JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2021, VOL. 00, NO. 0, 1–15: Synthetic Control Methods
https://doi.org/10.1080/01621459.2021.1891924

Taylor & Francis
Taylor & Francis Group

Matrix Completion Methods for Causal Panel Data Models

Susan Athey[a], Mohsen Bayati[b], Nikolay Doudchenko[b], Guido Imbens[c], and Khashayar Khosravi[d]

[a]Graduate School of Business, Stanford University, SIEPR, and NBER, Stanford, CA; [b]Graduate School of Business, Stanford University, Stanford, CA; [c]Graduate School of Business, and Department of Economics, Stanford University, SIEPR, and NBER, Stanford, CA; [d]Department of Electrical Engineering, Stanford University, Stanford, CA

ABSTRACT

In this article, 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.

ARTICLE HISTORY
Received October 2018
Accepted February 2021

KEYWORDS
Causality; Interactive fixed effects; Low-rank matrix estimation; Synthetic controls; Unconfoundedness

1. Introduction

In this article, 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, Diamond, and Hainmueller 2010, 2015; Doudchenko and Imbens 2016; Chernozhukov, Wuthrich, and Zhu 2017; Ben-Michael, Feller, and Rothstein 2018; Arkhangelsky et al. 2019; Ferman and Pinto 2019; Li 2020; Amjad, Shah, and Shen 2018, 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 performance 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

CONTACT Guido Imbens ✉ *imbens@stanford.edu* Graduate School of Business, and Department of Economics, Stanford University, SIEPR, and NBER, Stanford, CA 94305.

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) extended 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, Hastie, and Tibshirani 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 article, 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 onward, 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 datasets 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 article. 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 (Hirano, Imbens, and Ridder 2003; Rubin 2006), 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 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. 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 Section 8.2.

For any positive integer n, we use notation $[n]$ to refer to the set $\{1, \dots, n\}$ 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 of 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,

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

$$\mathbf{W} = \begin{pmatrix} 0 & 0 & 1 & \dots & 0 \\ 1 & 1 & 0 & \dots & 1 \\ 0 & 1 & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & 0 & \dots & 1 \end{pmatrix}, \tag{1}$$

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 article 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 interest, 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 (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 $\checkmark$ 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, Diamond, and Hainmueller 2010; Abadie 2019) focuses primarily on the case with a single treated unit which are treated for a number of periods from period $T_0 + 1$ onward, the single-treated-unit block structure:

$$\mathbf{Y} = \begin{pmatrix} \checkmark & \checkmark & \checkmark & \dots & \checkmark & \checkmark \\ \checkmark & \checkmark & \checkmark & \dots & \checkmark & \checkmark \\ \checkmark & \checkmark & \checkmark & \dots & \checkmark & ? \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \checkmark & \checkmark & \checkmark & \dots & \checkmark & ? \\ & & & & & \uparrow \\ & & & \multicolumn{3}{r}{\text{treated period}} \end{pmatrix} \quad \text{and}$$

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

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 $N \approx T$:

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

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

or

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

3.3. Horizontal and Vertical Regressions

Two special combinations of missing data patterns and 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

$$\hat{Y}_{iT} = \hat{\beta}_0 + \sum_{s=1}^{T-1} \hat{\beta}_s Y_{is}, \quad \text{where}$$

$$\hat{\beta} = \arg\min_{\beta} \sum_{i:(i,T)\in\mathcal{O}} \left(Y_{iT} - \beta_0 - \sum_{s=1}^{T-1} \beta_s Y_{is} \right)^2. \qquad (2)$$

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 pretreatment periods $t = 1, \dots, T-1$.

3.3.2. Vertical Regression and the Synthetic Control Literature

The synthetic control literature (Abadie, Diamond, and Hainmueller 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 pretreatment periods:

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

Doudchenko and Imbens (2016) and Ferman and Pinto (2019) showed 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 $t = T_0, \dots, T$, the predicted outcome is

$$\hat{Y}_{Nt} = \hat{\gamma}_0 + \sum_{i=1}^{N-1} \hat{\gamma}_i Y_{it}, \quad \text{where}$$

$$\hat{\gamma} = \arg\min_{\gamma} \sum_{t:(N,t)\in\mathcal{O}} \left(Y_{Nt} - \gamma_0 - \sum_{i=1}^{N-1} \gamma_i Y_{it} \right)^2. \qquad (3)$$

