Journal of Machine Learning Research 25 (2024) 1-71 Submitted 7/22; Published 7/24

Low-Rank Matrix Estimation in the Presence of Change-Points

Lei Shi LEISHI@BERKELEY.EDU
Department of Biostatistics
University of California, Berkeley
California 94704, United States

Guanghui Wang GHWANG.NK@GMAIL.COM
School of Statistics and Data Science, LPMC, KLMDASR, and LEBPS
Nankai University
Tianjin 300071, China

Changliang Zou NK.CHLZOU@GMAIL.COM
NITFID, School of Statistics and Data Science, LPMC, KLMDASR, and LEBPS
Nankai University
Tianjin 300071, China

Editor: Zaid Harchaoui

Abstract

We consider a general trace regression model with multiple structural changes and propose a universal approach for simultaneous exact or near-low-rank matrix recovery and change-point detection. It incorporates nuclear norm penalized least-squares minimization into a grid search scheme that determines the potential structural break. Under a set of general conditions, we establish the non-asymptotic error bounds with a nearly-oracle rate for the matrix estimators as well as the super-consistency rate for the change-point localization. We use concrete random design instances to justify the appropriateness of the proposed conditions. Numerical results demonstrate the validity and effectiveness of the proposed scheme.

Keywords: High-dimensional data, low-rank estimation, multiple change-point detection, non-asymptotic bounds, rate-optimal estimators

1. Introduction

High-dimensional low-rank matrix recovery has witnessed rapid development as well as tremendous success in both theoretical analysis and practical application. It appears in a wide variety of real-life scenarios, including recommendation systems (Ramlatchan et al., 2018), compressed sensing (Golbabaee and Vandergheynst, 2012), surveillance and environmental monitoring (Nobre and Stroup, 1994), economics and finance (Espinosa-Vega and Sole, 2010), and causal inference (Athey et al., 2021), to name a few. Suppose we have N observations $\{(y_i, \boldsymbol{X}_i)\}_{i=1}^N$, where $y_i \in \mathbb{R}$ is a response variable and $\boldsymbol{X}_i \in \mathbb{R}^{m_1 \times m_2}$ is a matrix of covariates. Consider the trace regression model

$$y_i = \langle \boldsymbol{X}_i \, , \, \boldsymbol{\Theta}^\star \rangle + \epsilon_i, \ i = 1, \ldots, N,$$

where $\mathbf{\Theta}^\star \in \mathbb{R}^{m_1 \times m_2}$ is the unknown low-rank matrix to be estimated, and ϵ_i is some unobserved noise. It is worth mentioning that a great number of interesting setups, such as multivariate regression, matrix completion, compressed sensing and vector auto-regressive processes can be encoded into this model (Negahban and Wainwright, 2011; Rohde and Tsybakov, 2011; Koltchinskii et al., 2011).

In real-life high-dimensional or big data applications, the underlying data-generating mechanism may encounter abrupt changes or transitions along time or some other variable. For instance, in a recommendation system, user preference for some products and services could change with time or vary with their age or income. In public health surveillance, reported case occurrences from multiple sites (which often implies a low-rank structure) may encounter sudden changes due to some policy interventions. To accommodate such scenarios, we consider the framework of matrix estimation in the presence of change-points or threshold effects, to wit,

$$y_i = \langle \boldsymbol{X}_i \,,\, \mathbf{\Theta}_s^\star \rangle + \epsilon_i, \;\; \tau_s^\star < t_i \le \tau_{s+1}^\star, \;\; s = 0, \ldots, s^\star; \;\; i = 1, \ldots, N, \tag{1}$$

where $t_i \in [0, 1]$ is some threshold variable (e.g., $t_i = i/N$ being the time index), $s^\star$ and $0 < \tau_1^\star < \cdots < \tau_{s^\star}^\star < 1$ denote respectively the number and locations of the change-points, with the convention of $\tau_0^\star = 0$ and $\tau_{s^\star+1}^\star = 1$, and $\mathbf{\Theta}_s^\star$ is the unknown *exact* or *near* low-rank matrix in the data segment corresponding to $t_i \in (\tau_s^\star, \tau_{s+1}^\star]$ for $s = 0, 1, \ldots, s^\star$. Of interest is to simultaneously recover $\mathbf{\Theta}_s^\star$'s and $\tau_s^\star$'s from the observations $\{(y_i, \boldsymbol{X}_i, t_i)\}_{i=1}^N$. Below we illustrate these definitions with some concrete examples.

Example 1 (Multivariate regression with change-points) *Suppose we have n observations $\{(\boldsymbol{y}_a, \boldsymbol{x}_a, t_a)\}_{a=1}^n$, where $t_a \in [0, 1]$ is the threshold variable, $\boldsymbol{x}_a \in \mathbb{R}^{m_1}$ is the variable of covariates and $\boldsymbol{y}_a \in \mathbb{R}^{m_2}$ is the multidimensional response variable. Each response-covariates-threshold triple is linked via the model*

$$\boldsymbol{y}_a = \mathbf{\Theta}_s^{\star\top} \boldsymbol{x}_a + \boldsymbol{w}_a, \;\; \tau_s^\star < t_a \le \tau_{s+1}^\star, \;\; s = 0, \ldots, s^\star; \;\; a = 1, \ldots, n,$$

where $\tau_s^\star$'s are the change-points, $\mathbf{\Theta}_s^\star \in \mathbb{R}^{m_1 \times m_2}$ are the corresponding low-rank matrices, and $\boldsymbol{w}_a \in \mathbb{R}^{m_2}$ are the noises. This model can be formulated into Model (1) by setting

$$t_i = t_a, \boldsymbol{X}_i = \boldsymbol{x}_a \boldsymbol{e}_b^\top, y_i = \boldsymbol{e}_b^\top \boldsymbol{y}_a, \epsilon_i = \boldsymbol{e}_b^\top \boldsymbol{w}_a, \;\; i = 1, \ldots, N (= n m_2),$$

where we use the map $(a, b) \mapsto i = (a-1)m_2 + b$, and $\boldsymbol{e}_b \in \mathbb{R}^{m_2}$ denotes the canonical basis vector with a single one in position b, for $a = 1, \ldots, n$ and $b = 1, \ldots, m_2$.

Example 2 (Compressed sensing with change-points) *Working with Model (1), suppose that the design matrices $\boldsymbol{X}_i \in \mathbb{R}^{m_1 \times m_2}$ are drawn i.i.d. from a standard Gaussian ensemble, meaning that each entry is an i.i.d. draw from the $N(0, 1)$ distribution.*

Example 3 (Vector auto-regressive (VAR) process with change-points) *Suppose we have n observations $\{(\boldsymbol{z}_a, t_a)\}_{a=1}^n$, where $t_a \in [0, 1]$ is the threshold variable, and $\boldsymbol{z}_a \in \mathbb{R}^m$ are generated by firstly choosing $\boldsymbol{z}_a$ according to some initial distribution, and then recursively setting*

$$\boldsymbol{z}_a = \mathbf{\Theta}_s^\star \boldsymbol{z}_{a-1} + \boldsymbol{w}_a, \;\; \tau_s^\star < t_a \le \tau_{s+1}^\star, \;\; s = 0, \ldots, s^\star; \;\; a = 2, \ldots, n,$$

where $\tau_s^\star$'s are the change-points, $\boldsymbol{\Theta}_s^\star \in \mathbb{R}^{m\times m}$ are the corresponding low-rank matrices, and $\boldsymbol{w}_a$'s are the noises. This model can be formulated as a particular instance of Model (1) with

$$t_i = t_a, \boldsymbol{X}_i = \boldsymbol{e}_b \boldsymbol{z}_{a-1}^\top, y_i = \boldsymbol{e}_b^\top \boldsymbol{z}_a, \epsilon_i = \boldsymbol{e}_b^\top \boldsymbol{w}_{i-1},\ i = 2, \ldots, N,$$

where i indexes the sample (a, b) and $\{\boldsymbol{e}_b \in \mathbb{R}^m\}_{b=1}^N$ are the basis vectors.

For vector-valued covariates, Model (1) is reduced to the linear regression model with structural breaks, and the goal here is to detect changes in the sparse regression coefficient, which has attracted considerable attention recently, see, for example, Lee et al. (2016), Leonardi and Bühlmann (2016), Kaul et al. (2019), Rinaldo et al. (2021) and Wang et al. (2021b). Despite the popularity of huge volumes of data collected in matrix form nowadays, there are only a limited number of estimation schemes designed for Model (1). For the VAR change model in Example 3, if the regression matrices $\boldsymbol{\Theta}_s^\star$'s are assumed to be sparse instead of low-rank, Safikhani and Shojaie (2022) and Safikhani et al. (2022) proposed a fused LASSO method and Wang et al. (2019a) suggested a dynamic programming approach. Bai et al. (2020) assumed that each regression matrix is a superposition of a stable low-rank component and a time-varying sparse component, and proposed a fused LASSO type estimation scheme. By allowing both the low-rank and sparse components to exhibit changes, Bai et al. (2023) developed a rolling window detection strategy.

In this paper, we attempt to develop a theoretically guaranteed methodology for low-rank matrix recovery in the presence of multiple change-points under the framework of Model (1). We first propose a *joint minimization* procedure for simultaneous matrix estimation and change detection if there is at most one change-point occurring to the data sequence. To be specific, we minimize the nuclear-norm-penalized least-squares over all feasible choices of the regression matrices and change-point. The idea of joint minimization is motivated by Lee et al. (2016), which studied the LASSO for high-dimensional linear regression with a possible change-point. However, tackling the nuclear norm incurs more technical difficulties due to its *inseparability*. Several conditions and techniques used in Lee et al. (2016) rely heavily on the separability of the ℓ_1-norm, and thus appear restrictive and hard to generalize. Fortunately, our proposed scheme provably yields not only desirable matrix estimators that match the optimal error rate of those obtained without any changes (e.g., Negahban and Wainwright (2011)), but also *super-consistent* estimation of the change-point (Chan, 1993; Lee et al., 2016). We further extend this scheme to the scenario with multiple change-points by considering a two-stage procedure.

1.1 Our contributions

From the methodological aspect, we propose a universal approach for simultaneous low-rank matrix estimation and multiple change-point detection for the general trace regression model with threshold effects (i.e., Model (1)). It builds on a recovery scheme that incorporates least-squares minimization with the nuclear norm penalty. To tailor for multiple change-points scenarios, we provide a novel thresholding rule followed by additional refinements to achieve desirable estimation and detection accuracy simultaneously.

From the theoretical aspect, we formulate general conditions under which our estimation and detection procedure is valid. Those conditions stand as non-trial extensions compared with classical results in the literature of low-rank matrix recovery or change-point detection. They are established under a fixed design setup and aim to incorporate a broad class of designs. When those conditions hold, we have theoretical guarantee for both the change-point localization and matrix estimation, that is, the convergence rate for the matrix estimators provably achieves the optimal rate for high-dimensional low-rank recovery without threshold effects, and the detected change-points have the super-consistency property. Moreover, using multivariate regression (Example 1) as a running example, we establish concrete results to justify the appropriateness of the general conditions as well as the validity of the proposed scheme.

1.2 Related literature

In the absence of change-points, a variety of powerful low-rank matrix estimation frameworks have been developed during the past decades, which cover many real-life application instances as well as different model setups. For example, Candès and Recht (2009) and Recht et al. (2010) studied a nuclear norm convex relaxation framework for noiseless matrix completion under the sampling-without-replacement scheme and different bases. They also explored reasonable conditions for successful recovery, like incoherence assumptions, which built up the foundation of the theoretical guarantee. When noises are inevitable, Keshavan et al. (2010) and Candès and Plan (2011) followed the thread of nuclear norm convex relaxation framework, while Negahban and Wainwright (2011) and Koltchinskii et al. (2011), among others, developed the nuclear norm penalization least-squares estimation, which is akin to LASSO in vector-based optimizations. These works also established the convergence rates of the proposed estimator under general conditions such as restricted strong convexity and (generalized) restricted isometry property. Following works made extensions and adaptation to other aspects, such as robustness (Elsener and van de Geer, 2018), non-Gaussian data (Fan et al., 2019), missingness quantification (Fithian and Mazumder, 2018), nonconvex optimization (Chen and Chi, 2018) and so on.

On the other hand, change-point detection also constitutes a canonical problem with numerous applications and has witnessed the development of many mature schemes. It dates back to 1950s (Page, 1954), and has gained increasing attention recently for modeling high-dimensional data, which is often exposed to some degree of heterogeneity in the form of abrupt changes in the parameters of the underlying data generating process. In particular, it has been used in the context of high-dimensional mean and covariance models (Cho and Fryzlewicz, 2015; Wang and Samworth, 2018; Wang et al., 2018; Yu and Chen, 2021; Liu et al., 2020; Dette et al., 2022), graphical models (Bybee and Atchadé, 2018; Londschien et al., 2021; Liu et al., 2021), networks (Wang et al., 2021a), and regression models (Lee et al., 2016; Leonardi and Bühlmann, 2016; Kaul et al., 2019; Wang et al., 2021b; Safikhani and Shojaie, 2022; Bai et al., 2020, 2023), to name a few.

1.3 Structure of the paper

The remainder of our paper is structured as follows. In Section 2, we first introduce the joint minimization scheme, together with its theoretical properties and implementation, if there exists at most one change-point. Then this estimation and detection procedure is extended to multiple change-points scenarios in Section 3. Numerical studies are presented in Section 4. Section 5 concludes the paper. All proofs regarding the theoretical results, together with additional numerical supports, are deferred to Appendix.

1.4 Notations

For a matrix $\boldsymbol{X}$, let X_{ij} be its (i,j)-th entry. Likewise, for a vector $\boldsymbol{x}$, let x_i be its ith component. For a matrix $\boldsymbol{X} \in \mathbb{R}^{m_1 \times m_2}$, we use $\operatorname{rank}(\boldsymbol{X})$ and $\rho_k(\boldsymbol{X})$ to denote respectively the rank and the k-th singular value of a given matrix $\boldsymbol{X}$ for $k = 1, \ldots, m := \min\{m_1, m_2\}$. The Schattern-$q$ norm of $\boldsymbol{X}$ is defined as $\|\boldsymbol{X}\|_{S_q} = \left\{\sum_{k=1}^{\operatorname{rank}(\boldsymbol{X})} \varrho_k(\boldsymbol{X})^q\right\}^{1/q}$. When $q = 2, \infty, 1$, the Schattern-q norm reduces to the commonly used Frobenius, operator and nuclear norm, which are denoted as $\|\boldsymbol{X}\|_F$, $\|\boldsymbol{X}\|_{\mathrm{op}}$ and $\|\boldsymbol{X}\|_*$, respectively. For two matrices $\boldsymbol{X}_1, \boldsymbol{X}_2 \in \mathbb{R}^{m_1 \times m_2}$, we denote their inner product as $\langle \boldsymbol{X}_1\,,\, \boldsymbol{X}_2\rangle = \operatorname{tr}\left(\boldsymbol{X}_1^\top \boldsymbol{X}_2\right)$, where $\operatorname{tr}(\cdot)$ is the trace operator. For vectors, we use $\|\cdot\|_1$ and $\|\cdot\|_2$ for the ℓ_1 and ℓ_2 norms, respectively. The capital letter N is used to denote the sample size under the general trace regression model (2) and (12). The small letter n, on the contrary, denotes the sample size in the specific multivariate regression model (Example 1).

2. Matrix estimation with a possible change-point

2.1 Joint minimization scheme

2.1.1 Model and reparameterization

We first confine attention to the at most one change-point (AMOC) scenario, i.e., Model (1) with $s^\star \le 1$. To be specific, suppose we have observations $\{(y_i, \boldsymbol{X}_i, t_i)\}_{i=1}^N$ such that

$$y_i = \langle \boldsymbol{X}_i\,,\, \boldsymbol{\Theta}_0^\star\rangle\, \mathbf{1}\left\{t_i \le \tau_1^\star\right\} + \langle \boldsymbol{X}_i\,,\, \boldsymbol{\Theta}_1^\star\rangle\, \mathbf{1}\left\{t_i > \tau_1^\star\right\} + \epsilon_i,$$

where $y_i \in \mathbb{R}$ is a response, $\boldsymbol{X}_i \in \mathbb{R}^{m_1 \times m_2}$ is a matrix of covariates, $t_i \in [0,1]$ represents a threshold variable with an unknown change-point $\tau_1^\star$ splitting the sample into two segments, $\boldsymbol{\Theta}_0^\star, \boldsymbol{\Theta}_1^\star \in \mathbb{R}^{m_1 \times m_2}$ are unknown matrices to be estimated in both segments, and ϵ_i is a noise. After reparameterizing $\boldsymbol{\Theta}^\star = \boldsymbol{\Theta}_0^\star$, $\boldsymbol{\Delta}^\star = \boldsymbol{\Theta}_1^\star - \boldsymbol{\Theta}_0^\star$ and $\tau^\star = \tau_1^\star$, and collecting $\boldsymbol{\Gamma}^\star = \left(\boldsymbol{\Theta}^{\star\top}, \boldsymbol{\Delta}^{\star\top}\right)^\top$, the AMOC model is equivalent to

$$\begin{aligned} y_i &= \langle \boldsymbol{X}_i\,,\, \boldsymbol{\Theta}^\star\rangle + \langle \boldsymbol{X}_i\,,\, \boldsymbol{\Delta}^\star\rangle\, \mathbf{1}\left\{t_i > \tau^\star\right\} + \epsilon_i, \\ &= \langle \boldsymbol{\mathcal{X}}_i(\tau^\star)\,,\, \boldsymbol{\Gamma}^\star\rangle + \epsilon_i, \end{aligned} \tag{2}$$

where we denote $\boldsymbol{\mathcal{X}}_i(\tau) = \left(\boldsymbol{X}_i^\top, \boldsymbol{X}_i(\tau)^\top\right)^\top$ with $\boldsymbol{X}_i(\tau) := \boldsymbol{X}_i \mathbf{1}\left\{t_i > \tau\right\}$ for any $0 < \tau < 1$.

2.1.2 Low-rank structure

In many applications, the regression matrices $\boldsymbol{\Theta}_s^\star$'s ($s = 0$ and 1) are either low-rank, or well approximated by low-rank matrices. If we impose low-rank restriction on $\boldsymbol{\Theta}_s^\star$'s, then $\boldsymbol{\Delta}^\star$ and $\boldsymbol{\Gamma}^\star$ are also of low-rank since

$$\max\{\operatorname{rank}(\boldsymbol{\Delta}^\star), \operatorname{rank}(\boldsymbol{\Gamma}^\star)\} \le 2\max\left\{\operatorname{rank}(\boldsymbol{\Theta}_0^\star), \operatorname{rank}(\boldsymbol{\Theta}_1^\star)\right\};$$

see Theorem 25. If $\boldsymbol{\Theta}_s^\star$'s have a more generally near low-rank structure (Negahban and Wainwright, 2011), i.e., their singular values fall within an ℓ_q-ball $\mathbb{B}_q(R_q) := \{\boldsymbol{\varrho} \in \mathbb{R}^m :$

$\sum_{k=1}^{m} |\varrho_k|^q \leq R_q\}$ for some $q \in (0,1)$ and $R_q > 0$, where $m = \min\{m_1, m_2\}$, then the transition matrix $\mathbf{\Delta}^\star$ should belong to $\mathbb{B}_q(2R_q)$ due to the additive property of the Schatten-q norm; see Rohde and Tsybakov (2011) and the references therein. Note that, by taking $q \to 0$, $\mathbb{B}_q(R_q)$ approaches the low-rank matrix space. Thus we can unify the exact and near low-rank matrix spaces with the notion of ℓ_q-balls by setting $q \in [0,1)$.

2.1.3 Penalized least-squares estimation

The above fact suggests a natural nuclear norm penalized least-squares estimator for $\mathbf{\Gamma}^\star$ if the chang-point is known as $\tau^\star = \tau$ for some $0 < \tau < 1$, that is,

$$\widehat{\mathbf{\Gamma}}(\tau) = \underset{\mathbf{\Gamma} \in \mathbb{R}^{(2m_1)\times m_2}}{\arg\min} \left\{ S_N(\mathbf{\Gamma}; \tau) + \lambda_N \|\mathbf{\Gamma}\|_* \right\}, \tag{3}$$

where

$$S_N(\mathbf{\Gamma}; \tau) = (2N)^{-1} \sum_{i=1}^{N} \left(y_i - \langle \boldsymbol{\mathcal{X}}_i(\tau) \, , \mathbf{\Gamma} \rangle\right)^2, \tag{4}$$

and $\lambda_N > 0$ is a regularization parameter that will be specified later. Then we can estimate the change-point $\tau^\star$ by searching for the best τ that yields the minimal value of penalized least-squares, namely,

$$\widehat{\tau} = \underset{\tau \in \mathbb{T}}{\arg\min} \left\{ S_N\left(\widehat{\mathbf{\Gamma}}(\tau); \tau\right) + \lambda_N \left\|\widehat{\mathbf{\Gamma}}(\tau)\right\|_* \right\},$$

where $\mathbb{T} = [\rho, 1-\rho] \subset [0,1]$ represents a parameter space for $\tau^\star$, and ρ is some boundary removal parameter that is frequently considered in the change-point detection literature (Csörgő and Horváth, 1997). At last, we obtain the estimator of $\mathbf{\Gamma}^\star$ as $\widehat{\mathbf{\Gamma}}(\widehat{\tau})$. In fact, the proposed estimator of $(\mathbf{\Gamma}^\star, \tau^\star)$ can be regarded as a joint minimization problem, i.e.,

$$\left(\widehat{\mathbf{\Gamma}}(\widehat{\tau}), \widehat{\tau}\right) = \underset{\mathbf{\Gamma} \in \mathbb{R}^{(2m_1)\times m_2}, \tau \in \mathbb{T}}{\arg\min} \left\{ S_N(\mathbf{\Gamma}; \tau) + \lambda_N \|\mathbf{\Gamma}\|_* \right\}. \tag{5}$$

Remark 1 *Since the nuclear norm is not separable, i.e., $\|\mathbf{\Gamma}\|_* \neq \|\mathbf{\Theta}\|_* + \|\mathbf{\Delta}\|_*$ for $\mathbf{\Gamma} = \left(\mathbf{\Theta}^\top, \mathbf{\Delta}^\top\right)^\top$, another form of penalization one might consider is $\|\mathbf{\Theta}\|_* + \|\mathbf{\Delta}\|_*$. Theoretically speaking, these two choices are equivalent to each other if we rescale the penalization factor by some constant, which can be established via the fact $(\|\mathbf{\Theta}\|_* + \|\mathbf{\Delta}\|_*)/\sqrt{2} \leq \|(\mathbf{\Theta}^\top, \mathbf{\Delta}^\top)^\top\|_* \leq \|\mathbf{\Theta}\|_* + \|\mathbf{\Delta}\|_*$, see Theorem 25. Alternatively, one might penalize $\mathbf{\Theta}_0^\star$ and $\mathbf{\Theta}_1^\star$ instead of $\mathbf{\Theta}^\star = \mathbf{\Theta}_0^\star$ and $\mathbf{\Delta}^\star = \mathbf{\Theta}_1^\star - \mathbf{\Theta}_0^\star$, which leads to solutions with similar theoretical properties (more precisely, non-asymptotic bounds with the same rates up to some constants). This is suggested by the fact that*

$$\begin{pmatrix} \mathbf{\Theta} \\ \mathbf{\Delta} \end{pmatrix} = \begin{pmatrix} \boldsymbol{I}_{m_1} & \boldsymbol{O} \\ -\boldsymbol{I}_{m_1} & \boldsymbol{I}_{m_1} \end{pmatrix} \begin{pmatrix} \mathbf{\Theta}_0 \\ \mathbf{\Theta}_1 \end{pmatrix}.$$

The transformation matrix is invertible and has only two distinct (but repeated) singular values, i.e., 1 and $\sqrt{2}$. By Theorem 26, both the penalization factor of the objective function and the non-asymptotic bounds can be rescaled up to some constants.

2.1.4 Multivariate regression example

Regarding Example 1, let $\boldsymbol{\mathcal{X}}_a(\tau) = \big(\boldsymbol{x}_a^\top, \boldsymbol{x}_a^\top \mathbf{1}\{t_a > \tau\}\big)^\top$ for some $\tau \in \mathbb{T}$. This change-point model can be rewritten as $\boldsymbol{y}_a = \boldsymbol{\Gamma}^{\star\top} \boldsymbol{\mathcal{X}}_a(\tau^\star) + \boldsymbol{w}_a$, where $\boldsymbol{\Gamma}^\star = \big(\boldsymbol{\Theta}_0^{\star\top}, \boldsymbol{\Theta}_1^{\star\top} - \boldsymbol{\Theta}_0^{\star\top}\big)^\top$. In this case our procedure proceeds as

$$\Big(\widehat{\boldsymbol{\Gamma}}(\widehat{\tau}), \widehat{\tau}\Big) = \underset{\boldsymbol{\Gamma} \in \mathbb{R}^{(2m_1) \times m_2}, \tau \in \mathbb{T}}{\arg\min} \left\{ \frac{1}{2n} \sum_{a=1}^{n} \Big\| \boldsymbol{y}_a - \boldsymbol{\Gamma}^\top \boldsymbol{\mathcal{X}}_a(\tau) \Big\|_2^2 + \lambda_n \|\boldsymbol{\Gamma}\|_* \right\},$$

for some $\lambda_n > 0$.

2.2 Theoretical analysis

2.2.1 Preview

In this section, we will perform a thorough analysis of the statistical properties of the regularized estimator $\big(\widehat{\boldsymbol{\Gamma}}(\widehat{\tau}), \widehat{\tau}\big)$ in (5). In case the discussion becomes too involved due to its theoretical essence, we give a block of preview of the core results at the beginning as well as a flow map to explain how the results are organized and guide the audience through the reading process. Furthermore, we will employ the multivariate regression setting as a special running example to make a short demonstration of the key results throughout the process.

Preview of the core results. From a high level, the goal in this section is to establish finite sample bounds on the estimation error of the estimator pair $\big(\widehat{\boldsymbol{\Gamma}}(\widehat{\tau}), \widehat{\tau}\big)$ as well as the in-sample prediction error under an appropriate set of assumptions. The whole analysis can be dissected into several components:

- *Main results.* There are two main results that involve different conditions and techniques for justification.

  The first result (Theorem 4) targets the setting where there is no structural change along the sequence of data points and provides the following finite sample rates on estimation and prediction error with some conditions and an appropriate choice of tuning parameter λ_N:

  $$\|\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^\star\|_F^2 \lesssim \lambda_N^2 r \vee \lambda_N \sum_{k=r+1}^{m} \rho_k(\boldsymbol{\Gamma}^\star), \quad \|\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^\star\|_* \lesssim \lambda_N r \vee \sum_{k=r+1}^{m} \rho_k(\boldsymbol{\Gamma}^\star),$$
  $$\frac{1}{2N} \sum_{i=1}^{N} \Big\langle \boldsymbol{\mathcal{X}}_i(\widehat{\tau})\,,\, \widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^\star \Big\rangle^2 \lesssim \lambda_N^2 r \vee \lambda_N \sum_{k=r+1}^{m} \rho_k(\boldsymbol{\Gamma}^\star).$$

  The results match those of an “oracle” estimator with the no-change prior knowledge.

  The second result (Theorem 7) is tailored to the data-generating process containing exactly one change-point. It establishes the following finite-sample rates for estimation and prediction error as well as change-point detection accuracy:

$$\|\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}^\star\|_F^2 \lesssim \lambda_N^2 r \vee \lambda_N \sum_{k=r+1}^m \rho_k(\mathbf{\Gamma}^\star), \quad \|\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}^\star\|_* \lesssim \lambda_N r \vee \sum_{k=r+1}^m \rho_k(\mathbf{\Gamma}^\star),$$

$$\frac{1}{2N}\left\|\mathfrak{X}\left(\widehat{\mathbf{\Gamma}};\widehat{\tau}\right)-\mathfrak{X}(\mathbf{\Gamma}^\star;\tau^\star)\right\|_2^2 \lesssim \lambda_N \sum_{k=r+1}^m \rho_k(\mathbf{\Gamma}^\star) \vee \lambda_N^2 r,$$

$$|\widehat{\tau}-\tau^\star| \lesssim \lambda_N \sum_{k=r+1}^m \rho_k(\mathbf{\Gamma}^\star) \vee \lambda_N^2 r.$$

The results give the same bounds (up to some constants) as those in Theorem 4 for the matrix estimation error as well as the prediction error in the presence of threshold effect. Moreover, Theorem 7 builds the error bound for change-point detection, which can be viewed as a non-asymptotic version of the super-consistency of $\widehat{\tau}$ to $\tau^\star$ for general low-rank matrix recovery in the presence of a change-point.

Additionally, we provide Corollary 9 and 10, which are concrete statements of Theorem 7 under exact and approximate low-rank cases, respectively.

- *Statement of conditions and assumptions.* The results rely on some distributional and structural assumptions on the model, regarding a unique restricted strong convexity nature of the optimization program (Assumption 1), identifiability and smoothness of the model with one change-point (Assumptions 2–3), tail of the noise (Assumption 4), etc.

- *Intermediate results to complete the logic flow.* Towards proving the main results, we fully exploit the property of the minimization program and establish several intermediate lemmas and corollaries that facilitate understanding and discussion. To name a few, Lemma 2 gives a deterministic inequality that provides instruction on choosing the correct level of penalization. Corollary 3 establishes prediction consistency for both settings. With a single change-point, Lemma 5 shows change-point detection consistency. These results are crucial ingredients for the main theorems.

- *Illustration under a concrete random design running example.* Assumptions 1–5 shall be verified under a multivariate regression example in Section 2.2.6. Under appropriate assumptions on the design and noise (see Assumption 6) and the scenario that $\mathbf{\Gamma}^\star$ has exact low rank r, if we choose $\lambda_n \asymp \sqrt{(m_1+m_2)/n}$, then with high probability,

$$\left\|\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}^\star\right\|_F^2 \lesssim r\lambda_n^2, \quad \left\|\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}^\star\right\|_* \lesssim r\lambda_n, \quad \frac{1}{2n}\left\|\mathfrak{X}\left(\widehat{\mathbf{\Gamma}};\widehat{\tau}\right)-\mathfrak{X}(\mathbf{\Gamma}^\star;\tau^\star)\right\|_2^2 \lesssim r\lambda_n^2,$$

and

$$|\widehat{\tau}-\tau^\star| \lesssim r\lambda_n^2.$$

Discussion on these rates is deferred to Section 2.2.6, together with the rates derived under near low-rank scenarios.

Before moving to the formal presentation of the results, we further introduce some quick notations. Let $\boldsymbol{y} = (y_1, \dots, y_N)^\top$ and $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_N)^\top$. Given $\tau \in \mathbb{T}$, define an observation operator $\mathfrak{X}(\cdot;\tau) : \mathbb{R}^{(2m_1)\times m_2} \mapsto \mathbb{R}^N$, with elements $[\mathfrak{X}(\boldsymbol{\Gamma};\tau)]_i = \langle \boldsymbol{\mathcal{X}}_i(\tau)\,,\, \boldsymbol{\Gamma}\rangle$ for $\boldsymbol{\Gamma} \in \mathbb{R}^{(2m_1)\times m_2}$. Intuitively, this linear operator measures the noiseless output signal through the AMOC model (2) with any given input $\boldsymbol{\Gamma}$. Therefore, with the operator $\mathfrak{X}(\cdot;\tau)$, we can reformulate Model (2) as $\boldsymbol{y} = \mathfrak{X}(\boldsymbol{\Gamma}^\star;\tau^\star)+\boldsymbol{\epsilon}$. The adjoint of the observation operator, denoted by $\mathfrak{X}^\star(\cdot;\tau)$, is the linear mapping from $\mathbb{R}^N$ to $\mathbb{R}^{(2m_1)\times m_2}$ given by $\mathfrak{X}^\star(\boldsymbol{v};\tau) = \sum_{i=1}^N v_i \boldsymbol{\mathcal{X}}_i(\tau)$ for $\boldsymbol{v} \in \mathbb{R}^N$.

2.2.2 Decomposable subspaces and prediction consistency

The crucial ingredient in our analysis is the specification of certain subspaces onto which we can project the regression matrices and utilize the low-rank structure. To formalize the idea, consider the singular value decomposition of the target matrix $\boldsymbol{\Gamma}^\star$. For each integer $r \in \{1, \dots, m\}$, let $\mathbb{U}^r := [\boldsymbol{u}_1, \dots, \boldsymbol{u}_r] \in \mathbb{R}^{m_1\times r}$ and $\mathbb{V}^r := [\boldsymbol{v}_1, \dots, \boldsymbol{v}_r] \in \mathbb{R}^{m_2\times r}$ be the subspaces spanned by the top r left and right singular vectors of $\boldsymbol{\Gamma}^\star$. We introduce the orthogonal decomposition $\mathbb{R}^{m_1\times m_2} = \mathcal{S}^r \oplus \mathcal{S}^{r\perp}$, where $\mathcal{S}^r$ is the linear space spanned by the elements of the form $\boldsymbol{u}_k\boldsymbol{x}^\top$ and $\boldsymbol{y}\boldsymbol{v}_k^\top$, $k = 1, \dots, r$, where $\boldsymbol{x}$ and $\boldsymbol{y}$ are arbitrary, and $\mathcal{S}^{r\perp}$ is its orthogonal complement. The orthogonal projection $\Pi^r_{\boldsymbol{\Gamma}^\star}$ onto $\mathcal{S}^r$ is given by $\Pi^r_{\boldsymbol{\Gamma}^\star}(\boldsymbol{M}) = \boldsymbol{P}_{\mathbb{U}^r}\boldsymbol{M} + \boldsymbol{M}\boldsymbol{P}_{\mathbb{V}^r} - \boldsymbol{P}_{\mathbb{U}^r}\boldsymbol{M}\boldsymbol{P}_{\mathbb{V}^r}$ for any matrix $\boldsymbol{M} \in \mathbb{R}^{m_1\times m_2}$, where $\boldsymbol{P}_{\mathbb{U}^r}$ and $\boldsymbol{P}_{\mathbb{V}^r}$ are orthogonal projections onto $\mathbb{U}^r$ and $\mathbb{V}^r$. The orthogonal projection $\Pi^{r\perp}_{\boldsymbol{\Gamma}^\star}$ onto $\mathcal{S}^{r\perp}$ is given by $\Pi^{r\perp}_{\boldsymbol{\Gamma}^\star}(\boldsymbol{M}) = (\boldsymbol{I}_{m_1} - \boldsymbol{P}_{\mathbb{U}^r})\boldsymbol{M}(\boldsymbol{I}_{m_2} - \boldsymbol{P}_{\mathbb{V}^r})$. These projection operators have appeared in many literature of low-rank matrix estimation, see, for example, Candès and Recht (2009), Recht (2011) and Negahban and Wainwright (2011).

We start the formal theoretical discussion by providing a preliminary lemma, which is an inequality that builds up the foundation of our theory.

Lemma 2 (Basic inequality) *If $\lambda_N \geq \sup_{\tau\in\mathbb{T}} 2\|\mathfrak{X}^\star(\boldsymbol{\epsilon};\tau)\|_{\mathrm{op}}/N$, then*

$$\begin{aligned}&\frac{1}{2N}\sum_{i=1}^N \left(\left\langle \boldsymbol{\mathcal{X}}_i(\widehat{\tau})\,,\, \widehat{\boldsymbol{\Gamma}}\right\rangle - \langle \boldsymbol{\mathcal{X}}_i(\tau^\star)\,,\, \boldsymbol{\Gamma}^\star\rangle\right)^2 + \frac{\lambda_N}{2}\left\|\Pi^{r\perp}_{\boldsymbol{\Gamma}^\star}\left(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^\star\right)\right\|_* \\ \leq &2\lambda_N\left\|\Pi^{r\perp}_{\boldsymbol{\Gamma}^\star}(\boldsymbol{\Gamma}^\star)\right\|_* + \frac{3\lambda_N}{2}\left\|\Pi^{r}_{\boldsymbol{\Gamma}^\star}\left(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^\star\right)\right\|_* + \mathcal{R}_N(\boldsymbol{\Gamma}^\star, \widehat{\tau}, \tau^\star), \qquad (6)\end{aligned}$$

where for given $\tau, \tau' \in \mathbb{T}$, $\mathcal{R}_N(\boldsymbol{\Gamma}^\star, \tau, \tau')$ is defined as

$$\mathcal{R}_N(\boldsymbol{\Gamma}^\star, \tau, \tau') = N^{-1}\sum_{i=1}^N \epsilon_i \left\langle \boldsymbol{\mathcal{X}}_i(\tau) - \boldsymbol{\mathcal{X}}_i(\tau')\,,\, \boldsymbol{\Gamma}^\star\right\rangle = N^{-1}\sum_{i=1}^N \epsilon_i \left\langle \boldsymbol{X}_i(\tau) - \boldsymbol{X}_i(\tau')\,,\, \boldsymbol{\Delta}^\star\right\rangle.$$

We add some remarks on Lemma 2. First, Lemma 2 is a deterministic result. The left-hand side of (6) in Lemma 2 contains two terms. The first one corresponds to the *prediction error.* The second term, $(\lambda_N/2)\left\|\Pi^{r\perp}_{\boldsymbol{\Gamma}^\star}\left(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^\star\right)\right\|_*$, combined with a direct projection term $(\lambda_N/2)\left\|\Pi^{r}_{\boldsymbol{\Gamma}^\star}\left(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^\star\right)\right\|_*$, measures the magnitude of *matrix estimation error* in nuclear norm. (6) holds under the requirement $\lambda_N \geq \sup_{\tau\in\mathbb{T}} 2\|\mathfrak{X}^\star(\boldsymbol{\epsilon};\tau)\|_{\mathrm{op}}/N$, which puts a restriction on the specification of the regularization parameter λ_N. This is a generalization of the no-threshold-effect result in Negahban and Wainwright (2011), where they used

$\lambda_N \geq 2\| \sum_{i=1}^N \epsilon_i \boldsymbol{X}_i \|_{\mathrm{op}}/N$. Our choice here incorporates the change structure information. We shall use the random multivariate regression example in Section 2.2.6 to show that with choice λ_N of the order $O(\sqrt{(m_1+m_2)/n})$, this requirement holds with high probability. Second, the remainder term $\mathcal{R}_N$ plays an important role in our analysis. Its scale reflects the noise level (ϵ part), discontinuity of the design for the change-points ($\mathbf{X}(\tau)$ part), and the size of the break $\boldsymbol{\Delta}^\star$. In the no-change-point setting, the remainder $\mathcal{R}_N$ is zero because $\boldsymbol{\Delta}^\star$ is zero so the term does not affect the analysis. In the presence of one single change-point, in general, $\mathcal{R}_N$ will not vanish. However, we are able to control its scale under a suitable set of conditions; see Lemma 6 later.

If we further assume that the operator norms of $\boldsymbol{\Gamma}^\star$ and $\widehat{\boldsymbol{\Gamma}}$ have an upper bound, say $\gamma_{max}/2$, then Lemma 2 provides the following upper bound on the prediction error.

