Doubly Robust Inference in Causal Latent Factor Models

Alberto Abadie
Department of Economics, MIT

Anish Agarwal
Department of IEOR, Columbia University

Raaz Dwivedi
Department of ORIE, Cornell Tech

Abhin Shah
Department of EECS, MIT

January 2, 2026

Abstract

This article introduces a new estimator of average treatment effects under unobserved confounding in modern data-rich environments featuring large numbers of units and outcomes. The proposed estimator is doubly robust, combining outcome imputation, inverse probability weighting, and a novel cross-fitting procedure for matrix completion. We derive finite-sample and asymptotic guarantees, and show that the error of the new estimator converges to a mean-zero Gaussian distribution at a parametric rate. Simulation results demonstrate the relevance of the formal properties of the estimators analyzed in this article.

Keywords: Average treatment effects, unobserved confounding, matrix completion, cross-fitting

---

Alberto Abadie, Department of Economics, MIT, abadie@mit.edu. Anish Agarwal, Department of Industrial Engineering and Operations Research, Columbia University, aa5194@columbia.edu. Raaz Dwivedi, Department of Operations Research and Information Engineering, Cornell Tech, dwivedi@cornell.edu. Abhin Shah, Department of Electrical Engineering and Computer Science, MIT, abhin@mit.edu. We are grateful to Haruki Kono, Guido Imbens, James Robins, Stefan Wager, and seminar participants at Columbia, MIT, the Online Causal Inference Seminar, and Stanford for helpful comments and discussion. Abadie gratefully acknowledges ONR support under grant N00014-24-1-2687.

1 Introduction

This article presents a novel framework for the estimation of average treatment effects in modern data-rich environments in the presence of unobserved confounding. We define modern data-rich environments as those featuring many outcome measurements across a wide range of units. Our interest in data-rich environments stems from the emergence of digital platforms (e.g., internet retailers, social media companies, and ride-sharing companies), electronic medical records systems, IoT devices, and other real-time digitized data systems, which gather economic and social behavior data with unprecedented scope and granularity.

Take the example of an internet retailer. The platform collects not only information on purchases of many customers across many products or product categories, but also on glance views, impressions, conversions, engagement metrics, navigation paths, shipping choices, payment methods, returns, reviews, and more. While some variables, such as geo-location and type of device or browser, can be safely treated as pre-determined relative to the platform's treatments (advertisements, discounts, web-page design, etc.), most are outcomes affected by the treatments, latent customer preferences, and unobserved product features. We leverage the availability of many outcome measures in modern data-rich environments to estimate average treatment effects in the presence of unobserved confounding. The core identification concept is that if each element of a high-dimensional outcome vector is influenced by a common low-dimensional vector of unobserved confounders, it becomes possible to remove the influence of the confounders and identify treatment effects.

Two primary approaches to the estimation of treatment effects are outcome-based and assignment-based methods. Consider again the example of an internet-retail platform where customers interact with various product categories. For each consumer-category pair, the platform makes decisions to either offer a discount or not, and records whether the consumer purchased a product in the category. Outcome-based methods operate by imputing the missing potential outcomes for each consumer-product category pair. This process involves predicting whether a consumer, who received a discount, would have made the purchase without the discount (i.e., the potential outcome without discount), and conversely, if a consumer who did not receive the discount would have purchased the product had they received the discount (i.e., the potential outcome with discount). In contrast, assignment-based methods estimate the probabilities of consumers receiving discounts in each product category and adjust for missing potential outcomes by weighting observed outcomes inversely to the probability of missingness.

A substantial body of literature has explored outcome-based methods, particularly in settings where all confounding factors are measured (see, e.g., Cochran, 1968; Rosenbaum and Rubin, 1983; Angrist, 1998; Abadie and Imbens, 2006, among many others). Imputing potential outcomes in the presence of unobserved confounders poses a more complex challenge. In this context, a commonly adopted framework is the synthetic control method and its variants (see, e.g., Abadie and Gardeazabal, 2003; Abadie et al., 2010; Cattaneo et al., 2021; Arkhangelsky et al., 2021). An alternative but related approach to outcome imputation under unobserved confounding is the latent factor framework (Bai and Ng, 2002; Bai, 2009; Xiong and Pelger, 2023), wherein each element of the large-dimensional outcome vector is influenced by the same low-dimensional vector of unobserved confounders. Matrix completion methods (see, e.g., Chatterjee, 2015; Athey et al., 2021; Bai and Ng, 2021; Dwivedi et al., 2022a; Agarwal et al., 2023a) which have found widespread applications in recommendation systems and panel data models, are closely related to latent factor models. Similarly, existing assignment-based procedures to estimate average treatment effects rely on the assumption of no unmeasured confounding (see, e.g., Robins et al., 2000; Hirano et al., 2003; Wooldridge, 2007), common trends restrictions (Abadie, 2005), or the availability of an instrumental variable (Abadie, 2003; Sloczynski et al., 2024).

In this article, we propose a doubly-robust estimator (see Robins et al., 1994; Bang and

Robins, 2005; Chernozhukov et al., 2018) of average treatment effects in the presence of unobserved confounding. This estimator leverages information on both the outcome process and the treatment assignment mechanism under a latent factor framework. It combines outcome imputation and inverse probability weighting with a new cross-fitting approach for matrix completion. We show that the proposed doubly-robust estimator has better finite-sample guarantees than alternative outcome-based and assignment-based estimators. Furthermore, the doubly-robust estimator is approximately Gaussian, asymptotically unbiased, and converges at a parametric rate, under provably valid error rates for matrix completion, irrespective of other properties of the matrix completion algorithm used.

