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

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-points detection; Non-asymptotic bounds; Rate-optimal estimators

1 Introduction

High-dimensional low-rank matrix recovery has witnessed a rapid development as well as a 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 Solé, 2011), 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 $\boldsymbol{\Theta}^\star \in \mathbb{R}^{m_1 \times m_2}$ is the unknown low-rank matrix to be estimated, and ϵ_i is some unobserved noise. It is worthy of 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 (Koltchinskii et al., 2011; Negahban and Wainwright, 2011; Rohde et al., 2011).

In real-life high-dimensional or big data application, the underlying data generating mechanism may encounter abrupt changes or transition along time or some other variable. For instance, in a recommendation system, user preference to 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,

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

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 $\boldsymbol{\Theta}_s^\star$ 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 $\boldsymbol{\Theta}_s^\star$'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 are linked via the model*

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

where $\tau_s^\star$'s are the change-points, $\boldsymbol{\Theta}_s^\star \in \mathbb{R}^{m_1 \times m_2}$ 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

$$t_i = t_a, \boldsymbol{X}_i = \boldsymbol{x}_a \boldsymbol{e}_b^\top, y_i = \boldsymbol{e}_b^\top \boldsymbol{y}_a, \epsilon_i = \boldsymbol{e}_b^\top \boldsymbol{w}_a, \ i = 1, \dots, N(= nm_2),$$

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 $a = 1, \dots, n$ and $b = 1, \dots, m_2$.

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*

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

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

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

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 there 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, Zhao, Lin and Willett (2021). Despite the popularity of huge volumes of data collected in matrix form nowadays, there are only limited number of estimation schemes designed for Model (1). For the VAR change model in Example 3, if the regression matrices $\mathbf{\Theta}_s^\star$'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, Yu, Rinaldo and Willett (2019) 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. (2021) developed a rolling window detection strategy.

In this paper, we attempt to develop 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 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-points detection for the general trace regression model with threshold effects (i.e., Model (1)). It builds on an recovery scheme that incorporates least-squares minimization with the nuclear norm penalty. To tailor for multiple change-points scenario, 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 at incorporating 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 Candes 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; Dette et al., 2022; Liu et al., 2020; Wang et al., 2018; Wang and Samworth, 2018; Yu and Chen, 2021), graphical models (Bybee and Atchadé, 2018; Liu et al., 2021; Londschien et al., 2021), networks (Wang, Yu and Rinaldo, 2021), and regression models (Bai et al., 2020, 2021; Kaul et al., 2019; Lee et al., 2016; Leonardi and Bühlmann, 2016; Safikhani and Shojaie, 2022; Wang, Zhao, Lin and Willett, 2021), 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 scenario 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 Supplementary Material.

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 $\mathrm{rank}(\boldsymbol{X})$ 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 $\|\boldsymbol{X}\|_{S_q} = \left\{\sum_{k=1}^{\mathrm{rank}(\boldsymbol{X})} \varrho_k(\boldsymbol{X})^q\right\}^{1/q}$. 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 $\langle \boldsymbol{X}_1 , \boldsymbol{X}_2 \rangle = \mathrm{tr}\left(\boldsymbol{X}_1^\top \boldsymbol{X}_2\right)$, where $\mathrm{tr}(\cdot)$ is the trace operator. For vectors, we use $\|\cdot\|_1$ and $\|\cdot\|_2$ for the ℓ_1 and ℓ_2 norms, respectively.

2 Matrix estimation with a possible change-point

2.1 Joint minimization scheme

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

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

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

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

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

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

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

see Proposition S.4. 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 Schattern-q norm; see Rohde et al. (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 $q \in [0, 1)$.

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 $S_N(\mathbf{\Gamma}; \tau) = (2N)^{-1} \sum_{i=1}^{N} \left(y_i - \langle \boldsymbol{\mathcal{X}}_i(\tau), \mathbf{\Gamma} \rangle \right)^2$, 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 $\mathbb{T} \subset [0, 1]$ represents a parameter space for $\tau^\star$. 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.,

$$\left(\widehat{\mathbf{\Gamma}}(\widehat{\tau}), \widehat{\tau} \right) = \underset{\mathbf{\Gamma} \in \mathbb{R}^{(2m_1)\times m_2}, \tau \in \mathbb{T}}{\arg\min} \left\{ S_N(\mathbf{\Gamma}; \tau) + \lambda_N \|\mathbf{\Gamma}\|_* \right\}. \tag{4}$$

Remark 1. *Since the nuclear norm is not separable, another form of penalization one might consider is* $\|\mathbf{\Theta}\|_* + \|\mathbf{\Delta}\|_*$, *for* $\mathbf{\Gamma} = \left(\mathbf{\Theta}^\top, \mathbf{\Delta}^\top \right)^\top$. *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* $(\|\mathbf{\Theta}\|_* + \|\mathbf{\Delta}\|_*)/\sqrt{2} \le \|(\mathbf{\Theta}^\top, \mathbf{\Delta}^\top)^\top\|_* \le \|\mathbf{\Theta}\|_* + \|\mathbf{\Delta}\|_*$, *see Proposition S.4. 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 Proposition S.5, both the penalization factor of the objective function and the non-asymptotic bounds can be rescaled up to some constants.*

