta) dN_k(t-\Delta')\right] dt \\
=&\int_0^\infty\int_0^\infty \omega_{s+1,k}(\Delta)\big\{V_{k,k}(\Delta-\Delta')+\Lambda_k\delta(\Delta-\Delta')+\Lambda_k^2\big\}\omega_{s+1,k}(\Delta')d\Delta' d\Delta.
\end{aligned}
$$

We can rewrite the statement of Assumption 21 in Chapter 4 as

$$
\int_0^\infty\int_0^\infty \omega_{s+1,k}(\Delta)\big\{V_{k,k}(\Delta-\Delta')+\Lambda_k\delta(\Delta-\Delta')+\Lambda_k^2\big\}\omega_{s+1,k}(\Delta')d\Delta' d\Delta \ge \beta_{\min}^2, \tag{C.47}
$$

which holds for $(j,k) \in \mathcal{E}$.

In summary, the compatibility and Beta-min conditions imply that

$$\begin{aligned}&\xi_1\beta_{\min}^2\text{card}(\mathcal{E}^*)\\ \leq& s\int_0^\infty\int_0^\infty \boldsymbol{\omega}_{s+1,\mathcal{E}^*}^{\mathrm{T}}(\Delta)\big\{\boldsymbol{V}_{\mathcal{E}^*,\mathcal{E}^*}(\Delta-\Delta')+\text{diag}\big(\boldsymbol{\Lambda}_{\mathcal{E}^*}\big)\delta(\Delta-\Delta')+\boldsymbol{\Lambda}_{\mathcal{E}^*}\boldsymbol{\Lambda}_{\mathcal{E}^*}^{\mathrm{T}}\big\}\boldsymbol{\omega}_{s+1,\mathcal{E}^*}(\Delta')d\Delta d\Delta',\end{aligned} \tag{C.48}$$

which holds for any $\mathcal{E}^*\subset\mathcal{E}_{s+1}$.

In addition, we will need the following inequality in the next section.

$$\frac{\xi_1\beta_{\min}^2\text{card}(\mathcal{E}^*)}{\text{card}(\mathcal{E}^*)s\lambda_{\max}b+s}\leq\iint \boldsymbol{\omega}_{s+1,\mathcal{E}^*}^{\mathrm{T}}(\Delta)\big\{\boldsymbol{V}_{\mathcal{E}^*,\mathcal{E}^*}(\Delta-\Delta')+\text{diag}\big(\boldsymbol{\Lambda}_{\mathcal{E}^*}\big)\delta(\Delta-\Delta')\big\}\boldsymbol{\omega}_{s+1,\mathcal{E}^*}(\Delta')d\Delta d\Delta'. \tag{C.49}$$

To see this, first verify that

$$\begin{aligned}&\int_0^\infty\int_0^\infty \boldsymbol{\omega}_{s+1,\mathcal{E}^*}^{\mathrm{T}}(\Delta)\boldsymbol{\Lambda}_{\mathcal{E}^*}\boldsymbol{\Lambda}_{\mathcal{E}^*}^{\mathrm{T}}\boldsymbol{\omega}_{s+1,\mathcal{E}^*}(\Delta')d\Delta d\Delta'\\ =&\left[\sum_{k\in\mathcal{E}^*}\Lambda_k\int_0^\infty\omega_{s+1,k}(\Delta)d\Delta\right]^2\\ \leq&\left\{\sum_{k\in\mathcal{E}^*}\Lambda_k\left[\int_0^\infty\omega_{s+1,k}(\Delta)d\Delta\right]^2\right\}\left\{\sum_{k\in\mathcal{E}^*}\Lambda_k\right\}\\ \leq&\text{card}(\mathcal{E}^*)\lambda_{\max}\left\{\sum_{k\in\mathcal{E}^*}\Lambda_k\left[\int_0^\infty\omega_{s+1,k}(\Delta)d\Delta\right]^2\right\}\\ \leq&\text{card}(\mathcal{E}^*)\lambda_{\max}\left\{\sum_{k\in\mathcal{E}^*}\Lambda_k b\int_0^\infty\omega_{s+1,k}^2(\Delta)d\Delta\right\}\\ =&\text{card}(\mathcal{E}^*)\lambda_{\max}b\int_0^\infty\sum_{k\in\mathcal{E}^*}\Lambda_k\omega_{s+1,k}^2(\Delta)d\Delta\\ =&\text{card}(\mathcal{E}^*)\lambda_{\max}b\int_0^\infty\boldsymbol{\omega}_{s+1,\mathcal{E}^*}^{\mathrm{T}}(\Delta)\text{diag}(\boldsymbol{\Lambda}_{\mathcal{E}^*})\boldsymbol{\omega}_{s+1,\mathcal{E}^*}(\Delta)d\Delta\\ =&\text{card}(\mathcal{E}^*)\lambda_{\max}b\int_0^\infty\int_0^\infty\boldsymbol{\omega}_{s+1,\mathcal{E}^*}^{\mathrm{T}}(\Delta)\text{diag}(\boldsymbol{\Lambda}_{\mathcal{E}^*})\delta(\Delta-\Delta')\boldsymbol{\omega}_{s+1,\mathcal{E}^*}(\Delta')d\Delta d\Delta',\end{aligned}$$

where the first inequality follows from the Cauchy-Schwartz inequality, the second from the fact that $\Lambda_k\leq\lambda_{\max}$, and the third from Cauchy-Schwartz inequality on $\int\omega_{j,k}(\Delta)d\Delta$ and $\text{supp}(\omega_{j,k})=[0,b]$ from Assumption 19. Also, we know that

$$\iint\boldsymbol{\omega}_{s+1,\mathcal{E}^*}^{\mathrm{T}}(\Delta)\boldsymbol{V}_{\mathcal{E}^*,\mathcal{E}^*}(\Delta-\Delta')\boldsymbol{\omega}_{s+1,\mathcal{E}^*}(\Delta')d\Delta d\Delta'\geq 0,$$

because

$$
\begin{aligned}
&\iint \boldsymbol{\omega}_{s+1,\mathcal{E}^*}^{\mathrm{T}}(\Delta)\boldsymbol{V}_{\mathcal{E}^*,\mathcal{E}^*}(\Delta-\Delta')\boldsymbol{\omega}_{s+1,\mathcal{E}^*}(\Delta')d\Delta d\Delta' \\
=&\iint \boldsymbol{\omega}_{s+1,\mathcal{E}^*}^{\mathrm{T}}(\Delta)\mathbb{E}\left\{[d\boldsymbol{N}_{\mathcal{E}^*}(\Delta)/d\Delta-\boldsymbol{\Lambda}_{\mathcal{E}^*}][d\boldsymbol{N}_{\mathcal{E}^*}(\Delta')/d\Delta'-\boldsymbol{\Lambda}_{\mathcal{E}^*}]^{\mathrm{T}}\right\}\boldsymbol{\omega}_{s+1,\mathcal{E}^*}(\Delta')d\Delta d\Delta' \\
=&\mathbb{E}\left\{\left[\int_0^\infty \boldsymbol{\omega}_{s+1,\mathcal{E}^*}^{\mathrm{T}}(\Delta)\big(d\boldsymbol{N}_{\mathcal{E}^*}(\Delta)-\boldsymbol{\Lambda}_{\mathcal{E}^*}d\Delta\big)\right]\left[\int_0^\infty \boldsymbol{\omega}_{s+1,\mathcal{E}^*}^{\mathrm{T}}(\Delta')\big(d\boldsymbol{N}_{\mathcal{E}^*}(\Delta')-\boldsymbol{\Lambda}_{\mathcal{E}^*}d\Delta'\big)\right]\right\} \\
\geq&0.
\end{aligned}
$$

Finally, adding $\mathrm{card}(\mathcal{E}^*)s\lambda_{\max}b\iint \boldsymbol{\omega}_{s+1,\mathcal{E}^*}^{\mathrm{T}}(\Delta)\boldsymbol{V}_{\mathcal{E}^*,\mathcal{E}^*}(\Delta-\Delta')\boldsymbol{\omega}_{s+1,\mathcal{E}^*}(\Delta')d\Delta d\Delta'$ on the right-hand side of (C.46) gives (C.49).

C.3.2 Proof of Theorem 7

Proof. In this proof, we assume that B is chosen sufficiently large so that $B \geq b$. To show the main result, we prove that following two statements hold under the event $\mathcal{M} \equiv \{\|\widehat{V}_{j,k} - V_{j,k}\|_{2,[-B,B]} \leq c_5 T^{-1/5}\}$.

i Every screened connected component is contained in one of the screened connected component, i.e., $\forall l$, $\exists l'$ such that $\widehat{\mathcal{C}}_l(\zeta) \subset \mathcal{C}_{l'}$.

ii Every connected component is contained in one of the screened connected component, i.e., if $j, k \in \mathcal{C}_l$ for some l, there exists an l' such that $j, k \in \widehat{\mathcal{C}}_{l'}(\zeta)$.

Together these imply that $\big\{\widehat{\mathcal{C}}_l(\zeta)\big\}_{l=1}^p = \big\{\mathcal{C}_l\big\}_{l=1}^p$.

Statement (i)

For any two connected components $\mathcal{C}_l$ and $\mathcal{C}_{l'}$, $l \neq l'$, we have $V_{j,k} \equiv 0$ for any $j \in \mathcal{C}_l$ and $k \in \mathcal{C}_{l'}$. Given the event $\mathcal{M}$, we have that $\big\|\widehat{V}_{j,k}\|_{2,[-B,B]} \leq 0 + \big\|\widehat{V}_{j,k} - V_{j,k}\big\|_{2,[-B,B]} \leq c_5 T^{-1/5}) \leq \zeta$, which holds for any pairs of $j \in \mathcal{C}_l$ and $k \in \mathcal{C}_{l'}$ of nodes. As a result, there are no screened edges between nodes in $\mathcal{C}_l$ and $\mathcal{C}_{l'}$. Applying this argument on all $l' \neq l$, we know that the nodes in $\mathcal{C}_l$ are isolated from the rest of the nodes using the threshold ζ. Hence, $\widehat{\mathcal{C}}_l(\zeta) \subset \mathcal{C}_{l'}$.

Statement (ii)

We will show that, for any node, its neighbours belong to the same screened connected component with itself given the event $\mathcal{M}$. This implies that, for any l, there exists an l' such that $\mathcal{C}_l \subset \widehat{\mathcal{C}}_{l'}(\zeta)$.