Corollary 3 (Prediction consistency) *If $\lambda_N \geq \sup_{\tau\in\mathbb{T}} 2\|\mathfrak{X}^\star(\boldsymbol{\epsilon};\tau)\|_{\mathrm{op}}/N$, then*

$$\frac{1}{2N}\sum_{i=1}^N \left(\left\langle \boldsymbol{\mathcal{X}}_i(\widehat{\tau}) \, , \, \widehat{\boldsymbol{\Gamma}} \right\rangle - \langle \boldsymbol{\mathcal{X}}_i(\tau^\star) \, , \, \boldsymbol{\Gamma}^\star \rangle \right)^2 \leq 2\lambda_N \sum_{k=r+1}^m \rho_k(\boldsymbol{\Gamma}^\star) + 6\lambda_N r \gamma_{max} + \lambda_N \|\boldsymbol{\Delta}^\star\|_*.$$

Corollary 3 can be translated into the prediction consistency property under a wide range of asymptotic regimes. For example, if we consider the scaling

$$\lambda_N \sum_{k=r+1}^m \rho_k(\boldsymbol{\Gamma}^\star) \to 0, \quad \lambda_N r \gamma_{max} \to 0, \quad \lambda_N \|\boldsymbol{\Delta}^\star\|_* \to 0, \tag{7}$$

then the prediction error vanishes asymptotically. Such scaling can be validated in many concrete examples. For example, in the random multivariate regression study in Section 2.3, with penalization level $\lambda_N \asymp \sqrt{(m_1+m_2)/n}$, a constant order for $\gamma_{\max}$, and the exact low-rank assumption, the results in (7) will hold when $n \to \infty$ and $\max\{m_1, m_2\} = o(n)$.

2.2.3 Restricted strong convexity

To control over the certain norm of the matrix estimation error $\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^\star$, we introduce the second ingredient in our analysis, viz., *restricted strong convexity* of the loss function while taking into account the possible existence of one single change-point.

Assumption 1 (Restricted strong convexity, RSC) *The restricted strong convexity condition holds with curvature $\kappa(\mathfrak{X}) > 0$ if*

$$\frac{1}{2N}\|\mathfrak{X}(\boldsymbol{M};\tau)\|_2^2 \geq \kappa(\mathfrak{X})\|\boldsymbol{M}\|_F^2, \text{ for all } \boldsymbol{M} \in \mathcal{C}(r, \delta, \boldsymbol{\Gamma}^\star, \mathbb{T}), \ \tau \in \mathbb{T}, \tag{8}$$

where for some $\delta \geq 0$,

$$\mathcal{C}(r,\delta,\boldsymbol{\Gamma}^\star,\mathbb{T}) = \Bigg\{ \boldsymbol{M} \in \mathbb{R}^{(2m_1)\times m_2} : \\ \|\boldsymbol{M}\|_F \geq \delta, \ \|\Pi_{\boldsymbol{\Gamma}^\star}^{r\perp}(\boldsymbol{M})\|_* \leq 3\|\Pi_{\boldsymbol{\Gamma}^\star}^{r}(\boldsymbol{M})\|_* + 4\sum_{k=r+1}^m \rho_k(\boldsymbol{\Gamma}^\star) + 2\|\boldsymbol{\Delta}^\star\|_F \Bigg\}. \tag{9}$$

We add some elaboration on Assumption 1. First, the present RSC condition follows the spirit of that in the context of regularized matrix estimation without any change-point (Negahban and Wainwright, 2011), to wit, in our notation, there exists some curvature constant $\kappa > 0$ such that $(2N)^{-1}\sum_{i=1}^{N}\langle \boldsymbol{X}_i\,,\,\boldsymbol{M}\rangle^2 \geq \kappa\|\boldsymbol{M}\|_F^2$, for all $\boldsymbol{M} \in \mathcal{C}(r,\delta,\boldsymbol{\Theta}^\star)$, where

$$\mathcal{C}(r,\delta,\boldsymbol{\Theta}^\star) = \left\{\boldsymbol{M} \in \mathbb{R}^{m_1\times m_2} : \|\boldsymbol{M}\|_F \geq \delta,\ \|\Pi_{\boldsymbol{\Theta}^\star}^{r\perp}(\boldsymbol{M})\|_* \leq 3\|\Pi_{\boldsymbol{\Theta}^\star}^{r}(\boldsymbol{M})\|_* + 4\sum_{k=r+1}^{m}\rho_k(\boldsymbol{\Theta}^\star)\right\}.$$

Nevertheless, due to the presence of a change-point, it demands that the curvature condition holds in a unified manner, i.e., for every possible position of the change-point $\tau \in \mathbb{T}$. This unification guarantees a local strong convexity property and eliminates the scenario where the “bad” positioning of the change-point ruins the behavior of the estimator. It’s worthy of noticing that (8) serves as an analog of the unified restricted eigenvalue condition proposed as in Assumption 2 of Lee et al. (2016), which studied the LASSO for high-dimensional linear regression with a possible change-point. Second, for the specification of the particular set where the RSC should hold, (9) has an additional term in the right-hand side of the second inequality, i.e., $2\|\boldsymbol{\Delta}^\star\|_F$, which accounts for the uncertainty of the change-point positioning as well as the change magnitude. When there’s no change, $\boldsymbol{\Delta}^\star = \mathbf{0}$ and thus (9) is reduced to the classic $\mathcal{C}(r,\delta,\boldsymbol{\Theta}^\star)$. Moreover, it is remarkable to point out that the δ in the set (9) is used to account for the term $\sum_{k=r+1}^{m}\rho_k(\boldsymbol{\Gamma}^\star)$ in the near low-rank situation. This means that for the exact low-rank cases, we can safely set $\delta = 0$. We shall show in the random multivariate regression example in Section 2.2.6 that this RSC holds with high probability (see Section 2.2.6 and Proposition 20).

2.2.4 Error rates without threshold effect

With Assumption 1 and the basic inequality (6) in Lemma 2, we can readily obtain some interesting bounds on the matrix estimation and change-point detection error. A natural question is whether the proposed scheme still behaves satisfactorily if no threshold effect exists. If one has the prior information that there’s no change, the $\boldsymbol{\Theta}^\star$ can be optimally recovered by using a direct trace norm penalized least-squares minimization. If this prior information is unavailable, it is of great interest whether the proposed scheme can adapt to such a situation. The answer is actually positive as summarized below.

Theorem 4 (Matrix estimation without threshold effect) *Assume $\boldsymbol{\Delta}^\star = \mathbf{0}$, and that Assumption 1 holds for some $\kappa(\mathfrak{X}) > 0$. If $\lambda_N \geq \sup_{\tau\in\mathbb{T}} 2\|\mathfrak{X}^\star(\boldsymbol{\epsilon};\tau)\|_{\text{op}}/N$, then*

$$\|\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^\star\|_F \leq \delta \vee \frac{6\lambda_N\sqrt{r}}{\kappa(\mathfrak{X})} \vee \left(\frac{4\lambda_N\sum_{k=r+1}^{m}\rho_k(\boldsymbol{\Gamma}^\star)}{\kappa(\mathfrak{X})}\right)^{1/2},$$

$$\left\|\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^\star\right\|_* \leq 16\sqrt{r}\delta \vee \frac{128\lambda_N r}{\kappa(\mathfrak{X})} \vee 8\sum_{k=r+1}^{m}\rho_k(\boldsymbol{\Gamma}^\star),$$

$$\frac{1}{2N}\sum_{i=1}^{N}\left\langle \boldsymbol{\mathcal{X}}_i(\widehat{\tau})\,,\,\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^\star\right\rangle^2 \leq 6\lambda_N\sqrt{r}\delta \vee \frac{36\lambda_N^2 r}{\kappa(\mathfrak{X})} \vee 4\lambda_N\sum_{k=r+1}^{m}\rho_k(\boldsymbol{\Gamma}^\star).$$

Theorem 4 gives compelling non-asymptotic bounds on the matrix estimation error (in the Frobenius and nuclear norms) and prediction error when no threshold effect or change-point exists. These bounds have a natural interpretation. Firstly the terms involving δ are admissible errors. In the exact low-rank scenarios it would no longer be necessary. The terms containing $\sum_{k=r+1}^{m} \rho_k(\mathbf{\Gamma}^\star)$ are known as *approximation errors*, which account for the expense to approximate the true matrix using a low-rank estimate. Then the remaining terms correspond to *estimation errors*, which measure the accuracy of our estimator for the low-rank approximation. In particular, comparing the Frobenius bound with the one given in Theorem 1 of Negahban and Wainwright (2011), i.e.,

$$\left\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^\star\right\|_F \le \delta \vee \frac{32\lambda_N\sqrt{r}}{\kappa} \vee \left(\frac{16\lambda_N \sum_{k=r+1}^{m} \rho_k(\mathbf{\Theta}^\star)}{\kappa}\right)^{1/2},$$

which can be regarded as a result of an "oracle" estimator with the no-change prior knowledge, we find that these two bounds coincide with each other up to some constants.

2.2.5 Error rates with threshold effect

Next, we turn to the scenario where there indeed exists a change-point in the threshold variables $\{t_i\}$ with $\mathbf{\Delta}^\star \neq \mathbf{0}$. We need the following assumption to depict the identifiability under low-rank and discontinuity of the model structure.

Assumption 2 (Identifiability and discontinuity) *Assume $\mathbf{\Gamma}^\star \in \mathbb{B}_q(R_q)$ for some $R_q > 0$ with $q \in [0, 1)$, and $\mathbf{\Delta}^\star \neq \mathbf{0}$. For a given $R_q' \geq R_q$ and some $\eta(N, m_1, m_2) > 0$, there exists some constant $c > 0$ such that for any $\tau \in \mathbb{T}$ with $|\tau - \tau^\star| > \eta(N, m_1, m_2)$ and $\mathbf{\Gamma} \in \{\mathbf{\Gamma} : \|\mathbf{\Gamma}\|_{S_q}^q \leq R_q'\}$ with $\mathbf{\Gamma} - \mathbf{\Gamma}^\star \in \mathcal{C}(r, \delta, \mathbf{\Gamma}^\star, \mathbb{T})$, it holds that*

$$\frac{1}{2N}\|\mathfrak{X}(\mathbf{\Gamma}; \tau) - \mathfrak{X}(\mathbf{\Gamma}^\star; \tau^\star)\|_2^2 > c\phi(\mathbf{\Delta}^\star)|\tau - \tau^\star|,$$

where $\phi(\mathbf{\Delta}^\star) > 0$ is some monotonically increasing function in certain norm of $\mathbf{\Delta}^\star$.

Assumption 2 implies that there is no low-rank representation that is equivalent to $\mathfrak{X}(\mathbf{\Gamma}^\star; \tau^\star)$ when the sample is split by $\tau \neq \tau^\star$. That is to say, when considering a splitting point τ located around the true change-point $\tau^\star$, the resulting prediction difference should be bounded strictly away from zero, thus rendering $\tau^\star$ identifiable. Furthermore, Assumption 2 specifies a linear growth rate in the prediction error as τ deviates from $\tau^\star$. The function $\phi(\mathbf{\Delta}^\star)$ is some curvature function that measures the effect of the change on detection ability, to wit, a change with a larger value of a certain norm of $\mathbf{\Delta}^\star$ corresponds to a higher level of detection performance. In many cases, it suffices to choose $\phi(\mathbf{\Delta}^\star) = \|\mathbf{\Delta}^\star\|_F$. One thing to note is that we only require this rate to hold for τ locating from $\tau^\star$ farther than a factor $\eta(N, m_1, m_2)$, which measures the change-point detection ability of the current scheme; more interpretation on $\eta(N, m_1, m_2)$ is provided in Remark 8. Lastly, Assumption 2 is also justifiable under the random multivariate regression example; see Section 2.2.6 as well as Proposition 22.

With Assumption 2, we can establish the following Lemma 5 to depict the consistency of the change-point detection scheme.

Lemma 5 (Change detection consistency with threshold effect) *Suppose Assumption 2 holds. If $\lambda_N \geq \sup_{\tau \in \mathbb{T}} 2\|\mathfrak{X}^\star(\boldsymbol{\epsilon}; \tau)\|_{\mathrm{op}}/N$, then $|\widehat{\tau} - \tau^\star| \leq \eta^\star$, where*

$$\eta^\star = \max\left\{\eta(N, m_1, m_2), \{c\phi(\mathbf{\Delta}^\star)\}^{-1}\left(2\lambda_N \sum_{k=r+1}^{m} \rho_k(\mathbf{\Gamma}^\star) + 6\lambda_N r\gamma_{max} + \lambda_N\|\mathbf{\Delta}^\star\|_*\right)\right\}.$$

Lemma 5 is sufficient to establish the estimation consistency of $\widehat{\tau}$ if

$$\phi(\mathbf{\Delta}^\star)^{-1}\lambda_N \sum_{k=r+1}^{m} \rho_k(\mathbf{\Gamma}^\star) \to 0,\ \phi(\mathbf{\Delta}^\star)^{-1}\lambda_N r\gamma_{max} \to 0,\ \phi(\mathbf{\Delta}^\star)^{-1}\lambda_N\|\mathbf{\Delta}^\star\|_* \to 0.$$

However, we assert here that this is not the best bound we can expect, but will serve as an initialization step in tightening the detection rate via *iteration* in further theoretical analysis. To this end, we need another assumption to guarantee a certain type of smoothness in the design.

Assumption 3 (Smoothness of design) *There exists some constant $C > 0$, such that for any $\tau \in \mathbb{T}$ with $|\tau - \tau^\star| > \eta(N, m_1, m_2)$ and for any $\mathbf{\Gamma}$ satisfying $\mathbf{\Gamma} - \mathbf{\Gamma}^\star \in \mathcal{C}(r, \delta, \mathbf{\Gamma}^\star, \mathbb{T})$, it holds that*

$$|\mathcal{T}_N(\mathbf{\Gamma}, \mathbf{\Gamma}^\star, \tau, \tau^\star)| \leq C|\tau - \tau^\star| \cdot \|\mathbf{\Gamma} - \mathbf{\Gamma}^\star\|_* \cdot \|\mathbf{\Delta}^\star\|_*,$$

where $\mathcal{T}_N(\mathbf{\Gamma}, \mathbf{\Gamma}^\star, \tau, \tau^\star) = N^{-1}\left\langle \mathfrak{X}(\mathbf{\Gamma} - \mathbf{\Gamma}^\star; \tau), \mathfrak{X}(\mathbf{\Gamma}^\star; \tau^\star) - \mathfrak{X}(\mathbf{\Gamma}^\star; \tau)\right\rangle$.

Intuitively speaking, by controlling $\mathcal{T}_N$ we are enforcing some smoothness on the threshold variables $\{t_i\}$ such that no extreme cases like point masses take place. This is suggested by the second element, $\mathfrak{X}(\mathbf{\Gamma}^\star; \tau^\star) - \mathfrak{X}(\mathbf{\Gamma}^\star; \tau)$, in the inner product we used to define $\mathcal{T}_N$, for which we wish a Lipchitz type of bound with respect to τ. Besides, through this condition, we can also control the smoothness over $\mathbf{\Gamma}$, when we consider the first element, $\mathfrak{X}(\mathbf{\Gamma} - \mathbf{\Gamma}^\star; \tau)$, in the inner product. These bounds implicitly restrict the magnitude of the design matrix $\boldsymbol{X}_i$. While mathematically complicated, this assumption is proved to be valid with high probability under certain random design circumstances; see Section 2.2.6 and Proposition 23.

Assumption 4 (Sub-Gaussian noises) *The noises ϵ_i are i.i.d. copies of a mean zero sub-Gaussian random variable ϵ, i.e., there exists some $K > 0$, such that $\mathbb{E}\{\exp\left(\epsilon^2/K^2\right)\} \leq e$.*

Starting from this assumption we begin to introduce a probabilistic structure for the noise. Now our choice of λ_N, i.e. $\lambda_N \geq \sup_{\tau \in \mathbb{T}} 2\|\mathfrak{X}^\star(\boldsymbol{\epsilon}; \tau)\|_{\mathrm{op}}/N$, becomes a random event. We will hereafter perform our analysis on this event, which bears a probability greater than $1 - \alpha_N$ for some $\alpha_N < 1$. For many concrete designs $\boldsymbol{X}_i$, either deterministic or random, it is often possible to show that α_N vanishes as $N \to \infty$, leading to a high probability guarantee for our analysis over the randomness; see, for example, Section 2.2.6.

The next lemma demonstrates a high probability control over the stochastic remainder $\mathcal{R}_N(\mathbf{\Gamma}^\star, \tau, \tau^\star)$.

Lemma 6 *Let $h_N(c_\tau) = (2c_\tau N)^{-1}\sum_{i:|t_i-\tau^\star|\le c_\tau}\langle \boldsymbol{X}_i\,,\,\boldsymbol{\Delta}^\star\rangle^2$ for some $c_\tau > 0$. Suppose Assumption 4 holds. Then, with probability greater than $1-2e\cdot\exp\left(-c'N\lambda_N^2/\{K^2\|\boldsymbol{\Delta}^\star\|_F^{-2}h_N(c_\tau)\}\right)$ for some constant $c' > 0$, we have*

$$\sup_{\tau:|\tau-\tau^\star|<c_\tau} |\mathcal{R}_N(\boldsymbol{\Gamma}^\star,\tau,\tau^\star)| \le \lambda_N\sqrt{c_\tau}\|\boldsymbol{\Delta}^\star\|_F.$$

Note the quantity $\|\boldsymbol{\Delta}^\star\|_F^{-2}h_N(c_\tau)$ in Lemma 6 is in the style of a sample mean. Under some structure conditions for $\boldsymbol{X}_i$ and $\boldsymbol{\Delta}^\star$, this term is bounded or grows rather slowly compared to $N\lambda_N^2$. For example, if we consider fixed design $\boldsymbol{X}_i$ with bounded operator norm, say $\|\boldsymbol{X}_i\|_{\rm op} \le \gamma'_{max}$ for some $\gamma'_{max} > 0$, then $\|\boldsymbol{\Delta}^\star\|_F^{-2}h_N(c_\tau) \le \operatorname{rank}(\boldsymbol{\Delta}^\star)\gamma'^2_{max}$, while in low-rank matrix recovery literature we can usually set $N\lambda_N^2 \asymp m$. Hence it results in a high probability guarantee. Similar results can be derived for large N under certain random designs, see, for example, Section 2.2.6.

Now based on Lemma 6 and the comment about the choice of λ_N, we can condition our analysis on a high-probability event where several stochastic terms of interest are well controlled. Before presenting our main result, we further impose one more technical assumption for the involved parameters.

Assumption 5 (Interplay between parameters) *The following conditions hold:*

$$\begin{gathered}
120C\{c\phi(\boldsymbol{\Delta}^\star)\}^{-1}\|\boldsymbol{\Delta}^\star\|_*\|\Pi^{r\perp}_{\boldsymbol{\Gamma}^\star}(\boldsymbol{\Gamma}^\star)\|_* < 1,\\
5\{c\phi(\boldsymbol{\Delta}^\star)\}^{-1}\|\boldsymbol{\Delta}^\star\|_F\kappa(\mathfrak{X})/16 < r,\\
\frac{1728\{c\phi(\boldsymbol{\Delta}^\star)\}^{-1}C\lambda_N r\|\boldsymbol{\Delta}^\star\|_*}{\kappa(\mathfrak{X})} < 1,\\
\frac{\{c\phi(\boldsymbol{\Delta}^\star)\}^{-2}\kappa(\mathfrak{X})\|\boldsymbol{\Delta}^\star\|_F^2}{320[1-1728\{c\phi(\boldsymbol{\Delta}^\star)\}^{-1}C\lambda_N r\|\boldsymbol{\Delta}^\star\|_*/\kappa(\mathfrak{X})]^2} < r,\\
\frac{\{c\phi(\boldsymbol{\Delta}^\star)\}^{-2}\lambda_N C\|\boldsymbol{\Delta}^\star\|_*\|\boldsymbol{\Delta}^\star\|_F^2}{96[1-1728\{c\phi(\boldsymbol{\Delta}^\star)\}^{-1}C\lambda_N r\|\boldsymbol{\Delta}^\star\|_*/\kappa(\mathfrak{X})]^2} < 1.
\end{gathered}$$

Basically Assumption 5 guarantees small magnitudes for several key quantities in our analysis, such as λ_N, $\|\Pi^{r\perp}_{\boldsymbol{\Gamma}^\star}(\boldsymbol{\Gamma}^\star)\|_*$, etc. We have commented before that proper scaling of these quantities can contribute significantly to controlling the errors of interest. These inequalities can hold simultaneously in many regimes. As one example, consider the regime where $\|\boldsymbol{\Delta}^\star\|_F$ is fixed and $\boldsymbol{\Gamma}^\star$ has an exact low rank r. This regime implies that $\|\boldsymbol{\Delta}^\star\|_* \le \sqrt{r}\|\boldsymbol{\Delta}^\star\|_F$ has the same order as $\sqrt{r}$, and $\|\Pi^{r\perp}_{\boldsymbol{\Gamma}^\star}(\boldsymbol{\Gamma}^\star)\|_* = 0$. The parameter $\kappa(\mathfrak{X})$ is also a bounded constant when evaluated in many concrete examples (such as the random design example in Section 2.2.6). The tuning parameter, λ_N, usually scales with $O(N^{-s})$ for some $s > 0$ thus converges to zero as $N \to \infty$. Therefore, we can check that Assumption 5 is satisfied under such scaling. More details are explained in Section A.2.

Theorem 7 (Recovery accuracy with threshold effect) *Suppose that Assumption 1–Assumption 5 hold. If $\lambda_N \ge \sup_{\tau\in\mathbb{T}}\frac{2}{N}\|\mathfrak{X}^\star(\boldsymbol{\epsilon};\tau)\|_{\rm op}$ holds with probability greater than $1-\alpha_N$, then there is some integer $m^\star > 0$ and a decreasing sequence $\{c_\tau^{(k)}\}_{k=1}^{m^\star}$ such that the following* *bounds hold with probability greater than* $1-\alpha_N-2e\sum_{k=1}^{m^\star}\exp\left(-c'N\lambda_N^2/\{K^2\|\mathbf{\Delta}^\star\|_F^{-2}h_N(c_\tau^{(k)})\}\right)$*:*

$$\left\|\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}^\star\right\|_F^2 \le \delta^2 \vee \frac{8\lambda_N\sum_{k=r+1}^m \rho_k(\mathbf{\Gamma}^\star)}{\kappa(\mathfrak{X})} \vee \frac{128\lambda_N^2 r}{\kappa(\mathfrak{X})^2},$$

$$\left\|\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}^\star\right\|_* \le 12\sqrt{2r}\delta \vee 12\sum_{k=r+1}^m \rho_k(\mathbf{\Gamma}^\star) \vee \frac{192\lambda_N r}{\kappa(\mathfrak{X})},$$

$$\frac{1}{2N}\left\|\mathfrak{X}\left(\widehat{\mathbf{\Gamma}};\widehat{\tau}\right)-\mathfrak{X}(\mathbf{\Gamma}^\star;\tau^\star)\right\|_2^2 \le 6\lambda_N\sqrt{2r}\delta \vee 6\lambda_N\sum_{k=r+1}^m \rho_k(\mathbf{\Gamma}^\star) \vee \frac{96\lambda_N^2 r}{\kappa(\mathfrak{X})},$$

$$|\widehat{\tau}-\tau^\star| \le 20\{c\phi(\mathbf{\Delta}^\star)\}^{-1}\lambda_N\sqrt{2r}\delta \vee 20\{c\phi(\mathbf{\Delta}^\star)\}^{-1}\lambda_N\sum_{k=r+1}^m \rho_k(\mathbf{\Gamma}^\star) \vee \frac{320\{c\phi(\mathbf{\Delta}^\star)\}^{-1}\lambda_N^2 r}{\kappa(\mathfrak{X})}.$$

Theorem 7 gives the same bounds (up to some constants) as those in Theorem 4 for the matrix estimation error as well as the prediction error in the presence of threshold effect. In addition, Theorem 7 builds the error bound for change-point detection, which generally refines that obtained in Lemma 5. To see this, consider the exact low-rank scenario where we conclude that $|\widehat{\tau}-\tau^\star| \lesssim \lambda_N^2 r$ for fixed $\kappa(\mathfrak{X})$ and $\phi(\mathbf{\Delta}^\star)$. Hence an improvement occurs by noticing that Lemma 5 gives $|\widehat{\tau}-\tau^\star| \lesssim \lambda_N r$ under such scaling. In fact, this result can be viewed as a non-asymptotic version of the super-consistency of $\widehat{\tau}$ to $\tau^\star$ for general low-rank matrix recovery in the presence of a change-point.

The most technical part of the proof of Theorem 7 is to entangle the Frobenius and nuclear norm-based estimation error bounds and the prediction error bound, as well as the change detection error bound, to push forward the tightening iteration using Theorem 18 and Theorem 19. This procedure requires more techniques due to the complexity of matrix formulation (especially that based on near-low-rank matrices).

Remark 8 *Theorem 7 is proven in an iteration scheme based on nonlinear system analysis (Vidyasagar, 2002), which accounts for the introduction of* $m^\star$ *and decreasing sequence* $\left(c_\tau^{(k)}\right)_{k=1}^{m^\star}$*. These quantities are generally dependent on* N, m_1, m_2 *as well as some model parameters. To ensure a high probability guarantee on the error bounds, it is remarkable to point out the term* $\sum_{k=1}^{m^\star}\exp\left(-c'N\lambda_N^2/\{K^2\|\mathbf{\Delta}^\star\|_F^{-2}h_N(c_\tau^{(k)})\}\right)$ *should not be too large. We consider the exact low-rank case with fixed* r *and* $\kappa(\mathfrak{X})$*. By the comment following Lemma 6,* $N\lambda_N^2/\{\|\mathbf{\Delta}^\star\|_F^{-2}h_N(c_\tau^{(k)})\}$ *generally grows linearly with* m*. Suppose the iteration is terminated at step* $m^\star+1$ *(meaning that we have the rate* $\gtrsim \lambda_N^2 r$ *at the* $m^\star$*-th iteration). Now we choose* $\eta(N,m_1,m_2) \asymp \lambda_N^2 r/\kappa(\mathfrak{X})^2$*. It can be checked that the nonlinear systems involved have a linear convergence rate, which entails the number of iterations* $m^\star \lesssim \log(\lambda_N^{-2}r^{-1})$*. In many concrete examples* $\lambda_N^{-2}r^{-1} \asymp N/m$ *(see Section 2.2.6), so that* $m^\star \lesssim \log(N/m)$*. Hence it renders a high probability result if* $m \gtrsim \log\log N$*.*

To better appreciate Theorem 7, we restate it in two concrete scenarios, namely, the exact and near low-rank matrix recovery.

Corollary 9 (Exact low-rank matrix recovery) *Suppose the conditions in Theorem 7 hold. In particular, assume* $\mathbf{\Gamma}^\star$ *is an exact low-rank matrix with rank* r *and Assumption 1* *holds with $\mathcal{C}(r, 0, \boldsymbol{\Gamma}^\star, \mathbb{T})$ and some $\kappa(\mathfrak{X}) > 0$. Then there is some integer $m^\star > 0$ and a decreasing sequence $\{c_\tau^{(k)}\}_{k=1}^{m^\star}$ such that the following bounds hold with probability greater than $1 - \alpha_N - 2e\sum_{k=1}^{m^\star} \exp\left(-c' N\lambda_N^2 / \{K^2 \|\boldsymbol{\Delta}^\star\|_F^{-2} h_N(c_\tau^{(k)})\}\right)$:*

$$\left\|\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^\star\right\|_F^2 \le \frac{128\lambda_N^2 r}{\kappa(\mathfrak{X})^2}, \quad \left\|\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^\star\right\|_* \le \frac{192\lambda_N r}{\kappa(\mathfrak{X})},$$

$$\frac{1}{2N}\left\|\mathfrak{X}\left(\widehat{\boldsymbol{\Gamma}}; \widehat{\tau}\right) - \mathfrak{X}(\boldsymbol{\Gamma}^\star; \tau^\star)\right\|_2^2 \le \frac{96\lambda_N^2 r}{\kappa(\mathfrak{X})},$$

$$|\widehat{\tau} - \tau^\star| \le \frac{320\{c\phi(\boldsymbol{\Delta}^\star)\}^{-1}\lambda_N^2 r}{\kappa(\mathfrak{X})}.$$

Corollary 10 (Near low-rank matrix recovery) *Suppose the conditions in Theorem 7 hold. In particular, assume $\boldsymbol{\Gamma}^\star \in \mathbb{B}_q(R_q)$ for some $q \in [0, 1)$ and Assumption 1 holds with $\mathcal{C}(R_q\lambda_N^{-q}, \delta, \boldsymbol{\Gamma}^\star, \mathbb{T})$ and some $\kappa(\mathfrak{X}) \in (0, 1]$. Then there is some integer $m^\star > 0$ and a decreasing sequence $\{c_\tau^{(k)}\}_{k=1}^{m^\star}$ such that the following bounds hold with probability greater than $1 - \alpha_N - 2e\sum_{k=1}^{m^\star} \exp\left(-c' N\lambda_N^2 / \{K^2 \|\boldsymbol{\Delta}^\star\|_F^{-2} h_N(c_\tau^{(k)})\}\right)$:*

$$\left\|\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^\star\right\|_F^2 \le \delta^2 \vee \frac{128\lambda_N^{2-q} R_q}{\kappa(\mathfrak{X})^{2-q}}, \quad \left\|\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^\star\right\|_* \le 12\sqrt{2R_q}\lambda_N^{-q/2}\delta \vee \frac{192 R_q \lambda_N^{1-q}}{\kappa(\mathfrak{X})^{1-q}},$$

$$\frac{1}{2N}\left\|\mathfrak{X}\left(\widehat{\boldsymbol{\Gamma}}; \widehat{\tau}\right) - \mathfrak{X}(\boldsymbol{\Gamma}^\star; \tau^\star)\right\|_2^2 \le 6\lambda_N^{1-q/2}\sqrt{2R_q}\delta \vee \frac{96\lambda_N^{2-q} R_q}{\kappa(\mathfrak{X})^{2-q}},$$

$$|\widehat{\tau} - \tau^\star| \le 20\{c\phi(\boldsymbol{\Delta}^\star)\}^{-1}\lambda_N^{1-q/2}\sqrt{2R_q}\delta \vee \frac{320\{c\phi(\boldsymbol{\Delta}^\star)\}^{-2}\lambda_N^{2-q} R_q}{\kappa(\mathfrak{X})^{2-q}}.$$

Proof of Corollary 9 is quite straightforward by noticing that $\delta = 0$ and $\|\Pi_{\boldsymbol{\Gamma}^\star}^{r\perp}(\boldsymbol{\Gamma}^\star)\|_* = 0$ under the exact low-rank assumption. The error bounds in Corollary 10 reduces to those in Corollary 9 when $q = 0$ and $\delta = 0$. The quantity $R_q\lambda_N^{-q}$ acts as the "effective rank" (Negahban and Wainwright, 2011), which is selected to achieve a trade-off between the estimation error and approximation error.

2.2.6 A random design study: multivariate regression with a possible change-point

Up to now, we are mainly investing our efforts in fixed design cases for general estimation and detection results. The assumptions we proposed have natural theoretical and practical interpretations, which serve as indispensable foundations for our main theorems. However, some of them involve complex data structure and mathematical formulation, thus raising an interesting question: whether these assumptions are realistic and verifiable in practice? In this section, we use multivariate regression to show how those assumptions can be justified with high probability. We introduce the following assumption on the random design and noise.

Assumption 6 (Random design and noise) *Suppose $\{(\boldsymbol{\epsilon}_a, \boldsymbol{x}_a, t_a)\}_{a=1}^n$ are independent random elements satisfying $t_a \sim U(0, 1)$, $\boldsymbol{x}_a$ are i.i.d. sub-gaussian random vectors with* *parameter $\overline{\sigma}^2$ and covariance spectral conditions $\underline{\sigma}^2 \leq \rho_{min}(\mathbf{\Sigma}) \leq \rho_{max}(\mathbf{\Sigma}) \leq \overline{\sigma}^2$, and $\boldsymbol{\epsilon}_a$ are i.i.d. sub-gaussian random vectors with parameter σ^2.*

Theorem 11 *Assume $\mathbf{\Gamma}^\star \in \mathbb{B}_q(R_q)$ for some $q \in [0, 1)$. If the regularization parameter λ_n is chosen such that $\lambda_n = 20\sigma\overline{\sigma}\sqrt{(m_1 + m_2)/n}$, then there are a sequence of positive constants $C, \{C_k\}_{k=0}^5$ and an integer $m^\star \asymp (1 - q/2)\log\{n/(m_1 + m_2)\}$ such that, for $n > Cm_1$, with probability at least*

$$1 - 3C_1\exp\{-C_2(m_1 + m_2)\} - C_3\exp(-C_4 n) - 2em^\star\exp\left\{-C_5\|\mathbf{\Delta}^\star\|_F^{-2}(m_1 + m_2)\right\},$$

we have

$$\begin{aligned}
\left\|\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^\star\right\|_F^2 &\leq C_0 R_q\left(\frac{\sigma\overline{\sigma}}{\underline{\sigma}^2}\right)^{2-q}\left(\frac{m_1 + m_2}{n}\right)^{(1-q/2)}, \\
\left\|\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^\star\right\|_* &\leq C_0 R_q\left(\frac{\sigma\overline{\sigma}}{\underline{\sigma}^2}\right)^{1-q}\left(\frac{m_1 + m_2}{n}\right)^{(1/2-q/2)}, \\
\frac{1}{2n}\left\|\mathfrak{X}\left(\widehat{\mathbf{\Gamma}}; \widehat{\tau}\right) - \mathfrak{X}(\mathbf{\Gamma}^\star; \tau^\star)\right\|_2^2 &\leq C_0 R_q\left(\frac{\sigma\overline{\sigma}}{\underline{\sigma}^2}\right)^{2-q}\left(\frac{m_1 + m_2}{n}\right)^{(1-q/2)}, \\
|\widehat{\tau} - \tau^\star| &\leq C_0 R_q\left(\frac{\sigma\overline{\sigma}}{\underline{\sigma}^2}\right)^{2-q}\left(\frac{m_1 + m_2}{n}\right)^{(1-q/2)}.
\end{aligned}$$

Theorem 11 establishes the non-asymptotic bounds on the matrix estimation error and prediction error for both exact and near low-rank scenarios. These bounds align perfectly with classical results in low-rank multivariate regression (Negahban and Wainwright, 2011). Besides, it also gives the change-point detection error bound, which is reduced to $r(m_1 + m_2)/n$ for the exact low-rank circumstances (i.e., $q = 0$). This rate entails the super-consistency phenomenon for change-point estimation in low-rank multivariate regression, extending the well-known results for linear regression under both low dimension (Chan, 1993) and high dimension (Lee et al., 2016); see Table 1.

Table 1: Convergence rates for change-point detection in existing literature

Model	Convergence Rate	Literature
Linear regression, low dimension	$O(1/n)$	Chan (1993)
Linear regression, high dimension	$O(s\log(p)/n)$, s: exact sparsity	Lee et al. (2016)
Reduced rank VAR	$O(rm/n)$, r: exact rank	Bai et al. (2023)
Low-rank multivariate regression	$O(R_q\{(m_1 + m_2)/n\}^{1-q/2})$, $q \in [0, 1)$	Current work

2.3 Implementation: proximal gradient descent

The implementation of the proposed method involves solving a sequence of optimization problems (3) at all feasible values of change-point $\tau \in \mathbb{T}$, each of which is composed of a smooth loss function (i.e., the least-squares loss) and a non-smooth penalty term (i.e., the nuclear norm penalty). The solution of (3) has been widely discussed in the literature and one can typically apply the proximal gradient descent method, see, for example, Nesterov (2013), Ji and Ye (2009) and Toh and Yun (2010).

To wit, for any $\mathbf{\Gamma}'$, we introduce the majorization quadratic approximation of $S_N(\mathbf{\Gamma}) := S_N(\mathbf{\Gamma}; \tau)$ at $\mathbf{\Gamma}'$, i.e.,

$$S_{\text{Major}}(\mathbf{\Gamma}; \mathbf{\Gamma}') := S_N(\mathbf{\Gamma}') + \left\langle \nabla S_N(\mathbf{\Gamma}') \, , \, \mathbf{\Gamma} - \mathbf{\Gamma}' \right\rangle + \frac{L}{2} \|\mathbf{\Gamma} - \mathbf{\Gamma}'\|_F^2 \tag{10}$$

for some $L > 0$. Then solving (3) can be done in an iterative way, where at each iteration, we update $\mathbf{\Gamma}'$ by $\mathbf{\Gamma}'' := \arg\min_{\mathbf{\Gamma}} \ \{S_{\text{Major}}(\mathbf{\Gamma}; \mathbf{\Gamma}') + \lambda \|\mathbf{\Gamma}\|_*\}$. In fact, the minimizer $\mathbf{\Gamma}''$ can be expressed using the singular value soft-thresholding operator (Toh and Yun, 2010), namely, $\mathbf{\Gamma}'' = \text{Soft}\left(\mathbf{\Gamma}' - L^{-1} \nabla S_N(\mathbf{\Gamma}'); L^{-1}\lambda\right)$, where for any matrix $\boldsymbol{G}$ with singular value decomposition $\boldsymbol{G} = \mathbf{U}_{\boldsymbol{G}}^\top \text{diag}\{(\rho_i(\boldsymbol{G}))\} \mathbf{V}_{\boldsymbol{G}}$,

$$\text{Soft}(\boldsymbol{G}; \xi) = \mathbf{U}_{\boldsymbol{G}}^\top \, \text{diag}\{((\varrho_i(\boldsymbol{G}) - \xi)_+)\} \mathbf{V}_{\boldsymbol{G}} \tag{11}$$

$x_+ = \max\{x, 0\}$.

Remark 12 (Algorithmic convergence) *In the current manuscript, we are focusing on the statistical properties of the estimators obtained from the joint minimization program. In general cases, the algorithmic behavior of the proximal gradient descent is not covered by the current manuscript. Nevertheless, in Proposition 32, we provide rigorous algorithmic justification for the random multivariate regression example studied in Section 2.2.6. Concretely, Proposition 32 states that, for the random design setting, for each given grid point* τ_k*, it takes* $O(\log \zeta^{-2})$ *steps to reach the tolerance* ζ^2*, which demonstrates a fast exponential convergence to the global minimum. When* ζ^2 *is set to be properly small, the detected change-point coincides with the global optimum too. See Section D.1 for a more detailed exposition.*

In practice, the regularization parameter λ_N is chosen through cross-validation. See Section D.2 for more details.

3. Extension to multiple change-points scenario

In this section, we extend the proposed procedure to the scenario with multiple change-points, to wit, $y_i = \langle \boldsymbol{X}_i \, , \, \mathbf{\Theta}_i \rangle + \epsilon_i$, where

$$\mathbf{\Theta}_i = \mathbf{\Theta}_s^\star, \ \tau_s^\star < t_i \le \tau_{s+1}^\star, \ s = 0, \dots, s^\star; \ i = 1, \dots, N. \tag{12}$$

Of interest is to simultaneously recover the low-rank matrices $\mathbf{\Theta}_s^\star$'s and change-points $\tau_s^\star$'s (with the convention of $\tau_0^\star = 0$ and $\tau_{s^\star+1}^\star = 1$), together with the number of change-points $s^\star$, from the response-covariates-threshold triple observations $\{(y_i, \boldsymbol{X}_i, t_i)\}_{i=1}^N$.

To handle multiple change-points, we shall first find some rough estimators of change-points, and then refine them to deliver a desirable error rate. The spirit of refinements over inefficient or sub-optimal initial change-point estimators is popular in the literature of multiple change-point detection (Harchaoui and Lévy-Leduc, 2010; Zou et al., 2014), and has been further explored for high-dimensional change detection, see, for example, Wang et al. (2021a) and Bai et al. (2023). However, to obtain consistent and (near) rate-optimal estimators of the regression matrices, existing methods typically need the removal of the detected change-points together with large enough neighborhoods (Safikhani and Shojaie, 2022; Safikhani et al., 2022; Bai et al., 2023). In other words, change detection and parameter estimation are performed separately, which may be inefficient in practice.

