Matrix Completion Methods for Causal Panel Data Models *

Susan Athey † Mohsen Bayati‡ Nikolay Doudchenko§
Guido Imbens¶ Khashayar Khosravi‖

Abstract

In this paper we study methods for estimating causal effects in settings with panel data, where some units are exposed to a treatment during some periods and the goal is estimating counterfactual (untreated) outcomes for the treated unit/period combinations. We propose a class of matrix completion estimators that uses the observed elements of the matrix of control outcomes corresponding to untreated unit/periods to impute the "missing" elements of the control outcome matrix, corresponding to treated units/periods. This leads to a matrix that well-approximates the original (incomplete) matrix, but has lower complexity according to the nuclear norm for matrices. We generalize results from the matrix completion literature by allowing the patterns of missing data to have a time series dependency structure that is common in social science applications. We 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. We show that all these estimators can be viewed as focusing on the same objective function. They differ solely in the way they deal with identification, in some cases solely through regularization (our proposed nuclear norm matrix completion estimator) and in other cases primarily through imposing hard restrictions (the unconfoundedness and synthetic control approaches). The proposed method outperforms unconfoundedness-based or synthetic control estimators in simulations based on real data.

Keywords: Causality, Synthetic Controls, Unconfoundedness, Interactive Fixed Effects, Low-Rank Matrix Estimation

*We are grateful for comments by Alberto Abadie and participants at the NBER Summer Institute and at seminars at Stockholm University and the 2017 California Econometrics Conference. This research was generously supported by ONR grant N00014-17-1-2131 and NSF grant CMMI:1554140.

†Professor of Economics, Graduate School of Business, Stanford University, SIEPR, and NBER, athey@stanford.edu.

‡Associate Professor, Graduate School of Business, Stanford University, bayati@stanford.edu.

§Google Research, 111 8th Ave, New York, NY 10011, nikolayd@google.com.

¶Professor of Economics, Graduate School of Business, and Department of Economics, Stanford University, SIEPR, and NBER, imbens@stanford.edu.

‖Department of Electrical Engineering, Stanford University, 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 some units are exposed to a binary treatment during some periods. To estimate the average causal effect of the treatment on the treated units in this setting, we impute the missing potential control outcomes.

The statistics and econometrics causal inference literatures have taken two general approaches to this problem. The literature on unconfoundedness (Rosenbaum and Rubin (1983); Imbens and Rubin (2015)) can be interpreted as imputing missing potential control outcomes for treated units using observed control outcomes for control units with similar values for observed outcomes in previous periods. In contrast, the recent synthetic control literature (Abadie and Gardeazabal (2003); Abadie et al. (2010, 2015); Doudchenko and Imbens (2016); Ben-Michael et al. (2018); Li (2019); Ferman and Pinto (2019); Arkhangelsky et al. (2019); Chernozhukov et al. (2017), see Abadie (2019) for a review) imputes missing control outcomes for treated units using weighted average outcomes for control units with the weights chosen so that the weighted lagged control outcomes match the lagged outcomes for treated units. Although at first sight similar, the two approaches are conceptually quite different in terms of the correlation patterns in the data they exploit to impute the missing potential outcomes. The unconfoundedness approach assumes that patterns over time are stable across units, and the synthetic control approach assumes that patterns across units are stable over time. In empirical work the two sets of methods have primarily been applied in settings with different structures on the missing data or assignment mechanism. In the case of the unconfoundedness literature the typical setting is one with the treated units all treated in the same periods, typically only the last period, and with a substantial number of control and treated units. The synthetic control literature has primarily focused on the setting with one or a small number of treated units observed prior to the treatment over a substantial number of periods. We argue that once regularization methods are used, the two approaches, unconfoundedness and synthetic controls, are applicable in the same settings, leaving the researcher with a real choice in terms of methods. In addition this insight allows for a more systematic comparison of their perfomance than has been appreciated in the literature.

In this study we 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 fact, we show that it can be viewed as nesting both. In the literature on factor models and interactive effects (Bai and Ng (2002); Bai (2003)) researchers model the observed outcome 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) extends these ideas to causal settings where a subset of units is treated from a common period onward, so that complete data methods for estimating the factors and factor loadings can be exploited. 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 that: (i) the complete matrix is the sum of a low rank matrix plus noise and (ii), the missingness is

completely at random (except Gamarnik and Misra (2016) that study a stylized rank one case). 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 three contributions. First, we present formal results for settings where the missing data patterns are not completely at random and have a structure that allows for correlation over time, generalizing the results from the matrix completion literature. In particular we allow for the possibility of staggered adoption (*e.g.,* Athey and Imbens (2018); Shaikh and Toulis (2019)), where units are treated from some initial adoption date onwards, but the adoption dates vary between units. We also modify the estimators from the matrix completion and factor model literatures to allow for unregularized unit and time fixed effects. Although these can be incorporated in the low rank matrix, in practice the performance of the estimator with the unregularized two-way fixed effects is substantially better. Compared to the factor model literature in econometrics 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 (or ℓ_1 regularization, Tibshirani (1996)) is computationally attractive relative to subset selection (or 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. In contrast, the proposed method does not impose any restrictions but uses regularization to characterize the estimator. In our third contribution we apply our methods to two real data sets where we observe the complete matrix. We artificially designate outcomes for some units and time periods to be missing, and then compare the performance of different imputation estimators. We find that the nuclear norm matrix completion estimator does well in a range of cases, including when T is small relative to N, when T is large relative to N, and when T and N are comparable. In contrast, the unconfoundedness and synthetic control approaches break down in some of these settings in the expected pattern (the unconfoundedness approach does not work very well if $T \gg N$, and the synthetic control approach does not work very well if $N \gg T$).

We discuss some extensions in the final part of the paper. In particular we consider extensions to settings where the probability of assignment to the treatment may vary systematically with observed characteristics. In the program evaluation literature such settings have often been addressed using inverse propensity score weighting (Rubin (2006); Hirano et al. (2003)), which can be applied here as well.

2 Set Up

We start by stating the causal problem. Consider a setting with N units observed over T periods. In each period each unit is characterized by two potential outcomes, $Y_{it}(0)$ and $Y_{it}(1)$. In period t unit i is exposed or not to a binary treatment, with $W_{it} = 1$ indicating

that the unit is exposed to the treatment and $W_{it} = 0$ otherwise. We observe for each unit and period the pair (W_{it}, Y_{it}) where the realized outcome is $Y_{it} = Y_{it}(W_{it})$. In addition to observing the matrix $\mathbf{Y}$ of realized outcomes and the matrix of treatment assignments $\mathbf{W}$, 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 $V_{it} \in \mathbb{R}^J$. Implicit in our set up is that we rule out dynamic effects and make the stable-unit-treatment-value assumption (Rubin (2006); Imbens and Rubin (2015)): the potential outcomes are indexed only by the contemporaneous treatment for that unit and not by past treatments or treatments for other units. Cases where such assumptions are restrictive include those analyzed in the dynamic treatment regime literature (Chamberlain et al. (1993); Hernan and Robins (2010)). In the case where units are only exposed to the treatment in the last period this issue is not material. Also, in the case with staggered adoption violations of the no-dynamics assumption simply changes the interpretation of the estimand, but does not in general invalidate a causal interpretation.

Here we focus on estimating the average effect for the treated, $\tau = \sum_{(i,t):W_{it}=1}[Y_{it}(1) - Y_{it}(0)]/\sum_{i,t} W_{it}$, although other averages such as the overall average causal effect, $\sum_{i,t}[Y_{it}(1) - Y_{it}(0)]/(NT)$, could be of interest too. In order to estimate such average treatment effects, one approach is to impute the missing potential outcomes. Because we focus on estimating the average effect for the treated, all the relevant values for $Y_{it}(1)$ are observed, and thus we only need to impute the missing entries in the $\mathbf{Y}(0)$ matrix for treated units with $W_{it} = 1$. For the moment we focus on the problem of imputing the missing entries in $\mathbf{Y}(0)$ given the observed values of $\mathbf{Y}(0)$ and the observed matrix $\mathbf{W}$. To ease the notation and facilitate the connection to the matrix completion literature we drop from here on the (0) part of $\mathbf{Y}(0)$ and simply focus on imputing the missing values of a partially observed matrix $\mathbf{Y}$ (with the understanding that this would be the matrix of control outcomes $\mathbf{Y}(0)$), with $\mathbf{W}$ the matrix of missing data (treatment assignment) indicators. One may also wish to use the observed values of $\mathbf{Y}(1)$ for imputing the missing values for $\mathbf{Y}(0)$, but we do not do so here. In setting with few values of $\mathbf{Y}(1)$ observed it is unlikely that the information in these values is important. (In particular in the case we focus on for part of this study, with only a single treated unit/period pair there would be no information in this value.). Extension to the cases that leverage also data from $\mathbf{Y}(1)$ require assumptions on the treatment effect and are briefly discussed in §8.2.

For any positive integer n, we use notation $[n]$ to refer to the set $\{1, \ldots, n\}$ and use $\mathbf{1}_n$ to denote the n by 1 vector of all ones. We define $\mathcal{M}$ to be the set of pairs of indices (i, t), $i \in [N]$, $t \in [T]$, corresponding to the missing entries with $W_{it} = 1$ and $\mathcal{O}$ to be the set os pairs of indices corresponding to the observed entries in $\mathbf{Y}$ with $W_{it} = 0$. 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,

$$\mathbf{Y} = \begin{pmatrix} Y_{11} & Y_{12} & ? & \ldots & Y_{1T} \\ ? & ? & Y_{23} & \ldots & ? \\ Y_{31} & ? & Y_{33} & \ldots & ? \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{N1} & ? & Y_{N3} & \ldots & ? \end{pmatrix}, \quad \text{and} \quad \mathbf{W} = \begin{pmatrix} 0 & 0 & 1 & \ldots & 0 \\ 1 & 1 & 0 & \ldots & 1 \\ 0 & 1 & 0 & \ldots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & 0 & \ldots & 1 \end{pmatrix}, \tag{2.1}$$

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 the event that the corresponding component of $\mathbf{Y}$, that is, Y_{it}, is missing. The main part of the paper is about the statistical problem of imputing the missing values in $\mathbf{Y}$. Once these are imputed we can then estimate the average causal effect of interst, τ.

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 discussion 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 and the matrix completion literature. Note that the matrix completion literature has focused primarily on the case where $\mathbf{W}$ is completely random, as in Equation (2.1), and where both dimensions of $\mathbf{Y}$ and $\mathbf{W}$ are large. First, we consider patterns of missing data, that is, distributions for $\mathbf{W}$ that differ from completely random. Second, we consider different shapes of the matrix $\mathbf{Y}$ where the relative size of the dimensions N and T may be very different and one or both may be modest in magnitude. Third, we consider a number of specific analyses in the econometrics literature that focus on particular combinations of missing data patterns and shapes of the matrices.

3.1 Patterns of Missing Data

In the statistics and computer science literatures on matrix completion the focus is typically on settings with randomly missing values, allowing for general patterns on the matrix of missing data indicators (Candès and Tao, 2010; Recht, 2011). In contrast in causal social science applications the missingness arises from treatment assignments and the choices that lead to these assignments. As a result are often specific structures on the missing data that depart substantially from complete randomness.

3.1.1 Block Structure

A leading example is a block structure, with a subset of the units adopting an irreversible treatment at a particular point in time $T_0 + 1$. In the example below the ✓ marks indicate

observed values and the ? indicate missing values.

