NBER WORKING PAPER SERIES

MATRIX COMPLETION METHODS FOR CAUSAL PANEL DATA MODELS

Susan Athey
Mohsen Bayati
Nikolay Doudchenko
Guido Imbens
Khashayar Khosravi

Working Paper 25132
http://www.nber.org/papers/w25132

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2018

We are grateful for comments by Alberto Abadie and participants at the NBER Summer Institute and at seminars at Stockholm University and the California Econometrics Conference. This research was generously supported by ONR grant N00014-17-1-2131 and NSF grant CMMI:1554140. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed a financial relationship of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w25132.ack

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2018 by Susan Athey, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Matrix Completion Methods for Causal Panel Data Models
Susan Athey, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi
NBER Working Paper No. 25132
October 2018
JEL No. C01,C21,C23

ABSTRACT

In this paper we study methods for estimating causal effects in settings with panel data, where a subset of units are exposed to a treatment during a subset of periods, and the goal is estimating counterfactual (untreated) outcomes for the treated unit/period combinations. We develop a class of matrix completion estimators that uses the observed elements of the matrix of control outcomes corresponding to untreated unit/periods to predict the "missing" elements of the matrix, corresponding to treated units/periods. The approach estimates a matrix that well-approximates the original (incomplete) matrix, but has lower complexity according to the nuclear norm for matrices. From a technical perspective, we generalize results from the matrix completion literature by allowing the patterns of missing data to have a time series dependency structure. We also present novel insights concerning the connections between the matrix completion literature, the literature on interactive fixed effects models and the literatures on program evaluation under unconfoundedness and synthetic control methods.

Susan Athey
Graduate School of Business
Stanford University
655 Knight Way
Stanford, CA 94305
and NBER
athey@stanford.edu

Mohsen Bayati
Graduate School of Business
Stanford University
Stanford, CA 94305
bayati@stanford.edu

Nikolay Doudchenko
Graduate School of Business
Stanford University
Stanford, CA 94305
nikolayd@stanford.edu

Guido Imbens
Graduate School of Business
Stanford University
655 Knight Way
Stanford, CA 94305
and NBER
Imbens@stanford.edu

Khashayar Khosravi
Department of Electrical Engineering
Stanford University
Stanford, CA 94305
khosravi@stanford.edu

1 Introduction

In this paper we develop new methods for estimating average causal effects in settings with panel or longitudinal data, where a subset of units is exposed to a binary treatment during a subset of periods, and we observe the realized outcome for each unit in each time period. To estimate the (average) effect of the treatment on the treated units in this setting, we focus on imputing the missing potential outcomes. The statistics and econometrics literatures have taken two general approaches to this problem. The literature on unconfoundedness (Rosenbaum and Rubin (1983); Imbens and Rubin (2015)) imputes missing potential outcomes using observed outcomes for units with similar values for observed outcomes in previous periods. The synthetic control literature (Abadie and Gardeazabal (2003); Abadie et al. (2010, 2015); Doudchenko and Imbens (2016)) imputes missing control outcomes for treated units by finding weighted averages of control units that match the treated units in terms of lagged outcomes. Although at first sight similar, the two approaches are conceptually quite different in terms of the patterns in the data they exploit to impute the missing potential outcomes. The unconfoundedness approach estimates patterns over time that are assumed to be stable across units, and the synthetic control approach estimates patterns across units that are assumed to be stable over time. Both sets of methods also primarily focus on settings with different structures on the missing data or assignment mechanism. In the case of the unconfoundedness literature typically the assumption is that the treated units are all treated in the same periods, typically only the last period, and there are a substantial number of control units. The synthetic control literature has primarily focused on the case where one or a small number of treated units are observed prior to the treatment over a substantial number of periods.

In this study we also draw on the econometric literature on factor models and interactive fixed effects, and the computer science and statistics literatures on matrix completion, to

take an approach to imputing the missing potential outcomes that is different from the unconfoundedness and synthetic control approaches. In the literature on factor models and interactive effects (Bai and Ng (2002); Bai (2003)) researchers model the observed outcome, in a balanced panel setting, as the sum of a linear function of covariates and an unobserved component that is a low rank matrix plus noise. Estimates are typically based on minimizing the sum of squared errors given the rank of the matrix of unobserved components, sometimes with the rank estimated. Xu (2017) applies this to causal settings where a subset of units is treated from common period onward, so that the complete data methods for estimating the factors and factor loadings can be used. The matrix completion literature (Candès and Recht (2009); Candès and Plan (2010); Mazumder et al. (2010)) focuses on imputing missing elements in a matrix assuming the complete matrix is the sum of a low rank matrix plus noise and the missingness is completely at random. The rank of the matrix is implicitly determined by the regularization through the addition of a penalty term to the objective function. Especially with complex missing data patterns using the nuclear norm as the regularizer is attractive for computational reasons.

In the current paper we make two contributions. First, we generalize the methods from the matrix completion literature to settings where the missing data patterns are not completely at random. In particular we allow for the possibility of staggered adoption (Athey and Imbens (2018)), where units are treated from some initial adoption date onwards, but the adoption dates vary between units. Compared to the factor model literature the proposed estimator focuses on nuclear norm regularization to avoid the computational difficulties that would arise for complex missing data patterns with the fixed-rank methods in Bai and Ng (2002) and Xu (2017), similar to the way LASSO (ℓ_1 regularization, Tibshirani (1996)) is computationally attractive relative to subset selection (ℓ_0 regularization) in linear regression models. The second contribution is to show that the synthetic control and unconfoundedness approaches, as well as our proposed method, can all be viewed as matrix

completion methods based on matrix factorization, all with the same objective function based on the Fröbenius norm for the difference between the latent matrix and the observed matrix. Given this common objective function the unconfoundedness and synthetic control approaches impose different sets of restrictions on the factors in the matrix factorization, whereas the proposed method does not impose any restrictions but uses regularization to define the estimator.

2 Set Up

Consider an $N \times T$ matrix $\mathbf{Y}$ of outcomes with typical element Y_{it}. We only observe Y_{it} for some units and some time periods. We define $\mathcal{M}$ to be the set of pairs of indices (i, t), $i \in \{1, \ldots, N\}$, $t \in \{1, \ldots, T\}$, corresponding to the missing entries and $\mathcal{O}$ to be the set corresponding to the observed entries: Y_{it} is missing if $(i, t) \in \mathcal{M}$ and observed if $(i, t) \in \mathcal{O}$. We wish to impute the missing Y_{it}. Our motivation for this problem arises from a causal potential outcome setting (e.g., Rubin (1974); Imbens and Rubin (2015)), where for each of N units and T time periods there exists a pair of potential outcomes, $Y_{it}(0)$ and $Y_{it}(1)$, with unit i exposed in period t to treatment $W_{it} \in \{0, 1\}$, and the realized outcome equal to $Y_{it} = Y_{it}(W_{it})$. In that case the primary object of interest may be the average causal effect of the treatment, $\tau = \sum_{i,t}[Y_{it}(1) - Y_{it}(0)]/(NT)$, or some other average treatment effect. In order to estimate such average treatment effects, one approach is to impute the missing potential outcomes. In this paper we focus directly on the problem of imputing the missing entries in the $\mathbf{Y}(0)$ matrix for treated units with $W_{it} = 1$.

In addition to partially observing the matrix $\mathbf{Y}$, we may also observe covariate matrices $\mathbf{X} \in \mathbb{R}^{N \times P}$ and $\mathbf{Z} \in \mathbb{R}^{T \times Q}$ where columns of $\mathbf{X}$ are unit-specific covariates, and columns of $\mathbf{Z}$ are time-specific covariates. We may also observe unit/time specific covariates $V_{it} \in \mathbb{R}^J$.

Putting aside the covariates for the time being, the data can be thought of as consisting

of two $N \times T$ matrices, one incomplete and one complete,

$$\mathbf{Y} = \begin{pmatrix} Y_{11} & Y_{12} & ? & \dots & Y_{1T} \\ ? & ? & Y_{23} & \dots & ? \\ Y_{31} & ? & Y_{33} & \dots & ? \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{N1} & ? & Y_{N3} & \dots & ? \end{pmatrix}, \qquad \text{and} \quad \mathbf{W} = \begin{pmatrix} 0 & 0 & 1 & \dots & 0 \\ 1 & 1 & 0 & \dots & 1 \\ 0 & 1 & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & 0 & \dots & 1 \end{pmatrix},$$

where

$$W_{it} = \begin{cases} 1 & \text{if } (i,t) \in \mathcal{M}, \\ 0 & \text{if } (i,t) \in \mathcal{O}, \end{cases}$$

is an indicator for Y_{it} being missing.

3 Patterns of Missing Data, Thin and Fat Matrices, and Horizontal and Vertical Regression

In this section, we discuss a number of particular configurations of the matrices $\mathbf{Y}$ and $\mathbf{W}$ that are the focus of distinct parts of the general literature. This serves to put in context the problem, and to motivate previously developed methods from the literature on causal inference under unconfoundedness, the synthetic control literature, and the interactive fixed effect literature, and subsequently to develop formal connections between all three. First, we consider patterns of missing data. Second, we consider different shapes of the matrix $\mathbf{Y}$. Third, we consider a number of specific analyses that focus on particular combinations of missing data patterns and shapes of the matrices.

3.1 Patterns of Missing Data

In the statistics literature on matrix completion the focus is on settings with randomly missing values, allowing for general patterns on the matrix of missing data indicators (Candès and Tao (2010); Recht (2011)). In many social science applications, however, there is a specific structure on the missing data.

3.1.1 Block Structure

A leading example is a block structure, with a subset of the units treated during every period from a particular point in time T_0 onwards.

$$\mathbf{Y}_{N\times T} = \begin{pmatrix} \checkmark & \checkmark & \checkmark & \checkmark & \dots & \checkmark \\ \checkmark & \checkmark & \checkmark & \checkmark & \dots & \checkmark \\ \checkmark & \checkmark & \checkmark & \checkmark & \dots & \checkmark \\ \checkmark & \checkmark & \checkmark & ? & \dots & ? \\ \checkmark & \checkmark & \checkmark & ? & \dots & ? \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \checkmark & \checkmark & \checkmark & ? & \dots & ? \end{pmatrix}.$$

There are two special cases of the block structure. Much of the literature on estimating average treatment effects under unconfoundedness focuses on the case where $T_0 = T$, so that the only treated units are in the last period. We will refer to this as the single-treated-period block structure. In contrast, the synthetic control literature focuses on the case of with a single treated unit which are treated for a number of periods from period T_0

onwards, the single-treated-unit block structure:

$$\mathbf{Y} = \begin{pmatrix} \checkmark & \checkmark & \checkmark & \dots & \checkmark & \checkmark \\ \checkmark & \checkmark & \checkmark & \dots & \checkmark & \checkmark \\ \checkmark & \checkmark & \checkmark & \dots & \checkmark & ? \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \checkmark & \checkmark & \checkmark & \dots & \checkmark & ? \\ & & & & & \uparrow \\ & & & & & \text{treated period} \end{pmatrix} \quad \text{and} \quad \mathbf{Y} = \begin{pmatrix} \checkmark & \checkmark & \checkmark & \dots & \checkmark & \\ \checkmark & \checkmark & \checkmark & \dots & \checkmark & \\ \checkmark & \checkmark & \checkmark & \dots & \checkmark & \\ \vdots & \vdots & \vdots & \ddots & \vdots & \\ \checkmark & \checkmark & \checkmark & \dots & \checkmark & \\ \checkmark & \checkmark & ? & \dots & ? & \leftarrow \text{treated unit} \end{pmatrix}.$$

3.1.2 Staggered Adoption

Another setting that has received attention is characterized by staggered adoption of the treatment (Athey and Imbens (2018)). Here units may differ in the time they are first exposed to the treatment, but once exposed they remain in the treatment group forever after. This naturally arises in settings where the treatment is some new technology that units can choose to adopt (e.g., Athey and Stern (2002)). Here:

$$\mathbf{Y}_{N\times T} = \begin{pmatrix} \checkmark & \checkmark & \checkmark & \checkmark & \dots & \checkmark & \text{(never adopter)} \\ \checkmark & \checkmark & \checkmark & \checkmark & \dots & ? & \text{(late adopter)} \\ \checkmark & \checkmark & ? & ? & \dots & ? & \\ \checkmark & \checkmark & ? & ? & \dots & ? & \text{(medium adopter)} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \\ \checkmark & ? & ? & ? & \dots & ? & \text{(early adopter)} \end{pmatrix}.$$

3.2 Thin and Fat Matrices

A second classification concerns the shape of the matrix $\mathbf{Y}$. Relative to the number of time periods, we may have many units, few units, or a comparable number. These data configurations may make particular analyses more attractive. For example, $\mathbf{Y}$ may be a thin matrix, with $N \gg T$, or a fat matrix, with $N \ll T$, or an approximately square matrix, with $N \approx T$:

$$\mathbf{Y} = \begin{pmatrix} ? & \checkmark & ? \\ \checkmark & ? & \checkmark \\ ? & ? & \checkmark \\ \checkmark & ? & \checkmark \\ ? & ? & ? \\ \vdots & \vdots & \vdots \\ ? & ? & \checkmark \end{pmatrix} \quad (\textbf{thin}) \qquad \mathbf{Y} = \begin{pmatrix} ? & ? & \checkmark & \checkmark & \checkmark & \dots & ? \\ \checkmark & \checkmark & \checkmark & \checkmark & ? & \dots & \checkmark \\ ? & \checkmark & ? & \checkmark & ? & \dots & \checkmark \end{pmatrix} \quad (\textbf{fat}),$$

or

$$\mathbf{Y} = \begin{pmatrix} ? & ? & \checkmark & \checkmark & \dots & ? \\ \checkmark & \checkmark & \checkmark & \checkmark & \dots & \checkmark \\ ? & \checkmark & ? & \checkmark & \dots & \checkmark \\ \checkmark & \checkmark & ? & \checkmark & \dots & \checkmark \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ ? & ? & \checkmark & \checkmark & \dots & \checkmark \end{pmatrix} \quad (\textbf{approximately square}).$$

3.3 Horizontal and Vertical Regressions

Two special combinations of missing data patterns and the shape of the matrices deserve particular attention because they are the focus of substantial separate literatures.