3.1 Preview of the results

The primary results in this section are two-fold:

- *Algorithm development*: we attempt to fulfill the refinements of both the change-point and regression matrix estimators in a joint manner in the proposed Algorithm 1. In the first stage, we obtain some initial change-point estimators based on a sequence of maximally selected change differences in conjunction with a novel thresholding rule, which are built on the consistency results on estimated low-rank matrices as developed in Section 2. It does not necessarily produce consistent change-point estimators (in their locations) but should identify the correct number of change-points with high probability. In the second stage, we suggest a joint refinement procedure for both change-point and regression matrix estimators with desirable error bounds by recasting the original problem into a sequence of sub-problems each with a single change-point, thus making the proposed joint minimization scheme in Section 2.1 applicable.

- *Theoretical investigation*: Theorem 14 provides theoretical justification of the proposed Algorithm 1. It establishes the estimation error bounds for matrix recovery and change-point detection:

  $$\|\widehat{\mathbf{\Gamma}}_s - \mathbf{\Gamma}_s^\star\|_F^2 \lesssim \underline{\lambda}_N^2 r, \qquad |\widehat{\tau}_s - \tau_s^\star| \lesssim \underline{\lambda}_N^2 r, \quad s = 1, \ldots, s^\star.$$

  The error rates for matrix recovery again match the oracle rates, and the bound for change-point detection demonstrates the super consistency phenomenon as in the single change-point setting.

  Corollary 15 further applies Theorem 14 to the multivariate regression setting and establishes the following error rates:

  $$\|\widehat{\mathbf{\Gamma}}_s - \mathbf{\Gamma}_s^\star\|_F^2 \lesssim \frac{r(m_1+m_2)}{n}, \quad |\widehat{\tau}_s - \tau_s^\star| \lesssim \frac{r(m_1+m_2)}{n}, \quad s = 1, \ldots, s^\star.$$

3.2 An algorithm for joint multiple change-point detection and matrix estimation

Algorithm 1 previews the two-stage procedure for joint multiple change-point detection and matrix estimation. Stage I is composed of two steps, by which we shall find $\widetilde{s}$ initial change-point estimators, i.e., $\widetilde{\tau}_1, \ldots, \widetilde{\tau}_{\widetilde{s}}$. In Step (i), it collects a set of rough change-point estimators by using a moving-window strategy. Each window $\mathcal{T}_i = [t_i - \omega, t_i + \omega]$ is of length 2ω. If ω is selected not too large, we can expect that there is at most one change-point occurring in $\mathcal{T}_i$. Hence we can apply the joint minimization scheme proposed in Section 2.1 to the data set corresponding to threshold variables in $\mathcal{T}_i$. The resulting estimator of the

Algorithm 1: Joint multiple change-point detection and matrix estimation

```
Input: Response-covariates-threshold triple observations $\mathcal{D} := \{(y_i, \boldsymbol{X}_i, t_i)\}_{i=1}^{N}$,
       moving-window parameter $0 < \omega < 1$, regularization parameter $\lambda_N > 0$ and
       stopping threshold $\zeta_N > 0$
Output: Estimated change-points $\{\widehat{\tau}_s\}_{s=1}^{\widetilde{s}}$ and the associated low-rank matrices
        $\{\widehat{\boldsymbol{\Theta}}_s\}_{s=1}^{\widetilde{s}}$
/* Stage I: Rough change-point estimators */
/*   Step (i): Change-point indicators */
Set the searching grid $\mathcal{G} = \{t_i\}_{i=1}^{G} \cap [\omega, 1-\omega]$ such that
    $$G = \lceil \frac{1-2\omega}{\omega/2} \rceil;\ t_0 = \omega,\ t_G = 1-\omega;\ t_i = \omega + \frac{\omega}{2} \cdot i,\ \text{for } i = 1, \ldots, G.$$
for $t_i \in \mathcal{G}$ do
    (1) Set $\mathcal{T}_i := [t_i - \omega, t_i + \omega]$ and $\mathcal{D}_{\mathcal{T}_i} = \{(y_j, \boldsymbol{X}_j, t_j) \in \mathcal{D} : t_j \in \mathcal{T}_i\}$
    (2) Apply the joint minimization scheme in Section 2.1 to $\mathcal{D}_{\mathcal{T}_i}$ with the
        regularization parameter $\lambda_{2\omega N}$
    (3) Record the resulting estimator of the change magnitude by $\widehat{\boldsymbol{\Delta}}_i$
/*   Step (ii): Sequential maximizers */
Set $s = 1$ and $\widetilde{\tau}_1 := \arg\max_{t_i \in \mathcal{G}} \|\widehat{\boldsymbol{\Delta}}_i\|_F$
while $\|\widehat{\boldsymbol{\Delta}}_{\widetilde{\tau}_s}\|_F > \zeta_N$ do
    $s \leftarrow s + 1$
    $\widetilde{\tau}_s := \arg\max_{t_i \in \mathcal{G} \setminus \cup_{j=1}^{s-1} [\widetilde{\tau}_j - 2\omega, \widetilde{\tau}_j + 2\omega]} \|\widehat{\boldsymbol{\Delta}}_i\|_F$
Record the resulting change-points until stopping as $\{\widetilde{\tau}_s\}_{s=1}^{\widetilde{s}}$
/* Stage II: Local refinements */
for $s = 1, \ldots, \widetilde{s}$ do
    (1) Set $\mathcal{I}_s = [(\widetilde{\tau}_{s-1} + \widetilde{\tau}_s)/2, (\widetilde{\tau}_s + \widetilde{\tau}_{s+1})/2]$
    (2) Apply the joint minimization scheme in Section 2.1 to $\{(y_j, \boldsymbol{X}_j, t_j) : t_j \in \mathcal{I}_s\}$
        with the regularization parameter $\lambda_{|\mathcal{I}_s|N}$
    (3) Record the detected change-point as $\widehat{\tau}_s$ and the estimated low-rank matrices
        as $\widehat{\boldsymbol{\Theta}}_s$
```

change magnitude is denoted by $\widehat{\mathbf{\Delta}}_i$. According to Theorem 4, if $\mathcal{T}_i$ contains no change-point, then $\widehat{\mathbf{\Delta}}_i$ would in general be small in either the Frobenius or nuclear norm. On the other hand, by Theorem 7, a large value of $\widehat{\mathbf{\Delta}}_i$ may indicate that t_i is located around some change-point, provided that the change signal is not too weak. Hence $\widehat{\mathbf{\Delta}}_i$ serves as a very good indicator of whether there exists certain change. To fix ideas, here we adopt $\|\widehat{\mathbf{\Delta}}_i\|_F$. However, we cannot select all t_i's corresponding to large values in $\|\widehat{\mathbf{\Delta}}_i\|_F$'s, which could generally result in redundant change-point estimates; that is why Step (ii) comes in. In Step (ii), we propose searching for a sequence of maximizers in conjunction with a thresholding rule to avoid overestimation. It is obvious that $\widetilde{\tau}_1 = \arg\max_{t_i \in \mathcal{G}} \|\widehat{\mathbf{\Delta}}_i\|_F$ can be set as the most "significant" change-point. Upon the determination of the first $s-1$ ($s \geq 2$) change-point candidates, we identify the next one as

$$\widetilde{\tau}_s = \arg\max_{t_i \in \mathcal{G} \setminus \cup_{j=1}^{s-1}[\widetilde{\tau}_j - 2\omega, \widetilde{\tau}_j + 2\omega]} \|\widehat{\mathbf{\Delta}}_i\|_F,$$

where in each step some neighborhoods (of length 4ω) of previously detected change-points have been removed to screen out redundant change-points. This is essentially a "forward" detection procedure, and similar to the binary segmentation algorithm in the change-point literature. To consistently recover the number of change-points, after each recursive, we stop if $\|\widehat{\mathbf{\Delta}}_{\widetilde{\tau}_s}\|_F < \zeta_N$ for some threshold ζ_N that will be specified later.

In Stage II, we perform local refinements over the change-points $\{\widetilde{\tau}_s\}_{s=1}^{\widetilde{s}}$ detected previously. For this purpose, let $\mathcal{I}_s = [(\widetilde{\tau}_{s-1} + \widetilde{\tau}_s)/2, (\widetilde{\tau}_s + \widetilde{\tau}_{s+1})/2]$ for $s = 1, \ldots, \widetilde{s}$, with the convention of $\widetilde{\tau}_0 = 0$ and $\widetilde{\tau}_{\widetilde{s}+1} = 1$. Then, for each s, we again apply the joint minimization scheme (cf. Section 2.1) to the data set corresponding to threshold variables in $\mathcal{I}_s$. The proposed refinement scheme simultaneously results in a new change-point estimator (i.e., $\widehat{\tau}_s$) and an estimator of the associated low-rank matrices (i.e., $\widehat{\mathbf{\Theta}}_s$), for $s = 1, \ldots, \widetilde{s}$.

Remark 13 (Computational cost of Algorithm 1) *Algorithm 1 requires $O(N)$ times to solve the nuclear-norm penalized least-squares minimization, even when multiple change-points are present. In Step (i) in Stage I, each moving-window $\mathcal{T}_i$ is of length 2ω, thus requiring solving $O(2\omega N)$ rounds of minimization. In total there are $O(\omega^{-1})$ running windows. Hence the total number of minimization programs becomes $O(N)$. Step (ii) requires searching for the maximums of a sequence of values, which does not involve solving the optimization program. Similarly, Stage II amounts to $O(N)$ rounds of minimization.*

3.3 Theoretical inverstigation of Algorithm 1

To facilitate theoretical analysis, we confine attention to the exact low-rank circumstances. Let $d_{min} = \min_{s=1,\ldots,s^\star+1}\{\tau_s^\star - \tau_{s-1}^\star\}$ be the minimal distance between two consecutive change-points, and $\Delta_{min} = \min_{s=1,\ldots,s^\star} \|\mathbf{\Delta}_s^\star\|_F^2$ and $\Delta_{max} = \max_{s=1,\ldots,s^\star} \|\mathbf{\Delta}_s^\star\|_F^2$ be the minimal and maximal change magnitude in the Frobenius norm, respectively. We define an event

$$\mathcal{E}_N := \{\widetilde{s} = s^\star \text{ and } \max_{s=1,\ldots,\widetilde{s}} |\widetilde{\tau}_s - \tau_s^\star| \leq d_{min}/6\}. \tag{13}$$

By the construction of our procedure, it can be shown that, on $\mathcal{E}_N$, $|\mathcal{I}_s| \geq 2d_{min}/3$.

Theorem 14 *Suppose Assumption 7–Assumption 12 in Appendix (parallel to those in Corollary 9) hold. Assume there exists some* $\underline{\lambda}_N > 0$ *such that* $\lambda_{2\omega N} = (2\omega)^{-1/2}\underline{\lambda}_N$, $\lambda_{|\mathcal{I}_s|N} \leq (2d_{min}/3)^{-1/2}\underline{\lambda}_N$, *and*

$$\underline{\lambda}_N \geq \sup_{0<t_{(i)}<t_{(j)}<1} \sup_{\tau\in[t_{(i)},t_{(j)}]} \frac{2}{N(t_{(j)}-t_{(i)})} \| \sum_{k:t_k\in[t_{(i)},t_{(j)}]} \epsilon_k \boldsymbol{\mathcal{X}}_k(\tau)\|_{\text{op}}$$

holds with probability greater than $1-\alpha_N$ *for some* $\alpha_N > 0$. *If the threshold* ζ_N *is selected such that* $\zeta_N = C'\underline{\lambda}_N^2 r/\kappa(\mathfrak{X})^2$ *for large enough* $C' > 0$ *and the minimal change magnitude* $\Delta_{min} > \zeta_N$, *then*

(i) the event $\mathcal{E}_N$ *holds with probability greater than* $1-\alpha_N-2em^\star N^2\exp\{-cN\lambda_N^2/(K^2\Delta_{max})\}$ *for some constant* $c > 0$ *and* $m^\star > 0$*;*

(ii) there exist some constants $C_1, C_2 > 0$ *such that*

$$\left\|\widehat{\boldsymbol{\Gamma}}_s - \boldsymbol{\Gamma}_s^\star\right\|_F^2 \leq \frac{C_1\underline{\lambda}_N^2 r}{\kappa(\mathfrak{X})^2},\ |\widehat{\tau}_s - \tau_s^\star| \leq \frac{C_2\{\phi(\boldsymbol{\Delta}^\star)\}^{-1}\underline{\lambda}_N^2 r}{\kappa(\mathfrak{X})}$$

hold uniformly for $s = 1, \ldots, \widetilde{s}$ *with probability greater than* $1-\alpha_N-2e\tilde{m}^\star N^2\exp\{-\tilde{c}N\lambda_N^2/(K^2\Delta_{max})\}$ *for some constant* $\tilde{c} > 0$ *and* $\tilde{m}^\star > 0$.

Theorem 14 implies that the estimation rate $d_{\min}/6$ for event $\mathcal{E}_N$ is the guarantee with high probability for Stage I of Algorithm 1, which we do not require to decrease as $N \to \infty$. The goal of Stage I is to construct a sequence of non-overlapping windows, each of which covers a true change-point. Then in Stage II, Algorithm 1 performs local refinement based on the sequence of intervals $\{\mathcal{I}_s\}_{s\in[\tilde{s}]}$ and gives the estimation rates that decrease with the sample size N. Further, if we consider the multivariate regression model with multiple change points and a data generating process under Assumption 6, we can prove the following result:

Corollary 15 *If the regularization parameter* $\underline{\lambda}_n$ *is chosen such that* $\underline{\lambda}_n = C\sigma\sigma_0\sqrt{(m_1+m_2)/n}$ *for some* $C > 0$, *then there are a sequence of positive constants* $\{C_k\}_{k=0}^7$ *and an integer* $m^\star \asymp (1-q/2)\log\{n/(m_1+m_2)\}$ *such that, for* $n > C_0 m_1$, *with probability at least*

$$1-3C_1n^2\exp\{-C_2(m_1+m_2)\}-C_3n^2\exp(-C_4n)-2C_5m^\star n^2\exp\left\{-C_6\|\boldsymbol{\Delta}^\star\|_F^{-2}(m_1+m_2)\right\},$$

we have

$$\left\|\widehat{\boldsymbol{\Gamma}}_s - \boldsymbol{\Gamma}_s^\star\right\|_F^2 \leq C_7\frac{r(m_1+m_2)}{n} \text{ and } |\widehat{\tau}_s - \tau_s^\star| \leq C_8\frac{r(m_1+m_2)}{n}.$$

Remark 16 (Alternative choices in Stage I) *The thresholding rule-based procedure provides a consistent selection of the number of change-points by exploiting the low-rank structure of the underlying regression matrices. Other choices that ensure a high probability result for the event* $\mathcal{E}_N$ *in (13) are also possible. For example, we may consider a score method by transferring the target problem into high-dimensional mean change detection, upon which state-of-the-art mean change detection methods (Cho and Fryzlewicz, 2015; Wang and Samworth, 2018; Wang et al., 2019b; Yu and Chen, 2021) can be leveraged to obtain initial* *change-point estimators. Let $\{\boldsymbol{Z}_i\}_{i=1}^N$ be the scores such that detecting changes in $\boldsymbol{\Theta}_i$'s can be framed into detecting changes in $\mathbb{E}(\boldsymbol{Z}_i)$'s. In some scenarios such as compressed sensing or phase retrieval, the scores can be directly set as $\boldsymbol{Z}_i = y_i \mathrm{vec}(\boldsymbol{X}_i)$ if $\boldsymbol{X}_i$'s are i.i.d.. To see this, we observe that $\mathbb{E}(\boldsymbol{Z}_i) = \Xi \mathrm{vec}(\boldsymbol{\Theta}_i)$, where $\Xi = \mathbb{E}\left\{\mathrm{vec}(\boldsymbol{X}_i)\mathrm{vec}(\boldsymbol{X}_i)^\top\right\}$. In certain cases $\boldsymbol{X}_i$'s are not i.i.d.; for example, in multivariate regression (cf. Example 1), $\boldsymbol{X}_i = \boldsymbol{x}_a \boldsymbol{e}_b^\top$, whose distribution varies with the position $b \in \{1, \ldots, m_2\}$. Fortunately, we can directly deal with $\{(\boldsymbol{y}_a, \boldsymbol{x}_a)\}_{a=1}^n$ and define scores as $\boldsymbol{Z}_a = \mathrm{vec}(\boldsymbol{x}_a \boldsymbol{y}_a^\top)$ for $a = 1, \ldots, n$. Observe that $\mathbb{E}(\boldsymbol{Z}_a) = \boldsymbol{\Xi}\mathrm{vec}(\boldsymbol{\Theta}_a)$ where $\boldsymbol{\Xi} = \boldsymbol{I}_{m_2} \otimes \mathbb{E}(\boldsymbol{X}_a \boldsymbol{X}_a^\top)$. As a consequence, detecting changes in $\boldsymbol{\Theta}_a$'s amounts to detecting changes in $\mathbb{E}(\boldsymbol{Z}_a)$'s. Let $\mathcal{A}$ be a prescribed mean change detection algorithm which will be applied to $\{\boldsymbol{Z}_i\}_{i=1}^N$. The output of $\mathcal{A}(\{\boldsymbol{Z}_i\}_{i=1}^N)$ can be used as the initializers in Stage I. However, existing theories could not be directly applied to provide a high-probability guarantee over $\mathcal{E}_N$, since the underlying covariance matrix of $\boldsymbol{Z}_i$ also shifts. It is of independent interest to study the high-dimensional mean change detection problem in the presence of heterogeneous covariances.*

Remark 17 (Number of change-points) *In Theorem 14 and Corollary 15, we assume a fixed number of change-points $s^\star$ to ease the presentation of theoretical results. However, we highlight that this is not required to prove the estimation consistency of the change-points and low-rank matrix signals. Take Corollary 15 as an example. When the number of change-points $s^\star$ grows, the moving window detection approach adopted in Algorithm 1 Stage I will give a sequence of data segments $\{\mathcal{I}_s\}_{s\in[s^\star]}$ around each true change-point, with high probability. Each $\mathcal{I}_s$ has a size of order $O(n/s^\star)$, which results in the following local estimation rates with high probability:*

$$\left\|\widehat{\boldsymbol{\Gamma}}_s - \boldsymbol{\Gamma}_s^\star\right\|_F^2 \leq C_7 \frac{s^\star r(m_1+m_2)}{n} \text{ and } |\widehat{\tau}_s - \tau_s^\star| \leq C_8 \frac{s^\star r(m_1+m_2)}{n}. \tag{14}$$

To guarantee consistency, (14) *requires $s^\star r(m_1+m_2)/n \to 0$. This suggests a trade-off among sample size n, number of change-points $s^\star$, rank r and dimension m_1+m_2. Asymptotically, if $s^\star \asymp n^a$, $r \asymp n^b$, $m_1+m_2 \asymp n^c$, where $a, b \geq 0$ and $c > 0$, then* (14) *requires $a+b+c<1$ for consistency.*

4. Numerical study

In this section, we run several synthetic experiments to show the validity and effectiveness of the proposed scheme in change-point detection as well as low-rank matrix recovery. A real-data example is also investigated, which reveals the benefit of incorporating structural changes for matrix estimation. The algorithm is implemented in `MATLAB` and the source code can be accessed through the public `GitHub` repository: `https://github.com/LeiShi-rocks/LowRank_ChangePoints`.

4.1 Single change-point scenario

We consider two simulation settings for low-rank matrix recovery with a single change-point, i.e., multivariate regression (Example 1) and compressed sensing (Example 2).

4.1.1 multivariate regression

The true change-point is set as $\tau^\star = 0.5$. The matrix signals are square matrices with rank $r = 5$. In Example 1, the thresholding variables are simply taken as $\boldsymbol{x}_a = a/n$, the covariates are generated independently from a multivariate standard Gaussian distribution $N_m(\mathbf{0}, \boldsymbol{I}_m)$, and the noises are i.i.d. copies from $N_m(\mathbf{0}, 0.1^2\boldsymbol{I}_m)$. We vary the configuration of several synthetic parameters to present a comprehensive numeric study. More concretely, we focus the following settings respectively: (i) the dimension is fixed as $m_1 = m_2 = m = 50$ and the sample size n ranges over $\{500, 1000, 2000\}$; (ii) the dimension $m_1 = m_2 = m$ takes values in $m \in \{50, 75, 100\}$ while the sample size scales with the dimension, i.e., $n = 5mr$. The true signals are generated from the singular vectors of standard Gaussian ensembles (see Section D.3 for more details) with $\|\boldsymbol{\Theta}_1^\star\|_F = \|\boldsymbol{\Theta}_2^\star\|_F = 1$ and a break $\|\boldsymbol{\Theta}_1^\star - \boldsymbol{\Theta}_2^\star\|_F = 0.1$. We introduce some benchmark procedures. The first one is to directly perform matrix estimation by ignoring the change-point (NC, for no-change). The second is to run matrix estimation with the known of the true change-point (Oracle). The third is first to vectorize each matrix covariate and then to apply the LASSO-based change detection method proposed by Lee et al. (2016) (Vec). The following criteria are reported, i.e., distance of the estimated change-point and the truth, estimation error of the low-rank matrices in both Frobenius norm and nuclear norm and estimated rank. Results over 100 replications are summarized in Table 2.

For change-point detection, our method is more accurate and more stable than the Vec based detection method in all experiments. In terms of matrix recovery, it achieves high accuracy in both Frobenius and nuclear norms and performs comparably well as the Oracle. On the contrary, the Vec behaves poorly since it distorts the low-rank structure. Note that in this setting the NC gives more accurate matrix estimation results, which is due to the fact that $\boldsymbol{\Theta}_1^\star$ and $\boldsymbol{\Theta}_2^\star$ share the same first four singular vectors and demonstrate a small break size (see Section D.3). In Appendix we also presented results under a relatively large break situation where the NC method becomes inferior. Besides, our method also demonstrates a satisfactory result on rank recovery.

4.1.2 compressed sensing

Similar to last setting, we set $\tau^\star = 0.5$ and the true signals are square matrices with $r = 5$. We consider two different specifications of the sample size and dimension, i.e., $m = 40$, $N \in \{1500, 2000, 2500\}$) and $m \in \{20, 35, 50\}\}, N = 10mr$. The covariates are generated independently from standard Gaussian ensembles and the noise are i.i.d. Gaussian variables from $N(0, 0.1^2)$. Results over 100 replications are summarized in Table 3. Similar to the multivariate regression setting, our method demonstrates high accuracy in both change-point detection and matrix recovery in a wide range of settings.

4.2 Multiple change-points scenario

In this section, we present the numerical results of matrix estimation with multiple change-points under the multivariate regression setting. We set $m_1 = m_2 = m = 40$ and $r = 5$. Then we generate $n = 2000$ independent covariates from $N_m(0, \boldsymbol{I}_m)$ and i.i.d. noise from $N_m(\mathbf{0}, 0.1^2\boldsymbol{I}_m)$. Three change-points are introduced, i.e., $\tau_1^\star = 0.25, \tau_2^\star = 0.50$ and $\tau_3^\star = 0.75$. For change-point detection, we report the number of estimated change-points as well

Table 2: Multivariate regression with a single change-point

Method	$\lvert\widehat{\tau}-\tau^{\star}\rvert$	$\widehat{\boldsymbol{\Theta}}_1$			$\widehat{\boldsymbol{\Theta}}_2$		
		$\lVert\widehat{\boldsymbol{\Theta}}_1-\boldsymbol{\Theta}_1^{\star}\rVert_F^2$	$\lVert\widehat{\boldsymbol{\Theta}}_1-\boldsymbol{\Theta}_1^{\star}\rVert_*$	rank	$\lVert\widehat{\boldsymbol{\Theta}}_2-\boldsymbol{\Theta}_2^{\star}\rVert_F^2$	$\lVert\widehat{\boldsymbol{\Theta}}_2-\boldsymbol{\Theta}_2^{\star}\rVert_*$	rank
Regime: Varying n with $(m,n)=(50,500)$							
Ours	0.031(0.028)	0.352(0.036)	1.497(0.068)	5.40(0.55)	0.347(0.030)	1.484(0.060)	5.37(0.53)
Oracle	-	0.347(0.027)	1.484(0.048)	5.33(0.49)	0.346(0.024)	1.479(0.044)	5.29(0.50)
NC	-	0.225(0.014)	1.201(0.033)	6.03(4.47)	0.225(0.013)	1.198(0.032)	6.03(4.47)
Vec	0.040(0.033)	0.899(0.102)	5.581(0.307)	50.00(0)	0.939(0.106)	5.706(0.314)	50.00(0)
Regime: Varying n with $(m,n)=(50,1000)$							
Ours	0.017(0.016)	0.206(0.017)	1.146(0.043)	5.13(0.39)	0.202(0.016)	1.134(0.038)	5.05(0.22)
Oracle	-	0.203(0.014)	1.138(0.035)	5.10(0.33)	0.202(0.014)	1.134(0.033)	5.03(0.17)
NC	-	0.135(0.007)	0.930(0.024)	5.88(0.46)	0.135(0.007)	0.929(0.024)	5.88(0.46)
Vec	0.025(0.025)	0.451(0.035)	3.981(0.155)	50.00(0)	0.454(0.037)	3.996(0.159)	50.00(0)
Regime: Varying n with $(m,n)=(50,2000)$							
Ours	0.006(0.006)	0.107(0.007)	0.831(0.025)	5.00(0)	0.108(0.007)	0.838(0.025)	5.00(0)
Oracle	-	0.107(0.007)	0.831(0.024)	5.00(0)	0.108(0.007)	0.836(0.026)	5.00(0)
NC	-	0.084(0.004)	0.732(0.019)	5.99(0.30)	0.085(0.004)	0.734(0.018)	5.99(0.30)
Vec	0.010(0.010)	0.229(0.011)	2.847(0.071)	50.00(0)	0.228(0.009)	2.842(0.059)	50.00(0)
Regime: Varying m with $(m,n)=(25,625)$							
Ours	0.024(0.022)	0.233(0.032)	3.317(0.240)	5.00(0)	0.233(0.026)	3.335(0.081)	5.00(0)
Oracle	-	0.214(0.022)	3.359(0.239)	5.00(0)	0.218(0.019)	3.366(0.069)	5.00(0)
NC	-	0.662(0.040)	3.047(0.107)	8.16(0.55)	0.670(0.036)	3.050(0.095)	8.16(0.55)
Vec	0.028(0.027)	0.256(0.022)	5.042(0.336)	25.00(0)	0.257(0.025)	5.092(0.145)	25.00(0)
Regime: Varying m with $(m,n)=(50,1250)$							
Ours	0.016(0.018)	0.224(0.019)	3.460(0.051)	5.00(0)	0.224(0.024)	3.464(0.072)	5.00(0)
Oracle	-	0.213(0.014)	3.486(0.042)	5.00(0)	0.213(0.016)	3.491(0.056)	5.00(0)
NC	-	0.668(0.021)	3.225(0.048)	9.45(0.50)	0.666(0.024)	3.225(0.049)	9.45(0.50)
Vec	0.022(0.022)	0.457(0.026)	7.022(0.170)	50.00(0)	0.457(0.026)	7.022(0.162)	50.00(0)
Regime: Varying m with $(m,n)=(75,1875)$							
Ours	0.014(0.014)	0.226(0.022)	3.486(0.056)	5.00(0)	0.226(0.023)	3.484(0.060)	5.00(0)
Oracle	-	0.213(0.013)	3.519(0.036)	5.00(0)	0.213(0.013)	3.514(0.037)	5.00(0)
NC	-	0.667(0.017)	3.306(0.034)	9.99(0.17)	0.665(0.018)	3.304(0.035)	9.99(0.17)
Vec	0.019(0.018)	0.642(0.023)	8.815(0.183)	75.00(0)	0.655(0.035)	8.887(0.170)	75.00(0)

as the accuracy of detection, measured by the following two criteria

$$\mathrm{OE}=\sup_{s=1,\cdots,s^{\star}}\ \inf_{s'=1,\cdots,\widehat{s}}|\widehat{\tau}_{s'}-\tau_s^{\star}|,\ \ \mathrm{UE}=\sup_{s'=1,\cdots,\widehat{s}}\ \inf_{s=1,\cdots,s^{\star}}|\widehat{\tau}_{s'}-\tau_s^{\star}|.$$

This pair of quantities measures the over- and under-segmentation errors, respectively, for which a desirable estimator should strike a balance. For matrix recovery, we introduce analogous concepts to measure the estimation error, i.e.,

$$\mathrm{MOE}=\sup_{s=1,\cdots,s^{\star}}\ \inf_{s'=1,\cdots,\widehat{s}}\lVert\widehat{\boldsymbol{\Theta}}_{s'}-\boldsymbol{\Theta}_s^{\star}\rVert_F^2,\ \ \mathrm{MUE}=\sup_{s'=1,\cdots,\widehat{s}}\ \inf_{s=1,\cdots,s^{\star}}\lVert\widehat{\boldsymbol{\Theta}}_{s'}-\boldsymbol{\Theta}_s^{\star}\rVert_F^2.$$

Besides, we report the maximal and minimal estimated rank across segments. Results over 100 replications are summarized in Table 4 and Figure 1.

When the magnitude of the change signal is small, detection and estimation are in general harder. Nevertheless, our method can recover the number and location of change-points with high accuracy. Besides, we can see that the refinement step plays an indispensable role in augmenting and stabilizing the performance of the roughly selected change-points.

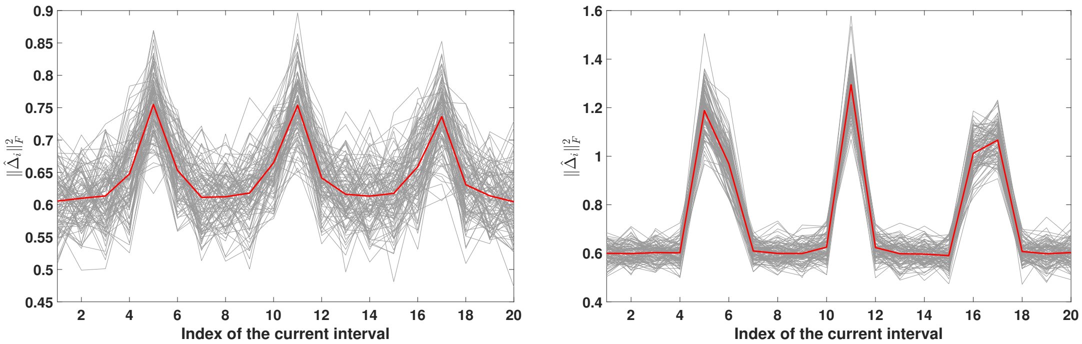
Table 3: Compressed sensing with a single change-point

Method	$\lvert\widehat{\tau}-\tau^\star\rvert$	$\mathbf{\Theta}_1$			$\mathbf{\Theta}_2$		
		$\lVert\widehat{\mathbf{\Theta}}_1-\mathbf{\Theta}_1^\star\rVert_F^2$	$\lVert\widehat{\mathbf{\Theta}}_1-\mathbf{\Theta}_1^\star\rVert_*$	rank	$\lVert\widehat{\mathbf{\Theta}}_2-\mathbf{\Theta}_2^\star\rVert_F^2$	$\lVert\widehat{\mathbf{\Theta}}_2-\mathbf{\Theta}_2^\star\rVert_*$	rank
Regime: Varying N with $(m,N)=(40,1500)$							
Ours	0.007(0.003)	0.246(0.029)	1.318(0.092)	5.41(1.06)	0.255(0.029)	1.358(0.104)	5.83(1.55)
Oracle	-	0.240(0.027)	1.298(0.083)	5.33(0.92)	0.239(0.022)	1.295(0.068)	5.23(0.85)
NC	-	0.798(0.041)	3.125(0.086)	17.52(0.73)	0.797(0.042)	3.124(0.099)	17.52(0.73)
Vec	0.103(0.085)	0.937(0.101)	4.696(0.151)	40.00(0)	1.074(0.211)	5.267(0.578)	40.00(0)
Regime: Varying N with $(m,N)=(40,2000)$							
Ours	0.006(0)	0.161(0.015)	1.050(0.049)	5.00(0)	0.165(0.016)	1.065(0.049)	5.00(0)
Oracle	-	0.157(0.015)	1.038(0.047)	5.00(0)	0.159(0.014)	1.042(0.045)	5.00(0)
NC	-	0.730(0.043)	3.031(0.092)	19.07(0.77)	0.744(0.031)	3.060(0.075)	19.07(0.77)
Vec	0.020(0.029)	0.677(0.055)	4.210(0.152)	40.00(0)	0.720(0.119)	4.366(0.368)	40.00(0)
Regime: Varying N with $(m,N)=(40,2500)$							
Ours	0.006(0)	0.120(0.010)	0.902(0.037)	5.00(0)	0.123(0.011)	0.916(0.041)	5.00(0)
Oracle	-	0.117(0.010)	0.887(0.037)	5.00(0)	0.117(0.010)	0.891(0.037)	5.00(0)
NC	-	0.700(0.034)	2.977(0.073)	19.93(0.70)	0.699(0.038)	2.975(0.089)	19.93(0.70)
Vec	0.008(0.006)	0.507(0.028)	3.715(0.109)	40.00(0)	0.530(0.044)	3.810(0.161)	40.00(0)
Regime: Varying m with $(m,N)=(20,1000)$							
Ours	0.006(0)	0.158(0.021)	1.005(0.066)	5.00(0)	0.159(0.020)	1.010(0.060)	5.00(0)
Oracle	-	0.153(0.020)	0.986(0.062)	5.00(0)	0.156(0.019)	0.999(0.060)	5.00(0)
NC	-	0.675(0.051)	2.487(0.102)	11.21(0.57)	0.682(0.055)	2.504(0.105)	11.21(0.57)
Vec	0.007(0.006)	0.232(0.028)	1.806(0.109)	20.00(0)	0.239(0.045)	1.831(0.166)	19.99(0.10)
Regime: Varying m with $(m,N)=(35,1750)$							
Ours	0.006(0)	0.162(0.017)	1.050(0.056)	5.00(0)	0.167(0.017)	1.066(0.054)	5.00(0)
Oracle	-	0.158(0.017)	1.034(0.055)	5.00(0)	0.162(0.015)	1.045(0.044)	5.00(0)
NC	-	0.725(0.040)	2.924(0.094)	17.06(0.71)	0.730(0.039)	2.937(0.083)	17.06(0.71)
Vec	0.012(0.013)	0.583(0.040)	3.712(0.127)	35.00(0)	0.617(0.068)	3.817(0.215)	35.00(0)
Regime: Varying m with $(m,N)=(50,2500)$							
Ours	0.006(0)	0.163(0.014)	1.063(0.044)	5.00(0)	0.166(0.015)	1.076(0.050)	5.04(0.40)
Oracle	-	0.159(0.014)	1.049(0.042)	5.00(0)	0.158(0.013)	1.047(0.040)	5.00(0)
NC	-	0.758(0.032)	3.233(0.079)	22.32(0.80)	0.762(0.033)	3.242(0.077)	22.32(0.80)
Vec	0.067(0.072)	0.863(0.090)	5.065(0.158)	50.00(0)	0.965(0.201)	5.545(0.626)	49.99(0.10)

Table 4: Multivariate regression with multiple change-points

Criterion		Small breaks		Large breaks	
		Rough	Refined	Rough	Refined
Change detection	$\widehat{s}$	3.12(0.36)	-	3.00(0)	-
	OE	0.027(0.038)	0.009(0.027)	0.002(0.001)	0.001(0.001)
	UE	0.041(0.055)	0.024(0.050)	0.002(0.001)	0.001(0.001)
Matrix recovery	MOE	-	0.291(0.029)	-	0.115(0.006)
	MUE	-	0.313(0.088)	-	0.115(0.006)
	$\max\widehat{r}_k$	-	5.07(0.26)	-	5.00(0)
	$\min\widehat{r}_k$	-	5.00(0)	-	5.00(0)

Meanwhile, thanks to the success of change-point localization, the matrix recovery tasks can be completed with high accuracy as well, in terms of both Frobenius error and rank recovery. On the other hand, when the signal is large, it is not surprising that the scheme can handle both change-point detection and matrix estimation more easily. The trajectory of $\lVert\widehat{\mathbf{\Delta}}\rVert_F^2$ in Figure 1 reflects the contrast of difficulty with different magnitudes of change signal.



(a) Small breaks

(b) Large breaks

Figure 1: *Trajectories of* $\|\widehat{\boldsymbol{\Delta}}_i\|_F^2$ *across intervals under the multivariate regression model with multiple change-points*

4.3 Real-data analysis

In this section, we study the air pollution problem induced by inhalable particulate matter (PM). According to California Air Resources Board[1], PM is a complex mixture of many chemical species, including solids and aerosols composed of small droplets of liquid, dry solid fragments, and solid cores with liquid coatings. Particles are defined by their diameter for air quality regulatory purposes. Those with a diameter of 10 microns or less (PM10) are inhalable into the lungs and can induce adverse health effects, such as repository disease and cardiovascular disorders. Fine particulate matter is defined as particles that are 2.5 microns or less in diameter (PM2.5). PM may be either directly emitted from sources (primary particles) or formed in the atmosphere through chemical reactions of gases (secondary particles) such as sulfur dioxide (SO_2), nitrogen oxides (NO_x), and certain organic compounds.

We investigate the relationship between the concentration of PM and four air pollutants: sulfur dioxide (SO_2), carbon monoxide (CO), nitrogen dioxide (NO_2), and ozone (O_3). Our study is based on an hourly air pollutants dataset from 12 nationally controlled air-quality monitoring sites collected by the Beijing Municipal Environmental Monitoring Center. The time period is from March 1st, 2013 to February 28th, 2017. The original data file and descriptions are available at the UCI Machine Learning Repository: `https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data`.

The original dataset contains a small portion of missing values, which are scattered in a relatively random pattern across time, sites, and pollution. For simplicity, we remove the days with missing measurements. The dataset is standardized to have mean 0 and variance 1. Then we aggregate the PM2.5 and PM10 concentrations across 12 sites to create the outcome matrix

$$\mathbf{Y} = (\underbrace{Y_1, \cdots, Y_{12}}_{\text{PM2.5}} \mid \underbrace{Y_{13}, \cdots, Y_{24}}_{\text{PM10}}) \in \mathbb{R}^{1100 \times 24}.$$

[1] `https://ww2.arb.ca.gov/resources/inhalable-particulate-matter-and-health`

The covariate matrix X can be constructed similarly:

$$\mathbf{X} = (\underbrace{X_1, \cdots, X_{12}}_{\text{SO}_2} \mid \underbrace{X_{13}, \cdots, X_{24}}_{\text{CO}} \mid \underbrace{X_{25}, \cdots, X_{36}}_{\text{NO}_2} \mid \underbrace{X_{37}, \cdots, X_{48}}_{\text{O}_3}) \in \mathbb{R}^{1100\times 48}.$$

We assume the multivariate linear regression structure with potential change-points (Example 1) to model the dataset, and the goal is to detect the possible breaks as well as recover the mechanism matrices $\mathbf{\Theta}_s^\star \in \mathbb{R}^{48\times 24}$ of interest.

