이 학 박 사 학 위 논 문

Nonconvex penalized matrix completion methods for causal inference in panel data

패널 자료의 인과 추론을 위한 비볼록 벌점화 행렬 완성 방법

2023년 8월

서울대학교 대학원

통계학과

김 보 영

Nonconvex penalized matrix completion methods for causal inference in panel data

패널 자료의 인과 추론을 위한 비볼록 벌점화 행렬 완성 방법

지도교수 김 용 대

이 논문을 이학박사 학위논문으로 제출함

2023년 4월

서울대학교 대학원

통계학과

김 보 영

김보영의 이학박사 학위논문을 인준함

2023년 6월

위 원 장	임 요 한	(인)
부위원장	김 용 대	(인)
위 원	정 성 규	(인)
위 원	이 권 상	(인)
위 원	온 일 상	(인)

Nonconvex penalized matrix completion methods for causal inference in panel data

By

Bo Young Kim

A Thesis
Submitted in fulfillment of the requirement
for the degree of
Doctor of Philosophy
in Statistics

Department of Statistics
College of Natural Sciences
Seoul National University
August, 2023

ABSTRACT

Nonconvex penalized matrix completion methods for causal inference in panel data

Bo Young Kim
The Department of Statistics
The Graduate School
Seoul National University

Low-rank matrix completion is a widely used approach for imputing missing entries of a matrix. The nuclear norm penalty, which shrinks the singular values of a matrix, is often employed due to its computational convenience. However, it introduces bias in the estimation. To address this issue, nonconvex penalties such as SCAD are utilized, which provide sparse and unbiased estimators.

In this thesis, we study the nonconvex penalized matrix completion methods for estimating causal effects in panel data with time-dependent treatment adoption. We first derive an upper bound for the estimation error of our proposed estimator for the

potential control outcomes, which improves upon existing methods that rely on the nuclear norm penalty. Remarkably, this upper bound matches the one obtained by the oracle estimator, under an additional condition on the magnitudes of true singular values. Furthermore, we establish the asymptotic normality of the corresponding estimator for the treatment effect, which exhibits a smaller asymptotic variance compared to an existing method. We perform numerical studies to assess the recovery of the potential control outcomes and the estimation of the average treatment effect. Simulations validate our theoretical results, and experiments using real data further demonstrate the promising performance of our proposed method.

Keywords: Time-dependent treatment adoption, Potential control outcomes, SCAD, Unbiased estimator, Upper bound, Oracle estimator, Causal effect, Asymptotic normality

Student Number: 2013 – 22898

To Myeongjong and my family,

For their love, patience, and belief at every step of the way.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Causal inference has been studied in statistics aiming to uncover the true cause-and-effect relationships between variables rather than relying solely on correlations. It enables us to understand how changes in one variable impact another, which is valuable for policy evaluation [Abadie and Cattaneo, 2018]. Policy evaluation involves the implementation of policies (or treatments, interventions, events) by governments, organizations, or natural factors. Causal inference provides insights into the effectiveness of these policies, guiding future policy directions and resource allocation. Typically, aggregated data collected over time [Abadie et al., 2010] are analyzed. For example, one might examine annual cigarette sales data across multiple states to assess the impact of state government-initiated tobacco taxes. When studying causal effects in panel data, a common approach is to compare outcomes before and after policy implementation. The Synthetic Control [Abadie et al., 2010] method is widely used and considers patterns

across units to estimate treatment effects. Additionally, fixed effects models [Abadie, 2005; Arkhangelsky et al., 2021] and factor models [Bai and Ng, 2017; Xu, 2017] have been introduced to capture both time-series and cross-sectional patterns. Recently, the matrix completion (MC) methods using a nuclear norm penalty [Athey et al., 2021] has been suggested. These assume a low-rank structure in the true potential control matrix. Furthermore, Farias et al. [2021] proposed a de-biased estimator that directly estimates the treatment effect while accounting for low-rank, addressing the bias issue associated with the nuclear norm penalty.

The MC methods aim to fill in missing entries in a partially observed matrix using the available information. By assuming a low-rank structure, where the matrix can be represented as a product of two lower-dimensional matrices, we can model the matrix with significantly fewer parameters compared to the original form. This dimension reduction offers advantages in terms of storage and interpretation, and the separation of the underlying structure (pattern) from the noise. In the MC, the nuclear norm penalty [Cai et al., 2010; Mazumder et al., 2010; Negahban and Wainwright, 2012] is commonly used, which shrinks the singular values of the matrix. This penalty is computationally convenient for low-rank modeling. However, the nuclear norm penalty can introduce bias in the estimated values. To mitigate this issue, researchers have proposed the use of nonconvex penalties [Gui et al., 2016; Lu et al., 2014; Song et al., 2018].

The main goals of this thesis are: (1) to propose an estimator

for the potential control matrix having desirable properties, and (2) to establish the asymptotic normality of the corresponding estimator for the treatment effect.

In this thesis, we investigate the application of nonconvex penalties in the MC methods for estimating causal effects in panel data with time-dependent treatment adoption patterns. Our study offers several contributions. Firstly, we provide theoretical results concerning the recovery of the potential control outcomes. We demonstrate that our proposed estimator achieves faster convergence rates compared to the previous method that utilized the nuclear norm penalty [Athey et al., 2021]. We also show that the oracle estimator becomes a local minimum of the nonconvex problem, and our upper bound aligns with the upper bound for the oracle estimator under certain conditions. Secondly, we establish the asymptotic normality of the estimator for the causal effect. We additionally verify that this estimator exhibits a smaller asymptotic variance compared to the existing method. Furthermore, we validate our theoretical findings through simulations. In the analysis of real data, we compare the results obtained from our nonconvex penalized estimator and the estimator for the causal effect with those from other methods.

The thesis is organized as follows. Chapter 2 offers a comprehensive review of the low-rank MC, covering different penalties, theoretical studies, and algorithms. In Chapter 3, we explain the setup and notation and review existing MC approaches specifically for the causal panel analysis. We introduce our proposed estima-

tor and present our theoretical results. The chapter also includes details on simulation studies and data analysis using real-world datasets. Finally, we conclude the thesis in Chapter 4.

Chapter 2

Review: low-rank matrix completion

2.1 Introduction

Low-rank matrix modeling is commonly employed when the main information in a data matrix is primarily represented by a few dominant singular values, while the smaller singular values can be considered negligible without losing major information. Matrix completion (MC) is a promising technique that can recover an intact matrix with a low-rank structure from undersampled, incomplete, or corrupted data. The MC has found numerous applications in various domains [Li et al., 2019] including recommender systems, gene expression data [Kapur et al., 2016; Mongia et al., 2019] analysis, image processing [Aggarwal and Gupta, 2016; Balachandrasekaran et al., 2016; Gu et al., 2014], network analysis [Mahindre et al., 2019; Wang, 2017], and signal processing [Du

et al., 2013; Yang et al., 2014].

The nuclear norm, also known as the trace norm, is a convex surrogate for the rank of a matrix. It is equivalent to the ℓ-1 norm (also known as Lasso) of its singular values vector. The nuclear norm is widely used in MC applications, where the objective is to recover a low-rank matrix [Athey et al., 2021; Mongia et al., 2019; Zhou et al., 2014]. While the nuclear norm offers computational advantages and convexity [Cai et al., 2010; Mazumder et al., 2010], it is known to introduce estimation bias for parameters with large absolute values. To address this issue, nonconvex penalties have been introduced, such as the smoothly clipped absolute deviation (SCAD) penalty [Fan and Li, 2001] and the minimax concave (MCP) penalty [Zhang, 2010]. These nonconvex penalties have gained popularity due to their desirable statistical properties, including continuity, sparsity, and unbiasedness [Kim et al., 2008; Wang et al., 2014; Zhang and Zhang, 2012]. Lu et al. [2014, 2015a,b]; Mazumder et al. [2020]; Wen et al. [2018]; Yao and Kwok [2016] have empirically shown that the MC with nonconvex penalties can yield favorable results. Furthermore, Gui et al. [2016] derived tighter upper bounds for nonconvex penalties, and demonstrated that the unique global optimal solution is equivalent to the oracle estimator, despite the computational challenges associated with solving the nonconvex problem.

The remainder of this Chapter is structured as follows. In Section 2.2, we provide an overview of the setup and notation for the low-rank MC and discuss low-rank matrix approximation,

which serves as the foundation of the MC approach, in Section 2.3. Section 2.4 delves into the low-rank MC, including an examination of the nuclear norm and nonconvex penalties. In Section 2.5, we outline the necessary assumptions and review existing theoretical studies. Finally, in Section 2.6, we outline the algorithms used for solving the convex and nonconvex optimization problems.

2.2 Setup and notation

We first introduce the notations used throughout this thesis. We use lowercase letters for scalars, bold lowercase letters for vectors, and bold uppercase letters for matrices. Let $\mathbf{L}^* \in \mathbb{R}^{N\times T}$ be the true low-rank matrix. We assume that $\mathbf{L}^*$ is a matrix of rank r with a compact singular value decomposition (SVD) given by:

$$\mathbf{L}^* = \mathbf{U}^*\Xi^*\mathbf{V}^{*\top},$$

where $\mathbf{U}^* \in \mathbb{R}^{N\times r}, \mathbf{V}^* \in \mathbb{R}^{T\times r}$ and $\Xi^* = \text{diag}(\xi_1^*, \cdots, \xi_r^*) \in \mathbb{R}^{r\times r}$. Here, columns of $\mathbf{U}^*$ and $\mathbf{V}^*$ represent the left and the right singular vectors, respectively, and diagonal entries of Ξ^* are the singular values of $\mathbf{L}^*$. In this Chapter, let $\mathbf{Y} \in \mathbb{R}^{N\times T}$ be the matrix with completely observed entries. $\mathcal{O}$ is the set of index pairs (i,t) corresponding to the observed entries in $\mathbf{Y}$ and let $n = |\mathcal{O}|$. We define the projection operator for any matrix $\mathbf{A}$, given the set $\mathcal{O}$:

$$\mathcal{P}_{\mathcal{O}}(\mathbf{A})_{it} = \begin{cases} A_{it} & \text{if } (i,t) \in \mathcal{O}, \\ 0 & \text{if } (i,t) \notin \mathcal{O} \end{cases}$$

$\mathcal{P}^{\perp}_{\mathcal{O}}(\mathbf{Y})$ is the complementary projection that satisfies $\mathcal{P}^{\perp}_{\mathcal{O}}(\mathbf{Y}) + \mathcal{P}_{\mathcal{O}}(\mathbf{Y}) = \mathbf{Y}$. Note that $\sum_{(i,t)\in\mathcal{O}}(Y_{it} - L_{it})^2$ can be easily rewritten

as $\|\mathcal{P}_{\mathcal{O}}(\mathbf{Y}-\mathbf{L})\|_F^2$.

We also employ various matrix norms. These include the nuclear norm, denoted as $\|\mathbf{A}\|_{tr}$, which is the sum of the singular values of $\mathbf{A}$: $\|\mathbf{A}\|_{tr} = \text{trace}(\sqrt{\mathbf{A}^\top\mathbf{A}}) = \sum_{i=1}^{min(N,T)} \xi_i(\mathbf{A})$ where $\xi_i(\mathbf{A})$ represents i-th singular value of $\mathbf{A}$. Another norm is the operator (spectral) norm, denoted as $\|\mathbf{A}\|_{op}$, which is defined as the maximum singular value of $\mathbf{A}$: $\|\mathbf{A}\|_{op} = \xi_1(\mathbf{A})$. The Frobenius norm denoted as $\|\mathbf{A}\|_F$, is the square root of the sum of the squares of all the singular values of $\mathbf{A}$: $\|\mathbf{A}\|_F = \sqrt{\text{trace}(\mathbf{A}^\top\mathbf{A})} = \sqrt{\sum_{i=1}^{min(N,T)} \xi_i^2(\mathbf{A})}$. Lastly, we consider the max norm, denoted as $\|\mathbf{A}\|_{\max}$, which is the maximum absolute value of the entries of $\mathbf{A}$: $\|\mathbf{A}\|_{\max} = \max_{1\le j\le N, 1\le k\le T} |A_{jk}|$. It is worth noting that the trace norm, the Frobenius norm, and the operator norm correspond to ℓ_1, ℓ_2, and ℓ_∞ norms of singular values, respectively. Additionally, we define the trace inner product as $\langle \mathbf{A}, \mathbf{B}\rangle = \text{trace}(\mathbf{A}^T\mathbf{B})$. We employ the symbols $a \vee b = \max(a,b)$ and $a \wedge b = \min(a,b)$, representing the maximum and minimum values between a and b, respectively. We define a standard basis vector $e_i(a)$ as a vector of dimension a, where only the i-th element is 1 and all other elements are 0. The notation $\mathbf{1}_a$ represents an a-dimensional vector consisting of all ones. The function $I(\cdot)$ is used as an indicator function, and I_a represents an a-dimensional identity matrix. Finally, we use $[N]$ to denote the set $\{1, 2, \cdots, N\}$, representing a sequence of numbers from 1 to N.

5	1	2	2	?	3	5	?
3	?	?	4	5	?	2	1
1	5	3	?	?	3	?	2
?	4	2	?	1	2	3	?
?	1	?	2	5	?	1	5

5	1	2	2	3	3	5	2
3	4	3	4	5	4	2	1
1	5	3	5	2	3	1	2
2	4	2	2	1	2	3	1
3	1	4	2	5	3	1	5

Figure 2.1: Examples of **(Left)** an incomplete observed matrix and **(Right)** a result of MC.

2.3 Low-rank matrix approximation

Before delving into the MC problem, we review a low-rank approximation problem when all elements of a matrix are observed. Matrix approximation using SVD forms the basis for algorithms for low-rank MC problems.

Let $\mathbf{Y} = \mathbf{U}_T \Xi_T \mathbf{V}_T^T$ be the SVD of a fully observed $\mathbf{Y}$. Assuming $N \geq T$, $\mathbf{U}_T$ is an $N \times T$ dimensional orthogonal matrix ($\mathbf{U}_T^\top \mathbf{U}_T = I_T$), and $\mathbf{V}_T$ is an $T \times T$ dimensional orthogonal matrix ($\mathbf{V}_T^\top \mathbf{V}_T = I_T$). The columns of $\mathbf{U}_T$ and $\mathbf{V}_T$ are the left and right singular vectors, respectively. Ξ_T is a diagonal matrix of $T \times T$ dimensions and the diagonal elements $\xi_1, \xi_2, \cdots, \xi_T$, the singular values, of Ξ_T satisfy $\xi_1 \geq \xi_2 \geq \cdots \geq \xi_T \geq 0$. First, for $r \leq \text{rank}(\mathbf{Y})$, the optimization problem to find the approximate matrix closest to $\mathbf{Y}$ and with rank r is

$$\min_{\mathbf{L}} \|\mathbf{Y} - \mathbf{L}\|_F \quad \text{s.t.} \quad \text{rank}(\mathbf{L}) = r. \tag{2.1}$$

Its closed form solution is $\mathbf{U}_T \,\text{diag}(\xi_1, \cdots, \xi_r, 0, \cdots, 0)\mathbf{V}_T^\top$ [Hastie et al., 2015]. We denote this approximation matrix as $\mathcal{D}_r^H(\mathbf{Y})$.

This is called rank-r SVD (or reduced-rank SVD). The diagonal matrix, $\text{diag}(\xi_1, \cdots, \xi_r, 0, \cdots, 0)$, denotes a matrix in which the remainder is 0 except for the first r diagonal elements of $\mathbf{Y}$.

When the rank of $\mathbf{Y}$ is r, we denote the compact SVD of $\mathbf{Y}$ as $\mathbf{Y} = \mathbf{U}\Xi\mathbf{V}^T$ where columns of $\mathbf{U} \in \mathbb{R}^{N\times r}$ and $\mathbf{V} \in \mathbb{R}^{T\times r}$ represent the first r left and right singular vectors, respectively, and $\Xi = \text{diag}(\xi_1, \cdots, \xi_r)$ represents a diagonal matrix whose entries are non-zero singular values of $\mathbf{Y}$. Consider now the problem of adopting the nuclear norm, instead of restricting rank as in (2.1):

$$\min_{\mathbf{L}} \frac{1}{2} \|\mathbf{Y} - \mathbf{L}\|_F^2 + \lambda \|\mathbf{L}\|_{tr} \tag{2.2}$$

where $\|\mathbf{L}\|_{tr}$ is the nuclear norm for matrix $\mathbf{L}$, and see Section 2.4.2 for details. The closed-form solution of (2.2) for $\lambda \geq 0$ is $\mathcal{D}_\lambda^S(\mathbf{Y}) = \mathbf{U}\,\text{diag}\left(\max(\xi_1 - \lambda, 0), \cdots, \max(\xi_r - \lambda, 0)\right)\mathbf{V}^T$ [Cai et al., 2010]. We call it the singular value shrinkage operator. Note that this operator is equivalent to the proximal operator for the nuclear norm [Hastie et al., 2015, Exercise 5.8], which applies the soft-thresholding rule to the singular values of the matrix.

2.4 Low-rank matrix completion

2.4.1 Rank constraint

Now suppose that some elements of $\mathbf{Y}$ are missing. The MC problem requires additional assumptions about $\mathbf{Y}$. We assume the low rank of the matrix. First, the objective function with the rank

constraint can be written as

$$\min_{\mathbf{L}} \ \text{rank}(\mathbf{L}) \quad \text{s.t. } Y_{it} = L_{it} \text{ for all } (i,t) \in \mathcal{O}. \tag{2.3}$$

However, it is well known that the problem (2.3) is NP-hard. Moreover, estimating $\mathbf{L}$ accurately to the observations of $\mathbf{Y}$ causes overfitting. Allowing some errors of $\mathbf{L}$ on $\mathbf{Y}$ have empirically had better performance results:

$$\min_{\mathbf{L}} \ \text{rank}(\mathbf{L}) \quad \text{s.t. } \|\mathcal{P}_{\mathcal{O}}(\mathbf{Y} - \mathbf{L})\|_F^2 \leq \delta. \tag{2.4}$$

We note that (2.4) and (2.5) below

$$\min_{\text{rank}(\mathbf{L}) \leq r} \|\mathcal{P}_{\mathcal{O}}(\mathbf{Y} - \mathbf{L})\|_F^2 \tag{2.5}$$

are equivalent. This is because the solution family for a certain $\delta \geq 0$ in (2.4) is the same as that for a certain r in (2.5). However, finding an exact solution (global minimum) of equations (2.4) and (2.5) is not guaranteed, although finding the local minima is possible since they are nonconvex problems. The convex relaxation of rank constraint for getting a global solution is to adopt the nuclear norm penalty.

2.4.2 Nuclear norm penalty

The nuclear norm penalty is defined by $\|\mathbf{L}\|_{tr} = tr(\sqrt{\mathbf{L}^\top \mathbf{L}})$. Let $\xi_i(\mathbf{L})$ (or ξ_i for convenience) be the ith largest singular value of $\mathbf{L}$, and $\boldsymbol{\xi}$ be the vector $(\xi_1, \ldots, \xi_r)^\top$. It can be expressed in the form below [Lee et al., 2010]:

$$\|\mathbf{L}\|_{tr} = \inf\{\|\boldsymbol{\xi}\|_1 : \mathbf{L} = \sum_j \xi_j \mathbf{u}_j \mathbf{v}_j^{'} \text{ where } \|\mathbf{u}_j\|_2 = 1 \text{ and } \|\mathbf{v}_j\|_2 = 1\} \tag{2.6}$$

where $\mathbf{u}_j$ and $\mathbf{v}_j$ are jth left and right singular vectors of $\mathbf{L}$, respectively. The nuclear norm is also known as the trace norm, the Schatten 1-norm, or the Ky-Fan norm. Since $\text{rank}(\mathbf{L}) = \|\boldsymbol{\xi}\|_0 = |supp(\boldsymbol{\xi})|$ holds, the nuclear norm is used as a surrogate for the rank constraint and it promotes the sparsity of singular values. Furthermore, the unit-ball of the nuclear norm is the convex hull of a set of matrices that are unit-norm and rank-one [Srebro and Shraibman, 2005]:

$$\begin{aligned}\{\mathbf{L} \in \mathbb{R}^{N\times T} :&\|\mathbf{L}\|_{tr} \leq 1\} \\ &=\text{conv}\{\mathbf{u}\mathbf{v}^T : \mathbf{u} \in \mathbb{R}^N, \mathbf{v} \in \mathbb{R}^T, \|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = 1\}.\end{aligned} \tag{2.7}$$

Therefore, the convex relaxation of equation (2.3) is:

$$\min_{\mathbf{L}} \|\mathbf{L}\|_{tr} \quad \text{s.t. } Y_{it} = L_{it} \text{ for all } (i,t) \in \mathcal{O}. \tag{2.8}$$

The figure in Hastie et al. [2015, Figure 7.3] shows the level set of the nuclear norm unit-ball of a 2×2 symmetric matrix of the convex problem (2.8). The equation (2.9) below allows noise in (2.8):

$$\min_{\mathbf{L}} \frac{1}{2} \|\mathcal{P}_{\mathcal{O}}(\mathbf{Y} - \mathbf{L})\|_F^2 + \lambda\|\mathbf{L}\|_{tr}. \tag{2.9}$$

The objective function of (2.9) is a convex function, so the local solution becomes a global solution. Since the nuclear norm is non-differentiable, solving the problem using convex optimization algorithms is common.

2.4.3 Nonconvex penalty

Even though solving (2.9) offers computational advantages, it yields a biased estimator. The rank of a matrix, which represents the number of nonzero singular values, can be effectively reduced by setting only a few singular values to zero. However, the nuclear norm penalty applies a constant shrinkage to all singular values regardless of their magnitudes. Nonconvex penalty methods have been introduced as alternatives to address this issue. These methods have demonstrated improved performance compared to convex penalties in various applications [Lu et al., 2014, 2015b; Song et al., 2018].

The nonconvex penalty is typically formulated in the following manner:

$$G_\lambda(\mathbf{L}) = \sum_{i=1}^{min(N,T)} g_\lambda(\xi_i(\mathbf{L})), \tag{2.10}$$

where $g_\lambda(\cdot)$ is a nonconvex penalty such as SCAD and MCP. Then the problem (2.9) becomes

$$\min_{\mathbf{L}} \frac{1}{2} \|\mathcal{P}_\mathcal{O}(\mathbf{Y} - \mathbf{L})\|_F^2 + G_\lambda(\mathbf{L}) := F_\lambda(\mathbf{L}) \tag{2.11}$$

where we denote the empirical loss function $\frac{1}{n} \|\mathcal{P}_\mathcal{O}(\mathbf{Y} - \mathbf{L})\|_F^2$ as $f_n(\mathbf{L})$. It is important to note that the function in (2.11) is nonconvex, which makes finding the global minimum challenging. Furthermore, there exist multiple local minima.

The smoothly clipped absolute deviation (SCAD) penalty [Fan and Li, 2001] and the minimax concave (MCP) penalty [Zhang, 2010] have gained wide acceptance due to their desirable statistical

properties, such as continuity, sparsity, and unbiasedness [Kim et al., 2008; Wang et al., 2014; Zhang and Zhang, 2012]. The penalty function of SCAD [Fan and Li, 2001] is defined as follows:

$$\begin{aligned} g_\lambda(|x|) =& \lambda |x| I(|x| < \lambda) \\ &+ \left(\frac{-|x|^2 + 2\gamma\lambda|x| - \lambda^2}{2(\gamma-1)} \right) I(\lambda \le |x| \le \gamma\lambda) \\ &+ \frac{\lambda^2(\gamma+1)}{2} I(|x| > \gamma\lambda) \end{aligned}$$

where $\gamma > 2$ is a constant. The penalty function of MCP [Zhang, 2010] is defined as follows:

$$g_\lambda(|x|) = \left(\lambda|x| - \frac{|x|^2}{2\gamma} \right) I(|x| \le \gamma\lambda) + \frac{\gamma\lambda^2}{2} I(|x| > \gamma\lambda)$$

where $\gamma > 0$ is a constant. Figure 2.2 shows the graph of some penalty functions. We see that the LASSO is convex and the MCP and the SCAD are nonconvex.

In addition to SCAD and MCP penalties, other penalty methods have been introduced and shown to be effective in various applications. These include the Truncated nuclear norm penalty [Hu et al., 2012], the Weighted nuclear norm penalties [Gu et al., 2014, 2017], and the Schatten capped p-norm penalty [Li et al., 2020].

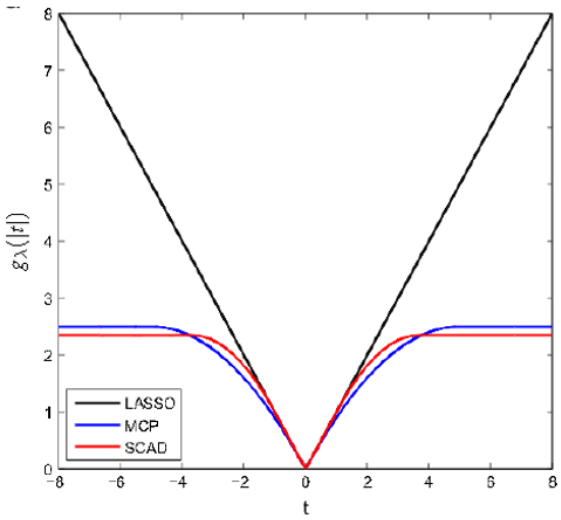

Figure 2.2: The graph of some penalty functions: lasso, MCP, and SCAD.

2.5 Theoretical studies for low-rank matrix completion

2.5.1 Assumption 1: Coherence

The canonical assumption for the MC is referred to as "coherence" [Candès and Recht, 2009; Hastie et al., 2015] and is particularly applicable in the "Exact" case as described in (2.3) and (2.8).

Assumption 2.5.1 (Coherence)**.** $\mathbf{L}^*$ *is* μ*-incoherent:*

$$\frac{N}{r}\max_{1\le i\le N}\|\mathbf{U}^{*\top}e_i(N)\|_2^2\le\mu \quad \textit{and} \quad \frac{T}{r}\max_{1\le i\le T}\|\mathbf{V}^{*\top}e_i(T)\|_2^2\le\mu$$

where $\|\cdot\|_2$ *denotes* l_2*-norm of a vector.*

This refers to the extent to which any singular vector of $\mathbf{L}^*$

matches a standard basis vector. The coherence parameter μ is expected to be small, with $1 \leq \mu \leq (N \vee T)/r$. Let us illustrate the necessity of the incoherence constraint with two examples. Consider the following matrices:

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \tag{2.12}$$

Both matrices have dimensions $N = T = 4$ and rank $r = 1$. However, the coherence of these matrices differs significantly.