Our article is related to Feng (2021), which also exploits latent structure in both treatment assignment and outcome processes to derive a doubly-robust estimator for average treatment effects in the presence of unobserved confounding. Relative to Feng (2021), our approach does not rely on a large block of outcomes that load on the latent confounders but remain unaffected by treatment. Moreover, we derive general finite-sample error bounds and a Gaussian approximation for the doubly robust estimator, introduce a meta cross-fitting algorithm that satisfies the required conditions, and establish primitive assumptions that accommodate broad forms of nonstationarity in the outcome process. Another closely related work is Choi et al. (2024), which derives inferential guarantees for matrix completion under dependent observation patterns and heterogeneous treatment probabilities. The paper establishes asymptotic normality for an estimator that combines nuclear-norm penalization with eigenvector estimation, under an approximate low-rank factor model for outcomes and a rank-one structure for the propensity matrix. Our results allow for a broader class of matrix-completion algorithms and do not impose a rank-one restriction on the propensity matrix.

Arkhangelsky and Imbens (2022) study doubly-robust identification with longitudinal data under the assumption that conditioning of a function of the treatment assignments over time (e.g., the fraction of times an individual is exposed to treatment) is enough to remove confounding. Athey et al. (2021), Bai and Ng (2021), Dwivedi et al. (2022a), Agarwal et al. (2023a), Xiong and Pelger (2023), and Choi and Yuan (2024) develop matrix-completion estimators for imputing potential outcomes. Among these, only Dwivedi et al. (2022a) allows for randomized treatment assignment; the remaining papers require large sub-blocks of units exposed to the same treatment, as in staggered-adoption designs. These approaches impose a low-rank structure on the outcome process but do not model analogous latent structure in the assignment mechanism. Our article fills this gap and shows that leveraging structure in the assignment process can yield substantial gains.

Terminology and notation. For any real number $b \in \mathbb{R}$, $\lfloor b \rfloor$ is the greatest integer less than or equal to b. For any positive integer b, $[b]$ denotes the set of integers from 1 to b, i.e., $[b] \triangleq \{1, \cdots, b\}$. We use c to denote any generic universal constant, whose value may change between instances. For any $c > 0$, $m(c) = \max\{c, \sqrt{c}\}$ and $\ell_c = \log(2/c)$. For any two deterministic sequences a_n and b_n where b_n is positive, $a_n = O(b_n)$ means that there exist a finite $c > 0$ and a finite $n_0 > 0$ such that $|a_n| \leq c\, b_n$ for all $n \geq n_0$. Similarly, $a_n = o(b_n)$ means that for every $c > 0$, there exists a finite $n_0 > 0$ such that $|a_n| < c\, b_n$ for all $n \geq n_0$. Further, $a_n = \Omega(b_n)$ means that there exist a finite $c > 0$ and a finite $n_0 > 0$ such that $|a_n| \geq c\, b_n$ for all $n \geq n_0$. For a sequence of random variables, $x_n = O_p(1)$ means that the sequence $|x_n|$ is stochastically bounded, i.e., for every $\varepsilon > 0$, there exists a finite $\delta > 0$ and a finite $n_0 > 0$ such that $\mathbb{P}\Big(|x_n| > \delta\Big) < \varepsilon$ for all $n \geq n_0$. Similarly, $x_n = o_p(1)$ means that the sequence $|x_n|$ converges to zero in probability, i.e., for every $\varepsilon > 0$ and $\delta > 0$, there exists a finite $n_0 > 0$ such that $\mathbb{P}\Big(|x_n| > \delta\Big) < \varepsilon$ for all $n \geq n_0$. For sequences of random variables x_n and b_n, $x_n = O_p(b_n)$ means $x_n = \overline{x}_n b_n$ where the sequence $\overline{x}_n = O_p(1)$. Likewise, $x_n = o_p(b_n)$ means $x_n = \overline{x}_n b_n$ where the sequence $\overline{x}_n = o_p(1)$.

A mean-zero random variable x is subGaussian if there exists some $b > 0$ such that $\mathbb{E}[\exp(sx)] \leq \exp(b^2 s^2/2)$ for all $s \in \mathbb{R}$. Then, the subGaussian norm of x is given by

$\|x\|_{\psi_2} = \inf\{t > 0 : \mathbb{E}[\exp(x^2/t^2)] \leq 2\}$. A mean-zero random variable x is subExponential if there exist some $b_1, b_2 > 0$ such that $\mathbb{E}[\exp(sx)] \leq \exp(b_1^2 s^2/2)$ for all $-1/b_2 < s < 1/b_2$. Then, the subExponential norm of x is given by $\|x\|_{\psi_1} = \inf\{t > 0 : \mathbb{E}[\exp(|x|/t)] \leq 2\}$. $\texttt{Uniform}(a, b)$ denotes the uniform distribution over the interval $[a, b]$ for $a, b \in \mathbb{R}$ such that $a < b$. $\mathcal{N}(\mu, \sigma^2)$ denotes the Gaussian distribution with mean μ and variance σ^2.

