Journal of Machine Learning Research 22 (2021) 1-62 Submitted 4/20; Revised 7/21; Published 10/21

Simultaneous Change Point Inference and Structure Recovery for High Dimensional Gaussian Graphical Models

Bin Liu LIUBIN0145@GMAIL.COM
Department of Statistics and Data Science, School of Management
Fudan University
Shanghai, 200433, China

Xinsheng Zhang XSZHANG@FUDAN.EDU.CN
Department of Statistics and Data Science, School of Management
Fudan University
Shanghai, 200433, China

Yufeng Liu YFLIU@EMAIL.UNC.EDU
Department of Statistics and Operations Research
Department of Genetics
Department of Biostatistics
Carolina Center for Genome Sciences
Lineberger Comprehensive Caner Center
University of North Carolina at Chapel Hill, U.S.A

Editor: Zaid Harchaoui

Abstract

In this article, we investigate the problem of simultaneous change point inference and structure recovery in the context of high dimensional Gaussian graphical models with possible abrupt changes. In particular, motivated by neighborhood selection, we incorporate a threshold variable and an unknown threshold parameter into a joint sparse regression model which combines p ℓ_1-regularized node-wise regression problems together. The change point estimator and the corresponding estimated coefficients of precision matrices are obtained together. Based on that, a classifier is introduced to distinguish whether a change point exists. To recover the graphical structure correctly, a data-driven thresholding procedure is proposed. In theory, under some sparsity conditions and regularity assumptions, our method can correctly choose a homogeneous or heterogeneous model with high accuracy. Furthermore, in the latter case with a change point, we establish estimation consistency of the change point estimator, by allowing the number of nodes being much larger than the sample size. Moreover, it is shown that, in terms of structure recovery of Gaussian graphical models, the proposed thresholding procedure achieves model selection consistency and controls the number of false positives. The validity of our proposed method is justified via extensive numerical studies. Finally, we apply our proposed method to the S&P 500 dataset to show its empirical usefulness.

Keywords: Change point detection, Gaussian graphical models, Hard thresholding, Neighborhood selection, Sparsity

1. Introduction

Networks are fundamental in representing dependence relationships among the nodes and have various real applications such as biology, finance, and social science. It is of great importance to explore the networks and uncover the underlying data generating models associated with networks. A network can be described by a graph $G = (V, E)$, where $V = \{1, \ldots, p\}$ is a vertex set and $E \subseteq V \times V$ is an edge set. Each node in V corresponds to an element of a p-dimensional random vector $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$. In a typical network, the edge set E captures the conditional dependence among the nodes in V. The Gaussian graphical model (GGM) is a popular method to describe the networks, where we assume $\boldsymbol{X} = (X_1, \ldots, X_p)^\top \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Specifically, the node pair $(a, b) \in E$ if X_a and X_b are conditionally dependent given the remaining variables $\boldsymbol{X}_{\setminus\{a,b\}} := \{X_k : 1 \leq k \leq p, k \neq a, b\}$. In other words, $(a, b) \notin E$ if and only if $X_a \perp X_b | \boldsymbol{X}_{\setminus\{a,b\}}$. Under the GGM, it is well known (Lauritzen 1996) that there is a deterministic relationship between the conditional independence and the $p \times p$ precision matrix $\boldsymbol{\Omega} = (\omega_{ab})_{1 \leq a,b \leq p}$, where $\boldsymbol{\Omega} := \boldsymbol{\Sigma}^{-1}$. In particular, $X_a \perp X_b | \boldsymbol{X}_{\setminus\{a,b\}} \Leftrightarrow \omega_{ab} = 0$.

In the past few years, motivated by the high throughput data analysis, a number of papers on time-invariant networks have appeared, especially in high dimensional settings. In those cases, the samples are assumed independently and identically distributed (i.i.d) draws from p-dimensional Gaussian distributions with the dimension being much larger than the sample size. Specifically, Meinshausen and Bühlmann (2006) proposed a neighborhood selection approach for the structure recovery, in a row-by-row fashion. Peng et al. (2009) extended Meinshausen and Bühlmann (2006) by estimating all neighborhoods jointly, and their procedure has an improvement on several networks such as those with hubs. Another popular approach to estimate the graphical structure is based on the ℓ_1 penalized likelihood (Yuan and Lin (2007); Banerjee et al. (2008); Friedman et al. (2008)). Instead of likelihood-based methods, Liu (2013) considered to estimate the Gaussian graphical model based on a multiple testing procedure, and Cai et al. (2011) proposed a constrained ℓ_1 minimization method for estimating the precision matrix under a broader distributional assumption. In addition, there are some other extensions for estimating GGMs including score matching for non-negative data (Yu et al. (2019)), random forests for discrete, continuous, and mixed variables (Fellinghauer et al. (2013)), as well as semiparametric methods for graph estimation with joint additive models (Voorman et al. (2014)).

In many real applications, however, the corresponding networks are typically non-stationary over time, known as time-varying networks. For example, in a gene regulatory network, a particular drug treatment can result in significant changes of the dependence structure among the associated genes; in political science, it is likely that the relationships among the campaigners undergo great changes before and after the election; in neuroscience, a stimulus can change the associations among different parts of the brain. Therefore, the aforementioned methods designed for i.i.d cases are no longer applicable in those examples, and it is desirable to construct a method for time-varying networks.

In this article, we consider the problem of high dimensional Gaussian graphical models with one possible abrupt change, where the number of nodes can be much larger than the sample size. Specifically, let $\boldsymbol{\Omega}^{(1)}$ and $\boldsymbol{\Omega}^{(2)}$ be two well-defined precision matrices. Let $\boldsymbol{X}^t = (X_1^t, \ldots, X_p^t)^\top$ be the t-th observation of a p-dimensional time-ordered data sequence,

where we assume $\boldsymbol{X}^t \sim N(\mathbf{0}, \boldsymbol{\Omega}_t^{-1})$ for $1 \leq t \leq T$. Suppose there exists a possible but unknown change point location $\tau_* \in (0,1)$ such that

$$\boldsymbol{\Omega}_t = \boldsymbol{\Omega}^{(1)} \mathbf{1}\big\{1 \leq t \leq \lfloor T\tau_* \rfloor\big\} + \boldsymbol{\Omega}^{(2)} \mathbf{1}\big\{\lfloor T\tau_* \rfloor + 1 \leq t \leq T\big\}. \tag{1}$$

In other words, by (1), if $\boldsymbol{\Omega}^{(1)} = \boldsymbol{\Omega}^{(2)}$, the data are homogeneous and τ_* is not identifiable; if $\boldsymbol{\Omega}^{(1)} \neq \boldsymbol{\Omega}^{(2)}$, the data are heterogeneous and a change point exists. The objective of this paper is to: (1) distinguish whether $\boldsymbol{\Omega}^{(1)}$ equals to $\boldsymbol{\Omega}^{(2)}$ (or whether τ_* exists); (2) once a heterogeneous model is selected, identify the location of τ_* as well as recover the two underlying networks, i.e., the non-zero elements of $\boldsymbol{\Omega}^{(1)}$ and $\boldsymbol{\Omega}^{(2)}$, simultaneously.

In contrast to the large literature on time-invariant networks, much less attention has been paid on the analysis of time-varying networks. By assuming $\boldsymbol{\Omega}^{(1)} \neq \boldsymbol{\Omega}^{(2)}$, a few papers exist in the literature on this topic. For example, by assuming that the covariances change smoothly over time, Zhou et al. (2010) proposed a nonparametric method for estimating time-varying Gaussian graphical structure using the ℓ_1 regularization method. In a different setting that the graphical structure is piece-wise constant, Kolar and Xing (2012) considered the Gaussian graphical structure recovery node-by-node, based on a time-coupled neighborhood selection procedure using the fused-type penalty (Harchaoui and Lévy-Leduc (2010)). These methods are designed for low dimensional problems in the sense that the dimension p is smaller than the sample size T. Recently, driven by modern statistical applications, there is a great need for the high dimensional time-varying network analysis with both discrete and continuous observations. For example, in the discrete case, Kolar et al. (2010) considered the structure recovery of time-varying Ising graphical models with smooth changes, using a pseudo-likelihood approach. As an extension, Roy et al. (2017) investigated the change point estimation in the context of high dimensional Markov random-field models with abrupt changes. In the continuous case, using penalized likelihood, Bybee and Atchadé (2018) proposed a majorize-minimize algorithm for estimating the change point in Gaussian graphical models. Gibberd and Nelson (2017) estimated piece-wise constant Gaussian graphical models using the group-fused graphical lasso, and Gibberd and Roy (2017) provided some theoretical results. Different from the settings in Bybee and Atchadé (2018); Gibberd and Nelson (2017), Yang and Peng (2020) proposed local group graphical lasso estimation under the assumption that the graph topology changes gradually over time.

In this paper, we consider simultaneous change point inference and structure recovery for Gaussian graphical models with a possible abrupt change, in a high dimensional setting with $p \succeq T$. To identify the change point τ_*, motivated by the neighborhood selection procedure (Meinshausen and Bühlmann (2006)), we incorporate a threshold variable $Q_t := t/T$ with $1 \leq t \leq T$ and an unknown threshold parameter $\tau \in (0,1)$ into a joint sparse regression model, by considering the p nodes simultaneously. The change point estimator $\hat{\tau}$ and the corresponding estimated coefficients of precision matrices are obtained via minimizing the joint ℓ_2 loss function with an ℓ_1 penalty. Based on that, a classifier is proposed to distinguish a homogeneous model from a heterogeneous one. Under some sparsity conditions and regularity assumptions, our proposed method can select a true model with high accuracy. Furthermore, once a heterogeneous model is selected, we establish that, with a high probability, $|\hat{\tau} - \tau_*| = O_p(s\log(p)/T)$ as $p, T \to \infty$, where s denotes the overall sparsity of $\boldsymbol{\Omega}^{(1)}$ and $\boldsymbol{\Omega}^{(2)}$. Hence, we allow the dimension p and the sparsity s to grow with T as long as $s\log(p) = o(T)$ holds. Note that our proposed method does not require prior knowledge of

τ_*. In other words, our technique applies to data with or without a change point, which is fundamentally different from the existing works where the i.i.d assumption or the existence of τ_* is typically imposed.

For the structure recovery of the two graphs, a data-driven hard thresholding procedure is constructed, built on the initial coefficient estimation. In particular, we tend to specify a relatively small regularization parameter for obtaining the change point estimator $\hat{\tau}$ as well as two estimated dense graphs. Then, a hard threshold variable is introduced to filter the "noisy edges" (false positives) of initially estimated graphs and yield relatively sparse graphs. To choose the "best" parameter in the threshold variable, a data-driven method is proposed. Theoretically, in terms of the structure recovery of the two graphs, we prove that, with a high probability, the proposed thresholding procedure achieves model selection consistency, which is crucial for controlling the type II error (number of false negatives). Furthermore, our theoretical results guarantee that the number of false positives (type I error) can be bounded by some universal constants. Empirically, we justify the validity of our method via extensive simulation studies. It is shown that the proposed method can efficiently recover various graphical structures. It is worth mentioning that for the structural recovery of GGM, false discovery proportion (FDP) or false discovery rate (FDR) control may be of interest in practice. If there is no change point or the change point location τ_* is known, we can use the multiple testing procedures (Liu (2013)) directly for estimating the graphs with FDR control. For GGM with a possible change point, however, recovering the graphs is a challenging problem due to the unknown change point. Hence, we propose a hard thresholding procedure that utilizes the estimators of the change point as well as the coefficients of precision matrices to recover the graphs. As a by-product of our theory, we can show that FDP $= o_p(1)$ as $(p, T) \to \infty$. To our limited knowledge, it is an open question for the FDR control of GGM with change points.

Note that in this paper, we consider abrupt changes in the context of high dimensional Gaussian graphical models. This is different from the settings of Zhou et al. (2010); Yang and Peng (2020), designed for smooth changes; or Kolar et al. (2010); Roy et al. (2017), designed for discrete observations. Other related papers include Guo et al. (2011); Danaher et al. (2014); Lee and Liu (2015); Cai et al. (2016), where they considered the problem of simultaneously estimating multiple graphs belonging to distinct classes separated at a pre-known location τ_*, which is very different from our setting with unknown τ_*. The closest settings in the literature to the current paper are considered in Kolar and Xing (2012), Bybee and Atchadé (2018), and Gibberd and Nelson (2017). In Kolar and Xing (2012), the number of nodes is assumed fixed and smaller than the sample size. In contrast, we consider a high dimensional problem. Furthermore, it is unclear how to combine the estimated change points obtained from each node as proposed by Kolar and Xing (2012) for consistent estimation. Bybee and Atchadé (2018) focused on proposing a fast algorithm for change point estimation based on the Gaussian penalized likelihood. They did not study any theoretical guarantees of parameter estimation or structure recovery of the graphs. Gibberd and Nelson (2017) considered estimating T graphs simultaneously, which can be computationally intensive especially when p is very large. As compared to Gibberd and Nelson (2017), our method is able to deal with large-scale graphs.

As pointed out by one reviewer, this paper mainly adopts a brute search method for localizing a change point. Hence, as an alternative method, it is possible to use the Fused

Graphical Lasso (FGL) in Danaher et al. (2014) to construct a corresponding loss function and find a change point that minimizes the losses. However, there are essential differences between FGL and our proposed method. First, FGL mainly uses the log-likelihood of GGM and focuses on settings where the change occurring is global in the sense that it affects the joint distribution of all nodes. In contrast, we propose our method in a node-wise way which focuses on settings where the conditional distribution of a single node or a few nodes may see a change. Second, under the framework of our methodology, we can estimate the difference between the two graphs (e.g. $\|\mathbf{\Omega}^{(1)} - \mathbf{\Omega}^{(2)}\|_1$) and obtain the same estimation error bound no matter the change point exists or not. This result motivates us to construct a classifier that can determine whether $\mathbf{\Omega}^{(1)} = \mathbf{\Omega}^{(2)}$ or $\mathbf{\Omega}^{(1)} \neq \mathbf{\Omega}^{(2)}$. Moreover, once a change point is detected, using our algorithms, we can obtain consistent estimation in terms of change point identification and structural recovery. It is unclear whether the above results apply to FGL. Third, from the computational point of view, the overall computational costs for FGL and our method are $O(Tp^3)$ and $O(T^2p^2)$, respectively. Hence, for large graphs with $p \gg T$, it is more computationally efficient to use our algorithms than FGL for change point detection. More importantly, for our proposed method, the computational cost can be further reduced via parallel computing. In the numerical studies, we provide a thorough empirical comparison between our method and the existing techniques.

The rest of this paper is organized as follows. In Section 2, we introduce our methodology for identifying τ_* as well as recovering the underlying networks of $\mathbf{\Omega}^{(1)}$ and $\mathbf{\Omega}^{(2)}$. In Section 3, we derive some theoretical results in terms of the change point identification and structure recovery. In Section 4, we justify the validity of our proposed method via extensive numerical studies. In Section 5, we apply our proposed method to the S&P 500 dataset for analyzing the networks during the financial crisis. Conclusions are provided in Section 6. The proofs of the main results are given in the appendix.

2. Methodology

In this section, we introduce our methodology for identifying the unknown change point location τ_* as well as recovering the graphical structures of $\mathbf{\Omega}^{(1)}$ and $\mathbf{\Omega}^{(2)}$ simultaneously. In Section 2.1, we introduce some definitions and notations. In Section 2.2, we first introduce a threshold variable and an unknown threshold parameter which capture the time-varying networks. Based on that, we propose a joint ℓ_2-loss function by considering the p nodes simultaneously. The change point estimator and the corresponding coefficients are then obtained via minimizing the joint loss function with an ℓ_1 penalty. Furthermore, we also introduce a classifier to detect whether a change point exists. In terms of the structure recovery of graphs, we find that the initial coefficient estimation tends to select "too many" components with non-zero estimated coefficients, resulting in a large type I error (number of false positives). To avoid this problem, a thresholding procedure is proposed.

2.1 Notations

We set X_a^t as the t-th observation for coordinate a with $1 \leq t \leq T$ and $1 \leq a \leq p$. Denote $\mathbf{X}$ as the $T \times p$ observation matrix. We set $\mathbf{X}^t = (X_1^t, \ldots, X_p^t)^\top$ as the t-th row of $\mathbf{X}$ with $1 \leq t \leq T$ and $\mathbf{X}_a = (X_a^1, \ldots, X_a^T)^\top$ as the a-th column of $\mathbf{X}$ with $1 \leq a \leq p$.

For a set S, denote $|S|$ as the cardinality of S. By letting $[p] = \{1, \ldots, p\}$, we define $[p] \setminus S = \{a : 1 \leq a \leq p \text{ and } a \notin S\}$. To simplify notations, we write $[p] \setminus \{a\}$ as $\setminus a$. For a T-dimensional vector $\boldsymbol{w} = (w_1, \ldots, w_T)^\top$, its empirical norm is defined as $\|\boldsymbol{w}\|_T = (T^{-1}\sum_{t=1}^T w_i^2)^{1/2}$. For a vector $\boldsymbol{v} \in \mathbb{R}^p$, denote $|\boldsymbol{v}|_p$ as its ℓ_p-norm with $1 \leq p \leq \infty$. We denote $J(\boldsymbol{v}) = \{1 \leq j \leq p : v_j \neq 0\}$ as the set of non-zero elements of $\boldsymbol{v}$. We set $\mathcal{M}(\boldsymbol{v}) = |J(\boldsymbol{v})|$ as the number of non-zero elements of $\boldsymbol{v}$. For a set J and $\boldsymbol{v} \in \mathbb{R}^p$, denote $\boldsymbol{v}_J$ as the vector in $\mathbb{R}^p$ that has the same coordinates as $\boldsymbol{v}$ on J and zero coordinates on the complement J^c of J. For a matrix $\mathbf{A} = (a_{ij})$, denote $\max_{i,j}|a_{ij}|$ by $\|\mathbf{A}\|_\infty$. For two real numbers a and b, we set $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. For any $x > 0$, denote $\lfloor x \rfloor$ as the largest integer smaller than or equal to x.

2.2 Change point inference

We introduce our methodology to estimate the unknown change point location τ_*. To this end, we first introduce the neighborhood selection procedure. Consider a p-dimensional random vector $\boldsymbol{X} = (X_1, \ldots, X_p)^\top \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. For each node $a \in \{1, \ldots, p\}$, consider the optimal prediction problem for X_a given the remaining variables $\boldsymbol{X}_{\setminus a}$. Let $\tilde{\boldsymbol{\theta}}^a = (\tilde{\theta}_1^a, \ldots, \tilde{\theta}_p^a)^\top \in \mathbb{R}^p$ be the vector of coefficients for the optimal prediction, which is defined as:

$$\tilde{\boldsymbol{\theta}}^a = \underset{\boldsymbol{\theta} \in \mathbb{R}^p, \theta_a = 0}{\operatorname{argmin}} \mathbb{E}\big(X_a - \sum_{k \neq a} \theta_k X_k\big)^2. \tag{2}$$

Let $\boldsymbol{\Omega} = (\omega_{ab})_{a,b=1}^p \in \mathbb{R}^{p \times p}$ be the inverse matrix of $\boldsymbol{\Sigma}$. It is well known (Meinshausen and Bühlmann (2006)) that $\tilde{\theta}_b^a = -\omega_{ab}/\omega_{aa}$ for $b \in V \setminus \{a\}$. In other words, the set of non-zero coordinates in $\tilde{\boldsymbol{\theta}}^a$ corresponds to the set $\{b \in V \setminus \{a\} : \omega_{ab} \neq 0\}$. Furthermore, it is also well known (Lauritzen, 1996) that for Gaussian distributions, $X_a \perp X_b | \boldsymbol{X}_{\setminus\{a,b\}} \Leftrightarrow \omega_{ab} = 0$. Therefore, the b-th coordinate in $\tilde{\boldsymbol{\theta}}^a$ corresponds to the conditional independence of X_a and X_b. We call $\mathrm{NE}_a = \{b \in V \setminus \{a\} : \tilde{\theta}_b^a \neq 0\}$ the neighborhood of node a.

For Gaussian distributions, using $\tilde{\boldsymbol{\theta}}^a$ as defined in (2), for each node a, we can rewrite X_a as the following regression model:

$$X_a = \sum_{k=1}^p \tilde{\theta}_k^a X_k + \epsilon_a, \tag{3}$$

where ϵ_a is independent of $\boldsymbol{X}_{\setminus a}$ and $\epsilon_a \sim N(0, 1/\omega_{aa})$.

Motivated by the above neighborhood selection procedure, we suppose that the graph G has a change point at the location τ_*. Then there exists at least one node whose optimal prediction coefficients as in (2) also have a change at τ_*. To be specific, define $\tilde{\boldsymbol{\delta}}^a = (\tilde{\delta}_1^a, \ldots, \tilde{\delta}_p^a) \in \mathbb{R}^p$ with $\tilde{\delta}_a^a = 0$ as the corresponding change of the optimal prediction coefficients for node a. By definition, $\tilde{\boldsymbol{\delta}}^a = \mathbf{0}$ if there is no change point for node a. Let $Q_t = t/T$ be a threshold variable. Then, similar to (3), we can assume that each node a follows the following regression model:

$$X_a^t = (\mathbf{X}^t)^\top \tilde{\boldsymbol{\theta}}^a + (\mathbf{X}^t)^\top \tilde{\boldsymbol{\delta}}^a \mathbf{1}\{Q_t \leq \tau_*\} + \epsilon_a^t, \text{ for } t = 1, \ldots, T, \text{ and } a = 1, \ldots, p, \tag{4}$$

where $\tilde{\boldsymbol{\theta}}^a = (\tilde{\theta}_1^a, \ldots, \tilde{\theta}_p^a) \in \mathbb{R}^p$ with $\tilde{\theta}_a^a = 0$, and $(\epsilon_a^t)_{1 \leq t \leq T} \sim N(0, 1/\omega_{aa}^t)$ with ϵ_a^t being independent of $\mathbf{X}_{\setminus a}^t$ and $\epsilon_a^{t_1}$ being independent of $\epsilon_a^{t_2}$ for $t_1 \neq t_2$.

By Model (4), for each node $a \in V$, the vector $\tilde{\boldsymbol{\theta}}^a + \tilde{\boldsymbol{\delta}}^a$ corresponds to the vector of the optimal prediction coefficients before τ_*, and $\tilde{\boldsymbol{\theta}}^a$ corresponds to the vector after τ_*, respectively. Consequently, for each node a, during the T observations, we define its neighborhood as:

$$\begin{aligned} \mathrm{NE}_a^{(1)} &= \Big\{ b \in V \setminus \{a\} : \tilde{\theta}_b^a + \tilde{\delta}_b^a \neq 0 \Big\}, \text{ for } 1 \leq t \leq \lfloor T\tau_* \rfloor, \\ \mathrm{NE}_a^{(2)} &= \Big\{ b \in V \setminus \{a\} : \tilde{\theta}_b^a \neq 0 \Big\}, \text{ for } \lfloor T\tau_* \rfloor + 1 \leq t \leq T. \end{aligned}$$

With $\mathrm{NE}_a^{(1)}$ and $\mathrm{NE}_a^{(2)}$ for $1 \leq a \leq p$, we define the edge sets before and after τ_* as:

$$\begin{aligned} E^{(1)} &= \Big\{ (a,b) : a \in \mathrm{NE}_b^{(1)} \cup b \in \mathrm{NE}_a^{(1)} \Big\}, \text{ for } 1 \leq t \leq \lfloor T\tau_* \rfloor, \\ E^{(2)} &= \Big\{ (a,b) : a \in \mathrm{NE}_b^{(2)} \cup b \in \mathrm{NE}_a^{(2)} \Big\}, \text{ for } \lfloor T\tau_* \rfloor + 1 \leq t \leq T. \end{aligned}$$

Note that $E^{(1)}$ and $E^{(2)}$ are the supports of $\boldsymbol{\Omega}^{(1)}$ and $\boldsymbol{\Omega}^{(2)}$, respectively. Our goal of this article is to identify the existence of τ_* as well as recover $E^{(1)}$ and $E^{(2)}$ simultaneously.

After introducing the oracle Model (4), we now present our methodology for identifying τ_*. To this end, some additional notations are needed. Let $\mathbf{X}^t(\tau) = ((\mathbf{X}^t)^\top, (\mathbf{X}^t)^\top \mathbf{1}\{Q_t \leq \tau\})^\top$ be a $2p \times 1$ vector for $1 \leq t \leq T$. Let $\mathbf{X}(\tau)$ be a $T \times 2p$ matrix whose t-th row is $\mathbf{X}^t(\tau)$. We also define the $2p \times 2p$ diagonal matrix $\mathbf{D}(\tau) = \mathrm{diag}\{\|\mathbf{X}_j(\tau)\|_T : j = 1, \ldots, 2p\}$. With these notations, we then introduce our methodology.

Define $\boldsymbol{\beta}^a = ((\boldsymbol{\theta}^a)^\top, (\boldsymbol{\delta}^a)^\top)^\top \in \mathbb{R}^{2p}$. For a fixed τ, define the loss function for node a as:

$$\begin{aligned} \mathcal{L}^a(\boldsymbol{\theta}^a, \boldsymbol{\delta}^a, \tau) &= \frac{1}{T} \sum_{t=1}^{T} \Big(X_a^t - (\mathbf{X}^t)^\top \boldsymbol{\theta}^a - (\mathbf{X}^t)^\top \boldsymbol{\delta}^a \mathbf{1}\{Q_t \leq \tau\} \Big)^2, \\ &= \big\| \mathbf{X}_a - \mathbf{X}(\tau) \boldsymbol{\beta}^a \big\|_T^2. \end{aligned}$$

Then, for a fixed τ, we define the lasso solution $\hat{\boldsymbol{\beta}}^a(\tau) = ((\hat{\boldsymbol{\theta}}^a(\tau))^\top, (\hat{\boldsymbol{\delta}}^a(\tau))^\top)^\top$ as:

$$\hat{\boldsymbol{\beta}}^a(\tau) = \underset{\boldsymbol{\theta}^a, \boldsymbol{\delta}^a \in \mathbb{R}^p; \theta_a^a = 0, \delta_a^a = 0}{\operatorname{argmin}} \|\mathbf{X}_a - \mathbf{X}(\tau)\boldsymbol{\beta}^a\|_T^2 + \lambda_T |\mathbf{D}(\tau)\boldsymbol{\beta}^a|_1, \text{ for } 1 \leq a \leq p, \tag{5}$$

where the non-negative λ_T is the regularization parameter to be specified.

Based on $(\hat{\boldsymbol{\beta}}^a(\tau))_{a=1}^p$, we define

$$H(\tau) = \sum_{a=1}^{p} \Big\{ \big\| \mathbf{X}_a - \mathbf{X}(\tau)\hat{\boldsymbol{\beta}}^a(\tau) \big\|_T^2 + \lambda_T \big| \mathbf{D}(\tau)\hat{\boldsymbol{\beta}}^a(\tau) \big|_1 \Big\}. \tag{6}$$

Finally, our estimator for τ_* is defined as follows:

$$\hat{\tau} = \underset{\tau \in \mathcal{T}}{\operatorname{argmin}}\, H(\tau), \tag{7}$$

where $\mathcal{T} = [t_0, t_1]$ is a prespecified search domain for the change point location τ_*.

Note that if $\tilde{\boldsymbol{\delta}}^a = \mathbf{0}$ for $1 \le a \le p$, the change point location τ_* is not identifiable. Hence, we need to identify its existence. To this end, we first put $\hat{\tau}$ into (5) and obtain $\hat{\boldsymbol{\beta}}^a(\hat{\tau}) = \big((\hat{\boldsymbol{\theta}}^a(\hat{\tau}))^\top, (\hat{\boldsymbol{\delta}}^a(\hat{\tau}))^\top\big)^\top$ with $1 \le a \le p$. Then we define the following classifier:

$$\Phi(\hat{\tau}, \{\hat{\boldsymbol{\beta}}^a(\hat{\tau}), 1 \le a \le p\}) = \mathbf{1}\Big\{ \sum_{a=1}^{p} |\hat{\boldsymbol{\delta}}^a(\hat{\tau})|_1 \ge K_0 s_1 \lambda_T \Big\}, \tag{8}$$

where K_0 is some constant, s_1 is the overall sparsity of the graphs (see Assumption 2), and λ_T is defined in (5).

The main idea for constructing $\Phi(\cdot)$ is that if there exists no change point in Model (1), we can prove that $\sum_{a=1}^{p} |\hat{\boldsymbol{\delta}}^a(\hat{\tau})|_1 < K_0 s_1 \lambda_T$ with high probability. Hence, once we observe that $\sum_{a=1}^{p} |\hat{\boldsymbol{\delta}}^a(\hat{\tau})|_1 \ge K_0 s_1 \lambda_T$, it is more likely that a change point exists. Based on (8), we detect the existence of the change point τ_* if $\Phi(\cdot) = 1$. In other words, we regard the graphs are homogeneous if $\Phi(\cdot) = 0$. As shown in our theory (Theorem 5 and Proposition 10), with probability tending to one, the above classifier can correctly detect whether the graph undergoes a change point. Furthermore, once we detect a change point, the estimator $\hat{\tau}$ in (7) is proven to be consistent (Theorem 7). Hence, our method can detect and identify the change point simultaneously. This is different from the existing works in Bybee and Atchadé (2018); Gibberd and Nelson (2017), where they assumed the change point exists and only focused on the estimation. Note that in practice, s_1 is unknown and we need to specify it to implement (8). In what follows, we will use a hard thresholding method to obtain $\hat{s}_1$ which is shown to enjoy good performance in our numerical studies.

Remark 1 *We note that Lee et al. (2016) considered the lasso for high dimensional regression with a (possible) change point, where a threshold variable is adopted in the regression model. A naive idea for the Gaussian graphical change point model is to apply the procedure in Lee et al. (2016) directly to each node and identify τ_*. However, this naive procedure has two main drawbacks: Firstly, the change point location τ_* is not identifiable for nodes whose neighborhood does not change during the observations; Secondly, it is also difficult to combine the corresponding estimators obtained from the p nodes. Therefore, to identify the change point location τ_* for the Gaussian graphical model, we need to consider the p nodes simultaneously. Note that a similar idea was previously adopted by Peng et al. (2009) to estimate the partial correlation coefficients using a joint sparse regression model. Moreover, thanks to the availability of parallel computations, we can calculate $\hat{\boldsymbol{\beta}}^a(\tau)$ with $1 \le a \le p$ separately. This makes our method apply to very large scale networks in real applicability.*

Remark 2 *The penalty function in (5) is a weighted ℓ_1 penalty for $\boldsymbol{\beta}^a := ((\boldsymbol{\theta}^a)^\top, (\boldsymbol{\delta}^a)^\top)^\top$. To see this, by letting*

$$\boldsymbol{X}_j = (X_j^1, \ldots, X_j^T)^\top, \quad \boldsymbol{X}_j(\tau) = \big(X_j^1 \mathbf{1}\{Q_1 \le \tau\}, \ldots, X_j^T \mathbf{1}\{Q_T \le \tau\}\big)^\top,$$

we have $\mathrm{diag}(\mathbf{D}(\tau)) = (D_1(\tau), \ldots, D_p(\tau), D_{p+1}(\tau), \ldots, D_{2p}(\tau))^\top$ *with* $D_j(\tau) = \|\boldsymbol{X}_j\|_T$ *and* $D_{j+p}(\tau) = \|\boldsymbol{X}_j(\tau)\|_T$ *for* $1 \le j \le p$. *Then, we can rewrite the penalty as*

$$\lambda_T \big|\mathbf{D}(\tau)\boldsymbol{\beta}^a(\tau)\big|_1 = \lambda_T \sum_{j=1}^{p} \big(\|\boldsymbol{X}_j\|_T |\theta_j^a| + \|\boldsymbol{X}_j(\tau)\|_T |\delta_j^a|\big). \tag{9}$$

Note that the weight $D_j(\tau)$ regarding $\boldsymbol{\theta}^a$ does not depend on τ, while the weight $D_j(\tau)$ with respect to $\boldsymbol{\delta}^a$ does. As pointed out by one reviewer, one may consider the following optimization problem without weights:

$$\hat{\boldsymbol{\beta}}^a(\tau) = \underset{\boldsymbol{\theta}^a, \boldsymbol{\delta}^a \in \mathbb{R}^p; \theta_a^a = 0, \delta_a^a = 0}{\operatorname{argmin}} \|\mathbf{X}_a - \mathbf{X}(\tau)\boldsymbol{\beta}^a\|_T^2 + \lambda_T |\boldsymbol{\beta}^a|_1, \text{ for } 1 \le a \le p. \tag{10}$$

The motivation of our weighted penalty in (9) mainly comes from the consideration that, in practice, the p-dimensional random vector $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ may have different scales. Hence, the weight $\mathbf{D}(\tau)$ helps to balance the regressors as well as the effect of threshold τ on $\boldsymbol{\delta}^a$. In other words, one can see that using weights in the penalty (9) for solving the optimization problem (5) is equivalent to the normalization of $\boldsymbol{X}_j$ and $\boldsymbol{X}_j(\tau)$ for solving (10). Moreover, from the theoretical perspective, under our following assumptions, we can show that there exist positive constants c_1 and C_1 such that with high probability,

$$0 < c_1 \le \min_{\tau \in \mathcal{T}} \min_{1 \le j \le p} D_j(\tau) \le \max_{\tau \in \mathcal{T}} \max_{1 \le j \le p} D_j(\tau) \le C_1 < \infty$$

holds. Hence, if we solve the unweighted problem (10), our theoretical results in Section 3 still hold without changing the orders.

Algorithm 1 describes our procedure for obtaining $\hat{\tau}$ as well as the coefficients $(\hat{\boldsymbol{\beta}}^a(\hat{\tau}))_{a=1}^p$.

Algorithm 1 The neighborhood selection procedure for the change point inference.

Input: Given the dataset $\mathcal{X} = \{\boldsymbol{X}^1, \ldots, \boldsymbol{X}^T\}$, set the search domain $\mathcal{T} = [t_0, t_1]$, the non-negative regularization parameter λ_T, an estimate $\hat{s}$ for s, and the constant K_0.

Step 1: For a fixed $\tau \in [t_0, t_1]$, calculate $\big(\hat{\boldsymbol{\beta}}^a(\tau)\big)_{a=1}^p$ as defined in (5).

Step 2: Based on $\big(\hat{\boldsymbol{\beta}}^a(\tau)\big)_{a=1}^p$ obtained in Step 1, calculate $H(\tau)$ as in (6).

Step 3: Return $\hat{\tau} = \underset{\tau \in \mathcal{T}}{\operatorname{argmin}} H(\tau)$ and $\hat{\boldsymbol{\beta}}^a(\hat{\tau}) = \big((\hat{\boldsymbol{\theta}}^a(\hat{\tau}))^\top, (\hat{\boldsymbol{\delta}}^a(\hat{\tau}))^\top\big)^\top$ for $a \in \{1, \ldots, p\}$.

Step 4: Calculate $\Phi(\hat{\tau}, \{\hat{\boldsymbol{\beta}}^a(\hat{\tau}), 1 \le a \le p\})$ in (8).