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

There are two special cases of the block structure. Much of the literature on estimating average treatment effects under unconfoundedness (*e.g.,* Imbens and Rubin (2015)) focuses on the case where $T_0 = T - 1$, 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 (*e.g.*, Abadie et al. (2010); Abadie (2019)) focuses primarily on the case of with a single treated unit which are treated for a number of periods from period $T_0 + 1$ onwards, the single-treated-unit block structure:

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

A special case that fits in both these settings is that with a single missing unit/time pair:

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

This specific setting is useful to contrast methods developed for the single-treated period (unconfoundedness) case with those developed for the single-treated unit (synthetic control) case because both sets of methods are potentially applicable.

3.1.2 Staggered Adoption

Another setting that has received attention is the staggered adoption design (Athey and Imbens (2018); Shaikh and Toulis (2019)). Here units may differ in the time they are first exposed to the treatment, but the treatment is irreversible. 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 of the problem 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 partly by removing the need for regularization. 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 $N \approx T$:

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

or

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

3.3 Horizontal and Vertical Regressions

Two special combinations of missing data patterns and matrix shape deserve particular attention because they are the focus of large, mostly separate, literatures.

3.3.1 Horizontal Regression and the Unconfoundedness Literature

The unconfoundedness literature (Rosenbaum and Rubin (1983); Rubin (2006); Imbens and Wooldridge (2009); Abadie and Cattaneo (2018)) focuses primarily on the single-treated-period block structure with a thin matrix ($N \gg T$), a substantial number of treated and control units, and imputes the missing potential outcomes in the last period using control

units with similar lagged outcomes:

$$\mathbf{Y} = \begin{pmatrix} \checkmark & \checkmark & \checkmark \\ \vdots & \vdots & \vdots \\ \checkmark & \checkmark & \checkmark \\ \checkmark & \checkmark & ? \\ \vdots & \vdots & \vdots \\ \checkmark & \checkmark & ? \end{pmatrix},$$

A simple version of the unconfoundedness 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 (Abadie et al. (2010)) focuses primarily on the single-treated-unit block structure with a relatively fat ($T \gg N$) or approximately square matrix ($T \approx N$), and a substantial number of pre-treatment periods:

$$\mathbf{Y} = \begin{pmatrix} \checkmark & \checkmark & \ldots & \checkmark & \checkmark & \ldots & \checkmark \\ \checkmark & \checkmark & \ldots & \checkmark & \checkmark & \ldots & \checkmark \\ \checkmark & \checkmark & \ldots & \checkmark & ? & \ldots & ? \end{pmatrix}.$$

Doudchenko and Imbens (2016) and Ferman and Pinto (2019) show 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 generalization of the original Abadie et al. (2010) synthetic control estimator, relaxing two restriction: (i) that the coefficients are nonnegative and (ii) that the intercept in this regression is zero. Note that these restrictions may well be substantively plausible and they can greatly improve precision.

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 $s \in \{1, \ldots, T_0\}$ 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 may serve to clarify 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, period T. 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 in (3.1). Note also that the restrictions that the coefficients are nonnegative and sum to one are common in the synthetic control literature, but could also be imposed in the unconfoundedness literature, although they do not appear to have been used there.

Juxtaposing the unconfoundedness and synthetic control approaches as we have done here raises the question how they are related, and whether there is an approach that avoids the choice between focusing on the cross-section and time-series correlation patterns. We further elaborate on the connection between the horizontal and vertical regression in §5 after introducing a third approach.

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 two-way fixed effects, and more generally, factor and interactive fixed effect models (e.g., Chamberlain (1984); Angrist and Pischke (2008); Arellano and Honoré (2001); Liang and Zeger (1986); Bai (2003, 2009); Bai and Ng (2002); 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), the common two-way fixed effect model is

$$Y_{it} = \gamma_i + \delta_t + \epsilon_{it}. \tag{3.3}$$

More general factor models can be written as

$$Y_{it} = \sum_{r=1}^{R} \gamma_{ir}\delta_{tr} + \varepsilon_{it}, \qquad \text{or} \quad \mathbf{Y} = \mathbf{U}\mathbf{V}^\top + \boldsymbol{\varepsilon}, \tag{3.4}$$

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 it is not necessarily known to the researcher. Methods for estimating the rank R are discussed in Bai and Ng (2002) and Moon and Weidner (2015).

Xu (2017) adapts 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). The block structure greatly simplifies the computation of the fixed rank estimators. However, this approach is not efficient, nor computationally attractive, in settings with more complex missing data patterns.

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; Wang et al., 2018). In this literature the starting point is an incompletely observed matrix $\mathbf{Y}$, and researchers have proposed low-rank matrix models as the basis for matrix completion, similar to (3.4). The focus is not on estimating $\mathbf{U}$ and $\mathbf{V}$ consistently, but on imputing the missing elements of $\mathbf{Y}$. Instead of fixing the rank R of the underlying matrix, a family of these estimators rely on regularization, and in particular nuclear norm regularization.

4 The Matrix Completion with Nuclear Norm Minimization Estimator

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

Assumption 1. *$\boldsymbol{\varepsilon}$ is independent of $\mathbf{L}^*$, and the elements of $\boldsymbol{\varepsilon}$ are σ-sub-Gaussian and independent of each other. Recall that a real-valued random variable ε is σ-sub-Gaussian if for all real numbers t we have $\mathbb{E}[\exp(t\varepsilon)] \leq \exp(\sigma^2 t^2/2)$.*

The goal is to estimate the matrix $\mathbf{L}^*$. We note that here the fixed effects are absorbed in $\mathbf{L}^*$ since they are two rank 1 matrices and their addition does not affect our low-rank assumption on $\mathbf{L}^*$.

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) $\mathbf{L}_{N\times T} = \mathbf{S}_{N\times N}\mathbf{\Sigma}_{N\times T}\mathbf{R}^{\top}_{T\times T}$, the singular values $\sigma_i(\mathbf{L})$ are the ordered diagonal elements of $\mathbf{\Sigma}$. Now consider the problem of estimating $\mathbf{L}^*$. Directly minimizing the sum of squared

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

Schatten Norm:	$\|\mathbf{L}\|^{\mathrm{S}}_p \equiv \left(\sum_{i\in[N]} \sigma_i(\mathbf{L})^p\right)^{1/p}$, $p \in (0,\infty)$
Fröbenius Norm:	$\|\mathbf{L}\|_F \equiv \|\mathbf{L}\|^{\mathrm{S}}_2 = \left(\sum_{i\in[N]} \sigma_i(\mathbf{L})^2\right)^{1/2} = \left(\sum_{i\in[N]}\sum_{t\in[T]} L^2_{it}\right)^{1/2}$
Rank Norm:	$\|\mathbf{L}\|_0 \equiv \lim_{p\downarrow 0} \|\mathbf{L}\|^{\mathrm{S}}_p = \sum_{i\in[N]} \mathbf{1}_{\sigma_i(\mathbf{L})>0}$
Nuclear Norm:	$\|\mathbf{L}\|_* \equiv \|\mathbf{L}\|^{\mathrm{S}}_1 = \sum_{i\in[N]} \sigma_i(\mathbf{L})$
Operator Norm:	$\|\mathbf{L}\|_{\mathrm{op}} \equiv \lim_{p\to\infty} \|\mathbf{L}\|^{\mathrm{S}}_p = \max_{i\in[N]} \sigma_i(\mathbf{L}) = \sigma_1(\mathbf{L})$
Max Norm:	$\|\mathbf{L}\|_{\max} \equiv \max_{(i,t)\in[N]\times[T]} \|L_{it}\|$
Element-Wise ℓ_1 Norm:	$\|\mathbf{L}\|_{1,e} \equiv \sum_{(i,t)\in[N]\times[T]} \|L_{it}\|$

differences,

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

does not lead to a useful estimator: if $(i,t) \in \mathcal{M}$ the objective function does not depend on L_{it}, and for pairs $(i,t) \in \mathcal{O}$ the estimator would simply be Y_{it}. To address this we regularize the problem by adding a penalty term $\lambda\|\mathbf{L}\|$ to the objective function in (4.2), for some choice of the norm $\|\cdot\|$ and a scalar λ. However, since we don not wish to regularize the fixed effects (that are included in $\mathbf{L}^*$), we estimate them explicitly by introducing variables $\Gamma \in \mathbb{R}^{N\times 1}$ and $\Delta \in \mathbb{R}^{T\times 1}$, and the variable $\mathbf{L}$ will be used for estimating the remaining low-rank components of $\mathbf{L}^*$. This is conceptually similar to not regularizing the intercept term in LASSO estimator, to reduce the bias created by the regularization term (Hastie et al., 2009).

The estimator: The general form of our proposed estimator for $\mathbf{L}^*$ is $\hat{\mathbf{L}} + \hat{\Gamma}\mathbf{1}^{\top}_T + \mathbf{1}_N\hat{\Delta}^{\top}$ where

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

Compared to the setting studied by Candès and Recht (2009); Candès and Plan (2010); Mazumder et al. (2010), we include the fixed effects Γ and Δ. Although formally the fixed effects can be subsumed in the matrix $\mathbf{L}$ ($\Gamma\mathbf{1}^{\top}_T$ and $\mathbf{1}_N\Delta^{\top}$ are both rank one matrices), in

practice, including these fixed effects separately and not regularizing them greatly improves the quality of the imputations. This is partly because compared to the settings studied in the matrix completion literature the fraction of observed values is relatively high, and so these fixed effects can be estimated accurately. The penalty factor λ will be 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 $\|\mathbf{L}\|_F = \sum_{(i,t)\in[N]\times[T]} L_{it}^2$). 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 optimization problem withs the rank norm 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: For simplicity, let us first assume that there are no fixed effects (so that we do not need to estimate Γ and Δ). The algorithm for calculating our estimator 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,

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

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. For the case that we are estimating fixed effects, after each iteration of obtaining $\mathbf{L}_{k+1}$, we can estimate Γ_{k+1} and Δ_{k+1} by using the first order conditions since they only appear in the squared error term. We would also replace the $\mathbf{P}_\mathcal{O}(\mathbf{Y})$ term in (4.5) by $\mathbf{P}_\mathcal{O}(\mathbf{Y} - \Gamma_k \mathbf{1}_T^\top - \mathbf{1}_N \Delta_k^\top)$.

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.

Confidence intervals: Studying asymptotic distribution of $\mathbf{L}^* - \hat{\mathbf{L}}$ in order to construct confidence intervals is beyond the scope of this paper and is an interesting future research question. However, one can use re-sampling methods to view statistical fluctuations of the imputed matrix. For example, one can again choose K random subsets $\mathcal{O}_k \subset \mathcal{O}$ and construct a cross-validated estimator $\hat{\mathbf{L}}^{(k)}$ for each set $\mathcal{O}_k$. Then, for each entry (i,t) use statistical fluctuations of $\{\hat{L}_{it}^{(k)}\}_{k\in[K]}$ to construct a confidence interval for L_{it}^*, related to the use of permutation metods in the synthetic control literature (Abadie et al. (2010)).

5 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), vertical (synthetic control) and difference-in-differences approaches. To faciliate the discussion, we focus on the case with the set of missing pairs $\mathcal{M}$ 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 depending on the relative magnitude of N and T.

The observed data are $\mathbf{Y}$, an $N \times T$ matrix with the (N,T) entry missing. We can partition this matrix as

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_0 & \mathbf{y}_1 \\ \mathbf{y}_2^\top & ? \end{pmatrix},$$

where $\mathbf{Y}_0$ is a $(N-1)\times(T-1)$ matrix, and $\mathbf{y}_1$ and $\mathbf{y}_2$ are $(N-1)$ and $(T-1)$ component vectors, respectively.

In this case the matrix completion, horizontal regression, vertical regression, synthetic control regression, the elastic net version, and difference-in-differences estimators are very closely related. They can all be charactized as focusing on the exact same objective function, but differing in the regularization and additional restrictions imposed on the parameters of the objective function.