For a vector $u \in \mathbb{R}^n$, we denote its t^{th} coordinate by u_t and its 2-norm $\|u\|_2$. For a matrix $U \in \mathbb{R}^{n_1 \times n_2}$, we denote the element in i^{th} row and j^{th} column by $u_{i,j}$, the i^{th} row by $U_{i,\cdot}$, the j^{th} column by $U_{\cdot,j}$, the largest eigenvalue by $\lambda_{\max}(U)$, and the smallest by $\lambda_{\min}(U)$. Given a set of indices $\mathcal{R} \subseteq [n_1]$ and $\mathcal{C} \subseteq [n_2]$, $U_{\mathcal{I}} \in \mathbb{R}^{|\mathcal{R}| \times |\mathcal{C}|}$ is a sub-matrix of U corresponding to the entries in $\mathcal{I} \triangleq \mathcal{R} \times \mathcal{C}$, and $U_{-\mathcal{I}} = \{u_{i,j} : (i,j) \in \{[n_1] \times [n_2]\} \setminus \mathcal{I}\}$. Further, we denote the Frobenius norm by $\|U\|_{\mathrm{F}} \triangleq \left(\sum_{i \in [n_1], j \in [n_2]} u_{i,j}^2\right)^{1/2}$, the $(1,2)$ operator norm by $\|U\|_{1,2} \triangleq \max_{j \in [n_2]} \left(\sum_{i \in [n_1]} u_{i,j}^2\right)^{1/2}$, the $(2,\infty)$ operator norm by $\|U\|_{2,\infty} \triangleq \max_{i \in [n_1]} \left(\sum_{j \in [n_2]} u_{i,j}^2\right)^{1/2}$, and the maximum norm by $\|U\|_{\max} \triangleq \max_{i \in [n_1], j \in [n_2]} |u_{i,j}|$. Given two matrices $U, V \in \mathbb{R}^{n_1 \times n_2}$, the operators $\odot$ and $\oslash$ denote element-wise multiplication and division, respectively, i.e., $t_{i,j} = u_{i,j} \cdot v_{i,j}$ when $T = U \odot V$, and $t_{i,j} = u_{i,j}/v_{i,j}$ when $T = U \oslash V$. When V is a binary matrix, i.e., $V \in \{0,1\}^{n_1 \times n_2}$, the operator $\otimes$ is defined such that $t_{i,j} = u_{i,j}$ if $v_{i,j} = 1$ and $t_{i,j} = ?$ if $v_{i,j} = 0$ for $T = U \otimes V$. Given two matrices $U \in \mathbb{R}^{n_1 \times n_2}$ and $V \in \mathbb{R}^{n_1 \times n_3}$, the operator $*$ denotes the (transposed column-wise) Khatri-Rao product of U and V, i.e., $T = U * V \in \mathbb{R}^{n_1 \times n_2 n_3}$ such that $t_{i,j} = u_{i,j - n_2 \bar{j}} \cdot v_{i,1+\bar{j}}$ where $\bar{j} = \lfloor (j-1)/n_2 \rfloor$. For random objects U and V, $U \perp\!\!\!\perp V$ means that U is independent of V.

2 Setup

Consider a setting with N units and M measurements per unit. For each unit-measurement pair $i \in [N]$ and $j \in [M]$, we observe a treatment assignment $a_{i,j} \in \{0,1\}$ and the value of the outcome $y_{i,j} \in \mathbb{R}$. Although our results can be easily generalized to multi-ary treatments, for the ease of exposition, we focus on binary treatments.

We operate within the Neyman-Rubin potential outcomes framework and denote the potential outcome for unit $i \in [N]$ and measurement $j \in [M]$ under treatment $a \in \{0, 1\}$ by $y_{i,j}^{(a)} \in \mathbb{R}$. A no-spillover assumption is implicit in the notation, i.e., the potential outcome $y_{i,j}^{(a)}$ does not depend on the treatment assignment for any other unit-measurement pair. In the context of online retail data, the assumption of no spillovers across measurements is justified if the cross-elasticity of demand across product categories, j, is low. Our framework allows for the possibility that the same treatment affects multiple outcomes (e.g., $a_{i,j} = a_{i,j'}$ with probability one, for some j and j' in $[M]$). Realized outcomes, $y_{i,j}$, depend on potential outcomes and treatment assignments,

$$y_{i,j} = y_{i,j}^{(0)}(1 - a_{i,j}) + y_{i,j}^{(1)} a_{i,j}, \tag{1}$$

for all $i \in [N]$ and $j \in [M]$. Section 4.4 and the supplementary appendix extend the framework proposed in this article to a panel data setting with lagged treatment effects.

2.1 Sources of stochastic variation

In the setup of this article, each unit $i \in [N]$ is characterized by a set of unknown parameters, $\{(\theta_{i,j}^{(0)}, \theta_{i,j}^{(1)}, p_{i,j}) \in \mathbb{R}^2 \times [0, 1]\}_{j \in [M]}$, which we treat as fixed. Potential outcomes and treatment assignments are generated as follows: for all $i \in [N], j \in [M]$, and $a \in \{0, 1\}$,

$$y_{i,j}^{(a)} = \theta_{i,j}^{(a)} + \varepsilon_{i,j}^{(a)} \tag{2}$$

and

$$a_{i,j} = p_{i,j} + \eta_{i,j}, \tag{3}$$

where $\varepsilon_{i,j}^{(a)}$ and $\eta_{i,j}$ are mean-zero random variables, and

$$\eta_{i,j} = \begin{cases} -p_{i,j} & \text{with probability} \quad 1 - p_{i,j} \\ 1 - p_{i,j} & \text{with probability} \quad p_{i,j}. \end{cases} \tag{4}$$