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, Diamond, and Hainmueller (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 $s \in \{1, \dots, T_0\}$ such that Y_{is} is approximately equal to Y_{Ns} for all control units $i = 1, \dots, N-1$. 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 (2), it is still the case that imputed $\hat{Y}_{NT}$ is linear in the observed $Y_{1T}, \dots, Y_{N-1,T}$, just with different weights than those obtained from the vertical regression. Similarly, if we estimate the vertical regression in (3), it is still the case that $\hat{Y}_{NT}$ is linear in $Y_{N1}, \dots, Y_{N,T-1}$, just with different weights from the horizontal regression in (2). 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 Section 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 $t = 1, \dots, T-1$, 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 across 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; Liang and Zeger 1986; Arellano and Honoré 2001; Bai and Ng 2002; Bai 2003, 2009; Pesaran 2006; Angrist and Pischke 2008; Moon and Weidner 2015, 2017; Amjad, Shah, and Shen 2018). 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

$$Y_{it} = \gamma_i + \delta_t + \epsilon_{it}. \tag{4}$$

More general factor models can be written as

$$Y_{it} = \sum_{r=1}^{R} u_{ir} v_{tr} + \varepsilon_{it}, \qquad \text{or } \mathbf{Y} = \mathbf{U}\mathbf{V}^\top + \boldsymbol{\varepsilon}, \tag{5}$$

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) adapted this interactive fixed effect approach to the matrix completion problem in the special case with blocked assignment, with additional applications in Gobillon and Magnac (2016), Kim and Oka (2014), and Hsiao, Steve Ching, and Ki Wan (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, Alon, and Jaakkola 2005; Candès and Recht 2009; Candès and Tao 2010; Keshavan, Montanari, and Oh 2010a), 2010b; Gross 2011; Koltchinskii, Lounici, and Tsybakov 2011; Negahban and Wainwright 2011, 2012; Recht 2011; Rohde and Tsybakov 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 (5). 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

$$\mathbf{Y} = \mathbf{L}^* + \boldsymbol{\varepsilon}, \qquad \text{where} \quad \mathbb{E}[\boldsymbol{\varepsilon} | \mathbf{L}^*] = \mathbf{0}\,. \tag{6}$$

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 $\mathbb{E}[\exp(t\varepsilon)] \le \exp(\sigma^2 t^2/2)$.

Table 1. Matrix norms for matrix L.

Schatten norm: $\|\mathbf{L}\|_p^S \equiv \left(\sum_{i\in[N]}\sigma_i(\mathbf{L})^p\right)^{1/p}, p \in (0,\infty)$

Fröbenius norm: $\|\mathbf{L}\|_F \equiv \|\mathbf{L}\|_2^S = \left(\sum_{i\in[N]}\sigma_i(\mathbf{L})^2\right)^{1/2} = \left(\sum_{i\in[N]}\sum_{t\in[T]} L_{it}^2\right)^{1/2}$

Rank norm: $\|\mathbf{L}\|_0 \equiv \lim_{p\downarrow 0}\|\mathbf{L}\|_p^S = \sum_{i\in[N]} \mathbf{1}_{\sigma_i(\mathbf{L})>0}$

Nuclear norm: $\|\mathbf{L}\|_* \equiv \|\mathbf{L}\|_1^S = \sum_{i\in[N]}\sigma_i(\mathbf{L})$

Operator norm: $\|\mathbf{L}\|_{op} \equiv \lim_{p\to\infty}\|\mathbf{L}\|_p^S = \max_{i\in[N]}\sigma_i(\mathbf{L}) = \sigma_1(\mathbf{L})$

Max norm: $\|\mathbf{L}\|_{\max} \equiv \max_{(i,t)\in[N]\times[T]}|L_{it}|$

Element-wise ℓ_1 norm: $\|\mathbf{L}\|_{1,e} \equiv \sum_{(i,t)\in[N]\times[T]}|L_{it}|$

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 $\mathbf{P}_{\mathcal{O}}^{\perp}(\mathbf{A})$ with typical elements:

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

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

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) $\mathbf{L}_{N\times T} = \mathbf{S}_{N\times N}\mathbf{\Sigma}_{N\times T}\mathbf{R}_{T\times T}^{\top}$, 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 differences,

$$\min_{\mathbf{L}} \frac{1}{|\mathcal{O}|}\sum_{(i,t)\in\mathcal{O}}(Y_{it}-L_{it})^2 = \min_{\mathbf{L}} \frac{1}{|\mathcal{O}|}\|\mathbf{P}_{\mathcal{O}}(\mathbf{Y}-\mathbf{L})\|_F^2\,, \quad (7)$$

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 (7), for some choice of the norm $\|\cdot\|$ and a scalar λ. However, since we do 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, Tibshirani, and Friedman 2009).