To make this precise, define for a given positive integer R, an $N \times R$ matrix $\mathbf{A}$, an $T \times R$ matrix $\mathbf{B}$, an N-vector γ and a T-vector δ the objective function

$$Q(\mathbf{Y}; R, \mathbf{A}, \mathbf{B}, \gamma, \delta) = \frac{1}{|\mathcal{O}|} \left\| P_{\mathcal{O}} \left(\mathbf{Y} - \mathbf{A}\mathbf{B}^\top - \gamma \mathbf{1}_T^\top - \mathbf{1}_N \delta^\top \right) \right\|_F^2 \tag{5.1}$$

For any pair of positive integers K and L, let $\mathbb{M}^{K,L}$ be the set of all $K \times L$ real-valued matrices. When $R = 0$, we take the product $\mathbf{A}\mathbf{B}^\top$ to be the $N \times T$ matrix with all elements equal to zero. First note that simply minimizing $Q(\mathbf{Y}; R, \mathbf{A}, \mathbf{B}, \gamma, \delta)$ over the rank R, the matrices $\mathbf{A}$, $\mathbf{B}$ and the vectors γ and δ,

$$\min_{R\in\{0,1,\ldots,\min(N,T)\}} \quad \min_{\mathbf{A}\in\mathbb{M}^{N,R},\mathbf{B}\in\mathbb{M}^{T,R},\gamma\in\mathbb{M}^{N,1},\delta\in\mathbb{M}^{T,1}} Q(\mathbf{Y}; R, \mathbf{A}, \mathbf{B}, \gamma, \delta),$$

has multiple solutions for the imputations $\hat{\mathbf{Y}}_{NT}$ where $\hat{\mathbf{Y}} = \mathbf{A}\mathbf{B}^\top + \gamma\mathbf{1}_T^\top + \mathbf{1}_N\delta^\top$. By choosing the rank R to the minimum of N and T, we can find for any pair γ and δ a solution for $\mathbf{A}$

and $\mathbf{B}$ such that $P_{\mathcal{O}}\left(\mathbf{Y} - \mathbf{A}\mathbf{B}^{\top} - \gamma \mathbf{1}_T^{\top} - \mathbf{1}_N \delta^{\top}\right)$ has all elements equal to zero, with different values for $\hat{\mathbf{Y}}_{NT}$.

The implication is that we need to add some structure to the optimization problem. The next result shows that horizontal regression, vertical regression, the Abadie-Diamond-Hainmueller synthetic control estimator, the difference-in-differences estimator, and the nuclear norm minimization matrix completion can all be expressed as minimizing $Q(\mathbf{Y}; R, \mathbf{A}, \mathbf{B}, \gamma, \delta)$ under different restrictions on, or with different approaches to regularization of the unknown parameters $(R, \mathbf{A}, \mathbf{B}, \gamma, \delta)$. The following theorem lays out these differences in hard restrictions and regularization approaches. Here the minimization for R is over the set $\{0, 1, 2, \ldots, \min(T, N)\}$, and the minimization for $\mathbf{A}$ and $\mathbf{B}$ is over the sets $\mathbb{M}^{N,R}$ and $\mathbb{M}^{T,R}$ respectively.

Theorem 1. *In the case with only the (N, T) entry missing, we have,*
(i) (nuclear norm matrix completion)

$$(R^{\text{mc-nnm}}, \mathbf{A}_{\lambda}^{\text{mc-nnm}}, \mathbf{B}_{\lambda}^{\text{mc-nnm}}, \gamma_{\lambda}^{\text{mc-nnm}}, \delta_{\lambda}^{\text{mc-nnm}}) = \text{argmin}_{R,\mathbf{A},\mathbf{B},\gamma,\delta} \left\{ Q(\mathbf{Y}; R, \mathbf{A}, \mathbf{B}, \gamma, \delta) + \frac{\lambda}{2}\|\mathbf{A}\|_F^2 + \frac{\lambda}{2}\|\mathbf{B}\|_F^2 \right\},$$

(ii) (horizontal regression, defined if $N > T$)

$$(R^{\text{hr}}, \mathbf{A}^{\text{hr}}, \mathbf{B}^{\text{hr}}, \gamma^{\text{hr}}, \delta^{\text{hr}}) = \text{argmin}_{R,\mathbf{A},\gamma,\delta} Q(\mathbf{Y}; R, \mathbf{A}, \mathbf{B}, \gamma, \delta),$$

subject to

$$R = T - 1, \quad \mathbf{A} = \begin{pmatrix} \mathbf{Y}_0 \\ \mathbf{y}_2^{\top} \end{pmatrix}, \quad \gamma = 0, \quad \delta_1 = \delta_2 = \ldots = \delta_{T-1} = 0,$$

(iii) (vertical regression, defined if $T > N$),

$$(R^{\text{vt}}, \mathbf{A}^{\text{vt}}, \mathbf{B}^{\text{vt}}, \gamma^{\text{vt}}, \delta^{\text{vt}}) = \text{argmin}_{R,\mathbf{A},\mathbf{B},\gamma,\delta} Q(\mathbf{Y}; R, \mathbf{A}, \mathbf{B}, \gamma, \delta),$$

subject to

$$R = N - 1, \quad \mathbf{B} = \begin{pmatrix} \mathbf{Y}_0^{\top} \\ \mathbf{y}_1^{\top} \end{pmatrix}, \quad \gamma_1 = \gamma_2 = \ldots = \gamma_{N-1} = 0, \quad \delta = 0,$$

(iv) (synthetic control),

$$(R^{\text{sc-adh}}, \mathbf{A}^{\text{sc-adh}}, \mathbf{B}^{\text{sc-adh}}, \gamma^{\text{sc-adh}}, \delta^{\text{sc-adh}}) = \text{argmin}_{R,\mathbf{A},\mathbf{B},\gamma,\delta} Q(\mathbf{Y}; R, \mathbf{A}, \mathbf{B}, \gamma, \delta),$$

subject to

$$R = N - 1, \quad \mathbf{B} = \begin{pmatrix} \mathbf{Y}_0^{\top} \\ \mathbf{y}_1^{\top} \end{pmatrix}, \quad \delta = 0, \quad \gamma = 0, \quad \forall\ i, A_{iT} \geq 0, \ \sum_{i=1}^{N-1} A_{iT} = 1,$$

(*v*) *(vertical regression, elastic net),*

$$(R^{\text{vt-en}}, \mathbf{A}^{\text{vt-en}}, \mathbf{B}^{\text{vt-en}}, \gamma^{\text{vt-en}}, \delta^{\text{vt-en}}) = \text{argmin}_{R,\mathbf{A},\mathbf{B},\gamma,\delta} \left\{ Q(\mathbf{Y}; R, \mathbf{A}, \mathbf{B}, \gamma, \delta) + \lambda \left[\frac{1-\alpha}{2} \left\| \begin{pmatrix} \mathbf{a}_2 \\ \mathbf{a}_3 \end{pmatrix} \right\|_F^2 + \alpha \left\| \begin{pmatrix} \mathbf{a}_2 \\ \mathbf{a}_3 \end{pmatrix} \right\|_1 \right] \right\},$$

subject to

$$R = N - 1, \quad \mathbf{B} = \begin{pmatrix} \mathbf{Y}_0^\top \\ \mathbf{y}_1^\top \end{pmatrix}, \quad \gamma_1 = \gamma_2 = \ldots = \gamma_{N-1} = 0, \quad \delta = 0,$$

where $\mathbf{A}$ *is partitioned as*

$$\mathbf{A} = \begin{pmatrix} \tilde{\mathbf{A}} & \mathbf{a}_1 \\ \mathbf{a}_2^\top & \mathbf{a}_3 \end{pmatrix},$$

(*vi*) *(difference-in-differences regression),*

$$(R^{\text{did}}, \mathbf{A}^{\text{did}}, \mathbf{B}^{\text{did}}, \gamma^{\text{did}}, \delta^{\text{did}}) = \text{argmin}_{R,\mathbf{A},\mathbf{B},\gamma,\delta} Q(\mathbf{Y}; R, \mathbf{A}, \mathbf{B}, \gamma, \delta),$$

subject to

$$R = 0.$$

The proof for this result is in §A.1.

Comment 1. There is no unique solution to minimizing $Q(\mathbf{Y}; \mathbf{A}, \mathbf{B})$ if we also minimize over the rank R. The nuclear norm estimator uses regularization to get around this by regularizing $\mathbf{A}$ and $\mathbf{B}$. The other estimators impose restrictions instead of (or in combination with) regularizing the estimators, while fixing R as a function of N and T. The restrictions for the horizontal regression on the one hand, and for the vertical regression, synthetic control and elastic net regression on the other hand, are quite different, and not directly comparable. However in other settings researchers have found that it is often better to regularize estimators than to impose hard restrictions. We find the same in our simulations below.

Comment 2. For nuclear norm matrix completion representation a key insight is that (Lemma 6, Mazumder et al. (2010))

$$\|\mathbf{L}\|_* = \min_{\mathbf{A},\mathbf{B}:\mathbf{L}=\mathbf{A}\mathbf{B}^\top} \frac{1}{2} \left(\|\mathbf{A}\|_F^2 + \|\mathbf{A}\|_F^2 \right).$$

In addition, if $\hat{\mathbf{L}}$ is the solution to Equation (4.3) that has rank $\hat{R}$, then one solution for $\mathbf{A}$ and $\mathbf{B}$ is given by

$$\mathbf{A} = \mathbf{S}\mathbf{\Sigma}^{1/2}, \quad \mathbf{B} = \mathbf{R}\mathbf{\Sigma}^{1/2} \tag{5.2}$$

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

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

where $\hat{\beta}$ is

$$(\hat{\beta}, \hat{\delta}_T) = \arg\min_{\beta, \delta_T} \sum_{i=1}^{N-1} \left(Y_{iT} - \delta_T - \sum_{t=1}^{T-1} \beta_t Y_{it} \right)^2.$$

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

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

where

$$(\hat{\alpha}, \hat{\gamma}_N) = \arg\min_{\alpha, \gamma_N} \sum_{t=1}^{T-1} \left(Y_{Nt} - \gamma_N - \sum_{i=1}^{N-1} \alpha_i Y_{it} \right)^2.$$

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. The synthetic control estimator further restricts the values of the γ_N and α_i. □

Comment 4. 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. □

6 Theoretical Bounds for the Estimation Error

In this section we focus on the case that there are no covariates or fixed effects, 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 $\mathbf{L}^*$.

$$\hat{\mathbf{L}} = \underset{\mathbf{L}: \|\mathbf{L}\|_{\max} \leq L_{\max}}{\arg\min} \left\{ \frac{1}{|\mathcal{O}|} \|\mathbf{P}_{\mathcal{O}}(\mathbf{Y} - \mathbf{L})\|_F^2 + \lambda \|\mathbf{L}\|_* \right\}. \tag{6.1}$$

6.1 Additional Notation

First, we start by introducting some new notation. Recall that for each positive integer n notation $[n]$ refers to the set of integers $\{1, 2, \ldots, n\}$. For any two real numbers a and b, we denote their maximum by $a \vee b$. 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 i^{th} 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 $\{t_i\}_{i \in [N]}$ on $[T]$ with distributions $\{\pi^{(i)}\}_{i \in [N]}$. 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 $\{\pi^{(i)}\}_{i \in [N]}$. Now, $\mathcal{O}$ can be written as $\mathcal{O} = \bigcup_{i=1}^N \Big\{(i, 1), (i, 2), \ldots, (i, t_i)\Big\}$. The equivalent of the unconfoundedness assumption in the program evaluation literature is that the adoption dates are independent of each other and of the idiosyncratic part of the outcomes, conditional on the systematic part. Formally, we make the following assumption:

Assumption 2. *Conditional on $\mathbf{L}^*$, the adoption dates t_i are independent of each other and of $\boldsymbol{\varepsilon}$.*

Remark 6.1. *This assumption is similar to the unconfoundedness assumption. In the setting where researchers use that assumption, with a single treated period, the only stochastic component of $\mathbf{W}$ is the last column. In that case the assumption is that conditional on the first $T-1$ rows of $\mathbf{Y}$, the last column of the assignment $\mathbf{W}$ is independent of the last column of $\mathbf{Y}$. As we show in §5, in the unconfoundedness approach the first $T-1$ columns of the matrix $\mathbf{L}$ are taken to be identical to the first $T-1$ columns of the matrix $\mathbf{Y}$ (and the last column of $\mathbf{L}$ is a linear combination of the first $T-1$ columns), so the conditioning on the first $T-1$ columns of $\mathbf{Y}$ is identical to conditioning on $\mathbf{L}$.*

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} satisfy Assumptions 1-2.