We prove by contradiction. Suppose that Nodes $1, \ldots, s$ are the neighbours of Node $s+1$, i.e., $\omega_{s+1,i} \neq 0$ for $i = 1, \ldots, s$. Given the event $\mathcal{M}$, $\|V_{j,k}\|_{2,[-B,B]} > 2c_5 T^{-1/5}$ means that Node j and Node k are in the same connected component. Assume that Node 1 and Node s+1 are not in the same connected component. As a direct consequence of this assumption, we have $\|\widehat{V}_{s+1,1}\|_{2,[-B,B]} \leq \zeta = c_5 T^{-1/5}$. Given the event $\mathcal{M}$, we also know that

$$\|V_{s+1,1}\|_{2,[-B,B]} \leq 2c_5 T^{-1/5}. \tag{C.50}$$

Consider the Wiener-Hopf integral equation (4.6) in Chapter 4, for $V_{s+1,1}$.

$$V_{s+1,1}(\Delta) = \omega_{s+1,1} * (V_{1,1} + \Lambda_1 \delta)(\Delta) + \sum_{l=2}^{s} \big[\omega_{s+1,l} * V_{l,1}\big](\Delta). \tag{C.51}$$

Multiplying $\omega_{s+1,1}(\Delta)$ on both sides of the equation and integral it on $[0, b]$ gives

$$\begin{aligned} \underbrace{\int_0^b \omega_{s+1,1}(\Delta) V_{s+1,1}(\Delta) d\Delta}_{\text{I}} = &\underbrace{\int_0^b \omega_{s+1,1}(\Delta)\, \omega_{s+1,1} * (V_{1,1} + \Lambda_1 \delta)(\Delta) d\Delta}_{\text{II}} + \\ &\underbrace{\int_0^b \omega_{s+1,1}(\Delta) \big[\boldsymbol{\omega}_{s+1,2:(s+1)} * \boldsymbol{V}_{2:(s+1),1}\big](\Delta) d\Delta}_{\text{III}}, \end{aligned} \tag{C.52}$$

where we use the fact that $\text{supp}(\omega_{j,k}) = [0, b]$ and $B \geq b$.

Applying the Cauchy-Schwarz inequality on I gives

$$\text{I} \leq \left[\int_0^b \omega_{s+1,1}^2(\Delta) d\Delta\right]^{1/2} \left[\int_0^b V_{s+1,1}^2(\Delta) d\Delta\right]^{1/2}.$$

From Assumption 19, we know that $\omega_{j,k}$ belongs a Sobolev class $W(\theta_1, L_1)$ on a bounded support $[0, b]$, which implies that $\{\omega_{j,k}\}$ are bounded. Letting C be the upper bound of $\omega_{j,k}$, we have

$$\int_0^b \omega_{s+1,1}^2(\Delta) d\Delta \leq \int_0^b \|\omega_{s+1,1}\|_\infty |\omega_{s+1,1}(\Delta)| d\Delta \leq C\Omega_{s+1,1}.$$

Therefore,

$$\mathrm{I} \leq 2C^{1/2}\Omega_{s+1,1}^{1/2}c_5T^{-1/5} \leq 2C^{1/2}\gamma_{\Omega}^{1/2}c_5T^{-1/5}, \tag{C.53}$$

where we use (C.50).

On the right-hand side, the first term II can rewritten as, from (C.47) and (C.49),

$$\mathrm{II} = \iint \omega_{s+1,1}(\Delta)(V_{1,1} + \Lambda_1\delta)(\Delta - \Delta')\omega_{s+1,1}(\Delta')d\Delta' d\Delta \geq (s\lambda_{\max}b + s)^{-1}\xi_1\beta_{\min}^2. \tag{C.54}$$

Now use the triangle inequailty that

$$|\mathrm{I}| = |\mathrm{II} + \mathrm{III}| \geq |\mathrm{II}| - |\mathrm{III}|.$$

Therefore, we have that

$$|\mathrm{III}| = \left| \int_0^b \omega_{s+1,1}(\Delta)\big[\boldsymbol{\omega}_{s+1,2:(s+1)} * \boldsymbol{V}_{2:(s+1),1}\big](\Delta)d\Delta \right| \geq (s\lambda_{\max}b+s)^{-1}\xi_1\beta_{\min}^2 - 2C^{1/2}\gamma_{\Omega}^{1/2}c_5T^{-1/5}. \tag{C.55}$$

We extract all the $l \in \{2, \ldots, s\}$ that satisfies

$$\int_0^b \omega_{s+1,1}(\Delta)\omega_{s+1,l} * V_{l,1}(\Delta)d\Delta > 2C^{1/2}\zeta\gamma_{\Omega}^{3/2}, \tag{C.56}$$

and we denote the set as $\mathrm{ne}(1)$. We know that $\mathrm{ne}(1)$ is non-empty set for sufficiently large T, since $\zeta = c_5T^{-1/5} = o(\beta_{\min}^2)$ as $T \to \infty$. Without loss of generality, we assume that $\mathrm{ne}(1) = \{2\}$. This proof applies directly if the set $\mathrm{ne}(1)$ has more than one element. The sum of remaining terms in III is bounded from above by

$$\left| \int_0^b \omega_{s+1,1}(\Delta) \sum_{l \in \{3:(s+1)\}} \big[\omega_{s+1,l} * V_{l,1}\big](\Delta)d\Delta \right| \leq 2(s-1)C^{1/2}c_5T^{-1/5}\gamma_{\Omega}^{3/2}. \tag{C.57}$$

For Node 2, we can see that, from (C.56)

$$\begin{aligned} 2C^{1/2}\zeta\gamma_{\Omega}^{3/2} &< \int_0^b \omega_{s+1,1}(\Delta)\omega_{s+1,2} * V_{2,1}(\Delta)d\Delta \\ 2C^{1/2}\zeta\gamma_{\Omega}^{3/2} &< \left[\int_0^b \omega_{s+1,1}^2(\Delta)d\Delta\right]^{1/2} \left[\int_0^b \left|\omega_{s+1,2} * V_{2,1}(\Delta)\right|^2 d\Delta\right]^{1/2}, \end{aligned}$$

where we use the Cauchy-Schwartz inequality. Using the fact that

$$\int_0^b \omega_{s+1,1}^2(\Delta)d\Delta < \int_0^b \|\omega_{s+1,1}\|_\infty |\omega_{s+1,1}(\Delta)|d\Delta \leq C\gamma_\Omega,$$

we get

$$2\zeta\gamma_\Omega < \left[\int_0^b \left|\omega_{s+1,2} * V_{2,1}(\Delta)\right|^2 d\Delta\right]^{1/2}.$$

Then, Young's inequality of convolution gives

$$\left[\int_0^b \left|\omega_{s+1,2} * V_{2,1}(\Delta)\right|^2 d\Delta\right]^{1/2} \leq \int_0^b |\omega_{s+1,2}(\Delta)|d\Delta \left[\int_0^b V_{2,1}^2(\Delta)d\Delta\right]^{1/2}.$$

Finally, using the fact that $\int_0^b |\omega_{s+1,2}(\Delta)|d\Delta = \Omega_{s+1,2} \leq \gamma_\Omega$, we have

$$2\zeta < \left\|V_{2,1}(\Delta)\right\|_{2,[-B,B]}. \tag{C.58}$$

By our choice of threshold, (C.58) means that Nodes 1 and 2 are in the same screened connected components,

We now look at the cross-covariance between Node $s+1$ and Nodes 1 and 2, where

$$\begin{pmatrix} V_{s+1,1} \\ V_{s+1,2} \end{pmatrix}(\Delta) = \begin{pmatrix} V_{1,1} + \Lambda_1\delta & V_{1,2} \\ V_{2,1} & V_{2,2} + \Lambda_2\delta \end{pmatrix} * \begin{pmatrix} \omega_{s+1,1} \\ \omega_{s+1,2} \end{pmatrix}(\Delta) + \begin{pmatrix} \boldsymbol{V}_{1,3:(s+1)} \\ \boldsymbol{V}_{2,3:(s+1)} \end{pmatrix} * \boldsymbol{\omega}_{s+1,3:(s+1)}(\Delta). \tag{C.59}$$

If Node 2 is in the same screened connected component with Node $s+1$, then Node 1 is also in the same screened connected component with $s+1$, which contradicts the assumption.

If Node 2 is not in the same connected component with Node $s+1$, then we know that $\|V_{s+1,2}\|_{2,[-B,B]} \leq 2c_5T^{-1/5}$. Now multiplying $\boldsymbol{\omega}_{s+1,1:2}^{\mathrm{T}}(\Delta)$ on both sides of the equation and integrating them over $[0,b]$.

$$\begin{aligned} &\underbrace{\int_0^b \boldsymbol{\omega}_{s+1,1:2}(\Delta)\cdot\boldsymbol{V}_{s+1,1:2}(\Delta)d\Delta}_{\text{I}} \\ &= \underbrace{\int_0^b \boldsymbol{\omega}_{s+1,1:2}^{\mathrm{T}}(\Delta)\big(\boldsymbol{V}_{1:2,1:2} + \mathrm{diag}(\boldsymbol{\Lambda}_{1:2})\delta\big) * \boldsymbol{\omega}_{s+1,1:2}(\Delta)d\Delta}_{\text{II}} \\ &\quad + \underbrace{\int_0^b \boldsymbol{\omega}_{s+1,1:2}^{\mathrm{T}}(\Delta)\boldsymbol{V}_{1:2,3:(s+1)} * \boldsymbol{\omega}_{s+1,3:(s+1)}(\Delta)d\Delta}_{\text{III}}. \end{aligned} \tag{C.60}$$

As in (C.53), I is bounded from above

$$\int_0^b \boldsymbol{\omega}_{s+1,1:2}(\Delta) \cdot \boldsymbol{V}_{s+1,1:2}(\Delta) d\Delta \leq 4\gamma_\Omega^{1/2} C^{1/2} c_5 T^{-1/5}. \tag{C.61}$$

And II is lower bounded using (C.49) with $\mathcal{E}^* = \{1, 2\}$

$$\begin{aligned}
\text{II} =& \int_0^b \boldsymbol{\omega}_{s+1,1:2}^{\mathrm{T}}(\Delta)\big(\boldsymbol{V}_{1:2,1:2} + \operatorname{diag}(\boldsymbol{\Lambda}_{1:2})\delta\big) * \boldsymbol{\omega}_{s+1,1:2}(\Delta) d\Delta \\
\geq& (2s\lambda_{\max} b + s)^{-1} \xi_1 \sum_{l=1}^{2} \int_0^b \int_0^b \omega_{s+1,l}(\Delta)(V_{l,l} + \Lambda_l \delta)(\Delta - \Delta')\omega_{s+1,l}(\Delta') d\Delta' d\Delta \\
\geq& 2(2s\lambda_{\max} b + s)^{-1} \xi_1 \beta_{\min}^2.
\end{aligned}$$

We use the triangle inequality to get that $|\text{III}| \geq |\text{II}| - |\text{I}|$. Hence,

$$\begin{aligned}
\left| \int_0^b \boldsymbol{\omega}_{s+1,1:2}^{\mathrm{T}}(\Delta) \boldsymbol{V}_{1:2,3:(s+1)} * \boldsymbol{\omega}_{s+1,3:(s+1)}(\Delta) d\Delta \right| \geq& 2(2s\lambda_{\max} b + s)^{-1} \xi_1 \beta_{\min}^2 - 4\gamma_\Omega^{1/2} C^{1/2} c_5 T^{-1/5} \\
\sum_{k=1}^{2} \left| \int_0^b \omega_{s+1,k}(\Delta) \sum_{l \in \{3:(s+1)\}} \big[\omega_{s+1,l} * V_{l,k}\big](\Delta) d\Delta \right| \geq& 2(2s\lambda_{\max} b + s)^{-1} \xi_1 \beta_{\min}^2 - 4\gamma_\Omega^{1/2} C^{1/2} c_5 T^{-1/5},
\end{aligned}$$

where we use the triangle inequality $|a + b| \leq |a| + |b|$. From (C.57), we know that

$$\begin{aligned}
&\left| \int_0^b \omega_{s+1,2}(\Delta) \sum_{l \in \{3:(s+1)\}} \big[\omega_{s+1,l} * V_{l,2}\big](\Delta) d\Delta \right| \\
\geq& 2(2s\lambda_{\max} b + s)^{-1} \xi_1 \beta_{\min}^2 - 4\gamma_\Omega^{1/2} C^{1/2} c_5 T^{-1/5} - 2(s-1) C^{1/2} c_5 T^{-1/5} \gamma_\Omega^{3/2}.
\end{aligned} \tag{C.62}$$