Output: This algorithm provides $\hat{\tau}$ and $\big\{\hat{\boldsymbol{\beta}}^a(\hat{\tau}) = \big((\hat{\boldsymbol{\theta}}^a(\hat{\tau}))^\top, (\hat{\boldsymbol{\delta}}^a(\hat{\tau}))^\top\big)^\top, \text{ for } 1 \le a \le p\big\}$.

2.3 Structure recovery

After identifying τ_*, we next consider the recovery of the graphs. In particular, we aim to recover the support of $\boldsymbol{\Omega}^{(1)}$ and $\boldsymbol{\Omega}^{(2)}$, respectively. Note that in Section 2.2, we have obtained the change point estimator $\hat{\tau}$. For each node a, we also get the corresponding lasso estimator $\hat{\boldsymbol{\beta}}^a(\hat{\tau}) = \big((\hat{\boldsymbol{\theta}}^a(\hat{\tau}))^\top, (\hat{\boldsymbol{\delta}}^a(\hat{\tau}))^\top\big)^\top$ with

$$\hat{\boldsymbol{\theta}}^a(\hat{\tau}) = \big(\hat{\theta}_1^a(\hat{\tau}), \ldots, \hat{\theta}_p^a(\hat{\tau})\big)^\top, \text{ and } \hat{\boldsymbol{\delta}}^a(\hat{\tau}) = \big(\hat{\delta}_1^a(\hat{\tau}), \ldots, \hat{\delta}_p^a(\hat{\tau})\big)^\top, \text{ for } 1 \le a \le p.$$

Based on $\hat{\boldsymbol{\beta}}^a(\hat{\tau})$ and Model (4), for each node a, we then set $\hat{\mathrm{NE}}_a^{(1)}$ and $\hat{\mathrm{NE}}_a^{(2)}$ as its neighborhood's estimation before τ_* and after τ_*, respectively, which are defined as:

$$\hat{\mathrm{NE}}_a^{(1)} = \Big\{b \in V \setminus \{a\} : \hat{\theta}_b^a(\hat{\tau}) + \hat{\delta}_b^a(\hat{\tau}) \neq 0\Big\}, \quad \hat{\mathrm{NE}}_a^{(2)} = \Big\{b \in V \setminus \{a\} : \hat{\theta}_b^a(\hat{\tau}) \neq 0\Big\}.$$

After introducing $\hat{\mathrm{NE}}_a^{(1)}$ and $\hat{\mathrm{NE}}_a^{(2)}$ with $1 \leq a \leq p$, by adopting the method in Meinshausen and Bühlmann (2006), the naive estimators for $E^{(1)}$ and $E^{(2)}$ can be defined as:

$$\hat{E}^{(1)} = \Big\{(a,b) : a \in \hat{\mathrm{NE}}_b^{(1)} \cup b \in \hat{\mathrm{NE}}_a^{(1)}\Big\}, \quad \hat{E}^{(2)} = \Big\{(a,b) : a \in \hat{\mathrm{NE}}_b^{(2)} \cup b \in \hat{\mathrm{NE}}_a^{(2)}\Big\}. \tag{11}$$

Note that we can also use "$\cap$" instead of "$\cup$" in (11) for the structure recovery. The discrepancy between "$\cup$" and "$\cap$" vanishes with a high probability. More importantly, we tend to include spurious edges instead of omitting the relevant ones. Hence, we adopt "$\cup$" throughout this paper.

In the literature, some studies (Zhou (2010)) show that lasso may select "too many" variables with non-zero estimated coefficients. Consequently, the naive estimators $\hat{E}^{(1)}$ and $\hat{E}^{(2)}$ as in (11) include too many "noisy" edges, resulting in a large type I error (false positives) in terms of the recovery of $E^{(1)}$ and $E^{(2)}$. To avoid this problem, we introduce the following two-step thresholding procedure:

Step 1: Get initial estimators for $E^{(1)}$ and $E^{(2)}$. In particular, set

$$\begin{aligned} \hat{E}_{\text{init}}^{(1)} &= \Big\{(a,b) : |\hat{\theta}_b^a(\hat{\tau}) + \hat{\delta}_b^a(\hat{\tau})| \geq r_0\lambda_T \cup |\hat{\theta}_a^b(\hat{\tau}) + \hat{\delta}_a^b(\hat{\tau})| \geq r_0\lambda_T\Big\}, \\ \hat{E}_{\text{init}}^{(2)} &= \Big\{(a,b) : |\hat{\theta}_b^a(\hat{\tau})| \geq r_0\lambda_T \cup |\hat{\theta}_a^b(\hat{\tau})| \geq r_0\lambda_T\Big\}, \end{aligned} \tag{12}$$

where λ_T is the non-negative regularization parameter in (5) and $r_0 > 0$ is a user prespecified parameter.

In Step 1, we use $r_0\lambda_T$ to obtain the initial estimation of graphical structures. The main purpose of Step 1 is to obtain $\hat{E}_{\text{init}}^{(1)}$ and $\hat{E}_{\text{init}}^{(1)}$ which have the same order of the overall sparsity of the true graphs $E^{(1)}$ and $E^{(2)}$. In other words, we have $|\hat{E}_{\text{init}}^{(1)}| = O(|E^{(1)}|)$ and $|\hat{E}_{\text{init}}^{(2)}| = O(|E^{(2)}|)$. Hence, $|\hat{E}_{\text{init}}^{(1)}|$ (or $|\hat{E}_{\text{init}}^{(2)}|$) can be an estimator of s_1 for the classifier in (8). Moreover, according to our numerical studies, our method is not very sensitive to the choice of r_0. In practice, setting $r_0 \in [1,4]$ works well. Using the initial estimators, we next introduce the final structural estimation as in Step 2.

Step 2: Let $t_{\text{thr}}^{(1)} = r_1^*\lambda_T|\hat{E}_{\text{init}}^{(1)}|$ and $t_{\text{thr}}^{(2)} = r_2^*\lambda_T|\hat{E}_{\text{init}}^{(2)}|$ for some positive constants r_1^* and r_2^*, where $|\hat{E}_{\text{init}}^{(1)}|$ and $|\hat{E}_{\text{init}}^{(2)}|$ denote the cardinality of $\hat{E}_{\text{init}}^{(1)}$ and $\hat{E}_{\text{init}}^{(2)}$, respectively.

The final estimators are defined as:

$$\begin{aligned} \breve{E}^{(1)} &= \Big\{(a,b) \in \hat{E}_{\text{init}}^{(1)} : |\hat{\theta}_b^a(\hat{\tau}) + \hat{\delta}_b^a(\hat{\tau})| \geq t_{\text{thr}}^{(1)} \cup |\hat{\theta}_a^b(\hat{\tau}) + \hat{\delta}_a^b(\hat{\tau})| \geq t_{\text{thr}}^{(1)}\Big\}, \\ \breve{E}^{(2)} &= \Big\{(a,b) \in \hat{E}_{\text{init}}^{(2)} : |\hat{\theta}_b^a(\hat{\tau})| \geq t_{\text{thr}}^{(2)} \cup |\hat{\theta}_a^b(\hat{\tau})| \geq t_{\text{thr}}^{(2)}\Big\}. \end{aligned} \tag{13}$$

Note that Zhou (2010) also investigated the thresholding technique for high dimensional variable selection in regression models. As shown in Theorem 11, under some regular

conditions, the above thresholding procedure as in (12) – (13) can achieve model selection consistency in the sense that $E^{(1)} \subset \breve{E}^{(1)}$ (or $E^{(2)} \subset \breve{E}^{(2)}$), and also control the number of false positives in the sense that $|\breve{E}^{(1)} \cap (E^{(1)})^c| = O(1)$ (or $|\breve{E}^{(2)} \cap (E^{(2)})^c| = O(1)$). Our numerical studies also show that the thresholding procedure is very efficient to recover $E^{(1)}$ and $E^{(2)}$. See Section 4 for more details.

2.3.1 A data-driven thresholding procedure

Note that we need to specify r_1^* (or r_2^*) for the threshold parameter $t_{\text{thr}}^{(1)}$ (or $t_{\text{thr}}^{(2)}$). By definition of $\breve{E}^{(1)}$ (or $\breve{E}^{(2)}$) as in (13), the recovery of the underlying networks depends on the choice of r_1^* (or r_2^*). In particular, large values of r_1^* (or r_2^*) yield sparse graphs and small values of r_1^* (or r_2^*) yield dense graphs. In practice, we choose the "best" r_1^* (or r_2^*) from a given candidate subset $\mathcal{R} = \{r_1, \ldots, r_M\}$. The goal is to find r_1^* (or r_2^*) from $\mathcal{R}$ in a data-driven way. The main idea is based on the observation that the initial estimator $\hat{E}_{\text{init}}^{(1)}$ (or $\hat{E}_{\text{init}}^{(2)}$) includes "noisy" edges (false positives) and "true" edges (true positives), and there is a "gap" between the "noisy" and "true" edges.

Next, we introduce our method to find the "gap" between "noisy" and "true" edges. Given $r_j \in \mathcal{R}$, let $\breve{E}^{(1)}(r_j)$ (or $\breve{E}^{(2)}(r_j)$) be the thresholded estimator associated with r_j. Define the ratio of the total number of edges to the total number of non-edges in $\breve{E}^{(k)}(r_j)$:

$$\text{Ratio}^{(k)}(r_j) = \frac{|\breve{E}^{(k)}(r_j)|}{(p^2-p)/2 - |\breve{E}^{(k)}(r_j)|}, \text{ for } k = 1, 2 \text{ and } 1 \le j \le M. \tag{14}$$

We also define

$$\text{DIF}^{(k)}(r_j) = \frac{\text{Ratio}^{(k)}(r_j) - \text{Ratio}^{(k)}(r_{j-1})}{r_j - r_{j-1}}, \text{ for } k = 1, 2 \text{ and } 2 \le j \le M \tag{15}$$

as the approximation for the first-order derivative of $\text{Ratio}^{(k)}(r_j)$. For $k = 1, 2$, define the cumulative sum (CUSUM) statistic (see Csörgö and Horváth (1997)) for $\text{DIF}^{(k)}(r_j)$ as:

$$\text{CUS}^{(k)}(r_j) = \sqrt{M}\frac{j}{M}\Big(1 - \frac{j}{M}\Big)\Big(\frac{1}{j}\sum_{\ell=1}^{j}\text{DIF}^{(k)}(r_\ell) - \frac{1}{M-j}\sum_{\ell=j+1}^{M}\text{DIF}^{(k)}(r_\ell)\Big). \tag{16}$$

Finally, for $k = 1, 2$, we set r_k^* as:

$$r_k^* = r_{j^*}, \text{ with } j^* := \underset{2 \le j \le M-1}{\text{argmax}} \big|\text{CUS}^{(k)}(r_j)\big|. \tag{17}$$

For a simple interpretation of the above process, we randomly generate a 100×100 symmetric matrix $\mathbf{A} = (a_{ij})$ with 30% non-zero off-diagonal elements. We set the diagonal entry $a_{ii} = 1$. For the non-zero off-diagonal entry a_{ij}, we set $a_{ij} \overset{\text{i.i.d}}{\sim} \text{U}(0.05, 0.1)$ with probability 0.4, and $a_{ij} \overset{\text{i.i.d}}{\sim} \text{U}(0.1, 0.5)$ with probability 0.6. In other words, $\mathbf{A}$ can be regarded as the coefficient matrix with estimated values. The non-zero entry a_{ij} corresponds to an estimated edge between i and j. Among the 30% edges in $\mathbf{A}$, 40% edges are "noisy" edges and 60% are "true" edges. The "gap" between the magnitude of "noisy" and "true" edges is 0.1. To find the "gap", we generate a series of threshold variables ranging from

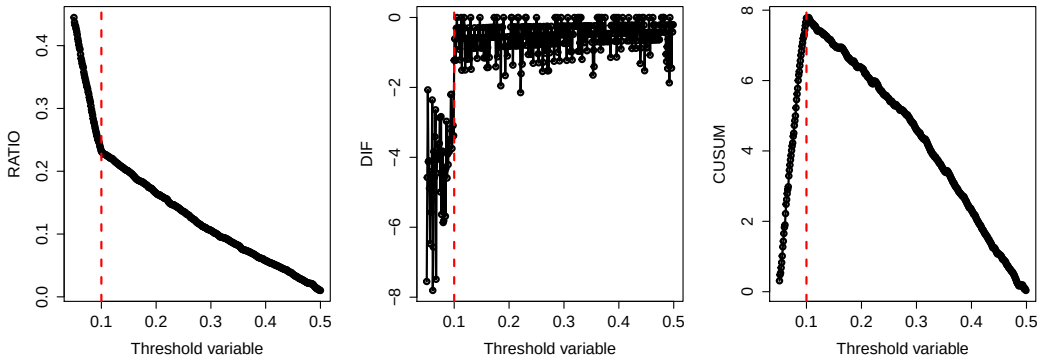

Figure 1: A simple illustrating example of the data-driven thresholding procedure

0.05 to 0.5, and calculate the corresponding Ratio, DIF, and CUSUM as defined in (14) – (16). The results are shown in Figure 1.

Figure 1 (left) shows that as the value of the threshold variable increases from 0.05 to 0.5, the Ratio decreases with different rates. In particular, from 0.05 to 0.1, the "noisy" edges are first "filtered". From 0.1 to 0.5, the "true" edges are then "filtered". Note that the Ratio decreases faster before 0.1 than that after 0.1. Hence, we can regard the "gap" as a "change-point" for the derivative of the Ratio as shown in Figure 1 (middle). To identify the "change-point" or "gap", we construct the well-known CUSUM statistic for DIF. Figure 1 (right) shows that CUSUM is maximized at the "gap". Algorithm 2 shows our method for the structure recovery using the data-driven thresholding procedure.

Algorithm 2 A data-driven thresholding procedure for the structure recovery.

Input: Given the estimated coefficients $\big\{\hat{\boldsymbol{\beta}}^a(\hat{\tau}) = \big((\hat{\boldsymbol{\theta}}^a(\hat{\tau}))^\top, (\hat{\boldsymbol{\delta}}^a(\hat{\tau}))^\top\big)^\top,\ \text{for } 1 \le a \le p\big\}$ obtained in Algorithm 1, set the candidate subset $\mathcal{R} = \{r_1, \ldots, r_M\}$.

Step 1: Obtain the initial estimator $\hat{E}^{(1)}_{\text{init}}$ (or $\hat{E}^{(2)}_{\text{init}}$) as defined in (12).

Step 2: Given $\mathcal{R}$, obtain the "best" r_1^* (or r_2^*) using the procedures as in (14) – (17).

Step 3: Based on r_1^* (or r_2^*) obtained in Step 2, get $\breve{E}^{(1)}(r_1^*)$ (or $\breve{E}^{(2)}(r_2^*)$) as in (13).

Output: This algorithm provides the estimator $\breve{E}^{(1)}$ for $E^{(1)}$ (or $\breve{E}^{(2)}$ for $E^{(2)}$).

3. Theoretical results

In this section, we present the theoretical results for our proposed method. In Section 3.1, some basic assumptions are introduced. Based on that, in Section 3.2, we derive some theoretical results in terms of the change point identification and structure recovery.

3.1 Basic assumptions

We introduce some basic assumptions needed for our theoretical results. In particular, Assumption 1 is the non-degenerate condition on the two covariance matrices $\mathbf{\Sigma}^{(1)}$ and $\mathbf{\Sigma}^{(2)}$, where $\mathbf{\Sigma}^{(1)} := (\mathbf{\Omega}^{(1)})^{-1}$ and $\mathbf{\Sigma}^{(2)} := (\mathbf{\Omega}^{(2)})^{-1}$. Assumption 2 requires some sparsity conditions on the parameter space. Assumptions 3 and 4 impose conditions on the signal strength and the search domain $\mathcal{T} = [t_0, t_1]$, respectively, for identifying the change point location. Assumption 5 is a technical condition on the regularization parameter λ_T. Specifically, the assumptions are summarized as follows:

- **Assumption 1** (Non-degenerate condition): (a) We require, without loss of generality, $\mathrm{Var}(X_a^t) = 1$ for $1 \le a \le p$ and $1 \le t \le T$. (b) $\mathrm{Var}(X_a^t | \mathbf{X}_{\setminus a}^t) = 1/\omega_{aa}^t > 0$ for $1 \le t \le T$ and $1 \le a \le p$. Furthermore, we require $1/\omega_{aa}^t \le \omega^2$ for some constant ω^2 with $1 \le a \le p$ and $1 \le t \le T$. (c) Let $\phi_{\min}^{(1)} = \lambda_{\min}(\mathbf{\Sigma}^{(1)})$, $\phi_{\min}^{(2)} = \lambda_{\min}(\mathbf{\Sigma}^{(2)})$, $\phi_{\max}^{(1)} = \lambda_{\max}(\mathbf{\Sigma}^{(1)})$, and $\phi_{\max}^{(2)} = \lambda_{\max}(\mathbf{\Sigma}^{(2)})$, where $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ denote the smallest and largest eigenvalues of a matrix. We require there exist some positive constants k_1 and K_1 such that $0 < k_1 \le \min(\phi_{\min}^{(1)}, \phi_{\min}^{(2)}) \le \max(\phi_{\max}^{(1)}, \phi_{\max}^{(2)}) \le K_1 < \infty$ holds.

- **Assumption 2** (Parameter space condition): Let

$$\mathcal{B}(s_1, M_0) = \Big\{ (\boldsymbol{\beta}^a)_{a=1}^p : \boldsymbol{\beta}^a = ((\boldsymbol{\theta}^a)^\top, (\boldsymbol{\delta}^a)^\top)^\top \text{ with } \boldsymbol{\theta}^a \in \mathbb{R}^p, \boldsymbol{\delta}^a \in \mathbb{R}^p \text{ satisfying : } |\boldsymbol{\beta}^a|_\infty \le M_0 \text{ for } 1 \le a \le p, \text{ and } \textstyle\sum_{a=1}^p \mathcal{M}(\boldsymbol{\beta}^a) \le s_1 \le p^2 \Big\}.$$

  We require that $(\tilde{\boldsymbol{\beta}}^a)_{a=1}^p \in \mathcal{B}(s_1, M_0)$.

- **Assumption 3** (Signal strength): Suppose there exists a positive constant δ_* such that

$$\min\Big(\sum_{a=1}^p (\tilde{\boldsymbol{\delta}}^a)^\top \mathbf{\Sigma}^{(1)} \tilde{\boldsymbol{\delta}}^a, \sum_{a=1}^p (\tilde{\boldsymbol{\delta}}^a)^\top \mathbf{\Sigma}^{(2)} \tilde{\boldsymbol{\delta}}^a \Big) > \delta_* \tag{18}$$

  holds.

- **Assumption 4** (Search domain): Define $\underline{\kappa} = \min(\phi_{\min}^{(1)}, \phi_{\min}^{(2)})$ and $\overline{\kappa} := \max(\phi_{\max}^{(1)}, \phi_{\max}^{(2)})$. For the search domain $\mathcal{T} = [t_0, t_1]$, we require

$$\min\Big(\lfloor T\tau \rfloor, \lfloor T(1-\tau) \rfloor \Big) \ge C_0 \log(pT), \quad \text{for} \quad \tau \in \mathcal{T}, \tag{19}$$

  and some large enough constant $C_0 > 0$. Furthermore, we also require

$$\lfloor T\tau_* \rfloor - \lfloor T\tau \rfloor < \frac{\underline{\kappa}}{2\overline{\kappa}} \big(T - \lfloor T\tau \rfloor \big), \quad \text{for } \tau \in [t_0, \tau_*], \tag{20}$$

  and

$$\lfloor T\tau \rfloor - \lfloor T\tau_* \rfloor < \frac{\underline{\kappa}}{2\overline{\kappa}} \lfloor T\tau \rfloor, \quad \text{for } \tau \in [\tau_*, t_1] \tag{21}$$

  hold, where $\tau_* \in (0, 1)$ is the true change point location.

- **Assumption 5** (Regularization parameter): We require the regularization parameter λ_T in (5) satisfies
$$\lambda_T = A\omega\Big(\frac{\log(p)}{T}\Big)^{1/2}$$
for some constant $A > \max(A_1^*/\mu, A_2^*/\mu)$, where $\mu \in (0,1)$ is a fixed constant and A_1^* and A_2^* are some universal positive constants not depending on p or T. More details about A_1^* and A_2^* can be found in the appendix.

Assumption 1 requires that $\mathbf{\Sigma}^{(1)}$ and $\mathbf{\Sigma}^{(2)}$ are non-degenerate in the sense that the eigenvalues of $\mathbf{\Sigma}^{(1)}$ and $\mathbf{\Sigma}^{(2)}$ are strictly bounded away from 0 and ∞. It is also crucial for proving the uniform restricted eigenvalue condition (URE) as shown in Proposition 4 below. Assumption 2 requires that the magnitude of parameters is bounded by the constant M_0. Assumption 2 also requires that both the networks $\mathbf{\Omega}^{(1)}$ and $\mathbf{\Omega}^{(2)}$ are at most s_1-sparse, which is a common assumption in the literature (Peng et al. (2009)). Assumptions 3 and 4 are important for the change point identification. In particular, Assumption 3 requires that the minimum signal strength satisfies (18). Assumption 4 provides basic requirements for the sample size T as well as the search domain $\mathcal{T}$. Lastly, Assumption 5 requires the regularization parameter $\lambda_T = O(\sqrt{\log(p)/T})$.

Remark 3 *Assumption 4 is commonly imposed for change point detection in high dimensional settings. In particular, Assumption (19) requires that we have the search domain $\mathcal{T}$ such that* $\min\Big(\lfloor T\tau\rfloor, \lfloor T(1-\tau)\rfloor\Big) \geq C_0\log(pT)$ *holds for all $\tau \in \mathcal{T}$. This implies that the minimum sample size T for change point detection of GGM is at least $O(\log p)$. Moreover, Assumptions (20) and (21) require that the search domain $\tilde{\mathcal{T}} := \{\lfloor Tt_0\rfloor, \lfloor Tt_0\rfloor+1, \ldots, \lfloor Tt_1\rfloor\}$ is not far away from the true change point location $\lfloor T\tau_*\rfloor$ with an order of $O(T)$. More specifically, by Assumptions (20), (21), and Assumption 1 with $0 < k_1 < \underline{\kappa} < \overline{\kappa} < K_1 < \infty$, we have*

$$\min(\lfloor T\tau_*\rfloor - \lfloor Tt_0\rfloor, \lfloor Tt_1\rfloor - \lfloor T\tau_*\rfloor) = O(T).$$

The motivation of Assumptions (20) and (21) comes from the fact that for change point estimation with a desired theoretical result, the following identifiability condition is needed:

$$\sum_{a=1}^{p} \|\mathbf{X}(\tau)\boldsymbol{\beta}^a - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\|_T^2 > c_*|\tau - \tau_*|, \quad \text{for any } \boldsymbol{\beta}^a \in \mathcal{B}(s_1, M_0) \quad \text{and} \quad \tau \in [t_0, t_1], \quad (22)$$

where $c_* > 0$ is some constant. Hence, to guarantee (22) holds, we need a search domain $\mathcal{T}$ not far away from τ_*, which requires Assumptions (20) and (21) hold. Moreover, compared with the existing literature with similar assumptions, we find that our requirement for $\tilde{\mathcal{T}}$ is much weaker. For example, Roy et al. (2017) and Bybee and Atchadé (2018) required

$$\min(\lfloor T\tau_*\rfloor - \lfloor Tt_0\rfloor, \lfloor Tt_1\rfloor - \lfloor T\tau_*\rfloor) = O(\sqrt{T\log(pT)}),$$

which imposes a much stronger condition for the search domain.

As pointed out by one reviewer, the search domain $\mathcal{T}$ in Assumptions (20) and (21) rely on $\underline{\kappa}/\overline{\kappa}$ and τ_*. For moderately large graphs, the ratio between the smallest and largest

eigenvalues, $\underline{\kappa}/\overline{\kappa}$, can be very small. In this extreme case, in theory, we need a search interval that is narrowly centered around the true change point, making our method difficult to use. In practice, however, we find that our method is not very sensitive to those parameters. For example, in our numerical studies with banded precision matrices $\mathbf{\Omega}^{(1)}$ and $\mathbf{\Omega}^{(2)}$, $\underline{\kappa}/\overline{\kappa}$ can be a small constant with an order of 10^{-3}. In this case, taking $\mathcal{T} = [0.15, 0.85]$ works quite well even though it is much wider than what is required by our theory.

3.2 Main results

3.2.1 Change point detection

After introducing basic assumptions, we now derive the theoretical results. We first consider the case of no change point. To this end, we introduce the following Proposition 4, which shows that the uniform restricted eigenvalue condition holds with a high probability. It is crucial to derive the desired bounds of estimation errors in $(\tilde{\boldsymbol{\beta}}^a)_{a=1}^p$ as well as τ_*.

Proposition 4 *(Uniform restricted eigenvalue (URE)$(s, c_0, \mathcal{S})$): For some integer $1 \le s \le 2p$, a positive number c_0, and some set $\mathcal{S} \subset [0,1] \subset \mathbb{R}$, we define*

$$\kappa(s, c_0, \mathcal{S}) := \min_{\tau \in \mathcal{S}} \min_{\substack{J_0 \subset \{1, \cdots, 2p\} \\ |J_0| \le s}} \min_{\substack{\boldsymbol{\gamma} \neq 0 \\ |\boldsymbol{\gamma}_{J_0^c}|_1 \le c_0 |\boldsymbol{\gamma}_{J_0}|_1}} \frac{|\mathbf{X}(\tau)\boldsymbol{\gamma}|_2}{\sqrt{T}|\boldsymbol{\gamma}_{J_0}|_2}. \tag{23}$$

Then under Assumptions 1, by letting $s = o(\sqrt{T/\log(pT)})$, the constant $\kappa(s, c_0, \mathcal{S}) > 0$ holds with probability at least $1 - (pT)^{-C}$ for some constant $C > 0$.

Note that for $\mathcal{S} = \{\tau_*\}$, $\kappa(s, c_0, \mathcal{S})$ as defined in (23) reduces to the restricted eigenvalue (RE) condition initially proposed b Bickel et al. (2009) for analyzing theoretical properties of both the lasso (Tibshirani (1996)) and Dantzig selector (Candes and Tao (2007)). It is well known that the RE condition is among the weakest assumptions on the design matrix (fixed or random) for deriving desired bounds of estimation errors in coefficients. Furthermore, as τ_* is unknown, it is necessary to impose the RE condition over $\tau \in \mathcal{S}$, which is called the uniform restricted eigenvalue condition proposed by Lee et al. (2016). It is worth mentioning that there are some differences between Lee et al. (2016) and this article with respect to URE. In particular, Lee et al. (2016) considered settings with a fixed design matrix while we consider a random design. Furthermore, since there is a possible change point during the observations, the design matrix $\mathbf{X}(\tau)$ is constructed using data with heteroscedasticity. More details can be found in the supplementary materials.

Using Proposition 4, the following Theorem 5 shows that the classifier $\Phi(\cdot)$ defined in (8) can correctly identify a homogeneous model with a high probability.

Theorem 5 *Suppose the data come from a homogeneous model with $\tilde{\boldsymbol{\delta}}^a = \mathbf{0}$ for all $1 \le a \le p$. Then under Assumptions 1, 2, and 5, we have*

$$\mathbb{P}\Big(\Phi(\hat{\tau}, \{\hat{\boldsymbol{\beta}}^a(\hat{\tau}), 1 \le a \le p\}) = 0\Big) \ge 1 - o(1).$$

Theorem 5 is built on the result of the estimation error in $(\tilde{\boldsymbol{\beta}}^a)_{a=1}^p$. More specifically, we can prove that our estimators obtained in Algorithm 1 satisfy $\sum_{a=1}^p \left|\hat{\boldsymbol{\beta}}^a(\hat{\tau}) - \tilde{\boldsymbol{\beta}}^a\right|_1 = O_p(\lambda_T s_1)$ even if τ_* is not identifiable. This result directly implies Theorem 5. More details can be found in the appendix.

3.2.2 Change point estimation

We now consider the case where τ_* is identifiable. In other words, we assume that $\tilde{\boldsymbol{\delta}}^a \neq 0$ for some $a \in \{1, \ldots, p\}$. We mainly focus on the estimation of τ_*. To this end, we first introduce the following Proposition 6, which plays a key role in identifying τ_* as well as deriving the desired bound of its estimation error.

Proposition 6 (Identifiability condition) *Suppose $(\boldsymbol{\beta}^a)_{a=1}^p \in \mathcal{B}(s_1, M_0)$ and $(\tilde{\boldsymbol{\beta}}^a)_{a=1}^p \in \mathcal{B}(s_1, M_0)$. Moreover, suppose Assumptions 1, 3, and 4 hold. Then for any η and τ such that $|\tau - \tau_*| > \eta > 0$,*

$$\sum_{a=1}^{p} \|\mathbf{X}(\tau)\boldsymbol{\beta}^a - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\|_T^2 > c_*\eta \tag{24}$$

holds with probability at least $1 - (pT)^{-C}$ for some universal constant $C > 0$, where c_* is also a universal positive constant not depending on p or T.

Note that Proposition 6 can be regarded as the identifiability condition for change point detection in Gaussian graphical models. Essentially, (24) characterizes the distance between the misspecified model $\{(\boldsymbol{\beta}^a)_{a=1}^p, \tau\}$ and the true underlying model $\{(\tilde{\boldsymbol{\beta}}^a)_{a=1}^p, \tau_*\}$ with $\tau \neq \tau_*$. It is shown by our proof that (24) relies on the difference between the two graphs (Assumption 3), the change point location (Assumption 4), and other technical conditions such as Assumptions 1 and 2. More details about Proposition 6 are provided in the supplementary materials.

With basic assumptions as well as Propositions 4 and 6, Theorem 7 below justifies the validity of Algorithm 1 for the change point identification.

Theorem 7 *Let $\hat{\tau}$ and $(\hat{\boldsymbol{\beta}}^a(\hat{\tau}))_{a=1}^p$ be the lasso solutions obtained via Algorithm 1. Suppose that Assumptions 1 – 5 hold with $\lambda_T s_1 \sum_{a=1}^{p} |\tilde{\boldsymbol{\delta}}^a|_1 \to 0$ (as $(p, T) \to \infty$) and $\sum_{a=1}^{p} |\tilde{\boldsymbol{\delta}}^a|_1^2 = o(\sqrt{T/\log(pT)})$ (as $(p, T) \to \infty$). Suppose the change point τ_* exists, then with probability at least $1 - C_1 p^{2-2A^2\mu^2/(A_1^*)^2} - C_2 p^{1-A^2\mu^2/(A_2^*)^2} - C_3(pT)^{-C_4}$, we have*

$$|\hat{\tau} - \tau_*| \leq M_1 A\omega\Big(\frac{\log(p)}{T}\Big)s_1, \tag{25}$$

where M_1 and $C_1, \ldots, C_4$ are universal positive constants not depending on p or T.

Theorem 7 provides a non-asymptotic bound of $|\hat{\tau} - \tau_*|$. As shown in (25), $|\hat{\tau} - \tau_*| \to 0$ as long as the scaling relationships among p, T, and s_1 satisfy $s_1 \log(p)/T \to 0$. Therefore, our paper allows the dimension p being much larger than the sample size T.

Remark 8 *According to the proof of Theorem 7, with very mild modifications, we can show that the change point estimation result in (25) is adaptive to the degree of changes between the two graphs. More specifically, we allow the minimum signal strength δ_* in Assumption 3 depends on n and p, say $\delta_*(n, p)$. In this case, if we require $\delta_*(n, p) \geq C_* s_1^{-1}$ for some large enough universal positive constant $C_* > 0$, our change point result can be modified to*

$$|\hat{\tau} - \tau_*| = O_p\Big(\frac{\log(p)}{T\delta_*(n, p)}s_1\Big). \tag{26}$$

Note that Result (26) is a generalization of (7) and it reduces to (7) by taking $\delta_*(n,p)=\delta_*$. Our change point estimator is still consistent as long as $\log(p)s_1/(T\delta_*(n,p))=o(1)$ holds.

It is worth mentioning that Enikeeva and Harchaoui (2019) considered change point detection for high dimensional mean vectors of Gaussian distributions. From the hypothesis testing point of view, they derived the minimax bounds for detecting sparse changes between two mean vectors with an order of $\|\boldsymbol{\delta}\|_2^2 = s_1\log(p)/T$, where $\boldsymbol{\delta}=\boldsymbol{\mu}_1-\boldsymbol{\mu}_2$ is the signal jump and s_1 is the overall sparsity of $\boldsymbol{\mu}_1-\boldsymbol{\mu}_2$. According to Enikeeva and Harchaoui (2019), no α-level test can detect a change point with probability tending to one if $\|\boldsymbol{\delta}\|_2^2 < cs_1\log(p)/T$ for some very small constant c. Note that our paper mainly focuses on change point estimation, which is essentially different from the problem of Enikeeva and Harchaoui (2019). Specifically, for change point estimation, we claim that the signal jump $\boldsymbol{\delta}$ obtains the minimax bound if no method can consistently estimate the true change point when a signal jump is smaller than that bound. In our paper, we require $\delta_*(n,p)\geq C_*s_1^{-1}$ for consistent change point estimation, which is typically a bigger order than that of Enikeeva and Harchaoui (2019) in the sense that $s_1^{-1}\gg s_1\log(p)/T$. To our limited knowledge, whether $s_1\log(p)/T$ is minimax optimal for change point estimation of high dimensional GGM is still an open question.

Proposition 9 is a by-product of Theorem 7, showing that we can control the prediction loss as well as estimation errors in coefficients. It is also crucial for verifying the two-step thresholding procedure.

Proposition 9 *Under the conditions of Theorem 7, with probability at least $1-C_1p^{2-2A^2\mu^2/(A_1^*)^2}-C_2p^{1-A^2\mu^2/(A_2^*)^2}-C_3(pT)^{-C_4}$, we have*

$$\sum_{a=1}^{p}\big\|\mathbf{X}(\hat{\tau})\hat{\boldsymbol{\beta}}^a(\hat{\tau})-\mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2\leq M_2A\omega\Big(\frac{\log(p)}{T}\Big)s_1,$$

and

$$\sum_{a=1}^{p}\big|\hat{\boldsymbol{\beta}}^a(\hat{\tau})-\tilde{\boldsymbol{\beta}}^a\big|_1\leq M_3A\omega\Big(\frac{\log(p)}{T}\Big)^{1/2}s_1,$$

where M_2, M_3, and $C_1,\ldots,C_4$ are universal positive constants not depending on p or T.

The following Proposition 10 shows that with high probability, our classifier $\Phi(\hat{\tau},\{\hat{\boldsymbol{\beta}}^a(\hat{\tau}),1\leq a\leq p\}$ defined in (8) can identify a heterogeneous model.

Proposition 10 *Suppose the assumptions in Theorem 7 hold. Furthermore, suppose additionally $\sum_{a=1}^{p}|\tilde{\boldsymbol{\delta}}^a|_1\geq(M_0+M_3)\lambda_Ts_1$ holds. Then, for a heterogeneous model with a change point, we have*

$$\mathbb{P}\Big(\Phi(\hat{\tau},\{\hat{\boldsymbol{\beta}}^a(\hat{\tau}),1\leq a\leq p\})=1\Big)\geq 1-o(1). \tag{27}$$

Combining the results in Theorem 5 and Proposition 10, we have established the validity of the classifier $\Phi(\cdot)$. This result is novel in the context of high dimensional dynamic Gaussian graphical models, since our method can automatically identify a homogeneous or heterogeneous model. This is very different from the existing works, where they assumed a change point exists and proposed to estimate its location.