3.3.1 Horizontal Regression and the Unconfoundedness Literature

The unconfoundedness literature focuses primarily on the single-treated-period block structure with a thin matrix, and imputes the missing potential outcomes in the last period using control units with similar lagged outcomes. A simple version of that approach is to regress the last period outcome on the lagged outcomes and use the estimated regression to predict the missing potential outcomes. That is, for the units with $(i, T) \in \mathcal{M}$, the predicted outcome is

$$\hat{Y}_{iT} = \hat{\beta}_0 + \sum_{s=1}^{T-1} \hat{\beta}_s Y_{is}, \quad \text{where } \hat{\beta} = \arg\min_{\beta} \sum_{i:(i,T)\in\mathcal{O}} \left(Y_{iT} - \beta_0 - \sum_{s=1}^{T-1} \beta_s Y_{is} \right)^2. \tag{3.1}$$

We refer to this as a **horizontal** regression, where the rows of the $\mathbf{Y}$ matrix form the units of observation. A more flexible, nonparametric, version of this estimator would correspond to matching where we find for each treated unit i a corresponding control unit j with Y_{jt} approximately equal to Y_{it} for all pre-treatment periods $t = 1, \ldots, T-1$.

3.3.2 Vertical Regression and the Synthetic Control Literature

The synthetic control literature focuses primarily on the single-treated-unit block structure with a fat or approximately square matrix. Doudchenko and Imbens (2016) discuss how the Abadie-Diamond-Hainmueller synthetic control method can be interpreted as regressing the outcomes for the treated unit prior to the treatment on the outcomes for the control units in the same periods. That is, for the treated unit in period t, for $t = T_0, \ldots, T$, the predicted outcome is

$$\hat{Y}_{Nt} = \hat{\gamma}_0 + \sum_{i=1}^{N-1} \hat{\gamma}_i Y_{it}, \quad \text{where } \hat{\gamma} = \arg\min_{\gamma} \sum_{t:(N,t)\in\mathcal{O}} \left(Y_{Nt} - \gamma_0 - \sum_{i=1}^{N-1} \gamma_i Y_{it} \right)^2. \tag{3.2}$$

We refer to this as a **vertical** regression, where the columns of the $\mathbf{Y}$ matrix form the units of observation. As shown in Doudchenko and Imbens (2016) this is a special case of the Abadie et al. (2015) estimator, without imposing their restrictions that the coefficients are nonnegative and that the intercept is zero.

Although this does not appear to have been pointed out previously, a matching version of this estimator would correspond to finding, for each period t where unit N is treated, a corresponding period $s \in \{1, \ldots, T_0 - 1\}$ such that Y_{is} is approximately equal to Y_{Ns} for all control units $i = 1, \ldots, N-1$. This matching version of the synthetic control estimator clarifies the link between the treatment effect literature under unconfoundedness and the synthetic control literature.

Suppose that the only missing entry is in the last period for unit N. In that case if we estimate the horizontal regression in (3.1), it is still the case that imputed $\hat{Y}_{NT}$ is linear in the observed $Y_{1T}, \ldots, Y_{N-1,T}$, just with different weights than those obtained from the vertical regression. Similarly, if we estimate the vertical regression in (3.2), it is still the case that $\hat{Y}_{NT}$ is linear in $Y_{N1}, \ldots, Y_{N,T-1}$, just with different weights from the horizontal regression.

3.4 Fixed Effects and Factor Models

The horizontal regression focuses on a pattern in the time path of the outcome Y_{it}, specifically the relation between Y_{iT} and the lagged Y_{it} for $t = 1, \ldots, T-1$, for the units for whom these values are observed, and assumes that this pattern is the same for units with missing outcomes. The vertical regression focuses on a pattern between units at times when we observe all outcomes, and assumes this pattern continues to hold for periods when some outcomes are missing. However, by focusing on only one of these patterns, cross-section or time series, these approaches ignore alternative patterns that may help in imputing the

missing values. An alternative is to consider approaches that allow for the exploitation of both stable patterns over time, and stable patterns accross units. Such methods have a long history in the panel data literature, including the literature on fixed effects, and more generally, factor and interactive fixed effect models (e.g., Chamberlain (1984); Arellano and Honoré (2001); Liang and Zeger (1986); Bai (2003, 2009); Pesaran (2006); Moon and Weidner (2015, 2017)). In the absence of covariates (although in this literature the coefficients on these covariates are typically the primary focus of the analyses), such models can be written as

$$Y_{it} = \sum_{r=1}^{R} \gamma_{ir}\delta_{tr} + \varepsilon_{it}, \qquad \text{or} \quad \mathbf{Y} = \mathbf{U}\mathbf{V}^{\top} + \boldsymbol{\varepsilon}, \tag{3.3}$$

where $\mathbf{U}$ is $N \times R$ and $\mathbf{V}$ is $T \times R$. Most of the early literature, Anderson (1958) and Goldberger (1972)), focused on the thin matrix case, with $N \gg T$, where asymptotic approximations are based on letting the number of units increase with the number of time periods fixed. In the modern part of this literature (Bai (2003, 2009); Pesaran (2006); Moon and Weidner (2015, 2017); Bai and Ng (2017)) researchers allow for more complex asymptotics with both N and T increasing, at rates that allow for consistent estimation of the factors $\mathbf{V}$ and loadings $\mathbf{U}$ after imposing normalizations. In this literature it is typically assumed that the number of factors R is fixed, although not necessarily known. Methods for estimating the rank R are discussed in Bai and Ng (2002) and Moon and Weidner (2015).

Xu (2017) implements this interactive fixed effect approach to the matrix completion problem in the special case with blocked assignment, with additional applications in Gobillon and Magnac (2013); Kim and Oka (2014) and Hsiao et al. (2012). Suppose the first N_C units are in the control group, and the last $N_T = N - N_C$ units are in the treatment group. The treatment group is exposed to the control treatment in the first $T_0 - 1$ pre-treatment periods, and exposed to the active treatment in the post-treatment periods $T_0, \ldots, T$. In

that case we can partition $\mathbf{U}$ and $\mathbf{V}$ accordingly and write

$$\mathbf{U}\mathbf{V}^\top = \begin{pmatrix} \mathbf{U}_C \\ \mathbf{U}_T \end{pmatrix} \begin{pmatrix} \mathbf{V}_{\text{pre}} \\ \mathbf{V}_{\text{post}} \end{pmatrix}^\top .$$

Using the data from the control group pre and post, and the pre data only for the treatment group, we have

$$\mathbf{Y}_C = \mathbf{U}_C \begin{pmatrix} \mathbf{V}_{\text{pre}} \\ \mathbf{V}_{\text{post}} \end{pmatrix}^\top + \varepsilon_C, \qquad \text{and} \quad \mathbf{Y}_{T,\text{pre}} = \mathbf{U}_T \mathbf{V}_{\text{pre}}^\top + \varepsilon_{T,\text{pre}}$$

where the first equation can be used to estimate $\mathbf{U}_C$, $\mathbf{V}_{\text{pre}}$, and $\mathbf{V}_{\text{post}}$, and the second is used to estimate $\mathbf{U}_T$, both by least squares after normalizing $\mathbf{U}$ and $\mathbf{V}$. Note that this is not necessarily efficient, because $\mathbf{Y}_{T,\text{pre}}$ is not used to estimate $\mathbf{V}_{\text{pre}}$.

Independently, a closely related literature has emerged in machine learning and statistics on matrix completion (Srebro et al. (2005); Candès and Recht (2009); Candès and Tao (2010); Keshavan et al. (2010a,b); Gross (2011); Recht (2011); Rohde et al. (2011); Negahban and Wainwright (2011, 2012); Koltchinskii et al. (2011); Klopp (2014)). In this literature the starting point is an incompletely observed matrix, and researchers have proposed matrix-factorization approaches to matrix completion, similar to (3.3). The focus is not on estimating $\mathbf{U}$ and $\mathbf{V}$ consistently, only on imputing the missing elements of $\mathbf{Y}$. Instead of fixing the rank of the underlying matrix, estimators rely on regularization, and in particular nuclear norm regularization.

4 The Nuclear Norm Matrix Completion Estimator

In the absence of covariates we model the $N \times T$ matrix of outcomes $\mathbf{Y}$ as

$$\mathbf{Y} = \mathbf{L}^* + \boldsymbol{\varepsilon}, \qquad \text{where} \quad \mathbb{E}[\boldsymbol{\varepsilon}|\mathbf{L}^*] = \mathbf{0}\,. \tag{4.1}$$

The ε_{it} can be thought of as measurement error. The goal is to estimate the matrix $\mathbf{L}^*$.

To facilitate the characterization of the estimator, define for any matrix $\mathbf{A}$, and given a set of pairs of indices $\mathcal{O}$, the two matrices $\mathbf{P}_{\mathcal{O}}(\mathbf{A})$ and $\mathbf{P}^{\perp}_{\mathcal{O}}(\mathbf{A})$ with typical elements:

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

A critical role is played by various matrix norms, summarized in Table 1. Some of these depend on the singular values, where, given the full Singular Value Decomposition (SVD) $\mathbf{L}_{N\times T} = \mathbf{S}_{N\times N}\boldsymbol{\Sigma}_{N\times T}\mathbf{R}^{\top}_{T\times T}$, the singular values $\sigma_i(\mathbf{L})$ are the ordered diagonal elements of Σ.

Table 1: Matrix Norms for Matrix $\mathbf{L}$

Schatten Norm	$\|\mathbf{L}\|_p$	$\left(\sum_i \sigma_i(\mathbf{L})^p\right)^{1/p}$
Fröbenius Norm	$\|\mathbf{L}\|_F$	$\left(\sum_i \sigma_i(\mathbf{L})^2\right)^{1/2} = \left(\sum_{i=1}^N \sum_{t=1}^T L_{it}^2\right)^{1/2}$
Rank Norm	$\|\mathbf{L}\|_0$	$\sum_i \mathbf{1}_{\sigma_i(\mathbf{L})>0}$
Nuclear Norm	$\|\mathbf{L}\|_*$	$\sum_i \sigma_i(\mathbf{L})$
Operator Norm	$\|\mathbf{L}\|_{\text{op}}$	$\max_i \sigma_i(\mathbf{L}) = \sigma_1(\mathbf{L})$
Max Norm	$\|\mathbf{L}\|_{\max}$	$\max_{1\le i\le N, 1\le t\le T} \|L_{it}\|$
Element Wise ℓ_1 Norm	$\|\mathbf{L}\|_{1,e} = \sum_{i,t} \|L_{it}\|$	

Now consider the problem of estimating $\mathbf{L}$. Directly minimizing the sum of squared

differences,

$$\min_{\mathbf{L}} \frac{1}{|\mathcal{O}|} \sum_{(i,t)\in\mathcal{O}} (Y_{it} - L_{it})^2 = \min_{\mathbf{L}} \frac{1}{|\mathcal{O}|} \|\mathbf{P}_{\mathcal{O}}(\mathbf{Y} - \mathbf{L})\|_F^2 \,, \tag{4.2}$$

does not lead to a useful estimator: if $(i,t) \in \mathcal{M}$ the objective function does not depend on L_{it}, and for other pairs (i,t) the estimator would simply be Y_{it}. Instead, we regularize the problem by adding a penalty term $\lambda\|\mathbf{L}\|$, for some choice of the norm $\|\cdot\|$.