Again, we extract all the $l \in \{3, \ldots, s\}$ that satisfies

$$\int_0^b \omega_{s+1,l}(\Delta) \omega_{s+1,l} * V_{l,2}(\Delta) d\Delta > 2C^{1/2} \zeta \gamma_\Omega^{3/2}. \tag{C.63}$$

We can show that this new group $\text{ne}(2)$ is non-empty and $\text{ne}(2)$ belongs to the same connected components with Node 2 using the same argument as for $\text{ne}(1)$. We can repeat the process until there are no remaining nodes in $\{2, \ldots, s\}$ are in the same screened connected component with Node 1. And these nodes cannot be in the same screened connected component with Node $s + 1$, which means that, for all $l \in \{1, \ldots, s\}$,

$$\int_0^b \omega_{s+1,l}(\Delta) V_{s+1,l}(\Delta) d\Delta \leq 2C^{1/2} \gamma_\Omega^{1/2} c_5 T^{-1/5}. \tag{C.64}$$

Recall the Wiener-Hopf integral equation (4.6)

$$\boldsymbol{V}_{s+1,\cdot}(\Delta) = \big[(\boldsymbol{V} + \text{diag}(\boldsymbol{\Lambda})\delta) * \boldsymbol{\omega}_{s+1,\cdot}\big](\Delta). \tag{C.65}$$

Again, we multiply $\boldsymbol{\omega}_{s+1,\cdot}^{\mathrm{T}}(\Delta)$ on both sides of the equation and integral them over $[0, b]$. On the right-hand side of the integrated equation, we have that

$$\int_0^b \boldsymbol{\omega}_{s+1,\cdot}^{\mathrm{T}}(\Delta)\big(\boldsymbol{V} + \text{diag}(\boldsymbol{\Lambda})\delta\big) * \boldsymbol{\omega}_{s+1,\cdot}(\Delta)d\Delta \geq s(s^2\lambda_{\max}b + s)^{-1}\xi_1, \tag{C.66}$$

where we use (C.49) with $\mathcal{E}^* = \{1, \ldots, s\}$. However, the left-hand side is upper bounded by $2sC^{1/2}\gamma_\Omega^{1/2}c_5T^{-1/5}$, which is a contradiction since $T^{-1/5} = o(\beta_{\min}^2)$.

To complete the proof, we examine the probability that the event $\mathcal{M}$ holds. From Corollary 2, we know that the event $\mathcal{M}$ holds probability converging to unity if

$$p^2T^{6/5}\exp\left(-c_6T^{1/5}\right) = o(1).$$

□

C.3.3 Proof of Theorem 8

Proof. Again, assume that the event $\mathcal{M} \equiv \{\|\widehat{V}_{j,k} - V_{j,k}\|_{2,[-B,B]} \leq c_5T^{-1/5}\}$ holds.

We first show that the size of the screened edge set is on the scale $sp\gamma_\Omega^2(1-\gamma_\Omega)^{-2}\lambda_{\max}^2T^{5/2}$.

Recall the Wiener-Hopf integral equation

$$\boldsymbol{V}(\Delta) = \boldsymbol{\omega}(\Delta)\text{diag}(\boldsymbol{\Lambda}) + (\boldsymbol{\omega} * \boldsymbol{V})(\Delta). \tag{C.67}$$

For each (j, k), we can see that

$$V_{j,k} = \omega_{j,k}\Lambda_k + \boldsymbol{\omega}_{j,\cdot} * \boldsymbol{V}_{\cdot,k}. \tag{C.68}$$

We have

$$\begin{aligned}
\|V_{j,k}\|_{2,[-B,B]} &\leq \Lambda_k\|\omega_{j,k}\|_{2,[-B,B]} + \|\boldsymbol{\omega}_{j,\cdot} * \boldsymbol{V}_{\cdot,k}\|_{2,[-B,B]} \\
&\leq \Lambda_k\|\omega_{j,k}\|_{2,[-B,B]} + \sum_{l=1}^{p}\|\omega_{j,l} * V_{l,k}\|_{2,[-B,B]} \\
&\leq \Lambda_k\|\omega_{j,k}\|_{2,[-B,B]} + \sum_{l=1}^{p}\|V_{l,k}\|_{2,[-B,B]}\|\omega_{j,l}\|_{1,[-B,B]},
\end{aligned} \tag{C.69}$$

where the inequalities follow from triangle inequality, norm inequality, and Young's inequality. Young's inequality takes the form that

$$\|f * g\|_r \leq \|f\|_p \|g\|_q, \quad \frac{1}{p} + \frac{1}{q} = \frac{1}{r} + 1, \tag{C.70}$$

where $\|f\|_p \equiv \left[\int_{-\infty}^{\infty} f^p(x)dx\right]^{1/p}$. Here we let $r = p = 2$, $q = 1$, $f = \mathbf{1}_{[-B,B]}V_{l,k}$, and $g = \omega_{j,l}$. Recall that supp$(\omega_{j,l}) = [0, b] \subset [-B, B]$.

From Assumption 19, we know that $\omega_{j,k}$ belongs a Sobolev class $W(\theta_1, L_1)$ on a bounded support $[0, b]$, which implies that $\{\omega_{j,k}\}$ are bounded. Letting C be the bound, we have

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

Let $\bar{V}_{j,k} \equiv \|V_{j,k}\|_{2,[-B,B]}$. We have

$$\bar{V}_{j,k} \leq C^{1/2}\Omega_{j,k}^{1/2}\Lambda_k + \mathbf{\Omega}_{j,\cdot} \cdot \bar{\boldsymbol{V}}_{\cdot,k}. \tag{C.71}$$

The ℓ_2-norm of the vector $\bar{\boldsymbol{V}}_{\cdot,k}$ can be bounded using the triangle inequality

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

We know from Assumption 15 that

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

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

$$\|\bar{\boldsymbol{V}}_{\cdot,k}\|_2 \leq C^{1/2}(1 - \gamma_\Omega)^{-1}\lambda_{\max} \|\mathbf{\Omega}_{\cdot,k}\|_1^{1/2}. \tag{C.72}$$

Hence,

$$\sum_{j,k} \bar{V}_{j,k}^2 = \sum_k \|\bar{\boldsymbol{V}}_{\cdot,k}\|_2^2 \leq C(1 - \gamma_\Omega)^{-2}\lambda_{\max}^2 \sum_{k=1}^{p} \|\mathbf{\Omega}_{\cdot,k}\|_1. \tag{C.73}$$

Further recall that the number of non-zero elements in $\Omega_{\cdot,k}$ is upper bounded by s and $\Omega_{j,k} \leq \gamma_\Omega$. The inequality becomes

$$\sum_{j,k} \bar{V}_{j,k}^2 \leq C(1 - \gamma_\Omega)^{-2}\lambda_{\max}^2 sp\gamma_\Omega^2. \tag{C.74}$$

Hence, among the $\bar{\boldsymbol{V}}$, there are at most $(C/c_5)spT^{5/2}\gamma_\Omega^2(1-\gamma_\Omega)^{-2}\lambda_{\max}^2$ elements are on the scale $c_5T^{-1/5}$ or higher.

Given event $\mathcal{M}$, only edges in the set $\{(j,k) : \|V_{j,k}\|_{2,[-B,B]} \geq c_5T^{-1/5}\}$ can be selected given the threshold $\zeta = 2c_5T^{-1/5}$. This implies that the size of the selected set is on the scale of $spT^{5/2}\gamma_\Omega^2(1-\gamma_\Omega)^{-2}\lambda_{\max}^2$.

For Claim (b), we know, from Assumption 22, that $\|V_{j,k}\|_{2,[-B,B]} \geq c_6T^{-\kappa}$ for $(j,k) \in \mathcal{E}$, where $\kappa < 1/5$. Hence, for a sufficiently large T, we have $\|\widehat{V}_{j,k}\|_{2,[-B,B]} > 2c_5T^{-1/5} = \zeta$ for $(j,k) \in \mathcal{E}$ given the event $\mathcal{M}$. Therefore, we have that $\mathcal{E}_j \subset \widehat{\mathcal{E}}_j^{ss}(\zeta)$.

To complete the proof, we examine the probability that the event $\mathcal{M}$ holds. From Corollary 2, we know that the event $\mathcal{M}$ holds probability converging to unity if $p^2T^{6/5}\exp\left(-c_6T^{1/5}\right) = o(1)$. □

C.4 Proofs of Results in Section 4.5

In this section, we provide the proofs of Theorem 9 and Corollary 2 in Chapter 4. The theorem gives a concentration inequality of the smoothing estimator $\bar{y}_{j,k}$ taking the form (Equation 4.29 in Chapter 4)

$$\bar{y}_{j,k} \equiv \frac{1}{T}\int_0^T \int_0^T f(t-t')dN_j(t')dN_k(t).$$

In Appendix C.4.1, we discuss the coupling construction for $d\boldsymbol{N}$ and establish its properties. In Appendix C.4.2, we state and prove a corollary that helps establish Theorem 9. In Appendix C.4.3, we show that $\bar{y}_{j,k}$ can be written as the mean of a dependent sequence, and obtain bounds on the τ-dependence coefficient of this sequence. In Appendix C.4.4, we apply the result in Merlevède et al. (2011) to derive a concentration inequality for $\bar{y}_{j,k}$. The remaining two sections are devoted to prove Corollary 2. In Appendix C.4.5, we show that the true cross-covariate $V_{j,k}$ is smooth under Assumption 19 in Chapter 4, which helps bound the bias in kernel smoothing. In Appendix C.4.6, we prove Corollary 2 by combining the pieces together.

C.4.1 Coupling Construction of the Hawkes process

In this section, we construct a coupling process for the multivariate Hawkes process that will be useful in showing Proposition 6 in the next section. The key to the construction is the Poisson embedding technique; we refer the readers to Brémaud and Massoulié (1996) for a details (see Lemma 3 for the univariate case and the proof of Theorem 7 for the multivariate case). Note that, when the transfer functions are assumed to be non-negative (i.e., $\omega_{j,k} \geq 0$ for all j, k), a simple construction is available using the Poisson branching process representation (Hawkes and Oakes, 1974).

The main result is summarized in the following theorem.