It follows that $\theta_{i,j}^{(a)}$ is the mean of the potential outcome $y_{i,j}^{(a)}$, and $p_{i,j}$ is the unknown assignment probability or latent propensity score. Let $\Theta^{(0)} \triangleq \{\theta_{i,j}^{(0)}\}_{i\in[N],j\in[M]}$, $\Theta^{(1)} \triangleq \{\theta_{i,j}^{(1)}\}_{i\in[N],j\in[M]}$, and $P \triangleq \{p_{i,j}\}_{i\in[N],j\in[M]}$ collect mean potential outcomes and assignment probabilities. For the rest of the article, we condition on $\Theta^{(0)}, \Theta^{(1)}$, and P, implying that $E^{(0)} \triangleq \{\varepsilon_{i,j}^{(0)}\}_{i\in[N],j\in[M]}$, $E^{(1)} \triangleq \{\varepsilon_{i,j}^{(1)}\}_{i\in[N],j\in[M]}$, and $W \triangleq \{\eta_{i,j}\}_{i\in[N],j\in[M]}$ capture all sources of randomness in potential outcomes and treatment assignments.

Our setup allows $\Theta^{(0)}, \Theta^{(1)}$ to be arbitrarily associated with P, inducing unobserved confounding. The assumptions in Section 4 imply that $\Theta^{(0)}, \Theta^{(1)}$, and P include all confounding factors, and require $(\varepsilon_{i,j}^{(0)}, \varepsilon_{i,j}^{(1)}) \perp\!\!\!\perp \eta_{i,j}$ for every $i \in [N]$ and $j \in [M]$.

2.2 Target causal estimand

For any given measurement $j \in [M]$, we aim to estimate the effect of the treatment averaged over all units,

$$\text{ATE}_{\cdot,j} \triangleq \mu_{\cdot,j}^{(1)} - \mu_{\cdot,j}^{(0)} \quad \text{where} \quad \mu_{\cdot,j}^{(a)} \triangleq \frac{1}{N} \sum_{i\in[N]} \theta_{i,j}^{(a)}. \tag{5}$$

$\text{ATE}_{\cdot,j}$ is akin to the conditional average treatment effect of Abadie and Imbens (2006), but based on the latent means, $\theta_{i,j}^{(a)}$, in Eq. (2) rather than on conditional means that depend on observed covariates only. It is straightforward to adapt the methods in this article to the estimation of alternative parameters, like the average treatment effect across measurements for each unit i, or the estimation of treatment effects over a subset of the units, $S \subset [N]$.

3 Estimation

In this section, we propose a procedure that uses the treatment assignment matrix A and the observed outcomes matrix Y to estimate $\text{ATE}_{\cdot,j}$, where

$$Y \triangleq \{y_{i,j}\}_{i\in[N],j\in[M]} \quad \text{and} \quad A \triangleq \{a_{i,j}\}_{i\in[N],j\in[M]}.$$

The estimator proposed in this section leverages matrix completion as a key subroutine. We start the section with a brief overview of matrix completion methods.

3.1 Matrix completion: A primer

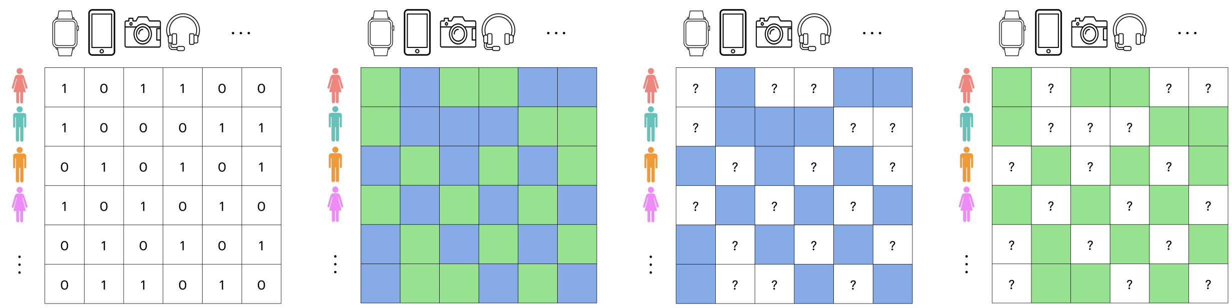
Consider a matrix of parameters $T \in \mathbb{R}^{N\times M}$. While T is unobserved, we observe $S \in \{\mathbb{R} \cup \{\,?\}\}^{N\times M}$ where $?$ denotes a missing value. The relationship between S and T is:

$$S = (T + H) \otimes F. \tag{6}$$

Here, $H \in \mathbb{R}^{N\times M}$ is a noise matrix, and $F \in \{0, 1\}^{N\times M}$ is a masking matrix with ones for the recorded entries of S and zeros for the missing entries.

A matrix completion algorithm, denoted by $\mathtt{MC}$, takes the S as its input, and returns an estimate of T, which we denote by $\widehat{T}$ or $\mathtt{MC}(S)$. In other words, $\mathtt{MC}$ produces an estimate of a matrix from noisy observations of a subset of all the elements of the matrix.

The matrix completion literature is rich with algorithms $\mathtt{MC}$ that provide error guarantees, namely bounds on $\|\mathtt{MC}(S) - T\|$ for a suitably chosen norm/metric $\|\cdot\|$, under a variety of assumptions on the triplet (T, H, F). Typical assumptions are (i) T is low-rank, (ii) the entries of H are independent, mean-zero and sub-Gaussian random variables, and (iii) the entries of F are independent Bernoulli random variables. Though matrix completion is commonly associated with the imputation of missing values, a typically underappreciated aspect is that it also denoises the observed matrix. Even when each entry of S is observed, $\mathtt{MC}(S)$ subtracts the effects of H from S, i.e., it performs matrix denoising. Nguyen et al.