The estimator: The general form of our proposed estimator for $\mathbf{L}^*$ is

$$\hat{\mathbf{L}} = \arg\min_{\mathbf{L}} \left\{ \frac{1}{|\mathcal{O}|} \|\mathbf{P}_{\mathcal{O}}(\mathbf{Y} - \mathbf{L})\|_F^2 + \lambda \|\mathbf{L}\|_* \right\} \,, \tag{4.3}$$

studied by Mazumder et al. (2010), with the penalty factor λ chosen through cross-validation that will be described at the end of this section. We will call this the Matrix-Completion with Nuclear Norm Minimization (MC-NNM) estimator.

Other commonly used Schatten norms would not work as well for this specific problem. For example, the Fröbenius norm on the penalty term would not have been suitable for estimating $\mathbf{L}^*$ in the case with missing entries because the solution for L_{it} for $(i,t) \in \mathcal{M}$ is always zero (which follows directly from the representation of $\|\mathbf{L}\|_F = \sum_{i,t} L_{it}^2$). The rank norm is not computationally feasible for large N and T if the cardinality and complexity of the set $\mathcal{M}$ are substantial. Formally, the problem is NP-hard. In contrast, a major advantage of using the nuclear norm is that the resulting estimator can be computed using fast convex optimization programs, e.g. the SOFT-IMPUTE algorithm by Mazumder et al. (2010) that will be described next.

Calculating the Estimator: The algorithm for calculating our estimator (in the case without additional covariates) goes as follows. Given the SVD for $\mathbf{A}$, $\mathbf{A} = \mathbf{S}\mathbf{\Sigma}\mathbf{R}^\top$, with

singular values $\sigma_1(\mathbf{A}), \ldots, \sigma_{\min(N,T)}(\mathbf{A})$, define the matrix shrinkage operator

$$\text{shrink}_\lambda(\mathbf{A}) = \mathbf{S}\tilde{\mathbf{\Sigma}}\mathbf{R}^\top, \tag{4.4}$$

where $\tilde{\mathbf{\Sigma}}$ is equal to $\mathbf{\Sigma}$ with the i-th singular value $\sigma_i(\mathbf{A})$ replaced by $\max(\sigma_i(\mathbf{A}) - \lambda, 0)$. Now start with the initial choice $\mathbf{L}_1(\lambda, \mathcal{O}) = \mathbf{P}_\mathcal{O}(\mathbf{Y})$. Then for $k = 1, 2, \ldots,$ define,

$$\mathbf{L}_{k+1}(\lambda, \mathcal{O}) = \text{shrink}_{\frac{\lambda|\mathcal{O}|}{2}}\left\{\mathbf{P}_\mathcal{O}(\mathbf{Y}) + \mathbf{P}_\mathcal{O}^\perp\Big(\mathbf{L}_k(\lambda)\Big)\right\}, \tag{4.5}$$

until the sequence $\{\mathbf{L}_k(\lambda, \mathcal{O})\}_{k\geq 1}$ converges. The limiting matrix $\hat{\mathbf{L}}(\lambda, \mathcal{O}) = \lim_{k\to\infty} \mathbf{L}_k(\lambda, \mathcal{O})$ is our estimator given the regularization parameter λ.

Cross-validation: The optimal value of λ is selected through cross-validation. We choose K (e.g., $K = 5$) random subsets $\mathcal{O}_k \subset \mathcal{O}$ with cardinality $\lfloor |\mathcal{O}|^2/NT \rfloor$ to ensure that the fraction of observed data in the cross-validation data sets, $|\mathcal{O}_k/|\mathcal{O}|$, is equal to that in the original sample, $|\mathcal{O}|/(NT)$. We then select a sequence of candidate regularization parameters $\lambda_1 > \cdots > \lambda_L = 0$, with a large enough λ_1, and for each subset $\mathcal{O}_k$ calculate $\hat{\mathbf{L}}(\lambda_1, \mathcal{O}_k), \ldots, \hat{\mathbf{L}}(\lambda_L, \mathcal{O}_k)$ and evaluate the average squared error on $\mathcal{O} \setminus \mathcal{O}_k$. The value of λ that minimizes the average squared error (among the K produced estimators corresponding to that λ) is the one chosen. It is worth noting that one can expedite the computation by using $\hat{\mathbf{L}}(\lambda_i, \mathcal{O}_k)$ as a warm-start initialization for calculating $\hat{\mathbf{L}}(\lambda_{i+1}, \mathcal{O}_k)$ for each i and k.

5 Theoretical Bounds for the Estimation Error

In this section we focus on the case that there are no covariates and provide theoretical results for the estimation error. Let $L_{\max}$ be a positive constant such that $\|\mathbf{L}^*\|_{\max} \leq L_{\max}$ (recall that $\|\mathbf{L}^*\|_{\max} = \max_{i,t} |\mathbf{L}^*_{it}|$). We also assume that $\mathbf{L}^*$ is a deterministic matrix.

Then consider the following estimator for $\mathbf{L}^*$.

$$\hat{\mathbf{L}} = \underset{\mathbf{L}:\|\mathbf{L}\|_{\max} \leq L_{\max}}{\arg\min} \left\{ \frac{1}{|\mathcal{O}|} \|\mathbf{P}_{\mathcal{O}}(\mathbf{Y} - \mathbf{L})\|_F^2 + \lambda \|\mathbf{L}\|_* \right\} . \tag{5.1}$$

5.1 Additional Notation

First, we start by introduction some new notation. For each positive integer n let $[n]$ be the set of integers $\{1, 2, \ldots, n\}$. In addition, for any pair of integers i, n with $i \in [n]$ define $e_i(n)$ to be the n dimensional column vector with all of its entries equal to 0 except the ith entry that is equal to 1. In other words, $\{e_1(n), e_2(n), \ldots, e_n(n)\}$ forms the standard basis for $\mathbb{R}^n$. For any two matrices $\mathbf{A}, \mathbf{B}$ of the same dimensions define the inner product $\langle \mathbf{A}, \mathbf{B} \rangle \equiv \text{trace}(\mathbf{A}^\top \mathbf{B})$. Note that with this definition, $\langle \mathbf{A}, \mathbf{A} \rangle = \|\mathbf{A}\|_F^2$.

Next, we describe a random observation process that defines the set $\mathcal{O}$. Consider N independent random variables $t_1, \ldots, t_N$ on $[T]$ with distributions $\pi^{(i)}$. Specifically, for each $(i, t) \in [N] \times [T]$, define $\pi_t^{(i)} \equiv \mathbb{P}[t_i = t]$. We also use the short notation $\mathbb{E}_\pi$ when taking expectation with respect to all distributions $\pi^{(1)}, \ldots, \pi^{(N)}$. Now, $\mathcal{O}$ can be written as $\mathcal{O} = \bigcup_{i=1}^N \Big\{ (i, 1), (i, 2), \ldots, (i, t_i) \Big\}$.

Also, for each $(i, t) \in \mathcal{O}$, we use the notation $\mathbf{A}_{it}$ to refer to $e_i(N) e_t(T)^\top$ which is a N by T matrix with all entries equal to zero except the (i, t) entry that is equal to 1. The data generating model can now be written as

$$Y_{it} = \langle \mathbf{A}_{it}, \mathbf{L}^* \rangle + \varepsilon_{it} \,, \quad \forall \ (i, t) \in \mathcal{O} \,,$$

where noise variables ε_{it} are independent σ-sub-Gaussian random variables that are also independent of $\mathbf{A}_{it}$. Recall that a random variable ε is σ-sub-Gaussian if for all real numbers t we have $\mathbb{E}[\exp(t\varepsilon)] \leq \exp(\sigma^2 t^2 / 2)$.

Note that the number of control units (N_{c}) is equal to the number of rows that have all

entries observed (i.e., $N_{\mathrm{c}} = \sum_{i=1}^{N} \mathbb{I}_{t_i=T}$). Therefore, the expected number of control units can be written as $\mathbb{E}_{\pi}[N_{\mathrm{c}}] = \sum_{i=1}^{N} \pi_T^{(i)}$. Defining

$$p_{\mathrm{c}} \equiv \min_{1 \le i \le N} \pi_T^{(i)} ,$$

we expect to have (on average) at least Np_{c} control units. The parameter p_{c} will play an important role in our main theoretical results. Specifically, assuming N and T are of the same order, we will show that the average per entry error (i.e., $\|\hat{\mathbf{L}} - \mathbf{L}^*\|_F / \sqrt{NT}$) converges to 0 if p_{c} grows larger than $\log^{3/2}(N)/\sqrt{N}$ up to a constant. To provide some intuition for such assumption on p_{c}, assume $\mathbf{L}^*$ is a matrix that is zero everywhere except in its ith row. Such $\mathbf{L}^*$ is clearly low-rank. But recovering the entry L^*_{iT} is impossible when $i_t < T$. Therefore, $\pi_T^{(i)}$ cannot be too small. Since i is arbitrary, in general p_{c} cannot be too small.

Remark 5.1. *It is worth noting that the sources of randomness in our observation process $\mathcal{O}$ are the random variables $\{t_i\}_{i=1}^N$ that are assumed to be independent of each other. But we allow that distributions of these random variables to be functions of $\mathbf{L}^*$. We also assume that the noise variables $\{\varepsilon_{it}\}_{it \in [N] \times [T]}$ are independent of each other and are independent of $\{t_i\}_{i=1}^N$. In §8 we discuss how our results could generalize to the cases with correlations among these noise variables.*

Remark 5.2. *The estimator* (5.1) *penalizes the error terms $(Y_{it} - L_{it})^2$, for $(i,t) \in \mathcal{O}$, equally. But the ex ante probability of missing entries in each row, the propensity score, increases as t increases. In §8.3, we discuss how the estimator can be modified by considering a weighted loss function based on propensity scores for the missing entries.*

5.2 Main Result

The main result of this section is the following theorem (proved in §A.1) that provides an upper bound for $\|\mathbf{L}^* - \hat{\mathbf{L}}\|_F/\sqrt{NT}$, the root-mean-squared-error (RMSE) of the estimator $\hat{\mathbf{L}}$. In literature on theoretical analysis of empirical risk minimization this type of upper bound is called an *oracle inequality*.

Theorem 1. *If the rank of* $\mathbf{L}^*$ *is* R*, then there is a constant* C *such that with probability greater than* $1 - 2(N+T)^{-2}$*,*

$$\frac{\|\mathbf{L}^* - \hat{\mathbf{L}}\|_F}{\sqrt{NT}} \leq C \max\left[L_{\max}\sqrt{\frac{\log(N+T)}{N\,p_c^2}}, \sigma\sqrt{\frac{R\,\log(N+T)}{T\,p_c^2}}, \sigma\sqrt{\frac{R\,\log^3(N+T)}{N\,p_c^2}} \right], \tag{5.2}$$

when the parameter λ *is a constant multiple of* $\sigma \max\left[\sqrt{N\log(N+T)}, \sqrt{T}\log^{3/2}(N+T)\right]/|\mathcal{O}|$.

Interpretation of Theorem 1: In order to see when the RMSE of $\hat{\mathbf{L}}$ converges to zero as N and T grow, we note that the right hand side of (5.2) converges to 0 when $\mathbf{L}^*$ is low-rank (R is constant) and $p_{\mathrm{c}} \gg \log^{3/2}(N+T)/\sqrt{\min(N,T)}$. A sufficient condition for the latter, when N and T are of the same order, is that the lower bound for the average number of control units (Np_{c}) grows larger than a constant times $\sqrt{N}\log^{3/2}(N)$. In §8 we will discuss how the estimator $\hat{\mathbf{L}}$ should be modified to obtain a sharper result that would hold for a smaller number of control units.

Comparison with existing theory on matrix-completion: Our estimator and its theoretical analysis are motivated by and generalize existing research on matrix-completion Srebro et al. (2005); Mazumder et al. (2010); Candès and Recht (2009); Candès and Tao (2010); Keshavan et al. (2010a,b); Gross (2011); Recht (2011); Rohde et al. (2011); Negahban and Wainwright (2011, 2012); Koltchinskii et al. (2011); Klopp (2014). The main

difference is in our observation model $\mathcal{O}$. Existing papers assume that entries $(i,t) \in \mathcal{O}$ are independent random variables whereas we allow for a dependency structure including staggered adoption where if $(i,t) \in \mathcal{O}$ then $(i,t') \in \mathcal{O}$ for all $t' < t$.