In the case of matrix $\mathbf{A}$, the coherence is maximal, with a value of $\mu = N/r * 1 = 4$. The singular vector corresponding to the non-zero singular value is $(1, 0, 0, 0)^T$. In this scenario, it is more likely to observe zero entries rather than non-zero entries. Consequently, the matrix becomes challenging to recover. In contrast, matrix $\mathbf{B}$ has minimal coherence, with a value of $\mu = N/r * 0.25 = 1$. The singular vector corresponding to the non-zero singular value is $(0.5, 0.5, 0.5, 0.5)^T$. In this case, the missing entries can be successfully recovered because the matrix has a uniform pattern with all its elements having the same value. By limiting the coherence, we ensure that the missing entries can be accurately estimated, leading to a more reliable recovery of the original matrix.

2.5.2 Assumption 2: Spikiness

In the "noise" setting (2.9), the incoherence condition is not robust, even for small perturbations [Hastie et al., 2015; Negahban

and Wainwright, 2012]. Instead, a less stringent assumption called "spikiness" is often considered.

Definition 1 (Spikiness ratio)**.** *For any nonzero matrix* $\mathbf{L}$*, the spikiness ratio of the matrix is defined as*

$$\alpha_{sp}(\mathbf{L}) = \sqrt{NT}\frac{\|\mathbf{L}\|_{max}}{\|\mathbf{L}\|_F}.$$

The spikiness ratio measures the uniformity in the spread of matrix elements. It satisfies $1 \leq \alpha_{sp}(\mathbf{L}) \leq \sqrt{NT}$, and smaller values indicate better performance. For example, the spikiness ratios of the left and right matrices in (2.12) are equal to $\sqrt{NT}$ and 1, respectively.

Let's consider an example [Hastie et al., 2015; Negahban and Wainwright, 2012] to illustrate that the coherence condition can be violated even with small perturbations. We start with a matrix $\mathbf{D}$ of dimensions $N = T = 4$, rank 1, and Frobenius norm one given as:

$$\mathbf{D} = \begin{pmatrix} 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{pmatrix} \tag{2.13}$$

The coherence of $\mathbf{D}$, denoted as $\mu(\mathbf{D})$, is equal to 1, which represents minimal coherence, and the spikiness ratio $\alpha_{sp}(\mathbf{D})$ is equal to $\sqrt{4 * 4} * 0.25/1 = 1$, indicating minimal spikiness.

We now consider the matrix $\mathbf{L}^* = \mathbf{D} + \delta * \mathbf{A}$ for some $\delta > 0$, where $\mathbf{A}$ is a matrix given in (2.12). $\mathbf{L}^*$ can have a high coherence even for sufficiently small values of δ. For example, when $\delta = 1/16$,

the coherence $\mu(\mathbf{L}^*)$ is approximately 1.4509, and when $\delta = 1/160$, the coherence $\mu(\mathbf{L}^*)$ is approximately 1.4953. On the other hand, the spikiness ratio $\alpha_{sp}(\mathbf{L}^*)$ decreases when δ is sufficiently small. For instance, when $\delta = 1/16$, $\alpha_{sp}(\mathbf{L}^*)$ is approximately 1.2286, and when $\delta = 1/160$, $\alpha_{sp}(\mathbf{L}^*)$ is approximately 1.0234. It is worth noting that the maximum coherence for $\mathbf{L}^*$ is equal to $4/2 = 2$, while the maximum spikiness ratio is equal to 4. This example highlights that the spikiness ratio is a useful measure for assessing the models with noise.

Recent works [Athey et al., 2021; Davenport et al., 2014; Gui et al., 2016; Klopp, 2014] assume spikiness condition, without loss of generality, to demonstrate theoretical results in the presence of noise.

Assumption 2.5.2 (spikiness condition)**.** *There exists a known* $\alpha^* > 0$*, such that*

$$\|\mathbf{L}^*\|_{max} = \frac{\alpha_{sp}(\mathbf{L}^*)\|\mathbf{L}^*\|_F}{\sqrt{NT}} \leq \alpha^*.$$

As we see in Section 2.5.3, the upper bound is expressed using the spikiness constants α^* or $\alpha_{sp}(\mathbf{L}^*)$.

2.5.3 Review of existing studies

In this section, we will review the theoretical results of existing studies for the MC problems with noise. Gunasekar et al. [2014]; Koltchinskii et al. [2011]; Negahban and Wainwright [2012] obtained minimax-optimal upper bounds (up to logarithmic factor)

for the problem (2.9) when assuming the spikiness conditions (Assumption 2.5.2):

$$\frac{\|\widehat{\mathbf{L}} - \mathbf{L}^*\|_F}{\sqrt{NT}} = \mathcal{O}\left((\alpha^* \vee \sigma)\sqrt{\frac{rM\log(M)}{n}}\right) \tag{2.14}$$

where $M = (N \vee T)$. It is worth noting that this upper bound can also be expressed in terms of $\alpha_{sp}(\mathbf{L}^*)$ [Hastie et al., 2015] using the following expression:

$$\frac{\|\widehat{\mathbf{L}} - \mathbf{L}^*\|_F}{\|\mathbf{L}^*\|_F} = \mathcal{O}\left((\alpha_{sp}(\mathbf{L}^*) \vee \sigma')\sqrt{\frac{rM\log(M)}{n}}\right)$$

where σ' is properly scaled variance proxy of noise.

Chen et al. [2020] derived upper bounds for the MC with the nuclear norm penalty in the max norm, spectral norm, and Frobenius norm. Among these, the upper bound obtained in the Frobenius norm is near-optimal. However, their analysis assumes the incoherence condition.

Gui et al. [2016] derived a upper bound for the nonconvex penalized estimator (2.11). They proceed by defining two sets that correspond to the non-zero singular values of the matrix $\mathbf{L}^*$:

$$\begin{aligned} S_1 &= \left\{\xi_1^*, \cdots, \xi_{r_1}^* | \xi_1^* \geq \cdots \geq \xi_{r_1}^* \geq \nu\right\} \text{ and} \\ S_2 &= \left\{\xi_{r_1+1}^*, \cdots, \xi_r^* | \nu > \xi_{r_1+1}^* \geq \cdots \geq \xi_r^* > 0\right\}. \end{aligned} \tag{2.15}$$

The first set, denoted as S_1, consists of the singular values greater than or equal to the constant ν. On the other hand, the second set denoted as S_2, includes the relatively small singular values. Let $r_1 := |S_1|$ and $r_2 := |S_2|$. r_1 represents the cardinality of set S_1, and r_2 represents the cardinality of set S_2. It is important to note that $\nu = \gamma\lambda$ for both the SCAD and MCP penalties.

They established the upper bound under the spikiness condition, which is expressed as:

$$\frac{\|\widehat{\mathbf{L}} - \mathbf{L}^*\|_F}{\sqrt{NT}} \leq (\alpha^* \vee \sigma) \left(C_1 r_1 \sqrt{\frac{\log M}{n}} + C_2 \sqrt{\frac{r_2 M \log M}{n}} \right)$$

where C_1 and C_2 are constants. This upper bound is tighter than (2.14) obtained using the convex penalty when some of the true singular values have large magnitudes. Additionally, they demonstrated that the unique global optimal solution is identical to the oracle estimator under certain conditions. However, finding this unique global optimal solution is challenging due to the nonconvexity of (2.11). It is worth noting that the traditional MC literature mentioned in this Section assumes that missing entries are sampled completely at random.

2.6 Algorithms for low-rank matrix completion

This section describes algorithms: *SOFT-IMPUTE* for the convex problem (2.9) and *PGH* for the nonconvex problem (2.11).

2.6.1 *SOFT-IMPUTE* algorithm

The SOFT-IMPUTE algorithm [Mazumder et al., 2010] performs a singular value shrinkage operator and replaces missing values at each step. The following is performed until the stopping criterion is reached for $k = 0, 1, 2, \ldots$, with an initial value $\mathbf{L}^0$:

$$\mathbf{L}^{k+1} \leftarrow \mathcal{D}_\lambda^S(\mathcal{P}_\mathcal{O}(\mathbf{Y}) + \mathcal{P}_\mathcal{O}^\perp(\mathbf{L}^k)). \tag{2.16}$$

The initial value $\mathbf{L}^0$ can be set to $\mathbf{Y}$ in which the missing values are replaced with 0. When we tune the regularization parameter, we can use a matrix trained with a close parameter as a warm starter. Note that we do not need to set the step size unlike the SVT algorithm [Cai et al., 2010].

Mazumder et al. [2010, Section 4.1] proved that $\mathbf{L}^k$ generated by the algorithm asymptotically converges to the global solution $\mathbf{L}^*$ of (2.9) when $k \to \infty$. It also shows that the non-asymptotic (worst) rate of convergence is of the order of $\mathcal{O}(1/k)$. The convergence rate in a specific setting is shown in Agarwal et al. [2010]. Hastie et al. [2015, Section 7.3.2] is also referred. Mazumder et al. [2010, Section 5] explains that the computational complexity order is the same as that of SVT algorithm [Cai et al., 2010].

2.6.2 *PGH* algorithm

We introduce one of the algorithms that find local minima of the problem (2.11): the Proximal gradient homotopy (PGH) Algorithm [Gui et al., 2016; Wang et al., 2014; Xiao and Zhang, 2013]. The algorithm is a variant of the Proximal gradient method [Nesterov, 2013].

Suppose that the penalty function can be decomposed into two parts as follows:

$$g_\lambda(|\xi_i(\mathbf{L})|) = \bar{g}_\lambda(|\xi_i(\mathbf{L})|) + \lambda|\xi_i(\mathbf{L})| \tag{2.17}$$

where $\bar{g}_\lambda(\cdot)$ is the differential concave function and $\lambda|\cdot|$ is the convex function. Note that $g_\lambda(|\xi_i(\mathbf{L})|) = g_\lambda(\xi_i(\mathbf{L}))$ since singular

values of a matrix are nonnegative. The concave part $\bar{g}_\lambda(\cdot)$ for the SCAD penalty function is given as

$$\begin{aligned}\bar{g}_\lambda(|x|) = &-\frac{x^2 - 2\lambda|x| + \lambda^2}{2(\gamma-1)} I(\lambda \le |x| \le \gamma\lambda) \\ &+ \left[\frac{(\gamma+1)\lambda^2}{2} - \lambda|x|\right] I(|x| > \gamma\lambda),\end{aligned}$$

and for the MCP penalty function, the concave part $\bar{g}_\lambda(\cdot)$ is given as

$$\bar{g}_\lambda(|x|) = -\frac{x^2}{2\gamma} I(|x| \le \gamma\lambda) + \left[\frac{\gamma\lambda^2}{2} - \lambda|x|\right] I(|x| > \gamma\lambda).$$

By denoting $\bar{G}_\lambda(\mathbf{L}) := \sum_{i=1}^{min(N,T)} \bar{g}_\lambda(|\xi_i(\mathbf{L})|) = \sum_{i=1}^{min(N,T)} g_\lambda(\xi_i(\mathbf{L})) - \lambda\|\mathbf{L}\|_{tr}$ and $\bar{f}_{n,\lambda}(\mathbf{L}) := f_n(\mathbf{L}) + \bar{G}_\lambda(\mathbf{L})$, we can rewrite the objective function (2.11):

$$F_\lambda(\mathbf{L}) = f_n(\mathbf{L}) + G_\lambda(\mathbf{L}) = \bar{f}_{n,\lambda}(\mathbf{L}) + \lambda\|\mathbf{L}\|_{tr}. \tag{2.18}$$

By taking a sufficiently large value of the initial regularization parameter $\lambda = \lambda_0$, and then gradually decreasing it until it reaches the target value, the algorithm (Algorithm 1) tries to find a solution that minimizes the nonconvex objective function (2.18) for each fixed λ.

When $\mathbf{L}_t^{k-1}$ is the update value of the *k-1*-th iteration in the t-th path, the following is the local quadratic approximation of F_{λ_t} with respect to λ_t:

$$\begin{aligned}\widetilde{F}_{l_t^{k-1},\lambda_t}(\mathbf{L}; \mathbf{L}_t^{k-1}) = &\bar{f}_{n,\lambda}(\mathbf{L}_t^{k-1}) + \langle \nabla \bar{f}_{n,\lambda}(\mathbf{L}_t^{k-1}), \mathbf{L} - \mathbf{L}_t^{k-1} \rangle \\ &+ \frac{l_t^{k-1}}{2}\|\mathbf{L} - \mathbf{L}_t^{k-1}\|_F^2 + \lambda_t\|\mathbf{L}\|_{tr}\end{aligned} \tag{2.19}$$

Algorithm 1 $\{\mathbf{L}^t\}_{t=1}^{K+1} \leftarrow \mathrm{PGH}\,(\lambda_0, \lambda_{tgt}, \epsilon_{opt}, l_{min})$

input $\lambda_0 > 0, \lambda_{tgt} > 0, \epsilon_{opt} > 0, l_{min} > 0$

1: **parameters** $\eta \in [0.9, 1)$

2: **initialize** $\mathbf{L}^0 \leftarrow 0, l_0 \leftarrow l_{min}, K \leftarrow \lfloor \frac{\ln(\lambda_0/\lambda_{tgt})}{\ln(1/\eta)} \rfloor$

3: **for** $t = 0, 1, 2, \cdots, K-1$ **do**

4: $\quad \lambda_{t+1} \leftarrow \eta\lambda_t$

5: $\quad \epsilon_{t+1} \leftarrow \lambda_{t+1}/4$

6: $\quad \{\mathbf{L}_{t+1}, l_{t+1}\} \leftarrow \mathrm{ProxGrad}\,(\lambda_{t+1}, \epsilon_{t+1}, \mathbf{L}_t, l_t)$

7: **end for**

8: $\{\mathbf{L}_{K+1}, l_{K+1}\} \leftarrow \mathrm{ProxGrad}\,(\lambda_{tgt}, \epsilon_{opt}, \mathbf{L}_K, l_K)$

9: **return** $\{\mathbf{L}_t\}_{t=1}^{K+1}$

where l_t^{k-1} is a Lipschitz constant of the *k-1*-th iteration in the t-th path. $\mathbf{L}_t^k$ is updated to minimize (2.19):

$$\mathbf{L}_t^k = \underset{\mathbf{L}\in\mathbb{R}^{N\times T}}{\operatorname{argmin}} \widetilde{F}_{l_t^{k-1},\lambda_t}(\mathbf{L}; \mathbf{L}_t^{k-1}). \tag{2.20}$$

We use the singular value shrinkage operator to solve this problem. Note that the low-rank matrix and Lipschitz constant (quadratic coefficient) estimated at the *t-1*-th path are used as the initial values of $\mathbf{L}_t$ and l_t at the t-th path (Line 6 and 8 in Algorithm 1). $\mathbf{L}_t^k$ converges toward the exact local solution of the optimization problem that minimizes (2.18) [Wang et al., 2014].

Xiao and Zhang [2013] presented the stopping criterion of the ProxGrad algorithm in the ℓ_1-regularized least-squares (ℓ_1-LS), which can be extended to the MC problem [Gui et al., 2016]. Assume that $\widehat{\mathbf{L}}$ is the optimal solution to the problem (2.11). Based

on the optimality condition, there exists $\boldsymbol{\Upsilon} \in \partial\|\widehat{\mathbf{L}}\|_{tr}$ such that, for all $\mathbf{L} \in \mathbb{R}^{N\times T}$, the inequality

$$\langle \widehat{\mathbf{L}} - \mathbf{L}, \nabla \bar{f}_{n,\lambda}(\widehat{\mathbf{L}}) + \lambda\boldsymbol{\Upsilon}\rangle \leq 0 \tag{2.21}$$

holds. Hence, we measure the suboptimality of $\mathbf{L}$ using

$$\begin{aligned}\omega_\lambda(\widehat{\mathbf{L}}) &= \min_{\Upsilon\in\partial\|\widehat{\mathbf{L}}\|_{tr}} \max_{\mathbf{L}} \left\{ \frac{\langle \widehat{\mathbf{L}} - \mathbf{L}, \nabla \bar{f}_{n,\lambda}(\widehat{\mathbf{L}}) + \lambda\boldsymbol{\Upsilon}\rangle}{\|\widehat{\mathbf{L}} - \mathbf{L}\|_{tr}} \right\} \\ &= \min_{\Upsilon\in\partial\|\widehat{\mathbf{L}}\|_{tr}} \left\{ \left\| \nabla \bar{f}_{n,\lambda}(\widehat{\mathbf{L}}) + \lambda\boldsymbol{\Upsilon} \right\|_{op} \right\}.\end{aligned} \tag{2.22}$$

The second equality is a consequence of the duality between the nuclear norm and the spectral norm. The optimality condition (2.21) guarantees that if $\widehat{\mathbf{L}}$ is an exact optimum, we have $\omega_\lambda(\widehat{\mathbf{L}}) < 0$. If $\widehat{\mathbf{L}}$ is close to the optimum, $\omega_\lambda(\widehat{\mathbf{L}})$ is likely to be a small positive value.

Algorithm 2 is a process of performing the proximal gradient method to obtain the t-th approximate local solution of (2.20). We set the initial value of l_t^k on Line 3 slightly smaller than l_t^{k-1} because we expect that $\mathbf{L}_t^k$ and $\mathbf{L}_t^{k-1}$ become closer as it is repeated. $\widehat{\epsilon}$ is the desired optimization precision.

Algorithm 3 determines l_t^k and calculates $\mathbf{L}_t^k$ with the chosen value. In Line 3-6, we increase the Lipschitz constant until $\widetilde{F}_{l,\lambda}(\mathbf{L}_t^k; \mathbf{L}_t^{k-1})$ becomes the tight upper bound of the objective function $F_\lambda(\mathbf{L}_t^k)$.

Algorithm 2 $\{\widetilde{\mathbf{L}}, \widehat{l}\} \leftarrow \text{ProxGrad}\left(\lambda, \widehat{\epsilon}, \mathbf{L}_t^0, l_t^0\right)$

input $\lambda > 0, \widehat{\epsilon} > 0, \mathbf{L}_t^0 \in \mathbb{R}^{N \times T}, l_t^0 > 0, k = 0$

repeat
  $k \leftarrow k + 1$
  $l_{init} \leftarrow \max\left\{l_{\min}, l_t^{k-1}/2\right\}$
  $\left\{\mathbf{L}_t^k, l_t^k\right\} \leftarrow \text{LineSearch}\left(\lambda, \mathbf{L}_t^{k-1}, l_{init}\right)$
until $\omega_\lambda\left(\mathbf{L}_t^k\right) \leq \widehat{\epsilon}$
$\widetilde{\mathbf{L}} \leftarrow L_t^k, \widehat{l} \leftarrow l_t^k$
return $\{\widetilde{\mathbf{L}}, \widehat{l}\}$

Algorithm 3 $\{\mathbf{L}_t^k, N\} \leftarrow \text{LineSearch}\left(\lambda, \mathbf{L}_t^{k-1}, l\right)$

input $\lambda > 0, \mathbf{L}_t^{k-1} \in \mathbb{R}^{N \times T}, l > 0$

repeat
  $\mathbf{L} \leftarrow \underset{\mathbf{L} \in \mathbb{R}^{N \times T}}{\operatorname{argmin}} \widetilde{F}_{l,\lambda}(\mathbf{L}; \mathbf{L}_t^{k-1})$
  if $F_\lambda(\mathbf{L}) > \widetilde{F}_{l,\lambda}(\mathbf{L}; \mathbf{L}_t^{k-1})$ **then**
    $l \leftarrow 2l$
  end if
until $F_\lambda(\mathbf{L}) \leq \widetilde{F}_{l,\lambda}(\mathbf{L}; \mathbf{L}_t^{k-1})$
$N \leftarrow l$
return $\{\mathbf{L}, N\}$

Chapter 3

Nonconvex penalized matrix completion for causal inference in panel data

3.1 Introduction

In program evaluation, there has been extensive research aimed at estimating average treatment causal effects. The data commonly takes the form of aggregated panel data, where a binary treatment is applied. Comparative case studies involve comparing outcomes before and after interventions. However, it is challenging because either control outcomes (before intervention) or treated outcomes (after intervention) are observed for each data point. That means

that potential control outcomes (counterfactuals) for units and times exposed to the treatment are missing. The typical approach to tackle this task is to first impute the counterfactuals and then estimate the average causal effect by comparing them with the corresponding observed treated outcomes.

Traditionally, regression models have been employed to estimate counterfactuals. Synthetic control literature [Abadie and Cattaneo, 2018; Abadie et al., 2010] deals with data with a single treated unit and assumes that patterns observed across different units remain stable over time. To capture more complex patterns, researchers have explored two-way fixed effects models [Abadie, 2005; Arkhangelsky et al., 2021; Doudchenko and Imbens, 2016], which consider both cross-sectional and time-series patterns, as well as factor models [Bai and Ng, 2017; Xu, 2017], which incorporate interactions between these patterns. Athey et al. [2021] recently proposed the nuclear norm matrix completion (MC) methods. They provided an upper bound on the estimation error for the nuclear-norm estimator. They established this under the condition that the true matrix is not excessively spiky, considering both stochastic and time-dependent treatment adoption patterns. Furthermore, they conducted empirical investigations that revealed the limitations of the synthetic control approach and demonstrated the superior performance of MC methods in scenarios where the number of units is larger than the number of time periods, or where the treatment adoption date is substantially earlier compared to the number of time periods.

Meanwhile, Farias et al. [2021] proposed a de-biased estimator for the average treatment effect. Their focus was on reducing the bias in the causal effect estimator that arises from using the nuclear norm penalty when they formulate the problem as a matrix approximation problem. They demonstrated the asymptotic normality of their estimator. Unlike Athey et al. [2021], they assumed a deterministic treatment adoption process and required the pattern of treatment adoption to be dissimilar to that of the true counterfactual matrix. Additionally, they imposed an incoherence condition on the true low-rank matrix, which is not robust to observations with noise.

The rest of this Chapter is organized as follows. In Section 3.2, we introduce the setup and notation. We then review existing MC methods used for causal inference in panel data in Section 3.3. Our proposed estimator is presented in Section 3.4, and the theoretical results are discussed in Section 3.5. We finally present the numerical studies, which include simulations and the analysis of semi-synthetic data, in Section 3.6.

3.2 Setup and notation

We deal with panel data with dimensions $N \times T$, where N represents the number of units and T represents the number of time periods. If a unit i is exposed to a treatment at time t, we denote it as $W_{it} = 1$; otherwise, $W_{it} = 0$. The observed outcome for the (i, t)-th element of the data matrix is denoted as

$Y_{it} = W_{it}Y_{it}(1) + (1 - W_{it})Y_{it}(0)$. $Y_{it}(0)$ represents the outcome when unit i is not exposed to treatment at time t, while $Y_{it}(1)$ represents the outcome when it is exposed to treatment. In causal panel data, we only observe either $Y_{it}(0)$ or $Y_{it}(1)$. If $Y_{it}(1)$ is observed, $Y_{it}(0)$ is missing. Our main question of interest is "What would have happened if the treated units did not adopt the treatment?". To estimate their potential control outcomes, we utilize the MC methods. To assess the impact of the treatment, we compare the observed treated outcomes with the estimated potential control outcomes for the same treated units and times.

When it comes to the adoption of treatment, we consider two commonly used time-dependent structures in the economics literature: a block structure (simultaneous adoption) and a staggered structure (staggered adoption) [Athey et al., 2021]. In the block structure, certain units adopt the treatment simultaneously at a specific time. A special case of the block structure is the single-treated-unit block structure [Abadie, 2021; Abadie et al., 2010], where the treatment is applied to a single unit. In the staggered structure, different units adopt the treatment at different points in time [Athey et al., 2021; Shaikh and Toulis, 2021]. For both structures, once treatment is adopted, the units remain treated. The matrices below illustrate examples of the treatment structures: (a) block structure and (b) staggered structure. A checkmark (✓) indicates the control outcome, while a question mark (?) represents the missing data (the treated outcome).

$$(a)\begin{pmatrix} \checkmark & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark \\ \checkmark & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark \\ \checkmark & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark \\ \checkmark & \checkmark & \checkmark & ? & ? & ? \\ \checkmark & \checkmark & \checkmark & ? & ? & ? \end{pmatrix} \qquad (b)\begin{pmatrix} \checkmark & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark \\ \checkmark & \checkmark & ? & ? & ? & ? \\ \checkmark & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark \\ \checkmark & \checkmark & \checkmark & \checkmark & ? & ? \\ \checkmark & \checkmark & \checkmark & ? & ? & ? \end{pmatrix}$$

For example, let's consider the Cigarette sales data [Abadie et al., 2010]. This dataset includes observations on cigarette sales per capita for 39 states, including California, in the U.S. from 1970 to 2000. California passed a tobacco control law known as "Proposition 99" in 1988, which took effect in 1989. Note that the data exhibits the single-treated-unit block structure. We are interested in estimating the effect of this law on cigarette sales. Using the MC methods with data from other states and California until 1988, we can impute the missing control outcomes (cigarette sales assuming the law had not been passed) for California from 1989 to 2000. By comparing the observed cigarette sales under the law with the imputed values, we can assess the impact of the tobacco control law.

3.3 Review of existing matrix completion methods

3.3.1 Nuclear norm penalized estimator for the potential control outcomes

Athey et al. [2021] developed matrix completion methods to in-

corporate both cross-sectional and time-series patterns to recover the potential control (counterfactual) matrix. They modeled the potential control matrix as follows:

$$\mathbf{Y}(\mathbf{0}) = \mathbf{L}^* + \boldsymbol{\varepsilon} \tag{3.1}$$

where $\mathbf{Y}(\mathbf{0}) \in \mathbb{R}^{N \times T}$ represents the potential control matrix, $\mathbf{L}^* \in \mathbb{R}^{N \times T}$ denotes the true low-rank matrix, and elements of $\boldsymbol{\varepsilon}$ are σ-sub-Gaussian random variables with mean 0.

