Journal of Machine Learning Research 26 (2025) 1-66 Submitted 12/22; Revised 6/25; Published 7/25

Exponential Family Graphical Models: Correlated Replicates and Unmeasured Confounders, with Applications to fMRI Data

Yanxin Jin YANXINJ@UMICH.EDU
Department of Statistics
University of Michigan
Ann Arbor MI, 48109, USA

Yang Ning YN265@CORNELL.EDU
Department of Statistics and Data Science
Cornell University
Ithaca NY, 14850, USA

Kean Ming Tan KEANMING@UMICH.EDU
Department of Statistics
University of Michigan
Ann Arbor MI, 48109, USA

Editor: Jie Peng

Abstract

Graphical models have been used extensively for modeling brain connectivity networks. However, unmeasured confounders and correlations among measurements are often overlooked during model fitting, which may lead to spurious scientific discoveries. Motivated by functional magnetic resonance imaging (fMRI) studies, we propose a novel method for constructing brain connectivity networks with correlated replicates and latent effects. In a typical fMRI study, each participant is scanned and fMRI measurements are collected across a period of time. In many cases, subjects may have different states of mind that cannot be measured during the brain scan: for instance, some subjects may be awake during the first half of the brain scan, and may fall asleep during the second half of the brain scan. To model the correlation among replicates and latent effects induced by the different states of mind, we assume that the correlated replicates within each independent subject follow a one-lag vector autoregressive model, and that the latent effects induced by the unmeasured confounders are piecewise constant. Theoretical guarantees are established for parameter estimation. We demonstrate via extensive numerical studies that our method is able to estimate latent variable graphical models with correlated replicates more accurately than existing methods.

Keywords: Convex optimization; correlated replicates; latent variables; fused lasso; piecewise constant.

1. Introduction

Undirected graphical models have been used extensively in various scientific domains to represent conditional dependence relationships between pairs of variables. In a graph, each node represents a random variable, and an edge connecting a pair of nodes indicates that the

pair of variables is conditionally dependent, given all of the other variables. For instance, in a brain connectivity network, each node represents a brain region, and an edge between two nodes indicates that the two brain regions are conditionally dependent. Many methods were proposed for estimating graphical models under various model assumptions. In particular, Gaussian graphical models have been studied extensively (Meinshausen and Bühlmann, 2006; Yuan and Lin, 2007; Friedman et al., 2008; Rothman et al., 2008; Cai et al., 2011; Sun and Zhang, 2013; Tan et al., 2015; Lin et al., 2016; Cai et al., 2016). To relax the Gaussianity assumption, exponential graphical models in which the node-conditional distribution for each variable belongs to an exponential family distribution were proposed (Ravikumar et al., 2010; Yang et al., 2015; Chen et al., 2015; Yang et al., 2018, 2013). To accommodate mixed data types, various work have proposed the mixed graphical model in which each variable can belong to different distribution (She et al., 2019; Chen et al., 2015; Yang et al., 2014; Lee and Hastie, 2015). Several authors have considered nonparametric graphical models without imposing any distributional assumption on the random variables (Sun et al., 2015; Tan et al., 2019). The literature on graphical models is vast: we refer the reader to Maathuis et al. (2018) for a comprehensive list of references.

In this paper, we focus on estimating brain connectivity networks using fMRI data. There are two major challenges presented by fMRI data: correlated replicates for each independent subject and the presence of unmeasured confounders. Firstly, each independent subject is scanned over a period of time, resulting a series of correlated brain scans. Moreover, while the fMRI brain scans are taken over time, the subjects may have different states of mind or head motion, which can be interpreted as unmeasured confounders. For instance, certain subjects may be awake during the first half of the brain scan, and may fall asleep during the second half of the brain scan. Different brain regions may be active or inactive, depending on whether the subject is awake or asleep. Thus, it is of utmost importance to model correlation across replicates and the latent effects induced by the unmeasured confounders simultaneously to obtain an accurate conditional independence graph.

However, most existing methods for estimating conditional independence graph assume that all relevant variables are observed, which will yield biased graph due to the existence of unmeasured confounders in fMRI data. In the context of Gaussian graphical models, Chandrasekaran et al. (2010) showed that marginalizing over the unmeasured confounders will yield a dense conditional independence graph of the observed variables even when the true underlying graph for the observed variables is sparse. To address this issue, various methods were proposed for modeling latent variable graphical models under various assumptions on the unmeasured confounders (Chandrasekaran et al., 2010; Tan et al., 2016; Fan et al., 2017; Wu et al., 2017). Besides, the aforementioned work mainly focused on estimating a conditional independence graph based on independent realizations of a common random vector. Such methods are not suitable for fMRI data, which typically involves highly correlated replicates. Some authors assumes that the graph evolves across time, i.e., time-varying graphical models, but these work do not model the correlation across replicates (Kolar et al., 2010; Hanneke et al., 2010; Sarkar and Moore, 2006; Guo et al., 2007; Zhou et al., 2010). To account for correlated replicates, several authors have assumed that the replicates follow a vector autoregressive (VAR) process, with the resulting graphical model remaining invariant over time (Qiu et al., 2016; Hall et al., 2016; Basu and Michailidis, 2015). In contrast, the literature on functional graphical models estimates the graphical

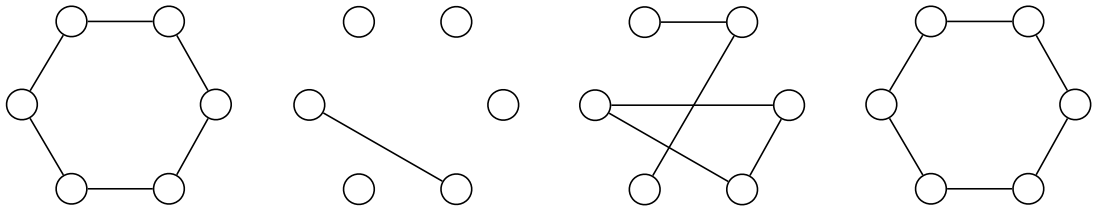


Figure 1: A toy example on a Gaussian graphical model with unmeasured confounders and correlated replicates. Panels (a), (b), (c), and (d) correspond to the true underlying graph, estimated graphs by Friedman et al. (2008), Chandrasekaran et al. (2010), and our proposed method, respectively.

model by directly treating correlated replicates as observations drawn from the underlying smoothed realizations of random functions (Qiao et al., 2019, 2020; Zapata et al., 2022).

In this paper, we consider modeling both the effect of unmeasured confounders and the temporal dependence of the replicates simultaneously. Figure 1 shows a toy example on Gaussian graphical models with unmeasured confounders and correlated replicates. We compare our proposed method with Friedman et al. (2008) that ignores both the unmeasured confounders and correlated replicates, and Chandrasekaran et al. (2010) that models the unmeasured confounders but ignores the correlated replicates. We generate the data using the same data generating mechanism as described in Section 5.3 with 50 subjects, 20 replicates for each subject, six observed variables, and 10 unmeasured confounders. We select the tuning parameters for the different methods using the stability approach (Liu et al., 2010). See also Section 3.2 for a more detailed description of the stability approach for tuning parameter selection for our proposed method. From Figure 1, we see that Friedman et al. (2008) and Chandrasekaran et al. (2010) have one and three spurious edges, respectively. Moreover, both methods fail to recover most of the true underlying edges. In contrast, our proposed method recovers the exact true underlying graph.

Recently, Tan et al. (2016) proposed to estimate a semiparametric exponential family graphical model with unmeasured confounders under the setting in which multiple replicates are collected for each subject. The main crux of their proposed method is based on a nuisance-free loss function that does not depend on the unmeasured confounders. The proposed method relies on two crucial assumptions: (i) the unmeasured confounders are constant across replicates within each subject; (ii) given the unmeasured confounders, the observed replicates within each subject are mutually independent. However, in many scientific settings, these assumptions may be violated. For instance, in the aforementioned fMRI study, unmeasured different states of mind will induce different latent effects across the brain scans and violate the constant unmeasured confounders assumption in Tan et al. (2016). Moreover, brain scans are taken every 1.5 seconds and thus the replicates are correlated.

We relax the two aforementioned key assumptions in Tan et al. (2016). Instead of assuming the unmeasured confounders are the same for all replicates, we assume that the effect induced by the unmeasured confounders is piecewise constant across replicates within

each independent subject. This is a reasonable assumption for fMRI data, since the latent effect can be always approximated by a constant in a small time interval (e.g., within 1.5 seconds). To model the correlation across replicates, we assume a one-lag vector autoregressive model on the replicates. Under the relaxed assumption, we propose a novel method for modeling exponential family graphical models with correlated replicates and unmeasured confounders. Our proposal incorporates a lasso penalty for estimating a sparse graph among the observed variables, a lasso penalty for modeling the correlation between two successive replicates, and a fused lasso penalty for modeling the piecewise constant latent effect induced by the unmeasured confounders. The resulting optimization problem is convex and can be solved using existing coordinate descent type algorithm.

Theoretically, we establish the non-asymptotic error bound for the proposed estimator. Due to the use of both lasso and fused lasso penalty, the error bound consists of both the estimation error of the lasso term and the fused lasso term. Thus, standard proof for lasso type problem in Bühlmann and Van De Geer (2011) will lead to a slower rate of convergence. To obtain a sharper rate, one needs to carefully balance these two terms by selecting the respective tuning parameters in an optimal way. By selecting the appropriate set of tuning parameters, our theoretical results reveal an interesting phenomenon on the interplay between the number of independent samples n and the number of replicates T. Finally, we show that the proposed estimator is adaptive to the absence of unmeasured confounders, i.e., our estimator matches the rate of convergence obtained by solving a lasso problem using the oracle knowledge that there are no unmeasured confounders.

An `R` package `latentgraph` will be made publicly available on CRAN.

2. Latent Variable Graphical Models with Correlated Replicates

2.1 A Review on Exponential Family Graphical Models

In this section, we provide a brief overview of the exponential family graphical model. The exponential family graphical model ensures that the joint probability density is strictly positive and contains many commonly used graphical models such as the Ising model and Poisson graphical model as its special cases (Geiger et al., 2001). Let $\boldsymbol{X} = (X_1, \ldots, X_p)^{\mathrm{T}}$ be a p-dimensional random vector, corresponding to p nodes in a graph. Then, the pairwise exponential family graphical model has the following joint density function:

$$p(\boldsymbol{x}) = \exp\left\{\sum_{j=1}^{p} f_j(x_j;\boldsymbol{\zeta}) + \frac{1}{2}\sum_{j=1}^{p}\sum_{k\neq j}\theta_{jk}x_j x_k - A(\boldsymbol{\Theta},\boldsymbol{\zeta})\right\},\ \boldsymbol{x}\in\Omega, \tag{1}$$

where $f_j(\cdot)$ is a node potential function, $\boldsymbol{\Theta} = \{\theta_{jk}\}_{1\le j<k\le p}$ is a symmetric square matrix, $\boldsymbol{\zeta}$ is a matrix of parameters for $f_j(\cdot)$, and the set Ω is the support of $\boldsymbol{X}$. The function $A(\cdot)$ is the log-partition function such that the joint density function (1) integrates to one. We provide three examples of commonly used exponential family graphical models in Section A.

The parameter $\theta_{jk} \in \mathbb{R}$ in (1) encodes the conditional dependence relationship between the jth and the kth variables, i.e., $\theta_{jk} = 0$ if and only if the jth and the kth variables are conditionally independent. Thus, estimating the exponential family graphical models amounts to estimating θ_{jk}. In principle, given n independent samples, an estimator of θ_{jk} can be obtained by maximizing the joint density of (1) for n independent samples. However,

$A(\boldsymbol{\Theta}, \boldsymbol{\zeta})$ is computationally intractable even for moderate p in general, especially when $\boldsymbol{\Theta}$ cannot be decomposed to smaller blocks. For example, for an Ising model, the log-partition function involves evaluating the summation of 2^p terms. To address the aforementioned challenge, many authors have proposed to maximize the conditional distribution of each variable, and then combine the resulting estimates to form a single graphical model (Meinshausen and Bühlmann, 2006; Ravikumar et al., 2010; Allen and Liu, 2012; Yang et al., 2012; Chen et al., 2015).

More specifically, for any node j, let $\boldsymbol{X}_{-j} = (X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p)^{\mathrm{T}} \in \mathbb{R}^{p-1}$. Then, $\boldsymbol{X}$ follows the exponential family graphical model if for any node j, the conditional density of X_j given $\boldsymbol{X}_{-j}$ is

$$p\left(x_j \mid \boldsymbol{x}_{-j}\right) = \exp\left\{f_j(x_j) + x_j \boldsymbol{\theta}_{j,-j}^{\mathrm{T}} \boldsymbol{x}_{-j} - D_j\left(\boldsymbol{\theta}_{j,-j}, f_j\right)\right\}, \tag{2}$$

where $\boldsymbol{\theta}_{j,-j} = (\theta_{j1}, \ldots, \theta_{j(j-1)}, \theta_{j(j+1)}, \ldots, \theta_{jp})^{\mathrm{T}}$ and $D_j(\boldsymbol{\theta}_{j,-j}, f_j)$ is the log-partition function that depends on $\boldsymbol{\theta}_{j,-j}$ and f_j. The exponential family graphical model can then be constructed by estimating $\boldsymbol{\theta}_{j,-j}$ for $j \in \{1, \ldots, p\}$ through fitting p generalized linear models.

2.2 Exponential Family Graphical Models with Correlated Replicates and Unmeasured Confounders

The pairwise exponential family graphical model in (1) assumes that all variables are observed and that there are no unmeasured confounders. Moreover, (1) does not accommodate correlated measurements or replicates. In this section, we propose an extension of the exponential family graphical model to accommodate both the correlated replicates and unmeasured confounders. Let $\boldsymbol{X}_t \in \mathbb{R}^p$ and $\boldsymbol{U}_t \in \mathbb{R}^q$ be vectors of the observed and unmeasured confounding random variables for the tth replicate, respectively. For notational simplicity, we assume that there are a total of T replicates. We start with the following assumption on the joint density of the replicates.

Assumption 1 *The joint conditional density of the T replicates, given the unmeasured confounders, takes the form*

$$p\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T \mid \boldsymbol{u}_1, \ldots, \boldsymbol{u}_T\right) = p\left(\boldsymbol{x}_1 \mid \boldsymbol{u}_1\right) \prod_{t=2}^{T} p\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{u}_t\right).$$

In other words, conditioned on the unmeasured confounders, the T replicates are assumed to follow a one-lag vector autoregressive model. That is, the tth replicate depends only on the $(t-1)$th replicate of the observed random variables. Moreover, the observed variables are conditionally independent of the unmeasured confounders across different replicates. One scientific motivation for the above assumption is the wake-sleep example described in the Introduction section. Brain activities (observed variables) are conditionally independent under different states of mind (the unmeasured confounders across replicates), i.e., when the subjects are awake or fall asleep. Moreover, we relax the one-lag autoregressive model for the replicates in Assumption 1 to the o-lag vector autoregressive model with $1 < o \leq T-1$ in Section B.

Definition 2 *A $(p+q)$-dimensional random vector $(\boldsymbol{X}_t^{\mathrm{T}}, \boldsymbol{U}_t^{\mathrm{T}})^{\mathrm{T}}$ follows an exponential family graphical model with correlated replicates and unmeasured confounders if the conditional density of $\boldsymbol{X}_t$ given $\boldsymbol{X}_{t-1}$ and $\boldsymbol{U}_t$ is*

$$p\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{u}_t\right) = \exp\left\{\sum_{j=1}^{p} f_{tj}(x_{tj};\boldsymbol{\zeta}) + \frac{1}{2}\sum_{j=1}^{p}\sum_{k\neq j}\theta_{jk}x_{tk}x_{tj} + \sum_{j=1}^{p}\sum_{k=1}^{p}\alpha_{jk}x_{(t-1)k}x_{tj} + \sum_{j=1}^{p}\sum_{m=1}^{q}\delta_{jm}u_{tm}x_{tj} - A\left(\boldsymbol{\Theta},\boldsymbol{\zeta}\right)\right\}, \tag{3}$$

where $f_{tj}(\cdot)$ is a node potential function, $\boldsymbol{\Theta} = \{\theta_{jk}\}_{1\leq j<k\leq p}$ is a symmetric square matrix, $\boldsymbol{\zeta}$ is a matrix of parameters for $f_{tj}(\cdot)$, and $A(\cdot)$ is the log-partition function such that the conditional density function integrates to one. Then for each node j, the conditional density of X_{tj} given $\boldsymbol{X}_{t(-j)}$, $\boldsymbol{X}_{t-1}$, and $\boldsymbol{U}_t$ is

$$p\left(x_{tj} \mid \boldsymbol{x}_{t(-j)}, \boldsymbol{x}_{t-1}, \boldsymbol{u}_t\right) = \exp\left\{f_{tj}\left(x_{tj}\right) + \sum_{k\neq j}\theta_{jk}x_{tk}x_{tj} + \sum_{k=1}^{p}\alpha_{jk}x_{(t-1)k}x_{tj} + \sum_{m=1}^{q}\delta_{jm}u_{tm}x_{tj} - D_{tj}\left(\theta_{jk},\alpha_{jk},\delta_{jm},f_{tj}\right)\right\}, \tag{4}$$

where $D_{tj}(\cdot)$ is the log-partition function such that the conditional density integrates to one.

Under the formulation in Definition 2, θ_{jk} encodes the conditional dependence relationship between the kth and jth nodes. That is, $\theta_{jk} = 0$ if and only if X_{tj} and X_{tk} are conditionally independent, given $\boldsymbol{X}_{t(-j)}$, $\boldsymbol{X}_{t-1}$, and $\boldsymbol{U}_t$ for all replicates $t = 1, \ldots, T$. The parameter α_{jk} quantifies the strength of dependency between $X_{(t-1)k}$ and X_{tj}. Finally, δ_{jm} encodes the conditional dependence relationship between the mth latent variable and the jth observed variable.

The proposed density functions (3) and (4) in Definition 2 generalize the exponential family graphical models in (1) and (2) to accommodate unmeasured confounders and correlated replicates. As suggested in Section 2.1, computing the partition function $A(\cdot)$ is computationally challenging and thus we will focus on estimating the parameter of interest θ_{jk} by utilizing the conditional density (4).

The form of the node potential function $f_{tj}(\cdot)$ and the log-partition function $D_{tj}(\cdot)$ is distribution specific. Let $f_{tj}(x_{tj}) = B_{1tj}x_{tj} + B_{2tj}x_{tj}^2 + \sum_{k=3}^{K} B_{ktj}G_{ktj}(x_{tj})$ for some scalar B_{ktj} and function $G_{ktj}(x_{tj})$. For notational simplicity, let $\eta_{tj} = B_{1tj} + \sum_{k\neq j}\theta_{jk}x_{tk} + \sum_{k=1}^{p}\alpha_{jk}x_{(t-1)k} + \sum_{m=1}^{q}\delta_{jm}u_{tm}$. In the following, we provide three examples on Gaussian graphical models, the Ising model, and the Poisson graphical models.

Example 1 *The Gaussian graphical model with correlated replicates and unmeasured confounders. The conditional distribution of X_{tj} given $\mathbf{X}_{t(-j)}$, $\mathbf{X}_{t-1}$ and $\boldsymbol{U}_t$ with $B_{2tj} = -1/2$ is given by:*

$$p\left(x_{tj} \mid \mathbf{x}_{t(-j)}, \mathbf{x}_{t-1}, \mathbf{u}_t\right) = \exp\left\{-\frac{1}{2}x_{tj}^2 + \eta_{tj}x_{tj} - D_{tj}\left(\eta_{tj}\right)\right\} \quad (x_{tj} \in \mathbb{R}), \tag{5}$$

where $f_{tj}(x_{tj}) = B_{1tj}x_{tj} - x_{tj}^2/2$ and $D_{tj}(\eta_{tj}) = \eta_{tj}^2/2 + \log(2\pi)/2$. The mean is η_{tj} and the variance is 1. Moreover, $\boldsymbol{\Theta} = [\theta_{jk}]$ is positive definite for a valid distribution.

Example 2 *The Ising model with correlated replicates and unmeasured confounders. The conditional distribution of X_{tj} given $\mathbf{X}_{t(-j)}$, $\mathbf{X}_{t-1}$ and $\boldsymbol{U}_t$ is:*

$$p\left(x_{tj} \mid \mathbf{x}_{t(-j)}, \mathbf{x}_{t-1}, \mathbf{u}_t\right) = \exp\left\{\eta_{tj} x_{tj} - D_{tj}\left(\eta_{tj}\right)\right\} \quad \left(x_{tj} \in \{0,1\}\right), \tag{6}$$

where $f_{tj}(x_{tj}) = 0$ and $D_{tj}(\eta_{tj}) = \log(1 + e^{\eta_{tj}})$.

Example 3 *The Poisson graphical model with correlated replicates and unmeasured confounders. The conditional distribution of X_{tj} given $\mathbf{X}_{t(-j)}$, $\mathbf{X}_{t-1}$ and $\boldsymbol{U}_t$ is:*

$$p\left(x_{tj} \mid \mathbf{x}_{t(-j)}, \mathbf{x}_{t-1}, \mathbf{u}_t\right) = \exp\left\{\eta_{tj} x_{tj} - \log\left(x_{tj}!\right) - D_{tj}\left(\eta_{tj}\right)\right\} \quad \left(x_{tj} \in \{0,1,\dots\}\right), \tag{7}$$

where $f_{tj}(x_{tj}) = B_{1tj} x_{tj} - \log(x_{tj}!)$ and $D_{tj}(\eta_{tj}) = \exp(\eta_{tj})$. The parameter θ_{jk} is constrained to be less than zero.

3. Method

3.1 Problem Formulation and Parameter Estimation

Suppose that there are n independent subjects $i = 1, \dots, n$ and each subject has $t = 1, \dots, T$ replicates. For notational simplicity, we assume that all subjects have the same number of replicates. The proposed method can be modified to accommodate different number of replicates across the subjects. Let $\boldsymbol{X}_{it} \in \mathbb{R}^p$ and $\boldsymbol{U}_{it} \in \mathbb{R}^q$ be the random observed variables and unmeasured confounders corresponding to the tth replicate of the ith subject, respectively. The primary goal is to estimate the conditional dependence relationships among the observed variables given the latent variables.

Inspired by the literature on measurement error models (Carroll et al., 2006), we use a functional approach to deal with the unmeasured confounders. Specifically, we treat the realization of the unmeasured confounders U_{itj} as nonrandom incidental nuisance parameters, which may differ from subject to subject. Such an approach is dated back to the so-called Neyman and Scott's problem in 1948; see Lancaster (2000) for a survey. However, in this functional approach, the graphical model involves a large number of unknown nuisance parameters such that the estimation of θ_{jk} in (4) is often inconsistent. To alleviate this problem, we further assume that for the same subject, the value of U_{itj} is piecewise constant across $t = 1, \dots, T$. In theory, this assumption may improve the estimation accuracy by reducing intrinsic dimension of the unknown incidental nuisance parameters. In practice, it is much less restrictive than assuming that the latent variables are constant as assumed in Tan et al. (2016). For example, in the analysis of brain development for children with attention deficit hyperactivity disorder, unmeasured covariates such as demographic factors can be treated as unmeasured confounders and their variation over scans are piecewise constant.

Assumption 3 *The unmeasured confounders are piecewise constant across replicates. That is, we assume for the ith sample, we have l knots with unknown location denoted as $k_{i1}, k_{i2}, \dots, k_{il}$ and let $k_{i0} = 1$, $k_{i(l+1)} = T$. Then the jth unmeasured confounder at the tth replicate for the ith subject satisfies*

$$U_{itj} = \sum_{a=1}^{l+1} g_{iaj} \mathbb{1}\left(k_{i(a-1)} \le t \le k_{ia}\right),$$

where g_{iaj} is an unknown constant and $\mathbb{1}(\cdot)$ is an indicator function.

Figure 2 provides a schematic of the assumptions in Tan et al. (2016) and our proposal for the ith sample. Figure 2(a) represents the assumptions in Tan et al. (2016): the tth and $(t-1)$th replicates for the observed variables are independent and the unmeasured confounders are constant across replicates. Figure 2(b) depicts the assumptions for our proposed method: the tth replicate of the observed variables are conditionally dependent on the $(t-1)$th replicate, and the unmeasured confounders are piecewise constant that may change across replicates.

We now reformulate the conditional density in (4) under Assumption 3. Let $\Delta_{tj} = \sum_{m=1}^{q} \delta_{jm} u_{tm}$. Then, (4) in Definition 2 can be rewritten as:

$$\begin{aligned} & p\left(x_{tj} \mid \mathbf{x}_{t(-j)}, \mathbf{x}_{t-1}, \Delta_{tj}\right) \\ & = \exp\left\{ f_{tj}\left(x_{tj}\right) + \sum_{k \neq j} \theta_{jk} x_{tk} x_{tj} + \sum_{k=1}^{p} \alpha_{jk} x_{(t-1)k} x_{tj} + \Delta_{tj} x_{tj} - D_{tj}\left(\theta_{jk}, \alpha_{jk}, \Delta_{tj}, f_{tj}\right) \right\}. \end{aligned} \tag{8}$$

We now construct a joint likelihood function for n subjects, each of which has T replicates using (8). Let $\boldsymbol{\alpha}_j = (\alpha_{j1}, \alpha_{j2}, \ldots, \alpha_{jp})^{\mathrm{T}} \in \mathbb{R}^p$, $\boldsymbol{\theta}_{j,-j} = (\theta_{j1}, \ldots, \theta_{j,j-1}, \theta_{j,j+1}, \ldots, \theta_{jp})^{\mathrm{T}} \in \mathbb{R}^{p-1}$, and $\boldsymbol{\Delta}_j = (\Delta_{11j}, \Delta_{12j}, \ldots, \Delta_{1Tj}, \Delta_{21j}, \Delta_{22j}, \ldots, \Delta_{nTj})^{\mathrm{T}} \in \mathbb{R}^{nT}$. Thus, we estimate $\boldsymbol{\theta}_{j,-j}$, $\boldsymbol{\alpha}_j$, and $\boldsymbol{\Delta}_j$ by solving

$$\underset{\boldsymbol{\theta}_{j,-j}, \boldsymbol{\alpha}_j, \boldsymbol{\Delta}_j}{\text{minimize}} \quad \left\{ -\frac{1}{nT} l\left(\boldsymbol{\theta}_{j,-j}, \boldsymbol{\alpha}_j, \boldsymbol{\Delta}_j\right) + \lambda \left\|\boldsymbol{\theta}_{j,-j}\right\|_1 + \beta \left\|\boldsymbol{\alpha}_j\right\|_1 + \gamma \left\| \left(\mathbf{I}_n \otimes \mathbf{C}\right) \boldsymbol{\Delta}_j \right\|_1 \right\}, \tag{9}$$

where $l(\boldsymbol{\theta}_{j,-j}, \boldsymbol{\alpha}_j, \boldsymbol{\Delta}_j) = \sum_{i=1}^{n} \sum_{t=1}^{T} \log p(x_{itj} | \boldsymbol{x}_{it(-j)}, \boldsymbol{x}_{i(t-1)}, \Delta_{itj})$. Here, λ, β, and γ are the sparsity inducing tuning parameters, $\mathbf{I}_n$ is an n-dimensional identity matrix, and $\mathbf{C} \in \mathbb{R}^{(T-1)\times T}$ is the discrete first derivative matrix defined as follows:

$$\mathbf{C} = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix}.$$

The penalty term $\left\| \left(\mathbf{I}_n \otimes \mathbf{C}\right) \boldsymbol{\Delta}_j \right\|_1$ is essentially a fused lasso penalty on $\boldsymbol{\Delta}_{ij} = (\Delta_{i1j}, \ldots, \Delta_{iTj})^{\mathrm{T}}$ for each subject. In other words, we assume that the difference between two consecutive elements of $\boldsymbol{\Delta}_{ij}$ is sparse, i.e., most of the values of $\Delta_{i(t+1)j} - \Delta_{itj}$ for $t = 1, \ldots, T-1$ is zero. Imposing the fused lasso penalty on $\boldsymbol{\Delta}_{ij}$ is motivated by the piece-wise constant assumption on the unmeasured confounders u_{itj}. Note that we do not need to assume that the association between the unmeasured confounders and the observed variables, δ_{jm}, are sparse. Additionally, all subjects share the same $\boldsymbol{\theta}_{j,-j}$ and $\boldsymbol{\alpha}_j$. However, for a given t and j, we allow Δ_{itj} to be different across different subjects.

Remark 4 *In practice, it may be more practical to allow the number of replicates to vary across subjects. Our proposed method can be easily generalized to accommodate this. Specifically, let T_i be the number of replicates for the ith subject. The loss function can then be*

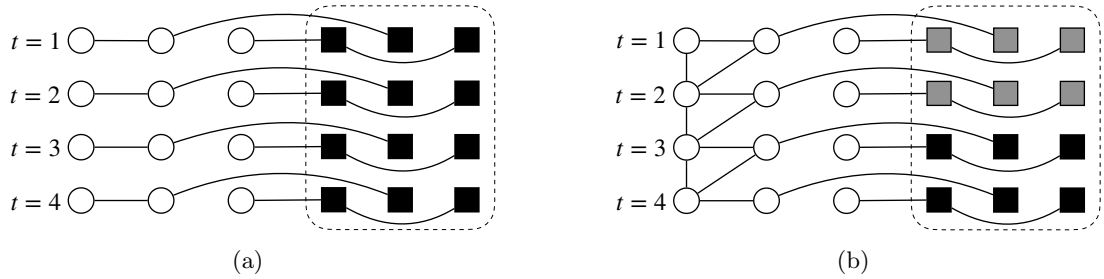


Figure 2: Panels (a) and (b) correspond to the assumptions on the replicates and unmeasured confounders of Tan et al. (2016) and our proposed method, respectively. There are four replicates for each subject, i.e., $t = \{1, 2, 3, 4\}$. Hollow circles and squares represent the observed variables and unmeasured confounders, respectively. In panel (b), the color of the unmeasured confounders changes from gray to black, indicating that the value of unmeasured confounders are allowed to change across replicates.

rewritten as

$$l(\boldsymbol{\theta}_{j,-j}, \boldsymbol{\alpha}_j, \boldsymbol{\Delta}_j) = \sum_{i=1}^{n} \sum_{t=1}^{T_i} \log p(x_{itj} | \boldsymbol{x}_{it(-j)}, \boldsymbol{x}_{i(t-1)}, \Delta_{itj}).$$

The same algorithm described in the following section can be used to estimate $\boldsymbol{\theta}_{j,-j}, \boldsymbol{\alpha}_j$, *and* $\boldsymbol{\Delta}_j$.

Remark 5 *The piecewise constant assumption on the unmeasured confounders in Assumption 3 can be restrictive to data applications in which the unmeasured confounders vary across replicates rapidly. Assumption 3 can be relaxed to piecewise linear, quadratic, or higher-order polynomials on the unmeasured confounders. This can be achieved by substituting the fused lasso penalty* $\|(\mathbf{I}_n \otimes \mathbf{C})\boldsymbol{\Delta}_j\|_1$ *in* (9) *by* $\|(\mathbf{I}_n \otimes \mathbf{C}^{(k+1)})\boldsymbol{\Delta}_j\|_1$, *where* $\mathbf{C}^{(k+1)}$ *is the discrete difference operator of order* $k+1$ *(Tibshirani, 2014; Wang et al., 2016). Using the aforementioned penalty with* $k = \{0, 1, 2, 3\}$ *encourages the effects induced by the unmeasured confounders to be piecewise constant, linear, quadratic, and cubic, respectively. We leave this for future work and refer the reader to Tibshirani (2014) for details on the discrete difference operator.*

3.2 An Algorithm for Solving (9)

In this section, we show that (9) can be transformed into a lasso problem and solved using the `R` package `glmnet`. Our proposed method yields non-symmetric estimates of $\boldsymbol{\Theta} = [\theta_{jk}]$. The intersection or union rules described in Meinshausen and Bühlmann (2006) can be used to construct a symmetric estimator for $\boldsymbol{\Theta}$. We start with the Gaussian graphical model in Example 1.

Let $\mathbf{x}_j = (x_{11j}, x_{12j}, \ldots, x_{1Tj}, x_{21j}, x_{22j}, \ldots, x_{nTj})^{\mathrm{T}} \in \mathbb{R}^{nT}$, $\mathbf{X}_{i(-j)} = (\mathbf{x}_{i1(-j)}, \mathbf{x}_{i2(-j)}, \ldots, \mathbf{x}_{iT(-j)})^{\mathrm{T}} \in \mathbb{R}^{T \times (p-1)}$, and let $\mathbf{X}_{i(T-1)} = (\mathbf{x}_{i0}, \mathbf{x}_{i1}, \ldots, \mathbf{x}_{i(T-1)})^{\mathrm{T}} \in \mathbb{R}^{T \times p}$. In addition, let $\mathbf{X}^{\otimes}_{-j} = (\mathbf{X}^{\mathrm{T}}_{1(-j)}, \ldots, \mathbf{X}^{\mathrm{T}}_{n(-j)})^{\mathrm{T}} \in \mathbb{R}^{(nT) \times (p-1)}$ and let $\mathbf{X}^{\otimes}_{T-1} = (\mathbf{X}^{\mathrm{T}}_{1(T-1)}, \ldots, \mathbf{X}^{\mathrm{T}}_{n(T-1)})^{\mathrm{T}} \in \mathbb{R}^{(nT) \times p}$.

Then, from Example 1, the canonical parameter $\boldsymbol{\eta}_j = \mathbf{B}_j + \mathbf{X}^{\otimes}_{-j}\boldsymbol{\theta}_{j,-j} + \mathbf{X}^{\otimes}_{T-1}\boldsymbol{\alpha}_j + \boldsymbol{\Delta}_j$, where $\mathbf{B}_j = \mathbf{1}_n \otimes (B_{11j}, B_{12j}, \ldots, B_{1Tj})^{\mathrm{T}}$ and $\mathbf{1}_n$ is an n-dimensional vector of ones. For simplicity, we assume that the data is centered, such that $\mathbf{B}_j = \mathbf{0}$. In the context of Gaussian graphical models, (9) reduces to

$$\underset{\boldsymbol{\theta}_{j,-j},\boldsymbol{\alpha}_j,\boldsymbol{\Delta}_j}{\text{minimize}} \quad \left\{ \frac{1}{2nT} \left\| \mathbf{x}_j - \boldsymbol{\eta}_j \right\|_2^2 + \lambda \left\| \boldsymbol{\theta}_{j,-j} \right\|_1 + \beta \left\| \boldsymbol{\alpha}_j \right\|_1 + \gamma \left\| \left(\mathbf{I}_n \otimes \mathbf{C} \right) \boldsymbol{\Delta}_j \right\|_1 \right\}. \tag{10}$$

Optimization problem (10) involves a fused lasso type penalty on $\boldsymbol{\Delta}_j$, and can be rewritten into a lasso problem by a change of variable. To this end, let $\mathbf{E} = \mathbf{I}_n \otimes \mathbf{1}_T^{\mathrm{T}} \in \mathbb{R}^{n \times (nT)}$, $\widetilde{\mathbf{C}} = ((\mathbf{I}_n \otimes \mathbf{C})^{\mathrm{T}}, \mathbf{E}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{(nT)\times(nT)}$, and $\mathbf{H}_j = \widetilde{\mathbf{C}}\boldsymbol{\Delta}_j = [\{(\mathbf{I}_n \otimes \widetilde{\mathbf{C}}^{-1}\mathbf{C})\boldsymbol{\Delta}_j\}^{\mathrm{T}}, (\mathbf{E}\boldsymbol{\Delta}_j)^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{nT}$. Then, (10) can be rewritten as

$$\underset{\boldsymbol{\theta}_{j,-j},\boldsymbol{\alpha}_j,\mathbf{H}_j}{\text{minimize}} \quad \left\{ \frac{1}{2nT} \| \mathbf{x}_j - \boldsymbol{\eta}_j \|_2^2 + \lambda \| \boldsymbol{\theta}_{j,-j} \|_1 + \beta \| \boldsymbol{\alpha}_j \|_1 + \gamma \| \mathbf{H}_{j1} \|_1 \right\}, \tag{11}$$

where $\boldsymbol{\eta}_j = \mathbf{X}^{\otimes}_{-j}\boldsymbol{\theta}_{j,-j} + \mathbf{X}^{\otimes}_{T-1}\boldsymbol{\alpha}_j + \widetilde{\mathbf{C}}^{-1}\mathbf{H}_j$ and $\mathbf{H}_{j1} = (\mathbf{I}_n \otimes \mathbf{C})\boldsymbol{\Delta}_j \in \mathbb{R}^{n(T-1)}$. To further simplify (11), let $\widetilde{\mathbf{x}}_j = (\mathbf{X}^{\otimes}_{-j}, \mathbf{X}^{\otimes}_{T-1}, \widetilde{\mathbf{C}}^{-1})$ and $\boldsymbol{\Theta}_j = (\boldsymbol{\theta}_{j,-j}^{\mathrm{T}}, \boldsymbol{\alpha}_j^{\mathrm{T}}, \mathbf{H}_j^{\mathrm{T}})^{\mathrm{T}}$. Then, (11) can be rewritten as

$$\underset{\boldsymbol{\theta}_{j,-j},\boldsymbol{\alpha}_j,\mathbf{H}_j}{\text{minimize}} \quad \left\{ \frac{1}{2nT} \| \mathbf{x}_j - \widetilde{\mathbf{x}}_j\boldsymbol{\Theta}_j \|_2^2 + \lambda \| \boldsymbol{\theta}_{j,-j} \|_1 + \beta \| \boldsymbol{\alpha}_j \|_1 + \gamma \| \mathbf{H}_{j1} \|_1 \right\}, \tag{12}$$

which we solve using `glmnet` package in `R` by treating $\mathbf{x}_j$ as response variable and $\widetilde{\mathbf{x}}_j$ as covariates.