3.2.3 Structure recovery

After presenting the results for the change point detection and estimation, we now consider the structure recovery of the two graphs. To this end, some additional conditions are needed.

- **Assumption 6**: We require the following conditions hold:

$$\min_{1\le a\le p}\min_{1\le b\le p, b\neq a}|\tilde{\theta}_b^a+\tilde{\delta}_b^a|\ge r_1^*(1+2\frac{M_3}{r_0})\lambda_T s_1+M_3\lambda_T s_1,$$

and

$$\min_{1\le a\le p}\min_{1\le b\le p, b\neq a}|\tilde{\theta}_b^a|\ge r_2^*(1+2\frac{M_3}{r_0})\lambda_T s_1+M_3\lambda_T s_1,$$

where $r_0>0$ is the same constant as in (12).

Note that Assumption 6 is a requirement on the minimum signal strength of $\mathbf{\Omega}^{(1)}$ and $\mathbf{\Omega}^{(2)}$. It is important for proving model selection consistency as shown in (28) and (29) below. Using Assumptions 1 – 6, Theorem 11 below justifies the validity of the two-step thresholding procedure as in (12) – (13) for recovering $E^{(1)}$ and $E^{(2)}$.

Theorem 11 *Let $\breve{E}^{(1)}$ and $\breve{E}^{(2)}$ be the thresholded estimators for $E^{(1)}$ and $E^{(2)}$, respectively. Suppose that Assumptions 1 – 6 hold with $\lambda_T s_1\sum_{a=1}^p|\tilde{\boldsymbol{\delta}}^a|_1\to 0$ and $\sum_{a=1}^p|\tilde{\boldsymbol{\delta}}^a|_1^2=o(\sqrt{T/\log(pT)})$. Suppose the change point τ_* exists, then with probability at least $1-C_1p^{2-2A^2\mu^2/(A_1^*)^2}-C_2p^{1-A^2\mu^2/(A_2^*)^2}-C_3(pT)^{-C_4}$, we have*

$$E^{(1)}\subset\breve{E}^{(1)},\ \text{and}\ |\breve{E}^{(1)}\cap(E^{(1)})^c|\le 2M_3/r_1^*, \tag{28}$$

and

$$E^{(2)}\subset\breve{E}^{(2)},\ \text{and}\ |\breve{E}^{(2)}\cap(E^{(2)})^c|\le 2M_3/r_2^*, \tag{29}$$

where M_3 is defined in (9) and $C_1,\ldots,C_4$ are universal positive constants.

Theorem 11 is built on the results of Proposition 9. In particular, to implement the thresholding procedure, "good" initial estimators for $(\tilde{\boldsymbol{\beta}}^a)_{a=1}^p$ are needed. Theorem 11 shows that the proposed thresholding procedure achieves model selection consistency in the sense that $E^{(1)}\subset\breve{E}^{(1)}$ or $E^{(2)}\subset\breve{E}^{(2)}$. In other words, the type II errors (false negatives) can be controlled with high probabilities. Furthermore, Theorem 11 also demonstrates that the number of false positives can be bounded by a finite number. The numerical studies in Section 4 provide a strong support for Theorem 11.

4. Simulation studies

In this section, we examine the empirical performance of our proposed method in terms of the change point identification and structure recovery, and compare its performance with existing techniques.

The data are generated as follows: let $\mathbf{\Omega}^{(1)}\in\mathbb{R}^{p\times p}$ and $\mathbf{\Omega}^{(2)}\in\mathbb{R}^{p\times p}$ be two different precision matrices. Let $\tau_*\in(0,1)$ be the true change point location. Then we generate

T independent observations $\boldsymbol{X}^1, \ldots, \boldsymbol{X}^T$, where $\boldsymbol{X}^t \overset{\text{i.i.d}}{\sim} N(\mathbf{0}, (\boldsymbol{\Omega}^{(1)})^{-1})$ for $1 \le t \le \lfloor T\tau_* \rfloor$, and $\boldsymbol{X}^t \overset{\text{i.i.d}}{\sim} N(\mathbf{0}, (\boldsymbol{\Omega}^{(2)})^{-1})$ for $\lfloor T\tau_* \rfloor + 1 \le t \le T$. To justify the broad applicability of our method, we generate $\boldsymbol{\Omega}^{(1)}$ and $\boldsymbol{\Omega}^{(2)}$ from the following two models:

- **Model 1** (Banded $\boldsymbol{\Omega}$): We generate banded precision matrices $\boldsymbol{\Omega}^{(1)}$ and $\boldsymbol{\Omega}^{(2)}$. This type of precision matrices corresponds to the chain networks (Fan et al. (2009)). The procedure is as follows: We first set $\boldsymbol{\pi}_1 = \{1, \ldots, p\}$ and $\boldsymbol{\pi}_2 = \{\pi_1, \ldots, \pi_p\}$, where $\boldsymbol{\pi}_2$ is a random permutation of $\boldsymbol{\pi}_1$. Then for $j \in \{1, 2\}$, we generate the covariance matrix $\boldsymbol{\Sigma}^{(j)} = (\sigma_{ab}^{(j)})$, where $\sigma_{ab}^{(1)} = \exp(-|t_a - t_b|/5)$, $\sigma_{ab}^{(2)} = \exp(-|t_{\pi_a} - t_{\pi_b}|/5)$, and $t_i - t_{i-1} \overset{\text{i.i.d}}{\sim} \mathrm{U}(0.5, 1) \cup \mathrm{U}(-1, -0.5)$ with $t_1 < t_2 \cdots < t_p$. Finally, we set the precision matrix as $\boldsymbol{\Omega}^{(j)} = (\boldsymbol{\Sigma}^{(j)})^{-1}$ for $j \in \{1, 2\}$. We further standardize $\boldsymbol{\Omega}^{(j)}$ to have unit diagonals. We repeat the above process for each replication. Note that Kolar and Xing (2012) previously used this model.
- **Model 2** (Block diagonal $\boldsymbol{\Omega}$): We generate block diagonal precision matrices $\boldsymbol{\Omega}^{(1)}$ and $\boldsymbol{\Omega}^{(2)}$. We first set $\boldsymbol{\pi}_1 = \{1, \ldots, p\}$ and $\boldsymbol{\pi}_2 = \{\pi_1, \ldots, \pi_p\}$, where $\boldsymbol{\pi}_2$ is a random permutation of $\boldsymbol{\pi}_1$. Then we generate $\boldsymbol{\Omega}^{(1)} = (\omega_{ab}^{(1)})$, where $\omega_{ab}^{(1)} = 0.8$ for $5(k-1)+1 \le a \ne b \le 5k$ $(k = 1, \ldots, \lfloor p/5 \rfloor)$, and $\omega_{ab}^{(1)} = 0$ otherwise. We generate $\boldsymbol{\Omega}^{(2)} = (\omega_{\pi_a \pi_b}^{(2)})$, where $\omega_{\pi_a \pi_b}^{(2)} = 0.8$ for $5(k-1)+1 \le \pi_a \ne \pi_b \le 5k$ $(k = 1, \ldots, \lfloor p/5 \rfloor)$, and $\omega_{\pi_a \pi_b}^{(2)} = 0$ otherwise. Finally, we standardize $\boldsymbol{\Omega}^{(j)}$ to have unit diagonals. We repeat the above process for each replication.

Note that for controlling the difference between $\boldsymbol{\Omega}^{(1)}$ and $\boldsymbol{\Omega}^{(2)}$, for the above two models, we set $\boldsymbol{\pi}_2 = (1, \ldots, p-\alpha, \pi_{p-\alpha+1}, \pi_{p-\alpha+2}, \ldots, \pi_p)$, where $(\pi_{p-\alpha+1}, \pi_{p-\alpha+2}, \ldots, \pi_p)$ is a random permutation of $(p-\alpha+1, p-\alpha+2, \ldots, p)$. In other words, the last α nodes in $\boldsymbol{\Omega}^{(1)}$ have a change point. A direct illustration is provided in Figure 2.

For Models 1 and 2, we set $T = 200$ with the dimension $p \in \{100, 200, 300\}$. Under these two models, we also consider different similarities, e.g. $\alpha \in \{10, 15, 20, 25\}$, between $\boldsymbol{\Omega}^{(1)}$ and $\boldsymbol{\Omega}^{(2)}$. For all models, we consider different change point locations by setting $\tau_* \in \{0.3, 0.4, 0.5\}$.

To identify τ_* in Algorithm 1, we set the search domain $\mathcal{T} = [0.15, 0.85]$. We set the regularization parameter $\lambda_T = C_0\sqrt{2\log(p/T)}$ with $C_0 = 0.3$ for both Models 1 and 2. With a specified λ_T, we use the R package "glmnet" to obtain $\hat{\boldsymbol{\beta}}^a(\tau)$ as defined in (5) for each τ and a. To use the classifier $\Phi(\hat{\tau}, \{\hat{\boldsymbol{\beta}}^a(\hat{\tau}), 1 \le a \le p\})$ in (8), we choose the constant K_0 via cross-validation (see Section 4.1) and set $s_1 = \hat{s} := |\hat{E}_{\text{init}}^{(1)}| \vee |\hat{E}_{\text{init}}^{(2)}|$, where $\hat{E}_{\text{init}}^{(1)}$ and $\hat{E}_{\text{init}}^{(2)}$ are defined in (12). For the structure recovery in Algorithm 2, we choose the candidate search domain $\mathcal{R}$ such that $t_{\text{thr}}^{(1)} \in [\hat{\beta}_{\min}^{(1)}, \hat{\beta}_{\max}^{(1)}]$ (or $t_{\text{thr}}^{(2)} \in [\hat{\beta}_{\min}^{(2)}, \hat{\beta}_{\max}^{(2)}]$, where $\hat{\beta}_{\min}^{(1)}$ (or $\hat{\beta}_{\min}^{(2)}$) denotes the corresponding minimum value in $\{|\hat{\theta}_b^a(\hat{\tau}) + \hat{\delta}_b^a(\hat{\tau})|, 1 \le a, b \le p\}$ (or $\{|\hat{\theta}_b^a(\hat{\tau})|, 1 \le a, b \le p\}$), and $\hat{\beta}_{\max}^{(1)}$ and $\hat{\beta}_{\max}^{(2)}$ denote the corresponding maximum values. All numerical results (without special instructions) are based on 200 replications.

4.1 Change point detection

With the above settings, we first consider the change point detection. To this end, for each replication, we randomly generate a data sequence $\boldsymbol{X}^1, \ldots, \boldsymbol{X}^T$ such that $\boldsymbol{X}^t \overset{\text{i.i.d}}{\sim}$

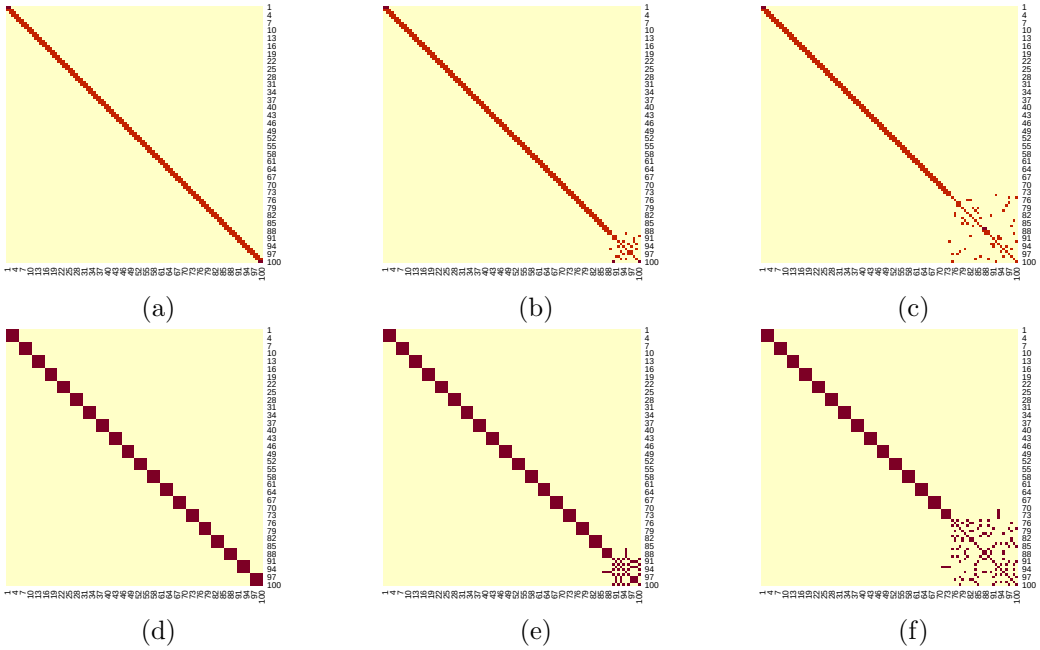

Figure 2: Heatmaps for networks in Models 1 and 2 with $p = 100$: (a) Banded $\mathbf{\Omega}^{(1)}$; (b) Banded $\mathbf{\Omega}^{(2)}$ with $\alpha = 10$; (c) Banded $\mathbf{\Omega}^{(2)}$ with $\alpha = 25$; (d) Block diagonal $\mathbf{\Omega}^{(1)}$; (e) Block diagonal $\mathbf{\Omega}^{(2)}$ with $\alpha = 10$; (f) Block diagonal $\mathbf{\Omega}^{(2)}$ with $\alpha = 25$.

$N(\mathbf{0}, (\mathbf{\Omega}_*^{(1)})^{-1})$ for $1 \le t \le \lfloor T\tau_* \rfloor$, and $\boldsymbol{X}^t \overset{\text{i.i.d}}{\sim} N(\mathbf{0}, (\mathbf{\Omega}_*^{(2)})^{-1})$ for $\lfloor T\tau_* \rfloor + 1 \le t \le T$. Let Z be a random variable following Bernoulli distribution $B(1, 0.5)$. Then we set $\mathbf{\Omega}_*^{(1)} = \mathbf{\Omega}^{(1)}$ and $\mathbf{\Omega}_*^{(2)} = \mathbf{\Omega}^{(1)} * Z + \mathbf{\Omega}^{(2)} * (1 - Z)$. In other words, with a probability 0.5, the data have a change point at $\tau_* \in \{0.5, 0.4, 0.3\}$. Within each replication, we use $\Phi(\hat{\tau}, \{\hat{\boldsymbol{\beta}}^a(\hat{\tau}), 1 \le a \le p\})$ defined in (8) to detect whether there exists a change point.

Note that the classifier $\Phi(\hat{\tau}, \{\hat{\boldsymbol{\beta}}^a(\hat{\tau}), 1 \le a \le p\})$ involves selection of K_0, which may affect the accuracy of change point detection. In other words, a large value of K_0 tends to identify the model as homogeneous and a small one tends to detect a change point. To select K_0 in a data-driven way, we use cross-validation. Specifically, in this numerical experiment, we set $T = 300$ and use the dataset $\{\boldsymbol{X}^1, \boldsymbol{X}^2, \boldsymbol{X}^4, \boldsymbol{X}^5, \ldots, \boldsymbol{X}^{298}, \boldsymbol{X}^{299}\}$ as the training set with a sample size 200 and use the remaining 100 samples, $\{\boldsymbol{X}^3, \boldsymbol{X}^6, \boldsymbol{X}^9, \ldots, \boldsymbol{X}^{297}, \boldsymbol{X}^{300}\}$, as the testing set. The cross-validation procedure is summarized as following three steps:
Step 1: We use the training data to obtain the estimation of $\hat{\tau}$, $\{\hat{\boldsymbol{\beta}}^a(\hat{\tau}), 1 \le a \le p\}$, as well as $\hat{s} = |\hat{E}_{\text{init}}^{(1)}| \vee |\hat{E}_{\text{init}}^{(2)}|$.
Step 2: For a given candidate value of K_0, say k_0, we can use (8) to decide the existence of a change point. If a change point is detected, we calculate the validation loss as :

$$CV\big(k_0, \hat{\tau}, \{\hat{\boldsymbol{\beta}}^a(\hat{\tau}), 1 \le a \le p\}\big)$$
$$= \sum_{a=1}^{p} \sum_{t:\ t \text{ mode } 3 == 0} \Big[X_a^t - (\boldsymbol{X}^t)^\top \hat{\boldsymbol{\theta}}^a(\hat{\tau}) - (\boldsymbol{X}^t)^\top \hat{\boldsymbol{\delta}}^a(\hat{\tau}) \mathbf{1}\{Q_t \le \hat{\tau}\} \Big]^2.$$

If a change point is not detected, we calculate the validation loss as :

$$CV\big(k_0, \hat{\tau}, \{\hat{\boldsymbol{\beta}}^a(\hat{\tau}), 1 \le a \le p\}\big) = \sum_{a=1}^{p} \sum_{t:\ t \text{ mode } 3==0} \Big[X_a^t - (\boldsymbol{X}^t)^\top \hat{\boldsymbol{\theta}}^a(\hat{\tau})\Big]^2.$$

Step 3: We repeat Step 2 for all candidate values of K_0 from a given candidate subset. In our numerical study, we set the candidate subset from a sequence ranging from 2^{-1} to 2^5. Lastly, we choose the best K_0 which has the minimum cross-validation loss.

It is worth mentioning that the above cross-validation procedure is computationally efficient since there is no need to do an additional brute search and calculate node-wise lasso. Table 1 records the type I and II errors for Models 1 and 2 using cross-validation with different dimensions and change point locations, where

$$\text{Type I error} = \frac{\text{number of false rejections}}{\text{number of replications}},\ \text{Type II error} = \frac{\text{number of false acceptions}}{\text{number of replications}}.$$

As shown in Table 1, our proposed method can detect a change point with high accuracy. Furthermore, we note that as τ_* gets closer to the boundary of data observations, it becomes more difficult to identify its existence. Since we fix the number of nodes having a change point at 100, the type II error increases as the dimension increases (especially for Model 2). One reasonable explanation is that, in this case, we can regard the 100 nodes as "signal" and the remaining nodes as "noise". The increasing dimension results in a decrease of the signal-noise ratio.

Table 1: Type I and II errors of change point detection for Models 1-2 with different dimensions and change point locations. The number of nodes having a change point is fixed at 100 ($\alpha = 100$). The reported results are based on 200 replications.

		$\tau_* = 0.5$		$\tau_* = 0.4$		$\tau_* = 0.3$	
Model	p	Type I	Type II	Type I	Type II	Type I	Type II
Model 1	100	0	0	0	0	0	0
	200	0	0	0	0	0	0
	300	0	0	0	0	0	0
Model 2	100	0	0	0	0	0	0
	200	0	0	0	0.	0	0
	300	0	0	0	0	0	0.04

4.2 Change point estimation and structure recovery

We next consider the change point estimation and structure recovery. As pointed out by one reviewer, as an alternative method, it is possible to use the fused graphical lasso method (denoted by FGL) (Danaher et al. (2014)) for estimating a single change point and recovering the graphical structures. More specifically, suppose there are K known number of classes or graphical structures, the fused graphical lasso solves the following optimization

problem:

$$\min_{\mathbf{\Omega}^{(k)}>0,k=1,2,\ldots,K}\Big\{\sum_{k=1}^{K}n_k\big[\mathrm{Tr}(\boldsymbol{S}^{(k)}\mathbf{\Omega}^{(k)})-\log|\mathbf{\Omega}^{(k)}|\big]+\lambda_1\sum_{k=1}^{K}\sum_{i\neq j}|\omega_{ij}^{(k)}|+\lambda_2\sum_{k<k'}\sum_{i,j}|\omega_{ij}^{(k)}-\omega_{ij}^{(k')}|\Big\}, \tag{30}$$

where K is the number of graphs, $\boldsymbol{S}^{(k)}$ and n_k are the sample covariance matrix estimation and the sample size for the k-the graph, λ_1 and λ_2 are two non-negative tuning parameters which control the overall sparsity of the K precision matrices and differences between corresponding elements of each pair of precision matrices, respectively. Note that for a change point model, there are at most two different graphs. Hence, to use FGL for identifying a change point, similar to Steps 1-3 in Algorithm 1, we can proceed as follows:
Step 1: For each fixed search point $\tau\in\mathcal{T}=[t_0,t_1]$, obtain $\hat{\mathbf{\Omega}}^{(\tau)}$ and $\hat{\mathbf{\Omega}}^{(1-\tau)}$ by solving

$$\begin{aligned}&\Big\{\hat{\mathbf{\Omega}}^{(\tau)},\hat{\mathbf{\Omega}}^{(1-\tau)}\Big\}\\&=\underset{\mathbf{\Omega}^{(\tau)}>0,\mathbf{\Omega}^{(1-\tau)}>0}{\mathrm{argmin}}\Big\{[T\tau]\big(\mathrm{Tr}(\boldsymbol{S}^{(\tau)}\mathbf{\Omega}^{(\tau)})-\log|\mathbf{\Omega}^{(\tau)}|\big)+(T-[T\tau])\big(\mathrm{Tr}(\boldsymbol{S}^{(1-\tau)}\mathbf{\Omega}^{(1-\tau)})-\log|\mathbf{\Omega}^{(1-\tau)}|\big)\\&+\lambda_1\sum_{i\neq j}\big(|\omega_{ij}^{(\tau)}|+|\omega_{ij}^{(1-\tau)}|\big)+\lambda_2\sum_{i,j}|\omega_{ij}^{(\tau)}-\omega_{ij}^{(1-\tau)}|\Big\},\end{aligned}$$

where $\boldsymbol{S}^{(\tau)}$ and $\boldsymbol{S}^{(1-\tau)}$ are the sample covariance matrix estimation using first $[T\tau]$ and last $T-[T\tau]$ observations, $\hat{\mathbf{\Omega}}^{(\tau)}=(\hat{\omega}_{ij}^{(\tau)})$ and $\hat{\mathbf{\Omega}}^{(1-\tau)}=(\hat{\omega}_{ij}^{(1-\tau)})$ are the precision matrix estimators before τ and after τ, respectively, and λ_1 and λ_2 are non-negative tuning parameters.
Step 2: Based on $\hat{\mathbf{\Omega}}^{(\tau)}$ and $\hat{\mathbf{\Omega}}^{(1-\tau)}$, define

$$\begin{aligned}\mathrm{FGL}(\tau)=&\Big\{[T\tau]\big(\mathrm{Tr}(\boldsymbol{S}^{(\tau)}\hat{\mathbf{\Omega}}^{(\tau)})-\log|\hat{\mathbf{\Omega}}^{(\tau)}|\big)+(T-[T\tau])\big(\mathrm{Tr}(\boldsymbol{S}^{(1-\tau)}\hat{\mathbf{\Omega}}^{(1-\tau)})-\log|\hat{\mathbf{\Omega}}^{(1-\tau)}|\big)\\&+\lambda_1\sum_{i\neq j}\big(|\hat{\omega}_{ij}^{(\tau)}|+|\hat{\omega}_{ij}^{(1-\tau)}|\big)+\lambda_2\sum_{i,j}|\hat{\omega}_{ij}^{(\tau)}-\hat{\omega}_{ij}^{(1-\tau)}|\Big\}.\end{aligned}$$

Step 3: Find $\hat{\tau}$ that minimizes $\mathrm{FGL}(\tau)$:

$$\hat{\tau}=\underset{\tau\in\mathcal{T}}{\mathrm{argmin}}\,\mathrm{FGL}(\tau).$$

Step 4: Once $\hat{\tau}$ is estimated, use $\hat{\mathbf{\Omega}}^{(\hat{\tau})}=(\hat{\omega}_{ij}^{(\hat{\tau})})$ and $\hat{\mathbf{\Omega}}^{(1-\hat{\tau})}=(\hat{\omega}_{ij}^{(1-\hat{\tau})})$ as the precision matrix estimation before and after the change point, respectively.

To implement FGL, we use the R package"JGL" for obtaining $\hat{\mathbf{\Omega}}^{(\tau)}$ and $\hat{\mathbf{\Omega}}^{(1-\tau)}$. Note that the implementation of FGL involves two tuning parameters λ_1 and λ_2. According to our numerical experiments, we find that setting (λ_1,λ_2) as $(0.25,0.1)$ enjoys good performance across various model settings.

To illustrate the performance in estimating τ_*, we record the mean (Mean), root mean squared errors (Rmse) for each simulation. We also adopt three indices (Precision, Recall, and F-score) for measuring the performance in recovering $E^{(1)}$ and $E^{(2)}$. The three indices are defined as follows:

- $\mathrm{Precision}=\dfrac{\sum_{a=1}^{p}\sum_{b=a+1}^{p}\mathbf{1}\{(a,b)\in E\}\wedge\mathbf{1}\{(a,b)\in\hat{E}\}}{\sum_{a=1}^{p}\sum_{b=a+1}^{p}\mathbf{1}\{(a,b)\in\hat{E}\}}$,

Table 2: Empirical results of Models 1 and 2 for various degrees of changes (α) between $\mathbf{\Omega}^{(1)}$ and $\mathbf{\Omega}^{(2)}$ with $p = 100$ and $\tau_* = 0.5$. The reported results are based on 200 replications.

			τ_*		$E^{(1)}$			$E^{(2)}$		
Model	α	Method	Mean	Rmse	Precision	Recall	F-score	Precision	Recall	F-score
Model 1	10	Oracle 1	-	-	0.98	0.98	0.98	0.98	0.98	0.98
		Oracle 2	-	-	0.82	1.00	0.90	0.94	1.00	0.97
		FGL	0.461	0.133	0.18	1.00	0.30	0.20	1.00	0.33
		Node-wise	0.486	0.047	0.82	0.99	0.90	0.94	1.00	0.97
	15	Oracle 1	-	-	0.98	0.98	0.98	0.98	0.98	0.98
		Oracle 2	-	-	0.80	0.99	0.88	0.93	1.00	0.97
		FGL	0.500	0.002	0.19	1.00	0.32	0.19	1.00	0.32
		Node-wise	0.503	0.007	0.80	1.00	0.89	0.94	1.00	0.97
	20	Oracle 1	-	-	0.98	0.98	0.98	0.98	0.98	0.98
		Oracle 2	-	-	0.78	0.99	0.87	0.93	1.00	0.96
		FGL	0.500	0.005	0.19	1.00	0.32	0.19	1.00	0.32
		Node-wise	0.504	0.007	0.78	1.00	0.88	0.93	1.00	0.97
	25	Oracle 1	-	-	0.98	0.98	0.98	0.98	0.98	0.98
		Oracle 2	-	-	0.78	1.00	0.87	0.92	1.00	0.96
		FGL	0.500	0.001	0.19	1.00	0.32	0.19	1.00	0.32
		Node-wise	0.505	0.007	0.77	0.99	0.86	0.93	1.00	0.96
Model 2	10	Oracle 1	-	-	1.00	1.00	1.00	1.00	1.00	1.00
		Oracle 2	-	-	0.77	1.00	0.87	0.95	1.00	0.97
		FGL	0.471	0.123	0.20	1.00	0.33	0.21	1.00	0.34
		Node-wise	0.504	0.008	0.78	1.00	0.87	0.95	1.00	0.98
	15	Oracle 1	-	-	1.00	1.00	1.00	1.00	1.00	1.00
		Oracle 2	-	-	0.75	1.00	0.86	0.94	1.00	0.97
		FGL	0.496	0.051	0.20	1.00	0.34	0.20	1.00	0.34
		Node-wise	0.504	0.007	0.75	1.00	0.86	0.94	1.00	0.97
	20	Oracle 1	-	-	1.00	1.00	1.00	1.00	1.00	1.00
		Oracle 2	-	-	0.74	1.00	0.85	0.93	1.00	0.96
		FGL	0.499	0.029	0.20	1.00	0.34	0.20	1.00	0.34
		Node-wise	0.504	0.007	0.73	1.00	0.85	0.93	1.00	0.96
	25	Oracle 1	-	-	1.00	1.00	1.00	1.00	1.00	1.00
		Oracle 2	-	-	0.72	1.00	0.84	0.91	1.00	0.95
		FGL	0.500	0.008	0.20	1.00	0.34	0.20	1.00	0.34
		Node-wise	0.503	0.006	0.72	1.00	0.84	0.92	1.00	0.96

- Recall $= \dfrac{\sum_{a=1}^{p}\sum_{b=a+1}^{p}\mathbf{1}\{(a,b)\in E\}\wedge\mathbf{1}\{(a,b)\in\hat{E}\}}{\sum_{a=1}^{p}\sum_{b=a+1}^{p}\mathbf{1}\{(a,b)\in E\}}$,

- F-score $= \dfrac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$.

By definitions, Precision is the number of correctly estimated edges divided by the total number of estimated edges, Recall is the number of correctly estimated edges divided by the total number of true edges, and F-score is a combination of Precision and Recall. Note that FGL estimates $\mathbf{\Omega}^{(1)}$ and $\mathbf{\Omega}^{(2)}$ instead of $E^{(1)}$ and $E^{(2)}$. Hence, according to FGL, we can identify that Node i and Node j has an edge if $|\hat{\omega}_{ij}^{(\hat{\tau})}| > 10^{-2}$ or $|\hat{\omega}_{ij}^{(1-\hat{\tau})}| > 10^{-2}$.

Table 2 demonstrates the results of Models 1 and 2 with $p = 100$ and $\tau_* = 0.5$, under various degrees of changes between $\mathbf{\Omega}^{(1)}$ and $\mathbf{\Omega}^{(2)}$. In addition to FGL, we consider three cases in the current setting:

Case 1: Assuming τ_* is known, we apply the neighborhood selection method in Meinshausen and Bühlmann (2006) to $\{\boldsymbol{X}^1, \ldots, \boldsymbol{X}^{\lfloor T\tau_* \rfloor}\}$ and $\{\boldsymbol{X}^{\lfloor T\tau_* \rfloor + 1}, \ldots, \boldsymbol{X}^T\}$, respectively, and recover the corresponding networks $E^{(1)}$ and $E^{(2)}$. We adopt a tuning parameter $\lambda = \sqrt{\log(p)/\lfloor T\tau_* \rfloor}$.

Case 2: Assuming τ_* is known, for each node a, we obtain $\hat{\boldsymbol{\beta}}^a(\tau_*)$ by (5). Then based on $\{\hat{\boldsymbol{\beta}}^a(\tau_*), 1 \le a \le p\}$, we use the thresholding procedure in Algorithm 2 to recover the corresponding two networks $E^{(1)}$ and $E^{(2)}$.

Case 3: Suppose the change point location is unknown. We use both Algorithms 1 and 2 to simultaneously estimate τ_* and recover the two networks $E^{(1)}$ and $E^{(2)}$.

The above three cases are denoted as Oracle 1, Oracle 2, and Node-wise, respectively. We can regard Case 2 as a special case of Case 3 by setting the search domain as $\mathcal{T} = \{\tau_*\}$. Note that since both Oracles 1 and 2 assume τ_* is known, the Mean and Rmse are not reported for these two cases. As can be seen from Table 2, by assuming τ_* is known, both Oracles 1 and 2 have better performance than other methods in terms of the recovery of $E^{(1)}$ and $E^{(2)}$. Interestingly, Oracle 1 performs better than Oracle 2. This is due to the known information of τ_*. Given τ_*, separate estimation performs better. When τ_* is not known, we can see that our node-wise approach (Node-wise) generally performs better than FGL in terms of structural recovery, which is demonstrated by higher precision. This result provides a strong support for our proposed two step thresholding based procedure in Algorithm 2. As for the change point identification, some interesting observations can be made. First, Node-wise has much higher accuracy than FGL, especially for Model 2 and cases when the degree of changes between $\mathbf{\Omega}^{(1)}$ and $\mathbf{\Omega}^{(2)}$ is small (e.g. $\alpha = 10$ or 15). A reasonable explanation is that FGL finds a change point by minimizing the overall likelihood function with a fused penalty $\lambda_2 \sum_{i,j} |\omega_{ij}^{(\tau)} - \omega_{ij}^{(1-\tau)}|$. When the number of nodes having a change point is small, FGL tends to identify the two graphs $\mathbf{\Omega}^{(1)}$ and $\mathbf{\Omega}^{(2)}$ as the same. As a result, its objective function FGL(τ) is not sensitive to such small changes, which fails to correctly identify the true change point. On the contrary, Node-wise is more sensitive to small changes since we construct the objective function in a node-wise way. Hence, even when the number of nodes with a change point is very small, it can still capture such information. Second, as the similarity between $\mathbf{\Omega}^{(1)}$ and $\mathbf{\Omega}^{(2)}$ increases, Rmses for both FGL and Node-wise increase, indicating that it is more difficult to identify the change point. Lastly, as the two graphs become more different with a larger α, we see that both FGL and Node-wise have similar and satisfactory results.

Table 3: Empirical results of Models 1 – 2 for various p with $\tau_* = 0.5$. For both Models 1 and 2, the degree of changes (α) between $\mathbf{\Omega}^{(1)}$ and $\mathbf{\Omega}^{(2)}$ is 20. The reported results are based on 200 replications

			τ_*		$E^{(1)}$			$E^{(2)}$		
Model	p	Method	Mean	Rmse	Precision	Recall	F-score	Precision	Recall	F-score
Model 1	100	FGL	0.500	0.001	0.19	1.00	0.32	0.19	1.00	0.32
		Node-wise	0.504	0.007	0.79	0.99	0.88	0.93	1.00	0.96
	200	FGL	0.497	0.045	0.12	1.00	0.21	0.12	1.00	0.21
		Node-wise	0.503	0.006	0.84	0.98	0.89	0.95	1.00	0.97
	300	FGL	0.480	0.074	0.08	1.00	0.15	0.09	1.00	0.16
		Node-wise	0.502	0.007	0.88	0.85	0.80	0.96	1.00	0.98
Model 2	100	FGL	0.498	0.015	0.20	1.00	0.34	0.20	1.00	0.34
		Node-wise	0.504	0.006	0.74	1.00	0.85	0.93	1.00	0.96
	200	FGL	0.493	0.031	0.12	1.00	0.22	0.12	1.00	0.22
		Node-wise	0.506	0.008	0.73	1.00	0.84	0.90	1.00	0.95
	300	FGL	0.495	0.022	0.09	1.00	0.17	0.09	1.00	0.17
		Node-wise	0.517	0.025	0.75	1.00	0.86	0.89	1.00	0.94

Table 3 shows the empirical results of Models 1 – 2 for $p \in \{100, 200, 300\}$ with a change point at $\tau_* = 0.5$, where the number of nodes having a change point is 20 for both models. We see that when p is small (e.g. $p = 100$), FGL and Node-wise have very similar performance in change point identification. As p increases from 100 to 300, Node-wise becomes better than FGL. The reason is that as p becomes larger, the changes between the two graphs become smaller. In such cases, FGL fails to detect such small changes. As for our method, it has good performance in estimating τ_* under different dimensions. This suggests that our method is applicable to relatively large-scale graphs with small changes. In terms of structural recovery, for both methods, the F-score decreases (especially for Model 1) as the dimension p increases, indicating that it is more difficult to recover the two underlying true graphs $E^{(1)}$ and $E^{(2)}$ with a larger p.