The potential control outcomes are estimated using MC methods enabling the estimation of the average treatment effect. The estimation process consists of two steps. In the first step, the counterfactual matrix is estimated as follows:

$$\widehat{\mathbf{L}} = \underset{\mathbf{L} \in \mathbb{R}^{N \times T}}{\operatorname{argmin}} \left\{ \frac{1}{n} \| \mathcal{P}_{\mathcal{O}} \left(\mathbf{Y}(\mathbf{0}) - \mathbf{L} \right) \|_F^2 + \lambda \| \mathbf{L} \|_{tr} \right\} \tag{3.2}$$

where $\| \cdot \|_{tr}$ represents the nuclear norm penalty. The average treatment effect for the treated (ATT) is then calculated:

$$\widehat{\theta} = \sum_{(i,t):W_{it}=1} \left[Y_{it}(1) - \widehat{Y}_{it}(0) \right] / \sum_{(i,t)} W_{it}. \tag{3.3}$$

Recall that $\mathcal{O}$ is the set of index pairs corresponding to the observed entries in $\mathbf{Y}(\mathbf{0})$ and $n = |\mathcal{O}|$. The use of the nuclear norm penalty offers computational advantages since the objective function (3.2) is a convex problem. When incorporating two-way fixed effects, the objective function (3.2) is modified as follows:

$$\underset{\mathbf{L} \in \mathbb{R}^{N \times T}, \boldsymbol{\eta} \in \mathbb{R}^N, \boldsymbol{\beta} \in \mathbb{R}^T}{\operatorname{argmin}} \left\{ \frac{1}{n} \left\| \mathcal{P}_{\mathcal{O}} \left(\mathbf{Y}(\mathbf{0}) - \mathbf{L} - \boldsymbol{\eta} \mathbf{1}_T^\top - \mathbf{1}_N \boldsymbol{\beta}^\top \right) \right\|_F^2 + \lambda \| \mathbf{L} \|_{tr} \right\}$$

where $\boldsymbol{\eta} \in \mathbb{R}^N$ is the vector whose elements represent the unit effects and $\boldsymbol{\beta} \in \mathbb{R}^T$ is the vector whose elements represent the

time effects. Note that incorporating the unit and time effects, without applying any regularization to them, has shown better empirical performance compared to the approach without fixed effects.

Under the assumptions of sub-Gaussian noise (Assumption 3.4.1), unconfoundedness (Assumption 3.5.2), and spikiness (Assumption 2.5.2), they derived theoretical results for the following estimator of the true low-rank matrix L^*:

$$\widehat{\mathbf{L}} = \underset{\|\mathbf{L}\|_{\max} \leq \alpha^*}{\operatorname{argmin}} \left\{ \frac{1}{n} \|\mathcal{P}_{\mathcal{O}}(\mathbf{Y}(\mathbf{0}) - \mathbf{L})\|_F^2 + \|\mathbf{L}\|_{tr} \right\}. \tag{3.4}$$

(3.4) is a slightly modified version of (3.2) introducing an additional constraint $\|\widehat{\mathbf{L}}\|_{\max} \leq \alpha^*$.

A random observation process determining the set $\mathcal{O}$ [Athey et al., 2021] is defined. Let $\{t_i\}_{i\in[N]}$ on $[T]$ represent N independent random variables that indicate the times at which the units adopt the treatment. Each variable has a distribution $\left\{\pi^{(i)}\right\}_{i\in[N]}$, where $\pi^{(i)} \equiv (P(t_i = 1), \cdots, P(t_i = T))$. The set $\mathcal{O}$ is defined as $\mathcal{O} = \bigcup_{i=1}^{N} \{(i,1),(i,2),\ldots,(i,t_i)\}$. The expectation with respect to all distributions $\left\{\pi^{(i)}\right\}_{i\in[N]}$ is denoted as $\mathbb{E}_\pi$. Additionally, we define p_{c} as

$$p_{\mathrm{c}} := \min_{1\leq i\leq N} \pi_T^{(i)} \tag{3.5}$$

where $\pi_T^{(i)}$ represents the probability of unit i never adopt the treatment. The parameter p_{c} is related to the number of observations (i.e., the number of control entries) and plays a crucial role in establishing the upper bound for the estimation error of the

counterfactuals. The upper bound is given by:

$$\frac{\left\|\mathbf{L}^* - \widehat{\mathbf{L}}\right\|_F}{\sqrt{NT}} \leq C\sqrt{\left(\frac{\sigma^2 r \log(N+T)}{T p_c^2} \vee \frac{\sigma^2 r \log^3(N+T)}{N p_c^2}\right) \vee \left(\frac{(\alpha^*)^2}{\sqrt{N} p_c} \vee \frac{(\alpha^*)^2 \log(N+T)}{N p_c}\right)}. \tag{3.6}$$

It should be noted that the time-dependency sampling structure results in a deterioration of the convergence result (3.6) compared to the result (2.14) obtained with a completely random sampling structure.

3.3.2 De-biased estimator for the average treatment effect

Farias et al. [2021] proposed the de-biased estimator for the average treatment effect for the treated. The observed outcome matrix $\mathbf{Y} \in \mathbb{R}^{N \times T}$ is modeled as follows:

$$\mathbf{Y} = \mathbf{L}^* + \boldsymbol{\varepsilon} + \boldsymbol{\Theta} \circ \mathbf{W}$$

$$\Theta_{it} = \theta^* + \delta_{it}$$

where $\circ$ is the Hadamard or 'entrywise' product. The counterfactual matrix is the sum of the low-rank matrix $\mathbf{L}^*$ and the error matrix $\boldsymbol{\varepsilon}$. $\mathbf{W}$ is the treatment matrix with elements $W_{it}, i \in [N]$ and $t \in [T]$. $\boldsymbol{\Theta}$ represents an unknown matrix of treatment effects, whose elements consist of the sum of the average treatment effect θ^* and the residual treatment effect δ_{it}.

First, the low-rank matrix and the treatment effect are jointly estimated using the nuclear norm penalty. To address the bias

issue resulting from the penalization of the nuclear norm, a de-biasing step is incorporated. The estimation procedure consists of the following steps:

$$(\hat{\mathbf{L}}^{(init)}, \hat{\theta}^{(init)}) \in \underset{\mathbf{L}\in\mathbb{R}^{N\times T}, \theta\in\mathbb{R}}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{Y}-\mathbf{L}-\theta\mathbf{W}\|_F^2 + \lambda\|\mathbf{L}\|_{tr} \quad (3.7)$$

$$\widehat{\theta}^{d} := \widehat{\theta}^{(init)} - \lambda \frac{\langle \mathbf{W}, \widehat{\mathbf{U}}\widehat{\mathbf{V}}^\top \rangle}{\left\|\Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{W})\right\|_F^2} \quad (3.8)$$

where $\widehat{\mathcal{T}}^\perp := \mathcal{T}^\perp(\widehat{\mathbf{U}}, \widehat{\mathbf{V}})$ (please refer to (3.15)) and $\widehat{\mathbf{U}}\widehat{\Xi}\widehat{\mathbf{V}}$ be the compact SVD of $\widehat{\mathbf{L}}^{(init)}$. The de-biasing step (3.8) is based on Lemma B.1.6 [Farias et al., 2021, Lemma 1]. To solve the convex problem (3.7), an alternating optimization method can be employed, which iteratively updates the estimates of the low-rank matrix and the treatment effect until convergence is achieved.

In their theoretical studies, the authors made certain assumptions to facilitate the analysis. They assumed incoherence (Assumption 2.5.1) of the true counterfactual matrix $\mathbf{L}^*$. They also assumed a deterministic treatment pattern for the treatment matrix $\mathbf{W}$, allowing for more general patterns beyond the block or the staggered structures. However, additional conditions (Assumption 3.3.1) on $\mathbf{W}$ are necessary for theoretical development. For brevity, we denote $\mathcal{T}^{*\perp} = \mathcal{T}^\perp(\mathbf{U}^*, \mathbf{V}^*)$ (please refer to (3.15)).

Assumption 3.3.1. *There exist positive constants C_{r_1}, C_{r_2} such that*

(a) $\|\mathbf{W}\mathbf{V}^*\|_{\mathrm{F}}^2 + \left\|\mathbf{W}^\top\mathbf{U}^*\right\|_{\mathrm{F}}^2 \le (1 - C_{r_1}/\log(N\wedge T))\,\|\mathbf{W}\|_{\mathrm{F}}^2$.

(b) $\left|\left\langle \mathbf{W}, \mathbf{U}^*\mathbf{V}^{*\top}\right\rangle\right| \|\Pi_{\mathcal{T}^{*\perp}}(\mathbf{W})\| \le$
$(1 - C_{r_2}/\log(N\wedge T))\,\|\Pi_{\mathcal{T}^{*\perp}}(\mathbf{W})\|_{\mathrm{F}}^2$.

These conditions ensure that $\mathbf{W}$ does not exhibit a pattern similar to that of $\mathbf{L}^*$, as such similarity would hinder the accurate estimation of the causal effect. The equivalent form of Assumption 3.3.1 (a) [Farias et al., 2021, Appendix C] is as follows:

$$\frac{C'_{r_1}}{\log(N \wedge T)} \|\mathbf{W}\|_{\mathrm{F}}^2 \leq \|\Pi_{\mathcal{T}^{*\perp}}(\mathbf{W})\|_{\mathrm{F}}^2 - \left\|\mathbf{U}^{*\top}\mathbf{W}\mathbf{V}^*\right\|_{\mathrm{F}}^2.$$

This states that the size of the projection of W onto the space $\mathcal{T}^{*\perp}$ should be sufficiently large.

They showed the asymptotic normality of the de-biased estimator under certain conditions:

$$\left(\widehat{\theta}^d - \theta^*\right) / V_{d,\theta}^{1/2} \to \mathcal{N}(0,1),$$

$$V_{d,\theta} = \sum_{i,t} \Pi_{\mathcal{T}^{*\perp}}(\mathbf{W})_{it}^2 \operatorname{Var}\left(\varepsilon_{it} + \delta_{it} W_{it}\right) / \left(\sum_{i,t} \Pi_{\mathcal{T}^{*\perp}}(\mathbf{W})_{it}^2\right)^2,$$

where $\Pi_{\mathcal{T}^{*\perp}}(\mathbf{W}) = (I_N - \mathbf{U}^*\mathbf{U}^{*\top})\mathbf{W}(I_T - \mathbf{V}^*\mathbf{V}^{*\top})$. It holds when the estimator $\widehat{\mathbf{L}}$ converges to $\mathbf{L}^*$ as $N, T \to \infty$.

3.4 The proposed estimator

We model the observed outcomes data matrix $\mathbf{Y} \in \mathbb{R}^{N \times T}$ as

$$\begin{aligned} \mathbf{Y} &= \mathbf{L}^* + \boldsymbol{\varepsilon} + \boldsymbol{\Theta} \circ \mathbf{W} \\ \Theta_{it} &= \theta^* + \delta_{it} \end{aligned} \tag{3.9}$$

where $\circ$ is the Hadamard or 'entrywise' product. Specifically, $\mathbf{Y}(\mathbf{0}) = \mathbf{L}^* + \boldsymbol{\varepsilon}$ represents the potential control matrix where $\boldsymbol{\varepsilon}$ is the noise matrix, while $\mathbf{Y}(\mathbf{1})$ represents the potential treated matrix. The matrix $\boldsymbol{\Theta}$ represents an unknown matrix of heterogeneous treatment effects. $\mathbf{W}$ is the treatment matrix with entries

in the set $\{0, 1\}$. We model the heterogeneous treatment effects with θ^* representing the average treatment effect and $\boldsymbol{\delta}$ representing the residual matrix of the treatment effect. Our goal is to estimate the average treatment effect:

$$\theta^* = \mathbb{E}\left(\mathbf{Y}(\mathbf{1}) - \mathbf{Y}(\mathbf{0})\right).$$

Assumptions for the noise matrices are as follows:

Assumption 3.4.1. *The elements of $\boldsymbol{\varepsilon}$ are σ-sub-Gaussian with mean 0 and independent of each other.*

Assumption 3.4.2. *The elements of $\boldsymbol{\delta}$ are τ-sub-Gaussian with mean 0 and independent of each other.*

We estimate the parameters of interest in two steps. In the first step, we adopt nonconvex penalties to estimate the potential control outcomes (counterfactuals) as follows:

$$\begin{aligned}\widehat{\mathbf{L}} \in \underset{\mathbf{L}\in\mathbb{R}^{N\times T}}{\operatorname{argmin}} &\left\{ \frac{1}{n}\left\|\mathcal{P}_{\mathcal{O}}\left(\mathbf{Y}(\mathbf{0}) - \mathbf{L}\right)\right\|_F^2 + \sum_{i=1}^{min(N,T)} g_\lambda\left(\xi_i(\mathbf{L})\right) \right\} \\ &:= F_\lambda(\mathbf{L})\end{aligned} \tag{3.10}$$

where $g_\lambda(\cdot)$ is a nonconvex penalty such as SCAD and MCP. In the second step, we calculate the average treatment effect for the treated (ATT):

$$\widehat{\theta} = \sum_{(i,t):W_{it}=1} \left[Y_{it}(1) - \widehat{Y}_{it}(0)\right] / \sum_{(i,t)} W_{it}. \tag{3.11}$$

We can enhance the performance by incorporating two-way fixed effects. (3.10) would be modified accordingly:

$$
\begin{aligned}
&(\widehat{\mathbf{L}}, \widehat{\boldsymbol{\eta}}, \widehat{\boldsymbol{\beta}}) \in \\
&\underset{\mathbf{L}\in\mathbb{R}^{N\times T}, \boldsymbol{\eta}\in\mathbb{R}^{N}, \boldsymbol{\beta}\in\mathbb{R}^{T}}{\operatorname{argmin}} \frac{1}{n} \left\| \mathcal{P}_{\mathcal{O}} \left(\mathbf{Y}(\mathbf{0}) - \mathbf{L} - \boldsymbol{\eta}\mathbf{1}_T^\top - \mathbf{1}_N \boldsymbol{\beta}^\top \right) \right\|_F^2 \\
&\qquad + \sum_{i=1}^{min(N,T)} g_\lambda \left(\xi_i(\mathbf{L}) \right)
\end{aligned}
\tag{3.12}
$$

where $\boldsymbol{\eta} \in \mathbb{R}^N$ is the vector whose elements represent the unit effects and $\boldsymbol{\beta} \in \mathbb{R}^T$ is the vector whose elements represent the time effects. Note that we do not impose regularization on the unit effects and time effects.

For the estimation, we adopt the Proximal Gradient Homotopy (PGH) Algorithm [Gui et al., 2016; Wang et al., 2014; Xiao and Zhang, 2013]. Refer to Section 2.6.2 for the algorithm. The details of the calculation and the stopping criterion of the algorithm are presented in Appendix A. When estimating two-way fixed effects in (3.12), we can use first-order conditions once $\mathbf{L}_t$ is updated at each t-th path of the algorithm.

3.5 Theoretical results

3.5.1 Recovery of the potential control outcomes

In this section, we provide theoretical results regarding the recovery of the counterfactual matrix. We demonstrate that, given certain conditions, the nonconvex penalized estimator achieves a faster convergence rate compared to the previous method that uses

the convex penalty. Under specific conditions, the oracle estimator can be a local solution to the nonconvex penalized problem, accurately recovering the true rank. Furthermore, the upper bound of the oracle estimator is equivalent to that of the nonconvex penalized estimator.

Under Assumption 2.5.2, we focus on the theoretical results related to the following estimator for $\mathbf{L}^*$:

$$\widehat{\mathbf{L}} = \underset{\|\mathbf{L}\|_{\max} \leq \alpha^*}{\operatorname{argmin}} \left\{ \frac{1}{n} \|\mathcal{P}_{\mathcal{O}}(\mathbf{Y}(\mathbf{0}) - \mathbf{L})\|_F^2 + G_\lambda(\mathbf{L}) \right\}, \tag{3.13}$$

which is a slightly modified version of (3.10). Here, we have introduced the additional constraint $\|\widehat{\mathbf{L}}\|_{\max} \leq \alpha^*$.

The assumption stated below corresponds to the regularity condition of the $g_\lambda(\cdot)$ and $\bar{g}_\lambda(\cdot)$ functions. The nonconvex penalties SCAD and MCP satisfy this assumption.

Assumption 3.5.1.

(i) There exists a constant $\nu > 0$ on the nonnegative real line such that the function $g_\lambda(t)$ satisfies $g'_\lambda(t) = 0$ for all $t \geq \nu$.

(ii) On the nonnegative real line, $\bar{g}'_\lambda(t)$ is both monotone and Lipschitz continuous. In other words, for any $t' \geq t$, there exists a constant $\zeta_- \geq 0$ such that $\bar{g}'_\lambda(t') - \bar{g}'_\lambda(t) \geq -\zeta_-(t' - t)$.

(iii) Both the function $\bar{g}_\lambda(t)$ and its derivative $\bar{g}'_\lambda(t)$ pass through the origin, meaning that $\bar{g}_\lambda(0) = \bar{g}'_\lambda(0) = 0$.

(iv) On the nonnegative real line, the absolute value of $\bar{g}'_\lambda(t)$ is bounded above by λ, specifically $|\bar{g}'_\lambda(t)| \leq \lambda$.

In 3.5.1 (i), the constant $\nu = \gamma\lambda$ holds for both SCAD and MCP penalties. The constant ζ_- in Assumption 3.5.1 (ii) determines the curvature of the concavity of the penalty functions. For SCAD penalty function, $\zeta_- = \frac{1}{\gamma-1}$, while for MCP penalty function, $\zeta_- = \frac{1}{\gamma}$.

Recall the random observation process that determines the set $\mathcal{O}$ and definition of p_{c} in (3.5) in Section 3.3.1. Traditional literature on the MC [Chen et al., 2020; Gui et al., 2016; Gunasekar et al., 2014; Koltchinskii et al., 2011; Negahban and Wainwright, 2012] assumes randomly sampled missing entries without any specific structure. In our work, we build upon the approach introduced by Athey et al. [2021], which extends the MC framework to incorporate the time-dependence structure in the observation process. This modification enables us to capture more realistic scenarios and address challenges posed by causal panel data.

The next assumption, known as unconfoundedness, is crucial for identifying the average treatment effect.

Assumption 3.5.2. *The adoption dates t_i for units are independent of each other and of $\boldsymbol{\varepsilon}$ conditional on $\mathbf{L}^*$.*

Assumption 3.5.2 means that the error terms are independent of the assignment of treatment given the systematic component. Further details on the unconfoundedness in the program evaluation can be found in Athey et al. [2021].

Recall the two sets (2.15) corresponding to the non-zero true singular values, and their cardinalities, r_1 and r_2. The following

theorem provides the upper bound for the estimation error of our proposed estimator. For the detailed proof of this theorem, please refer to Appendix B.1.1.

Theorem 3.5.1. *Assume that* $\mathbf{L}^*$ *satisfies* $\|\mathbf{L}^*\|_{\max} \leq \alpha^*$. *For any optimal solution* $\widehat{\mathbf{L}}$ *to the problem of* (3.13) *with SCAD or MCP penalty, there exist constants* C_1, C_2 *and* C_3 *such that, for a penalty parameter* $\lambda \geq C_2\sigma\frac{\left[\sqrt{N\log(N+T)}\vee\sqrt{T\log^3(N+T)}\right]}{n}$, *the following inequality holds with a probability at least* $1-(N+T)^{-2}$*:*

$$\frac{\left\|\mathbf{L}^* - \widehat{\mathbf{L}}\right\|_F}{\sqrt{NT}} \leq C_1\sqrt{\left[\frac{\sigma^2 r_1\log^3(N+T)}{Np_c^2} + \left(\frac{\sigma^2 r_2\log(N+T)}{Tp_c^2} \vee \frac{\sigma^2 r_2\log^3(N+T)}{Np_c^2}\right)\right] \vee \left(\frac{(\alpha^*)^2}{\sqrt{N}p_c} \vee \frac{(\alpha^*)^2\log(N+T)}{Np_c}\right)}. \tag{3.14}$$

In Theorem 3.5.1, (3.14) shows that when $r_1 > 0$, if the conditions $(N\wedge T)p_c^2 \gg \log^3(N+T)$ and $Np_c \gg (\sqrt{N}\vee\log(N+T))$ are satisfied, the right-hand side converges to 0 as $N, T \to \infty$ in the normal setting where $r_1 = r_2 = \mathcal{O}(1)$. In particular, when $r_2 = 0$, the estimator $\widehat{\mathbf{L}}$ converges to the true counterfactual matrix $\mathbf{L}^*$ if $Np_c^2 \gg \log^3(N+T)$ and $Np_c \gg (\sqrt{N}\vee\log(N+T))$. This means that as long as N grows faster than T and tends to infinity, $\widehat{\mathbf{L}}$ converges to $\mathbf{L}^*$.

The proposed estimator exhibits a faster convergence rate compared to the nuclear norm-based estimator. In the nuclear norm-based estimator (3.6) [Athey et al., 2021], when $r_1 = 0$, the upper bounds (3.14) and (3.6) are identical. However, when $r_1 > 0$ and

$N \gg T\log^2(N+T)$ (3.14) can provide a tighter compared to (3.6). Moreover, when $r_2 = 0$ and $N \gg T\log^2(N+T)$ (3.14) can be much tighter than (3.6). Furthermore, it is worth noting that (3.14) becomes equivalent to (3.18) in Theorem 3.5.3. As mentioned by Athey et al. [2021], the estimation error gets worse when the observations are time-dependent compared to that in previous study [Gui et al., 2016] on the MC, where the sampling of missing entries is assumed to be completely at random. This implies that a larger amount of data is required for achieving consistent estimation in the presence of time-dependent observations.

To define the oracle estimator, we introduce two subspaces. Recall that the compact SVD of $\mathbf{L}^*$ is given by $\mathbf{L}^* = \mathbf{U}^*\Xi^*\mathbf{V}^{*\top}$, where $\mathbf{U}^* \in \mathbb{R}^{N\times r}$, $\mathbf{V}^* \in \mathbb{R}^{T\times r}$, and $\Xi^* = \mathrm{diag}(\boldsymbol{\xi}^*) \in \mathbb{R}^{r\times r}$. The $\mathrm{row}(\cdot) \subseteq \mathbb{R}^T$ and $\mathrm{col}(\cdot) \subseteq \mathbb{R}^N$ represent the row and the column space of an $N \times T$ matrix, respectively. It is worth noting that $\mathrm{col}(\mathbf{L}^*) = \mathrm{span}\{u_i^*\}$ and $\mathrm{row}(\mathbf{L}^*) = \mathrm{span}\{v_i^*\}$ where u_i^* and v_i^* are the i-th left and right singular vectors of $\mathbf{L}^*$, respectively. Based on this, we can construct the subspace $\mathcal{F}$ and $\mathcal{T}^\perp$ of $\mathbb{R}^{N\times T}$ [Gui et al., 2016; Gunasekar et al., 2014; Negahban et al., 2012]:

$$\begin{aligned}\mathcal{F}(\mathbf{U}^*, \mathbf{V}^*) =& \{\Delta \mid \mathrm{row}(\Delta) \subseteq \mathrm{row}(\mathbf{L}^*) \text{ and } \mathrm{col}(\Delta) \subseteq \mathrm{col}(\mathbf{L}^*)\} \\ \mathcal{T}^\perp(\mathbf{U}^*, \mathbf{V}^*) =& \{\Delta \mid \mathrm{row}(\Delta) \perp \mathrm{row}(\mathbf{L}^*) \text{ and } \mathrm{col}(\Delta) \perp \mathrm{col}(\mathbf{L}^*)\}.\end{aligned} \tag{3.15}$$

Note that $\mathcal{F} \neq \mathcal{T}$, but $\mathcal{F} \subseteq \mathcal{T}$. For any $\mathbf{L}_1 \in \mathcal{F}(\mathbf{U}^*, \mathbf{V}^*)$ and $\mathbf{L}_2 \in \mathcal{T}^\perp(\mathbf{U}^*, \mathbf{V}^*)$, it can be verified that $\mathbf{L}_1^T\mathbf{L}_2 = 0$ by definition (consequently, $\langle \mathbf{L}_1, \mathbf{L}_2 \rangle = 0$), which implies that $\mathbf{L}_1$ is orthogonal to the space $\mathcal{T}^\perp(\mathbf{U}^*, \mathbf{V}^*)$. The decomposability property of

regularizers can be found in Negahban et al. [2012]. It is worth mentioning that $\mathcal{F}$ is a rank-r subspace consisting of matrices with rank r. A projection operator $\Pi_{\mathcal{T}^\perp}(\cdot)$ onto the $\mathcal{T}^\perp$ is denoted as follows:

$$\Pi_{\mathcal{T}^\perp}(\mathbf{A}) = \left(I - \mathbf{U}^*\mathbf{U}^{*\top}\right)\mathbf{A}\left(I - \mathbf{V}^*\mathbf{V}^{*\top}\right). \tag{3.16}$$

Now we can introduce the oracle estimator as follows:

$$\widehat{\mathbf{L}}_O = \underset{\mathbf{L}\in\mathcal{F}(\mathbf{U}^*,\mathbf{V}^*),\|\mathbf{L}\|_{\max}\leq\alpha^*}{\operatorname{argmin}} f_n(\mathbf{L}). \tag{3.17}$$

The oracle estimator $\widehat{\mathbf{L}}_O$ is the estimator when the true rank is known. In this case, the objective function only considers the loss function without any penalty functions, and the solution is restricted to the rank-r subspace $\mathcal{F}(\mathbf{U}^*, \mathbf{V}^*)$. It is important to note that the functions to be optimized in (3.13) are nonconvex, which makes finding the global minimum challenging. Furthermore, there exist multiple local minima. In the following theorem, we prove that the oracle estimator can be the local minimum of the problem under certain conditions. The proof of this result can be found in Appendix B.1.2.

Theorem 3.5.2. *Assuming that* $\operatorname{rank}(\mathbf{L}^*) = r$*, let* $\widehat{\Delta}_O = \widehat{\mathbf{L}}_O - \mathbf{L}^*$*. Suppose* $\alpha_{sp}(\widehat{\Delta}_O) \leq \frac{1}{c_0}\sqrt{\frac{Np_c}{\log(N+T)}}$*. For positive constants* $C_1, \ldots C_3$ *and* C_4*, if* $\boldsymbol{\xi}^*$ *satisfies the condition*

$$\min_{i\in S}|(\boldsymbol{\xi}^*)_i| \geq \gamma\lambda + C_1\sqrt{\frac{\sigma^2 rT\log^3(N+T)}{p_c^2} \vee \frac{(\alpha^*)^2T\sqrt{N}}{p_c}},$$

where $S = \operatorname{supp}(\boldsymbol{\xi}^*)$*, then for the estimator of* (3.13) *with SCAD*

or MCP penalty and penalty parameter

$$\lambda \geq C_2 \frac{\sigma\left[\sqrt{N\log(N+T)} \vee \sqrt{T}\log^{3/2}(N+T)\right]}{n} + C_3\sqrt{\frac{\sigma^2 r\log(N+T)}{T} \vee \frac{p_c(\alpha^*)^2\sqrt{N}}{T\log^2(N+T)}},$$

the following result holds with a probability greater than $1-(N+T)^{-2}$, $\widehat{\mathbf{L}}_O$ *is one of the local minima of the problem* (3.13)*. Additionally, we have the inequality*

$$\frac{\left\|\mathbf{L}^* - \widehat{\mathbf{L}}_O\right\|_F}{\sqrt{NT}} \leq C_4\sqrt{\frac{\sigma^2 r\log^3(N+T)}{Np_c^2} \vee \frac{(\alpha^*)^2}{\sqrt{N}p_c}}.$$

It is worthy note about the condition of $\alpha_{sp}(\widehat{\Delta}_O) \leq \frac{1}{c_0}\sqrt{\frac{Np_c}{\log(N+T)}}$ in Theorem 3.5.2. This condition can be equivalently expressed as $\frac{\|\widehat{\mathbf{L}}_O - \mathbf{L}^*\|_F^2}{NT} \geq c_0'(\alpha^*)^2\frac{\log(N+T)}{Np_c}$, which holds when N is significantly larger than T and p_c is sufficiently large. Alternatively, a weaker condition can be stated as $\frac{1}{NT}\mathbb{E}_\pi\left[\sum_{i=1}^N\sum_{t=1}^{t_i}\left((\widehat{\mathbf{L}}_O)_{it} - L_{it}^*\right)^2\right] \geq c_0'(\alpha^*)^2\frac{\log(N+T)}{N}$. This is based on the proof Lemma B.1.3, which can be found in Athey et al. [2021].