2.2 Theoretical analysis

We will perform a thorough analysis on statistical properties of the regularized estimator $\big(\widehat{\mathbf{\Gamma}}(\widehat{\tau}), \widehat{\tau}\big)$. Let $\boldsymbol{y} = (y_1, \ldots, y_N)^\top$ and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_N)^\top$. Given $\tau \in \mathbb{T}$, define an observation operator $\mathfrak{X}(\cdot; \tau) : \mathbb{R}^{(2m_1)\times m_2} \to \mathbb{R}^N$, with elements $[\mathfrak{X}(\mathbf{\Gamma}; \tau)]_i = \langle \boldsymbol{\mathcal{X}}_i(\tau), \mathbf{\Gamma} \rangle$ for $\mathbf{\Gamma} \in \mathbb{R}^{(2m_1)\times m_2}$, and thus Model (2) can be reformulated as $\boldsymbol{y} = \mathfrak{X}(\mathbf{\Gamma}^\star; \tau^\star) + \boldsymbol{\epsilon}$. 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 $\boldsymbol{v} \in \mathbb{R}^N$. For $\tau, \tau' \in \mathbb{T}$, let

$$\begin{aligned}\mathcal{R}_N(\mathbf{\Gamma}^\star, \tau, \tau') &= N^{-1} \sum_{i=1}^N \epsilon_i \langle \boldsymbol{\mathcal{X}}_i(\tau) - \boldsymbol{\mathcal{X}}_i(\tau'), \mathbf{\Gamma}^\star \rangle \\ &= N^{-1} \sum_{i=1}^N \epsilon_i \langle \boldsymbol{X}_i(\tau) - \boldsymbol{X}_i(\tau'), \mathbf{\Delta}^\star \rangle,\end{aligned}$$

which will play a crucial role in our analysis.

The first ingredient in our analysis is the specification of certain subspaces onto which we can project the matrices and utilize the low-rank structure. To formalize the idea, consider the singular value decomposition of the target matrix $\mathbf{\Gamma}^\star$. For each integer $r \in \{1, \ldots, m\}$, let $\mathbb{U}^r := [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r] \in \mathbb{R}^{m_1 \times r}$ and $\mathbb{V}^r := [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r] \in \mathbb{R}^{m_2 \times r}$ be the subspaces spanned by the top r left and right singular vectors of $\mathbf{\Gamma}^\star$. 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 $\boldsymbol{y}\boldsymbol{v}_k^\top$, $k = 1, \ldots, r$, where $\boldsymbol{x}$ and $\boldsymbol{y}$ are arbitrary, and $\mathcal{S}^{r\perp}$ is its orthogonal complement. The orthogonal projection $\Pi^r_{\mathbf{\Gamma}^\star}$ onto $\mathcal{S}^r$ is given by $\Pi^r_{\mathbf{\Gamma}^\star}(\boldsymbol{M}) = \boldsymbol{P}_{\mathbb{U}^r}\boldsymbol{M} + \boldsymbol{M}\boldsymbol{P}_{\mathbb{V}^r} - \boldsymbol{P}_{\mathbb{U}^r}\boldsymbol{M}\boldsymbol{P}_{\mathbb{V}^r}$ 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 $\Pi^{r\perp}_{\mathbf{\Gamma}^\star}$ onto $\mathcal{S}^{r\perp}$ is given by $\Pi^{r\perp}_{\mathbf{\Gamma}^\star}(\boldsymbol{M}) = (\boldsymbol{I}_{m_1} - \boldsymbol{P}_{\mathbb{U}^r})\boldsymbol{M}(\boldsymbol{I}_{m_2} - \boldsymbol{P}_{\mathbb{V}^r})$. 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 now proceed to provide an inequality that builds up the foundation of our theory.

Lemma 1 (Basic inequality). *If* $\lambda_N \geq \sup_{\tau \in \mathbb{T}} 2\|\mathfrak{X}^\star(\boldsymbol{\epsilon}; \tau)\|_{\mathrm{op}}/N$, *then*

$$\begin{aligned}&\frac{1}{2N} \sum_{i=1}^N \Big(\Big\langle \boldsymbol{\mathcal{X}}_i(\widehat{\tau}), \widehat{\mathbf{\Gamma}} \Big\rangle - \langle \boldsymbol{\mathcal{X}}_i(\tau^\star), \mathbf{\Gamma}^\star \rangle \Big)^2 + \frac{\lambda_N}{2} \Big\| \Pi^{r\perp}_{\mathbf{\Gamma}^\star} \Big(\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^\star \Big) \Big\|_* \\ \leq &2\lambda_N \big\| \Pi^{r\perp}_{\mathbf{\Gamma}^\star} (\mathbf{\Gamma}^\star) \big\|_* + \frac{3\lambda_N}{2} \Big\| \Pi^r_{\mathbf{\Gamma}^\star} \Big(\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^\star \Big) \Big\|_* + \mathcal{R}_N(\mathbf{\Gamma}^\star, \widehat{\tau}, \tau^\star). \end{aligned} \tag{5}$$

Lemma 1 is a deterministic result, but conditioned on the event $\lambda_N \geq \sup_{\tau\in\mathbb{T}} 2\|\mathfrak{X}^\star(\boldsymbol{\epsilon};\tau)\|_{\text{op}}/N$, 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 $\lambda_N \geq 2\|\sum_{i=1}^N \epsilon_i \boldsymbol{X}_i\|_{\text{op}}/N$. Our choice here incorporates the change structure information. We shall show in Section 2.3 that with suitable choice of λ_N, this event holds with high probability.

The left-hand side of (5) in Lemma 1 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. 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 we have the following upper bound on the prediction error.

Corollary 1 (Prediction consistency)**.** *If* $\lambda_N \geq \sup_{\tau\in\mathbb{T}} 2\|\mathfrak{X}^\star(\boldsymbol{\epsilon};\tau)\|_{\text{op}}/N$*, then*

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

Corollary 1 gives the consistency property of the prediction error under the scaling that $\lambda_N\sum_{k=r+1}^m \rho_k(\boldsymbol{\Gamma}^\star)\to 0$, $\lambda_N r\gamma_{max}\to 0$ and $\lambda_N\|\boldsymbol{\Delta}^\star\|_*\to 0$.