4.1. The Estimator

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

$$(\hat{\mathbf{L}},\hat{\Gamma},\hat{\Delta}) = \arg\min_{\mathbf{L},\Gamma,\Delta}\left\{\frac{1}{|\mathcal{O}|}\|\mathbf{P}_{\mathcal{O}}(\mathbf{Y}-\mathbf{L}-\Gamma\mathbf{1}_T^{\top}-\mathbf{1}_N\Delta^{\top})\|_F^2+\lambda\|\mathbf{L}\|_*\right\}. \quad (8)$$

Compared to the setting studied by Candès and Recht (2009), Candès and Plan (2010), and Mazumder, Hastie, and Tibshirani (2010), we include the fixed effects Γ and Δ. Although formally the fixed effects can be subsumed in the matrix $\mathbf{L}$ ($\Gamma\mathbf{1}_T^{\top}$ 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 with 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, for example, the SOFT-IMPUTE algorithm by Mazumder, Hastie, and Tibshirani (2010) that will be described next.

4.2. 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 $\mathbf{A}$, $\mathbf{A}=\mathbf{S\Sigma R}^{\top}$, with singular values $\sigma_1(\mathbf{A}),\ldots,\sigma_{\min(N,T)}(\mathbf{A})$, define the matrix shrinkage operator

$$\text{shrink}_{\lambda}(\mathbf{A}) = \mathbf{S}\tilde{\mathbf{\Sigma}}\mathbf{R}^{\top}\,, \quad (9)$$

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

$$\mathbf{L}_{k+1}(\lambda,\mathcal{O}) = \text{shrink}_{\frac{\lambda|\mathcal{O}|}{2}}\left\{\mathbf{P}_{\mathcal{O}}(\mathbf{Y})+\mathbf{P}_{\mathcal{O}}^{\perp}\Big(\mathbf{L}_k(\lambda,\mathcal{O})\Big)\right\}, \quad (10)$$

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 (10) by $\mathbf{P}_{\mathcal{O}}(\mathbf{Y}-\Gamma_k\mathbf{1}_T^{\top}-\mathbf{1}_N\Delta_k^{\top})$.

4.3. 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 datasets, $|\mathcal{O}_k/|\mathcal{O}|$, 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.

4.4. Confidence Intervals

Studying asymptotic distribution of $\mathbf{L}^* - \hat{\mathbf{L}}$ to construct confidence intervals is beyond the scope of this article and is an interesting future research question. However, one can use resampling 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 methods in the synthetic control literature (Abadie, Diamond, and Hainmueller 2010).

5. The Relationship With Horizontal and Vertical Regressions

In the second contribution of this article, we discuss the relation between the matrix completion estimator and the horizontal (unconfoundedness), vertical (synthetic control), and difference-in-differences approaches. To facilitate 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 characterized 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

$$Q(\mathbf{Y}; R, \mathbf{A}, \mathbf{B}, \gamma, \delta) = \frac{1}{|\mathcal{O}|} \left\| P_{\mathcal{O}} \left(\mathbf{Y} - \mathbf{A}\mathbf{B}^\top - \gamma \mathbf{1}_T^\top - \mathbf{1}_N \delta^\top \right) \right\|_F^2. \tag{11}$$

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)

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

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

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

subject to

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

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

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

subject to

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

(iv) (synthetic control),

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

subject to

$$R = N-1, \qquad \mathbf{B} = \begin{pmatrix} \mathbf{Y}_0^\top \\ \mathbf{y}_1^\top \end{pmatrix}, \qquad \delta = 0, \qquad \gamma = 0,$$

$$\forall\, i, A_{iT} \geq 0, \ \sum_{i=1}^{N-1} A_{iT} = 1,$$

(v) (vertical regression, elastic net),

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

subject to

$$R = N-1, \qquad \mathbf{B} = \begin{pmatrix} \mathbf{Y}_0^\top \\ \mathbf{y}_1^\top \end{pmatrix},$$

$$\gamma_1 = \gamma_2 = \cdots = \gamma_{N-1} = 0, \qquad \delta = 0,$$

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),

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

subject to

$$R = 0.$$

The proof for this result is in Section 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 (Mazumder, Hastie, and Tibshirani 2010, Lemma 6)

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

In addition, if $\hat{\mathbf{L}}$ is the solution to Equation (8) that has rank $\hat{R}$, then one solution for $\mathbf{A}$ and $\mathbf{B}$ is given by

$$\mathbf{A} = \mathbf{S}\boldsymbol{\Sigma}^{1/2}, \quad \mathbf{B} = \mathbf{R}\boldsymbol{\Sigma}^{1/2}, \tag{12}$$

where $\hat{\mathbf{L}} = \mathbf{S}_{N\times\hat{R}}\boldsymbol{\Sigma}_{\hat{R}\times\hat{R}}\mathbf{R}_{T\times\hat{R}}^\top$ is singular value decomposition of $\hat{\mathbf{L}}$. The proof of this fact is provided in (Mazumder, Hastie, and Tibshirani 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 nonregularized 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 $\mathbf{L}^*$.