6 The Relationship with Horizontal and Vertical Regressions

In the second contribution of this paper we discuss the relation between the matrix completion estimator and the horizontal (unconfoundedness) and vertical (synthetic control) approaches. To faciliate the discussion, we focus on the case with $\mathcal{M}$ containing a single pair, unit N in period T, $\mathcal{M} = \{(N,T)\}$. In that case the various previously proposed versions of the vertical and horizontal regressions are both directly applicable, although estimating the coefficients may require regularization.

The observed data are $\mathbf{Y}$, an $N \times T$ matrix that can be partitioned as

$$\mathbf{Y} = \begin{pmatrix} \tilde{\mathbf{Y}} & \mathbf{y}_1 \\ \mathbf{y}_2^\top & ? \end{pmatrix},$$

where $\tilde{\mathbf{Y}}$ is $(N-1) \times (T-1)$, $\mathbf{y}_1$ is $(N-1) \times 1$, and $\mathbf{y}_2$ is $(T-1) \times 1$.

The matrix completion solution to imputing Y_{NT} can be characterized, for a given regularization parameter λ, as

$$\mathbf{L}^{\text{mc-nnm}}(\lambda) = \arg\min_{\mathbf{L}} \left\{ \frac{1}{|\mathcal{O}|} \left\| P_{\mathcal{O}} \left(\mathbf{Y} - \mathbf{L} \right) \right\|_F^2 + \lambda \|\mathbf{L}\|_* \right\}. \tag{6.1}$$

The predicted value for the missing entry Y_{NT} is then

$$\hat{Y}_{NT}^{\text{mc-nnm}} = \mathbf{L}_{NT}^{\text{mc-nnm}}(\lambda). \tag{6.2}$$

We are interested in comparing this estimator to horizontal regression estimator. Let us initially assume that the horizontal regression is well defined, without regularization, so that $N > T$. First define

$$\hat{\beta}^{\text{hr}} = \left(\tilde{\mathbf{Y}}^\top \tilde{\mathbf{Y}}\right)^{-1} \left(\tilde{\mathbf{Y}}^\top \mathbf{y}_1\right).$$

Then the horizontal regression based prediction is

$$\hat{Y}_{NT}^{\text{hr}} = \mathbf{y}_2^\top \hat{\beta}^{\text{hr}} = \mathbf{y}_2^\top \left(\tilde{\mathbf{Y}}^\top \tilde{\mathbf{Y}}\right)^{-1} \left(\tilde{\mathbf{Y}}^\top \mathbf{y}_1\right).$$

For the vertical (synthetic control) regression, initially assuming $T > N$, we start with

$$\hat{\gamma}^{\text{vt}} = \left(\tilde{\mathbf{Y}} \tilde{\mathbf{Y}}^\top\right)^{-1} \left(\tilde{\mathbf{Y}} \mathbf{y}_2\right),$$

leading to the horizontal regression based prediction

$$\hat{Y}_{NT}^{\text{vt}} = \mathbf{y}_1^\top \hat{\gamma}^{\text{hr}} = \mathbf{y}_1^\top \left(\tilde{\mathbf{Y}} \tilde{\mathbf{Y}}^\top\right)^{-1} \left(\tilde{\mathbf{Y}} \mathbf{y}_2\right).$$

The original (Abadie et al. (2010)) synthetic control estimator imposes the additional restrictions $\gamma_i \geq 0$, and $\sum_{i=1}^{N-1} \gamma_i = 1$, leading to

$$\hat{\gamma}^{\text{sc-adh}} = \arg\min_{\gamma} \left\| \mathbf{y}_2 - \tilde{\mathbf{Y}}^\top \gamma \right\|_F^2, \qquad \text{subject to } \forall i \ \gamma_i \geq 0, \ \sum_{i=1}^{N-1} \gamma_i = 1.$$

Then the synthetic control based prediction is

$$\hat{Y}_{NT}^{\text{sc-adh}} = \mathbf{y}_1^\top \hat{\gamma}^{\text{sc-adh}}.$$

The Doudchenko and Imbens (2016) modification allows for the possibility that $N \geq T$ and regularizes the estimator for γ. Focusing here on an elastic net regularization, their proposed estimator is

$$\hat{\gamma}^{\text{vt-en}} = \arg\min_{\gamma} \left\{ \left\| \mathbf{y}_2 - \tilde{\mathbf{Y}}^\top \gamma \right\|_F^2 + \lambda \left(\alpha \left\| \gamma \right\|_1 + \frac{1-\alpha}{2} \left\| \gamma \right\|_F^2 \right) \right\}.$$

Then the vertical elastic net prediction is

$$\hat{Y}_{NT}^{\text{vt-en}} = \mathbf{y}_1^\top \hat{\gamma}^{\text{vt-en}}.$$

We can modify the horizontal regression in the same way to allow for restrictions on the β, and regularization, although such methods have not been used in practice.

The question in this section concerns the relation between the various predictors, $\hat{Y}_{NT}^{\text{mc-nnm}}$, $\hat{Y}_{NT}^{\text{hr}}$, $\hat{Y}_{NT}^{\text{vt}}$, $\hat{Y}_{NT}^{\text{sc-adh}}$, and $\hat{Y}_{NT}^{\text{vt-en}}$. The first result states that all these estimators can be viewed as particular cases of matrix factorization estimators, with the difference coming in the way the estimation of the components of the matrix factorization is carried out.

Theorem 2. *All five estimators* $\hat{Y}_{NT}^{\text{mc-nnm}}$, $\hat{Y}_{NT}^{\text{hr}}$, $\hat{Y}_{NT}^{\text{vt}}$, $\hat{Y}_{NT}^{\text{sc-adh}}$, *and* $\hat{Y}_{NT}^{\text{vt-en}}$, *can be written in the form* $\hat{Y}_{NT}^{\text{est}} = \hat{L}_{NT}^{\text{est}}$, *for* $\text{est} \in \{\text{mc} - \text{nnm}, \text{hr}, \text{vt}, \text{sc} - \text{adh}, \text{vt} - \text{en}\}$, *where*

$$\hat{\mathbf{L}}^{\text{est}} = \mathbf{A}^{\text{est}} \mathbf{B}^{\text{est}\top},$$

with $\mathbf{L}^{\text{est}}$, $\mathbf{A}^{\text{est}}$, *and* $\mathbf{B}^{\text{est}}$ $N \times T$, $N \times R$ *and* $T \times R$ *dimensional matrices, and* $\mathbf{A}^{\text{est}}$ *and*

$\mathbf{B}^{\text{est}}$ *estimated as*

$$\left(\mathbf{A}^{\text{est}}, \mathbf{B}^{\text{est}}\right) = \arg\min_{\mathbf{A},\mathbf{B}} \left\{ \frac{1}{|\mathcal{O}|} \left\| P_{\mathcal{O}}\left(\mathbf{Y} - \mathbf{A}\mathbf{B}^{\top}\right) \right\|_F^2 + \text{penalty terms on } (\mathbf{A}, \mathbf{B}) \right\},$$

subject to restrictions on $\mathbf{A}$ *and* $\mathbf{B}$*, with the penalty terms and the restrictions specific to the estimator.*

Theorem 2 follows from the following result.

Theorem 3. *We have,*

(*i*) *(nuclear norm matrix completion)*

$$(\mathbf{A}_{\lambda}^{\text{mc-nnm}}, \mathbf{B}_{\lambda}^{\text{mc-nnm}}) = \arg\min_{\mathbf{A},\mathbf{B}} \left\{ \frac{1}{|\mathcal{O}|} \left\| P_{\mathcal{O}}\left(\mathbf{Y} - \mathbf{A}\mathbf{B}^{\top}\right) \right\|_F^2 + \lambda' \|\mathbf{A}\|_F^2 + \lambda' \|\mathbf{B}\|_F^2 \right\},$$

for $\lambda' = \lambda/2$.

(*ii*) *(horizontal regression, defined if* $N > T$*),* $R = T - 1$

$$(\mathbf{A}^{\text{hr}}, \mathbf{B}^{\text{hr}}) = \lim_{\lambda \downarrow 0} \arg\min_{\mathbf{A},\mathbf{B}} \left\{ \frac{1}{|\mathcal{O}|} \left\| P_{\mathcal{O}}\left(\mathbf{Y} - \mathbf{A}\mathbf{B}^{\top}\right) \right\|_F^2 + \lambda \|\mathbf{A}\|_F^2 + \lambda \|\mathbf{B}\|_F^2 \right\},$$

$$\text{subject to } \mathbf{A}^{\text{hr}} = \begin{pmatrix} \tilde{\mathbf{Y}} \\ \mathbf{y}_2^{\top} \end{pmatrix},$$

(*iii*) *(vertical regression, defined if* $T > N$*),* $R = N - 1$

$$(\mathbf{A}^{\text{vt}}, \mathbf{B}^{\text{vt}}) = \lim_{\lambda \downarrow 0} \arg\min_{\mathbf{A},\mathbf{B}} \left\{ \frac{1}{|\mathcal{O}|} \left\| P_{\mathcal{O}}\left(\mathbf{Y} - \mathbf{A}\mathbf{B}^{\top}\right) \right\|_F^2 + \lambda \|\mathbf{A}\|_F^2 + \lambda \|\mathbf{B}\|_F^2 \right\},$$

subject to

$$\mathbf{B}^{\text{vt}} = \begin{pmatrix} \tilde{\mathbf{Y}}^{\top} \\ \mathbf{y}_1^{\top} \end{pmatrix}.$$

(iv) (synthetic control), $R = N - 1$

$$(\mathbf{A}^{\text{sc-adh}}, \mathbf{B}^{\text{sc-adh}}) = \lim_{\lambda \downarrow 0} \arg\min_{\mathbf{A},\mathbf{B}} \left\{ \frac{1}{|\mathcal{O}|} \left\| P_{\mathcal{O}} \left(\mathbf{Y} - \mathbf{A}\mathbf{B}^\top \right) \right\|_F^2 + \lambda \|\mathbf{A}\|_F^2 + \lambda \|\mathbf{B}\|_F^2 \right\},$$

subject to

$$\mathbf{B}^{\text{sc-adh}} = \begin{pmatrix} \tilde{\mathbf{Y}}^\top \\ \mathbf{y}_1^\top \end{pmatrix}, \quad \forall\ i, A_{iT} \geq 0,\ \sum_{i=1}^{N-1} A_{iT} = 1,$$

(v) (elastic net), $R = N - 1$

$$(\mathbf{A}^{\text{vt-en}}, \mathbf{B}^{\text{vt-en}}) = \lim_{\lambda \downarrow 0} \arg\min_{\mathbf{A},\mathbf{B}} \left\{ \frac{1}{|\mathcal{O}|} \left\| P_{\mathcal{O}} \left(\mathbf{Y} - \mathbf{A}\mathbf{B}^\top \right) \right\|_F^2 + \lambda \left[\frac{1-\alpha}{2} \left\| \begin{pmatrix} \mathbf{a}_2 \\ \mathbf{a}_3 \end{pmatrix} \right\|_F^2 + \alpha \left\| \begin{pmatrix} \mathbf{a}_2 \\ \mathbf{a}_3 \end{pmatrix} \right\|_1 \right] \right\},$$

subject to

$$\mathbf{B}^{\text{vt-en}} = \begin{pmatrix} \tilde{\mathbf{Y}}^\top \\ \mathbf{y}_1^\top \end{pmatrix}, \qquad \text{where} \quad \mathbf{A} = \begin{pmatrix} \tilde{\mathbf{A}} & \mathbf{a}_1 \\ \mathbf{a}_2^\top & \mathbf{a}_3 \end{pmatrix}.$$

The proof is straightforward algebra and is omitted.