To control over 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 in the presence of a change-point.

Assumption 1 (Restricted strong convexity, RSC)**.** *Let* $\mathbb{S}=[\tau^\star-c_\tau,\tau^\star+c_\tau]\subset\mathbb{T}$ *be a neighborhood of the change-point* $\tau^\star$ *with* $c_\tau\geq 0$*. The restricted strong convexity condition holds with curvature* $\kappa(\mathfrak{X})>0$ *if*

$$\frac{1}{2N}\|\mathfrak{X}(\boldsymbol{M};\tau)\|_2^2 \geq \kappa(\mathfrak{X})\|\boldsymbol{M}\|_F^2,\ \textit{for all}\ \boldsymbol{M}\in\mathcal{C}(r,\delta,\boldsymbol{\Gamma}^\star,\mathbb{S}),\ \tau\in\mathbb{S}, \tag{6}$$

where for some $\delta\geq 0$*,*

$$\begin{aligned}&\mathcal{C}(r,\delta,\boldsymbol{\Gamma}^\star,\mathbb{S})\\ =&\Bigg\{\boldsymbol{M}\in\mathbb{R}^{(2m_1)\times m_2}:\|\boldsymbol{M}\|_F\geq\delta,\ \|\Pi^{r\perp}_{\boldsymbol{\Gamma}^\star}(\boldsymbol{M})\|_*\leq 3\|\Pi^{r}_{\boldsymbol{\Gamma}^\star}(\boldsymbol{M})\|_*\\ &+4\sum_{k=r+1}^m\rho_k(\boldsymbol{\Gamma}^\star)+2\min\{\sqrt{c_\tau}\|\boldsymbol{\Delta}^\star\|_F,\|\boldsymbol{\Delta}^\star\|_*\}\Bigg\}.\end{aligned} \tag{7}$$

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

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

First, 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 $\tau \in \mathbb{S}$. This unification guarantees a local strong convexity property and eliminates the scenario where "bad" positioning of the change-point ruins the behavior of the estimator. It's worthy of noticing that (6) 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, (7) has an additional term in the right-hand side of the second inequality, i.e., $2\min\{\sqrt{c_\tau}\|\boldsymbol{\Delta}^\star\|_F, \|\boldsymbol{\Delta}^\star\|_*\}$, 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 (7) is reduced to the classic $\mathcal{C}(r, \delta, \boldsymbol{\Theta}^\star)$. At last, it is remarkable to point out that the δ in the set (7) is used to account for the term $\sum_{k=r+1}^{m}\rho_k(\boldsymbol{\Gamma}^\star)$ in near low-rank situation. This means that for the exact low-rank cases, we can safely set $\delta = 0$. We shall show in further examples that this RSC holds under some random design scenarios with high probability.

With Assumption 1 and the basic inequality (5) in Lemma 1, we can readily obtain some interesting bounds on the matrix estimation and change-point detection error. Before going further, 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 situation. The answer is actually positive as summarized below.

Theorem 1 (Matrix estimation with no threshold effect)**.** *Assume* $\boldsymbol{\Delta}^\star = \mathbf{0}$*, and that*

Assumption 1 holds for some $\kappa(\mathfrak{X}) > 0$ *with* $\mathbb{S} = \mathbb{T}$. *If* $\lambda_N \geq \sup_{\tau\in\mathbb{T}} 2\|\mathfrak{X}^\star(\boldsymbol{\epsilon};\tau)\|_{\mathrm{op}}/N$, *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 1 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 $\sum_{k=r+1}^m \rho_k(\boldsymbol{\Gamma}^\star)$ 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.,

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

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

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

Assumption 2 (Identifiability and discontinuity)**.** *Assume* $\boldsymbol{\Gamma}^\star \in \mathbb{B}_q(R_q)$ *for some* $R_q > 0$ *with* $q \in [0,1)$, *and* $\boldsymbol{\Delta}^\star \neq \mathbf{0}$. *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* $\tau \in \mathbb{S} = [\tau^\star - c_\tau, \tau^\star + c_\tau] \subset \mathbb{T}$ *with* $|\tau - \tau^\star| > \eta(N, m_1, m_2)$ *and* $\boldsymbol{\Gamma} \in \{\boldsymbol{\Gamma} : \|\boldsymbol{\Gamma}\|_{S_q}^q \leq R_q'\}$ *with* $\boldsymbol{\Gamma} - \boldsymbol{\Gamma}^\star \in \mathcal{C}(r, \delta, \boldsymbol{\Gamma}^\star, \mathbb{S})$, *it holds that*

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

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

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 larger value of certain norm of $\mathbf{\Delta}^\star$ corresponds to 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 2.

Lemma 2 (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 2 is sufficient to establish the estimation consistency of $\widehat{\tau}$ if $\phi(\mathbf{\Delta}^\star)^{-1}\lambda_N \sum_{k=r+1}^{m} \rho_k(\mathbf{\Gamma}^\star) \to 0$, $\phi(\mathbf{\Delta}^\star)^{-1}\lambda_N r\gamma_{max} \to 0$ and $\phi(\mathbf{\Delta}^\star)^{-1}\lambda_N\|\mathbf{\Delta}^\star\|_* \to 0$. 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 certain type of smoothness on the design.