Theorem 11. *Suppose that Assumptions 15 and 19 hold for the point process $\boldsymbol{N}$. For a given $z > 0$, there exists a point process $\widetilde{\boldsymbol{N}}$ such that $\widetilde{\boldsymbol{N}}$ has the same distribution as $\boldsymbol{N}$, and $\widetilde{\boldsymbol{N}}$ is independent of the history of $\boldsymbol{N}$ up to time z (i.e., $\mathcal{H}_z$). Moreover, for $(n-1)b < u$, $n = 1, 2, \ldots$*

$$\mathbb{E}\big|d\widetilde{\boldsymbol{N}}(z+u) - d\boldsymbol{N}(z+u)\big|/du \preceq 2v\boldsymbol{\Omega}^n\mathbf{1},$$

and

$$\mathbb{E}\big|d\widetilde{\boldsymbol{N}}(t')d\widetilde{\boldsymbol{N}}(z+u) - d\boldsymbol{N}(t')d\boldsymbol{N}(z+u)\big|/(dudt') \preceq 2v^2\sum_{i=1}^{n+1}\boldsymbol{\Omega}^{i-1}[\mathbf{1}\mathbf{1}^{\mathrm{T}}]\big[\boldsymbol{\Omega}^{\mathrm{T}}\big]^{n-i+1},$$

where $\preceq$ denotes the element-wise inequality, and v is a constant that satisfies $v^2 = \max_{j,k}(\|V_{j,k}\|_\infty + \Lambda_j\Lambda_k, (\mu_j + \|\omega_{j,k}\|_\infty]^2)$.

Loosely speaking, $\mathbb{E}\big|d\widetilde{\boldsymbol{N}}(z+u) - d\boldsymbol{N}(z+u)\big|/du$ captures the temporal dependence in the first-moment, and $\mathbb{E}\big|d\widetilde{\boldsymbol{N}}(t')d\widetilde{\boldsymbol{N}}(z+u) - d\boldsymbol{N}(t')d\boldsymbol{N}(z+u)\big|/(dudt')$ characterizes the temporal dependence in the second-moment (see Proposition 6 in the next section or Lemma 3 in Dedecker and Prieur (2004)).

Proof. Following Brémaud and Massoulié (1996), we first construct a sequence of point processes that converge to $\boldsymbol{N}$. Let $\boldsymbol{N}^{(0)}$ be a p-variate Poisson process with intensity $\lambda_{\max}$, where $\lambda_{\max}$ is a large constant. For $n = 1$, construct a process $\boldsymbol{N}^{(1)}(t)$ as

$$dN_j^{(1)}(t) = \mathbf{1}_{[r_j^{(1)}(t) \leq \mu_j/\lambda_{\max}]}dN_j^{(0)}(t), \quad j = 1, \ldots, p, \tag{C.75}$$

where $r_j^{(1)}(t) \sim \text{Unif}([0,1])$. For $n \geq 2$, let $d\boldsymbol{N}^{(n)}$ be

$$\begin{aligned} \boldsymbol{\lambda}^{(n)}(t) &= \boldsymbol{\mu} + \big(\boldsymbol{\omega} * d\boldsymbol{N}^{(n-1)}\big)(t) \\ dN_j^{(n)}(t) &= \mathbf{1}_{[r_j^{(n)}(t) \leq \lambda_j^{(n)}(t)/\lambda_{\max}]} dN_j^{(0)}(t), \quad j = 1, \ldots, p, \end{aligned} \tag{C.76}$$

where $r_j^{(n)}(t) \sim \text{Unif}([0,1])$.

From Lemma 3 in Brémaud and Massoulié (1996), we know that $\boldsymbol{N}^{(n)}$ admits the intensity $\boldsymbol{\lambda}^{(n)}$. It is shown in Brémaud and Massoulié (1996) that the sequence $\{\boldsymbol{N}^{(n)}\}_{n=1}^{\infty}$ converges to a process that has the same distribution as $\boldsymbol{N}$ under Assumption 15 in Chapter 4.

Next, we construct the sequence of processes that converges to the coupling process $\widetilde{\boldsymbol{N}}$. Let $\widetilde{\boldsymbol{N}}^{(0)}$ be a p-variate Poisson process with intensity $\lambda_{\max}$ and independent of $\boldsymbol{N}^{(0)}$. Define the first process as

$$d\widetilde{N}_j^{(1)}(t) = \mathbf{1}_{[t \leq z]} \mathbf{1}_{[\tilde{r}_j^{(1)}(t) \leq \mu_j/\lambda_{\max}]} d\widetilde{N}_j^{(0)}(t) + \mathbf{1}_{[t>z]} \mathbf{1}_{[r_j^{(1)}(t) \leq \mu_j/\lambda_{\max}]} dN_j^{(0)}(t), \quad j = 1, \ldots, p, \tag{C.77}$$

where $\tilde{r}_j^{(1)}(t) \sim \text{Unif}([0,1])$ and $r_j^{(1)}(t)$ is the same variable defined in (C.75). For $n \geq 2$, we construct $\widetilde{\boldsymbol{N}}^{(n)}$ as

$$\begin{aligned} \widetilde{\boldsymbol{\lambda}}^{(n)}(t) &= \boldsymbol{\mu} + \big(\boldsymbol{\omega} * d\widetilde{\boldsymbol{N}}^{(n-1)}\big)(t) \\ d\widetilde{N}_j^{(n)}(t) &= \mathbf{1}_{[t \leq z]} \mathbf{1}_{[\tilde{r}_j^{(n)}(t) \leq \widetilde{\lambda}_j^{(n)}/M]} d\widetilde{N}_j^{(0)}(t) + \mathbf{1}_{[t>z]} \mathbf{1}_{[r_j^{(n)}(t) \leq \widetilde{\lambda}_j^{(n)}/M]} dN_j^{(0)}(t), \quad j = 1, \ldots, p, \end{aligned} \tag{C.78}$$

where $\tilde{r}_j^{(n)}(t) \sim \text{Unif}([0,1])$ and $r_j^{(n)}(t)$ defined as in (C.76).

The sequence $\{\widetilde{\boldsymbol{N}}^{(n)}\}_{n=1}^{\infty}$ converges to $\widetilde{\boldsymbol{N}}$ using the same argument in Brémaud and Massoulié (1996) because the hybrid process $d\bar{\boldsymbol{N}}^{(0)} \equiv \mathbf{1}_{[t \leq z]} d\widetilde{\boldsymbol{N}}^{(0)} + \mathbf{1}_{[t>z]} d\boldsymbol{N}^{(0)}$ is still a Poisson process. To see this, one can check that, for any Borel set $A \in \mathcal{B}(\mathbb{R}^+)$, $\bar{\boldsymbol{N}}^{(0)}(A)$ follows a Poisson distribution of expectation $Mm(A)$ where $m(A)$ is the Lebesgue measure of A.

Moreover, for every n, $\widetilde{\boldsymbol{N}}$ is independent of $\mathcal{H}_z$. It is clear that, by construction, $\widetilde{\boldsymbol{N}}^{(n-1)}$ is independent of the history of $\boldsymbol{N}^{(n-1)}$ before time z (denoted as $\mathcal{H}_z^{(n-1)}$) by construction. Combining this with the fact that $\boldsymbol{\omega}(\Delta) = 0$ for $\Delta \leq 0$ from Assumption 19, we see that $\widetilde{\boldsymbol{\lambda}}^{(n)}$ is independent of $\mathcal{H}_z^{(n-1)}$. Recall that $\widetilde{\boldsymbol{N}}^{(n)}$ is determined by $\widetilde{\boldsymbol{\lambda}}^{(n)}$ and $\widetilde{\boldsymbol{N}}^{(0)}$, we know that $\widetilde{\boldsymbol{N}}^{(n)}$ is independent of

$\mathcal{H}_z^{(n-1)}$, and hence independent of $\mathcal{H}_z^{(n)}$. Therefore, we know that the iterative construction preserve the independence between $\widetilde{\boldsymbol{N}}^{(n)}$ and $\mathcal{H}_z^{(n)}$. As a result, $\widetilde{\boldsymbol{N}} \equiv \widetilde{\boldsymbol{N}}^{(\infty)}$ is independent of $\mathcal{H}_z^{(\infty)} \equiv \mathcal{H}_z$.

So far, we have shown the first part of Theorem 11 that there exist identically distributed $\boldsymbol{N}$ and $\widetilde{\boldsymbol{N}}$ such that $\widetilde{\boldsymbol{N}}$ is independent of $\mathcal{H}_z$, and the two processes satisfy the following sets of equations

$$\begin{aligned} \boldsymbol{\lambda}(t) &= \boldsymbol{\mu} + \big(\boldsymbol{\omega} * d\boldsymbol{N}\big)(t) \\ dN_j(t) &= \mathbf{1}_{[r_j(t) \le \lambda_j(t)/M]} dN_j^{(0)}(t), \quad j = 1, \dots, p, \end{aligned} \tag{C.79}$$

and

$$\begin{aligned} \widetilde{\boldsymbol{\lambda}}(t) &= \boldsymbol{\mu} + \big(\boldsymbol{\omega} * d\widetilde{\boldsymbol{N}}\big)(t) \\ d\widetilde{N}_j(t) &= \mathbf{1}_{[t \le z]} \mathbf{1}_{[\tilde{r}_j(t) \le \widetilde{\lambda}_j(t)/M]} d\widetilde{N}_j^{(0)}(t) + \mathbf{1}_{[t > z]} \mathbf{1}_{[r_j(t) \le \widetilde{\lambda}_j(t)/M]} dN_j^{(0)}(t), \quad j = 1, \dots, p. \end{aligned} \tag{C.80}$$

Our next task is to bound $\mathbb{E}\big|d\widetilde{N}_j(z+u) - dN_j(z+u)\big|$ using (C.79) and (C.80).

We claim that, for $n = 1, 2, \dots$,

$$\mathbb{E}\big|d\widetilde{\boldsymbol{N}}(z+u) - d\boldsymbol{N}(z+u)\big|/du \preceq 2v\boldsymbol{\Omega}^n \mathbf{1}, \quad u \in ((n-1)b, \infty). \tag{C.81}$$

We prove (C.81) by induction.

For $u \in \mathbb{R}$, we have a crude bound from the triangle inequality

$$\mathbb{E}\big|dN_j(u+z) - d\widetilde{N}_j(u+z)\big|/du \le \mathbb{E}\big|dN_j(u+z)\big|/du + \mathbb{E}\big|d\widetilde{N}_j(u+z)\big|/du = 2\Lambda_j \le 2v. \tag{C.82}$$

Jointly for all $j = 1, \dots, p$, $\mathbb{E}\big|d\boldsymbol{N}(u+z) - d\widetilde{\boldsymbol{N}}(u+z)\big|/du \preceq 2v\mathbf{1}$ for any u.