Table 4 shows the empirical results of Models 1 – 2 for various p and τ_*. In most cases with different change point locations, Node-wise performs better than FGL (see Figure 3). For all models with a fixed p, as the change point location gets closer to the boundary of the observations, e.g. $\tau_* = 0.3$, it is more difficult to identify the true change point location, which is illustrated by the increasing Rmse. For both methods, we note that, as τ_* decreases from 0.5 to 0.3, the sample size for estimating $\mathbf{\Omega}^{(1)}$ decreases while that for estimating $\mathbf{\Omega}^{(2)}$ increases, resulting in a worse recovery of $E^{(1)}$ and a better recovery of $E^{(2)}$.

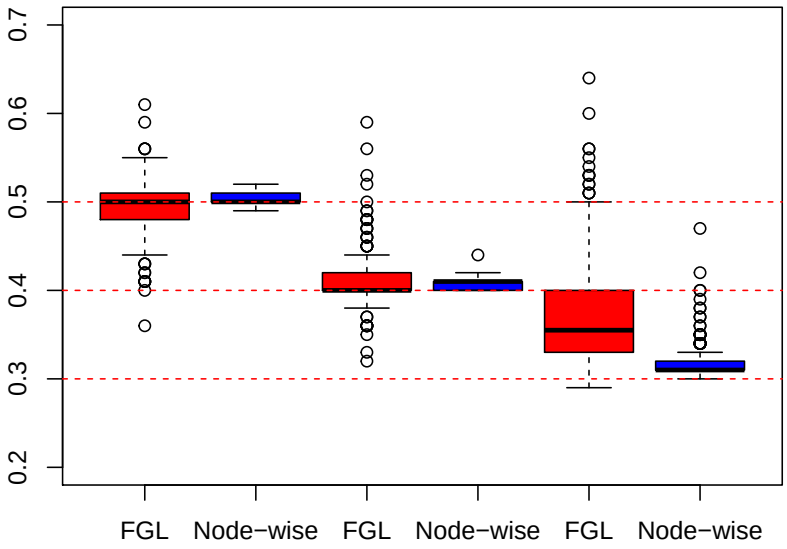

Figure 3: Boxplots of change point estimation for Model 2 with $p = 200$ and $\tau_* \in \{0.5, 0.4, 0.3\}$. The degree of changes (α) between $\mathbf{\Omega}^{(1)}$ and $\mathbf{\Omega}^{(2)}$ is 20. The reported results are based on 200 replications.

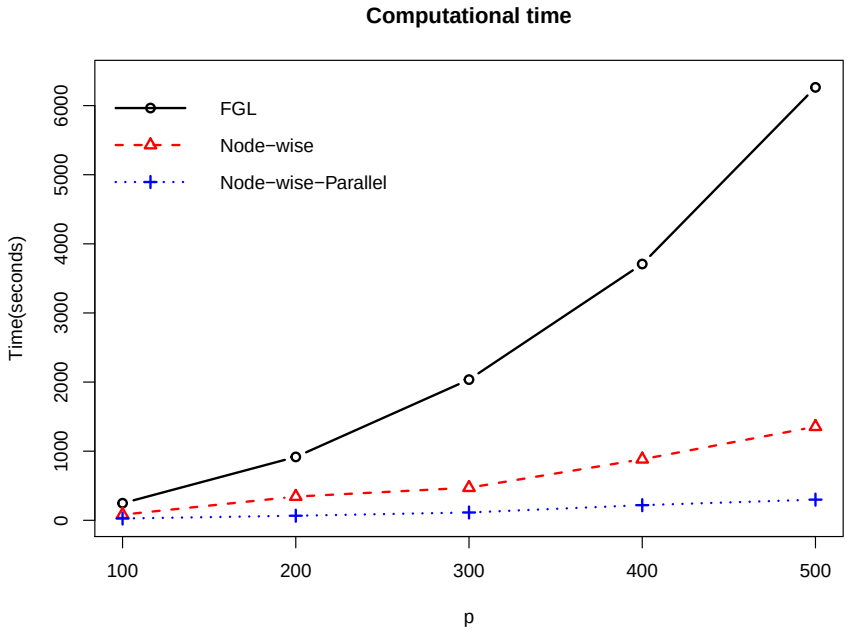

Figure 4: Computational time (seconds) for one replication with $p \in \{100, 200, 300, 400, 500\}$ and $T = 200$.

Table 4: Empirical results of Models 1 – 2 for various p and τ_*. For both Models 1 and 2, the degree of changes (α) between $\mathbf{\Omega}^{(1)}$ and $\mathbf{\Omega}^{(2)}$ is 20. The reported results are based on 200 replications.

				τ_*		$E^{(1)}$			$E^{(2)}$		
Model	p	τ_*	Method	Mean	Rmse	Precision	Recall	F-score	Precision	Recall	F-score
Model 1	100	0.5	FGL	0.500	0.001	0.19	1.00	0.32	0.19	1.00	0.32
			Node-wise	0.505	0.007	0.78	1.00	0.88	0.93	1.00	0.96
		0.4	FGL	0.400	0.001	0.16	1.00	0.28	0.21	1.00	0.35
			Node-wise	0.403	0.006	0.76	0.99	0.86	0.94	1.00	0.97
		0.3	FGL	0.300	0.002	0.14	1.00	0.24	0.23	1.00	0.37
			Node-wise	0.306	0.008	0.74	0.99	0.85	0.95	1.00	0.97
	200	0.5	FGL	0.498	0.048	0.12	1.00	0.21	0.12	1.00	0.21
			Node-wise	0.503	0.007	0.83	0.99	0.90	0.95	1.00	0.97
		0.4	FGL	0.396	0.026	0.10	1.00	0.18	0.13	1.00	0.24
			Node-wise	0.403	0.006	0.82	0.99	0.89	0.95	1.00	0.98
		0.3	FGL	0.296	0.012	0.08	1.00	0.15	0.15	1.00	0.26
			Node-wise	0.308	0.010	0.80	0.98	0.87	0.95	1.00	0.98
Model 2	100	0.5	FGL	0.500	0.016	0.20	1.00	0.34	0.20	1.00	0.34
			Node-wise	0.503	0.005	0.73	1.00	0.85	0.93	1.00	0.96
		0.4	FGL	0.396	0.020	0.18	1.00	0.30	0.23	1.00	0.37
			Node-wise	0.401	0.004	0.70	1.00	0.82	0.94	1.00	0.97
		0.3	FGL	0.297	0.016	0.15	0.99	0.26	0.25	1.00	0.40
			Node-wise	0.306	0.008	0.65	1.00	0.79	0.95	1.00	0.98
	200	0.5	FGL	0.494	0.033	0.12	1.00	0.22	0.12	1.00	0.22
			Node-wise	0.504	0.007	0.73	1.00	0.85	0.90	1.00	0.95
		0.4	FGL	0.409	0.035	0.11	1.00	0.20	0.13	1.00	0.24
			Node-wise	0.406	0.008	0.71	1.00	0.83	0.91	1.00	0.95
		0.3	FGL	0.375	0.100	0.10	0.98	0.19	0.14	1.00	0.24
			Node-wise	0.321	0.030	0.67	0.99	0.80	0.92	1.00	0.96

4.3 Computational cost

Lastly, we report the computational time for FGL and Node-wise for one replication. We implement the corresponding program independently on a CPU (Linux) with 2.50GHz, 6 cores, and 4GB of RAM. Note that for FGL, the computational cost is $O(T\times\text{FGLasso}(T,p))$, where $\text{FGLasso}(T,p)$ denotes the computational cost for solving fused graphical lasso with a sample size T and a data dimensionality p. As shown in Danaher et al. (2014), it typically requires a cost of $O(p^3)$ for each iteration for the algorithm in FGL. As for Node-wise, its computational cost is $O(Tp\times\text{Lasso}(T,p)$, where $\text{Lasso}(T,p)$ is the computational cost for

solving lasso with a sample size T and a data dimensionality p. For example, Lasso(T, p) is $O(pT)$ for the coordinate descent algorithm used in the R package "glmnet" for one iteration. Hence, the overall computational costs for FGL and Node-wise are $O(Tp^3)$ and $O(T^2p^2)$, respectively. For change point detection, it is more computationally expensive for using FGL than Node-wise when $p > T$. This is demonstrated in Figure 4. When p is relatively small (e.g. $p = 100$), FGL and Node-wise have similar computational costs. As p becomes larger, it requires a bigger computational burden for FGL than Node-wise. This suggests that Node-wise is more computationally efficient than FGL for change point detection. Furthermore, Figure 4 also provides the computational cost for Node-wise with parallel computing, where for each fixed $\tau \in [t_0, t_1]$, we calculate $\{\hat{\boldsymbol{\beta}}^a(\tau), 1 \le a \le p\}$ in (5) by dividing the p regression problems into six ndependent cores. We see from Figure 4 that the computational cost can be reduced significantly via parallel computing.

Note that the computational cost of our method in Algorithm 1 increases linearly in terms of numbers of grid points $\{Q_t := t/T, t = 1, \ldots, T\} \cap \mathcal{T}$, which can be expensive when T is very large. To solve this problem, in practice, we may use a two-stage-based method. Specifically, in the first stage, we use coarser grid points $\{Q_t := t/T, t = 1, 1 + k_0, 1 + 2k_0, \ldots, 1 + Mk_0\} \cap \mathcal{T}$, where $k_0 \ge 1$ is the user prespecified interval between two search points and M is the largest integer such that $1 + Mk_0 \le T$. Using these coarser grid points and Algorithm 1, we obtain an initial change point estimator $\hat{\tau}$. Then, in the second stage, we use new fine-resolution grid points $\{Q_t := t/T, t = \lfloor T\hat{\tau} \rfloor - 100, \lfloor T\hat{\tau} \rfloor - 100 + 1, \ldots, \lfloor T\hat{\tau} \rfloor + 100\} \cap \mathcal{T}$, which are constructed around the initial change point $\hat{\tau}$. Lastly, the final change point estimator $\hat{\hat{\tau}}$ is obtained using the above grid points and Algorithm 1. This leads to a more efficient algorithm when T is extremely large.

5. Real data analysis

In this section, we apply our proposed method to the S&P 500 index for analyzing the networks among the stocks. We consider a three-year period from January 4th, 2007 to December 31st, 2009, covering the recent financial crisis beginning in 2008. During this period, 372 stocks are considered, resulting in a dataset with $T = 755$ and $p = 372$. We obtain the final dataset from Yahoo Finance! (https://finance.yahoo.com/). Let $p_{t,i}$ be the closing price of the company i at the date t with $i = 1, \ldots, 372$ and $t = 1, \ldots, 755$. Then the log return for the company i at the date t is defined as: $r_{t,i} = \log(p_{t,i}) - \log(p_{t-1,i})$. Our analysis is based on the variables $r_{t,i}$. The 372 stocks are categorized into 10 Global Industry Classification Standard (GICS) sectors, including Consumer Discretionary (62 stocks), Consumer Staples (34 stocks), Energy (24 stocks), Financials (61 stocks), Health Care (40 stocks), Industrials (42 stocks), Information Technology (55 stocks), Materials (23 stocks), Telecommunications (6 stocks), and Utilities (25 stocks). Our goal is to investigate how the networks among the stocks evolve during the financial crisis.

We first identify the change point location. To this end, we apply Algorithm 1 to log return variables $r_{t,i}$ with the search domain $\mathcal{T} = [0.15, 0.85]$ and regularization parameter $\lambda_T = 10^{-3}\sqrt{2\log(p/T)}$. As can be seen from Figure 5 (left), the corresponding plot of $H(\tau)$ as defined in (6) is minimized at the location July 2, 2008, suggesting that the graphical structure among the stocks may undergo an abrupt change after that day. To interpret this result, we refer to the time series plot of the S&P 500 index as shown in Figure 5 (middle).

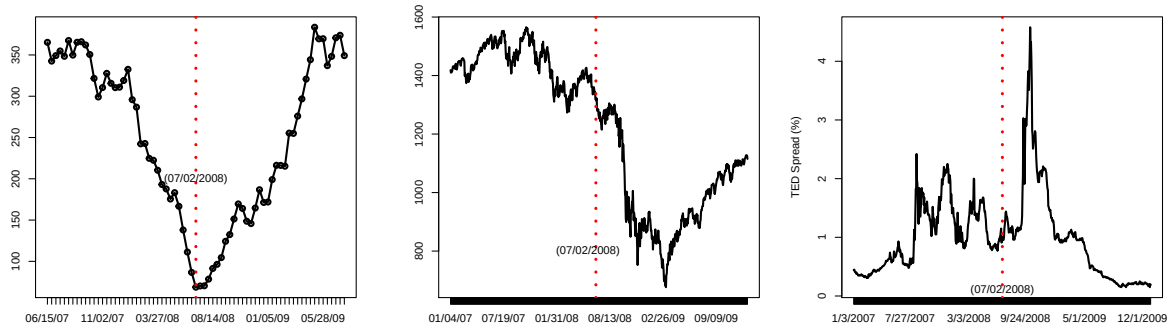

Figure 5: Plots of $H(\tau)$ (left), S&P 500 index (middle), and TED spread (right), during the period from January 4th, 2007 to December 31st, 2009.

We see that from the beginning of 2007 to the beginning of 2009, the S&P 500 index experiences a big slump. Furthermore, we note that the index declines with a relatively fast speed after the change point, as compared to that before the change point. To investigate the results further, we study the T-bills and "ED" (TED) spread, as shown in Figure 5 (right), where TED spread is short for the difference between the 3-month of London Inter-Bank Offer Rate (LIBOR), and the 3-month short-term U.S. government debt ("T-bills"). It is known that the TED spread is an indicator of credit risk in the general economy and an increase of the TED spread suggests an increased risk. The TED spread experiences a big fluctuation during the considered period, indicating that the entire economy is unstable during the financial crisis. We also note that the biggest fluctuation begins around July 2, 2008, which is consistent with our identified change point location. The above analysis provides some strong supports that there is an abrupt change of the graphical structure among the stocks after July 2, 2008, during the financial crisis.

Next we analyze the graphical structure among the stocks. We use Algorithm 2 to recover the two graphs, before and after the change point, respectively. To measure the change in graphical structure, we record the estimated number of edges for each sector. We divide the edges into two cases: edges from connected stocks which belong to the same sector (within-sector) and those from stocks belonging to different sectors (cross-sector). Figure 6 shows the number of edges for each sector for both cases.

We first consider the graphical structure within each sector. Figure 6 (top left) shows that there are more connections among stocks in the same sector after the change point. Furthermore, Figure 6 (top left) indicates that, compared to other sectors, the stocks in Health Care, Information Technology, and Consumer Staples are more connected to each other. Moreover, Figure 6 (bottom left) illustrates that, there are more increased edges of stocks belonging to Consumer Discretionary, Financials, and Information Technology, indicating that those three sectors are more affected by the financial crisis.

We next consider the graphical structure among sectors. It is shown in Figure 6 (top right) that stocks from different sectors are more related to each other during the financial crisis. Figure 6 (top right) shows that stocks belonging to Financials have more connections to other sectors during the financial crisis. To further investigate the financial crisis's

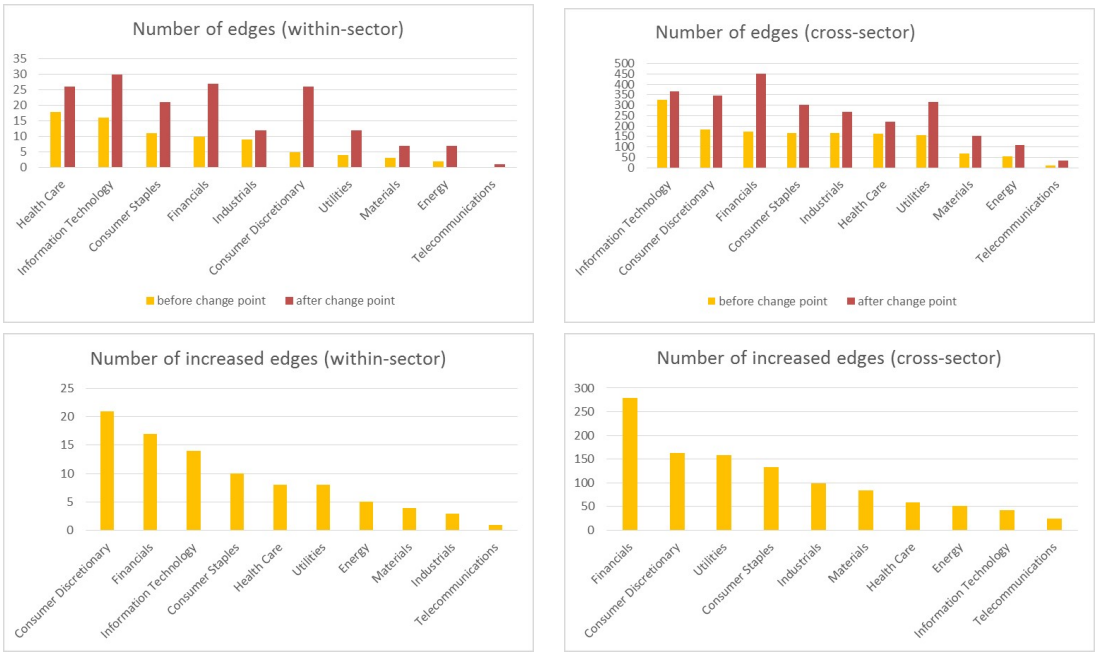

Figure 6: The number of edges before and after the change point for each sector. Top left: number of edges (within-sector); Top right: number of edges (cross-sector); Bottom left: number of increased edges (within-sector); Bottom right: number of increased edges (cross-sector).

Table 5: The first three sectors accounting for the largest, second and third largest proportions in increased (cross-sector) edges for the ten sectors.

Sector	First	Second	Third
Consumer Discretionary	Utilities	Consumer Staples	Financials
Consumer Staples	Financials	Consumer Discretionary	Industrials
Energy	Financials	Utilities	Consumer Discretionary
Financials	Consumer Discretionary	Utilities	Consumer Staples
Health Care	Financials	Consumer Discretionary	Information Technology
Industrials	Financials	Consumer Discretionary	Utilities
Information Technology	Health Care	Consumer Discretionary	Utilities
Materials	Financials	Consumer Staples	Health Care
Telecommunications	Financials	Consumer Discretionary	Consumer Staples
Utilities	Financials	Consumer Discretionary	Industrials

influence, for each sector, among its increased edges connected to other sectors, we record three corresponding sectors accounting for the largest, second and third largest proportions, respectively. The results are provided in Table 5. We see from Table 5 that, during the financial crisis, the sectors of Financials, Consumer Discretionary, and Utilities tend to have

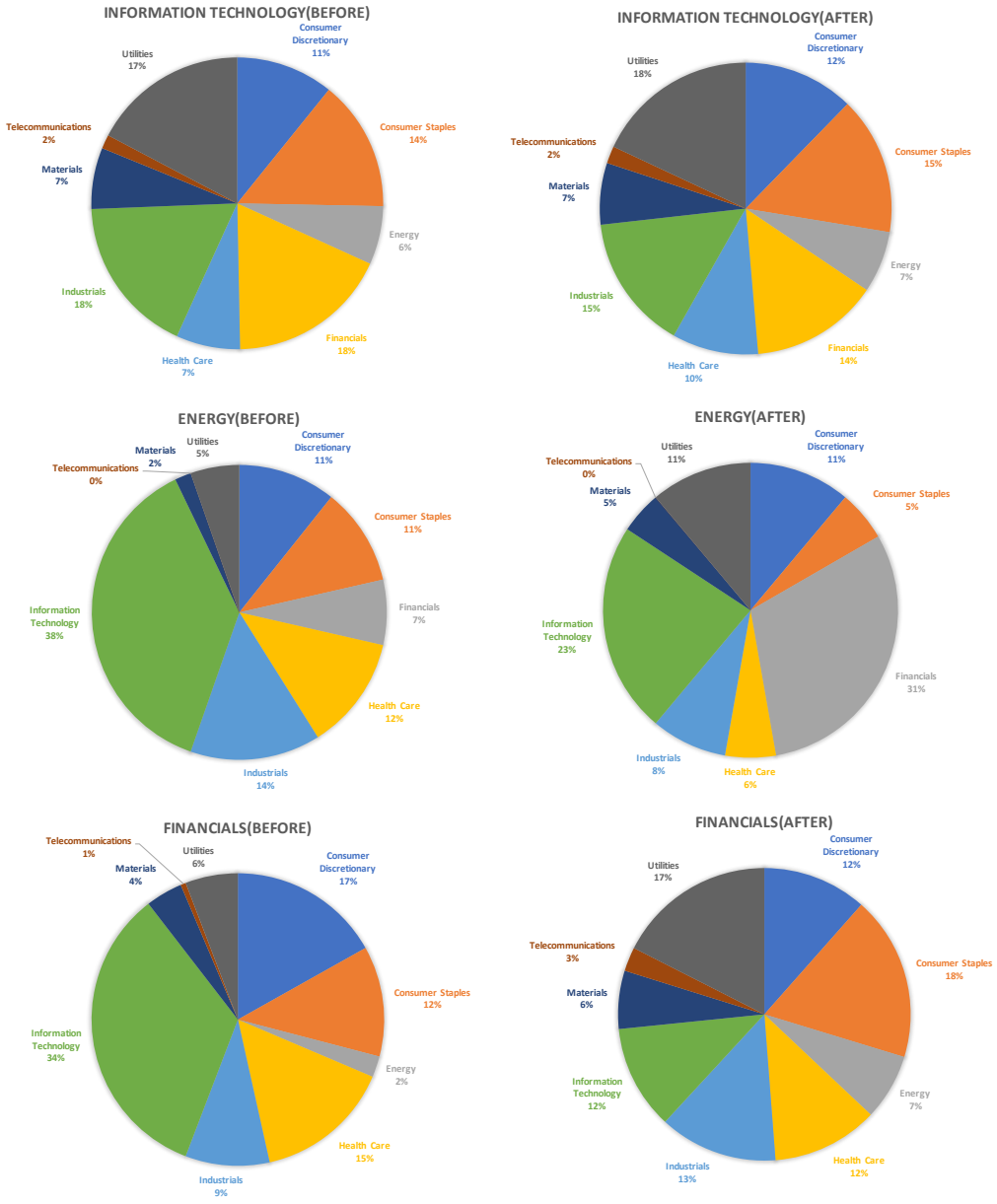

Figure 7: Estimated connections for the sectors of Information Technology (top), Energy (middle), and Financials (bottom), during the financial crisis. Left and right columns correspond to before and after the change point, respectively. Bigger proportions present more connections.

more connections to other sectors. This is consistent with that of Figure 6 (bottom right), where the above three sectors have more increased (cross-sector) edges than other sectors.

Finally, we illustrate the relationship between the sector of Information Technology (Energy, or Financials) with other sectors, before and after the change point, respectively. Figure 7 provides the corresponding results. Figure 7 (top) shows that for Information Technology, there is no big change of its connections to other sectors during the financial crisis, and it is mainly related to the sectors of Financials, Industrials, Consumer Staples, and Utilities. In contrast, Figure 7 (middle) indicates that there exist some changes in the neighborhood of Energy. For example, before the change point, Energy is mainly connected to Information Technology; after the change point, there are increasing connections to Financials and Utilities. Some other changes can also be found in Financials.

6. Conclusions

In this article, we present new methods for simultaneous change point inference and structure recovery in the context of high dimensional Gaussian graphical models with (possible) abrupt changes. For the change point identification, motivated by the neighborhood selection, we introduce a joint sparse regression model by considering the p nodes simultaneously, and incorporate a threshold variable and an unknown threshold parameter into the regression model to characterize the time-varying networks. The change point estimator and the estimated coefficients are obtained via minimizing the joint ℓ_2 loss function with an ℓ_1 penalty. Based on the above estimators, we also introduce a method for detecting whether the data are homogeneous. For the structure recovery, a data-driven hard thresholding procedure is proposed. Theoretically, under some regular conditions, we prove that the proposed method can select a true model (homogeneous or heterogeneous) with high accuracy. Once a heterogeneous model is identified, the change point estimator is proven to be consistent, by allowing the number of nodes being much larger than the sample size. Furthermore, in terms of the structure recovery, we prove that the thresholding procedure achieves model selection consistency and controls the number of false positives. The proposed method is relatively efficient to implement, and its validity is justified via both extensive numerical studies and a real data application.

Note that this paper focuses on the single-change-point setting. One possible extension is to consider multiple change points. For example, we may combine our node-wise-based loss function with the dynamic programming or binary segmentation techniques as in Leonardi and Bühlmann (2016) for localizing multiple change points. The main difficulty for extending to the case of multiple change points is that, in the multiple case, the candidate search interval $[s, e]$ may contain no change point, one change point, or more. Hence, to derive the desired theoretical results, we need detailed discussions about the lasso properties for each case and provide an oracle inequality in the general case. We leave this extension as a future research direction.

Acknowledgments

The authors would like to thank the action editor Professor Zaid Harchaoui and the referees for their helpful comments and suggestions. This research was supported in part by the National Natural Science Foundation of China Grant 12101132 (Bin Liu), 11971116 (Zhang)

and US National Institute of Health Grant R01GM126550 and National Science Foundation Grants DMS1821231 and 2100729 (Yufeng Liu).

Appendix

The appendix provides additional results for the main paper. In Section A, some notations are introduced. In Section B, we provide several useful lemmas needed for the main theorems. In Section C, we give detailed proofs of the main results. Proofs of the useful lemmas are given in Section D.

Appendix A. Notations

We present some additional notations. Recall that $\mathbf{X}(\tau)$ is the $T \times 2p$ matrix whose t-th row is $\mathbf{X}^t(\tau) \in \mathbb{R}^{2p}$ with $\mathbf{X}^t(\tau) = ((\mathbf{X}^t)^\top, (\mathbf{X}^t)^\top \mathbf{1}\{Q_t \le \tau\})^\top$, and whose a-th column is $\mathbf{X}_a(\tau) \in \mathbb{R}^T$, where $Q_t := t/T$ is the threshold variable and

$$\begin{aligned} &\mathbf{X}_a(\tau) = \boldsymbol{X}_a = (X_a^1, \ldots, X_a^T)^\top, && \text{for } 1 \le a \le p, \\ &\mathbf{X}_a(\tau) = \boldsymbol{X}_a(\tau) := \big(X_a^1 \mathbf{1}\{Q_1 \le \tau\}, \ldots, X_a^T \mathbf{1}\{Q_T \le \tau\}\big)^\top, && \text{for } p+1 \le a \le 2p. \end{aligned} \tag{31}$$

Define

$$X_{\max} = \max_{\tau \in \mathcal{T}} \max_{1 \le a \le 2p} \big\{ \|\mathbf{X}_a(\tau)\|_T, a = 1, \ldots, 2p, \tau \in \mathcal{T} \big\}, \tag{32}$$

and

$$X_{\min} = \min_{1 \le a \le 2p} \big\{ \|\mathbf{X}_a(t_0)\|_T, a = 1, \ldots, 2p \big\}, \tag{33}$$

where t_0 comes from $\mathcal{T} = [t_0, t_1]$. Let $r_T = \min_{1 \le a \le p} \dfrac{\|\boldsymbol{X}_a(t_0)\|_T^2}{\|\boldsymbol{X}_a\|_T^2}$. Recall $\mathbf{D}(\tau) := \mathrm{diag}\{\|\mathbf{X}_a(\tau)\|_T : a = 1, \ldots, 2p\}$. We set $\hat{\mathbf{D}} = \mathbf{D}(\hat{\tau})$ and $\mathbf{D} = \mathbf{D}(\tau_*)$. For two matrices $\mathbf{V}_1$ and $\mathbf{V}_2$, define their maximum distance as $\|\mathbf{V}_1 - \mathbf{V}_2\|_\infty = \max_{i,j} |(\mathbf{V}_1)_{ij} - (\mathbf{V}_2)_{ij}|$. Denote $\otimes$ as the Kronecker product for two matrices. We use $C_1, C_2, \ldots$ to denote constants that may vary from line to line.

Appendix B. Useful lemmas

The following Lemma 12 shows that there exist some constants K_1 and K_2 such that the two events $\{X_{\max} \le K_1\}$ and $\{X_{\min} \ge K_2\}$ hold with a high probability. In other words, r_T is strictly bounded away from zero with a high probability. The proof of Lemma 12 is provided in Section D.1.

Lemma 12 *Let $\{\boldsymbol{X}^t\}_{t=1}^T$ be independent p-dimensional Gaussian random vectors with $\boldsymbol{X}^t = (X_1^t, \ldots, X_p^t)^\top$ for $1 \le t \le T$. Suppose $\mathbb{E}X_a^t = 0$ and $Var(X_a^t) = 1$ for $1 \le t \le T$ and $1 \le a \le p$. Then for $X_{\max}$ as defined in (32), there exists a constant $K_1 > 1$ such that*

$$\mathbb{P}\big(X_{\max} \le K_1\big) \ge 1 - p \exp(-C_1 T) \tag{34}$$

holds, where $C_1 > 0$ is a constant only depending on K_1.

For $X_{\min}$ as defined in (33), there is a constant $K_2 > 0$ such that

$$\mathbb{P}\big(X_{\min} \geq K_2\big) \geq 1 - p\exp(-C_2 T), \tag{35}$$

where $C_2 > 0$ is a constant only depending on K_2.

The following Lemma 13 shows that the design matrix is smooth with respect to the threshold variable Q_t. The proof of Lemma 13 is provided in Section D.2.

Lemma 13 *Suppose $\boldsymbol{X}^t \overset{i.i.d}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma}^{(1)})$ for $1 \leq t \leq \lfloor T\tau_* \rfloor$, and $\boldsymbol{X}^t \overset{i.i.d}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma}^{(2)})$ for $\lfloor T\tau_* \rfloor + 1 \leq t \leq T$ with $\boldsymbol{X}^t = (X_1^t, \ldots, X_p^t)^\top$, where $\boldsymbol{\Sigma}^{(1)} = (\boldsymbol{\Omega}^{(1)})^{-1}$ and $\boldsymbol{\Sigma}^{(2)} = (\boldsymbol{\Omega}^{(2)})^{-1}$. Then there is a universal constant $C^* > 0$ such that for any $\eta > 0$*

$$\sup_{1\leq a\leq p} \sup_{|\tau-\tau_*|\leq \eta} \frac{1}{T}\sum_{t=1}^{T} |X_a^t|^2 \big|\mathbf{1}\{Q_t \leq \tau_*\} - \mathbf{1}\{Q_t \leq \tau\}\big| \leq C^*\eta \tag{36}$$

holds with probability at least $1 - (pT)^{-C}$ for some constant $C > 0$.

The following Lemma 14 is the basic inequality for proving the main results. Before presenting Lemma 14, we introduce some definitions and notations.

For each node $a \in V := \{1, \ldots, p\}$ and $\tau \in \mathcal{T} := [t_0, t_1]$, we define

$$\begin{aligned} V_{1b}^a &:= \frac{1}{T}\sum_{t=1}^{T} \frac{\epsilon_a^t}{\sigma_a^t} \frac{X_b^t}{\|\boldsymbol{X}_b\|_T}, \\ V_{2b}^a(\tau) &:= \frac{1}{T}\sum_{t=1}^{T} \frac{\epsilon_a^t}{\sigma_a^t} \frac{X_b^t \mathbf{1}\{Q_t \leq \tau\}}{\|\boldsymbol{X}_b(\tau)\|_T}, \quad \text{for } 1 \leq b \leq p, \text{ and } b \neq a, \end{aligned} \tag{37}$$

where ϵ_a^t is defined in (4) with $\mathrm{Var}(\epsilon_a^t) = 1/\omega_{aa}^t := (\sigma_a^t)^2$, and $\boldsymbol{X}_b(\tau) := (X_b^1 \mathbf{1}\{Q_1 \leq \tau\}, \ldots, X_b^T \mathbf{1}\{Q_T \leq \tau\})^\top$. Note that by Assumption 1, we have $(\sigma_a^t)^2 \leq \omega^2$ for some constant $\omega > 0$. For a constant $\mu \in (0, 1)$, we also define the following two node-wise events:

$$\begin{aligned} \mathcal{A}^a &= \bigcap_{\substack{b=1 \\ b\neq a}}^{p} \big\{2|V_{1b}^a| \leq \mu\lambda_T/\omega\big\}, \\ \mathcal{B}^a &= \bigcap_{\substack{b=1 \\ b\neq a}}^{p} \Big\{ \sup_{\tau\in\mathcal{T}} 2|V_{2b}^a(\tau)| \leq \mu\lambda_T/\omega\Big\}, \text{ for } 1 \leq a \leq p. \end{aligned} \tag{38}$$

Recall $\boldsymbol{\beta}^a = ((\boldsymbol{\theta}^a)^\top, (\boldsymbol{\delta}^a)^\top)^\top$ and $\tilde{\boldsymbol{\beta}}^a = ((\tilde{\boldsymbol{\theta}}^a)^\top, (\tilde{\boldsymbol{\delta}}^a)^\top)^\top$, we then define

$$R = \frac{2}{T}\sum_{a=1}^{p}\sum_{t=1}^{T} \epsilon_a^t (\mathbf{X}^t)^\top \tilde{\boldsymbol{\delta}}^a \big(\mathbf{1}\{Q_t \leq \hat{\tau}\} - \mathbf{1}\{Q_t \leq \tau_*\}\big). \tag{39}$$

After introducing basic notations, we now present Lemma 14. Lemma 14 is the basic inequality for proving our main results. The proof of Lemma 14 is provided in Section D.3.

Lemma 14 *Let $\hat{\tau}$ and $(\hat{\boldsymbol{\beta}}^a(\hat{\tau}))_{a=1}^p$ be the solutions obtained from (5) – (7). Then conditional on the event $\bigcap_{1\le a\le p}\{\mathcal{A}^a\cap\mathcal{B}^a\}$, we have*

$$\begin{aligned}&\sum_{a=1}^p \big\|\mathbf{X}(\hat{\tau})\hat{\boldsymbol{\beta}}^a-\mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2+\lambda_T(1-\mu)\sum_{a=1}^p\big|\hat{\mathbf{D}}(\hat{\boldsymbol{\beta}}^a-\tilde{\boldsymbol{\beta}}^a)\big|_1\\&\le 2\lambda_T\sum_{a=1}^p\big|\hat{\mathbf{D}}(\hat{\boldsymbol{\beta}}^a-\tilde{\boldsymbol{\beta}}^a)_{J_0^a}\big|_1+\lambda_T\sum_{a=1}^p\big|\|\hat{\mathbf{D}}\tilde{\boldsymbol{\beta}}^a\|_1-|\mathbf{D}\tilde{\boldsymbol{\beta}}^a|_1\big|+R,\end{aligned}\tag{40}$$

and

$$\begin{aligned}&\sum_{a=1}^p \big\|\mathbf{X}(\hat{\tau})\hat{\boldsymbol{\beta}}^a-\mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2+\lambda_T(1-\mu)\sum_{a=1}^p\big|\hat{\mathbf{D}}(\hat{\boldsymbol{\beta}}^a-\tilde{\boldsymbol{\beta}}^a)\big|_1\\&\le 2\lambda_T\sum_{a=1}^p\big|\hat{\mathbf{D}}(\hat{\boldsymbol{\beta}}^a-\tilde{\boldsymbol{\beta}}^a)_{J_0^a}\big|_1+\sum_{a=1}^p\big\|\mathbf{X}(\hat{\tau})\tilde{\boldsymbol{\beta}}^a-\mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2,\end{aligned}\tag{41}$$

where $J_0^a:=J(\tilde{\boldsymbol{\beta}}^a)$, $\hat{\boldsymbol{\beta}}^a:=\hat{\boldsymbol{\beta}}^a(\hat{\tau})$, $\mathbf{D}:=\mathbf{D}(\tau_*)$, and $\hat{\mathbf{D}}:=\mathbf{D}(\hat{\tau})$.