Note that the number of control units (N_c) is equal to the number of rows that have all entries observed (i.e., $N_c = \sum_{i=1}^N \mathbb{I}_{\{t_i = T\}}$). Therefore, the expected number of control units can be written as $\mathbb{E}_\pi[N_c] = \sum_{i=1}^N \pi_T^{(i)}$. Defining

$$p_c \equiv \min_{1 \le i \le N} \pi_T^{(i)}\,,$$

we expect to have (on average) at least $N p_c$ control units. The parameter p_c will play an important role in our main theoretical results. To provide some intuition, assume $\mathbf{L}^*$ is a matrix that is zero everywhere except in its i^{th} row. Such $\mathbf{L}^*$ is clearly low-rank. But recovering the entry L^*_{iT} is impossible when $i_t < T$ which means $\pi_T^{(i)}$ cannot be too small. Since i is arbitrary, in general, p_c cannot be too small.

Remark 6.2. *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 6.3. *The estimator* (6.1) *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 §8.4, we discuss how the estimator can be modified by considering a weighted loss function based on propensity scores for the missing entries.*

6.2 Main Result

The main result of this section is the next theorem (proved in §A.2) that provides an upper bound for $\|\mathbf{L}^* - \hat{\mathbf{L}}\|_F/\sqrt{NT}$, the root-mean-squared-error (RMSE) of the estimator $\hat{\mathbf{L}}$.

Theorem 2. *Suppose Assumptions 1 and 2 hold, rank of $\mathbf{L}^*$ is R, $T \geq C_0 \log(N+T)$ for a constant C_0, and the penalty parameter λ is a constant multiple of*

$$\frac{\sigma\left[\sqrt{N\log(N+T)} \vee \sqrt{T\log^3(N+T)}\right]}{|\mathcal{O}|}.$$

Then there is a constant C such that with probability greater than $1-2(N+T)^{-2}$,

$$\frac{\|\mathbf{L}^* - \hat{\mathbf{L}}\|_F}{\sqrt{NT}} \leq C\left[\left(\sqrt{\frac{L_{\max}^2 \log(N+T)}{N\,p_c}} \vee \sqrt{\frac{\sigma^2 R\log(N+T)}{T\,p_c^2}} \vee \sqrt{\frac{\sigma^2 R\log^3(N+T)}{N\,p_c^2}}\right) + \sqrt{\frac{L_{\max}^2 RT}{Np_c^2}}\right]. \quad (6.2)$$

Interpretation of Theorem 2: 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 (6.2) converges to 0 when $\mathbf{L}^*$ is low-rank (R is constant), $N \geq T$, and $p_c \gg (\sqrt{1/T} \vee \sqrt{T/N})\log^{3/2}(N+T)$. For example, when T is the same order as $N^{1/3}$, a sufficient condition for the latter is that the lower bound for the average number of control units (Np_c) grows larger than a constant times $N^{5/6}\log^{3/2}(N)$. 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 time series dependency structure. In particular this includes the staggered adoption setting where if $(i,t) \in \mathcal{O}$ then $(i,t') \in \mathcal{O}$ for all $t' < t$. The impact of this additional correlation is that the estimation error deteriorates significantly compared to the ones in prior literature. For example, as discussed above, in the case of $N^{1/3} = T$, in order to have a consistent estimation we need more data. Specifically, a factor $N^{5/6}$ (up to logarithmic factors) more entries per column should be observed, than in the matrix completion literature.

Remark 6.4. *We note that in statement of Theorem 2, the lower bound on λ depends on $\mathcal{O}$ which is a random variable. The left hand side of the inequality* (6.2) *is also random, depending on $\mathcal{O}$ and the noise, but the right hand side of* (6.2) *is deterministic. In order to understand the role of randomness, we describe the main three steps of the proof. First, in Lemma 1, we prove a deterministic upper bound for $\sum_{(i,t)\in\mathcal{O}} \langle \mathbf{A}_{it}, \mathbf{L}^* - \hat{\mathbf{L}} \rangle^2 / |\mathcal{O}|$ that holds for every realization of the random variable $\mathcal{O}$, when λ grows by operator norm of a certain error matrix, $\sum_{(i,t)\in\mathcal{O}} \varepsilon_{it} \mathbf{A}_{it}$. Next, in Lemma 2, we use randomness of $\mathcal{O}$ and noise to prove a probabilistic bound on the operator norm of this error matrix. The final step, Lemma 3, also uses randomness of $\mathcal{O}$ and noise to show that $\sum_{(i,t)\in\mathcal{O}} \langle \mathbf{A}_{it}, \mathbf{L}^* - \hat{\mathbf{L}} \rangle^2 / |\mathcal{O}|$ concentrates and (with high probability) is larger than a constant fraction of its expectation up to an additive constant.*

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 five 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. Although this estimator is not necessarily well-defined if $N \gg T$, the restrictions ensured that it was well-defined in all the settings we used.

- **MC-NNM**: Our proposed matrix completion approached via nuclear norm minimization, explained in §4 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.

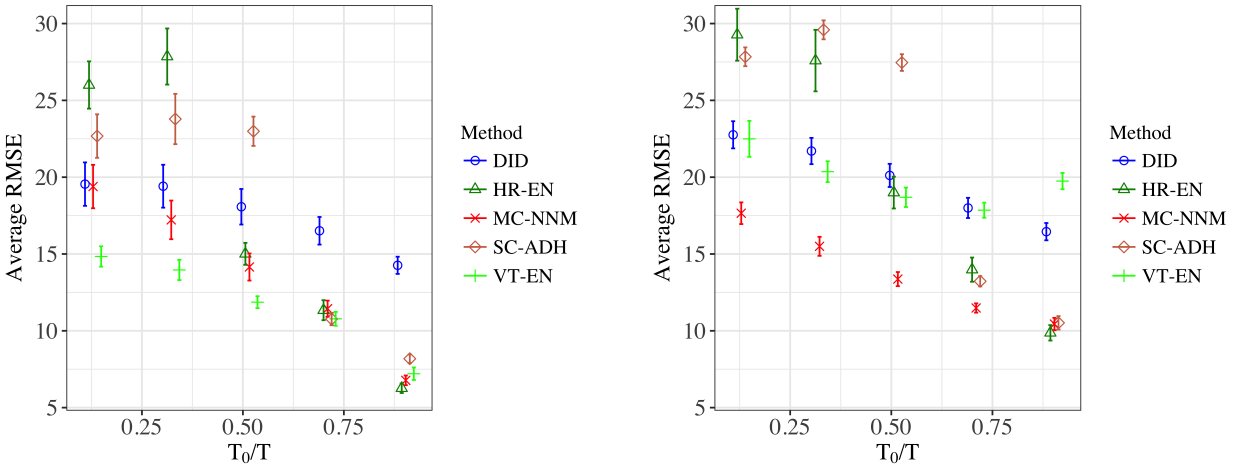
7.1 The Abadie-Diamond-Hainmueller California Smoking Data

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 randomly selected N_t units adopt the treatment in period $T_0 + 1$, and the remaining units never adopt the treatment.
- Case 2: Staggered adoption where randomly N_t units adopt the treatment in some period after period T, with the actual adoption date varying randomly 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 95% sampling 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, $(N, T) = (490, 10)$, ..., $(N, T) = (70, 70)$, ..., $(N, T) = (10, 490)$, 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.

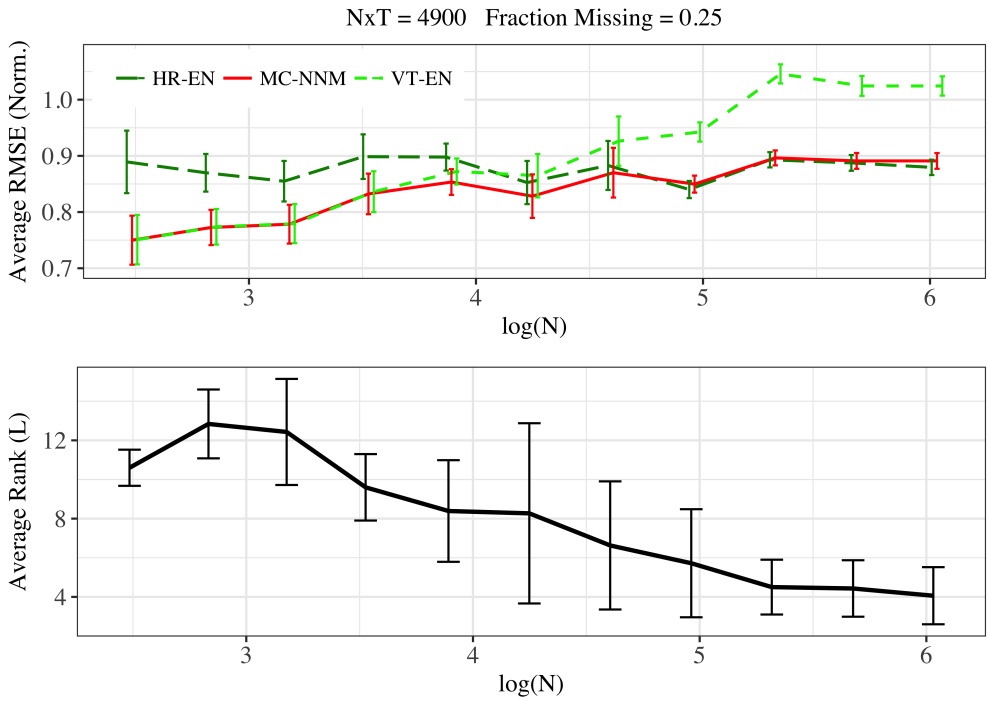


Figure 2: Stock Market Data

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 for the same reason: it is trying to do the horizontal regression with too few observations relative to the number of regressors. 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 §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 Leveraging Data From Treated Units

In previous sections we only focused on imputing $\mathbf{Y}(0)$ to solve the treatment effect estimation problem. We note that this approach allows for very general assumptions on the treatment effect. For example if treatment effect has no (low-dimensional) patterns, imputing $\mathbf{Y}(0)$ is the best one can do because $\mathbf{Y}(1)$ would not have any pattern that can be used for imputation. We also note that in many of the applications there are very few treated unit/periods, so imputing the missing entries in $\mathbf{Y}(1)$ would be much more challenging in practice.

However, when the treatment effect is constant or has a low-rank pattern we can extend our approach and leverage the additional data from $\mathbf{Y}(1)$. We describe these next.

(a) **When treatment effect is constant.** If the treatment effect is constant for every pair (i,t), then we can consider the following natural extension of our estimator (4.3).

$$(\hat{\mathbf{L}}, \hat{\Gamma}, \hat{\Delta}, \hat{\tau}) = \arg\min_{\mathbf{L},\Gamma,\Delta,\tau} \left\{ \frac{1}{NT} \|\mathbf{Y} - \mathbf{L} - \Gamma \mathbf{1}_T^\top - \mathbf{1}_N \Delta^\top - \tau \mathbf{W}\|_F^2 + \lambda \|\mathbf{L}\|_* \right\}, \quad (8.5)$$

where variable $\tau \in \mathbb{R}$ is used for estimating the constant treatment effect. Also, recall that $\mathbf{W}$ is the binary treatment matrix. Note that here the squared error term includes all entries $(i,t) \in [N] \times [T]$.