In the context of the exponential family graphical models, the loss function $l(\cdot)$ in (9) does not take the form of a squared error loss and is different for each distribution. We derive a general algorithm for fitting exponential family graphical models by a quadratic approximation on $l(\cdot)$ through the second-order Taylor expansion. That is, starting with an initial value $\widehat{\boldsymbol{\eta}}_j^0$, we consider solving the following optimization problem iteratively until convergence:

$$\begin{aligned} &(\boldsymbol{\theta}_{j,-j}^k, \boldsymbol{\alpha}_j^k, \mathbf{H}_j^k) = \\ &\underset{\boldsymbol{\theta}_{j,-j},\boldsymbol{\alpha}_j,\mathbf{H}_j}{\text{argmin}} \quad \frac{L}{2nT} \left\| \frac{1}{L}\mathbf{x}_j - \boldsymbol{\eta}_j + \widehat{\boldsymbol{\eta}}_j^{k-1} - \frac{1}{L}\boldsymbol{D}_j'(\widehat{\boldsymbol{\eta}}_j^{k-1}) \right\|_2^2 + \lambda \left\| \boldsymbol{\theta}_{j,-j} \right\|_1 + \beta \left\| \boldsymbol{\alpha}_j \right\|_1 + \gamma \left\| \mathbf{H}_{j1} \right\|_1, \end{aligned} \tag{13}$$

where $\boldsymbol{\eta}_j = \mathbf{X}^{\otimes}_{-j}\boldsymbol{\theta}_{j,-j} + \mathbf{X}^{\otimes}_{T-1}\boldsymbol{\alpha}_j + \widetilde{\mathbf{C}}^{-1}\mathbf{H}_j$ and L is chosen such that $l''(\boldsymbol{\eta}_j) \preceq L\boldsymbol{I}$. Note that at the kth iteration, $\widehat{\boldsymbol{\eta}}_j^{k-1}$ and $\boldsymbol{D}_j'(\widehat{\boldsymbol{\eta}}_j^{k-1})/L$ are both constants. Thus, the loss function is quadratic in $\boldsymbol{\theta}_j$. Let $\mathbf{y}_j = \mathbf{x}_j/L + \widehat{\boldsymbol{\eta}}_j^{k-1} - \boldsymbol{D}_j'(\widehat{\boldsymbol{\eta}}_j^{k-1})/L$, then (13) can be rewritten as

$$(\boldsymbol{\theta}_{j,-j}^k, \boldsymbol{\alpha}_j^k, \mathbf{H}_j^k) = \underset{\boldsymbol{\theta}_{j,-j},\boldsymbol{\alpha}_j,\mathbf{H}_j}{\text{argmin}} \quad \frac{L}{2nT} \left\| \mathbf{y}_j - \widetilde{\mathbf{x}}_j\boldsymbol{\Theta}_j \right\|_2^2 + \lambda \left\| \boldsymbol{\theta}_{j,-j} \right\|_1 + \beta \left\| \boldsymbol{\alpha}_j \right\|_1 + \gamma \left\| \mathbf{H}_{j1} \right\|_1, \tag{14}$$

which can be solved directly using the `glmnet` package in `R` by treating $\mathbf{y}_j$ to be the response and $\widetilde{\mathbf{x}}_j$ to be the covariates.

Optimization problem (14) involves the specification of three tuning parameters: λ, β, and γ. In practice, we propose to select these tuning parameters using a stability metric, recommended by Lim and Yu (2016). Specifically, we partition the independent subjects into five subsets, each of which consists of 80% of the total number of samples. Subsequently, for each subset, we estimate the parameters and compute the estimation stability metric for each variable as follows:

$$\widehat{\mathbf{m}}_j = \frac{1}{5}\sum_{\ell=1}^{5} \widehat{\boldsymbol{\eta}}_j^{\ell}, \qquad \mathrm{ES}_j = \frac{(1/5)\sum_{l=1}^{5} \left\|\widehat{\boldsymbol{\eta}}_j^{\ell} - \widehat{\mathbf{m}}_j\right\|_2^2}{\left\|\widehat{\mathbf{m}}_j\right\|_2^2}, \tag{15}$$

where $\widehat{\boldsymbol{\eta}}_j^{\ell} = \mathbf{X}_{-j}^{\otimes}\widehat{\boldsymbol{\theta}}_{j,-j}^{\ell} + \mathbf{X}_{T-1}^{\otimes}\widehat{\boldsymbol{\alpha}}_j^{\ell} + \widehat{\boldsymbol{\Delta}}_j^{\ell}$. Here, $\widehat{\boldsymbol{\theta}}_{j,-j}^{\ell}$, $\widehat{\boldsymbol{\alpha}}_j^{\ell}$, and $\widehat{\boldsymbol{\Delta}}_j^{\ell}$ are estimators obtained by solving (14) using the ℓth subset of the data. Finally, we compute the mean estimation stability metric $\mathrm{ES} = \sum_{j=1}^{p} \mathrm{ES}_j/p$ and select the set of tuning parameters that yields the smallest mean estimation stability.

4. Theoretical Results

In this section, we derive non-asymptotic upper bounds for the estimation error of $\widehat{\boldsymbol{\theta}}_{j,-j}$, $\widehat{\boldsymbol{\alpha}}_j$, and $\widehat{\boldsymbol{\Delta}}_j$. In particular, we aim to provide upper bounds for $\left\|\widehat{\boldsymbol{\theta}}_{j,-j} - \boldsymbol{\theta}^*_{j,-j}\right\|_1 + \left\|\widehat{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}^*_j\right\|_1 + (nT)^{-1/2}\left\|\widehat{\boldsymbol{\Delta}}_j - \boldsymbol{\Delta}^*_j\right\|_2$ under two scenarios in which the number of samples n is less than and greater than the number of replicates T, respectively. Throughout this section, we analyze the theoretical properties of the proposed estimator in the context of Gaussian graphical models. While the general exponential family graphical models can be analyzed in a similar way, the theoretical results are suboptimal compared to the Gaussian setting, partly because it lacks a sharp bound due to the fused lasso penalty. We defer the theoretical results for general exponential family graphical models and further discussion to Section E of the Appendix.

Recall that, for the ith subject, tth replicate, and jth variable, we assume the model

$$X_{itj} = \mathbf{X}_{it(-j)}\boldsymbol{\theta}^*_{j,-j} + \mathbf{X}_{i(t-1)}\boldsymbol{\alpha}^*_j + \Delta^*_{itj} + \epsilon_{itj}, \tag{16}$$

where the random noise $\epsilon_{itj} \sim N\{0, (\sigma^{\epsilon}_{jj,t})^2\}$ is independent of $\mathbf{X}_{it(-j)}$ and $\mathbf{X}_{i(t-1)}$. Note that the random noise is independent but may not be identically distributed, i.e., the random noise in (16) can have different variance. For notational simplicity, throughout the manuscript, let $(\sigma^{\epsilon}_m)^2 = \max_{t,j}\{(\sigma^{\epsilon}_{jj,t})^2\}$. Let $\Delta_m = \max_{i,t,j} |\Delta^*_{itj} - \Delta^*_{i(t-1)j}|$ be the maximum difference between two consecutive elements of the sequence $\Delta^*_{i1j}, \Delta^*_{i2j}, \ldots, \Delta^*_{iTj}$, and let $\tau = \max_{i,j}\sum_{t=2}^{T} I(\Delta^*_{itj} \neq \Delta^*_{i(t-1)j})$ be the maximum number of jumps between the consecutive elements Δ^*_{itj} and $\Delta^*_{i(t-1)j}$. Let $\Delta_{\max} = \Delta_m + 1$.

We start with imposing an assumption on the mean and the covariance matrix of the replicates for each independent subject.

Assumption 6 *For the ith subject and jth variable, let $\boldsymbol{X}_{ij} = (X_{i1j}, \ldots, X_{iTj})^{\mathrm{T}} \sim N(\boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_{jj})$. Assume that the mean of X_{itj} is bounded by a constant, i.e., $|\mu_{itj}| \leq \mu_{\max}$. In addition, assume that the ℓ_2-norm of $\boldsymbol{\mu}_{ij}$ satisfies $\left\|\boldsymbol{\mu}_{ij}\right\|_2 \leq \mu_m \min(c_5^{\frac{1}{3}} \log^{\frac{1}{3}}(2pnT) n^{\frac{1}{3}} T^{\frac{1}{6}}, \sqrt{T})$ with $c_5 = 2\Delta_{\max}\tau/\pi$. Finally, assume that there exists a constant $\kappa > 0$ such that $\max_{1\leq j\leq p}\left\|\boldsymbol{\Sigma}_{jj}\right\|_{\mathrm{op}} \leq \kappa$, where $\|\boldsymbol{\Sigma}_{jj}\|_{\mathrm{op}}$ is the operator norm of $\boldsymbol{\Sigma}_{jj}$.*

Recall that the mean $\boldsymbol{\mu}_{ij}$ depends on the latent effect Δ^*_{itj} in (16). For technical convenience, similar to Hall et al. (2016), we assume that μ_{itj} is bounded in order to control X_{itj}. In addition, we further require that the ℓ_2-norm of $\boldsymbol{\mu}_{ij}$ cannot grow too fast with T and n, which is mainly used to control the magnitude of $\sum_{t=1}^{T} X_{itj}(\bar{\Delta}_{itj} - \Delta^*_{itj})$ with some intermediate estimator $\bar{\Delta}_{itj}$. In Section F.4 of the Appendix, we provide some examples in which Assumption 6 holds with high probability.

Next, we define some additional notation. Let $\boldsymbol{\omega}^*_j = \{(\boldsymbol{\theta}^*_{j,-j})^{\mathrm{T}}, (\boldsymbol{\alpha}^*_j)^{\mathrm{T}}\}^{\mathrm{T}}$ and let $\mathcal{S}_j = \{k : \omega^*_{jk} \neq 0\}$ be the active set. We denote $s_j = |\mathcal{S}_j|$ as the cardinality of $\mathcal{S}_j$. Let $\mathbf{Y}_{itj} = (\mathbf{X}_{it(-j)}, \mathbf{X}_{i(t-1)}) \in \mathbb{R}^{1\times(2p-1)}$ and $\mathbf{Y}_j = (\mathbf{Y}^{\mathrm{T}}_{11j}, \mathbf{Y}^{\mathrm{T}}_{12j}, \ldots, \mathbf{Y}^{\mathrm{T}}_{1Tj}, \mathbf{Y}^{\mathrm{T}}_{21j}, \ldots, \mathbf{Y}^{\mathrm{T}}_{nTj})^{\mathrm{T}} \in \mathbb{R}^{nT\times(2p-1)}$. Moreover, let $\boldsymbol{\omega}_j^{\mathcal{S}_j}$ and $\boldsymbol{\omega}_j^{\mathcal{S}_j^c}$ be subvectors of $\boldsymbol{\omega}_j$ with indices $\mathcal{S}_j$ and $\mathcal{S}_j^c$, respectively.

Assumption 7 *Let $\widehat{\boldsymbol{\Sigma}}_j = (\mathbf{Y}_j - \mathbf{U}_j)^{\mathrm{T}}(\mathbf{Y}_j - \mathbf{U}_j)/(nT)$, where $\mathbf{U}_j = \mathbb{E}(\mathbf{Y}_j)$. With some constants $\mathcal{C}$ and ϕ_0, assume that*

$$\left\|(\boldsymbol{\omega}^*_j)^{\mathcal{S}_j}\right\|_1^2 \leq \frac{\left(\boldsymbol{\omega}^*_j\right)^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_j \boldsymbol{\omega}^*_j s_j}{\phi_0} + \frac{2\mathcal{C}s_j \log\{nT(2p-1)\}}{\phi_0\sqrt{nT}} \left\|\boldsymbol{\omega}^*_j\right\|_1^2.$$

This assumption can be viewed as a variation of the compatibility assumption (Bühlmann and Van De Geer, 2011), which is commonly used to show the ℓ_1 convergence rate of the lasso estimator. Since the rows of $\mathbf{Y}_j$ are dependent, verifying this assumption is technically more challenging than that with independent and identically distributed data. In Section F.5 of the Appendix, we have shown that Assumption 7 holds with probability at least $1 - 2/(nT)^2$.

In the following, we present our main results on the estimation error of our proposed estimator under two scenarios, depending on the relative magnitude of the number of replicates and the number of independent samples. For notational simplicity, let $\sigma_m = \max\{\sqrt{2\kappa}, \sqrt{2}\sigma^{\epsilon}_m, 1\}$ where κ is as defined in Assumption 6. Moreover, we will use the notation $\mathcal{C}_i$ for $i = 1, \ldots, 8$ to denote some constants that do not depend on n, p, T, τ, $\Delta_{\max}$, and σ_m; see the proof in Section F.2 of the Appendix for the specific values of $\mathcal{C}_i$.

Theorem 8 *Assume that $T > c_5 \log^{-1}(2pn) \log^{\frac{1}{2}}(2pnT) n$ holds. Set the tuning parameters as*

$$\gamma = \mathcal{C}_1 \sigma_m \sqrt{\log(2pnT)/\max\{1, \lfloor\{c_5 \log^{\frac{1}{2}}(2pnT)/\log(2pn)\}^{\frac{2}{3}} T^{\frac{1}{3}} n^{\frac{2}{3}}\rfloor\}} n^{-1} T^{-\frac{1}{2}},$$

$$\lambda = \beta = 2\log(T)\log(nTp) n^{-\frac{1}{6}} T^{-\frac{1}{3}},$$

in (11). When

$$\sqrt{s_j} c_5^{\frac{1}{3}} \log^{\frac{1}{3}}(2pnT) \mu_{\max} n^{-\frac{1}{6}} T^{-\frac{1}{3}} + \sqrt{2\mathcal{C}s_j \log\{nT(2p-1)\}}\,(nT)^{-\frac{1}{4}} \leq \frac{1}{16}\phi_0^{\frac{1}{2}},$$

$$2\max\left(c_1 n^{-\frac{1}{6}} T^{-\frac{1}{3}}, c_2 n^{-\frac{1}{6}} T^{-\frac{1}{3}} + c_3 n^{-\frac{1}{2}} T^{-\frac{1}{2}}\right) \leq 1,$$

$n, T, p \geq 6$, and under Assumptions 1–7, we have

$$\left\|\widehat{\boldsymbol{\theta}}_{j,-j} - \boldsymbol{\theta}^*_{j,-j}\right\|_1 + \left\|\widehat{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}^*_j\right\|_1 + \frac{1}{\sqrt{nT}}\left\|\widehat{\boldsymbol{\Delta}}_j - \boldsymbol{\Delta}^*_j\right\|_2 \leq 2\max\left(c_1 n^{-\frac{1}{6}}T^{-\frac{1}{3}}, c_2 n^{-\frac{1}{6}}T^{-\frac{1}{3}} + c_3 n^{-\frac{1}{2}}T^{-\frac{1}{2}}\right),$$

with probability at least $1-3\exp[-\lfloor\{\log(2pnT)\Delta^2_{\max}\tau^2 Tn^2/\pi^2\}^{1/3}\rfloor]-[(n+1)p+\sqrt{\log(nT)}]/[n(T-1)\sqrt{\log\{n(T-1)\}}]-2\exp(-\min[\log^2(T)/\{2\sigma^\epsilon_m\sqrt{\mu^2_{\max}+(\sigma^X_m)^2}\}^2, \log(T)/|2\sigma^\epsilon_m\sqrt{\mu^2_{\max}+(\sigma^X_m)^2}|]/2)$, *where*

$$\begin{aligned}
c_1 &= \mathcal{C}_2(1+\sqrt{\phi_0})s_j\log(T)\log(nTp)/\phi_0,\\
c_2 &= \mathcal{C}_3\sigma^2_m c_5^{\frac{2}{3}}(\mu_{\max}+1)^2\log^{-1}(T)\log^{-\frac{1}{3}}(nTp) + \mathcal{C}_5\sigma_m(\mu_{\max}+1)c_5^{\frac{1}{3}}\log^{\frac{1}{3}}(2pnT),\\
c_3 &= \mathcal{C}_5\sigma_m\sqrt{\log(2pnT)} + \mathcal{C}_3\sigma^2_m\log^{-1}(T) + \mathcal{C}_4 c_5^{\frac{1}{3}}\sigma^2_m(\mu_{\max}+1)\log^{-1}(T)\log^{-\frac{1}{6}}(nTp).
\end{aligned}$$

Theorem 9 *Assume that* $T \leq c_5\log^{-1}(2pn)\log^{\frac{1}{2}}(2pnT)n$ *holds. Set the tuning parameters as*

$$\gamma = \mathcal{C}_1\sigma_m\sqrt{\log(2pnT)/(T-1)}n^{-1}T^{-1/2}, \text{ and } \lambda = \beta = 2\log(T)\log(nTp)T^{-1/2},$$

in (11). When

$$\sqrt{s_j}c_5^{\frac{1}{3}}\mu_{\max}\log^{\frac{1}{3}}(2pnT)n^{-\frac{1}{6}}T^{-\frac{1}{3}} + \sqrt{2\mathcal{C}s_j\log\left\{nT\left(2p-1\right)\right\}}\left(nT\right)^{-\frac{1}{4}} \leq \frac{1}{16}\phi_0^{\frac{1}{2}},$$

$$T^{\frac{1}{2}} \geq 2\max\left\{c_1^{'}, c_2^{'}+c_3^{'}\right\},$$

$n, T, p \geq 6$, and under Assumptions 1–7, we obtain

$$\left\|\widehat{\boldsymbol{\theta}}_{j,-j} - \boldsymbol{\theta}^*_{j,-j}\right\|_1 + \left\|\widehat{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}^*_j\right\|_1 + \frac{1}{\sqrt{nT}}\left\|\widehat{\boldsymbol{\Delta}}_j - \boldsymbol{\Delta}^*_j\right\|_2 \leq 2\max\left\{c_1^{'}, c_2^{'}+c_3^{'}\right\}T^{-\frac{1}{2}},$$

with probability at least $1-[(n+1)p+\sqrt{\log(nT)}]/[n(T-1)\sqrt{\log\{n(T-1)\}}]-\exp\{-(T-1)\}-2\exp(-\min[\log^2(T)/\{2\sigma^\epsilon_m\sqrt{\mu^2_{\max}+(\sigma^X_m)^2}\}^2, \log(T)/|2\sigma^\epsilon_m\sqrt{\mu^2_{\max}+(\sigma^X_m)^2}|]/2)$, *where*

$$\begin{aligned}
c_1^{'} &= \mathcal{C}_2(1+\sqrt{\phi_0})s_j\log(T)\log(nTp)/\phi_0,\\
c_2^{'} &= \mathcal{C}_3\Delta_{\max}\tau\sigma^2_m(\mu_{\max}+3)^2\log^{\frac{1}{2}}(2nTp)/\{\log(2np)\log(T)\},\\
c_3^{'} &= \mathcal{C}_5\Delta^{\frac{1}{2}}_{\max}\tau^{\frac{1}{2}}\sigma_m(\mu_{\max}+4)\log^{\frac{3}{4}}(2npT)/\log^{\frac{1}{2}}(2np).
\end{aligned}$$

The results in Theorems 8–9 are non-asymptotic results with explicit probability. In the asymptotic setting in which $n, T, p \to \infty$, with probability 1, our estimation error is $o_p(1)$.

Remark 10 *Since the estimator* $(\widehat{\boldsymbol{\theta}}_{j,-j}, \widehat{\boldsymbol{\alpha}}_j, \widehat{\boldsymbol{\Delta}}_j)$ *is obtained by solving a lasso type problem in (11), one may follow the standard proof in Bühlmann and Van De Geer (2011) to establish the error bound of* $(\widehat{\boldsymbol{\theta}}_{j,-j}, \widehat{\boldsymbol{\alpha}}_j, \widehat{\boldsymbol{\Delta}}_j)$. *However, this will lead to slower rates of convergence than those obtained in Theorems 8–9 due to the structure of the fused lasso penalty not being fully exploited. In a recent paper, Wang et al. (2016) established the sharp rates for the fused lasso estimator based on the incoherence property of the discrete difference operator; see*

also Tibshirani (2014). Our proof strategy is partially inspired by their technique. However, there are several important differences. First, we decouple the temporal dependence among random variables using martingales. Second, due to the combination of lasso and fused lasso penalties in (11), the error bound consists of both the estimation error of lasso $(\widehat{\boldsymbol{\theta}}_{j,-j} - \boldsymbol{\theta}^*_{j,-j}, \widehat{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}^*_j)$ *and the error of fused lasso* $\widehat{\boldsymbol{\Delta}}_j - \boldsymbol{\Delta}^*_j$*. To obtain a sharp rate, one needs to carefully quantify and balance these two terms in the proof by choosing their tuning parameters* γ, λ*, and* β *in an optimal way. Our theorems reveal that the optimal choices of* γ, λ*, and* β *differ depending on whether* T *exceeds* $c_5 \log^{-1}(2pn) \log^{\frac{1}{2}}(2pnT)n$ *and vice versa.*

To further simplify the results in Theorems 8 and 9, assume that ϕ_0, σ_m, $\Delta_{\max}$, and τ are all constants. Then, Theorems 8 and 9 imply that with probability tending to 1,

$$\big\|\widehat{\boldsymbol{\theta}}_{j,-j} - \boldsymbol{\theta}^*_{j,-j}\big\|_1 + \big\|\widehat{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}^*_j\big\|_1 + \frac{1}{\sqrt{nT}}\big\|\widehat{\boldsymbol{\Delta}}_j - \boldsymbol{\Delta}^*_j\big\|_2 \lesssim \begin{cases} s_j \log(T) \log(nTp) n^{-1/6} T^{-1/3}, & \text{if } T \gg n, \\ s_j \log(T) \log(nTp) T^{-1/2}, & \text{if } T \ll n, \end{cases} \tag{17}$$

where the notation $a_n \lesssim b_n$ stands for $a_n = Cb_n$ with C being a constant and $a_n \gtrsim b_n$ is defined similarly. Following the standard proof in Bühlmann and Van De Geer (2011), if the incidental nuisance parameter $\boldsymbol{\Delta}_j$ is known, we can obtain the following error bound for the lasso estimator

$$\big\|\bar{\boldsymbol{\theta}}_{j,-j} - \boldsymbol{\theta}^*_{j,-j}\big\|_1 + \big\|\bar{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}^*_j\big\|_1 \lesssim s_j \log^{1/2}(p)(nT)^{-1/2}, \tag{18}$$

where $(\bar{\boldsymbol{\theta}}_{j,-j}, \bar{\boldsymbol{\alpha}}_j)$ minimizes the loss function $(2nT)^{-1}\|\mathbf{x}_j - \boldsymbol{\eta}_j\|_2^2 + \lambda\|\boldsymbol{\theta}_{j,-j}\|_1 + \beta\|\boldsymbol{\alpha}_j\|_1$ with $\boldsymbol{\Delta}_j$ fixed and $\boldsymbol{\eta}_j$ is as defined in (11), and nT can be viewed as the sample size. Due to the presence of a large amount of unknown incidental nuisance parameters $\boldsymbol{\Delta}^*_j$, the rate in (17) is nonstandard and slower than the regular lasso estimator, $s_j \log^{1/2}(p)(nT)^{-1/2}$.

In the literature on incidental nuisance parameters, it is often of interest to study the estimator under the following two scenarios: (1) n is fixed and $T \to \infty$; (2) T is fixed and $n \to \infty$. Under the first case, we have $T \gg n$ and therefore the estimation error in (17) is of the order $s_j \log(T) \log(Tp) T^{-1/3}$. Moreover, if $s_j = O(1)$ and ignoring the logarithmic factors, the rate reduces to $O_p(T^{-1/3})$, which agrees with the minimax optimal rate of the fused lasso estimator (ignoring the logarithmic factors) (Tibshirani, 2014). Thus, the estimation error in (17) is dominated by the fused lasso term $\frac{1}{\sqrt{nT}}\big\|\widehat{\boldsymbol{\Delta}}_j - \boldsymbol{\Delta}^*_j\big\|_2$. Given the results in Tibshirani (2014), the upper bound in (17) is non-improvable. Under the second case, we have $T \ll n$ and the upper bound in (17) reduces to $O\{s_j \log(np)\}$, which does not converge to zero, yielding an inconsistent estimator. The current setting corresponds to the classical Neyman and Scott's problem, in which the number of nuisance parameters $\boldsymbol{\Delta}_j$ increases too fast relative to the amount of data points nT. Moreover, we include a numerical study to compare the scenarios where $T \gg n$ and $T \ll n$ in Appendix K, whose conclusion is aligned with our discussion about (18) in this paragraph. Appendix I contains more detailed discussions of the restrictiveness of assuming that ϕ_0, σ_m, $\Delta_{\max}$, and τ are constants.

To conclude this section, we will show that our estimator is adaptive to the absence of unmeasured confounders. Recall that if we know a priori that there are no unmeasured

confounders, i.e., $\boldsymbol{\Delta}_j^* = \mathbf{0}$, we can estimate $(\boldsymbol{\theta}_{j,-j}, \boldsymbol{\alpha}_j)$ by the oracle lasso estimator with $\boldsymbol{\Delta}_j^* = \mathbf{0}$, which has an error bound in (18). The following proposition shows that if our approach is applied to the setting when there are no unmeasured confounders, the rate of convergence of our proposed estimator matches that of the oracle lasso estimator (18) up to logarithmic factors.

Proposition 11 *Assume that (16) does not contain any unmeasured confounders, i.e.,* $\boldsymbol{\Delta}_j^* = 0$. *Set the tuning parameters as*

$$\gamma = \mathcal{C}_1 \sigma_m \sqrt{\log(2pnT)\log(2np)/\log(T)} n^{-1} T^{-1/2}, \text{ and } \lambda = \beta = 2\log(T)\log(nTp) n^{-1/2} T^{-1/2},$$

in (11). When

$$\sqrt{2\mathcal{C}_8 s_j \log\{nT(2p-1)\}} \phi_0^{-\frac{1}{2}} (nT)^{-\frac{1}{4}} \leq \frac{1}{16},$$
$$n^{\frac{1}{2}} T^{\frac{1}{2}} \geq 2\max\left\{ c_1'' \log(T)\log(nTp) s_j, \mathcal{C}_6 \sigma_m^2 \log^{\frac{1}{2}}(2pnT) \right\},$$

$n, T, p \geq 6$, *and under Assumptions 1–7, we obtain*

$$\left\|\widehat{\boldsymbol{\theta}}_{j,-j} - \boldsymbol{\theta}_{j,-j}^*\right\|_1 + \left\|\widehat{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}_j^*\right\|_1 + \frac{1}{\sqrt{nT}} \left\|\widehat{\boldsymbol{\Delta}}_j - \boldsymbol{\Delta}_j^*\right\|_2 \lesssim s_j \log(T)\log(nTp)(nT)^{-\frac{1}{2}},$$

with probability at least $1 - 2\exp\{-\lfloor \log(2pnT)/\log(2pn) \rfloor\} - [n(T-1)\sqrt{\log\{n(T-1)\}}]^{-1} - 4/(nTp) - 2T^{-1/(2\mathcal{C}_7)\min\{\log(T)/\mathcal{C}_7, 1\}} - 2/(nT)^2$.

5. Numerical Studies

We conduct extensive numerical studies to evaluate the performance of our proposal on two different types of conditional independence graph: (i) Gaussian graphical models, and (ii) binary Ising models (Section H of the Appendix). For each model, we compare our proposed method to some existing methods on latent variable graphical models. To evaluate the performance across different methods, we define the true and false positive rates as the proportion of correctly estimated edges and the proportion of incorrectly estimated edges in the underlying graph, respectively.

For Gaussian graphical models, we compare our proposal with four different existing methods: the graphical lasso (Friedman et al., 2008); the neighborhood selection approach (Meinshausen and Bühlmann, 2006); the low-rank plus sparse latent variable Gaussian graphical model (Chandrasekaran et al., 2010); and latent variable graphical models with conditionally independent replicates (Tan et al., 2016). Friedman et al. (2008), Meinshausen and Bühlmann (2006), and Chandrasekaran et al. (2010) do not explicitly model the replicates: we therefore apply these methods by treating the replicates as independent samples. Moreover, our proposal, Meinshausen and Bühlmann (2006), and Tan et al. (2016) yield asymmetric estimates of the edge set. To obtain a symmetric edge set, we consider intersection and union rules in Meinshausen and Bühlmann (2006), and report the best results for the competing methods. We report our results using the intersection rule.

All of the aforementioned methods have a sparsity tuning parameter: we apply all methods using a fine grid of the sparsity tuning parameter values to obtain the curves shown in Figures 4–5. There is an additional tuning parameter for Chandrasekaran et al. (2010),

which models the confounding bias introduced by the unmeasured confounders. We set this tuning parameter to equal a constant multiplied by the sparsity tuning parameter, and we consider different values of constants and report the best results for Chandrasekaran et al. (2010). Our proposal has two additional tuning parameters which model the correlated data and the effect introduced by the unmeasured confounders. We detail the choice of tuning parameters for different settings on replicates and unmeasured confounders in the corresponding sections.

To assess the effects of correlated data and latent variables on graph estimation, we consider three different data generating mechanisms: (i) correlated replicates without latent variables; (ii) independent replicates with latent variables; and (iii) correlated replicates with latent variables. Out of the aforementioned approaches, our proposed method is the only method that models both correlated replicates and latent variables. Both Chandrasekaran et al. (2010) and Tan et al. (2016) model only the latent variables and do not take into account correlated replicates. We have also included additional numerical studies for sensitivity analysis in the Appendix: (i) Comparison between our proposed method to that of Tan et al. (2016) (assuming that location of piecewise constants are known) in Appendix J; (ii) Comparison between our proposed method to that of Zapata et al. (2022) that considers dependent data as functional data as described in Appendix L.

Recall that for Gaussian graphical models, the inverse covariance matrix encodes the conditional dependence relationships among the variables. Let $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$. We generate the inverse covariance matrix $\mathbf{\Theta}$ by randomly setting 10% of the off-diagonal elements in $\mathbf{\Theta}$ to equal 0.3, and setting the others to zero. To ensure the positive definiteness of $\mathbf{\Theta}$, we set $\Theta_{jj} = |\Lambda_{\min}(\mathbf{\Theta})| + 0.1$ for $j = 1, 2, \ldots, p$, where $\Lambda_{\min}(\mathbf{\Theta})$ is the minimum eigenvalue of $\mathbf{\Theta}$. Such a transformation ensures that $\mathbf{\Theta}$ is positive definite. We will use the aforementioned to generate $\mathbf{\Theta}$, unless otherwise is specified. For all of the numerical studies, we set $n = 50$, $T = 20$, and $p = 100$. The results, averaged over 100 independent data sets, are summarized in Figures 4–5.

5.1 Independent Replicates with Unmeasured Confounders

We now consider the case when there are unmeasured confounders with independent replicates. Let $\mathbf{U}_{it}$ be unmeasured confounders for the tth replicate for subject i. We consider two settings:

1. The unmeasured confounders are constant across replicates within each subject, that is $\mathbf{U}_{i1} = \mathbf{U}_{i2} = \cdots = \mathbf{U}_{iT}$. This simulation setting is considered in Tan et al. (2016).

2. The unmeasured confounders are piecewise constant. That is, we assume that $\mathbf{U}_{it} = \mathbf{U}_{it_1}$ when $t \leq \lfloor T/2 \rfloor$ and $\mathbf{U}_{it} = \mathbf{U}_{it_2}$ when $t > \lfloor T/2 \rfloor$, where $\lfloor T/2 \rfloor$ is the largest integer that is less than or equal to $T/2$, and $\mathbf{U}_{it_1} \neq \mathbf{U}_{it_2}$.

Similar to Tan et al. (2016), we generate the data by first partitioning $\mathbf{\Sigma}$ and $\mathbf{\Theta}$ into

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{XX} & \mathbf{\Sigma}_{XU} \\ \mathbf{\Sigma}_{UX} & \mathbf{\Sigma}_{UU} \end{pmatrix} \qquad \text{and} \qquad \mathbf{\Theta} = \begin{pmatrix} \mathbf{\Theta}_{XX} & \mathbf{\Theta}_{XU} \\ \mathbf{\Theta}_{UX} & \mathbf{\Theta}_{UU} \end{pmatrix},$$

where $\mathbf{\Theta}_{XX}$, $\mathbf{\Theta}_{XU}$, and $\mathbf{\Theta}_{UU}$ quantify the conditional independence relationships among the observed variables, between the observed variables and unmeasured confounders, and of

the unmeasured confounders, respectively. We set 10% of the off-diagonal entries in $\mathbf{\Theta}_{X,X}$ and 80% of the off-diagonal entries in $\mathbf{\Theta}_{X,U}$ and $\mathbf{\Theta}_{U,U}$ to equal 0.3. To ensure positive definiteness of $\mathbf{\Theta}$, we set $\Theta_{jj} = |\Lambda_{\min}(\mathbf{\Theta})| + 0.2$ for $j = 1, 2, \ldots, p$.

For the scenario in which the unmeasured confounders are constant across replicates within each subject, we first generate $\mathbf{U}_i \sim N_q(\mathbf{0}, \mathbf{\Sigma}_{UU})$. Then, we generate T replicates for each subject from a conditional normal distribution, i.e., $\mathbf{X}_{it} \mid \mathbf{U}_i \sim N_p(\mathbf{\Sigma}_{XU}\mathbf{\Sigma}_{UU}^{-1}\mathbf{U}_i, \mathbf{\Sigma}_{XX} - \mathbf{\Sigma}_{XU}\mathbf{\Sigma}_{UU}^{-1}\mathbf{\Sigma}_{UX})$. For the second scenario in which the unmeasured confounders are piecewise constant within each subject, we generate $\mathbf{U}_i^1, \mathbf{U}_i^2 \sim N_q(\mathbf{0}, \mathbf{\Sigma}_{UU})$. Similarly to the first setting, when $t \leq \lfloor T/2 \rfloor$, we generate the $\lfloor T/2 \rfloor$ replicates for each subject from the conditional distribution depend on $\mathbf{U}_i^1$, then generate the rest replicates according to $\mathbf{U}_i^2$. Besides, let $q = 5$, which means that we have 5 unmeasured confounders in total. Recall that our proposal has two additional tuning parameters: we set $\beta \in \{0.05, 0.1, 0.15\}$ and $\gamma \in \{1, 1.5, 2\}$. Due to the high degree of overlap among the 9 lines representing our proposal, making them difficult to distinguish, we have chosen to display only three lines in Figure 3, corresponding to $\gamma = 1$ with $\beta \in \{1, 1.5, 2\}$. The complete set of results is provided in Appendix M.

From Figure 3, we see that methods that account for unmeasured confounders outperform methods that do not model the unmeasured confounders. Specifically, Tan et al. (2016) has the best performance in the case of independent replicates and constant unmeasured confounders in Figure 3(a). This is not surprising since Tan et al. (2016) is explicitly designed to model such a setting. Our proposal reduces to that of Tan et al. (2016) as $\gamma, \beta \to \infty$. Thus, our proposal has very similar performance to that of Tan et al. (2016). However, when the unmeasured confounders are piecewise constant, our proposed method is much better than that of Tan et al. (2016) and is comparable to that of Chandrasekaran et al. (2010).

5.2 Correlated Replicates without Unmeasured Counfounders

In this section, we evaluate the effect of correlated replicates on graph estimation. We assume that the replicates within each subject are correlated under an AR(1) process, i.e., we assume that

$$\mathbf{X}_{i1} \sim N_p(\mathbf{0}, \mathbf{\Sigma}), \qquad \mathbf{X}_{it} \mid \mathbf{X}_{i(t-1)} \sim N_p(\mathbf{A}\mathbf{X}_{i(t-1)}, \mathbf{\Sigma}), \quad \text{for } t = 2, \ldots, T, \tag{19}$$

where $\mathbf{A}$ is a transition matrix that quantifies the correlation between $\mathbf{X}_{it}$ and $\mathbf{X}_{i(t-1)}$. We consider two different types of transition matrix:

1. Diagonal transition matrix $\mathbf{A}$ with $A_{jj} = 0.9$ for $j = 1, \ldots, p$. In other words, each variable at the tth replicate is conditionally dependent only with itself for the $(t-1)$th replicate.

2. Sparse transition matrix $\mathbf{A}$ with 5% elements of $\mathbf{A}$ set to equal 0.3. That is, the jth variable at the tth replicate is conditionally dependent with other variables at the $(t-1)$th replicate.

We generate the data according to (19). Similar to Section 5.1, we vary tuning parameter $\beta \in \{0.05, 0.1, 0.15\}$ and $\gamma \in \{1, 1.5, 2\}$ to assess the performance of our proposal relative to

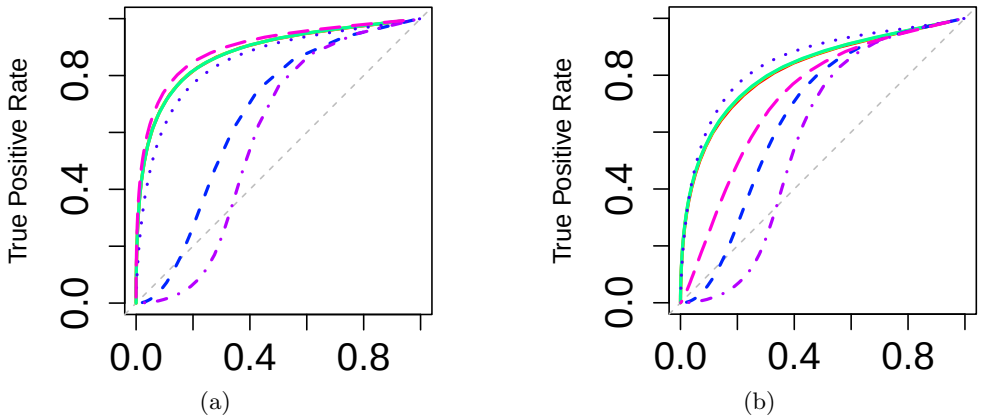


Figure 3: Results for independent replicates with unmeasured confounders in Section 5.1. Panels (a) and (b) correspond to the results for constant and piecewise constant unmeasured confounders, respectively. For our proposal, we set $\gamma = 1$. We consider three different values of β: $\beta = 0.05$ (red solid), $\beta = 0.1$ (yellow-green solid), and $\beta = 0.15$ (green solid). The other curves represent Friedman et al. (2008) (purple dot-dashed); Meinshausen and Bühlmann (2006) (blue short-dashed); Chandrasekaran et al. (2010) (indigo dots); and Tan et al. (2016) (pink long-dashed).

existing methods. The complete set of results is provided in Appendix M, and we only show the results with $\gamma = 1$ and $\beta \in \{0.05, 0.1, 0.15\}$ in Figure 4. From Figure 4, we see that our proposed method using different set of tuning parameters dominate all of the competing methods that assume independent replicates. The results illustrate that not modeling the correlation among the replicates can have a significant impact on the estimated graph structure. This is especially apparent in Figure 4(b) when the correlation between two replicates is modeled using a sparse transition matrix.