The following Lemma 15 shows that the event $\bigcap_{1\le a\le p}\{\mathcal{A}^a\cap\mathcal{B}^a\}$ occurs with a high probability. The proof of Lemma 15 is provided in Section D.4.

Lemma 15 *Let $\Phi(x)$ be the cumulative distribution function (CDF) of the standard normal random variable. Recall $\mathcal{A}^a$ and $\mathcal{B}^a$ as defined in (38). Then conditional on the event*

$$\mathcal{E}^{(1)}=\{X_{\max}\le K_1,\ \textit{and}\ X_{\min}\ge K_2\},\tag{42}$$

we have

$$\mathbb{P}\Big(\bigcap_{1\le a\le p}\{\mathcal{A}^a\cap\mathcal{B}^a\}\Big)\ge 1-6p^2\Phi\Big(\frac{-\mu K_2\sqrt{T}}{2\omega K_1}\lambda_T\Big),$$

where the constants K_1 and K_2 come from Lemma 12.

Before presenting Lemma 17, we need to introduce the following event: for some constant $\eta>0$, we define

$$\mathcal{C}(\eta)=\Big\{\sup_{|\tau-\tau_*|\le\eta}\Big|\frac{2}{T}\sum_{a=1}^p\sum_{t=1}^T\epsilon_a^t(\mathbf{X}^t)^\top\tilde{\boldsymbol{\delta}}^a\big(\mathbf{1}\{Q_t\le\tau\}-\mathbf{1}\{Q_t\le\tau_*\}\big)\Big|\le\lambda_T\sqrt{\eta}\Big\}.\tag{43}$$

The following Lemma 16 shows that the event $\cap_{j=1}^m\mathcal{C}(\eta_j)$ occurs with a high probability. The proof of Lemma 16 is provided in Section D.5.

Lemma 16 *For a given integer m, let $\eta_1,\dots,\eta_m$ be some positive constants. Suppose $\sum_{a=1}^p|\tilde{\boldsymbol{\delta}}^a|_1^2=o(\sqrt{T/\log(pT)})$ holds. Then conditional on the events*

$$\begin{aligned}\mathcal{E}^{(2)}\ &=\Big\{\Big\|\frac{1}{\big(\lfloor T\tau_*\rfloor-\lfloor T(\tau_*-\eta_j)\rfloor\big)}\sum_{t=\lfloor T(\tau_*-\eta_j)\rfloor}^{\lfloor T\tau_*\rfloor}\big(\boldsymbol{\Sigma}^{(1)}-\mathbf{X}^t(\mathbf{X}^t)^\top\big)\Big\|_\infty\\&\qquad\le C_1\sqrt{\frac{\log(pT)}{T}},\ \textit{for}\ j=1,\dots,m\Big\},\end{aligned}\tag{44}$$

and

$$\begin{aligned}\mathcal{E}^{(3)} &= \Big\{ \Big\| \frac{1}{\big([T(\tau_*+\eta_j)]-\lfloor T\tau_*\rfloor\big)} \sum_{t=\lfloor T\tau_*\rfloor}^{[T(\tau_*+\eta_j)]} \big(\mathbf{\Sigma}^{(2)} - \mathbf{X}^t(\mathbf{X}^t)^\top\big) \Big\|_\infty \\ &\qquad \le C_2\sqrt{\frac{\log(pT)}{T}},\ \textit{for } j=1,\ldots,m \Big\},\end{aligned} \tag{45}$$

we have

$$\begin{aligned}&\mathbb{P}\Big(\textstyle\bigcap_{j=1}^m \mathcal{C}(\eta_j)\Big) \\ &\ge 1 - 4\sum_{j=1}^m \Phi\left(\frac{-\lambda_T K_2 \mu\sqrt{T}}{2K_1\omega^2\sqrt{C_1^{(j)}\underline{\kappa}^{-1}\delta^*}}\right) - 4\sum_{j=1}^m \Phi\left(\frac{-\lambda_T K_2 \mu\sqrt{T}}{2K_1\omega^2\sqrt{C_2^{(j)}\underline{\kappa}^{-1}\delta^*}}\right),\end{aligned}$$

where K_1 and K_2 come from Lemma 12, ω^2 comes from Assumption 1 (b), $\underline{\kappa}$ comes from Assumption 4, $C_1^{(j)}$ and $C_2^{(j)}$ are positive constants only depending on η_j for $1\le j\le m$, and $\delta^* := \max\big(\sum_{a=1}^p (\tilde{\boldsymbol{\delta}}^a)^\top \mathbf{\Sigma}^{(1)}\tilde{\boldsymbol{\delta}}^a, \sum_{a=1}^p (\tilde{\boldsymbol{\delta}}^a)^\top \mathbf{\Sigma}^{(2)}\tilde{\boldsymbol{\delta}}^a\big)$.

The following Lemmas 17 and 18 show that, if we have prior estimation error bounds in τ_* as well as $(\tilde{\boldsymbol{\beta}}^a)_{a=1}^p$, we can further tighten the corresponding estimation error bounds using the prior bounds. These two lemmas are crucial for proving our main results. Their proofs are provided in Sections D.6 and D.7, respectively.

Lemma 17 *Let $\hat{\tau}$ and $(\hat{\boldsymbol{\beta}}^a(\hat{\tau}))_{a=1}^p$ be the solutions obtained from (5) – (7). Suppose that $|\hat{\tau}-\tau_*|\le c_\tau$ and $\sum_{a=1}^p |\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a|_1 \le c_{\boldsymbol{\beta}}$ hold, for some c_τ and $c_{\boldsymbol{\beta}}$. Suppose further that Assumptions 1 – 2 hold. Then conditional on $\bigcap_{1\le a\le p}\{\mathcal{A}^a\cap\mathcal{B}^a\}\bigcap\mathcal{C}(c_\tau)$, we have*

$$\begin{aligned}\sum_{a=1}^p \big\|\mathbf{X}(\hat{\tau})\hat{\boldsymbol{\beta}}^a - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2 \le 3\lambda_T\Big\{&\frac{6X_{\max}^2}{\kappa'^2}\lambda_T s_1 \bigvee \frac{2X_{\max}}{\kappa'}\Big(s_1 c_{\boldsymbol{\beta}} c_\tau C^* \sum_{a=1}^p |\tilde{\boldsymbol{\delta}}^a|_1\Big)^{1/2} \\ &\bigvee\Big(\sqrt{c_\tau} + (2X_{\min})^{-1} c_\tau C^* \sum_{a=1}^p |\tilde{\boldsymbol{\delta}}^a|_1\Big)\Big\},\end{aligned} \tag{46}$$

and

$$\begin{aligned}\sum_{a=1}^p \big|\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a\big|_1 \le \frac{3}{(1-\mu)X_{\min}}\Big\{&\frac{6X_{\max}^2}{\kappa'^2}\lambda_T s_1 \bigvee \frac{2X_{\max}}{\kappa'}\Big(s_1 c_{\boldsymbol{\beta}} c_\tau C^* \sum_{a=1}^p |\tilde{\boldsymbol{\delta}}^a|_1\Big)^{1/2} \\ &\bigvee\Big(\sqrt{c_\tau} + (2X_{\min})^{-1} c_\tau C^* \sum_{a=1}^p |\tilde{\boldsymbol{\delta}}^a|_1\Big)\Big\},\end{aligned} \tag{47}$$

where κ' is a constant related to the URE condition, the constant C^* is defined in (36), and s_1 comes from Assumption 2.

Lemma 18 *Let $\hat{\tau}$ and $(\hat{\boldsymbol{\beta}}^a(\hat{\tau}))_{a=1}^p$ be the solutions obtained from (5) – (7). Suppose that $|\hat{\tau}-\tau_*|\le c_\tau$ and $\sum_{a=1}^p |\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a|_1 \le c_{\boldsymbol{\beta}}$ hold, for some c_τ and $c_{\boldsymbol{\beta}}$. Define*

$$\tilde{c}_\tau = c_*^{-1}\lambda_T\Big((1+\mu)c_{\boldsymbol{\beta}}X_{\max} + \sqrt{c_\tau} + (2X_{\min})^{-1}c_\tau C^* \sum_{a=1}^p |\tilde{\boldsymbol{\delta}}^a|_1\Big).$$

If Assumptions 1 – 4 hold, then conditional on $\bigcap_{1\le a\le p}\{\mathcal{A}^a\cap\mathcal{B}^a\}\bigcap\mathcal{C}(c_\tau)$, we have

$$|\hat{\tau}-\tau_*| \le \tilde{c}_\tau, \tag{48}$$

where the constant c_* is defined in (24), and the constant C^* is defined in (36).

Appendix C. Proof of main results

C.1 Proof of Proposition 4

Proof Since the proof for a homogeneous model is easier than that of a heterogenous one, for simplicity, we only consider the latter case. Recall $\mathbf{X}(\tau)$ is a $T \times 2p$ random design matrix whose t-th row is defined as

$$\mathbf{X}^t(\tau) = \big((\mathbf{X}^t)^\top, (\mathbf{X}^t)^\top \mathbf{1}\{Q_t \le \tau\}\big)$$

with $Q_t := t/T$. Then we have $\frac{1}{T}\mathbf{X}(\tau)^\top \mathbf{X}(\tau) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{X}^t(\tau)(\mathbf{X}^t(\tau))^\top$. Define

$$\hat{\mathbf{V}}(\tau) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{X}^t(\tau)(\mathbf{X}^t(\tau))^\top, \quad \mathbf{V}(\tau) = \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\mathbf{X}^t(\tau)(\mathbf{X}^t(\tau))^\top, \ \text{ for } \tau \in \mathcal{T}.$$

With these notations, the proof of Proposition 4 proceeds in three steps. In Step 1, we prove $\mathbf{V}(\tau)$ satisfies the URE condition uniformly over $\tau \in \mathcal{T}$. In Step 2, we prove that the maximum distance between $\mathbf{V}(\tau)$ and $\hat{\mathbf{V}}(\tau)$ can be bounded by $C\sqrt{\log(pT)/T}$. In Step 3, we combine the previous two steps and finish the proof. Now, we consider the three steps in detail.

Step 1 : Note that $\mathbf{X}^t \overset{\text{i.i.d}}{\sim} N(\mathbf{0}, \mathbf{\Sigma}^{(1)})$ for $1 \le t \le \lfloor T\tau_* \rfloor$, and $\mathbf{X}^t \overset{\text{i.i.d}}{\sim} N(\mathbf{0}, \mathbf{\Sigma}^{(2)})$ for $\lfloor T\tau_* \rfloor + 1 \le t \le T$, where $\mathbf{\Sigma}^{(1)} := (\mathbf{\Omega}^{(1)})^{-1}$ and $\mathbf{\Sigma}^{(2)} := (\mathbf{\Omega}^{(2)})^{-1}$. Straightforward calculations show that

$$\mathbf{V}(\tau) = \begin{cases} \mathbf{\Sigma}^{(1)} \otimes \underbrace{\begin{pmatrix} \tau_*, & \tau \\ \tau, & \tau \end{pmatrix}}_{\mathbf{A}_1(\tau)} + \mathbf{\Sigma}^{(2)} \otimes \underbrace{\begin{pmatrix} 1-\tau_*, & 0 \\ 0, & 0 \end{pmatrix}}_{\mathbf{A}_2(\tau)}, & \text{if } \ \tau \le \tau_*, \\ \mathbf{\Sigma}^{(1)} \otimes \underbrace{\begin{pmatrix} \tau_*, & \tau_* \\ \tau_*, & \tau_* \end{pmatrix}}_{\mathbf{A}_3(\tau)} + \mathbf{\Sigma}^{(2)} \otimes \underbrace{\begin{pmatrix} 1-\tau_*, & \tau-\tau_* \\ \tau-\tau_*, & \tau-\tau_* \end{pmatrix}}_{\mathbf{A}_4(\tau)}, & \text{if } \ \tau > \tau_*. \end{cases}$$

Note that $\lambda_{\min}(\mathbf{A}_1(\tau)) > 0$, $\lambda_{\min}(\mathbf{A}_2(\tau)) = 0$ for $\tau \le \tau_*$, and $\lambda_{\min}(\mathbf{A}_3(\tau)) = 0$, $\lambda_{\min}(\mathbf{A}_4(\tau)) > 0$ for $\tau > \tau_*$. Furthermore, by Assumption 1 (c), we also have $\lambda_{\min}(\mathbf{\Sigma}^{(1)}) > 0$ and $\lambda_{\min}(\mathbf{\Sigma}^{(2)}) > 0$. Using the fact that all eigenvalues of the Kronecker product of two matrices can be written as the product between their eigenvalues and by Weyl's Theorem, we have $\lambda_{\min}(\mathbf{V}(\tau)) > 0$ over $\tau \in [t_0, t_1]$. Therefore, $\mathbf{V}(\tau)$ satisfies the URE condition uniformly over $\tau \in \mathcal{T} = [t_0, t_1]$.

Step 2 : We prove $\sup_{\tau\in\mathcal{T}} \|\hat{\mathbf{V}}(\tau) - \mathbf{V}(\tau)\|_\infty = O_p(\sqrt{\log(pT)/T})$. To this end, we define

$$\tilde{\mathbf{V}}(\tau) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\mathbf{X}^t(\tau)(\mathbf{X}^t(\tau))^\top, \ \text{ for } \tau \in \mathcal{T}. \tag{49}$$

Hence, by introducing $\tilde{\mathbf{V}}(\tau)$ as in (49), to bound $\sup_{\tau\in\mathcal{T}} \|\hat{\mathbf{V}}(\tau) - \mathbf{V}(\tau)\|_\infty$, we need to consider $\sup_{\tau\in\mathcal{T}} \|\tilde{\mathbf{V}}(\tau) - \mathbf{V}(\tau)\|_\infty$ and $\sup_{\tau\in\mathcal{T}} \|\hat{\mathbf{V}}(\tau) - \tilde{\mathbf{V}}(\tau)\|_\infty$, respectively. We first consider $\sup_{\tau\in\mathcal{T}} \|\hat{\mathbf{V}}(\tau) - \mathbf{V}(\tau)\|_\infty$. By definition, we can write $\hat{\mathbf{V}}(\tau)$ as:

$$\tilde{\mathbf{V}}(\tau) = \begin{cases} \boldsymbol{\Sigma}^{(1)} \otimes \begin{pmatrix} \frac{\lfloor T\tau_* \rfloor}{T}, & \frac{\lfloor T\tau \rfloor}{T} \\ \frac{\lfloor T\tau \rfloor}{T}, & \frac{\lfloor T\tau \rfloor}{T} \end{pmatrix} + \boldsymbol{\Sigma}^{(2)} \otimes \begin{pmatrix} \frac{T - \lfloor T\tau_* \rfloor}{T}, & 0 \\ 0, & 0 \end{pmatrix}, & \text{if } \tau \le \tau_*, \\ \boldsymbol{\Sigma}^{(1)} \otimes \begin{pmatrix} \frac{\lfloor T\tau_* \rfloor}{T}, & \frac{\lfloor T\tau_* \rfloor}{T} \\ \frac{\lfloor T\tau_* \rfloor}{T}, & \frac{\lfloor T\tau_* \rfloor}{T} \end{pmatrix} + \boldsymbol{\Sigma}^{(2)} \otimes \begin{pmatrix} \frac{T - \lfloor T\tau_* \rfloor}{T}, & \frac{\lfloor T\tau \rfloor - \lfloor T\tau_* \rfloor}{T} \\ \frac{\lfloor T\tau \rfloor - \lfloor T\tau_* \rfloor}{T}, & \frac{\lfloor T\tau \rfloor - \lfloor T\tau_* \rfloor}{T} \end{pmatrix}, & \text{if } \tau > \tau_*. \end{cases}$$

Hence, by Assumption 1 (a) and using the fact that $|\lfloor T\tau \rfloor/T - \tau| \le C/T$ over $\tau \in \mathcal{T}$, we have

$$\sup_{\tau \in \mathcal{T}} \|\tilde{\mathbf{V}}(\tau) - \mathbf{V}(\tau)\|_\infty = O(1/T). \tag{50}$$

Next, we bound $\sup_{\tau \in \mathcal{T}} \|\hat{\mathbf{V}}(\tau) - \tilde{\mathbf{V}}(\tau)\|_\infty$. To this end, for $1 \le j, k \le p$, we define

$$\begin{aligned} \hat{V}_{jk} = & \frac{1}{T} \sum_{t=1}^{T} \big(X_j^t X_k^t - \mathbb{E}(X_j^t X_k^t)\big), \\ = & \underbrace{\frac{1}{T} \sum_{t=1}^{\lfloor T\tau_* \rfloor} \big(X_j^t X_k^t - \mathbb{E}(X_j^t X_k^t)\big)}_{\hat{V}_{jk}^{(1)}} + \underbrace{\frac{1}{T} \sum_{t=\lfloor T\tau_* \rfloor + 1}^{T} \big(X_j^t X_k^t - \mathbb{E}(X_j^t X_k^t)\big)}_{\hat{V}_{jk}^{(2)}}, \end{aligned} \tag{51}$$

and

$$\begin{aligned} \hat{V}_{jk}(\tau) = & \frac{1}{T} \sum_{t=1}^{T} \big(X_j^t X_k^t \mathbf{1}\{Q_t \le \tau\} - \mathbb{E}(X_j^t X_k^t \mathbf{1}\{Q_t \le \tau\})\big), \\ = & \frac{1}{T} \sum_{t=1}^{\lfloor T\tau \rfloor} \big(X_j^t X_k^t - \mathbb{E}(X_j^t X_k^t)\big). \end{aligned} \tag{52}$$

By defining $\hat{V}_{jk}$ and $\hat{V}_{jk}(\tau)$ as in (51) and (52), to bound $\sup_{\tau \in \mathcal{T}} \|\hat{\mathbf{V}}(\tau) - \tilde{\mathbf{V}}(\tau)\|_\infty$, it is sufficient to consider $\max_{1 \le j,k \le p} |\hat{V}_{jk}|$ and $\sup_{\tau \in [t_0, t_1]} \max_{1 \le j,k \le p} |\hat{V}_{jk}(\tau)|$, respectively.

Firstly, we consider $\max_{j,k} |\hat{V}_{jk}|$. For any $x > 0$, using the triangle inequality, we have

$$\mathbb{P}\big(\max_{1 \le j,k \le p} |\hat{V}_{jk}| > x\big) \le p^2 \max_{1 \le j,k \le p} \Big(\mathbb{P}\big(|\hat{V}_{jk}^{(1)}| > x/2\big) + \mathbb{P}\big(|\hat{V}_{jk}^{(2)}| > x/2\big)\Big).$$

Note that X_j^t and X_k^t follow Gaussian distributions, which implies $X_j^t X_k^t$ follows sub-exponential distributions. Using Bernstein's inequality for sub-exponential distributions, for each j and k we have

$$\begin{aligned} & p^2 \mathbb{P}\big(|\hat{V}_{jk}^{(1)}| > x/2\big) \\ & \le p^2 \mathbb{P}\Big(\Big|\frac{1}{\lfloor T\tau_* \rfloor} \sum_{t=1}^{\lfloor T\tau_* \rfloor} \big(X_j^t X_k^t - \mathbb{E}(X_j^t X_k^t)\big)\Big| > x/2\Big), \\ & \le C_1 p^2 \exp(-C_2 \lfloor T\tau_* \rfloor x^2). \end{aligned} \tag{53}$$

For $p^2\mathbb{P}\big(|\hat{V}_{jk}^{(2)}| > x/2\big)$, using Bernstein's inequality again, we have

$$p^2\mathbb{P}\big(|\hat{V}_{jk}^{(2)}| > x/2\big) \leq C_1 p^2 \exp(-C_2\lfloor T(1-\tau_*)\rfloor x^2). \tag{54}$$

After bounding $\max_{j,k}|\hat{V}_{jk}|$ as in (53) and (54), we now consider $\sup_{\tau\in[t_0,t_1]}\max_{j,k}|\hat{V}_{jk}(\tau)|$. Note that there is a change point at $\tau_* \in [t_0, t_1]$. Hence, it is sufficient to consider

$$D_1 := \sup_{\tau\in[t_0,\tau_*]}\max_{1\leq j,k\leq p}|\hat{V}_{jk}(\tau)|, \quad D_2 := \sup_{\tau\in[\tau_*,t_1]}\max_{1\leq j,k\leq p}|\hat{V}_{jk}(\tau)|$$

respectively. Firstly, we consider D_1. For any $x > 0$, using Bernstein's inequality, we have

$$\begin{aligned}
&\mathbb{P}\big(\sup_{\tau\in[t_0,\tau_*]}\max_{1\leq j,k\leq p}|\hat{V}_{jk}(\tau)| > x\big) \\
&\leq \lfloor T(\tau_* - t_0)\rfloor p^2 \sup_{\tau\in[t_0,\tau_*]}\max_{1\leq j,k\leq p}\mathbb{P}\big(|\hat{V}_{jk}(\tau)| > x\big), \\
&\leq C_1 T p^2 \exp(-C_2\lfloor Tt_0\rfloor x^2),
\end{aligned} \tag{55}$$

where the second inequality comes from the fact that $\tau \in [t_0, \tau_*]$. Next, we consider D_2. Note that for $\tau \in [\tau_*, t_1]$, we can write $\hat{V}_{jk}(\tau)$ as in (52) as:

$$\hat{V}_{jk}(\tau) = \underbrace{\frac{1}{T}\sum_{t=1}^{\lfloor T\tau_*\rfloor}\Big(X_j^tX_k^t - \mathbb{E}(X_j^tX_k^t)\Big)}_{\hat{V}_{jk}^{(1)}(\tau)} + \underbrace{\frac{1}{T}\sum_{t=\lfloor T\tau_*\rfloor+1}^{\lfloor T\tau\rfloor}\Big(X_j^tX_k^t - \mathbb{E}(X_j^tX_k^t)\Big)}_{\hat{V}_{jk}^{(2)}(\tau)}.$$

Then, similar to the previous procedure, for $\hat{V}_{jk}^{(1)}(\tau)$ and $\hat{V}_{jk}^{(2)}(\tau)$, we have

$$\mathbb{P}\big(\sup_{\tau\in[\tau_*,t_1]}\max_{1\leq j,k\leq p}|\hat{V}_{jk}^{(1)}(\tau)| > x\big) \leq C_1 T p^2 \exp(-C_2\lfloor T\tau_*\rfloor x^2), \tag{56}$$

and

$$\mathbb{P}\big(\sup_{\tau\in[\tau_*,t_1]}\max_{1\leq j,k\leq p}|\hat{V}_{jk}^{(2)}(\tau)| > x\big) \leq C_1 T p^2 \exp(-C_2 T x^2). \tag{57}$$

Combining the results in (50), (53), (54), (55), (56), and (57), with probability at least $1-(pT)^{-C_1}$, we have

$$\sup_{\tau\in\mathcal{T}}\|\hat{\mathbf{V}}(\tau) - \mathbf{V}(\tau)\|_\infty \leq C_2\sqrt{\frac{\log(pT)}{T}}. \tag{58}$$

Step 3 : For any $\tau \in \mathcal{T}$, $J_0 \subset \{1,\ldots,2p\}$ with $|J_0| \leq s$, and $\boldsymbol{\gamma} \in \mathbb{R}^{2p}$ satisfying $|\boldsymbol{\gamma}_{J_0^c}|_1 \leq c_0|\boldsymbol{\gamma}_{J_0}|_1$, we have

$$\begin{aligned}
\frac{\boldsymbol{\gamma}^\top\hat{\mathbf{V}}(\tau)\boldsymbol{\gamma}}{|\boldsymbol{\gamma}_{J_0}|_2^2} &= \frac{\boldsymbol{\gamma}^\top\mathbf{V}(\tau)\boldsymbol{\gamma}}{|\boldsymbol{\gamma}_{J_0}|_2^2} + \frac{\boldsymbol{\gamma}^\top(\mathbf{V}(\tau)-\hat{\mathbf{V}}(\tau))\boldsymbol{\gamma}}{|\boldsymbol{\gamma}_{J_0}|_2^2}, \\
&\geq \frac{\boldsymbol{\gamma}^\top\mathbf{V}(\tau)\boldsymbol{\gamma}}{|\boldsymbol{\gamma}_{J_0}|_2^2} - \frac{\sup_{\tau\in\mathcal{T}}\|\hat{\mathbf{V}}(\tau)-\mathbf{V}(\tau)\|_\infty}{|\boldsymbol{\gamma}_{J_0}|_2^2}|\boldsymbol{\gamma}|_1^2, \\
&\geq \frac{\boldsymbol{\gamma}^\top\mathbf{V}(\tau)\boldsymbol{\gamma}}{|\boldsymbol{\gamma}_{J_0}|_2^2} - \frac{\sup_{\tau\in\mathcal{T}}\|\hat{\mathbf{V}}(\tau)-\mathbf{V}(\tau)\|_\infty}{|\boldsymbol{\gamma}_{J_0}|_2^2}(1+c_0)^2|\boldsymbol{\gamma}_{J_0}|_1^2, \\
&\geq \frac{\boldsymbol{\gamma}^\top\mathbf{V}(\tau)\boldsymbol{\gamma}}{|\boldsymbol{\gamma}_{J_0}|_2^2} - \sup_{\tau\in\mathcal{T}}\|\hat{\mathbf{V}}(\tau)-\mathbf{V}(\tau)\|_\infty(1+c_0)^2 s.
\end{aligned} \tag{59}$$

Note that $\mathbf{V}(\tau)$ satisfies the URE condition uniformly over $\tau \in \mathcal{T}$. Hence, by (58) and the last inequality as in (59), choosing $s = o(\sqrt{T/\log(pT)})$, we complete the proof. ■

C.2 Proof of Theorem 5

Proof In this section, we aim to prove

$$\mathbb{P}\Big(\Phi(\hat{\tau}, \{\hat{\boldsymbol{\beta}}^a(\hat{\tau}), 1 \le a \le p\}) = 0\Big) \ge 1 + o(1), \tag{60}$$

where $\Phi(\hat{\tau}, \{\hat{\boldsymbol{\beta}}^a(\hat{\tau}), 1 \le a \le p\}) := \mathbf{1}\Big\{\sum_{a=1}^p |\hat{\boldsymbol{\delta}}^a(\hat{\tau})|_1 \ge K_0 s_1 \lambda_T\Big\}$. To prove (60), we need to establish the estimation error bound of $\tilde{\boldsymbol{\beta}}^a$ with $1 \le a \le p$. In particular, we will prove

$$\sum_{a=1}^p \big|\hat{\boldsymbol{\beta}}^a(\hat{\tau}) - \tilde{\boldsymbol{\beta}}^a\big|_1 \le K_0 \lambda_T s_1.$$

Note that for a homogeneous model with $\tilde{\boldsymbol{\delta}}^a = \mathbf{0}$ for $1 \le a \le p$, we have $\sum_{a=1}^p \big\|\mathbf{X}(\hat{\tau})\tilde{\boldsymbol{\beta}}^a - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2 = 0$. Combining this result with (41) in Lemma 14, conditional on the event $\bigcap_{1\le a\le p}\{\mathcal{A}^a \cap \mathcal{B}^a\}$, we have

$$\begin{aligned} &\sum_{a=1}^p \big\|\mathbf{X}(\hat{\tau})\hat{\boldsymbol{\beta}}^a - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2 + \lambda_T(1-\mu)\sum_{a=1}^p \big|\hat{\mathbf{D}}(\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a)\big|_1 \\ &\le 2\lambda_T \sum_{a=1}^p \big|\hat{\mathbf{D}}(\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a)_{J_0^a}\big|_1, \end{aligned} \tag{61}$$

where $J_0^a := J(\tilde{\boldsymbol{\beta}}^a)$, $\hat{\boldsymbol{\beta}}^a := \hat{\boldsymbol{\beta}}^a(\hat{\tau})$, and $\hat{\mathbf{D}} := \mathbf{D}(\hat{\tau})$. To deal with (61), we define the following $2p^2$-dimensional vectors:

$$\tilde{\boldsymbol{\beta}}^{pp} = \big((\tilde{\boldsymbol{\beta}}^1)^\top, \ldots, (\tilde{\boldsymbol{\beta}}^p)^\top\big)^\top,\ \hat{\boldsymbol{\beta}}^{pp} = \big((\hat{\boldsymbol{\beta}}^1)^\top, \ldots, (\hat{\boldsymbol{\beta}}^p)^\top\big)^\top. \tag{62}$$

We also define the following $2p^2 \times 2p^2$ block diagonal matrix $\breve{\mathbf{D}}$ and the $Tp \times 2p^2$ design matrix $\tilde{\mathbf{X}}(\tau)$ as follows:

$$\breve{\mathbf{D}} = \mathrm{diag}\big\{\underbrace{\hat{\mathbf{D}}, \ldots, \hat{\mathbf{D}}}_{p}\big\},\ \tilde{\mathbf{X}}(\tau) = \mathrm{diag}\big\{\underbrace{\mathbf{X}(\tau), \ldots, \mathbf{X}(\tau)}_{p}\big\}. \tag{63}$$

Let $J_0 = J(\tilde{\boldsymbol{\beta}}^{pp})$ be the set of non-zero elements of $\tilde{\boldsymbol{\beta}}^{pp}$. By Assumption 2, we have $|J_0| = \sum_{a=1}^p |J_0^a| \le s_1$. Furthermore, combining (61), (62), and (63), we have

$$\big|\breve{\mathbf{D}}(\hat{\boldsymbol{\beta}}^{pp} - \tilde{\boldsymbol{\beta}}^{pp})_{J_0^c}\big|_1 \le \frac{1+\mu}{1-\mu}\big|\breve{\mathbf{D}}(\hat{\boldsymbol{\beta}}^{pp} - \tilde{\boldsymbol{\beta}}^{pp})_{J_0}\big|_1. \tag{64}$$

Define

$$\kappa'(s_1, c_0, \mathcal{S}) = \min_{\tau\in\mathcal{S}} \min_{\substack{J_0\subset\{1,\ldots,2p^2\}\\ |J_0|\le s_1}} \min_{\substack{\boldsymbol{\gamma}\neq 0\\ |\boldsymbol{\gamma}_{J_0^c}|_1 \le c_0|\boldsymbol{\gamma}_{J_0}|_1}} \frac{|\tilde{\mathbf{X}}(\tau)\boldsymbol{\gamma}|_2}{\sqrt{T}|\boldsymbol{\gamma}_{J_0}|_2}. \tag{65}$$

With a proof procedure similar to Proposition 3, we can show that $\tilde{\mathbf{X}}(\tau)$ satisfies the URE condition uniformly over $\tau \in \mathcal{T}$ by setting $s_1 = o(\sqrt{T/\log(pT)})$, i.e. $\kappa'(s_1, c_0, \mathcal{S}) > 0$ holds. Set $\kappa' := \kappa'(s_1, c_0, \mathcal{S})$ with $c_0 := (1+\mu)/(1-\mu)$ and $S := \mathcal{T} = [t_0, t_1]$. We then have

$$\begin{aligned}
\kappa'^2 \big|\breve{\mathbf{D}}(\hat{\boldsymbol{\beta}}^{pp} - \tilde{\boldsymbol{\beta}}^{pp})_{J_0}\big|_2^2 &\leq \frac{1}{T}\big|\tilde{\mathbf{X}}(\hat{\tau})\breve{\mathbf{D}}(\hat{\boldsymbol{\beta}}^{pp} - \tilde{\boldsymbol{\beta}}^{pp})\big|_2^2 \\
&\leq \frac{\max(\breve{\mathbf{D}})^2}{T}(\hat{\boldsymbol{\beta}}^{pp} - \tilde{\boldsymbol{\beta}}^{pp})^\top \tilde{\mathbf{X}}(\hat{\tau})^\top \tilde{\mathbf{X}}(\hat{\tau})(\hat{\boldsymbol{\beta}}^{pp} - \tilde{\boldsymbol{\beta}}^{pp}) \qquad (66) \\
&= \max(\breve{\mathbf{D}})^2 \sum_{a=1}^{p} \big\|\mathbf{X}(\hat{\tau})\hat{\boldsymbol{\beta}}^a - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2,
\end{aligned}$$

where the last equality comes from the assumption that $\tilde{\boldsymbol{\delta}}^a = \mathbf{0}$ for $1 \leq a \leq p$. Combining (61) and (66), we have

$$\begin{aligned}
\sum_{a=1}^{p} \big\|\mathbf{X}(\hat{\tau})\hat{\boldsymbol{\beta}}^a - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2 &\leq 2\lambda_T \big|\breve{\mathbf{D}}(\hat{\boldsymbol{\beta}}^{pp} - \tilde{\boldsymbol{\beta}}^{pp})_{J_0}\big|_1 \\
&\leq 2\lambda_T\sqrt{s_1}\big|\breve{\mathbf{D}}(\hat{\boldsymbol{\beta}}^{pp} - \tilde{\boldsymbol{\beta}}^{pp})_{J_0}\big|_2 \qquad (67) \\
&\leq 2\lambda_T\sqrt{s_1}\frac{\max(\breve{\mathbf{D}})}{\kappa'} \sum_{a=1}^{p} \big\|\mathbf{X}(\hat{\tau})\hat{\boldsymbol{\beta}}^a - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T.
\end{aligned}$$

The above inequality implies

$$\sum_{a=1}^{p} \big\|\mathbf{X}(\hat{\tau})\hat{\boldsymbol{\beta}}^a - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T \leq 2\lambda_T\sqrt{s_1}\frac{\max(\breve{\mathbf{D}})}{\kappa'}. \qquad (68)$$

On the other hand, using (64), we have

$$\begin{aligned}
\sum_{a=1}^{p} \big|\hat{\mathbf{D}}(\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a)\big|_1 &= \big|\breve{\mathbf{D}}(\hat{\boldsymbol{\beta}}^{pp} - \tilde{\boldsymbol{\beta}}^{pp})_{J_0^c}\big|_1 + \big|\breve{\mathbf{D}}(\hat{\boldsymbol{\beta}}^{pp} - \tilde{\boldsymbol{\beta}}^{pp})_{J_0}\big|_1 \\
&\leq \frac{2}{1-\mu}\big|\breve{\mathbf{D}}(\hat{\boldsymbol{\beta}}^{pp} - \tilde{\boldsymbol{\beta}}^{pp})_{J_0}\big|_1 \\
&\leq \frac{2}{1-\mu}\sqrt{s_1}\big|\breve{\mathbf{D}}(\hat{\boldsymbol{\beta}}^{pp} - \tilde{\boldsymbol{\beta}}^{pp})_{J_0}\big|_2 \\
&\leq \frac{2}{1-\mu}\sqrt{s_1}\frac{\max(\breve{\mathbf{D}})}{\kappa'} \sum_{a=1}^{p} \big\|\mathbf{X}(\hat{\tau})\hat{\boldsymbol{\beta}}^a - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T.
\end{aligned}$$

Combining (67) and (68), we have

$$\sum_{a=1}^{p} \big|\hat{\mathbf{D}}(\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a)\big|_1 \leq \frac{4\max(\breve{\mathbf{D}})^2}{(1-\mu)\kappa'^2}\lambda_T s_1 \leq \frac{4X_{\max}}{(1-\mu)\kappa'^2}\lambda_T s_1.$$

Using the fact that $\sum_{a=1}^{p} \big|\hat{\mathbf{D}}(\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a)\big|_1 \geq \min(\hat{\mathbf{D}}) \sum_{a=1}^{p} \big|\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a\big|_1$, we have

$$\sum_{a=1}^{p} \big|\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a\big|_1 = \sum_{a=1}^{p} \big|\hat{\boldsymbol{\theta}}^a - \tilde{\boldsymbol{\theta}}^a\big|_1 + \sum_{a=1}^{p} \big|\hat{\boldsymbol{\delta}}^a\big|_1 \leq \frac{4X_{\max}}{(1-\mu)X_{\min}\kappa'^2}\lambda_T s_1. \qquad (69)$$

The last inequality is due to $\tilde{\boldsymbol{\delta}}^a = \mathbf{0}$ with $1 \leq a \leq p$. Hence, choosing $K_0 = \dfrac{4X_{\max}}{(1-\mu)X_{\min}\kappa'^2}$ in (60), we complete the proof. ■

C.3 Proof of Proposition 6

Proof In this section, we prove (24) in Proposition 6. By the triangle inequality, we have

$$\sum_{a=1}^{p} \big\|\mathbf{X}(\tau)\boldsymbol{\beta}^a - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2 \geq \mathbb{E}\Big(\sum_{a=1}^{p} \big\|\mathbf{X}(\tau)\boldsymbol{\beta}^a - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2\Big) - \Delta, \tag{70}$$

where $\Delta := \big|\sum_{a=1}^{p} \|\mathbf{X}(\tau)\boldsymbol{\beta}^a - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\|_T^2 - \mathbb{E}\big(\sum_{a=1}^{p} \|\mathbf{X}(\tau)\boldsymbol{\beta}^a - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\|_T^2\big)\big|$. Therefore, by (70), to prove (24), we need two steps. In Step 1, we prove that there exists $c_* > 0$ such that

$$\mathbb{E}\Big(\sum_{a=1}^{p} \big\|\mathbf{X}(\tau)\boldsymbol{\beta}^a - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2\Big) > c_*\eta \tag{71}$$

holds for sufficiently large T and p. In Step 2, we prove $\Delta = o_p(1)$ holds uniformly over $\tau \in \mathcal{T}$. Next, we consider the two steps in detail.