(b) **When treatment effect has a low-rank pattern.** Assume the treatment effect is not constant but is such that the matrix $\mathbf{Y}(1)$ has a low-rank expectation. Then we can impute $\mathbf{Y}(1)$ the same way we impute $\mathbf{Y}(0)$, using our estimator (4.3) applied to treated entries. Then we can use imputed matrix $\hat{\mathbf{Y}}(0)$ and $\hat{\mathbf{Y}}(1)$ to estimate the treatment effect matrix $\mathbf{Y}(1) - \mathbf{Y}(0)$.

8.3 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 $\boldsymbol{\varepsilon}_{it}$ are correlated over time. 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 i^{th} 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 $\Omega_{ts} = \sigma^2 \rho^{|t-s|}$, 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

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

8.4 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:

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

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

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

8.5 Relaxing the Dependence of Theorem 2 on $p_{\mathbf{c}}$

Recall from §6.1 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 2 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 §6.1 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 $\sum_{i=1}^N \pi_T^{(i)}/N$.

8.6 Nearly Low-rank Matrices

Another possible extension of Theorem 2 is to the cases where $\mathbf{L}^*$ may have high rank, but most of its singular values are small. More formally, if $\sigma_1 \ge \cdots > \sigma_{\min(N,T)}$ 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).

8.7 Additional Missing Entries

In §6.1 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. In this case, one can extend Theorem 2 to the setting with $\mathcal{O} = \left[\bigcup_{i=1}^{N} \Big\{(i,1),(i,2),\ldots,(i,t_i)\Big\}\right] \setminus \mathcal{O}_{\text{miss}}$, 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 proposed Abadie et al. (2010) and the elastic net estimators proposed by Doudchenko and Imbens (2016).

References

A Abadie and MD Cattaneo. Econometric methods for program evaluation. *Annual Review of Economics*, 18, 2018.

Alberto Abadie. Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, 2019.

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

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*, pages 495–510, 2015.

Theodore Wilbur Anderson. *An introduction to multivariate statistical analysis*, volume 2. Wiley New York, 1958.

Joshua Angrist and Steve Pischke. *Mostly Harmless Econometrics: An Empiricists' Companion*. Princeton University Press, 2008.

Manuel Arellano and Bo Honoré. Panel data models: some recent developments. *Handbook of econometrics*, 5:3229–3296, 2001.

Dmitry Arkhangelsky, Susan Athey, David A Hirshberg, Guido W Imbens, and Stefan Wager. Synthetic difference in differences. Technical report, National Bureau of Economic Research, 2019.

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

Susan Athey and Scott Stern. The impact of information technology on emergency health care outcomes. *The RAND Journal of Economics*, 33(3):399–432, 2002.

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

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

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

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

Eli Ben-Michael, Avi Feller, and Jesse Rothstein. The augmented synthetic control method. *arXiv preprint arXiv:1811.04170*, 2018.

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

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

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

Emmanuel J. Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inf. Theor.*, 56(5):2053–2080, 2010. ISSN 0018-9448.

Gary Chamberlain. Panel data. *Handbook of econometrics*, 2:1247–1318, 1984.

Gary Chamberlain et al. Feedback in panel data medels. Technical report, Harvard-Institute of Economic Research, 1993.

Victor Chernozhukov, Kaspar Wuthrich, and Yinchu Zhu. An exact and robust conformal inference method for counterfactual and synthetic controls. *arXiv preprint arXiv:1712.09089*, 2017.

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.

Bruno Ferman and Cristine Pinto. Synthetic controls with imperfect pre-treatment fit. *arXiv preprint arXiv:1911.08521*, 2019.

D. Gamarnik and S. Misra. A note on alternating minimization algorithm for the matrix completion problem. *IEEE Signal Processing Letters*, 23(10):1340–1343, 2016.

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

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

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

Nima Hamidi and Mohsen Bayati. On Low-rank Trace Regression under General Sampling Distribution. *arXiv e-prints*, art. arXiv:1904.08576, Apr 2019.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. New York: Springer, 2009.

Trevor Hastie, Rahul Mazumder, Jason D. Lee, and Reza Zadeh. Matrix completion and low-rank svd via fast alternating least squares. *J. Mach. Learn. Res.*, 16(1):3367–3402, 2015. ISSN 1532-4435.

Miguel A Hernan and James M Robins. *Causal inference*. CRC Boca Raton, FL:, 2010.

Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

Cheng Hsiao, H Steve Ching, and Shui Ki Wan. 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, 2012.

Guido Imbens and Jeffrey Wooldridge. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1):5–86, 2009.

Guido W Imbens and Donald B Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

Raghunandan H. Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE Trans. Inf. Theor.*, 56(6):2980–2998, June 2010a.

Raghunandan H. Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from noisy entries. *J. Mach. Learn. Res.*, 11:2057–2078, August 2010b.

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

Olga Klopp. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 20(1):282–303, 2014. ISSN 1350-7265.

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

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

Kathleen T Li. Statistical inference for average treatment effects estimated by synthetic control methods. *Journal of the American Statistical Association*, pages 1–16, 2019.

Kung-Yee Liang and Scott L Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

Pascal Massart et al. About the constants in talagrand's concentration inequalities for empirical processes. *The Annals of Probability*, 28(2):863–884, 2000.

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.

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

Hyungsik Roger Moon and Martin Weidner. Dynamic linear panel regression models with interactive fixed effects. *Econometric Theory*, 33(1):158–195, 2017.

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. *Journal of Machine Learning Research*, 13(May):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.

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

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

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

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

Donald B Rubin. *Matched sampling for causal effects*. Cambridge University Press, 2006.

Azeem Shaikh and Panagiotis Toulis. Randomization tests in observational studies with staggered adoption of treatment. *University of Chicago, Becker Friedman Institute for Economics Working Paper*, (2019-144), 2019.

Nathan Srebro, Noga Alon, and Tommi S. Jaakkola. Generalization error bounds for collaborative prediction with low-rank matrices. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 1321–1328. 2005.

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

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

Yixin Wang, Dawen Liang, Laurent Charlin, and David M. Blei. The Deconfounded Recommender: A Causal Inference Approach to Recommendation. *arXiv e-prints*, art. arXiv:1808.06581, Aug 2018.

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

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

A.1 Proof of Theorem 1

To prove part (i), we first state that if

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

and

$$(\hat{R},\hat{\mathbf{A}},\hat{\mathbf{B}}) = \arg\min_R \arg\min_{\mathbf{A}\in\mathbb{M}^{N,R},\mathbf{B}\in\mathbb{M}^{T,R}} Q(\mathbf{Y};R,\mathbf{A},\mathbf{B},\gamma,\delta) + \frac{\lambda}{2}\|\mathbf{A}\|_F^2 + \frac{\lambda}{2}\|\mathbf{B}\|_F^2,$$

then

$$\hat{\mathbf{L}} = \hat{\mathbf{A}}\hat{\mathbf{B}}^\top. \tag{A.1}$$

This follows from the fact (Lemma 6, Mazumder et al. (2010)) that

$$\|\mathbf{L}\|_* = \min_{\mathbf{A},\mathbf{B}:\mathbf{L}=\mathbf{A}\mathbf{B}^\top} \frac{1}{2}\left(\|\mathbf{A}\|_F^2 + \|\mathbf{A}\|_F^2\right).$$

Now define

$$\hat{\mathbf{L}}(\gamma,\delta) = \underset{\mathbf{L}:\|\mathbf{L}\|_{\max}\leq L_{\max}}{\arg\min} \left\{ \frac{1}{|\mathcal{O}|}\|\mathbf{P}_{\mathcal{O}}(\mathbf{Y}-\mathbf{L}-\gamma\mathbf{1}_T^\top-\mathbf{1}_N\delta^\top)\|_F^2 + \lambda\|\mathbf{L}\|_* \right\},$$

and

$$(\hat{R}(\gamma,\delta),\hat{\mathbf{A}}(\gamma,\delta),\hat{\mathbf{B}}(\gamma,\delta)) =$$

$$\arg\min_R \arg\min_{\mathbf{A}\in\mathbb{M}^{N,R},\mathbf{B}\in\mathbb{M}^{T,R}} Q(\mathbf{Y};R,\mathbf{A},\mathbf{B},\gamma,\delta) + \frac{\lambda}{2}\|\mathbf{A}\|_F^2 + \frac{\lambda}{2}\|\mathbf{B}\|_F^2$$

$$= \arg\min_R \arg\min_{\mathbf{A}\in\mathbb{M}^{N,R},\mathbf{B}\in\mathbb{M}^{T,R}} Q(\mathbf{Y}-\gamma\mathbf{1}_T^\top-\mathbf{1}_N\delta^\top;R,\mathbf{A},0,0) + \frac{\lambda}{2}\|\mathbf{A}\|_F^2 + \frac{\lambda}{2}\|\mathbf{B}\|_F^2,$$

which, by (A.1) implies

$$\mathbf{L}(\gamma,\delta) = \mathbf{A}(\gamma,\delta)\mathbf{B}(\gamma,\delta)^\top).$$

Define

$$S(\mathbf{Y};\mathbf{L},\gamma,\delta) = \frac{1}{|\mathcal{O}|}\|\mathbf{P}_{\mathcal{O}}(\mathbf{Y}-\mathbf{L}-\gamma\mathbf{1}_T^\top-\mathbf{1}_N\delta^\top)\|_F^2 + \lambda\|\mathbf{L}\|_*.$$

Also define

$$(\hat{\mathbf{L}},\hat{\gamma},\hat{\delta}) = \underset{\mathbf{L}:\|\mathbf{L}\|_{\max}\leq L_{\max},\gamma,\delta}{\arg\min} \left\{ \frac{1}{|\mathcal{O}|}\|\mathbf{P}_{\mathcal{O}}(\mathbf{Y}-\mathbf{L}-\gamma\mathbf{1}_T^\top-\mathbf{1}_N\delta^\top)\|_F^2 + \lambda\|\mathbf{L}\|_* \right\},$$

$$(\tilde{R},\tilde{\mathbf{A}},\tilde{\mathbf{B}},\tilde{\gamma},\tilde{\delta}) =$$

$$\arg\min_R \arg\min_{\mathbf{A}\in\mathbb{M}^{N,R},\mathbf{B}\in\mathbb{M}^{T,R},\gamma,\delta} Q(\mathbf{Y};R,\mathbf{A},\mathbf{B},\gamma,\delta) + \frac{\lambda}{2}\|\mathbf{A}\|_F^2 + \frac{\lambda}{2}\|\mathbf{B}\|_F^2,$$

$$\hat{Q} = \frac{1}{|\mathcal{O}|}\|\mathbf{P}_{\mathcal{O}}(\mathbf{Y} - \hat{\mathbf{L}} - \hat{\gamma}\mathbf{1}_T^\top - \mathbf{1}_N\hat{\delta}^\top)\|_F^2 + \lambda\|\mathbf{L}\|_*,$$

and

$$\tilde{Q} = Q(\mathbf{Y}; \tilde{R}, \tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\gamma}, \tilde{\delta}).$$

In order to prove that $\hat{\mathbf{L}} = \tilde{\mathbf{A}}\tilde{\mathbf{B}}^\top$, we prove that $(\hat{\gamma}, \hat{\delta}) = (\tilde{\gamma}, \tilde{\delta})$.