5.3 Correlated Replicates with Unmeasured Confounders

In this section, we allow replicates within each subject to be correlated, and that there are unmeasured confounders. Throughout the numerical studies in this section, we assume that the correlated replicates are modeled according to the sparse transition matrix $\mathbf{A}$ as described in Section 5.2. We consider constant and piecewise constant unmeasured confounders as described in Section 5.1. Specifically, we assume the model

$$\begin{aligned} \mathbf{X}_{i1} \mid \mathbf{U}_{i1} &\sim N_p(\mathbf{\Sigma}_{OH}\mathbf{\Sigma}_{HH}^{-1}\mathbf{U}_{i1}, \mathbf{\Sigma}_{XX} - \mathbf{\Sigma}_{XU}\mathbf{\Sigma}_{UU}^{-1}\mathbf{\Sigma}_{UX}), \\ \mathbf{X}_{it} \mid \mathbf{X}_{i(t-1)}, \mathbf{U}_{it} &\sim N_p(\mathbf{A}\mathbf{X}_{i(t-1)} + \mathbf{\Sigma}_{OH}\mathbf{\Sigma}_{HH}^{-1}\mathbf{U}_{it}, \mathbf{\Sigma}_{XX} - \mathbf{\Sigma}_{XU}\mathbf{\Sigma}_{UU}^{-1}\mathbf{\Sigma}_{UX}). \end{aligned} \tag{20}$$

We generate the data according to (20) using the same data generating mechanisms as described in Sections 5.1–5.2. For the two additional tuning parameters in our proposal, we vary $\beta \in \{0.01, 0.02, 0.03\}$ and $\gamma \in \{1, 1.5, 2\}$. We show the complete set of results in Appendix M, and only show the results with $\gamma = 1$ and $\beta \in \{0.01, 0.02, 0.03\}$ in Figure 5.

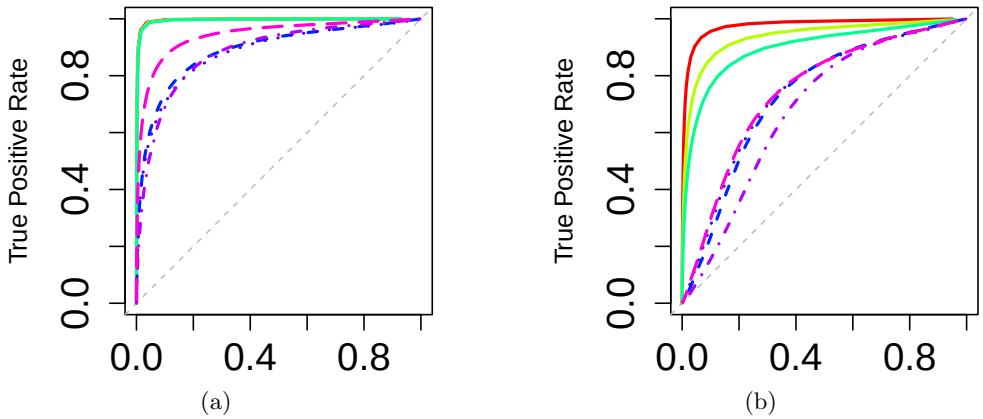


Figure 4: Results for correlated replicates without unmeasured confounders. Panels (a) and (b) correspond to diagonal and sparse transition matrices, respectively. Other details are as in Figure 3.

We can see from both Figures 5(a)–(b) that our proposal outperforms all existing methods when there are correlated replicates and unmeasured confounders. In fact, all existing methods have area under the curves of approximately 0.5. From Figure 5(a), we see that even when the unmeasured confounders are constant, Tan et al. (2016) can no longer estimate the graph accurately since the conditional independent replicates assumption is violated. We see that not modeling either the correlated replicates or unmeasured confounders can lead to biased estimation of the graph.

6. Data Application

In this section, we apply the proposed method to construct a brain connectivity network for ADHD-200 data (Biswal et al., 2010). This dataset consists of resting state brain images and the phenotypic information of the subjects, such as age, gender, and intelligence quotient. After removing missing data from the original data set, we have 465 subjects, and each subject has between 76 and 276 images. We select 150 independent subjects from the groups of children and adolescent, respectively. Moreover, for computational convenience, we select 10 consecutive images from each subject as replicates. Similar to Power et al. (2011), we consider 264 brain regions of interest as the variables of interest.

Although the data set consists of several phenotypic variables, there may also be some unmeasured phenotypic variables that can potentially serve as confounders. Ignoring the unmeasured confounders or the observed phenotypic variables and directly fitting a Gaussian graphical model using Meinshausen and Bühlmann (2006) may lead to a biased conditional independence graph. Besides, the replicates from the same subject may be correlated and ignoring the correlation among replicates may also lead to a biased estimation of graph. To assess whether the unmeasured confounders and correlated replicates have effects on the conditional dependency structure of the observed variables, we implement four methods with different assumptions on the replicates and unmeasured confounders: (1) correlated

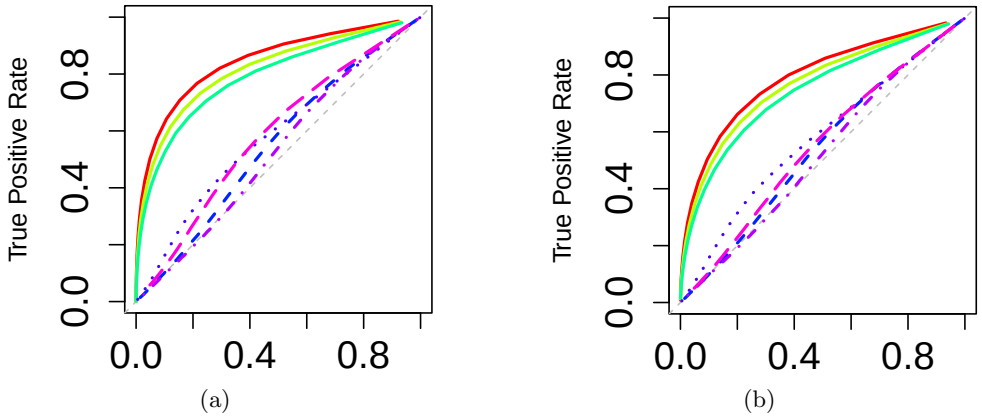


Figure 5: Results for constant and piecewise constant unmeasured confounders with sparse transition matrix $\mathbf{A}$ in Section 5.3. Panels (a) and (b) correspond to constant and piecewise constant unmeasured confounders, respectively. For our proposal, we set $\gamma = 1$. We consider three different values of β: $\beta = 0.01$ (red solid), $\beta = 0.02$ (yellow-green solid), and $\beta = 0.03$ (green solid).

Other details are as in Figure 3.

replicates with piecewise constant unmeasured confounders (our proposal); (2) independent replicates with piecewise constant unmeasured confounders (our proposal by setting β to be arbitrary large); (3) correlated replicates with constant unmeasured confounders (our proposal by setting γ to be arbitrary large); (4) independent replicates without unmeasured confounders (Meinshausen and Bühlmann, 2006).

Our proposed method involves three tuning parameters, i.e., λ, β, and γ. As suggested by both Theorems 8 and 9, we set $\lambda = \beta$, reducing the tuning parameters from three to two. We consider a fine-grid of tuning parameters that yields the number of estimated edges in the range of $100 - 200$. Then use the stability metrics described in Section 3.2 to select the tuning parameters. The selected tuning parameters for the children subsets of data are $\lambda = \beta = 0.23$ and $\gamma = 0.5$, yielding a total number of 164 edges. On the other hand, the tuning parameters for the adolescent subsets of data are $\lambda = \beta = 0.1$ and $\gamma = 0.2$, which yields 189 edges.

We also implement the proposed method with either β or γ set arbitrary large, and use the aforementioned stability selection procedure to choose the remaining tuning parameters. For model that ignores the correlation among replicates, the selected tuning parameters for the children and adolescent subsets are $\lambda = 0.23, \gamma = 0.5$ and $\lambda = 0.05, \gamma = 0.1$, respectively. For model that ignores the unmeasured confounders, the selected tuning parameters for the children and adolescent subsets are $\lambda = \beta = 0.34$ and $\lambda = \beta = 0.32$, respectively. Finally, we implement the proposal of Meinshausen and Bühlmann (2006) with the sparsity tuning parameter selected using stability selection as described in Lim and Yu (2016). The selected tuning parameter for the children and adolescent subsets are $\lambda = 0.46$ and $\lambda = 0.41$, respectively.

To evaluate the results across the four different methods, we treat our proposed method as the baseline approach and compare the estimated graph across the different methods. Let $\widehat{E}_1, \ldots, \widehat{E}_4$ be the set of edges obtained from methods (1)–(4). For $k = \{2, 3, 4\}$, we define $\text{Eint}_k = \widehat{E}_1 \cap \widehat{E}_k$ as a set of edges that are present in both $\widehat{E}_1$ and $\widehat{E}_k$; $\text{Eext}_k = \widehat{E}_1^c \cap \widehat{E}_k$ as a set of edges that are not present in $\widehat{E}_1$ but present in $\widehat{E}_k$; $\text{Elack}_k = \widehat{E}_1 \cap \widehat{E}_k^c$ as a set of edges that are present in $\widehat{E}_1$ but not in $\widehat{E}_k$. The results are summarized in Table 1.

Table 1: Number of edges in $\widehat{E}_k$, Eint_k, Eext_k, and Elack_k for the children and adolescent groups.

Nunmber of edges	$\widehat{E}_k$	Eint_k	Eext_k	Elack_k
our proposal, children	164			
our proposal, adolescent	189			
our proposal with $\beta = 10^6$, children	164	164	0	0
our proposal with $\beta = 10^6$, adolescent	203	189	14	0
our proposal with $\gamma = 10^6$, children	244	164	80	0
our proposal with $\gamma = 10^6$, adolescent	247	182	65	7
Meinshausen and Bühlmann (2006), children	155	143	12	21
Meinshausen and Bühlmann (2006), adolescent	188	169	19	20

From Table 1, we see that the number of edges for adolescent is larger than that of the children group across all methods. Using the estimated graph for our proposed method as the baseline, we see that ignoring unmeasured confounders (our proposal with $\gamma = 10^6$) can lead to very different estimated graphs with high Eext_k, i.e., a high number of additional edges are estimated that are not present in $\widehat{E}_1$. Moreover, the number of edges in the set Elack_k is the largest for Meinshausen and Bühlmann (2006), indicating that ignoring both the unmeasured confounders and correlated replicates can be detrimental.

To further compare the Meinshausen and Bühlmann (2006) with our proposed method, we plot the difference between the estimated graphs in Figure 6. The estimated graphs between the two methods are quite different for both children and adolescents. In Figure 6, most edges in Elack_4 are located within the occipital lobe, the brain's visual processing region associated with sight, image recognition, and perception. Previous studies suggest that individuals with ADHD may exhibit hyperactivation in the occipital lobe (Wang et al., 2007), and abnormal activation or maturation in this region may contribute to visuospatial difficulties and inattention commonly observed in ADHD (Chang et al., 2020). Therefore, we anticipate a greater number of edges in the occipital lobe, consistent with our method's correct identification of more edges within this area. The edges in Eext_4 for children are distributed across various brain regions, whereas for adolescents, these edges are more concentrated in Wernicke's area, a region linked to the comprehension of written and spoken language. However, as demonstrated by Lee et al. (2017), individuals with ADHD often exhibit reduced connectivity in brain areas related to language processing. This aligns well with our proposed method, which accurately estimates fewer edges in Wernicke's area. Our

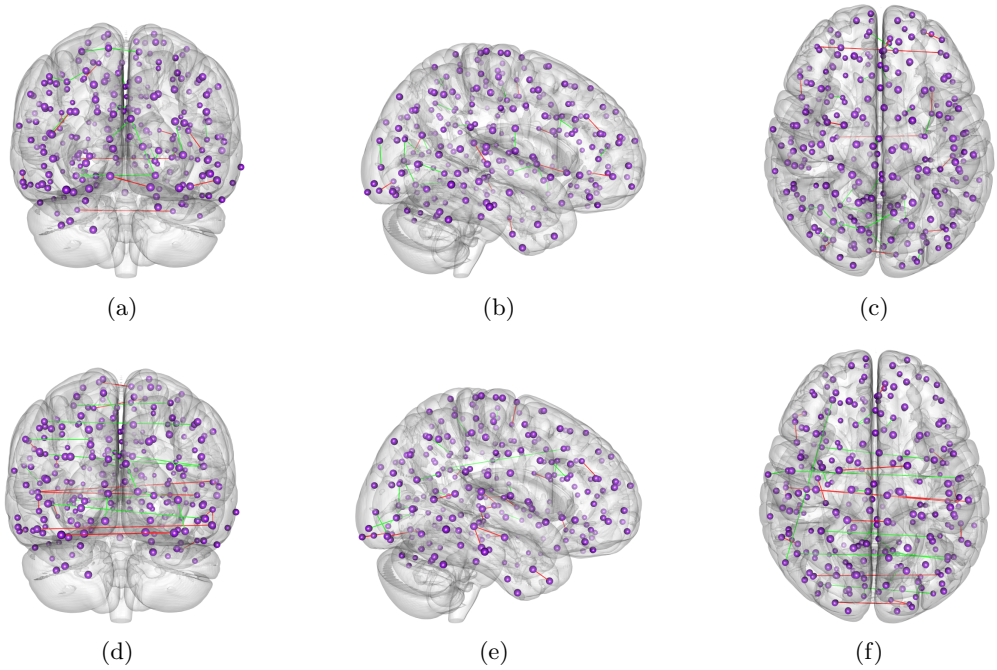


Figure 6: Coronal, sagittal, and transverse snapshots of the difference between our proposal and that of Meinshausen and Bühlmann (2006). Panels (a)–(c) are coronal, sagittal, and transverse snapshots for children and panels (d)–(f) are coronal, sagittal, and transverse snapshots for adolescent. The red and green lines represent the edges in Eext_4 and Elack_4, respectively.

results suggest that the potential bias introduced by the correlation across replicates and unmeasured confounders can be large, and care must be taken when estimating a conditional dependence graph. Finally, we acknowledge that Gaussian graphical models may be inadequate for modeling the ADHD-200 data due to potential heavy-tailed noise and outliers, and a more sophisticated techniques such as that of Lee and Xue (2018) can be helpful.

7. Discussion

In some applications, one may be only interested in a subset of parameters such as $\boldsymbol{\theta}_{j,-j}$. By carefully going through our analysis, we realize that the three estimation errors $\|\widehat{\boldsymbol{\theta}}_{j,-j} - \boldsymbol{\theta}^*_{j,-j}\|_1$, $\|\widehat{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}^*_j\|_1$ and $\|\widehat{\boldsymbol{\Delta}}_j - \boldsymbol{\Delta}^*_j\|_2$ are entangled together in the proof and cannot be separated. To the best of our knowledge, we are not aware of any analysis of lasso type estimator that gives sharp convergence rate for a subset of high dimensional parameters.

While our analysis cannot provide a sharper rate for a subset of parameters such as $\boldsymbol{\theta}_{j,-j}$, it is possible to achieve this goal by using a different approach via thresholding debiased/decorrelated estimators (Van de Geer et al., 2014; Zhang and Zhang, 2014). To be specific, for each component of $\boldsymbol{\theta}_{j,-j}$, say $(\boldsymbol{\theta}_{j,-j})_\ell$, we can construct the debiased/decorrelated

estimators $(\widetilde{\boldsymbol{\theta}}_{j,-j})_\ell$ by treating the rest of parameters $(\boldsymbol{\theta}_{j,-j})_{-\ell}$ and $\boldsymbol{\alpha}_j$ and $\boldsymbol{\Delta}_j$ as the nuisance parameter. Under some regularity conditions, it can be shown that $(\widetilde{\boldsymbol{\theta}}_{j,-j})_\ell$ is asymptotically normal with mean $(\boldsymbol{\theta}^*_{j,-j})_\ell$, which can provide valid inference for $(\boldsymbol{\theta}^*_{j,-j})_\ell$. To obtain a sharper rate for the subvector $\boldsymbol{\theta}_{j,-j}$, we can further leverage its sparsity assumption by thresholding the debiased/decorrelated estimators $|(\widetilde{\boldsymbol{\theta}}_{j,-j})_\ell|$ at some appropriate level, e.g., $C\sqrt{\log(p)/(nT)}$ with some constant $C > 0$, for $1 \le \ell \le p-1$. We expect that the resulting estimator of $\boldsymbol{\theta}_{j,-j}$ can be shown to attain a faster convergence rate in L_1 norm. The detailed analysis is beyond the scope of this work. We leave it for future investigations.

Appendix A. Examples of Exponential Family Graphical Models

In this section, we provide three examples of commonly used exponential family graphical models, which are given as follows:

Example 4 *The Gaussian graphical model with mean zero has the joint density function*

$$p(\boldsymbol{x}) = \exp\left[-\frac{1}{2}\sum_{j=1}^{p}\zeta_j x_j^2 + \frac{1}{2}\sum_{j=1}^{p}\sum_{k\neq j}\theta_{jk}x_j x_k + \log\left\{\frac{|\widetilde{\boldsymbol{\Theta}}|^{\frac{1}{2}}}{(2\pi)^{\frac{p}{2}}}\right\}\right], \quad \boldsymbol{x}\in\mathbb{R}^p, \tag{21}$$

where $f_j(x_j;\boldsymbol{\zeta}) = -\zeta_j x_j^2/2$ *and* $A(\boldsymbol{\Theta},\boldsymbol{\zeta}) = -\log\{|\widetilde{\boldsymbol{\Theta}}|^{1/2}/(2\pi)^{p/2}\}$. *For the matrix* $\widetilde{\boldsymbol{\Theta}}$, *the diagonal element* $\widetilde{\theta}_{jj} = \zeta_j$, *and the off-diagonal element* $\widetilde{\theta}_{jk} = -\theta_{jk}$ *for* $j \neq k$. *Besides,* $\widetilde{\boldsymbol{\Theta}}$ *should be positive definite for a valid distribution.*

Example 5 *The Ising graphical model can be written as:*

$$p(\boldsymbol{x}) = \exp\left\{\frac{1}{2}\sum_{j=1}^{p}\sum_{k\neq j}\theta_{jk}x_j x_k - A(\boldsymbol{\Theta};\boldsymbol{\zeta})\right\}, \quad \boldsymbol{x}\in\{0,1\}^p, \tag{22}$$

where $f_j(x_j;\boldsymbol{\zeta}) = 0$ *and* $\theta_{jk}\in\mathbb{R}$.

Example 6 *The Poisson graphical model can be written as:*

$$p(\boldsymbol{x}) = \exp\left[\sum_{j=1}^{p}\{\zeta_j x_j - \log(x_j!)\} + \frac{1}{2}\sum_{j=1}^{p}\sum_{k\neq j}\theta_{jk}x_j x_k - A(\boldsymbol{\Theta};\boldsymbol{\zeta})\right] \quad \boldsymbol{x}\in\{0,1,2,\ldots\}^p, \tag{23}$$

where $f_j(x_j;\boldsymbol{\zeta}) = \zeta_j x_j - \log(x_j!)$. *To ensure that the log-partition function is finite, i.e.,* $A(\boldsymbol{\Theta},\boldsymbol{\zeta}) < \infty$, θ_{jk} *is constrained to be negative. In other words, Poisson graphical models capture negative conditional relationships between pairs of nodes. We refer the reader to Yang et al. (2013) for variants of Poisson graphical models that are able to capture a richer conditional dependence relationships between pairs of nodes.*

Appendix B. Extension of One-lag Autoregressive Model

The one-lag autoregressive model for the replicates in Assumption 1 can be easily relaxed to the o-lag vector autoregressive model with $1 < o \le T-1$. Specifically, given the unmeasured

confounders, the joint conditional density of the T replicates can be written as

$$p\left(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_T \mid \boldsymbol{u}_1,\ldots,\boldsymbol{u}_T\right) = p\left(\boldsymbol{x}_1 \mid \boldsymbol{u}_1\right)\cdot\prod_{t=2}^{o} p\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1},\ldots,\boldsymbol{x}_1,\boldsymbol{u}_t\right)\cdot\prod_{t=o+1}^{T} p\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1},\ldots,\boldsymbol{x}_{t-o},\boldsymbol{u}_t\right).$$

Thus, for each node j, the conditional density of X_{tj} given $\boldsymbol{X}_{t(-j)}$, $\boldsymbol{U}_t$, $\boldsymbol{X}_{t-1}$, …, $\boldsymbol{X}_{t-o}$ is

$$\begin{aligned} &p\left(x_{tj} \mid \boldsymbol{x}_{t(-j)},\boldsymbol{u}_t,\boldsymbol{x}_{t-1},\ldots,\boldsymbol{x}_{t-o}\right) \\ &= \exp\left\{ f_{tj}\left(x_{tj}\right) + \sum_{k\neq j}\theta_{jk}x_{tk}x_{tj} + \sum_{k=1}^{p}\sum_{r=1}^{o}\alpha_{jh}x_{(t-r)k}x_{tj} + \sum_{m=1}^{q}\delta_{jm}u_{tm}x_{tj} - D_{tj}\left(\theta_{jk},\alpha_{jh},\delta_{jm},f_{tj}\right)\right\}, \end{aligned} \tag{24}$$

where $h=(k-1)o+r$ and $x_{(t-r)k}=0$ when $t-r\leq 0$.

Appendix C. Discussion about Extension to Mixed Graphical Models

The exponential family graphical models enforce all variables to belong to the same distribution. However, such an assumption may be violated in many scientific applications. For instance, in genomics data, variables from RNA-sequencing are count-valued and variables from SNP-arrays are binary (Yang et al., 2014). To allow for mixed data types, the mixed graphical models were proposed (Cheng et al., 2017; She et al., 2019; Chen et al., 2015; Yang et al., 2014; Lee and Hastie, 2015). Using similar ideas in She et al. (2019) or Chen et al. (2015), our proposed method can also be extended to the context of mixed graphical models, and we leave this for future work.

Appendix D. Notation

In this section, we compile notation that are used throughout the main manuscript. We start with notation in the proposed optimization problem. Let $\boldsymbol{\alpha}_j = (\alpha_{j1},\alpha_{j2},\ldots,\alpha_{jp})^{\mathrm{T}} \in \mathbb{R}^p$, $\boldsymbol{\theta}_{j,-j} = (\theta_{j1},\ldots,\theta_{j,j-1},\theta_{j,j+1},\ldots,\theta_{jp})^{\mathrm{T}} \in \mathbb{R}^{p-1}$, and $\boldsymbol{\Delta}_j = (\Delta_{11j},\Delta_{12j},\ldots,\Delta_{1Tj},\Delta_{21j},\Delta_{22j},\ldots,\Delta_{nTj})^{\mathrm{T}} \in \mathbb{R}^{nT}$. For notational convenience, let $\boldsymbol{\Theta}_j = (\boldsymbol{\theta}_{j,-j}^{\mathrm{T}},\boldsymbol{\alpha}_j^{\mathrm{T}},\mathbf{H}_j^{\mathrm{T}})^{\mathrm{T}}$. Let $\mathbf{I}_n$ be an n-dimensional identity matrix, and define $\mathbf{C}\in\mathbb{R}^{(T-1)\times T}$ as

$$\mathbf{C} = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix}.$$

We now present notation that are commonly used in Section 2 of the main manuscript. Let $\boldsymbol{X}=(X_1,\ldots,X_p)^{\mathrm{T}}$ be the observed variables. Moreover, let $\boldsymbol{X}_{-j}=(X_1,\ldots,X_{j-1},X_{j+1},\ldots,X_p)^{\mathrm{T}} \in \mathbb{R}^{p-1}$, $\boldsymbol{X}_t=(X_{t1},\ldots,X_{tp})\in\mathbb{R}^p$, $\boldsymbol{X}_{t(-j)}=(X_t1,\ldots,X_{t(j-1)},X_{t(j+1)},\ldots,X_{tp})^{\mathrm{T}}\in\mathbb{R}^{p-1}$. Denote $\boldsymbol{U}_t=(U_{t1},\ldots,U_{tq})\in\mathbb{R}^q$ as the unmeasured variables.

We now proceed to notation used in Section 3 of the main manuscript. We define the following notation for convenience:

$$\begin{aligned}
\mathbf{x}_j &= (x_{11j}, x_{12j}, \ldots, x_{1Tj}, x_{21j}, x_{22j}, \ldots, x_{nTj})^{\mathrm{T}} \in \mathbb{R}^{nT};\\
\mathbf{X}_{i(-j)} &= (\mathbf{x}_{i1(-j)}, \mathbf{x}_{i2(-j)}, \ldots, \mathbf{x}_{iT(-j)})^{\mathrm{T}} \in \mathbb{R}^{T\times(p-1)};\\
\mathbf{X}_{i(T-1)} &= (\mathbf{x}_{i0}, \mathbf{x}_{i1}, \ldots, \mathbf{x}_{i(T-1)})^{\mathrm{T}} \in \mathbb{R}^{T\times p};\\
\mathbf{X}^{\otimes}_{-j} &= (\mathbf{X}^{\mathrm{T}}_{1(-j)}, \ldots, \mathbf{X}^{\mathrm{T}}_{n(-j)})^{\mathrm{T}} \in \mathbb{R}^{(nT)\times(p-1)};\\
\mathbf{X}^{\otimes}_{T-1} &= (\mathbf{X}^{\mathrm{T}}_{1(T-1)}, \ldots, \mathbf{X}^{\mathrm{T}}_{n(T-1)})^{\mathrm{T}} \in \mathbb{R}^{(nT)\times p}.
\end{aligned}$$

Moreover, let $\widetilde{\mathbf{x}}_j = (\mathbf{X}^{\otimes}_{-j}, \mathbf{X}^{\otimes}_{T-1}, \widetilde{\mathbf{C}}^{-1})$. Let $\mathbf{E} = \mathbf{I}_n \otimes \mathbf{1}^{\mathrm{T}}_T \in \mathbb{R}^{n\times(nT)}$, where $\mathbf{1}_n$ is an n-dimensional vector of ones. Additionally, we have $\widetilde{\mathbf{C}} = ((\mathbf{I}_n \otimes \mathbf{C})^{\mathrm{T}}, \mathbf{E}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{(nT)\times(nT)}$, $\mathbf{H}_j = \widetilde{\mathbf{C}}\boldsymbol{\Delta}_j = [\{(\mathbf{I}_n \otimes \widetilde{\mathbf{C}}^{-1}\mathbf{C})\boldsymbol{\Delta}_j\}^{\mathrm{T}}, (\mathbf{E}\boldsymbol{\Delta}_j)^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{nT}$, and $\mathbf{H}_{j1} = (\mathbf{I}_n \otimes \mathbf{C})\boldsymbol{\Delta}_j \in \mathbb{R}^{n(T-1)}$.

Finally, we present notation that are used in presenting the theoretical results in Section 4 in the context of Gaussian graphical models. Let $(\sigma^{\epsilon}_m)^2 = \max_{t,j}\{(\sigma^{\epsilon}_{jj,t})^2\}$, $\Delta_m = \max_{i,t,j} |\Delta^*_{itj} - \Delta^*_{i(t-1)j}|$, $\tau = \max_{i,j}\sum_{t=2}^{T} I(\Delta^*_{itj} \neq \Delta^*_{i(t-1)j})$ and $\Delta_{\max} = \Delta_m + 1$. Let $\boldsymbol{X}_{ij} = (X_{i1j}, \ldots, X_{iTj})^{\mathrm{T}} \sim N(\boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_{jj})$, $|\mu_{itj}| \leq \mu_{\max}$, and $\max_{1\leq j\leq p} \left\|\boldsymbol{\Sigma}_{jj}\right\|_{\mathrm{op}} \leq \kappa$, where $\|\boldsymbol{\Sigma}_{jj}\|_{\mathrm{op}}$ is the operator norm of $\boldsymbol{\Sigma}_{jj}$. Let $\boldsymbol{\omega}^*_j = \{(\boldsymbol{\theta}^*_{j,-j})^{\mathrm{T}}, (\boldsymbol{\alpha}^*_j)^{\mathrm{T}}\}^{\mathrm{T}}$, $\mathcal{S}_j = \{k : \omega^*_{jk} \neq 0\}$ be the active set and $s_j = |\mathcal{S}_j|$. let $\boldsymbol{\omega}_j^{\mathcal{S}_j}$ and $\boldsymbol{\omega}_j^{\mathcal{S}_j^c}$ be subvectors of $\boldsymbol{\omega}_j$ with indices $\mathcal{S}_j$ and $\mathcal{S}_j^c$, respectively. Let $\mathbf{Y}_{itj} = (\mathbf{X}_{it(-j)}, \mathbf{X}_{i(t-1)}) \in \mathbb{R}^{1\times(2p-1)}$ and $\mathbf{Y}_j = (\mathbf{Y}^{\mathrm{T}}_{11j}, \mathbf{Y}^{\mathrm{T}}_{12j}, \ldots, \mathbf{Y}^{\mathrm{T}}_{1Tj}, \mathbf{Y}^{\mathrm{T}}_{21j}, \ldots, \mathbf{Y}^{\mathrm{T}}_{nTj})^{\mathrm{T}} \in \mathbb{R}^{nT\times(2p-1)}$. Let $\boldsymbol{\Sigma}_{jj'} = \mathrm{Cov}(\boldsymbol{X}_{ij}, \boldsymbol{X}_{ij'})$ with $\boldsymbol{X}_{ij} = (X_{i1j}, \ldots, X_{iTj})^{\mathrm{T}}$, and $\widetilde{\boldsymbol{\Sigma}}_{jj'} = \mathrm{Cov}(\boldsymbol{X}_{ij}, \breve{\boldsymbol{X}}_{ij'})$ with $\breve{\boldsymbol{X}}_{ij'} = (X_{i0j}, \ldots, X_{i(T-1)j})^{\mathrm{T}}$. We define κ' as $\max\{\max_{j,j'} \|\boldsymbol{\Sigma}_{jj'}\|_{\mathrm{op}}, \max_{j,j'} \|\boldsymbol{\Sigma}_{jj'}\|_{\mathrm{op}}\} \leq \kappa'$. Let $\mathbf{U}_j = \mathbb{E}(\mathbf{Y}_j)$ and $\boldsymbol{\Sigma}_j = \mathbb{E}\{(\mathbf{Y}_j - \mathbf{U}_j)^{\mathrm{T}}(\mathbf{Y}_j - \mathbf{U}_j)\}/(nT)$. Let ϕ_0 is a positive constant and the upper bound of the smallest eigenvalue of $\boldsymbol{\Sigma}_j$. Let $\sigma_m = \max\{\sqrt{2\kappa}, \sqrt{2}\sigma^{\epsilon}_m, 1\}$ where κ is as defined in Assumption 6. The notation $a_n \lesssim b_n$ stands for $a_n = Cb_n$ with C is a constant and $a_n \gtrsim b_n$ is defined similarly.

Appendix E. Theoretical Results for Exponential Family Graphical Models

In this section, we obtain non-asymptotic upper bounds for $\|\widehat{\boldsymbol{\theta}}_{j,-j} - \boldsymbol{\theta}^*_{j,-j}\|_1 + \|\widehat{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}^*_j\|_1 + \|\widehat{\boldsymbol{\Delta}}_j - \boldsymbol{\Delta}^*_j\|_2/\sqrt{nT}$ in the context of exponential family graphical models. Based on the conditional density (8), the joint density function for the exponential family graphical models can be written as:

$$p\left(\widetilde{\boldsymbol{x}}\right) = \exp\left\{ f\left(\widetilde{\boldsymbol{x}}\right) + \sum_{t=1}^{T}\sum_{j=1}^{p}\sum_{k\neq j}\theta_{jk}x_{tj}x_{tk} + \sum_{t=1}^{T}\sum_{j=1}^{p}\sum_{k=1}^{p}\alpha_{jk}\boldsymbol{x}_{(t-1)k}x_{tj} + \sum_{t=1}^{T}\sum_{j=1}^{p}\Delta_{tj}x_{tj} - A\left(\widetilde{\boldsymbol{\theta}}\right)\right\} \tag{25}$$

where $\widetilde{\boldsymbol{x}} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T] \in \mathbb{R}^{Tp}$ and $\widetilde{\boldsymbol{\theta}} = (\boldsymbol{\theta}_{1,-1}^{\mathrm{T}}, \ldots, \boldsymbol{\theta}_{p,-p}^{\mathrm{T}}, \boldsymbol{\alpha}_1^{\mathrm{T}}, \ldots, \boldsymbol{\alpha}_p^{\mathrm{T}}, \boldsymbol{\Delta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\Delta}_p^{\mathrm{T}})^{\mathrm{T}}$. The log-partition function $A(\widetilde{\boldsymbol{\theta}})$ is

$$A\left(\widetilde{\boldsymbol{\theta}}\right) := \log \int_{\widetilde{\boldsymbol{x}}} \exp\left\{ f\left(\widetilde{\boldsymbol{x}}\right) + \sum_{t=1}^{T}\sum_{j=1}^{p}\sum_{k\neq j} \theta_{jk} x_{tj} x_{tk} + \sum_{t=1}^{T}\sum_{j=1}^{p}\sum_{k=1}^{p} \alpha_{jk} \boldsymbol{x}_{(t-1)k} x_{tj} + \sum_{t=1}^{T}\sum_{j=1}^{p} \Delta_{tj} x_{tj} \right\} d\widetilde{\boldsymbol{x}}.$$

Assume that we have n subjects. Similar to Yang et al. (2015), we impose Assumptions 12–14 hold, which are standard conditions for exponential family graphical models about the log-partition function for both the joint and conditional distributions.

Assumption 12 *For notational simplicity, we write $D(\eta_{itj}^*) = D_{tj}\left(\theta_{jk}, \alpha_{jk}, \Delta_{tj}, f_{tj}\right)$, where $D_{tj}\left(\theta_{jk}, \alpha_{jk}, \Delta_{tj}, f_{tj}\right)$ is the log-partition function in (8). We assume $D^{''}(\cdot) \geq c_d > 0$ and $\max_{i,t,j} D^{''}(\eta_{itj}^*) \leq c_u$, where $\eta_{itj}^* = \sum_{k\neq j} \theta_{jk}^* x_{itk} + \sum_{k=1}^{p} \alpha_{jk}^* x_{i(t-1)k} + \Delta_{itj}^*$.*

Assumption 13 *Assume that $\max_{t,j} \mathbb{E}(X_{tj}^2) \leq c_v$.*

Assumption 14 *Let*

$$\widetilde{A}_{tj}\left(u; \widetilde{\theta}\right)$$
$$= \log \int_{\widetilde{\boldsymbol{x}}} \exp\left\{ u x_{tj}^2 + f\left(\widetilde{\boldsymbol{x}}\right) + \sum_{t=1}^{T}\sum_{j=1}^{p}\sum_{k\neq j} \theta_{jk} x_{tj} x_{tk} + \sum_{t=1}^{T}\sum_{j=1}^{p}\sum_{k=1}^{p} \alpha_{jk} \boldsymbol{x}_{(t-1)k} x_{tj} + \sum_{t=1}^{T}\sum_{j=1}^{p} \Delta_{tj} x_{tj} \right\} d\widetilde{\boldsymbol{x}}.$$

Assume that for $j = 1, \ldots, p$ and $t = 1, \ldots, T$,

$$\max_{u:|u|\leq 1} \frac{\partial^2}{\partial u^2} \widetilde{A}_{tj}\left(u; \widetilde{\theta}\right) \leq c_h.$$

Let $\epsilon_{itj}^{\triangledown} = X_{itj} - D^{'}(\eta_{itj}^*)$, $\boldsymbol{\epsilon}_{ij}^{\triangledown} = (\epsilon_{i1j}^{\triangledown}, \ldots, \epsilon_{itj}^{\triangledown})^{\mathrm{T}}$, and $\boldsymbol{\epsilon}_{ij}^{X} = \boldsymbol{X}_{ij} - \boldsymbol{\mu}_{ij}$. Since $D^{'}(\eta_{itj}^*)$ and μ_{itj} are the conditional and marginal mean of X_{itj}, respectively, we can treat $\boldsymbol{\epsilon}_{ij}^{\triangledown}$ and $\boldsymbol{\epsilon}_{ij}^{X}$ as error terms. Assumption 15 controls the norm of $\boldsymbol{\epsilon}_{ij}^{\triangledown}$ and $\boldsymbol{\epsilon}_{ij}^{X}$. Define the ψ_1-norm of a random vector $\boldsymbol{\phi} \in \mathbb{R}^T$ as

$$\|\boldsymbol{\phi}\|_{\psi_1} := \sup_{\|\boldsymbol{t}\|_\infty = 1} \left\| \langle \boldsymbol{\phi}, \boldsymbol{t} \rangle \right\|_{\psi_1} = \inf\left\{ K > 0 : \sup_{\|\mathbf{t}\|_\infty = 1} \mathbb{E} \exp\left(\left| \frac{\langle \boldsymbol{\phi}, \boldsymbol{t} \rangle}{K} \right| \right) \leq 2 \right\}.$$

We define another norm as

$$\|\boldsymbol{\phi}\|_{E_1} := \left\| |\boldsymbol{\phi}|_1 \right\|_{\psi_1} = \inf\left\{ K > 0 : \mathbb{E} \exp\left(\left| \frac{\boldsymbol{\phi}}{K} \right|_1 \right) = \mathbb{E} \exp\left(\frac{\sum_{t=1}^{T} |\phi_t|}{K} \right) \leq 2 \right\}.$$

Assumption 15 *For $\forall i, j$, $\max\{\|\boldsymbol{\epsilon}_{ij}^{\triangledown}\|_{\psi_1}, \|\boldsymbol{\epsilon}_{ij}^{X}\|_{\psi_1}\} \leq K\sqrt{T}$, for some positive constant K.*

When the data are independent, such bounds are reasonable; see Zajkowski (2019) for further details. Next, we impose Assumptions 16–17. Assumption 16 is about the mean and covariance matrix of $\boldsymbol{X}_{ij}$, which is similar to Assumption 6 under the Gaussian setting and Assumption 17 is the compatibility condition similar to Assumption 7. Let $\widetilde{C} = \max\{16Cc_dK, 2c_d, 8CK\}$, where C is a absolute constant, c_d and K are constants appearing in Assumptions 12 and 15.

Assumption 16 *For the ith subject and jth variable, let $\boldsymbol{X}_{ij} = (X_{i1j}, \ldots, X_{iTj})^{\mathrm{T}}$, $\boldsymbol{\mu}_{ij} = \mathbb{E}(\boldsymbol{X}_{ij})$ and $\boldsymbol{\Sigma}_{jj} = \mathrm{Var}(\boldsymbol{X}_{ij})$. Assume that the mean of X_{itj} is bounded by a constant, i.e., $|\mu_{itj}| \leq \mu_{\max}$. In addition, assume that the ℓ_2-norm of $\boldsymbol{\mu}_{ij}$ satisfies $\|\boldsymbol{\mu}_{ij}\|_2 \leq \mu_{\max}\sqrt{\widetilde{C}\Delta_{\max}\tau}n^{1/2}T^{1/4}$.*

Assumption 17 *Let $\widehat{\boldsymbol{\Sigma}}_j = (\mathbf{Y}_j - \mathbf{U}_j)^{\mathrm{T}}(\mathbf{Y}_j - \mathbf{U}_j)/(nT)$. For some constant $\phi_0 > 0$ and $\boldsymbol{\omega}_j^*$ satisfying $\|(\boldsymbol{\omega}_j^*)^{\mathcal{S}_j^c}\|_1 \leq 7\|(\boldsymbol{\omega}_j^*)^{\mathcal{S}_j}\|_1$, we have*

$$\left\|(\boldsymbol{\omega}_j^*)^{\mathcal{S}_j}\right\|_1^2 \leq \frac{\left(\boldsymbol{\omega}_j^*\right)^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_j\boldsymbol{\omega}_j^* s_j}{\phi_0^2}.$$

Theorem 18 establishes the estimation error in the context of exponential family graphical models.