(a) A (b) Y (c) $Y^{(0),\text{obs}}$ (d) $Y^{(1),\text{obs}}$

Figure 1: Schematic of the treatment assignment matrix A, the observed outcomes matrix Y (where green and blue fills indicate observations under $a = 1$ and $a = 0$, respectively), and the observed component of the potential outcomes matrices, i.e., $Y^{(0),\text{obs}}$ and $Y^{(1),\text{obs}}$ (where ? indicates a missing value). All matrices are $N \times M$ where N is the number of customers and M is the number of products.

(2019) provide a survey of various matrix completion algorithms.

3.2 Key building blocks

We now define and express matrices that are related to the quantities of interest $\Theta^{(0)}, \Theta^{(1)}$, and P in a form similar to Eq. (6). See Figure 1 for a visual representation of these matrices.

- **Outcomes**: Let $Y^{(0),\text{obs}} = Y \otimes (\mathbf{1} - A) \in \{\mathbb{R} \cup \{\,?\}\}^{N \times M}$ be a matrix with (i, j)-th entry equal to $y_{i,j}$ if $a_{i,j} = 0$, and equal to ? otherwise. Here, $\mathbf{1}$ is the $N \times M$ matrix with all entries equal to one. Analogously, let $Y^{(1),\text{obs}} = Y \otimes A \in \{\mathbb{R} \cup \{\,?\}\}^{N \times M}$ be a matrix with (i, j)-th entry equal to $y_{i,j}$ if $a_{i,j} = 1$, and equal to ? otherwise. In other words, $Y^{(0),\text{obs}}$ and $Y^{(1),\text{obs}}$ capture the observed components of $\{y^{(0)}_{i,j}\}_{i\in[N],j\in[M]}$ and $\{y^{(1)}_{i,j}\}_{i\in[N],j\in[M]}$, respectively, with missing entries denoted by ?. Then, we can write

$$Y^{(0),\text{obs}} = (\Theta^{(0)} + E^{(0)}) \otimes (\mathbf{1} - A) \quad \text{and} \quad Y^{(1),\text{obs}} = (\Theta^{(1)} + E^{(1)}) \otimes A. \tag{7}$$

- **Treatments**: From Eq. (3), we can write

$$A = (P + W).$$

Building on the earlier discussion, the application of matrix completion yields estimates:

$$\widehat{\Theta}^{(0)} = \texttt{MC}(Y^{(0),\text{obs}}), \quad \widehat{\Theta}^{(1)} = \texttt{MC}(Y^{(1),\text{obs}}), \quad \text{and} \quad \widehat{P} = \texttt{MC}(A), \tag{8}$$

where the algorithm MC may vary for $\widehat{\Theta}^{(0)}$, $\widehat{\Theta}^{(1)}$, and $\widehat{P}$. Because all entries of A are observed, $\texttt{MC}(A)$ denoises A but does not need to impute missing entries. From Eq. (7) and Eq. (8), it follows that $\widehat{\Theta}^{(0)}$ and $\widehat{\Theta}^{(1)}$ depend on A and Y, whereas $\widehat{P}$ depends only on A.

In this section, we deliberately leave the matrix completion algorithm MC as a "black-box". In Section 4, we establish finite-sample and asymptotic guarantees for our proposed estimator, contingent on specific properties for MC. In Section 5, we propose a novel end-to-end matrix completion algorithm that satisfies these properties.

Given matrix completion estimates of $(\widehat{\Theta}^{(0)}, \widehat{\Theta}^{(1)}, \widehat{P})$, we formulate two preliminary estimators for $\text{ATE}_{\cdot,j}$: (i) an outcome imputation estimator, which uses $\widehat{\Theta}^{(0)}$ and $\widehat{\Theta}^{(1)}$ only, and (ii) an inverse probability weighting estimator, which uses $\widehat{P}$ only. Then, we combine these to obtain a doubly-robust estimator of $\text{ATE}_{\cdot,j}$.

Outcome imputation (OI) estimator. Let $\widehat{\theta}_{i,j}^{(a)}$ denote the (i,j)-th entry of $\widehat{\Theta}^{(a)}$ for $i \in [N], j \in [M]$, and $a \in \{0,1\}$. The OI estimator for $\text{ATE}_{\cdot,j}$ is defined as follows:

$$\widehat{\text{ATE}}_{\cdot,j}^{\text{OI}} \triangleq \widehat{\mu}_{\cdot,j}^{(1,\text{OI})} - \widehat{\mu}_{\cdot,j}^{(0,\text{OI})}, \tag{9}$$

where

$$\widehat{\mu}_{\cdot,j}^{(a,\text{OI})} \triangleq \frac{1}{N} \sum_{i \in [N]} \widehat{\theta}_{i,j}^{(a)} \quad \text{for} \quad a \in \{0,1\}.$$

That is, the OI estimator is obtained by taking the difference of the average value of the j-th column of the estimates $\widehat{\Theta}^{(0)}$ and $\widehat{\Theta}^{(1)}$. The quality of OI depends on how well $\widehat{\Theta}^{(0)}$ and $\widehat{\Theta}^{(1)}$ approximate the mean potential outcome matrices $\Theta^{(0)}$ and $\Theta^{(1)}$, respectively.