Assumption 3 (Smoothness of design)**.** *There exists some constant* $C > 0$*, such that for any* $\tau \in \mathbb{S}' = [\tau^\star - c'_\tau, \tau^\star + c'_\tau] \subset \mathbb{T}$ *with* $|\tau - \tau^\star| > \eta(N, m_1, m_2)$ *and* $0 \leq c'_\tau < c_\tau$ *and for any* $\mathbf{\Gamma}$ *satisfying* $\mathbf{\Gamma} - \mathbf{\Gamma}^\star \in \mathcal{C}(r, \delta, \mathbf{\Gamma}^\star, \mathbb{S}') \cap \{\boldsymbol{M} : \|\boldsymbol{M}\|_* \leq c_{\mathbf{\Gamma}}\}$ *with some* $c_{\mathbf{\Gamma}} > 0$*, it holds that*

$$|\mathcal{T}_N(\mathbf{\Gamma}, \mathbf{\Gamma}^\star, \tau, \tau^\star)| \leq Cc'_\tau c_{\mathbf{\Gamma}}\|\mathbf{\Delta}^\star\|_*,$$

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 circumstance; see Section 2.3.

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 $\mathbb{E}\{\exp(\epsilon^2/K^2)\} \leq e$.*

Starting from this assumption we begin to introduce probabilistic structure for the noise. Now our choice of λ_N, i.e. $\lambda_N \geq \sup_{\tau\in\mathbb{T}} 2\|\mathfrak{X}^\star(\boldsymbol{\epsilon};\tau)\|_{\text{op}}/N$, 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 2.3.

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

Lemma 3. *Let $h_N(c_\tau) = (2c_\tau N)^{-1}\sum_{i:|t_i-\tau^\star|\leq c_\tau}\langle \boldsymbol{X}_i\,,\,\boldsymbol{\Delta}^\star\rangle^2$. 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*

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

Note the quantity $\|\boldsymbol{\Delta}^\star\|_F^{-2}h_N(c_\tau)$ in Lemma 3 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 $\|\boldsymbol{X}_i\|_{\text{op}} \leq \gamma'_{max}$ for some $\gamma'_{max} > 0$, then $\|\boldsymbol{\Delta}^\star\|_F^{-2}h_N(c_\tau) \leq \text{rank}(\boldsymbol{\Delta}^\star)\gamma'^2_{max}$, 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 design, see, for example, Section 2.3.

Now based on Lemma 3 and the comment about 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 involving parameters.

Assumption 5. *The following conditions hold:*

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

Basically Assumption 5 guarantees small magnitudes for several key quantities in our analysis, such as $\lambda_N, \|\Pi^{r\perp}_{\boldsymbol{\Gamma}^\star}(\boldsymbol{\Gamma}^\star)\|_*$, etc. These inequalities can hold simultaneously when N is sufficiently large, under the scaling that $\lambda_N \to 0$, $r \to \infty$ and $\lambda_N r \to 0$ for fixed $\boldsymbol{\Delta}^\star$. We have commented before that proper scaling of these quantities can contribute significantly to controlling the errors of interest.

Theorem 2 (Recovery accuracy with threshold effect)**.** *Suppose that Assumption 1–Assumption 5 hold. Assume* $\lambda_N \geq \sup_{\tau\in\mathbb{T}} \frac{2}{N}\|\mathfrak{X}^\star(\boldsymbol{\epsilon};\tau)\|_{\mathrm{op}}$ *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* $1-\alpha_N-2e\sum_{k=1}^{m^\star}\exp\left(-c'N\lambda_N^2/\{K^2\|\boldsymbol{\Delta}^\star\|_F^{-2}h_N(c_\tau^{(k)})\}\right)$*:*

$$
\begin{aligned}
\left\|\widehat{\boldsymbol{\Gamma}}-\boldsymbol{\Gamma}^\star\right\|_F^2 &\leq \delta^2 \vee \frac{8\lambda_N\sum_{k=r+1}^m \rho_k(\boldsymbol{\Gamma}^\star)}{\kappa(\mathfrak{X})} \vee \frac{128\lambda_N^2 r}{\kappa(\mathfrak{X})^2},\\
\left\|\widehat{\boldsymbol{\Gamma}}-\boldsymbol{\Gamma}^\star\right\|_* &\leq 12\sqrt{2r}\delta \vee 12\sum_{k=r+1}^m \rho_k(\boldsymbol{\Gamma}^\star) \vee \frac{192\lambda_N r}{\kappa(\mathfrak{X})},\\
\frac{1}{2N}\left\|\mathfrak{X}\left(\widehat{\boldsymbol{\Gamma}};\widehat{\tau}\right)-\mathfrak{X}(\boldsymbol{\Gamma}^\star;\tau^\star)\right\|_2^2 &\leq 6\lambda_N\sqrt{2r}\delta \vee 6\lambda_N\sum_{k=r+1}^m \rho_k(\boldsymbol{\Gamma}^\star) \vee \frac{96\lambda_N^2 r}{\kappa(\mathfrak{X})},\\
|\widehat{\tau}-\tau^\star| &\leq 20\{c\phi(\boldsymbol{\Delta}^\star)\}^{-1}\lambda_N\sqrt{2r}\delta \vee 20\{c\phi(\boldsymbol{\Delta}^\star)\}^{-1}\lambda_N\sum_{k=r+1}^m \rho_k(\boldsymbol{\Gamma}^\star) \vee \frac{320\{c\phi(\boldsymbol{\Delta}^\star)\}^{-1}\lambda_N^2 r}{\kappa(\mathfrak{X})}.
\end{aligned}
$$

Theorem 2 gives the same bounds (up to some constants) as those in Theorem 1 for the matrix estimation error as well as the prediction error in the presence of threshold effect. In addition, Theorem 2 builds the error bound for change-point detection, which generally refines that obtained in Lemma 2. 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 $\phi(\boldsymbol{\Delta}^\star)$. Hence an improvement incurs by noticing that Lemma 2 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 2 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 Lemma S.1 and Lemma S.2 in Supplementary Material. This procedure requires more techniques due to the complexity of matrix formulation (especially that based on near low-rank matrices).