For $u \in (0, \infty)$, we have

$$
\begin{aligned}
&\mathbb{E}\big|d\boldsymbol{N}(u+z) - d\widetilde{\boldsymbol{N}}(u+z)\big|/du \\
\leq &\mathbb{E}\big|\mathbf{1}_{[r_j(u+z)\leq\lambda_j(u+z)/M]}dN_j^{(0)}(u+z) - \mathbf{1}_{[r_j(u+z)\leq\widetilde{\lambda}_j(u+z)/M]}dN_j^{(0)}(u+z)\big|/du \\
= &\mathbb{E}\big|\lambda_j(u+z) - \widetilde{\lambda}_j(u+z)\big| \\
= &\mathbb{E}\left|\mu_j + \sum_{l=1}^{p}\int_0^\infty \omega_{j,l}(\Delta)dN_l(u+z-\Delta) - \mu_j - \sum_{l=1}^{p}\int_0^\infty \omega_{j,l}(\Delta)d\widetilde{N}_l(u+z-\Delta)\right| \\
= &\mathbb{E}\left|\sum_{l=1}^{p}\int_0^\infty \omega_{j,l}(\Delta)\big[dN_l(u+z-\Delta) - d\widetilde{N}_l(u+z-\Delta)\big]\right| \\
\leq &\sum_{l=1}^{p}\int_0^b |\omega_{j,l}(\Delta)|\mathbb{E}\big|dN_l(u+z-\Delta) - d\widetilde{N}_l(u+z-\Delta)\big| \\
\leq &\sum_{l=1}^{p}\int_0^b |\omega_{j,l}(\Delta)|2vd\Delta \\
\leq &2v\boldsymbol{\Omega}_{j,\cdot}^{\mathrm{T}}\mathbf{1},
\end{aligned}
$$

where we use the crude bound (C.82) in the second-to-last inequality. Jointly for all j, we have, for $u \in (0, \infty)$

$$\mathbb{E}\big|d\boldsymbol{N}(u+z) - d\widetilde{\boldsymbol{N}}(u+z)\big|/du \preceq 2v\boldsymbol{\Omega}\mathbf{1}.$$

Now assume that for $u \in ((n-2)b, \infty)$, it holds that

$$\mathbb{E}\big|d\boldsymbol{N}(u+z) - d\widetilde{\boldsymbol{N}}(u+z)\big|/du \preceq 2v\boldsymbol{\Omega}^{n-1}\mathbf{1}.$$

Then, for $u \in ((n-1)b, \infty)$, we can see that

$$
\begin{aligned}
&\mathbb{E}\big|d\boldsymbol{N}(u+z) - d\widetilde{\boldsymbol{N}}(u+z)\big|/du \\
\leq &\sum_{l=1}^{p}\int_0^b |\omega_{j,l}(\Delta)|\mathbb{E}\big|dN_l(u+z-\Delta) - d\widetilde{N}_l(u+z-\Delta)\big| \\
\leq &\sum_{l=1}^{p}\int_0^b |\omega_{j,l}(\Delta)|2\big[\boldsymbol{\Omega}^{n-1}\big]_{l,\cdot}\boldsymbol{\Lambda}d\Delta \\
\leq &2v\boldsymbol{\Omega}_{j,\cdot}^{\mathrm{T}}\boldsymbol{\Omega}^{n-1}\mathbf{1}.
\end{aligned}
$$

In other words, we have, for $u \in ((n-1)b, \infty)$

$$\mathbb{E}\big|d\boldsymbol{N}(u+z) - d\widetilde{\boldsymbol{N}}(u+z)\big|/du \preceq 2v\boldsymbol{\Omega}^{n}\mathbf{1}.$$

We have completed the induction for $\mathbb{E}\big|d\boldsymbol{N}(u+z)-d\widetilde{\boldsymbol{N}}(u+z)\big|/du$. Next, we bound $\mathbb{E}\big|d\boldsymbol{N}(t')d\boldsymbol{N}^{\mathrm{T}}(u+z)-d\widetilde{\boldsymbol{N}}(t')d\widetilde{\boldsymbol{N}}^{\mathrm{T}}(u+z)\big|/(dt'du)$. Noting that $\mathbb{E}\big|d\boldsymbol{N}(t')d\boldsymbol{N}^{\mathrm{T}}(u+z)-d\widetilde{\boldsymbol{N}}(t')d\widetilde{\boldsymbol{N}}^{\mathrm{T}}(u+z)\big|/(dt'du)$ is symmetric, we discuss the case when $t' \geq u+z$. Furthermore, when $t' = u+z$, the following derivation holds with an exception on the diagonal elements

$$\mathbb{E}\big|dN_k(u+z)dN_k(u+z) - d\widetilde{N}_k(u+z)d\widetilde{N}_k(u+z)\big| = \mathbb{E}\big|dN_k(u+z) - d\widetilde{N}_k(u+z)\big|.$$

This has been taken care of by our previous discussion. Therefore, in what follows, we focus on the case when $t' > u+z$.

We claim that, for $u \in ((n-1)b, \infty)$,

$$\mathbb{E}\big|d\boldsymbol{N}(t')d\boldsymbol{N}^{\mathrm{T}}(u+z) - d\widetilde{\boldsymbol{N}}(t')d\widetilde{\boldsymbol{N}}^{\mathrm{T}}(u+z)\big|/(dt'du) \preceq 2v^2 \sum_{i=1}^{n+1} \boldsymbol{\Omega}^{i-1}\mathbf{1}\mathbf{1}^{\mathrm{T}}\big[\boldsymbol{\Omega}^{\mathrm{T}}\big]^{n+1-i}, \quad \text{(C.83)}$$

where $\boldsymbol{\Omega}^0 \equiv \boldsymbol{I}$ the identity matrix. Again, we prove (C.83) using a similar induction.

For an arbitrary u, we can obtain a crude bound

$$\begin{aligned}
&\mathbb{E}\big|dN_j(t')dN_k(u+z) - d\widetilde{N}_j(t')d\widetilde{N}_k(u+z)\big|/(dt'du)\\
\preceq&2\mathbb{E}\big|dN_j(t')dN_k(u+z)\big|\\
=&2[V_{j,k}(t'-u-z) + \Lambda_j\Lambda_k]\\
\leq&2v^2.
\end{aligned} \quad \text{(C.84)}$$

Jointly for j, k, $\mathbb{E}\big|d\boldsymbol{N}(t')d\boldsymbol{N}^{\mathrm{T}}(u+z) - d\widetilde{\boldsymbol{N}}(t')d\widetilde{\boldsymbol{N}}^{\mathrm{T}}(u+z)\big|/(dt'du) \leq 2v^2\mathbf{1}\mathbf{1}^{\mathrm{T}}$ for any u.

For $u \in (0, \infty)$, using the construction of the coupling process, we can see that

$$
\begin{aligned}
&\mathbb{E}\big|dN_j(t')dN_k(u+z) - d\widetilde{N}_j(t')d\widetilde{N}_k(u+z)\big|/(dt'du) \\
\leq &\mathbb{E}\big|\big[1_{[r_j(t')\leq\lambda_j(t')]}dN_k(u+z) - 1_{[r_j(t')\leq\widetilde{\lambda}_j(t')]}d\widetilde{N}_k(u+z)\big]d\widetilde{N}_k^{(0)}(t')\big|/(dt'du) \\
= &\mathbb{E}\big|dN_k(u+z)\lambda_j(t') - d\widetilde{N}_k(u+z)\widetilde{\lambda}_j(t')\big|/du \\
\leq &\mathbb{E}|\mu_j\big[dN_k(u+z) - d\widetilde{N}_k(u+z)\big]|/du + \\
&\mathbb{E}\left|\sum_{l=1}^{p}\int_0^\infty \omega_{j,l}(\Delta)\big[dN_l(t'-\Delta)dN_k(u+z) - d\widetilde{N}_l(t'-\Delta)d\widetilde{N}_k(u+z)\big]\right|/du \\
\leq &\mu_j\mathbb{E}\big|dN_k(u+z) - d\widetilde{N}_k(u+z)\big|/du + \\
&\sum_{l=1}^{p}\int_0^\infty |\omega_{j,l}(\Delta)|\mathbb{E}\big|dN_l(t'-\Delta)dN_k(u+z) - d\widetilde{N}_l(t'-\Delta)d\widetilde{N}_k(u+z)\big|/du \\
= &\mu_j\mathbb{E}\big|dN_k(u+z) - d\widetilde{N}_k(u+z)\big|/du + \\
&|\omega_{j,k}(t'-u-z)|\mathbb{E}\big|dN_k(u+z) - d\widetilde{N}_k(u+z)\big|/du + \\
&\sum_{l=1}^{p}\int_0^\infty |\omega_{j,l}(\Delta)|[1-1_{[l=k,\Delta=t'-u-z]}]\mathbb{E}\big|dN_l(t'-\Delta)dN_k(u+z) - d\widetilde{N}_l(t'-\Delta)d\widetilde{N}_k(u+z)\big|/du \\
\leq &(\mu_j+w)\mathbb{E}\big|dN_k(u+z) - d\widetilde{N}_k(u+z)\big|/du \\
&+\sum_{l=1}^{p}\int_0^\infty |\omega_{j,l}(\Delta)|[1-1_{[l=k,\Delta=t'-u-z]}]\mathbb{E}\big|dN_l(t'-\Delta)dN_k(u+z) - d\widetilde{N}_l(t'-\Delta)d\widetilde{N}_k(u+z)\big|/du.
\end{aligned}
$$

We notice that, when $l = k$ and $t'-\Delta = u+z$, $\mathbb{E}\big|dN_l(t'-\Delta)dN_k(u+z) - d\widetilde{N}_l(t'-\Delta)d\widetilde{N}_k(u+z)\big| = \mathbb{E}\big|dN_k(u+z) - d\widetilde{N}_k(u+z)\big|$. Therefore, we separate this term from the integral. In the mean time $\{\Delta : \Delta = t'-u-z\}$ is a set of measure zero. We use (C.81) to bound the first term and (C.84) to bound the second term, which gives

$$
\begin{aligned}
&\mathbb{E}\big|dN_j(t')dN_k(u+z) - d\widetilde{N}_j(t')d\widetilde{N}_k(u+z)\big|/(dt'du) \\
\leq &2v(\mu_j+w)\mathbf{\Omega}_{k,\cdot}^{\mathrm{T}}\mathbf{1} + 2v^2\mathbf{\Omega}_{j,\cdot}^{\mathrm{T}}\mathbf{1} \\
\leq &2v^2\mathbf{\Omega}_{k,\cdot}^{\mathrm{T}}\mathbf{1} + 2v^2\mathbf{\Omega}_{j,\cdot}^{\mathrm{T}}\mathbf{1}.
\end{aligned}
$$

where we use $v \geq \mu_j + w$ and the fact that all entries in $\mathbf{\Omega}$ are non-negative in the last inequality.