Inverse probability weighting (IPW) estimator. Let $\widehat{p}_{i,j}$ denote the (i,j)-th entry of

$\widehat{P}$ for $i \in [N]$ and $j \in [M]$. The IPW estimate for $\text{ATE}_{\cdot,j}$ is defined as follows:

$$\widehat{\text{ATE}}_{\cdot,j}^{\text{IPW}} \triangleq \widehat{\mu}_{\cdot,j}^{(1,\text{IPW})} - \widehat{\mu}_{\cdot,j}^{(0,\text{IPW})}, \tag{10}$$

where

$$\widehat{\mu}_{\cdot,j}^{(0,\text{IPW})} \triangleq \frac{1}{N} \sum_{i \in [N]} \frac{y_{i,j}\left(1 - a_{i,j}\right)}{1 - \widehat{p}_{i,j}} \quad \text{and} \quad \widehat{\mu}_{\cdot,j}^{(1,\text{IPW})} \triangleq \frac{1}{N} \sum_{i \in [N]} \frac{y_{i,j} a_{i,j}}{\widehat{p}_{i,j}}.$$

That is, the IPW estimator is obtained by taking the difference of the average value of the j-th column of the matrices $Y^{(0),\text{obs}}$ and $Y^{(1),\text{obs}}$, replacing unobserved entries with zeros, and weighting each outcome by the inverse of the estimated assignment probability to account for confounding. The quality of the IPW estimate depends on how well $\widehat{P}$ approximates the probability matrix P.

The matrix completion-based OI and IPW estimators in Eq. (9) and Eq. (10) have the same form as the classical OI and IPW estimators, which are derived for settings where all confounders are observed (e.g., Imbens and Rubin, 2015). In contrast to the classical setting, our framework is one with unmeasured confounding.

3.3 Doubly-robust (DR) estimator

The DR estimator of $\text{ATE}_{\cdot,j}$ combines the estimates $\widehat{\Theta}^{(0)}, \widehat{\Theta}^{(1)}$, and $\widehat{P}$ from Eq. (8). It is defined as follows:

$$\widehat{\text{ATE}}_{\cdot,j}^{\text{DR}} \triangleq \widehat{\mu}_{\cdot,j}^{(1,\text{DR})} - \widehat{\mu}_{\cdot,j}^{(0,\text{DR})}, \tag{11}$$

where

$$\widehat{\mu}_{\cdot,j}^{(0,\text{DR})} \triangleq \frac{1}{N} \sum_{i \in [N]} \widehat{\theta}_{i,j}^{(0,\text{DR})} \quad \text{with} \quad \widehat{\theta}_{i,j}^{(0,\text{DR})} \triangleq \widehat{\theta}_{i,j}^{(0)} + \left(y_{i,j} - \widehat{\theta}_{i,j}^{(0)}\right) \frac{1 - a_{i,j}}{1 - \widehat{p}_{i,j}},$$

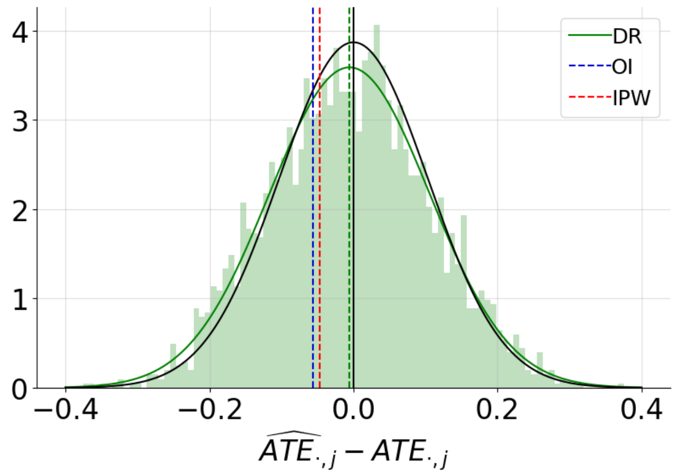


Figure 2: Simulation evidence of the convergence of the error of the doubly-robust (DR) estimator to a mean-zero Gaussian distribution. The histogram represents $\widehat{\text{ATE}}^{\text{DR}}_{\cdot,j} - \text{ATE}_{\cdot,j}$, the green curve represents the (best) fitted Gaussian distribution, and the black curve represents the Gaussian approximation from Theorem 2 in Section 4. Histogram counts are normalized so that the area under the histogram integrates to one. Unlike DR, the outcome imputation (OI) and inverse probability weighting (IPW) estimators have non-trivial biases, as evidenced by the means of the distributions in dashed green, blue, and red, respectively. Section S3 in the supplementary appendix reports complete simulation results.

and

$$\widehat{\mu}^{(1,\text{DR})}_{\cdot,j} \triangleq \frac{1}{N} \sum_{i\in[N]} \widehat{\theta}^{(1,\text{DR})}_{i,j} \quad \text{with} \quad \widehat{\theta}^{(1,\text{DR})}_{i,j} \triangleq \widehat{\theta}^{(1)}_{i,j} + \Big(y_{i,j} - \widehat{\theta}^{(1)}_{i,j}\Big)\frac{a_{i,j}}{\widehat{p}_{i,j}}. \tag{12}$$

In Section 4, we prove that $\widehat{\text{ATE}}^{\text{DR}}_{\cdot,j}$ consistently estimates $\text{ATE}_{\cdot,j}$ as long as either $(\widehat{\Theta}^{(0)}, \widehat{\Theta}^{(1)})$ is consistent for $(\Theta^{(0)}, \Theta^{(1)})$ or $\widehat{P}$ is consistent for P, i.e., it is doubly-robust. Furthermore, we show that the DR estimator provides superior finite-sample guarantees to the OI and IPW estimators, and that it satisfies a central limit theorem at a parametric rate under weak conditions on the convergence rate of the matrix completion routine. Using simulated data, Figure 2 demonstrates the improved performance of DR, relative to OI and IPW. Despite substantial biases observed in both OI and IPW estimates, the error of the DR estimate closely follows a mean-zero Gaussian distribution. We provide a detailed description of the simulation setup in supplementary appendix Section S3.