Suppose $(\hat{\gamma}, \hat{\delta}) \neq (\tilde{\gamma}, \tilde{\delta})$. Then

$$\tilde{Q} = Q(\mathbf{Y}; \tilde{R}, \tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\gamma}, \tilde{\delta}) < Q(\mathbf{Y}; \hat{R}(\hat{\gamma}, \hat{\delta}), \hat{\mathbf{A}}(\hat{\gamma}, \hat{\delta}), \hat{\mathbf{B}}(\hat{\gamma}, \hat{\delta}), \hat{\gamma}, \hat{\delta}) = \hat{Q}.$$

But, also

$$\hat{Q} = S(\mathbf{Y}; \hat{\mathbf{L}}, \hat{\gamma}, \hat{\delta}) < S(\mathbf{Y}; \hat{\mathbf{L}}(\tilde{\gamma}, \tilde{\delta}), \tilde{\gamma}, \tilde{\delta}) = Q(\mathbf{Y}; \tilde{R}, \tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\gamma}, \tilde{\delta}),$$

which leads to a contradiction. Hence $(\hat{\gamma}, \hat{\delta}) = (\tilde{\gamma}, \tilde{\delta})$.

Next, consider part (ii). Consider minimizing

$$\min_{R} \min_{\mathbf{A}\in\mathbb{M}^{N,T-1}, \mathbf{B}\in\mathbb{M}^{T,T-1}, \gamma, \delta} Q(\mathbf{Y}; R, \mathbf{A}, \mathbf{B}, \gamma, \delta), \tag{A.2}$$

subject to

$$R = T - 1, \quad \mathbf{A} = \begin{pmatrix} \mathbf{Y}_0 \\ \mathbf{y}_2^\top \end{pmatrix}, \quad \gamma = 0, \quad \delta_1 = \delta_2 = \ldots = \delta_{T-1} = 0.$$

Partition $\mathbf{B}$ into

$$\mathbf{B} = \begin{pmatrix} \mathbf{B}_0 \\ \mathbf{b}_2^\top \end{pmatrix}.$$

Substituting for the $Q(\cdot)$ and for the restricted parameters, the minimization problem in (A.2) is identical to

$$\min_{\mathbf{B}_0\in\mathbb{M}^{T-1,T-1}, \mathbf{b}_2\in\mathbb{M}^{T-1,1}, \delta_T} \left\| P_{\mathcal{O}}\left(\begin{pmatrix} \mathbf{Y}_0 & \mathbf{y}_1 \\ \mathbf{y}_2^\top & ? \end{pmatrix} - \begin{pmatrix} \mathbf{Y}_0 \\ \mathbf{y}_2^\top \end{pmatrix} \begin{pmatrix} \mathbf{B}_0 \\ \mathbf{b}_2^\top \end{pmatrix}^\top - \mathbf{1}_N \begin{pmatrix} 0 & 0 & \ldots & 0 & \delta_T \end{pmatrix} \right) \right\|_F^2$$

$$= \min_{\mathbf{B}_0\in\mathbb{M}^{T-1,T-1}, \mathbf{b}_2\in\mathbb{M}^{T-1,1}, \delta_T} \left\| P_{\mathcal{O}}\left(\begin{pmatrix} \mathbf{Y}_0 - \mathbf{Y}_0\mathbf{B}_0 & \mathbf{y}_1 - \mathbf{Y}_0\mathbf{b}_2^\top - \mathbf{1}_{N-1}\delta_T \\ \mathbf{y}_2^\top - \mathbf{y}_2^\top\mathbf{B}_0 & ? \end{pmatrix} \right) \right\|_F^2$$

$$= \min_{\mathbf{B}_0\in\mathbb{M}^{T-1,T-1}, \mathbf{b}_2\in\mathbb{M}^{T-1,1}, \delta_T} \left\{ \|\mathbf{Y}_0 - \mathbf{Y}_0\mathbf{B}_0\|_F^2 + \left\|\mathbf{y}_2^\top - \mathbf{y}_2^\top\mathbf{B}_0\right\|_F^2 + \left\|\mathbf{y}_1 - \mathbf{Y}_0\mathbf{b}_2^\top - \mathbf{1}_{N-1}\delta_T\right\|_F^2 \right\}$$

$$= \min_{\mathbf{B}_0\in\mathbb{M}^{T-1,T-1}} \left\{ \|\mathbf{Y}_0 - \mathbf{Y}_0\mathbf{B}_0\|_F^2 + \left\|\mathbf{y}_2^\top - \mathbf{y}_2^\top\mathbf{B}_0\right\|_F^2 \right\} + \min_{\mathbf{b}_2\in\mathbb{M}^{T-1,1}, \delta_T} \left\|\mathbf{y}_1 - \mathbf{Y}_0\mathbf{b}_2^\top - \mathbf{1}_{N-1}\delta_T\right\|_F^2.$$

The solution for $\mathbf{B}_0$ is the $(T-1)\times(T-1)$ dimensional identity matrix. The solution for $\mathbf{B}_2$ and δ_T are the solution to the regression of $\mathbf{y}_1$ on a constant and $\mathbf{Y}_0$, which proves the second part.

The remaining parts follow the same argument. □

A.2 Proof of Theorem 2

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* $\lambda \geq 3\|\mathfrak{E}\|_{\text{op}}/|\mathcal{O}|$,

$$\sum_{(i,t)\in\mathcal{O}} \frac{\langle \mathbf{A}_{it}, \mathbf{L}^* - \hat{\mathbf{L}}\rangle^2}{|\mathcal{O}|} \leq 10\lambda\sqrt{R}\,\|\mathbf{L}^* - \hat{\mathbf{L}}\|_F\,. \tag{A.3}$$

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).

Remark A.1. *We also note that, while* $\mathcal{O}$ *is a random variable, Lemma 1 makes a deterministic statement. Specifically, its result holds for any realization of the random variable* $\mathcal{O}$*. In fact, its proof is based on linear algebra facts and is not using any probabilistic argument.*

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.3) is larger than a constant fraction of $\|\hat{\mathbf{L}} - \mathbf{L}^*\|_F^2$ up to an additive term. 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 is similar to the proof of Theorem 1 in (Negahban and Wainwright, 2012), Lemma 12 in (Klopp, 2014), or Corollary 3.1 in (Hamidi and Bayati, 2019) for the cases that observation process $\mathcal{O}$ does not have a dependency structure, like in our setting. Therefore, for completeness we provide a summary of this proof (adapted to our setting) in §A.5.

Lemma 3. *If the estimator* $\hat{\mathbf{L}}$ *satisfies* $\|\hat{\mathbf{L}}-\mathbf{L}^*\|_F^2 \geq 4L_{\max}^2\theta/p_c$ *for a positive number* θ*, then for constants* C *and* C'*,*

$$\mathbb{P}_\pi\left\{\frac{p_c}{2}\|\hat{\mathbf{L}}-\mathbf{L}^*\|_F^2 > \sum_{(i,t)\in\mathcal{O}}\langle\mathbf{A}_{it},\hat{\mathbf{L}}-\mathbf{L}^*\rangle^2 + C\frac{L_{\max}^2RT^2}{p_c}\right\} \leq 2\exp\left(-\frac{C'\theta}{T}\right),$$

whenever $C'\theta > T$*.*

Now we are ready to prove the main theorem.

Proof of Theorem 2. Let $\boldsymbol{\Delta} = \mathbf{L}^* - \hat{\mathbf{L}}$. 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

$$\sum_{(i,t)\in\mathcal{O}}\frac{\langle\mathbf{A}_{it},\boldsymbol{\Delta}\rangle^2}{|\mathcal{O}|} \leq \frac{30C_1\sigma\sqrt{R}\left[\sqrt{N\,\log(N+T)}\vee\sqrt{T}\,\log^{3/2}(N+T)\right]}{|\mathcal{O}|}\|\boldsymbol{\Delta}\|_F. \tag{A.4}$$

Now, we use Lemma 3 to find a lower bound for the left hand side of (A.4). But first note that if $p_{\text{c}}\|\boldsymbol{\Delta}\|_F^2 \leq 8L_{\max}^2T\log(N+T)$ then

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

holds which proves Theorem 2. Otherwise, for $\theta \equiv C'T\log(N+T)$ with a large enough constant C', we have

$$\|\boldsymbol{\Delta}\|_F^2 \geq \frac{4L_{\max}^2\theta}{p_{\text{c}}},$$

and the condition $C'\theta > T$ is also satisfied. Therefore, we can invoke Lemma 3 and obtain for a constant C'',

$$\mathbb{P}_\pi\left\{\frac{p_{\text{c}}}{2}\|\boldsymbol{\Delta}\|_F^2 - C''\frac{L_{\max}^2RT^2}{p_{\text{c}}} \leq \sum_{(i,t)\in\mathcal{O}}\langle\mathbf{A}_{it},\boldsymbol{\Delta}\rangle^2\right\} \geq 1-\frac{1}{(N+T)^2}. \tag{A.5}$$

Combining (A.5), (A.4), and the union bound we have, with probability greater than $1-2(N+T)^{-2}$,

$$\begin{aligned}\frac{p_{\text{c}}\|\boldsymbol{\Delta}\|_F^2}{2} &\leq C'''\sigma\sqrt{R}\left[\sqrt{N\,\log(N+T)}\vee\sqrt{T}\,\log^{3/2}(N+T)\right]\|\boldsymbol{\Delta}\|_F + C''\frac{L_{\max}^2RT^2}{p_{\text{c}}}\\ &\leq \frac{4C'''\sigma^2R\left[N\,\log(N+T)\vee T\,\log^3(N+T)\right]}{p_{\text{c}}} + \frac{p_{\text{c}}\|\boldsymbol{\Delta}\|_F^2}{4} + C''\frac{L_{\max}^2RT^2}{p_{\text{c}}},\end{aligned}$$

where the last step also used the inequality $2ab \leq a^2+b^2$. Therefore, we obtain

$$\|\boldsymbol{\Delta}\|_F^2 \leq \frac{4C'''\sigma^2R\left[N\,\log(N+T)\vee T\,\log^3(N+T)\right]}{p_{\text{c}}^2} + 4C''\frac{L_{\max}^2RT^2}{p_{\text{c}}^2}$$

The main result now follows after dividing both sides with $\sqrt{NT}\|\boldsymbol{\Delta}\|_F$, using $T \geq C_0\log(N+T)$, and choosing a large enough constant C. □

A.3 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

$$\sum_{(i,t)\in\mathcal{O}} \frac{\langle \mathbf{L}^* - \hat{\mathbf{L}}, \mathbf{A}_{it}\rangle^2}{|\mathcal{O}|} + 2\sum_{(i,t)\in\mathcal{O}} \frac{\varepsilon_{it}\langle \mathbf{L}^* - \hat{\mathbf{L}}, \mathbf{A}_{it}\rangle}{|\mathcal{O}|} + \lambda\|\hat{\mathbf{L}}\|_* \leq \lambda\|\mathbf{L}^*\|_* \,. \tag{A.6}$$

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

$$\begin{aligned}
\sum_{(i,t)\in\mathcal{O}} \frac{\langle \boldsymbol{\Delta}, \mathbf{A}_{it}\rangle^2}{|\mathcal{O}|} &\leq -\frac{2}{|\mathcal{O}|}\langle \boldsymbol{\Delta}, \mathfrak{E}\rangle + \lambda\|\mathbf{L}^*\|_* - \lambda\|\hat{\mathbf{L}}\|_* && \text{(A.7)}\\
&\overset{(a)}{\leq} \frac{2}{|\mathcal{O}|}\|\boldsymbol{\Delta}\|_*\|\mathfrak{E}\|_{\text{op}} + \lambda\|\mathbf{L}^*\|_* - \lambda\|\hat{\mathbf{L}}\|_* && \text{(A.8)}\\
&\leq \frac{2}{|\mathcal{O}|}\|\boldsymbol{\Delta}\|_*\|\mathfrak{E}\|_{\text{op}} + \lambda\|\boldsymbol{\Delta}\|_* && \text{(A.9)}\\
&\overset{(b)}{\leq} \frac{5}{3}\lambda\|\boldsymbol{\Delta}\|_* \,. && \text{(A.10)}
\end{aligned}$$

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

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

(ii) $rank(\boldsymbol{\Delta}_1) \leq 2R$.