Theorem 18 *Set the tuning parameters as*

$$\gamma = 2\widetilde{C}\log(2npT)/(n\sqrt{T}), \quad \lambda = \beta = 2\log(T)\log(nTp)T^{-1/4}.$$

When

$$\max\{\widetilde{c}_1 s_j \log(T)\log(nTp), \widetilde{c}_2\tau\log(2npT)\log^{-1}(T) + \widetilde{c}_3\sqrt{\tau}\log(2npT)\}T^{-\frac{1}{4}} \leq 1,$$

$$\frac{\mu_{\max}\sqrt{s_j\tau\widetilde{C}\Delta_{\max}}}{\phi_0 T^{\frac{1}{4}}} \leq \frac{1}{16},$$

$n, T, p \geq 6$, and under Assumptions 12–17, we have

$$\begin{aligned}&\left\|\widehat{\boldsymbol{\theta}}_{j,-j} - \boldsymbol{\theta}^*_{j,-j}\right\|_1 + \left\|\widehat{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}_j^*\right\|_1 + \frac{1}{\sqrt{nT}}\left\|\widehat{\boldsymbol{\Delta}}_j - \boldsymbol{\Delta}_j^*\right\|_2 \\ &\leq \max\{\widetilde{c}_1 s_j \log(T)\log(nTp), \widetilde{c}_2\tau\log(2npT)\log^{-1}(T) + \widetilde{c}_3\sqrt{\tau}\log(2npT)\}T^{-\frac{1}{4}}, \qquad (26)\end{aligned}$$

with probability at least $1 - 4(c_v c_u c_d + c_h c_u^2)/(c_d^2 T^{1/c_d}) - 4/(nTp) - 5/(n^2T^3)$, where $\widetilde{c}_1 = 16/(c_d\phi_0^2) + 2/(c_d\phi_0)$, $\widetilde{c}_2 = 112\widetilde{C}^2/c_d + 4(2 + \mu_{\max}\sqrt{7/c_d})^2\widetilde{C}\Delta_{\max} + 16\sqrt{7\Delta_{\max}/c_d}\widetilde{C}^{2/3}(2 + \mu_{\max}\sqrt{7/c_d})$, $\widetilde{c}_3 = 14\widetilde{C}/c_d + \sqrt{7/c_d}(2 + \mu_{\max}\sqrt{7/c_d})\sqrt{\widetilde{C}\Delta_{\max}}$, $\widetilde{C} = \max\{16Cc_dK, 2c_d, 8CK\}$ and C is a absolute constant.

If we assume that ϕ_0, σ_m, $\Delta_{\max}$, $\mu_{\max}$, and τ are all constants and $n, T \to \infty$, Theorems 18 is equivalent to that with probability tending to 1,

$$\left\|\widehat{\boldsymbol{\theta}}_{j,-j} - \boldsymbol{\theta}^*_{j,-j}\right\|_1 + \left\|\widehat{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}_j^*\right\|_1 + \frac{1}{\sqrt{nT}}\left\|\widehat{\boldsymbol{\Delta}}_j - \boldsymbol{\Delta}_j^*\right\|_2 \lesssim s_j\log(T)\log(nTp)T^{-1/4}, \qquad (27)$$

where the notation $a_n \lesssim b_n$ stands for $a_n = Cb_n$ with C is a constant. We can see that this rate is slower than Theorems 8 and 9 in the context of Gaussian graphical models. This is mainly due to technical difficulties in the analysis of the fused lasso estimator under the general exponential family distribution. Specifically, when the data are Gaussian, Lemma 22 provides a sharp bound for the inner product of the Gaussian error and fussed lasso error $\bar{\boldsymbol{\Delta}}_{ij} - \boldsymbol{\Delta}_{ij}$, by carefully truncating the eigenvalues of the difference operator $\mathbf{C}$ and applying the Hanson-Wright inequality subsequently. However, such techniques are not available for the general exponential family distribution, leading to a slower rate in the analysis of the fused lasso error. Developing new techniques for the proof of such cases are out of the scope of this manuscript and we leave it for future work.

Appendix F. Proof for Gaussian Graphical Models

F.1 Technical Lemmas

We first provide some technical lemmas to facilitate the proof of Theorems 8–9. Lemmas 19 and 20 control the tail behavior of interaction terms between the observed variable and the random noise ϵ_{itj}. The proof of Lemma 19 is provided in Section F.3.1. The proof of Lemma 20 is similar to Lemma 19 and is omitted. Recall from Section 4 that $X_{itj} \sim N(\mu_{itj}, \sigma^2_{jj,t})$ with $|\mu_{itj}| \leq \mu_{\max}$ and $(\sigma^X_m)^2 = \max_{j,t}(\sigma^2_{jj,t})$. Besides, let $\epsilon_{itj} \sim N\{0, (\sigma^\epsilon_{jj,t})^2\}$ and $(\sigma^\epsilon_m)^2 = \max_{j,t}\{(\sigma^\epsilon_{jj,t})^2\}$.

Lemma 19 *Assume $X_{itj} \sim N(\mu_{itj}, \sigma^2_{jj,t})$ and $\epsilon_{itj} \sim N\{0, (\sigma^\epsilon_{jj,t})^2\}$. We have*

$$\max_{1\leq k\leq p, k\neq j} \frac{1}{nT}\left|\sum_{i=1}^{n}\sum_{t=1}^{T}\epsilon_{itj}X_{itk}\right| \leq \lambda_0,$$

with probability at least $1-\exp(-\min[\log^2(T)/\{2\sigma^\epsilon_m\sqrt{\mu^2_{\max}+(\sigma^X_m)^2}\}^2, \log(T)/|2\sigma^\epsilon_m\sqrt{\mu^2_{\max}+(\sigma^X_m)^2}|]/2) - \exp[\log\{2(p-1)\} - 3\lambda_0^2 nT/\{2\lambda_0\log(T) + 6\log^2(T)\}]$.

Lemma 20 *Assume $X_{i(t-1)k} \sim N(\mu_{i(t-1)k}, \sigma^2_{kk,t-1})$ and $\epsilon_{itj} \sim N\{0, (\sigma^\epsilon_{jj,t})^2\}$. We have*

$$\max_{1\leq k\leq p} \frac{1}{nT}\left|\sum_{i=1}^{n}\sum_{t=1}^{T}\epsilon_{itj}X_{i(t-1)k}\right| \leq \beta_0,$$

with probability at least $1-\exp(-\min[\log^2(T)/\{2\sigma^\epsilon_m\sqrt{\mu^2_{\max}+(\sigma^X_m)^2}\}^2, \log(T)/|2\sigma^\epsilon_m\sqrt{\mu^2_{\max}+(\sigma^X_m)^2}|]/2) - \exp[\log(2p) - 3\beta_0^2 nT/\{2\beta_0\log(T) + 6\log^2(T)\}]$.

Lemmas 21–22 establishes upper bounds for terms related to $\boldsymbol{\Delta}_{ij}$. The proof of Lemma 21 is provided in Section F.3.2. The results in Lemmas 21 and 22 are similar, except that Lemma 22 is a maximum bound. The proof of Lemma 22 is similar to Lemma 21, and is hence omitted.

Lemma 21 *Let $\boldsymbol{\eta} \sim N(\mathbf{0}, \mathbf{Q})$ and $D = 8\sqrt{T\log(nT)/(\pi^2 i_0)}$. For $i_0 \in \{1, \ldots, T-1\}$ and any deterministic vector $\boldsymbol{\Delta}_{ij}$, we have*

$$\frac{1}{nT}\boldsymbol{\eta}^{\mathrm{T}}(\bar{\boldsymbol{\Delta}}_{ij} - \boldsymbol{\Delta}_{ij}) \leq \frac{\sqrt{2\|\mathbf{Q}\|_{\mathrm{op}}}\left\{\sqrt{i_0 + \log(n)} + \sqrt{\log(nT)}\right\}}{nT}\|\bar{\boldsymbol{\Delta}}_{ij} - \boldsymbol{\Delta}_{ij}\|_2 + \frac{\sqrt{\|\mathbf{Q}\|_{\mathrm{op}}}D}{nT}\left(\|\mathbf{C}\bar{\boldsymbol{\Delta}}_{ij}\|_1 + \|\mathbf{C}\boldsymbol{\Delta}_{ij}\|\right.$$

with probability at least $1-[2n^2(T-1)\sqrt{\log\{n(T-1)\}}]^{-1}-2\exp(-i_0)/n-1/\{nT\sqrt{2\log(nT)}\}$, *where* $\|\mathbf{Q}\|_{\text{op}}$ *is the operator norm of* $\mathbf{Q}$.

Lemma 22 *Let* $\boldsymbol{\eta}_j \sim N(\mathbf{0}, \mathbf{Q})$ *and* $D' = 8\sqrt{T\log(2pnT)/(\pi^2 i_0)}$. *For* $i_0 \in \{1, \ldots, T-1\}$ *and all deterministic vector* $\boldsymbol{\Delta}_{ij}$, *we have that uniformly over* $1 \le j \le p$,

$$\begin{aligned}
&\frac{1}{nT}\boldsymbol{\eta}_j^{\mathrm{T}}\left(\bar{\boldsymbol{\Delta}}_{ij} - \boldsymbol{\Delta}_{ij}\right) \\
&\le \frac{\sqrt{2\|\mathbf{Q}\|_{\text{op}}}\left\{\sqrt{i_0 + \log(2pn)} + \sqrt{\log(2pnT)}\right\}}{nT}\|\bar{\boldsymbol{\Delta}}_{ij} - \boldsymbol{\Delta}_{ij}\|_2 + \frac{\sqrt{\|\mathbf{Q}\|_{\text{op}}}D'}{nT}\left(\|\mathbf{C}\bar{\boldsymbol{\Delta}}_{ij}\|_1 + \|\mathbf{C}\boldsymbol{\Delta}_{ij}\|_1\right),
\end{aligned}$$

with probability at least $1-[8(T-1)p^2n^2\sqrt{\log\{2pn(T-1)\}}]^{-1}-\exp(-i_0)/(np)-1/\{2pnT\sqrt{2\log(2pnT)}\}$, *where* $\|\mathbf{Q}\|_{\text{op}}$ *is the operator norm of* $\mathbf{Q}$.

F.2 Proof of Theorems

F.2.1 Proof of Theorem 8

Let $\boldsymbol{\theta}^*_{j,-j} \in \mathbb{R}^{p-1}$, $\boldsymbol{\alpha}^*_j \in \mathbb{R}^p$, and $\boldsymbol{\Delta}^*_j \in \mathbb{R}^{nT}$ be the true underlying parameters, and let $\widehat{\boldsymbol{\theta}}_{j,-j}$, $\widehat{\boldsymbol{\alpha}}_j$, and $\widehat{\boldsymbol{\Delta}}_j$ be the solution obtained from solving (9) under the Gaussian loss. For notational convenience, we write $\boldsymbol{\omega}^*_j = \{(\boldsymbol{\theta}^*_{j,-j})^{\mathrm{T}}, (\boldsymbol{\alpha}^*_j)^{\mathrm{T}}\}^{\mathrm{T}}$ and $\widehat{\boldsymbol{\omega}}_j = (\widehat{\boldsymbol{\theta}}^{\mathrm{T}}_{j,-j}, \widehat{\boldsymbol{\alpha}}^{\mathrm{T}}_j)^{\mathrm{T}}$. Let $\mathcal{S}_j = \{k : \omega^*_{jk} \neq 0\}$ be the active set and let $s_j = |\mathcal{S}_j|$ be the cardinality of $\mathcal{S}_j$. To establish an upper bound on the estimation error, we start with defining

$$N = \|\widehat{\boldsymbol{\theta}}_{j,-j} - \boldsymbol{\theta}^*_{j,-j}\|_1 + \|\widehat{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}^*_j\|_1 + \frac{1}{\sqrt{nT}}\|\widehat{\boldsymbol{\Delta}}_j - \boldsymbol{\Delta}^*_j\|_2.$$

The goal is to show that $N \le M$, where

$$\begin{aligned}
M = 2\max[&c_1 n^{-\frac{1}{6}}T^{-\frac{1}{3}}s_j\log(T)\log(nTp), \\
&n^{-\frac{1}{6}}T^{-\frac{1}{3}}\left\{c_2\tau^{\frac{2}{3}}\log^{-1}(T)\log^{-\frac{1}{3}}(nTp) + c_5\tau^{\frac{1}{3}}\log^{\frac{1}{3}}(2pnT)\right\} \\
&+ n^{-\frac{1}{2}}T^{-\frac{1}{2}}\left\{448\sigma_m\sqrt{\log(2pnT)} + 3584\sigma_m^2\log^{-1}(T) + c_3\tau^{\frac{1}{3}}\log^{-1}(T)\log^{-\frac{1}{6}}(nTp)\right\}],
\end{aligned}$$

where $c_1 = (32 + 4\sqrt{\phi_0})/\phi_0$, $c_2 = 3584\sigma_m^2 c_4^{4/3}(\mu_{\max}+1)^2$, $c_3 = 7168c_4^{2/3}\sigma_m^2(\mu_{\max}+1)$, $c_4 = [4\Delta^2_{\max}/\pi^2]^{\frac{1}{4}}$ and $c_5 = 448\sigma_m(\mu_{\max}+1)c_4^{2/3}$. Note that the constant $\phi_0 > 0$ is the same compatability-type constant that appears in Assumption 7. Let $\zeta = M/(M+N)$ such that $0 < \zeta < 1$. Set

$$\begin{aligned}
\bar{\boldsymbol{\theta}}_{j,-j} &= \zeta\widehat{\boldsymbol{\theta}}_{j,-j} + (1-\zeta)\boldsymbol{\theta}^*_{j,-j}; \\
\bar{\boldsymbol{\alpha}}_j &= \zeta\widehat{\boldsymbol{\alpha}}_j + (1-\zeta)\boldsymbol{\alpha}^*_j; \\
\bar{\boldsymbol{\Delta}}_j &= \zeta\widehat{\boldsymbol{\Delta}}_j + (1-\zeta)\boldsymbol{\Delta}^*_j.
\end{aligned}$$

Then, it can be shown that $\zeta N = \|\bar{\boldsymbol{\theta}}_{j,-j} - \boldsymbol{\theta}^*_{j,-j}\|_1 + \|\bar{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}^*_j\|_1 + \|\bar{\boldsymbol{\Delta}}_j - \boldsymbol{\Delta}^*_j\|_2/\sqrt{nT}$. In the following, we show that $\zeta N \le M/2$, which implies $N \le M$.

Let $Q\left(\boldsymbol{\omega}_j, \boldsymbol{\Delta}_j\right)$ be the loss function in (9) under the assumption that the random variables are Gaussian, that is,

$$Q\left(\boldsymbol{\omega}_j, \boldsymbol{\Delta}_j\right) = \frac{1}{2nT}\left\|\mathbf{X}_j - \mathbf{Y}_j\boldsymbol{\omega}_j - \boldsymbol{\Delta}_j\right\|_2^2 + \lambda\left\|\boldsymbol{\theta}_{j,-j}\right\|_1 + \beta\left\|\boldsymbol{\alpha}_j\right\|_1 + \gamma\sum_{i=1}^{n}\left\|\mathbf{C}\boldsymbol{\Delta}_{ij}\right\|_1, \quad (28)$$

where $\boldsymbol{\omega}_j = (\boldsymbol{\theta}_{j,-j}^{\mathrm{T}}, \boldsymbol{\alpha}_j^{\mathrm{T}})^{\mathrm{T}}$, $\mathbf{Y}_{itj} = (\mathbf{X}_{it(-j)}, \mathbf{X}_{i(t-1)})$, and $\mathbf{Y}_j = (\mathbf{Y}_{11j}^{\mathrm{T}}, \mathbf{Y}_{12j}^{\mathrm{T}}, \ldots, \mathbf{Y}_{1Tj}^{\mathrm{T}}, \mathbf{Y}_{21j}^{\mathrm{T}}, \ldots, \mathbf{Y}_{nTj}^{\mathrm{T}})^{\mathrm{T}}$. Since $Q(\cdot)$ is a convex loss, by convexity, we have

$$\begin{aligned} Q\left(\bar{\boldsymbol{\omega}}_j, \bar{\boldsymbol{\Delta}}_j\right) &= Q\left\{\zeta\widehat{\boldsymbol{\omega}}_j + (1-\zeta)\,\boldsymbol{\omega}_j^*, \zeta\widehat{\boldsymbol{\Delta}}_j + (1-\zeta)\boldsymbol{\Delta}_j^*\right\} \\ &\le \zeta Q\left(\widehat{\boldsymbol{\omega}}_j, \widehat{\boldsymbol{\Delta}}_j\right) + (1-\zeta)\,Q\left(\boldsymbol{\omega}_j^*, \boldsymbol{\Delta}_j^*\right) \\ &\le Q\left(\boldsymbol{\omega}_j^*, \boldsymbol{\Delta}_j^*\right), \end{aligned} \quad (29)$$

where the last inequality follows from the fact that $Q(\widehat{\boldsymbol{\omega}}_j, \widehat{\boldsymbol{\Delta}}_j) \le Q(\boldsymbol{\omega}_j^*, \boldsymbol{\Delta}_j^*)$. Substituting (16) and (28) into (29), and upon rearranging the terms, we have

$$\begin{aligned} &\frac{1}{2nT}\left\|\mathbf{Y}_j\left(\bar{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*\right)\right\|_2^2 + \frac{1}{2nT}\left\|\bar{\boldsymbol{\Delta}}_j - \boldsymbol{\Delta}_j^*\right\|_2^2 \\ &\le \underbrace{\frac{1}{nT}\boldsymbol{\epsilon}_j^{\mathrm{T}}\mathbf{Y}_j\left(\bar{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*\right) + \lambda\left(\|\boldsymbol{\theta}_{j,-j}^*\|_1 - \|\bar{\boldsymbol{\theta}}_{j,-j}\|_1\right) + \beta\left(\|\boldsymbol{\alpha}_j^*\|_1 - \|\bar{\boldsymbol{\alpha}}_j\|_1\right)}_{\mathbb{I}_1} \\ &+ \sum_{i=1}^{n}\left\{\underbrace{\frac{1}{nT}\boldsymbol{\epsilon}_{ij}^{\mathrm{T}}\left(\bar{\boldsymbol{\Delta}}_{ij} - \boldsymbol{\Delta}_{ij}^*\right) + \frac{1}{nT}\left(\bar{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*\right)^{\mathrm{T}}\mathbf{Y}_{ij}^{\mathrm{T}}\left(\boldsymbol{\Delta}_{ij}^* - \bar{\boldsymbol{\Delta}}_{ij}\right)}_{\mathbb{I}_2} + \gamma\left(\|\mathbf{C}\boldsymbol{\Delta}_{ij}^*\|_1 - \|\mathbf{C}\bar{\boldsymbol{\Delta}}_{ij}\|_1\right)\right\}. \end{aligned} \quad (30)$$

We now establish upper bounds for $\mathbb{I}_1$ and $\mathbb{I}_2$, respectively.

Upper Bound for $\mathbb{I}_1$: from the definition of $\mathbf{Y}_j$, we have

$$\boldsymbol{\epsilon}_j^{\mathrm{T}}\mathbf{Y}_j(\bar{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*) = \underbrace{\boldsymbol{\epsilon}_j^{\mathrm{T}}\mathbf{X}_{-j}^{\otimes}(\bar{\boldsymbol{\theta}}_{j,-j} - \boldsymbol{\theta}_{j,-j}^*)}_{\mathbb{I}_{11}} + \underbrace{\boldsymbol{\epsilon}_j^{\mathrm{T}}\mathbf{X}_{T-1}^{\otimes}(\bar{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}_j^*)}_{\mathbb{I}_{12}}.$$

It suffices to obtain upper bounds for $\mathbb{I}_{11}$ and $\mathbb{I}_{12}$. By the Holder's inequality, Lemma 19, and picking $\lambda = 2\log(T)\log(nTp)n^{-1/6}T^{-1/3}$, when $n, T, p \ge 6$ we have

$$\frac{1}{nT}\boldsymbol{\epsilon}_j^{\mathrm{T}}\mathbf{X}_{-j}^{\otimes}\left(\bar{\boldsymbol{\theta}}_{j,-j} - \boldsymbol{\theta}_{j,-j}^*\right) \le \frac{1}{nT}\left\|\boldsymbol{\epsilon}_j^{\mathrm{T}}\mathbf{X}_{-j}^{\otimes}\right\|_\infty \cdot \left\|\bar{\boldsymbol{\theta}}_{j,-j} - \boldsymbol{\theta}_{j,-j}^*\right\|_1 \le \frac{\lambda}{2}\left\|\bar{\boldsymbol{\theta}}_{j,-j} - \boldsymbol{\theta}_{j,-j}^*\right\|_1, \quad (31)$$

with probability at least $1-\exp(-\min[\log^2(T)/\{2\sigma_m^\epsilon\sqrt{\mu_{\max}^2 + (\sigma_m^X)^2}\}^2, \log(T)/|2\sigma_m^\epsilon\sqrt{\mu_{\max}^2 + (\sigma_m^X)^2}|]/2) - 2/(nTp)$. Similarly, by an application of Lemma 20 and picking $\beta = \lambda$, when $n, T, p \ge 6$ we obtain

$$\frac{1}{nT}\boldsymbol{\epsilon}_j^{\mathrm{T}}\mathbf{X}_{T-1}^{\otimes}(\bar{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}_j^*) \le \frac{1}{nT}\left\|\boldsymbol{\epsilon}_j^{\mathrm{T}}\mathbf{X}_{T-1}^{\otimes}\right\|_\infty \cdot \|\bar{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}_j^*\|_1 \le \frac{\beta}{2}\|\bar{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}_j^*\|_1, \quad (32)$$

with probability at least $1-\exp(-\min[\log^2(T)/\{2\sigma_m^\epsilon\sqrt{\mu_{\max}^2+(\sigma_m^X)^2}\}^2, \log(T)/|2\sigma_m^\epsilon\sqrt{\mu_{\max}^2+(\sigma_m^X)^2}|]/2)-2/(nTp)$. Since $\lambda=\beta$, substituting (31) and (32) into $\mathbb{I}_1$ yields

$$\mathbb{I}_1 \leq \frac{\beta}{2}\|\bar{\boldsymbol{\omega}}_j-\boldsymbol{\omega}_j^*\|_1+\beta\left(\|\boldsymbol{\omega}_j^*\|_1-\|\bar{\boldsymbol{\omega}}_j\|_1\right), \tag{33}$$

with probability at least $1-2\exp(-\min[\log^2(T)/\{2\sigma_m^\epsilon\sqrt{\mu_{\max}^2+(\sigma_m^X)^2}\}^2, \log(T)/|2\sigma_m^\epsilon\sqrt{\mu_{\max}^2+(\sigma_m^X)^2}|]/2)-4/(nTp)$. Let $\boldsymbol{\omega}_j^{\mathcal{S}_j}$ and $\boldsymbol{\omega}_j^{\mathcal{S}_j^c}$ be subvectors of $\boldsymbol{\omega}_j$ with indices $\mathcal{S}_j$ and $\mathcal{S}_j^c$, respectively. Then, upon rearranging the terms, (33) can be rewritten as

$$\mathbb{I}_1 \leq \frac{3\beta}{2}\left\|\bar{\boldsymbol{\omega}}_j^{\mathcal{S}_j}-(\boldsymbol{\omega}_j^*)^{\mathcal{S}_j}\right\|_1-\frac{\beta}{2}\left\|\bar{\boldsymbol{\omega}}_j^{\mathcal{S}_j^c}\right\|_1, \tag{34}$$

with probability at least $1-2\exp(-\min[\log^2(T)/\{2\sigma_m^\epsilon\sqrt{\mu_{\max}^2+(\sigma_m^X)^2}\}^2, \log(T)/|2\sigma_m^\epsilon\sqrt{\mu_{\max}^2+(\sigma_m^X)^2}|]/2)-4/(nTp)$.

Upper Bound for $\mathbb{I}_2$: we start with providing an upper bound for $\boldsymbol{\epsilon}_{ij}^{\mathrm{T}}(\bar{\boldsymbol{\Delta}}_{ij}-\boldsymbol{\Delta}_{ij}^*)/(nT)$. For $i_0\in\{1,\ldots,T-1\}$, let $D=8\sqrt{T\log(nT)/(\pi^2 i_0)}$. Recall that $\epsilon_{itj}\sim N\{0,(\sigma_{tj}^\epsilon)^2\}$ and $(\sigma_m^\epsilon)^2=\max_{t,j}\{(\sigma_{tj}^\epsilon)^2\}$. By Lemma 21, we have

$$\frac{1}{nT}\boldsymbol{\epsilon}_{ij}^{\mathrm{T}}\left(\bar{\boldsymbol{\Delta}}_{ij}-\boldsymbol{\Delta}_{ij}^*\right)\leq\frac{2\sigma_m^\epsilon\left\{\sqrt{i_0\log(n)}+\sqrt{\log(nT)}\right\}}{nT}\left\|\bar{\boldsymbol{\Delta}}_{ij}-\boldsymbol{\Delta}_{ij}^*\right\|_2+\frac{\sigma_m^\epsilon D}{nT}\left(\left\|\mathbf{C}\bar{\boldsymbol{\Delta}}_{ij}\right\|_1+\left\|\mathbf{C}\boldsymbol{\Delta}_{ij}^*\right\|_1\right), \tag{35}$$

with probability at least $1-[2n^2(T-1)\sqrt{\log\{n(T-1)\}}]^{-1}-2\exp(-i_0)/n-1/\{nT\sqrt{2\log(nT)}\}$.

Next, we provide an upper bound for $(\bar{\boldsymbol{\omega}}_j-\boldsymbol{\omega}_j^*)^{\mathrm{T}}\mathbf{Y}_{ij}^{\mathrm{T}}(\boldsymbol{\Delta}_{ij}^*-\bar{\boldsymbol{\Delta}}_{ij})/(nT)$ in $\mathbb{I}_2$. Let $\mathbb{E}(\mathbf{Y}_{ij})=\mathbf{U}_{ij}$ and $B(n,T)=\mu_{\max}\min(c_4^{2/3}n^{1/3}T^{1/6},\sqrt{T})$. By Assumption 6, $\|\boldsymbol{\mu}_{ij}\|_2\leq B(n,T)$. Recall that $\|\bar{\boldsymbol{\omega}}_j-\boldsymbol{\omega}_j^*\|_1+\|\bar{\boldsymbol{\Delta}}_j-\boldsymbol{\Delta}_j^*\|_2/\sqrt{nT}=\zeta N\leq M$ with $\zeta=M/(N+M)$. This implies that $\left\|\bar{\boldsymbol{\omega}}_j-\boldsymbol{\omega}_j^*\right\|_1\leq M$. Coupling the above with the Holder's inequality, we obtain

$$\begin{aligned}
&\frac{1}{nT}\left(\bar{\boldsymbol{\omega}}_j-\boldsymbol{\omega}_j^*\right)^{\mathrm{T}}\mathbf{Y}_{ij}^{\mathrm{T}}\left(\boldsymbol{\Delta}_{ij}^*-\bar{\boldsymbol{\Delta}}_{ij}\right)\\
&\leq\frac{1}{nT}\left\|\bar{\boldsymbol{\omega}}_j-\boldsymbol{\omega}_j^*\right\|_1\cdot\left\|\mathbf{Y}_{ij}^{\mathrm{T}}\left(\boldsymbol{\Delta}_{ij}^*-\bar{\boldsymbol{\Delta}}_{ij}\right)\right\|_\infty\\
&\leq\frac{M}{nT}\left\|\mathbf{Y}_{ij}^{\mathrm{T}}\left(\boldsymbol{\Delta}_{ij}^*-\bar{\boldsymbol{\Delta}}_{ij}\right)\right\|_\infty\\
&\leq\frac{M}{nT}\left\|\left(\mathbf{Y}_{ij}-\mathbf{U}_{ij}\right)^{\mathrm{T}}\left(\boldsymbol{\Delta}_{ij}^*-\bar{\boldsymbol{\Delta}}_{ij}\right)\right\|_\infty+\frac{M}{nT}\left\|\mathbf{U}_{ij}^{\mathrm{T}}\left(\boldsymbol{\Delta}_{ij}^*-\bar{\boldsymbol{\Delta}}_{ij}\right)\right\|_\infty\\
&\leq\frac{M}{nT}\left\|\left(\mathbf{Y}_{ij}-\mathbf{U}_{ij}\right)^{\mathrm{T}}\left(\boldsymbol{\Delta}_{ij}^*-\bar{\boldsymbol{\Delta}}_{ij}\right)\right\|_\infty+\frac{M}{nT}\max_{k\neq j}\left\|\boldsymbol{\mu}_{ik}\right\|_2\left\|\boldsymbol{\Delta}_{ij}^*-\bar{\boldsymbol{\Delta}}_{ij}\right\|_2\\
&\leq\frac{2M\sqrt{\kappa}\left\{\sqrt{i_0\log(2pn)}+\sqrt{\log(2pnT)}+B(n,T)\right\}}{nT}\left\|\bar{\boldsymbol{\Delta}}_{ij}-\boldsymbol{\Delta}_{ij}^*\right\|_2+\frac{M\sqrt{\kappa}D'}{nT}\left(\left\|\mathbf{C}\bar{\boldsymbol{\Delta}}_{ij}\right\|_1+\left\|\mathbf{C}\boldsymbol{\Delta}_{ij}^*\right\|_1\right),
\end{aligned} \tag{36}$$

with probability at least $1-[8n^2p(T-1)\sqrt{\log\{n(T-1)\}}]^{-1}-\exp\{-i_0\}/n-1/\{2nT\sqrt{2\log(2npT)}\}$. Note that the last inequality follows by an application of Lemma 22 and Assumption 6. Let

$\sigma_m = \max\{2\sqrt{\kappa}, 2\sigma_m^{\epsilon}, 1\}$. We have $M \leq 1$ by the condition that $2\max(c_1 n^{-1/6}T^{-1/3}, c_2 n^{-1/6}T^{-1/3} + c_3 n^{-1/2}T^{-1/2}) \leq 1$. Combining (35) and (36), we have

$$\mathbb{I}_2 \leq \frac{2\left\{\sqrt{i_0 \log(2pn)} + \sqrt{\log(2pnT)} + B(n,T)\right\}\sigma_m}{nT}\|\bar{\boldsymbol{\Delta}}_{ij} - \boldsymbol{\Delta}_{ij}^*\|_2 + \frac{2\sigma_m D'}{nT}\left(\|\mathbf{C}\boldsymbol{\Delta}_{ij}^*\|_1 + \|\mathbf{C}\bar{\boldsymbol{\Delta}}_{ij}\|_1\right), \tag{37}$$

with probability at least $1 - [n^2(T-1)\sqrt{\log\{n(T-1)\}}]^{-1} - 3\exp\{-i_0\}/n - 2/\{nT\sqrt{\log(nT)}\}$.

Moreover, recall that γ is the tuning parameter for $\|(\mathbf{I}_n \otimes \mathbf{C})\boldsymbol{\Delta}_j^*\|_1$. Let $\gamma = 2\sigma_m D'/(nT)$ and substituting (34) and (37) into (30), we have

$$\begin{aligned}
&\frac{1}{2nT}\|\mathbf{Y}_j(\bar{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*)\|_2^2 + \frac{1}{2nT}\|\bar{\boldsymbol{\Delta}}_j - \boldsymbol{\Delta}_j^*\|_2^2 + \frac{\beta}{2}\|\bar{\boldsymbol{\omega}}_j^{\mathcal{S}_j^c}\| \\
&\leq \frac{3\beta}{2}\|\bar{\boldsymbol{\omega}}_j^{\mathcal{S}_j} - (\boldsymbol{\omega}_j^*)^{\mathcal{S}_j}\|_1 + \frac{2\sigma_m\left\{\sqrt{i_0\log(2pn)} + \sqrt{\log(2pnT)} + B(n,T)\right\}}{nT}\|\bar{\boldsymbol{\Delta}}_j - \boldsymbol{\Delta}_j^*\|_2 \\
&\quad + \frac{4\sigma_m D'}{nT}\|(\mathbf{I}_n \otimes \mathbf{C})\boldsymbol{\Delta}_j^*\|_1,
\end{aligned} \tag{38}$$

with probability at least $1 - [n(T-1)\sqrt{\log\{n(T-1)\}}]^{-1} - 3\exp\{-i_0\} - 2/\{T\sqrt{\log(nT)}\} - 4/(nTp) - \exp(-\min[\log^2(T)/\{2\sigma_m^{\epsilon}\sqrt{\mu_{\max}^2 + (\sigma_m^X)^2}\}^2, \log(T)/|2\sigma_m^{\epsilon}\sqrt{\mu_{\max}^2 + (\sigma_m^X)^2}|]/2)$.

We now consider (38) under the following two cases:

1. $\frac{2\sigma_m\left\{\sqrt{i_0\log(2pn)} + \sqrt{\log(2pnT)} + B(n,T)\right\}}{nT}\|\bar{\boldsymbol{\Delta}}_j - \boldsymbol{\Delta}_j^*\|_2 + \frac{4\sigma_m D'}{nT}\|(\mathbf{I}_n \otimes \mathbf{C})\boldsymbol{\Delta}_j^*\|_1 \leq \frac{1}{4}\beta\|\bar{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*\|_1$;

2. $\frac{2\sigma_m\left\{\sqrt{i_0\log(2pn)} + \sqrt{\log(2pnT)} + B(n,T)\right\}}{nT}\|\bar{\boldsymbol{\Delta}}_j - \boldsymbol{\Delta}_j^*\|_2 + \frac{4\sigma_m D'}{nT}\|(\mathbf{I}_n \otimes \mathbf{C})\boldsymbol{\Delta}_j^*\|_1 > \frac{1}{4}\beta\|\bar{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*\|_1$.

Recall that $\zeta N = \|\bar{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*\|_1 + \|\bar{\boldsymbol{\Delta}}_j - \boldsymbol{\Delta}_j^*\|_2/\sqrt{nT}$ and the goal is to obtain $\zeta N \leq M/2$. To this end, we will derive upper bounds for $\|\bar{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*\|_1$ and $\|\bar{\boldsymbol{\Delta}}_j - \boldsymbol{\Delta}_j^*\|_2/\sqrt{nT}$ separately.

Case 1: in this case, (38) can be simplified to

$$\frac{1}{2nT}\|\mathbf{Y}_j(\bar{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*)\|_2^2 + \frac{1}{2nT}\|\bar{\boldsymbol{\Delta}}_j - \boldsymbol{\Delta}_j^*\|_2^2 + \frac{\beta}{4}\|\bar{\boldsymbol{\omega}}_j^{\mathcal{S}_j^c}\| \leq \frac{7\beta}{4}\|\bar{\boldsymbol{\omega}}_j^{\mathcal{S}_j} - (\boldsymbol{\omega}_j^*)^{\mathcal{S}_j}\|_1. \tag{39}$$

Since $\|\bar{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*\|_1 = \|\bar{\boldsymbol{\omega}}_j^{\mathcal{S}_j} - (\boldsymbol{\omega}_j^*)^{\mathcal{S}_j}\|_1 + \|\bar{\boldsymbol{\omega}}_j^{\mathcal{S}_j^c}\|_1$, following an argument similar to Lemma 6.3 in Bühlmann and Van De Geer (2011), we have

$$
\begin{aligned}
&\frac{2}{nT}\left\|\mathbf{Y}_j\left(\bar{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*\right)\right\|_2^2 + \frac{2}{nT}\left\|\bar{\boldsymbol{\Delta}}_j - \boldsymbol{\Delta}_j^*\right\|_2^2 + \beta\left\|\bar{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*\right\|_1 \\
&= \frac{2}{nT}\left\|\mathbf{Y}_j\left(\bar{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*\right)\right\|_2^2 + \frac{2}{nT}\left\|\bar{\boldsymbol{\Delta}}_j - \boldsymbol{\Delta}_j^*\right\|_2^2 + \beta\left\|\bar{\boldsymbol{\omega}}_j^{\mathcal{S}_j} - (\boldsymbol{\omega}_j^*)^{\mathcal{S}_j}\right\|_1 + \beta\left\|\bar{\boldsymbol{\omega}}_j^{\mathcal{S}_j^c}\right\|_1 \\
&\leq 8\beta\left\|\bar{\boldsymbol{\omega}}_j^{\mathcal{S}_j} - (\boldsymbol{\omega}_j^*)^{\mathcal{S}_j}\right\|_1 \\
&\leq 8\beta\sqrt{\frac{s_j}{\phi_0 nT}}\left\|\mathbf{Y}_j\left(\bar{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*\right)\right\|_2 + 8\beta\sqrt{\frac{s_j}{\phi_0 nT}}\left\|\mathbf{U}_j\left(\bar{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*\right)\right\|_2 + 8\beta\frac{\sqrt{2\mathcal{C}s_j\log\left\{nT\left(2p-1\right)\right\}}}{\sqrt{\phi_0}\left(nT\right)^{\frac{1}{4}}}\left\|\bar{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*\right\|_1 \\
&\leq 8\beta\sqrt{\frac{s_j}{\phi_0 nT}}\left\|\mathbf{Y}_j\left(\bar{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*\right)\right\|_2 + 8\beta\left[\frac{\sqrt{s_j}c_4^{\frac{2}{3}}\mu_{\max}\tau^{\frac{1}{3}}\log^{\frac{1}{3}}\left(2pnT\right)}{\sqrt{\phi_0}n^{\frac{1}{6}}T^{\frac{1}{3}}} + \frac{\sqrt{2\mathcal{C}s_j\log\left\{nT\left(2p-1\right)\right\}}}{\sqrt{\phi_0}\left(nT\right)^{\frac{1}{4}}}\right]\left\|\bar{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*\right\|_1 \\
&\leq \frac{2}{nT}\left\|\mathbf{Y}_j\left(\bar{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*\right)\right\|_2^2 + \frac{8\beta^2 s_j}{\phi_0} + 8\beta\left[\frac{\sqrt{s_j}c_4^{\frac{2}{3}}\mu_{\max}\tau^{\frac{1}{3}}\log^{\frac{1}{3}}\left(2pnT\right)}{\sqrt{\phi_0}n^{\frac{1}{6}}T^{\frac{1}{3}}} + \frac{\sqrt{2\mathcal{C}s_j\log\left\{nT\left(2p-1\right)\right\}}}{\sqrt{\phi_0}\left(nT\right)^{\frac{1}{4}}}\right]\left\|\bar{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*\right\|_1,
\end{aligned}
\tag{40}
$$

with probability at least $1-[n(T-1)\sqrt{\log\{n(T-1)\}}]^{-1} - 3\exp\{-i_0\} - 2/\{T\sqrt{\log(nT)}\} - 4/(nTp) - \exp(-\min[\log^2(T)/\{2\sigma_m^\epsilon\sqrt{\mu_{\max}^2+(\sigma_m^X)^2}\}^2, \log(T)/|2\sigma_m^\epsilon\sqrt{\mu_{\max}^2+(\sigma_m^X)^2}|]/2)$. The first inequality follows from (39), the second inequality follows from Lemma 28, the third inequality follows the Holder's inequality and Assumption 6, and the last inequality follows from the fact that $uv \leq u^2+v^2/4$ for any $u, v \geq 0$. To simplify (40), if $8\sqrt{s_j}c_4^{2/3}\mu_{\max}\tau^{1/3}\log^{1/3}(2pnT)/(\sqrt{\phi_0}n^{1/6}$ $8\sqrt{2\mathcal{C}s_j\log\{nT(2p-1)\}}/\{\sqrt{\phi_0}(nT)^{1/4}\} \leq 1/2$ and with probability at least $1-[n(T-1)\sqrt{\log\{n(T-1)\}}]^{-1} - 3\exp\{-i_0\} - 2/\{T\sqrt{\log(nT)}\} - \exp(-\min[\log^2(T)/\{2\sigma_m^\epsilon\sqrt{\mu_{\max}^2+(\sigma_m^X)^2}\}^2, \log(T)/|2\sigma_m^\epsilon$ $4/(nTp)$,

$$
\frac{2}{nT}\left\|\bar{\boldsymbol{\Delta}}_j - \boldsymbol{\Delta}_j^*\right\|_2^2 + \frac{\beta}{2}\left\|\bar{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*\right\|_1 \leq \frac{8\beta^2 s_j}{\phi_0},
\tag{41}
$$

which directly implies $\|\bar{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*\|_1 \leq 16\beta s_j/\phi_0$ and $\|\bar{\boldsymbol{\Delta}}_j - \boldsymbol{\Delta}_j^*\|_2/\sqrt{nT} \leq 2\beta\sqrt{s_j/\phi_0}$. Recall that $\beta = 2\log(T)\log(nTp)n^{-1/6}T^{-1/3}$, and thus we have

$$
\zeta N = \left\|\bar{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*\right\|_1 + \frac{1}{\sqrt{nT}}\left\|\bar{\boldsymbol{\Delta}}_j - \boldsymbol{\Delta}_j^*\right\|_2 \leq c_1 s_j\log(T)\log(nTp)\,n^{-\frac{1}{6}}T^{-\frac{1}{3}},
\tag{42}
$$

with probability at least $1-[n(T-1)\sqrt{\log\{n(T-1)\}}]^{-1} - 3\exp\{-i_0\} - 2/\{T\sqrt{\log(nT)}\} - 4/(nTp) - \exp(-\min[\log^2(T)/\{2\sigma_m^\epsilon\sqrt{\mu_{\max}^2+(\sigma_m^X)^2}\}^2, \log(T)/|2\sigma_m^\epsilon\sqrt{\mu_{\max}^2+(\sigma_m^X)^2}|]/2)$ and $c_1 = (32+4\sqrt{\phi_0})/\phi_0$.