4 Main Results

This section presents the formal results of the article. Section 4.1 details assumptions, Section 4.2 discusses finite-sample guarantees, and Section 4.3 presents a central limit theorem for $\widehat{\text{ATE}}^{\text{DR}}_{\cdot,j}$.

4.1 Assumptions

Requirements on data generating process. We make two assumptions on how the data is generated. First, we impose a positivity condition on the assignment probabilities.

Assumption 1 (Positivity on true assignment probabilities)**.** *The unknown assignment probability matrix P is such that*

$$\lambda \leq p_{i,j} \leq 1 - \lambda, \tag{13}$$

for all $i \in [N]$ and $j \in [M]$, where $0 < \lambda \leq 1/2$.

Assumption 1 requires that the propensity score for each unit-outcome pair is bounded away from 0 and 1, implying that any unit-item pair can be assigned either of the two treatments. An analogous assumption is pervasive in causal inference models with no-unmeasured confounding. For simplicity of exposition and to avoid notational clutter, Assumption 1 requires Eq. (13) for all outcomes, $j \in [M]$. In practical applications, however, $\text{ATE}_{\cdot,j}$ may be estimated for a select group of those outcomes. In that case, the positivity assumption applies only for the selected subset of outcomes for which $\text{ATE}_{\cdot,j}$ is estimated.

Next, we formalize the requirements on the noise variables.

Assumption 2 (Zero-mean, independent, and subGaussian noise)**.** *Fix any $j \in [M]$. Then,*

(a) $\{(\varepsilon^{(0)}_{i,j}, \varepsilon^{(1)}_{i,j}, \eta_{i,j}) : i \in [N]\}$ are mean zero and independent (across i);

(b) for every $i \in [N]$ and $j \in [M]$, $(\varepsilon_{i,j}^{(0)}, \varepsilon_{i,j}^{(1)}) \perp\!\!\!\perp \eta_{i,j}$; moreover, the distribution of $(\varepsilon_{i,j}^{(0)}, \varepsilon_{i,j}^{(1)})$ depends on $(\Theta^{(0)}, \Theta^{(1)}, P)$ only through $(\theta_{i,j}^{(0)}, \theta_{i,j}^{(1)})$, and the distribution of $\eta_{i,j}$ depends on $(\Theta^{(0)}, \Theta^{(1)}, P)$ only through $p_{i,j}$; and

(c) $\varepsilon_{i,j}^{(a)}$ has subGaussian norm bounded by a constant $\overline{\sigma}$ for every $i \in [N]$ and $a \in \{0, 1\}$.

Assumption 2(a) defines $(\Theta^{(0)}, \Theta^{(1)}, P)$ as matrices collecting the means of the potential outcomes and treatment assignments in Eqs. (2) and (3). Further, for every measurement, it imposes independence across units in the noise variables. Assumption 2(b) imposes independence between the noise in the potential outcomes and noise in treatment assignment, and implies that for each particular unit i and measurement j, confounding emerges only from the interplay between $(\theta_{i,j}^{(0)}, \theta_{i,j}^{(1)})$ and $p_{i,j}$. Finally, Assumption 2(c) is mild and useful to derive finite-sample guarantees. For the central limit theorem in Section 4.3, subGaussianity could be dispensed with by restricting the moments of $\varepsilon_{i,j}^{(a)}$. Assumption 2 does not restrict the dependence between $\varepsilon_{i,j}^{(0)}$ and $\varepsilon_{i,j}^{(1)}$. Neither Assumption 2 restricts the dependence of $\eta_{i,j}$ across outcomes. In particular, Assumption 2 allows for the existence of pairs of outcomes (j, j') such that $\mathbb{E}[\eta_{i,j}^2] = \mathbb{E}[\eta_{i,j'}^2] = \mathbb{E}[\eta_{i,j}\eta_{i,j'}]$, in which case $a_{i,j} = a_{i,j'}$ with probability one.

Requirements on matrix completion estimators. First, we assume the estimate $\widehat{P}$ is consistent with Assumption 1.

Assumption 3 (Positivity on estimated assignment probabilities)**.** *The estimated probability matrix $\widehat{P}$ is such that*

$$\bar{\lambda} \leq \widehat{p}_{i,j} \leq 1 - \bar{\lambda},$$

for all $i \in [N]$ and $j \in [M]$, where $0 < \bar{\lambda} \leq \lambda$.

Assumption 3 holds when the entries of $\widehat{P}$ are truncated to the range $[\bar{\lambda}, 1-\bar{\lambda}]$, provided $\bar{\lambda}$ is not greater than λ. Second, our theoretical analysis requires independence between certain elements of the estimates $(\widehat{P}, \widehat{\Theta}^{(0)}, \widehat{\Theta}^{(1)})$ from Eq. (8), and the noise matrices $(W, E^{(0)}, E^{(1)})$.