Remark 2. *Theorem 2 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 3, $N\lambda_N^2/\{\|\mathbf{\Delta}^\star\|_F^{-2} h_N(c_\tau^{(k)})\}$ 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 2.3), 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 2, we restate it in two concrete scenarios, namely, the exact and near low-rank matrix recovery.

Corollary 2 (Exact low-rank matrix recovery)**.** *Suppose the conditions in Theorem 2 hold. In particular, assume $\mathbf{\Gamma}^\star$ is an exact low-rank matrix with rank r and Assumption 1 holds with $\mathcal{C}(r, 0, \mathbf{\Gamma}^\star, \mathbb{S})$ 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 $1-\alpha_N-2e\sum_{k=1}^{m^\star}\exp\left(-c'N\lambda_N^2/\{K^2\|\mathbf{\Delta}^\star\|_F^{-2}h_N(c_\tau^{(k)})\}\right)$*:*

$$\left\|\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}^\star\right\|_F^2\le\frac{128\lambda_N^2 r}{\kappa(\mathfrak{X})^2},\quad\left\|\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}^\star\right\|_*\le\frac{192\lambda_N r}{\kappa(\mathfrak{X})},$$
$$\frac{1}{2N}\left\|\mathfrak{X}\left(\widehat{\mathbf{\Gamma}};\widehat{\tau}\right)-\mathfrak{X}(\mathbf{\Gamma}^\star;\tau^\star)\right\|_2^2\le\frac{96\lambda_N^2 r}{\kappa(\mathfrak{X})},$$
$$|\widehat{\tau}-\tau^\star|\le\frac{320\{c\phi(\mathbf{\Delta}^\star)\}^{-1}\lambda_N^2 r}{\kappa(\mathfrak{X})}.$$

Corollary 3 (Near low-rank matrix recovery)**.** *Suppose the conditions in Theorem 2 hold. In particular, assume* $\mathbf{\Gamma}^\star\in\mathbb{B}_q(R_q)$ *for some* $q\in[0,1)$ *and Assumption 1 holds with* $\mathcal{C}(R_q\lambda_N^{-q},\delta,\mathbf{\Gamma}^\star,\mathbb{S})$ *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* $1-\alpha_N-2e\sum_{k=1}^{m^\star}\exp\left(-c'N\lambda_N^2/\{K^2\|\mathbf{\Delta}^\star\|_F^{-2}h_N(c_\tau^{(k)})\}\right)$*:*

$$\left\|\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}^\star\right\|_F^2\le\delta^2\vee\frac{128\lambda_N^{2-q}R_q}{\kappa(\mathfrak{X})^{2-q}},\quad\left\|\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}^\star\right\|_*\le12\sqrt{2R_q}\lambda_N^{-q/2}\delta\vee\frac{192R_q\lambda_N^{1-q}}{\kappa(\mathfrak{X})^{1-q}},$$
$$\frac{1}{2N}\left\|\mathfrak{X}\left(\widehat{\mathbf{\Gamma}};\widehat{\tau}\right)-\mathfrak{X}(\mathbf{\Gamma}^\star;\tau^\star)\right\|_2^2\le6\lambda_N^{1-q/2}\sqrt{2R_q}\delta\vee\frac{96\lambda_N^{2-q}R_q}{\kappa(\mathfrak{X})^{2-q}},$$
$$|\widehat{\tau}-\tau^\star|\le20\{c\phi(\mathbf{\Delta}^\star)\}^{-1}\lambda_N^{1-q/2}\sqrt{2R_q}\delta\vee\frac{320\{c\phi(\mathbf{\Delta}^\star)\}^{-2}\lambda_N^{2-q}R_q}{\kappa(\mathfrak{X})^{2-q}}.$$

Proof of Corollary 2 is quite straightforward by noticing that $\delta=0$ and $\|\Pi_{\mathbf{\Gamma}^\star}^{r\perp}(\mathbf{\Gamma}^\star)\|_*=0$ under the exact low-rank assumption. The error bounds in Corollary 3 reduces to those in Corollary 2 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.

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

Up to now we are mainly investing our efforts in fixed design case for general estimation and detection results. The assumptions we proposed have natural theoretical and practical interpretation, 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.

Regarding Example 1, let $\boldsymbol{\mathcal{X}}_a(\tau) = \left(\boldsymbol{x}_a^\top, \boldsymbol{x}_a^\top \mathbf{1}\{t_a > \tau\}\right)^\top$ for some $\tau \in \mathbb{T} = [\rho, 1-\rho] \subset [0,1]$, where ρ is some boundary removal parameter that is frequently considered in the change-point detection literature (Csörgő and Horváth, 1997). 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 $\boldsymbol{\Gamma}^\star = \left(\boldsymbol{\Theta}_0^{\star\top}, \boldsymbol{\Theta}_1^{\star\top} - \boldsymbol{\Theta}_0^{\star\top}\right)^\top$. In this case our procedure proceeds as

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

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* $t_a \sim U(0,1)$, $\boldsymbol{x}_a \sim \mathcal{N}(0, \sigma_0^2 \boldsymbol{I}_{m_1})$ *and* $\boldsymbol{\epsilon}_a \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_{m_2})$.