$$\hat{\mathbf{L}} = \operatorname*{arg\,min}_{\mathbf{L}:\|\mathbf{L}\|_{\max}\leq L_{\max}} \left\{ \frac{1}{|\mathcal{O}|} \|\mathbf{P}_{\mathcal{O}}(\mathbf{Y}-\mathbf{L})\|_F^2 + \lambda\|\mathbf{L}\|_* \right\}. \tag{13}$$

6.1. Additional Notation

First, we start by introducing some new notation. Recall that for each positive integer n notation $[n]$ refers to the set of integers $\{1, 2, \dots, n\}$. 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 ith entry that is equal to 1. In other words, $\{e_1(n), e_2(n), \ldots, e_n(n)\}$ forms the standard basis for $\mathbb{R}^n$. For any two matrices $\mathbf{A}, \mathbf{B}$ of the same dimensions define the inner product $\langle \mathbf{A}, \mathbf{B}\rangle \equiv \operatorname{trace}(\mathbf{A}^\top \mathbf{B})$. 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 $\mathcal{O} = \bigcup_{i=1}^N \left\{(i,1),(i,2),\ldots,(i,t_i)\right\}$. 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 Section 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 and 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 Np_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 ith row. Such $\mathbf{L}^*$ is clearly low-rank. But recovering the entry L_{iT}^* is impossible when $t_i < T$ 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 Section 8 we discuss how our results could generalize to the cases with correlations among these noise variables.

Remark 6.3. The estimator (13) 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 Section 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 Section 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, $T \ge C_0\log(N+T)$ 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 $1-2(N+T)^{-2}$,

$$\frac{\|\mathbf{L}^*-\hat{\mathbf{L}}\|_F}{\sqrt{NT}} \le C\left[\left(\sqrt{\frac{L_{\max}^2\log(N+T)}{N p_c}} \vee \sqrt{\frac{\sigma^2 R\log(N+T)}{T p_c^2}} \vee \sqrt{\frac{\sigma^2 R\log^3(N+T)}{N p_c^2}}\right) + \sqrt{\frac{L_{\max}^2 RT}{N p_c^2}}\right]. \qquad (14)$$

6.2.1. Interpretation of Theorem 2

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 (14) converges to 0 when $\mathbf{L}^*$ is low-rank (R is constant), $N \ge T$, 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 Section 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.

6.2.2. Comparison With Existing Theory on Matrix-Completion

Our estimator and its theoretical analysis are motivated by and generalize existing research on matrix-completion (Srebro, Alon, and Jaakkola 2005; Candès and Recht 2009; Candès and Tao 2010; Keshavan, Montanari, and Oh 2010a, 2010b; Mazumder, Hastie, and Tibshirani 2010; Gross 2011; Koltchinskii, Lounici, and Tsybakov 2011; Negahban and Wainwright 2011, 2012; Recht 2011; Rohde and Tsybakov 2011; Klopp 2014). The main difference is in our observation model $\mathcal{O}$. Existing articles 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 $N^{1/3} = T$, 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 (14) is also random, depending on $\mathcal{O}$ and the noise, but the right-hand side of (14) is deterministic. 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, Diamond, and Hainmueller (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 Section 4.

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, Diamond, and Hainmueller (2010) with $N = 38, T = 31$. Note that in the original dataset 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.

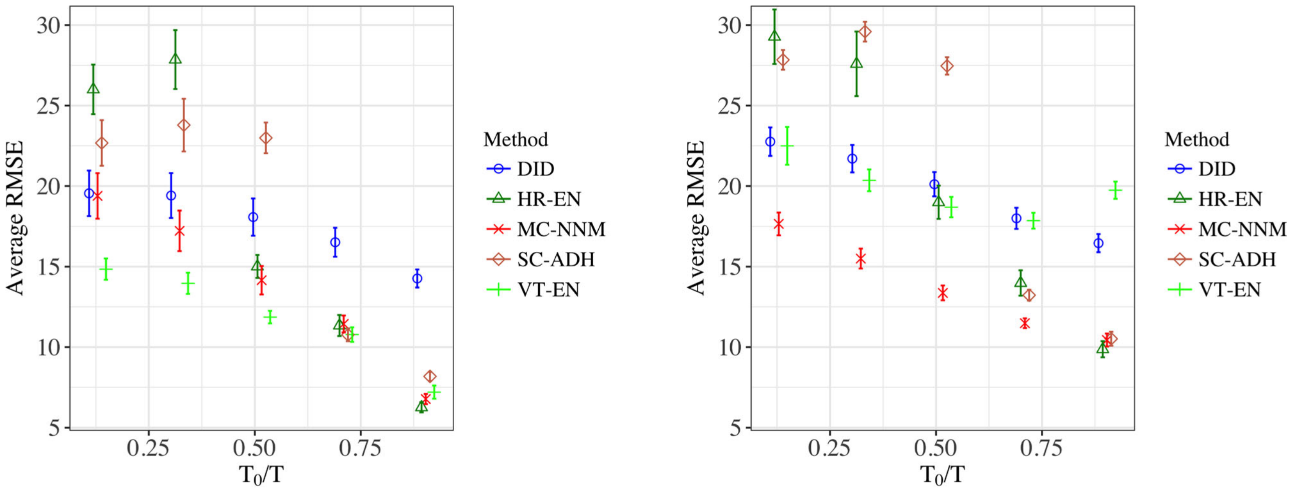
We consider two settings for the treatment adoption:

- 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 (pretreatment values of treatment units).