Jointly for j, k, we have

$$\begin{aligned}&\mathbb{E}\big|d\boldsymbol{N}(t')d\boldsymbol{N}^{\mathrm{T}}(u+z) - d\widetilde{\boldsymbol{N}}(t')d\widetilde{\boldsymbol{N}}_k^{\mathrm{T}}(u+z)\big|/(dt'du)\\ \leq& 2v^2\mathbf{1}\mathbf{1}^{\mathrm{T}}\boldsymbol{\Omega}^{\mathrm{T}} + 2v^2\boldsymbol{\Omega}\mathbf{1}\mathbf{1}^{\mathrm{T}} = 2v^2\sum_{i=1}^{2}\boldsymbol{\Omega}^{i-1}\mathbf{1}\mathbf{1}^{\mathrm{T}}\big[\boldsymbol{\Omega}^{\mathrm{T}}\big]^{2-i}.\end{aligned}$$

Now, assume that (C.83) holds for $u \in ((n-2)b, \infty)$. Using a similar argument, we get that, for $u \in ((n-1)b, \infty)$,

$$\begin{aligned}&\mathbb{E}\big|dN_j(t')dN_k(u+z) - d\widetilde{N}_j(t')d\widetilde{N}_k(u+z)\big|/(dt'du)\\ \leq&(\mu_j+w)\mathbb{E}\big|dN_k(u+z) - d\widetilde{N}_k(u+z)\big|/du+\\ &\sum_{l=1}^{p}\int_0^{\infty}|\omega_{j,l}(\Delta)|[1-1_{[l=k,\Delta=t'-u-z]}]\mathbb{E}\big|dN_l(t'-\Delta)dN_k(u+z) - d\widetilde{N}_l(t'-\Delta)d\widetilde{N}_k(u+z)\big|/du\\ \leq&2v(\mu_j+w)\boldsymbol{\Omega}_{k,\cdot}^{\mathrm{T}}\boldsymbol{\Omega}^{n-1}\mathbf{1} + \boldsymbol{\Omega}_{j,\cdot}^{\mathrm{T}}\left\{v^2\sum_{i=1}^{n}\boldsymbol{\Omega}^{i-1}[\mathbf{1}\mathbf{1}^{\mathrm{T}}]\big[\boldsymbol{\Omega}^{\mathrm{T}}\big]^{n-i-1}\boldsymbol{\Omega}_{k,\cdot}\right\}\\ \leq&2v^2\mathbf{1}^{\mathrm{T}}\big[\boldsymbol{\Omega}^{\mathrm{T}}\big]^{n-1}\boldsymbol{\Omega}_{k,\cdot} + 2v^2\sum_{i=1}^{n}\boldsymbol{\Omega}_{j,\cdot}^{\mathrm{T}}\boldsymbol{\Omega}^{i-1}[\mathbf{1}\mathbf{1}^{\mathrm{T}}]\big[\boldsymbol{\Omega}^{\mathrm{T}}\big]^{n-i-1}\boldsymbol{\Omega}_{k,\cdot}.\end{aligned}$$

Jointly for j, k, we have, for $u \in ((n-1)b, \infty)$,

$$\mathbb{E}\big|d\boldsymbol{N}(t')d\boldsymbol{N}^{\mathrm{T}}(u+z) - d\widetilde{\boldsymbol{N}}(t')d\widetilde{\boldsymbol{N}}^{\mathrm{T}}(u+z)\big|/(dt'du) \leq 2v^2\sum_{i=1}^{n+1}\boldsymbol{\Omega}^{i}[\mathbf{1}\mathbf{1}^{\mathrm{T}}]\big[\boldsymbol{\Omega}^{\mathrm{T}}\big]^{n-i+1}.$$

We have completed the proof. □

C.4.2 A corollary of Theorem 11

When the dimension of $\boldsymbol{\Omega}$ (i.e., p) is allowed to be very large, a large gap u might be required in order for the upper bound in Theorem 11 to vanish. This can be particularly problematic in neuroscience application where the number of neurons p are often very large. Using Assumption 16, the following result stablishes direct element-wise upper bounds for $\mathbb{E}\big|d\widetilde{N}_j(z+u) - dN_j(z+u)\big|/du$ and $\mathbb{E}\big|d\widetilde{N}_j(t')d\widetilde{N}_k(z+u) - dN_j(t')dN_k(z+u)\big|/(dt'du)$.

Corollary 3. *Suppose that Assumptions 15, 16, and 19 hold. For a given $z > 0$ and the process $\widetilde{\boldsymbol{N}}$ constructed in Theorem 11, we have, for each j,*

$$\mathbb{E}\big|d\widetilde{N}_j(z+u) - dN_j(z+u)\big|/du \leq a_1 \exp(-a_2 u),$$

where a_1 is a constant depends on v, ρ, d_0, and s, and $a_2 = -\log\big(\gamma_\Omega\big)/b$. Furthermore, for any j, k and $t' > z + u$, we have

$$\mathbb{E}\big|d\widetilde{N}_j(t')d\widetilde{N}_k(z+u) - dN_j(t')dN_k(z+u)\big|/(dt'du) \leq a_3 \exp(-a_4 u),$$

where $a_3 = 2(n_0+1)a_1^2$, $a_4 = a_2/2$, and n_0 is a constant that depends only on ρ_Ω.

Proof. From Assumption 16, we know that $\|\boldsymbol{\Omega}^{d_0}\mathbf{1}\|_2 \leq \rho$, where d_0 and ρ are introduced in Assumption 16.

For $0 \leq u < b$, we know from Theorem 11, for any j,

$$\mathbb{E}\big|d\widetilde{N}_j(z+u) - dN_j(z+u)\big|/du \leq 2v\boldsymbol{\Omega}_{j,\cdot}^{\mathrm{T}}\mathbf{1} \leq 2v\sum_{k\in\mathcal{E}_j}\Omega_{j,k} \leq 2vs^{1/2}\gamma_\Omega.$$

And for $2 \leq n < d_0$, we know that

$$\mathbb{E}\big|d\widetilde{N}_j(z+u) - dN_j(z+u)\big|/du \leq 2v\boldsymbol{\Omega}_{j,\cdot}\boldsymbol{\Omega}^{n-1}\mathbf{1} \leq 2vs^{n/2}\gamma_\Omega^n.$$

For $n \geq d_0$, we have

$$\mathbb{E}\big|d\widetilde{N}_j(z+u) - dN_j(z+u)\big|/du \leq 2v\|\boldsymbol{\Omega}^{n-d_0}\boldsymbol{\Omega}^{d_0}\mathbf{1}\|_2 \leq 2v\gamma_\Omega^{n-d_0}\rho = [2v\rho\gamma_\Omega^{-d_0}]\gamma_\Omega^n.$$

Letting $a_1 \equiv \max(2v\rho\gamma_\Omega^{-d_0}, 2vs^{d_0/2})$ gives

$$\mathbb{E}\big|d\widetilde{N}_j(z+u) - dN_j(z+u)\big|/du \leq a_1 \exp(-a_2 u), \tag{C.85}$$

where $a_2 = -\log(\gamma_\Omega)/b$.

Similarly, for $u \in ((n-1)b, nb]$,

$$\begin{aligned}
&\mathbb{E}\big|d\widetilde{N}_j(t')d\widetilde{N}_k(z+u) - dN_j(t')dN_k(z+u)\big|/(dt'du) \\
\leq& 2v^2\left[\sum_{i=1}^{n+1}\mathbf{\Omega}^{i-1}[\mathbf{1}\mathbf{1}^{\mathrm{T}}][\mathbf{\Omega}^{\mathrm{T}}]^{n-i+1}\right]_{j,k} = 2\left[\sum_{i=1}^{n+1}[v\mathbf{\Omega}^{i-1}\mathbf{1}][v\mathbf{\Omega}^{n-i+1}\mathbf{1}]^{\mathrm{T}}\right]_{j,k} \\
=& 2\sum_{i=1}^{n+1}[v\mathbf{\Omega}^{i-1}\mathbf{1}]_j[v\mathbf{\Omega}^{n-i+1}\mathbf{1}]_k \leq 2\sum_{i=1}^{n+1}\|v\mathbf{\Omega}^{i-1}\mathbf{1}\|_\infty\|v\mathbf{\Omega}^{n-i+1}\mathbf{1}\|_\infty \qquad \text{(C.86)} \\
\leq& 2\sum_{i=1}^{n+1}a_1\exp(-a_2(i-1)b)a_1\exp(-a_2(n-i+1)b) \\
=& 2a_1^2(n+1)\exp(-a_2nb).
\end{aligned}$$

Since $log(n+1)/n = o(1)$, we know that there exists a constant n_0 such that $\log(n+1)/n \leq a_2b/2$ for any $n \geq n_0$. This constant n_0 depends only on γ_Ω since $a_2b = -\log(\gamma_\Omega)$. Therefore,

$$\mathbb{E}\big|d\widetilde{N}_j(t')d\widetilde{N}_k(z+u) - dN_j(t')dN_k(z+u)\big|/(dt'du) \leq 2a_1^2\exp(-a_2nb/2) \leq 2a_1^2\exp(-a_2u/2),$$

for $u \in ((n-1)b, nb]$ and $n \geq n_0$. When $u \leq (n_0-1)b$, we have

$$\mathbb{E}\big|d\widetilde{N}_j(t')d\widetilde{N}_k(z+u) - dN_j(t')dN_k(z+u)\big|/(dt'du) \leq 2(n_0+1)a_1^2\exp(-a_2nb) \leq 2(n_0+1)a_1^2\exp(-a_2u/2).$$

In summary, for any $u \geq 0$, we have $\tau_{j,k}^2(u) \leq a_3\exp(-a_4u)$, where $a_3 = 2(n_0+1)a_1^2$ and $a_4 = a_2/2$. □

C.4.3 Weak dependence

In this section, we rewrite $\bar{y}_{j,k}$ as the mean of a discrete sequence, and show that the sequence has (i) a sub-Gaussian tail and (ii) τ-dependence coefficient decays exponentially fast as the time gap increases. Without loss of generality, let $b_1 \leq b_2 \leq 0$. Otherwise, one can split the function f into f_- defined on $[b_1, 0]$ and f_+ defined on $[0, b_2]$, and analyzes them accordingly (note that f_- and f_+ are not the positive and negative part of the function).

We define the following series $\{y_{j,k,i}\}$ for $i = 1, \ldots, T/(2\epsilon)$

$$y_{j,k,i} \equiv \frac{1}{2\epsilon}\int_{2\epsilon(i-1)}^{2\epsilon i}\int_0^T f(t-t')dN_k(t')dN_j(t),$$

where ϵ is the smallest number satisfies that $\epsilon \geq \max\{|b_1|, b\}$ and $T/(2\epsilon)$ is an integer. We immediately know that

$$\frac{2\epsilon}{T}\sum_{i=1}^{T/(2\epsilon)} y_{j,k,i} = \frac{2\epsilon}{T}\sum_{i=1}^{T/(2\epsilon)} \frac{1}{2\epsilon}\int_{2\epsilon(i-1)}^{2\epsilon i}\int_0^T f(t-t')dN_k(t')dN_j(t) = \bar{y}_{j,k}.$$

The first two lemmas are useful in showing (i). The first lemma quantifies the tail behaviour of a Poisson random variables.

Lemma 16. *Suppose that x is a Poisson random variable with mean m. Then for any $n > 0$*

$$\mathbb{P}(x - m \geq n) \leq \exp(-m - \log(n/em)n). \tag{C.87}$$

The next lemma deals with the tail behaviour of the product of two random variables, which is an adaptation of Lemma A.2 in Fan et al. (2014).