To study the performance of our method, we split the dataset into two parts: a test set $\{\mathbf{Y}_{\text{test}}, \mathbf{X}_{\text{test}}\}$ with 20% of the total observations ($N_{\text{test}} = 220$) and a training set $\{\mathbf{Y}_{\text{train}}, \mathbf{X}_{\text{train}}\}$ with the remaining 80% ($N_{\text{train}} = 880$). Then the training set is further divided into 5 folds and we apply cross-validation to tune the number of change-points, with 4 folds for model training and 1 fold for validation. More specifically, we apply Algorithm 1 by choosing different stopping thresholds ζ_N and construct models with varying numbers of change-points. Let "train-cv" and "validation-cv" represent the sample in training folds and validation folds, respectively, and let "test-cv" represent the testing sample. The training, validation, and test errors based on the k-th split are measured respectively by

$$\begin{aligned}
\mathtt{Err}_{\text{train-cv},k} &= \frac{1}{m_2 N_{\text{train-cv}}} \|\mathbf{Y}_{\text{train-cv},k} - \widehat{\mathbf{Y}}_{\text{train-cv},k}\|_F^2, \\
\mathtt{Err}_{\text{validation-cv},k} &= \frac{1}{m_2 N_{\text{validation-cv}}} \|\mathbf{Y}_{\text{validation-cv},k} - \widehat{\mathbf{Y}}_{\text{validation-cv},k}\|_F^2, \\
\mathtt{Err}_{\text{test},k} &= \frac{1}{m_2 N_{\text{test}}} \|\mathbf{Y}_{\text{test},k} - \widehat{\mathbf{Y}}_{\text{test},k}\|_F^2.
\end{aligned}$$

Table 5 reports the training, validation, and test errors of the algorithm. We see that there is a natural trade-off between the number of change-points selected $\widehat{s}$ and the prediction error: when $\widehat{s}$ is small, the model is too simple and can not fully capture the structure of the underlying mechanism; when $\widehat{s}$ is too large, the test error will be inflated due to overfitting. In our case, $\widehat{s} = 2$ achieves an ideal balance between the two edges. In this case, the selected change-points are $\widehat{s}_1 = 0.3928$, $\widehat{s}_2 = 0.9160$, corresponding to the middle of February in 2015 and the end of November 2016, respectively. The first time point possibly marks a critical moment when the air pollutants began to impact the formulation of PM in Beijing more significantly. The second change-point might imply the improvement of air pollution conditions, since the Chinese government took many actions in 2016 to improve the air quality, including improving the law system, promoting clean energy, encouraging the development of green industries, etc.[2]

5. Conclusion

In this paper, we study the trace regression model with a threshold variable and multiple change-points. We first develop a grid-search based nuclear norm penalized least-squares scheme for simultaneous change-point detection and high-dimensional low-rank matrix recovery under the AMOC circumstances, and then extend it to the multiple change-points

[2] For example, see the official "13th Five-Year Plan Outline" released in 2016 by the Chinese government: `https://www.uschina.org/policy/official-13th-five-year-plan-outline-released`.

Table 5: Train-validation-test errors for the air pollution data

#change-points	0	1	2	3	4
Train Error	0.1842	0.1770	0.1591	0.1427	0.1268
Validation Error	0.2836	0.2092	**0.2079**	0.2202	0.2258
Test Error	0.1925	0.1746	**0.1728**	0.1761	0.1772

scenarios. Under a set of general sufficient conditions, we establish consistency of the change-point localization and the convergence upper bound on matrix signal recovery for the proposed procedure, which aligns well with the classic results in both worlds.

The present work imposes Gaussian or sub-Gaussian distributional assumptions, which are quite common in the literature. However, real-life data typically possess less satisfactory moment or tail properties such as Cauchy or log-Gaussian noise or could be contaminated by outliers. It is thus of great importance to incorporate robustness into the proposed scheme, for example, by using some robust loss function or truncation-based procedures (Tan et al., 2023; Fan et al., 2021). In addition, it is also of great interest to develop a pre-estimation procedure for testing the existence of any change-point, by exploiting the low-rank structures. We save these interesting questions for future endeavors.

Acknowledgments

All authors contributed equally to this work and are listed in alphabetical order. We would like to acknowledge the action editor, Zaid Harchaoui, and anonymous referees for their valuable comments and suggestions, which have improved the manuscript greatly. Zou was supported by the National Key R&D Program of China (Grant Nos. 2022YFA1003703, 2022YFA1003800) and the National Natural Science Foundation of China (Grant Nos. 11925106, 12231011, 11931001, 12226007, 12326325). Wang was supported by the National Key R&D Program of China (Grant Nos. 2021YFA1000100, 2021YFA1000101, 2022YFA1003800) and the Natural Science Foundation of Shanghai (Grant No. 23ZR1419400).

Appendix A. Key results

Appendix A outlines the key results regarding the proofs of Theorem 7, Corollary 10 and Theorem 11 for single change-point scenario and Theorem 14 and Corollary 15 for multiple change-point scenario. Further technical details are deferred to Appendix B and some well-known facts that will be used in the proofs are presented in Appendix C.

A.1 Single change-point scenario

Lemma 18 (Error bounds for matrix estimation with threshold effect) *Suppose that Assumption 1, Assumption 3 and Assumption 4 hold. If $|\widehat{\tau}-\tau^\star| \le c_\tau$ and $\left\|\widehat{\boldsymbol{\Gamma}}-\boldsymbol{\Gamma}^\star\right\|_* \le c_{\boldsymbol{\Gamma}}$, then, with probability greater than $1-\alpha_N-2e\cdot\exp\left(-c'N\lambda_N^2/\{K^2\|\boldsymbol{\Delta}^\star\|_F^{-2}h_N(c_\tau)\}\right)$ for some constant $c'>0$, it holds that*

$$\begin{aligned}\left\|\widehat{\boldsymbol{\Gamma}}-\boldsymbol{\Gamma}^\star\right\|_F^2 &\le \delta^2 \vee \frac{8\lambda_N\|\Pi^{r\perp}_{\boldsymbol{\Gamma}^\star}(\boldsymbol{\Gamma}^\star)\|_*}{\kappa(\mathfrak{X})}\\ &\qquad \vee \frac{128\lambda_N^2 r}{\kappa(\mathfrak{X})^2} \vee \frac{4\lambda_N\sqrt{c_\tau}\|\boldsymbol{\Delta}^\star\|_F}{\kappa(\mathfrak{X})} \vee \frac{4Cc_{\boldsymbol{\Gamma}}c_\tau\|\boldsymbol{\Delta}^\star\|_*}{\kappa(\mathfrak{X})}, \qquad (15)\\ \left\|\widehat{\boldsymbol{\Gamma}}-\boldsymbol{\Gamma}^\star\right\|_* &\le 12\sqrt{2r}\delta \vee 12\|\Pi^{r\perp}_{\boldsymbol{\Gamma}^\star}(\boldsymbol{\Gamma}^\star)\|_*\\ &\qquad \vee \frac{192\lambda_N r}{\kappa(\mathfrak{X})} \vee 6\sqrt{c_\tau}\|\boldsymbol{\Delta}^\star\|_F \vee 24\sqrt{\frac{2Cc_{\boldsymbol{\Gamma}}c_\tau r\|\boldsymbol{\Delta}^\star\|_*}{\kappa(\mathfrak{X})}}, \qquad (16)\\ \frac{1}{2N}\left\|\mathfrak{X}\left(\widehat{\boldsymbol{\Gamma}};\widehat{\tau}\right)-\mathfrak{X}(\boldsymbol{\Gamma}^\star;\tau^\star)\right\|_2^2 &\le 6\lambda_N\sqrt{2r}\delta \vee 6\lambda_N\|\Pi^{r\perp}_{\boldsymbol{\Gamma}^\star}(\boldsymbol{\Gamma}^\star)\|_* \vee \frac{96\lambda_N^2 r}{\kappa(\mathfrak{X})}\\ &\qquad \vee 3\lambda_N\sqrt{c_\tau}\|\boldsymbol{\Delta}^\star\|_F \vee 12\lambda_N\sqrt{\frac{2Cc_{\boldsymbol{\Gamma}}c_\tau r\|\boldsymbol{\Delta}^\star\|_*}{\kappa(\mathfrak{X})}}. \qquad (17)\end{aligned}$$

Lemma 19 (Improved error bound for change-point detection with threshold effect) *Suppose Assumption 2 and Assumption 4 hold. If $|\widehat{\tau}-\tau^\star| \le c_\tau$ and $\left\|\widehat{\boldsymbol{\Gamma}}-\boldsymbol{\Gamma}^\star\right\|_* \le c_{\boldsymbol{\Gamma}}$, then, with probability greater than $1-\alpha_N-2e\cdot\exp\left(-c'N\lambda_N^2/\{K^2\|\boldsymbol{\Delta}^\star\|_F^{-2}h_N(c_\tau)\}\right)$ for some constant $c'>0$, it holds that*

$$|\widehat{\tau}-\tau^\star| \le \eta^\star,$$

where

$$\eta^\star = \max\left\{\eta(N,m_1,m_2), \{c\phi(\boldsymbol{\Delta}^\star)\}^{-1}\left(\frac{3\lambda_N}{2}c_{\boldsymbol{\Gamma}}+\lambda_N\sqrt{c_\tau}\|\boldsymbol{\Delta}^\star\|_F\right)\right\}.$$

The proofs of Theorem 18 and Theorem 19 are deferred to Appendix B.

Proof of Theorem 7

Proof Part I. First assume $\left\|\widehat{\boldsymbol{\Gamma}}-\boldsymbol{\Gamma}^\star\right\|_F \ge \delta$. Under such circumstances, the terms associated with δ in the bounds provided in Theorem 18 can be omitted. At the beginning we take a glimpse of the bounds in Theorem 18. Each of the three bounds consist of five terms. The first term is an admissible error term determined primarily by δ. For the prediction error (17) and the nuclear norm error (16), each pair of corresponding terms of the last four are proportional with respect to a factor $\lambda_N/2$. For example, as to the second term, i.e., $6\lambda_N \|\Pi_{\boldsymbol{\Gamma}^\star}^{r\perp}(\boldsymbol{\Gamma}^\star)\|_*$ for the prediction error and $12\|\Pi_{\boldsymbol{\Gamma}^\star}^{r\perp}(\boldsymbol{\Gamma}^\star)\|_*$ for nuclear norm error, it holds $\{6\lambda_N \|\Pi_{\boldsymbol{\Gamma}^\star}^{r\perp}(\boldsymbol{\Gamma}^\star)\|_*\}/\{12\|\Pi_{\boldsymbol{\Gamma}^\star}^{r\perp}(\boldsymbol{\Gamma}^\star)\|_*\} = \lambda_N/2$. While similar proportion relation also holds for the second to fourth terms of the Frobenius norm error bound (15) and the nuclear norm error, it fails for the last term pair in this case. However, another relation holds between the last term and the third term:

$$\left(\frac{192\lambda_N r}{\kappa(\mathfrak{X})} \Big/ 24\sqrt{\frac{2Cc_{\boldsymbol{\Gamma}}c_\tau r\|\boldsymbol{\Delta}^\star\|_*}{\kappa(\mathfrak{X})}}\right)^2 = \frac{128\lambda_N r}{\kappa(\mathfrak{X})^2} \Big/ \frac{4Cc_{\boldsymbol{\Gamma}}c_\tau\|\boldsymbol{\Delta}^\star\|_*}{\kappa(\mathfrak{X})}.$$

These relations show that, once the maximum is attained at a certain term for the nuclear norm bound (16), the same order can be assumed for the terms in the Frobenius bound (15) as well as the prediction error (17). Using this fact we divide our proof into four cases:

Case I.

$$12\|\Pi_{\boldsymbol{\Gamma}^\star}^{r\perp}(\boldsymbol{\Gamma}^\star)\|_* \vee \frac{192\lambda_N r}{\kappa(\mathfrak{X})} \vee 6\sqrt{c_\tau}\|\boldsymbol{\Delta}^\star\|_F \vee 24\sqrt{\frac{2Cc_{\boldsymbol{\Gamma}}c_\tau r\|\boldsymbol{\Delta}^\star\|_*}{\kappa(\mathfrak{X})}} = 12\|\Pi_{\boldsymbol{\Gamma}^\star}^{r\perp}(\boldsymbol{\Gamma}^\star)\|_*.$$

In this case it's not hard to verify that the bounds in Theorem 18 reduce respectively to

$$\left\|\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^\star\right\|_* \le 12\|\Pi_{\boldsymbol{\Gamma}^\star}^{r\perp}(\boldsymbol{\Gamma}^\star)\|_*, \ \frac{1}{2N}\left\|\mathfrak{X}\left(\widehat{\boldsymbol{\Gamma}};\widehat{\tau}\right) - \mathfrak{X}(\boldsymbol{\Gamma}^\star;\tau^\star)\right\|_2^2 \le 6\lambda_N\|\Pi_{\boldsymbol{\Gamma}^\star}^{r\perp}(\boldsymbol{\Gamma}^\star)\|_*.$$

Using Theorem 19, we can also derive a bound on the detection error, i.e.,

$$|\widehat{\tau} - \tau^\star| \le 20\{c\phi(\boldsymbol{\Delta}^\star)\}^{-1}\lambda_N\|\Pi_{\boldsymbol{\Gamma}^\star}^{r\perp}(\boldsymbol{\Gamma}^\star)\|_*.$$

Using the above results, we can update $c_\tau = 20\{c\phi(\boldsymbol{\Delta}^\star)\}^{-1}\lambda_N\|\Pi_{\boldsymbol{\Gamma}^\star}^{r\perp}(\boldsymbol{\Gamma}^\star)\|_*$ and $c_{\boldsymbol{\Gamma}} = 12\|\Pi_{\boldsymbol{\Gamma}^\star}^{r\perp}(\boldsymbol{\Gamma}^\star)\|_*$. Since the last term in the bound (15) does not follow the linear proportion relation, a special treatment is required. By Assumption 5 that $120C\{c\phi(\boldsymbol{\Delta}^\star)\}^{-1}\|\boldsymbol{\Delta}^\star\|_*\|\Pi_{\boldsymbol{\Gamma}^\star}^{r\perp}(\boldsymbol{\Gamma}^\star)\|_* \le 1$, we have

$$\begin{aligned}\left\|\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^\star\right\|_F^2 &\le \frac{8\lambda_N\|\Pi_{\boldsymbol{\Gamma}^\star}^{r\perp}(\boldsymbol{\Gamma}^\star)\|_*}{\kappa(\mathfrak{X})} \vee \frac{4Cc_{\boldsymbol{\Gamma}}c_\tau\|\boldsymbol{\Delta}^\star\|_*}{\kappa(\mathfrak{X})} \\ &= \frac{8\lambda_N\|\Pi_{\boldsymbol{\Gamma}^\star}^{r\perp}(\boldsymbol{\Gamma}^\star)\|_*}{\kappa(\mathfrak{X})} \vee \left(\frac{4C\|\boldsymbol{\Delta}^\star\|_*}{\kappa(\mathfrak{X})}\cdot 12\|\Pi_{\boldsymbol{\Gamma}^\star}^{r\perp}(\boldsymbol{\Gamma}^\star)\|_* \cdot 20\{c\phi(\boldsymbol{\Delta}^\star)\}^{-1}\lambda_N\|\Pi_{\boldsymbol{\Gamma}^\star}^{r\perp}(\boldsymbol{\Gamma}^\star)\|_*\right) \\ &\le \frac{8\lambda_N\|\Pi_{\boldsymbol{\Gamma}^\star}^{r\perp}(\boldsymbol{\Gamma}^\star)\|_*}{\kappa(\mathfrak{X})}.\end{aligned}$$

Case II.

$$12\|\Pi_{\boldsymbol{\Gamma}^\star}^{r\perp}(\boldsymbol{\Gamma}^\star)\|_* \vee \frac{192\lambda_N r}{\kappa(\mathfrak{X})} \vee 6\sqrt{c_\tau}\|\boldsymbol{\Delta}^\star\|_F \vee 24\sqrt{\frac{2Cc_{\boldsymbol{\Gamma}}c_\tau r\|\boldsymbol{\Delta}^\star\|_*}{\kappa(\mathfrak{X})}} = \frac{192\lambda_N r}{\kappa(\mathfrak{X})}.$$

Similar to Case I, the bounds in Theorem 18 now reduce respectively to

$$\begin{aligned}&\left\|\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^\star\right\|_F^2 \le \frac{128\lambda_N^2 r}{\kappa(\mathfrak{X})^2}, \ \left\|\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^\star\right\|_* \le \frac{192\lambda_N r}{\kappa(\mathfrak{X})}, \\ &\frac{1}{2N}\left\|\mathfrak{X}\left(\widehat{\boldsymbol{\Gamma}};\widehat{\tau}\right) - \mathfrak{X}(\boldsymbol{\Gamma}^\star;\tau^\star)\right\|_2^2 \le \frac{96\lambda_N^2 r}{\kappa(\mathfrak{X})}.\end{aligned}$$

Moreover, the detection error rate is

$$|\widehat{\tau} - \tau^\star| \leq \frac{320\{c\phi(\mathbf{\Delta}^\star)\}^{-1}\lambda_N^2 r}{\kappa(\mathfrak{X})}.$$

Next we hope to apply Theorem 18 and Theorem 19 to cope with the rest cases. Let $c_\tau^{(m)}$, $c_{\mathbf{\Gamma}}^{(m)}$ denote the bounds on $|\widehat{\tau} - \tau^\star|$ and $\left\|\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^\star\right\|_*$ in the m-th iteration, respectively. In light of (22) and Lemma 5, we start the iteration with

$$\begin{aligned} c_{\mathbf{\Gamma}}^{(1)} &= 4\|\Pi_{\mathbf{\Gamma}^\star}^{r\perp}(\mathbf{\Gamma}^\star)\|_* + 2\|\mathbf{\Delta}^\star\|_* + 16r\gamma_{max}, \\ c_\tau^{(1)} &= \{c\phi(\mathbf{\Delta}^\star)\}^{-1}\left(2\lambda_N \sum_{k=r+1}^{m} \rho_k(\mathbf{\Gamma}^\star) + 6\lambda_N r\gamma_{max} + \lambda_N\|\mathbf{\Delta}\|_*\right). \end{aligned}$$

Case III.

$$c_{\mathbf{\Gamma}}^{(m)} = 6\sqrt{c_\tau^{(m-1)}}\|\mathbf{\Delta}^\star\|_F.$$

This implies by Theorem 19 that

$$c_\tau^{(m)} = \{c\phi(\mathbf{\Delta}^\star)\}^{-1}\left(\frac{3\lambda_N}{2}c_{\mathbf{\Gamma}}^{(m)} + \lambda_N\sqrt{c_\tau^{(m-1)}}\|\mathbf{\Delta}^\star\|_F\right) = 10\{c\phi(\mathbf{\Delta}^\star)\}^{-1}\lambda_N\sqrt{c_\tau^{(m-1)}}\|\mathbf{\Delta}^\star\|_F.$$

This system has exactly one converging fixed point beyond 0, which is

$$c_{\mathbf{\Gamma}}^{\infty} := 60\{c\phi(\mathbf{\Delta}^\star)\}^{-1}\lambda_N\|\mathbf{\Delta}^\star\|_F,\ c_\tau^{\infty} := 100\{c\phi(\mathbf{\Delta}^\star)\}^{-2}\lambda_N^2\|\mathbf{\Delta}^\star\|_F^2.$$

Let $c_{\mathbf{\Gamma}}^\star$ and $c_\tau^\star$ be the bounds on $\|\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^\star\|_*$ and $|\widehat{\tau} - \tau^\star|$ as appeared in Theorem 7. We find that $c_{\mathbf{\Gamma}}^{\infty} < c_{\mathbf{\Gamma}}^*$ and $c_\tau^{\infty} < c_\tau^\star$ hold strictly when $r > 5\{c\phi(\mathbf{\Delta}^\star)\}^{-1}\|\mathbf{\Delta}^\star\|_F\kappa(\mathfrak{X})/16$, which is guaranteed by Assumption 5.

Case IV.

$$c_{\mathbf{\Gamma}}^{(m)} = 24\sqrt{\frac{2Cc_\tau^{(m-1)}c_{\mathbf{\Gamma}}^{(m-1)}r\|\mathbf{\Delta}^\star\|_*}{\kappa(\mathfrak{X})}} := B_1\sqrt{c_\tau^{(m-1)}}\sqrt{c_{\mathbf{\Gamma}}^{(m-1)}}.$$

Again, using Theorem 19 we obtain

$$\begin{aligned} c_\tau^{(m)} &= \{c\phi(\mathbf{\Delta}^\star)\}^{-1}\left(\frac{3\lambda_N}{2}c_{\mathbf{\Gamma}}^{(m)} + \lambda_N\sqrt{c_\tau^{(m-1)}}\|\mathbf{\Delta}^\star\|_F\right) \\ &= \{c\phi(\mathbf{\Delta}^\star)\}^{-1}\frac{3\lambda_N}{2}B_1\sqrt{c_\tau^{(m-1)}}\sqrt{c_{\mathbf{\Gamma}}^{(m-1)}} + \{c\phi(\mathbf{\Delta}^\star)\}^{-1}\lambda_N\sqrt{c_\tau^{(m-1)}}\|\mathbf{\Delta}^\star\|_F \\ &:= B_2\sqrt{c_\tau^{(m-1)}}\sqrt{c_{\mathbf{\Gamma}}^{(m-1)}} + B_3\sqrt{c_\tau^{(m-1)}}. \end{aligned}$$

This system has one pair of converging fixed points provided that

$$B_1B_2 = \{c\phi(\mathbf{\Delta}^\star)\}^{-1}\frac{3\lambda_N}{2}B_1^2 = \frac{1728\{c\phi(\mathbf{\Delta}^\star)\}^{-1}C\lambda_N r\|\mathbf{\Delta}^\star\|_*}{\kappa(\mathfrak{X})} < 1, \text{ (guaranteed by Assumption 5)}$$

which is

$$c_{\boldsymbol{\Gamma}}^{\infty} := B_1^2\left(\frac{B_3}{1-B_1B_2}\right)^2,\ c_{\tau}^{\infty} := \left(\frac{B_3}{1-B_1B_2}\right)^2.$$

Based on Assumption 5, simple algebra shows that the limits are strictly smaller than $c_{\boldsymbol{\Gamma}}^{\infty} < c_{\boldsymbol{\Gamma}}^{\star}$ and $c_{\tau}^{\infty} < c_{\tau}^{\star}$.

Now, as a summary of the results in Case III and Case IV, we've shown that, if the dominating term is not $c_{\boldsymbol{\Gamma}}^{\star}$, then starting from our initial points, we are guaranteed to reach a bound lower than $c_{\boldsymbol{\Gamma}}^{\star}$ within certain steps (say $m^{\star}$) through the iteration scheme we proposed in view of Theorem 18 and Theorem 19. Finally, combining the four cases along with the assumption of $\left\|\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^{\star}\right\|_F \geq \delta$ in the beginning, we reach the conclusion.

Part II. Then we consider $\left\|\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^{\star}\right\|_F \leq \delta$. According to (27) and (28), we have

$$\begin{aligned}\frac{1}{2N}\left\|\mathfrak{X}\left(\widehat{\boldsymbol{\Gamma}};\widehat{\tau}\right) - \mathfrak{X}(\boldsymbol{\Gamma}^{\star};\tau^{\star})\right\|_2^2 \leq& 6\lambda_N\sqrt{2r}\delta \vee 6\lambda_N\|\Pi_{\boldsymbol{\Gamma}^{\star}}^{r\perp}(\boldsymbol{\Gamma}^{\star})\|_* \vee 3\lambda_N\sqrt{c_\tau}\|\boldsymbol{\Delta}^{\star}\|_F,\\ \left\|\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^{\star}\right\|_* \leq& 12\sqrt{2r}\delta \vee 12\|\Pi_{\boldsymbol{\Gamma}^{\star}}^{r\perp}(\boldsymbol{\Gamma}^{\star})\|_* \vee 6\sqrt{c_\tau}\|\boldsymbol{\Delta}^{\star}\|_F.\end{aligned}$$

Note that these two inequalities are derived using the basic inequality and has nothing to do with the RSC assumption, therefore still hold when $\left\|\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^{\star}\right\|_F \leq \delta$. With an analysis similar to Cases I, II and III in Part I, we can derive that, with possible $m^{\star}$ rounds of iteration,

$$\begin{aligned}\frac{1}{2N}\left\|\mathfrak{X}\left(\widehat{\boldsymbol{\Gamma}};\widehat{\tau}\right) - \mathfrak{X}(\boldsymbol{\Gamma}^{\star};\tau^{\star})\right\|_2^2 \leq& 6\lambda_N\sqrt{2r}\delta \vee 6\lambda_N\|\Pi_{\boldsymbol{\Gamma}^{\star}}^{r\perp}(\boldsymbol{\Gamma}^{\star})\|_* \vee \frac{96\lambda_N^2 r}{\kappa(\mathfrak{X})},\\ \left\|\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^{\star}\right\|_* \leq& 12\sqrt{2r}\delta \vee 12\|\Pi_{\boldsymbol{\Gamma}^{\star}}^{r\perp}(\boldsymbol{\Gamma}^{\star})\|_* \vee \frac{192\lambda_N r}{\kappa(\mathfrak{X})}.\end{aligned}$$

The change in the third term is attributed to an iteration scheme when the original third term dominates the bound. Further, we can obtain the detection error bound

$$|\widehat{\tau} - \tau^{\star}| \leq 20\{c\phi(\boldsymbol{\Delta}^{\star})\}^{-1}\lambda_N\sqrt{2r}\delta \vee 20\{c\phi(\boldsymbol{\Delta}^{\star})\}^{-1}\lambda_N\|\Pi_{\boldsymbol{\Gamma}^{\star}}^{r\perp}(\boldsymbol{\Gamma}^{\star})\|_* \vee \frac{320\{c\phi(\boldsymbol{\Delta}^{\star})\}^{-1}\lambda_N^2 r}{\kappa(\mathfrak{X})}.$$

■

Proof of Corollary 10

Proof We begin by finding a threshold value γ for the singular values of $\boldsymbol{\Gamma}^{\star}$, and set the "effective rank" to be $r := |\{j : \rho_j(\boldsymbol{\Gamma}^{\star}) > \gamma\}|$, that is, the number of singular values greater than γ. With this choice, we have

$$\|\Pi_{\boldsymbol{\Gamma}^{\star}}^{r\perp}(\boldsymbol{\Gamma}^{\star})\|_* = \sum_{k=r+1}^{m}\rho_k(\boldsymbol{\Gamma}^{\star}) = \gamma\sum_{k=r+1}^{m}\rho_k(\boldsymbol{\Gamma}^{\star})\gamma^{-1} \leq \gamma\sum_{k=r+1}^{m}\{\rho_k(\boldsymbol{\Gamma}^{\star})\gamma^{-1}\}^q \leq \gamma^{1-q}R_q.$$

Meanwhile we have $R_q \ge \sum_{k=1}^{m} \rho_k(\mathbf{\Gamma}^\star)^q \ge r\gamma^q$, which gives $r \le R_q\gamma^{-q}$. Now letting $\gamma = \lambda_N/\kappa(\mathfrak{X})$, we further have $\|\Pi^{r\perp}_{\mathbf{\Gamma}^\star}(\mathbf{\Gamma}^\star)\|_* \le \lambda_N^{1-q}R_q/\kappa(\mathfrak{X})^{1-q}$ and $r \le \lambda_N^{-q}R_q\kappa(\mathfrak{X})^q$. Substituting these quantities into the general bounds in Theorem 7 concludes the proof. ■

Proof of Theorem 11

To prove Theorem 11, we will apply Corollary 10 by first verifying that the involving assumptions are satisfied for multivariate regression. In particular, Theorem 20, Theorem 22 and Theorem 23 show that Assumption 1, Assumption 2 and Assumption 3 hold with high probability, respectively. In particular, Theorem 21 is required in the verification of Assumption 2. Their proofs are deferred to Appendix B.

Proposition 20 (RSC for multivariate regression) *If $n > Cm_1$, then*

$$\mathbb{P}\left(\kappa(\mathfrak{X})\|\mathbf{\Gamma}\|_F^2 \le \frac{1}{2n}\|\mathfrak{X}(\mathbf{\Gamma};\tau)\|_2^2 \le \kappa'(\mathfrak{X})\|\mathbf{\Gamma}\|_F^2,\ \forall\ \tau \in \mathbb{T} \text{ and } \mathbf{\Gamma} \in \mathbb{R}^{(2m_1)\times m_2}\right) \ge 1 - C_1\exp(-C_2 n), \tag{18}$$

where $\kappa(\mathfrak{X}) = \underline{\rho}\underline{\sigma}^2/2$ for some constant $\underline{\rho} > 0$ and $\kappa'(\mathfrak{X}) = 3\overline{\rho}\overline{\sigma}^2/2$.

Lemma 21 (Uniform spectral concentration for a sample covariance process) *Suppose $\{\boldsymbol{x}_a\}_{a=1}^n$ are i.i.d. copies of some mean zero sub-Gaussian random vector $\boldsymbol{x}$ with parameter $\overline{\sigma}^2$. Assume that the covariance matrix $\mathbf{\Sigma}$ satisfies $\underline{\sigma}^2 \le \rho_{min}(\mathbf{\Sigma}) \le \rho_{max}(\mathbf{\Sigma}) \le \overline{\sigma}^2$, and $\{B_a(\tau), \tau \in \mathbb{T} := (0, \tau_0^\star]\}_{a=1}^n$ are i.i.d copies of the stochastic process $B(\tau) = \mathbf{1}\{U < \tau\}, 0 < \tau \le \tau_0^\star < 1$, where $U \sim$ Uniform$(0,1)$. Assume the following Bernstein-type inequality holds:*

$$\forall\ \boldsymbol{v} \in \mathbb{S}^{m_1-1} \text{ and } \tau \in \mathbb{T},\ \mathbb{P}\left(\left|\sum_{a=1}^n \left(\boldsymbol{v}^\top\boldsymbol{x}_a\boldsymbol{x}_a^\top\boldsymbol{v} - \boldsymbol{v}^\top\mathbf{\Sigma}\boldsymbol{v}\right)\cdot B_a(\tau)\right| > \sqrt{c_1 n\tau\delta} + c_2\delta\right) < \exp(-\delta).$$

Consider $\log(n) = o(m_1)$. For large enough n, m_1 and constants $c, C > 0$, under the condition $C_0R_q > 1$, with probability greater than $1 - C(1 + n + m_1^{-1}\log(n)n^3)\exp(-cm_1)$, it holds uniformly for all $\tau \in \mathbb{T}, \tau \ge \frac{C_0R_qm_1}{n}$ that

$$\left\|n^{-1}\sum_{a=1}^n \left\{\boldsymbol{x}_a\boldsymbol{x}_a^\top B_a(\tau) - \tau\mathbf{\Sigma}\right\}\right\|_{\mathrm{op}} \le c'\sqrt{\frac{\tau m_1}{n}} \le \frac{c'}{\sqrt{C_0R_q}}\tau.$$

Proposition 22 (Identifiability and discountinuity for multivariate regression) *If $n > Cm_1$, then*

$$\mathbb{P}\Big(\frac{1}{2n}\|\mathfrak{X}(\mathbf{\Gamma},\tau) - \mathfrak{X}(\mathbf{\Gamma}^\star,\tau^\star)\|_2^2 > \frac{3\underline{\sigma}^2\|\mathbf{\Delta}^\star\|_F^2}{160}|\tau - \tau^\star|,\ \forall\ \tau \in \mathbb{T} \text{ such that } |\tau - \tau^\star| > \frac{cm_1}{n} \text{ and } \forall\ \mathbf{\Gamma} \in \mathbb{R}^{(2m_1)\times m_2}\Big) \ge 1 - C_1\exp(-C_2 n).$$

Proposition 23 (Smoothness of multivariate regression design) *If $n > Cm_1$, then*

$$\mathbb{P}\left(|\mathcal{T}_N(\mathbf{\Gamma},\mathbf{\Gamma}^\star,\tau,\tau^\star)| \le C\overline{\sigma}^2|\tau - \tau^\star|\cdot\|\mathbf{\Gamma} - \mathbf{\Gamma}^\star\|\cdot\|\mathbf{\Delta}^\star\|_*,\ \forall\ \tau \in \mathbb{T}, \mathbf{\Gamma} \in \mathbb{R}^{m_1\times m_2}\right) \ge 1 - C_1\exp(-C_2 n).$$

Then we need to specify several quantities appeared in the statement of Corollary 10 under the multivariate regression model.

- *The convexity parameter $\kappa(\mathfrak{X})$.* By (18), we establish $\kappa(\mathfrak{X}) = \underline{\rho}\underline{\sigma}^2/2$.
- *Tuning parameter λ_n and the associated probabilistic rate α_n.* Note that
$$\sup_{\tau\in\mathbb{T}} \frac{2}{n}\|\mathfrak{X}^\star(\boldsymbol{\epsilon};\tau)\|_{\text{ op}} = \sup_{\tau\in\mathbb{T}} \left\|\frac{2}{n}\sum_{a=1}^n \boldsymbol{\mathcal{X}}_a(\tau)\boldsymbol{\epsilon}_a^\top\right\|_{\text{op}} \leq \left\|\frac{2}{n}\sum_{a=1}^n \boldsymbol{x}_a\boldsymbol{\epsilon}_a^\top\right\|_{\text{op}} + \sup_{\tau\in(0,1)}\left\|\frac{2}{n}\sum_{a=1}^n \boldsymbol{x}_a(\tau)\boldsymbol{\epsilon}_a^\top\right\|_{\text{op}}.$$
Now the first term is just the analogous stochastic term for tuning parameter selection under the no-change case as in Negahban and Wainwright (2011). Mimicking the proof of Lemma 3 therein, we can establish that
$$\mathbb{P}\left(\left\|\frac{2}{n}\sum_{a=1}^n \boldsymbol{x}_a\boldsymbol{\epsilon}_a^\top\right\|_{\text{op}} \geq 10\sigma\overline{\sigma}\sqrt{\frac{m_1+m_2}{n}}\right) \leq c_1\exp(-c_2(m_1+m_2)).$$
As τ ranges, the second term is formulated as a partial sum process. According to Lévy's inequality,
$$\begin{aligned}&\mathbb{P}\left(\sup_{\tau\in\mathbb{T}}\left\|\frac{2}{n}\sum_{a=1}^n \boldsymbol{x}_a(\tau)\boldsymbol{\epsilon}_a^\top\right\|_{\text{op}} \geq 10\sigma\overline{\sigma}\sqrt{\frac{m_1+m_2}{n}}\right)\\ \leq&2\mathbb{P}\left(\left\|\frac{2}{n}\sum_{a=1}^n \boldsymbol{x}_a\boldsymbol{\epsilon}_a^\top\right\|_{\text{op}} \geq 10\sigma\overline{\sigma}\sqrt{\frac{m_1+m_2}{n}}\right)\\ \leq&2c_1\exp(-c_2(m_1+m_2)).\end{aligned}$$
To sum up, with probability at least $1-3c_1\exp(-c_2(m_1+m_2))$ we have
$$\sup_{\tau\in\mathbb{T}}\frac{2}{n}\|\mathfrak{X}^\star(\boldsymbol{E};\tau)\|_{\text{ op}} \leq 20\sigma\overline{\sigma}\sqrt{\frac{m_1+m_2}{n}}.$$
Thus we pick
$$\lambda_n = 20\sigma\overline{\sigma}\sqrt{\frac{m_1+m_2}{n}},\ \text{with } \alpha_n = 3c_1\exp(-c_2(m_1+m_2)).$$
- *The minimal detection length $\eta(N, m_1, m_2)$.* Comparing with our final bounds, it can be set as $\eta(N,m_1,m_2) = \frac{cm_1}{n}$ for some $c>0$.
- *The averaging term $h_n(c_\tau^{(k)})$ as in Lemma 6.* For a general c_τ, apply
$$\begin{aligned}h_n(c_\tau) &= (2c_\tau n)^{-1}\sum_{a=\lceil(\tau^\star-c_\tau)n\rceil}^{\lfloor(\tau^\star+c_\tau)n\rfloor}\left\|\boldsymbol{\Delta}^{\star\top}\boldsymbol{x}_a\right\|_2^2\\ &\leq \left\|(2c_\tau n)^{-1}\sum_{a=\lceil(\tau^\star-c_\tau)n\rceil}^{\lfloor(\tau^\star+c_\tau)n\rfloor}\boldsymbol{x}_a\boldsymbol{x}_a^\top\right\|_{\text{op}}\|\boldsymbol{\Delta}^\star\|_F^2.\end{aligned}$$

Applying Theorem 6.5 of Wainwright (2019), we can establish a high probability bound on the operator norm of Wishart matrix (also see more details in the proof of Theorem 20) provided that $c_\mathcal{T} n > Cm_1$:

$$\mathbb{P}\left(\left\|(2c_\mathcal{T} n)^{-1}\sum_{a=\lceil(\tau^\star-c_\mathcal{T})n\rceil}^{\lfloor(\tau^\star+c_\mathcal{T})n\rfloor}\boldsymbol{x}_a\boldsymbol{x}_a^\top\right\|_{\mathrm{op}}\le 2\overline{\sigma}^2\right)\ge 1-C_1\exp(-C_2 n),$$

which suggests with probability greater than $1-C_1\exp(-C_2 n)$

$$h_n(c_\mathcal{T})\le 2\overline{\sigma}^2\|\boldsymbol{\Delta}^\star\|_F^2.$$

Note this conclusion is based on a prerequisite involving $c_\mathcal{T}$, i.e., $c_\mathcal{T} n > Cm_1$. Since we are considering a finite decreasing sequence $\left(c_\mathcal{T}^{(k)}\right)_{k=1}^{m_\star}$, it suffices to have $c_\mathcal{T}^{(m^\star)} n > Cm_1$. If no such C existed, we would have $c_\mathcal{T}^{(m^\star)} = o(m_1/n)$. This aligns with (even outperforms in near low-rank case) the desired rate we are establishing. Checking the iteration steps in the proof of Theorem 7, this is impossible since the iteration would have stopped. Thus we conclude it is enough to consider such a unified C.

By managing the constants it suffices to have $n > Cm_1$ for some constant $C > 0$.

- *The rate of the step size $m^\star$.* This has been discussed in Remark 8.

Finally, it remains to check Assumption 5. Aggregating the above results, clearly these inequalities hold if n, m_1 and r (choosing the effective rank in the near low-rank case) is large enough.

A.2 Verify Assumption 5 in concrete examples

As one example, consider the regime where $\|\boldsymbol{\Delta}^\star\|_F$ is fixed and $\boldsymbol{\Gamma}^\star$ has exact low rank r. This regime implies that $\|\boldsymbol{\Delta}^\star\|_* \le \sqrt{r}\|\boldsymbol{\Delta}^\star\|_F$ has the same order as $\sqrt{r}$, and $\|\Pi_{\boldsymbol{\Gamma}^\star}^{r\perp}(\boldsymbol{\Gamma}^\star)\|_* = 0$. The parameter $\kappa(\mathfrak{X})$ is also a bounded constant when evaluated in many concrete examples (such as the random design example in Section 2.2.6). The tuning parameter, λ_N, usually scales with $O(N^{-s})$ for some $s > 0$ thus converges to zero as $N \to \infty$. Therefore, we can check that Assumption 5 is satisfied under such scaling. By introducing some additional universal constants C_1, C_2, C_3, Assumption 5 can be simplified to those given in Table 6.