7.2. Stock Market Data

In the next illustration, we use a financial dataset—daily returns for 2453 stocks over 10 years (3082 days). Since we only have access to a single instance of the data, to observe statistical fluctuations of the RMSE, for each N and T we create 50 subsamples by looking at the first T daily returns of N randomly sampled stocks for a range of pairs of (N, T), always with $N \times T = 4900$, ranging from very thin to very fat, $(N, T) = (490, 10), \ldots, (N, T) = (70, 70), \ldots, (N, T) = (10, 490)$, with in each case the second half the entries missing for a randomly selected half the units (so 25% of the entries missing overall), in a block design. Here we focus on the comparison between the



(a) Simultaneous adoption, $N_t = 8$

(b) Staggered adoption, $N_t = 35$

Figure 1. California smoking data.

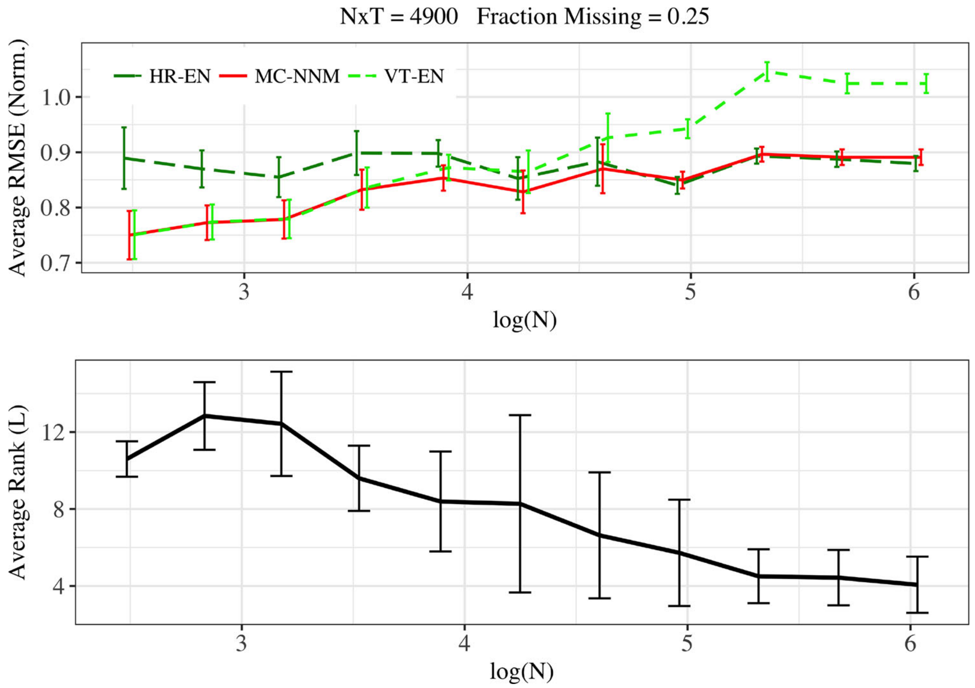


Figure 2. Stock market data.

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. 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 4–12, where smaller ranks are used as N grows relative to T. This validates the fact that there is a stronger correlation between daily return of different stocks than between returns for different time periods of the same stock.

8. Generalizations

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

8.1. The Model With Covariates

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

$$Y_{it} = L^*_{it} + \sum_{p=1}^{P}\sum_{q=1}^{Q} X_{ip}H^*_{pq}Z_{qt} + \gamma^*_i + \delta^*_t + V_{it}^\top \beta^* + \varepsilon_{it} \quad (15)$$

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

$$\mathbf{Y} = \mathbf{L}^* + \mathbf{X}\mathbf{H}^*\mathbf{Z}^\top + \Gamma^*\mathbf{1}_T^\top + \mathbf{1}_N(\Delta^*)^\top + \left[V_{it}^\top\beta^*\right]_{it} + \boldsymbol{\varepsilon}. \quad (16)$$