The following theorem presents the upper bound for the oracle estimator. The proof can be found in Appendix B.1.3.

Theorem 3.5.3. *Suppose that* $\mathbf{L}^*$ *has rank* r *and satisfies* $\|\mathbf{L}^*\|_{\max} \leq \alpha^*$*. Then, for a constant* C*, the upper bound between the oracle estimator* $\widehat{\mathbf{L}}_O$ *(the solution to* (3.17)*) and* $\mathbf{L}^*$ *satisfies*

$$\frac{\left\|\mathbf{L}^* - \widehat{\mathbf{L}}_O\right\|_F}{\sqrt{NT}} \leq C\sqrt{\frac{\sigma^2 r\log^3(N+T)}{Np_{\mathrm{c}}^2} \vee \left(\frac{(\alpha^*)^2}{\sqrt{N}p_{\mathrm{c}}} \vee \frac{(\alpha^*)^2\log(N+T)}{Np_{\mathrm{c}}}\right)}. \tag{3.18}$$

The inequality holds with a probability of at least $1-(N+T)^{-2}$.

The upper bound of the oracle estimator coincides with the upper bound of our proposed estimator (3.14) when $r_2 = 0$. Although the oracle estimator cannot be obtained in practice, it serves as an ideal estimator with an optimal convergence rate.

3.5.2 Estimation of the average treatment effect

In this section, we investigate the asymptotic normality of our estimator for the causal effect. Additionally, we compare it to that of the de-biased estimator [Farias et al., 2021] under the assumption of spikiness and observe that our estimator has a smaller asymptotic variance.

We first show the following theorem, which is the asymptotic normality of our estimator for the causal effect. The proof of this theorem can be found in Appendix B.1.4. Let $n_t = NT - n$ be the number of treated observations.

Theorem 3.5.4 (Asymptotic Normality)**.** *Suppose that* $\mathbf{L}^*$ *has rank* r *and satisfies* $\|\mathbf{L}^*\|_{max} \leq \alpha^*$*. Additionally, assume that each* δ_{it} *is a mean-zero, independent random variable with a sub-Gaussian norm* $\|\delta_{it}\|_{\psi_2} = O(1)$*. We assume that* $r = \alpha^* = \sigma = \mathcal{O}(1)$ *and* $n_t = \Omega(NT)$*. Under these conditions, for the estimator of* (3.11) *with a value of* $\widehat{\mathbf{Y}}(0) = \widehat{\mathbf{L}}_O$*,*

$$\widehat{\theta} - \theta^* = \frac{\langle \boldsymbol{\varepsilon} + \delta \circ \mathbf{W}, \mathbf{W} \rangle}{\|\mathbf{W}\|_{\mathrm{F}}^2} + O\left(\frac{1}{\sqrt{N} p_c} \vee \frac{\log(N+T)}{N p_c} \vee \frac{\log^3(N+T)}{N p_c^2} \right) \tag{3.19}$$

holds with a probability greater than $1 - (N+T)^{-2}$*. Furthermore,*

if $p_c \gg \frac{1}{\sqrt{N}} \log^{\frac{3}{2}}(N+T)$,

$$\begin{aligned} \left(\widehat{\theta} - \theta^*\right) / V_\theta^{1/2} &\to \mathcal{N}(0,1), \\ V_\theta = \sum_{it} W_{it}^2 \operatorname{Var}\left(\varepsilon_{it} + \delta_{it} W_{it}\right) &/ \left(\sum_{it} W_{it}^2\right)^2 \end{aligned} \tag{3.20}$$

holds as $N, T \to \infty$.

It is important to note that the assumption $n_t = \|\mathbf{W}\|_F^2 = \Omega(NT)$ in Theorem 3.5.4 encompasses the block and the staggered structure, but it accounts for neither the single-treated-unit nor the single-treated-time block structure.

In the remainder of this section, we establish the asymptotic normality of the de-biased estimator under the spikiness condition of the true matrix, allowing us to compare the asymptotic variance of our proposed estimator to that of the de-biased estimator. The original de-biased estimator (Section 3.3.2, [Farias et al., 2021]) assumed incoherence of $\mathbf{L}^*$ (Assumption 2.5.1), which is an extreme assumption in MC models with noise. Therefore, we extend their work and provide a more general result for asymptotic normality under the milder assumption of the low-rank matrix.

Let $\chi := \xi_1^*/\xi_r^*$ be the condition number of $\mathbf{L}^*$. The following theorem demonstrates the asymptotic normality of the de-biased estimator. The proof can be found in Appendix B.1.5.

Theorem 3.5.5 (Asymptotic Normality of the de-biased estimator)**.** *Under Assumption 3.3.1, suppose each* δ_{it} *is a mean-zero, independent random variable with a sub-Gaussian norm* $\|\delta_{it}\|_{\psi_2} = O(1)$. *Assume that* $\chi = r = \sigma = O(1)$, $\xi_r^* = \Omega(N \wedge T)$ *and*

$n_t = \Omega(NT)$, *we have the following:*

$$\widehat{\theta}^d - \theta^* = \frac{\left\langle \boldsymbol{\varepsilon} + \delta \circ \mathbf{W}, \Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{W}) \right\rangle}{\left\| \Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{W}) \right\|_{\mathrm{F}}^2} + O\left(\frac{\sqrt{\log(N \wedge T)} \left\| \mathbf{L}^* - \widehat{\mathbf{L}}^{(init)} \right\|_F}{\sqrt{NT}} \right). \tag{3.21}$$

Furthermore, if $\frac{\sqrt{\log(N \wedge T)} \left\| \mathbf{L}^* - \widehat{\mathbf{L}}^{(init)} \right\|_F}{\sqrt{NT}} \to 0$ *as* $N, T \to \infty$,

$$\left(\widehat{\theta}^d - \theta^* \right) / V_{d,\theta}^{1/2} \to \mathcal{N}(0,1),$$

$$V_{d,\theta} = \sum_{i,t} \Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{W})_{it}^2 \operatorname{Var}\left(\varepsilon_{it} + \delta_{it} W_{it} \right) / \left(\sum_{i,t} \Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{W})_{it}^2 \right)^2.$$

In Theorem 3.5.5, we do not provide a proof for the convergence of $\widehat{\mathbf{L}}^{(init)}$ to $\mathbf{L}^*$ in the joint estimation problem of $\mathbf{L}^*$ and θ^* as it is not the main focus of our study. Previous work [Athey et al., 2021] demonstrated convergence results for the nuclear norm-based estimation of $\mathbf{L}^*$. Additionally, Farias et al. [2021, Lemma 8] characterized the estimation error under the assumption of incoherence (Assumption 2.5.1) of $\mathbf{L}^*$. In the MC literature [Negahban and Wainwright, 2012], the upper bound for the nuclear norm-based estimator was obtained when the missing pattern is completely at random. To establish the upper bound in the joint estimation problem for $\mathbf{L}^*$ and θ^*, one can extend or combine the proofs from previous literature.

The following proposition demonstrates that the asymptotic variance V_θ of our estimator in (3.19) is smaller than $V_{d,\theta}$ (3.21) when the elements of error matrices have constant variances. This implies that our estimator exhibits less variation and provides more accurate estimations compared to the de-biased estimator. The proof can be found in Appendix B.2.3.

Proposition 3.5.6. *Assume that* $\mathrm{Var}(\varepsilon_{it}) = \sigma_\varepsilon^2$ *and* $\mathrm{Var}(\delta_{it}) = \sigma_\delta^2$. *For the asymptotic variance of our estimator, denoted as* V_θ, *and that of the de-biased estimator, denoted as* $V_{d,\theta}$, *we have the following inequality:*

$$V_\theta \leq V_{d,\theta}.$$

3.6 Numerical studies

In our numerical studies, we focus on two main tasks. Firstly, our goal is to recover the potential control matrix using our proposed method and evaluate their performance. To achieve this, we conduct simulations to investigate the upper bounds and analyze real-world data. Secondly, our focus shifts toward estimating the average treatment effect. We perform simulations to assess the asymptotic normality of the estimators and analyze semi-synthetic data.

3.6.1 Recovery of the potential control outcomes

We begin by conducting the simulations to validate our theoretical result regarding the upper bound. In this simulation, we consider a true matrix with dimensions $N = 40, T = 10$, and a rank of 5. We generate the true matrix $\mathbf{L}^*$ using the compact SVD: $\mathbf{L}^* = \mathbf{U}^* \operatorname{diag}(\boldsymbol{\xi}^*)\mathbf{V}^{*\top}$, where columns of $\mathbf{U}^*$ and $\mathbf{V}^*$ are the left and right singular vectors of a random matrix and we set the singular values $\boldsymbol{\xi}^*$ to be $(32.5, 31.8, 29.2, 26.5, 20.7)$. Furthermore, we sample observation noise from $\mathcal{N}(0, 0.5^2)$.

In the scenario of the block structure, we randomly select the number of control units (N_c) to ensure that the ratio of control units to the total number of units is $N_c/N = \{0.2, 0.4, 0.6, 0.8, 0.9\}$. For treated units, the treatment is adopted at time $T_0 = 7$, where T_0 represents the treatment adoption time. We repeated the sampling process 10 times for each N_c/N ratio. We calculate the average of the mean squared error in Frobenius norm (MSE) $\|\mathbf{L}^* - \widehat{\mathbf{L}}\|_F^2/(NT)$ and the estimated rank (Rank) from the 10 repetitions. The SCAD estimator is obtained with a parameter value of $\gamma = 5$. The tuning parameter λ is selected using 5-fold cross-validation.

Figure 3.1 illustrates the results obtained by the MC methods using the SCAD and the nuclear norm penalty. In Figure 3.1 (**Left**), we observe that as the number of controlled units increases (i.e., fewer missing elements in the matrix) the estimation error becomes smaller. Moreover, the proposed estimator with the SCAD penalty outperforms the nuclear-norm estimator in terms of estimation accuracy for all N_c/N ratios. This finding aligns with our theoretical result regarding the upper bound. Figure 3.1 (**Right**) displays that the SCAD penalty provides estimates that are closer to the true rank compared to the nuclear norm penalty across all N_c/N settings. Note that the true rank is denoted by the black horizontal line at 5.

In the staggered structure, we follow a similar approach. For treated units, the treatment is adopted at various times starting from T_0. Figure 3.2 illustrates that the MSE is smaller and the true rank is better estimated when using the SCAD penalty compared

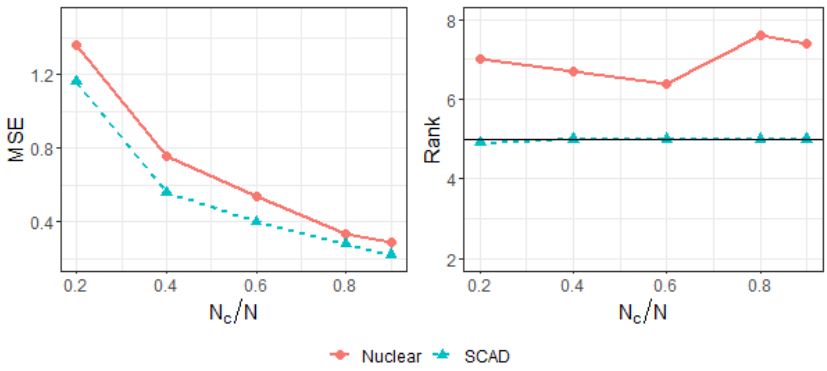

Figure 3.1: The simulation results for the block structure with the nuclear norm and the SCAD penalty: the average MSE (**Left**) and the average Rank (**Right**). Note that the black horizontal line in (**Right**) represents the true rank of 5.

to using the nuclear norm penalty, which is similar to the results for the block structure.

In the analysis using real data, we aim to assess the accuracy of our proposed method and other existing approaches in estimating the counterfactual matrix. We work with data matrices that contain control units, and we deliberately designate certain entries as treated (i.e., missing). We impute the missing values with various methods including our own and then compare the imputed values with the actual control outcomes. Two treatment adoption scenarios are considered: the block structure and the staggered structure. In both scenarios, the tuning parameter λ is selected using 5-fold cross-validation to minimize the average root mean squared error (RMSE). The parameter γ of the SCAD penalty is

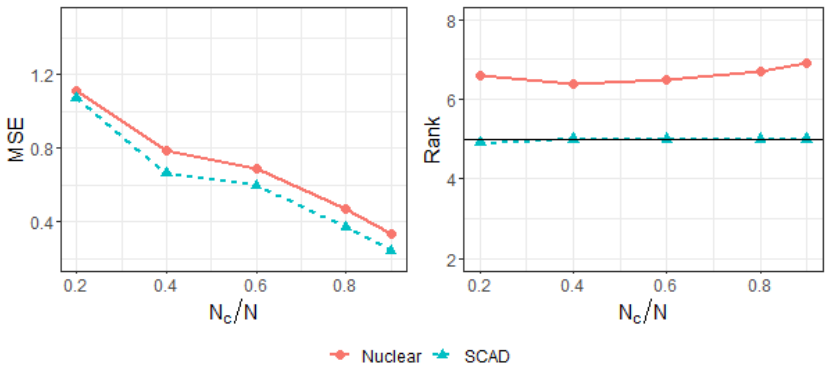

Figure 3.2: The simulation results for the staggered structure with the nuclear norm and the SCAD penalty: the average MSE (**Left**) and the average Rank (**Right**). Note that the black horizontal line in (**Right**) represents the true rank of 5.

set as fixed, following Fan and Li [2001]. The treatment adoption process is repeated 10 times, and we report the average RMSE.

We compare the performance of the following estimators in our evaluation: (1) Fixed: A two-way fixed effects model, where the outcome variable $Y_{it}(0)$ is modeled as $Y_{it}(0) = \eta_i + \beta_t + \varepsilon_{it}$. (2) Synthetic [Abadie et al., 2010; Athey et al., 2021]: The synthetic control method. The weights are estimated using pre-intervention outcomes. (3) MC (Nuclear) [Athey et al., 2021]: The MC method with the nuclear norm penalty. (4) MC+fixed (Nuclear) [Athey et al., 2021]: The MC method with the nuclear norm penalty, incorporating fixed effects on both units and times. (5) MC (SCAD): Our MC method with the SCAD penalty. (6) MC+fixed (SCAD): Our MC method with the SCAD penalty, incorporating fixed ef-

fects on both units and times. Note that for the nuclear norm-based methods, namely MC (Nuclear) and MC+fixed (Nuclear), we set maximum estimated ranks. These maximum ranks are determined based on the spectrums of the control matrices, where the missing entries are imputed with the averages of the matrices.

We first utilized the Cigarette sales data [Abadie et al., 2010] for our analysis. California enacted its state law called Proposition 99 to control tobacco sales in 1988. The original dataset contains treated outcomes starting from 1989 for California, while only control outcomes are observed for other states. However, for our experiment, we excluded the data for California as the untreated values are unavailable. Instead, we focused on the control per-capita cigarette sales of the remaining 38 states, covering the period from 1970 to 2000 ($N = 38$ and $T = 31$). We artificially designated certain units and time periods as treated, and then compared the actual observations with the imputed values. In the block structure scenario, we set $N_c/N \approx 0.7$ ($N_c = 27$) and $T_0 = 19$ to align with the year when California passed its tobacco control law. For the staggered structure scenario, we set $N_c/N \approx 0.7$ ($N_c = 27$). The treatment adoption times for treated units were randomly assigned after $T_0 = 13$.

Table 3.1 presents the average test RMSE and the average estimated rank (Rank) for low-rank-based models based on 10 repetitions in the block and the staggered structures for the Cigarette sales data. Among the low-rank models, the one incorporating fixed effects, particularly when utilizing SCAD penalties, yields

the best results. On the other hand, the model that only considers fixed effects performs poorly for both structures. While the synthetic control approach outperforms the low-rank models without fixed effects, its results are inferior to those obtained with the MC with fixed effects for both structures.

Method	Block	Staggered
Fixed effects	18.5289 (-)	16.5367 (-)
Synthetic	14.7203 (-)	13.3319 (-)
MC (Nuclear)	16.3344 (6)	16.0396 (6)
MC+fixed (Nuclear)	14.2500 (6)	12.9798 (5)
MC (SCAD)	15.1553 (1)	14.4285 (1)
MC+fixed (SCAD)	**12.0015 (1)**	**11.9644 (1)**

Table 3.1: The average RMSE and the average Rank of the low-rank matrix in the parentheses for the analysis of the Cigarette sales data with the block and the staggered structures

The second dataset analyzed in this thesis comprises the annual GDP of seventeen countries [Abadie et al., 2015]. The Berlin Wall fell in November 1989, and the official reunification of West and Eastern Germany took place in October 1990. We excluded the data for West Garmany due to the unavailability of untreated values. Our analysis focuses on the remaining 16 countries over the years from 1960 to 2003 ($N = 16$ and $T = 44$). In the block structure scenario, we set $N_c/N \approx 0.6$ ($N_c = 10$) and $T_0 = 31$ to align with the year of reunification. For the staggered struc-

ture scenario, we also set $N_c/N \approx 0.6$ ($N_c = 10$), and the treatment adoption times for treated units were randomly assigned after $T_0 = 21$.

Table 3.2 displays the average test RMSE and the average Rank for low-rank-based models, obtained from 10 repetitions, in the block and the staggered structures using the GDP data. We observe that the SCAD estimator with fixed effects produces the best outcomes for the block structures. On the other hand, for the staggered structures, the SCAD estimator without fixed effects outperforms the other methods. It is worth noting that the nuclear-norm-based estimators exhibit poor performance in both scenarios.

Method	Block	Staggered
Fixed effects	3450.5 (-)	2638.2 (-)
Synthetic	2407.1 (-)	2306.1 (-)
MC (Nuclear)	9788.6 (3.5)	7778.3 (3.5)
MC+fixed (Nuclear)	2386.9 (3.9)	1884.9 (3.7)
MC (SCAD)	2512.0 (1.9)	**1863.5 (2.7)**
MC+fixed (SCAD)	**2138.0 (1.6)**	1922.5 (1.2)

Table 3.2: The average RMSE and the average Rank of the low-rank matrix in the parentheses for the analysis of the GDP data with the block and the staggered structures

We briefly discuss the low-rank models and the synthetic control method which is a traditional approach in program evaluation.

The low-rank models take into account both unit and time patterns and their interactions. Specifically, the models incorporating fixed effects can capture patterns that the models only with a low-rank matrix might overlook [Athey et al., 2021]. On the other hand, the synthetic control focuses solely on patterns among units. In our experiment, we observed that the synthetic control method performed poorly when there was a significant deviation between the observed outcomes of the treated states and the control units. This finding aligns with the observations in Abadie et al. [2010, Section 3.4], where it was noted that certain states with extreme values during the pre-intervention period could not be accurately reconstructed as a convex combination of cigarette sales from other states. In summary, MC methods are more robust in the selection of treated and control units compared to the synthetic control method. For a more detailed discussion, please refer to Appendix C.

3.6.2 Estimation of the average treatment effect

To verify the theoretical results concerning the asymptotic normality of the estimators, we implement the simulations. We consider a true matrix with dimensions $N = 40, T = 40$, and a rank of 5. To generate the true matrix $\mathbf{L}^*$, we employ the compact SVD: $\mathbf{L}^* = \mathbf{U}^* \operatorname{diag}(\boldsymbol{\xi}^*)\mathbf{V}^{*\top}$. Here, columns of $\mathbf{U}^*$ and $\mathbf{V}^*$ represent the left and right singular vectors of a randomly generated matrix, and we set the singular values $\boldsymbol{\xi}^*$ to be $(32.5, 31.8, 29.2, 26.5, 20.7)$. Additionally, we set the true average treatment effect θ^* to be 1. We

independently sample the observation noise ϵ_{it} and the treatment noise δ_{it} from $\mathcal{N}(0, 0.5^2)$.

In the block structure scenario, we randomly select the control units and treatment adoption times that satisfy the conditions $20 \leq N_c \leq 39$ and $20 \leq T_0 \leq 39$, respectively [Farias et al., 2021]. To assess the asymptotic normalities of our estimator and the de-biased estimator, we conduct 200 instances. We present the empirical distributions of $(\widehat{\theta} - \theta^*)/V_\theta$ (3.20) and $(\widehat{\theta}^d - \theta^*)/V_{\theta,d}$ (3.21), where V_θ and $V_{\theta,d}$ represent the asymptotic variances of our estimator and the de-biased estimator, respectively. The tuning parameter λ is chosen using 5-fold cross-validation.

Figure 3.3 displays histograms depicting the empirical distributions of our estimator and the de-biased estimator. The histograms are overlaid with the $\mathcal{N}(0, 1)$ density function. The average asymptotic variances for 200 repetitions are 0.016 and 0.033 for our estimator and the de-biased estimator, respectively. These results provide support for our theoretical findings in Section 3.5.2.

In the analysis using real datasets, the objective is to assess the accuracy of our method and other existing approaches in estimating the average treatment effect. We deliberately designate certain entries of data matrices as treated, as described in Section 3.6.1. Additionally, we introduce artificial treatments based on Farias et al. [2021]. We set the true treatment effect $\theta^* = \sigma_\delta = \overline{\mathbf{Y}(\mathbf{0})}/10$, where $\mathbf{Y}(\mathbf{0})$ denotes the collected data and $\overline{\mathbf{Y}(\mathbf{0})}$ represents its average. The noise is randomly selected from $\mathcal{N}(0, \sigma_\delta^2)$. We report the average normalized error (ANE) $|(\widehat{\theta} - \theta^*)/\theta^*|$ over

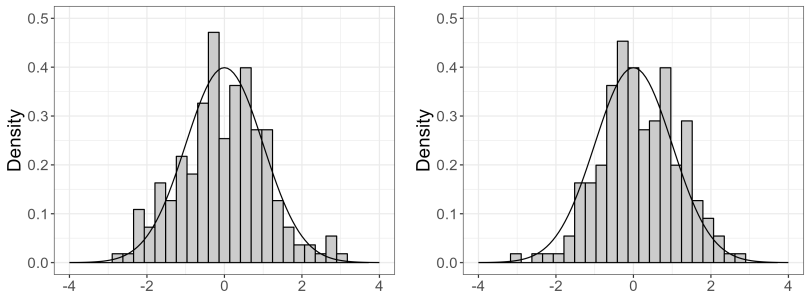

Figure 3.3: Empirical distributions of $(\widehat{\theta}-\theta^*)/V_\theta$ for our estimator (**Left**) and of $(\widehat{\theta}^d-\theta^*)/V_{\theta,d}$ for the de-biased estimator (**Right**). The lines represent the $\mathcal{N}(0,1)$ density functions.

10 repetitions. The tuning parameter is chosen through 5-fold cross-validation. We compare the results for our estimator with the de-biased estimator (De-biased) [Farias et al., 2021] and the debiased estimator incorporating the two-way fixed effect (De-biased+fixed). The other estimators for the average treatment effect, except for De-biased and De-biased+fixed, are estimated based on (3.11) using the corresponding estimators for the counterfactuals. These estimators for the average treatment effect are denoted with the same acronym as methods of estimating the counterfactuals. Note that the methods for recovering the counterfactual are described in Section 3.6.1. Additionally, we set the maximum estimated ranks for the low-rank matrices in the nuclear norm-based methods, specifically MC (Nuclear), MC+fixed (Nuclear), De-based and Debiased+fixed.

The first dataset is the Cigarette sales data [Abadie et al., 2010]. Table 3.3 presents the average ANE with the block and the

staggered structures. Among the various estimators, the estimator corresponding to the SCAD estimator incorporating fixed effects exhibits the best performance. However, it is worth noting that the de-biased estimator with fixed effects produces results that are not significantly different from the best-performing estimator for both structures.

Method	Block	Staggered
Fixed effects	0.4158	0.4710
Synthetic	0.3098	0.3906
MC (Nuclear)	0.9221	0.9193
MC+fixed (Nuclear)	0.4503	0.3721
MC (SCAD)	0.4158	0.4296
MC+fixed (SCAD)	**0.2601**	**0.3132**
De-biased	0.3585	0.4537
De-biased+fixed	0.2725	0.3259

Table 3.3: The average ANE on the Cigarette sales data with the block and the staggered structures

We utilize the GDP data [Abadie et al., 2015] for the second dataset. Table 3.4 displays the average ANE with the block and the staggered structures. The SCAD estimator without fixed effects demonstrates superior performance for the block structures, while the de-biased estimator incorporating fixed effects yields the best result for the staggered structures.

Method	Block	Staggered
Fixed effects	1.1384	0.8466
Synthetic	0.8661	0.5714
MC (Nuclear)	7.6874	5.4102
MC+fixed (Nuclear)	0.8696	0.6081
MC (SCAD)	**0.5156**	0.4335
MC+fixed (SCAD)	0.6666	0.5458
De-biased	62.3592	0.5588
De-biased+fixed	0.5687	**0.3116**

Table 3.4: The average ANE on the GDP data with the block and the staggered structures

Chapter 4

Conclusions

In this thesis, we studied causal inference in panel data. We proposed an estimator for recovering potential control outcomes by incorporating nonconvex penalties into the MC for the two treatment adoption scenarios: simultaneous and staggered. Our proposed estimator demonstrated improved convergence rates compared to the existing approach based on the convex penalty and achieved the same convergence rate as the oracle estimator under certain conditions. Moreover, we established the asymptotic normality of the estimator for the treatment effect, with a smaller variance compared to the de-biased method. In our numerical studies, we conducted extensive experiments to evaluate the performance of our estimator and the corresponding estimator for the treatment effects. We empirically demonstrated that our estimator is more robust compared to the synthetic control method when it comes to the selection of control and treatment units in treatment adoption structures. An important direction of future

research is to extend our approach to settings where treatment exposure depends on other observed covariates or prior outcomes.

Bibliography

Alberto Abadie. Semiparametric difference-in-differences estimators. *The Review of Economic Studies*, 72(1):1–19, 2005.

Alberto Abadie. Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, 59(2):391–425, 2021.