Case 2: we first derive the upper bound of $\|\bar{\boldsymbol{\Delta}}_j - \boldsymbol{\Delta}_j^*\|_2/\sqrt{nT}$. From the condition of case (2), we have

$$
\begin{aligned}
\frac{3\beta}{2}\left\|\bar{\boldsymbol{\omega}}_j^{\mathcal{S}_j} - (\boldsymbol{\omega}_j^*)^{\mathcal{S}_j}\right\|_1 \leq{}& \frac{12\sigma_m\left\{\sqrt{i_0\log(2pn)} + \sqrt{\log(2pnT)} + B(n,T)\right\}}{nT}\left\|\bar{\boldsymbol{\Delta}}_j - \boldsymbol{\Delta}_j^*\right\|_2 \\
&+ \frac{24\sigma_m D'}{nT}\left\|\left(\mathbf{I}_n\otimes\mathbf{C}\right)\boldsymbol{\Delta}_j^*\right\|_1.
\end{aligned}
\tag{43}
$$

From (38), we obtain

$$\frac{1}{2nT}\left\|\bar{\boldsymbol{\Delta}}_j-\boldsymbol{\Delta}_j^*\right\|_2^2 \leq \frac{3\beta}{2}\left\|\bar{\boldsymbol{\omega}}_j^{\mathcal{S}_j}-(\boldsymbol{\omega}_j^*)^{\mathcal{S}_j}\right\|_1+\frac{2\sigma_m\left\{\sqrt{i_0\log(2pn)}+\sqrt{\log(2pnT)}+B(n,T)\right\}}{nT}\left\|\bar{\boldsymbol{\Delta}}_j-\boldsymbol{\Delta}_j^*\right\|_2 + \frac{4\sigma_m D'}{nT}\left\|(\mathbf{I}_n\otimes\mathbf{C})\boldsymbol{\Delta}_j^*\right\|_1. \tag{44}$$

Substituting (43) into (44), we have

$$\left\|\bar{\boldsymbol{\Delta}}_j-\boldsymbol{\Delta}_j^*\right\|_2^2 \leq 28\sigma_m\left\{\sqrt{i_0\log(2pn)}+\sqrt{\log(2pnT)}+B(n,T)\right\}\left\|\bar{\boldsymbol{\Delta}}_j-\boldsymbol{\Delta}_j^*\right\|_2+56\sigma_m D'\left\|(\mathbf{I}_n\otimes\mathbf{C})\boldsymbol{\Delta}_j^*\right\|_1. \tag{45}$$

Let $x=\|\bar{\boldsymbol{\Delta}}_j-\boldsymbol{\Delta}_j^*\|_2$, $b=28\sigma_m\{\sqrt{i_0\log(2pn)}+\sqrt{\log(2pnT)}+B(n,T)\}$ and $c=56\sigma_m D'\|(\mathbf{I}_n\otimes\mathbf{C})\boldsymbol{\Delta}_j^*\|_1$. Then (45) can be rewritten as $x^2-bx-c\leq 0$. Since x is bounded by the larger root of $x^2-bx-c\leq 0$, we have

$$x\leq\frac{b+\sqrt{b^2+4c}}{2}\leq\frac{b+\sqrt{b^2}+\sqrt{4c}}{2}\leq b+\sqrt{c}.$$

Thus, the upper bound for $\|\bar{\boldsymbol{\Delta}}_j-\boldsymbol{\Delta}_j^*\|_2/\sqrt{nT}$ takes the form

$$\frac{1}{\sqrt{nT}}\left\|\bar{\boldsymbol{\Delta}}_j-\boldsymbol{\Delta}_j^*\right\|_2\leq\frac{28\sigma_m\left\{\sqrt{i_0\log(2pn)}+\sqrt{\log(2pnT)}+B(n,T)\right\}}{\sqrt{nT}}+\frac{\sqrt{56\sigma_m D'\left\|(\mathbf{I}_n\otimes\mathbf{C})\boldsymbol{\Delta}_j^*\right\|_1}}{\sqrt{nT}}. \tag{46}$$

Next, we derive the upper bound for $\|\bar{\boldsymbol{\omega}}_j-\boldsymbol{\omega}_j^*\|_1$ under case (2). Recall $\sigma_m\geq 1$, we obtain

$$\begin{aligned}\left\|\bar{\boldsymbol{\omega}}_j-\boldsymbol{\omega}_j^*\right\|_1 &< \frac{4}{\beta}\left[\frac{2\sigma_m\left\{\sqrt{i_0\log(2pn)}+\sqrt{\log(2pnT)}+B(n,T)\right\}}{nT}\left\|\bar{\boldsymbol{\Delta}}_j-\boldsymbol{\Delta}_j^*\right\|_2+\frac{4\sigma_m D'}{nT}\left\|(\mathbf{I}_n\otimes\mathbf{C})\boldsymbol{\Delta}_j^*\right\|_1\right]\\ &\leq\frac{224\sigma_m^2}{\beta nT}\left\{\sqrt{i_0\log(2pn)}+\sqrt{\log(2pnT)}+B(n,T)+\sqrt{D'\left\|(\mathbf{I}_n\otimes\mathbf{C})\boldsymbol{\Delta}_j^*\right\|_1}\right\}^2,\end{aligned} \tag{47}$$

where the first inequality follows the assumption of Case (2) and the last inequality follows from (46). Let $\Delta_{\max}=\max_{i,t,j}\ |\Delta_{itj}^*-\Delta_{i(t-1)j}^*|+1$ and assume that $\boldsymbol{\Delta}_{ij}^*$ are piecewise constants with at most τ different constants across the T replicates for each subject. Thus, $\|(\mathbf{I}_n\otimes\mathbf{C})\boldsymbol{\Delta}_j^*\|_1\leq\Delta_{\max}\tau n$. Combining (46) and (47), we have

$$\begin{aligned}\zeta N &= \left\|\bar{\boldsymbol{\omega}}_j-\boldsymbol{\omega}_j^*\right\|_1+\frac{1}{\sqrt{nT}}\left\|\bar{\boldsymbol{\Delta}}_j-\boldsymbol{\Delta}_j^*\right\|_2\\ &\leq\frac{224\sigma_m^2}{\beta nT}\left\{\sqrt{i_0\log(2pn)}+\sqrt{\log(2pnT)}+B(n,T)+\sqrt{D'\Delta_{\max}\tau n}\right\}^2\\ &+\frac{112\sigma_m}{\sqrt{nT}}\left\{\sqrt{i_0\log(2pn)}+\sqrt{\log(2pnT)}+B(n,T)+\sqrt{D'\Delta_{\max}\tau n}\right\},\end{aligned} \tag{48}$$

where $D^{'} = 8\sqrt{T\log(2pnT)/(\pi^2 i_0)}$. Since (48) holds for any value of i_0, next, we identify i_0 such that the upper bound ζN at (48) is tight. For notational convenience, let $h = i_0^{1/4}$ and $z(h) = \sqrt{\log(2pn)}h^2 + 2c_4\tau^{1/2}\log^{1/4}(2pnT)T^{1/4}n^{1/2}h^{-1} + \sqrt{\log(2pnT)} + B(n,T)$, where $c_4 = [2\Delta_{\max}/\pi]^{1/2}$. Then, (48) can be rewritten as

$$\zeta N \leq f(z) = \frac{224\sigma_m^2}{\beta nT}z^2 + \frac{112\sigma_m}{\sqrt{nT}}z.$$

The fact that $f^{'}(z) = 448\sigma_m^2 z/(\beta nT) + 112\sigma_m/\sqrt{nT} > 0$ implies that $f(z)$ is an increasing function of z, and thus it suffices to find the of h such that the value of $z(h)$ is minimized. Since $z^{''}(h) = 2\sqrt{\log(2pn)} + 4c_4\tau^{1/2}\log^{1/4}(2pnT)T^{1/4}n^{1/2}h^{-3} > 0$, z is a strictly convex function of h. It can be shown that the minimum of $z(h)$ is achieved when $h = \{c_4\tau^{1/2}\log^{1/4}(2pnT)T^{1/4}n^{1/2}/\sqrt{\log(2pn)}\}^{1/3}$. Since $i_0 \in \{1, 2, \ldots, T-1\}$, we need to carefully select the value of i_0 on its range. When $T > \tau\log^{1/2}(2pnT)c_4^2 n/\log(2pn)$, we have $i_0 = \max\{1, \lfloor(c_4\tau^{1/2}\log^{1/4}(2pnT)T^{1/4}n^{1/2}/\sqrt{\log(2pn)})^{4/3}\rfloor\} \leq T-1$. Thus, it can be shown that

$$\begin{aligned}
&\sqrt{i_0\log(2pn)} + \sqrt{\log(2pnT)} + B(n,T) + \sqrt{D'\Delta_{\max}\tau n} \\
&\leq 4c_4^{\frac{2}{3}}\tau^{\frac{1}{3}}\log^{\frac{1}{3}}(2pnT)\, n^{\frac{1}{3}}T^{\frac{1}{6}} + 4\sqrt{\log(2pnT)} + 4B(n,T).
\end{aligned} \tag{49}$$

Recall that $B(n,T) = \mu_{\max}\min(c_4^{2/3}\tau^{1/3}\log^{1/3}(2pnT)n^{1/3}T^{1/6}, \sqrt{T})$. Then the upper bound for ζN at (48) is

$$\begin{aligned}
\min f(z) \leq{}& \frac{224\sigma_m^2}{\beta nT}\left\{4(\mu_{\max}+1)c_4^{\frac{2}{3}}\tau^{\frac{1}{3}}\log^{\frac{1}{3}}(2pnT)n^{\frac{1}{3}}T^{\frac{1}{6}} + 4\sqrt{\log(2pnT)}\right\}^2 \\
&+ \frac{112\sigma_m}{\sqrt{nT}}\left\{4(\mu_{\max}+1)c_4^{\frac{2}{3}}\tau^{\frac{1}{3}}\log^{\frac{1}{3}}(2pnT)n^{\frac{1}{3}}T^{\frac{1}{6}} + 4\sqrt{\log(2pnT)}\right\}.
\end{aligned} \tag{50}$$

Recall that $\beta = 2\log(T)\log(nTp)n^{-1/6}T^{-1/3}$, then (50) can be written as

$$\begin{aligned}
&\min f(z) \\
&\leq \left\{c_2\tau^{\frac{2}{3}}\log^{\frac{1}{3}}(2pnT)\log^{-1}(T)\log^{-\frac{2}{3}}(nTp) + c_5\tau^{\frac{1}{3}}\log^{\frac{1}{3}}(2pnT)\right\}n^{-\frac{1}{6}}T^{-\frac{1}{3}} \\
&+ \left\{448\sigma_m\sqrt{\log(2pnT)} + 1792\sigma_m^2\log(2pnT)\log^{-1}(T)\log^{-1}(nTp)\right\}n^{-\frac{1}{2}}T^{-\frac{1}{2}} \\
&+ \left\{c_3\tau^{\frac{1}{3}}\log^{-1}(T)\log^{-\frac{5}{6}}(nTp)\log^{\frac{2}{3}}(2pnT)\right\}n^{-\frac{1}{2}}T^{-\frac{1}{2}},
\end{aligned} \tag{51}$$

where $c_2 = 1792\sigma_m^2 c_4^{4/3}(\mu_{\max}+1)^2$, $c_3 = 3584c_4^{2/3}\sigma_m^2(\mu_{\max}+1)$, $c_5 = 448\sigma_m(\mu_{\max}+1)c_4^{2/3}$.

Recall the definition of M. Combining the upper bound for ζN in (42) and (51) yields

$$\zeta N \leq \frac{M}{2}. \tag{52}$$

Since $\zeta = M/(M+N)$, by (52), we obtain $N \leq M$. Recall that $N = \|\widehat{\boldsymbol{\theta}}_{j,-j} - \boldsymbol{\theta}^*_{j,-j}\|_1 + \|\widehat{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}^*_j\|_1 + \|\widehat{\boldsymbol{\Delta}}_j - \boldsymbol{\Delta}^*_j\|_2/\sqrt{nT}$, then

$$
\begin{aligned}
&\big\|\widehat{\boldsymbol{\theta}}_{j,-j} - \boldsymbol{\theta}^*_{j,-j}\big\|_1 + \big\|\widehat{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}^*_j\big\|_1 + \frac{1}{\sqrt{nT}}\big\|\widehat{\boldsymbol{\Delta}}_j - \boldsymbol{\Delta}^*_j\big\|_2 \\
&\leq M \\
&\leq 2\max[c_1 n^{-\frac{1}{6}} T^{-\frac{1}{3}} s_j \log(T)\log(nTp), \\
&\quad n^{-\frac{1}{6}} T^{-\frac{1}{3}} \Big\{c_2 \tau^{\frac{2}{3}} \log^{-1}(T)\log^{-\frac{1}{3}}(nTp) + c_5 \tau^{\frac{1}{3}} \log^{\frac{1}{3}}(2pnT)\Big\} \\
&\quad + n^{-\frac{1}{2}} T^{-\frac{1}{2}} \Big\{448\sigma_m \sqrt{\log(2pnT)} + 3584\sigma_m^2 \log^{-1}(T) + c_3 \tau^{\frac{1}{3}} \log^{-1}(T)\log^{-\frac{1}{6}}(nTp)\Big\}],
\end{aligned}
$$

with probability at least $1-[n(T-1)\sqrt{\log\{n(T-1)\}}]^{-1} - 3\exp[-\lfloor\{\log(2pnT)\Delta^2_{\max}\tau^2 T n^2/\pi^2\}^{1/3}\rfloor] - 2/\{T\sqrt{\log(nT)}\} - 4/(nTp) - 2\exp(-\min[\log^2(T)/\{2\sigma^\epsilon_m\sqrt{\mu^2_{\max} + (\sigma^X_m)^2}\}^2, \log(T)/|2\sigma^\epsilon_m\sqrt{\mu^2_{\max} + (\sigma^X_m)^2}|]/2)$, as desired.

F.2.2 Proof of Theorem 9

The proof of Theorem 9 is similar to the proof of Theorem 8. In particular, let

$$
N = \big\|\widehat{\boldsymbol{\theta}}_{j,-j} - \boldsymbol{\theta}^*_{j,-j}\big\|_1 + \big\|\widehat{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}^*_j\big\|_1 + \frac{1}{\sqrt{nT}}\big\|\widehat{\boldsymbol{\Delta}}_j - \boldsymbol{\Delta}^*_j\big\|_2.
$$

The goal is to show that $N \leq M'$, where

$$
M' = 2\max\Big\{c_1' \log(T)\log(nTp)s_j, c_2'\tau\log^{-1}(2np)\log^{-1}(T)\log^{\frac{1}{2}}(2nTp) + c_3'\tau^{\frac{1}{2}}\log^{\frac{3}{4}}(2npT)\log^{-\frac{1}{2}}(2np)\Big\}T^{-\frac{1}{2}},
$$

where $c_1' = 4(8+2\sqrt{\phi_0})/\phi_0$, $c_2' = 3584\sigma^2_m\Delta_{\max}(\mu_{\max}+3)^2$ and $c_3' = 448\Delta^{1/2}_{\max}\sigma_m(\mu_{\max}+4)$. Similar to the proof of Theorem 8, we require $M' \leq 1$. Thus, we assume the condition $T^{\frac{1}{2}} \geq 2\max\left(c_1', c_2'\right)$.

The proof is similar to that of Theorem 8 with the main difference being the choice of β, λ, and i_0. First, we choose $\beta = \lambda = 2\log(T)\log(nTp)T^{-1/2}$ to obtain the optimal upper bound of ζN, then (42) will reduce to

$$
\zeta N \leq c_1' \log(T)\log(nTp)s_j T^{-\frac{1}{2}}, \tag{53}
$$

where $c_1' = 4(8+2\sqrt{\phi_0})/\phi_0$. Besides, under the condition $T \leq 2\log^{1/2}(2npT)\Delta_{\max}\tau n/\{\pi\log(2np)\}$, we choose $i_0 = T-1$, then (49) can be rewritten as

$$
\sqrt{i_0\log(2pn)} + \sqrt{\log(2pnT)} + B(n,T) + \sqrt{D'\Delta_{\max}\tau n} \leq (\mu_{\max}+2)\sqrt{T\log(2pnT)} + 4\sqrt{\Delta_{\max}\tau n\log^{\frac{1}{2}}(2npT)}. \tag{54}
$$

Thus, the upper bound for ζN in (48) will be

$$
\zeta N \leq \Big\{c_2'\tau\log^{-1}(2np)\log^{-1}(T)\log^{\frac{1}{2}}(2nTp) + c_3'\tau^{\frac{1}{2}}\log^{\frac{3}{4}}(2npT)\log^{-\frac{1}{2}}(2np)\Big\}T^{-\frac{1}{2}}, \tag{55}
$$

where $c_2' = 3584\sigma^2_m\Delta_{\max}(\mu_{\max}+3)^2$ and $c_3' = 448\Delta^{1/2}_{\max}\sigma_m(\mu_{\max}+4)$.

Combining (53) and (55), the upper bound for ζN is

$$\begin{aligned}\zeta N &\leq \max\left\{c_1^{'}\log(T)\log(nTp)s_j, c_2^{'}\tau\log^{-1}(2np)\log^{-1}(T)\log^{\frac{1}{2}}(2nTp) + c_3^{'}\tau^{\frac{1}{2}}\log^{\frac{3}{4}}(2npT)\log^{-\frac{1}{2}}(2np)\right\}T^{-\frac{1}{2}} \\ &\leq \frac{M^{'}}{2}. \end{aligned} \tag{56}$$

Recall the definition of ζ and N, we can obtain that

$$\begin{aligned}&\left\|\widehat{\boldsymbol{\theta}}_{j,-j} - \boldsymbol{\theta}^*_{j,-j}\right\|_1 + \left\|\widehat{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}^*_j\right\|_1 + \frac{1}{\sqrt{nT}}\left\|\widehat{\boldsymbol{\Delta}}_j - \boldsymbol{\Delta}^*_j\right\|_2 \\ &\leq 2\max\left\{c_1^{'}\log(T)\log(nTp)s_j, c_2^{'}\tau\log^{-1}(2np)\log^{-1}(T)\log^{\frac{1}{2}}(2nTp) + c_3^{'}\tau^{\frac{1}{2}}\log^{\frac{3}{4}}(2npT)\log^{-\frac{1}{2}}(2np)\right\}T^{-\frac{1}{2}},\end{aligned} \tag{57}$$

with probability at least $1-[n(T-1)\sqrt{\log\{n(T-1)\}}]^{-1}-\exp\{-(T-1)\}-2/\{T\sqrt{\log(nT)}\}-4/(nTp)-2\exp(-\min[\log^2(T)/\{2\sigma_m^{\epsilon}\sqrt{\mu^2_{\max}+(\sigma_m^X)^2}\}^2, \log(T)/|2\sigma_m^{\epsilon}\sqrt{\mu^2_{\max}+(\sigma_m^X)^2}|]/2)$.

F.3 Proof of Technical Lemmas

F.3.1 Proof of Lemma 19

This proof is similar to the proof of Lemma 6 in Hall et al. (2016). Recall that for $k \neq j$, $X_{itk} \sim N(\mu_{itk}, \sigma^2_{kk,t})$ with $(\sigma_m^X)^2 = \max_{k,t}(\sigma^2_{kk,t})$, and $\epsilon_{itj} \sim N\{0, (\sigma^{\epsilon}_{jj,t})^2\}$ with $(\sigma_m^{\epsilon})^2 = \max_{t,j}\{(\sigma^{\epsilon}_{jj,t})^2\}$. Let $\boldsymbol{\epsilon}_j = (\epsilon_{11j}, \epsilon_{12j}, \ldots, \epsilon_{1Tj}, \epsilon_{21j}, \ldots, \epsilon_{nTj})^{\mathrm{T}}$ and let $\boldsymbol{X}_k = (X_{11k}, X_{12k}, \ldots, X_{1Tk}, X_{21k}, \ldots, X_{nTk})^{\mathrm{T}}$. For simplicity, we rewrite $\boldsymbol{\epsilon}_j$ and $\boldsymbol{X}_k$ as $\boldsymbol{\epsilon}_j = (\epsilon^{'}_{1j}, \epsilon^{'}_{2j}, \ldots, \epsilon^{'}_{(nT)j})^{\mathrm{T}}$ and $\boldsymbol{X}_k = (X^{'}_{1k}, X^{'}_{2k}, \ldots, X^{'}_{(nT)k})^{\mathrm{T}}$, where $\epsilon^{'}_{lj} = \epsilon_{itj}$ and $X^{'}_{lk} = X_{itk}$ with $l = (i-1)T + t$. Then, we have

$$\max_{1\leq k\leq p, k\neq j}\frac{1}{nT}\left|\sum_{i=1}^{n}\sum_{t=1}^{T}\epsilon_{itj}X_{itk}\right| = \max_{1\leq k\leq p, k\neq j}\frac{1}{nT}\left|\sum_{l=1}^{nT}\epsilon^{'}_{lj}X^{'}_{lk}\right|. \tag{58}$$

In this proof, our goal is to bound (58) by Lemma 23. First, we define notation Z_m, G_m^k and R_m needed by Lemma 23. Denote the sequence Z_m as

$$Z_m = \frac{1}{nT}\sum_{l=1}^{m}\epsilon^{'}_{lj}X^{'}_{lk}.$$

Then we have

$$\begin{aligned}\mathbb{E}\left(Z_m - Z_{m-1}\middle|\epsilon^{'}_{1j}, \ldots, \epsilon^{'}_{(m-1)j}, X^{'}_{1k}, \ldots, X^{'}_{(m-1)k}\right) &= \frac{1}{nT}\mathbb{E}\left(\epsilon^{'}_{mj}X^{'}_{mk}\middle|\epsilon^{'}_{1j}, \ldots, \epsilon^{'}_{(m-1)j}, X^{'}_{1k}, \ldots, X^{'}_{(m-1)k}\right) \\ &= \frac{1}{nT}\mathbb{E}\left(\epsilon^{'}_{mj}\right)\mathbb{E}\left(X^{'}_{mk}\middle|\epsilon^{'}_{1j}, \ldots, \epsilon^{'}_{(m-1)j}, X^{'}_{1k}, \ldots, X^{'}_{(m-1)k}\right) \\ &= 0,\end{aligned}$$

where the second equality holds since $\epsilon^{'}_{mj}$ is independent with $\boldsymbol{X}_k$ for $j \neq k$ and $\epsilon^{'}_{mj}$ is independent with $\epsilon^{'}_{lj}$, for $l = 1, \ldots, m-1$. Thus, we conclude that Z_m is a martingale.

Recall that $X_{mk}^{'} \sim N(\mu_{mk}^{'}, \sigma_{kk,m}^2)$. Let $|r| \leq 1/|2\sigma_m^{\epsilon}\sqrt{\mu_{\max}^2 + (\sigma_m^X)^2}|$, then by smoothing we have

$$\begin{aligned}
\mathbb{E}\left(e^{r\epsilon_{mj}^{'}X_{mk}^{'}}\right) &= \mathbb{E}\left\{\mathbb{E}\left(e^{r\epsilon_{mj}^{'}X_{mk}^{'}}\middle|\epsilon_{mj}^{'}\right)\right\} \\
&= \mathbb{E}\left\{e^{r\mu_{mk}^{'}\epsilon_{mj}^{'}+\frac{1}{2}\left(r\sigma_{kk,m}\epsilon_{mj}^{'}\right)^2}\right\} \\
&\leq \sqrt{\mathbb{E}\left(e^{2r\mu_{mk}^{'}\epsilon_{mj}^{'}}\right)}\sqrt{\mathbb{E}\left\{e^{\left(r\sigma_{kk,m}\epsilon_{mj}^{'}\right)^2}\right\}} \\
&= \frac{\exp\left\{2\left(r\mu_{mk}^{'}\sigma_{jj,m}^{\epsilon}\right)^2\right\}}{\sqrt{1-2\left(r\sigma_{jj,m}^{\epsilon}\sigma_{kk,m}\right)^2}} \\
&\leq e^{\frac{1}{2}\left\{2r\sigma_m^{\epsilon}\sqrt{\mu_{\max}^2+(\sigma_m^X)^2}\right\}^2},
\end{aligned} \tag{59}$$

where the second equality holds with $\epsilon_{mj}^{'}$ is independent with $X_{mk}^{'}$ and the third equality holds with $\epsilon_{mj}^{'} \sim N\{0, (\sigma_{jj,m}^{\epsilon})^2\}$ and $(\epsilon_{mj}^{'}/\sigma_{jj,m}^{\epsilon})^2 \sim \chi_1^2$. Therefore, we obtain that $\epsilon_{mj}^{'}X_{mk}^{'}$ follows sub-exponential with parameter $|2\sigma_m^{\epsilon}\sqrt{\mu_{\max}^2 + (\sigma_m^X)^2}|$, which we denote as $\epsilon_{mj}^{'}X_{mk}^{'} \sim \text{subE}(|2\sigma_m^{\epsilon}\sqrt{\mu_{\max}^2 + (\sigma_m^X)^2}|)$. By Lemma 24, we have

$$\mathbb{P}\left\{\left|\epsilon_{mj}^{'}X_{mk}^{'}\right| \geq \log(T)\right\} \leq \exp\left(-\frac{1}{2}\min\left[\frac{\log^2(T)}{\left\{2\sigma_m^{\epsilon}\sqrt{\mu_{\max}^2+(\sigma_m^X)^2}\right\}^2}, \frac{\log(T)}{\left|2\sigma_m^{\epsilon}\sqrt{\mu_{\max}^2+(\sigma_m^X)^2}\right|}\right]\right). \tag{60}$$

Let $B = \log(T)/(nT)$ and define sequence G_m^k as

$$G_m^k = \sum_{l=1}^{m}\mathbb{E}\left\{\left(\frac{1}{nT}\epsilon_{lj}^{'}X_{lk}^{'}\right)^k\middle|\epsilon_{1j}^{'},\ldots,\epsilon_{(l-1)j}^{'},X_{1k}^{'},\ldots,X_{(l-1)k}^{'}\right\} \leq |G_m^k| \leq mB^k, \tag{61}$$

with probability at least $1-\exp(-\min[\log^2(T)/\{2\sigma_m^{\epsilon}\sqrt{\mu_{\max}^2+(\sigma_m^X)^2}\}^2, \log(T)/|2\sigma_m^{\epsilon}\sqrt{\mu_{\max}^2+(\sigma_m^X)^2}|]/2)$. The inequality in (61) follows (60). Then for $\rho > 0$, let

$$R_m = \sum_{k=2}^{\infty}\frac{\rho^k G_m^k}{k!}; \quad R_m^{'} = \sum_{k=2}^{\infty}\frac{(-1)^k\rho^k G_m^k}{k!}; \quad R_m^{''} = m\left(e^{\rho B}-1-\rho B\right).$$

Besides, we have

$$R_m^{'} \text{ and } R_m \leq m\sum_{k=2}^{\infty}\frac{(\rho B)^k}{k!} = R_m^{''}, \tag{62}$$

where the second inequality follows (61). The upper bound of $|Z_m|$ is

$$\begin{aligned}\mathbb{P}(|Z_m| \geq z) &= \mathbb{P}(Z_m \geq z) + \mathbb{P}(-Z_m \geq z) \\ &\leq \mathbb{E}\left(e^{\rho Z_m}\right) e^{-\rho z} + \mathbb{E}\left(e^{-\rho Z_m}\right) e^{-\rho z} \\ &= \mathbb{E}\left(e^{\rho Z_m - R_m + R_m}\right) e^{-\rho z} + \mathbb{E}\left(e^{-\rho Z_m - R_m^{'} + R_m^{'}}\right) e^{-\rho z} \\ &\leq \mathbb{E}\left(e^{\rho Z_m - R_m}\right) e^{R_m^{''} - \rho z} + \mathbb{E}\left(e^{-\rho Z_m - R_m^{'}}\right) e^{R_m^{''} - \rho z} + p_0 \\ &\leq 2e^{R_m^{''} - \rho z} + p_0, \end{aligned} \tag{63}$$

with $p_0 = \exp(-\min[\log^2(T)/\{2\sigma_m^{\epsilon}\sqrt{\mu_{\max}^2 + (\sigma_m^X)^2}\}^2, \log(T)/|2\sigma_m^{\epsilon}\sqrt{\mu_{\max}^2 + (\sigma_m^X)^2}|]/2)$. The first inequality follows Markov's inequality, the second inequality follows (62) and the last inequality follows Lemma 23. The next step of this proof is to find the value of ρ to minimize the right hand side of (63). Recall the $R_m^{''} = m(e^{\rho B} - 1 - \rho B)$. Denote the right hand side of (63) as

$$f(\rho) = 2e^{R_m^{''} - \rho z} + p_0 = 2\exp\left\{m\left(e^{\rho B} - 1 - \rho B\right) - \rho z\right\} + p_0.$$

Since $f(\rho)$ is strictly convex, $f(\rho)$ obtains its minimizer at the root of $f^{'}(\rho) = 0$, which is $\rho^* = \log\{z/(mB) + 1\}/B$. Then (63) will be

$$\mathbb{P}\left(\left|Z_m\right| \geq z\right) \leq f(\rho^*) = 2\exp\left\{-mg\left(\frac{z}{mB}\right)\right\} + p_0,$$

where $g(x) = (1+x)\log(1+x) - x$. Since $g(x) \geq 3x^2/\{2(x+3)\}$ for $x \geq 0$, we have

$$\mathbb{P}\left(\left|Z_m\right| \geq z\right) \leq 2\exp\left(-\frac{3z^2}{2zB + 6mB^2}\right) + p_0 = 2\exp\left\{-\frac{3z^2n^2T^2}{2znT\log(T) + 6m\log^2(T)}\right\} + p_0, \tag{64}$$

where the equality follows the fact that $B = \log(T)/(nT)$. Let $m = nT$ and $z = \lambda_0$. By (64) we have

$$\mathbb{P}\left[\max_{1\leq k\leq p, k\neq j}\left|\frac{1}{nT}\sum_{l=1}^{nT}\epsilon_{lj}^{'}X_{lk}^{'}\right| \geq \lambda_0\right] \leq 2(p-1)\exp\left\{-\frac{3\lambda_0^2 nT}{2\lambda_0\log(T) + 6\log^2(T)}\right\} + p_0. \tag{65}$$

Recall p_0, then we have

$$\max_{1\leq k\leq p, k\neq j}\left|\frac{1}{nT}\sum_{l=1}^{nT}\epsilon_{lj}^{'}X_{lk}^{'}\right| < \lambda_0,$$

with probability at least $1 - \exp(-\min[\log^2(T)/\{2\sigma_m^{\epsilon}\sqrt{\mu_{\max}^2 + (\sigma_m^X)^2}\}^2, \log(T)/|2\sigma_m^{\epsilon}\sqrt{\mu_{\max}^2 + (\sigma_m^X)^2}|]/2) - \exp[\log\{2(p-1)\} - 3\lambda_0^2 nT/\{2\lambda_0\log(T) + 6\log^2(T)\}]$.

F.3.2 Proof of Lemma 21

Let $\boldsymbol{\eta} \sim N(\mathbf{0}, \boldsymbol{Q})$. The goal is to obtain an upper bound for $\boldsymbol{\eta}^{\mathrm{T}}(\bar{\boldsymbol{\Delta}}_{ij} - \boldsymbol{\Delta}_{ij})/(nT)$. Recall that $\mathbf{C}$ is the discrete first derivative matrix,

$$\mathbf{C} = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix}.$$

Let $S_c(1) = \{\boldsymbol{W} \in \text{row}(\mathbf{C}) : \|\mathbf{C}\boldsymbol{W}\|_1 \leq 1\}$, where row($\mathbf{C}$) is the row space of $\mathbf{C}$. The steps in this proof are:

1. get the upper bound of $\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{W}$ for $\forall \boldsymbol{W} \in S_c(1)$.
2. substitute $\boldsymbol{W}$ with a specific value related to $\bar{\boldsymbol{\Delta}}_{ij}$ and $\boldsymbol{\Delta}_{ij}$.