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 $\mathbb{R}^{T\times 1}$. A slightly richer version of this model that allows linear terms in covariates can be defined as by

$$\mathbf{Y} = \mathbf{L}^* + \tilde{\mathbf{X}}\tilde{\mathbf{H}}^*\tilde{\mathbf{Z}}^\top + \Gamma^*\mathbf{1}_T^\top + \mathbf{1}_N(\Delta^*)^\top + \left[V_{it}^\top\beta^*\right]_{it} + \boldsymbol{\varepsilon}, \quad (17)$$

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,

$$\mathbf{Y} = \mathbf{L}^* + \tilde{\mathbf{X}}\tilde{\mathbf{H}}^*_{X,Z}\tilde{\mathbf{Z}}^\top + \tilde{\mathbf{H}}^*_Z\tilde{\mathbf{Z}}^\top + \mathbf{X}\tilde{\mathbf{H}}^*_X + \Gamma^*\mathbf{1}_T^\top + \mathbf{1}_N(\Delta^*)^\top + \left[V_{it}^\top\beta^*\right]_{it} + \boldsymbol{\varepsilon}. \quad (18)$$

From now on, we will use the richer model (18) 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_λ, defined in Section 2, that performs coordinate descent with respect to $\mathbf{H}$. Then we can apply this operator after each step of using shrink_λ. 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 (8).

$$(\hat{\mathbf{L}}, \hat{\Gamma}, \hat{\Delta}, \hat{\tau}) = \arg\min_{\mathbf{L},\Gamma,\Delta,\tau}\left\{\frac{1}{NT}\|\mathbf{Y} - \mathbf{L} - \Gamma\mathbf{1}_T^\top - \mathbf{1}_N\Delta^\top - \tau\mathbf{W}\|_F^2 + \lambda\|\mathbf{L}\|_*\right\}, \quad (19)$$

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 (8) 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 ith row of $\mathbf{Y}$ and $\mathbf{L}$, respectively. The original objective function for $\mathcal{O} = [N] \times [T]$ is

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

We can modify this to $\sum_{i=1}^{N}(Y_{i\cdot} - L_{i\cdot})\boldsymbol{\Omega}^{-1}(Y_{i\cdot} - L_{i\cdot})^\top/|\mathcal{O}| + \lambda_L\|\mathbf{L}\|_*$, where the choice for the $T \times T$ matrix $\boldsymbol{\Omega}$ would reflect the autocorrelation in the $\boldsymbol{\varepsilon}_{it}$. For example, with a first-order autoregressive process, we would use $\Omega_{ts} = \sigma^2\rho^{|t-s|}$, 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 modeling this probability in terms of the covariates and a latent factor structure. Let the propensity score be $e_{it} = \mathbb{P}(W_{it} = 1|X_i, Z_t, V_{it})$, and let $\mathbf{E}$ be the $N \times T$ matrix with typical element e_{it}. Let us again consider the case without covariates. In that case we may wish to model the assignment $\mathbf{W}$ as

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

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

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

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

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

8.5. *Relaxing the Dependence of Theorem 2 on p_c*

Recall from Section 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 $p_c = \min_{1\le i\le N} \pi_T^{(i)}$ rather than $\sum_{i=1}^N \pi_T^{(i)}/N$. The reason for this, as discussed in Section 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 nonzero). 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 nonzero 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 Section 6.1, we assumed that all entries (i, t) of $\mathbf{Y}$ for $t \le t_i$ 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 pretreatment period. Rather, such scenario can occur when measuring outcome values is costly. In this case, one can extend Theorem 2 to the setting with $\mathcal{O} = \left[\bigcup_{i=1}^N \left\{(i, 1), (i, 2), \dots , (i, t_i)\right\}\right] \backslash \mathcal{O}_{\text{miss}}$, where each $(i, t) \in \cup_{i=1}^N \{(i, 1), (i, 2), \dots , (i, t_i)\}$ 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, Diamond, and Hainmueller (2010) and the elastic net estimators proposed by Doudchenko and Imbens (2016).

Acknowledgments

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.

Funding

This research was generously supported by ONR grant N00014-19-1-2468 and NSF grant CMMI:1554140.

References

Abadie, A. (2019), "Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects," *Journal of Economic Literature* (forthcoming). [1,3]

Abadie, A., and Cattaneo, M. D. (2018), "Econometric Methods for Program Evaluation," *Annual Review of Economics*, 10, 465–503. [4]

Abadie, A., Diamond, A., and Hainmueller, J. (2010), "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program," *Journal of the American Statistical Association*, 105, 493–505. [1,3,4,5,7,10,13]

——— (2015), "Comparative Politics and the Synthetic Control Method," *American Journal of Political Science*, 59, 495–510. [1]

Abadie, A., and Gardeazabal, J. (2003), "The Economic Costs of Conflict: A Case Study of the Basque Country," *American Economic Review*, 93, 113–132. [1]