Lemma 17. *Suppose that Z_1 and Z_2 satisfies that*

$$P(|Z_i| > n) \leq \exp(1 - (t/K_i)), \quad i = 1, 2 \tag{C.88}$$

for all $n \geq 0$. Then for any $n \geq 0$ and $K^* = \sqrt{K_1 K_2}(\log 2 + 1)$,

$$P(|Z_1 Z_2| > n) \leq \exp\left(1 - (t/K^*)^{1/2}\right). \tag{C.89}$$

The proof follows the proof of Lemma A.2 in Fan et al. (2014) by choosing $r_1 = r_2 = 1$ and $r^* = 1/2$.

The next lemma provides an approach to evaluate the τ-dependence of the sequence $\{y_{j,k,i}\}$. We refer the readers to Merlevède et al. (2011) for the definition of the τ-dependence of a sequence. In this paper, we mainly use the coupling result in Dedecker and Prieur (2004) to bound the τ-dependence coefficient (Lemma 3 in Dedecker and Prieur (2004)), stated in the following lemma.

Lemma 18. *For a sequence $\{y_{j,k,i}\}_{i=1}^n$, let $\mathcal{H}_z^y$ be the σ-filed generated by $\{y_{j,k,i}\}_{i=1}^z$ and u be a positive integer. The τ-dependence coefficient $\tau_y(u)$ of the sequence $\{y_{j,k,i}\}_{i=1}^n$ satisfies that*

$$\tau_y(u) = \sup_z \tau(\mathcal{H}_z, y_{j,k,z+u}) \leq \sup_z \mathbb{E}|y_{j,k,z+u} - \tilde{y}_{j,k,z+u}|,$$

where $\tilde{y}_{j,k,z+u}$ is a random variable distributed identically as $y_{j,k,z+u}$ and $\tilde{y}_{j,k,z+u}$ is independent of $\mathcal{H}_z^y$.

We are now ready to give the properties of the sequence $\{y_{j,k,i}\}$ that we will use in proving Theorem 9.

Proposition 6. *Suppose that Assumptions 15, 16, and 19 hold. Further assume that there exists* $\lambda_{\max}$ *so that* $\lambda_j(t) \leq \lambda_{\max}$ *for all* $j = 1, \ldots, p$. *Let* $f(\cdot)$ *be a bounded function on a bounded support, i.e., supp*$(f) = [b_1, b_2]$ *with* $b_2 \leq 0$, *and* $\|f\|_\infty \equiv \max_x |f(x)| \leq C_f$. *We have that*

i) the τ*-dependence coefficient of* $y_{j,k,i}$ *satisfies that, for any positive integer* l,

$$\tau_y(u) \leq a_5 \exp(-a_6 l), l \geq 1. \tag{C.90}$$

ii) the sequence has an exponential tail of order $1/2$, *i.e.,*

$$\sup_{i>0} \mathbb{P}(|y_{j,k,i}| \geq x) \leq \exp\left(1 - a_7 x^{1/2}\right). \tag{C.91}$$

Proof. We first show that the tail bound (C.91) holds on $y_{j,k,i}$.

Since $\|f\|_\infty = C_f$ and Supp$(K) = [b_1, b_2]$ by definition, we know that

$$y_{j,k,i} \leq \frac{C_f}{2\epsilon} N_j\big([2\epsilon(i-1), 2\epsilon i]\big) N_k\big([2\epsilon i - 2\epsilon + b_1, 2\epsilon i + b_2]\big).$$

Using the Poisson embedding technique and the assumption that $\lambda_j(t) \leq \lambda_{\max}$, we know that both random variables on the right-hand side are marginally dominated by two Poisson random variables of mean $2\epsilon\lambda_{\max}$ and $(2\epsilon + b_2 - b_1)\lambda_{\max}$. From Lemma 16, we see that both are Poisson random variables. Applying Lemma 17, we see that y_i has an exponential tail of order $1/2$.

Next, we evaluate the τ-dependence coefficient of the sequence $\{y_{j,k,i}\}$. For any z, we construct a sequence $\{\tilde{y}_{j,k,i}\}_{i=1}^{T/(2\epsilon)}$ that is distributed identically as $\{y_{j,k,i}\}_{i=1}^{T/(2\epsilon)}$ while $\tilde{y}_{j,k,z+l}$ is independent of $\{y_{j,k,i}\}_{i=1}^{z}$ for all positive integer l. Let $\boldsymbol{N}$ and $\widetilde{\boldsymbol{N}}$ be the two processes constructed in Theorem 11 such that $\boldsymbol{N} \sim \widetilde{\boldsymbol{N}}$ and $\widetilde{\boldsymbol{N}}$ is independent of $\mathcal{H}_{2\epsilon z}$. This process exists as a result of Corollary 3. Hence, in what follows, we will let $y_{j,k,i}$ be defined on $d\boldsymbol{N}$ as

$$y_{j,k,i} \equiv \frac{1}{2\epsilon} \int_{2\epsilon(i-1)}^{2\epsilon i} \int_0^T f(t - t') dN_k(t') dN_j(t). \tag{C.92}$$

And we define $\tilde{y}_{j,k,i}$ on $\widetilde{\boldsymbol{N}}$ as

$$\tilde{y}_{j,k,i} \equiv \frac{1}{2\epsilon} \int_{2\epsilon(i-1)}^{2\epsilon i} \int_0^T f(t-t') d\tilde{N}_k(t') d\tilde{N}_j(t). \tag{C.93}$$

As a result of the construction, we know that $\{\tilde{y}_{j,k,i}\}_{i=1}^{T/(2\epsilon)}$ is distributed identically as $\{y_{j,k,i}\}_{i=1}^{T/(2\epsilon)}$ and $\{\tilde{y}_{j,k,i}\}_{i=1}^{T/(2\epsilon)}$ is independent of $\{y_{j,k,i}\}_{i=1}^{z}$.

We now bound the quantity $\mathbb{E}\big|\tilde{y}_{j,k,z+l} - y_{j,k,z+l}\big|$.

$$\begin{aligned}
&\mathbb{E}\big|\tilde{y}_{j,k,z+l} - y_{j,k,z+l}\big| \\
=&\frac{1}{2\epsilon}\mathbb{E}\left|\int_{2\epsilon(i+l-1)}^{2\epsilon(z+l)} \int_{t+b_1}^{t+b_2} f(t-t')\big[d\tilde{N}_k(t')d\tilde{N}_j(t) - dN_k(t')dN_j(t)\big]\right| \\
\leq&\frac{1}{2\epsilon}\int_{2\epsilon(z+l-1)}^{2\epsilon(z+l)} \int_{t+b_1}^{t+b_2} f(t-t')\mathbb{E}\left|d\tilde{N}_k(t')d\tilde{N}_j(t) - dN_k(t')dN_j(t)\right| \\
\leq&\frac{C_f}{2\epsilon}\int_{2\epsilon(z+l-1)}^{2\epsilon(z+l)} \int_{t+b_1}^{t+b_2} \mathbb{E}\left|d\widetilde{N}_k(t')d\widetilde{N}_j(t) - dN_k(t')dN_j(t)\right| \\
\leq&\frac{C_f}{2\epsilon}\int_{2\epsilon(z+l-1)}^{2\epsilon(z+l)} \int_{t+b_1}^{t+b_2} a_3 \exp(-a_4[\min(t',t) - 2\epsilon z])dt'dt
\end{aligned}$$

where we use the fact that $\|f\|_\infty \leq C_f$ in the second inequality, and Corollary 3 in the last inequality. Notice that $\min(t',t) - 2\epsilon z = 2\epsilon(z+l-1) + b_2 - 2\epsilon z = 2\epsilon(l-1) + b_2$. Thus, we have

$$\mathbb{E}\big|\tilde{y}_{j,k,z+l} - y_{j,k,z+l}\big| \leq a_5 \exp(-a_6 l),$$

where $a_5 = C_f(b_2 - b_1)\exp(b_2 - 2\epsilon)a_3$ and $a_6 = 2\epsilon a_4$. This completes the proof. □

C.4.4 Proof of Theorem 9

Proof. In this proof, we use C_1, C_2, and C_3 to denote global constants whose value might change from line to line. As in the previous section, we set

$$y_{j,k,i} \equiv \frac{1}{2\epsilon} \int_{2\epsilon(i-1)}^{2\epsilon i} \int_0^T f(t-t') dN_k(t') dN_j(t).$$

From the conclusion of Proposition 6, we know that $\{y_{j,k,i}\}_{i=1}^{T/(2\epsilon)}$ satisfies the conditions of Theorem 1 of Merlevède et al. (2011) with $\gamma_1 = 1$ and $\gamma_2 = 1/2$. Thus, applying Theorem 1 in

Merlevède et al. (2011) on $\{y_{j,k,i} - \mathbb{E}y_{j,k,i}\}_{i=1}^{T/(2\epsilon)}$ gives

$$\mathbb{P}\left(\left|\sum_{i=1}^{T/(2\epsilon)} y_{j,k,i} - \frac{T}{2\epsilon}\mathbb{E}y_{j,k,i}\right| \geq T\epsilon_1\right) \leq \frac{T}{2\epsilon}\exp\left(-\frac{(\epsilon_1 T)^{1/3}}{C_1}\right) + \exp\left(-\frac{\epsilon_1^2 T^2}{C_2(1+Tv_y/2\epsilon)}\right) + \exp\left[-\frac{\epsilon_1^2 T^2}{C_3 T/2\epsilon}\exp\left(\frac{(\epsilon_1 T)^{2/9}}{C_4(\log \epsilon_1 T)^{1/3}}\right)\right], \tag{C.94}$$

where ϵ_1 is to be specified later. Here v_y is a measure of "variance" defined in (1.11) of Merlevède et al. (2011). We can see that v_y is upper bounded since

$$\begin{aligned} v_y \leq & \sup_{i>0}\left\{\mathbb{E}\big[(y_{j,k,i} - \mathbb{E}y_i)^2\big] + 2\sum_{l\geq 1}\mathbb{E}\big[(y_{j,k,i} - \mathbb{E}y_{j,k,i})(y_{j,k,i+l} - \mathbb{E}y_{j,k,i+l})\big]\right\} \\ = & \sup_{i>0}\left\{\mathbb{E}\big[(y_{j,k,i} - \mathbb{E}y_{j,k,i})^2\big] + 2\sum_{l\geq 1}\mathbb{E}\big[\mathbb{E}[y_{j,k,i+l} - \mathbb{E}y_{j,k,i+l} \mid y_{j,k,i}](y_{j,k,i} - \mathbb{E}y_{j,k,i})\big]\right\} \\ \leq & \sup_{i>0}\left\{\mathbb{E}\big[(y_{j,k,i} - \mathbb{E}y_{j,k,i})^2\big] + 2\sum_{l\geq 1}\mathbb{E}\big[|\mathbb{E}[y_{j,k,i+l} - \mathbb{E}y_{j,k,i+l} \mid y_{j,k,i}||y_{j,k,i} - \mathbb{E}y_{j,k,i})|\big]\right\} \\ \leq & \sup_{i>0}\left\{\mathbb{E}\big[(y_{j,k,i} - \mathbb{E}y_{j,k,i})^2\big] + 2\sum_{l\geq 1} a_5\exp(-a_6 l)\mathbb{E}\big[|y_{j,k,i} - \mathbb{E}y_{j,k,i}|\big]\right\}, \end{aligned} \tag{C.95}$$

which is finite. Here the second inequality use the definition of the τ-dependence where $\tau_y(j-i) \geq |\mathbb{E}[y_j \mid y_i] - \mathbb{E}y_j|$ (see Equation 2.1 in Merlevède et al. (2011)) . Letting $\epsilon_1 = c_5 T^{-2/5}/2$ gives

$$\mathbb{P}\left(|\bar{y}_{j,k} - \mathbb{E}\bar{y}_{j,k}| \geq c_5 T^{-2/5}\right) \leq c_7 T\exp(-c_6 T^{1/5}).$$

□

C.4.5 Smoothness of cross-covariances

We now move on to the proof of Corollary 2. The following lemma shows that the cross-covariance is Lipschitz given the smoothness assumption on $\omega_{j,k}$ (Assumption 19 in Chapter 4). This helps bound the bias of the kernel smoothing estimator.