Alberto Abadie and Matias D Cattaneo. Econometric methods for program evaluation. *Annual Review of Economics*, 10(1), 2018.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association*, 105(490):493–505, 2010.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Comparative politics and the synthetic control method. *American Journal of Political Science*, 59(2):495–510, 2015.

Alekh Agarwal, Sahand Negahban, and Martin J Wainwright. Fast global convergence rates of gradient methods for high-

dimensional statistical recovery. *Advances in Neural Information Processing Systems*, 23, 2010.

Priya Aggarwal and Anubha Gupta. Accelerated fmri reconstruction using matrix completion with sparse recovery via split bregman. *Neurocomputing*, 216:319–330, 2016.

Dmitry Arkhangelsky, Susan Athey, David A Hirshberg, Guido W Imbens, and Stefan Wager. Synthetic difference-in-differences. *American Economic Review*, 111(12):4088–4118, 2021.

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

Jushan Bai and Serena Ng. Principal components and regularized estimation of factor models. *arXiv preprint arXiv:1708.08137*, 2017.

Arvind Balachandrasekaran, Greg Ongie, and Mathews Jacob. Accelerated dynamic mri using structured low rank matrix completion. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 1858–1862. IEEE, 2016.

Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on optimization*, 20(4):1956–1982, 2010.

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

Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, and Yuling Yan. Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *SIAM journal on optimization*, 30(4):3098–3121, 2020.

Mark A Davenport, Yaniv Plan, Ewout Van Den Berg, and Mary Wootters. 1-bit matrix completion. *Information and Inference: A Journal of the IMA*, 3(3):189–223, 2014.

Nikolay Doudchenko and Guido W Imbens. Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. Technical report, National Bureau of Economic Research, 2016.

Rong Du, Cailian Chen, Bo Yang, and Xinping Guan. Vanet based traffic estimation: A matrix completion approach. In *2013 IEEE Global Communications Conference (GLOBECOM)*, pages 30–35. IEEE, 2013.

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

Vivek Farias, Andrew Li, and Tianyi Peng. Learning treatment effects in panels with general intervention patterns. *Advances in Neural Information Processing Systems*, 34, 2021.

Shuhang Gu, Lei Zhang, Wangmeng Zuo, and Xiangchu Feng. Weighted nuclear norm minimization with application to image denoising. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2862–2869, 2014.

Shuhang Gu, Qi Xie, Deyu Meng, Wangmeng Zuo, Xiangchu Feng, and Lei Zhang. Weighted nuclear norm minimization and its applications to low level vision. *International journal of computer vision*, 121(2):183–208, 2017.

Huan Gui, Jiawei Han, and Quanquan Gu. Towards faster rates and oracle property for low-rank matrix estimation. In *International Conference on Machine Learning*, pages 2300–2309. PMLR, 2016.

Suriya Gunasekar, Pradeep Ravikumar, and Joydeep Ghosh. Exponential family matrix completion under structural constraints. In *International Conference on Machine Learning*, pages 1917–1925. PMLR, 2014.

Nima Hamidi and Mohsen Bayati. On low-rank trace regression under general sampling distribution. *arXiv preprint arXiv:1904.08576*, 2019.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

Yao Hu, Debing Zhang, Jieping Ye, Xuelong Li, and Xiaofei He. Fast and accurate matrix completion via truncated nuclear norm

regularization. *IEEE transactions on pattern analysis and machine intelligence*, 35(9):2117–2130, 2012.

Arnav Kapur, Kshitij Marwah, and Gil Alterovitz. Gene expression prediction using low-rank matrix completion. *BMC bioinformatics*, 17(1):1–13, 2016.

Yongdai Kim, Hosik Choi, and Hee-Seok Oh. Smoothly clipped absolute deviation on high dimensions. *Journal of the American Statistical Association*, 103(484):1665–1673, 2008.

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

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

Jason D Lee, Ben Recht, Nathan Srebro, Joel Tropp, and Russ R Salakhutdinov. Practical large-scale optimization for max-norm regularization. In *Advances in neural information processing systems*, pages 1297–1305, 2010.

Guorui Li, Guang Guo, Sancheng Peng, Cong Wang, Shui Yu, Jianwei Niu, and Jianli Mo. Matrix completion via schatten capped p norm. *IEEE Transactions on Knowledge and Data Engineering*, 2020.

Xiao Peng Li, Lei Huang, Hing Cheung So, and Bo Zhao. A survey

on matrix completion: Perspective of signal processing. *arXiv preprint arXiv:1901.10885*, 2019.

Canyi Lu, Jinhui Tang, Shuicheng Yan, and Zhouchen Lin. Generalized nonconvex nonsmooth low-rank minimization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4130–4137, 2014.

Canyi Lu, Jinhui Tang, Shuicheng Yan, and Zhouchen Lin. Nonconvex nonsmooth low rank minimization via iteratively reweighted nuclear norm. *IEEE Transactions on Image Processing*, 25(2):829–839, 2015a.

Canyi Lu, Changbo Zhu, Chunyan Xu, Shuicheng Yan, and Zhouchen Lin. Generalized singular value thresholding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015b.

Gunjan Mahindre, Anura P Jayasumana, Kelum Gajamannage, and Randy Paffenroth. On sampling and recovery of topology of directed social networks–a low-rank matrix completion based approach. In *2019 IEEE 44th Conference on Local Computer Networks (LCN)*, pages 324–331. IEEE, 2019.

Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of machine learning research*, 11(Aug):2287–2322, 2010.

Rahul Mazumder, Diego Saldana, and Haolei Weng. Matrix com-

pletion with nonconvex regularization: Spectral operators and scalable algorithms. *Statistics and Computing*, pages 1–26, 2020.

Aanchal Mongia, Debarka Sengupta, and Angshul Majumdar. Mcimpute: Matrix completion based imputation for single cell rna-seq data. *Frontiers in genetics*, 10:9, 2019.

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

Sahand Negahban and Martin J Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research*, 13(1): 1665–1697, 2012.

Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical science*, 27(4):538–557, 2012.

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

Azeem M Shaikh and Panos Toulis. Randomization tests in observational studies with staggered adoption of treatment. *Journal of the American Statistical Association*, 116(536):1835–1848, 2021.

Yipeng Song, Johan A Westerhuis, Nanne Aben, Lodewyk FA

Wessels, Patrick JF Groenen, and Age K Smilde. Generalized simultaneous component analysis of binary and quantitative data. *arXiv preprint arXiv:1807.04982*, 2018.

Nathan Srebro and Adi Shraibman. Rank, trace-norm and max-norm. In *International Conference on Computational Learning Theory*, pages 545–560. Springer, 2005.

Gilbert W Stewart. Perturbation theory for the singular value decomposition. Technical report, 1998.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

Yiran Wang. Matrix completion algorithms with applications in biomedicine, e-commerce and social science. 2017.

Zhaoran Wang, Han Liu, and Tong Zhang. Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *Annals of statistics*, 42(6):2164, 2014.

Fei Wen, Lei Chu, Peilin Liu, and Robert C Qiu. A survey on nonconvex regularization-based sparse and low-rank recovery in signal processing, statistics, and machine learning. *IEEE Access*, 6:69883–69906, 2018.

Lin Xiao and Tong Zhang. A proximal-gradient homotopy method for the sparse least-squares problem. *SIAM Journal on Optimization*, 23(2):1062–1091, 2013.

Yiqing Xu. Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis*, 25 (1):57–76, 2017.

Dong Yang, Guisheng Liao, Shengqi Zhu, Xi Yang, and Xuepan Zhang. Sar imaging with undersampled data via matrix completion. *IEEE Geoscience and Remote Sensing Letters*, 11(9): 1539–1543, 2014.

Quanming Yao and James Kwok. Efficient learning with a family of nonconvex regularizers by redistributing nonconvexity. In *International Conference on Machine Learning*, pages 2645–2654. PMLR, 2016.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942, 2010.

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

Cun-Hui Zhang and Tong Zhang. A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, 27(4):576–593, 2012.

Xiaowei Zhou, Can Yang, Hongyu Zhao, and Weichuan Yu. Low-rank modeling and its applications in image analysis. *ACM Computing Surveys (CSUR)*, 47(2):1–33, 2014.

Appendix A

Appendix A.

A.1 Computation of the *PGH* algorithm

In this section, we discuss necessary computations for implementing the *PGH* algorithm in our problem. First, we calculate the closed-form solution of (2.20). By performing simple calculations, and letting $\widetilde{\mathbf{L}} := \mathbf{L}_t^{k-1}$, we can obtain the following:

$$\begin{aligned}
\mathbf{L}_t^k &= \underset{\mathbf{L}\in\mathbb{R}^{N\times T}}{\operatorname{argmin}} \widetilde{F}_{l_t^{k-1},\lambda_t}(\mathbf{L};\widetilde{\mathbf{L}}) \\
&= \underset{\mathbf{L}\in\mathbb{R}^{N\times T}}{\operatorname{argmin}} \frac{l_t^{k-1}}{2}\left\|\mathbf{L}-\left\{\widetilde{\mathbf{L}}-\frac{1}{l_t^{k-1}}\nabla\bar{f}_{n,\lambda}(\widetilde{\mathbf{L}})\right\}\right\|_F^2 + \lambda_t\|\mathbf{L}\|_{tr} \\
&= \underset{\mathbf{L}\in\mathbb{R}^{N\times T}}{\operatorname{argmin}} \frac{1}{2}\left\|\mathbf{L}-\left\{\widetilde{\mathbf{L}}-\frac{1}{l_t^{k-1}}\nabla\bar{f}_{n,\lambda}(\widetilde{\mathbf{L}})\right\}\right\|_F^2 + \frac{\lambda_t}{l_t^{k-1}}\|\mathbf{L}\|_{tr}.
\end{aligned}$$

The matrix approximation problem with the nuclear norm penalty can be achieved using the singular value shrinkage operator defined in Section 2.3. Therefore the solution is obtained as follows:

$$\mathbf{L}_t^k = \mathcal{D}^S_{\frac{\lambda_t}{l_t^{k-1}}}\left(\widetilde{\mathbf{L}}-\frac{1}{l_t^{k-1}}\nabla\bar{f}_{n,\lambda}(\widetilde{\mathbf{L}})\right). \tag{A.1}$$

Recall that $\nabla \bar{f}_{n,\lambda}(\widetilde{\mathbf{L}}) = \nabla f_n(\widetilde{\mathbf{L}}) + \nabla \bar{G}_\lambda(\widetilde{\mathbf{L}})$. We can compute $\nabla f_n(\widetilde{\mathbf{L}}) = -\frac{2}{n}\mathcal{P}_{\mathcal{O}}(\mathbf{Y}(\mathbf{0}) - \widetilde{\mathbf{L}})$ based on the definition of f_n. For $\nabla \bar{G}_\lambda$, if we denote the SVD of $\widetilde{\mathbf{L}}$ as $\mathbf{U}_{\widetilde{\mathbf{L}}} \operatorname{diag}\left(\boldsymbol{\xi}(\widetilde{\mathbf{L}})\right) \mathbf{V}_{\widetilde{\mathbf{L}}}^\top$, then $\nabla \bar{G}(\widetilde{\mathbf{L}}) = \mathbf{U}_{\widetilde{\mathbf{L}}} \operatorname{diag}\left(\bar{g}'_\lambda(\xi_1(\widetilde{\mathbf{L}})), \cdots, \bar{g}'_\lambda(\xi_{\min(N,T)}(\widetilde{\mathbf{L}}))\right) \mathbf{V}_{\widetilde{\mathbf{L}}}^\top$ [Yao and Kwok, 2016, Lemma 21]. Specifically, for the SCAD penalty, we have:

$$\bar{g}'_\lambda(x) = -\frac{|x| - \lambda}{(\gamma - 1)} I\left(\lambda \leq |x| \leq \gamma\lambda\right) - \lambda I\left(|x| > \gamma\lambda\right),$$

and for the MCP penalty, we can derive:

$$\bar{g}'_\lambda(x) = -\frac{|x|}{\gamma} I\left(|x| \leq \gamma\lambda\right) - \lambda I\left(|x| > \gamma\lambda\right).$$

A stopping criterion for the *PGH* algorithm in MC problems is presented in (2.22). However, calculating $\omega_\lambda(\widehat{\mathbf{L}})$, which takes into account $\partial\|\widehat{\mathbf{L}}\|_{tr}$, is not as straightforward as in the case of the ℓ_1-LS problem. Therefore, an alternative stopping criterion, as presented in Xiao and Zhang [2013], can be used. The alternative criterion is as follows:

$$l_t^k \left\| \mathbf{L}_t^{k-1} - \mathbf{L}_t^k \right\|_F \leq \widehat{\epsilon},$$

where $\widehat{\epsilon}$ represents the desired optimization precision (convergence tolerance) in the t-th path of the *PGH* algorithm.

Appendix B

Appendix B.

B.1 Proof of Theorems

B.1.1 Proof of Theorem 3.5.1

We define the error matrix for the observed entries as $\mathfrak{E} = \sum_{(i,t)\in\mathcal{O}} \varepsilon_{it} A_{it}$, where A_{it} refers to $e_i(N)e_t(T)^\top$. To demonstrate the proof of Theorem 3.5.1, we make use of several lemmas. The first lemma is necessary for determining the value of λ and provides an upper bound for $\|\mathfrak{E}\|_{\text{op}}$. This lemma is taken from Athey et al. [2021].

Lemma B.1.1. *There exists a constant C such that*

$$\|\mathfrak{E}\|_{\text{op}} \leq C\sigma \max\left[\sqrt{N\log(N+T)}, \sqrt{T}\log^{3/2}(N+T)\right],$$

with a probability greater than $1-(N+T)^{-2}$.

This result relies on a concentration inequality for the sum of random matrices. Note that the correlation assumption of $\mathcal{O}$ leads to a larger upper bound than that of the independent assumption.

The next lemma plays a crucial role in obtaining the faster convergence rate for the nonconvex penalized estimator by providing an upper bound for $\|\Pi_{\mathcal{F}}(\nabla f_n(\mathbf{L}^*))\|_{op}$. The detailed proof of this lemma can be found in Section B.2.1.

Lemma B.1.2. *For an r-dimensional subspace $\mathcal{F}(\mathbf{U}^*, \mathbf{V}^*)$, there exists a constant C such that*

$$\|\Pi_{\mathcal{F}}(\mathfrak{E})\|_{op} \leq C\sigma\sqrt{T}\log^{\frac{3}{2}}(N+T).$$

This inequality holds with a probability of at least $1-(N+T)^{-2}$.

The next lemma, derived from Athey et al. [2021], is closely related to the Restricted Strong Convexity (RSC) condition [Negahban and Wainwright, 2011, 2012; Negahban et al., 2012] with high probability.

Lemma B.1.3. *If the estimator $\widehat{\mathbf{L}}$ satisfies $\left\|\widehat{\mathbf{L}}-\mathbf{L}^*\right\|_F^2 \geq 4(\alpha^*)^2\theta/p_c$ (or equivalently $\alpha_{sp}(\widehat{\mathbf{L}}-\mathbf{L}^*) \leq \sqrt{\frac{NTp_c}{4\theta}}$) for a positive number θ, then there exists a constant $C \geq 0.001$ such that, when $C\theta > T$, we have*

$$\frac{p_c}{2}\left\|\widehat{\mathbf{L}}-\mathbf{L}^*\right\|_F^2 \leq \sum_{(i,t)\in\mathcal{O}}\left\langle A_{it}, \widehat{\mathbf{L}}-\mathbf{L}^*\right\rangle^2 + 8(\alpha^*)^2T\sqrt{N}$$

with a probability greater than $1-2\exp\left(-\frac{C\theta}{T}\right)$.

First, for $\widehat{\Delta} = \widehat{\mathbf{L}}-\mathbf{L}^*$, it should be noted that the condition $\|\widehat{\Delta}\|_F^2 \geq 4(\alpha^*)^2\theta/p_c$ is equivalent to $\alpha_{sp}(\widehat{\Delta}) \leq \sqrt{NTp_c/4\theta}$ since $\|\widehat{\Delta}\|_{max} = \frac{\alpha_{sp}(\widehat{\Delta})\|\widehat{\Delta}\|_F}{\sqrt{NT}} \leq \alpha^*$ holds. Therefore, the condition can be interpreted as the spikiness of $\widehat{\Delta}$, indicating that the maximum magnitude of the entries of $\widehat{\Delta}$, is limited.

The Restricted Strong Convexity (RSC) condition [Gui et al., 2016; Negahban and Wainwright, 2012] is a useful property that helps control the estimation error by ensuring the strong convexity of loss functions within a given set. The loss function f_n is said to satisfy the RSC condition [Gui et al., 2016, Assumption 3.1] with a positive curvature κ_f (without a tolerance function) if the following inequality holds for any Δ in a constraint set:

$$f_n(\mathbf{L}^* + \Delta) \leq f_n(\mathbf{L}^*) + \langle \nabla f_n(\mathbf{L}^*), \Delta \rangle + \kappa_f \|\Delta\|_F^2.$$

In the MC problems, the parameter relies on the sampling operator, i.e., the random observation process [Hamidi and Bayati, 2019]. The constraint set in this work is defined to control the spikiness of $\widehat{\mathbf{L}} - \mathbf{L}^*$.

If we apply the appropriate $\theta = c_0' T \log(N+T)$ to Lemma B.1.3, it is evident that when condition $\left\|\widehat{\mathbf{L}} - \mathbf{L}^*\right\|_F^2 \geq c_0'(\alpha^*)^2 T \log(N+T)/p_c$ (or equivalently $\alpha_{sp}(\widehat{\Delta}) \leq \frac{1}{c_0}\sqrt{\frac{Np_c}{\log(N+T)}}$), is satisfied, the following inequality holds with a probability greater than $1-(N+T)^{-2}$ for $\widehat{\Delta} = \widehat{\mathbf{L}} - \mathbf{L}^*$:

$$\begin{aligned}
&f_n(\mathbf{L}^* + \widehat{\Delta}) - f_n(\mathbf{L}^*) - \langle \nabla f_n(\mathbf{L}^*), \widehat{\Delta} \rangle \\
=&\frac{1}{n} \sum_{(i,t)\in\mathcal{O}} \left(Y_{it}(0) - \left\langle A_{it}, \mathbf{L}^* + \widehat{\Delta} \right\rangle \right)^2 - \frac{1}{n} \sum_{(i,t)\in\mathcal{O}} (Y_{it}(0) - \langle A_{it}, \mathbf{L}^* \rangle)^2 \\
&+ \frac{2}{n} \sum_{(i,t)\in\mathcal{O}} (Y_{it}(0) - \langle A_{it}, \mathbf{L}^* \rangle) \langle A_{it}, \widehat{\Delta} \rangle \\
=&\frac{1}{n} \sum_{(i,t)\in\mathcal{O}} \left\langle A_{it}, \widehat{\mathbf{L}} - \mathbf{L}^* \right\rangle^2 \\
\geq&\frac{p_c}{2n} \left\|\widehat{\mathbf{L}} - \mathbf{L}^*\right\|_F^2 - \frac{8}{n}(\alpha^*)^2 T\sqrt{N}.
\end{aligned}$$

These equations are derived using the fact that $f_n(\mathbf{L}) = \frac{1}{n}\left\|\mathcal{P}_{\mathcal{O}}(\mathbf{Y}(\mathbf{0}) - \mathbf{L})\right\|_F^2 = \frac{1}{n}\sum_{(i,t)\in\mathcal{O}}\left(Y_{it}(0) - \langle A_{it}, \mathbf{L}\rangle\right)^2$ and $\nabla f_n(\mathbf{L}) = -\frac{2}{n}\sum_{(i,t)\in\mathcal{O}}\left(Y_{it}(0) - \langle A_{it}, \mathbf{L}\rangle\right)A_{it}$. Additionally, if $\|\widehat{\Delta}\|_F^2 \geq 32(\alpha^*)^2 T\sqrt{N}/p_c$, we can establish the inequality:

$$\begin{aligned} f_n(\mathbf{L}^* + \widehat{\Delta}) - f_n(\mathbf{L}^*) - \langle \nabla f_n(\mathbf{L}^*), \widehat{\Delta}\rangle &\geq \frac{p_c}{2n}\|\widehat{\Delta}\|_F^2 - \frac{8}{n}(\alpha^*)^2 T\sqrt{N} \\ &\geq \frac{p_c}{2n}\|\widehat{\Delta}\|_F^2 - \frac{p_c}{4n}\|\widehat{\Delta}\|_F^2 \\ &= \frac{p_c}{4n}\|\widehat{\Delta}\|_F^2, \end{aligned}$$

which holds with a probability greater than $1 - (N+T)^{-2}$. In this case, the loss function f_n satisfies the RSC condition with a positive curvature $\kappa_f = \frac{p_c}{2n}$.

The key to prove Theorem 3.5.1 lies in the next lemma, adapted from Gui et al. [2016], in which the upper bound for the estimation error using nonconvex penalties in the context of trace regression is provided. Note that $\zeta_- = \frac{1}{\gamma - 1}$ for the SCAD penalty and $\zeta_- = \frac{1}{\gamma}$ for the MCP penalty.

Lemma B.1.4 (Adapted from Gui et al. [2016])**.** *Suppose that* $\alpha_{sp}(\widehat{\Delta}) \leq \frac{1}{c_1}\left[\sqrt{\frac{Np_c}{\log(N+T)}} \wedge N^{1/4}\sqrt{p_c}\right]$ *(which is equivalent to* $\|\widehat{\Delta}\|_F^2 \geq c_1'(\alpha^*)^2 T\left[\log(N+T) \vee \sqrt{N}\right]/p_c$*) and the nonconvex penalty satisfies Assumption 3.5.1. Under the condition that* $\kappa_f := \frac{p_c}{2n} > \zeta_-$*, for any optimal solution* $\widehat{\mathbf{L}}$ *of* (3.13) *with a penalty parameter* $\lambda \geq 2\|\mathfrak{E}\|_{op}/n$*, the following inequality holds with a probability greater than* $1 - (N+T)^{-2}$*:*

$$\left\|\widehat{\mathbf{L}} - \mathbf{L}^*\right\|_F \leq \frac{\tau\sqrt{r_1}}{\kappa_f - \zeta_-} + \frac{3\lambda\sqrt{r_2}}{\kappa_f - \zeta_-}. \tag{B.1}$$

Here, $r_1 = |S_1|$ and $r_2 = |S_2|$. The paramter τ is defined as $\tau = \|\Pi_{\mathcal{F}_{S_1}}(\nabla f_n(\mathbf{L}^*))\|_{op} = \|\Pi_{\mathcal{F}_{S_1}}(\frac{2}{n}\mathfrak{E})\|_{op}$, where $\mathcal{F}_{S_1}$ represents a subspace of $\mathcal{F}$ associated with S_1.

Proof of Theorem 3.5.1.
If $\|\widehat{\mathbf{L}} - \mathbf{L}^*\|_F^2 \geq (\alpha^*)^2 \left[C' \log(N+T) \vee C''\sqrt{N}\right] T/p_c$ for appropriate constants C' and C'', Lemma B.1.4 is applicable. By using Lemma B.1.2 and Lemma B.1.1, we obtain the following inequalities:

$$\frac{1}{n}\left\|\Pi_{\mathcal{F}_{S_1}}(\mathfrak{E})\right\|_{op} \leq C_1 \frac{\sigma}{n}\sqrt{T}\log^{\frac{3}{2}}(N+T), \tag{B.2}$$

$$\frac{1}{n}\|\mathfrak{E}\|_{op} \leq C_2 \frac{\sigma}{n}\max\left[\sqrt{N\log(N+T)}, \sqrt{T}\log^{\frac{3}{2}}(N+T)\right]. \tag{B.3}$$

Using equations (B.2) and (B.3) in Lemma B.1.4 with $\zeta_- = \kappa_f/2$, there are positive constants C_1' and C_2' such that:

$$\begin{aligned}\frac{\|\widehat{\Delta}\|_F}{\sqrt{NT}} \leq & C_1'\sigma\sqrt{\frac{r_1\log^3(N+T)}{Np_c^2}} + \\ & C_2'\sigma\left(\sqrt{\frac{r_2\log(N+T)}{Tp_c^2}} \vee \sqrt{\frac{r_2\log^3(N+T)}{Np_c^2}}\right).\end{aligned} \tag{B.4}$$

If $C'\log(N+T) \geq C''\sqrt{N}$ and $\|\widehat{\mathbf{L}} - \mathbf{L}^*\|_F^2 \leq C'(\alpha^*)^2 T\log(N+T)/p_c$ holds, then we have:

$$\frac{\|\widehat{\Delta}\|_F}{\sqrt{NT}} \leq C'\alpha^*\sqrt{\frac{\log(N+T)}{Np_c}}. \tag{B.5}$$

If $C''\sqrt{N} \geq C'\log(N+T)$ and $\|\widehat{\mathbf{L}} - \mathbf{L}^*\|_F^2 \leq C''(\alpha^*)^2 T\sqrt{N}/p_c$ holds, then the following inequality holds:

$$\frac{\|\widehat{\Delta}\|_F}{\sqrt{NT}} \leq C''\alpha^*\sqrt{\frac{1}{\sqrt{N}p_c}}. \tag{B.6}$$

By combining equations (B.4), (B.5) and (B.6), we obtain the upper bound:

$$\frac{\left\|\mathbf{L}^* - \widehat{\mathbf{L}}\right\|_F}{\sqrt{NT}} \leq C\sqrt{\frac{(\alpha^*)^2}{p_{\mathrm{c}}}\left(\frac{1}{\sqrt{N}} \vee \frac{\log(N+T)}{N}\right) \vee \left[\frac{\sigma^2 r_1 \log^3(N+T)}{Np_{\mathrm{c}}^2} + \left(\frac{\sigma^2 r_2 \log(N+T)}{Tp_{\mathrm{c}}^2} \vee \frac{\sigma^2 r_2 \log^3(N+T)}{Np_{\mathrm{c}}^2}\right)\right]}.$$

□

B.1.2 Proof of Theorem 3.5.2

To prove Theorem 3.5.2, we first introduce the lemma, which is demonstrated in Section B.2.2. Recall that $\widehat{\Delta}_O = \widehat{\mathbf{L}}_O - \mathbf{L}^*$.

Lemma B.1.5. *$\mathbf{L}^* \in \mathbb{R}^{N\times T}$ has rank r, and suppose $\alpha_{sp}(\widehat{\Delta}_O) \leq \frac{1}{c_0}\sqrt{\frac{Np_c}{\log(N+T)}}$. Then for a constant C, the following upper bound holds between the oracle estimator $\widehat{\mathbf{L}}_O$ (the solution to equation (3.17)) and the true matrix $\mathbf{L}^*$:*

$$\frac{\left\|\widehat{\mathbf{L}}_O - \mathbf{L}^*\right\|_F}{\sqrt{NT}} \leq C\sqrt{\frac{\sigma^2 r \log^3(N+T)}{Np_c^2} \vee \frac{(\alpha^*)^2}{\sqrt{N}p_c}}$$

with a probability greater than $1-(N+T)^{-2}$.