Step (i): first, we would like to introduce some new notation. The singular value decomposition of $\mathbf{C}$ is

$$\mathbf{C} = \mathbf{U}\boldsymbol{\Xi}\mathbf{V}^{\mathrm{T}},$$

where both $\mathbf{U} \in \mathbb{R}^{(T-1)\times(T-1)}$ and $\mathbf{V} \in \mathbb{R}^{T\times(T-1)}$ are orthogonal matrixes and $\boldsymbol{\Xi} \in \mathbb{R}^{(T-1)\times(T-1)}$ is a diagonal matrix with diagonal ξ_i, $i \in \{1, 2, \ldots, T-1\}$. Then the pseudoinverse of $\mathbf{C}$ is

$$\mathbf{C}^{+} = \mathbf{V}\boldsymbol{\Xi}^{-1}\mathbf{U}^{\mathrm{T}}.$$

For $i_0 \in \{1, \ldots, T-1\}$, let $[i_0] = \{1, \ldots, i_0\}$ and $\mathbf{P}_{[i_0]} = \mathbf{V}_{[i_0]}\mathbf{V}_{[i_0]}^{\mathrm{T}}$, where $\mathbf{V}_{[i_0]}$ is a matrix containing the first i_0 columns of $\mathbf{V}$. Then $\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{W}$ can be written as

$$\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{W} = \underbrace{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{P}_{[i_0]}\boldsymbol{W}}_{\mathbb{I}_1} + \underbrace{\boldsymbol{\eta}^{\mathrm{T}}\left(\mathbf{I} - \mathbf{P}_{[i_0]}\right)\boldsymbol{W}}_{\mathbb{I}_2}. \tag{66}$$

To bound the term $\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{W}$, we would consider $\mathbb{I}_1$ and $\mathbb{I}_2$ separately. Upper Bound for $\mathbb{I}_1$ in (66): by Holder's inequality, we have

$$\mathbb{I}_1 \leq \left\|\mathbf{V}_{[i_0]}^{\mathrm{T}}\boldsymbol{\eta}\right\|_2 \cdot \left\|\mathbf{V}_{[i_0]}^{\mathrm{T}}\boldsymbol{W}\right\|_2. \tag{67}$$

Now, we would like to further bound the term $\|\mathbf{V}_{[i_0]}^{\mathrm{T}}\boldsymbol{\eta}\|_2$ in (67). Recall $\boldsymbol{\eta} \sim N(\mathbf{0}, \mathbf{Q})$. Since $\mathbf{V}_{[i_0]}^{\mathrm{T}}\boldsymbol{\eta} \sim N(\mathbf{0}, \mathbf{V}_{[i_0]}^{\mathrm{T}}\mathbf{Q}\mathbf{V}_{[i_0]})$, $\|\mathbf{V}_{[i_0]}^{\mathrm{T}}\boldsymbol{\eta}\|_2$ can be written as

$$\left\|\mathbf{V}_{[i_0]}^{\mathrm{T}}\boldsymbol{\eta}\right\|_2^2 = \mathbf{Z}^{\mathrm{T}}\mathbf{V}_{[i_0]}^{\mathrm{T}}\mathbf{Q}\mathbf{V}_{[i_0]}\mathbf{Z},$$

where $\mathbf{Z} \sim N(\mathbf{0}, \mathbf{I})$. By Lemma 25, we have

$$\begin{aligned}
&\mathbb{P}\left[\left\|\mathbf{V}_{[i_0]}^{\mathrm{T}}\boldsymbol{\eta}\right\|_2^2 - \mathbb{E}\left(\left\|\mathbf{V}_{[i_0]}^{\mathrm{T}}\boldsymbol{\eta}\right\|_2^2\right) > \{i_0 + \log(n)\}\,\nu\right] \\
&\leq \mathbb{P}\left[\left|\mathbf{Z}^{\mathrm{T}}\mathbf{V}_{[i_0]}^{\mathrm{T}}\mathbf{Q}\mathbf{V}_{[i_0]}\mathbf{Z} - \mathbb{E}\mathbf{Z}^{\mathrm{T}}\mathbf{V}_{[i_0]}^{\mathrm{T}}\mathbf{Q}\mathbf{V}_{[i_0]}\mathbf{Z}\right| > \{i_0 + \log(n)\}\,\nu\right] \\
&\leq 2\exp\left(-\min\left[\frac{\{i_0 + \log(n)\}^2\,\nu^2}{\left\|\mathbf{V}_{[i_0]}^{\mathrm{T}}\mathbf{Q}\mathbf{V}_{[i_0]}\right\|_{\mathrm{F}}^2}, \frac{\{i_0 + \log(n)\}\,\nu}{\left\|\mathbf{V}_{[i_0]}^{\mathrm{T}}\mathbf{Q}\mathbf{V}_{[i_0]}\right\|_{\mathrm{op}}}\right]\right),
\end{aligned} \tag{68}$$

where $\|\cdot\|_F$ is the Frobenius norm and $\|\cdot\|_{\mathrm{op}}$ is the operator norm. Set $\nu = \|\mathbf{Q}\|_{\mathrm{op}}$. Since $\mathbb{E}(\|\mathbf{V}_{[i_0]}\boldsymbol{\eta}\|_2^2) = \mathrm{tr}(\mathbf{V}_{[i_0]}^{\mathrm{T}}\mathbf{Q}\mathbf{V}_{[i_0]}) \leq i_0\|\mathbf{V}_{[i_0]}^{\mathrm{T}}\mathbf{Q}\mathbf{V}_{[i_0]}\|_{\mathrm{op}}$, $\|\mathbf{V}_{[i_0]}^{\mathrm{T}}\mathbf{Q}\mathbf{V}_{[i_0]}\|_{\mathrm{F}}^2 \leq i_0\|\mathbf{V}_{[i_0]}^{\mathrm{T}}\mathbf{Q}\mathbf{V}_{[i_0]}\|_{\mathrm{op}}^2$ and $\|\mathbf{V}_{[i_0]}^{\mathrm{T}}\mathbf{Q}\mathbf{V}_{[i_0]}\|_{\mathrm{op}}^2 \leq \|\mathbf{Q}\|_{\mathrm{op}}^2$, (68) will be

$$\mathbb{P}\left[\left\|\mathbf{V}_{[i_0]}^{\mathrm{T}}\boldsymbol{\eta}\right\|_2^2 \geq 2\,\{i_0 + \log(n)\}\left\|\mathbf{Q}\right\|_{\mathrm{op}}\right] \leq 2\exp\left(-i_0\right)/n. \tag{69}$$

Substituting (69) into (67), we have

$$\mathbb{I}_1 \leq \sqrt{2\left\{i_0+\log(n)\right\}\left\|\mathbf{Q}\right\|_{\text{op}}}\left\|\mathbf{V}_{[i_0]}^{\text{T}}\boldsymbol{W}\right\|_2 \leq \sqrt{2\left\{i_0+\log(n)\right\}\left\|\mathbf{Q}\right\|_{\text{op}}}\left\|\boldsymbol{W}\right\|_2, \tag{70}$$

with probability at least $1-2\exp(-i_0)/n$.

Upper Bound for $\mathbb{I}_2$ in (66): Recall that $\text{row}(\mathbf{C})$ is the row space of $\mathbf{C}$. Let $\mathbf{P}_{\text{row}(\mathbf{C})} = \mathbf{C}^{+}\mathbf{C}$ be the projection onto $\text{row}(\mathbf{C})$, and for $\forall \boldsymbol{W} \in S_c(1)$, $\exists$ vector $\boldsymbol{L}$ that $\boldsymbol{W} = \mathbf{P}_{\text{row}(\mathbf{C})}\boldsymbol{L}$. Then we have

$$\mathbb{I}_2 = \boldsymbol{\eta}^{\text{T}}\left(\mathbf{I}-\mathbf{P}_{[i_0]}\right)\mathbf{P}_{\text{row}(\mathbf{C})}\boldsymbol{L} \leq \left\|\boldsymbol{\eta}^{\text{T}}\left(\mathbf{I}-\mathbf{P}_{[i_0]}\right)\mathbf{C}^{+}\right\|_{\infty}\cdot\left\|\mathbf{C}\boldsymbol{L}\right\|_1 \leq \left\|\boldsymbol{\eta}^{\text{T}}\left(\mathbf{I}-\mathbf{P}_{[i_0]}\right)\mathbf{C}^{+}\right\|_{\infty}, \tag{71}$$

where the first inequality holds with Holder's inequality and the second inequality holds with the fact that $\|\mathbf{C}\boldsymbol{L}\|_1 = \|\mathbf{C}\mathbf{C}^{+}\mathbf{C}\boldsymbol{W}\|_1 = \|\mathbf{C}\boldsymbol{W}\|_1 \leq 1$. To further bound $\mathbb{I}_2$, let $\boldsymbol{e}_j$ be the jth canonical basis vector and $\boldsymbol{g}_j = (\mathbf{I}-\mathbf{P}_{[i_0]})\mathbf{C}^{+}\boldsymbol{e}_j$. let $\boldsymbol{u}_j = (u_{j1}, u_{j2}, \ldots, u_{j(T-1)})^{\text{T}}$, $j = 1, \ldots, T-1$, as the jth column of $\mathbf{U}$, then we obtain

$$\left\|\boldsymbol{g}_j\right\|_2^2 = \left\|\left[\mathbf{0}, \mathbf{V}_{[T-1]\setminus[i_0]}\right]\boldsymbol{\Xi}^{-1}\mathbf{U}^{\text{T}}\boldsymbol{e}_j\right\|_2^2 = \sum_{i=i_0+1}^{T-1}\frac{u_{ji}^2}{\xi_i^2}, \tag{72}$$

where $[\mathbf{0}, \mathbf{V}_{[T-1]\setminus[i_0]}]$ can be obtained by substituting first i_0 columns of $\mathbf{V}$ with $\mathbf{0}$. By relating $\mathbf{C}$ with finite difference operator, Wang et al. (2016) shows that $u_{ij} = \sqrt{2/T}\sin(\pi ij/T)$ and $\xi_i = 2\sin\{\pi(i-1)/(2T)\}$. Then the upper bound for $\|\boldsymbol{g}_j\|_2^2$ is

$$\begin{aligned}
\sum_{i=i_0+1}^{T-1}\frac{u_{ji}^2}{\xi_i^2} &\leq \frac{2}{T}\sum_{i=i_0+1}^{T-1}\frac{1}{\xi_i^2} \\
&= \frac{2}{T}\sum_{i=i_0+1}^{T-1}\frac{1}{4\sin^2\left(\pi\left(i-1\right)/\left(2T\right)\right)} \\
&\leq 2\int_{(i_0-1)/T}^{(T-2)/T}\frac{1}{4\sin^2\left(\pi x/2\right)}dx \\
&= \frac{\cot\left\{\pi\left(i_0-1\right)/\left(2T\right)\right\}}{\pi} \\
&\leq \frac{4T}{\pi^2 i_0},
\end{aligned} \tag{73}$$

where the first equality holds by $\sin(\pi ij/T) \leq 1$ and the last inequality holds by $\cot(x) \leq 1/x$ and $i_0/(i_0-1) \leq 2$. Recall that $\boldsymbol{\eta} \sim N(\mathbf{0}, \mathbf{Q})$, then $\boldsymbol{g}_j^{\text{T}}\boldsymbol{\eta} \sim N(\mathbf{0}, \boldsymbol{g}_j^{\text{T}}\mathbf{Q}\boldsymbol{g}_j)$. Since $\|\boldsymbol{\eta}^{\text{T}}(\mathbf{I}-\mathbf{P}_{[i_0]})\mathbf{C}^{+}\|_{\infty} = \max_{1\leq j\leq T-1}|\boldsymbol{\eta}^{\text{T}}\boldsymbol{g}_j| = \max_{1\leq j\leq T-1}|\boldsymbol{g}_j^{\text{T}}\boldsymbol{\eta}|$,

$$\begin{aligned}
P\left[\max_{1\leq j\leq T-1}\left|\boldsymbol{g}_j^{\text{T}}\boldsymbol{\eta}\right| > 4\sqrt{\frac{\left\|\mathbf{Q}\right\|_{\text{op}}T\log\left\{n(T-1)\right\}}{\pi^2 i_0}}\right] &\leq \sum_{j=1}^{T-1}P\left[\left|\boldsymbol{g}_j^{\text{T}}\boldsymbol{\eta}\right| > 2\sqrt{\left\|\mathbf{Q}\right\|_{\text{op}}\left\|\boldsymbol{g}_j\right\|_2^2\log\left\{n(T-1)\right\}}\right] \\
&\leq \sum_{j=1}^{T-1}P\left[\left|\boldsymbol{g}_j^{\text{T}}\boldsymbol{\eta}\right| > 2\sqrt{\boldsymbol{g}_j^{\text{T}}\mathbf{Q}\boldsymbol{g}_j\log\left\{n(T-1)\right\}}\right] \\
&\leq \frac{1}{2n^2(T-1)\sqrt{\log\left\{n(T-1)\right\}}},
\end{aligned} \tag{74}$$

where the first inequality follows that $\boldsymbol{g}_j^{\mathrm{T}}\mathbf{Q}\boldsymbol{g}_j = \|\mathbf{Q}^{\frac{1}{2}}\boldsymbol{g}_j\|_2^2 \le \|\mathbf{Q}\|_{\mathrm{op}}\|\boldsymbol{g}_j\|_2^2$ and the last inequality follows Lemma 26.

Let $D = 8\sqrt{T\log(nT)/(\pi^2 i_0)}$, then the upper bound for $\mathbb{I}_2$ is

$$\mathbb{I}_2 \le \sqrt{\|\mathbf{Q}\|_{\mathrm{op}}}D, \tag{75}$$

with probability at least $1 - 1/[2n^2(T-1)\sqrt{\log\{n(T-1)\}}]$. Substitute (70) and (75) into (66), we have

$$\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{W} \le \sqrt{2\{i_0 + \log(n)\}\|\mathbf{Q}\|_{\mathrm{op}}}\|\boldsymbol{W}\|_2 + \sqrt{\|\mathbf{Q}\|_{\mathrm{op}}}D, \tag{76}$$

with probability at least $1 - [2n^2(T-1)\sqrt{\log\{n(T-1)\}}]^{-1} - 2\exp(-i_0)/n$.

Step (ii): since

$$\frac{1}{nT}\boldsymbol{\eta}^{\mathrm{T}}\left(\bar{\boldsymbol{\Delta}}_{ij} - \boldsymbol{\Delta}_{ij}\right) = \underbrace{\frac{1}{nT}\boldsymbol{\eta}^{\mathrm{T}}\mathbf{P}_{\mathrm{row}(\mathbf{C})}\left(\bar{\boldsymbol{\Delta}}_{ij} - \boldsymbol{\Delta}_{ij}\right)}_{\mathbb{I}_3} + \underbrace{\frac{1}{nT}\boldsymbol{\eta}^{\mathrm{T}}\left(\mathbf{I} - \mathbf{P}_{\mathrm{row}(\mathbf{C})}\right)\left(\bar{\boldsymbol{\Delta}}_{ij} - \boldsymbol{\Delta}_{ij}\right)}_{\mathbb{I}_4}. \tag{77}$$

We bound $\mathbb{I}_3$ and $\mathbb{I}_4$ in (77) separately. Upper Bound for $\mathbb{I}_3$ in (77): substituting $\boldsymbol{W} = \mathbf{P}_{\mathrm{row}(\mathbf{C})}(\bar{\boldsymbol{\Delta}}_{ij} - \boldsymbol{\Delta}_{ij})/\|\mathbf{C}(\bar{\boldsymbol{\Delta}}_{ij} - \boldsymbol{\Delta}_{ij})\|_1$ into (76) and applying Holder's inequality several times, we have

$$\begin{aligned}
\mathbb{I}_3 &\le \frac{\sqrt{\|\mathbf{Q}\|_{\mathrm{op}}}D}{nT}\|\mathbf{C}\bar{\boldsymbol{\Delta}}_{ij}\|_1 + \frac{\sqrt{\|\mathbf{Q}\|_{\mathrm{op}}}D}{nT}\|\mathbf{C}\boldsymbol{\Delta}_{ij}\|_1 + \frac{\sqrt{2\{i_0 + \log(n)\}\|\mathbf{Q}\|_{\mathrm{op}}}}{nT}\|\mathbf{C}^{+}\mathbf{C}\|_2 \cdot \|\bar{\boldsymbol{\Delta}}_{ij} - \boldsymbol{\Delta}_{ij}\|_2 \\
&\le \frac{\sqrt{\|\mathbf{Q}\|_{\mathrm{op}}}D}{nT}\|\mathbf{C}\bar{\boldsymbol{\Delta}}_{ij}\|_1 + \frac{\sqrt{\|\mathbf{Q}\|_{\mathrm{op}}}D}{nT}\|\mathbf{C}\boldsymbol{\Delta}_{ij}\|_1 + \frac{\sqrt{2\{i_0 + \log(n)\}\|\mathbf{Q}\|_{\mathrm{op}}}}{nT}\|\bar{\boldsymbol{\Delta}}_{ij} - \boldsymbol{\Delta}_{ij}\|_2,
\end{aligned} \tag{78}$$

with probability at least $1 - [2n^2(T-1)\sqrt{\log\{n(T-1)\}}]^{-1} - 2\exp(-i_0)/n$. The second inequality follows the fact that $\mathbf{C}^{+}\mathbf{C}$ is idempotent.

Upper Bound for $\mathbb{I}_4$ in (77):

$$\mathbb{I}_4 \le \frac{1}{nT}\|\boldsymbol{\eta}^{\mathrm{T}}\left(\mathbf{I} - \mathbf{P}_{\mathrm{row}(\mathbf{C})}\right)\|_2 \cdot \|\bar{\boldsymbol{\Delta}}_{ij} - \boldsymbol{\Delta}_{ij}\|_2 \le \frac{\sqrt{2\|\mathbf{Q}\|_{\mathrm{op}}\log(nT)}}{nT}\|\bar{\boldsymbol{\Delta}}_{ij} - \boldsymbol{\Delta}_{ij}\|_2, \tag{79}$$

with probability at least $1 - 1/\{nT\sqrt{2\log(nT)}\}$. The second inequality is obtained by,

$$\begin{aligned}
\mathbb{P}\left\{\|\boldsymbol{\eta}^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_{\mathrm{row}(\mathbf{C})})\|_2 > \sqrt{2\|\mathbf{Q}\|_{\mathrm{op}}\log(nT)}\right\} &= \mathbb{P}\left\{\frac{|\mathbf{1}^{\mathrm{T}}\boldsymbol{\eta}|}{\sqrt{T}} > \sqrt{2\|\mathbf{Q}\|_{\mathrm{op}}\log(nT)}\right\} \\
&= \mathbb{P}\left\{|\mathbf{1}^{\mathrm{T}}\boldsymbol{\eta}| > \sqrt{2\|\mathbf{1}\|_2^2\|\mathbf{Q}\|_{\mathrm{op}}\log(nT)}\right\} \\
&\le \mathbb{P}\left\{|\mathbf{1}^{\mathrm{T}}\boldsymbol{\eta}| > \sqrt{2\mathbf{1}^{\mathrm{T}}\mathbf{Q}\mathbf{1}\log(nT)}\right\} \\
&\le \frac{1}{nT\sqrt{2\log(nT)}}.
\end{aligned}$$

Substituting (78) and (79) into (77), we have the desired conclusion

$$\frac{1}{nT}\boldsymbol{\eta}^{\mathrm{T}}\left(\bar{\boldsymbol{\Delta}}_{ij}-\boldsymbol{\Delta}_{ij}\right)$$
$$\leq \frac{\sqrt{2\|\mathbf{Q}\|_{\mathrm{op}}}\left\{\sqrt{i_0+\log(n)}+\sqrt{\log(nT)}\right\}}{nT}\|\bar{\boldsymbol{\Delta}}_{ij}-\boldsymbol{\Delta}_{ij}\|_2+\frac{\sqrt{\|\mathbf{Q}\|_{\mathrm{op}}}D}{nT}\left(\|\mathbf{C}\bar{\boldsymbol{\Delta}}_{ij}\|_1+\|\mathbf{C}\boldsymbol{\Delta}_{ij}\|_1\right),$$

with probability at least $1-[2n^2(T-1)\sqrt{\log\{n(T-1)\}}]^{-1}-2\exp(-i_0)/n-1/\{nT\sqrt{2\log(nT)}\}$.

F.3.3 Some Technical Lemmas

In this section, we provide several lemmas that are useful for proving Theorems 8–9 and Lemmas 19–21.

Lemma 23 (Lemma 3.3 in Houdré and Reynaud-Bouret (2003)) *Let $(Z_m, m\in N)$ be a martingale. For all $k\geq 2$, let $G_m^k=\sum_{l=1}^m \mathbb{E}\{(Z_l-Z_{l-1})^k|\mathcal{F}_{l-1}\}$, where $\mathcal{F}_{l-1}$ is the filter containing all the information up to $l-1$. For $\forall\rho>0$, let $R_m=\sum_{k=2}^\infty \rho^k G_m^k/k!$ and $R_m^{'}=\sum_{k=2}^\infty(-1)^k\rho^k G_m^k/k!$. If $Z_0=0$, then*

$$\mathbb{E}\left\{\exp\left(\rho Z_m-R_m\right)\right\}\leq 1;\quad \mathbb{E}\left\{\exp\left(-\rho Z_m-R_m^{'}\right)\right\}\leq 1.$$

Lemma 24 (Bernstein's inequality in Rigollet and Hütter (2015)) *Let $X\sim \mathrm{subE}(\nu)$ and $\mathbb{E}(X)=0$, then for any $t>0$,*

$$\mathbb{P}\left(\left|X\right|>t\right)\leq \exp\left\{-\frac{1}{2}\min\left(\frac{t^2}{\nu^2},\frac{t}{\nu}\right)\right\}.$$

Lemma 25 (Hanson-Wright inequality in Rudelson and Vershynin (2013)) *Let $\mathbf{Z}=(Z_1,\ldots,Z_n)\in\mathbb{R}^n$ be a random vector with independent components Z_i such that $\mathbb{E}(Z_i)=0$ and $\|Z_i\|_{\psi_2}\leq K$, where $\|\cdot\|_{\psi_2}$ is the sub-gaussian norm. Let $\mathbf{Q}$ be an $n\times n$ matrix. Then, for every $t\geq 0$, we have*

$$\mathbb{P}\left\{\left|\mathbf{Z}^{\mathrm{T}}\mathbf{Q}\mathbf{Z}-\mathbb{E}(\mathbf{Z}^{\mathrm{T}}\mathbf{Q}\mathbf{Z})\right|>t\right\}\leq 2\exp\left\{-c\min\left(\frac{t^2}{K^4\|\mathbf{Q}\|_{\mathrm{F}}^2},\frac{t}{K^2\|\mathbf{Q}\|_{\mathrm{op}}}\right)\right\},$$

where $\|\mathbf{Q}\|_{\mathrm{F}}$ is the Frobenius norm and $\|\mathbf{Q}\|_{\mathrm{op}}$ is the operator norm.

Lemma 26 (Proposition 1.1 in Rigollet and Hütter (2015)) *Let $X\sim N(\mu,\sigma^2)$, then for any $t>0$,*

$$\mathbb{P}\left(\left|X-\mu\right|>t\right)\leq \frac{\sigma}{t}\exp\left(-\frac{t^2}{2\sigma^2}\right).$$

F.4 Verifying Assumption 6

In this subsection, we provide an example in which Assumption 6 holds under (16). Denote $\boldsymbol{X}_{it}=(X_{it1},\ldots,X_{itp})^{\mathrm{T}}$. Let us consider the following form

$$\boldsymbol{X}_{it}=\widetilde{\mathbf{A}}\boldsymbol{X}_{i(t-1)}+\widetilde{\boldsymbol{\Delta}}_{it}+\widetilde{\boldsymbol{\epsilon}}_{it},\tag{80}$$

where $\widetilde{\mathbf{A}} \in \mathbb{R}^{p\times p}$, $\widetilde{\boldsymbol{\Delta}}_{it} \in \mathbb{R}^p$ and $\widetilde{\boldsymbol{\epsilon}}_{it} \sim N(0, \widetilde{\boldsymbol{\Sigma}})$ independent of $\boldsymbol{X}_{i(t-1)}$. Since given $\boldsymbol{X}_{i(t-1)}$, $\boldsymbol{X}_{it}$ is multivariate Gaussian. We can easily rewrite (80) as the model in (16), where

$$\begin{aligned}\boldsymbol{\theta}^*_{j,-j} &= (\widetilde{\boldsymbol{\Sigma}}_{-j,-j})^{-1}\widetilde{\boldsymbol{\Sigma}}_{-j,j},\\ \boldsymbol{\alpha}^*_j &= \widetilde{\mathbf{A}}_{j\cdot} - \widetilde{\boldsymbol{\Sigma}}_{j,-j}(\widetilde{\boldsymbol{\Sigma}}_{-j,-j})^{-1}\widetilde{\mathbf{A}}_{-j\cdot},\\ \Delta^*_{itj} &= \widetilde{\Delta}_{itj} - \widetilde{\boldsymbol{\Sigma}}_{j,-j}(\widetilde{\boldsymbol{\Sigma}}_{-j,-j})^{-1}\widetilde{\Delta}_{it(-j)}.\end{aligned}$$

Thus, it suffices to verify Assumption 6 under the model (80). Assume that $\widetilde{\Delta}_{i1j} = \cdots = \widetilde{\Delta}_{iqj} = 0$ and $\widetilde{\Delta}_{i(q+1)j} = \cdots = \widetilde{\Delta}_{iTj} = \Delta_{ij}$ for some $1 < q < T$. So, the sequence $\Delta^*_{i1j}, \Delta^*_{i2j}, \ldots, \Delta^*_{iTj}$ has a single jump at $t = q$, which yields $\Delta_m = \max_{i,t,j} |\Delta^*_{itj} - \Delta^*_{i(t-1)j}| = \max_{i,j} |\Delta_{ij} - \widetilde{\boldsymbol{\Sigma}}_{j,-j}(\widetilde{\boldsymbol{\Sigma}}_{-j,-j})^{-1}\Delta_{i(-j)}|$ and $\tau = \max_{i,j}\sum_{t=2}^T I(\Delta^*_{itj} \neq \Delta^*_{i(t-1)j}) = 1$. Thus, under mild conditions such as $\|\widetilde{\boldsymbol{\Sigma}}_{j,-j}(\widetilde{\boldsymbol{\Sigma}}_{-j,-j})^{-1}\|_2$ and $\|\boldsymbol{\Delta}_i\|_2$ are bounded by some constants, we can claim that $\Delta_{\max} = \Delta_m + 1$ is bounded by a constant.

Next, to compute $\mu_{itj} = \mathbb{E}(X_{itj})$, we further assume $\widetilde{\mathbf{A}} = a\mathbf{I}_p$. Then (80) yields $\mu_{itj} = \mathbb{E}(X_{itj}) = a\mathbb{E}(X_{i(t-1)j}) + \widetilde{\Delta}_{itj}$, where for notational simplicity we let $X_{i0j} = 0$. We have $\mu_{itj} = 0$ for $t \le q$ and $\mu_{itj} = \sum_{s=q+1}^t a^{s-q-1}\Delta_{ij}$ for $t \ge q+1$. In the latter case, it is easily shown that $\max|\mu_{itj}| \le \frac{1-|a|^{T-q}}{1-|a|}\max|\Delta_{ij}|$. Thus, for any $|a| \le c$ for some constant $c < 1$, we obtain that $\max|\mu_{itj}|$ is bounded by a constant. Apparently, the ℓ_2 norm of $\boldsymbol{\mu}_{ij}$ satisfies $\|\boldsymbol{\mu}_{ij}\|_2 \le \sqrt{T-q}\mu_{\max}$.

Finally, we can similarly compute the covariance matrix of $\boldsymbol{X}_{ij}$, i.e., $\boldsymbol{\Sigma}_{jj}$. For simplicity, assume $\widetilde{\Sigma}_{jj} = 1$. It can be shown that $(\boldsymbol{\Sigma}_{jj})_{11} = 1$ and $(\boldsymbol{\Sigma}_{jj})_{tt} = a^2(\boldsymbol{\Sigma}_{jj})_{(t-1)(t-1)} + 1$. For the off-diagonal entries, we have $(\boldsymbol{\Sigma}_{jj})_{tt'} = a^{t'-t}(\boldsymbol{\Sigma}_{jj})_{tt}$ for $t' > t$. As a result, following the similar calculation, we can show that the matrix ℓ_1 (and ℓ_∞) norm of $\boldsymbol{\Sigma}_{jj}$ is bounded by a constant. By the matrix norm interpolation inequality ($\|\mathbf{M}\|_{op}^2 \le \|\mathbf{M}\|_{\ell_1}\|\mathbf{M}\|_{\ell_\infty}$), the operator norm of $\boldsymbol{\Sigma}_{jj}$ is also bounded. Thus, the condition on the covariance matrix in Assumption 6 holds.

F.5 Assumption 7 Holds with High Probability

In this subsection, we will show that Assumption 7 holds with high probability. Some mild assumptions are needed, which are stated in Assumption 27.

Assumption 27 *Let* $\boldsymbol{\Sigma}_{jj'} = \mathrm{Cov}(\boldsymbol{X}_{ij}, \boldsymbol{X}_{ij'})$ *with* $\boldsymbol{X}_{ij} = (X_{i1j}, \ldots, X_{iTj})^{\mathrm{T}}$, *and* $\widetilde{\boldsymbol{\Sigma}}_{jj'} = \mathrm{Cov}(\boldsymbol{X}_{ij}, \check{\boldsymbol{X}}_{ij'})$ *with* $\check{\boldsymbol{X}}_{ij'} = (X_{i1j}, \ldots, X_{i(T-1)j})^{\mathrm{T}}$. *We assume that* $\max\{\max_{j,j'}\|\widetilde{\boldsymbol{\Sigma}}_{jj'}\|_{\mathrm{op}}, \max_{j,j'}\|\boldsymbol{\Sigma}_{jj'}\|_{\mathrm{op}}\} \le \kappa'$. *Besides, let* $\mathbf{U}_j = \mathbb{E}(\mathbf{Y}_j)$ *and* $\boldsymbol{\Sigma}_j = \mathbb{E}\{(\mathbf{Y}_j - \mathbf{U}_j)^{\mathrm{T}}(\mathbf{Y}_j - \mathbf{U}_j)\}/(nT)$. *Assume* $\min_{j\in\{1,\ldots,p\}}\Lambda_{\min}(\boldsymbol{\Sigma}_j) \ge \phi_0$, *where* ϕ_0 *is a positive constant and* $\Lambda_{\min}(\boldsymbol{\Sigma}_j)$ *is the smallest eigenvalue of* $\boldsymbol{\Sigma}_j$.

Lemma 28 shows that Assumption 7 holds with probability at least $1 - 2/(nT)^2$ under Assumption 27. The proof of Lemma 28 is showed in Section F.5.1.

Lemma 28 *Let* $\widehat{\boldsymbol{\Sigma}}_j = (\mathbf{Y}_j - \mathbf{U}_j)^{\mathrm{T}}(\mathbf{Y}_j - \mathbf{U}_j)/(nT)$. *With sufficiently large positive constant* $\mathcal{C}_0$ *and under Assumption 27, we have*

$$\left\|(\boldsymbol{\omega}^*_j)^{\mathcal{S}_j}\right\|_1^2 \le \frac{\left(\boldsymbol{\omega}^*_j\right)^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_j\boldsymbol{\omega}^*_j s_j}{\phi_0} + \frac{2\mathcal{C}s_j\log\{nT(2p-1)\}}{\phi_0\sqrt{nT}}\left\|\boldsymbol{\omega}^*_j\right\|_1^2.$$

with probability at least $1-2/(nT)^2$.

F.5.1 Proof of Lemma 28

Under Assumption 7, we can easily obtain that

$$\begin{aligned}\left\|(\boldsymbol{\omega}_j^*)^{\mathcal{S}_j}\right\|_1^2 &\le \frac{\left(\boldsymbol{\omega}_j^*\right)^{\mathrm{T}}\boldsymbol{\Sigma}_j\boldsymbol{\omega}_j^* s_j}{\phi_0} \\ &= \frac{\left(\boldsymbol{\omega}_j^*\right)^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_j\boldsymbol{\omega}_j^* s_j}{\phi_0} + \frac{\left(\boldsymbol{\omega}_j^*\right)^{\mathrm{T}}\left(\boldsymbol{\Sigma}_j-\widehat{\boldsymbol{\Sigma}}_j\right)\boldsymbol{\omega}_j^* s_j}{\phi_0} \\ &\le \frac{\left(\boldsymbol{\omega}_j^*\right)^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_j\boldsymbol{\omega}_j^* s_j}{\phi_0} + \frac{\left\|\boldsymbol{\Sigma}_j-\widehat{\boldsymbol{\Sigma}}_j\right\|_{\max}\left\|\boldsymbol{\omega}_j^*\right\|_1^2 s_j}{\phi_0}, \end{aligned} \tag{81}$$

where the last inequality follows holder's inequality and $\|\mathbf{M}\|_{\max}=\max_{j,k}|M_{jk}|$ with $\mathbf{M}$ is a matrix. For the next step, we will bound the term $\|\boldsymbol{\Sigma}_j-\widehat{\boldsymbol{\Sigma}}_j\|_{\max}$. Recall $\widehat{\boldsymbol{\Sigma}}_j$. Let $\widetilde{\boldsymbol{X}}_{it(-j)}=\boldsymbol{X}_{it(-j)}-\boldsymbol{U}_{it(-j)}$ and $\widetilde{\boldsymbol{X}}_{i(t-1)}=\boldsymbol{X}_{i(t-1)}-\boldsymbol{U}_{i(t-1)}$ with $\boldsymbol{U}_{it(-j)}=\mathbb{E}\{\boldsymbol{X}_{it(-j)}\}$ and $\boldsymbol{U}_{i(t-1)}=\mathbb{E}\{\boldsymbol{X}_{i(t-1)}\}$. Then we have

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}}_j &= \frac{1}{nT}\left(\mathbf{Y}_j-\mathbf{U}_j\right)^{\mathrm{T}}\left(\mathbf{Y}_j-\mathbf{U}_j\right) \\ &= \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\begin{pmatrix}\widetilde{\boldsymbol{X}}_{it(-j)}\widetilde{\boldsymbol{X}}_{it(-j)}^{\mathrm{T}} & \widetilde{\boldsymbol{X}}_{it(-j)}\widetilde{\boldsymbol{X}}_{i(t-1)}^{\mathrm{T}} \\ \widetilde{\boldsymbol{X}}_{i(t-1)}\widetilde{\boldsymbol{X}}_{it(-j)}^{\mathrm{T}} & \widetilde{\boldsymbol{X}}_{i(t-1)}\widetilde{\boldsymbol{X}}_{i(t-1)}^{\mathrm{T}}\end{pmatrix} \\ &\triangleq \begin{pmatrix}\widehat{\boldsymbol{\Sigma}}_{tt(-j)(-j)} & \widehat{\boldsymbol{\Sigma}}_{t(t-1)(-j)} \\ \widehat{\boldsymbol{\Sigma}}_{t(t-1)(-j)}^{\mathrm{T}} & \widehat{\boldsymbol{\Sigma}}_{(t-1)(t-1)}\end{pmatrix}. \end{aligned} \tag{82}$$

We will discuss the upper bound for each element in $\widehat{\boldsymbol{\Sigma}}_{tt(-j)(-j)}-\boldsymbol{\Sigma}_{tt(-j)(-j)}$, $\widehat{\boldsymbol{\Sigma}}_{t(t-1)(-j)}-\boldsymbol{\Sigma}_{tt(-j)(-j)}$ and $\widehat{\boldsymbol{\Sigma}}_{(t-1)(t-1)}-\boldsymbol{\Sigma}_{tt(-j)(-j)}$, separately. The matrices $\boldsymbol{\Sigma}_{tt(-j)(-j)}=\mathbb{E}\{\widehat{\boldsymbol{\Sigma}}_{tt(-j)(-j)}\}$, $\boldsymbol{\Sigma}_{tt(-j)(-j)}=\mathbb{E}\{\widehat{\boldsymbol{\Sigma}}_{t(t-1)(-j)}\}$ and $\boldsymbol{\Sigma}_{tt(-j)(-j)}=\mathbb{E}\{\widehat{\boldsymbol{\Sigma}}_{(t-1)(t-1)}\}$.

First, the $(j,j^{'})$th element in the matrix $\widehat{\boldsymbol{\Sigma}}_{tt(-j)(-j)}$ is

$$\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\widetilde{X}_{itj}\widetilde{X}_{itj'} \triangleq \frac{1}{nT}\widetilde{\boldsymbol{X}}_j^{\mathrm{T}}\widetilde{\boldsymbol{X}}_{j'},$$

where $\widetilde{\boldsymbol{X}}_j=(\widetilde{X}_{11j},\ldots,\widetilde{X}_{1Tj},\ldots,\widetilde{X}_{nTj})^{\mathrm{T}}$ and $\widetilde{\boldsymbol{X}}_{j'}=(\widetilde{X}_{11j'},\ldots,\widetilde{X}_{1Tj'},\ldots,\widetilde{X}_{nTj'})^{\mathrm{T}}$. We mainly use Lemma 25 to bound $\sum_{i=1}^{n}\sum_{t=1}^{T}\widetilde{X}_{itj}\widetilde{X}_{itj'}/(nT)$. To convert $\sum_{i=1}^{n}\sum_{t=1}^{T}\widetilde{X}_{itj}\widetilde{X}_{itj'}/(nT)$ into a formate of $\boldsymbol{Z}\mathbf{Q}\boldsymbol{Z}^{\mathrm{T}}$, let $\widetilde{\mathbf{X}}_{jj'}=(\widetilde{\boldsymbol{X}}_j^{\mathrm{T}},\widetilde{\boldsymbol{X}}_{j'}^{\mathrm{T}})^{\mathrm{T}}$. By some simple calculations, the random vector $\widetilde{\mathbf{X}}_{jj'}$ has zero mean and variance

$$\widetilde{\mathbf{Q}}=\begin{pmatrix}\mathbf{I}_n\otimes\boldsymbol{\Sigma}_{jj} & \mathbf{I}_n\otimes\boldsymbol{\Sigma}_{jj'} \\ \left(\mathbf{I}_n\otimes\boldsymbol{\Sigma}_{jj'}\right)^{\mathrm{T}} & \mathbf{I}_n\otimes\boldsymbol{\Sigma}_{j'j'}\end{pmatrix}, \tag{83}$$

where $\mathbf{\Sigma}_{jj} = \mathrm{Cov}(\boldsymbol{X}_{ij}, \boldsymbol{X}_{ij})$, $\mathbf{\Sigma}_{jj'} = \mathrm{Cov}(\boldsymbol{X}_{ij}, \boldsymbol{X}_{ij'})$ and $\mathbf{\Sigma}_{j'j'} = \mathrm{Cov}(\boldsymbol{X}_{ij'}, \boldsymbol{X}_{ij'})$. Next, let $\widetilde{\mathbf{Q}}^{1/2}\boldsymbol{Z} = (\widetilde{\boldsymbol{X}}_j^{\mathrm{T}}, \widetilde{\boldsymbol{X}}_{j'}^{\mathrm{T}})^{\mathrm{T}}$. We divide $\boldsymbol{Z}$ and $\widetilde{\mathbf{Q}}^{1/2}$ corresponding to $\widetilde{\boldsymbol{X}}_j$ and $\widetilde{\boldsymbol{X}}_{j'}$. Let $\boldsymbol{Z} = (\boldsymbol{Z}_1^{\mathrm{T}}, \boldsymbol{Z}_2^{\mathrm{T}})^{\mathrm{T}}$ and

$$\widetilde{\mathbf{Q}}^{\frac{1}{2}} = \begin{pmatrix} \widetilde{\mathbf{Q}}_{11} & \widetilde{\mathbf{Q}}_{12} \\ \widetilde{\mathbf{Q}}_{12}^{\mathrm{T}} & \widetilde{\mathbf{Q}}_{22} \end{pmatrix}. \tag{84}$$

Then we have

$$\begin{pmatrix} \widetilde{\boldsymbol{X}}_j \\ \widetilde{\boldsymbol{X}}_{j'} \end{pmatrix} = \begin{pmatrix} \widetilde{\mathbf{Q}}_{11}\boldsymbol{Z}_1 + \widetilde{\mathbf{Q}}_{12}\boldsymbol{Z}_2 \\ \widetilde{\mathbf{Q}}_{12}^{\mathrm{T}}\boldsymbol{Z}_1 + \widetilde{\mathbf{Q}}_{22}\boldsymbol{Z}_2 \end{pmatrix} \quad \text{and} \quad \widetilde{\boldsymbol{X}}_j^{\mathrm{T}}\widetilde{\boldsymbol{X}}_{j'} = \boldsymbol{Z}^{\mathrm{T}} \begin{pmatrix} \widetilde{\mathbf{Q}}_{11} \\ \widetilde{\mathbf{Q}}_{12} \end{pmatrix} \begin{pmatrix} \widetilde{\mathbf{Q}}_{12}^{\mathrm{T}} & \widetilde{\mathbf{Q}}_{22} \end{pmatrix} \boldsymbol{Z} \triangleq \boldsymbol{Z}^{\mathrm{T}}\mathbf{Q}\boldsymbol{Z}. \tag{85}$$

Besides, to link the $\widetilde{\mathbf{Q}}$ with $\widetilde{\mathbf{Q}}_{11}, \widetilde{\mathbf{Q}}_{22}$ and $\widetilde{\mathbf{Q}}_{12}$,

$$\widetilde{\mathbf{Q}} = \widetilde{\mathbf{Q}}^{\frac{1}{2}}\widetilde{\mathbf{Q}}^{\frac{1}{2}} = \begin{pmatrix} \widetilde{\mathbf{Q}}_{11}^2 + \widetilde{\mathbf{Q}}_{12}\widetilde{\mathbf{Q}}_{12}^{\mathrm{T}} & \widetilde{\mathbf{Q}}_{11}\widetilde{\mathbf{Q}}_{12} + \widetilde{\mathbf{Q}}_{12}\widetilde{\mathbf{Q}}_{22} \\ \widetilde{\mathbf{Q}}_{12}^{\mathrm{T}}\widetilde{\mathbf{Q}}_{11} + \widetilde{\mathbf{Q}}_{22}\widetilde{\mathbf{Q}}_{12}^{\mathrm{T}} & \widetilde{\mathbf{Q}}_{12}^{\mathrm{T}}\widetilde{\mathbf{Q}}_{12} + \widetilde{\mathbf{Q}}_{22}^2 \end{pmatrix}. \tag{86}$$

By Lemma 25, with sufficiently large $\mathcal{C}_0$, the upper bound of $\widetilde{\boldsymbol{X}}_j^{\mathrm{T}}\widetilde{\boldsymbol{X}}_{j'}$ is

$$\begin{aligned} & p\left\{ \frac{1}{nT}\left| \widetilde{\boldsymbol{X}}_j^{\mathrm{T}}\widetilde{\boldsymbol{X}}_{j'} - \mathbb{E}\left(\widetilde{\boldsymbol{X}}_j^{\mathrm{T}}\widetilde{\boldsymbol{X}}_{j'} \right) \right| > \frac{2\mathcal{C}_0\kappa^{'}\log\left\{nT\left(2p-1\right)\right\}}{\sqrt{nT}} \right\} \\ & \leq 2\exp\left(-c\min\left[\frac{4\mathcal{C}_0^2(\kappa^{'})^2\log^2\left\{nT\left(2p-1\right)\right\}nT}{\|\mathbf{Q}\|_F^2}, \frac{2\mathcal{C}_0\kappa^{'}\log\left\{nT\left(2p-1\right)\right\}\sqrt{nT}}{\|\mathbf{Q}\|_{\mathrm{op}}} \right] \right) \\ & \leq \frac{2}{\left\{nT\left(2p-1\right)\right\}^2}, \end{aligned} \tag{87}$$

where the last inequality follows that

$$\|\mathbf{Q}\|_{\mathrm{op}} = \left\| \begin{pmatrix} \widetilde{\mathbf{Q}}_{11} \\ \widetilde{\mathbf{Q}}_{12} \end{pmatrix} \begin{pmatrix} \widetilde{\mathbf{Q}}_{12}^{\mathrm{T}} & \widetilde{\mathbf{Q}}_{22} \end{pmatrix} \right\|_{\mathrm{op}} = \left\| \begin{pmatrix} \widetilde{\mathbf{Q}}_{12}^{\mathrm{T}} & \widetilde{\mathbf{Q}}_{22} \end{pmatrix} \begin{pmatrix} \widetilde{\mathbf{Q}}_{11} \\ \widetilde{\mathbf{Q}}_{12} \end{pmatrix} \right\|_{\mathrm{op}} = \left\| \mathbf{I}_n \otimes \mathbf{\Sigma}_{jj'} \right\|_{\mathrm{op}} = \left\| \mathbf{\Sigma}_{jj'} \right\|_{\mathrm{op}} \leq \kappa^{'}, \tag{88}$$

where the third equality holds by (83) and (86) and the inequality holds by Assumption 7.