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

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

$$\|\boldsymbol{\Delta}\|_* \overset{(c)}{\leq} 4\|\boldsymbol{\Delta}_1\|_* \overset{(d)}{\leq} 4\sqrt{2R}\|\boldsymbol{\Delta}_1\|_F \overset{(e)}{\leq} 4\sqrt{2R}\|\boldsymbol{\Delta}\|_F \,. \tag{A.11}$$

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.10) we obtain

$$\sum_{(i,t)\in\mathcal{O}} \frac{\langle \mathbf{\Delta}, \mathbf{A}_{it}\rangle^2}{|\mathcal{O}|} \leq 10\lambda\sqrt{R}\,\|\mathbf{\Delta}\|_F\,, \tag{A.12}$$

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 $u_1,\ldots,u_R$, 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

$$\begin{aligned}\mathbf{\Delta}_1 &= (\mathbf{P}_U + \mathbf{P}_{U^\perp})\mathbf{\Delta}(\mathbf{P}_V + \mathbf{P}_{V^\perp}) - \mathbf{P}_{U^\perp}\mathbf{\Delta}\mathbf{P}_{V^\perp} &\text{(A.13)}\\ &= \mathbf{P}_U\mathbf{\Delta} + \mathbf{P}_{U^\perp}\mathbf{\Delta}\mathbf{P}_V\,. &\text{(A.14)}\end{aligned}$$

Using this fact we have

$$\begin{aligned}\langle \mathbf{\Delta}_1, \mathbf{\Delta}_2\rangle &= \mathrm{trace}\left(\mathbf{\Delta}^\top\mathbf{P}_U\mathbf{P}_{U^\perp}\mathbf{\Delta}\mathbf{P}_{V^\perp} + \mathbf{P}_V\mathbf{\Delta}^\top\mathbf{P}_{U^\perp}\mathbf{P}_{U^\perp}\mathbf{\Delta}\mathbf{P}_{V^\perp}\right) &\text{(A.15)}\\ &= \mathrm{trace}\left(\mathbf{P}_V\mathbf{\Delta}^\top\mathbf{P}_{U^\perp}\mathbf{\Delta}\mathbf{P}_{V^\perp}\right) &\text{(A.16)}\\ &= \mathrm{trace}\left(\mathbf{\Delta}^\top\mathbf{P}_{U^\perp}\mathbf{\Delta}\mathbf{P}_{V^\perp}\mathbf{P}_V\right) = 0 &\text{(A.17)}\end{aligned}$$

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

Looking at (A.14), 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,

$$\|\mathbf{L}^* + \mathbf{\Delta}_2\|_* = \|\mathbf{L}^*\|_* + \|\mathbf{\Delta}_2\|_*\,. \tag{A.18}$$

On the other hand, using inequality (A.8), for $\lambda \geq 3\|\mathfrak{E}\|_{\mathrm{op}}/|\mathcal{O}|$ we have

$$\begin{aligned}\lambda\left(\|\hat{\mathbf{L}}\|_* - \|\mathbf{L}^*\|_*\right) &\leq \frac{2}{|\mathcal{O}|}\|\mathbf{\Delta}\|_*\|\mathfrak{E}\|_{\mathrm{op}}\\ &\leq \frac{2}{3}\lambda\|\mathbf{\Delta}\|_*\\ &\leq \frac{2}{3}\lambda\left(\|\mathbf{\Delta}_1\|_* + \|\mathbf{\Delta}_2\|_*\right)\,. &\text{(A.19)}\end{aligned}$$

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

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

Here (f) follows from (A.18). Now, combining the last inequality with (A.19) we get

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

That finishes proof of part (iii). □

A.4 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 $i \in [N]$. 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 $\alpha \geq 0$

$$\mathbb{P}\left\{ \left\|\sum_{i=1}^N \mathbf{Z}_i\right\|_{\text{op}} \geq \alpha \right\} \leq (d_1 + d_2) \exp\left[\frac{-\alpha^2}{2\sigma_Z^2 + (2D\alpha)/3}\right] . \tag{A.20}$$

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 $\alpha \geq 0$,

$$\begin{aligned}\mathbb{P}\{\ \|\mathfrak{E}\|_{\text{op}} \geq \alpha\} &\leq \mathbb{P}\left\{\ \left\|\sum_{i=1}^{N}\overline{\mathbf{B}}_i\right\|_{\text{op}} \geq \alpha\right\} + \sum_{(i,t)\in\mathcal{O}}\mathbb{P}\{|\varepsilon_{it}| \geq D\} \\ &\leq \mathbb{P}\left\{\ \left\|\sum_{i=1}^{N}\overline{\mathbf{B}}_i\right\|_{\text{op}} \geq \alpha\right\} + 2|\mathcal{O}|\exp\left\{\frac{-D^2}{2\sigma^2}\right\} \\ &\leq \mathbb{P}\left\{\ \left\|\sum_{i=1}^{N}\overline{\mathbf{B}}_i\right\|_{\text{op}} \geq \alpha\right\} + \frac{1}{(N+T)^3}\,. \end{aligned} \tag{A.21}$$

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,

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

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

$$\left\|\sum_{i=1}^{N}\overline{\mathbf{B}}_i\right\|_{\text{op}} \leq \left\|\sum_{i=1}^{N}\mathbf{Z}_i\right\|_{\text{op}} + \frac{\sigma}{(N+T)^3}\,. \tag{A.22}$$

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,

$$\left\|\sum_{i=1}^{N}\mathbb{E}[\mathbf{Z}_i\mathbf{Z}_i^\top]\right\|_{\text{op}} \leq \max_{(i,t)\in\mathcal{O}}\left\{\mathbb{E}\left[(\bar{\varepsilon}_{it} - E[\bar{\varepsilon}_{it}])^2\right]\right\}\ \left\|\sum_{i=1}^{N}\mathbb{E}\left[\sum_{t=1}^{t_i}e_i(N)e_i(N)^\top\right]\right\|_{\text{op}} \tag{A.23}$$

$$\leq 2\sigma^2\max_{i\in[N]}\left(\sum_{t\in[T]}t\pi_t^{(i)}\right) \leq 2T\sigma^2 \tag{A.24}$$

and

$$\left\|\sum_{i=1}^{N}\mathbb{E}[\mathbf{Z}_i^\top\mathbf{Z}_i]\right\|_{\text{op}} \le 2\sigma^2\left\|\sum_{i=1}^{N}\mathbb{E}\left[\sum_{t=1}^{t_i}e_t(T)e_t(T)^\top\right]\right\|_{\text{op}} \tag{A.25}$$

$$= 2\sigma^2\max_{t\in[T]}\left(\sum_{i\in[N]}\sum_{t'=t}^{T}\pi_{t'}^{(i)}\right) = 2N\sigma^2\,. \tag{A.26}$$

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

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

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

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

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

Using this for a t that is a large enough constant times $\log(N+T)$, together with (A.21) and (A.22), shows with probability larger than $1-2(N+T)^{-3}$

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

for a constant C_1. □

A.5 Proof of Lemma 3

Proof of Lemma 3 is similar to the proof of Theorem 1 in (Negahban and Wainwright, 2012), Lemma 12 in (Klopp, 2014), or Corollary 3.1 in (Hamidi and Bayati, 2019). However, for completeness, below we provide a summary of this proof (adapted to our setting).

Recall that our aim is to prove that when $\hat{\mathbf{L}}$ satisfies $\|\hat{\mathbf{L}}-\mathbf{L}^*\|_F^2 \ge 4L_{\max}^2\theta/p_{\text{c}}$ for a positive number θ, then for constants C and C',

$$\mathbb{P}_\pi\left\{\frac{p_{\text{c}}}{2}\|\hat{\mathbf{L}}-\mathbf{L}^*\|_F^2 > \sum_{(i,t)\in\mathcal{O}}\langle\mathbf{A}_{it}, \hat{\mathbf{L}}-\mathbf{L}^*\rangle^2 + C\frac{L_{\max}^2RT^2}{p_{\text{c}}}\right\} \le 2\exp\left(-\frac{C'\theta}{T}\right)\,,$$

whenever $C'\theta > T$.

First, we define some additional notation. Given the observation set $\mathcal{O}$, for every N by T matrix $\mathbf{M}$ define $\mathcal{X}_{\mathcal{O}}(M)$ to be an $|O|$ by 1 vector that is obtained by stacking observed entries of all rows of $\mathbf{M}$ vertically. Specifically, for all $i \in [N]$ we define

$$\mathcal{X}_{\mathcal{O}}^{(i)}(\mathbf{M}) \equiv [\langle \mathbf{A}_{i1}, \mathbf{M} \rangle, \ldots, \langle \mathbf{A}_{it_i}, \mathbf{M} \rangle]^\top,$$

and then define,

$$\mathcal{X}_{\mathcal{O}}(\mathbf{M}) \equiv \begin{bmatrix} \mathcal{X}_{\mathcal{O}}^{(1)}(\mathbf{M}) \\ \vdots \\ \mathcal{X}_{\mathcal{O}}^{(N)}(\mathbf{M}) \end{bmatrix}.$$

In addition, we define $L^2(\Pi)$ norm of $\mathbf{M}$ to be defined by

$$\|\mathbf{M}\|_{L^2(\Pi)} \equiv \sqrt{\mathbb{E}_\pi\Big(\|\mathcal{X}_{\mathcal{O}}(\mathbf{M})\|_2^2\Big)}.$$

We also define the constraint set

$$\mathcal{C}(\theta, \eta) \equiv \Big\{ \mathbf{M} \in \mathbb{R}^{N \times T} \mid \|\mathbf{M}\|_{\max} \le 1\,,\ \|\mathbf{M}\|_{L^2(\Pi)}^2 \ge \theta\,,\ \|\mathbf{M}\|_* \le \sqrt{\eta}\|\mathbf{M}\|_F \Big\}.$$

Now we are ready to prove Lemma 3.