Table 6: Simplified version of Assumption 5

Original	Simplified
$120C\{c\phi(\boldsymbol{\Delta}^\star)\}^{-1}\Vert\boldsymbol{\Delta}^\star\Vert_*\Vert\Pi_{\boldsymbol{\Gamma}^\star}^{r\perp}(\boldsymbol{\Gamma}^\star)\Vert_* < 1$	$0 < 1$
$5\{c\phi(\boldsymbol{\Delta}^\star)\}^{-1}\Vert\boldsymbol{\Delta}^\star\Vert_F\kappa(\mathfrak{X})/16 < r$	$R_0 < r$
$1728\{c\phi(\boldsymbol{\Delta}^\star)\}^{-1}C\lambda_N r\Vert\boldsymbol{\Delta}^\star\Vert_*/\kappa(\mathfrak{X}) < 1$	$C_1\lambda_N r^{3/2} < 1$
$\frac{\{c\phi(\boldsymbol{\Delta}^\star)\}^{-2}\kappa(\mathfrak{X})\Vert\boldsymbol{\Delta}^\star\Vert_F^2}{320[1-1728\{c\phi(\boldsymbol{\Delta}^\star)\}^{-1}C\lambda_N r\Vert\boldsymbol{\Delta}^\star\Vert_*/\kappa(\mathfrak{X})]^2} < r$	$C_2/(1-C_1\lambda_N r^{3/2}) < r$
$\frac{\{c\phi(\boldsymbol{\Delta}^\star)\}^{-2}\lambda_N C\Vert\boldsymbol{\Delta}^\star\Vert_*\Vert\boldsymbol{\Delta}^\star\Vert_F^2}{96[1-1728\{c\phi(\boldsymbol{\Delta}^\star)\}^{-1}C\lambda_N r\Vert\boldsymbol{\Delta}^\star\Vert_*/\kappa(\mathfrak{X})]^2} < 1$	$C_3\lambda_N/(1-C_1\lambda_N r^{3/2}) < 1$

We can choose r to be an integer (either bounded or moderately growing) and c to be a constant that is large enough to meet the second and fourth condition of Table 6. When $\lambda_N \to 0$, all the conditions can still be justified.

A.3 Assumptions and Proofs of Theorem 14 and Corollary 15

Let $t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(N)}$ be the order statistics of $\{t_i\}_{i=1}^N$. Let $\mathcal{T}_{ij} = [t_{(i)}, t_{(j)}]$ be the interval such that $d_{ij} := t_{(j)} - t_{(i)} \geq \underline{d}$ for some $\underline{d} > 0$ and it contains at most one change-point. If T_{ij} contains no change-point, let $\mathbf{\Theta}_{ij}$ be the corresponding regression matrix and define $\mathbf{\Delta}_{ij} = \mathbf{0}$. If T_{ij} contains some change-point $\tau_k^\star \in \mathcal{T}_{ij}$, let $\mathbf{\Theta}_{ij}^\top$ and $\mathbf{\Theta'}_{ij}^\top$ be the regression matrices before and after the change-point, and define $\mathbf{\Delta}_{ij} = \mathbf{0}$. For each case, we denote $\mathbf{\Gamma}_{ij} = (\mathbf{\Theta}_{ij}, \mathbf{\Delta}_{ij}^\top)^\top$. We consider without loss of generality that there exist some t_j such that $t_j = t_i \pm \omega$ in Stage I and one can find t_{j_1} and t_{j_2} such that $\mathcal{I}_s = [t_{j_1}, t_{j_2}]$ in Stage II. Otherwise, we can modify our two-stage procedure by replacing each end of $\mathcal{T}_i$ (and $\mathcal{I}_s$) by the closest t_j. Let $\mathbb{T}_{ij} = [t_{(i)} + \rho, t_{(j)} - \rho] \in \mathcal{T}_{ij}$ for some boundary removal parameter $\rho > 0$. We assume that ω, ρ and the number of change-points $s^\star$ are all fixed. We first introduce some assumptions that are parallel to those in Corollary 9 but tailored for multiple change-point scenario.

Assumption 7 *The restricted strong convexity condition holds with curvature $\kappa(\mathfrak{X}) > 0$ in the sense that*

$$\frac{1}{2d_{ij}N} \sum_{k:t_k \in \mathcal{T}_{ij}} \langle \boldsymbol{\mathcal{X}}_k(\tau) \, , \, \boldsymbol{M} \rangle^2 \geq \kappa(\mathfrak{X}) \|\boldsymbol{M}\|_F^2, \text{ for all } \boldsymbol{M} \in \mathcal{C}(r, \mathbf{\Gamma}_{ij}), \ \tau \in \mathbb{T}_{ij},$$

where

$$\mathcal{C}(r, \mathbf{\Gamma}_{ij}) = \left\{ \boldsymbol{M} \in \mathbb{R}^{(2m_1) \times m_2} : \ \|\Pi_{\mathbf{\Gamma}_{ij}}^{r\perp}(\boldsymbol{M})\|_* \leq 3\|\Pi_{\mathbf{\Gamma}_{ij}}^r(\boldsymbol{M})\|_* + 2\|\mathbf{\Delta}_{ij}\|_F \right\}.$$

Assumption 8 *Consider $\mathbf{\Delta}_{ij} \neq \mathbf{0}$ with the change-point $\tau_k^\star$. There exists some constants $\eta(N, m_1, m_2) > 0$ and $c > 0$ such that for any $\tau \in \mathbb{T}_{ij}$ with $|\tau - \tau_k^\star| > \eta(N, m_1, m_2)$ and for any $\mathbf{\Gamma}$ with $\mathbf{\Gamma} - \mathbf{\Gamma}_{ij} \in \mathcal{C}(r, \mathbf{\Gamma}_{ij})$, it holds that*

$$\frac{1}{2d_{ij}N} \sum_{k:t_k \in \mathcal{T}_{ij}} (\langle \boldsymbol{\mathcal{X}}_k(\tau) \, , \, \mathbf{\Gamma} \rangle - \langle \boldsymbol{\mathcal{X}}_k(\tau_k^\star) \, , \, \mathbf{\Gamma}_{ij} \rangle)^2 > c\phi(\mathbf{\Delta}_{ij})|\tau - \tau_k^\star|.$$

Assumption 9 *Consider $\mathbf{\Delta}_{ij} \neq \mathbf{0}$ with the change-point $\tau_k^\star$. There exists some constant $C > 0$ such that for any $\tau \in \mathbb{T}_{ij}$ with $\eta(N, m_1, m_2) < |\tau - \tau_k^\star| < c_\tau$ for some $\eta(N, m_1, m_2) > 0$ and $c_\tau > 0$, and for any $\mathbf{\Gamma}$ with $\mathbf{\Gamma} - \mathbf{\Gamma}_{ij} \in \mathcal{C}(r, \mathbf{\Gamma}_{ij}) \cap \{\boldsymbol{M} : \ \|\boldsymbol{M}\|_* \leq c_{\mathbf{\Gamma}}\}$ for some $c_{\mathbf{\Gamma}} > 0$, it holds that*

$$|\mathcal{T}_N(\mathbf{\Gamma}, \mathbf{\Gamma}_{ij}, \tau, \tau_k^\star)| \leq C c_\tau c_{\mathbf{\Gamma}} \|\mathbf{\Delta}_{ij}\|_*,$$

where

$$\mathcal{T}_N(\mathbf{\Gamma}, \mathbf{\Gamma}_{ij}, \tau, \tau_k^\star) = (d_{ij}N)^{-1} \sum_{k:t_k \in \mathcal{T}_{ij}} \langle \boldsymbol{\mathcal{X}}_k(\tau) \, , \, \mathbf{\Gamma} - \mathbf{\Gamma}_{ij} \rangle \, \langle \boldsymbol{\mathcal{X}}_k(\tau_k^\star) - \boldsymbol{\mathcal{X}}_k(\tau) \, , \, \mathbf{\Gamma}_{ij} \rangle \, .$$

Assumption 10 *The noises ϵ_i are i.i.d. copies of a mean zero sub-Gaussian random variable ϵ, i.e., there exists some $K > 0$, such that $\mathbb{E}\{\exp\left(\epsilon^2/K^2\right)\} \leq e$.*

Assumption 11 *There exist some constants $\{C_i\}_{i=1}^{6}$ such that*

$$C_1\{\phi(\boldsymbol{\Delta}_{ij})\}^{-1}\|\boldsymbol{\Delta}_{ij}\|_F\kappa(\mathfrak{X}) < r,$$
$$\frac{C_2\{\phi(\boldsymbol{\Delta}_{ij})\}^{-1}\lambda_N r\|\boldsymbol{\Delta}_{ij}\|_*}{\kappa(\mathfrak{X})} < 1,$$
$$\frac{C_3\{\phi(\boldsymbol{\Delta}_{ij})\}^{-2}\kappa(\mathfrak{X})\|\boldsymbol{\Delta}_{ij}\|_F^2}{[1 - C_4\{\phi(\boldsymbol{\Delta}_{ij})\}^{-1}\lambda_N r\|\boldsymbol{\Delta}_{ij}\|_*/\kappa(\mathfrak{X})]^2} < r,$$
$$\frac{C_5\{\phi(\boldsymbol{\Delta}_{ij})\}^{-2}\lambda_N\|\boldsymbol{\Delta}_{ij}\|_*\|\boldsymbol{\Delta}_{ij}\|_F^2}{[1 - C_6\{\phi(\boldsymbol{\Delta}_{ij})\}^{-1}\lambda_N r\|\boldsymbol{\Delta}_{ij}\|_*/\kappa(\mathfrak{X})]^2} < 1.$$

Assumption 12 *For $c_\tau > 0$, $h_N(c_\tau) = (2c_\tau N)^{-1}\sum_{i:|t_i-\tau^\star|\leq c_\tau}\langle \boldsymbol{X}_i\,,\,\boldsymbol{\Delta}^\star\rangle^2$ is bounded.*

Note that Assumption 12 is parallel to Assumption 5 in Lee et al. (2016) which studied the LASSO estimation in linear regression models with a single change-point, and greatly facilities our theoretical analysis. In fact, this assumption could be removed and the resulting high probability result then depends on these h_N in a similar way like that appears in Theorem 7.

In Stage I, we choose $\mathcal{T}_{ij} = \mathcal{T}_i = [t_i - \omega, t_i + \omega]$. Under Assumption 7–Assumption 11, by using arguments similar to those in the proof of Corollary 9, we can show that, with probability greater than $1 - \alpha_N - 2e\tilde{m}^\star N^2\exp\{-\tilde{c}N\lambda_N^2/(K^2\Delta_{max})\}$ for some constant $\tilde{c} > 0$ and $\tilde{m}^\star > 0$

$$\|\widehat{\boldsymbol{\Delta}}_i - \boldsymbol{\Delta}_{ij}\|_F^2 \leq \frac{C_1\lambda_{2\omega N}^2 r}{\kappa(\mathfrak{X})^2}$$

for some $C_1 > 0$. Hence if $\boldsymbol{\Delta}_{ij} = \mathbf{0}$, then $\|\widehat{\boldsymbol{\Delta}}_i\|_F^2 \leq \frac{C_1\lambda_{2\omega N}^2 r}{\kappa(\mathfrak{X})^2}$, and if $\boldsymbol{\Delta}_{ij} \neq \mathbf{0}$ with some change-point $\tau_k^\star$, then $\|\widehat{\boldsymbol{\Delta}}_i\|_F^2 \geq \|\boldsymbol{\Delta}_{ij}\|_F^2 - \frac{C_1\lambda_{2\omega N}^2 r}{\kappa(\mathfrak{X})^2}$. As a result, if we select $\zeta_N = \frac{C'\lambda_{2\omega N}^2 r}{\kappa(\mathfrak{X})^2}$ for some $C' > C_1$, then we can conclude that $\widehat{s} = s^\star$. In other words, the event $\mathcal{E}_N$ holds with high probability. In Stage II, by choosing $\mathcal{T}_{ij} = \mathcal{I}_s$ and using similar arguments the conclusion follows.

By using arguments similar to those in the proofs of Theorem 11 and Theorem 14, Corollary 15 follows directly. Hence we omit the proof.

Appendix B. Technical details

Proof of Lemma 2

Proof By the definition of $\widehat{\boldsymbol{\Gamma}}$ and $\widehat{\tau}$, for any $\boldsymbol{\Gamma}$ and $\tau \in \mathbb{T}$, we have

$$\frac{1}{2N}\left\|\boldsymbol{y} - \mathfrak{X}\left(\widehat{\boldsymbol{\Gamma}};\widehat{\tau}\right)\right\|_2^2 + \lambda_N\|\widehat{\boldsymbol{\Gamma}}\|_* \leq \frac{1}{2N}\left\|\boldsymbol{y} - \mathfrak{X}\left(\boldsymbol{\Gamma};\tau\right)\right\|_2^2 + \lambda_N\|\boldsymbol{\Gamma}\|_*.$$

Using (2), the inequality becomes

$$\begin{aligned}
&\frac{1}{2N}\sum_{i=1}^{N}\left(\langle \boldsymbol{\mathcal{X}}_i(\tau^\star)\,,\,\mathbf{\Gamma}^\star\rangle - \left\langle \boldsymbol{\mathcal{X}}_i(\widehat{\tau})\,,\,\widehat{\mathbf{\Gamma}}\right\rangle + \epsilon_i\right)^2 + \lambda_N\|\widehat{\mathbf{\Gamma}}\|_* \\
\leq &\frac{1}{2N}\sum_{i=1}^{N}\left(\langle \boldsymbol{\mathcal{X}}_i(\tau^\star)\,,\,\mathbf{\Gamma}^\star\rangle - \langle \boldsymbol{\mathcal{X}}_i(\tau)\,,\,\mathbf{\Gamma}\rangle + \epsilon_i\right)^2 + \lambda_N\|\mathbf{\Gamma}\|_*.
\end{aligned}$$

Some basic algebra yields

$$\begin{aligned}
&\frac{1}{2N}\sum_{i=1}^{N}\left(\langle \boldsymbol{\mathcal{X}}_i(\tau^\star)\,,\,\mathbf{\Gamma}^\star\rangle - \left\langle \boldsymbol{\mathcal{X}}_i(\widehat{\tau})\,,\,\widehat{\mathbf{\Gamma}}\right\rangle\right)^2 - \frac{1}{2N}\sum_{i=1}^{N}\left(\langle \boldsymbol{\mathcal{X}}_i(\tau^\star)\,,\,\mathbf{\Gamma}^\star\rangle - \langle \boldsymbol{\mathcal{X}}_i(\tau)\,,\,\mathbf{\Gamma}\rangle\right)^2 \\
&\leq \frac{1}{N}\sum_{i=1}^{N}\epsilon_i\left(\left\langle \boldsymbol{\mathcal{X}}_i(\widehat{\tau})\,,\,\widehat{\mathbf{\Gamma}}\right\rangle - \langle \boldsymbol{\mathcal{X}}_i(\tau)\,,\,\mathbf{\Gamma}\rangle\right) + \lambda_N\|\mathbf{\Gamma}\|_* - \lambda_N\|\widehat{\mathbf{\Gamma}}\|_* \\
&= \frac{1}{N}\sum_{i=1}^{N}\epsilon_i\left(\left\langle \boldsymbol{\mathcal{X}}_i(\widehat{\tau})\,,\,\widehat{\mathbf{\Gamma}}\right\rangle - \langle \boldsymbol{\mathcal{X}}_i(\widehat{\tau})\,,\,\mathbf{\Gamma}\rangle\right) + \frac{1}{N}\sum_{i=1}^{N}\epsilon_i\left(\langle \boldsymbol{\mathcal{X}}_i(\widehat{\tau})\,,\,\mathbf{\Gamma}\rangle - \langle \boldsymbol{\mathcal{X}}_i(\tau)\,,\,\mathbf{\Gamma}\rangle\right) + \lambda_N\|\mathbf{\Gamma}\|_* - \lambda_N\|\widehat{\mathbf{\Gamma}}\|_* \\
&\leq (\lambda_N/2)\|\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}\|_* + \lambda_N\|\mathbf{\Gamma}\|_* - \lambda_N\|\widehat{\mathbf{\Gamma}}\|_* + \mathcal{R}_N(\mathbf{\Gamma}, \widehat{\tau}, \tau).
\end{aligned}$$

Here for the last inequality we used the dual norm inequality, i.e., for any $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m_1, m_2}$, $\langle \boldsymbol{A}\,,\,\boldsymbol{B}\rangle \leq \|\boldsymbol{A}\|_{\mathrm{op}}\|\boldsymbol{B}\|_*$. Substituting (τ, Γ) with $(\tau^\star, \mathbf{\Gamma}^\star)$, we obtain

$$\frac{1}{2N}\sum_{i=1}^{N}\left(\langle \boldsymbol{\mathcal{X}}_i(\tau^\star)\,,\,\mathbf{\Gamma}^\star\rangle - \left\langle \boldsymbol{\mathcal{X}}_i(\widehat{\tau})\,,\,\widehat{\mathbf{\Gamma}}\right\rangle\right)^2 \leq (\lambda_N/2)\|\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^\star\|_* + \lambda_N\|\mathbf{\Gamma}^\star\|_* - \lambda_N\|\widehat{\mathbf{\Gamma}}\|_* + \mathcal{R}_N(\mathbf{\Gamma}^\star, \widehat{\tau}, \tau^\star). \tag{19}$$

Note that

$$\begin{aligned}
\|\mathbf{\Gamma}^\star\|_* - \|\widehat{\mathbf{\Gamma}}\|_* &= \|\mathbf{\Gamma}^\star\|_* - \|\Pi^{r}_{\mathbf{\Gamma}^\star}\mathbf{\Gamma}^\star + \Pi^{r\perp}_{\mathbf{\Gamma}^\star}(\mathbf{\Gamma}^\star) + \Pi^{r}_{\mathbf{\Gamma}^\star}(\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^\star) + \Pi^{r\perp}_{\mathbf{\Gamma}^\star}(\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^\star)\|_* \\
&\leq \|\mathbf{\Gamma}^\star\|_* - \|\Pi^{r}_{\mathbf{\Gamma}^\star}\mathbf{\Gamma}^\star + \Pi^{r\perp}_{\mathbf{\Gamma}^\star}(\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^\star)\|_* + \|\Pi^{r\perp}_{\mathbf{\Gamma}^\star}(\mathbf{\Gamma}^\star) + \Pi^{r}_{\mathbf{\Gamma}^\star}(\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^\star)\|_* \\
&\leq \|\mathbf{\Gamma}^\star\|_* - \|\Pi^{r}_{\mathbf{\Gamma}^\star}\mathbf{\Gamma}^\star\|_* - \|\Pi^{r\perp}_{\mathbf{\Gamma}^\star}(\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^\star)\|_* + \|\Pi^{r\perp}_{\mathbf{\Gamma}^\star}(\mathbf{\Gamma}^\star)\|_* + \|\Pi^{r}_{\mathbf{\Gamma}^\star}(\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^\star)\|_* \\
&\leq 2\|\Pi^{r\perp}_{\mathbf{\Gamma}^\star}(\mathbf{\Gamma}^\star)\|_* + \|\Pi^{r}_{\mathbf{\Gamma}^\star}(\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^\star)\|_* - \|\Pi^{r\perp}_{\mathbf{\Gamma}^\star}(\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^\star)\|_*.
\end{aligned} \tag{20}$$

Here the second inequality is due to the triangle inequality, and the third inequality is an application of the decomposibility of nuclear norm with respect to projection $\Pi^{r}_{\mathbf{\Gamma}^\star}(\cdot)$ plus the triangle inequality (see the proof of Negahban and Wainwright (2011) and Klopp (2014)). Now combining this with (19), we have

$$\begin{aligned}
&\frac{1}{2N}\sum_{i=1}^{N}\left(\langle \boldsymbol{\mathcal{X}}_i(\tau^\star)\,,\,\mathbf{\Gamma}^\star\rangle - \left\langle \boldsymbol{\mathcal{X}}_i(\widehat{\tau})\,,\,\widehat{\mathbf{\Gamma}}\right\rangle\right)^2 \\
&\leq (\lambda_N/2)\|\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^\star\|_* + \lambda_N\|\mathbf{\Gamma}^\star\|_* - \lambda_N\|\widehat{\mathbf{\Gamma}}\|_* + \mathcal{R}_N(\mathbf{\Gamma}^\star, \widehat{\tau}, \tau^\star) \\
&\leq (\lambda_N/2)\|\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^\star\|_* + 2\lambda_N\|\Pi^{r\perp}_{\mathbf{\Gamma}^\star}(\mathbf{\Gamma}^\star)\|_* + \lambda_N\|\Pi^{r}_{\mathbf{\Gamma}^\star}(\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^\star)\|_* - \lambda_N\|\Pi^{r\perp}_{\mathbf{\Gamma}^\star}(\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^\star)\|_* \\
&\quad + \mathcal{R}_N(\mathbf{\Gamma}^\star, \widehat{\tau}, \tau^\star) \text{ (applying (20))} \\
&\leq 2\lambda_N\|\Pi^{r\perp}_{\mathbf{\Gamma}^\star}(\mathbf{\Gamma}^\star)\|_* + (3\lambda_N/2)\|\Pi^{r}_{\mathbf{\Gamma}^\star}(\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^\star)\|_* - (\lambda_N/2)\|\Pi^{r\perp}_{\mathbf{\Gamma}^\star}(\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^\star)\|_* \\
&\quad + \mathcal{R}_N(\mathbf{\Gamma}^\star, \widehat{\tau}, \tau^\star), \text{ (decomposing } \widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^\star \text{ by projection and applying the triangle inequality)}
\end{aligned}$$

which gives

$$\begin{aligned}&\frac{1}{2N}\sum_{i=1}^{N}\Big(\langle \boldsymbol{\mathcal{X}}_i(\tau^\star)\,,\,\boldsymbol{\Gamma}^\star\rangle-\Big\langle \boldsymbol{\mathcal{X}}_i(\widehat{\tau})\,,\,\widehat{\boldsymbol{\Gamma}}\Big\rangle\Big)^2+\frac{\lambda_N}{2}\|\Pi^{r\perp}_{\boldsymbol{\Gamma}^\star}(\widehat{\boldsymbol{\Gamma}}-\boldsymbol{\Gamma}^\star)\|_*\\ \le&2\lambda_N\|\Pi^{r\perp}_{\boldsymbol{\Gamma}^\star}(\boldsymbol{\Gamma}^\star)\|_*+\mathcal{R}_N(\boldsymbol{\Gamma},\widehat{\tau},\tau)+\frac{3\lambda_N}{2}\|\Pi^{r}_{\boldsymbol{\Gamma}^\star}(\widehat{\boldsymbol{\Gamma}}-\boldsymbol{\Gamma}^\star)\|_*.\end{aligned}$$

■

Proof of Corollary 3

Proof First, we notice that, for all $\boldsymbol{A}\in\mathbb{R}^{(2m_1)\times m_2}$,

$$\Pi^{r}_{\boldsymbol{\Gamma}^\star}(\boldsymbol{A})=\boldsymbol{P}_{\mathbb{U}^r}\boldsymbol{A}+\boldsymbol{A}\boldsymbol{P}_{\mathbb{V}^r}-\boldsymbol{P}_{\mathbb{U}^r}\boldsymbol{A}\boldsymbol{P}_{\mathbb{V}^r}=\boldsymbol{P}_{\mathbb{U}^r}\boldsymbol{A}(I-\boldsymbol{P}_{\mathbb{V}^r})+\boldsymbol{A}\boldsymbol{P}_{\mathbb{V}^r}. \tag{21}$$

Therefore, it follows that $\operatorname{rank}\big(\Pi^{r}_{\boldsymbol{\Gamma}^\star}(\boldsymbol{A})\big)\le 2r$, $\|\Pi^{r}_{\boldsymbol{\Gamma}^\star}(\boldsymbol{A})\|_{\mathrm{op}}\le 2\|\boldsymbol{A}\|_{\mathrm{op}}$ and $\|\Pi^{r}_{\boldsymbol{\Gamma}^\star}(\boldsymbol{A})\|_*\le\sqrt{2r}\|\boldsymbol{A}\|_F$.

By the basic inequality (6), we have

$$\frac{1}{2N}\sum_{i=1}^{N}\Big(\langle \boldsymbol{\mathcal{X}}_i(\tau^\star)\,,\,\boldsymbol{\Gamma}^\star\rangle-\Big\langle \boldsymbol{\mathcal{X}}_i(\widehat{\tau})\,,\,\widehat{\boldsymbol{\Gamma}}\Big\rangle\Big)^2\le2\lambda_N\|\Pi^{r\perp}_{\boldsymbol{\Gamma}^\star}(\boldsymbol{\Gamma}^\star)\|_*+\mathcal{R}_N(\boldsymbol{\Gamma}^\star,\widehat{\tau},\tau^\star)+\frac{3\lambda_N}{2}\left\|\Pi^{r}_{\boldsymbol{\Gamma}^\star}(\widehat{\boldsymbol{\Gamma}}-\boldsymbol{\Gamma}^\star)\right\|_*.$$

We consider the three terms on the right-hand side respectively. Apparently, we have $\|\Pi^{r\perp}_{\boldsymbol{\Gamma}^\star}(\boldsymbol{\Gamma}^\star)\|_*=\sum_{k=r+1}^{m}\sigma_j(\boldsymbol{\Gamma}^\star)$. Using the definition of $\mathcal{R}_N(\boldsymbol{\Gamma}^\star,\widehat{\tau},\tau^\star)$, it follows that

$$\begin{aligned}R_N(\boldsymbol{\Gamma}^\star,\widehat{\tau},\tau^\star)&=N^{-1}\sum_{i=1}^{N}\epsilon_i\left\langle \boldsymbol{X}_i(\widehat{\tau})-\boldsymbol{X}_i(\tau^\star)\,,\,\boldsymbol{\Delta}^\star\right\rangle\\&\le\|\boldsymbol{\Delta}^\star\|_*\cdot\left\|N^{-1}\sum_{i=1}^{N}\epsilon_i\left(\boldsymbol{X}_i(\widehat{\tau})-\boldsymbol{X}_i(\tau^\star)\right)\right\|_{\mathrm{op}}\quad\text{(the dual norm inequality)}\\&\le\|\boldsymbol{\Delta}^\star\|_*\cdot\left\|N^{-1}\sum_{i=1}^{N}\epsilon_i\left(\boldsymbol{\mathcal{X}}_i(\widehat{\tau})-\boldsymbol{\mathcal{X}}_i(\tau^\star)\right)\right\|_{\mathrm{op}}\\&\le\lambda_N\|\boldsymbol{\Delta}^\star\|_*.\end{aligned}$$

Meanwhile, applying the result we state following (21), we obtain

$$\frac{3\lambda_N}{2}\left\|\Pi^{r}_{\boldsymbol{\Gamma}^\star}\left(\widehat{\boldsymbol{\Gamma}}-\boldsymbol{\Gamma}^\star\right)\right\|_*\le\frac{3\lambda_N}{2}\operatorname{rank}\Big(\Pi^{r}_{\boldsymbol{\Gamma}^\star}\left(\widehat{\boldsymbol{\Gamma}}-\boldsymbol{\Gamma}^\star\right)\Big)\left\|\Pi^{r}_{\boldsymbol{\Gamma}^\star}\left(\widehat{\boldsymbol{\Gamma}}-\boldsymbol{\Gamma}^\star\right)\right\|_{\mathrm{op}}\le6\lambda_N r\gamma_{max}.$$

To sum up, we have

$$\frac{1}{2N}\sum_{i=1}^{N}\Big(\langle \boldsymbol{\mathcal{X}}_i(\tau^\star)\,,\,\boldsymbol{\Gamma}^\star\rangle-\Big\langle \boldsymbol{\mathcal{X}}_i(\widehat{\tau})\,,\,\widehat{\boldsymbol{\Gamma}}\Big\rangle\Big)^2\le2\lambda_N\|\Pi^{r\perp}_{\boldsymbol{\Gamma}^\star}(\boldsymbol{\Gamma}^\star)\|_*+\lambda_N\|\boldsymbol{\Delta}^\star\|_*+6\lambda_N r\gamma_{max}.$$

As a by-product, using the basic inequality again, we can also derive

$$\|\widehat{\boldsymbol{\Gamma}}-\boldsymbol{\Gamma}^\star\|_*\le4\|\Pi^{r\perp}_{\boldsymbol{\Gamma}^\star}(\boldsymbol{\Gamma}^\star)\|_*+2\|\boldsymbol{\Delta}^\star\|_*+16r\gamma_{max}. \tag{22}$$

■

Proof of Theorem 4

Proof When no threshold effect exists, $\mathbf{\Delta}^\star = \mathbf{0}$, and the term $\mathcal{R}_N(\mathbf{\Gamma}^\star, \widehat{\tau}, \tau^\star)$ in the right-hand side of the basic inequality (6) vanishes. Meanwhile, noticing that $\langle \boldsymbol{\mathcal{X}}_i(\tau^\star)\,,\, \mathbf{\Gamma}^\star \rangle = \langle \boldsymbol{\mathcal{X}}_i(\widehat{\tau})\,,\, \mathbf{\Gamma}^\star \rangle$, the prediction error term, i.e., the first term on the left-hand side, becomes

$$(2N)^{-1} \sum_{i=1}^N \left(\langle \boldsymbol{\mathcal{X}}_i(\tau^\star)\,,\, \mathbf{\Gamma}^\star \rangle - \left\langle \boldsymbol{\mathcal{X}}_i(\widehat{\tau})\,,\, \widehat{\mathbf{\Gamma}} \right\rangle \right)^2 = (2N)^{-1} \sum_{i=1}^N \left\langle \boldsymbol{\mathcal{X}}_i(\widehat{\tau})\,,\, \widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^\star \right\rangle^2 .$$

To sum up, the inequality (6) then becomes

$$\frac{1}{2N} \sum_{i=1}^N \left\langle \boldsymbol{\mathcal{X}}_i(\widehat{\tau})\,,\, \widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^\star \right\rangle^2 + \frac{\lambda_N}{2} \|\Pi^{r\perp}_{\mathbf{\Gamma}^\star}(\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^\star)\|_* \le 2\lambda_N \|\Pi^{r\perp}_{\mathbf{\Gamma}^\star}(\mathbf{\Gamma}^\star)\|_* + \frac{3\lambda_N}{2} \|\Pi^{r}_{\mathbf{\Gamma}^\star}(\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^\star)\|_* . \tag{23}$$

This implies that

$$\begin{aligned} \frac{\lambda_N}{2} \|\Pi^{r\perp}_{\mathbf{\Gamma}^\star}(\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^\star)\|_* &\le 2\lambda_N \|\Pi^{r\perp}_{\mathbf{\Gamma}^\star}(\mathbf{\Gamma}^\star)\|_* + \frac{3\lambda_N}{2} \|\Pi^{r}_{\mathbf{\Gamma}^\star}(\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^\star)\|_* \\ &= 2\lambda_N \sum_{k=r+1}^m \rho_k(\mathbf{\Gamma}^\star) + \frac{3\lambda_N}{2} \|\Pi^{r}_{\mathbf{\Gamma}^\star}(\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^\star)\|_*, \end{aligned}$$

which further shows that the error $\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^\star$ lies in $\mathcal{C}(r, \delta, \mathbf{\Gamma}^\star, \mathbb{T})$ defined in Assumption 1.

Therefore, for the case when $\left\|\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^\star\right\|_F > \delta$, we can apply the RSC condition, and obtain

$$\begin{aligned} \kappa(\mathfrak{X}) \left\|\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^\star\right\|_F^2 &\le \frac{1}{2N} \sum_{i=1}^N \left\langle \boldsymbol{\mathcal{X}}_i(\widehat{\tau})\,,\, \widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^\star \right\rangle^2 \\ &\le 2\lambda_N \left\|\Pi^{r\perp}_{\mathbf{\Gamma}^\star}(\mathbf{\Gamma}^\star)\right\|_* + \frac{3\lambda_N}{2} \left\|\Pi^{r}_{\mathbf{\Gamma}^\star}\left(\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^\star\right)\right\|_* \text{ (using (23))} \\ &\le 3\lambda_N \sqrt{r} \left\|\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^\star\right\|_F + 2\lambda_N \sum_{k=r+1}^m \rho_k(\mathbf{\Gamma}^\star) \text{ (using the results following (21))} \\ &\le 6\lambda_N \sqrt{r} \left\|\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^\star\right\|_F \vee 4\lambda_N \sum_{k=r+1}^m \rho_k(\mathbf{\Gamma}^\star). \end{aligned} \tag{24}$$

Applying the basic inequality again, we have

$$\begin{aligned} \frac{1}{2N} \sum_{i=1}^N \left\langle \boldsymbol{\mathcal{X}}_i(\widehat{\tau})\,,\, \widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^\star \right\rangle^2 + \frac{\lambda_N}{2} \left\|\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^\star\right\|_* &\le 2\lambda_N \|\Pi^{r\perp}_{\mathbf{\Gamma}^\star}(\mathbf{\Gamma}^\star)\|_* + 2\lambda_N \left\|\Pi^{r}_{\mathbf{\Gamma}^\star}\left(\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^\star\right)\right\|_* \\ &\le 2\lambda_N \sum_{k=r+1}^m \rho_k(\mathbf{\Gamma}^\star) + 4\lambda_N \sqrt{r} \left\|\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^\star\right\|_F \\ &\le 4\lambda_N \sum_{k=r+1}^m \rho_k(\mathbf{\Gamma}^\star) \vee 8\lambda_N \sqrt{r} \left\|\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^\star\right\|_F . \end{aligned} \tag{25}$$

Now we substitute $\left\|\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}^\star\right\|_F$ with the bound we've derived in (24), and obtain

$$\begin{aligned}8\lambda_N\sqrt{r}\left\|\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}^\star\right\|_F &\leq \frac{48\lambda_N^2 r}{\kappa(\mathfrak{X})}\vee 8\lambda_N\sqrt{r}\left(\frac{4\lambda_N\sum_{k=r+1}^m\rho_k(\mathbf{\Gamma}^\star)}{\kappa(\mathfrak{X})}\right)^{1/2}\\ &\leq \frac{48\lambda_N^2 r}{\kappa(\mathfrak{X})}\vee\left(\frac{64\lambda_N^2 r}{\kappa(\mathfrak{X})}\vee 4\lambda_N\sum_{k=r+1}^m\rho_k(\mathbf{\Gamma}^\star)\right)\\ &= \frac{64\lambda_N^2 r}{\kappa(\mathfrak{X})}\vee 4\lambda_N\sum_{k=r+1}^m\rho_k(\mathbf{\Gamma}^\star).\end{aligned}$$

Here the second inequality is derived using $\sqrt{ab}\leq a\vee b$ for all positive a and b. Summarizing these results, we have

$$\frac{1}{2N}\sum_{i=1}^N\left\langle\boldsymbol{\mathcal{X}}_i(\widehat{\tau})\,,\,\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}^\star\right\rangle^2+\frac{\lambda_N}{2}\left\|\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}^\star\right\|_*\leq\frac{64\lambda_N^2 r}{\kappa(\mathfrak{X})}\vee 4\lambda_N\sum_{k=r+1}^m\rho_k(\mathbf{\Gamma}^\star),$$

which further gives

$$\begin{aligned}\frac{1}{2N}\sum_{i=1}^N\left\langle\boldsymbol{\mathcal{X}}_i(\widehat{\tau})\,,\,\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}^\star\right\rangle^2 &\leq \frac{64\lambda_N^2 r}{\kappa(\mathfrak{X})}\vee 4\lambda_N\sum_{k=r+1}^m\rho_k(\mathbf{\Gamma}^\star),\\ \left\|\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}^\star\right\|_* &\leq \frac{128\lambda_N r}{\kappa(\mathfrak{X})}\vee 8\sum_{k=r+1}^m\rho_k(\mathbf{\Gamma}^\star).\end{aligned}$$

Recall that our above analysis starts from $\left\|\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}^\star\right\|_F\geq\delta$. For $\widehat{\mathbf{\Gamma}}$ such that $\left\|\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}^\star\right\|_F\leq\delta$, using (25) again, we have

$$\begin{aligned}\frac{1}{2N}\sum_{i=1}^N\left\langle\boldsymbol{\mathcal{X}}_i(\widehat{\tau})\,,\,\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}^\star\right\rangle^2 &\leq 8\lambda_N\sqrt{r}\delta\vee 4\lambda_N\sum_{k=r+1}^m\rho_k(\mathbf{\Gamma}^\star),\\ \left\|\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}^\star\right\|_* &\leq 16\sqrt{r}\delta\vee 8\sum_{k=r+1}^m\rho_k(\mathbf{\Gamma}^\star).\end{aligned}$$

Then we conclude that

$$\begin{aligned}\left\|\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}^\star\right\|_F &\leq \delta\vee\frac{6\lambda_N\sqrt{r}}{\kappa(\mathfrak{X})}\vee\left(\frac{4\lambda_N\sum_{k=r+1}^m\rho_k(\mathbf{\Gamma}^\star)}{\kappa(\mathfrak{X})}\right)^{1/2},\\ \left\|\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}^\star\right\|_* &\leq 16\sqrt{r}\delta\vee\frac{128\lambda_N r}{\kappa(\mathfrak{X})}\vee 8\sum_{k=r+1}^m\rho_k(\mathbf{\Gamma}^\star),\\ \frac{1}{2N}\sum_{i=1}^N\left\langle\boldsymbol{\mathcal{X}}_i(\widehat{\tau})\,,\,\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}^\star\right\rangle^2 &\leq 8\lambda_N\sqrt{r}\delta\vee\frac{64\lambda_N^2 r}{\kappa(\mathfrak{X})}\vee 4\lambda_N\sum_{k=r+1}^m\rho_k(\mathbf{\Gamma}^\star).\end{aligned}$$

■

Proof of Lemma 5

Proof Suppose that $|\widehat{\tau}-\tau^\star|>\eta^\star$. Then

$$
\begin{aligned}
&\frac{1}{2N}\left\|\boldsymbol{y}-\mathfrak{X}(\widehat{\boldsymbol{\Gamma}};\widehat{\tau})\right\|_2^2+\lambda_N\left\|\widehat{\boldsymbol{\Gamma}}\right\|_*-\frac{1}{2N}\|\boldsymbol{y}-\mathfrak{X}(\boldsymbol{\Gamma}^\star;\tau^\star)\|_2^2-\lambda_N\|\boldsymbol{\Gamma}^\star\|_*\\
=&\frac{1}{2N}\left\|\mathfrak{X}(\widehat{\boldsymbol{\Gamma}};\widehat{\tau})-\mathfrak{X}(\boldsymbol{\Gamma}^\star;\tau^\star)\right\|_2^2-\frac{1}{N}\sum_{i=1}^{N}\epsilon_i\left(\left\langle\boldsymbol{\mathcal{X}}_i(\widehat{\tau})\,,\,\widehat{\boldsymbol{\Gamma}}\right\rangle-\left\langle\boldsymbol{\mathcal{X}}_i(\widehat{\tau})\,,\,\boldsymbol{\Gamma}^\star\right\rangle\right)\\
-&\mathcal{R}_N(\boldsymbol{\Gamma}^\star,\widehat{\tau},\tau^\star)+\lambda_N\left\|\widehat{\boldsymbol{\Gamma}}\right\|_*-\lambda_N\|\boldsymbol{\Gamma}^\star\|_*\\
\geq&\frac{1}{2N}\left\|\mathfrak{X}(\widehat{\boldsymbol{\Gamma}};\widehat{\tau})-\mathfrak{X}(\boldsymbol{\Gamma}^\star;\tau^\star)\right\|_2^2-\frac{\lambda_N}{2}\left\|\widehat{\boldsymbol{\Gamma}}-\boldsymbol{\Gamma}^\star\right\|_*\\
&-\mathcal{R}_N(\boldsymbol{\Gamma}^\star,\widehat{\tau},\tau^\star)+\lambda_N\left\|\widehat{\boldsymbol{\Gamma}}\right\|_*-\lambda_N\|\boldsymbol{\Gamma}^\star\|_*\text{ (the dual norm inequality)}\\
\geq&\frac{1}{2N}\left\|\mathfrak{X}(\widehat{\boldsymbol{\Gamma}};\widehat{\tau})-\mathfrak{X}(\boldsymbol{\Gamma}^\star;\tau^\star)\right\|_2^2-2\lambda_N\left\|\Pi^{r\perp}_{\boldsymbol{\Gamma}^\star}(\boldsymbol{\Gamma}^\star)\right\|_*\\
&-\frac{3\lambda_N}{2}\left\|\Pi^{r}_{\boldsymbol{\Gamma}^\star}\left(\widehat{\boldsymbol{\Gamma}}-\boldsymbol{\Gamma}^\star\right)\right\|_*-\mathcal{R}_N(\boldsymbol{\Gamma}^\star,\widehat{\tau},\tau^\star)\text{ (using (20) and the triangle inequality)}\\
\geq&\frac{1}{2N}\left\|\mathfrak{X}(\widehat{\boldsymbol{\Gamma}};\widehat{\tau})-\mathfrak{X}(\boldsymbol{\Gamma}^\star;\tau^\star)\right\|_2^2-2\lambda_N\sum_{k=r+1}^{m}\rho_k(\boldsymbol{\Gamma}^\star)-6\lambda_N r\gamma_{max}-\lambda_N\left\|\boldsymbol{\Delta}\right\|_*\text{ (using the result following (21))}\\
>&c\phi(\boldsymbol{\Delta}^\star)\eta^\star-2\lambda_N\sum_{k=r+1}^{m}\rho_k(\boldsymbol{\Gamma}^\star)-6\lambda_N r\gamma_{max}-\lambda_N\|\boldsymbol{\Delta}\|_*\geq 0\text{ (by Assumption 2).}
\end{aligned}
$$