Following similar procedures, we could obtains the upper bounds for the other two terms $\widehat{\mathbf{\Sigma}}_{t(t-1)(-j)} - \mathbf{\Sigma}_{tt(-j)(-j)}$ and $\widehat{\mathbf{\Sigma}}_{(t-1)(t-1)} - \mathbf{\Sigma}_{tt(-j)(-j)}$. For $\widehat{\mathbf{\Sigma}}_{t(t-1)(-j)}$, the upper bounds for the $(j, j^{'})$th elements is

$$p\left\{ \frac{1}{nT}\left| \sum_{i=1}^{n}\sum_{t=1}^{T}\widetilde{X}_{itj}\widetilde{X}_{i(t-1)j^{'}} - \mathbb{E}\left(\sum_{i=1}^{n}\sum_{t=1}^{T}\widetilde{X}_{itj}\widetilde{X}_{i(t-1)j^{'}} \right) \right| > \frac{2\mathcal{C}_0\kappa^{'}\log\left\{nT\left(2p-1\right)\right\}}{\sqrt{nT}} \right\} \leq \frac{2}{\left\{nT\left(2p-1\right)\right\}^2}. \tag{89}$$

For $\widehat{\mathbf{\Sigma}}_{(t-1)(t-1)}$, the upper bounds for the $(j, j^{'})$th elements is

$$p\left\{\frac{1}{nT}\left|\sum_{i=1}^{n}\sum_{t=1}^{T}\widetilde{X}_{i(t-1)j}\widetilde{X}_{i(t-1)j^{'}}-\mathbb{E}\left(\sum_{i=1}^{n}\sum_{t=1}^{T}\widetilde{X}_{i(t-1)j}\widetilde{X}_{i(t-1)j^{'}}\right)\right|>\frac{2\mathcal{C}_0\kappa^{'}\log\left\{nT\left(2p-1\right)\right\}}{\sqrt{nT}}\right\}$$
$$\leq\frac{2}{\left\{nT\left(2p-1\right)\right\}^2}. \tag{90}$$

Combining (87), (89) and (90), the upper bound for

$$p\left[\|\mathbf{\Sigma}_j-\widehat{\mathbf{\Sigma}}_j\|_{\max}\leq\frac{2\mathcal{C}_0\kappa^{'}\log\left\{nT\left(2p-1\right)\right\}}{\sqrt{nT}}\right]>1-\frac{2}{n^2T^2}. \tag{91}$$

Substituting (91) into (81) yields

$$p\left[\left\|(\boldsymbol{\omega}_j^*)^{\mathcal{S}_j}\right\|_1^2\leq\frac{\left(\boldsymbol{\omega}_j^*\right)^{\mathrm{T}}\widehat{\mathbf{\Sigma}}_j\boldsymbol{\omega}_j^*s_j}{\phi_0}+\frac{2\mathcal{C}_0s_j\kappa^{'}\log\left\{nT\left(2p-1\right)\right\}}{\phi_0\sqrt{nT}}\left\|\boldsymbol{\omega}_j^*\right\|_1^2\right]>1-\frac{2}{n^2T^2}.$$

Appendix G. Proof of Theorem 18

G.1 Technical Lemmas

The proof of Theorem 18 will use Lemmas 29–30, which perform similar functions as Lemmas 19 and 20. Since the proof of Lemma 30 is similar to Lemma 29, we will only show the proof of Lemma 29.

Lemma 29 *Let* $\epsilon_{itj}^{\triangledown}=X_{itj}-D^{'}(\eta_{itj}^*)$. *We have*

$$\max_{1\leq k\leq p,k\neq j}\frac{1}{nT}\left|\sum_{i=1}^{n}\sum_{t=1}^{T}\epsilon_{itj}^{\triangledown}X_{itk}\right|\leq\lambda_0,$$

with probability at least $1-2(c_vc_uc_d+c_hc_u^2)/(c_d^2T^{1/c_d})-\exp[\log\{2(p-1)\}-3\lambda_0^2nT/\{2\lambda_0\log(T)+6\log^2(T)\}]$.

Lemma 30 *Let* $\epsilon_{itj}^{\triangledown}=X_{itj}-D^{'}(\eta_{itj}^*)$. *We have*

$$\max_{1\leq k\leq p}\frac{1}{nT}\left|\sum_{i=1}^{n}\sum_{t=1}^{T}\epsilon_{itj}^{\triangledown}X_{i(t-1)k}\right|\leq\beta_0,$$

with probability at least $1-2(c_vc_uc_d+c_hc_u^2)/(c_d^2T^{1/c_d})-\exp[\log(2p)-3\beta_0^2nT/\{2\beta_0\log(T)+6\log^2(T)\}]$.

G.2 Proof of Theorem 18

The log-likelihood function of exponential family distribution can be written as

$$l_{itj}\left(\boldsymbol{\omega}_j, \Delta_{itj}\right) = -\eta_{itj} X_{itj} + D(\eta_{itj}), \tag{92}$$

where $\eta_{itj} = \boldsymbol{Y}_{itj}\boldsymbol{\omega}_j + \Delta_{itj}$, $\boldsymbol{Y}_{itj} = (\boldsymbol{X}_{it(-j)}, \boldsymbol{X}_{i(t-1)})$ and $\boldsymbol{\omega}_j = (\boldsymbol{\theta}_{j,-j}^{\mathrm{T}}, \boldsymbol{\alpha}_j^{\mathrm{T}})^{\mathrm{T}}$. Then the optimization problem of exponential family distribution is

$$\underset{\boldsymbol{\theta}_{j,-j},\boldsymbol{\alpha}_j,\boldsymbol{\Delta}_j}{\text{minimize}} \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} l_{itj}\left(\boldsymbol{\omega}_j, \Delta_{itj}\right) + \lambda\left\|\boldsymbol{\theta}_{j,-j}\right\|_1 + \beta\left\|\boldsymbol{\alpha}_j\right\|_1 + \gamma\sum_{i=1}^{n}\left\|\mathbf{C}\boldsymbol{\Delta}_{ij}\right\|_1. \tag{93}$$

Next, we will apply Taylor expansion to obtain a similar equation as (30). By the convexity of $l_{itj}(\widehat{\boldsymbol{\omega}}_j, \widehat{\Delta}_{itj})$, we can easily have

$$B_{itj} \triangleq l_{itj}\left(\widehat{\boldsymbol{\omega}}_j, \widehat{\Delta}_{itj}\right) - l_{itj}\left(\boldsymbol{\omega}_j^*, \Delta_{itj}^*\right) - \bigtriangledown_1 l_{itj}\left(\boldsymbol{\omega}_j^*, \Delta_{itj}^*\right)\left(\widehat{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*\right) - \bigtriangledown_2 l_{itj}\left(\boldsymbol{\omega}_j^*, \Delta_{itj}^*\right)\left(\widehat{\Delta}_{itj} - \Delta_{itj}^*\right) > 0. \tag{94}$$

Here, for notational simplicity, we just use $\widehat{\boldsymbol{\omega}}_j, \widehat{\Delta}_{itj}$ to denote the convex combinations of the estimators and the truth as in the proof of Theorem 8.

Besides, by (93), we have

$$\begin{aligned}
&\frac{1}{nT}\sum_{i,t} l_{itj}\left(\widehat{\boldsymbol{\omega}}_j, \widehat{\Delta}_{itj}\right) + \lambda\left\|\widehat{\boldsymbol{\theta}}_{j,-j}\right\|_1 + \beta\left\|\widehat{\boldsymbol{\alpha}}_j\right\|_1 + \gamma\sum_{i=1}^{n}\left\|\mathbf{C}\widehat{\boldsymbol{\Delta}}_{ij}\right\|_1 \\
&\leq \frac{1}{nT}\sum_{i,t} l_{itj}\left(\boldsymbol{\omega}_j^*, \Delta_{itj}^*\right) + \lambda\left\|\boldsymbol{\theta}_{j,-j}^*\right\|_1 + \beta\left\|\boldsymbol{\alpha}_j^*\right\|_1 + \gamma\sum_{i=1}^{n}\left\|\mathbf{C}\boldsymbol{\Delta}_{ij}^*\right\|_1.
\end{aligned} \tag{95}$$

By simple calculation of (95), we can obtain

$$\begin{aligned}
&\frac{1}{nT}\sum_{i,t}\left\{l_{itj}\left(\widehat{\boldsymbol{\omega}}_j, \widehat{\Delta}_{itj}\right) - l_{itj}\left(\boldsymbol{\omega}_j^*, \Delta_{itj}^*\right)\right\} \\
&\leq \lambda\left(\left\|\boldsymbol{\theta}_{j,-j}^*\right\|_1 - \left\|\widehat{\boldsymbol{\theta}}_{j,-j}\right\|_1\right) + \beta\left(\left\|\boldsymbol{\alpha}_j^*\right\|_1 - \left\|\widehat{\boldsymbol{\alpha}}_j\right\|_1\right) + \gamma\sum_{i=1}^{n}\left(\left\|\mathbf{C}\boldsymbol{\Delta}_{ij}^*\right\|_1 - \left\|\mathbf{C}\widehat{\boldsymbol{\Delta}}_{ij}\right\|_1\right).
\end{aligned} \tag{96}$$

Combining (94) and (96) yields

$$\begin{aligned}
\frac{1}{nT}\sum_{i,t} B_{itj} \leq &-\frac{1}{nT}\sum_{i,t}\bigtriangledown_1 l_{itj}\left(\boldsymbol{\omega}_j^*, \Delta_{itj}^*\right)\left(\widehat{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*\right) - \frac{1}{nT}\sum_{i,t}\bigtriangledown_2 l_{itj}\left(\boldsymbol{\omega}_j^*, \Delta_{itj}^*\right)\left(\widehat{\Delta}_{itj} - \Delta_{itj}^*\right) \\
&+ \lambda\left(\left\|\boldsymbol{\theta}_{j,-j}^*\right\|_1 - \left\|\widehat{\boldsymbol{\theta}}_{j,-j}\right\|_1\right) + \beta\left(\left\|\boldsymbol{\alpha}_j^*\right\|_1 - \left\|\widehat{\boldsymbol{\alpha}}_j\right\|_1\right) + \gamma\sum_{i=1}^{n}\left(\left\|\mathbf{C}\boldsymbol{\Delta}_{ij}^*\right\|_1 - \left\|\mathbf{C}\widehat{\boldsymbol{\Delta}}_{ij}\right\|_1\right).
\end{aligned} \tag{97}$$

By Taylor expansion of $l_{itj}(\widehat{\boldsymbol{\omega}}_j, \widehat{\Delta}_{itj})$, we can also have

$$\begin{aligned}
&\frac{1}{nT}\sum_{i,t} B_{itj} \\
&\geq c_d \left\{ \left(\widehat{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*\right)^{\mathrm{T}} \frac{1}{nT}\sum_{i,t} \boldsymbol{Y}_{itj}^{\mathrm{T}} \boldsymbol{Y}_{itj} \left(\widehat{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*\right) + \frac{1}{nT}\sum_{i,t} \left(\widehat{\Delta}_{itj} - \Delta_{itj}^*\right)^2 + \frac{1}{nT}\sum_{i,t} \left(\widehat{\Delta}_{itj} - \Delta_{itj}^*\right) \boldsymbol{Y}_{itj} \left(\widehat{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*\right) \right\}
\end{aligned} \tag{98}$$

By combining (97) and (98), we obtain that

$$\begin{aligned}
&\frac{c_d}{nT}\sum_{i,t} \left\{ \boldsymbol{Y}_{itj} \left(\widehat{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*\right) \right\}^2 + \frac{c_d}{nT}\sum_{i,t} \left(\widehat{\Delta}_{itj} - \Delta_{itj}^*\right)^2 \\
&\leq \underbrace{-\frac{1}{nT}\sum_{i,t} \bigtriangledown_1 l_{itj} \left(\boldsymbol{\omega}_j^*, \Delta_{itj}^*\right) \left(\widehat{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*\right) + \lambda \left(\left\|\boldsymbol{\theta}_{j,-j}^*\right\|_1 - \left\|\widehat{\boldsymbol{\theta}}_{j,-j}\right\|_1 \right) + \beta \left(\left\|\boldsymbol{\alpha}_j^*\right\|_1 - \left\|\widehat{\boldsymbol{\alpha}}_j\right\|_1 \right)}_{\mathbb{I}_1} \\
&\quad \underbrace{- \frac{1}{nT}\sum_{i,t} \bigtriangledown_2 l_{itj} \left(\boldsymbol{\omega}_j^*, \Delta_{itj}^*\right) \left(\widehat{\Delta}_{itj} - \Delta_{itj}^*\right) - \frac{2c_d}{nT}\sum_{i,t} \left(\widehat{\Delta}_{itj} - \Delta_{itj}^*\right) \boldsymbol{Y}_{itj} \left(\widehat{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*\right)}_{\mathbb{I}_2} \\
&\quad + \gamma \sum_{i=1}^{n} \left(\left\|\mathbf{C}\boldsymbol{\Delta}_{ij}^*\right\|_1 - \left\|\mathbf{C}\widehat{\boldsymbol{\Delta}}_{ij}\right\|_1 \right).
\end{aligned} \tag{99}$$

We will establish upper bounds for $\mathbb{I}_1$ and $\mathbb{I}_2$, respectively.

Upper Bound for $\mathbb{I}_1$: Recall (92). Let $\epsilon_{itj}^{\bigtriangledown} = X_{itj} - D^{'}(\eta_{itj}^*)$, then we have

$$\begin{aligned}
&- \frac{1}{nT}\sum_{i,t} \bigtriangledown_1 l_{itj} \left(\boldsymbol{\omega}_j^*, \Delta_{itj}^*\right) \left(\widehat{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*\right) \\
&= \frac{1}{nT}\sum_{i,t} \epsilon_{itj}^{\bigtriangledown} \boldsymbol{Y}_{itj} \left(\widehat{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*\right) \\
&= \frac{1}{nT}\sum_{i,t} \epsilon_{itj}^{\bigtriangledown} \boldsymbol{X}_{it(-j)} \left(\widehat{\boldsymbol{\theta}}_{j,-j} - \boldsymbol{\theta}_{j,-j}^*\right) + \frac{1}{nT}\sum_{i,t} \epsilon_{itj}^{\bigtriangledown} \boldsymbol{X}_{i(t-1)} \left(\widehat{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}_j^*\right) \\
&\leq \frac{1}{nT} \left\| \sum_{i,t} \epsilon_{itj}^{\bigtriangledown} \boldsymbol{X}_{it(-j)} \right\|_\infty \left\| \widehat{\boldsymbol{\theta}}_{j,-j} - \boldsymbol{\theta}_{j,-j}^* \right\|_1 + \frac{1}{nT} \left\| \sum_{i,t} \epsilon_{itj}^{\bigtriangledown} \boldsymbol{X}_{i(t-1)} \right\|_\infty \left\| \widehat{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}_j^* \right\|_1 \\
&\leq \frac{\beta}{2} \left\| \widehat{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^* \right\|_1,
\end{aligned} \tag{100}$$

with probability at least $1 - 4(c_v c_u c_d + c_h c_u^2)/(c_d^2 T^{1/c_d}) - 4/(nTp)$. The first inequality follows Holder's inequality, and the second inequality follows Lemmas 29–30 with $n, T, p \geq 6$ and $\lambda = \beta = 2\log(T)\log(nTp)T^{-1/4}$. Then the upper bound for $\mathbb{I}_1$ can be written as

$$\mathbb{I}_1 \leq \frac{\beta}{2} \left\| \widehat{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^* \right\|_1 + \beta \left(\left\| \boldsymbol{\omega}_j^* \right\|_1 - \left\| \widehat{\boldsymbol{\omega}}_j \right\|_1 \right). \tag{101}$$

Define $\boldsymbol{\omega}_j^{\mathcal{S}_j}$ and $\boldsymbol{\omega}_j^{\mathcal{S}_j^c}$ be the subvectors of $\boldsymbol{\omega}_j$ with indices $\mathcal{S}_j$ and $\mathcal{S}_j^c$, respectively. The equation (101) can be converted to

$$\mathbb{I}_1 \leq \frac{3\beta}{2}\left\|\widehat{\boldsymbol{\omega}}_j^{\mathcal{S}_j} - (\boldsymbol{\omega}_j^*)^{\mathcal{S}_j}\right\|_1 - \frac{\beta}{2}\left\|\widehat{\boldsymbol{\omega}}_j^{\mathcal{S}_j^c}\right\|_1, \tag{102}$$

with probability at least $1 - 4(c_v c_u c_d + c_h c_u^2)/(c_d^2 T^{1/c_d}) - 4/(nTp)$.

Upper Bound for $\mathbb{I}_2$: $\mathbb{I}_2$ can be rewritten as

$$\mathbb{I}_2 = \underbrace{\frac{1}{nT}\sum_{i=1}^{n}\epsilon_{itj}^{\triangledown}\left(\widehat{\Delta}_{ij} - \Delta_{ij}^*\right)}_{\mathbb{I}_{21}} + \underbrace{\frac{2c_d}{nT}\sum_{i,t}\left(\widehat{\Delta}_{itj} - \Delta_{itj}^*\right)\boldsymbol{Y}_{itj}\left(\widehat{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j^*\right)}_{\mathbb{I}_{22}}. \tag{103}$$

We will discuss the upper bound for $\mathbb{I}_{21}$ and $\mathbb{I}_{22}$, separately.

For $\mathbb{I}_{21}$, let $\mathbf{P} = \mathbf{C}^+\mathbf{C} \in \mathbb{R}^{T\times T}$, then $\mathbb{I}_{21}$ is equal to

$$\begin{aligned}\frac{1}{nT}\sum_{i=1}^{n}\epsilon_{itj}^{\triangledown}\left(\widehat{\Delta}_{ij} - \Delta_{ij}^*\right) &= \frac{1}{nT}\sum_{i=1}^{n}\epsilon_{itj}^{\triangledown}\mathbf{P}\left(\widehat{\Delta}_{ij} - \Delta_{ij}^*\right) + \frac{1}{nT}\sum_{i=1}^{n}\epsilon_{itj}^{\triangledown}(\mathbf{I}_T - \mathbf{P})\left(\widehat{\Delta}_{ij} - \Delta_{ij}^*\right)\\ &\leq \sum_{i=1}^{n}\left\|\frac{1}{nT}\epsilon_{itj}^{\triangledown}\mathbf{C}^+\right\|_\infty\left\|\mathbf{C}\left(\Delta_{ij}^* - \widehat{\Delta}_{ij}\right)\right\|_1 + \sum_{i=1}^{n}\left\|\frac{1}{nT}\epsilon_{itj}^{\triangledown}(\mathbf{I}_T - \mathbf{P})\right\|_2\left\|\Delta_{ij}^* - \widehat{\Delta}_{ij}\right\|_2,\end{aligned} \tag{104}$$

where the last inequality holds by Holder's inequality. Next, we will bound $\|\epsilon_{itj}^{\triangledown}\mathbf{C}^+\|_\infty$ and $\|\epsilon_{itj}^{\triangledown}(\mathbf{I}_T - \mathbf{P})\|_2$, separately. Let $\boldsymbol{C}_k^+$ as the kth column of $\mathbf{C}^+$. We could bound $\|\epsilon_{itj}^{\triangledown}\mathbf{C}^+\|_\infty$ as

$$\begin{aligned}P\left\{\left\|\epsilon_{itj}^{\triangledown}\mathbf{C}^+\right\|_\infty \geq 8CK\sqrt{T}\log(nT)\right\} &= P\left\{\max_k\left|\epsilon_{itj}^{\triangledown}\boldsymbol{C}_k^+\right| \geq 8CK\sqrt{T}\log(nT)\right\}\\ &\leq \sum_{k=1}^{T}P\left\{\left|\epsilon_{itj}^{\triangledown}\boldsymbol{C}_k^+\right| \geq 8CK\sqrt{T}\log(nT)\right\}\\ &\leq \frac{2}{(nT)^3},\end{aligned} \tag{105}$$

where the last inequality holds by Lemma 31, Assumption 15 and the definition of ψ_1 norm. Following similar procedures, the bound of $\|\epsilon_{itj}^{\triangledown}(\mathbf{I}_T - \mathbf{P})\|_2$ is

$$P\left\{\left\|\epsilon_{itj}^{\triangledown}\left(\mathbf{I}_T - \mathbf{P}\right)\right\|_2 \geq 8CK\log(nT)\right\} \leq \frac{2}{(nT)^3}. \tag{106}$$

By the union bound over i and combining (105) and (106), the bound of $\mathbb{I}_{21}$ is

$$\mathbb{I}_{21} \leq \frac{8CK\log(nT)}{n\sqrt{T}}\sum_{i=1}^{n}\left(\left\|\mathbf{C}\boldsymbol{\Delta}_{ij}^*\right\|_1 + \left\|\mathbf{C}\widehat{\boldsymbol{\Delta}}_{ij}\right\|_1\right) + \frac{8CK\log(nT)}{nT}\sum_{i=1}^{n}\left\|\boldsymbol{\Delta}_{ij}^* - \widehat{\boldsymbol{\Delta}}_{ij}\right\|_2, \tag{107}$$

with probability $1-4/(n^2T^3)$. Let $B(n,T)=\mu_{\max}\sqrt{\widetilde{C}\Delta_{\max}\tau}n^{1/2}T^{1/4}$. We follow a similar procedures as (107) and (36) to bound $\mathbb{I}_{22}$, then we have

$$\mathbb{I}_{22}\leq\frac{16Cc_dK\log(2pnT)}{n\sqrt{T}}\sum_{i=1}^{n}\left(\left\|\mathbf{C}\boldsymbol{\Delta}_{ij}^{*}\right\|_1+\left\|\mathbf{C}\widehat{\boldsymbol{\Delta}}_{ij}\right\|_1\right)+\frac{16Cc_dK\log(2pnT)+2c_dB(n,T)}{nT}\sum_{i=1}^{n}\left\|\boldsymbol{\Delta}_{ij}^{*}-\widehat{\boldsymbol{\Delta}}_{ij}\right\|_2, \tag{108}$$

with probability $1-1/(4n^2p^2T^3)$. Substituting (107) and (108) into (103), the upper bound of $\mathbb{I}_2$ is

$$\mathbb{I}_2\leq\frac{2\widetilde{C}\log(2pnT)}{n\sqrt{T}}\sum_{i=1}^{n}\left(\left\|\mathbf{C}\boldsymbol{\Delta}_{ij}^{*}\right\|_1+\left\|\mathbf{C}\widehat{\boldsymbol{\Delta}}_{ij}\right\|_1\right)+\frac{\left\{2\widetilde{C}+B(n,T)\right\}\log(2pnT)}{nT}\sum_{i=1}^{n}\left\|\boldsymbol{\Delta}_{ij}^{*}-\widehat{\boldsymbol{\Delta}}_{ij}\right\|_2, \tag{109}$$

with probability $1-5/(n^2T^3)$. The constant $\widetilde{C}=\max\{16Cc_dK,2c_d,8CK\}$. Let $\gamma=2\widetilde{C}\log(2pnT)/(n\sqrt{T})$, then by (109) and (102), (99) can be written as

$$\begin{aligned}
&\frac{c_d}{nT}\sum_{i,t}\left\{\boldsymbol{Y}_{itj}\left(\widehat{\boldsymbol{\omega}}_j-\boldsymbol{\omega}_j^{*}\right)\right\}^2+\frac{c_d}{nT}\sum_{i,t}\left(\widehat{\Delta}_{itj}-\Delta_{itj}^{*}\right)^2\\
&\leq\frac{3\beta}{2}\left\|\widehat{\boldsymbol{\omega}}_j^{\mathcal{S}_j}-(\boldsymbol{\omega}_j^{*})^{\mathcal{S}_j}\right\|_1-\frac{\beta}{2}\left\|\widehat{\boldsymbol{\omega}}_j^{\mathcal{S}_j^c}\right\|_1+\frac{4\widetilde{C}\log(2pnT)}{n\sqrt{T}}\left\|\left(\mathbf{I}_n\otimes\mathbf{C}\right)\boldsymbol{\Delta}_j^{*}\right\|_1\\
&\quad+\frac{\left\{2\widetilde{C}+B(n,T)\right\}\log(2pnT)}{nT}\left\|\boldsymbol{\Delta}_j^{*}-\widehat{\boldsymbol{\Delta}}_j\right\|_2,
\end{aligned} \tag{110}$$

with probability at least $1-4(c_vc_uc_d+c_hc_u^2)/(c_d^2T^{1/c_d})-4/(nTp)-5/(n^2T^3)$. Next, we will consider (110) under two cases:

1. $\frac{4\widetilde{C}\log(2pnT)}{n\sqrt{T}}\left\|\left(\mathbf{I}_n\otimes\mathbf{C}\right)\boldsymbol{\Delta}_j^{*}\right\|_1+\frac{\left\{2\widetilde{C}+B(n,T)\right\}\log(2pnT)}{nT}\left\|\boldsymbol{\Delta}_j^{*}-\widehat{\boldsymbol{\Delta}}_j\right\|_2\leq\frac{1}{4}\beta\left\|\bar{\boldsymbol{\omega}}_j-\boldsymbol{\omega}_j^{*}\right\|_1$;
2. $\frac{4\widetilde{C}\log(2pnT)}{n\sqrt{T}}\left\|\left(\mathbf{I}_n\otimes\mathbf{C}\right)\boldsymbol{\Delta}_j^{*}\right\|_1+\frac{\left\{2\widetilde{C}+B(n,T)\right\}\log(2pnT)}{nT}\left\|\boldsymbol{\Delta}_j^{*}-\widehat{\boldsymbol{\Delta}}_j\right\|_2>\frac{1}{4}\beta\left\|\bar{\boldsymbol{\omega}}_j-\boldsymbol{\omega}_j^{*}\right\|_1$.

Following similar proof steps at **Cases 1–2** in the proof of Theorem 8, we could obtain that with probability at least $1-4(c_vc_uc_d+c_hc_u^2)/(c_d^2T^{1/c_d})-4/(nTp)-5/(n^2T^3)$,

$$\begin{aligned}
&\left\|\widehat{\boldsymbol{\omega}}_j-\boldsymbol{\omega}_j^{*}\right\|_1+\frac{1}{\sqrt{nT}}\left\|\widehat{\boldsymbol{\Delta}}_j-\boldsymbol{\Delta}_j^{*}\right\|_2\\
&\leq\max\{\widetilde{c}_1s_j\log(T)\log(nTp),\widetilde{c}_2\tau\log(2npT)\log^{-1}(T)+\widetilde{c}_3\sqrt{\tau}\log(2pnT)\}T^{-\frac{1}{4}},
\end{aligned} \tag{111}$$

where $\widetilde{c}_1=16/(c_d\phi_0^2)+2/(c_d\phi_0)$, $\widetilde{c}_2=112\widetilde{C}^2/c_d+4(2+\mu_{\max}\sqrt{7/c_d})^2\widetilde{C}\Delta_{\max}+16\sqrt{7\Delta_{\max}/c_d}\widetilde{C}^{2/3}(2+\mu_{\max}\sqrt{7/c_d})$ and $\widetilde{c}_3=14\widetilde{C}/c_d+\sqrt{7/c_d}(2+\mu_{\max}\sqrt{7/c_d})\sqrt{\widetilde{C}\Delta_{\max}}$.

G.3 Proof of Lemma 29

Similar to proof of Lemma 19, let $\epsilon_{lj}^{\triangledown} = \epsilon_{itj}^{\triangledown}$ and $X_{lk}^{\triangledown} = X_{itk}$ with $l = (i-1)T + t$. Besides, denote the sequence $Z_m^{\triangledown}$ as $Z_m^{\triangledown} = \frac{1}{nT}\sum_{l=1}^{m} \epsilon_{lj}^{\triangledown} X_{lk}^{\triangledown}$. The proof is similar to the proof of Lemma 19 except

1. Prove $Z_m^{\triangledown}$ is a martingale;
2. Bound the term $|\epsilon_{mj}^{\triangledown} X_{mk}^{\triangledown}|$.

To prove $Z_m^{\triangledown}$ is a martingale, by smoothing we have

$$\begin{aligned}
&\mathbb{E}\left\{Z_m^{\triangledown} - Z_{m-1}^{\triangledown} | \epsilon_{1j}^{\triangledown}, \ldots, \epsilon_{(m-1)j}^{\triangledown}, X_{1k}^{\triangledown}, \ldots, X_{(m-1)k}^{\triangledown}\right\} \\
&= \frac{1}{nT}\mathbb{E}\left\{\epsilon_{mj}^{\triangledown} X_{mk}^{\triangledown} | \epsilon_{1j}^{\triangledown}, \ldots, \epsilon_{(m-1)j}^{\triangledown}, X_{1k}^{\triangledown}, \ldots, X_{(m-1)k}^{\triangledown}\right\} \\
&= \frac{1}{nT}\mathbb{E}\left[X_{mk}^{\triangledown}\mathbb{E}\left\{\epsilon_{mj}^{\triangledown} | X_{mk}^{\triangledown}\right\} | \epsilon_{1j}^{\triangledown}, \ldots, \epsilon_{(m-1)j}^{\triangledown}, X_{1k}^{\triangledown}, \ldots, X_{(m-1)k}^{\triangledown}\right] \\
&= \frac{1}{nT}\mathbb{E}\left[X_{mk}^{\triangledown}\mathbb{E}\left[\mathbb{E}\left\{\epsilon_{mj}^{\triangledown} | X_{it(-j)}, X_{i(t-1)}, \Delta_{itj}\right\} | X_{mk}^{\triangledown}\right] | \epsilon_{1j}^{\triangledown}, \ldots, \epsilon_{(m-1)j}^{\triangledown}, X_{1k}^{\triangledown}, \ldots, X_{(m-1)k}^{\triangledown}\right] \\
&= 0, \qquad (112)
\end{aligned}$$

where the last equality holds by the fact that X_{itj} follows exponential family distribution with mean $D^{'}(\eta_{itj}^{*})$ and $\epsilon_{itj}^{\triangledown} = X_{itj} - D^{'}(\eta_{itj}^{*})$.

Next, to bound the term $|\epsilon_{mj}^{\triangledown} X_{mk}^{\triangledown}|$,

$$P\left\{|\epsilon_{mj}^{\triangledown} X_{mk}^{\triangledown}| \geq \log(T)\right\} \leq \underbrace{P\left\{\epsilon_{mj}^{\triangledown} X_{mk}^{\triangledown} \geq \log(T)\right\}}_{\mathbb{I}_1} + \underbrace{P\left\{\epsilon_{mj}^{\triangledown} X_{mk}^{\triangledown} \leq -\log(T)\right\}}_{\mathbb{I}_2}. \qquad (113)$$

For $\mathbb{I}_1$, we have

$$\begin{aligned}
P\left\{\epsilon_{mj}^{\triangledown} X_{mk}^{\triangledown} \geq \log(T)\right\} &= P\left\{\epsilon_{itj}^{\triangledown} X_{itk} \geq \log(T)\right\} \\
&\leq \frac{\mathbb{E}\left(e^{\lambda \epsilon_{itj}^{\triangledown} X_{itk}}\right)}{e^{\lambda \log(T)}} \\
&= \frac{\mathbb{E}\left[\mathbb{E}\left\{e^{\lambda X_{itj} X_{itk} - \lambda X_{itk} D^{'}(\eta_{itj}^{*})} \middle| \boldsymbol{X}_{it(-j)}, \boldsymbol{X}_{i(t-1)}, \Delta_{itj}\right\}\right]}{e^{\lambda \log(T)}} \\
&= \frac{\mathbb{E}\left\{e^{D(\eta_{itj}^{*} + \lambda X_{itk}) - D(\eta_{itj}^{*}) - \lambda X_{itk} D^{'}(\eta_{itj}^{*})}\right\}}{e^{\lambda \log(T)}} \\
&\leq \frac{\mathbb{E}\left\{e^{\lambda^2 X_{itk}^2 D^{''}(\eta_{itj}^{*})}\right\}}{e^{\lambda \log(T)}}, \qquad (114)
\end{aligned}$$

where the second inequality follows the definition of $\epsilon_{itj}^{\triangledown}$, the second equality follows the fact that given $\boldsymbol{X}_{it(-j)}, \boldsymbol{X}_{i(t-1)}, \Delta_{itj}$, X_{itj} follows exponential family distribution, the last inequality follows Taylor expansion. To further bound (114), we will bound $\mathbb{E}[\exp\{\lambda^2 X_{itk}^2 D^{''}(\eta_{itj}^{*})\}]$

first. Let $\lambda = \sqrt{1/c_d}$. For some $a \in [0,1]$, by a simple calculation of moment generating function of X_{itk}^2,

$$\begin{aligned}
&\mathbb{E}\left\{e^{X_{itk}^2 D^{''}(\eta_{itj}^*)/c_d}\right\} \\
&= \widetilde{A}_{itk}\left\{D^{''}\left(\eta_{itj}^*\right)/c_d; \boldsymbol{\omega}^*, \boldsymbol{\Delta}^*\right\} - \widetilde{A}_{itk}\left\{0; \boldsymbol{\omega}^*, \boldsymbol{\Delta}^*\right\} \\
&= \frac{D^{''}\left(\eta_{itj}^*\right)}{c_d}\frac{\partial}{\partial u}\widetilde{A}_{itk}\left\{0; \boldsymbol{\omega}^*, \boldsymbol{\Delta}^*\right\} + \frac{\left\{D^{''}\left(\eta_{itj}^*\right)\right\}^2}{c_d^2}\frac{\partial^2}{\partial u^2}\widetilde{A}_{itk}\left\{a\lambda^2 D^{''}\left(\eta_{itj}^*\right); \boldsymbol{\omega}^*, \boldsymbol{\Delta}^*\right\} \\
&\leq \frac{c_v}{c_d}D^{''}\left(\eta_{itj}^*\right) + \frac{c_h}{c_d^2}\left\{D^{''}\left(\eta_{itj}^*\right)\right\}^2 \\
&\leq \frac{c_v c_u}{c_d} + \frac{c_h c_u^2}{c_d^2}, \qquad (115)
\end{aligned}$$

where the second inequality follows Assumptions 13 and 14, and the last inequality follows Assumption 12. Substitute (115) into (114) yields

$$P\left\{\epsilon_{mj}^{\triangledown} X_{mk}^{\triangledown} \geq \log(T)\right\} \leq \frac{c_v c_u c_d + c_h c_u^2}{c_d^2 T^{1/c_d}}. \qquad (116)$$

Similarly, the bound of $\mathbb{I}_2$ is

$$P\left\{\epsilon_{mj}^{\triangledown} X_{mk}^{\triangledown} \leq -\log(T)\right\} \leq \frac{c_v c_u c_d + c_h c_u^2}{c_d^2 T^{1/c_d}}. \qquad (117)$$

Combining (116) with (117), we have

$$P\left\{\left|\epsilon_{mj}^{\triangledown} X_{mk}^{\triangledown}\right| \geq \log(T)\right\} \leq \frac{2c_v c_u c_d + c_h c_u^2}{c_d^2 T^{1/c_d}}. \qquad (118)$$

G.4 Some Technical Lemmas

Lemma 31 (Proposition 3.5 in Zajkowski (2019)) *Let* $S_d(x) = \sum_{i_1,\ldots,i_d}^n a_{i_1,\ldots,i_d}\phi_{i_1}\cdots\phi_{i_d}$ *and* $\mathbf{A}$ *be a multi-indexed array of* $[a_{i_1,\ldots,i_d}]_{i_1,\ldots,i_d=1}^n$ *with* $a_{i_1,\ldots,i_d} \in \mathbb{R}$. *When* $\|\boldsymbol{\phi}\|_{\psi_d} < \infty$,

$$P\left\{\left|S_d(\boldsymbol{\phi}) - \mathbb{E}S_d(\boldsymbol{\phi})\right| \geq t\right\} \leq 2e^{-g(t)},$$

where

$$g(t) = \begin{cases} \frac{t^2}{16C^2\|S_d(\boldsymbol{\phi})\|_{\psi_1}^2}, & \text{if } 0 \leq t \leq 4C\|S_d(\boldsymbol{\phi})\|_{\psi_1}, \\ \frac{t}{2C\|S_d(\boldsymbol{\phi})\|_{\psi_1}} - 1, & \text{if } t > 4C\|S_d(\boldsymbol{\phi})\|_{\psi_1}, \end{cases}$$

and C *is the absolute constant.*

Appendix H. Binary Ising Model

In this section, we perform numerical studies for the binary Ising model with correlated replicates and unmeasured confounders. We compare our proposal to that of Ravikumar et al. (2011). We first generate $\mathbf{\Theta}$ described in Section 5.1, but set non-zero entries in $\mathbf{\Theta}$ from a Uniform distribution with support $[-0.5, -0.25] \cup [0.25, 0.5]$. Then, we generate the piecewise constant unmeasured confounders $\mathbf{U}_i$ as described in Section 5.1. Given $\mathbf{U}_i$ and $\mathbf{\Theta}$, we apply Gibbs sampler to generate $\boldsymbol{X}_{11}, \boldsymbol{X}_{21}, \ldots, \boldsymbol{X}_{n1}$, i.e., the first replicate for all subjects. Suppose that $x_{l11}, x_{l12}, \ldots, x_{l1p}$ are generated from the lth iteration of Gibbs sampler and we have obtained $\boldsymbol{X}_{11}, \boldsymbol{X}_{21}, \ldots, \boldsymbol{X}_{(i-1)1}$, then

$$X_{(l+1)1j} \sim \text{Bernoulli} \left\{ \frac{\exp\left(\theta_{jj} + \sum_{k \neq j} \theta_{jk} x_{l1k} + \sum_{m=p+1}^{p+q} \theta_{jm} u_{i1m}\right)}{1 + \exp\left(\theta_{jj} + \sum_{k \neq j} \theta_{jk} x_{l1k} + \sum_{m=p+1}^{p+q} \theta_{jm} u_{i1m}\right)} \right\},$$

where $j = 1, \ldots, p$. Note that we take the first 10^4 generated samples as burn-in samples, and collect one sample every 10^3 iterations (Ravikumar et al., 2010; Tan et al., 2014).