The proof of Theorem 3.5.2 is derived based on the proof of Gui et al. [2016, Theorem 3.5].

Proof of Theorem 3.5.2.

We aim to demonstrate that the oracle estimator satisfies the first-order optimality condition of optimization problem (3.13), which

states that there exists some $\widehat{\mathbf{W}}_O \in \partial \left\| \widehat{\mathbf{L}}_O \right\|_{tr}$ such that, for any $\mathbf{L}' \in \mathbb{R}^{N \times T}$, the following holds:

$$\max_{\mathbf{L}'} \left\langle \widehat{\mathbf{L}}_O - \mathbf{L}', \nabla \bar{f}_{n,\lambda} \left(\widehat{\mathbf{L}}_O \right) + \lambda \widehat{\mathbf{W}}_O \right\rangle \leq 0. \tag{B.7}$$

We project the left-hand side of equation (B.7) onto the subspaces $\mathcal{F}$ and $\mathcal{T}^{\perp}$, resulting in the following:

$$\begin{aligned} &\left\langle \widehat{\mathbf{L}}_O - \mathbf{L}', \nabla \bar{f}_{n,\lambda} \left(\widehat{\mathbf{L}}_O \right) + \lambda \widehat{\mathbf{W}}_O \right\rangle \\ =& \underbrace{\left\langle \mathbf{\Pi}_{\mathcal{F}} \left(\widehat{\mathbf{L}}_O - \mathbf{L}' \right), \nabla \bar{f}_{n,\lambda} \left(\widehat{\mathbf{L}}_O \right) + \lambda \widehat{\mathbf{W}}_O \right\rangle}_{\mathcal{I}_1} + \\ & \underbrace{\left\langle \mathbf{\Pi}_{\mathcal{T}^{\perp}} \left(\widehat{\mathbf{L}}_O - \mathbf{L}' \right), \nabla \bar{f}_{n,\lambda} \left(\widehat{\mathbf{L}}_O \right) + \lambda \widehat{\mathbf{W}}_O \right\rangle}_{\mathcal{I}_2}. \end{aligned} \tag{B.8}$$

To analyze the term $\mathcal{I}_1$, let $\boldsymbol{\xi}^*$ be the vector of the singular values of $\mathbf{L}^*$ and let $\widehat{\boldsymbol{\xi}}_O = \boldsymbol{\xi} \left(\widehat{\mathbf{L}}_O \right)$ be the corresponding vector for $\widehat{\mathbf{L}}_O$. According to the perturbation bounds for singular values (Weyl's inequality) [Stewart, 1998, Theorem 1], we have the following inequality:

$$\max_i \left| (\boldsymbol{\xi}^*)_i - \left(\widehat{\boldsymbol{\xi}}_O \right)_i \right| \leq \left\| \mathbf{L}^* - \widehat{\mathbf{L}}_O \right\|_{op} \leq \left\| \mathbf{L}^* - \widehat{\mathbf{L}}_O \right\|_F.$$

By applying Lemma B.1.5, we obtain the inequality:

$$\max_i \left| (\boldsymbol{\xi}^*)_i - \left(\widehat{\boldsymbol{\xi}}_O \right)_i \right| \leq C_1 \sqrt{\frac{\sigma^2 r T \log^3(N+T)}{p_c^2} \vee \frac{(\alpha^*)^2 T \sqrt{N}}{p_c}}.$$

Using that $S = \operatorname{supp}(\boldsymbol{\xi}^*)$ with $|S| = r$ and applying the triangular

inequality, we have:

$$\begin{aligned}
\min_{i\in S}\left|\left(\widehat{\boldsymbol{\xi}}_O\right)_i\right| =& \min_{i\in S}\left|\left(\widehat{\boldsymbol{\xi}}_O\right)_i - (\boldsymbol{\xi}^*)_i + (\boldsymbol{\xi}^*)_i\right| \\
\geq& -\max_{i\in S}\left|\left(\widehat{\boldsymbol{\xi}}_O - \boldsymbol{\xi}^*\right)_i\right| + \min_{i\in S}|(\boldsymbol{\xi}^*)_i| \\
\geq& -C_1\sqrt{\frac{\sigma^2 rT\log^3(N+T)}{p_c^2}\vee\frac{(\alpha^*)^2T\sqrt{N}}{p_c}} + \gamma\lambda + \\
& C_1\sqrt{\frac{\sigma^2 rT\log^3(N+T)}{p_c^2}\vee\frac{(\alpha^*)^2T\sqrt{N}}{p_c}} \\
=& \gamma\lambda.
\end{aligned}$$

The second inequality is based on the assumption of the theorem. Since $\operatorname{rank}\left(\widehat{\mathbf{L}}_O\right) = r$ and $\widehat{\mathbf{L}}_O \in \mathcal{F}$, we have

$$\left(\widehat{\boldsymbol{\xi}}_O\right)_1 \geq \ldots \geq \left(\widehat{\boldsymbol{\xi}}_O\right)_r \geq \gamma\lambda > \left(\widehat{\boldsymbol{\xi}}_O\right)_{r+1} = \cdots = \left(\widehat{\boldsymbol{\xi}}_O\right)_{\min(N,T)} = 0.$$

The compact SVD of the oracle estimator is $\widehat{\mathbf{L}}_O = \mathbf{U}^* \operatorname{diag}\left((\widehat{\boldsymbol{\xi}}_O)_1, \cdots, (\widehat{\boldsymbol{\xi}}_O)_r\right) \mathbf{V}^{*\top}$. Recall that $G_\lambda(\mathbf{L}) = \bar{G}_\lambda(\mathbf{L}) + \lambda\|\mathbf{L}\|_{tr}$. Thus,

$$\begin{aligned}
&\boldsymbol{\Pi}_{\mathcal{F}}\left(\nabla G_\lambda\left(\widehat{\mathbf{L}}_O\right)\right) = \boldsymbol{\Pi}_{\mathcal{F}}\left(\nabla\bar{G}_\lambda\left(\widehat{\mathbf{L}}_O\right) + \lambda\partial\left\|\widehat{\mathbf{L}}_O\right\|_{tr}\right) \\
&= \boldsymbol{\Pi}_{\mathcal{F}}\left(\mathbf{U}^*\operatorname{diag}\left(\bar{g}'_\lambda((\widehat{\boldsymbol{\xi}}_O)_1), \cdots, \bar{g}'_\lambda((\widehat{\boldsymbol{\xi}}_O)_r)\right)\mathbf{V}^{*\top} + \lambda\mathbf{U}^*\mathbf{V}^{*\top} + \lambda\widehat{\mathbf{Z}}_O\right) \\
&= \mathbf{U}^*\left(\operatorname{diag}\left(\bar{g}'_\lambda((\widehat{\boldsymbol{\xi}}_O)_1), \cdots, \bar{g}'_\lambda((\widehat{\boldsymbol{\xi}}_O)_r)\right) + \lambda I_r\right)\mathbf{V}^{*\top},
\end{aligned}$$

where $\widehat{\mathbf{Z}}_O \in \mathcal{T}^\perp$ and $\left\|\widehat{\mathbf{Z}}_O\right\|_{op} \leq 1$. The second equality follows from $\nabla\bar{G}_\lambda(\mathbf{L}) = \mathbf{U}^*\operatorname{diag}\left(\bar{g}'_\lambda((\widehat{\boldsymbol{\xi}}_O)_1), \cdots, \bar{g}'_\lambda((\widehat{\boldsymbol{\xi}}_O)_r)\right)\mathbf{V}^{*\top}$ [Yao and Kwok, 2016, Lemma 21] and the definition of $\partial\|\cdot\|_{tr}$, and the last equality holds by projecting each component onto $\mathcal{F}$.

Now for $g_\lambda(t) = \bar{g}_\lambda(t) + \lambda|t|$ and $g'_\lambda(t) = \bar{g}'_\lambda(t) + \lambda t$ for all $t > 0$, the i-th element of the diagonal matrix $\operatorname{diag}\left(\bar{g}'_\lambda((\widehat{\boldsymbol{\xi}}_O)_1), \cdots, \bar{g}'_\lambda((\widehat{\boldsymbol{\xi}}_O)_r)\right) +$

λI_r is

$$\bar{g}'_\lambda\left(\left(\widehat{\boldsymbol{\xi}}_O\right)_i\right)+\lambda = g'_\lambda\left(\left(\widehat{\boldsymbol{\xi}}_O\right)_i\right).$$

Also, since $g'_\lambda(t)=0$ for all $t \geq \gamma\lambda$, we have $g'_\lambda\left(\left(\widehat{\boldsymbol{\xi}}_O\right)_i\right)=0$ for $i \in S$. Therefore, this results in

$$\boldsymbol{\Pi}_{\mathcal{F}}\left(\nabla G_\lambda\left(\widehat{\mathbf{L}}_O\right)\right)=\mathbf{0}. \tag{B.9}$$

Furthermore, by the optimality condition, for all $\mathbf{L}' \in \mathbb{R}^{N\times T}$,

$$\max_{\mathbf{L}'}\left\langle \boldsymbol{\Pi}_{\mathcal{F}}\left(\widehat{\mathbf{L}}_O-\mathbf{L}'\right), \nabla f_n\left(\widehat{\mathbf{L}}_O\right)\right\rangle \leq 0. \tag{B.10}$$

Using equations (B.9) and (B.10), we have the following for $\mathcal{I}_1$:

$$\begin{aligned}
&\max_{\mathbf{L}'}\left\langle \boldsymbol{\Pi}_{\mathcal{F}}\left(\widehat{\mathbf{L}}_O-\mathbf{L}'\right), \nabla \bar{f}_{n,\lambda}\left(\widehat{\mathbf{L}}_O\right)+\lambda\widehat{\mathbf{W}}_O\right\rangle \\
&\quad=\max_{\mathbf{L}'}\left\langle \boldsymbol{\Pi}_{\mathcal{F}}\left(\widehat{\mathbf{L}}_O-\mathbf{L}'\right), \nabla f_n\left(\widehat{\mathbf{L}}_O\right)\right\rangle + \\
&\qquad\max_{\mathbf{L}'}\left\langle \boldsymbol{\Pi}_{\mathcal{F}}\left(\widehat{\mathbf{L}}_O-\mathbf{L}'\right), \boldsymbol{\Pi}_{\mathcal{F}}\left(\nabla G_\lambda\left(\widehat{\mathbf{L}}_O\right)\right)\right\rangle \\
&\quad\leq 0.
\end{aligned} \tag{B.11}$$

To analyze the term $\mathcal{I}_2$, let's consider the SVD of $\nabla\bar{G}_\lambda(\widehat{\mathbf{L}}_O)$ as $\nabla\bar{G}_\lambda(\widehat{\mathbf{L}}_O) = \mathbf{U}^*\operatorname{diag}\left(\bar{g}'_\lambda((\widehat{\boldsymbol{\xi}}_O)_1),\cdots,\bar{g}'_\lambda((\widehat{\boldsymbol{\xi}}_O)_r)\right)\mathbf{V}^{*\top}$. When we project $\nabla\bar{G}_\lambda\left(\widehat{\mathbf{L}}_O\right)$ onto the subspace $\mathcal{T}^\perp$, we have:

$$\begin{aligned}
&\boldsymbol{\Pi}_{\mathcal{T}^\perp}\left(\nabla\bar{G}_\lambda\left(\widehat{\mathbf{L}}_O\right)\right) \\
&=\left(I_N-\mathbf{U}^*\mathbf{U}^{*\top}\right)\mathbf{U}^*\operatorname{diag}\left(\bar{g}'_\lambda((\widehat{\boldsymbol{\xi}}_O)_1),\cdots,\bar{g}'_\lambda((\widehat{\boldsymbol{\xi}}_O)_r)\right)\mathbf{V}^{*\top}\left(I_T-\mathbf{V}^*\mathbf{V}^{*\top}\right) \\
&=\left(\mathbf{U}^*-\mathbf{U}^*\right)\operatorname{diag}\left(\bar{g}'_\lambda((\widehat{\boldsymbol{\xi}}_O)_1),\cdots,\bar{g}'_\lambda((\widehat{\boldsymbol{\xi}}_O)_r)\right)\left(\mathbf{V}^{*\top}-\mathbf{V}^{*\top}\right) \\
&=\mathbf{0}.
\end{aligned}$$

If the condition $\left\|\widehat{\mathbf{L}}_O-\mathbf{L}^*\right\|_F^2 \geq c_0(\alpha^*)^2T\log(N+T)/p_c$ is satisfied,

then we can show that:

$$\left\|\nabla f_n(\mathbf{L}^*) - \nabla f_n\left(\mathbf{L}^* + \widehat{\mathbf{\Delta}}_O\right)\right\|_F \leq C_2 \frac{p_c}{T\log(N+T)} \left\|\widehat{\mathbf{\Delta}}_O\right\|_F. \tag{B.12}$$

Since $\widehat{\mathbf{L}}_{it} - \mathbf{L}^*_{it} \leq 2\alpha^*$, we have:

$$\frac{1}{n}\sum_{(i,t)\in\mathcal{O}} \langle A_{it}, \widehat{\mathbf{L}}_O - \mathbf{L}^*\rangle^2 = \frac{1}{n}\sum_{(i,t)\in\mathcal{O}} \left((\widehat{\mathbf{L}}_O)_{it} - \mathbf{L}^*_{it}\right)^2 \leq 4(\alpha^*)^2.$$

Using this, we can calculate:

$$\left\|\widehat{\Delta}_O\right\|_F^2 \geq c_0 \frac{1}{4n} \sum_{(i,t)\in\mathcal{O}} \langle A_{it}, \widehat{\mathbf{L}}_O - \mathbf{L}^*\rangle^2 T\log(N+T)/p_c$$

and thus

$$\frac{1}{n}\sum_{(i,t)\in\mathcal{O}} \langle A_{it}, \widehat{\mathbf{L}}_O - \mathbf{L}^*\rangle^2 \leq \frac{4}{c_0}\frac{p_c}{T\log(N+T)} \|\widehat{\Delta}\|_F^2.$$

Note that $f_n(\mathbf{L}^*+\widehat{\Delta}_O) - f_n(\mathbf{L}^*) - \langle \nabla f_n(\mathbf{L}^*), \widehat{\Delta}_O\rangle = \frac{1}{n}\sum_{(i,t)\in\mathcal{O}} \langle A_{it}, \widehat{\mathbf{L}}_O - \mathbf{L}^*\rangle$. Therefore f_n satisfies the strong smoothness condition with a parameter of $C_2 \frac{p_c}{T\log(N+T)}$. By the equivalent condition of the strong smoothness, equation (B.12) holds.

Applying the triangular inequality, we have:

$$\begin{aligned}
\left\|\nabla f_n\left(\widehat{\mathbf{L}}_O\right)\right\|_{op} &\leq \left\|\nabla f_n\left(\mathbf{L}^*\right)\right\|_{op} + \left\|\nabla f_n\left(\mathbf{L}^*\right) - \nabla f_n\left(\widehat{\mathbf{L}}_O\right)\right\|_{op} \\
&\leq \left\|\nabla f_n\left(\mathbf{L}^*\right)\right\|_{op} + \left\|\nabla f_n\left(\mathbf{L}^*\right) - \nabla f_n\left(\widehat{\mathbf{L}}_O\right)\right\|_{F} \\
&\leq \left\|\nabla f_n\left(\mathbf{L}^*\right)\right\|_{op} + C_2 \frac{p_c}{T\log(N+T)} \left\|\widehat{\mathbf{\Delta}}_O\right\|_F,
\end{aligned} \tag{B.13}$$

and by applying the inequality from Lemma B.1.5 to (B.13), we

obtain:

$$
\begin{aligned}
&\left\|\mathbf{\Pi}_{\mathcal{T}^\perp}\left(\nabla f_n\left(\widehat{\mathbf{L}}_O\right)\right)\right\|_{op} \\
&\leq \left\|\nabla f_n\left(\widehat{\mathbf{L}}_O\right)\right\|_{op} \leq \left\|\nabla f_n\left(\mathbf{L}^*\right)\right\|_{op} + C_2\frac{p_c}{T\log(N+T)}\left\|\widehat{\mathbf{\Delta}}_O\right\|_F \\
&\leq \frac{2}{n}\|\mathfrak{E}\|_{op} + C_2\frac{p_c}{T\log(N+T)}C_1\sqrt{\frac{\sigma^2 rT\log^3(N+T)}{Np_c^2} \vee \frac{(\alpha^*)^2T\sqrt{N}}{p_c}} \\
&\leq C_3\frac{\sigma\left[\sqrt{N\log(N+T)} \vee \sqrt{T}\log^{3/2}(N+T)\right]}{n} \\
&\quad + C_4\sqrt{\frac{\sigma^2 r\log(N+T)}{T} \vee \frac{p_c(\alpha^*)^2\sqrt{N}}{T\log^2(N+T)}} \\
&\leq \lambda.
\end{aligned}
$$

The last inequality holds based on the assumption regarding the magnitude of λ. By defining $\widehat{\mathbf{Z}}_O = -\lambda^{-1}\mathbf{\Pi}_{\mathcal{T}^\perp}\left(\nabla f_n\left(\widehat{\mathbf{L}}_O\right)\right)$ and using the fact that that $\widehat{\mathbf{Z}}_O \in \mathcal{T}^\perp, \left\|\widehat{\mathbf{Z}}_O\right\|_{op} \leq 1$, we have $\widehat{\mathbf{W}}_O = \mathbf{U}^*\mathbf{V}^{*\top} + \widehat{\mathbf{Z}}_O \in \partial\left\|\widehat{\mathbf{L}}_O\right\|_{tr}$. Furthermore,

$$
\begin{aligned}
&\mathbf{\Pi}_{\mathcal{T}^\perp}\left(\nabla f_n\left(\widehat{\mathbf{L}}_O\right) + \lambda\widehat{\mathbf{W}}_O\right) \\
&\quad = \mathbf{\Pi}_{\mathcal{T}^\perp}\left(\nabla f_n\left(\widehat{\mathbf{L}}_O\right) + \lambda\left(\mathbf{U}^*\mathbf{V}^{*\top} + \widehat{\mathbf{Z}}_O\right)\right) \\
&\quad = \mathbf{\Pi}_{\mathcal{T}^\perp}\left(\nabla f_n\left(\widehat{\mathbf{L}}_O\right)\right) - \lambda\frac{1}{\lambda}\mathbf{\Pi}_{\mathcal{T}^\perp}\left(\nabla f_n\left(\widehat{\mathbf{L}}_O\right)\right) \\
&\quad = \mathbf{0}.
\end{aligned}
$$

Now we calculate $\mathcal{I}_2$:

$$
\begin{aligned}
\mathcal{I}_2 &= \left\langle \mathbf{\Pi}_{\mathcal{T}^\perp}\left(\widehat{\mathbf{L}}_O - \mathbf{L}'\right), \nabla\bar{f}_{n,\lambda}\left(\widehat{\mathbf{L}}_O\right) + \lambda\widehat{\mathbf{W}}_O\right\rangle \\
&= \left\langle \mathbf{\Pi}_{\mathcal{T}^\perp}\left(\widehat{\mathbf{L}}_O - \mathbf{L}'\right), \nabla f_n\left(\widehat{\mathbf{L}}_O\right) + \lambda\widehat{\mathbf{W}}_O\right\rangle \\
&\qquad + \left\langle \mathbf{\Pi}_{\mathcal{T}^\perp}\left(\widehat{\mathbf{L}}_O - \mathbf{L}'\right), \nabla\bar{G}_\lambda\left(\widehat{\mathbf{L}}_O\right)\right\rangle \\
&= 0.
\end{aligned}
\tag{B.14}
$$

We can derive (B.7) by combining equations (B.11), (B.14) and (B.8). □

B.1.3 Proof of Theorem 3.5.3

Proof. If the inequality $\frac{\left\|\mathbf{L}^* - \widehat{\mathbf{L}}_O\right\|_F}{\sqrt{NT}} \leq c_0 \sqrt{\frac{(\alpha^*)^2 \log(N+T)}{N p_{\mathrm{c}}}}$ holds, then the following inequality holds as well:

$$\frac{\left\|\mathbf{L}^* - \widehat{\mathbf{L}}_O\right\|_F}{\sqrt{NT}} \leq C_1 \sqrt{\frac{(\alpha^*)^2 \log(N+T)}{N p_{\mathrm{c}}}}.$$

If the above inequality does not hold, we can use Lemma B.1.5 to derive the following inequality:

$$\frac{\left\|\mathbf{L}^* - \widehat{\mathbf{L}}_O\right\|_F}{\sqrt{NT}} \leq C_2 \sqrt{\frac{(\alpha^*)^2}{\sqrt{N} p_{\mathrm{c}}} \vee \frac{\sigma^2 r \log^3(N+T)}{N p_{\mathrm{c}}^2}}$$

with a probability greater than $1 - (N+T)^{-2}$.

By considering both cases, we complete the proof. □

B.1.4 Proof of Theorem 3.5.4

Let $n_t = NT - n$ be the number of treated observations. We can easily show that (3.11) is equivalent to:

$$\widehat{\theta} = \underset{\theta \in \mathbb{R}}{\operatorname{argmin}} \frac{1}{n_t} \left\| \mathcal{P}_{\mathcal{O}^{\mathrm{c}}} \left(\mathbf{Y}(1) - \widehat{\mathbf{Y}}(0) - \theta \mathbf{W} \right) \right\|_F^2. \tag{B.15}$$

By utilizing the equation (B.15), we can prove the theorem. The proof is inspired by Zhang and Zhang [2014] and Farias et al. [2021].

Proof. The first order condition of (B.15) is:

$$\langle \mathbf{W}, \mathcal{P}_{\mathcal{O}^{\mathrm{c}}} \left(\mathbf{Y}(\mathbf{1}) - \widehat{\mathbf{L}}_O - \widehat{\theta} \mathbf{W} \right) \rangle = 0.$$

Using the definition of $\mathbf{Y}(\mathbf{1})$, this can be rewritten as:

$$\langle \mathbf{W}, \mathcal{P}_{\mathcal{O}^c}\left(\mathbf{L}^* - \widehat{\mathbf{L}}_O\right)\rangle + \langle \mathbf{W}, \mathcal{P}_{\mathcal{O}^c}\left(\boldsymbol{\epsilon} + \delta \circ \mathbf{W}\right)\rangle + \left(\theta^* - \widehat{\theta}\right)\|\mathbf{W}\|_F^2 = 0.$$

Since $\langle \mathbf{W}, \mathcal{P}_{\mathcal{O}^c}(\mathbf{A})\rangle = \langle \mathbf{W}, \mathbf{A}\rangle$ for any matrix $\mathbf{A}$, based on the definition of $\mathbf{W}$ and the projection operator $\mathcal{P}$, we have:

$$(\widehat{\theta} - \theta^*) = \frac{\langle \mathbf{W}, \mathbf{L}^* - \widehat{\mathbf{L}}_O\rangle}{\|\mathbf{W}\|_F^2} + \frac{\langle \mathbf{W}, \boldsymbol{\epsilon} + \delta \circ \mathbf{W}\rangle}{\|\mathbf{W}\|_F^2}.$$

Let $\delta' = \frac{\langle \mathbf{W}, \mathbf{L}^* - \widehat{\mathbf{L}}_O\rangle}{\|\mathbf{W}\|_F^2}$. Note that for any matrices $\mathbf{A}$ and $\mathbf{B}$, we have

$$\begin{aligned} |\langle \mathbf{A}, \mathbf{B}\rangle| &\leq \|\mathbf{A}\|_{op}\|\mathbf{B}\|_{tr} \\ &\leq \|\mathbf{A}\|_{op}\sqrt{r}\|\mathbf{B}\|_F \\ &\leq \|\mathbf{A}\|_F\sqrt{r}\|\mathbf{B}\|_F. \end{aligned}$$

The first inequality holds due to the matrix Hölder inequality. Therefore,

$$\begin{aligned} |\delta'| &= \frac{|\langle \mathbf{W}, \mathbf{L}^* - \widehat{\mathbf{L}}_O\rangle|}{\|\mathbf{W}\|_F^2} \\ &\leq \frac{\sqrt{r}\left\|\mathbf{L}^* - \widehat{\mathbf{L}}_O\right\|_F \|\mathbf{W}\|_F}{\|\mathbf{W}\|_F^2} \\ &\leq \frac{\sqrt{r}\left\|\mathbf{L}^* - \widehat{\mathbf{L}}_O\right\|_F}{\sqrt{n_t}}. \end{aligned}$$

The last inequality holds because $\|\mathbf{W}\|_F^2 = n_t$ by the difinition of $\mathbf{W}$. Now, according to Theorem 3.5.3,

$$|\delta'| \lesssim \sqrt{\frac{(\alpha^*)^2 r\sqrt{N}T}{n_t p_{\mathrm{c}}} \vee \frac{(\alpha^*)^2 rT\log(N+T)}{n_t p_{\mathrm{c}}} \vee \frac{\sigma^2 r^2 T\log^3(N+T)}{n_t p_{\mathrm{c}}^2}}.$$

If $r = \alpha^* = \sigma = \mathcal{O}(1)$ and $n_t = \Omega(NT)$, then (3.19) holds. Additionally, if $p_c \gg \frac{1}{\sqrt{N}}\log^{\frac{3}{2}}(N+T)$, we can establish the asymptotic normality using the central limit theorem (CLT). □

B.1.5 Proof of Theorem 3.5.5

The second step for estimating the causal effect, as described in (3.8), is derived from the following lemma [Farias et al., 2021, Lemma 1]. The proof of this lemma is omitted.

Lemma B.1.6. *Suppose* $(\hat{\mathbf{L}}^{(init)}, \hat{\theta}^{(init)})$ *be a minimizer of* (3.7). *We denote the SVD of* $\hat{\mathbf{L}}^{(init)}$ *as* $\hat{\mathbf{L}}^{(init)} = \hat{\mathbf{U}}\widehat{\Xi}\hat{\mathbf{V}}^\top$, *and let* $\hat{\mathcal{T}}^\perp = \mathcal{T}^\perp(\widehat{\mathbf{U}}, \widehat{\mathbf{V}})$. *We define* $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{\varepsilon} + \delta \circ \mathbf{W}$. *Then, the following equation holds:*

$$\begin{aligned}\left(\hat{\theta} - \theta^*\right)&\|\Pi_{\hat{\mathcal{T}}^\perp}(\mathbf{W})\|_F^2 \\ &= \lambda\left\langle \mathbf{W}, \hat{\mathbf{U}}\hat{\mathbf{V}}^\top \right\rangle + \left\langle \Pi_{\hat{\mathcal{T}}^\perp}(\mathbf{W}), \hat{\boldsymbol{\varepsilon}} \right\rangle + \left\langle \mathbf{W}, \Pi_{\hat{\mathcal{T}}^\perp}(\mathbf{L}^*) \right\rangle.\end{aligned}$$

In the remaining part of this section, we introduce the non-convex proxy problem and establish its connection with the convex problem discussed in Farias et al. [2021]. We then describe the implication of Assumption 3.3.1, which is derived from this relationship. Finally, we utilize it in the proof of the theorem.