This immediately leads to a contradiction, since $\widehat{\boldsymbol{\Gamma}}$ is the minimizer of the loss. ■

Proof of Lemma 6

Proof Without loss of generality, we consider $t_i=i/N$. Since $\mathcal{R}_N(\boldsymbol{\Gamma}^\star,\tau,\tau^\star)=-\mathcal{R}_N(\boldsymbol{\Gamma}^\star,\tau^\star,\tau)$, we only need to consider one side, for example, $\tau\leq\tau^\star$. Then we have

$$
\mathcal{R}_N(\boldsymbol{\Gamma}^\star,\tau,\tau^\star)=N^{-1}\sum_{\tau<i/N\leq\tau^\star}\epsilon_i\left\langle\boldsymbol{X}_i\,,\,\boldsymbol{\Delta}^\star\right\rangle.
$$

Now applying Lévy's inequality (see Theorem 29), we have

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{\tau:\tau^\star-c_\tau<\tau\leq\tau^\star}N^{-1}\left|\sum_{\tau<i/N\leq\tau^\star}\epsilon_i\left\langle\boldsymbol{X}_i\,,\,\boldsymbol{\Delta}^\star\right\rangle\right|\geq\lambda_N\sqrt{c_\tau}\|\boldsymbol{\Delta}^\star\|_F\right)\\
\leq\;&2\mathbb{P}\left(\left|N^{-1}\sum_{\tau^\star-c_\tau<i/N\leq\tau^\star+c_\tau}\epsilon_i\left\langle\boldsymbol{X}_i\,,\,\boldsymbol{\Delta}^\star\right\rangle\right|\geq\lambda_N\sqrt{c_\tau}\|\boldsymbol{\Delta}^\star\|_F\right):=\text{I}.
\end{aligned}
$$

Applying Hoeffding's inequality (Vershynin, 2018), with some absolute constant $c > 0$, we have

$$\begin{aligned} \mathrm{I} \; & \leq 2e \cdot \exp\left(- \frac{c\lambda_N^2 \|\boldsymbol{\Delta}^\star\|_F^2 c_\tau}{K^2 N^{-2} \sum_{i:|t_i - \tau^\star| \leq c_\tau} \langle \boldsymbol{X}_i \, , \, \boldsymbol{\Delta}^\star \rangle^2} \right) \\ & = 2e \cdot \exp\left(- \frac{cN\lambda_N^2 \|\boldsymbol{\Delta}^\star\|_F^2}{2K^2 h_N(c_\tau)} \right) \end{aligned}$$

which concludes the proof. ■

Proof of Theorem 18

Proof Lemma 6 and the basic inequality (6) imply that, with probability greater than $1 - \alpha_N - 2e \cdot \exp\left(-c' N \lambda_N^2 / \{K^2 \|\boldsymbol{\Delta}^\star\|_F^{-2} h_N(c_\tau)\}\right)$ for some constant $c' > 0$,

$$\begin{aligned} & \frac{1}{2N} \left\| \mathfrak{X}\left(\widehat{\boldsymbol{\Gamma}}; \widehat{\tau}\right) - \mathfrak{X}(\boldsymbol{\Gamma}^\star; \tau^\star) \right\|_2^2 + \frac{\lambda_N}{2} \left\| \Pi_{\boldsymbol{\Gamma}^\star}^{r\perp} \left(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^\star\right) \right\|_* \\ \leq & 2\lambda_N \left\| \Pi_{\boldsymbol{\Gamma}^\star}^{r\perp}(\boldsymbol{\Gamma}^\star) \right\|_* + \mathcal{R}_N(\boldsymbol{\Gamma}^\star, \widehat{\tau}, \tau^\star) + \frac{3\lambda_N}{2} \left\| \Pi_{\boldsymbol{\Gamma}^\star}^{r} \left(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^\star\right) \right\|_* \\ \leq & 2\lambda_N \|\Pi_{\boldsymbol{\Gamma}^\star}^{r\perp}(\boldsymbol{\Gamma}^\star)\|_* + \frac{3\lambda_N}{2} \left\| \Pi_{\boldsymbol{\Gamma}^\star}^{r} \left(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^\star\right) \right\|_* + \lambda_N \sqrt{c_\tau} \|\boldsymbol{\Delta}^\star\|_F, \end{aligned}$$

which further suggests the error matrix $\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^\star \in \mathcal{C}(r, \delta, \boldsymbol{\Gamma}^\star, \mathbb{T})$ for $\left\| \widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^\star \right\|_F \geq \delta$. Therefore we can apply the RSC condition to obtain

$$\begin{aligned} \kappa(\mathfrak{X}) \left\| \widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^\star \right\|_F^2 & \leq \frac{1}{2N} \left\| \mathfrak{X}\left(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^\star; \widehat{\tau}\right) \right\|_2^2 \\ & \leq \frac{1}{2N} \left\| \mathfrak{X}\left(\widehat{\boldsymbol{\Gamma}}; \widehat{\tau}\right) - \mathfrak{X}(\boldsymbol{\Gamma}^\star; \tau^\star) + \mathfrak{X}(\boldsymbol{\Gamma}^\star; \tau^\star) - \mathfrak{X}(\boldsymbol{\Gamma}^\star; \widehat{\tau}) \right\|_2^2 \\ & \leq \frac{1}{2N} \left\| \mathfrak{X}\left(\widehat{\boldsymbol{\Gamma}}; \widehat{\tau}\right) - \mathfrak{X}(\boldsymbol{\Gamma}^\star; \tau^\star) \right\|_2^2 + \frac{1}{N} \left\langle \mathfrak{X}\left(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^\star; \widehat{\tau}\right) \, , \, \mathfrak{X}(\boldsymbol{\Gamma}^\star; \tau^\star) - \mathfrak{X}(\boldsymbol{\Gamma}^\star; \widehat{\tau}) \right\rangle \\ & \quad \text{(using } \langle \boldsymbol{a} + \boldsymbol{b} \, , \, \boldsymbol{a} + \boldsymbol{b} \rangle = \langle \boldsymbol{a} \, , \, \boldsymbol{a} \rangle + 2 \langle \boldsymbol{a} \, , \, \boldsymbol{b} \rangle + \langle \boldsymbol{b} \, , \, \boldsymbol{b} \rangle \leq \langle \boldsymbol{a} \, , \, \boldsymbol{a} \rangle + 2 \langle \boldsymbol{a} + \boldsymbol{b} \, , \, \boldsymbol{b} \rangle \text{)} \\ & \leq \frac{1}{2N} \left\| \mathfrak{X}\left(\widehat{\boldsymbol{\Gamma}}; \widehat{\tau}\right) - \mathfrak{X}(\boldsymbol{\Gamma}^\star; \tau^\star) \right\|_2^2 + C c_{\boldsymbol{\Gamma}} c_\tau \|\boldsymbol{\Delta}^\star\|_* \text{ (by Assumption 3)} \\ & \leq 2\lambda_N \left\| \Pi_{\boldsymbol{\Gamma}^\star}^{r\perp}(\boldsymbol{\Gamma}^\star) \right\|_* + \frac{3\lambda_N}{2} \left\| \Pi_{\boldsymbol{\Gamma}^\star}^{r} \left(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^\star\right) \right\|_* + \lambda_N \sqrt{c_\tau} \|\boldsymbol{\Delta}^\star\|_F + C c_{\boldsymbol{\Gamma}} c_\tau \|\boldsymbol{\Delta}^\star\|_* \\ & \quad \text{(using the basic inequality)} \\ & \leq 2\lambda_N \|\Pi_{\boldsymbol{\Gamma}^\star}^{r\perp}(\boldsymbol{\Gamma}^\star)\|_* + \frac{3\sqrt{2r}\lambda_N}{2} \|\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^\star\|_F + \lambda_N \sqrt{c_\tau} \|\boldsymbol{\Delta}^\star\|_F + C c_{\boldsymbol{\Gamma}} c_\tau \|\boldsymbol{\Delta}^\star\|_* \\ & \quad \text{(by the implication of (21)).} \end{aligned}$$

To sum up, we've proved the following bound:

$$\left\| \widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^\star \right\|_F^2 \leq \delta^2 \vee \frac{72\lambda_N^2 r}{\kappa(\mathfrak{X})^2} \vee \frac{8\lambda_N \|\Pi_{\boldsymbol{\Gamma}^\star}^{r\perp}(\boldsymbol{\Gamma}^\star)\|_*}{\kappa(\mathfrak{X})} \vee \frac{4\lambda_N \sqrt{c_\tau} \|\boldsymbol{\Delta}^\star\|_F}{\kappa(\mathfrak{X})} \vee \frac{4C c_{\boldsymbol{\Gamma}} c_\tau \|\boldsymbol{\Delta}^\star\|_*}{\kappa(\mathfrak{X})}. \tag{26}$$

For the prediction error, we obtain

$$\begin{aligned}
&\frac{1}{2N}\left\|\mathfrak{X}\left(\widehat{\mathbf{\Gamma}};\widehat{\tau}\right)-\mathfrak{X}(\mathbf{\Gamma}^\star;\tau^\star)\right\|_2^2\\
\leq&2\lambda_N\|\Pi_{\mathbf{\Gamma}^\star}^{r\perp}(\mathbf{\Gamma}^\star)\|_*+\frac{3\lambda_N}{2}\left\|\Pi_{\mathbf{\Gamma}^\star}^{r}\left(\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}^\star\right)\right\|_*+\lambda_N\sqrt{c_\tau}\|\mathbf{\Delta}^\star\|_F\\
\leq&2\lambda_N\|\Pi_{\mathbf{\Gamma}^\star}^{r\perp}(\mathbf{\Gamma}^\star)\|_*+\frac{3\lambda_N\sqrt{2r}}{2}\left\|\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}^\star\right\|_F+\lambda_N\sqrt{c_\tau}\|\mathbf{\Delta}^\star\|_F\\
\leq&6\lambda_N\|\Pi_{\mathbf{\Gamma}^\star}^{r\perp}(\mathbf{\Gamma}^\star)\|_*\vee 3\lambda_N\sqrt{c_\tau}\|\mathbf{\Delta}^\star\|_F\vee\frac{9\lambda_N\sqrt{2r}}{2}\left\|\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}^\star\right\|_F \qquad (27)\\
\leq&\frac{9\lambda_N\sqrt{2r}}{2}\delta\vee\frac{54\lambda_N^2 r}{\kappa(\mathfrak{X})}\vee 3\lambda_N\sqrt{c_\tau}\|\mathbf{\Delta}^\star\|_F\vee 6\lambda_N\|\Pi_{\mathbf{\Gamma}^\star}^{r\perp}(\mathbf{\Gamma}^\star)\|_*\vee 9\lambda_N\sqrt{\frac{2Cc_{\mathbf{\Gamma}}c_\tau r\|\mathbf{\Delta}^\star\|_*}{\kappa(\mathfrak{X})}} \text{ (applying (26)).}
\end{aligned}$$

For the estimation error in nuclear norm, we have

$$\begin{aligned}
\left\|\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}^\star\right\|_*\leq&\left\|\Pi_{\mathbf{\Gamma}^\star}^{r\perp}\left(\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}^\star\right)\right\|_*+\left\|\Pi_{\mathbf{\Gamma}^\star}^{r}\left(\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}^\star\right)\right\|_*\\
\leq&4\|\Pi_{\mathbf{\Gamma}^\star}^{r\perp}(\mathbf{\Gamma}^\star)\|_*+4\left\|\Pi_{\mathbf{\Gamma}^\star}^{r}\left(\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}^\star\right)\right\|_*+2\sqrt{c_\tau}\|\mathbf{\Delta}^\star\|_F\\
&\text{(by the basic inequality (6) and Lemma 6)}\\
\leq&4\|\Pi_{\mathbf{\Gamma}^\star}^{r\perp}(\mathbf{\Gamma}^\star)\|_*+4\sqrt{2r}\left\|\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}^\star\right\|_F+2\sqrt{c_\tau}\|\mathbf{\Delta}^\star\|_F\\
\leq&12\|\Pi_{\mathbf{\Gamma}^\star}^{r\perp}(\mathbf{\Gamma}^\star)\|_*\vee 12\sqrt{2r}\left\|\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}^\star\right\|_F\vee 6\sqrt{c_\tau}\|\mathbf{\Delta}^\star\|_F \qquad (28)\\
\leq&12\|\Pi_{\mathbf{\Gamma}^\star}^{r\perp}(\mathbf{\Gamma}^\star)\|_*\vee 6\sqrt{c_\tau}\|\mathbf{\Delta}^\star\|_F\vee 12\sqrt{2r}\delta\vee\frac{192\lambda_N r}{\kappa(\mathfrak{X})}\vee 24\sqrt{\frac{2Cc_{\mathbf{\Gamma}}c_\tau r\|\mathbf{\Delta}^\star\|_*}{\kappa(\mathfrak{X})}} \text{ (applying (26)).}
\end{aligned}$$

■

Proof of Theorem 19

Proof The proof proceeds in a similar spirit to that of Lemma 5. Recall we have shown that $\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}^\star\in\mathcal{C}(r,\delta,\mathbf{\Gamma}^\star,\mathbb{T})$. Suppose that $|\widehat{\tau}-\tau^\star|>\eta^\star$. Then

$$\begin{aligned}
&\frac{1}{2N}\left\|\boldsymbol{y}-\mathfrak{X}\left(\widehat{\mathbf{\Gamma}};\widehat{\tau}\right)\right\|_2^2+\lambda_N\left\|\widehat{\mathbf{\Gamma}}\right\|_*-\frac{1}{2N}\|\boldsymbol{y}-\mathfrak{X}(\mathbf{\Gamma}^\star;\tau^\star)\|_2^2-\lambda_N\|\mathbf{\Gamma}^\star\|_*\\
=&\frac{1}{2N}\left\|\mathfrak{X}\left(\widehat{\mathbf{\Gamma}};\widehat{\tau}\right)-\mathfrak{X}(\mathbf{\Gamma}^\star;\tau^\star)\right\|_2^2-\frac{1}{N}\sum_{i=1}^N\epsilon_i\left(\left\langle\boldsymbol{\mathcal{X}}_i(\widehat{\tau}),\widehat{\mathbf{\Gamma}}\right\rangle-\langle\boldsymbol{\mathcal{X}}_i(\widehat{\tau}),\mathbf{\Gamma}^\star\rangle\right)\\
&-\mathcal{R}_N(\mathbf{\Gamma}^\star,\widehat{\tau},\tau^\star)+\lambda_N\left\|\widehat{\mathbf{\Gamma}}\right\|_*-\lambda_N\|\mathbf{\Gamma}^\star\|_*\\
\geq&\frac{1}{2N}\left\|\mathfrak{X}\left(\widehat{\mathbf{\Gamma}};\widehat{\tau}\right)-\mathfrak{X}(\mathbf{\Gamma}^\star;\tau^\star)\right\|_2^2-\frac{\lambda_N}{2}\|\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}^\star\|_*-\mathcal{R}_N(\mathbf{\Gamma}^\star,\widehat{\tau},\tau^\star)+\lambda_N\left\|\widehat{\mathbf{\Gamma}}\right\|_*-\lambda_N\|\mathbf{\Gamma}^\star\|_*\\
\geq&\frac{1}{2N}\left\|\mathfrak{X}\left(\widehat{\mathbf{\Gamma}};\widehat{\tau}\right)-\mathfrak{X}(\mathbf{\Gamma}^\star;\tau^\star)\right\|_2^2-\frac{3\lambda_N}{2}\left\|\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}^\star\right\|_*-\mathcal{R}_N(\mathbf{\Gamma}^\star,\widehat{\tau},\tau^\star)\\
\geq&c\phi(\mathbf{\Delta}^\star)\eta^\star-\frac{3\lambda_N}{2}c_{\mathbf{\Gamma}}-\lambda_N\sqrt{c_\tau}\|\mathbf{\Delta}^\star\|_F.
\end{aligned}$$

where the second inequality follows from (20), and the last from Assumption 2. This immediately leads to a contradiction, since $\widehat{\mathbf{\Gamma}}$ is the minimizer of the loss.

■

Proof of Theorem 20

Proof Let $\rho_{max}(\boldsymbol{M})$ and $\rho_{min}(\boldsymbol{M})$ denote the maximal and minimal singular value of a matrix $\boldsymbol{M}$. Denote $\boldsymbol{\mathcal{X}}(\tau) = (\boldsymbol{\mathcal{X}}_i(\tau))$. We have

$$\begin{aligned}\frac{1}{2n}\|\mathfrak{X}(\mathbf{\Gamma};\tau)\|_2^2 &= \frac{1}{2n}\sum_{j=1}^{m_2}\|\boldsymbol{\mathcal{X}}(\tau)\mathbf{\Gamma}_{\cdot j}\|_2^2\\ &\geq \frac{1}{2n}\sum_{j=1}^{m_2}\rho_{min}(\boldsymbol{\mathcal{X}}(\tau))\,\|\mathbf{\Gamma}_{\cdot j}\|_2^2\\ &= \rho_{min}\left(\frac{1}{2n}\boldsymbol{\mathcal{X}}(\tau)\right)\|\mathbf{\Gamma}\|_F^2.\end{aligned}$$

It suffices to find a lower bound on $\rho_{min}\left((2n)^{-1}\boldsymbol{\mathcal{X}}(\tau)\right)$ that holds uniformly for $\tau \in \mathbb{T}$.

Let

$$\begin{aligned}\widehat{\mathbf{\Sigma}}(\tau) &= n^{-1}\boldsymbol{\mathcal{X}}(\tau)^\top\boldsymbol{\mathcal{X}}(\tau)\\ &= \begin{bmatrix} n^{-1}\sum_{a=1}^n \boldsymbol{x}_a\boldsymbol{x}_a^\top & n^{-1}\sum_{a=1}^n \boldsymbol{x}_a\boldsymbol{x}_a^\top\mathbf{1}\{t_a > \tau\}\\ n^{-1}\sum_{a=1}^n \boldsymbol{x}_a\boldsymbol{x}_a^\top\mathbf{1}\{t_a > \tau\} & n^{-1}\sum_{a=1}^n \boldsymbol{x}_a\boldsymbol{x}_a^\top\mathbf{1}\{t_a > \tau\}\end{bmatrix}\\ &:= \begin{bmatrix}\widehat{\mathbf{\Sigma}}_S & \widehat{\mathbf{\Sigma}}_S(\tau)\\ \widehat{\mathbf{\Sigma}}_S(\tau) & \widehat{\mathbf{\Sigma}}_S(\tau)\end{bmatrix}.\end{aligned}$$

It's not hard to show that

$$\mathbf{\Sigma}(\tau) := \mathrm{Var}\{\boldsymbol{\mathcal{X}}_a(\tau)\} = \begin{bmatrix}1 & 1-\tau\\ 1-\tau & 1-\tau\end{bmatrix}\otimes\mathbf{\Sigma} := \boldsymbol{V}(\tau)\otimes\mathbf{\Sigma}.$$

Over $\mathbb{T} = [\rho, 1-\rho]$ one can show that there exists $\underline{\rho}, \overline{\rho} > 0$, such that

$$\inf_{\tau\in\mathbb{T}}\rho_{min}\left\{\boldsymbol{V}(\tau)\right\} \geq \underline{\rho},\ \ \sup_{\tau\in\mathbb{T}}\rho_{max}\left\{\boldsymbol{V}(\tau)\right\} \leq \overline{\rho}.$$

The spectrum property of Kronecker product then guarantees a pair of uniform bounds on $\mathbf{\Sigma}(\tau)$:

$$\inf_{\tau\in\mathbb{T}}\rho_{\min}\left\{\mathbf{\Sigma}(\tau)\right\} \geq \underline{\rho\sigma}^2,\ \ \sup_{\tau\in\mathbb{T}}\rho_{\max}\left\{\mathbf{\Sigma}(\tau)\right\} \leq \overline{\rho\sigma}^2.$$

By Theorem 25,

$$\sup_{\tau\in\mathbb{T}}\left\|\widehat{\mathbf{\Sigma}}(\tau) - \mathbf{\Sigma}(\tau)\right\|_{\mathrm{op}} \leq \underbrace{\left\|\widehat{\mathbf{\Sigma}}_S - \mathbf{\Sigma}\right\|_{\mathrm{op}}}_{\text{(I)}} + 3\underbrace{\sup_{\tau\in\mathbb{T}}\left\|\widehat{\mathbf{\Sigma}}_S(\tau) - (1-\tau)\mathbf{\Sigma}\right\|_{\mathrm{op}}}_{\text{(II)}}. \tag{29}$$

Applying Theorem 28 (see Theorem 6.5 of Wainwright (2019)), for (I) we have

$$\mathbb{P}\left(\overline{\sigma}^{-2}\left\|\widehat{\mathbf{\Sigma}}_S-\mathbf{\Sigma}\right\|_{\mathrm{op}}\geq C_1\left\{\sqrt{\frac{m_1}{n}}+\frac{m_1}{n}\right\}+\delta\right)\leq C_2\exp(-C_3 n\min\{\delta,\delta^2\}).$$

Take $\delta=C\sqrt{m_1/n}$. When $n>Cm_1$, we have

$$\mathbb{P}\left(\overline{\sigma}^{-2}\left\|\widehat{\mathbf{\Sigma}}_S-\mathbf{\Sigma}\right\|_{\mathrm{op}}\geq C_1\sqrt{\frac{m_1}{n}}\right)\leq C_2\exp(-C_3 m_1).$$

This implies with probability greater than $1-C_2\exp(-C_3m_1)$

$$\left\|\widehat{\mathbf{\Sigma}}_S-\mathbf{\Sigma}\right\|_{\mathrm{op}}\leq C'\overline{\rho}\overline{\sigma}^2.$$

For (II), we first make one step of discretization. Let $\mathbb{T}_S=\{i\cdot 10^{-m_1},\ i=1,\dots,10^{m_1}\}\cap\mathbb{T}$. Notice that

$$\begin{aligned}(\mathrm{II})\leq&\underbrace{\sup_{\tau\in\mathbb{T}_S}\left\|\widehat{\mathbf{\Sigma}}_S(\tau)-(1-\tau)\mathbf{\Sigma}\right\|_{\mathrm{op}}}_{(\mathrm{III})}\\&+\underbrace{\sup_{\tau,\tau'\in\mathbb{T},|\tau-\tau'|\leq 10^{-m_1}}\left\|\left(\widehat{\mathbf{\Sigma}}_S(\tau)-(1-\tau)\mathbf{\Sigma}\right)-\left(\widehat{\mathbf{\Sigma}}_S(\tau')-(1-\tau')\mathbf{\Sigma}\right)\right\|_{\mathrm{op}}}_{(\mathrm{IV})}.\end{aligned}\tag{30}$$

For (III), we first show that for each fixed $\tau\in\mathbb{T}$, $\boldsymbol{x}_a(\tau)$ are i.i.d. mean zero sub-Guassian vectors with parameter $\overline{\sigma}^2$. For $v\in\mathbb{R}^{m_1}$ with $\|v\|_2=1$,

$$\begin{aligned}\mathbb{E}\left\{\exp\left(\lambda\boldsymbol{x}_a(\tau)^\top v\right)\right\}&=\tau+(1-\tau)\mathbb{E}\left\{\exp\left(\lambda\boldsymbol{x}_a^\top v\right)\right\}\\&\leq\tau+(1-\tau)\exp\left(\frac{\lambda^2\overline{\sigma}^2}{2}\right)\\&\leq\exp\left(\frac{\lambda^2\overline{\sigma}^2}{2}\right),\end{aligned}$$

which concludes that $\boldsymbol{x}_a(\tau)\sim\mathrm{SG}(\overline{\sigma}^2)$. Applying Theorem 28 to each $\tau\in\mathbb{T}_S$, we have

$$\mathbb{P}\left(\left(\overline{\rho}\overline{\sigma}^2\right)^{-1}\left\|\widehat{\mathbf{\Sigma}}_S(\tau)-(1-\tau)\mathbf{\Sigma}\right\|_{\mathrm{op}}\geq C_1\left\{\sqrt{\frac{m_1}{n}}+\frac{m_1}{n}\right\}+\delta\right)\leq C_2\exp(-C_3 n\min\{\delta,\delta^2\}).$$

When $n>Cm_1$, again choosing $\delta=C\sqrt{m_1/n}$, we have

$$\mathbb{P}\left(\left(\overline{\rho}\overline{\sigma}^2\right)^{-1}\left\|\widehat{\mathbf{\Sigma}}_S(\tau)-(1-\tau)\mathbf{\Sigma}\right\|_{\mathrm{op}}\geq C_1\sqrt{\frac{m_1}{n}}\right)\leq C_2\exp(-C_3 m_1).$$

Then with probability greater than $1-C_2\exp(-C_3m_1)$

$$\left\|\widehat{\mathbf{\Sigma}}_S(\tau)-(1-\tau)\mathbf{\Sigma}\right\|_{\mathrm{op}}\leq C'\overline{\rho}\overline{\sigma}^2.$$

Taking union bound over $\mathbb{T}_S$ with $|\mathbb{T}_S| = 10^{m_1}$, if we choose C_3 to be large enough, with probability at least $1 - C_2 \exp(-C_3 m_1)$

$$\sup_{\tau \in \mathbb{T}_S} \left\| \widehat{\mathbf{\Sigma}}_S(\tau) - (1-\tau)\mathbf{\Sigma} \right\|_{\text{op}} \leq C' \overline{\rho\sigma}^2.$$

It remains to bound (IV). Suppose $\tau > \tau'$. Note $\widehat{\mathbf{\Sigma}}_S(\tau) - \widehat{\mathbf{\Sigma}}_S(\tau')$ is equivalent to $\widehat{\mathbf{\Sigma}}_S\{1 - (\tau - \tau')\}$ in distribution. The intuition is (IV) only concentrates on small intervals whose length is controlled under 10^{-m_1}, which is negligible:

$$\begin{aligned} & \sup_{\tau,\tau' \in \mathbb{T}, |\tau-\tau'| \leq 10^{-m_1}} \left\| \left(\widehat{\mathbf{\Sigma}}_S(\tau) - (1-\tau)\mathbf{\Sigma} \right) - \left(\widehat{\mathbf{\Sigma}}_S(\tau') - (1-\tau')\mathbf{\Sigma} \right) \right\|_{\text{op}} \\ \overset{d}{=} & \sup_{\tau,\tau' \in \mathbb{T}, |\tau-\tau'| \leq 10^{-m_1}} \left\| \widehat{\mathbf{\Sigma}}_S\{1 - (\tau - \tau')\} - (\tau - \tau')\mathbf{\Sigma} \right\|_{\text{op}} \\ \leq & \underbrace{\sup_{\tau,\tau' \in \mathbb{T}, |\tau-\tau'| \leq 10^{-m_1}} \left\| \widehat{\mathbf{\Sigma}}_S\{1 - (\tau - \tau')\} \right\|_{\text{op}}}_{\text{(IV.1)}} + \sigma_0^2 10^{-m_1}. \end{aligned} \tag{31}$$

To bound the term (IV.1), we first find a $1/4$ net $\mathcal{B}$ for the unit sphere $\mathbb{S}^{m_1-1}$ with cardinality $\left|\mathbb{S}^{m_1-1}\right| \leq 9^{m_1}$. Exercise 4.4.3 in Vershynin (2018) asserts that

$$\sup_{\tau,\tau' \in \mathbb{T}, |\tau-\tau'| \leq 10^{-m_1}} \left\| \widehat{\mathbf{\Sigma}}_S\{1 - (\tau - \tau')\} \right\|_{\text{op}} \leq 2 \underbrace{\max_{v \in \mathcal{B}} \sup_{\tau,\tau' \in \mathbb{T}, |\tau-\tau'| \leq 10^{-m_1}} \frac{1}{n} \sum_{a=1}^{n} v^\top \boldsymbol{x}_a \boldsymbol{x}_a^\top v \mathbf{1}\{\tau' < t_a < \tau\}}_{\text{(IV.2)}}. \tag{32}$$

Let $\mathcal{E}_k$ denotes the event that there exists a subset of $\{t_a, a = 1, \ldots, n\}$ with k elements that belong to one interval $(\tau', \tau) \subset \mathbb{T}$ with length smaller than 10^{-m_1}, that is,

$$\mathcal{E}_k = \{\exists (\tau', \tau) \subset \mathbb{T}, \text{ such that } |(\tau', \tau) \cap \{t_a, a = 1, \ldots, n\}| = k\}.$$

Besides we let $\mathcal{E}_{n+1} = \varnothing$. Now for each $v \in \mathcal{B}$,

$$\begin{aligned} & \sup_{\tau,\tau' \in \mathbb{T}, |\tau-\tau'| \leq 10^{-m_1}} \frac{1}{n} \sum_{a=1}^{n} v^\top \boldsymbol{x}_a \boldsymbol{x}_a^\top v \mathbf{1}\{\tau' < t_a < \tau\} \\ \leq & \frac{\max_{a=1,\ldots,n}(v^\top \boldsymbol{x}_a \boldsymbol{x}_a^\top v)}{n} \cdot \sup_{\tau,\tau' \in \mathbb{T}, |\tau-\tau'| \leq 10^{-m_1}} \left(\sum_{a=1}^{n} \mathbf{1}\{\tau' < t_a < \tau\} \right) \\ = & \frac{\max_{a=1,\ldots,n}(v^\top \boldsymbol{x}_a \boldsymbol{x}_a^\top v)}{n} \cdot \left(\sum_{k=1}^{n} k \cdot \mathbf{1}\{\mathcal{E}_k \backslash \mathcal{E}_{k+1}\} \right) \\ \leq & \underbrace{\frac{\max_{a=1,\ldots,n}(v^\top \boldsymbol{x}_a \boldsymbol{x}_a^\top v)}{n}}_{\text{(V)}} + \underbrace{\frac{\max_{a=1,\ldots,n}(v^\top \boldsymbol{x}_a \boldsymbol{x}_a^\top v)}{n} \left(\sum_{k=1}^{n} k \cdot \mathbf{1}\{\mathcal{E}_k \backslash \mathcal{E}_{k+1}\} \right)}_{\text{(VI)}}. \end{aligned} \tag{33}$$

Clearly we have $\mathcal{E}_{n+1} \subset \mathcal{E}_n \subset \cdots \subset \mathcal{E}_2 \subset \mathcal{E}_1$. When $k \geq 2$, we have

$$\mathbb{E}\left(\mathbf{1}\{\mathcal{E}_k \backslash \mathcal{E}_{k+1}\}\right) \leq \mathbb{E}\left(\mathbf{1}\{\mathcal{E}_k\}\right) \leq \mathbb{E}\left(\mathbf{1}\{\mathcal{E}_2\}\right) \leq \binom{n}{2} \int_{|\tau-\tau'|<10^{-m_1}} 1 \; d\tau d\tau' \leq 10^{-m_1} n^2.$$

For (V),

$$\mathbb{P}\left(\frac{\max_{a=1,\ldots,n}(v^\top \boldsymbol{x}_a \boldsymbol{x}_a^\top v)}{n} \geq \overline{\sigma}^2 \delta\right) \leq n\mathbb{P}\left(\frac{v^\top \boldsymbol{x}_a \boldsymbol{x}_a^\top v}{n} \geq \overline{\sigma}^2 \delta\right) \leq nC_1 \exp(-C_2 n).$$

Taking $\delta = cm_1/n$ and C_2 large enough, we have

$$\mathbb{P}\left(\frac{\max_{a=1,\ldots,n}(v^\top \boldsymbol{x}_a \boldsymbol{x}_a^\top v)}{n} \geq \frac{c\overline{\sigma}^2 m_1}{n}\right) \leq C_1 \exp(-C_2 m_1).$$

When $n > Cm_1$, we have

$$\mathbb{P}\left(\frac{\max_{a=1,\ldots,n}(v^\top \boldsymbol{x}_a \boldsymbol{x}_a^\top v)}{n} \geq C_1 \overline{\rho\sigma}^2\right) \leq C_1 \exp(-C_2 m_1).$$

For (VI), we directly apply Markov's inequality

$$\mathbb{P}\left[\mathrm{VI} > \delta\right] \leq \delta^{-1} \mathbb{E}\left[\frac{\max_{a=1,\ldots,n}(v^\top \boldsymbol{x}_a \boldsymbol{x}_a^\top v)}{n} \cdot \left(\sum_{k=2}^{n} k \cdot \mathbf{1}\{\mathcal{E}_k \backslash \mathcal{E}_{k+1}\}\right)\right] \leq \delta^{-1} \frac{C\log(n)}{n} n^3 10^{-m_1}.$$

With $\delta = C'\overline{\rho\sigma}^2$ we have

$$\mathbb{P}\left[\mathrm{VI} > C'\overline{\rho\sigma}^2\right] \leq C(\overline{\sigma}^2)^{-1} \log(n) n^2 10^{-m_1}.$$

Combining (V) and (VI) we have

$$\begin{aligned}
&\mathbb{P}\left[\sup_{\tau,\tau' \in \mathbb{T}, |\tau-\tau'| \leq 10^{-m_1}} \frac{1}{n}\sum_{a=1}^{n} v^\top \boldsymbol{x}_a \boldsymbol{x}_a^\top v \mathbf{1}\{\tau' < t_a < \tau\} > C'\overline{\rho\sigma}^2 + C'\overline{\rho\sigma}^2\right] \\
\leq &\mathbb{P}(\mathrm{V} > C'\overline{\rho\sigma}^2) + \mathbb{P}(\mathrm{VI} > C'\overline{\rho\sigma}^2) \\
\leq &C\exp(-Cm_1) + Cm_1^{-1}\log(n) n^3 10^{-m_1}.
\end{aligned}$$

Furthermore, taking union over the net $\mathcal{B}$ (taking the constants and n to large enough if needed), we proved that with probability greater than $1 - C_1\exp(C_2 m_1) - C_3 m_1^{-1}\log(n) n^3 \exp(-C_4 m_1)$,

$$\sup_{\tau,\tau' \in \mathbb{T}, |\tau-\tau'| \leq 10^{-m_1}} \left\| \widehat{\boldsymbol{\Sigma}}_S\{1 - (\tau - \tau')\} \right\|_{\mathrm{op}} \leq C\overline{\sigma}^2.$$

To sum up, we've shown the following uniform control:

$$\begin{aligned}
\sup_{\tau \in \mathbb{T}} \left\| \widehat{\boldsymbol{\Sigma}}(\tau) - \boldsymbol{\Sigma}(\tau) \right\|_{\mathrm{op}} &\leq \mathrm{I} + 3 \times \mathrm{II} && \text{(see (29))} \\
&\leq \mathrm{I} + 3 \times (\mathrm{III} + \mathrm{IV}) && \text{(see (30))} \\
&\leq \mathrm{I} + 3 \times (\mathrm{III} + \mathrm{IV.1} + \sigma_0^2 10^{-m_1}) && \text{(see (31))} \\
&\leq \mathrm{I} + 3 \times (\mathrm{III} + 2 \times \mathrm{IV.2} + \sigma_0^2 10^{-m_1}) && \text{(see (32))} \\
&\leq \mathrm{I} + 3 \times \left\{\mathrm{III} + 2\max_{\mathcal{B}}(\mathrm{V} + \mathrm{VI}) + \overline{\sigma}^2 10^{-m_1}\right\} && \text{(see (33))} \\
&\leq C''\overline{\rho\sigma}^2.
\end{aligned}$$

with probability greater than $1 - C_1 \exp(C_2 m_1) - C_3 \log(n) n^2 \exp(-C_4 m_1)$ provided $n > C_0 m_1$. We can choose C_k to be large enough, we can control

$$C'' < \frac{\underline{\rho}}{2\overline{\rho}}.$$

Then following a similar argument as in Wainwright (2009) we have

$$\begin{aligned}
\rho_{\min}\left\{\widehat{\mathbf{\Sigma}}(\tau)\right\} &= \min_{\beta \in \mathbb{R}^{2m_1}, \|\beta\|_2 = 1} \beta^\top \widehat{\mathbf{\Sigma}}(\tau)\beta \\
&= \min_{\beta \in \mathbb{R}^{2m_1}, \|\beta\|_2 = 1} \left\{\beta^\top \mathbf{\Sigma}(\tau)\beta + \beta^\top \left(\widehat{\mathbf{\Sigma}}(\tau) - \mathbf{\Sigma}(\tau)\right)\beta\right\} \\
&\geq \rho_{\min}\left\{\mathbf{\Sigma}(\tau)\right\} - \left\|\widehat{\mathbf{\Sigma}}(\tau) - \mathbf{\Sigma}(\tau)\right\|_{\mathrm{op}} \geq (\underline{\rho} - C''\overline{\rho})\overline{\sigma}^2 \geq \frac{\underline{\rho}\overline{\sigma}^2}{2},
\end{aligned}$$

and on the other direction,

$$\begin{aligned}
\rho_{\max}\left\{\widehat{\mathbf{\Sigma}}(\tau)\right\} &= \max_{\beta \in \mathbb{R}^{2m_1}, \|\beta\|_2 = 1} \beta^\top \widehat{\mathbf{\Sigma}}(\tau)\beta \\
&= \max_{\beta \in \mathbb{R}^{2m_1}, \|\beta\|_2 = 1} \left\{\beta^\top \mathbf{\Sigma}(\tau)\beta + \beta^\top \left(\widehat{\mathbf{\Sigma}}(\tau) - \mathbf{\Sigma}(\tau)\right)\beta\right\} \\
&\leq \rho_{\max}\left\{\mathbf{\Sigma}(\tau)\right\} + \left\|\widehat{\mathbf{\Sigma}}(\tau) - \mathbf{\Sigma}(\tau)\right\|_{\mathrm{op}} \leq (\overline{\rho} + C''\overline{\rho})\overline{\sigma}^2 \leq \frac{3\overline{\rho}\overline{\sigma}^2}{2}.
\end{aligned}$$

■

Lemma 24 *Suppose $\{X_i\}_{i=1}^n$ are i.i.d. copies of some mean zero sub-exponential random variable X and $\{B_i(\tau), \tau \in \mathbb{T} := (0, \tau_0^\star]\}_{i=1}^n$ are i.i.d copies of the stochastic process $B(\tau) = \mathbf{1}\big\{U < \tau\big\}$, $0 < \tau \leq \tau_0^\star < 1$, where $U \sim$ Uniform$(0,1)$. Assume the following Bernstein type inequality holds:*

$$\forall\ \tau \in \mathbb{T},\ \mathbb{P}\left(\left|\sum_{i=1}^n X_i \cdot B_i(\tau)\right| > \sqrt{c_1 n\tau\delta} + c_2\delta\right) < 2\exp(-\delta).$$

1. *For any integer $d > 0$, with probability greater than $1 - 2\cdot 10^d \exp(-\delta) - 2n\exp(-c'\delta_1) - C\delta_2^{-1}\log(n)n^2 10^{-d}$, it holds uniformly for all $\tau \in \mathbb{T}$ that*

$$\left|\sum_{i=1}^n X_i \cdot B_i(\tau)\right| \leq \frac{1}{2}\left(\sqrt{c_1 n\tau\delta} + c_2\delta\right) + \frac{\delta_1 + \delta_2}{4}.$$

   Here C, c' are constants.