Proof of Lemma 3. First, define $\vartheta = \eta T^2 / p_{\mathrm{c}}$. We show that proof of Lemma 3 would be a corollary of the following statement. Whenever $\mathbf{M} \in \mathcal{C}(\theta, \eta)$ and constant C_3 is such that $C_3\theta > T$, for a constant C_2,

$$\mathbb{P}_\pi \left\{ \frac{1}{2}\|\mathbf{M}\|_{L^2(\Pi)}^2 > \|\mathcal{X}_{\mathcal{O}}(\mathbf{M})\|_2^2 + C_2\vartheta \right\} \le 2\exp\left(-\frac{C_3\theta}{T}\right). \tag{A.27}$$

The reason for the fact that Lemma 3 follows from this statement is as follows. It is straightforward to see that for all $\mathbf{M}$, $\|\mathbf{M}\|_{L^2(\Pi)}^2 \ge p_{\mathrm{c}}\|\mathbf{M}\|_F^2$ which means the assumption $\|\hat{\mathbf{L}} - \mathbf{L}^*\|_F^2 \ge 4L_{\max}^2\theta/p_{\mathrm{c}}$ gives $\|\hat{\mathbf{L}} - \mathbf{L}^*\|_{L^2(\Pi)}^2 \ge 4L_{\max}^2\theta$. Also, recall from Eq. (A.11) that $\|\hat{\mathbf{L}} - \mathbf{L}^*\|_* \le \sqrt{32R}\|\hat{\mathbf{L}} - \mathbf{L}^*\|_F$. Therefore, if we take $\eta = 32R$, then $\frac{1}{2L_{\max}}(\hat{\mathbf{L}} - \mathbf{L}^*) \in \mathcal{C}(\theta, \eta)$. We can now apply Eq. (A.27) to $\mathbf{M} = \frac{1}{2L_{\max}}(\hat{\mathbf{L}} - \mathbf{L}^*)$ and obtain, for a constants C, C_2,

$$\begin{aligned}
\mathbb{P}_\pi \left\{ \frac{p_{\mathrm{c}}}{2}\|\hat{\mathbf{L}} - \mathbf{L}^*\|_F^2 > \sum_{(i,t)\in\mathcal{O}} \langle \mathbf{A}_{it}, \hat{\mathbf{L}} - \mathbf{L}^* \rangle^2 + C\frac{L_{\max}^2 R T^2}{p_{\mathrm{c}}} \right\} &= \mathbb{P}_\pi \left\{ \frac{p_{\mathrm{c}}}{2}\|\mathbf{M}\|_F^2 > \|\mathcal{X}_{\mathcal{O}}(\mathbf{M})\|_2^2 + C_2\vartheta \right\} \\
&\le \mathbb{P}_\pi \left\{ \frac{1}{2}\|\mathbf{M}\|_{L^2(\Pi)}^2 > \|\mathcal{X}_{\mathcal{O}}(\mathbf{M})\|_2^2 + C_2\vartheta \right\} \\
&\le 2\exp\left(-\frac{C_3\theta}{T}\right),
\end{aligned}$$

which is what we needed.

Therefore, for the remaining part of this proof we will be working on proof of Eq. (A.27). Let us define the following bad event,

$$\mathcal{B} \equiv \left\{ \exists \mathbf{M} \in \mathcal{C}(\theta, \eta) \;\middle|\; \|\mathbf{M}\|^2_{L^2(\Pi)} - \|\mathcal{X}_{\mathcal{O}}(M)\|_2^2 \geq \frac{1}{2}\|\mathbf{M}\|^2_{L^2(\Pi)} + C_2\vartheta \right\} .$$

Our goal will be to bound probability the event $\mathcal{B}$. Let also ξ to be a constant larger than 1, and define for every $\rho \geq \theta$,

$$\mathcal{C}(\theta, \eta, \rho) \equiv \left\{ \mathbf{M} \in \mathcal{C}(\theta, \eta) \;\middle|\; \rho \leq \|\mathbf{M}\|^2_{L^2(\Pi)} \leq \rho\xi \right\} .$$

Therefore, $\mathcal{C}(\theta, \eta) = \cup_{\ell=1}^{\infty} \mathcal{C}(\theta, \eta, \theta\xi^{\ell-1})$. Now, assume that $\mathbf{M} \in \mathcal{C}(\theta, \eta)$ such that event $\mathcal{B}$ holds for $\mathbf{M}$. This means for some $\ell \geq 1$, $\mathbf{M} \in \mathcal{C}(\theta, \eta, \xi^{\ell-1}\theta)$ and event $\mathcal{B}$ holds for $\mathbf{M}$. Thus,

$$\begin{aligned} \|\mathbf{M}\|^2_{L^2(\Pi)} - \|\mathcal{X}_{\mathcal{O}}(\mathbf{M})\|_2^2 &\geq \frac{1}{2}\|\mathbf{M}\|^2_{L^2(\Pi)} + C_2\vartheta \\ &\geq \frac{1}{2\xi}\xi^{\ell}\theta + C_2\vartheta \,. \end{aligned}$$

Now, define the event $\mathcal{B}_\ell$ by,

$$\mathcal{B}_\ell \equiv \left\{ \exists \mathbf{M} \in \mathcal{C}(\theta, \eta, \xi^{\ell-1}\theta) \;\middle|\; \|\mathbf{M}\|^2_{L^2(\Pi)} - \|\mathcal{X}_{\mathcal{O}}(\mathbf{M})\|_2^2 \geq \frac{1}{2\xi}\xi^{\ell}\theta + C_2\vartheta \right\} .$$

This means $\mathcal{B} \subseteq \cup_{\ell=1}^{\infty} \mathcal{B}_\ell$. Our next step is to use the following concentration inequality for the random variables $\mathcal{X}_{\mathcal{O}}^{(1)}(\mathbf{M}), \ldots, \mathcal{X}_{\mathcal{O}}^{(N)}(\mathbf{M})$ and use the result to complete proof of Lemma 3.

Lemma 5. *Let* $Z_\rho \equiv \sup_{\mathbf{M} \in \mathcal{C}(\theta,\eta,\rho)} \left\{ \|\mathbf{M}\|^2_{L^2(\Pi)} - \|\mathcal{X}_{\mathcal{O}}(\mathbf{M})\|_2^2 \right\}$*, then there exist constants* C_2 *and* C_3 *such that*

$$\mathbb{P}_\pi \left\{ Z_\rho \geq \frac{1}{2\xi}\xi\rho + C_2\vartheta \right\} \leq \exp\left(-\frac{C_3\rho\xi}{T} \right) . \tag{A.28}$$

Proof of Lemma 5 is provided at the end of this section. Now, assuming Lemma 5, we have

$$\begin{aligned} \mathbb{P}_\pi(\mathcal{B}_\ell) &\leq \exp\left(-\frac{C_3\xi^{\ell}\theta}{T} \right) \\ &\leq \exp\left(-\frac{C_3\ell\log(\xi)\theta}{T} \right) , \end{aligned}$$

where the last step uses $x \geq \log(x)$ for $x > 1$. Therefore, via union bound,

$$\begin{aligned} \mathbb{P}_\pi(\mathcal{B}) &\leq \sum_{\ell=1}^{\infty} \exp\left(-\frac{C_4\ell\theta}{T} \right) \\ &= \frac{\exp\left(-\frac{C_4\theta}{T}\right)}{1 - \exp\left(-\frac{C_4\theta}{T}\right)} . \end{aligned}$$

If $C_4\theta > T$, we obtain

$$\mathbb{P}_\pi(\mathcal{B}) \leq 2\exp\left(-\frac{C_4\theta}{T}\right),$$

which finishes the proof. □

Proof of Lemma 5. Let

$$\widetilde{Z}_\rho \equiv \sup_{\mathbf{M}\in\mathcal{C}(\theta,\eta,\rho)} \left\{ \left| \|\mathbf{M}\|^2_{L^2(\Pi)} - \|\mathcal{X}_\mathcal{O}(\mathbf{M})\|_2^2 \right| \right\}.$$

Clearly $Z_\rho \leq \widetilde{Z}_\rho$ therefore, if we prove expression (A.28) holds for $\widetilde{Z}_\rho$, then it would hold for Z_ρ as well. We aim to use Massart's concentration inequality (Theorem 3 of Massart et al. (2000)). But, first we would need to find upper bounds for $\mathbb{E}(\widetilde{Z}_\rho)$ and a certain variance term. Specifically, for the variance part, we use

$$\begin{aligned}
\sup_{\mathbf{M}\in\mathcal{C}(\theta,\eta,\rho)} \sum_{i=1}^N \mathrm{Var}\left[\|\mathcal{X}_\mathcal{O}^{(i)}(\mathbf{M})\|_2^2\right] &\leq \sup_{\mathbf{M}\in\mathcal{C}(\theta,\eta,\rho)} \sum_{i=1}^N \mathbb{E}\left[\|\mathcal{X}_\mathcal{O}^{(i)}(\mathbf{M})\|_2^4\right] \\
&\leq T \sup_{\mathbf{M}\in\mathcal{C}(\theta,\eta,\rho)} \sum_{i=1}^N \mathbb{E}\left[\|\mathcal{X}_\mathcal{O}^{(i)}(\mathbf{M})\|_2^2\right] \\
&\leq T\rho\xi\,,
\end{aligned}$$

where the last inequality uses definition of $\mathcal{C}(\theta,\eta,\rho)$.

Next, we work on finding an upper bound for $\mathbb{E}(\widetilde{Z}_\rho)$. First, using a symmetrization argument (Lemma 6.3 of Ledoux and Talagrand (2013)) we have

$$\mathbb{E}(\widetilde{Z}_\rho) \leq 2\mathbb{E}\left[\sup_{\mathbf{M}\in\mathcal{C}(\theta,\eta,\rho)} \left| \sum_{i=1}^N \zeta_i \|\mathcal{X}_\mathcal{O}^{(i)}(\mathbf{M})\|_2^2 \right|\right],$$

where $(\zeta_i)_{i=1}^n$ are iid Rademacher random variables. Note that we used identity function for F and norm-infinity on an infinite dimensional vector indexed by matrices $\mathbf{M}$ in $\mathcal{C}(\theta,\eta,\rho)$. Next, we define the N by T matrix $\mathbf{M}^{\mathrm{Rad}}_\mathcal{O}$ such that its (it) entry is equal to:

$$\begin{cases} \zeta_i M_{it} & \text{If } (it)\in\mathcal{O} \\ 0 & \text{Otherwise.} \end{cases}$$

Now, we have,

$$\begin{aligned}
\mathbb{E}\left[\sup_{\mathbf{M}\in\mathcal{C}(\theta,\eta,\rho)}\Big|\sum_{i=1}^{N}\zeta_i\|\mathcal{X}_{\mathcal{O}}^{(i)}(\mathbf{M})\|_2^2\Big|\right] &= \mathbb{E}\left[\sup_{\mathbf{M}\in\mathcal{C}(\theta,\eta,\rho)}\Big|\langle\mathbf{M}_{\mathcal{O}}^{\mathrm{Rad}},\mathbf{M}\rangle\Big|\right] \\
&\leq \mathbb{E}\left[\sup_{\mathbf{M}\in\mathcal{C}(\theta,\eta,\rho)}\|\mathbf{M}_{\mathcal{O}}^{\mathrm{Rad}}\|_{\mathrm{op}}\|\mathbf{M}\|_*\right] \\
&\leq \sqrt{\eta}\,\mathbb{E}\left[\sup_{\mathbf{M}\in\mathcal{C}(\theta,\eta,\rho)}\|\mathbf{M}_{\mathcal{O}}^{\mathrm{Rad}}\|_{\mathrm{op}}\|\mathbf{M}\|_F\right] \\
&\leq T\sqrt{\frac{\eta}{p_{\mathrm{c}}}}\,\mathbb{E}\left[\sup_{\mathbf{M}\in\mathcal{C}(\theta,\eta,\rho)}\|\mathbf{M}\|_{L^2(\Pi)}\right] \\
&\leq T\sqrt{\frac{\eta\rho\xi}{p_{\mathrm{c}}}} \\
&\leq \frac{1}{2}\left(C_1\rho\xi+\frac{T^2\eta}{C_1p_{\mathrm{c}}}\right),
\end{aligned}$$

for any positive constant C_1.

Finally, noting that each term $\|\mathcal{X}_{\mathcal{O}}^{(i)}(\mathbf{M})\|_2^2$ is at most T, by taking the constant C_1 small enough and invoking Massart's inequality as discussed above (choosing $x=\rho\xi/T$), we have

$$\begin{aligned}
\mathbb{P}_\pi\left\{\widetilde{Z}_\rho\geq\frac{1}{2\xi}\rho\xi+C_2\vartheta\right\} &\leq \mathbb{P}_\pi\left\{\widetilde{Z}_\rho\geq C_1'\rho\xi+C_2'\frac{T^2\eta}{p_{\mathrm{c}}}\right\} \\
&\leq \exp\left(-\frac{C_3\rho\xi}{T}\right),
\end{aligned}$$

for suitable constants C_1, C_2, C_3, C_1', and C_2' which finished the proof. □