Step 1 : For $|\tau - \tau_*| > \eta$, we consider two cases: $\tau \in [t_0, \tau_*]$ with $\tau_* - \tau > \eta$, and $\tau \in [\tau_*, t_1]$ with $\tau - \tau_* > \eta$. We first consider $\tau \in [t_0, \tau_*]$ with $\tau_* - \tau > \eta$. Recall $Q_t = t/T$, and $\boldsymbol{\beta}^a = ((\boldsymbol{\theta}^a)^\top, (\boldsymbol{\delta}^a)^\top)^\top$. We can write $\mathbb{E}\big(\sum_{a=1}^{p} \big\|\mathbf{X}(\tau)\boldsymbol{\beta}^a - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2\big)$ into three parts:

$$\mathbb{E}\big(\sum_{a=1}^{p} \big\|\mathbf{X}(\tau)\boldsymbol{\beta}^a - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2\big) = L_1 + L_2 + L_3, \tag{72}$$

where L_1, L_2, and L_3 are defined as:

$$L_1 = \frac{\lfloor T\tau \rfloor}{T} \sum_{a=1}^{p} (\boldsymbol{\theta}^a + \boldsymbol{\delta}^a - \tilde{\boldsymbol{\theta}}^a - \tilde{\boldsymbol{\delta}}^a)^\top \boldsymbol{\Sigma}^{(1)} (\boldsymbol{\theta}^a + \boldsymbol{\delta}^a - \tilde{\boldsymbol{\theta}}^a - \tilde{\boldsymbol{\delta}}^a),$$

$$L_2 = \frac{\lfloor T\tau_* \rfloor - \lfloor T\tau \rfloor}{T} \sum_{a=1}^{p} (\boldsymbol{\theta}^a - \tilde{\boldsymbol{\theta}}^a - \tilde{\boldsymbol{\delta}}^a)^\top \boldsymbol{\Sigma}^{(1)} (\boldsymbol{\theta}^a - \tilde{\boldsymbol{\theta}}^a - \tilde{\boldsymbol{\delta}}^a),$$

$$L_3 = \frac{\lfloor T(1-\tau_*) \rfloor}{T} \sum_{a=1}^{p} (\boldsymbol{\theta}^a - \tilde{\boldsymbol{\theta}}^a)^\top \boldsymbol{\Sigma}^{(2)} (\boldsymbol{\theta}^a - \tilde{\boldsymbol{\theta}}^a).$$

Therefore, by (72) and considering $L_1 \geq 0$, we have

$$\mathbb{E}\big(\sum_{a=1}^{p} \big\|\mathbf{X}(\tau)\boldsymbol{\beta}^a - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2\big) \geq L_2 + L_3. \tag{73}$$

By Assumptions 1 and 2, we have $\max(\lambda_{\max}(\boldsymbol{\Sigma}^{(1)}), \lambda_{\max}(\boldsymbol{\Sigma}^{(2)}) \leq K_1$, $|\boldsymbol{\theta}^a|_\infty \leq M_0$, $|\boldsymbol{\delta}^a|_\infty \leq M_0$, $|\tilde{\boldsymbol{\theta}}^a|_\infty \leq M_0$, and $|\tilde{\boldsymbol{\delta}}^a|_\infty \leq M_0$ for $1 \leq a \leq p$. Straightforward calculations show that there exists a positive constant C_0 such that

$$\sum_{a=1}^{p} (\boldsymbol{\theta}^a - \tilde{\boldsymbol{\theta}}^a - \tilde{\boldsymbol{\delta}}^a)^\top \boldsymbol{\Sigma}^{(1)} (\boldsymbol{\theta}^a - \tilde{\boldsymbol{\theta}}^a - \tilde{\boldsymbol{\delta}}^a) \leq C_0 s_1, \quad \sum_{a=1}^{p} (\boldsymbol{\theta}^a - \tilde{\boldsymbol{\theta}}^a)^\top \boldsymbol{\Sigma}^{(2)} (\boldsymbol{\theta}^a - \tilde{\boldsymbol{\theta}}^a) \leq C_0 s_1, \tag{74}$$

where s_1 comes from Assumption 2, and C_0 only depends on K_1 and M_0. Note that for any $\tau \in [t_0, \tau_*]$, we have

$$\Big|\frac{\lfloor T\tau_* \rfloor - \lfloor T\tau \rfloor}{T} - (\tau_* - \tau)\Big| = O(1/T), \text{ and } \Big|\frac{\lfloor T(1-\tau_*) \rfloor}{T} - (1-\tau_*)\Big| = O(1/T).$$

Hence, by (72) – (74), for sufficiently large T and p, we have

$$\begin{aligned}
&\mathbb{E}\big(\sum_{a=1}^{p} \big\|\mathbf{X}(\tau)\boldsymbol{\beta}^a - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2\big) \\
&\geq (\tau_* - \tau)\sum_{a=1}^{p}(\boldsymbol{\theta}^a - \tilde{\boldsymbol{\theta}}^a - \tilde{\boldsymbol{\delta}}^a)^\top \boldsymbol{\Sigma}^{(1)}(\boldsymbol{\theta}^a - \tilde{\boldsymbol{\theta}}^a - \tilde{\boldsymbol{\delta}}^a) + (1-\tau_*)\sum_{a=1}^{p}(\boldsymbol{\theta}^a - \tilde{\boldsymbol{\theta}}^a)^\top \boldsymbol{\Sigma}^{(2)}(\boldsymbol{\theta}^a - \tilde{\boldsymbol{\theta}}^a) + O(s_1/T), \\
&= (\tau_* - \tau)\sum_{a=1}^{p}(\tilde{\boldsymbol{\delta}}^a)^\top \boldsymbol{\Sigma}^{(1)}\tilde{\boldsymbol{\delta}}^a + (\tau_* - \tau)\sum_{a=1}^{p}(\boldsymbol{\theta}^a - \tilde{\boldsymbol{\theta}}^a)^\top \boldsymbol{\Sigma}^{(1)}(\boldsymbol{\theta}^a - \tilde{\boldsymbol{\theta}}^a) - 2(\tau_* - \tau)\sum_{a=1}^{p}(\boldsymbol{\theta}^a - \tilde{\boldsymbol{\theta}}^a)^\top \boldsymbol{\Sigma}^{(1)}\tilde{\boldsymbol{\delta}}^a \\
&\quad + (1-\tau_*)\sum_{a=1}^{p}(\boldsymbol{\theta}^a - \tilde{\boldsymbol{\theta}}^a)^\top \boldsymbol{\Sigma}^{(2)}(\boldsymbol{\theta}^a - \tilde{\boldsymbol{\theta}}^a) + O(s_1/T).
\end{aligned} \tag{75}$$

By the definitions of $\underline{\kappa} := \min(\lambda_{\min}(\boldsymbol{\Sigma}^{(1)}), \lambda_{\min}(\boldsymbol{\Sigma}^{(2)}))$, and $\overline{\kappa} := \max(\lambda_{\max}(\boldsymbol{\Sigma}^{(1)}), \lambda_{\max}(\boldsymbol{\Sigma}^{(2)}))$, we have

$$\begin{aligned}
&(\tau_* - \tau)\sum_{a=1}^{p}(\boldsymbol{\theta}^a - \tilde{\boldsymbol{\theta}}^a)^\top \boldsymbol{\Sigma}^{(1)}(\boldsymbol{\theta}^a - \tilde{\boldsymbol{\theta}}^a) + (1-\tau_*)\sum_{a=1}^{p}(\boldsymbol{\theta}^a - \tilde{\boldsymbol{\theta}}^a)^\top \boldsymbol{\Sigma}^{(2)}(\boldsymbol{\theta}^a - \tilde{\boldsymbol{\theta}}^a), \\
&\geq (1-\tau)\underline{\kappa}\sum_{a=1}^{p}(\boldsymbol{\theta}^a - \tilde{\boldsymbol{\theta}}^a)^\top(\boldsymbol{\theta}^a - \tilde{\boldsymbol{\theta}}^a), \\
&\geq \frac{(1-\tau)\underline{\kappa}}{\overline{\kappa}}\sum_{a=1}^{p}(\boldsymbol{\theta}^a - \tilde{\boldsymbol{\theta}}^a)^\top \boldsymbol{\Sigma}^{(1)}(\boldsymbol{\theta}^a - \tilde{\boldsymbol{\theta}}^a).
\end{aligned} \tag{76}$$

Hence, by (75) and (76), to prove (71), we consider two cases:

Case 1: $\frac{(1-\tau)\underline{\kappa}}{\overline{\kappa}}\sum_{a=1}^{p}(\boldsymbol{\theta}^a - \tilde{\boldsymbol{\theta}}^a)^\top \boldsymbol{\Sigma}^{(1)}(\boldsymbol{\theta}^a - \tilde{\boldsymbol{\theta}}^a) - 2(\tau_* - \tau)\sum_{a=1}^{p}(\boldsymbol{\theta}^a - \tilde{\boldsymbol{\theta}}^a)^\top \boldsymbol{\Sigma}^{(1)}\tilde{\boldsymbol{\delta}}^a \geq 0$. In this case, considering (75), for sufficiently large T and p, we have

$$\mathbb{E}\big(\sum_{a=1}^{p}\big\|\mathbf{X}(\tau)\boldsymbol{\beta}^a - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2\big) \geq (\tau_* - \tau)\sum_{a=1}^{p}(\tilde{\boldsymbol{\delta}}^a)^\top \boldsymbol{\Sigma}^{(1)}\tilde{\boldsymbol{\delta}}^a.$$

Case 2: $\frac{(1-\tau)\underline{\kappa}}{\overline{\kappa}}\sum_{a=1}^{p}(\boldsymbol{\theta}^a - \tilde{\boldsymbol{\theta}}^a)^\top \boldsymbol{\Sigma}^{(1)}(\boldsymbol{\theta}^a - \tilde{\boldsymbol{\theta}}^a) < 2(\tau_* - \tau)\sum_{a=1}^{p}(\boldsymbol{\theta}^a - \tilde{\boldsymbol{\theta}}^a)^\top \boldsymbol{\Sigma}^{(1)}\tilde{\boldsymbol{\delta}}^a$. In this case, we first define the following p^2-dimensional vectors as:

$$\begin{aligned}
\boldsymbol{\theta}^{pp} &:= \Big(\big((\boldsymbol{\Sigma}^{(1)})^{1/2}\boldsymbol{\theta}^1\big)^\top, \ldots, \big((\boldsymbol{\Sigma}^{(1)})^{1/2}\boldsymbol{\theta}^p\big)^\top\Big)^\top, \\
\tilde{\boldsymbol{\theta}}^{pp} &:= \Big(\big((\boldsymbol{\Sigma}^{(1)})^{1/2}\tilde{\boldsymbol{\theta}}^1\big)^\top, \ldots, \big((\boldsymbol{\Sigma}^{(1)})^{1/2}\tilde{\boldsymbol{\theta}}^p\big)^\top\Big)^\top, \\
\tilde{\boldsymbol{\delta}}^{pp} &:= \Big(\big((\boldsymbol{\Sigma}^{(1)})^{1/2}\tilde{\boldsymbol{\delta}}^1\big)^\top, \ldots, \big((\boldsymbol{\Sigma}^{(1)})^{1/2}\tilde{\boldsymbol{\delta}}^p\big)^\top\Big)^\top.
\end{aligned} \tag{77}$$

By (77), we have $\sum_{a=1}^{p}(\boldsymbol{\theta}^a - \tilde{\boldsymbol{\theta}}^a)^\top \boldsymbol{\Sigma}^{(1)}(\boldsymbol{\theta}^a - \tilde{\boldsymbol{\theta}}^a) = |\boldsymbol{\theta}^{pp} - \tilde{\boldsymbol{\theta}}^{pp}|_2^2$. Hence, in Case 2, we have

$$\frac{(1-\tau)\underline{\kappa}}{\overline{\kappa}}|\boldsymbol{\theta}^{pp} - \tilde{\boldsymbol{\theta}}^{pp}|_2^2 < 2(\tau_* - \tau)(\boldsymbol{\theta}^{pp} - \tilde{\boldsymbol{\theta}}^{pp})^\top \tilde{\boldsymbol{\delta}}^{pp}. \tag{78}$$

Let J_0 be the set of non-zero elements of the p^2-dimensional vector $\tilde{\boldsymbol{\delta}}^{pp}$ as defined in (77). Considering (78), and by the Cauchy-Swartz inequality, we have

$$\frac{(1-\tau)\underline{\kappa}}{\overline{\kappa}}|\boldsymbol{\theta}^{pp}-\tilde{\boldsymbol{\theta}}^{pp}|_2^2 < 2(\tau_*-\tau)(\boldsymbol{\theta}^{pp}-\tilde{\boldsymbol{\theta}}^{pp})_{J_0}^\top \tilde{\boldsymbol{\delta}}_{J_0}^{pp} < 2(\tau_*-\tau)|(\boldsymbol{\theta}^{pp}-\tilde{\boldsymbol{\theta}}^{pp})_{J_0}|_2|\tilde{\boldsymbol{\delta}}_{J_0}^{pp}|_2. \quad (79)$$

For $|(\boldsymbol{\theta}^{pp}-\tilde{\boldsymbol{\theta}}^{pp})_{J_0}|_2$, by (79), we have

$$|(\boldsymbol{\theta}^{pp}-\tilde{\boldsymbol{\theta}}^{pp})_{J_0}|_2 \le \frac{|\boldsymbol{\theta}^{pp}-\tilde{\boldsymbol{\theta}}^{pp}|_2^2}{|(\boldsymbol{\theta}^{pp}-\tilde{\boldsymbol{\theta}}^{pp})_{J_0}|_2} < \frac{2(\tau_*-\tau)}{(1-\tau)\underline{\kappa}/\overline{\kappa}}|\tilde{\boldsymbol{\delta}}_{J_0}^{pp}|_2 < \frac{2(\tau_*-t_0)}{(1-t_0)\underline{\kappa}/\overline{\kappa}}|\tilde{\boldsymbol{\delta}}_{J_0}^{pp}|_2, \quad (80)$$

where the last inequality as in (80) comes from the fact that $\tau \in [t_0, \tau_*]$.

Based on (75), (76), (77), and (80), in Case 2, we further have

$$\begin{aligned}
&\mathbb{E}\big(\sum_{a=1}^{p}\big\|\mathbf{X}(\tau)\boldsymbol{\beta}^a-\mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2\big)\\
&\ge (\tau_*-\tau)|\tilde{\boldsymbol{\delta}}_{J_0}^{pp}|_2^2+\frac{(1-\tau)\underline{\kappa}}{\overline{\kappa}}|\boldsymbol{\theta}^{pp}-\tilde{\boldsymbol{\theta}}^{pp}|_2^2-2(\tau_*-\tau)(\boldsymbol{\theta}^{pp}-\tilde{\boldsymbol{\theta}}^{pp})_{J_0}^\top\tilde{\boldsymbol{\delta}}_{J_0}^{pp},\\
&=(\tau_*-\tau)|\tilde{\boldsymbol{\delta}}_{J_0}^{pp}-(\boldsymbol{\theta}^{pp}-\tilde{\boldsymbol{\theta}}^{pp})_{J_0}|_2^2+\frac{(1-\tau)\underline{\kappa}}{\overline{\kappa}}|\boldsymbol{\theta}^{pp}-\tilde{\boldsymbol{\theta}}^{pp}|_2^2,\\
&\quad-(\tau_*-\tau)|(\boldsymbol{\theta}^{pp}-\tilde{\boldsymbol{\theta}}^{pp})_{J_0}|_2^2.
\end{aligned} \quad (81)$$

By Assumption 4, we have $\frac{(1-\tau)\underline{\kappa}}{\overline{\kappa}} \ge (\tau_*-\tau)$ for $\tau \in [t_0, \tau_*]$. Considering that $|\boldsymbol{\theta}^{pp}-\tilde{\boldsymbol{\theta}}^{pp}|_2^2 \ge |(\boldsymbol{\theta}^{pp}-\tilde{\boldsymbol{\theta}}^{pp})_{J_0}|_2^2$, we have

$$\frac{(1-\tau)\underline{\kappa}}{\overline{\kappa}}|\boldsymbol{\theta}^{pp}-\tilde{\boldsymbol{\theta}}^{pp}|_2^2-(\tau_*-\tau)|(\boldsymbol{\theta}^{pp}-\tilde{\boldsymbol{\theta}}^{pp})_{J_0}|_2^2 \ge 0. \quad (82)$$

Define $c_0 = \frac{2(\tau_*-t_0)}{(1-t_0)\underline{\kappa}/\overline{\kappa}}$. Note that under Assumption 4, we have $0 < c_0 < 1$. Then, by (80) and (81), and (82), for sufficiently large T and p, we have

$$\begin{aligned}
\mathbb{E}\big(\sum_{a=1}^{p}\big\|\mathbf{X}(\tau)\boldsymbol{\beta}^a-\mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2\big) &\ge (\tau_*-\tau)(1-c_0)^2|\tilde{\boldsymbol{\delta}}_{J_0}^{pp}|_2^2\\
&=(\tau_*-\tau)(1-c_0)^2\sum_{a=1}^{p}(\tilde{\boldsymbol{\delta}}^a)^\top\boldsymbol{\Sigma}^{(1)}\tilde{\boldsymbol{\delta}}^a.
\end{aligned} \quad (83)$$

After getting the lower bound of $\mathbb{E}\big(\sum_{a=1}^{p}\big\|\mathbf{X}(\tau)\boldsymbol{\beta}^a-\mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2\big)$ as in (83) for $\tau \in [t_0, \tau_*]$ with $\tau_*-\tau > \eta$, we next consider the case for $\tau \in [\tau_*, t_1]$ with $\tau-\tau_* > \eta$. Define $c_1 = \frac{2(t_1-\tau_*)}{t_1\underline{\kappa}/\overline{\kappa}}$. Under Assumption 4, we have $0 < c_1 < 1$. Furthermore, under Assumptions 1 – 4, using a proof procedure similar to the case of $\tau \in [\tau_*, t_1]$, we can prove

$$\mathbb{E}\big(\sum_{a=1}^{p}\big\|\mathbf{X}(\tau)\boldsymbol{\beta}^a-\mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2\big) \ge (\tau-\tau_*)(1-c_1)^2\sum_{a=1}^{p}(\tilde{\boldsymbol{\delta}}^a)^\top\boldsymbol{\Sigma}^{(2)}\tilde{\boldsymbol{\delta}}^a \quad (84)$$

holds for sufficiently large T and p.

Note that under Assumption 3, we have $\min\Big(\sum_{a=1}^{p}(\tilde{\boldsymbol{\delta}}^a)^\top\boldsymbol{\Sigma}^{(1)}\tilde{\boldsymbol{\delta}}^a, \sum_{a=1}^{p}(\tilde{\boldsymbol{\delta}}^a)^\top\boldsymbol{\Sigma}^{(2)}\tilde{\boldsymbol{\delta}}^a\Big) > \delta_*$ for some $\delta_* > 0$. Finally, combining (83) and (84), for any τ and η with $|\tau - \tau_*| > \eta$, there exists $c_* > 0$ such that

$$\mathbb{E}\Big(\sum_{a=1}^{p}\big\|\mathbf{X}(\tau)\boldsymbol{\beta}^a - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2\Big) > c_*\eta \tag{85}$$

holds for sufficiently large T and p, where $c_* := \min\big(\delta_*(1-c_0)^2, \delta_*(1-c_1)^2\big)$.
Step 2: Next, we prove $\Delta = o_p(1)$ uniformly over $\tau \in [t_0, t_1]$, where

$$\Delta := \Big|\sum_{a=1}^{p}\|\mathbf{X}(\tau)\boldsymbol{\beta}^a - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\|_T^2 - \mathbb{E}\big(\sum_{a=1}^{p}\|\mathbf{X}(\tau)\boldsymbol{\beta}^a - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\|_T^2\big)\Big|. \tag{86}$$

We first consider $\tau \in [t_0, \tau_*]$. By definition, we can decompose Δ as in (86) into three parts:

$$\Delta = \Delta_1 + \Delta_2 + \Delta_3, \tag{87}$$

where

$$\begin{aligned}
\Delta_1 &:= \sum_{a=1}^{p}(\boldsymbol{\beta}^a - \tilde{\boldsymbol{\beta}}^a)^\top\Big\{\frac{1}{T}\sum_{t=1}^{\lfloor T\tau\rfloor}\big(\mathbf{X}^t(\mathbf{X}^t)^\top - \boldsymbol{\Sigma}^{(1)}\big)\Big\}(\boldsymbol{\beta}^a - \tilde{\boldsymbol{\beta}}^a),\\
\Delta_2 &:= \sum_{a=1}^{p}(\boldsymbol{\theta}^a - \tilde{\boldsymbol{\theta}}^a - \tilde{\boldsymbol{\delta}}^a)^\top\Big\{\frac{1}{T}\sum_{t=\lfloor T\tau\rfloor+1}^{\lfloor T\tau_*\rfloor}\big(\mathbf{X}^t(\mathbf{X}^t)^\top - \boldsymbol{\Sigma}^{(1)}\big)\Big\}(\boldsymbol{\theta}^a - \tilde{\boldsymbol{\theta}}^a - \tilde{\boldsymbol{\delta}}^a),\\
\Delta_3 &:= \sum_{a=1}^{p}(\boldsymbol{\theta}^a - \tilde{\boldsymbol{\theta}}^a)^\top\Big\{\frac{1}{T}\sum_{t=\lfloor T\tau_*\rfloor+1}^{T}\big(\mathbf{X}^t(\mathbf{X}^t)^\top - \boldsymbol{\Sigma}^{(2)}\big)\Big\}(\boldsymbol{\theta}^a - \tilde{\boldsymbol{\theta}}^a).
\end{aligned}$$

By (87), to prove $\Delta = o_p(1)$, it is sufficient to consider Δ_1, Δ_2, and Δ_3, respectively. We first consider Δ_1. Recall $\hat{V}_{jk}(\tau) = \frac{1}{T}\sum_{t=1}^{\lfloor T\tau\rfloor}\big(X_j^tX_k^t - \mathbb{E}(X_j^tX_k^t)\big)$ as defined in (52). Following the proof technique as in Section C.1, we have

$$\sup_{\tau\in[t_0,\tau_*]}\Big\|\frac{1}{T}\sum_{t=1}^{\lfloor T\tau\rfloor}\big(\mathbf{X}^t(\mathbf{X}^t)^\top - \boldsymbol{\Sigma}^{(1)}\big)\Big\|_\infty = \sup_{\tau\in[t_0,\tau_*]}\max_{1\le j,k\le p}|\hat{V}_{jk}(\tau)| = O_p\big(\sqrt{\log(pT)/T}\big). \tag{88}$$

Therefore, by (88), using $(a+b)^2 \le 2a^2 + 2b^2$, we have

$$\begin{aligned}
\Delta_1 &\le \sum_{a=1}^{p}|(\boldsymbol{\beta}^a - \tilde{\boldsymbol{\beta}}^a)|_1^2\sup_{\tau\in[t_0,\tau_*]}\max_{1\le j,k\le p}|\hat{V}_{jk}(\tau)|,\\
&\le \sum_{a=1}^{p}2(|\boldsymbol{\beta}^a|_1^2 + |\tilde{\boldsymbol{\beta}}^a|_1^2)\sup_{\tau\in[t_0,\tau_*]}\max_{1\le j,k\le p}|\hat{V}_{jk}(\tau)|,\\
&\le 4s_1M_0^2O_p(\sqrt{T/\log(pT)}),
\end{aligned} \tag{89}$$

where the last inequality in (89) comes from Assumption 2, i.e. $\sum_{a=1}^{p}\mathcal{M}(\boldsymbol{\beta}^a) \le s_1$, $\sum_{a=1}^{p}\mathcal{M}(\tilde{\boldsymbol{\beta}}^a) \le s_1$, $|\boldsymbol{\beta}^a|_\infty \le M_0$, and $|\tilde{\boldsymbol{\beta}}^a|_\infty \le M_0$ for $1 \le a \le p$. Considering $s_1 = o(\sqrt{T/\log(pT)})$, we have $\Delta_1 = o_p(1)$ uniformly over $\tau \in [t_0, \tau_*]$. Similarly, for Δ_2 and Δ_3,

we can also prove $\Delta_2 = o_p(1)$ and $\Delta_3 = o_p(1)$ for $\tau \in [t_0, \tau_*]$. Then, by (87), we have $\Delta = o_p(1)$ over $\tau \in [t_0, \tau_*]$. Using a similar proof procedure, we can prove $\Delta = o_p(1)$ uniformly for $\tau \in [\tau_*, t_1]$. Finally, we have proved that

$$\Delta = o_p(1), \quad \text{uniformly over} \quad \tau \in [t_0, t_1]. \tag{90}$$

Combining (70), (85), and (90), for sufficiently large T and p,

$$\sum_{a=1}^{p} \|\mathbf{X}(\tau)\boldsymbol{\beta}^a - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\|_T^2 > c_* \eta$$

holds with probability at least $1 - (pT)^{-C}$ for some constant $C > 0$, which finishes the proof. ■

C.4 Proof of Theorem 7 and Proposition 9

Proof The proof of Theorem 7 mainly relies on Lemmas 17 and 18. By Lemma 17, if we have prior upper bounds of $|\hat{\tau} - \tau_*|$ and $\sum_{a=1}^{p} |\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a|_1$, respectively, say c_τ and $c_{\boldsymbol{\beta}}$, then using (47), we can tighten the upper bound of $\sum_{a=1}^{p} |\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a|_1$. Furthermore, by (48) as in Lemma 18, with the tightened upper bound for estimation error in $(\tilde{\boldsymbol{\beta}})_{a=1}^{p}$, we can further tighten the bound of $|\hat{\tau} - \tau_*|$. The above analysis motivates us to adopt a chaining technique by iteratively applying Lemmas 17 and 18 to tighten the upper bound for estimation errors in τ_* and $(\tilde{\boldsymbol{\beta}})_{a=1}^{p}$, respectively. Note that the same idea is also adopted by Lee et al. (2016).

In particular, let $c_\tau^{(m)}$ and $c_{\boldsymbol{\beta}}^{(m)}$ denote the bounds of $|\hat{\tau} - \tau_*|$ and $\sum_{a=1}^{p} |\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a|_1$, respectively, in the m-th iteration. Our iterative procedure can be described briefly as follows:

$$\big(c_\tau^{(m)}, c_{\boldsymbol{\beta}}^{(m)}\big) \xrightarrow[(47)]{\textit{Lemma 17}} \big(c_\tau^{(m)}, c_{\boldsymbol{\beta}}^{(m+1)}\big) \xrightarrow[(48)]{\textit{Lemma 18}} \big(c_\tau^{(m+1)}, c_{\boldsymbol{\beta}}^{(m+1)}\big). \tag{91}$$

Note that by (47), there are three terms in the upper bound of $\sum_{a=1}^{p} |\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a|_1$. In this section, we aim to show that, after a finite number of iterations, the term $6X_{\max}^2 \kappa'^{-2} \lambda_T s_1$ dominates the other two terms. In other words, we aim to show that

$$\begin{aligned} &\Big\{ \frac{6X_{\max}^2}{\kappa'^2} \lambda_T s_1 \bigvee \frac{2X_{\max}}{\kappa'} \Big(s_1 c_{\boldsymbol{\beta}}^{(m)} c_\tau^{(m)} C^* \sum_{a=1}^{p} |\tilde{\boldsymbol{\delta}}^a|_1 \Big)^{1/2} \bigvee \Big(\sqrt{c_\tau^{(m)}} + (2X_{\min})^{-1} c_\tau^{(m)} C^* \sum_{a=1}^{p} |\tilde{\boldsymbol{\delta}}^a|_1 \Big) \Big\}, \\ &= \frac{6X_{\max}^2}{\kappa'^2} \lambda_T s_1. \end{aligned} \tag{92}$$

holds for some $m = m^*$ with $m^* < \infty$. Then, if (92) holds, by (91), (46), and (47), we can obtain the final upper bounds of $\sum_{a=1}^{p} \big\|\mathbf{X}(\hat{\tau})\hat{\boldsymbol{\beta}}^a(\hat{\tau}) - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2$ and $\sum_{a=1}^{p} \big|\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a\big|_1$ as follows:

$$\sum_{a=1}^{p} \big\|\mathbf{X}(\hat{\tau})\hat{\boldsymbol{\beta}}^a(\hat{\tau}) - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2 \leq \underbrace{\frac{18X_{\max}^2}{\kappa'^2}}_{M_2} \lambda_T^2 s_1, \tag{93}$$

and

$$\sum_{a=1}^{p} \left|\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a\right|_1 \leq \underbrace{\frac{3}{(1-\mu)X_{\min}} \frac{6X_{\max}^2}{\kappa'^2}}_{M_3} \lambda_T s_1 := c^*_{\boldsymbol{\beta}}. \tag{94}$$

Considering (93) and (94), we finish the proof of Proposition 9 . Furthermore, by (91), (92), (94), and (48), we can obtain the final upper bound of $|\hat{\tau} - \tau_*|$ as follows:

$$\begin{aligned} |\hat{\tau} - \tau_*| &\leq c_*^{-1}\lambda_T \Big((1+\mu) c^*_{\boldsymbol{\beta}} X_{\max} + \sqrt{c_\tau^{(m^*)}} + (2X_{\min})^{-1} c_\tau^{(m^*)} C^* \sum_{a=1}^{p} |\tilde{\boldsymbol{\delta}}^a|_1 \Big), \\ &\leq \underbrace{\Big(\frac{3(1+\mu)}{1-\mu} \frac{X_{\max}}{X_{\min}} + 1 \Big) \frac{6X_{\max}^2}{\kappa'^2 c_*}}_{M_1} \lambda_T^2 s_1 := c^*_\tau. \end{aligned} \tag{95}$$

Considering (95), we complete the proof of Theorem 7.

In what follows, we will prove that, after a finite number of iterations, say m^*, (92) holds. To this end, we need some notations and conditions. Define

$$H_1 = \frac{3(1+\mu)}{1-\mu} \frac{X_{\max}}{X_{\min}} + 1, H_2 = \frac{C^*}{2X_{\min} c_*}, H_3 = \frac{6X_{\max}^2 c_*}{\kappa'^2},\ H_4 = \frac{36(1+\mu)X_{\max}^3}{(1-\mu)^2 X_{\min}}.$$

With the above notations, we require the following conditions (96) – (100) hold:

$$H_1 H_2 \lambda_T \sum_{a=1}^{p} |\tilde{\boldsymbol{\delta}}^a|_1 < 1, \tag{96}$$

$$\frac{H_1}{\big(1 - H_1 H_2 \lambda_T \sum_{a=1}^{p} |\tilde{\boldsymbol{\delta}}^a|_1\big)^2} < H_3 s_1, \tag{97}$$

$$(2\kappa'^{-2} H_4 + 1) H_2 \lambda_T \sum_{a=1}^{p} |\tilde{\boldsymbol{\delta}}^a|_1 < 1, \tag{98}$$

$$\frac{1}{\big(1 - (2\kappa'^{-2} H_4 s_1 + 1) H_2 \lambda_T \sum_{a=1}^{p} |\tilde{\boldsymbol{\delta}}^a|_1\big)^2} < H_1 H_3 s_1, \tag{99}$$

$$\frac{H_2 \lambda_T \sum_{a=1}^{p} |\tilde{\boldsymbol{\delta}}^a|_1}{\big(1 - (2\kappa'^{-2} H_4 s_1 + 1) H_2 \lambda_T \sum_{a=1}^{p} |\tilde{\boldsymbol{\delta}}^a|_1\big)^2} < \frac{(1-\mu)c_*}{4}. \tag{100}$$

Note that if $\lambda_T s_1 \sum_{a=1}^{p} |\tilde{\boldsymbol{\delta}}^a|_1 \to 0$ as $p, T \to \infty$, the above conditions (96) – (100) hold. Furthermore, with a similar proof procedure using in Lee et al. (2016), under conditions (96) – (100), we can prove that, after a finite number of iterations, (92) holds, which yields the desired upper bounds as in (93), (94), and (95), respectively. To save space, we omit the details here.