Comment 1. For nuclear norm matrix completion, if $\hat{L}$ is the solution to Equation (4.3) that has rank $\hat{R}$, then one solution for $\mathbf{A}$ and $\mathbf{B}$ is given by

$$\mathbf{A} = \mathbf{S}\mathbf{\Sigma}^{1/2}, \quad \mathbf{B} = \mathbf{R}\mathbf{\Sigma}^{1/2} \tag{6.3}$$

where $\hat{\mathbf{L}} = \mathbf{S}_{N \times \hat{R}} \mathbf{\Sigma}_{\hat{R} \times \hat{R}} \mathbf{R}_{T \times \hat{R}}^\top$ is singular value decomposition of $\hat{\mathbf{L}}$. The proof of this fact is provided in (Mazumder et al. (2010); Hastie et al. (2015)). □

Comment 2. For the horizontal regression the solution for $\mathbf{B}$ is

$$\mathbf{B}^{\text{hr}} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \\ \hat{\beta}_1 & \hat{\beta}_2 & \dots & \hat{\beta}_{T-1} \end{pmatrix},$$

and similarly for the vertical regression the solution for $\mathbf{A}$ is

$$\mathbf{A}^{\text{vt}} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \\ \hat{\gamma}_1 & \hat{\gamma}_2 & \dots & \hat{\gamma}_{N-1} \end{pmatrix}.$$

The regularization in the elastic net version only affects the last row of this matrix, and replaces it with a regularized version of the regression coefficients. □

Comment 3. The horizontal and vertical regressions are fundamentally different approaches, and they cannot easily be nested. Without some form of regularization they cannot be applied in the same setting, because the non-regularized versions require $N > T$ or $N < T$ respectively. As a result there is also no direct way to test the two methods against each other. Given a particular choice for regularization, however, one can use cross-validation methods to compare the two approaches. □

7 Two Illustrations

The objective of this section is to compare the accuracy of imputation for the matrix completion method with previously used methods. In particular, in a real data matrix $\mathbf{Y}$ where no unit is treated (no entries in the matrix are missing), we choose a subset of units as hypothetical treated units and aim to predict their values (for time periods following a randomly selected initial time). Then, we report the average root-mean-squared-error (RMSE) of each algorithm on values for the pseudo-treated (time, period) pairs. In these cases there is not necessarily a single right algorithm. Rather, we wish to assess which of the algorithms generally performs well, and which ones are robust to a variety of settings, including different adoption regimes and different configurations of the data.

We compare the following estimators:

- **DID**: Difference-in-differences based on regressing the observed outcomes on unit and time fixed effects and a dummy for the treatment.
- **VT-EN**: The vertical regression with elastic net regularization, relaxing the restrictions from the synthetic control estimator.
- **HR-EN**: The horizontal regression with elastic net regularization, similar to unconfoundedness type regressions.
- **SC-ADH**: The original synthetic control approach by Abadie et al. (2010), based on the vertical regression with Abadie-Diamond-Hainmueller restrictions.
- **MC-NNM**: Our proposed matrix completion approached via nuclear norm minimization, explained in Section 2 above.

The comparison between **MC-NNM** and the two versions of the elastic net estimator, **HR-EN** and **VT-EN**, is particularly salient. In much of the literature researchers choose

ex ante between vertical and horizontal type regressions. The **MC-NNM** method allows one to sidestep that choice in a data-driven manner.

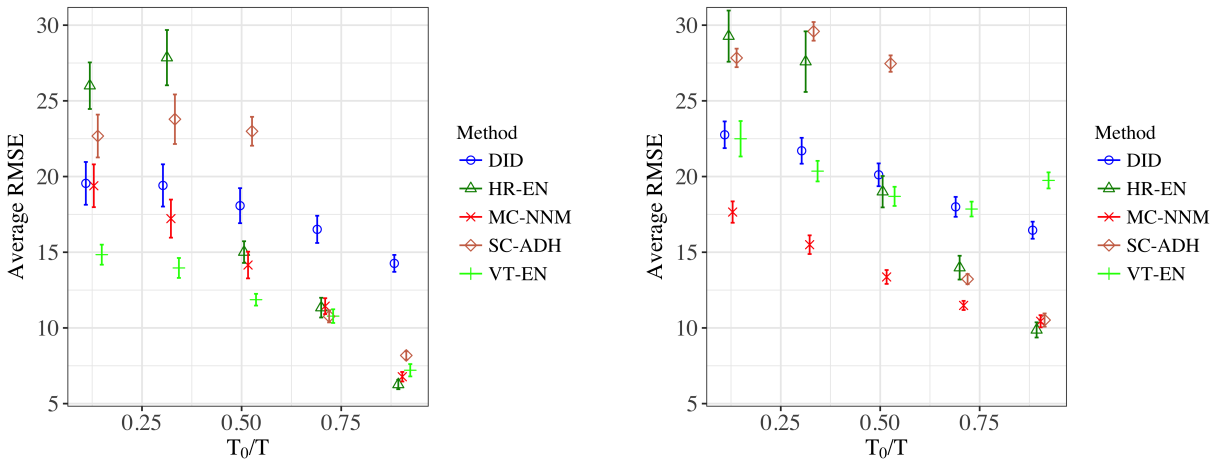
7.1 The Abadie-Diamond-Hainmueller California Smoking Data

We use the control units from the California smoking data studied in Abadie et al. (2010) with $N = 38, T = 31$. Note that in the original data set there are 39 units but one of them (state of California) is treated which will be removed in this section since the untreated values for that unit are not available. We then artificially designate some units and time periods to be treated, and compare predicted values for those unit/time-periods to the actual values.

We consider two settings for the treatment adoption:

- Case 1: Simultaneous adoption where N_t units adopt the treatment in period $T_0 + 1$, and the remaining units never adopt the treatment.
- Case 2: Staggered adoption where N_t units adopt the treatment in some period after period T, with the actual adoption date varying among these units.

In each case, the average RMSE for different ratios T_0/T is reported in Figure 1. For clarity of the figures, for each T_0/T, while all confidence intervals of various methods are calculated using the same ratio T_0/T, in the figure they are slightly jittered to the left or right. In the simultaneous adoption case, DID generally does poorly, suggesting that the data are rich enough to support more complex models. For small values of T_0/T, SC-ADH and HR-EN perform poorly while VT-EN is superior. As T_0/T grows closer to one, VT-EN, HR-EN, SC-ADH and MC-NNM methods all do well. The staggered adoption results are similar with some notable differences; VT-EN performs poorly (similar to DID) and MC-NNM is the superior approach. The performance improvement of MC-NNM can be attributed to its use of additional observations (pre-treatment values of treatment units).



(a) Simultaneous adoption, $N_t = 8$

(b) Staggered adoption, $N_t = 35$

Figure 1: California Smoking Data

7.2 Stock Market Data

In the next illustration we use a financial data set – daily returns for 2453 stocks over 10 years (3082 days). Since we only have access to a single instance of the data, in order to observe statistical fluctuations of the RMSE, for each N and T we create 50 sub-samples by looking at the first T daily returns of N randomly sampled stocks for a range of pairs of (N, T), always with $N \times T = 4900$, ranging from very thin to very fat, $(N, T) = (490, 10)$, $\ldots$ $(N, T) = (70, 70)$, $\ldots$ $(N, T) = (10, 490)$, with in each case the second half the entries missing for a randomly selected half the units (so 25% of the entries missing overall), in a block design. Here we focus on the comparison between the **HR-EN**, **VT-EN**, and **MC-NNM** estimators as the shape of the matrix changes. We report the average RMSE. Figure 2 shows the results.

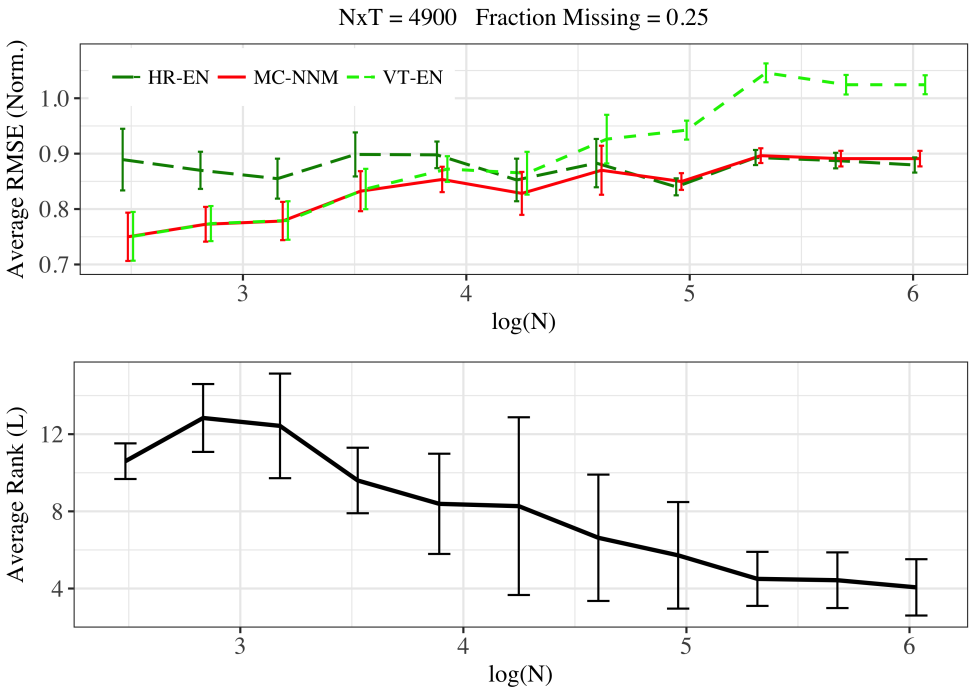


Figure 2: Stock Market Data

In the $T \ll N$ case the **VT-EN** estimator does poorly, not surprisingly because it attempts to do the vertical regression with too few time periods to estimate that well. When $N \ll T$, the **HR-EN** estimator does poorly. The most interesting finding is that the proposed **MC-NNM** method adapts well to both regimes and does as well as the best estimator in both settings, and better than both in the approximately square setting.

The bottom graph in Figure 2 shows that MC-NNM approximates the data with a matrix of rank 4 to 12, where smaller ranks are used as N grows relative to T. This validates the fact that there is a stronger correlation between daily return of different stocks than between returns for different time periods of the same stock.

8 Generalizations

Here we provide a brief discussion on how our estimator and its analysis should be adapted to more general settings.

8.1 The Model with Covariates

In Section 2 we described the basic model, and discussed the specification and estimation for the case without covariates. In this section we extend that to the case with unit-specific, time-specific, and unit-time specific covariates. For unit i we observe a vector of unit-specific covariates denoted by X_i, and $\mathbf{X}$ denoting the $N \times P$ matrix of covariates with ith row equal to $X_i^\top$. Similarly, Z_t denotes the time-specific covariates for period t, with $\mathbf{Z}$ denoting the $T \times Q$ matrix with t^{th} row equal to $Z_t^\top$. In addition we allow for a unit-time specific J by 1 vector of covariates V_{it}.

The model we consider is

$$Y_{it} = L_{it}^* + \sum_{p=1}^{P} \sum_{q=1}^{Q} X_{ip} H_{pq}^* Z_{qt} + \gamma_i^* + \delta_t^* + V_{it}^\top \beta^* + \varepsilon_{it} \,. \tag{8.1}$$

the ε_{it} is random noise. We are interested in estimating the unknown parameters $\mathbf{L}^*$, $\mathbf{H}^*$, γ^*, δ^* and β^*. This model allows for traditional econometric fixed effects for the units (the γ_i^*) and time effects (the δ_t^*). It also allows for fixed covariate (these have time varying coefficients) and time covariates (with individual coefficients) and time varying individual covariates. Note that although we can subsume the unit and time fixed effects into the matrix $\mathbf{L}^*$, we do not do so because we regularize the estimates of $\mathbf{L}^*$, but do not wish to regularize the estimates of the fixed effects.

The model can be rewritten as

$$\mathbf{Y} = \mathbf{L}^* + \mathbf{X}\mathbf{H}^*\mathbf{Z}^\top + \Gamma^*\mathbf{1}_T^\top + \mathbf{1}_N(\Delta^*)^\top + \left[V_{it}^\top \beta^*\right]_{it} + \boldsymbol{\varepsilon}\,. \tag{8.2}$$

Here $\mathbf{L}^*$ is in $\mathbb{R}^{N\times T}$, $\mathbf{H}^*$ is in $\mathbb{R}^{P\times Q}$, Γ^* is in $\mathbb{R}^{N\times 1}$ and Δ^* is in $\mathbb{R}^{T\times 1}$. An slightly richer version of this model that allows linear terms in covariates can be defined as by

$$\mathbf{Y} = \mathbf{L}^* + \tilde{\mathbf{X}}\tilde{\mathbf{H}}^*\tilde{\mathbf{Z}}^\top + \Gamma^*\mathbf{1}_T^\top + \mathbf{1}_N(\Delta^*)^\top + \left[V_{it}^\top \beta^*\right]_{it} + \boldsymbol{\varepsilon} \tag{8.3}$$

where $\tilde{\mathbf{X}} = [\mathbf{X}|\mathbf{I}_{N\times N}]$, $\tilde{\mathbf{Z}} = [\mathbf{Z}|\mathbf{I}_{T\times T}]$, and

$$\tilde{\mathbf{H}}^* = \begin{bmatrix} \mathbf{H}^*_{X,Z} & \mathbf{H}^*_X \\ \mathbf{H}^*_Z & \mathbf{0} \end{bmatrix}$$

where $\mathbf{H}^*_{XZ} \in \mathbb{R}^{P\times Q}$, $\mathbf{H}^*_Z \in \mathbb{R}^{N\times Q}$, and $\mathbf{H}^*_X \in \mathbb{R}^{P\times T}$. In particular,

$$\mathbf{Y} = \mathbf{L}^* + \tilde{\mathbf{X}}\tilde{\mathbf{H}}^*_{X,Z}\tilde{\mathbf{Z}}^\top + \tilde{\mathbf{H}}^*_Z\tilde{\mathbf{Z}}^\top + \mathbf{X}\tilde{\mathbf{H}}^*_X + \Gamma^*\mathbf{1}_T^\top + \mathbf{1}_N(\Delta^*)^\top + \left[V_{it}^\top \beta^*\right]_{it} + \boldsymbol{\varepsilon} \tag{8.4}$$

From now on, we will use the richer model (8.4) but abuse the notation and use notation $\mathbf{X}, \mathbf{H}^*, \mathbf{Z}$ instead of $\tilde{\mathbf{X}}, \tilde{\mathbf{H}}^*, \tilde{\mathbf{Z}}$. Therefore, the matrix $\mathbf{H}^*$ will be in $\mathbb{R}^{(N+P)\times(T+Q)}$.