Theorem 3. *Assume* $\boldsymbol{\Gamma}^\star \in \mathbb{B}_q(R_q)$ *for some* $q \in [0,1)$. *If the regularization parameter* λ_n *is chosen such that* $\lambda_n = 20\sigma\sigma_0\sqrt{(m_1+m_2)/n}$, *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*

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

we have

$$\begin{aligned}
\left\|\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^\star\right\|_F^2 &\le C_0 R_q \left(\frac{\sigma}{\sigma_0}\right)^{2-q} \left(\frac{m_1+m_2}{n}\right)^{(1-q/2)}, \\
\left\|\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^\star\right\|_* &\le C_0 R_q \left(\frac{\sigma}{\sigma_0}\right)^{1-q} \left(\frac{m_1+m_2}{n}\right)^{(1/2-q/2)}, \\
\frac{1}{2n}\left\|\mathfrak{X}\left(\widehat{\boldsymbol{\Gamma}}; \widehat{\tau}\right) - \mathfrak{X}(\boldsymbol{\Gamma}^\star; \tau^\star)\right\|_2^2 &\le C_0 R_q \left(\frac{\sigma}{\sigma_0}\right)^{2-q} \left(\frac{m_1+m_2}{n}\right)^{(1-q/2)}, \\
|\widehat{\tau} - \tau^\star| &\le C_0 R_q \left(\frac{\sigma}{\sigma_0}\right)^{2-q} \left(\frac{m_1+m_2}{n}\right)^{(1-q/2)}.
\end{aligned}$$

Theorem 3 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).

2.4 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., $S_{\text{Major}}(\mathbf{\Gamma}; \mathbf{\Gamma}') := S_N(\mathbf{\Gamma}') + \langle \nabla S_N(\mathbf{\Gamma}') \,, \mathbf{\Gamma} - \mathbf{\Gamma}' \rangle + \frac{L}{2} \|\mathbf{\Gamma} - \mathbf{\Gamma}'\|_F^2$ 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 $\boldsymbol{G} = \mathbf{U}_{\boldsymbol{G}}^\top \text{diag}\{(\rho_i(\boldsymbol{G}))\} \mathbf{V}_{\boldsymbol{G}}$, $\text{Soft}(\boldsymbol{G}; \xi) = \mathbf{U}_{\boldsymbol{G}}^\top \,\text{diag}\{((\varrho_i(\boldsymbol{G}) - \xi)_+)\} \mathbf{V}_{\boldsymbol{G}}$ with $x_+ = \max\{x, 0\}$.

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

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

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 desirable error rate. The spirit of refinements over inefficient or sub-optimal initial change-point estimators is popular in the literature of multiple change-points detection (Harchaoui and Lévy-Leduc, 2010; Zou et al., 2014), and has been further explored for high-dimensional change detection, see, for example, Wang, Yu and Rinaldo (2021) and Bai et al. (2021). 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 neighbourhoods (Bai et al., 2021; Safikhani et al., 2022; Safikhani and Shojaie, 2022). In other words, change detection and parameter estimation are performed separately, which may be inefficient in practice.

We attempt to fulfill the refinements of both the change-point and regression matrix estimators in a joint manner. 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 to 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.

Algorithm 1 previews the two-stage procedure for joint multiple change-points 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 change magnitude is denoted by $\widehat{\mathbf{\Delta}}_i$. According to Theorem 1, 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 2, a large value of $\widehat{\mathbf{\Delta}}_i$ may indicate that t_i is located around

Algorithm 1: Joint multiple change-points 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 $\mathcal{G} = \{t_i\}_{i=1}^N \cap [\omega, 1-\omega]$

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 $\widehat{\boldsymbol{\Theta}}_s$

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 $s = 1, \ldots, \widetilde{s}$.

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

$$\mathcal{E}_N := \{\widetilde{s} = s^\star \text{ and } \max_{s=1,\ldots,\widetilde{s}} |\widetilde{\tau}_s - \tau_s^\star| \leq d_{min}/6\}. \tag{8}$$

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

Theorem 4. *Suppose Assumption S1–Assumption S6 in Supplemental Material (parallel to those in Corollary 2) 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*

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

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* $\tilde{m}^\star > 0$.

Corollary 4. *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 3 (Alternative choices in Stage I)**.** *The thresholding rule based procedure provide 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 (8) 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, Zou, Wang and Yin, 2019; 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 $\mathbb{E}(\boldsymbol{Z}_i) = \boldsymbol{\Xi}\mathrm{vec}(\boldsymbol{\Theta}_i)$, where $\boldsymbol{\Xi} = \mathbb{E}\left\{\mathrm{vec}(\boldsymbol{X}_i)\mathrm{vec}(\boldsymbol{X}_i)^\top\right\}$. 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.

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.

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 $N_m(\boldsymbol{0}, \boldsymbol{I}_m)$, and the noises are i.i.d. copies from $N_m(\boldsymbol{0}, 0.1^2 \boldsymbol{I}_m)$. 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 S.4.1 for more details) with $\|\mathbf{\Theta}_1^\star\|_F = \|\mathbf{\Theta}_2^\star\|_F = 1$ and a break $\|\mathbf{\Theta}_1^\star - \mathbf{\Theta}_2^\star\|_F = 0.1$. 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 1.

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 $\mathbf{\Theta}_1^\star$ and $\mathbf{\Theta}_2^\star$ share the same first four singular vectors and demonstrate a small break size (see Section S.4.1). In Supplementary Material 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 specification 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 2. 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.