Lemma 19. *Under Assumption 19, the cross-covariance function is θ_2-Lipschitz for $1 \leq j,k \leq p$ for some $\theta_2 > 0$, i.e. $|V_{j,k}(x) - V_{j,k}(y)| \leq \theta_2|x-y|$.*

Proof. Let $g_{j,k}(\Delta) = V_{j,k}(\Delta)/\Lambda_k$. We have the following version of the integral equation

$$\boldsymbol{g} = \boldsymbol{\omega} + (\boldsymbol{\omega} * \boldsymbol{g})(\Delta). \tag{C.96}$$

Plugging (C.96) into itself gives that

$$\boldsymbol{g}(\Delta) = \sum_{i=1}^{\infty} \boldsymbol{\omega}^{(i)}(\Delta). \tag{C.97}$$

For each element in $\boldsymbol{g}$, we have

$$g_{j,k}(\Delta) = \omega_{j,k}(\Delta) + \sum_{i=2}^{\infty} \big[\boldsymbol{\omega}_{j,\cdot}^{(i-1)} * \boldsymbol{\omega}_{\cdot,k}\big](\Delta). \tag{C.98}$$

Then,

$$\big|g_{j,k}(x) - g_{j,k}(y)\big| \leq \big|\omega_{j,k}(x) - \omega_{j,k}(y)\big| + \sum_{i=2}^{\infty} \big|\big[\boldsymbol{\omega}_{j,\cdot}^{(i-1)} * \boldsymbol{\omega}_{\cdot,k}\big](x) - \boldsymbol{\omega}_{j,\cdot}^{(i-1)} * \boldsymbol{\omega}_{\cdot,k}\big](y)\big|. \tag{C.99}$$

Recall that we assume in Assumption 19 that $\omega_{j,k}$ lives in a Sobolev space $W(\theta_1, L_1)$ on $[0, b]$, which implies that $\omega_{j,k}$ is θ_0-Lipschitz for some positive constant θ_0.

For each $i \geq 2$, we have

$$\begin{aligned}
&\big|\big[\boldsymbol{\omega}_{j,\cdot}^{(i-1)} * \boldsymbol{\omega}_{\cdot,k}\big](x) - \boldsymbol{\omega}_{j,\cdot}^{(i-1)} * \boldsymbol{\omega}_{\cdot,k}\big](y)\big| \\
=&\left|\int_0^{\infty} \boldsymbol{\omega}_{j,\cdot}^{(i-1)}(\Delta) \cdot \big[\boldsymbol{\omega}_{\cdot,k}(x-\Delta) - \boldsymbol{\omega}_{\cdot,k}(y-\Delta)\big] d\Delta\right| \\
\leq&\sum_{l=1}^{p} \left|\int_0^{\infty} \omega_{j,l}^{(i-1)}(\Delta)\big[\boldsymbol{\omega}_{l,k}(x-\Delta) - \boldsymbol{\omega}_{l,k}(y-\Delta)\big] d\Delta\right| \\
\leq&\sum_{l=1}^{p} \theta_0|x-y| \int_0^{\infty} \big|\omega_{j,l}^{(i-1)}(\Delta)\big| d\Delta \\
\leq&\theta_0|x-y| \sum_{l=1}^{p} \Omega_{j,l}^{(i-1)} \leq \theta_0|x-y| \sum_{l=1}^{p} \Omega_{j,l}^{i-1} \\
\leq&\theta_0|x-y|\gamma_{\Omega}^{i-1},
\end{aligned} \tag{C.100}$$

where θ_0 is the Lipschitz constant for $\omega_{j,k}$. Here the third to last inequality follows from the property of the convolution. For $i = 2$, Young's inequality gives

$$\|\omega_{j,\cdot} * \omega_{\cdot,k}\|_1 \leq \sum_{l=1}^{p} \|\omega_{j,l}\|_1 \|\omega_{l,k}\|_1 = \Omega_{j,\cdot} \cdot \Omega_{\cdot,k}. \tag{C.101}$$

The last inequality of (C.100) follows from the fact that $\Gamma_{\max}(\Omega^i) \leq \gamma_\Omega^i$. As a result, we have

$$\left|g_{j,k}(x) - g_{j,k}(y)\right| \leq \theta_0 |x-y| \sum_{i=0}^{\infty} \gamma_\Omega^i = \frac{\theta_0}{1-\gamma_\Omega}|x-y|. \tag{C.102}$$

Let $\theta_2 = \max_j(\Lambda_j)\theta_0/(1-\gamma_\Omega)$ and we have

$$\left|V_{j,k}(x) - V_{j,k}(y)\right| \leq \theta_2 |x-y|.$$

□

C.4.6 Proof of Corollary 2

Proof. Recall that the estimator $\widehat{V}_{j,k}$ takes the form (Equation 4.24 in Chapter 4)

$$\widehat{V}_{j,k}(\Delta) = \underbrace{\frac{1}{Th}\iint_{[0,T]^2} K\big([\Delta - \{t'-t\}]/h\big)\, dN_j(t) dN_k(t')}_{\text{I}/h} - \underbrace{\frac{1}{T}N_j(T)\frac{1}{T}N_k(T)}_{\text{II}}. \tag{C.103}$$

For I, applying Theorem 9 with $f(x) = K\big([\Delta - \{t'-t\}]/h\big) \leq 1$ gives

$$\mathbb{P}\left(|\bar{y}_{j,k} - \mathbb{E}\bar{y}_{j,k}| \geq c_5 T^{-2/5}\right) \leq c_7 T \exp(-c_6 T^{1/5}). \tag{C.104}$$

We see that

$$\begin{aligned}\mathbb{E}\bar{y}_{j,k} =& \mathbb{E}\left[\frac{2\epsilon}{\int_{2\epsilon(i-1)}^{2\epsilon i}} \int_0^T K\big([\Delta - \{t-t'\}]/h\big) dN_k(t') dN_j(t)\right] \\ =& \int_{-\Delta-h}^{-\Delta+h} K\left(\frac{\Delta+t'}{h}\right)(V_{j,k}(-t') + \Lambda_j\Lambda_k) dt'.\end{aligned} \tag{C.105}$$

And the bias is

$$\begin{aligned}
&|\mathbb{E}\bar{y}_{j,k} - hV_{j,k}(\Delta) + h\Lambda_j\Lambda_k| \\
=&\left|\int_{-\Delta-h}^{-\Delta+h} K\left(\frac{\Delta+t}{h}\right) V_{j,k}(-t')dt' - hV_{j,k}(\Delta) + h\Lambda_j\Lambda_k\right| \\
=&\left|\int_{-\Delta-h}^{-\Delta+h} K\left(\frac{\Delta+t'}{h}\right) [V_{j,k}(-t') - V_{j,k}(\Delta)]dt'\right| \\
\leq&\int_{-\Delta-h}^{-\Delta+h} K\big([\Delta+t']/h\big)\theta_2| - t' - \Delta|dt' \qquad \text{(C.106)} \\
\leq&\int_{-\Delta-h}^{-\Delta+h} \theta_2|t'+\Delta|dt' \\
=&\theta_2\int_{-h}^{h}|z|dz \\
=&\theta_2 h^2,
\end{aligned}$$

where we use Lemma 19 in the second-to-last inequality and $K(\cdot) \leq 1$ in the last inequality.

Similarly, for each term in II, it is easy to check that

$$\mathbb{P}\left(|N_j(T)/T - \Lambda_j| \geq c_5 T^{-2/5}\right) \leq c_7 T\exp(-c_6 T^{1/5}). \qquad \text{(C.107)}$$

The proof of (C.107) is very similar to the proof of Theorem 9 and is thus omitted.

Combining (C.104), (C.106), and (C.107), we have

$$\begin{aligned}
\left|\widehat{V}_{j,k}(\Delta) - V_{j,k}(\Delta)\right| \leq&\left|h^{-1}\bar{y}_{j,k} - h^{-1}\mathbb{E}\bar{y}_{j,k}\right| + \left|h^{-1}\mathbb{E}\bar{y}_{j,k,i} - V_{j,k}(\Delta) + \Lambda_j\Lambda_k\right| \\
&+\left|\frac{1}{T^2}(N_j(T) - T\Lambda_j)N_k(T)\right| + \left|\Lambda_j\frac{1}{T}N_k(T) - \Lambda_j\Lambda_k\right| \\
\leq&c_5T^{-2/5}h^{-1} + \theta_2 h + (\|\mathbf{\Lambda}\|_\infty + c_5T^{-2/5})c_5T^{-2/5} + \|\mathbf{\Lambda}\|_\infty c_5T^{-2/5}. \qquad \text{(C.108)}
\end{aligned}$$

Now, let $h = T^{-1/5}$. We can see that

$$\left|\widehat{V}_{j,k}(\Delta) - V_{j,k}(\Delta)\right| \leq c_5' T^{-1/5}. \qquad \text{(C.109)}$$

Lastly, we need a uniform bound on $\widehat{V}_{j,k} - V_{j,k}$ on the region $[-B, B]$. We first see that the above probability statement holds for a grid of $[T^{1/5}]$ points on $[-B, B]$. The gap between adjacent point

is $2BT^{1/5}$. From basic calculus we get that, for all $\Delta \in [-B, B]$,

$$\begin{aligned}\left|\widehat{V}_{j,k}(\Delta) - V_{j,k}(\Delta)\right| =& \left|\widehat{V}_{j,k}(\Delta) - \widehat{V}_{j,k}(\Delta_k) + \widehat{V}_{j,k}(\Delta_k) - V_{j,k}(\Delta_k) + V_{j,k}(\Delta_k) - V_{j,k}(\Delta)\right| \\ \leq& 2BT^{-1/5} + c_5' T^{-1/5} + \theta_2 T^{-1/5} \equiv c_5 T^{-1/5}.\end{aligned} \tag{C.110}$$

Therefore, we have that

$$\left\|\widehat{V}_{j,k} - V_{j,k}\right\|_{2,[-B,B]} \leq c_5 T^{-1/5}. \tag{C.111}$$

By investigating (C.104) and (C.107), we can see that the probability of (C.111) to hold for $1 \leq j, k \leq p$ converges to unity as long as

$$p^2 T^{6/5} \exp\left(-c_6 T^{1/5}\right) = o(1)$$

. □