Then given the ith independent sample, we obtain $\boldsymbol{X}_{i2}, \boldsymbol{X}_{i3}, \ldots, \boldsymbol{X}_{iT}$ using similar Gibbs sampler procedure but the distribution for $(l+1)$th iteration is now

$$X_{i(l+1)j} \sim \text{Bernoulli} \left\{ \frac{\exp\left(\theta_{jj} + \sum_{k \neq j} \theta_{jk} x_{l1k} + \sum_{m=p+1}^{p+q} \theta_{jm} u_{itm} + \sum_{k=1}^{p} \alpha_{jk} x_{i(t-1)k}\right)}{1 + \exp\left(\theta_{jj} + \sum_{k \neq j} \theta_{jk} x_{l1k} + \sum_{m=p+1}^{p+q} \theta_{jm} u_{itm} + \sum_{k=1}^{p} \alpha_{jk} x_{i(t-1)k}\right)} \right\},$$

where $j = 1, \ldots, p$, $\mathbf{x}_{il}$ are samples obtained from the lth iterations and $\boldsymbol{\alpha}_j$ is the jth row of a diagonal transition matrix $\mathbf{A}$ described in Section 5.2.

We set $n = 200$, $T = 10$, $p = 20$, and $q = 5$ and the results are shown in Figure 7. For our proposal, we consider a fine-grid of λ, set $\beta = 0.01$, and vary γ in three different values: 0.5, 1, and 2. We see that our proposal outperforms Ravikumar et al. (2011), which ignores the correlated replicates and unmeasured confounders.

Appendix I. Restrictiveness of Forcing ϕ_0, σ_m, $\Delta_{\max}$, and τ as Constants

In this discussion, we address the implications and limitations of constraining ϕ_0, σ_m, $\Delta_{\max}$, and τ to be constants. The parameter σ_m serves as an upper bound for both σ_m^{ϵ}, the standard deviation of the error term, and κ, the upper bound for the operator norm of $\mathbf{\Sigma}_{jj}$. It is a prevalent practice in statistical literature to treat σ_m as a constant. The primary limitation of this assumption is that it implies a bounded variation in the error term and the data, a premise that is reasonable across numerous scientific disciplines.

The parameter ϕ_0, defined in Assumption 6, bounds the largest eigenvalue of the covariance matrix of $\boldsymbol{Y}_j$ for each j in the set $\{1, \ldots, p\}$. Assuming a constant eigenvalue for data covariance is also a common approach in statistical analyses. However, this assumption might not be suitable when dealing with data that exhibits highly heterogeneous features or where there is significant natural variability.

Regarding the parameters $\Delta_{\max}$ and τ, $\Delta_{\max} = \Delta_m + 1$, where Δ_m is the maximum difference between any two consecutive elements in the sequence $\Delta_{i1j}^*, \Delta_{i2j}^*, \ldots, \Delta_{iTj}^*$. Similarly, τ is defined as the maximum number of jumps in the sequence, mathematically

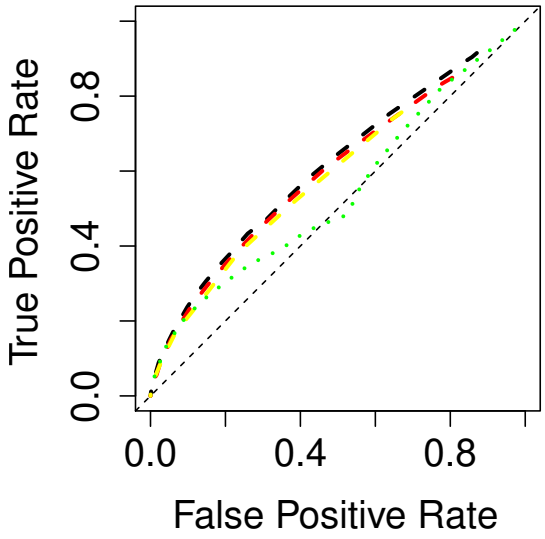


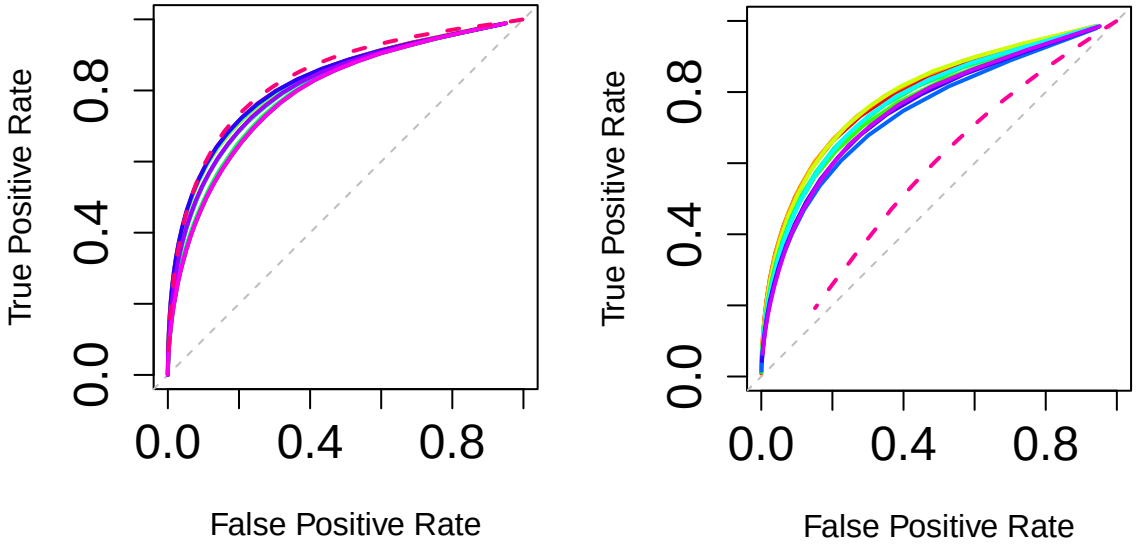
Figure 7: Result for binary Ising model with correlated replicates and unmeasured confounders. For our proposal, we set $\beta = 0.01$ and three different values of γ: $\gamma = 0.5$ (yellow short-dashed), $\gamma = 1$ (red short-dashed), and $\gamma = 2$ (black short-dashed). The line with green dots represent Ravikumar et al. (2011).

represented as $\tau = \max_{i,j} \sum_{t=2}^{T} I(\Delta_{itj} \neq \Delta^*_{i(t-1)j})$. The primary constraint in fixing $\Delta_{\max}$ and τ as constants lies in the assumption that unmeasured confounders do not exhibit excessive variability, and that the magnitude of these changes is finite. In conclusion, while simplifying assumptions on ϕ_0, σ_m, $\Delta_{\max}$, and τ enhance the tractability and comparability of our model, they bring certain limitations, particularly in scenarios involving high data variability and heterogeneity.

Appendix J. Numerical Study about Prior Knowledge of Knot Number and Location

In this section, we compare our proposal with Tan et al. (2016), particularly focusing on scenarios where Tan et al. (2016) have prior knowledge of knot number and location, while our proposal operates without such information. This numerical study is divided into two scenarios: independent replicates and correlated replicates, both with piecewise constant unmeasured confounders. The data generation process follows the methods described in Sections 5.1 and 5.3. Consistent with Section 5, we set $n = 50$, $T = 20$ and $p = 100$. In our proposal, β varies within $\{0.05, 0.1, 0.15\}$ and γ within $\{1, 1.5, 2\}$. For Tan et al. (2016), we estimate the graph between adjacent knots, adhering to their constant confounder

assumption, and assess the "or" and "and" rules for graph combination. We choose the "or" rule for superior performance, as shown in Figure 8. As depicted in Figure 8(a), Tan et al. (2016)'s method outperforms ours under conditions that perfectly align with their assumptions and their knowledge of knot details. Conversely, in 8(b), our approach is more effective in scenarios with correlated replicates, where Tan et al. (2016)'s assumption of independent replicates does not hold.



(a) (b)

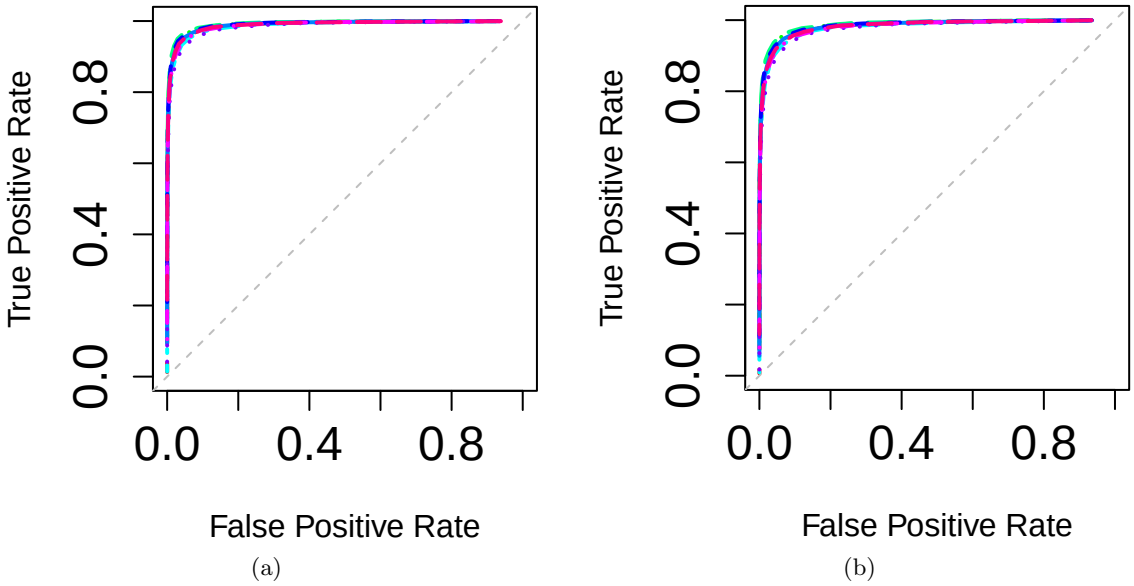
Figure 8: Results for comparison between our proposal and Tan et al. (2016) when Tan et al. (2016) have prior knowledge of knots number and location. Panels (a) and (b) illustrate results for independent and correlated replicates with piecewise constant unmeasured confounders, respectively. For our approach (solid line), we adjust β within $\{0.05, 0.1, 0.15\}$ and γ within $\{1, 1.5, 2\}$. The results from Tan et al. (2016) are depicted as a pink dashed line.

Appendix K. Numerical Study under Scenarios $T \gg n$ and $T \ll n$

In this section, we explore scenarios where $T \gg n$ and $T \ll n$. This study considers two scenarios: correlated replicates with both constant and piecewise constant unmeasured confounders. For $T \gg n$, the parameters are set to $n = 10$, $T = 100$, $p = 30$, $l = 5$ and $R = 2$, while for $T \ll n$, we use $n = 100$, $T = 10$, $p = 30$, $l = 5$ and $R = 2$. The data generation process follows the approach detailed in Section 5.3. Within this framework, β varies within $\{0.01, 0.02, 0.03\}$ and γ within $\{1, 1.5, 2\}$. The outcomes for constant and piecewise constant unmeasured confounders are presented in Figures 9 and 10, respectively.

Figure 9 indicates that our proposal's performance is similar in both $T \gg n$ and $T \ll n$ settings. This suggests that when unmeasured confounders are constant across replicates, the estimation error of our proposal is not significantly affected by the relationship between

n and T. However, Figure 10 reveals a notably poorer performance in the $T \ll n$ scenario compared to $T \gg n$ for piecewise constant unmeasured confounders. This finding aligns with our previous analysis (17), which indicated that a convergence to zero in estimation error requires T approaching infinity when $T \gg n$, and both T and n approaching infinity when $T \ll n$.



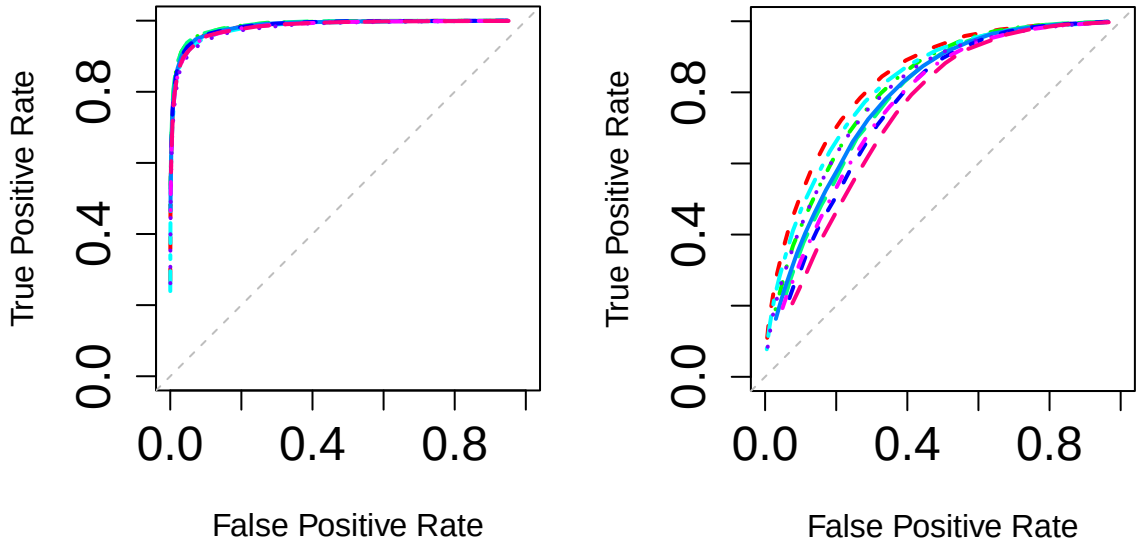
(a) (b)

Figure 9: Results for evaluation of the performance of our proposal when $T \gg n$ and $T \ll n$, both in the context of constant unmeasured confounders with a sparse transition matrix A. Panels (a) and (b) illustrate the results for $T \gg n$ and $T \ll n$, respectively. We vary β in the set $\{0.01, 0.02, 0.03\}$ and γ in the set $\{1, 1.5, 2\}$.

Appendix L. Numerical Study with Graphical Model Considering Dependent Data

In this section, we compare our proposal with Zapata et al. (2022), focusing on estimating graphical methods for dependent data. This study encompasses two scenarios: correlated replicates both without and with unmeasured confounders, paralleling the discussions in Sections 5.2 and 5.3.

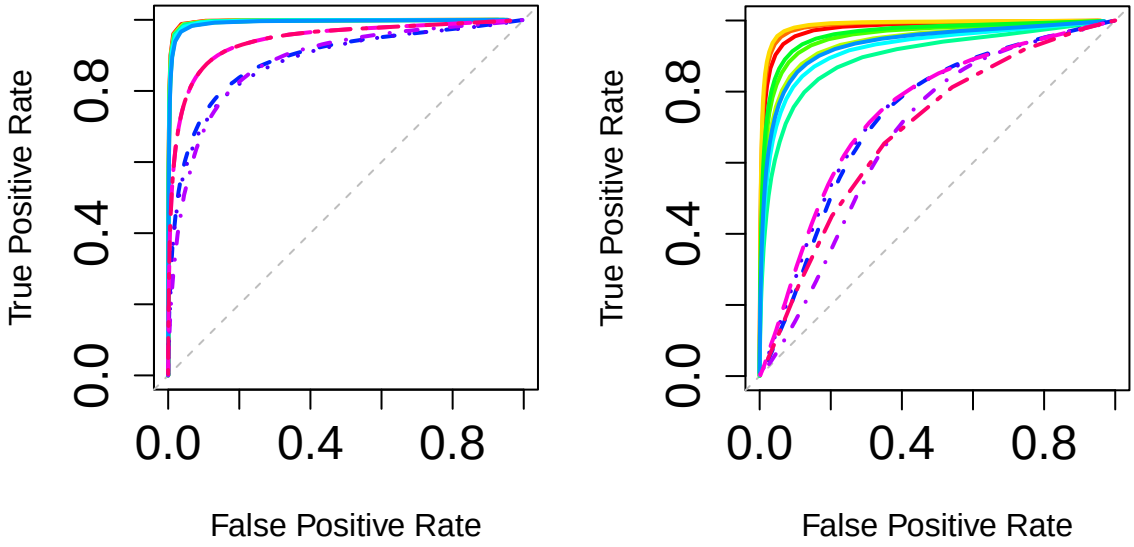
Adhering to the settings in Section 5, we set parameters $n = 50$, $T = 20$, $p = 100$, $l = 5$ and $R = 2$. The data generation process mirrors that used in Sections 5.2 and 5.3. For our approach, we adjust β within $\{0.05, 0.1, 0.15\}$ and γ within $\{1, 1.5, 2\}$. For Zapata et al. (2022), we utilize a fine grid to select the optimal tuning parameters and showcase its best performance. We also include the results of Friedman et al. (2008), Meinshausen and Bühlmann (2006), Chandrasekaran et al. (2010) and Tan et al. (2016) in this numerical study for comparison. The outcomes are presented in Figures 11 to 12. Specifically, Figure 11 illustrates the scenario with correlated replicates without unmeasured confounders, and



(a) (b)

Figure 10: Results for evaluation of the performance of our proposal when $T \gg n$ and $T \ll n$, both in the context of piecewise constant unmeasured confounders with a sparse transition matrix A. Panels (a) and (b) illustrate the results for $T \gg n$ and $T \ll n$, respectively. We vary β in the set $\{0.01, 0.02, 0.03\}$ and γ in the set $\{1, 1.5, 2\}$.

Figure 12 represents the scenario with correlated replicates with unmeasured confounders. In both scenarios, our proposal demonstrates superior performance compared to Zapata et al. (2022). Additionally, Zapata et al. (2022) outperforms other methods, with the exception of our proposal, as shown in Figure 11. It displays comparable performance to other methods in Figure 12. This outcome is reasonable, considering that only Zapata et al. (2022) and our proposal are tailored to handle dependent data, and Zapata et al. (2022) do not consider the existence of unmeasured confounders.

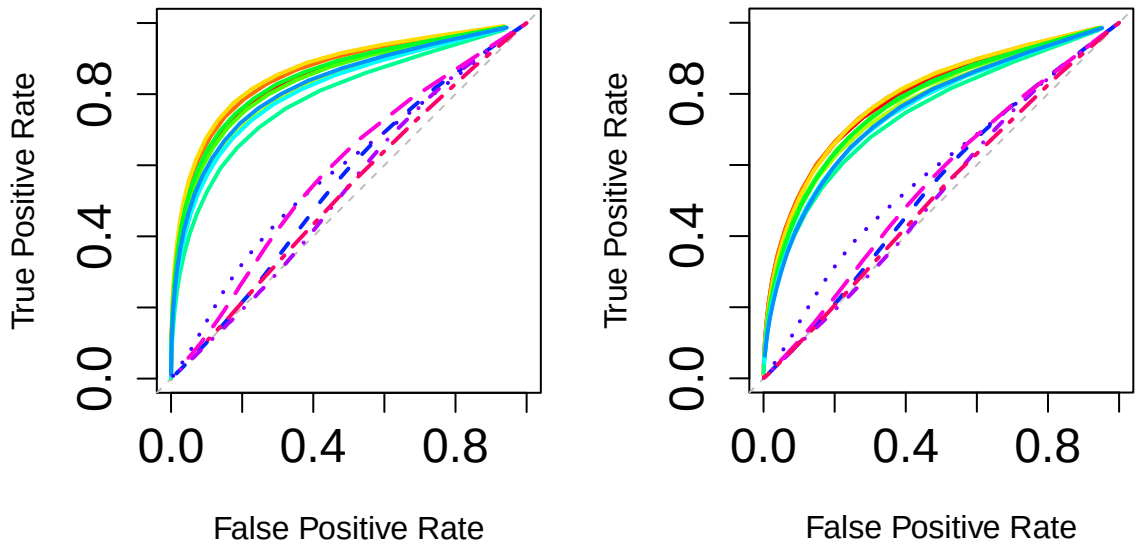


(a) (b)

Figure 11: Results for correlated replicates without unmeasured confounders. Panels (a) and (b) correspond to diagonal and sparse transition matrices, respectively. For our proposal (solid line), we set β within $\{0.05, 0.1, 0.15\}$ and γ within $\{1, 1.5, 2\}$. Other details are as in Figure 3.

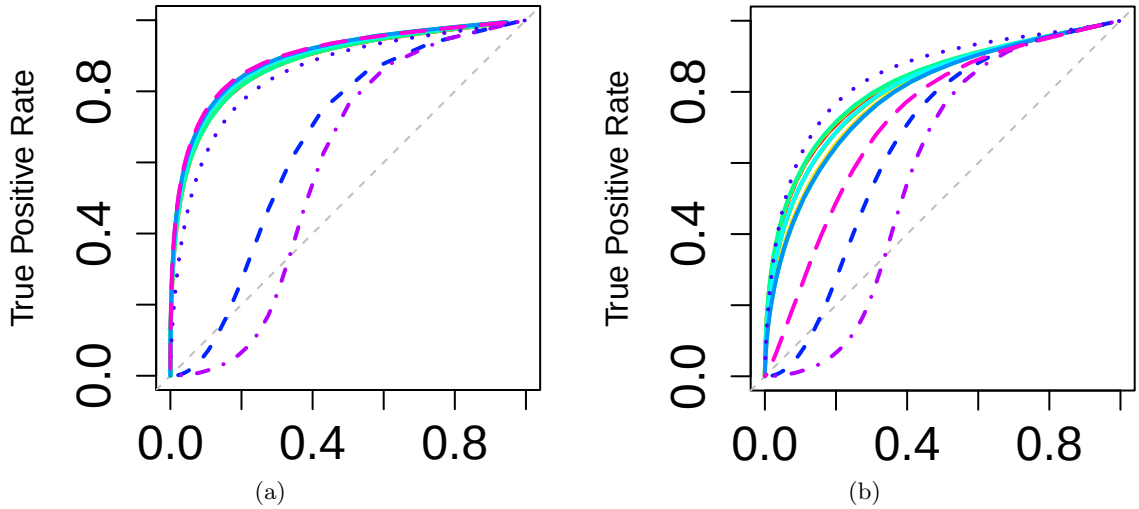
Appendix M. Numerical Study for Sensitivity Check

In this section, we conduct a sensitivity analysis for our tuning parameters β and γ. We vary the values of these parameters to assess if there are significant performance differences in our proposal under different tuning settings. The chosen sets of tuning parameters and the corresponding results are showcased in Figures 13–15. Our analysis reveals that the performance of our proposal is relatively unaffected by variations in β and γ.



(a) (b)

Figure 12: Results for constant and piecewise constant unmeasured confounders with sparse transition matrix. Panels (a) and (b) correspond to constant and piecewise constant unmeasured confounders, respectively. For our proposal (solid line), we set β within $\{0.05, 0.1, 0.15\}$ and γ within $\{1, 1.5, 2\}$. Other details are as in Figure 3.



(a) (b)

Figure 13: Results for independent replicates with unmeasured confounders in Section 5.1. Panels (a) and (b) correspond to the results for constant and piecewise constant unmeasured confounders, respectively. For our proposal (solid line), we set β within $\{0.05, 0.1, 0.15\}$ and γ within $\{1, 1.5, 2\}$. Other details are as in Figure 3.

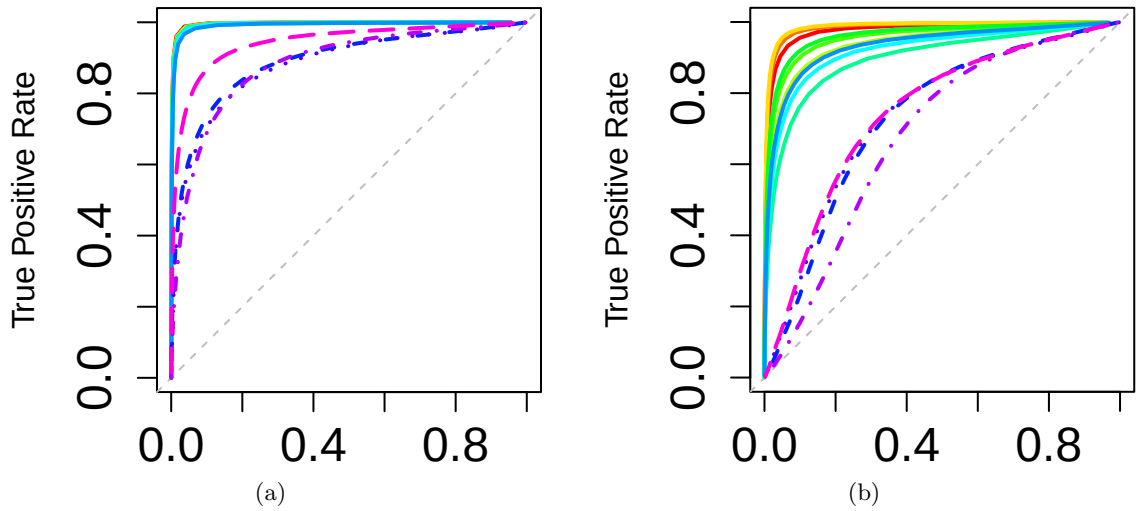


Figure 14: Results for correlated replicates without unmeasured confounders in Section 5.2. Panels (a) and (b) correspond to diagonal and sparse transition matrices, respectively. For our proposal (solid line), we set β within $\{0.05, 0.1, 0.15\}$ and γ within $\{1, 1.5, 2\}$. Other details are as in Figure 3.

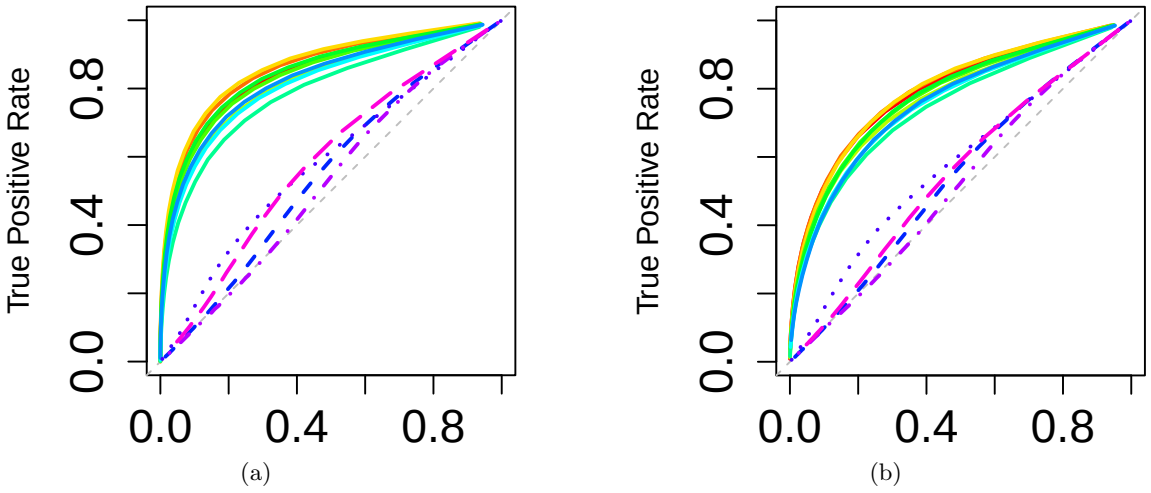


Figure 15: Results for constant and piecewise constant unmeasured confounders with sparse transition matrix $\mathbf{A}$ in Section 5.3. Panels (a) and (b) correspond to constant and piecewise constant unmeasured confounders, respectively. For our proposal (solid line), we set β within $\{0.01, 0.02, 0.3\}$ and γ within $\{1, 1.5, 2\}$. Other details are as in Figure 3.

References

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

Sumanta Basu and George Michailidis. Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, 43(4):1535–1567, 2015.

Bharat B Biswal, Maarten Mennes, Xi-Nian Zuo, Suril Gohel, Clare Kelly, Steve M Smith, Christian F Beckmann, Jonathan S Adelstein, Randy L Buckner, Stan Colcombe, et al. Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences*, 107(10):4734–4739, 2010.

Peter Bühlmann and Sara Van De Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media, 2011.

Tony Cai, Weidong Liu, and Xi Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494): 594–607, 2011.

Tony Cai, Zhao Ren, and Harrison H Zhou. Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electronic Journal of Statistics*, 10(1):1–59, 2016.

Raymond J Carroll, David Ruppert, Leonard A Stefanski, and Ciprian M Crainiceanu. *Measurement Error in Nonlinear Models: a Modern Perspective*. CRC press, 2006.

Venkat Chandrasekaran, Pablo A Parrilo, and Alan S Willsky. Latent variable graphical model selection via convex optimization. In *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, pages 1610–1613. IEEE, 2010.

Tung-Ming Chang, Rei-Cheng Yang, Ching-Tai Chiang, Chen-Sen Ouyang, Rong-Ching Wu, Sebastian Yu, and Lung-Chang Lin. Delay maturation in occipital lobe in girls with inattention subtype of attention-deficit hyperactivity disorder. *Clinical EEG and Neuroscience*, 51(5):325–330, 2020.

Shizhe Chen, Daniela Witten, and Ali Shojaie. Selection and estimation for mixed graphical models. *Biometrika*, 102(1):47–64, 2015.

Jie Cheng, Tianxi Li, Elizaveta Levina, and Ji Zhu. High-dimensional mixed graphical models. *Journal of Computational and Graphical Statistics*, 26(2):367–378, 2017.

Jianqing Fan, Han Liu, Yang Ning, and Hui Zou. High dimensional semiparametric latent graphical model for mixed data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2):405–421, 2017.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Dan Geiger, David Heckerman, Henry King, and Christopher Meek. Stratified exponential families: graphical models and model selection. *The Annals of Statistics*, 29(2):505–529, 2001.

Fan Guo, Steve Hanneke, Wenjie Fu, and Eric P Xing. Recovering temporally rewiring networks: A model-based approach. In *Proceedings of the 24th international conference on Machine learning*, pages 321–328, 2007.

Eric C Hall, Garvesh Raskutti, and Rebecca Willett. Inference of high-dimensional autoregressive generalized linear models. *arXiv preprint arXiv:1605.02693*, 2016.

Steve Hanneke, Wenjie Fu, and Eric P Xing. Discrete temporal models of social networks. *Electronic Journal of Statistics*, 4:585–605, 2010.

Christian Houdré and Patricia Reynaud-Bouret. Exponential inequalities, with constants, for u-statistics of order two. In *Stochastic Inequalities and Applications*, pages 55–69. Springer, 2003.

Mladen Kolar, Le Song, Amr Ahmed, and Eric P Xing. Estimating time-varying networks. *The Annals of Applied Statistics*, 4(1):94–123, 2010.

Tony Lancaster. The incidental parameter problem since 1948. *Journal of Econometrics*, 95(2):391–413, 2000.

Jason D Lee and Trevor J Hastie. Learning the structure of mixed graphical models. *Journal of Computational and Graphical Statistics*, 24(1):230–253, 2015.

Kevin H Lee and Lingzhou Xue. Nonparametric finite mixture of Gaussian graphical models. *Technometrics*, 60(4):511–521, 2018.

Yubu Lee, Bo-yong Park, Oliver James, Seong-Gi Kim, and Hyunjin Park. Autism spectrum disorder related functional connectivity changes in the language network in children, adolescents and adults. *Frontiers in Human Neuroscience*, 11:418, 2017.

Chinghway Lim and Bin Yu. Estimation stability with cross-validation (ES-CV). *Journal of Computational and Graphical Statistics*, 25(2):464–492, 2016.

Lina Lin, Mathias Drton, and Ali Shojaie. Estimation of high-dimensional graphical models using regularized score matching. *Electronic Journal of Statistics*, 10(1):806–854, 2016.

Han Liu, Kathryn Roeder, and Larry Wasserman. Stability approach to regularization selection (stars) for high dimensional graphical models. *Advances in Neural Information Processing Systems*, 23, 2010.

Marloes Maathuis, Mathias Drton, Steffen Lauritzen, and Martin Wainwright. *Handbook of graphical models*. CRC Press, 2018.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, pages 1436–1462, 2006.

Jonathan D Power, Alexander L Cohen, and Steven M Nelson. Functional network organization of the human brain. *Neuron*, 72(4):665–678, 2011.

Xinghao Qiao, Shaojun Guo, and Gareth M James. Functional graphical models. *Journal of the American Statistical Association*, 114(525):211–222, 2019.

Xinghao Qiao, Cheng Qian, Gareth M James, and Shaojun Guo. Doubly functional graphical models in high dimensions. *Biometrika*, 107(2):415–431, 2020.

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

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

Pradeep Ravikumar, Martin J Wainwright, Garvesh Raskutti, and Bin Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

Phillippe Rigollet and Jan-Christian Hütter. High dimensional statistics. *Lecture Notes for Course 18S997*, 2015.

Adam J Rothman, Peter J Bickel, Elizaveta Levina, and Ji Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

Mark Rudelson and Roman Vershynin. Hanson-wright inequality and Sub-Gaussian concentration. *Electronic Communications in Probability*, 18, 2013.

Purnamrita Sarkar and Andrew W Moore. Dynamic social network analysis using latent space models. In *Advances in Neural Information Processing Systems*, pages 1145–1152, 2006.

Yiyuan She, Shao Tang, and Qiaoya Zhang. Indirect gaussian graph learning beyond gaussianity. *IEEE Transactions on Network Science and Engineering*, 7(3):918–929, 2019.

Siqi Sun, Mladen Kolar, and Jinbo Xu. Learning structured densities via infinite dimensional exponential families. In *Advances in Neural Information Processing Systems*, pages 2287–2295, 2015.

Tingni Sun and Cun-Hui Zhang. Sparse matrix inversion with scaled lasso. *The Journal of Machine Learning Research*, 14(1):3385–3418, 2013.

Kean Ming Tan, Palma London, Karthik Mohan, Su-In Lee, Maryam Fazel, and Daniela Witten. Learning graphical models with hubs. *The Journal of Machine Learning Research*, 15(1):3297–3331, 2014.

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

Kean Ming Tan, Yang Ning, Daniela Witten, and Han Liu. Replicates in high dimensions, with applications to latent variable graphical models. *Biometrika*, 103(4):761–777, 2016.

Kean Ming Tan, Junwei Lu, Tong Zhang, and Han Liu. Layer-wise learning strategy for nonparametric tensor product smoothing spline regression and graphical models. *The Journal of Machine Learning Research*, 20(119):1–38, 2019.

Ryan J Tibshirani. Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics*, 42(1):285–323, 2014.

Sara Van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

JianZhe Wang, TianZi Jiang, Qingjiu Cao, and Yufeng Wang. Characterizing anatomic differences in boys with attention-deficit/hyperactivity disorder with the use of deformation-based morphometry. *American Journal of Neuroradiology*, 28(3):543–547, 2007.

Yu-Xiang Wang, James Sharpnack, Alexander J Smola, and Ryan J Tibshirani. Trend filtering on graphs. *The Journal of Machine Learning Research*, 17(1):3651–3691, 2016.

Changjing Wu, Hongyu Zhao, Huaying Fang, and Minghua Deng. Graphical model selection with latent variables. *Electronic Journal of Statistics*, 11(2):3485–3521, 2017.

Eunho Yang, Genevera Allen, Zhandong Liu, and Pradeep K Ravikumar. Graphical models via generalized linear models. In *Advances in Neural Information Processing Systems*, pages 1358–1366, 2012.

Eunho Yang, Pradeep Ravikumar, Genevera Allen, and Zhandong Liu. On poisson graphical models. In *NIPS*, pages 1718–1726, 2013.

Eunho Yang, Yulia Baker, Pradeep Ravikumar, Genevera Allen, and Zhandong Liu. Mixed graphical models via exponential families. In *Artificial intelligence and statistics*, pages 1042–1050. PMLR, 2014.

Eunho Yang, Pradeep Ravikumar, Genevera Allen, and Zhandong Liu. Graphical models via univariate exponential family distributions. *The Journal of Machine Learning Research*, 16(1):3813–3847, 2015.

Zhuoran Yang, Yang Ning, and Han Liu. On semiparametric exponential family graphical models. *The Journal of Machine Learning Research*, 19(1):2314–2372, 2018.

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

Krzysztof Zajkowski. Norms of sub-exponential random vectors. *Statistics & Probability Letters*, 152:147–152, 2019.

Javier Zapata, Sang-Yun Oh, and Alexander Petersen. Partial separability and functional graphical models for multivariate gaussian processes. *Biometrika*, 109(3):665–681, 2022.

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

Shuheng Zhou, John Lafferty, and Larry Wasserman. Time varying undirected graphs. *Machine Learning*, 80(2-3):295–319, 2010.