We consider the following original convex problem:

$$\min_{\mathbf{L}\in\mathbb{R}^{N\times T}, \theta\in\mathbb{R}} h(\mathbf{L}, \theta) := \frac{1}{2}\|\mathbf{Y} - \mathbf{L} - \theta\mathbf{W}\|_F^2 + \lambda\|\mathbf{L}\|_{tr}$$

and the non-convex proxy problem:

$$\begin{aligned}\min_{\mathbf{A}\in\mathbb{R}^{N\times r}, \mathbf{B}\in\mathbb{R}^{T\times r}, \theta\in\mathbb{R}} &\tilde{h}(\mathbf{A}, \mathbf{B}; \theta) \\ &:= \frac{1}{2}\left\|\mathbf{Y} - \mathbf{A}\mathbf{B}^T - \theta\mathbf{W}\right\|_F^2 + \frac{\lambda}{2}\|\mathbf{A}\|_F^2 + \frac{\lambda}{2}\|\mathbf{B}\|_F^2.\end{aligned} \tag{B.16}$$

It is well known that minimizing $\tilde{h}$ is equivalent to solving $\min h(\mathbf{L}; \theta)$ s.t. $\mathrm{rank}(\mathbf{L}) \leq r$. We hope that for the solution $\mathbf{A}$

and $\mathbf{B}$ of (B.16), $\mathbf{A} \approx \mathbf{A}^*$ and $\mathbf{B} \approx \mathbf{B}^*$ hold, where $\mathbf{L}^* = \mathbf{A}^*\mathbf{B}^{*\top}$ with $\mathbf{A}^* = \mathbf{U}^*\Xi^{*1/2}, \mathbf{B}^* = \mathbf{V}^*\Xi^{*1/2}$. Farias et al. [2021, Lemma 3] showed that the critical points of $\tilde{h}$ approximately satisfy the first-order condition of h, so that $(\mathbf{A}\mathbf{B}^\top, \theta)$ is close to the optimizer of h. Note that Assumption 3.3.1 is required for the establishment of Farias et al. [2021, Lemma 3].

The next lemma [Farias et al., 2021, Lemma 16] presents the implication of Assumption 3.3.1 (a).

Lemma B.1.7. *Suppose* $\mathbf{A}^*, \mathbf{A} \in \mathbb{R}^{N\times r}$ *and* $\mathbf{B}^*, \mathbf{B} \in \mathbb{R}^{T\times r}$. *Let* $\mathbf{A}^*\mathbf{B}^{*\top} = \mathbf{U}^*\Xi^*\mathbf{V}^{*\top}, \mathbf{A}\mathbf{B}^\top = \mathbf{U}\Xi\mathbf{V}^\top$ *be the SVD of* $\mathbf{A}^*\mathbf{B}^{*\top}, \mathbf{A}\mathbf{B}^\top$, *respectively. Assume* $\mathbf{A}^* = \mathbf{U}^*\Xi^{*1/2}$, $\mathbf{B}^* = \mathbf{V}^*\Xi^{*1/2}$, *and* $\chi = \xi_1(\Xi^*)/\xi_r(\Xi^*)$. *Suppose* $\frac{\sigma\sqrt{N\wedge T}}{\xi_{\min}} \le C_2 \frac{1}{\chi^2 r^2 \log^5(N\wedge T)}$, *and*

$$\|\mathbf{A}-\mathbf{A}^*\|_{\mathrm{F}} + \|\mathbf{B}-\mathbf{B}^*\|_{\mathrm{F}} \le C_{\mathrm{F}} \frac{\sigma\sqrt{N\wedge T}\log^{2.5}(N\wedge T)\sqrt{\xi_{\max} r}}{\xi_{\min}}. \tag{B.17}$$

Let $\mathcal{T}^{*\perp}$ *be the orthogonal space related to* $\mathbf{A}^*\mathbf{B}^{*\top}$ *and* $\mathcal{T}^\perp$ *be the orthogonal space related to* $\mathbf{A}\mathbf{B}^\top$. *If there exists a constant* C_{r_1} *such that* $\|\mathbf{W}\mathbf{V}^*\|_{\mathrm{F}}^2 + \left\|\mathbf{W}^\top\mathbf{U}^*\right\|_{\mathrm{F}}^2 \le \left(1 - \frac{C_{r_1}}{\log(N\wedge T)}\right)\|\mathbf{W}\|_{\mathrm{F}}^2$, *then for large enough* N *and* T, *the following holds:*

$$\|\mathbf{W}\mathbf{V}\|_{\mathrm{F}}^2 + \left\|\mathbf{W}^\top\mathbf{U}\right\|_{\mathrm{F}}^2 \le \left(1 - \frac{C_{r_1}}{2\log(N\wedge T)}\right)\|\mathbf{W}\|_{\mathrm{F}}^2$$
$$\|\Pi_{\mathcal{T}^\perp}(\mathbf{W})\|_{\mathrm{F}}^2 \ge \frac{C_{r_1}}{2\log(N\wedge T)}\|\mathbf{W}\|_{\mathrm{F}}^2.$$

Please refer to the definition (3.15) for the orthogonal spaces. Since (B.17) approximately holds according to Farias et al. [2021, Lemma 3], we can directly use the result of the lemma to prove the

theorem. Note that a similar result for the case of the de-biased estimator for Lasso appears in Zhang and Zhang [2014].

Proof of Theorem 3.5.5. By Lemma B.1.6, we have:

$$\widehat{\theta}^d - \theta^* = \frac{\langle \Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{W}), \epsilon + \delta \circ \mathbf{W} \rangle}{\|\Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{W})\|_F^2} + \frac{\langle \mathbf{W}, \Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{L}^*) \rangle}{\|\Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{W})\|_F^2}.$$

Let $\delta' = \frac{\langle \mathbf{W}, \Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{L}^*) \rangle}{\|\Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{W})\|_F^2}$. Since $|\langle \mathbf{A}, \mathbf{B} \rangle| \leq \|\mathbf{A}\|_{op} \|\mathbf{B}\|_{tr} = \sqrt{r}\|\mathbf{A}\|_F \|\mathbf{B}\|_F$ for any matrices $\mathbf{A}$ and $\mathbf{B}$, and

$$\langle \mathbf{W}, \Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{L}^*) \rangle = \langle \mathbf{W}, \Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{L}^* - \widehat{\mathbf{L}}^{(init)}) \rangle = \langle \Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{W}), \mathbf{L}^* - \widehat{\mathbf{L}}^{(init)} \rangle,$$

we can obtain:

$$\left|\delta'\right| \leq \frac{\sqrt{r}\|\Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{W})\|_F \left\|\mathbf{L}^* - \widehat{\mathbf{L}}^{(init)}\right\|_F}{\|\Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{W})\|_F^2} = \frac{\sqrt{r}\left\|\mathbf{L}^* - \widehat{\mathbf{L}}^{(init)}\right\|_F}{\|\Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{W})\|_F}.$$

According to Lemma B.1.7, we have $\frac{1}{\|\Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{W})\|_F} \leq C\frac{\sqrt{\log(N \wedge T)}}{\|\mathbf{W}\|_F}$. Hence,

$$\begin{aligned}\left|\delta'\right| &\lesssim \frac{\sqrt{r\log(N \wedge T)}\left\|\mathbf{L}^* - \widehat{\mathbf{L}}^{(init)}\right\|_F}{\|\mathbf{W}\|_F} \\ &= \frac{\sqrt{r\log(N \wedge T)}\left\|\mathbf{L}^* - \widehat{\mathbf{L}}^{(init)}\right\|_F}{\sqrt{n_t}}.\end{aligned}$$

If $n_t = \Omega(NT)$ and $r = O(1)$, (3.21) holds. Furthermore, if $\frac{\sqrt{\log(N \wedge T)}\left\|\mathbf{L}^* - \widehat{\mathbf{L}}^{(init)}\right\|_F}{\sqrt{NT}} \to 0$ as $N \to \infty$, then the asymptotic normality can be established using the central limit theorem (CLT).

□

B.2 Proof of Lemmas and Proposition

B.2.1 Proof of Lemma B.1.2

First, we introduce the Bernstein inequality of rectangular matrices, as presented by Tropp [2012].

Proposition B.2.1 (Matrix Bernstein Inequality)**.** *Let* $\mathbf{Z}_1, \ldots, \mathbf{Z}_N$ *be independent matrices in* $\mathbb{R}^{d_1 \times d_2}$ *such that* $\mathbb{E}[\mathbf{Z}_i] = \mathbf{0}$ *and* $\|\mathbf{Z}_i\|_{\text{op}} \leq D$ *almost surely for all* $i \in [N]$*. Let* σ_Z *be a parameter such that*

$$\sigma_Z^2 \geq \max\left\{ \left\| \sum_{i=1}^{N} \mathbb{E}\left[\mathbf{Z}_i \mathbf{Z}_i^\top\right] \right\|_{\text{op}}, \left\| \sum_{i=1}^{N} \mathbb{E}\left[\mathbf{Z}_i^\top \mathbf{Z}_i\right] \right\|_{\text{op}} \right\}.$$

For any $\alpha \geq 0$*, the following inequality holds:*

$$P\left\{ \left\| \sum_{i=1}^{N} \mathbf{Z}_i \right\|_{\text{op}} \geq \alpha \right\} \leq (d_1 + d_2) \exp\left[\frac{-\alpha^2}{2\sigma_Z^2 + (2D\alpha)/3} \right].$$

To prove Lemma B.1.2, we referenced the proof process used in Athey et al. [2021, Lemma 2] and Gui et al. [2016, Lemma E.3.]. We denote the projection operator onto the subspace $\mathcal{F}$ as $\Pi_{\mathcal{F}}(\cdot)$ and it follows that $\Pi_{\mathcal{F}}(\mathbf{A}) = \mathbf{U}^*\mathbf{U}^{*\top}\mathbf{A}\mathbf{V}^*\mathbf{V}^{*\top}$ [Gui et al., 2016].

Proof of Lemma B.1.2.

Since $\mathbf{U}^*$ and $\mathbf{V}^*$ are matrices corresponding to the left and right singular vectors of $\mathbf{L}^*$, for $\mathcal{S} = \mathcal{F}(\mathbf{U}^*, \mathbf{V}^*)$, we have the following relation:

$$\begin{aligned} \|\mathbf{\Pi}_{\mathcal{S}}(\mathfrak{E})\|_{op} &= \left\| \mathbf{U}^*\mathbf{U}^{*\top}\mathfrak{E}\mathbf{V}^*\mathbf{V}^{*\top} \right\|_{op} \\ &= \left\| \mathbf{U}^{*\top}\mathfrak{E}\mathbf{V}^* \right\|_{op}. \end{aligned}$$

We start by introducing the definitions. For every $i \in [N]$, we define $\mathbf{B}_i$ as follows:

$$\mathbf{B}_i = \mathbf{U}^{*\top} \left(\sum_{t=1}^{t_i} \varepsilon_{it} A_{it} \right) \mathbf{V}^* = \mathbf{U}^{*\top} \mathfrak{E} \mathbf{V}^*.$$

It should be noted that based on the definition of $\mathfrak{E}$, $\mathbf{U}^{*\top} \mathfrak{E} \mathbf{V}^* = \sum_{i=1}^{N} \mathbf{B}_i$ and $\mathbb{E}[\mathbf{B}_i] = \mathbf{0}$ for all $i \in [N]$. We set the bound D to be $D \equiv C_2 \sigma \sqrt{\log(N+T)}$, where C_2 is a sufficiently large constant. For each $(i,t) \in \mathcal{O}$, we define $\bar{\varepsilon}_{it} = \varepsilon_{it} I(|\varepsilon_{it}| \leq D)$ and $\overline{\mathbf{B}}_i = \mathbf{U}^{*\top} \left\{ \sum_{t=1}^{t_i} \bar{\varepsilon}_{it} A_{it} \right\} \mathbf{V}^*$ for all $i \in [N]$.

By applying the union bound and utilizing the property that ε_{it} is σ-sub-Gaussian random variable ($P(|\varepsilon_{it}| \geq t) \leq 2 \exp\left\{-t^2/\left(2\sigma^2\right)\right\}$), we can derive the following inequality for any $\alpha \geq 0$:

$$\begin{aligned} P\left(\|\mathbf{U}^{*\top} \mathfrak{E} \mathbf{V}^*\|_{\text{op}} \geq \alpha \right) &\leq P\left(\left\| \sum_{i=1}^{N} \overline{\mathbf{B}}_i \right\|_{\text{op}} \geq \alpha \right) + 2NT \exp\left(\frac{-D^2}{2\sigma^2} \right) \\ &\leq P\left(\left\| \sum_{i=1}^{N} \overline{\mathbf{B}}_i \right\|_{\text{op}} \geq \alpha \right) + \frac{1}{(N+T)^3}. \end{aligned} \tag{B.18}$$

Let us define $\mathbf{Z}_i$ as $\overline{\mathbf{B}}_i - \mathbb{E}\left[\overline{\mathbf{B}}_i\right]$ for each $i \in [N]$. Then, the following holds:

$$\left\| \sum_{i=1}^{N} \overline{\mathbf{B}}_i \right\|_{\text{op}} \leq \left\| \sum_{i=1}^{N} \mathbf{Z}_i \right\|_{\text{op}} + \left\| \mathbb{E}\left[\sum_{i=1}^{N} \overline{\mathbf{B}}_i \right] \right\|_{\text{op}}.$$

Since the mean of ε_{it} is 0, we can establish the following:

$$\begin{aligned}
|\mathbb{E}\left[\bar{\varepsilon}_{it}\right]| &= |\mathbb{E}\left[\varepsilon_{it} I(|\varepsilon_{it}| \leq D)\right]| = |\mathbb{E}\left[\varepsilon_{it} I(|\varepsilon_{it}| \geq D)\right]| \\
&\leq \sqrt{\mathbb{E}\left[\varepsilon_{it}^2\right] P\left(|\varepsilon_{it}| \geq D\right)} \\
&\leq \sqrt{2\sigma^2 \exp\left\{-D^2/\left(2\sigma^2\right)\right\}} \\
&\leq \frac{\sigma}{(N+T)^4}.
\end{aligned}$$

Given that $\mathbb{E}\left[\sum_{i=1}^{N} \overline{\mathbf{B}}_i\right] = \mathbf{U}^{*\top} \mathbb{E}\left[\sum_{i=1}^{N} \sum_{t=1}^{t_i} \bar{\varepsilon}_{it} A_{it}\right] \mathbf{V}^*$, we have

$$\begin{aligned}
\left\|\mathbb{E}\left[\sum_{i=1}^{N} \overline{\mathbf{B}}_i\right]\right\|_{\max} &= \left\|\mathbf{U}^{*\top}\right\|_{op} \left\|\sum_{i=1}^{N} \sum_{t=1}^{t_i} \bar{\varepsilon}_{it} A_{it}\right\|_{op} \left\|\mathbf{V}^*\right\|_{op} \\
&\leq \sqrt{NT} \left\|\sum_{i=1}^{N} \sum_{t=1}^{t_i} \overline{\varepsilon_{it}} A_{it}\right\|_{max} \\
&\leq \frac{\sigma\sqrt{NT}}{(N+T)^4} \leq \frac{\sigma}{(N+T)^3}.
\end{aligned}$$

Thus, we can establish the following result:

$$\left\|\sum_{i=1}^{N} \overline{\mathbf{B}}_i\right\|_{\text{op}} \leq \left\|\sum_{i=1}^{N} \mathbf{Z}_i\right\|_{\text{op}} + \frac{\sigma}{(N+T)^3}. \tag{B.19}$$

To find σ_Z in Proposition B.2.1, we start by examining $\mathbf{Z}_i \mathbf{Z}_i^\top$:

$$\begin{aligned}
\mathbf{Z}_i \mathbf{Z}_i^\top =& \mathbf{U}^{*\top} e_i(N) \sum_{t}^{t_i} \left(\bar{\varepsilon}_{it} - \mathbb{E}(\bar{\varepsilon}_{it})\right) e_t(T)^\top \mathbf{V}^* \mathbf{V}^{*\top} \\
& \sum_{t}^{t_i} \left(\bar{\varepsilon}_{it} - \mathbb{E}(\bar{\varepsilon}_{it})\right) e_t(T) e_i(N)^\top \mathbf{U}^*.
\end{aligned}$$

We denote $\mathbf{u}_i^*$ as the i-th row of $\mathbf{U}^*$ and $\mathbf{v}_i^*$ as the i-th row of $\mathbf{V}^*$. Using this notation, we can show that $\mathbf{U}^{*\top} e_i(N) = \mathbf{u}_i^*$ and $\mathbf{V}^{*\top} \left\{\sum_{t=1}^{t_i} \left(\bar{\varepsilon}_{it} - \mathbb{E}(\bar{\varepsilon}_{it})\right) e_t(T)\right\} = \sum_{t=1}^{t_i} \left(\bar{\varepsilon}_{it} - \mathbb{E}(\bar{\varepsilon}_{it})\right) \mathbf{v}_t^*$.

Since ε_{it} 's are independent, we have:

$$\mathbb{E}\left[\left\{\sum_{t=1}^{t_i}\left(\bar{\varepsilon}_{it}-\mathbb{E}(\bar{\varepsilon}_{it})e_t(T)^\top\right)\right\}\mathbf{V}^*\mathbf{V}^{*\top}\left\{\sum_{t=1}^{t_i}\left(\bar{\varepsilon}_{it}-\mathbb{E}(\bar{\varepsilon}_{it})\right)e_t(T)\right\}\right]$$
$$=\mathbb{E}\left[\sum_{t=1}^{t_i}\left(\bar{\varepsilon}_{it}-\mathbb{E}(\bar{\varepsilon}_{it})\right)^2\left\|\mathbf{v}_t^*\right\|_2^2\right].$$

The following also holds:

$$\sum_{i=1}^{N}\mathbb{E}\left[\mathbf{Z}_i\mathbf{Z}_i^\top\right]=\mathbb{E}\left[\sum_{i=1}^{N}\mathbf{u}_i^*\left\{\sum_{t=1}^{t_i}\left(\bar{\varepsilon}_{it}-\mathbb{E}(\bar{\varepsilon}_{it})\right)^2\left\|\mathbf{v}_t^*\right\|_2^2\right\}\mathbf{u}_i^{*\top}\right]$$
$$=\mathbb{E}\left[\sum_{i=1}^{N}\mathbf{u}_i^*\mathbf{u}_i^{*\top}\left\{\sum_{t=1}^{t_i}\left(\bar{\varepsilon}_{it}-\mathbb{E}(\bar{\varepsilon}_{it})\right)^2\left\|\mathbf{v}_t^*\right\|_2^2\right\}\right].$$

Now we establish that

$$\left\|\sum_{i=1}^{N}\mathbb{E}\left[\mathbf{Z}_i\mathbf{Z}_i^\top\right]\right\|_{\text{op}}$$
$$\leq\max_{(i,t)\in\mathcal{O}}\left\{\mathbb{E}\left[\left(\bar{\varepsilon}_{it}-E\left[\bar{\varepsilon}_{it}\right]\right)^2\right]\right\}\left\|\sum_{i=1}^{N}\mathbb{E}\left[\mathbf{u}_i^*\mathbf{u}_i^{*\top}\left(\sum_{t=1}^{t_i}\left\|\mathbf{v}_t^*\right\|_2^2\right)\right]\right\|_{\text{op}}$$
$$\leq 2\sigma^2\left(\sum_{t=1}^{T}\left\|\mathbf{v}_t^*\right\|_2^2\right)\left\|\sum_{i=1}^{N}\mathbf{u}_i^*\mathbf{u}_i^{*\top}\right\|_{\text{op}}$$
$$=2\sigma^2 r,$$

where the inequality holds because the random variable $\bar{\varepsilon}_{it}-\mathbb{E}\left[\bar{\varepsilon}_{it}\right]$ is centered and σ-sub-Gaussian. The equality holds since $\sum_{t=1}^{T}\left\|\mathbf{v}_t^*\right\|_2^2=r$ and $\sum_{i=1}^{N}\mathbf{u}_i^*\mathbf{u}_i^{*\top}=I_r$.

Similarly, we can also examine $\mathbf{Z}_i^\top\mathbf{Z}_i$:

$$\mathbf{Z}_i^\top\mathbf{Z}_i=\mathbf{V}^{*\top}\sum_{t}^{t_i}\left(\bar{\varepsilon}_{it}-\mathbb{E}(\bar{\varepsilon}_{it})\right)e_t(T)e_i(N)^\top\mathbf{U}^*\mathbf{U}^{*\top}$$
$$e_i(N)\sum_{t}^{t_i}\left(\bar{\varepsilon}_{it}-\mathbb{E}(\bar{\varepsilon}_{it})\right)e_t(T)^\top\mathbf{V}^*.$$

Using $\mathbf{U}^{*\top} e_i(N) = \mathbf{u}_i^*$ and

$$\mathbb{E}\left[\mathbf{V}^{*\top}\left\{\sum_{t=1}^{t_i}\left(\bar{\varepsilon}_{it} - \mathbb{E}(\bar{\varepsilon}_{it})e_t(T)\right)\right\}\left\{\sum_{t=1}^{t_i}\left(\bar{\varepsilon}_{it} - \mathbb{E}(\bar{\varepsilon}_{it})\right)e_t(T)^\top\right\}\mathbf{V}^*\right]$$
$$= \mathbb{E}\left\{\sum_{t=1}^{t_i}\left(\bar{\varepsilon}_{it} - \mathbb{E}(\bar{\varepsilon}_{it})\right)^2 \mathbf{v}_t^*\mathbf{v}_t^{*\top}\right\},$$

we obtain

$$\sum_{i=1}^{N}\mathbb{E}\left[\mathbf{Z}_i^\top\mathbf{Z}_i\right] = \mathbb{E}\left[\sum_{i=1}^{N}\|\mathbf{u}_i^*\|_2^2\left\{\sum_{t=1}^{t_i}\left(\bar{\varepsilon}_{it} - \mathbb{E}(\bar{\varepsilon}_{it})\right)^2\mathbf{v}_t^*\mathbf{v}_t^{*\top}\right\}\right].$$

Consequently, we prove the following:

$$\begin{aligned}
&\left\|\sum_{i=1}^{N}\mathbb{E}\left[\mathbf{Z}_i^\top\mathbf{Z}_i\right]\right\|_{\text{op}} \\
&\leq \max_{(i,t)\in\mathcal{O}}\left\{\mathbb{E}\left[\left(\bar{\varepsilon}_{it} - E\left[\bar{\varepsilon}_{it}\right]\right)^2\right]\right\}\left\|\sum_{i=1}^{N}\|\mathbf{u}_i^*\|_2^2\,\mathbb{E}\left[\sum_{t=1}^{t_i}\mathbf{v}_t^*\mathbf{v}_t^{*\top}\right]\right\|_{\text{op}} \\
&\leq 2\sigma^2\sum_{i=1}^{N}\|\mathbf{u}_i^*\|_2^2\ \left\|\sum_{t=1}^{T}\mathbf{v}_t^*\mathbf{v}_t^{*\top}\right\|_{op} \\
&= 2\sigma^2 r
\end{aligned}$$

where the second inequality holds because the spectral norm is monotone, and the equality holds from $\sum_{i=1}^{N}\|\mathbf{u}_i^*\|_2^2 = r$ and $\sum_{t=1}^{T}\mathbf{v}_t^*\mathbf{v}_t^{*\top} = I_r$. Therefore, we set $\sigma_Z^2 = 2\sigma^2 r$.

Furthermore, since $\left\|\overline{\mathbf{B}}_i\right\|_{op} \leq \left\|\sum_{i=1}^{N}\sum_{t=1}^{t_i}\bar{\varepsilon}_{it}A_{it}\right\|_{op} \leq D\sqrt{T}$ and $\left\|\mathbb{E}[\overline{\mathbf{B}}_i]\right\|_{op} \leq D\sqrt{T}$ for all $i \in [N]$, we have $\|\mathbf{Z}_i\|_{\text{op}} \leq 2D\sqrt{T}$. Finally, by applying Proposition B.2.1, we obtain the following inequality:

$$\begin{aligned}
P\left\{\left\|\sum_{i=1}^{N}\mathbf{Z}_i\right\|_{\text{op}} \geq \alpha\right\} &\leq (N+T)\exp\left[-\frac{\alpha^2}{4\sigma^2 r + (4D\alpha\sqrt{T})/3}\right] \\
&\leq (N+T)\exp\left[-\frac{3}{16}\min\left(\frac{\alpha^2}{\sigma^2 r}, \frac{\alpha}{D\sqrt{T}}\right)\right].
\end{aligned}$$

Consequently, there exists a constant C_3 with a probability greater than $1 - \exp(-t)$ such that:

$$\left\| \sum_{i=1}^{N} \mathbf{Z}_i \right\|_{\text{op}}$$
$$\leq C_3 \sigma \max(\sqrt{r[t + \log(N+T)]}, \sqrt{T \log(N+T)}[t + \log(N+T)]).$$

By choosing a sufficiently large constant $C \log(N+T)$ for a constant C to t in the above equation, and using (B.18) and (B.19), we can obtain the following inequality with a probability greater than $1 - 2(N+T)^{-3}$ for a constant C_1:

$$\begin{aligned} \left\| \mathbf{U}^{*\top} \mathfrak{E} \mathbf{V}^* \right\|_{\text{op}} &\leq C_1 \sigma \max \left[\sqrt{r \log(N+T)}, \sqrt{T} \log^{3/2}(N+T) \right] \\ &= C_1 \sigma \sqrt{T} \log^{3/2}(N+T) \end{aligned}$$

since $r \leq T$ and $\sqrt{\log(N+T)} < \log^{3/2}(N+T)$. □

B.2.2 Proof of Lemma B.1.5

The lemma establishes the deterministic upper bound of the oracle estimator under the spikiness condition of $\widehat{\Delta}_O = \widehat{\mathbf{L}}_O - \mathbf{L}^*$. The proof is inspired by Gui et al. [2016, Lemma D.3].