We estimate $\mathbf{H}^*$, $\mathbf{L}^*$, δ^*, γ^*, and β^* by solving the following convex program,

$$\min_{\mathbf{H},\mathbf{L},\delta,\gamma,\beta} \left[\sum_{(i,t)\in\mathcal{O}} \frac{1}{\|\mathcal{O}|} \left(Y_{it} - L_{it} - \sum_{p=1}^{P}\sum_{q=1}^{Q} X_{ip}H_{pq}Z_{qt} - \gamma_i - \delta_t - V_{it}\beta \right)^2 + \lambda_L\|\mathbf{L}\|_* + \lambda_H\|\mathbf{H}\|_{1,e} \right].$$

Here $\|\mathbf{H}\|_{1,e} = \sum_{i,t}|H_{it}|$ is the element-wise ℓ_1 norm. We choose λ_L and λ_H through cross-validation.

Solving this convex program is similar to the covariate-free case. In particular, by using

a similar operator to shrink$_\lambda$, defined in §2, that performs coordinate descent with respect to $\mathbf{H}$. Then we can apply this operator after each step of using shrink$_\lambda$. Coordinate descent with respect to γ, δ, and β is performed similarly but using a simpler operation since the function is smooth with respect to them.

8.2 Autocorrelated Errors

One drawback of MC-NNM is that it does not take into account the time series nature of the observations. It is likely that the columns of $\boldsymbol{\varepsilon}$ exhibit autocorrelation. We can take this into account by modifying the objective function. Let us consider this in the case without covariates, and, for illustrative purposes, let us use an autoregressive model of order one. Let $\mathbf{Y}_{i\cdot}$ and $\mathbf{L}_{i\cdot}$ be the ith row of $\mathbf{Y}$ and $\mathbf{L}$ respectively. The original objective function for $\mathcal{O} = [N] \times [T]$ is

$$\frac{1}{|\mathcal{O}|}\sum_{i=1}^{N}\sum_{t=1}^{T}(Y_{it} - L_{it})^2 + \lambda_L\|\mathbf{L}\|_* = \frac{1}{|\mathcal{O}|}\sum_{i=1}^{N}(Y_{i\cdot} - L_{i\cdot})(Y_{i\cdot} - L_{i\cdot})^\top + \lambda_L\|\mathbf{L}\|_*.$$

We can modify this to $\sum_{i=1}^{N}(Y_{i\cdot} - L_{i\cdot})\boldsymbol{\Omega}^{-1}(Y_{i\cdot} - L_{i\cdot})^\top/|\mathcal{O}| + \lambda_L\|\mathbf{L}\|_*$, where the choice for the $T \times T$ matrix $\boldsymbol{\Omega}$ would reflect the autocorrelation in the $\boldsymbol{\varepsilon}_{it}$. For example, with a first order autoregressive process, we would use $\Omega_{ts} = \rho^{|t-s|}$, with ρ an estimate of the autoregressive coefficient. Similarly, for the more general version $\mathcal{O} \subset [N] \times [T]$, we can use the function

$$\frac{1}{|\mathcal{O}|}\sum_{(i,t)\in\mathcal{O}}\sum_{(i,s)\in\mathcal{O}}(Y_{it} - L_{it})[\boldsymbol{\Omega}^{-1}]_{ts}(Y_{is} - L_{is}) + \lambda_L\|\mathbf{L}\|_*\,.$$

8.3 Weighted Loss Function

Another limitation of MC-NNM is that it puts equal weight on all observed elements of the difference $\mathbf{Y} - \mathbf{L}$ (ignoring the covariates). Ultimately we care solely about predictions of

the model for the missing elements of $\mathbf{Y}$, and for that reason it is natural to emphasize the fit of the model for elements of $\mathbf{Y}$ that are observed, but that are similar to the elements that are missing. In the program evaluation literature this is often achieved by weighting the fit by the propensity score, the probability of outcomes for a unit being missing.

We can do so in the current setting by modelling this probability in terms of the covariates and a latent factor structure. Let the propensity score be $e_{it} = \mathbb{P}(W_{it} = 1 | X_i, Z_t, V_{it})$, and let $\mathbf{E}$ be the $N \times T$ matrix with typical element e_{it}. Let us again consider the case without covariates. In that case we may wish to model the assignment $\mathbf{W}$ as

$$\mathbf{W}_{N\times T} = \mathbf{E}_{N\times T} + \boldsymbol{\eta}_{N\times T}.$$

We can estimate this using the same matrix completion methods as before, now without any missing values:

$$\hat{\mathbf{E}} = \arg\min_{\mathbf{E}} \frac{1}{NT} \sum_{(i,t)} (W_{it} - e_{it})^2 + \lambda_L \|\mathbf{E}\|_* .$$

Given the estimated propensity score we can then weight the objective function for estimating $\mathbf{L}^*$:

$$\hat{\mathbf{L}} = \arg\min_{\mathbf{L}} \frac{1}{|\mathcal{O}|} \sum_{(i,t)\in\mathcal{O}} \frac{\hat{e}_{it}}{1-\hat{e}_{it}} (Y_{it} - L_{it})^2 + \lambda_L \|\mathbf{L}\|_* .$$

8.4 Relaxing the Dependence of Theorem 1 on $p_{\mathbf{c}}$

Recall from §5.1 that the average number of control units is $\sum_{i=1}^{N} \pi_T^{(i)}$. Therefore, the fraction of control units is $\sum_{i=1}^{N} \pi_T^{(i)}/N$. However, the estimation error in Theorem 1 depends on $p_{\mathrm{c}} = \min_{1\le i\le N} \pi_T^{(i)}$ rather than $\sum_{i=1}^{N} \pi_T^{(i)}/N$. The reason for this, as discussed in §5.1 is due to special classes of matrices $\mathbf{L}^*$ where most of the rows are nearly zero (e.g, when only one row is non-zero). In order to relax this constraint we would need to restrict the family

of matrices $\mathbf{L}^*$. An example of such restriction is given by Negahban and Wainwright (2012) where they assume $\mathbf{L}^*$ is not too spiky. Formally, they assume the ratio $\|\mathbf{L}^*\|_{\max}/\|\mathbf{L}^*\|_F$ should be of order $1/\sqrt{NT}$ up to logarithmic terms. To see the intuition for this, in a matrix with all equal entries this ratio is $1/\sqrt{NT}$ whereas in a matrix where only the $(1,1)$ entry is non-zero the ratio is 1. While both matrices have rank 1, in the former matrix the value of $\|\mathbf{L}^*\|_F$ is obtained from most of the entries. In such situations, one can extend our results and obtain an upper bound that depends on $\sum_{i=1}^N \pi_T^{(i)}/N$.

8.5 Nearly Low-rank Matrices

Another possible extension of Theorem 1 is to the cases where $\mathbf{L}^*$ may have high rank, but most of its singular values are small. More formally, if $\sigma_1 \geq \cdots > \sigma_{\min(N,T)}$ are singular values of $\mathbf{L}^*$, one can obtain upper bounds that depend on k and $\sum_{r=k+1}^{\min(N,T)} \sigma_r$ for any $k \in [\min(N,T)]$. One can then optimize the upper bound by selecting the best k. In the low-rank case such optimization leads to selecting k equal to R. This type of more general upper bound has been proved in some of prior matrix completion literature, e.g. Negahban and Wainwright (2012). We expect their analyses would be generalize-able to our setting (when entries of $\mathcal{O}$ are not independent).

8.6 Additional Missing Entries

In §5.1 we assumed that all entries (i,t) of $\mathbf{Y}$ for $t \leq t_i$ are observed. However, it may be possible that some such values are missing due to lack of data collection. This does not mean that any treatment occurred in the pre-treatment period. Rather, such scenario can occur when measuring outcome values is costly and can be missed. In this case, one can extend Theorem 1 to the setting with $\mathcal{O} = \left[\bigcup_{i=1}^N \left\{(i,1),(i,2),\ldots,(i,t_i)\right\}\right] \setminus \mathcal{O}_{\text{miss}}$, where each $(i,t) \in \cup_{i=1}^N \{(i,1),(i,2),\ldots,(i,t_i)\}$ can be in $\mathcal{O}_{\text{miss}}$, independently, with probability p

for p that is not too large.

9 Conclusions

We present new results for estimation of causal effects in panel or longitudinal data settings. The proposed estimator, building on the interactive fixed effects and matrix completion literatures has attractive computational properties in settings with large N and T, and allows for a relatively large number of factors. We show how this set up relates to the program evaluation and synthetic control literatures. In illustrations we show that the method adapts well to different configurations of the data, and find that generally it outperforms the synthetic control estimators from Abadie et al. (2010) and the elastic net estimators from Doudchenko and Imbens (2016).

References

Abadie, A., A. Diamond, and J. Hainmueller (2010). 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.

Abadie, A., A. Diamond, and J. Hainmueller (2015). Comparative politics and the synthetic control method. *American Journal of Political Science*, 495–510.

Abadie, A. and J. Gardeazabal (2003). The economic costs of conflict: A case study of the basque country. *American Economic Review 93*(-), 113–132.

Anderson, T. W. (1958). *An introduction to multivariate statistical analysis*, Volume 2. Wiley New York.

Arellano, M. and B. Honoré (2001). Panel data models: some recent developments. *Handbook of econometrics 5*, 3229–3296.

Athey, S. and G. W. Imbens (2018). Design-based analysis in difference-in-differences settings with staggered adoption. Technical report, National Bureau of Economic Research.

Athey, S. and S. Stern (2002). The impact of information technology on emergency health care outcomes. *The RAND Journal of Economics 33*(3), 399432.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica 71*(1), 135–171.

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica 77*(4), 1229–1279.

Bai, J. and S. Ng (2002). Determining the number of factors in approximate factor models. *Econometrica 70*(1), 191–221.

Bai, J. and S. Ng (2017). Principal components and regularized estimation of factor models. *arXiv preprint arXiv:1708.08137*.

Bühlmann, P. and S. Van De Geer (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Candès, E. J. and Y. Plan (2010). Matrix completion with noise. *Proceedings of the IEEE 98*(6), 925–936.

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

Candès, E. J. and T. Tao (2010). The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inf. Theor. 56*(5), 2053–2080.

Chamberlain, G. (1984). Panel data. *Handbook of econometrics 2*, 1247–1318.

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

Gobillon, L. and T. Magnac (2013). Regional policy evaluation: Interactive fixed effects and synthetic controls. *Review of Economics and Statistics* (00).

Goldberger, A. S. (1972). Structural equation methods in the social sciences. *Econometrica: Journal of the Econometric Society*, 979–1001.

Gross, D. (2011). Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Information Theory 57*(3), 1548–1566.

Hastie, T., R. Mazumder, J. D. Lee, and R. Zadeh (2015). Matrix completion and low-rank svd via fast alternating least squares. *J. Mach. Learn. Res. 16*(1), 3367–3402.

Hsiao, C., H. Steve Ching, and S. Ki Wan (2012). A panel data approach for program evaluation: measuring the benefits of political and economic integration of hong kong with mainland china. *Journal of Applied Econometrics 27*(5), 705–740.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Keshavan, R. H., A. Montanari, and S. Oh (2010a, June). Matrix completion from a few entries. *IEEE Trans. Inf. Theor. 56*(6), 2980–2998.

Keshavan, R. H., A. Montanari, and S. Oh (2010b, August). Matrix completion from noisy entries. *J. Mach. Learn. Res. 11*, 2057–2078.

Kim, D. and T. Oka (2014). Divorce law reforms and divorce rates in the usa: An interactive fixed-effects approach. *Journal of Applied Econometrics 29*(2), 231–245.

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

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

Liang, K.-Y. and S. L. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika 73*(1), 13–22.