Finally, we show that Lemmas 17 and 18 hold with a high probability through the m^* iterations. In particular, we show that the event $\bigcap_{1 \leq a \leq p} \{\mathcal{A}^a \cap \mathcal{B}^a\} \bigcap \{\bigcap_{j=1}^{m^*} \mathcal{C}(\eta_j)\}$ occurs

with a high probability. Let $\Phi(x)$ be the CDF of the standard normal distribution. We have $2\Phi(-x) \leq \exp(-x^2/2)$ for $x \geq \sqrt{2/\pi}$. Hence, by Lemmas 15 and 16, conditional on the events $\mathcal{E}^{(1)}$, $\mathcal{E}^{(2)}$, and $\mathcal{E}^{(3)}$ as defined in (42), (44), and (45), we have

$$\begin{aligned}
&\mathbb{P}\Big(\bigcap\nolimits_{1\leq a\leq p}\{\mathcal{A}^a \cap \mathcal{B}^a\} \bigcap \big\{\bigcap\nolimits_{j=1}^{m*} \mathcal{C}(\eta_j)\big\}\Big) \\
&\geq 1 - 3p^2 \exp\left(\frac{-\mu^2 T}{8\omega^2}\frac{K_2^2}{K_1^2}\lambda_T^2\right) - 2\sum_{j=1}^{m^*}\exp\left(\frac{-\mu^2 T}{8\omega^4 C_1^{(j)}\underline{\kappa}^{-1}\delta^*}\frac{K_2^2}{K_1^2}\lambda_T^2\right) \\
&\quad - 2\sum_{j=1}^{m^*}\exp\left(\frac{-\mu^2}{8\omega^4 C_2^{(j)}\underline{\kappa}^{-1}\delta^*}\frac{K_2^2}{K_1^2}\lambda_T^2\right), \\
&\geq 1 - 3p^2 \exp\left(\frac{-\mu^2 T}{8\omega^2}\frac{K_2^2}{K_1^2}\lambda_T^2\right) - 4m^* \exp\left(\frac{-\mu^2 T}{8\omega^2 A^*}\frac{K_2^2}{K_1^2}\lambda_T^2\right),
\end{aligned} \tag{101}$$

where $A^* := \max_{1\leq j\leq m^*, 1\leq k\leq 2}(\omega^2 C_k^{(j)}\underline{\kappa}^{-1}\delta^*)$. Taking $A_1^* = 4K_1/K_2$ and $A_2^* = 2\sqrt{2A^*}K_1/K_2$ as in Assumption 5, and combining (101), conditional on the events $\mathcal{E}^{(1)}$, $\mathcal{E}^{(2)}$ and $\mathcal{E}^{(3)}$, we have

$$\mathbb{P}\Big(\bigcap_{1\leq a\leq p}\{\mathcal{A}^a \cap \mathcal{B}^a\} \bigcap \big\{\bigcap_{j=1}^{m*} \mathcal{C}(\eta_j)\big\}\Big) \geq 1 - 3(p^2)^{1-A^2\mu^2/(A_1^*)^2} - 4m^* p^{1-A^2\mu^2/(A_2^*)^2}.$$

Note that, the events $\mathcal{E}^{(1)}$, $\mathcal{E}^{(2)}$ and $\mathcal{E}^{(3)}$ occur with a probability at least $1 - Q_1(pT)^{-Q_2}$ for some universal constants $Q_1 > 0$ and $Q_2 > 0$

Finally, taking $C_1 \equiv 3$, $C_2 \equiv 4m^*$, $C_3 \equiv Q_1$, and $C_4 \equiv Q_2$ in Theorem 7 and Proposition 9, we complete the proof. ■

C.5 Proof of Theorem 11

Proof The proofs of (28) and (29) are quite similar. To save space, we only consider (28) here. The proof of (28) proceeds in two steps. In Step 1, we prove $E^{(1)} \subset \breve{E}^{(1)}$. In Step 2, we prove $|\breve{E}^{(1)} \cap (E^{(1)})^c| \leq 2M_3/r_1^*$. Now, we consider the two steps in details.

Step 1 : Recall $\hat{E}_{\text{init}}^{(1)}$ as defined in (12). By definitions, we have

$$\begin{aligned}
&|\hat{E}_{\text{init}}^{(1)} \cap (E^{(1)})^c| \\
&= \sum_{(a,b)\in(E^{(1)})^c} \mathbf{1}\big\{|\hat{\theta}_b^a(\hat{\tau}) + \hat{\delta}_b^a(\hat{\tau})| \geq r_0\lambda_T \cup |\hat{\theta}_a^b(\hat{\tau}) + \hat{\delta}_a^b(\hat{\tau})| \geq r_0\lambda_T\big\}, \\
&\leq \sum_{(a,b)\in(E^{(1)})^c} \mathbf{1}\big\{|\hat{\theta}_b^a(\hat{\tau}) + \hat{\delta}_b^a(\hat{\tau})| \geq r_0\lambda_T\big\} + \sum_{(a,b)\in(E^{(1)})^c} \mathbf{1}\big\{|\hat{\theta}_a^b(\hat{\tau}) + \hat{\delta}_a^b(\hat{\tau})| \geq r_0\lambda_T\big\},
\end{aligned} \tag{102}$$

where the last inequality comes from $\mathbf{1}\{A \cup B\} \leq \mathbf{1}\{A\} + \mathbf{1}\{B\}$ for two events A and B. Note that for $(a,b) \in (E^{(1)})^c$, we have $\tilde{\theta}_b^a + \tilde{\delta}_b^a = 0$. Hence, combining (102), we have

$$\begin{aligned} |\hat{E}_{\text{init}}^{(1)} \cap (E^{(1)})^c| &\leq \sum_{(a,b)\in(E^{(1)})^c} \mathbf{1}\big\{|\hat{\theta}_b^a(\hat{\tau}) - \tilde{\theta}_b^a + \hat{\delta}_b^a(\hat{\tau}) - \tilde{\delta}_b^a| \geq r_0\lambda_T\big\} \\ &\quad + \sum_{(a,b)\in(E^{(1)})^c} \mathbf{1}\big\{|\hat{\theta}_a^b(\hat{\tau}) - \tilde{\theta}_a^b + \hat{\delta}_a^b(\hat{\tau}) - \tilde{\delta}_a^b| \geq r_0\lambda_T\big\}, \\ &\leq \frac{2}{r_0\lambda_T} \sum_{a=1}^{p} \sum_{b\neq a} |\hat{\theta}_b^a(\hat{\tau}) - \tilde{\theta}_b^a| + |\hat{\delta}_b^a(\hat{\tau}) - \tilde{\delta}_b^a|, \\ &\leq 2\frac{M_3}{r_0} s_1, \end{aligned} \tag{103}$$

where the last inequality of (103) comes from $\sum_{a=1}^{p} \big|\hat{\boldsymbol{\beta}}^a(\hat{\tau}) - \tilde{\boldsymbol{\beta}}^a\big|_1 \leq M_3\lambda_T s_1$ obtained in Proposition 9. By (103), we have

$$|\hat{E}_{\text{init}}^{(1)}| \leq |\hat{E}_{\text{init}}^{(1)} \cap (E^{(1)})^c| + |E^{(1)}| \leq (1 + 2\frac{M_3}{r_0})s_1. \tag{104}$$

Define the following p^2-dimensional vectors:

$$\tilde{\boldsymbol{\alpha}}^{pp} = \big((\tilde{\boldsymbol{\alpha}}^1)^\top, (\tilde{\boldsymbol{\alpha}}^2)^\top, \ldots, (\tilde{\boldsymbol{\alpha}}^p)^\top\big)^\top,\ \hat{\boldsymbol{\alpha}}^{pp} = \big((\hat{\boldsymbol{\alpha}}^1)^\top, (\hat{\boldsymbol{\alpha}}^2)^\top, \ldots, (\hat{\boldsymbol{\alpha}}^p)^\top\big)^\top,$$

with $\tilde{\boldsymbol{\alpha}}^a$ and $\hat{\boldsymbol{\alpha}}^a$ being defined as

$$\tilde{\boldsymbol{\alpha}}^a = \tilde{\boldsymbol{\theta}}^a + \tilde{\boldsymbol{\delta}}^a = \big(\tilde{\theta}_1^a + \tilde{\delta}_1^a, \ldots, \tilde{\theta}_p^a + \tilde{\delta}_p^a\big)^\top,$$

$$\hat{\boldsymbol{\alpha}}^a = \hat{\boldsymbol{\theta}}^a(\hat{\tau}) + \hat{\boldsymbol{\delta}}^a(\hat{\tau}) = \big(\hat{\theta}_1^a(\hat{\tau}) + \hat{\delta}_1^a(\hat{\tau}), \ldots, \hat{\theta}_p^a(\hat{\tau}) + \hat{\delta}_p^a(\hat{\tau})\big)^\top, \quad \text{for } 1 \leq a \leq p.$$

Moreover, let $\boldsymbol{v}_{\text{init}} := \hat{\boldsymbol{\alpha}}^{pp} - \tilde{\boldsymbol{\alpha}}^{pp}$ and $\tilde{\boldsymbol{\alpha}}_{\min}^{pp} := \min_{1\leq a\leq p} \min_{1\leq b\leq p, b\neq a} |\tilde{\theta}_b^a + \tilde{\delta}_b^a|$. Recall $t_{\text{thr}}^{(1)} = r_1^* \lambda_T |\hat{E}_{\text{init}}^{(1)}|$. By (104), we have $t_{\text{thr}}^{(1)} \leq r_1^*(1 + 2\frac{M_3}{r_0})\lambda_T s_1$. For $(a,b) \in E^{(1)}$, we then have

$$\begin{aligned} \max\big(|\hat{\theta}_b^a(\hat{\tau}) + \hat{\delta}_b^a(\hat{\tau})|, |\hat{\theta}_a^b(\hat{\tau}) + \hat{\delta}_a^b(\hat{\tau})|\big) &\geq \tilde{\boldsymbol{\alpha}}_{\min}^{pp} - |\boldsymbol{v}_{\text{init}}|_1, \\ &\geq \tilde{\boldsymbol{\alpha}}_{\min}^{pp} - \sum_{a=1}^{p} |\hat{\boldsymbol{\beta}}^a(\hat{\tau}) - \tilde{\boldsymbol{\beta}}^a|_1, \\ &\geq \tilde{\boldsymbol{\alpha}}_{\min}^{pp} - M_3\lambda_T s_1 \ \text{(Proposition 7)}, \\ &\geq t_{\text{thr}}^{(1)} \ \text{(Assumption 6)}, \end{aligned} \tag{105}$$

which implies $(a,b) \in \breve{E}^{(1)}$. Hence, by (105) and the construction of $\hat{E}_{\text{init}}^{(1)}$ as in (2.9), we have $E^{(1)} \subset \breve{E}^{(1)} \subset \hat{E}_{\text{init}}^{(1)}$.

Step 2 : In this step, we prove $|\breve{E}^{(1)} \cap (E^{(1)})^c| \leq 2M_3/r_1^*$. By definitions, we have

$$\begin{aligned} &|\breve{E}^{(1)} \cap (E^{(1)})^c| \\ &= \sum_{(a,b)\in(E^{(1)})^c} \mathbf{1}\big\{|\hat{\theta}_b^a(\hat{\tau}) + \hat{\delta}_b^a(\hat{\tau})| \geq t_{\text{thr}}^{(1)} \cup |\hat{\theta}_a^b(\hat{\tau}) + \hat{\delta}_a^b(\hat{\tau})| \geq t_{\text{thr}}^{(1)}\big\}, \\ &\leq \sum_{(a,b)\in(E^{(1)})^c} \mathbf{1}\big\{|\hat{\theta}_b^a(\hat{\tau}) + \hat{\delta}_b^a(\hat{\tau})| \geq t_{\text{thr}}^{(1)}\big\} + \sum_{(a,b)\in(E^{(1)})^c} \mathbf{1}\big\{|\hat{\theta}_a^b(\hat{\tau}) + \hat{\delta}_a^b(\hat{\tau})| \geq t_{\text{thr}}^{(1)}\big\}. \end{aligned} \tag{106}$$

Note that for $(a,b) \in (E^{(1)})^c$, we have $\tilde{\theta}_b^a + \tilde{\delta}_b^a = 0$. By (106), we have

$$\begin{aligned} |\check{E}^{(1)} \cap (E^{(1)})^c| &\leq \sum_{(a,b)\in(E^{(1)})^c} \mathbf{1}\big\{|\hat{\theta}_b^a(\hat{\tau}) - \tilde{\theta}_b^a + \hat{\delta}_b^a(\hat{\tau}) - \tilde{\delta}_b^a| \geq t_{\text{thr}}^{(1)}\big\} \\ &\quad + \sum_{(a,b)\in(E^{(1)})^c} \mathbf{1}\big\{|\hat{\theta}_a^b(\hat{\tau}) - \tilde{\theta}_a^b + \hat{\delta}_a^b(\hat{\tau}) - \tilde{\delta}_a^b| \geq t_{\text{thr}}^{(1)}\big\}, \\ &\leq \frac{2}{t_{\text{thr}}^{(1)}} \sum_{a=1}^{p} \sum_{b\neq a} |\hat{\theta}_b^a(\hat{\tau}) - \tilde{\theta}_b^a| + |\hat{\delta}_b^a(\hat{\tau}) - \tilde{\delta}_b^a|, \\ &\leq 2M_3\lambda_T s_1 / t_{\text{thr}}^{(1)} \leq 2M_3/r_1^*, \end{aligned} \tag{107}$$

where the last inequality of (107) comes from $t_{\text{thr}}^{(1)} = r_1^*\lambda_T|\hat{E}_{\text{init}}^{(1)}| \geq r_1^*\lambda_T s_1$.

Finally, combining Steps 1 and 2, we complete the proof of (28). With a similar proof procedure, we can also prove (29), which finishes the proof of Theorem 11. ■

Appendix D. Proof of useful lemmas

D.1 Proof of Lemma 12

Proof We first prove (34). By the definition of $\|\mathbf{X}_a(\tau)\|_T$ as in (31), for any $\tau \in \mathcal{T}$, we have $\|\mathbf{X}_a(\tau)\|_T \geq \|\mathbf{X}_{a+p}(\tau)\|_T$ with $1 \leq a \leq p$. Hence, to prove (34), it is sufficient to consider $\max_{1\leq a\leq p} \|\mathbf{X}_a\|_T$. Note that $T\|\mathbf{X}_a\|_T^2 \sim \chi^2(T)$ for $1 \leq a \leq p$. For any $z > 1$, using the tail probability for $\chi^2(T)$, we have

$$\mathbb{P}\big(\max_{1\leq a\leq p} \|\mathbf{X}_a\|_T^2 > z\big) \leq p \max_{1\leq a\leq p} \mathbb{P}\big(\sum_{t=1}^{T} (X_a^t)^2 > Tz\big) \leq p\big(z\exp(1-z)\big)^{T/2}.$$

Therefore, choosing $z = K_1^2$ with some $K_1 > 1$, we finish the proof of (34).

Next, we prove (35). Note that $\|\mathbf{X}_a(t_0)\|_T \geq \|\mathbf{X}_{a+p}(t_0)\|_T$ for $1 \leq a \leq p$. To prove (35), it is sufficient to consider $\min_{p+1\leq a\leq 2p} \|\mathbf{X}_a(t_0)\|_T$. Recall $Q_t = t/T$. We then have $T\|\mathbf{X}_a(t_0)\|_T^2 \sim \chi^2(\lfloor Tt_0 \rfloor)$ for $p+1 \leq a \leq 2p$. Using the tail probability for $\chi^2(\lfloor Tt_0 \rfloor)$ distribution again, for any $0 < z < t_0$, we have

$$\mathbb{P}\big(\min_{p+1\leq a\leq 2p} \|\mathbf{X}_a(t_0)\|_T^2 < z\big) \leq p\mathbb{P}\big(\chi^2(\lfloor Tt_0 \rfloor) < zT\big) \leq p\big(z'\exp(1-z')\big)^{\lfloor Tt_0 \rfloor/2}$$

where $z' := z/t_0$. Choosing $z = K_2^2$ with some $0 < K_2 < t_0$, we finish the proof of (35). ■

D.2 Proof of Lemma 13

Proof We aim to prove (36). By the triangle inequality, we have

$$\sup_{1\leq a\leq p} \sup_{|\tau-\tau_*|\leq\eta} \frac{1}{T} \sum_{t=1}^{T} |X_a^t|^2 \big|\mathbf{1}\{Q_t \leq \tau_*\} - \mathbf{1}\{Q_t \leq \tau\}\big| \leq D_1 + D_2, \tag{108}$$

where

$$D_1 := \sup_{1\le a\le p}\sup_{|\tau-\tau_*|\le\eta}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\Big(|X_a^t|^2\big|\mathbf{1}\{Q_t\le\tau_*\}-\mathbf{1}\{Q_t\le\tau\}\big|\Big),$$
$$D_2 := \sup_{1\le a\le p}\sup_{|\tau-\tau_*|\le\eta}\Big|\frac{1}{T}\sum_{t=1}^{T}\Big(\big(|X_a^t|^2-\mathbb{E}(X_a^t)^2\big)\big|\mathbf{1}\{Q_t\le\tau_*\}-\mathbf{1}\{Q_t\le\tau\}\big|\Big)\Big|.$$

Therefore, by (108), to prove Lemma 13, it is sufficient to bound D_1 and D_2, respectively.

We first consider D_1. Note that under Assumption 1 (a), X_a^t follows $N(0,1)$ for $1\le t\le T$ and $1\le a\le p$. We have $\mathbb{E}(X_a^t)^2=1$ for $1\le a\le p$ and $1\le t\le T$. Recall $Q_t=t/T$. Note that

$$|\mathbf{1}\big\{Q_t\le\tau_*\big\}-\mathbf{1}\{Q_t\le\tau\}|=\mathbf{1}\big\{\{\tau\le Q_t\le\tau_*\}\cup\{\tau_*\le Q_t\le\tau\}\big\},\tag{109}$$

Then, by (109), for sufficiently large T, there exists a positive constant C_1 such that

$$D_1\le\sup_{|\tau-\tau_*|\le\eta}\frac{|\lfloor T\tau\rfloor-\lfloor T\tau_*\rfloor|}{T}\le C_1\eta.\tag{110}$$

Next, we bound D_2. Note that $(X_a^t)^2$ follows the sub-exponential distribution. For any $z>0$, by (109), we have

$$\mathbb{P}(D_2>z)\le D_{21}+D_{22},$$

where

$$D_{21} := pT\sup_{1\le a\le p}\sup_{\tau_*\le\tau\le\eta+\tau_*}\mathbb{P}\Big(\Big|\frac{1}{T}\sum_{t=\lfloor T\tau_*\rfloor}^{\lfloor T\tau\rfloor}(X_a^t)^2-\mathbb{E}(X_a^t)^2\Big|>z/2\Big),$$
$$D_{22} := pT\sup_{1\le a\le p}\sup_{\tau_*-\eta\le\tau\le\tau_*}\mathbb{P}\Big(\Big|\frac{1}{T}\sum_{t=\lfloor T\tau\rfloor}^{\lfloor T\tau_*\rfloor}(X_a^t)^2-\mathbb{E}(X_a^t)^2\Big|>z/2\Big).$$

For D_{21}, using Bernstein's inequality for sub-exponential distributions and considering $\tau_*\le\tau\le\eta+\tau_*$ there exists C_1 and C_2 such that

$$D_{21}\le C_1pT\exp\Big(-C_2\frac{z^2T^2}{\lfloor T(\tau_*+\eta)\rfloor-\lfloor T\tau_*\rfloor}\Big)\tag{111}$$

holds. Similarly, for D_{22}, using Bernstein's inequality again and considering $\tau_*-\eta\le\tau\le\tau_*$, we have

$$D_{22}\le C_1pT\exp\Big(-C_2\frac{z^2T^2}{\lfloor T\tau_*\rfloor-\lfloor T(\tau_*-\eta)\rfloor}\Big).\tag{112}$$

Therefore, by (111) and (112), choosing $z=C\sqrt{\log(pT)/T}$, we have $D_2\le C\sqrt{\log(pT)/T}$ with probability at least $1-(pT)^{-C_2}$ for sufficiently large T and p.

Finally, combining (108), (110), and the result that $D_2\le C\sqrt{\log(pT)/T}$, we finish the proof of Lemma 13. ■

D.3 Proof of Lemma 14

Proof By the definitions of $\hat{\tau}$ and $(\hat{\boldsymbol{\beta}}^a)_{a=1}^p$ as obtained from (5) – (7), we have

$$\sum_{a=1}^{p}\Big(\big\|\mathbf{X}_a-\mathbf{X}(\hat{\tau})\hat{\boldsymbol{\beta}}^a\big\|_T^2+\lambda_T|\hat{\mathbf{D}}\hat{\boldsymbol{\beta}}^a|_1\Big)\le\sum_{a=1}^{p}\Big(\big\|\mathbf{X}_a-\mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2+\lambda_T|\mathbf{D}\tilde{\boldsymbol{\beta}}^a|_1\Big). \tag{113}$$

Recall X_a^t as defined in (4). For $\big\|\mathbf{X}_a-\mathbf{X}(\hat{\tau})\hat{\boldsymbol{\beta}}^a\big\|_T^2$ and $\big\|\mathbf{X}_a-\mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2$ in (113), we have

$$\begin{aligned}
&\big\|\mathbf{X}_a-\mathbf{X}(\hat{\tau})\hat{\boldsymbol{\beta}}^a\big\|_T^2-\big\|\mathbf{X}_a-\mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2\\
&=T^{-1}\sum_{t=1}^{T}\big(X_a^t-\mathbf{X}^t(\hat{\tau})^\top\hat{\boldsymbol{\beta}}^a\big)^2-T^{-1}\sum_{t=1}^{T}\big(X_a^t-\mathbf{X}^t(\tau_*)^\top\tilde{\boldsymbol{\beta}}^a\big)^2,\\
&=T^{-1}\sum_{t=1}^{T}\Big(\epsilon_a^t-\big(\mathbf{X}^t(\hat{\tau})^\top\hat{\boldsymbol{\beta}}^a-\mathbf{X}^t(\tau_*)^\top\tilde{\boldsymbol{\beta}}^a\big)\Big)^2-T^{-1}\sum_{t=1}^{T}\Big(\epsilon_a^t-\big(\mathbf{X}^t(\tau_*)^\top\tilde{\boldsymbol{\beta}}^a-\mathbf{X}^t(\tau_*)^\top\tilde{\boldsymbol{\beta}}^a\big)\Big)^2,\\
&=\big\|\mathbf{X}(\hat{\tau})\hat{\boldsymbol{\beta}}^a-\mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2-2T^{-1}\sum_{t=1}^{T}\epsilon_a^t\big(\mathbf{X}^t(\hat{\tau})^\top\hat{\boldsymbol{\beta}}^a-\mathbf{X}^t(\tau_*)^\top\tilde{\boldsymbol{\beta}}^a\big).
\end{aligned} \tag{114}$$

Recall $\hat{\boldsymbol{\beta}}^a:=((\hat{\boldsymbol{\theta}}^a)^\top,(\hat{\boldsymbol{\delta}}^a)^\top)^\top$ and $\tilde{\boldsymbol{\beta}}^a:=((\tilde{\boldsymbol{\theta}}^a)^\top,(\tilde{\boldsymbol{\delta}}^a)^\top)^\top$. Then, for $2T^{-1}\sum_{t=1}^{T}\epsilon_a^t\big(\mathbf{X}^t(\hat{\tau})^\top\hat{\boldsymbol{\beta}}^a-\mathbf{X}^t(\tau_*)^\top\tilde{\boldsymbol{\beta}}^a\big)$ as in (114), we have

$$\begin{aligned}
\frac{2}{T}\sum_{t=1}^{T}\epsilon_a^t\big(\mathbf{X}^t(\hat{\tau})^\top\hat{\boldsymbol{\beta}}^a-\mathbf{X}^t(\tau_*)^\top\tilde{\boldsymbol{\beta}}^a\big)&=\frac{2}{T}\sum_{t=1}^{T}\epsilon_a^t(\mathbf{X}^t)^\top(\hat{\boldsymbol{\theta}}^a-\tilde{\boldsymbol{\theta}}^a)\\
&+\frac{2}{T}\sum_{t=1}^{T}\epsilon_a^t(\mathbf{X}^t)^\top\mathbf{1}\{Q_t\le\hat{\tau}\}(\hat{\boldsymbol{\delta}}^a-\tilde{\boldsymbol{\delta}}^a)+\frac{2}{T}\sum_{t=1}^{T}\epsilon_a^t(\mathbf{X}^t)^\top\tilde{\boldsymbol{\delta}}^a\big(\mathbf{1}\{Q_t\le\hat{\tau}\}-\mathbf{1}\{Q_t\le\tau_*\}\big).
\end{aligned}$$

Recall $\mathcal{A}^a$ and $\mathcal{B}^a$ as defined in (38). For each node a, under the events $\mathcal{A}^a$ and $\mathcal{B}^a$, we have

$$\Big|\frac{2}{T}\sum_{t=1}^{T}\epsilon_a^t(\mathbf{X}^t)^\top(\hat{\boldsymbol{\theta}}^a-\tilde{\boldsymbol{\theta}}^a)+\frac{2}{T}\sum_{t=1}^{T}\epsilon_a^t(\mathbf{X}^t)^\top\mathbf{1}\{Q_t\le\hat{\tau}\}(\hat{\boldsymbol{\delta}}^a-\tilde{\boldsymbol{\delta}}^a)\Big|\le\lambda_T\mu\big|\hat{\mathbf{D}}(\hat{\boldsymbol{\beta}}^a-\tilde{\boldsymbol{\beta}}^a)\big|_1. \tag{115}$$

Combining (113), (114), and (115), under the event $\bigcap_{1\le a\le p}\{\mathcal{A}^a\cap\mathcal{B}^a\}$, we have

$$\begin{aligned}
\sum_{a=1}^{p}\big\|\mathbf{X}(\hat{\tau})\hat{\boldsymbol{\beta}}^a-\mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2\ &\le\lambda_T\mu\sum_{a=1}^{p}\big|\hat{\mathbf{D}}(\hat{\boldsymbol{\beta}}^a-\tilde{\boldsymbol{\beta}}^a)\big|_1+\lambda_T\sum_{a=1}^{p}\big(|\mathbf{D}\tilde{\boldsymbol{\beta}}^a|_1-|\hat{\mathbf{D}}\hat{\boldsymbol{\beta}}^a|_1\big),\\
&+\frac{2}{T}\sum_{a=1}^{p}\sum_{t=1}^{T}\epsilon_a^t(\mathbf{X}^t)^\top\tilde{\boldsymbol{\delta}}^a\big(\mathbf{1}\{Q_t\le\hat{\tau}\}-\mathbf{1}\{Q_t\le\tau_*\}\big).
\end{aligned} \tag{116}$$

By adding $\lambda_T\sum_{a=1}^{p}\big|\hat{\mathbf{D}}(\hat{\boldsymbol{\beta}}^a-\tilde{\boldsymbol{\beta}}^a)\big|_1$ on the two sides of (116), we have

$$\begin{aligned}
&\sum_{a=1}^{p}\big\|\mathbf{X}(\hat{\tau})\hat{\boldsymbol{\beta}}^a-\mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2+\lambda_T(1-\mu)\sum_{a=1}^{p}\big|\hat{\mathbf{D}}(\hat{\boldsymbol{\beta}}^a-\tilde{\boldsymbol{\beta}}^a)\big|_1\\
&\le\lambda_T\sum_{a=1}^{p}\big|\hat{\mathbf{D}}(\hat{\boldsymbol{\beta}}^a-\tilde{\boldsymbol{\beta}}^a)\big|_1+\lambda_T\sum_{a=1}^{p}\big(|\mathbf{D}\tilde{\boldsymbol{\beta}}^a|_1-|\hat{\mathbf{D}}\hat{\boldsymbol{\beta}}^a|_1\big)+R.
\end{aligned} \tag{117}$$

where R is defined in (39). Let $J_0^a = J(\tilde{\boldsymbol{\beta}}^a)$ be the set of non-zero elements of $\tilde{\boldsymbol{\beta}}^a$. Note that

$$|\hat{\beta}_j^a - \tilde{\beta}_j^a| + |\tilde{\beta}_j^a| - |\hat{\beta}_j^a| = 0, \quad \text{if } j \in (J_0^a)^c. \tag{118}$$

Hence, by (117) and (118), and using the triangle inequality, we have

$$\begin{aligned}
&\sum_{a=1}^{p} \big\|\mathbf{X}(\hat{\tau})\hat{\boldsymbol{\beta}}^a - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2 + \lambda_T(1-\mu)\sum_{a=1}^{p}\big|\hat{\mathbf{D}}(\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a)\big|_1 \\
&\le 2\lambda_T \sum_{a=1}^{p}\big|\hat{\mathbf{D}}(\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a)_{J_0^a}\big|_1 + \lambda_T \sum_{a=1}^{p}\big||\hat{\mathbf{D}}\tilde{\boldsymbol{\beta}}^a|_1 - |\mathbf{D}\tilde{\boldsymbol{\beta}}^a|_1\big| + R,
\end{aligned}$$

which completes the proof of (40) in Lemma 14.

After proving (40), we next consider (41). The proof technique is similar. To save space, we omit the details. ■

D.4 Proof of Lemma 15

Proof We first consider $\mathcal{A}^a$. Let $\Phi(x)$ be the CDF of the standard normal distribution. Note that $\epsilon_a^t \sim N(0, (\sigma_a^t)^2)$ for $1 \le t \le T$. By the definition of V_{1b}^a as in (37), conditional on $\mathcal{X}$, we have $\sqrt{T}V_{1b}^a \sim N(0,1)$ for $1 \le b \le p$ with $b \ne a$. Therefore, for each node a,

$$\mathbb{P}((\mathcal{A}^a)^c) \le \sum_{b=1, b\ne a}^{p} \mathbb{P}\big(2\sqrt{T}|V_{1b}^a| > \frac{\sqrt{T}\mu\lambda_T}{\omega}\big) = 2(p-1)\Phi\big(-\frac{\sqrt{T}\mu\lambda_T}{2\omega}\big) \le 2p\Phi\big(-\frac{\sqrt{Tr_T}\mu\lambda_T}{2\omega}\big), \tag{119}$$

where the last inequality comes from $0 < r_T \le 1$.

We next consider $\mathcal{B}^a$. Note that $Q_t := t/T$ and $\|\boldsymbol{X}_a(\tau)\|_T \ge \|\boldsymbol{X}_a(t_0)\|_T$ for $\tau \in [t_0, t_1]$. Hence, we can write $V_{2b}^a(\tau)$ in (37) as a partial sum process and have

$$V_{2b}^a(\tau) := \frac{1}{T}\sum_{t=1}^{\lfloor T\tau\rfloor} \frac{\epsilon_a^t}{\sigma_a^t}\frac{X_b^t}{\|\boldsymbol{X}_b(\tau)\|_T} \le \frac{1}{T}\sum_{t=1}^{\lfloor T\tau\rfloor} \frac{\epsilon_a^t}{\sigma_a^t}\frac{X_b^t}{\|\boldsymbol{X}_b(t_0)\|_T}. \tag{120}$$

By (120), conditional on $\mathcal{X}$, we have

$$\begin{aligned}
&\mathbb{P}\big(\sup_{\tau\in\mathcal{T}}|\sqrt{T}V_{2b}^a(\tau)| > \sqrt{T}\mu\lambda_T/2\omega\big) \\
&\le \mathbb{P}\big(\sqrt{T}\sup_{\tau\in\mathcal{T}}\big|\frac{1}{T}\sum_{t=1}^{\lfloor T\tau\rfloor}\frac{\epsilon_a^t}{\sigma_a^t}\frac{X_b^t}{\|\boldsymbol{X}_b(t_0)\|_T}\big| > \sqrt{T}\mu\lambda_T/2\omega\big), \\
&\le \mathbb{P}\big(\sqrt{T}\sup_{1\le s\le T}\big|\frac{1}{T}\sum_{t=1}^{s}\frac{\epsilon_a^t}{\sigma_a^t}\frac{X_b^t}{\|\boldsymbol{X}_b\|_T}\big| > \frac{\sqrt{T}\mu\lambda_T\|\boldsymbol{X}_b(t_0)\|_T}{2\omega\|\boldsymbol{X}_b\|_T}\big), \\
&\le 2\mathbb{P}\big(\sqrt{T}|V_{1b}^a| > \frac{\sqrt{T}\mu\lambda_T\|\boldsymbol{X}_b(t_0)\|_T}{2\omega\|\boldsymbol{X}_b\|_T}\big),
\end{aligned} \tag{121}$$

where the last inequality in (121) comes from the fact that V_{1b}^a follows a Gaussian distribution and by Levy's inequality (see Proposition A.1.2 in van der Vaart and Wellner (1996)).

Therefore, by (121), we have

$$\begin{aligned}\mathbb{P}((\mathcal{B}^a)^c) &\le \sum_{b=1,b\neq a}^{p}\mathbb{P}\big(\sup_{\tau\in\mathcal{T}}|\sqrt{T}V_{2b}^a(\tau)|>\sqrt{T}\mu\lambda_T/2\omega\big),\\ &\le \sum_{b=1,b\neq a}^{p}2\mathbb{P}\big(\sqrt{T}|V_{1b}^a|>\frac{\sqrt{T}\mu\lambda_T\|\boldsymbol{X}_b(t_0)\|_T}{2\omega\|\boldsymbol{X}_b\|_T}\big),\\ &\le 4p\Phi\big(-\frac{\sqrt{Tr_T}\mu\lambda_T}{2\omega}\big),\end{aligned}\tag{122}$$

the last inequality in (122) comes from $r_T := \min_{1\le a\le p}\frac{\|\boldsymbol{X}_a(t_0)\|_T^2}{\|\boldsymbol{X}_a\|_T^2}$ and Assumption 1 (b).

Note that under the event $\mathcal{E}^{(1)}$ as defined in (42), we have $r_T \ge K_2^2/K_1^2$. Finally, combining (119) and (122), using $\mathbb{P}(\bigcap_{1\le a\le p}\mathcal{A}^a\cap\mathcal{B}^a)\ge 1-\sum_{a=1}^p\big(\mathbb{P}\big((\mathcal{A}^a)^c\big)+\mathbb{P}\big((\mathcal{B}^a)^c\big)\big)$, we complete the proof of Lemma 15. ■

D.5 Proof of Lemma 16

Proof Recall $Q_t := t/T$. By the definition of $C(\eta_j)$ as defined in (43), we have

$$\mathbb{P}\big((\mathcal{C}(\eta_j))^c\big)\le D_1(\eta_j)+D_2(\eta_j),\tag{123}$$

where

$$\begin{aligned}D_1(\eta_j) &:= \mathbb{P}\Big(\sup_{-\eta_j\le\tau-\tau_*\le 0}\Big|\frac{2}{T}\sum_{a=1}^{p}\sum_{t=\lfloor T\tau\rfloor}^{\lfloor T\tau_*\rfloor}\epsilon_a^t(\mathbf{X}^t)^\top\tilde{\boldsymbol{\delta}}^a\Big|>\lambda_T\sqrt{\eta_j}\Big),\\ D_2(\eta_j) &:= \mathbb{P}\Big(\sup_{0\le\tau-\tau_*\le\eta_j}\Big|\frac{2}{T}\sum_{a=1}^{p}\sum_{t=\lfloor T\tau_*\rfloor}^{\lfloor T\tau\rfloor}\epsilon_a^t(\mathbf{X}^t)^\top\tilde{\boldsymbol{\delta}}^a\Big|>\lambda_T\sqrt{\eta_j}\Big).\end{aligned}$$

Hence, by (123), we need to bound $D_1(\eta_j)$ and $D_2(\eta_j)$, respectively.