Table 1: 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)

Table 2: Compressed sensing with a single change-point

Method	$\lvert\widehat{\tau}-\tau^\star\rvert$	$\boldsymbol{\Theta}_1$			$\boldsymbol{\Theta}_2$		
		$\lVert\widehat{\boldsymbol{\Theta}}_1-\boldsymbol{\Theta}_1^\star\rVert_F^2$	$\lVert\widehat{\boldsymbol{\Theta}}_1-\boldsymbol{\Theta}_1^\star\rVert_*$	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 $(m,N)=(40,1500)$							
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)

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 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.,

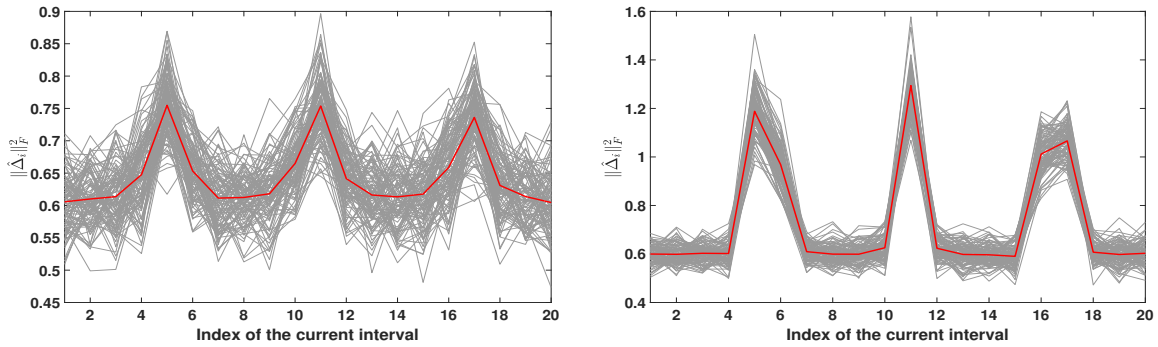
$$\mathrm{MOE} = \sup_{s=1,\cdots,s^\star} \inf_{s'=1,\cdots,\widehat{s}} \|\widehat{\boldsymbol{\Theta}}_{s'} - \boldsymbol{\Theta}_s^\star\|_F^2, \ \mathrm{MUE} = \sup_{s'=1,\cdots,\widehat{s}} \inf_{s=1,\cdots,s^\star} \|\widehat{\boldsymbol{\Theta}}_{s'} - \boldsymbol{\Theta}_s^\star\|_F^2.$$

Besides, we report the maximal and minimal estimated rank across segments. Results over 100 replications are summarized in Table 3 and Figure 1.

Table 3: 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)

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 for augmenting and stabilizing the performance of the roughly selected change-points. 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



(a) Small breaks (b) Large breaks

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

of $\|\widehat{\mathbf{\Delta}}\|_F^2$ in Figure 1 reflects the contrast of difficulty with different magnitudes of change signal.

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 concentration of PM and four air pollutants: sulfur dioxide (SO_2), carbon monoxide (CO), and 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

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

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 Learing 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

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

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% ($N_{\text{train}} = 880$). We apply Algorithm 1 by choosing different stopping thresholds ζ_N and construct models with varying number of change-points. The training and test errors are measured respectively by

$$\texttt{Err}_{\text{train}} = \frac{1}{m_2 N_{\text{train}}} \|\mathbf{Y}_{\text{train}} - \widehat{\mathbf{Y}}_{\text{train}}\|_F^2, \ \texttt{Err}_{\text{test}} = \frac{1}{m_2 N_{\text{test}}} \|\mathbf{Y}_{\text{test}} - \widehat{\mathbf{Y}}_{\text{test}}\|_F^2.$$

Table 4 reports the training 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 condition, since the Chinese government took many actions in 2016 to improve the air quality, including improving law system, promoting clean energy, encouraging the development of green industries, etc.[2]

Table 4: Training and test errors for the air pollution data

#Change-points	0	1	2	3	4
Test Error	0.1925	0.1746	**0.1728**	0.1761	0.1772
Training error	0.1976	0.1745	0.1592	0.1448	0.1360

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 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 align 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 (Fan et al., 2021; Tan et al., 2022). 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 endeavor.

SUPPLEMENTARY MATERIAL

[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.

The Supplementary Material contains the proofs of all theoretical results presented in this article and some necessary lemmas, together with additional numerical studies.

References

Athey, S., Bayati, M., Doudchenko, N., Imbens, G. and Khosravi, K. (2021), 'Matrix completion methods for causal panel data models', *Journal of the American Statistical Association* **116**(536), 1716–1730.

Bai, P., Safikhani, A. and Michailidis, G. (2020), 'Multiple change points detection in low rank and sparse high dimensional vector autoregressive models', *IEEE Transactions on Signal Processing* **68**, 3074–3089.

Bai, P., Safikhani, A. and Michailidis, G. (2021), 'Multiple change point detection in reduced rank high dimensional vector autoregressive models', *arXiv preprint arXiv:2109.14783* .

Bybee, L. and Atchadé, Y. (2018), 'Change-point computation for large graphical models: a scalable algorithm for gaussian graphical models with change-points', *The Journal of Machine Learning Research* **19**(1), 440–477.

Candes, E. J. and Plan, Y. (2011), '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.

Candès, E. J. and Recht, B. (2009), 'Exact matrix completion via convex optimization', *Foundations of Computational mathematics* **9**(6), 717–772.

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

Chen, Y. and Chi, Y. (2018), '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.

Cho, H. and Fryzlewicz, P. (2015), '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.

Csörgő, M. and Horváth, L. (1997), *Limit theorems in change-point analysis*, Wiley Series in Probability and Statistics, John Wiley & Sons, Ltd., Chichester. With a foreword by David Kendall.

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

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

Espinosa-Vega, M. A. and Solé, J. (2011), 'Cross-border financial surveillance: a network perspective', *Journal of Financial Economic Policy* .