Amjad, M., Shah, D., and Shen, D. (2018), "Robust Synthetic Control," *Journal of Machine Learning Research*, 19, 1–51. [1,5]

Anderson, T. W. (1958), *An Introduction to Multivariate Statistical Analysis* (Vol. 2), New York: Wiley. [5]

Angrist, J., and Pischke, S. (2008), *Mostly Harmless Econometrics: An Empiricists' Companion*, Princeton, NJ: Princeton University Press. [5]

Arellano, M., and Honoré, B. (2001), "Panel Data Models: Some Recent Developments," *Handbook of Econometrics*, 5, 3229–3296. [5]

Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., and Wager, S. (2019), "Synthetic Difference in Differences," forthcoming, *American Economic Review*. [1]

Athey, S., and Imbens, G. W. (2018), "Design-Based Analysis in Difference-in-Differences Settings With Staggered Adoption," Technical Report, National Bureau of Economic Research. [2,4]

Athey, S., and Stern, S. (2002), "The Impact of Information Technology on Emergency Health Care Outcomes," *The RAND Journal of Economics*, 33, 399–432. [4]

Bai, J. (2003), "Inferential Theory for Factor Models of Large Dimensions," *Econometrica*, 71, 135–171. [1,5]

——— (2009), "Panel Data Models With Interactive Fixed Effects," *Econometrica*, 77, 1229–1279. [5]

Bai, J., and Ng, S. (2002), "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70, 191–221. [1,2,5]

——— (2017), "Principal Components and Regularized Estimation of Factor Models," arXiv no. 1708.08137. [5]

Ben-Michael, E., Feller, A., and Rothstein, J. (2018), "The Augmented Synthetic Control Method," arXiv no. 1811.04170. [1]

Candès, E. J., and Plan, Y. (2010), "Matrix Completion With Noise," *Proceedings of the IEEE*, 98, 925–936. [2,6]

Candès, E. J., and Recht, B. (2009), "Exact Matrix Completion via Convex Optimization," *Foundations of Computational Mathematics*, 9, 717. [2,5,6,9]

Candès, E. J., and Tao, T. (2010), "The Power of Convex Relaxation: Near-Optimal Matrix Completion," *IEEE Transactions on Information Theory*, 56, 2053–2080. [3,5,9]

Chamberlain, G. (1984), "Panel Data," *Handbook of Econometrics*, 2, 1247–1318. [5]

——— (1993), "Feedback in Panel Data Models," Technical Report, Harvard-Institute of Economic Research. [2]

Chernozhukov, V., Wuthrich, K., and Zhu, Y. (2017), "An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls," arXiv no. 1712.09089. [1]

Doudchenko, N., and Imbens, G. W. (2016), "Balancing, Regression, Difference-in-Differences and Synthetic Control Methods: A Synthesis," Technical Report, National Bureau of Economic Research. [1,4,5,13]

Ferman, B., and Pinto, C. (2019), "Synthetic Controls With Imperfect Pre-Treatment Fit," arXiv no. 1911.08521. [1,4]

Gamarnik, D., and Misra, S. (2016), "A Note on Alternating Minimization Algorithm for the Matrix Completion Problem," *IEEE Signal Processing Letters*, 23, 1340–1343. [2]

Gobillon, L., and Magnac, T. (2016), "Regional Policy Evaluation: Interactive Fixed Effects and Synthetic Controls," *Review of Economics and Statistics*, 98, 535–551. [5]

Goldberger, A. S. (1972), "Structural Equation Methods in the Social Sciences," *Econometrica: Journal of the Econometric Society*, 40, 979–1001. [5]

Gross, D. (2011), "Recovering Low-Rank Matrices From Few Coefficients in Any Basis," *IEEE Transactions on Information Theory*, 57, 1548–1566. [5,9]

Hamidi, N., and Bayati, M. (2019), "On Low-Rank Trace Regression Under General Sampling Distribution," arXiv no. 1904.08576.

Hastie, T., Mazumder, R., Lee, J. D., and Zadeh, R. (2015), "Matrix Completion and Low-Rank SVD via Fast Alternating Least Squares," *Journal of Machine Learning Research*, 16, 3367–3402. [8]

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning*, New York: Springer. [6]

Hernan, M. A., and Robins, J. M. (2010), *Causal Inference*, Boca Raton, FL: CRC Press. [2]

Hirano, K., Imbens, G. W., and Ridder, G. (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71, 1161–1189. [2]

Hsiao, C., Steve Ching, H., and Ki Wan, S. (2012), "A Panel Data Approach for Program Evaluation: Measuring the Benefits of Political and Economic Integration of Hong Kong With Mainland China," *Journal of Applied Econometrics*, 27, 705–740. [5]