2. *For any integer $d > 0$, with probability greater than $1 - 10^d\exp(-\delta) - C\delta_1^{-1}10^{-d}n^3$, it holds uniformly for all $\tau \in \mathbb{T}$ that*

$$\left|\sum_{i=1}^n \{B_i(\tau) - \tau\}\right| \leq \frac{1}{2}\sqrt{c_1 n\tau\delta} + \frac{1}{2}\sqrt{\frac{c_1 n\delta}{10^d}} + \frac{n10^{-d}}{2} + \frac{1 + \delta_1}{2}.$$

Proof

Part 1. Let $\mathbb{T}_d = \{10^{-d} \cdot k,\ k = 1, \cdots, n\} \cap (0, \tau_0^\star]$. Taking one step discretization, for any $\tau \in \mathbb{T}$, there is $\tau' \in \mathbb{T}_d$, such that $|\tau - \tau'| < 10^{-d}$. Then

$$\left|\sum_{i=1}^n X_i \cdot B_i(\tau)\right| \leq \underbrace{\left|\sum_{i=1}^n X_i \cdot B_i(\tau')\right|}_{\text{(I)}} + \underbrace{\sup_{\substack{0<\tau<\tau'\leq\tau_0^\star,\\ |\tau-\tau'|<10^{-d}}} \left|\sum_{i=1}^n X_i \cdot \{B_i(\tau') - B_i(\tau)\}\right|}_{\text{(II)}}.$$

For (I), simply taking union bound, we have

$$\mathbb{P}\left(\forall\ \tau \in \mathbb{T}_d,\ \left|\sum_{i=1}^n X_i \cdot B_i(\tau)\right| > \sqrt{c_1 n\tau\delta} + c_2\delta\right) \leq 2 \cdot 10^d \exp(-\delta).$$

For (II), we have

$$\sup_{\substack{0<\tau<\tau'\leq\tau_0^\star,\\ |\tau-\tau'|<10^{-d}}} \left|\sum_{i=1}^n X_i \cdot \{B_i(\tau') - B_i(\tau)\}\right| \leq \sup_{\substack{0<\tau<\tau'\leq\tau_0^\star,\\ |\tau-\tau'|<10^{-d}}} \sum_{i=1}^n |X_i| \cdot \mathbf{1}\{\tau \leq U_i < \tau'\}.$$

Let $\mathcal{E}_k$ denotes the event that there exists a subset of $\{U_i, i = 1, \ldots, n\}$ with k elements that belong to one interval $(\tau', \tau) \subset \mathbb{T}$ with length smaller than 10^{-d}, that is,

$$\mathcal{E}_k = \{\exists (\tau', \tau) \subset \mathbb{T},\ \text{such that } |(\tau', \tau) \cap \{U_i, i = 1, \ldots, n\}| = k\}.$$

Besides we let $\mathcal{E}_{n+1} = \varnothing$. Now for each $v \in \mathcal{B}$,

$$\begin{aligned}
&\sup_{\substack{0<\tau<\tau'\leq\tau_0^\star,\\ |\tau-\tau'|<10^{-d}}} \sum_{i=1}^n |X_i| \cdot \mathbf{1}\{\tau \leq U_i < \tau'\} \\
&\leq \max_{i=1,\ldots,n} |X_i| \cdot \sup_{\tau,\tau'\in\mathbb{T}, |\tau-\tau'|\leq 10^{-d}} \left(\sum_{i=1}^n \mathbf{1}\{\tau' < U_i < \tau\}\right) \\
&= \max_{i=1,\ldots,n} |X_i| \cdot \left(\sum_{k=1}^n k \cdot \mathbf{1}\{\mathcal{E}_k \backslash \mathcal{E}_{k+1}\}\right) \\
&\leq \underbrace{\max_{i=1,\ldots,n} |X_i|}_{\text{(III)}} + \underbrace{\max_{i=1,\ldots,n} |X_i| \left(\sum_{k=1}^n k \cdot \mathbf{1}\{\mathcal{E}_k \backslash \mathcal{E}_{k+1}\}\right)}_{\text{(IV)}}.
\end{aligned}$$

Clearly we have $\mathcal{E}_{n+1} \subset \mathcal{E}_n \subset \cdots \subset \mathcal{E}_2 \subset \mathcal{E}_1$. When $k \geq 2$, we have

$$\mathbb{E}\left(\mathbf{1}\{\mathcal{E}_k \backslash \mathcal{E}_{k+1}\}\right) \leq \mathbb{E}\left(\mathbf{1}\{\mathcal{E}_k\}\right) \leq \mathbb{E}\left(\mathbf{1}\{\mathcal{E}_2\}\right) \leq \binom{n}{2} \int_{|\tau-\tau'|<10^{-d}} 1\ d\tau d\tau' \leq 10^{-d} n^2.$$

For (III), by the condition of sub-exponential tail for X,

$$\mathbb{P}\left(\max_{i=1,\ldots,n} |X_i| \geq \delta_1\right) \leq n\mathbb{P}\left(|X_i| \geq \delta_1\right) \leq 2n \exp(-c'\delta_1).$$

For (IV), we directly apply Markov's inequality

$$\mathbb{P}\left(\mathrm{IV} > \delta_2\right) \le \delta_2^{-1}\mathbb{E}\left[\max_{i=1,\cdots,n}|X_i| \cdot \left(\sum_{k=2}^{n} k \cdot \mathbf{1}\{\mathcal{E}_k \backslash \mathcal{E}_{k+1}\}\right)\right] \le \delta_2^{-1}C\log(n)n^3 10^{-d}.$$

To sum up, we've proved that with probability at least

$$1 - 10^d\exp(-\delta) - n\exp(-c'\delta_1) - C\delta_2^{-1}\log(n)n^3 10^{-d},$$

it holds uniformly for $\tau \in \mathbb{T}$ that

$$\left|\sum_{i=1}^{n} X_i \cdot B_i(\tau)\right| \le \frac{1}{2}\left(\sqrt{c_1 n\tau\delta} + c_2\delta\right) + \frac{1}{2}\sqrt{\frac{c_1 n\delta}{10^d}} + \frac{\delta_1 + \delta_2}{4}.$$

Part 2. Following a similar discretization procedure we can prove the second uniform bound, i.e., with probability greater than $1 - 10^d\exp(-\delta) - C\delta_1^{-1}10^{-d}n^3$, it hold uniformly for $\tau \in \mathbb{T}$ that

$$\left|\sum_{i=1}^{n}\{B_i(\tau) - \tau\}\right| \le \frac{1}{2}\sqrt{c_1 n\tau\delta} + \frac{1}{2}\sqrt{\frac{c_1 n\delta}{10^d}} + \frac{n10^{-d}}{2} + \frac{1+\delta_1}{2}.$$

■

Proof of Theorem 21

Proof Note we have

$$\begin{aligned}\left\|n^{-1}\sum_{a=1}^{n}\left(\boldsymbol{x}_a\boldsymbol{x}_a^\top B_a(\tau) - \tau\boldsymbol{\Sigma}\right)\right\|_{\mathrm{op}} &\le \left\|n^{-1}\sum_{a=1}^{n}\left(\boldsymbol{x}_a\boldsymbol{x}_a^\top - \boldsymbol{\Sigma}\right)B_a(\tau)\right\|_{\mathrm{op}} + \left\|n^{-1}\sum_{a=1}^{n}\left(B_a(\tau) - \tau\right)\boldsymbol{\Sigma}\right\|_{\mathrm{op}} \\ &= \left\|n^{-1}\sum_{a=1}^{n}\left(\boldsymbol{x}_a\boldsymbol{x}_a^\top - \boldsymbol{\Sigma}\right)B_a(\tau)\right\|_{\mathrm{op}} + \left|n^{-1}\sum_{a=1}^{n}\left(B_a(\tau) - \tau\right)\right|.\end{aligned}$$

We bound the above to terms respectively.

First we find a 1/4 net $\mathcal{B}$ for the unit sphere $\mathbb{S}^{m_1-1}$ with cardinality $\left|\mathbb{S}^{m_1-1}\right| \le 9^{m_1}$. Exercise 4.4.3 in Vershynin (2018) asserts that

$$\left\|\sum_{a=1}^{n}\left(\boldsymbol{x}_a\boldsymbol{x}_a^\top - \boldsymbol{\Sigma}\right)B_a(\tau)\right\|_{\mathrm{op}} \le 2\max_{\boldsymbol{v}\in\mathbb{S}^{m_1-1}}\left|\sum_{a=1}^{n}\boldsymbol{v}^\top\left(\boldsymbol{x}_a\boldsymbol{x}_a^\top - \boldsymbol{\Sigma}\right)\boldsymbol{v}B_a(\tau)\right|.$$

For each $\boldsymbol{v}$, note $Z_a = \boldsymbol{v}^\top\boldsymbol{x}_a\boldsymbol{x}_a^\top\boldsymbol{v}$ is a scaled chi-squared distribution. We verify the summation satisfies the Bernstein type inequality as in the statement of Theorem 24 for every $\tau \in \mathbb{T}$.

Following the proof of Theorem 6.5 in Wainwright (2019), the moment generating function (MGF) of the summation is bounded by

$$\begin{aligned}
&\mathbb{E}\exp\left\{n^{-1}\sum_{a=1}^{n}u\left(Z_a-\mathbb{E}Z_a\right)B_a(\tau)\right\}\\
=&(\mathbb{E}\exp\left\{un^{-1}\left(Z_a-\mathbb{E}Z_a\right)B_a(\tau)\right\})^n\\
=&(\mathbb{E}_{B_a(\tau)}\mathbb{E}\exp\left\{un^{-1}\left(Z_a-\mathbb{E}Z_a\right)B_a(\tau)\right\})^n\\
\leq&(\mathbb{E}_{B_a(\tau)}\mathbb{E}\exp\left\{2un^{-1}\epsilon Z_aB_a(\tau)\right\})^n.
\end{aligned}$$

Now by performing Taylor's expansion on the exponential function mimicking Wainwright (2019), we can show that

$$\mathbb{E}\exp\left\{n^{-1}\sum_{a=1}^{n}u\left(Z_a-\mathbb{E}Z_a\right)B_a(\tau)\right\}\leq\exp(\frac{C\tau\overline{\sigma}^4u^2}{n}),\quad\text{for all }|u|<\frac{n}{C'\overline{\sigma}^2},\tag{34}$$

which suggests for any positive δ,

$$\mathbb{P}\left(\left|\sum_{a=1}^{n}(Z_a-\mathbb{E}Z_a)\cdot B_a(\tau)\right|>\overline{\sigma}^2\sqrt{c_1n\tau\delta}+c_2\overline{\sigma}^2\delta\right)<2\exp(-\delta).$$

Now apply Theorem 24 with $d=m_1$, $\delta=\delta_1=\delta_2=cm_1$ for some constant $c>0$, and we then have

$$\begin{aligned}
&\mathbb{P}\left(\forall\ \tau\in\mathbb{T},\ \left|n^{-1}\sum_{a=1}^{n}\left(\boldsymbol{v}^\top\boldsymbol{x}_a\boldsymbol{x}_a^\top\boldsymbol{v}-\boldsymbol{v}^\top\boldsymbol{\Sigma}\boldsymbol{v}\right)B_a(\tau)\right|>c'\overline{\sigma}^2\left(\sqrt{\frac{\tau m_1}{n}}+\frac{m_1}{n}\right)\right)\\
<&2\cdot 10^{m_1}\exp(-cm_1)+2n\exp(-cm_1)+C{m_1}^{-1}\log(n)n^3 10^{-m_1}\\
<&C(1+n+m_1^{-1}\log(n)n^3)\exp(-cm_1)\ \text{(for large enough } c,C>0).
\end{aligned}$$

For $\tau\geq\frac{C_0R_qm_1}{n}$ for some $C_0R_q>1$, the above bound implies

$$\begin{aligned}
\mathbb{P}\left(\forall\ \tau\in\mathbb{T},\ \tau\geq\frac{C_0R_qm_1}{n},\ \left|n^{-1}\sum_{a=1}^{n}\left(\boldsymbol{v}^\top\boldsymbol{x}_a\boldsymbol{x}_a^\top\boldsymbol{v}-\boldsymbol{v}^\top\boldsymbol{\Sigma}\boldsymbol{v}\right)B_a(\tau)\right|>c'\overline{\sigma}^2\sqrt{\frac{\tau m_1}{n}}\right)&\\
<C(1+n+m_1^{-1}\log(n)n^3)\exp(-cm_1).&
\end{aligned}$$

Now taking union bound over $\mathcal{B}$ and picking large enough constants c',c,C, we have proved

$$\begin{aligned}
\mathbb{P}\left(\forall\ \tau\in\mathbb{T},\ \tau\geq\frac{C_0R_qm_1}{n},\ \left\|n^{-1}\sum_{a=1}^{n}\left(\boldsymbol{x}_a\boldsymbol{x}_a^\top-\boldsymbol{\Sigma}\right)B_a(\tau)\right\|_{\text{op}}>c'\overline{\sigma}^2\sqrt{\frac{\tau m_1}{n}}\right)&\\
<C(1+n+m_1^{-1}\log(n)n^3)\exp(-cm_1).&
\end{aligned}$$

In words, with probability at least $1-C(1+n+m_1^{-1}\log(n)n^3)\exp(-cm_1)$ it holds uniformly for $\tau\in\mathbb{T}$, $\tau\geq\frac{C_0R_qm_1}{n}$ that

$$\left\|n^{-1}\sum_{a=1}^{n}\left(\boldsymbol{x}_a\boldsymbol{x}_a^\top-\boldsymbol{\Sigma}\right)B_a(\tau)\right\|_{\text{op}}\leq c'\overline{\sigma}^2\sqrt{\frac{\tau m_1}{n}}\leq\frac{c'\overline{\sigma}^2}{\sqrt{C_0R_q}}\tau.$$

Now using Part 2 of Theorem 24 with $d = m_1$ and $\delta = \delta_1 = cm_1$, for the empirical distribution function class $\{B_a(\tau)\}$, with probability greater than

$$1 - 10^{m_1} \exp(-cm_1) - Cm_1^{-1} 10^{-m_1} n^3,$$

it hold uniformly for $\tau \in \mathbb{T}$ that

$$\left| n^{-1} \sum_{a=1}^{n} \{B_a(\tau) - \tau\} \right| \leq \frac{1}{2}\sqrt{\frac{c\tau m_1}{n}} + \frac{cm_1}{2n}.$$

With $\tau \geq \frac{C_0 R_q m_1}{n}$ we further have

$$\left| n^{-1} \sum_{a=1}^{n} \{B_a(\tau) - \tau\} \right| \leq \frac{c'}{\sqrt{C_0 R_q}} \tau.$$

To sum up, we've shown that, for large enough n, m_1 and constants $c, C > 0$, under the condition $C_0 R_q > 1$, with probability greater than $1 - C(1 + n + m_1^{-1} \log(n) n^3) \exp(-cm_1)$, it holds uniformly for all $\tau \in \mathbb{T}, \tau \geq \frac{C_0 R_q m_1}{n}$ that

$$\left\| n^{-1} \sum_{a=1}^{n} \left\{ \boldsymbol{x}_a \boldsymbol{x}_a^\top B_a(\tau) - \tau \boldsymbol{\Sigma} \right\} \right\|_{\text{op}} \leq c' \overline{\sigma}^2 \sqrt{\frac{\tau m_1}{n}} \leq \frac{c' \overline{\sigma}^2}{\sqrt{C_0 R_q}} \tau.$$

■

Proof of Theorem 22

Proof We first consider $\tau < \tau^\star$. The other direction can be proved analogously. Note that

$$\begin{aligned} \frac{1}{2n} \|\mathfrak{X}(\boldsymbol{\Gamma}, \tau) - \mathfrak{X}(\boldsymbol{\Gamma}^\star, \tau^\star)\|_2^2 = & \frac{1}{2n} \sum_{a=1}^{n} \|(\boldsymbol{\Theta} - \boldsymbol{\Theta}^\star)^\top \boldsymbol{x}_a\|^2 \mathbf{1}\{t_a \leq \tau\} && \text{(I)} \\ & + \frac{1}{2n} \sum_{a=1}^{n} \|(\boldsymbol{\Theta} - \boldsymbol{\Theta}^\star + \boldsymbol{\Delta})^\top \boldsymbol{x}_a\|^2 \mathbf{1}\{\tau < t_a \leq \tau^\star\} && \text{(II)} \\ & + \frac{1}{2n} \sum_{a=1}^{n} \|(\boldsymbol{\Theta} - \boldsymbol{\Theta}^\star + \boldsymbol{\Delta} - \boldsymbol{\Delta}^\star)^\top \boldsymbol{x}_a\|^2 \mathbf{1}\{t_a > \tau^\star\}. && \text{(III)} \end{aligned}$$

We hope to show the above summation is larger than $c|\tau^\star - \tau|$ for some positive constant c (independent of τ) with high probability.

Taking expectation, we get

$$\begin{aligned} \mathbb{E}(\text{I}) &= 0.5\tau \left\langle \boldsymbol{\Theta} - \boldsymbol{\Theta}^\star , \boldsymbol{\Theta} - \boldsymbol{\Theta}^\star \right\rangle_{\boldsymbol{\Sigma}}, \\ \mathbb{E}(\text{II}) &= 0.5(\tau^\star - \tau) \left\langle \boldsymbol{\Theta} - \boldsymbol{\Theta}^\star + \boldsymbol{\Delta} , \boldsymbol{\Theta} - \boldsymbol{\Theta}^\star + \boldsymbol{\Delta} \right\rangle_{\boldsymbol{\Sigma}}, \\ \mathbb{E}(\text{III}) &= 0.5(1 - \tau^\star) \left\langle \boldsymbol{\Theta} - \boldsymbol{\Theta}^\star + \boldsymbol{\Delta} - \boldsymbol{\Delta}^\star , \boldsymbol{\Theta} - \boldsymbol{\Theta}^\star + \boldsymbol{\Delta} - \boldsymbol{\Delta}^\star \right\rangle_{\boldsymbol{\Sigma}}. \end{aligned}$$

We consider the concentration of Term I. Similar to the proof of Theorem 20, it can be shown that $\boldsymbol{x}_a \mathbf{1}\{t_a \leq \tau\} \sim \mathrm{SG}(\overline{\sigma}^2)$ with covariance $\tau\boldsymbol{\Sigma}$. Provided $n > Cm_1$ for some $C > 0$, with probability greater than $1 - C_1 \exp(-C_2 n)$,

$$\left\| \frac{1}{2n} \sum_{a=1}^{n} \boldsymbol{x}_a \boldsymbol{x}_a^\top \mathbf{1}\{t_a \leq \tau\} - 0.5\tau\boldsymbol{\Sigma} \right\|_{\mathrm{op}} \leq C'\tau, \ \forall\ \tau \in \mathbb{T}.$$

which implies

$$\begin{aligned} |\mathrm{I} - \mathbb{E}(\mathrm{I})| &\leq \left\| \frac{1}{2n} \sum_{a=1}^{n} \boldsymbol{x}_a \boldsymbol{x}_a^\top \mathbf{1}\{t_a \leq \tau\} - 0.5\tau\boldsymbol{\Sigma} \right\|_{\mathrm{op}} \cdot \langle \boldsymbol{\Theta} - \boldsymbol{\Theta}^\star \,,\, \boldsymbol{\Theta} - \boldsymbol{\Theta}^\star \rangle_{\boldsymbol{\Sigma}} \\ &\leq C'\tau \,\langle \boldsymbol{\Theta} - \boldsymbol{\Theta}^\star \,,\, \boldsymbol{\Theta} - \boldsymbol{\Theta}^\star \rangle_{\boldsymbol{\Sigma}} \,. \end{aligned}$$

Now

$$\mathrm{I} = \mathbb{E}(\mathrm{I}) + \mathrm{I} - \mathbb{E}(\mathrm{I}) \geq \mathbb{E}(\mathrm{I}) - |\mathrm{I} - \mathbb{E}(\mathrm{I})| \geq C'\tau \,\langle \boldsymbol{\Theta} - \boldsymbol{\Theta}^\star \,,\, \boldsymbol{\Theta} - \boldsymbol{\Theta}^\star \rangle_{\boldsymbol{\Sigma}}$$

with probability at least $1 - C_1 \exp(-C_2 n)$. Similar results hold for the other terms too. To sum up we proved that, provided $n > Cm_1$, with probability at least $1 - C_1 \exp(-C_2 n)$

$$\begin{aligned} \mathrm{I} &\geq C'\tau \,\langle \boldsymbol{\Theta} - \boldsymbol{\Theta}^\star \,,\, \boldsymbol{\Theta} - \boldsymbol{\Theta}^\star \rangle_{\boldsymbol{\Sigma}} \,, \\ \mathrm{II} &\geq C'(\tau^\star - \tau) \,\langle \boldsymbol{\Theta} - \boldsymbol{\Theta}^\star + \boldsymbol{\Delta} \,,\, \boldsymbol{\Theta} - \boldsymbol{\Theta}^\star + \boldsymbol{\Delta} \rangle_{\boldsymbol{\Sigma}} \,, \\ \mathrm{III} &\geq C'(1 - \tau^\star) \,\langle \boldsymbol{\Theta} - \boldsymbol{\Theta}^\star + \boldsymbol{\Delta} - \boldsymbol{\Delta}^\star \,,\, \boldsymbol{\Theta} - \boldsymbol{\Theta}^\star + \boldsymbol{\Delta} - \boldsymbol{\Delta}^\star \rangle_{\boldsymbol{\Sigma}} \,. \end{aligned}$$

Conditioning on the above event and taking summation over the three lower bounds, we see that

$$\begin{aligned} \mathrm{I} + \mathrm{II} + \mathrm{III} &\geq C'\{(\tau^\star - \tau) \,\langle \boldsymbol{\Theta} - \boldsymbol{\Theta}^\star - \boldsymbol{\Delta}^\star \,,\, \boldsymbol{\Theta} - \boldsymbol{\Theta}^\star - \boldsymbol{\Delta}^\star \rangle_{\boldsymbol{\Sigma}} \\ &\quad + (1 - \tau^\star) \,\langle \boldsymbol{\Theta} - \boldsymbol{\Theta}^\star + \boldsymbol{\Delta} - \boldsymbol{\Delta}^\star \,,\, \boldsymbol{\Theta} - \boldsymbol{\Theta}^\star + \boldsymbol{\Delta} - \boldsymbol{\Delta}^\star \rangle_{\boldsymbol{\Sigma}}\} \\ &\geq C'\underline{\sigma}^2 \frac{(\tau^\star - \tau)(1 - \tau^\star)}{1 - \tau} \|\boldsymbol{\Delta}^\star\|_F^2 \text{ (using Theorem 30)} \\ &\geq C'\underline{\sigma}^2 \frac{\rho(\tau^\star - \tau)}{1 - \rho} \|\boldsymbol{\Delta}^\star\|_F^2 = \frac{C'\underline{\sigma}^2\rho}{1 - \rho} (\tau^\star - \tau) \|\boldsymbol{\Delta}^\star\|_F^2 \text{ (using } \tau \in \mathbb{T} = [\rho, 1 - \rho]). \end{aligned}$$

To sum up, we can pick $c = \frac{C'\underline{\sigma}^2\rho}{1-\rho}\|\boldsymbol{\Delta}^\star\|_F^2 > 0$ since we assumed $\boldsymbol{\Delta}^\star \neq \mathbf{0}$. Then the result holds with probability at least $1 - C_1 \exp(-C_2 n)$.

■

Proof of Theorem 23

Proof Without loss of generality, we only show the case where $\tau < \tau^\star$. Some algebra leads to

$$
\begin{aligned}
&|\mathcal{T}_N(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}^\star, \tau, \tau^\star)| \\
&\quad = \left| n^{-1} \sum_{a=1}^{n} \boldsymbol{x}_a^\top \{(\boldsymbol{\Theta} - \boldsymbol{\Theta}^\star) + (\boldsymbol{\Delta} - \boldsymbol{\Delta}^\star)\} \boldsymbol{\Delta}^{\star\top} \boldsymbol{x}_a \mathbf{1}\{\tau < t_i < \tau^\star\} \right| \\
&\quad = \left| \mathrm{tr}\left(\{(\boldsymbol{\Theta} - \boldsymbol{\Theta}^\star) + (\boldsymbol{\Delta} - \boldsymbol{\Delta}^\star)\}^\top \left[n^{-1} \sum_{i=1}^{n} \boldsymbol{x}_a \boldsymbol{x}_a^\top \mathbf{1}\{\tau < t_i < \tau^\star\} \right] \boldsymbol{\Delta}^\star \right) \right| \\
&\quad \le \left\| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_a \boldsymbol{x}_a^\top \mathbf{1}\{\tau < t_i < \tau^\star\} \right\|_{\mathrm{op}} \cdot \|(\boldsymbol{\Theta} - \boldsymbol{\Theta}^\star) + (\boldsymbol{\Delta} - \boldsymbol{\Delta}^\star)\|_F \cdot \|\boldsymbol{\Delta}^\star\|_F \qquad (35) \\
&\quad \le C'\overline{\sigma}^2 |\tau - \tau^\star| \cdot \|\boldsymbol{\Delta}^\star\|_* \cdot \|(\boldsymbol{\Theta} - \boldsymbol{\Theta}^\star) + (\boldsymbol{\Delta} - \boldsymbol{\Delta}^\star)\|_* \text{ (using the bounded moment condition)} \\
&\quad \le C'\overline{\sigma}^2 |\tau - \tau^\star| \cdot \|\boldsymbol{\Delta}^\star\|_* \cdot \sqrt{2} \|\boldsymbol{\Gamma} - \boldsymbol{\Gamma}^\star\|_* \text{ (by Theorem 25).}
\end{aligned}
$$

Note Step (35) applies previous results about uniform convergence of the sample matrix convergence, and the operator norm is bounded by $C'\overline{\sigma}^2|\tau - \tau^\star|$ with probability at least $1 - C_1 \exp(-C_2 n)$. That is, we have

$$
\mathbb{P}\left(|\mathcal{T}_N(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}^\star, \tau, \tau^\star)| \le C\overline{\sigma}^2 |\tau - \tau^\star| \cdot \|\boldsymbol{\Gamma} - \boldsymbol{\Gamma}^\star\| \cdot \|\boldsymbol{\Delta}^\star\|_*,\ \forall\, \tau \in \mathbb{T}, \boldsymbol{\Gamma} \in \mathbb{R}^{m_1 \times m_2} \right) \ge 1 - C_1 \exp(-C_2 n).
$$

■

Appendix C. Several useful facts

Proposition 25 (Inequalities on joint nuclear and operator norm) *For two matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m_1 \times m_2}$, it holds that*

$$
\begin{aligned}
&\max\{\|\boldsymbol{A}\|_{\mathrm{op}}, \|\boldsymbol{B}\|_{\mathrm{op}}\} \le \|(\boldsymbol{A}^\top, \boldsymbol{B}^\top)^\top\|_{\mathrm{op}} \le \sqrt{\|\boldsymbol{A}\|_{\mathrm{op}}^2 + \|\boldsymbol{B}\|_{\mathrm{op}}^2} \le \|\boldsymbol{A}\|_{\mathrm{op}} + \|\boldsymbol{B}\|_{\mathrm{op}} \\
&(\|\boldsymbol{A}\|_* + \|\boldsymbol{B}\|_*)/\sqrt{2} \le \|(\boldsymbol{A}^\top, \boldsymbol{B}^\top)^\top\|_* \le \|\boldsymbol{A}\|_* + \|\boldsymbol{B}\|_* \qquad (36)
\end{aligned}
$$

Proof First, it's easy to prove the right-hand side of both inequalities, noting that

$$
\|(\boldsymbol{A}^\top, \boldsymbol{B}^\top)^\top\|_{\mathrm{op}}^2 = \|\boldsymbol{A}\boldsymbol{A}^T + \boldsymbol{B}\boldsymbol{B}^T\|_{\mathrm{op}} \le \|\boldsymbol{A}\|_{\mathrm{op}}^2 + \|\boldsymbol{B}\|_{\mathrm{op}}^2 \le (\|\boldsymbol{A}\|_{\mathrm{op}} + \|\boldsymbol{B}\|_{\mathrm{op}})^2, \qquad (37)
$$

and

$$
\|(\boldsymbol{A}^\top, \boldsymbol{B}^\top)^\top\|_* \le \|(\boldsymbol{A}^\top, \mathbf{0})^\top + (\mathbf{0}, \boldsymbol{B}^\top)^\top\|_* \le \|\boldsymbol{A}\|_* + \|\boldsymbol{B}\|_*.
$$

Also, the left-hand side of the operator norm inequality is easy to derive using the equality in (37).

To prove the left-hand side of (36), we use the duality between nuclear norm and the operator norm. Note that

$$\begin{aligned}
\|(\boldsymbol{A}^\top, \boldsymbol{B}^\top)^\top\|_* &= \max_{\|(\boldsymbol{C}^\top, \boldsymbol{D}^\top)^\top\|_{\mathrm{op}} \le 1} \left\langle (\boldsymbol{A}^\top, \boldsymbol{B}^\top)^\top \,,\, (\boldsymbol{C}^\top, \boldsymbol{D}^\top)^\top \right\rangle \\
&= \max_{\|\boldsymbol{C}\boldsymbol{C}^T + \boldsymbol{D}\boldsymbol{D}^T\|_{\mathrm{op}} \le 1} \left\langle (\boldsymbol{A}^\top, \boldsymbol{B}^\top)^\top \,,\, (\boldsymbol{C}^\top, \boldsymbol{D}^\top)^\top \right\rangle \\
&\ge \max_{\substack{\|\boldsymbol{C}\boldsymbol{C}^T\|_{\mathrm{op}} \le 1/2, \\ \|\boldsymbol{D}\boldsymbol{D}^T\|_{\mathrm{op}} \le 1/2}} \langle \boldsymbol{A} \,,\, \boldsymbol{C} \rangle + \langle \boldsymbol{B} \,,\, \boldsymbol{D} \rangle \\
&= \max_{\substack{\|\boldsymbol{C}\|_{\mathrm{op}} \le 1/\sqrt{2}, \\ \|\boldsymbol{D}\|_{\mathrm{op}} \le 1/\sqrt{2}}} \langle \boldsymbol{A} \,,\, \boldsymbol{C} \rangle + \langle \boldsymbol{B} \,,\, \boldsymbol{D} \rangle \\
&= \frac{1}{\sqrt{2}} (\|\boldsymbol{A}\|_* + \|\boldsymbol{B}\|_*).
\end{aligned}$$

■

Proposition 26 (Inequalities on norm for the multiplication of matrices) *Suppose* $\boldsymbol{A} \in \mathbb{R}^{m_1 \times m_1}, \boldsymbol{B} \in \mathbb{R}^{m_1 \times m_2}$ *and* $\boldsymbol{A}$ *is inversible. It holds*

$$\begin{aligned}
\|\boldsymbol{A}^{-1}\|_{\mathrm{op}}^{-1} \|\boldsymbol{B}\|_* \le \|\boldsymbol{A}\boldsymbol{B}\|_* \le \|\boldsymbol{A}\|_{\mathrm{op}} \|\boldsymbol{B}\|_*, \\
\|\boldsymbol{A}^{-1}\|_{\mathrm{op}}^{-1} \|\boldsymbol{B}\|_F \le \|\boldsymbol{A}\boldsymbol{B}\|_F \le \|\boldsymbol{A}\|_{\mathrm{op}} \|\boldsymbol{B}\|_F.
\end{aligned}$$

Proof For $\|\boldsymbol{A}\boldsymbol{B}\|_*$ we have

$$\|\boldsymbol{A}\boldsymbol{B}\|_* = \sup_{\|\boldsymbol{C}\|_{\mathrm{op}}=1} \langle \boldsymbol{A}\boldsymbol{B} \,,\, \boldsymbol{C} \rangle = \sup_{\|\boldsymbol{C}\|_{\mathrm{op}}=1} \left\langle \boldsymbol{B} \,,\, \boldsymbol{A}^\top \boldsymbol{C} \right\rangle \le \|\boldsymbol{B}\|_* \|\boldsymbol{A}^\top \boldsymbol{C}\|_{\mathrm{op}} \le \|\boldsymbol{A}\|_{\mathrm{op}} \|\boldsymbol{B}\|_*.$$

Apply this result to $\boldsymbol{A}^{-1} \cdot \boldsymbol{A}\boldsymbol{B}$, we have

$$\|\boldsymbol{B}\|_* \le \|\boldsymbol{A}^{-1}\|_{\mathrm{op}} \|\boldsymbol{A}\boldsymbol{B}\|_*,$$

implying a lower bound

$$\|\boldsymbol{A}\boldsymbol{B}\|_* \ge \|\boldsymbol{A}^{-1}\|_{\mathrm{op}}^{-1} \|\boldsymbol{B}\|_*.$$

For $\|\boldsymbol{A}\boldsymbol{B}\|_F$, we have

$$\|\boldsymbol{A}\boldsymbol{B}\|_F^2 = \sum_{k=1}^{m_2} \|\boldsymbol{A}\boldsymbol{B}_{\cdot k}\|_2^2 \le \sum_{k=1}^{m_2} \|\boldsymbol{A}\|_{\mathrm{op}}^2 \|\boldsymbol{B}_{\cdot k}\|_2^2 \le \|\boldsymbol{A}\|_{\mathrm{op}}^2 \|\boldsymbol{B}\|_F^2.$$

The lower bound follows similarly.

■

Proposition 27 (Vectorization and Gaussianity) $\boldsymbol{X} \in \mathbb{R}^{m_1 \times m_2}$ *is a random matrix from Gaussian ensemble* $\mathcal{N}_{m_1 m_2}(\boldsymbol{0}, \boldsymbol{\Sigma})$*, i.e.,* $\boldsymbol{X}^{\mathrm{V}} \sim \mathcal{N}_{m_1 m_2}(\boldsymbol{0}, \boldsymbol{\Sigma})$*.* $\boldsymbol{U}_1 \in \mathbb{R}^{m_1 \times m_1}$ *and* $\boldsymbol{U}_2 \in \mathbb{R}^{m_2 \times m_2}$ *are orthogonal matrices. Then* $\boldsymbol{U}_1 \boldsymbol{X} \boldsymbol{U}_2$ *is a random matrix from Gaussian ensemble* $\mathcal{N}_{m_1 m_2}(\boldsymbol{0}, \boldsymbol{\Sigma}')$*, where* $\boldsymbol{\Sigma}' = (\boldsymbol{U}_1 \otimes \boldsymbol{U}_2) \boldsymbol{\Sigma} (\boldsymbol{U}_1 \otimes \boldsymbol{U}_2)$*.*

Proof Use $(\boldsymbol{U}_1\boldsymbol{X}\boldsymbol{U}_2)^{\mathrm{V}} = (\boldsymbol{U}_1 \otimes \boldsymbol{U}_2)\boldsymbol{X}^{\mathrm{V}}$. ■

The following proposition is taken from Theorem 6.5 of Wainwright (2019).

Proposition 28 (Spectral concentration of Wishart matrices) *There are universal constants $\{c_j\}_{j=0}^3$ such that, for any row-wise σ-sub-Gaussian random matrix $\boldsymbol{X} \in \mathbb{R}^{m_1\times m_2}$. Let $\boldsymbol{\Sigma}$ be the population covariance for each row, then the sample covariance matrix $\widehat{\boldsymbol{\Sigma}} = n^{-1}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top$ satifies the bounds*

$$\mathbb{E}\left\{\exp(\lambda\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\mathrm{op}})\right\} \le \exp\left(c_0\frac{\lambda^2\sigma^4}{n} + 4d\right),\ \textit{for all } |\lambda| < \frac{n}{64e^2\sigma^2},$$

and hence

$$\mathbb{P}\left\{\frac{\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\mathrm{op}}}{\sigma^2}\right\} \ge c_2\exp(-c_3 n\min\{\delta,\delta^2\}),\ \textit{for all } \delta \ge 0.$$

The following proposition for Lévy's inequality is taken from Proposition A.1.2 of van der Vaart and Wellner (1996):

Proposition 29 (Lévy's inequality) *Let $X_1,\dots,X_n$ be independent, symmetric stochastic processes indexed by an arbitrary set. Let $S_k = \sum_{i=1}^k X_i$ be the partial sum. Then for every $\lambda > 0$ we have the inequalities*

$$\mathbb{P}\left(\max_{k\le n}\|S_k\| > \lambda\right) \le 2\mathbb{P}\left(\|S_n\| > \lambda\right), \mathbb{P}\left(\max_{k\le n}\|X_k\| > \lambda\right) \le 2\mathbb{P}\left(\|S_n\| > \lambda\right).$$

Proposition 30 (Closed solution for one minimization program) *Let $\boldsymbol{A}_1, \boldsymbol{A}_2$ be some constant matrix in $\mathbb{R}^{m_1\times m_2}$. $\tau_1^\star$ and τ_2 are positive constants. Then it holds*

$$\tau_1^\star\|\boldsymbol{A}_1 - \boldsymbol{A}\|_F^2 + \tau_2\|\boldsymbol{A} - \boldsymbol{A}_2\|_F^2 \ge \frac{\tau_1^\star\tau_2}{\tau_1^\star + \tau_2}\|\boldsymbol{A}_1 - \boldsymbol{A}_2\|_F^2.$$

where the equality is attained at

$$\boldsymbol{A} = \frac{\tau_1^\star\boldsymbol{A}_1 + \tau_2\boldsymbol{A}_2}{\tau_1^\star + \tau_2}.$$

Proof Simply setting the derivative with respect to $\boldsymbol{A}$ to zero, one can obtain

$$-2\tau_1^\star\left(\boldsymbol{A}_1 - \boldsymbol{A}\right) + 2\tau_2\left(\boldsymbol{A} - \boldsymbol{A}_2\right) = 0,$$

which give

$$\boldsymbol{A} = \frac{\tau_1^\star\boldsymbol{A}_1 + \tau_2\boldsymbol{A}_2}{\tau_1^\star + \tau_2}.$$

Now taking this solution into the program generates the tight lower bounds. ■

Proposition 31 (Generalization of Hanson-Wright inequality) *Let $\{\boldsymbol{x}_i\}_{i=1}^n$ be i.i.d copies of $\boldsymbol{x} \sim N(\boldsymbol{0}, \boldsymbol{I}_m)$. Let $\boldsymbol{A}_i$ be $m \times m$ matrices. Then, for every $t > 0$, it holds*

$$\mathbb{P}\left\{\left|\sum_{i=1}^n \left(\boldsymbol{x}_i^\top \boldsymbol{A}_i \boldsymbol{x}_i - \mathbb{E}\left(\boldsymbol{x}^\top \boldsymbol{A}_i \boldsymbol{x}\right)\right)\right| > t\right\} \leq 2\exp\left\{-c\min\left(\frac{t^2}{\sum_{i=1}^n \|\boldsymbol{A}_i\|_F^2}, \frac{t}{\max_{i=1,\cdots,n}\|\boldsymbol{A}_i\|_{\mathrm{op}}}\right)\right\}$$

Proof The case where $n = 1$ is the famous Hanson-Wright inequality(see Rudelson and Vershynin (2013); Vershynin (2018)). The proof to the generalization is simple if we observe the summation

$$\sum_{i=1}^n \left(\boldsymbol{x}_i^\top \boldsymbol{A}_i \boldsymbol{x}_i - \mathbb{E}\left(\boldsymbol{x}^\top \boldsymbol{A}_i \boldsymbol{x}\right)\right)$$

can be aggregated into a large quadratic form in Gaussian vectors:

$$\sum_{i=1}^n \left(\boldsymbol{x}_i^\top \boldsymbol{A}_i \boldsymbol{x}_i - \mathbb{E}\left(\boldsymbol{x}^\top \boldsymbol{A}_i \boldsymbol{x}\right)\right) = \left(\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top, \cdots, \boldsymbol{x}_n^\top\right) \begin{pmatrix} \boldsymbol{A}_1 & & & \\ & \boldsymbol{A}_2 & & \\ & & \ddots & \\ & & & \boldsymbol{A}_n \end{pmatrix} \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_n \end{pmatrix}.$$

Now apply the original Hanson-Wright to complete the proof. ■

Appendix D. Additional algorithmic details and numerical supports

D.1 Proximal gradient descent for solving joint minimization scheme (5)

In this section, we summarize the implementation details for the proposed joint single change-point detection and matrix estimation (5) in Algorithm 2 and 3 below.