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

Moon, H. R. and M. Weidner (2015). Linear regression for panel with unknown number of factors as interactive fixed effects. *Econometrica 83*(4), 1543–1579.

Moon, H. R. and M. Weidner (2017). Dynamic linear panel regression models with interactive fixed effects. *Econometric Theory 33*(1), 158–195.

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

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

Negahban, S. N., P. Ravikumar, M. J. Wainwright, and B. Yu (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science 27*(4), 538–557.

Pesaran, M. H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica 74*(4), 967–1012.

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

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

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*(5), 688–701.

Srebro, N., N. Alon, and T. S. Jaakkola (2005). Generalization error bounds for collaborative prediction with low-rank matrices. In L. K. Saul, Y. Weiss, and L. Bottou (Eds.), *Advances in Neural Information Processing Systems 17*, pp. 1321–1328.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics 12*(4), 389–434.

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

A Online Appendix for " Matrix Completion Methods for Causal Panel Data Models": Proofs

A.1 Proof of Theorem 1

First, we will discuss three main steps that are needed for the proof.

Step 1: We show an upper bound for the sum of squared errors for all $(i,t) \in \mathcal{O}$ in terms of the regularization parameter λ, rank of $\mathbf{L}^*$, $\|\mathbf{L}^* - \hat{\mathbf{L}}\|_F$, and $\|\mathfrak{E}\|_{\text{op}}$ where $\mathfrak{E} \equiv \sum_{(i,t)\in\mathcal{O}} \varepsilon_{it}\mathbf{A}_{it}$.

Lemma 1 (Adapted from Negahban and Wainwright (2011))**.** *Then for all* $\lambda \geq 3\|\mathfrak{E}\|_{\text{op}}/|\mathcal{O}|$,

$$\sum_{(i,t)\in\mathcal{O}} \frac{\langle \mathbf{A}_{it}, \mathbf{L}^* - \hat{\mathbf{L}}\rangle^2}{|\mathcal{O}|} \leq 10\lambda\sqrt{R}\,\|\mathbf{L}^* - \hat{\mathbf{L}}\|_F\,. \tag{A.1}$$

This type of result has been shown before by Recht (2011); Negahban and Wainwright (2011); Koltchinskii et al. (2011); Klopp (2014). For convenience of the reader, we include its proof in §A. Similar results also appear in the analysis of LASSO type estimators (for example see Bühlmann and Van De Geer (2011) and references therein).

Step 2: The upper bound provided by Lemma 1 contains λ and also requires the condition $\lambda \geq 3\|\mathfrak{E}\|_{\text{op}}/|\mathcal{O}|$. Therefore, in order to have a tight bound, it is important to show an upper bound for $\|\mathfrak{E}\|_{\text{op}}$ that holds with high probability. Next lemma provides one such result.

Lemma 2. *There exist a constant* C_1 *such that*

$$\|\mathfrak{E}\|_{\text{op}} \leq C_1\sigma \max\left[\sqrt{N\log(N+T)}, \sqrt{T}\log^{3/2}(N+T)\right]\,,$$

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

This result uses a concentration inequality for sum of random matrices to find a bound for $\|\mathfrak{E}\|_{\text{op}}$. We note that previous papers, Recht (2011); Negahban and Wainwright (2011); Koltchinskii et al. (2011); Klopp (2014), contain a similar step but in their case $\mathcal{O}$ is obtained by independently sampling elements of $[N]\times[T]$. However, in our case observations from each row of the matrix are correlated. Therefore, prior results do not apply. In fact, the correlation structure deteriorates the type of upper bound that can be obtained for $\|\mathfrak{E}\|_{\text{op}}$.

Step 3: The last main step is to show that, with high probability, the random variable on the left hand side of (A.1) is larger than a constant fraction of $\|\hat{\mathbf{L}}-\mathbf{L}^*\|_F^2$. In high-dimensional statistics literature this property is also referred to as *Restricted Strong Convexity*, Negahban et al. (2012); Negahban and Wainwright (2011, 2012). The following Lemma states this property for our setting and its proof that is similar to the proof of Theorem 1 in (Negahban and Wainwright, 2012) or Lemma 12 in (Klopp, 2014) is omitted.

Lemma 3. *If the estimator* $\hat{\mathbf{L}}$ *defined above satisfies* $\|\hat{\mathbf{L}}-\mathbf{L}^*\|_F \geq \kappa$ *for a positive number* κ*, then,*

$$\mathbb{P}_\pi\left\{\frac{p_c}{2}\,\|\hat{\mathbf{L}}-\mathbf{L}^*\|_F^2 \leq \sum_{(i,t)\in\mathcal{O}} \langle \mathbf{A}_{it}, \hat{\mathbf{L}}-\mathbf{L}^*\rangle^2\right\} \geq 1-\exp\left(-\frac{p_c^2\kappa^2}{32\,T\,L_{\max}^2}\right).$$

Now we are equipped to prove the main theorem.

Proof of Theorem 1. Let $\mathbf{\Delta} = \mathbf{L}^* - \hat{\mathbf{L}}$. Then using Lemma 2 and selecting λ equal to

$3\|\mathfrak{E}\|_{\text{op}}/|\mathcal{O}|$ in Lemma 1, with probability greater than $1-(N+T)^{-2}$, we have

$$\sum_{(i,t)\in\mathcal{O}} \frac{\langle \mathbf{A}_{it}, \mathbf{\Delta}\rangle^2}{|\mathcal{O}|} \leq \frac{30C_1\sigma\sqrt{R}\max\left[\sqrt{N\,\log(N+T)}, \sqrt{T}\,\log^{3/2}(N+T)\right]}{|\mathcal{O}|}\,\|\mathbf{\Delta}\|_F\,. \quad \text{(A.2)}$$

Now, we use Lemma 3 to find a lower bound for the left hand side of (A.2). But first note that if $p_{\text{c}}^2\|\mathbf{\Delta}\|_F^2/(32\,T\,L_{\max}^2) \leq 2\log(N+T)$ then

$$\frac{\|\mathbf{\Delta}\|_F}{\sqrt{NT}} \leq 8L_{\max}\sqrt{\frac{\log(N+T)}{N\,p_{\text{c}}^2}}$$

holds which proves Theorem 1. Otherwise, using Lemma 3 for $\kappa = (8L_{\max}/p_{\text{c}})\sqrt{T\log(N+T)}$,

$$\mathbb{P}\left\{\frac{1}{2}p_{\text{c}}\|\mathbf{\Delta}\|_F^2 \leq \sum_{(i,t)\in\mathcal{O}} \langle \mathbf{A}_{it}, \mathbf{\Delta}\rangle^2\right\} \geq 1-\frac{1}{(N+T)^2}\,. \quad \text{(A.3)}$$

Combining this result, (A.2), and union bound we have, with probability greater than $1-2(N+T)^{-2}$,

$$\|\mathbf{\Delta}\|_F^2 \leq 60C_1\sigma\sqrt{R}\max\left(\sigma\sqrt{\frac{N\,\log(N+T)}{p_{\text{c}}^2}}, \sqrt{\frac{T}{p_{\text{c}}^2}}\,\log^{3/2}(N+T)\right)\|\mathbf{\Delta}\|_F\,.$$

The main result now follows after dividing both sides with $\sqrt{NT}\|\mathbf{\Delta}\|_F$. □

A.2 Proof of Lemma 1

Variants of this Lemma for similar models have been proved before. But for completeness we include its proof that is adapted from Negahban and Wainwright (2011).

Proof of Lemma 1. Let

$$f(\mathbf{L}) \equiv \sum_{(i,t)\in\mathcal{O}} \frac{(Y_{it} - L_{it})^2}{|\mathcal{O}|} + \lambda\|\mathbf{L}\|_* \,.$$

Now, using the definition of $\hat{\mathbf{L}}$,

$$f(\hat{\mathbf{L}}) \leq f(\mathbf{L}^*)\,,$$

which is equivalent to

$$\sum_{(i,t)\in\mathcal{O}} \frac{\langle \mathbf{L}^* - \hat{\mathbf{L}}, \mathbf{A}_{it}\rangle^2}{|\mathcal{O}|} + 2\sum_{(i,t)\in\mathcal{O}} \frac{\varepsilon_{it}\langle \mathbf{L}^* - \hat{\mathbf{L}}, \mathbf{A}_{it}\rangle}{|\mathcal{O}|} + \lambda\|\hat{\mathbf{L}}\|_* \leq \lambda\|\mathbf{L}^*\|_* \,. \tag{A.4}$$

Now, defining $\mathbf{\Delta} \equiv \mathbf{L}^* - \hat{\mathbf{L}}$ and using the definition of $\mathfrak{E}$, the above equation gives

$$\sum_{(i,t)\in\mathcal{O}} \frac{\langle \mathbf{\Delta}, \mathbf{A}_{it}\rangle^2}{|\mathcal{O}|} \leq -\frac{2}{|\mathcal{O}|}\langle \mathbf{\Delta}, \mathfrak{E}\rangle + \lambda\|\mathbf{L}^*\|_* - \lambda\|\hat{\mathbf{L}}\|_* \tag{A.5}$$

$$\overset{(a)}{\leq} \frac{2}{|\mathcal{O}|}\|\mathbf{\Delta}\|_*\|\mathfrak{E}\|_{\text{op}} + \lambda\|\mathbf{L}^*\|_* - \lambda\|\hat{\mathbf{L}}\|_* \tag{A.6}$$

$$\leq \frac{2}{|\mathcal{O}|}\|\mathbf{\Delta}\|_*\|\mathfrak{E}\|_{\text{op}} + \lambda\|\mathbf{\Delta}\|_* \tag{A.7}$$

$$\overset{(b)}{\leq} \frac{5}{3}\lambda\|\mathbf{\Delta}\|_* \,. \tag{A.8}$$

Here, (a) uses inequality $|\langle \mathbf{A}, \mathbf{B}\rangle| \leq \|\mathbf{A}\|_{\text{op}}\|\mathbf{B}\|_{\max}$ which is due to the fact that operator norm is dual norm to nuclear norm, and (b) uses the assumption $\lambda \geq 3\|\mathfrak{E}\|_{\text{op}}/|\mathcal{O}|$. Before continuing with the proof of Lemma 1 we state the following Lemma that is proved later in this section.

Lemma 4. *Let* $\mathbf{\Delta} \equiv \mathbf{L}^* - \hat{\mathbf{L}}$ *for* $\lambda \geq 3\|\mathfrak{E}\|_{\text{op}}/|\mathcal{O}|$ *Then there exist a decomposition* $\mathbf{\Delta} = \mathbf{\Delta}_1 + \mathbf{\Delta}_2$ *such that*

(i) $\langle \mathbf{\Delta}_1, \mathbf{\Delta}_2 \rangle = 0$.

(ii) $rank(\mathbf{\Delta}_1) \leq 2r$.

(iii) $\|\mathbf{\Delta}_2\|_* \leq 3\|\mathbf{\Delta}_1\|_*$.

Now, invoking the decomposition $\mathbf{\Delta} = \mathbf{\Delta}_1 + \mathbf{\Delta}_2$ from Lemma 4 and using the triangle inequality, we obtain

$$\|\mathbf{\Delta}\|_* \overset{(c)}{\leq} 4\|\mathbf{\Delta}_1\|_* \overset{(d)}{\leq} 4\sqrt{2r}\|\mathbf{\Delta}_1\|_F \overset{(e)}{\leq} 4\sqrt{2r}\|\mathbf{\Delta}\|_F \,. \tag{A.9}$$

where (c) uses Lemma 4(iii), (d) uses Lemma 4(ii) and Cauchy-Schwarz inequality, and (e) uses Lemma 4(i). Combining this with (A.8) we obtain

$$\sum_{(i,t)\in\mathcal{O}} \frac{\langle \mathbf{\Delta}, \mathbf{A}_{it} \rangle^2}{|\mathcal{O}|} \leq 10\lambda\sqrt{r}\,\|\mathbf{\Delta}\|_F \,, \tag{A.10}$$

which finishes the proof of Lemma 1. □

Proof of Lemma 4. Let $\mathbf{L}^* = \mathbf{U}_{N\times r}\mathbf{S}_{r\times r}(\mathbf{V}_{T\times r})^\top$ be the singular value decomposition for the rank r matrix $\mathbf{L}^*$. Let $\mathbf{P}_U$ be the projection operator onto column space of $\mathbf{U}$ and let $\mathbf{P}_{U^\perp}$ be the projection operator onto the orthogonal complement of the column space of $\mathbf{U}$. Let us recall a few linear algebra facts about these projection operators. If columns of $\mathbf{U}$ are denoted by $u_1, \ldots, u_0$, since $\mathbf{U}$ is unitary, $\mathbf{P}_U = \sum_{i=1}^r u_i u_i^\top$. Similarly, $\mathbf{P}_{U^\perp} = \sum_{i=r+1}^N u_i u_i^\top$ where $u_1, \ldots, u_0, u_{r+1}, \ldots, u_N$ forms an orthonormal basis for $\mathbb{R}^N$. In addition, the projector operators are idempotent (i.e., $\mathbf{P}_U^2 = \mathbf{P}_U, \mathbf{P}_{U^\perp}^2 = \mathbf{P}_{U^\perp}$), $\mathbf{P}_U + \mathbf{P}_{U^\perp} = \mathbf{I}_{N\times N}$.