Imbens, G. W. and Rubin, D. B. (2015), *Causal Inference in Statistics, Social, and Biomedical Sciences*, New York: Cambridge University Press. [1,2,3]

Imbens, G. W., and Wooldridge, J. M. (2009), "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47, 5–86. [4]

Keshavan, R. H., Montanari, A., and Oh, S. (2010a), "Matrix Completion From a Few Entries," *IEEE Transactions on Information Theory*, 56, 2980–2998. [5,9]

——— (2010b), "Matrix Completion From Noisy Entries," *Journal of Machine Learning Research*, 11, 2057–2078. [5,9]

Kim, D., and Oka, T. (2014), "Divorce Law Reforms and Divorce Rates in the USA: An Interactive Fixed-Effects Approach," *Journal of Applied Econometrics*, 29, 231–245. [5]

Klopp, O. (2014), "Noisy Low-Rank Matrix Completion With General Sampling Distribution," *Bernoulli*, 20, 282–303. [5,9]

Koltchinskii, V., Lounici, K., and Tsybakov, A. B. (2011), "Nuclear-Norm Penalization and Optimal Rates for Noisy Low-Rank Matrix Completion," *The Annals of Statistics*, 39, 2302–2329. [5,9]

Ledoux, M., and Talagrand, M. (2013), *Probability in Banach Spaces: Isoperimetry and Processes*, Berlin, Heidelberg: Springer.

Li, K. T. (2020), "Statistical Inference for Average Treatment Effects Estimated by Synthetic Control Methods," *Journal of the American Statistical Association*, 115, 2068–2083. [1]

Liang, K.-Y., and Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 13–22. [5]

Massart, P. (2000), "About the Constants in Talagrand's Concentration Inequalities for Empirical Processes," *The Annals of Probability*, 28, 863–884.

Mazumder, R., Hastie, T., and Tibshirani, R. (2010), "Spectral Regularization Algorithms for Learning Large Incomplete Matrices," *Journal of Machine Learning Research*, 11, 2287–2322. [2,6,8,9]

Moon, H. R., and Weidner, M. (2015), "Linear Regression for Panel With Unknown Number of Factors as Interactive Fixed Effects," *Econometrica*, 83, 1543–1579. [5]

——— (2017), "Dynamic Linear Panel Regression Models With Interactive Fixed Effects," *Econometric Theory*, 33, 158–195. [5]

Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. (2012), "A Unified Framework for High-Dimensional Analysis of *M*-Estimators With Decomposable Regularizers," *Statistical Science*, 27, 538–557.

Negahban, S. N., and Wainwright, M. J. (2011), "Estimation of (Near) Low-Rank Matrices With Noise and High-Dimensional Scaling," *The Annals of Statistics*, 39, 1069–1097. [5,9]

——— (2012), "Restricted Strong Convexity and Weighted Matrix Completion: Optimal Bounds With Noise," *Journal of Machine Learning Research*, 13, 1665–1697. [5,9,13]

Pesaran, M. H. (2006), "Estimation and Inference in Large Heterogeneous Panels With a Multifactor Error Structure," *Econometrica*, 74, 967–1012. [5]

Recht, B. (2011), "A Simpler Approach to Matrix Completion," *Journal of Machine Learning Research*, 12, 3413–3430. [3,5,9]

Rohde, A., and Tsybakov, A. B. (2011), "Estimation of High-Dimensional Low-Rank Matrices," *The Annals of Statistics*, 39, 887–930. [5,9]

Rosenbaum, P. R., and Rubin, D. B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55. [1,4]

Rubin, D. B. (2006), *Matched Sampling for Causal Effects*, Cambridge: Cambridge University Press. [2,4]

Shaikh, A., and Toulis, P. (2019), "Randomization Tests in Observational Studies With Staggered Adoption of Treatment," University of Chicago, Becker Friedman Institute for Economics Working Paper 2019-144. [2,4]

Srebro, N., Alon, N., and Jaakkola, T. S. (2005), “Generalization Error Bounds for Collaborative Prediction With Low-Rank Matrices,” in *Advances in Neural Information Processing Systems* (Vol. 17), eds. L. K. Saul, Y. Weiss, and L. Bottou, pp. 1321–1328. [5,9]

Tibshirani, R. (1996), “Regression Shrinkage and Selection via the Lasso,” *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [2]

Tropp, J. A. (2012), “User-Friendly Tail Bounds for Sums of Random Matrices,” *Foundations of Computational Mathematics*, 12, 389–434.

Wang, Y., Liang, D., Charlin, L., and Blei, D. M. (2018), “The Deconfounded Recommender: A Causal Inference Approach to Recommendation,” arXiv no. 1808.06581. [5]

Xu, Y. (2017), “Generalized Synthetic Control Method: Causal Inference With Interactive Fixed Effects Models,” *Political Analysis*, 25, 57–76. [2,5]