Fan, J., Gong, W. and Zhu, Z. (2019), 'Generalized high-dimensional trace regression via nuclear norm regularization', *Journal of Econometrics* **212**(1), 177–202.

Fan, J., Wang, W. and Zhu, Z. (2021), 'A shrinkage principle for heavy-tailed data: High-dimensional robust low-rank matrix recovery', *The Annals of Statistics* **49**(3), 1239–1266.

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

Golbabaee, M. and Vandergheynst, P. (2012), Hyperspectral image compressed sensing via low-rank and joint-sparse matrix recovery, *in* '2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', pp. 2741–2744.

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

Ji, S. and Ye, J. (2009), An accelerated gradient method for trace norm minimization, *in* 'Proceedings of the 26th annual international conference on machine learning', pp. 457–464.

Kaul, A., Jandhyala, V. K. and Fotopoulos, S. B. (2019), 'An efficient two step algorithm for high dimensional change point regression models without grid search.', *J. Mach. Learn. Res.* **20**, 111–1.

Keshavan, R. H., Montanari, A. and Oh, S. (2010), 'Matrix completion from noisy entries', *The Journal of Machine Learning Research* **11**, 2057–2078.

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

Lee, S., Seo, M. H. and Shin, Y. (2016), 'The lasso for high dimensional regression with a possible change point', *Journal of the Royal Statistical Society. Series B, Statistical methodology* **78**(1), 193.

Leonardi, F. and Bühlmann, P. (2016), 'Computationally efficient change point detection for high-dimensional regression', *arXiv preprint arXiv:1601.03704* .

Liu, B., Zhang, X. and Liu, Y. (2021), 'Simultaneous change point inference and structure recovery for high dimensional gaussian graphical models', *Journal of Machine Learning Research* **22**(274), 1–62.

Liu, B., Zhou, C., Zhang, X. and Liu, Y. (2020), 'A unified data-adaptive framework for high dimensional change point detection', *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82**(4), 933–963.

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

Negahban, S. and Wainwright, M. J. (2011), 'Estimation of (near) low-rank matrices with noise and high-dimensional scaling', *The Annals of Statistics* pp. 1069–1097.

Nesterov, Y. (2013), 'Gradient methods for minimizing composite functions', *Mathematical Programming* **140**(1), 125–161.

Nobre, F. F. and Stroup, D. F. (1994), 'A monitoring system to detect changes in public health surveillance data', *International journal of epidemiology* **23**(2), 408–418.

Page, E. S. (1954), 'Continuous inspection schemes', *Biometrika* **41**(1/2), 100–115.

Ramlatchan, A., Yang, M., Liu, Q., Li, M., Wang, J. and Li, Y. (2018), 'A survey of matrix completion methods for recommendation systems', *Big Data Mining and Analytics* **1**(4), 308–323.

Recht, B. (2011), 'A simpler approach to matrix completion.', *Journal of Machine Learning Research* **12**(12).

Recht, B., Fazel, M. and Parrilo, P. A. (2010), 'Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization', *SIAM review* **52**(3), 471–501.

Rinaldo, A., Wang, D., Wen, Q., Willett, R. and Yu, Y. (2021), Localizing changes in high-dimensional regression models, *in* 'International Conference on Artificial Intelligence and Statistics', PMLR, pp. 2089–2097.

Rohde, A., Tsybakov, A. B. et al. (2011), 'Estimation of high-dimensional low-rank matrices', *The Annals of Statistics* **39**(2), 887–930.

Safikhani, A., Bai, Y. and Michailidis, G. (2022), 'Fast and scalable algorithm for detection of structural breaks in big var models', *Journal of Computational and Graphical Statistics* **31**(1), 176–189.

Safikhani, A. and Shojaie, A. (2022), 'Joint Structural Break Detection and Parameter Estimation in High-Dimensional Nonstationary VAR Models', *J. Amer. Statist. Assoc.* **117**(537), 251–264.

Tan, K. M., Sun, Q. and Witten, D. (2022), 'Sparse reduced rank huber regression in high dimensions', *Journal of the American Statistical Association* pp. 1–11.

Toh, K.-C. and Yun, S. (2010), 'An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems', *Pacific Journal of Optimization* **6**(615-640), 15.

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

Wang, D., Yu, Y. and Rinaldo, A. (2021), 'Optimal change point detection and localization in sparse dynamic networks', *Ann. Statist.* **49**(1), 203–232.
URL: *https://doi.org/10.1214/20-AOS1953*

Wang, D., Yu, Y., Rinaldo, A. and Willett, R. (2019), 'Localizing changes in high-dimensional vector autoregressive processes', *arXiv preprint arXiv:1909.06359* .

Wang, D., Zhao, Z., Lin, K. Z. and Willett, R. (2021), 'Statistically and computationally efficient change point localization in regression settings', *Journal of Machine Learning Research* **22**(248), 1–46.

Wang, G., Zou, C. and Yin, G. (2018), 'Change-point detection in multinomial data with a large number of categories', *Ann. Statist.* **46**(5), 2020–2044.

Wang, T. and Samworth, R. J. (2018), 'High dimensional change point estimation via sparse projection', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80**(1), 57–83.

Wang, Y., Zou, C., Wang, Z. and Yin, G. (2019), 'Multiple change-points detection in high dimension', *Random Matrices Theory Appl.* **8**(4), 1950014, 35.

Yu, M. and Chen, X. (2021), 'Finite sample change point inference and identification for high-dimensional mean vectors', *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **83**(2), 247–270.

Zou, C., Yin, G., Feng, L. and Wang, Z. (2014), 'Nonparametric maximum likelihood approach to multiple change-point problems', *The Annals of Statistics* **42**(3), 970–1002.