Define $\mathbf{P}_V$ and $\mathbf{P}_{V^\perp}$ similarly. Now, we define $\mathbf{\Delta}_1$ and $\mathbf{\Delta}_2$ as follows:

$$\mathbf{\Delta}_2 \equiv \mathbf{P}_{U^\perp}\mathbf{\Delta}\mathbf{P}_{V^\perp} \quad , \quad \mathbf{\Delta}_1 \equiv \mathbf{\Delta} - \mathbf{\Delta}_2 \,.$$

It is easy to see that

$$\mathbf{\Delta}_1 = (\mathbf{P}_U + \mathbf{P}_{U^\perp})\mathbf{\Delta}(\mathbf{P}_V + \mathbf{P}_{V^\perp}) - \mathbf{P}_{U^\perp}\mathbf{\Delta}\mathbf{P}_{V^\perp} \tag{A.11}$$

$$= \mathbf{P}_U\mathbf{\Delta} + \mathbf{P}_{U^\perp}\mathbf{\Delta}\mathbf{P}_V\,. \tag{A.12}$$

Using this fact we have

$$\langle \mathbf{\Delta}_1, \mathbf{\Delta}_2 \rangle = \operatorname{trace}\left(\mathbf{\Delta}^\top \mathbf{P}_U \mathbf{P}_{U^\perp}\mathbf{\Delta}\mathbf{P}_{V^\perp} + \mathbf{P}_V \mathbf{\Delta}^\top \mathbf{P}_{U^\perp}\mathbf{P}_{U^\perp}\mathbf{\Delta}\mathbf{P}_{V^\perp}\right) \tag{A.13}$$

$$= \operatorname{trace}\left(\mathbf{P}_V \mathbf{\Delta}^\top \mathbf{P}_{U^\perp}\mathbf{\Delta}\mathbf{P}_{V^\perp}\right) \tag{A.14}$$

$$= \operatorname{trace}\left(\mathbf{\Delta}^\top \mathbf{P}_{U^\perp}\mathbf{\Delta}\mathbf{P}_{V^\perp}\mathbf{P}_V\right) = 0 \tag{A.15}$$

that gives part (i). Note that we used $\operatorname{trace}(\mathbf{AB}) = \operatorname{trace}(\mathbf{BA})$.

Looking at (A.12), part (ii) also follows since both $\mathbf{P}_U$ and $\mathbf{P}_V$ have rank r and sum of two rank r matrices has rank at most $2r$.

Before moving to part (iii), we note another property of the above decomposition of $\mathbf{\Delta}$ that will be needed next. Since the two matrices $\mathbf{L}^*$ and $\mathbf{\Delta}_2$ have orthogonal singular vectors to each other,

$$\|\mathbf{L}^* + \mathbf{\Delta}_2\|_* = \|\mathbf{L}^*\|_* + \|\mathbf{\Delta}_2\|_*\,. \tag{A.16}$$

On the other hand, using inequality (A.6), for $\lambda \geq 3\|\mathfrak{E}\|_{\mathrm{op}}/|\mathcal{O}|$ we have

$$\begin{aligned}
\lambda\left(\|\hat{\mathbf{L}}\|_* - \|\mathbf{L}^*\|_*\right) &\leq \frac{2}{|\mathcal{O}|}\|\mathbf{\Delta}\|_*\|\mathfrak{E}\|_{\mathrm{op}} \\
&\leq \frac{2}{3}\lambda\|\mathbf{\Delta}\|_* \\
&\leq \frac{2}{3}\lambda\left(\|\mathbf{\Delta}_1\|_* + \|\mathbf{\Delta}_2\|_*\right)\,.
\end{aligned} \tag{A.17}$$

Now, we can use the following for the left hand side

$$\begin{aligned}
\|\hat{\mathbf{L}}\|_* - \|\mathbf{L}^*\|_* &= \|\mathbf{L}^* + \mathbf{\Delta}_1 + \mathbf{\Delta}_2\|_* - \|\mathbf{L}^*\|_* \\
&\geq \|\mathbf{L}^* + \mathbf{\Delta}_2\|_* - \|\mathbf{\Delta}_1\|_* - \|\mathbf{L}^*\|_* \\
&\overset{(f)}{=} \|\mathbf{L}^*\|_* + \|\mathbf{\Delta}_2\|_* - \|\mathbf{\Delta}_1\|_* - \|\mathbf{L}^*\|_* \\
&= \|\mathbf{\Delta}_2\|_* - \|\mathbf{\Delta}_1\|_* .
\end{aligned}$$

Here (f) follows from (A.16). Now, combining the last inequality with (A.17) we get

$$\|\mathbf{\Delta}_2\|_* - \|\mathbf{\Delta}_1\|_* \leq \frac{2}{3}\left(\|\mathbf{\Delta}_1\|_* + \|\mathbf{\Delta}_2\|_*\right) .$$

That finishes proof of part (iii). □

A.3 Proof of Lemma 2

First we state the matrix version of Bernstein inequality for rectangular matrices (see Tropp (2012) for a derivation of it).

Proposition 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 $i \in [N]$ and a constant R. Let σ_Z be such that*

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

Then, for any $\alpha \geq 0$

$$\mathbb{P}\left\{ \left\| \sum_{i=1}^{N} \mathbf{Z}_i \right\|_{\text{op}} \geq \alpha \right\} \leq (d_1 + d_2) \exp\left[\frac{-\alpha^2}{2\sigma_Z^2 + (2D\alpha)/3} \right] . \tag{A.18}$$

Proof of Lemma 2. Our goal is to use Proposition 1. Define the sequence of independent random matrices $\mathbf{B}_1, \ldots, \mathbf{B}_N$ as follows. For every $i \in [N]$, define

$$\mathbf{B}_i = \sum_{t=1}^{t_i} \varepsilon_{it} \mathbf{A}_{it} .$$

By definition, $\mathfrak{E} = \sum_{i=1}^{N} \mathbf{B}_i$ and $\mathbb{E}[\mathbf{B}_i] = \mathbf{0}$ for all $i \in [N]$. Define the bound $D \equiv C_2 \sigma \sqrt{\log(N+T)}$ for a large enough constant C_2. For each $(i,t) \in \mathcal{O}$ define $\bar{\varepsilon}_{it} = \varepsilon_{it} \mathbb{I}_{|\varepsilon_{it}| \leq D}$. Also define $\overline{\mathbf{B}}_i = \sum_{t=1}^{t_i} \bar{\varepsilon}_{it} \mathbf{A}_{it}$ for all $i \in [N]$.

Using union bound and the fact that for σ-sub-Gaussian random variables ε_{it} we have $\mathbb{P}(|\varepsilon_{it}| \geq t) \leq 2\exp\{-t^2/(2\sigma^2)\}$ gives, for each $\alpha \geq 0$,

$$\begin{aligned}
\mathbb{P}\{ \|\mathfrak{E}\|_{\text{op}} \geq \alpha\} &\leq \mathbb{P}\left\{ \left\| \sum_{i=1}^{N} \overline{\mathbf{B}}_i \right\|_{\text{op}} \geq \alpha \right\} + \sum_{(i,t)\in\mathcal{O}} \mathbb{P}\{|\varepsilon_{it}| \geq D\} \\
&\leq \mathbb{P}\left\{ \left\| \sum_{i=1}^{N} \overline{\mathbf{B}}_i \right\|_{\text{op}} \geq \alpha \right\} + 2|\mathcal{O}| \exp\left\{ \frac{-D^2}{2\sigma^2} \right\} \\
&\leq \mathbb{P}\left\{ \left\| \sum_{i=1}^{N} \overline{\mathbf{B}}_i \right\|_{\text{op}} \geq \alpha \right\} + \frac{1}{(N+T)^3} . \qquad (A.19)
\end{aligned}$$

Now, for each $i \in [N]$, define $\mathbf{Z}_i \equiv \overline{\mathbf{B}}_i - \mathbb{E}[\overline{\mathbf{B}}_i]$. Then,

$$\begin{aligned}\left\|\sum_{i=1}^{N} \overline{\mathbf{B}}_i\right\|_{\text{op}} &\leq \left\|\sum_{i=1}^{N} \mathbf{Z}_i\right\|_{\text{op}} + \left\|\mathbb{E}\left[\sum_{1\leq i\leq N} \overline{\mathbf{B}}_i\right]\right\|_{\text{op}} \\ &\leq \left\|\sum_{i=1}^{N} \mathbf{Z}_i\right\|_{\text{op}} + \left\|\mathbb{E}\left[\sum_{1\leq i\leq N} \overline{\mathbf{B}}_i\right]\right\|_{F} \leq \left\|\sum_{i=1}^{N} \mathbf{Z}_i\right\|_{\text{op}} + \sqrt{NT}\left\|\mathbb{E}\left[\sum_{1\leq i\leq N} \overline{\mathbf{B}}_i\right]\right\|_{\max} .\end{aligned}$$

But since each ε_{it} has mean zero,

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

Therefore,

$$\sqrt{NT}\left\|\mathbb{E}\left[\sum_{1\leq i\leq N} \overline{\mathbf{B}}_i\right]\right\|_{\max} \leq \frac{\sigma\sqrt{NT}}{(N+T)^4} \leq \frac{\sigma}{(N+T)^3} ,$$

which gives

$$\left\|\sum_{i=1}^{N} \overline{\mathbf{B}}_i\right\|_{\text{op}} \leq \left\|\sum_{i=1}^{N} \mathbf{Z}_i\right\|_{\text{op}} + \frac{\sigma}{(N+T)^3} . \tag{A.20}$$

We also note that $\|\mathbf{Z}_i\|_{\text{op}} \leq 2D\sqrt{T}$ for all $i \in [N]$. The next step is to calculate σ_Z

defined in the Proposition 1. We have,

$$\left\|\sum_{i=1}^{N}\mathbb{E}[\mathbf{Z}_i\mathbf{Z}_i^\top]\right\|_{\text{op}} \leq \max_{(i,t)\in\mathcal{O}}\left\{\mathbb{E}\left[(\bar{\varepsilon}_{it}-E[\bar{\varepsilon}_{it}])^2\right]\right\} \left\|\sum_{i=1}^{N}\mathbb{E}\left[\sum_{t=1}^{t_i}e_i(N)e_i(N)^\top\right]\right\|_{\text{op}} \tag{A.21}$$

$$\leq 2\sigma^2\max_{i\in[N]}\left(\sum_{t\in[T]}t\pi_t^{(i)}\right) \leq 2T\sigma^2 \tag{A.22}$$

and

$$\left\|\sum_{i=1}^{N}\mathbb{E}[\mathbf{Z}_i^\top\mathbf{Z}_i]\right\|_{\text{op}} \leq 2\sigma^2\left\|\sum_{i=1}^{N}\mathbb{E}\left[\sum_{t=1}^{t_i}e_t(T)e_t(T)^\top\right]\right\|_{\text{op}} \tag{A.23}$$

$$= 2\sigma^2\max_{t\in[T]}\left(\sum_{i\in[N]}\sum_{t'=t}^{T}\pi_{t'}^{(i)}\right) = 2N\sigma^2\,. \tag{A.24}$$

Note that here we used the fact that random variables $\bar{\varepsilon}_{it}-E[\bar{\varepsilon}_{it}]$ are independent of each other and centered which means all cross terms of the type $\mathbb{E}\{(\bar{\varepsilon}_{it}-E[\bar{\varepsilon}_{it}])(\bar{\varepsilon}_{js}-E[\bar{\varepsilon}_{js}])\}$ are zero for $(i,t)\neq(j,s)$. Therefore, $\sigma_Z^2 = 2\sigma^2\max(N,T)$ works. Applying Proposition 1, we obtain

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

$$\leq (N+T)\exp\left[-\frac{3}{16}\min\left(\frac{\alpha^2}{\sigma^2\max(N,T)},\frac{\alpha}{D\sqrt{T})}\right)\right]\,.$$

Therefore, there is a constant C_3 such that with probability greater than $1 - \exp(-t)$,

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

Using this for a t that is a large enough constant times $\log(N+T)$, together with (A.19) and (A.20), shows with probability larger than $1 - 2(N+T)^{-3}$

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

for a constant C_1. □