Proof of Lemma B.1.5. Let $\widehat{\Delta}_O = \widehat{\mathbf{L}}_O - \mathbf{L}^*$. According to the observation model $Y_{it}(0) = \langle A_{it}, \mathbf{L}^* \rangle + \epsilon_{it}$ for $(i,t) \in \mathcal{O}$, we can

express the difference between $f_n\left(\widehat{\mathbf{L}}_O\right)$ and $f_n\left(\mathbf{L}^*\right)$ as follows:

$$
\begin{aligned}
& f_n\left(\widehat{\mathbf{L}}_O\right) - f_n\left(\mathbf{L}^*\right) \\
&= \frac{1}{n}\sum_{(i,t)\in\mathcal{O}}\left(Y_{it}(0) - \langle A_{it}, \widehat{\mathbf{L}}_O\rangle\right)^2 - \frac{1}{n}\sum_{(i,t)\in\mathcal{O}}\left(Y_{it}(0) - \langle A_{it}, \mathbf{L}^*\rangle\right)^2 \\
&= \frac{1}{n}\sum_{(i,t)\in\mathcal{O}}\left(Y_{it}(0) - \langle A_{it}, \mathbf{L}^* + \widehat{\mathbf{\Delta}}_O\rangle\right)^2 - \frac{1}{n}\sum_{(i,t)\in\mathcal{O}}\left(Y_{it}(0) - \langle A_{it}, \mathbf{L}^*\rangle\right)^2 \\
&= \frac{1}{n}\sum_{(i,t)\in\mathcal{O}}\left(\epsilon_{it} - \langle A_{it}, \widehat{\mathbf{\Delta}}_O\rangle\right)^2 - \frac{1}{n}\sum_{(i,t)\in\mathcal{O}}\epsilon_{it}^2 \\
&= \frac{1}{n}\sum_{(i,t)\in\mathcal{O}}\left(\epsilon_{it}^2 - 2\langle A_{it}, \widehat{\mathbf{\Delta}}_O\rangle\epsilon_{it} + \langle A_{it}, \widehat{\mathbf{\Delta}}_O\rangle^2 - \epsilon_{it}^2\right) \\
&= \frac{1}{n}\sum_{(i,t)\in\mathcal{O}}\langle A_{it}, \widehat{\mathbf{\Delta}}_O\rangle^2 - \frac{2}{n}\sum_{(i,t)\in\mathcal{O}}\langle \epsilon_{it}A_{it}, \widehat{\mathbf{\Delta}}_O\rangle \\
&= \frac{1}{n}\sum_{(i,t)\in\mathcal{O}}\langle A_{it}, \widehat{\mathbf{\Delta}}_O\rangle^2 - \frac{2}{n}\langle \mathfrak{E}, \widehat{\mathbf{\Delta}}_O\rangle,
\end{aligned}
$$

where $\mathfrak{E} = \sum_{(i,t)\in\mathcal{O}}\epsilon_{it}A_{it}$ and the third equality holds because $Y_{it}(0) = \langle A_{it}, \mathbf{L}^* + \widehat{\mathbf{\Delta}}_O\rangle = \epsilon_{it} - \langle A_{it}, \widehat{\mathbf{\Delta}}_O\rangle$. Since the oracle estimator $\widehat{\mathbf{L}}_O$ minimizes $f_n(\cdot)$ over the subspace $\mathcal{F}$, and $\mathbf{L}^* \in \mathcal{F}$, we have $f_n\left(\widehat{\mathbf{L}}_O\right) \leq f_n\left(\mathbf{L}^*\right)$, which yields:

$$
\frac{1}{n}\sum_{(i,t)\in\mathcal{O}}\langle A_{it}, \widehat{\mathbf{\Delta}}_O\rangle^2 \leq \frac{2}{n}\langle \mathfrak{E}, \widehat{\mathbf{\Delta}}_O\rangle. \tag{B.20}
$$

If $\alpha_{sp}(\widehat{\Delta}_O) \leq \frac{1}{c_0}\sqrt{\frac{Np_c}{\log(N+T)}}$ holds, then we establish the following inequality:

$$
\sum_{(i,t)\in\mathcal{O}}\left\langle A_{it}, \widehat{\mathbf{L}}_O - \mathbf{L}^*\right\rangle^2 \geq \frac{p_c}{2}\left\|\widehat{\mathbf{L}}_O - \mathbf{L}^*\right\|_F^2 - 8(\alpha^*)^2T\sqrt{N}. \tag{B.21}
$$

This inequality is derived from Lemma B.1.3, and it holds with a probability greater than $1 - (N+T)^{-2}$.

By substituting the inequality (B.20) into (B.21), we obtain:

$$\frac{p_c}{2n}\left\|\widehat{\mathbf{L}}_O - \mathbf{L}^*\right\|_F^2 - 8(\alpha^*)^2 T\sqrt{N}/n \leq \frac{2}{n}\langle \mathfrak{E}, \widehat{\mathbf{\Delta}}_O\rangle.$$

Then, we can derive the following inequality:

$$\begin{aligned}
\left\|\widehat{\mathbf{\Delta}}_O\right\|_F^2 &\leq \frac{4}{p_c}\langle \mathfrak{E}, \widehat{\mathbf{\Delta}}_O\rangle + \frac{16}{p_c}(\alpha^*)^2 T\sqrt{N} \\
&= \frac{4}{p_c}\langle \Pi_{\mathcal{F}}(\mathfrak{E}), \widehat{\mathbf{\Delta}}_O\rangle + \frac{16}{p_c}(\alpha^*)^2 T\sqrt{N} \\
&\leq \frac{4}{p_c}\left\|\Pi_{\mathcal{F}}(\mathfrak{E})\right\|_{op}\left\|\widehat{\mathbf{\Delta}}_O\right\|_{tr} + \frac{16}{p_c}(\alpha^*)^2 T\sqrt{N} \\
&\leq \frac{4\sqrt{r}}{p_c}\left\|\Pi_{\mathcal{F}}(\mathfrak{E})\right\|_{op}\left\|\widehat{\mathbf{\Delta}}_O\right\|_F + \frac{16}{p_c}(\alpha^*)^2 T\sqrt{N}.
\end{aligned}$$

Using the inequality $2ab \leq a^2 + b^2$, we have

$$\left\|\widehat{\mathbf{\Delta}}_O\right\|_F^2 \leq \frac{8r}{p_c^2}\left\|\Pi_{\mathcal{F}}(\mathfrak{E})\right\|_{op}^2 + \frac{1}{2}\left\|\widehat{\mathbf{\Delta}}_O\right\|_F^2 + \frac{16}{p_c}(\alpha^*)^2 T\sqrt{N},$$

and then

$$\frac{1}{2}\left\|\widehat{\mathbf{\Delta}}_O\right\|_F^2 \leq \frac{8r}{p_c^2}\left\|\Pi_{\mathcal{F}}(\mathfrak{E})\right\|_{op}^2 + \frac{16}{p_c}(\alpha^*)^2 T\sqrt{N}.$$

Then, applying Lemma B.1.2, we obtain the bound:

$$\frac{\left\|\widehat{\mathbf{\Delta}}_O\right\|_F}{\sqrt{NT}} \leq \sqrt{C_1\frac{\sigma^2 r\log^3(N+T)}{Np_c^2} + C_2\frac{(\alpha^*)^2}{\sqrt{N}p_c}},$$

which completes the proof. □

B.2.3 Proof of Proposition 3.5.6

Proof. Since ϵ_{it} and δ_{it} are independent of each other, we have:

$$\begin{aligned}
&V_{d,\theta} - V_\theta \\
&= \sum_{i,t} \Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{W})_{it}^2 \operatorname{Var}(\epsilon_{it}) \Bigg/ \left(\sum_{i,t} \Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{W})_{it}^2 \right)^2 \\
&\quad - \sum_{i,t} W_{it}^2 \operatorname{Var}(\epsilon_{it}) \Bigg/ \left(\sum_{i,t} W_{it}^2 \right)^2 \\
&\quad + \sum_{i,t} \Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{W})_{it}^2 \operatorname{Var}(\delta_{it} W_{it}) \Bigg/ \left(\sum_{i,t} \Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{W})_{it}^2 \right)^2 \\
&\quad - \sum_{i,t} W_{it}^2 \operatorname{Var}(\delta_{it} W_{it}) \Bigg/ \left(\sum_{i,t} W_{it}^2 \right)^2 .
\end{aligned} \tag{B.22}$$

For the first and second terms in (B.22), the following inequality holds since $\|\Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{W})\|_F^2 \le \|\mathbf{W}\|_F^2$:

$$\begin{aligned}
&\sum_{i,t} \Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{W})_{it}^2 \operatorname{Var}(\epsilon_{it}) \Bigg/ \left(\sum_{i,t} \Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{W})_{it}^2 \right)^2 \\
&\qquad - \sum_{i,t} W_{it}^2 \operatorname{Var}(\epsilon_{it}) \Bigg/ \left(\sum_{i,t} W_{it}^2 \right)^2 \\
&= \operatorname{Var}(\epsilon_{it}) \left\{ \frac{1}{\sum_{(i,t)} \Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{W})_{it}^2} - \frac{1}{\sum_{(i,t)} W_{it}^2} \right\} \ge 0.
\end{aligned} \tag{B.23}$$

For the last two terms in (B.22), note that $\mathbf{W}$ is not random

and $W_{it} = W_{it}^2 = W_{it}^3$. We have:

$$\frac{\sum_{i,t} \Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{W})_{it}^2 \operatorname{Var}(\delta_{it} W_{it})}{\left(\sum_{i,t} \Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{W})_{it}^2\right)^2} - \frac{\sum_{i,t} W_{it}^2 \operatorname{Var}(\delta_{it} W_{it})}{\left(\sum_{i,t} W_{it}^2\right)^2}$$

$$= \operatorname{Var}(\delta_{it}) \left\{ \frac{\sum_{i,t} \Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{W})_{it}^2 W_{it}}{\left(\sum_{i,t} \Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{W})_{it}^2\right)^2} - \frac{\sum_{i,t} W_{it}^3}{\left(\sum_{i,t} W_{it}^2\right)^2} \right\}$$

$$= \operatorname{Var}(\delta_{it}) \times$$

$$\frac{\left(\sum_{i,t} \Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{W})_{it}^2 W_{it}\right)\left(\sum_{i,t} W_{it}^2\right)^2 - \left(\sum_{i,t} W_{it}^3\right)\left(\sum_{i,t} \Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{W})_{it}^2\right)^2}{\left(\sum_{i,t} \Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{W})_{it}^2\right)^2 \left(\sum_{i,t} W_{it}^2\right)^2}$$

$$= \operatorname{Var}(\delta_{it}) \frac{\left(\sum_{i,t} \Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{W})_{it}^2 W_{it}\right)\left(\sum_{i,t} W_{it}^2\right) - \left(\sum_{i,t} \Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{W})_{it}^2\right)^2}{\left(\sum_{i,t} \Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{W})_{it}^2\right)^2 \left(\sum_{i,t} W_{it}^2\right)}$$

$$\geq \operatorname{Var}(\delta_{it}) \frac{\left(\sum_{i,t} \Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{W})_{it} W_{it}^2\right)^2 - \left(\sum_{i,t} \Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{W})_{it}^2\right)^2}{\left(\sum_{i,t} \Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{W})_{it}^2\right)^2 \left(\sum_{i,t} W_{it}^2\right)}$$

$$\geq \operatorname{Var}(\delta_{it}) \times$$

$$\frac{\left(\sum_{i,t} \Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{W})_{it} W_{it} - \sum_{i,t} \Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{W})_{it}^2\right)\left(\sum_{i,t} \Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{W})_{it} W_{it} + \sum_{i,t} \Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{W})_{it}^2\right)}{\left(\sum_{i,t} \Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{W})_{it}^2\right)^2 \left(\sum_{i,t} W_{it}^2\right)}$$

$$= 0. \text{(B.24)}$$

The first inequality holds by the Cauchy-Schwarz inequality, and the last equality is derived from

$$\sum_{i,t} \Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{W})_{it} \left(W_{it} - \Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{W})_{it}\right) = \left\langle \Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{W}), \mathbf{W} - \Pi_{\widehat{\mathcal{T}}^\perp}(\mathbf{W}) \right\rangle = 0.$$

Therefore, based on (B.23) and (B.24), we conclude that $V_{d,\theta} - V_\theta \geq 0$. □

Appendix C

Appendix C.

C.1 Comparison with the *Synthetic control* method

To begin, we provide a brief overview of the synthetic control method [Abadie et al., 2010; Athey et al., 2021]. We define $\mathcal{O}_c^U$ as the set of control units and $\mathcal{O}_t^U$ as the set of treated units. It is important to note that $|\mathcal{O}_c^U| = n_c$ and $|\mathcal{O}_t^U| = n_t$. In our treatment adoption scenarios, the number of treated units n_t is always greater than or equal to 1. Following the approach of Athey et al. [2021], the estimation of weights in the synthetic control method in our experiments relies solely on the use of $\mathbf{Y}(0)$ without considering any additional information. For each unit $i \in \mathcal{O}_t^U$ and for years $t > T_0$, we estimate the counterfactual outcome as

$$\widehat{Y}_{it}(0) = \sum_{j \in \mathcal{O}_c^U} \widehat{w}_j^{(i)} Y_{jt}(0),$$

where $\widehat{\mathbf{w}}^{(i)} = (\widehat{w}_1^{(i)}, \cdots, \widehat{w}_{n_c}^{(i)})^\top$ is chosen as follows:

$$\widehat{\mathbf{w}}^{(i)} = \underset{\mathbf{w}}{\operatorname{argmin}} \sum_{h=1}^{T_0} v_h \left(Y_{ih}(0) - \sum_{j \in \mathcal{O}_c^U} w_j^{(i)} Y_{jh}(0) \right)^2 \tag{C.1}$$

for the block structures. The positive constants v_h for $h = 1, \cdots, T_0$ represent the predictive power of each predictor on $\widehat{Y}_{ih}(0)$, and can be determined using data-driven methods. In the case of the staggered structures, where t_i denotes the treatment adoption point ($t_i \geq T_0$) for unit i in $\mathcal{O}_t^U$, the weight is selected as

$$\widehat{\mathbf{w}}^{(i)} = \underset{\mathbf{w}}{\operatorname{argmin}} \sum_{h=1}^{t_i} v_h \left(Y_{ih}(0) - \sum_{j \in \mathcal{O}_c^U} w_j^{(i)} Y_{jh}(0) \right)^2. \tag{C.2}$$

It is worth noting that settings with a small T_0 and a large n_c may introduce a significant risk of overfitting, as stated in Abadie et al. [2010]. Empirical evidence from Athey et al. [2021] demonstrates that the synthetic control method's performance deteriorates as T_0/T becomes smaller compared to the MC methods.

We examine situations where the results of the synthetic control are inferior to those of the MC methods. To illustrate this, we utilize experimental results with the Cigarette sales data from Section 3.6.1. Figure C.1 presents time-series plots displaying the observed cigarette sales for a specific repetition in the block structures. The blue lines represent the time-series plots for the control states, and the red lines represent the time series for the treated states. In this analysis, we used 27 states as control units, excluding the other 11 states shown in the figure. The treatment adoption date is indicated by the dashed vertical line.

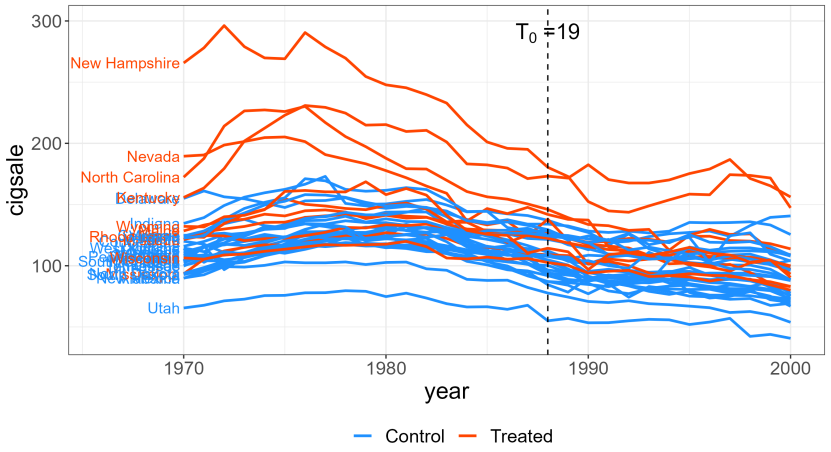

Figure C.1: Time series of the observed cigarette sales for a specific repetition in the block structures.

In Figure C.2, we show the observed and estimated outputs for treated states in the block structures, taken from the corresponding repetition in our analysis of real data. The closer the estimated lines align with the observed data, the better the performance. The figure illustrates that the results of the synthetic control for Kentucky, New Hampshire, and North Carolina are poor both before and after the intervention, whereas the MC methods exhibit relatively better patterns. These states have observed outcomes with extreme values compared to other states as shown in Figure C.1. This aligns with the findings in Abadie et al. [2010, Section 3.4], which states that the outcomes of states with extreme values during the pre-intervention period cannot be accurately replicated as a convex combination of other states' cigarette sales. Among the 38 states, it was observed that the synthetic control performs well

when the treated units do not differ significantly in characteristics (e.g., Maine, Ohio, and Wisconsin). Synthetic control models assume that patterns across units are stable over time, while low-rank models incorporate patterns both across units and over time allowing for interactions between them. Particularly, the MC with fixed effects can capture patterns that the traditional MC cannot. Additionally, the performance enhancement of the MC may be attributed to its ability to utilize additional observations (the values of treatment units during pre-intervention periods) [Athey et al., 2021].

Figure C.3 presents time-series plots of the observed cigarette sales for a specific repetition for the staggered structures. The time T_0 is indicated by the dashed vertical line. Recall that the treatment adaptation dates for each state were individually selected, after $T_0 = 13$. Figure C.4 displays the observed and estimated outputs of the treated units for the corresponding repetition in the staggered structures. The dashed vertical lines indicate the treatment adoption dates for each treated state. The results of the synthetic control for Kentucky, New Hampshire, and Utah appear to be poor, but the outcomes for Kentucky and New Hampshire are comparatively better than those in the block structures depicted in Figure C.2. This difference can be attributed to the inclusion of North Carolina, which also exhibits extreme outcomes, among the control units as shown in Figure C.3. In summary, the results of the synthetic control method vary considerably depending on the selection of treated and control units. On the other hand, the

MC methods demonstrate more robustness to variations in the treatment adoption pattern.

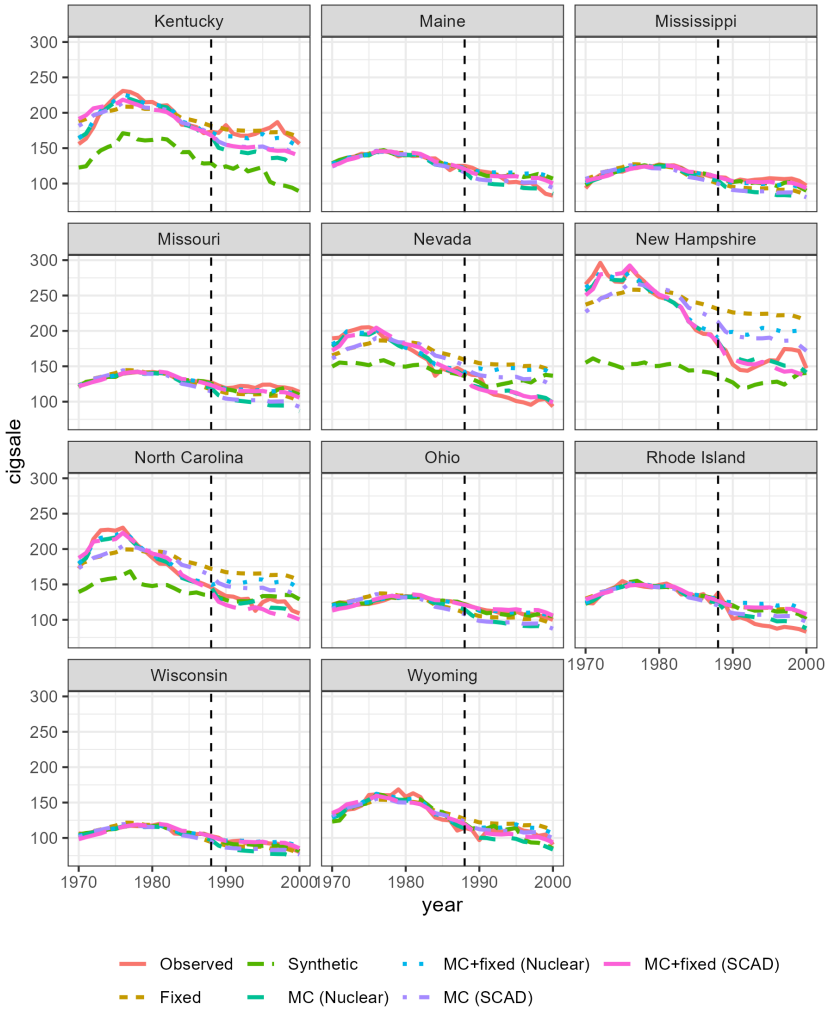

Figure C.2: Time series of the estimated and observed cigarette sales for all treated states in a specific repetition of the block structures. The dashed vertical lines indicate the year when Proposition 99 was passed.

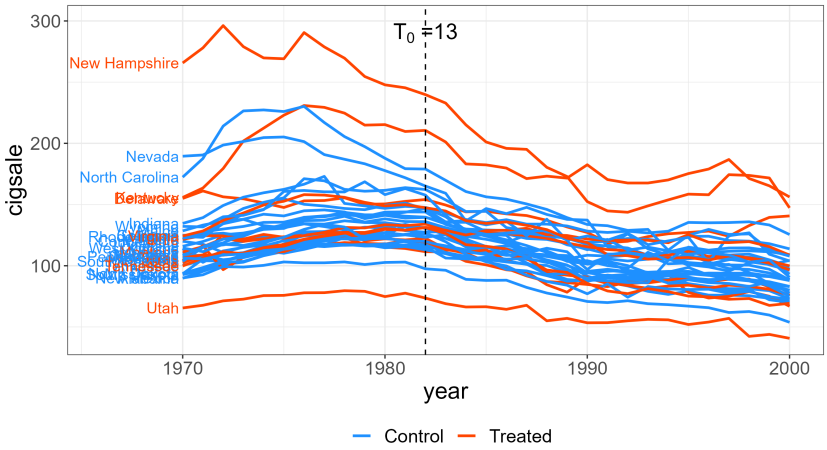

Figure C.3: Time series of the observed cigarette sales for a specific repetition in the staggered structures.

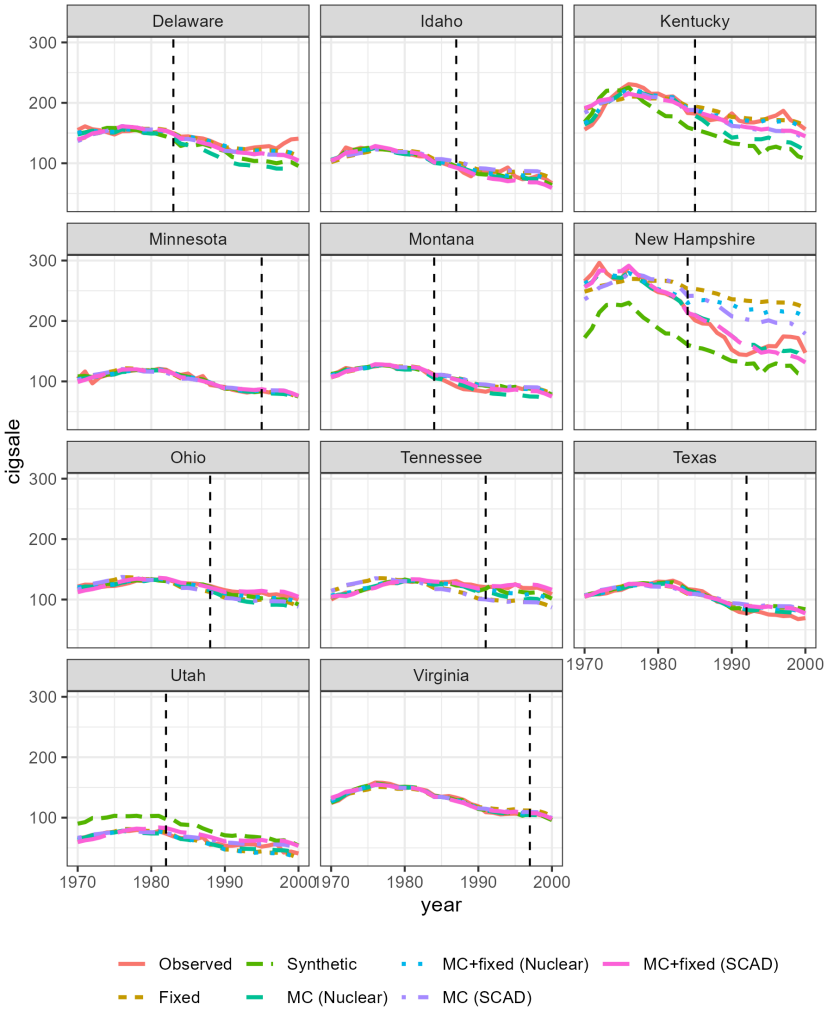

Figure C.4: Time series of the estimated and observed cigarette sales for all treated states in a specific repetition of the staggered structures. The dashed vertical lines indicate the year of the treatment adoption, which varies for each treated state.

국문초록

Low-rank 행렬 완성은 행렬의 누락된 성분을 채우기 위해 널리 사용되는 방법이다. 행렬의 특이값 크기를 축소하는 nuclear norm 벌점화는 계산 편의성 때문에 일반적으로 사용되나, 추정에 편향을 발생시킨다. 이 문제를 해결하기 위해 SCAD와 같은 비볼록 벌점화가 사용되며, 이는 성기고 편향되지 않은 추정량을 제공한다.

본 학위 논문은 시간 의존적 처리 (치료) 채택 구조를 갖는 패널 자료에서 인과 효과를 추정하기 위해 비볼록 벌점화 행렬 완성 방법을 연구한다. 우리는 먼저 nuclear norm 벌점화에 의존하는 기존 방법을 개선하는 잠재 제어 행렬에 대해 제안된 추정량의 추정 오류에 대한 상한을 도출한다. 놀랍게도, 이 상한은 참 특이값의 크기에 대한 추가 조건이 주어졌을 때 오라클 추정량이 얻는 상한과 일치한다. 또한 치료 효과에 대한 추정량의 점근적 정규성과, 이 추정량이 기존 방법에 비해 더 작은 점근 분산을 갖는다는 것을 증명한다. 우리는 잠재 제어 행렬의 복구와 평균 치료 효과의 추정을 평가하기 위해 수치 연구를 수행한다. 시뮬레이션은 우리의 이론적 결과를 검증하고, 실제 자료를 사용한 실험은 제안 방법의 유망한 성능을 입증한다.

주요어: 시간 의존적 처리 채택, 잠재 제어 행렬, SCAD, 불편 추정량, 상한, 오라클 추정량, 인과 효과, 점근적 정규성

학 번: 2013–22898