In the rest of this section, we consider the random multivariate regression example in Section 2.2.6 and perform an algorithmic convergence analysis. For ease of reading, we quickly summarize the following quantities appeared from Algorithm 2 and 3:

- τ_k: k-th candidate grid searching point;
- $\widehat{\boldsymbol{\Gamma}}_k$: minimizer of (3) given search point τ_k;
- $\widehat{\boldsymbol{\Gamma}}_k^{(l)}$: the l-th step of the proximal gradient descent algorithm for solving (3) given search point τ_k;
- $\widehat{k}$: the index where $S_N(\widehat{\boldsymbol{\Gamma}}_k, \tau_k)$ is minimized;
- $\widehat{k}^{(l)}$: the index where at the step l, $S_N(\widehat{\boldsymbol{\Gamma}}_k^{(l)}, \tau_k)$ is minimized;
- $\widehat{\tau}^{(l)}$: the time point that minimizes $S_N(\widehat{\boldsymbol{\Gamma}}_k^{(l)}, \tau_k)$.
- $\widehat{\tau}$, $\widehat{\Gamma}$: global minimizer of (5).

We state the result as the proposition below:

Algorithm 2: Joint single change-point detection and matrix estimation

```
Input: Observed data $(\boldsymbol{y}_i, \boldsymbol{X}_i, t_i)$, for $i = 1, \cdots, n$; regularization parameter $\lambda_N$;
       floor curvature $L^{(0)}$; ceiling curvature $L_{max}$; updating rate $\eta$; convergence
       tolerance tol and maximal iteration $T$.
Output: Estimator $\widehat{\boldsymbol{\Gamma}}, \widehat{\tau}$.
/* Step (i): pick $K$ candidate grid searching points                    */
Set candidate grid searching points:
    $$\tau_k = \frac{1}{2}\{(1-2\omega)/K \cdot (k-1) + (1-2\omega)/K \cdot k\},$$
  for $k = 1, \cdots, K$.
/* Step (ii): solve program (3) at each $\tau_k$ with accelerated proximal
   gradient descent (Algorithm 3)                                        */
for $k = 1, \ldots, K$ do
    Run Algorithm 3 with the given inputs and testing break position $\tau_k$;
    Return estimator $\widehat{\boldsymbol{\Gamma}}_k$ and obj$_k$.
/* Step (iii): minimization over grid searching points                   */
Set $\widehat{k} = \arg\min_{k=1,\cdots,K}$ obj$_k$.
Return $\widehat{\tau} = \tau_{\widehat{k}}$ and $\widehat{\boldsymbol{\Gamma}} = \boldsymbol{\Gamma}_{\widehat{k}}$.
```

Algorithm 3: Proximal gradient descent for (3)

```
Input: Same input as Algorithm 2; testing break position $\tau$.
Output: Estimator $\widehat{\boldsymbol{\Gamma}}$; objective value obj.
Set $l = 1$ and $\boldsymbol{\Gamma}^{(0)} = \mathbf{0}$;
Calculate the sub-gradient $\boldsymbol{G}^{(l)} = \nabla S_N(\boldsymbol{\Gamma}; \tau)|_{\boldsymbol{\Gamma}=\boldsymbol{\Gamma}^{(l)}}$.
Set $L = \min\{\eta L^{(l-1)}, L_{\max}\}$.
while $L < L_{\max}$ do
    Compute $\boldsymbol{\Omega} = \text{Soft}\left(\boldsymbol{\Gamma}^{(l)} - L^{-1}\boldsymbol{G}^{(l)}; L^{-1}\lambda_N\right)$ based on (11);
    Calculate $S_N(\boldsymbol{\Omega}; \tau)$ based on (4) and $S_{\text{Major}}(\boldsymbol{\Omega}; \boldsymbol{\Omega}^{(l)})$ based on (10);
    if $S_N(\boldsymbol{\Omega}; \tau) \le S_{\text{Major}}(\boldsymbol{\Omega}; \boldsymbol{\Omega}^{(l)})$ then
        Set $\boldsymbol{\Gamma}^{(l+1)} = \boldsymbol{\Omega}$, $L^{(l)} = L$
        break
    else
        Set $L = \min\{L/\eta, L_{\max}\}$.
Repeat above steps until the stop criterion is meet: $\|\boldsymbol{\Gamma}^{(l+1)} - \boldsymbol{\Gamma}^{(l)}\|_F / \|\boldsymbol{\Gamma}^{(l)}\|_F \le$ tol
  or the maximal number of iteration $T$ is hit.
Set $\widehat{\boldsymbol{\Gamma}} = \boldsymbol{\Gamma}^{(l+1)}$ and record objective value obj $= S_N(\boldsymbol{\Gamma}; \tau) + \lambda_N \|\boldsymbol{\Gamma}\|_*$.
```

Proposition 32 *Assume $\mathbf{\Gamma}^\star$ has exact low rank R with $\|\mathbf{\Gamma}^\star\|_* \le \bar{s}$. For $n > Cm_1$, with probability greater than*

$$1 - 3C_1 \exp\{-C_2 n\},$$

For $n > Cm_1$ and any given tolerance parameter ζ^2,

$$S_N(\widehat{\mathbf{\Gamma}}_k^{(l)}; \tau_k) - S_N(\widehat{\mathbf{\Gamma}}_k; \tau_k) \le \zeta^2, \quad \|\widehat{\mathbf{\Gamma}}_k^{(l)} - \widehat{\mathbf{\Gamma}}_k\|_F^2 \le \frac{4\zeta^2}{\underline{\rho}\sigma_0^2}, \tag{38}$$

for all steps l satisfying

$$l \ge \frac{2\log\{\zeta^{-2}(S_N(\mathbf{\Gamma}_k^{(0)}; \tau_k) - S_N(\widehat{\mathbf{\Gamma}}(\tau_k); \tau_k))\}}{\log(12/11)} + \log_2\log_2(\frac{\bar{s}\lambda_N}{\zeta^2})\left\{1 + \frac{\log 2}{\log(12/11)}\right\}. \tag{39}$$

Furthermore, recall $k^\star$ as the minimizing index for Step (iii) of Algorithm 2. For the tolerance smaller than the gap:

$$\zeta^2 < \left\{\min_{k \ne \widehat{k}} S_N(\widehat{\mathbf{\Gamma}}_k; \tau_k)) - S_N(\widehat{\mathbf{\Gamma}}; \widehat{\tau})\right\}, \tag{40}$$

and l with

$$l \ge \frac{2\log\{\zeta^{-2}\max_{k\in[K]}(S_N(\mathbf{\Gamma}_k^{(0)}; \tau_k) - S_N(\widehat{\mathbf{\Gamma}}(\tau_k); \tau_k))\}}{\log(12/11)} + \log_2\log_2(\frac{\bar{s}\lambda_N}{\zeta^2})\left\{1 + \frac{\log 2}{\log(12/11)}\right\}, \tag{41}$$

we have $\widehat{\tau}^{(l)} = \widehat{\tau}$.

Proposition 32 is adapted from Theorem 2 of Agarwal et al. (2010), which states that when using PGD for penalized M-estimation, the excess loss decays geometrically up to any squared error ζ^2 given certain conditions. In our setting, for each given τ_k, it takes $O(\log\zeta^{-2})$ steps to reach the tolerance ζ^2, which demonstrates a fast exponential convergence to the global minimum. When ζ^2 is set to be properly small, the detected change-point $\widehat{\tau}^{(l)}$ coincides with the global optimum too.

Proof [Proof of Proposition 32] Proposition 32 is based on Theorem 2 of Agarwal et al. (2010), which states that when using PGD for penalized M-estimation, the excess loss decays geometrically up to any squared error ζ^2 given the so-called restricted strong convexity (RSC) and restricted strong smoothness (RSM) conditions. Our Proposition 20 verifies the RSC and RSM conditions uniformly at a sequence of searching points with high probability in the multivariate regression setting:

If $n > Cm_1$, then

$$\mathbb{P}\left(\kappa(\mathfrak{X})\|\mathbf{\Gamma}\|_F^2 \le \frac{1}{2n}\|\mathfrak{X}(\mathbf{\Gamma}; \tau)\|_2^2 \le \kappa'(\mathfrak{X})\|\mathbf{\Gamma}\|_F^2, \ \forall\, \tau \in \mathbb{T} \text{ and } \mathbf{\Gamma} \in \mathbb{R}^{(2m_1)\times m_2}\right) \ge 1 - C_1\exp(-C_2 n),$$

where $\kappa(\mathfrak{X}) = \underline{\rho}\sigma_0^2/2$ for some constant $\underline{\rho} > 0$ and $\kappa'(\mathfrak{X}) = 3\bar{\rho}\sigma_0^2/2$.

Therefore, by substituting the quantities of Agarwal et al. (2010), Theorem 2 with the counterpart in our setting, we can conclude (38) for the steps (39).

To prove $\widehat{\tau}^{(l)} = \widehat{\tau}$, simply notice that for ζ^2 satisfying (40), l with (41), and any $k \neq \widehat{k}$

$$S_N(\widehat{\mathbf{\Gamma}}_k^{(l)}; \tau_k) \geq S_N(\widehat{\mathbf{\Gamma}}_k; \tau_k) \geq S_N(\widehat{\mathbf{\Gamma}}; \widehat{\tau}) + \zeta^2 \geq S_N(\widehat{\mathbf{\Gamma}}_{\widehat{k}}^{(l)}; \widehat{\tau}),$$

which concludes the proof. ■

D.2 Determining penalization level λ_N

In practice, the regularization parameter λ_N is chosen through cross-validation. Suppose the threshold variable t_i is reorganized in an increasing order. Our cross-validation procedure proceeds as follows: (i) Splitting: given data $\mathcal{D} = \{(y_i, \boldsymbol{X}_i, t_i) : i = 1, \ldots, N\}$, we split the data into K folds $\mathcal{D}_1, \ldots, \mathcal{D}_K$ in an incremental manner:

$$\mathcal{D}_k = \{(y_i, \boldsymbol{X}_i, t_i) : i = k + [N/K] * l,\ l = 0, \ldots, K-1\},$$

where $[N/K]$ is the largest integer smaller than N/K. (ii) Validating: we pick a sequence of K candidate values, $\lambda_N \in \{\lambda_{N,1}, \ldots, \lambda_{N,K}\}$. For $k = 1, \ldots, K$, we choose in turn $\mathcal{D}_k$ as the prediction set $\mathcal{D}_{k,\text{test}}$ and the union of the rest $\mathcal{D}_k$'s as the training set $\mathcal{D}_{k,\text{train}}$. Then we apply the proposed program with penalization λ_k on $\mathcal{D}_{k,\text{train}}$ and compute the prediction error on the testing data. The final penalization level $\lambda_N^\star$ is determined as the one that gives the smallest prediction error.

We add some additional remarks for the above cross-validation procedure. First, the splitting step (i) is completed in an incremental manner. This is crucial because it guarantees that, when there are change-points in the data, the relative location of the change-points in each of the subset $\mathcal{D}_k$ remains almost identical as the full data $\mathcal{D}$. Second, the candidate values for λ_N can be usually motivated by theory and picked in some principled way. For example, in multivariate regression, our Theorem 11 suggests the penalization of the order $c(\frac{m_1+m_2}{n})^{1/2}$. Therefore, in terms of tuning we can choose a sequence of $c \in \{c_1, \ldots, c_K\}$ to further construct $\lambda_N \in \{\lambda_{N,1}, \ldots, \lambda_{N,K}\}$.

D.3 Construction of signals

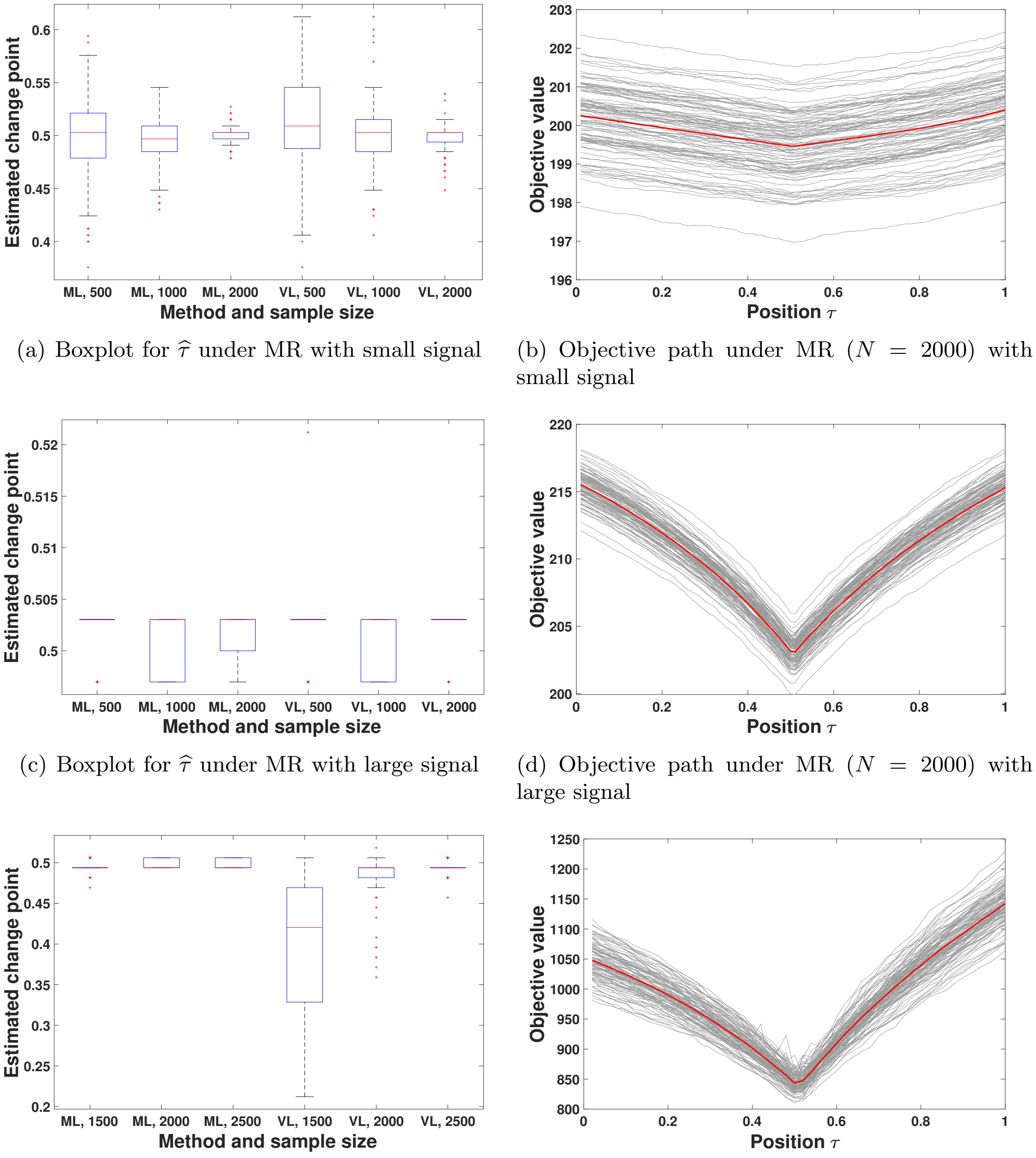
We generate the low-rank signals from the singular vectors of Gaussian ensembles. To ensure a large break, $\mathbf{\Theta}_0^\star$ and $\mathbf{\Theta}_1^\star$ are separately constructed in the following way: first generate a random matrix $\boldsymbol{M}_s \in \mathbb{R}^{m \times 100}$ $(s = 0, 1)$ with i.i.d. standard Gaussian variables. Let $\boldsymbol{U}_s \boldsymbol{S}_s \boldsymbol{V}_s^\top$ be the singular value decomposition of $\boldsymbol{M}_s$. Then $\mathbf{\Theta}_s$ is given by $\mathbf{\Theta}_s^\star = \boldsymbol{U}_s^r \boldsymbol{V}_s^{r\top} / \sqrt{r},\ s = 0, 1$. To generate signals with a small break, we take one single standard Gaussian ensemble $\boldsymbol{M}$ in $\mathbb{R}^{m \times 100}$ and get its SVD $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$. Aggregating some of the singular vectors into new matrices: $\boldsymbol{U}_0^r = \boldsymbol{U}_1^r = [u_1, \cdots, u_r]; \boldsymbol{V}_0^r = [v_1, \cdots, v_{r-1}, v_r];\ \boldsymbol{V}_1 = [v_1, \cdots, v_{r-1}, v_{r+1}]$. Also define $\boldsymbol{D} = \text{diag}\{\sqrt{4.75/4}, \sqrt{4.75/4}, \sqrt{4.75/4}, \sqrt{4.75/4}, \sqrt{0.25}\}$. Now construct $\mathbf{\Theta}_s^\star = \frac{1}{\sqrt{r}} \boldsymbol{U}_s \boldsymbol{D} \boldsymbol{V}_s^\top,\ s = 0, 1$. This way, it is easy to show that $\mathbf{\Theta}_0^\star$ and $\mathbf{\Theta}_1^\star$ share the same left singular vectors and $\|\mathbf{\Theta}_0^\star - \mathbf{\Theta}_1^\star\|_F^2 = 0.1$.

D.4 Multivariate regression with large signals

Table 7: Multivariate regression with one change-point and large signal

Regime	(m, N)	Method	$\lvert\widehat{\tau} - \tau^\star\rvert$	$\widehat{\boldsymbol{\Theta}}_1$			$\widehat{\boldsymbol{\Theta}}_2$		
				$\lVert\widehat{\boldsymbol{\Theta}}_1 - \boldsymbol{\Theta}_1^\star\rVert_F^2$	$\lVert\widehat{\boldsymbol{\Theta}}_1 - \boldsymbol{\Theta}_1^\star\rVert_*$	rank	$\lVert\widehat{\boldsymbol{\Theta}}_2 - \boldsymbol{\Theta}_2^\star\rVert_F^2$	$\lVert\widehat{\boldsymbol{\Theta}}_2 - \boldsymbol{\Theta}_2^\star\rVert_*$	rank
Varying N	(50, 500)	ML	0.003(0)	0.315(0.021)	1.468(0.053)	5.28(0.90)	0.313(0.022)	1.471(0.055)	5.41(1.12)
		VL	0.003(0.002)	0.896(0.030)	5.591(0.094)	50.00(0)	0.901(0.028)	5.600(0.093)	50.00(0)
		Oracle	-	0.314(0.021)	1.467(0.054)	5.31(0.96)	0.314(0.022)	1.474(0.055)	5.45(1.20)
		NC	-	0.737(0.031)	2.441(0.061)	10.59(0.59)	0.731(0.028)	2.429(0.057)	10.59(0.59)
	(50, 1000)	ML	0.003(0)	0.181(0.014)	1.108(0.037)	5.00(0)	0.181(0.012)	1.110(0.034)	5.00(0)
		VL	0.003(0)	0.449(0.014)	3.974(0.067)	50.00(0)	0.450(0.014)	3.980(0.064)	50.00(0)
		Oracle	-	0.179(0.014)	1.103(0.037)	5.00(0)	0.180(0.013)	1.108(0.034)	5.00(0)
		NC	-	0.645(0.025)	2.367(0.047)	10.72(0.83)	0.644(0.025)	2.366(0.048)	10.72(0.83)
	(50, 2000)	ML	0.003(0)	0.096(0.006)	0.810(0.025)	5.00(0)	0.096(0.006)	0.810(0.024)	5.00(0)
		VL	0.003(0)	0.229(0.007)	2.848(0.043)	50.00(0)	0.231(0.007)	2.856(0.045)	50.00(0)
		Oracle	-	0.094(0.006)	0.804(0.025)	5.00(0)	0.096(0.006)	0.808(0.024)	5.00(0)
		NC	-	0.583(0.019)	2.292(0.037)	10.05(0.26)	0.588(0.020)	2.303(0.036)	10.05(0.26)
Varying m	(25, 625)	ML	0.003(0.002)	0.182(0.019)	1.080(0.048)	5.00(0)	0.184(0.017)	1.083(0.047)	5.00(0)
		VL	0.003(0)	0.243(0.016)	2.062(0.072)	25.00(0)	0.244(0.017)	2.062(0.073)	25.00(0)
		Oracle	-	0.181(0.019)	1.078(0.048)	5.00(0)	0.185(0.017)	1.086(0.047)	5.00(0)
		NC	-	0.639(0.040)	2.238(0.069)	8.76(0.47)	0.647(0.036)	2.251(0.069)	8.76(0.47)
	(50, 1250)	ML	0.003(0)	0.183(0.012)	1.112(0.034)	5.00(0)	0.182(0.014)	1.109(0.037)	5.00(0)
		VL	0.003(0)	0.448(0.014)	3.970(0.068)	50.00(0)	0.448(0.014)	3.969(0.057)	50.00(0)
		Oracle	-	0.181(0.012)	1.107(0.034)	5.00(0)	0.181(0.014)	1.107(0.037)	5.00(0)
		NC	-	0.644(0.021)	2.366(0.043)	10.75(0.72)	0.641(0.024)	2.361(0.046)	10.75(0.72)
	(75, 1875)	ML	0.003(0)	0.182(0.010)	1.121(0.027)	5.00(0)	0.182(0.010)	1.120(0.028)	5.00(0)
		VL	0.003(0)	0.639(0.013)	5.816(0.060)	75.00(0)	0.643(0.015)	5.833(0.070)	75.00(0)
		Oracle	-	0.180(0.009)	1.117(0.027)	5.00(0)	0.181(0.010)	1.118(0.028)	5.00(0)
		NC	-	0.647(0.016)	2.424(0.038)	12.02(1.96)	0.644(0.017)	2.419(0.039)	12.02(1.96)

D.5 Plots of estimated change-points and objective trajectories



(a) Boxplot for $\widehat{\tau}$ under MR with small signal

(b) Objective path under MR ($N = 2000$) with small signal

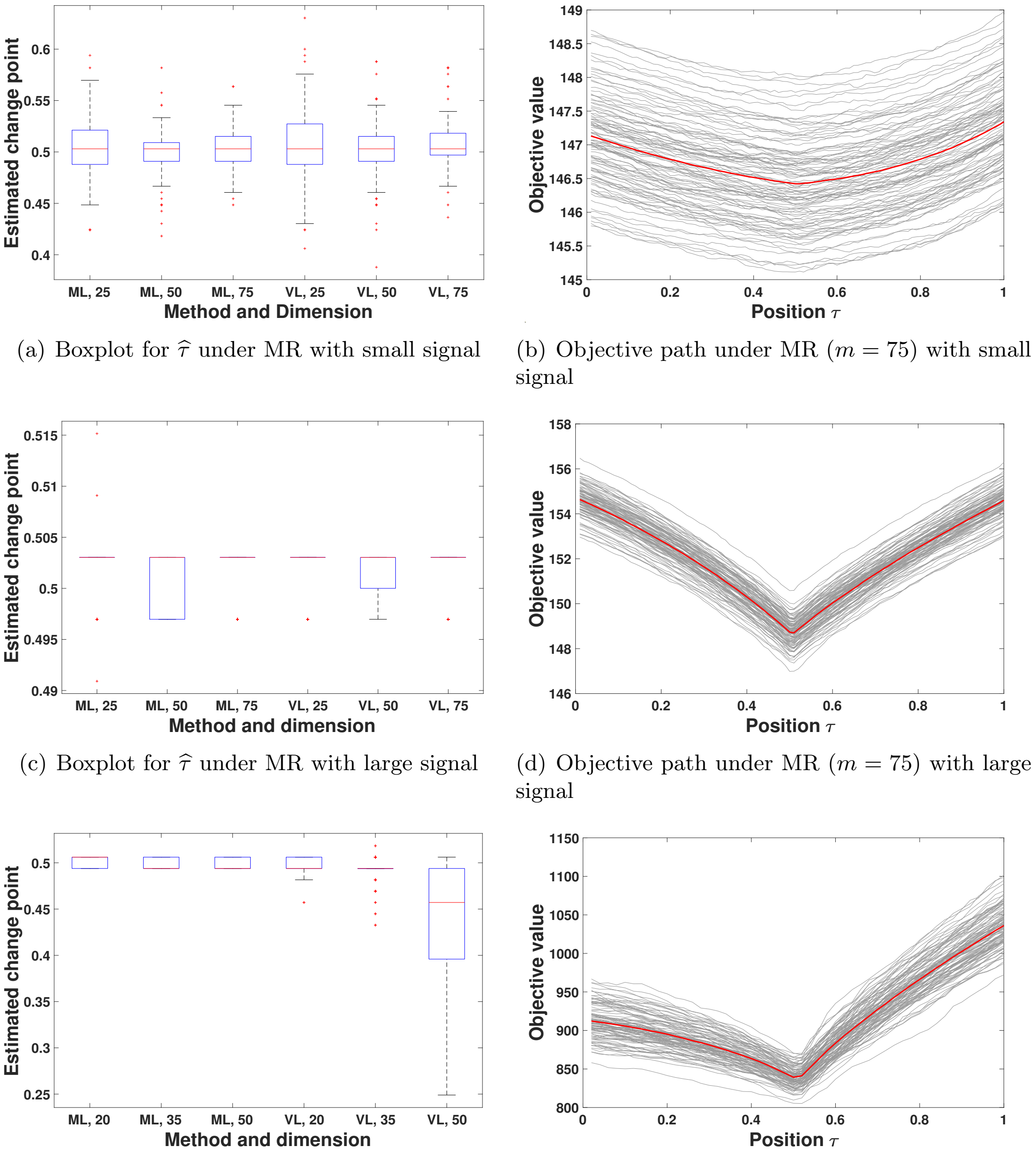
(c) Boxplot for $\widehat{\tau}$ under MR with large signal

(d) Objective path under MR ($N = 2000$) with large signal

(e) Boxplot for $\widehat{\tau}$ under CS with large signal

(f) Objective path under CS ($N = 2500$) with large signal

Figure 2: Boxplot for $\widehat{\tau}$ and objective path under different models with varying sample size



(a) Boxplot for $\widehat{\tau}$ under MR with small signal

(b) Objective path under MR ($m = 75$) with small signal

(c) Boxplot for $\widehat{\tau}$ under MR with large signal

(d) Objective path under MR ($m = 75$) with large signal

(e) Boxplot for $\widehat{\tau}$ under CS with large signal

(f) Objective path under CS ($m = 50$) with large signal

Figure 3: Boxplot for $\widehat{\tau}$ and objective path under different models with varying dimension

D.6 Data analysis: comparison with other methods

We make some additional numerical comparison for the air pollution data analysis conducted in Section 4.3.

In Table 8 below, we compare three straightforward estimation and prediction schemes without considering a change-point structure. “OLS” stands for a direct output from ordinary least squares; "ℓ_1 penalization" adds ℓ_1 penalty to induce sparsity pattern for the mechanism matrix; "Nuclear norm" stands for the nuclear norm penalization that is used to induce a low-rank structure. From the results we can see that, when no change-point is included, adding both ℓ_1 and nuclear norm penalty can slightly improve the prediction accuracy. Moreover, a low-rank model gives better prediction results than a sparse model. However, the improvement from either penalization does not demonstrate a significant edge.

Table 8: Test error for different methods without change-points for the air pollution data

Methods	OLS	ℓ_1 penalization	Nuclear norm
Test error	0.1931	0.1928	0.1925

In Table 9 below, we further compare ℓ_1 penalization and nuclear norm penalization with change-point structure incorporated. We can see that for each given number of change-points within the range of 1 to 4, nuclear norm penalization always demonstrates higher prediction accuracy. Overall, nuclear norm penalization with two change-points outperform the rest, which is still consistent with the conclusion for the real data analysis (Section 4.3) in the main paper.

Table 9: Test error for different methods with change-points for the air pollution data

# change-points		1	2	3	4
Methods	**Nuclear norm**	0.1746	**0.1728**	0.1761	0.1772
	ℓ_1 penalization	0.1825	0.1884	0.1797	0.1817

D.7 Multivariate regression with a single change-point near the boundary

In this section, we conduct some numerical experiments to test the performance of the algorithm when the change-point is close to the boundary. The setup is based on multivariate regression with one change-point, and the location of the change point is set to two locations: $\tau^\star = 0.1$ and $\tau^\star = 0.2$. The results are reported in Table 10. We can see that the proposed method can successfully detect the location of the change point and recover the true signal with a near-oracle performance. The no-change point algorithm gives a result that recovers the matrix signal in the segment that has more data but fail to recover the signal that has few data points. Meanwhile, LASSO based methods work poorly for recovering the true matrix signals as the matrix elements are not sparse due to construction.

D.8 Numerical experiments for multiple change-points detection with random locations

In this section, to test how robust the proposed algorithm is to the locations of the change-points, we present numerical experiments on multiple change-point detection where three change-points are generated randomly from $(0, 1/3]$, $(1/3, 2/3]$ and $(2/3, 1)$, respectively. The remaining setup is kept the same. Table 11 reports the results.

Table 10: Multivariate regression with a single change point near the boundary

Method	$\lvert\widehat{\tau}-\tau^{\star}\rvert$	$\widehat{\mathbf{\Theta}}_1$		$\widehat{\mathbf{\Theta}}_2$	
		$\lVert\widehat{\mathbf{\Theta}}_1-\mathbf{\Theta}_1^{\star}\rVert_F^2$	$\lVert\widehat{\mathbf{\Theta}}_1-\mathbf{\Theta}_1^{\star}\rVert_*$	$\lVert\widehat{\mathbf{\Theta}}_2-\mathbf{\Theta}_2^{\star}\rVert_F^2$	$\lVert\widehat{\mathbf{\Theta}}_2-\mathbf{\Theta}_2^{\star}\rVert_*$
Regime: $\tau^{\star}=0.1$					
Ours	0.0045(0)	0.475(0.027)	3.074(0.072)	0.113(0.003)	3.150(0.082)
Oracle	-	0.447(0.024)	2.995(0.066)	0.113(0.004)	3.136(0.090)
NC	-	1.671(0.023)	4.897(0.065)	0.106(0.003)	1.714(0.077)
Vec	0.0045(0)	0.751(0.035)	4.387(0.104)	0.133(0.007)	4.128(0.117)
Regime: $\tau^{\star}=0.2$					
Ours	0.0045(0)	0.272(0.013)	2.698(0.057)	0.127(0.004)	3.768(0.075)
Oracle	-	0.259(0.016)	2.650(0.058)	0.126(0.004)	3.775(0.079)
NC	-	1.351(0.012)	4.496(0.057)	0.171(0.004)	2.054(0.075)
Vec	0.0045(0)	0.481(0.032)	3.810(0.044)	0.151(0.006)	4.521(0.037)

Table 11: Mutiple change-points detection with random change-point locations

Criterion		Small breaks		Large breaks	
		Rough	Refined	Rough	Refined
Change detection	$\widehat{s}$	3.18(0.70)	-	3.00(0)	-
	OE	0.077(0.121)	0.060(0.110)	0.002(0.001)	0.001(0.001)
	UE	0.095(0.093)	0.095(0.104)	0.002(0.001)	0.001(0.001)
Matrix recovery	MOE	-	0.410(0.030)	-	0.315(0.024)
	MUE	-	0.460(0.047)	-	0.315(0.024)
	$\max\widehat{r}_k$	-	5.39(0.53)	-	6.60(1.12)
	$\min\widehat{r}_k$	-	5.00(0)	-	5.00(0)

References

Alekh Agarwal, Sahand Negahban, and Martin J Wainwright. Fast global convergence rates of gradient methods for high-dimensional statistical recovery. *Advances in Neural Information Processing Systems*, 23, 2010.

Susan Athey, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, 116(536):1716–1730, 2021.

Peiliang Bai, Abolfazl Safikhani, and George Michailidis. Multiple change points detection in low rank and sparse high dimensional vector autoregressive models. *IEEE Transactions on Signal Processing*, 68:3074–3089, 2020.

Peiliang Bai, Abolfazl Safikhani, and George Michailidis. Multiple change point detection in reduced rank high dimensional vector autoregressive models. *Journal of the American Statistical Association*, 118(544):2776–2792, 2023.

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(11):1–38, 2018.

Emmanuel J. Candès and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.

Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

K. S. Chan. Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model. *The Annals of Statistics*, 21(1):520–533, 1993.

Yudong Chen and Yuejie Chi. Harnessing structures in big data via guaranteed low-rank matrix estimation: Recent theory and fast algorithms via convex and nonconvex optimization. *IEEE Signal Processing Magazine*, 35(4):14–31, 2018.

Haeran Cho and Piotr Fryzlewicz. Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 77(2):475–507, 2015.

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

Holger Dette, Guangming Pan, and Qing Yang. Estimating a change point in a sequence of very high-dimensional covariance matrices. *Journal of the American Statistical Association*, 117(537):444–454, 2022.

Andreas Elsener and Sara van de Geer. Robust low-rank matrix estimation. *The Annals of Statistics*, 46(6B):3481–3509, 2018.

Marco A Espinosa-Vega and Mr Juan Sole. *Cross-border financial surveillance: a network perspective*. International Monetary Fund, 2010.

Jianqing Fan, Wenyan Gong, and Ziwei Zhu. Generalized high-dimensional trace regression via nuclear norm regularization. *Journal of Econometrics*, 212(1):177–202, 2019.

Jianqing Fan, Weichen Wang, and Ziwei Zhu. A shrinkage principle for heavy-tailed data: high-dimensional robust low-rank matrix recovery. *The Annals of Statistics*, 49(3):1239–1266, 2021.

William Fithian and Rahul Mazumder. Flexible low-rank statistical modeling with missing data and side information. *Statistical Science*, 33(2):238–260, 2018.

Mohammad Golbabaee and Pierre Vandergheynst. Hyperspectral image compressed sensing via low-rank and joint-sparse matrix recovery. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2741–2744, 2012.

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

Shuiwang Ji and Jieping Ye. An accelerated gradient method for trace norm minimization. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 457–464, 2009.

Abhishek Kaul, Venkata K. Jandhyala, and Stergios B. Fotopoulos. An efficient two step algorithm for high dimensional change point regression models without grid search. *Journal of Machine Learning Research*, 20(111):1–40, 2019.

Raghunandan H. Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from noisy entries. *Journal of Machine Learning Research*, 11:2057–2078, 2010.

Olga Klopp. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 20(1):282–303, 2014.

Vladimir Koltchinskii, Karim Lounici, and Alexandre B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

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.

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

Bin Liu, Cheng Zhou, Xinsheng Zhang, and Yufeng Liu. A unified data-adaptive framework for high dimensional change point detection. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 82(4):933–963, 2020.

Bin Liu, Xinsheng Zhang, and Yufeng Liu. Simultaneous change point inference and structure recovery for high dimensional Gaussian graphical models. *Journal of Machine Learning Research*, 22:Paper No. [274], 62, 2021.

Malte Londschien, Solt Kovács, and Peter Bühlmann. Change-point detection for graphical models in the presence of missing values. *Journal of Computational and Graphical Statistics*, 30(3):768–779, 2021.

Sahand Negahban and Martin J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2):1069–1097, 2011.

Yu. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1, Ser. B):125–161, 2013.

Flavio F Nobre and Donna F Stroup. A monitoring system to detect changes in public health surveillance data. *International Journal of Epidemiology*, 23(2):408–418, 1994.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41:100–115, 1954.

Andy Ramlatchan, Mengyun Yang, Quan Liu, Min Li, Jianxin Wang, and Yaohang Li. A survey of matrix completion methods for recommendation systems. *Big Data Mining and Analytics*, 1(4):308–323, 2018.

Benjamin Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(104):3413–3430, 2011.

Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

Alessandro Rinaldo, Daren Wang, Qin Wen, Rebecca Willett, and Yi Yu. Localizing changes in high-dimensional regression models. In *International Conference on Artificial Intelligence and Statistics*, pages 2089–2097. PMLR, 2021.

Angelika Rohde and Alexandre B. Tsybakov. Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930, 2011.

Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-Gaussian concentration. *Electronic Communications in Probability*, 18:no. 82, 9, 2013.

Abolfazl Safikhani and Ali Shojaie. Joint Structural Break Detection and Parameter Estimation in High-Dimensional Nonstationary VAR Models. *Journal of the American Statistical Association*, 117(537):251–264, 2022.

Abolfazl Safikhani, Yue Bai, and George Michailidis. Fast and scalable algorithm for detection of structural breaks in big VAR models. *Journal of Computational and Graphical Statistics*, 31(1):176–189, 2022.

Kean Ming Tan, Qiang Sun, and Daniela Witten. Sparse reduced rank Huber regression in high dimensions. *Journal of the American Statistical Association*, 118(544):2383–2393, 2023.

Kim-Chuan Toh and Sangwoon Yun. An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. *Pacific Journal of Optimization*, 6(3): 615–640, 2010.

Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes.* Springer, 1996.

Roman Vershynin. *High-dimensional probability.* Cambridge University Press, 2018.

M. Vidyasagar. *Nonlinear systems analysis.* SIAM, 2002.

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

Martin J. Wainwright. *High-dimensional statistics.* Cambridge University Press, 2019.

Daren Wang, Yi Yu, Alessandro Rinaldo, and Rebecca Willett. Localizing changes in high-dimensional vector autoregressive processes. *arXiv:1909.06359*, 2019a.

Daren Wang, Yi Yu, and Alessandro Rinaldo. Optimal change point detection and localization in sparse dynamic networks. *The Annals of Statistics*, 49(1):203–232, 2021a.

Daren Wang, Zifeng Zhao, Kevin Z. Lin, and Rebecca Willett. Statistically and computationally efficient change point localization in regression settings. *Journal of Machine Learning Research*, 22(248):1–46, 2021b.

Guanghui Wang, Changliang Zou, and Guosheng Yin. Change-point detection in multinomial data with a large number of categories. *The Annals of Statistics*, 46(5):2020–2044, 2018.

Tengyao Wang and Richard J. Samworth. High dimensional change point estimation via sparse projection. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 80(1):57–83, 2018.

Yunlong Wang, Changliang Zou, Zhaojun Wang, and Guosheng Yin. Multiple change-points detection in high dimension. *Random Matrices: Theory and Applications*, 8(4):1950014, 35, 2019b.

Mengjia Yu and Xiaohui Chen. Finite sample change point inference and identification for high-dimensional mean vectors. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 83(2):247–270, 2021.

Changliang Zou, Guosheng Yin, Long Feng, and Zhaojun Wang. Nonparametric maximum likelihood approach to multiple change-point problems. *The Annals of Statistics*, 42(3): 970–1002, 2014.