We first consider $D_1(\eta_j)$. Note that conditional on $\mathcal{X}$, $\sum_{a=1}^{p}\epsilon_a^t(\mathbf{X}^t)^\top\tilde{\boldsymbol{\delta}}^a$ follows a Gaussian distribution. Then, using Levy's inequality (van der Vaart and Wellner (1996)), we have

$$D_1(\eta_j)\le 2\mathbb{P}\Big(\Big|\frac{2}{T}\sum_{t=\lfloor T(\tau_*-\eta_j)\rfloor}^{\lfloor T\tau_*\rfloor}\sum_{a=1}^{p}\epsilon_a^t(\mathbf{X}^t)^\top\tilde{\boldsymbol{\delta}}^a\Big|>\lambda_T\sqrt{\eta_j}\Big).\tag{124}$$

Next, we examine the distribution of $\frac{2}{T}\sum_{t=\lfloor T(\tau_*-\eta_j)\rfloor}^{\lfloor T\tau_*\rfloor}\sum_{a=1}^{p}\epsilon_a^t(\mathbf{X}^t)^\top\tilde{\boldsymbol{\delta}}^a$. To this end, define

$$\boldsymbol{e}^t=(e_1^t,\ldots,e_p^t)^\top,\boldsymbol{\Lambda}^t=\big((\mathbf{X}^t)^\top\tilde{\boldsymbol{\delta}}^1,\ldots,(\mathbf{X}^t)^\top\tilde{\boldsymbol{\delta}}^p\big)^\top,\ \text{for } 1\le t\le T.\tag{125}$$

Note that by Assumption 1 (b) and Anderson (2003), we have $\mathrm{Cov}(\boldsymbol{e}^t)=\omega^4\boldsymbol{\Omega}^{(1)}$ for $1\le t\le\lfloor T\tau_*\rfloor$ and $\mathrm{Cov}(\boldsymbol{e}^t)=\omega^4\boldsymbol{\Omega}^{(2)}$ for $\lfloor T\tau_*\rfloor+1\le t\le T$. Furthermore, consider that $(\boldsymbol{e}^t)_{t=1}^T$

are independent random vectors. Then conditional on $\mathcal{X}$, we have

$$\begin{aligned}&\frac{2}{T}\sum_{t=\lfloor T(\tau_*-\eta_j)\rfloor}^{\lfloor T\tau_*\rfloor}\sum_{a=1}^{p}\epsilon_a^t(\mathbf{X}^t)^\top\tilde{\boldsymbol{\delta}}^a\\&\quad\sim N\Big(0,\underbrace{\frac{4\eta_j}{T}\frac{\omega^4}{\lfloor T\tau_*\rfloor-\lfloor T(\tau_*-\eta_j)\rfloor}\sum_{t=\lfloor T(\tau_*-\eta_j)\rfloor}^{\lfloor T\tau_*\rfloor}(\boldsymbol{\Lambda}^t)^\top\boldsymbol{\Omega}^{(1)}\boldsymbol{\Lambda}^t}_{q^2}\Big).\end{aligned}\tag{126}$$

By definitions of $\boldsymbol{\Lambda}^t$ and q^2 as in (125) and (126), conditional on $\mathcal{X}$, we have

$$\begin{aligned}q^2&\le\frac{\omega^4}{\big(\lfloor T\tau_*\rfloor-\lfloor T(\tau_*-\eta_j)\rfloor\big)\underline{\kappa}}\sum_{t=\lfloor T(\tau_*-\eta_j)\rfloor}^{\lfloor T\tau_*\rfloor}(\boldsymbol{\Lambda}^t)^\top\boldsymbol{\Lambda}^t,\\&=\frac{\omega^4}{\big(\lfloor T\tau_*\rfloor-\lfloor T(\tau_*-\eta_j)\rfloor\big)\underline{\kappa}}\sum_{t=\lfloor T(\tau_*-\eta_j)\rfloor}^{\lfloor T\tau_*\rfloor}\sum_{a=1}^{p}(\tilde{\boldsymbol{\delta}}^a)^\top\mathbf{X}^t(\mathbf{X}^t)^\top\tilde{\boldsymbol{\delta}}^a,\end{aligned}\tag{127}$$

where the first inequality in (127) comes from Assumption 1 (c) with $\underline{\kappa}:=\min(\phi_{\min}^{(1)},\phi_{\min}^{(2)})$. Furthermore, by plugging $\boldsymbol{\Sigma}^{(1)}$ into (127), and using the trianlge inequality, we have

$$\begin{aligned}q^2\le{}&\omega^4\underline{\kappa}^{-1}\sum_{a=1}^{p}(\tilde{\boldsymbol{\delta}}^a)^\top\boldsymbol{\Sigma}^{(1)}\tilde{\boldsymbol{\delta}}^a\\&+\omega^4\underline{\kappa}^{-1}\Big|\sum_{a=1}^{p}(\tilde{\boldsymbol{\delta}}^a)^\top\Big\{\frac{1}{\big(\lfloor T\tau_*\rfloor-\lfloor T(\tau_*-\eta_j)\rfloor\big)}\sum_{t=\lfloor T(\tau_*-\eta_j)\rfloor}^{\lfloor T\tau_*\rfloor}\big(\boldsymbol{\Sigma}^{(1)}-\mathbf{X}^t(\mathbf{X}^t)^\top\big)\Big\}\tilde{\boldsymbol{\delta}}^a\Big|.\end{aligned}\tag{128}$$

Note that, under the event $\mathcal{E}^{(2)}$ as defined in (44), we have

$$\Big\|\frac{1}{\big(\lfloor T\tau_*\rfloor-\lfloor T(\tau_*-\eta_j)\rfloor\big)}\sum_{t=\lfloor T(\tau_*-\eta_j)\rfloor}^{\lfloor T\tau_*\rfloor}\big(\boldsymbol{\Sigma}^{(1)}-\mathbf{X}^t(\mathbf{X}^t)^\top\big)\Big\|_\infty\le C_1\sqrt{\frac{\log(pT)}{T}}.\tag{129}$$

Recall $\delta^*:=\max\big(\sum_{a=1}^{p}(\tilde{\boldsymbol{\delta}}^a)^\top\boldsymbol{\Sigma}^{(1)}\tilde{\boldsymbol{\delta}}^a,\sum_{a=1}^{p}(\tilde{\boldsymbol{\delta}}^a)^\top\boldsymbol{\Sigma}^{(2)}\tilde{\boldsymbol{\delta}}^a\big)$. Hence, combining (128) and (129), there exists a positive constant $C_1^{(j)}$ only depending on η_j such that

$$q^2\le\omega^4\underline{\kappa}^{-1}\delta^*+\omega^4\Big(\underline{\kappa}^{-1}\sum_{a=1}^{p}|\tilde{\boldsymbol{\delta}}^a|_1^2\Big)C_1\sqrt{\frac{\log(pT)}{T}}\le C_1^{(j)}\omega^4\underline{\kappa}^{-1}\delta^*,\tag{130}$$

where the last inequality in (130) comes from the assumption that $\sum_{a=1}^{p}|\tilde{\boldsymbol{\delta}}^a|_1^2=o(\sqrt{T/\log(pT)})$. Combining (124), (126), and (130), conditional on the event $\mathcal{E}^{(2)}$, we have

$$D_1(\eta_j)\le4\Phi\left(\frac{-\lambda_T\sqrt{T}}{2\omega^2\sqrt{C_1^{(j)}\underline{\kappa}^{-1}\delta^*}}\right)\le4\Phi\left(\frac{-\lambda_T\mu K_2\sqrt{T}}{2K_1\omega^2\sqrt{C_1^{(j)}\underline{\kappa}^{-1}\delta^*}}\right),\tag{131}$$

where the last inequality in (131) comes from the fact that $\mu \in (0,1)$ and $0 < K_2 \leq K_1$.

After bounding $D_1(\eta_j)$ as in (131), we next consider $D_2(\eta_j)$. Using a similar proof procedure, under the event $\mathcal{E}^{(3)}$ as defined in (45), we can prove that

$$D_2(\eta_j) \leq 4\Phi\left(\frac{-\lambda_T\sqrt{T}}{2\omega^2\sqrt{C_2^{(j)}\underline{\kappa}^{-1}\delta^*}}\right) \leq 4\Phi\left(\frac{-\lambda_T\mu K_2\sqrt{T}}{2K_1\omega^2\sqrt{C_2^{(j)}\underline{\kappa}^{-1}\delta^*}}\right). \tag{132}$$

where $C_2^{(j)}$ is a constant only depending on η_j.

Finally, combining (123), (131), and (132), and using the fact that

$$\mathbb{P}\Big(\bigcap_{j=1}^{m}\mathcal{C}(\eta_j)\Big) \geq 1 - \sum_{j=1}^{m}\mathbb{P}\big((\mathcal{C}(\eta_j))^c\big),$$

we complete the proof of Lemma 16. ■

D.6 Proof of Lemma 17

D.6.1 Proof of (46) in Lemma 17

Proof We first prove (46) in Lemma 17. Note that, by assumptions, $|\hat{\tau} - \tau_*| \leq c_\tau$ and $\sum_{a=1}^{p}|\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a|_1 \leq c_{\boldsymbol{\beta}}$ hold. Moreover, by Lemma 14, we have

$$\begin{aligned}
&\sum_{a=1}^{p}\big\|\mathbf{X}(\hat{\tau})\hat{\boldsymbol{\beta}}^a - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2 + \lambda_T(1-\mu)\sum_{a=1}^{p}\big|\hat{\mathbf{D}}(\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a)\big|_1 \\
&\leq 2\lambda_T\sum_{a=1}^{p}\big|\hat{\mathbf{D}}(\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a)_{J_0^a}\big|_1 + \lambda_T\sum_{a=1}^{p}\big||\hat{\mathbf{D}}\tilde{\boldsymbol{\beta}}^a|_1 - |\mathbf{D}\tilde{\boldsymbol{\beta}}^a|_1\big| + R,
\end{aligned} \tag{133}$$

where $J_0^a := J(\tilde{\boldsymbol{\beta}}^a)$, and R is defined in (39). To derive the desired results, we need to bound $\lambda_T\sum_{a=1}^{p}\big||\hat{\mathbf{D}}\tilde{\boldsymbol{\beta}}^a|_1 - |\mathbf{D}\tilde{\boldsymbol{\beta}}^a|_1\big| + R$. For R, by the assumption $|\hat{\tau} - \tau_*| \leq c_\tau$ and under the event $\mathcal{C}(c_\tau)$, we have $|R| \leq \lambda_T\sqrt{c_\tau}$. For $\lambda_T\sum_{a=1}^{p}\big||\hat{\mathbf{D}}\tilde{\boldsymbol{\beta}}^a|_1 - |\mathbf{D}\tilde{\boldsymbol{\beta}}^a|_1\big|$, by the definitions of $\hat{\mathbf{D}}$ and $\mathbf{D}$ as in Section A, using Lemma 13, we have

$$\begin{aligned}
&\lambda_T\sum_{a=1}^{p}\big||\hat{\mathbf{D}}\tilde{\boldsymbol{\beta}}^a|_1 - |\mathbf{D}\tilde{\boldsymbol{\beta}}^a|_1\big|, \\
&\leq \lambda_T\sum_{a=1}^{p}\Big\{\sum_{j=1}^{p}\big|\big(\|\boldsymbol{X}_j(\hat{\tau})\|_T - \|\boldsymbol{X}_j(\tau_*)\|_T\big)|\tilde{\delta}_j^a|\big|\Big\}, \\
&\leq \lambda_T\sum_{a=1}^{p}\Big\{\sum_{j=1}^{p}\frac{1}{2\|\boldsymbol{X}_j(t_0)\|_T}|\tilde{\delta}_j^a|\frac{1}{T}\sum_{t=1}^{T}(X_j^t)^2\big|\mathbf{1}\{Q_t \leq \hat{\tau}\} - \mathbf{1}\{Q_t \leq \tau_*\}\big|\Big\}, \\
&\leq \lambda_T(2X_{\min})^{-1}C^*c_\tau\sum_{a=1}^{p}|\tilde{\boldsymbol{\delta}}^a|_1,
\end{aligned} \tag{134}$$

where C^* comes from Lemma 13. Therefore, by (133), $|R| \leq \lambda_T\sqrt{c_\tau}$, and (134), we consider two cases:

Case 1: $\sqrt{c_\tau} + (2X_{\min})^{-1}C^* c_\tau \sum_{a=1}^{p} |\tilde{\boldsymbol{\delta}}^a|_1 \le \sum_{a=1}^{p} \big|\hat{\mathbf{D}}(\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a)_{J_0^a}\big|_1$. In this case, considering (133), we have

$$\lambda_T(1-\mu)\sum_{a=1}^{p}\big|\hat{\mathbf{D}}(\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a)\big|_1 \le 3\lambda_T \sum_{a=1}^{p}\big|\hat{\mathbf{D}}(\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a)_{J_0^a}\big|_1.$$

Using $|\hat{\mathbf{D}}(\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a)\big|_1 = |\hat{\mathbf{D}}(\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a)_{J_0^a}\big|_1 + |\hat{\mathbf{D}}(\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a)_{(J_0^a)^c}\big|_1$, we then have

$$\sum_{a=1}^{p}|\hat{\mathbf{D}}(\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a)_{(J_0^a)^c}\big|_1 \le \frac{(2+\mu)}{1-\mu}\sum_{a=1}^{p}|\hat{\mathbf{D}}(\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a)_{J_0^a}\big|_1. \tag{135}$$

Similar to the proof in Section C.2, we define the following $2p^2$-dimensional vectors:

$$\tilde{\boldsymbol{\beta}}^{pp} = \big((\tilde{\boldsymbol{\beta}}^1)^\top, \ldots, (\tilde{\boldsymbol{\beta}}^p)^\top\big)^\top,\ \hat{\boldsymbol{\beta}}^{pp} = \big((\hat{\boldsymbol{\beta}}^1)^\top, \ldots, (\hat{\boldsymbol{\beta}}^p)^\top\big)^\top. \tag{136}$$

We also define the following $2p^2 \times 2p^2$ block diagonal matrix $\breve{\mathbf{D}}$ and the $Tp \times 2p^2$ design matrix $\tilde{\mathbf{X}}(\tau)$ as follows:

$$\breve{\mathbf{D}} = \mathrm{diag}\big\{\underbrace{\hat{\mathbf{D}}, \ldots, \hat{\mathbf{D}}}_{p}\big\},\ \tilde{\mathbf{X}}(\tau) = \mathrm{diag}\big\{\underbrace{\mathbf{X}(\tau), \ldots, \mathbf{X}(\tau)}_{p}\big\}, \tag{137}$$

where $\hat{\mathbf{D}} := \mathbf{D}(\hat{\tau})$. Let $J_0 = J(\tilde{\boldsymbol{\beta}}^{pp})$ be the set of non-zero elements of $\tilde{\boldsymbol{\beta}}^{pp}$. By Assumption 2, we have $|J_0| = \sum_{a=1}^{p}|J_0^a| \le s_1$. Furthermore, by (135), (136), and (137), we then have

$$\Big|\breve{\mathbf{D}}(\hat{\boldsymbol{\beta}}^{pp} - \tilde{\boldsymbol{\beta}}^{pp})_{(J_0)^c}\Big|_1 \le \frac{(2+\mu)}{1-\mu}\Big|\breve{\mathbf{D}}(\hat{\boldsymbol{\beta}}^{pp} - \tilde{\boldsymbol{\beta}}^{pp})_{J_0}\Big|_1.$$

Recall $\kappa'(s_1, c_0, \mathcal{S})$ in (65). Similar to Proposition 4, we can show that $\tilde{\mathbf{X}}(\tau)$ satisfies the URE condition uniformly over $\tau \in \mathcal{T}$ by setting $s_1 = o(\sqrt{T/\log(pT)})$. In other words, $\kappa'(s_1, c_0, \mathcal{S}) > 0$ holds. Set $\kappa' := \kappa'(s_1, c_0, \mathcal{S})$ with $c_0 := (2+\mu)/(1-\mu)$ and $\mathcal{S} := \{\tau : |\tau - \tau_*| \le c_\tau\}$. We then have

$$\begin{aligned}
&\kappa'^2\big|\breve{\mathbf{D}}(\hat{\boldsymbol{\beta}}^{pp} - \tilde{\boldsymbol{\beta}}^{pp})_{J_0}\big|_2^2 \\
&\le \frac{1}{T}\big|\tilde{\mathbf{X}}(\hat{\tau})\breve{\mathbf{D}}(\hat{\boldsymbol{\beta}}^{pp} - \tilde{\boldsymbol{\beta}}^{pp})\big|_2^2, \\
&\le \max(\hat{\mathbf{D}})^2\frac{1}{T}\big|\tilde{\mathbf{X}}(\hat{\tau})(\hat{\boldsymbol{\beta}}^{pp} - \tilde{\boldsymbol{\beta}}^{pp})\big|_2^2, \\
&\le \max(\hat{\mathbf{D}})^2\Big\{\sum_{a=1}^{p}\big\|\mathbf{X}(\hat{\tau})\hat{\boldsymbol{\beta}}^a - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2\Big\} \\
&\quad + \max(\hat{\mathbf{D}})^2\Big\{\sum_{a=1}^{p}\frac{1}{T}\sum_{t=1}^{T}\Big(2\Big(\mathbf{X}^t(\hat{\tau})^\top\hat{\boldsymbol{\beta}}^a - \mathbf{X}^t(\hat{\tau})^\top\tilde{\boldsymbol{\beta}}^a\Big)\Big((\mathbf{X}^t)^\top\tilde{\boldsymbol{\delta}}^a\big(\mathbf{1}\{Q_t \le \tau_*\} - \mathbf{1}\{Q_t \le \hat{\tau}\}\big)\Big)\Big)\Big\}, \\
&\le \max(\hat{\mathbf{D}})^2\Big\{\sum_{a=1}^{p}\Big(\big\|\mathbf{X}(\hat{\tau})\hat{\boldsymbol{\beta}}^a - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2 + 2c_{\boldsymbol{\beta}}|\tilde{\boldsymbol{\delta}}^a|_1\sup_{1\le a\le p}\frac{1}{T}\sum_{t=1}^{T}(X_a^t)^2\big|\mathbf{1}\{Q_t \le \tau_*\} - \mathbf{1}\{Q_t \le \hat{\tau}\}\big|\Big)\Big\}, \\
&\le X_{\max}^2\Big(\sum_{a=1}^{p}\big\|\mathbf{X}(\hat{\tau})\hat{\boldsymbol{\beta}}^a - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2 + 2c_{\boldsymbol{\beta}}c_\tau C^*\sum_{a=1}^{p}|\tilde{\boldsymbol{\delta}}^a|_1\Big),
\end{aligned} \tag{138}$$

where the last inequality in (138) comes from Lemma 13.

Combining the results in (133) and (138), we have

$$\begin{aligned}
&\sum_{a=1}^{p} \big\|\mathbf{X}(\hat{\tau})\hat{\boldsymbol{\beta}}^a - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2, \\
&\leq 3\lambda_T \sum_{a=1}^{p} \big|\hat{\mathbf{D}}(\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a)_{J_0^a}\big|_1, \\
&\leq 3\lambda_T\sqrt{s_1}\big|\hat{\mathbf{D}}(\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a)_{J_0^a}\big|_2, \\
&\leq 3\lambda_T\sqrt{s_1}\Big\{\kappa'^{-2}X_{\max}^2\Big(\sum_{a=1}^{p}\big\|\mathbf{X}(\hat{\tau})\hat{\boldsymbol{\beta}}^a - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2 + 2c_{\boldsymbol{\beta}}c_\tau C^* \sum_{a=1}^{p}|\tilde{\boldsymbol{\delta}}^a|_1\Big)\Big\}^{1/2},
\end{aligned} \tag{139}$$

where s_1 comes from Assumption 2. Note that $a+b \leq 2a \vee 2b$ for $a, b \geq 0$. Then by (139), conditional on $\bigcap_{1\leq a\leq p}\{\mathcal{A}^a \cap \mathcal{B}^a\}\bigcap \mathcal{C}(c_\tau)$, we have

$$\sum_{a=1}^{p}\big\|\mathbf{X}(\hat{\tau})\hat{\boldsymbol{\beta}}^a - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2 \leq \frac{18X_{\max}^2}{\kappa'^2}\lambda_T^2 s_1 \bigvee \frac{6X_{\max}}{\kappa'}\lambda_T\Big(s_1 c_{\boldsymbol{\beta}}c_\tau C^* \sum_{a=1}^{p}|\tilde{\boldsymbol{\delta}}^a|_1\Big)^{1/2}. \tag{140}$$

Case 2: $\sqrt{c_\tau} + (2X_{\min})^{-1}C^*c_\tau \sum_{a=1}^{p}|\tilde{\boldsymbol{\delta}}^a|_1 > \sum_{a=1}^{p}\big|\hat{\mathbf{D}}(\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a)_{J_0^a}\big|_1$. In this case, considering (133), we have

$$\sum_{a=1}^{p}\big\|\mathbf{X}(\hat{\tau})\hat{\boldsymbol{\beta}}^a - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2 \leq 3\lambda_T\Big(\sqrt{c_\tau} + (2X_{\min})^{-1}C^*c_\tau\sum_{a=1}^{p}|\tilde{\boldsymbol{\delta}}^a|_1\Big). \tag{141}$$

Combining (140) and (141), we have proved that

$$\begin{aligned}
&\sum_{a=1}^{p}\big\|\mathbf{X}(\hat{\tau})\hat{\boldsymbol{\beta}}^a - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2, \\
&\leq \lambda_T\Big\{\frac{18X_{\max}^2}{\kappa'^2}\lambda_T s_1 \bigvee \frac{6X_{\max}}{\kappa'}\Big(s_1 c_{\boldsymbol{\beta}}c_\tau C^*\sum_{a=1}^{p}|\tilde{\boldsymbol{\delta}}^a|_1\Big)^{1/2} \bigvee 3\Big(\sqrt{c_\tau} + (2X_{\min})^{-1}c_\tau C^*\sum_{a=1}^{p}|\tilde{\boldsymbol{\delta}}^a|_1\Big)\Big\}
\end{aligned}$$

which completes the proof of (46) in Lemma 17.

D.6.2 Proof of (47) in Lemma 17

After proving (46), we now prove (47). We also consider two cases:

Case 1: $\sqrt{c_\tau} + (2X_{\min})^{-1}C^*c_\tau \sum_{a=1}^{p}|\tilde{\boldsymbol{\delta}}^a|_1 \leq \sum_{a=1}^{p}\big|\hat{\mathbf{D}}(\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a)_{J_0^a}\big|_1$. In this case, by (133) and (138), we have

$$\begin{aligned}
&\sum_{a=1}^{p}\big|\hat{\mathbf{D}}(\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a)\big|_1 \\
&\leq \frac{3}{1-\mu}\sum_{a=1}^{p}\big|\hat{\mathbf{D}}(\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a)_{J_0^a}\big|_1, \\
&\leq \frac{3}{1-\mu}\sqrt{s_1}\sum_{a=1}^{p}\big|\hat{\mathbf{D}}(\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a)_{J_0^a}\big|_2, \\
&\leq \frac{3\sqrt{s_1}}{(1-\mu)\kappa'}X_{\max}\Big(\sum_{a=1}^{p}\big\|\mathbf{X}(\hat{\tau})\hat{\boldsymbol{\beta}}^a - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2 + 2c_{\boldsymbol{\beta}}c_\tau C^*\sum_{a=1}^{p}|\tilde{\boldsymbol{\delta}}^a|_1\Big)^{1/2}.
\end{aligned} \tag{142}$$

Note that $a+b \le 2a \vee 2b$ for $a, b \ge 0$. Combining the results in (140) and (142), we have

$$\sum_{a=1}^{p} \big|\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a\big|_1 \le \frac{1}{(1-\mu)X_{\min}} \Big\{ \frac{18X_{\max}^2}{\kappa'^2} \lambda_T s_1 \bigvee \frac{6X_{\max}}{\kappa'} \Big(s_1 c_{\boldsymbol{\beta}} c_\tau C^* \sum_{a=1}^{p} |\tilde{\boldsymbol{\delta}}^a|_1 \Big)^{1/2} \Big\}. \quad (143)$$

Case 2: $\sqrt{c_\tau} + (2X_{\min})^{-1} C^* c_\tau \sum_{a=1}^{p} |\tilde{\boldsymbol{\delta}}^a|_1 > \sum_{a=1}^{p} \big|\hat{\mathbf{D}}(\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a)_{J_0^a}\big|_1$. In this case, by (133), we have

$$\sum_{a=1}^{p} \big|\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a\big|_1 \le \frac{3}{(1-\mu)X_{\min}} \Big(\sqrt{c_\tau} + (2X_{\min})^{-1} C^* c_\tau \sum_{a=1}^{p} |\tilde{\boldsymbol{\delta}}^a|_1 \Big). \quad (144)$$

Finally, combining (143) and (144), we complete the proof of (47) in Lemma 17. ■

D.7 Proof of Lemma 18

Proof By the definitions of $\hat{\tau}$ and $(\hat{\boldsymbol{\beta}}^a)_{a=1}^p$ as obtained from (5) – (7), we have

$$\sum_{a=1}^{p} \Big\{ \|\mathbf{X}_a - \mathbf{X}(\hat{\tau})\hat{\boldsymbol{\beta}}^a\|_T^2 + \lambda_T |\hat{\mathbf{D}}\hat{\boldsymbol{\beta}}^a|_1 \Big\} \le \sum_{a=1}^{p} \Big\{ \|\mathbf{X}_a - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\|_T^2 + \lambda_T |\mathbf{D}\tilde{\boldsymbol{\beta}}^a|_1 \Big\}. \quad (145)$$

Conditional on $\bigcap_{1\le a\le p} \{\mathcal{A}^a \cap \mathcal{B}^a\} \bigcap \mathcal{C}(c_\tau)$, by the result in (114), we have

$$\begin{aligned}
&\sum_{a=1}^{p} \big\|\mathbf{X}_a - \mathbf{X}(\hat{\tau})\hat{\boldsymbol{\beta}}^a\big\|_T^2 - \sum_{a=1}^{p} \big\|\mathbf{X}_a - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2, \\
&= \sum_{a=1}^{p} \big\|\mathbf{X}(\hat{\tau})\hat{\boldsymbol{\beta}}^a - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2 - \sum_{a=1}^{p} \frac{2}{T} \sum_{t=1}^{T} \epsilon_a^t (\mathbf{X}^t)^\top (\hat{\boldsymbol{\theta}}^a - \tilde{\boldsymbol{\theta}}^a) \\
&\quad - \sum_{a=1}^{p} \frac{2}{T} \sum_{t=1}^{T} \epsilon_a^t (\mathbf{X}^t)^\top \mathbf{1}\{Q_t \le \hat{\tau}\} (\hat{\tilde{\boldsymbol{\delta}}}^a - \tilde{\boldsymbol{\delta}}^a) - R \\
&\ge \sum_{a=1}^{p} \big\|\mathbf{X}(\hat{\tau})\hat{\boldsymbol{\beta}}^a - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2 - \mu\lambda_T \sum_{a=1}^{p} \big|\hat{\mathbf{D}}(\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a)\big|_1 - \lambda_T \sqrt{c_\tau}.
\end{aligned} \quad (146)$$

Note that, by assumptions, $|\hat{\tau} - \tau_*| \le c_\tau$ and $\sum_{a=1}^{p} |\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a|_1 \le c_{\boldsymbol{\beta}}$ hold. Considering (145) and (146), under the event $\bigcap_{1\le a\le p} \{\mathcal{A}^a \cap \mathcal{B}^a\} \bigcap \mathcal{C}(c_\tau)$, we have

$$\begin{aligned}
&\sum_{a=1}^{p} \Big\{ \|\mathbf{X}_a - \mathbf{X}(\hat{\tau})\hat{\boldsymbol{\beta}}^a\|_T^2 + \lambda_T |\hat{\mathbf{D}}\hat{\boldsymbol{\beta}}^a|_1 \Big\} - \sum_{a=1}^{p} \Big\{ \|\mathbf{X}_a - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\|_T^2 + \lambda_T |\mathbf{D}\tilde{\boldsymbol{\beta}}^a|_1 \Big\}, \\
&\ge \sum_{a=1}^{p} \big\|\mathbf{X}(\hat{\tau})\hat{\boldsymbol{\beta}}^a - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2 - \sum_{a=1}^{p} \Big\{ \mu\lambda_T \big|\hat{\mathbf{D}}(\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a)\big|_1 + \lambda_T \big|\hat{\mathbf{D}}(\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a)\big|_1 \\
&+ \lambda_T |(\hat{\mathbf{D}} - \mathbf{D})\tilde{\boldsymbol{\beta}}^a|_1 \Big\} - \lambda_T \sqrt{c_\tau}, \\
&\ge \sum_{a=1}^{p} \big\|\mathbf{X}(\hat{\tau})\hat{\boldsymbol{\beta}}^a - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\big\|_T^2 - \lambda_T(1+\mu) c_{\boldsymbol{\beta}} X_{\max} - \lambda_T (2X_{\min})^{-1} c_\tau C^* \sum_{a=1}^{p} |\tilde{\boldsymbol{\delta}}^a|_1 - \lambda_T \sqrt{c_\tau},
\end{aligned} \quad (147)$$

where the last inequality in (147) comes from $\sum_{a=1}^{p} \left|\hat{\mathbf{D}}(\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a)\right|_1 \leq X_{\max} \sum_{a=1}^{p} |\hat{\boldsymbol{\beta}}^a - \tilde{\boldsymbol{\beta}}^a|_1 \leq X_{\max} c_{\boldsymbol{\beta}}$ and the result in (134). Recall

$$\tilde{c}_\tau := c_*^{-1} \lambda_T \Big((1+\mu) c_{\boldsymbol{\beta}} X_{\max} + (2X_{\min})^{-1} c_\tau C^* \sum_{a=1}^{p} |\tilde{\boldsymbol{\delta}}^a|_1 + \sqrt{c_\tau} \Big).$$

Suppose $\tilde{c}_\tau < |\hat{\tau} - \tau_*| \leq c_\tau$ holds. Then, by Proposition 6 and (147), we have

$$\begin{aligned}
&\sum_{a=1}^{p} \Big\{ \|\mathbf{X}_a - \mathbf{X}(\hat{\tau})\hat{\boldsymbol{\beta}}^a\|_T^2 + \lambda_T |\hat{\mathbf{D}}\hat{\boldsymbol{\beta}}^a|_1 \Big\} - \sum_{a=1}^{p} \Big\{ \|\mathbf{X}_a - \mathbf{X}(\tau_*)\tilde{\boldsymbol{\beta}}^a\|_T^2 + \lambda_T |\mathbf{D}\tilde{\boldsymbol{\beta}}^a|_1 \Big\}, \\
&> c_* \tilde{c}_\tau - \lambda_T (1+\mu) c_{\boldsymbol{\beta}} X_{\max} - \lambda_T (2X_{\min})^{-1} c_\tau C^* \sum_{a=1}^{p} |\tilde{\boldsymbol{\delta}}^a|_1 - \lambda_T \sqrt{c_\tau} = 0,
\end{aligned}$$

which contradicts with (145). Therefore, we have $|\hat{\tau} - \tau_*| \leq \tilde{c}_\tau$, which completes the proof. ■

References

Theodore Wilbur Anderson. *An Introduction to Multivariate Statistical Analysis.* Wiley, Hoboken, 3rd edition, 2003.

Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516, 2008.

Peter J Bickel, Yaacov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

Leland Bybee and Yves Atchadé. Change-point computation for large graphical models: A scalable algorithm for Gaussian graphical models with change-points. *Journal of Machine Learning Research*, 19:1–38, 2018.

T Tony Cai, Hongzhe Li, Weidong Liu, and Jichun Xie. Joint estimation of multiple high-dimensional precision matrices. *Statistica Sinica*, 26:445–464, 2016.

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.

Emmanuel Candes and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

Miklós Csörgö and Lajos Horváth. *Limit theorems in change-point analysis.* John Wiley and Sons Inc, 1997.

Patrick Danaher, Pei Wang, and Daniela M Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):373–397, 2014.

Farida Enikeeva and Zaid Harchaoui. High-dimensional change-point detection under sparse alternatives. *The Annals of Statistics*, 47(4):2051–2079, 2019.

Jianqing Fan, Yang Feng, and Yichao Wu. Network exploration via the adaptive lasso and scad penalties. *The Annals of Applied Statistics*, 3(2):521–541, 2009.

Bernd Fellinghauer, Peter Bühlmann, Martin Ryffel, Michael von Rhein, and Jan D. Reinhardt. Stable graphical model estimation with random forests for discrete, continuous, and mixed variables. *Computational Statistics and Data Analysis*, 64:132–152, 2013.

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

Alex J Gibberd and Sandipan Roy. Multiple changepoint estimation in high-dimensional Gaussian graphical models. *Preprint arXiv:1712.05786*, 2017.

Alexander J Gibberd and James DB Nelson. Regularized estimation of piecewise constant Gaussian graphical models: The group-fused graphical lasso. *Journal of Computational and Graphical Statistics*, 26(3):623–634, 2017.

Jian Guo, Elizaveta Levina, George Michailidis, and Ji Zhu. Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15, 2011.

Zaıd Harchaoui and Céline Lévy-Leduc. Multiple change-point estimation with a total variation penalty. *Journal of the American Statistical Association*, 105(492):1480–1493, 2010.

M Kolar and E. P. Xing. Estimating networks with jumps. *Electronic Journal of Statistics*, 6(1):2069–2106, 2012.

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

Steffen L Lauritzen. *Graphical models*. Oxford University Press, New York, 1996.

Sokbae Lee, Myung Hwan Seo, and Youngki Shin. The lasso for high dimensional regression with a possible change point. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(1):193–210, 2016.

Wonyul Lee and Yufeng Liu. Estimation of multiple graphical models with common structures. *Journal of Machine Learning Research*, 16(1):1035–1062, 2015.

Florencia Leonardi and Peter Bühlmann. Computationally efficient change point detection for high-dimensional regression. *Preprint arXiv:1601.03704*, 2016.

Weidong Liu. Gaussian graphical model estimation with false discovery rate control. *The Annals of Statistics*, 41(6):2948–2978, 2013.

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

Jie Peng, Pei Wang, Nengfeng Zhou, and Ji Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486):735–746, 2009.

S. Roy, Y. Atchadé, and G. Michailidis. Change point estimation in high dimensional Markov random-field models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1187–1206, 2017.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(1):267–288, 1996.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes with Applications to Statistics*. Springer, 1996.

Arend Voorman, Ali Shojaie, and Daniela Witten. Graph estimation with joint additive models. *Biometrika*, 101(1):85–101, 2014.

Jilei Yang and Jie Peng. Estimating time-varying graphical models. *Journal of Computational and Graphical Statistics*, 29(1):191–202, 2020.

Shiqing Yu, Mathias Drton, and Ali Shojaie. Generalized score matching for non-negative data. *Journal of Machine Learning Research*, 20:1–70, 2019.

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

Shuheng Zhou. Thresholded lasso for high dimensional variable selection and statistical estimation. *Preprint arXiv:1002.1583*, 2010.

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