We formally state this independence condition as an assumption below.

Assumption 4 (Independence between estimates and noise)**.** *Fix any $j \in [M]$. There exists a non-empty partition $(\mathcal{R}_0, \mathcal{R}_1)$ of the units $[N]$ such that*

$$\left\{\left(\widehat{p}_{i,j}, \widehat{\theta}^{(a)}_{i,j}\right)\right\}_{i\in\mathcal{R}_s} \perp\!\!\!\perp \left\{\eta_{i,j}\right\}_{i\in\mathcal{R}_s} \tag{14}$$

and

$$\left\{\widehat{p}_{i,j}\right\}_{i\in\mathcal{R}_s} \perp\!\!\!\perp \left\{\left(\eta_{i,j}, \varepsilon^{(a)}_{i,j}\right)\right\}_{i\in\mathcal{R}_s}, \tag{15}$$

for every $a \in \{0,1\}$ and $s \in \{0,1\}$.

Eq. (14) requires that within each of the two partitions of the units, estimated mean potential outcomes and estimated assignment probabilities are jointly independent of the error in assignment probabilities, for every measurement. Similarly, Eq. (15) requires that within each of the two partitions of the units, estimated assignment probabilities are independent jointly of the noise in assignment probabilities and potential outcomes, for every measurement. Conditions like Eq. (14) and Eq. (15) are familiar in the doubly-robust estimation literature. Chernozhukov et al. (2018) employ a cross-fitting device to enforce an assumption similar to Assumption 4 in a context with no unmeasured confounders. Section 5 provides a novel cross-fitting procedure for matrix estimation under which Assumption 4 holds for any `MC` algorithm (under additional assumptions on the noise variables).

Matrix completion error rates. The formal guarantees in this section depend on the normalized $(1,2)$-norms of the errors in estimating the unknown parameters $(\Theta^{(0)}, \Theta^{(1)}, P)$. We use the following notation for these errors:

$$\mathcal{E}\left(\widehat{P}\right) \triangleq \frac{\|\widehat{P}-P\|_{1,2}}{\sqrt{N}} \quad \text{and} \quad \mathcal{E}\left(\widehat{\Theta}\right) \triangleq \sum_{a\in\{0,1\}} \mathcal{E}\left(\widehat{\Theta}^{(a)}\right), \quad \text{with} \quad \mathcal{E}\left(\widehat{\Theta}^{(a)}\right) \triangleq \frac{\|\widehat{\Theta}^{(a)}-\Theta^{(a)}\|_{1,2}}{\sqrt{N}}. \tag{16}$$

A variety of matrix completion algorithms deliver $\mathcal{E}\left(\widehat{P}\right) = O_p(\min\{N,M\}^{-\alpha})$ and $\mathcal{E}\left(\widehat{\Theta}\right) = O_p(\min\{N,M\}^{-\beta})$, where $0 < \alpha, \beta \leq 1/2$. For simplicity, the conditions in this section track dependence on N only. We say that the normalized errors $\mathcal{E}\left(\widehat{P}\right)$ and $\mathcal{E}\left(\widehat{\Theta}\right)$ achieve the parametric rate when they have the same rate as $O_p(N^{-1/2})$. Section 5 explicitly characterizes how the rates of convergence $\mathcal{E}\left(\widehat{P}\right)$ and $\mathcal{E}\left(\widehat{\Theta}\right)$ depend on N and M for a particular matrix completion algorithm based on Bai and Ng (2021).

4.2 Non-asymptotic guarantees

The first main result of this section provides a non-asymptotic error bound for $\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,j} - \mathrm{ATE}_{\cdot,j}$ in terms of the errors $\mathcal{E}\left(\widehat{P}\right)$ and $\mathcal{E}\left(\widehat{\Theta}\right)$ defined in Eq. (16).

Theorem 1 (**Finite Sample Guarantees for DR**)**.** *Suppose Assumptions 1 to 4 hold. Fix $\delta \in (0,1)$ and $j \in [M]$. Then, with probability at least $1-\delta$, we have*

$$\left|\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,j} - \mathrm{ATE}_{\cdot,j}\right| \leq \mathrm{Err}^{\mathrm{DR}}_{N,\delta}, \tag{17}$$

where

$$\mathrm{Err}^{\mathrm{DR}}_{N,\delta} \triangleq \frac{2}{\bar{\lambda}}\left[\mathcal{E}\left(\widehat{\Theta}\right)\mathcal{E}\left(\widehat{P}\right) + \left(\frac{\sqrt{c\ell_{\delta/12}}}{\sqrt{\ell_1}}\mathcal{E}\left(\widehat{\Theta}\right) + 2\overline{\sigma}\sqrt{c\ell_{\delta/12}} + \frac{2\overline{\sigma}m(c\ell_{\delta/12})}{\sqrt{\ell_1}}\right)\frac{1}{\sqrt{N}}\right], \tag{18}$$

for $m(c)$ and ℓ_c as defined in Section 1.

The proof of Theorem 1 is given in supplementary appendix S1. Eqs. (17) and (18) bound the absolute error of the DR estimator by the rate of $\mathcal{E}\left(\widehat{\Theta}\right)(\mathcal{E}\left(\widehat{P}\right) + N^{-0.5}) + N^{-0.5}$. When $\mathcal{E}\left(\widehat{P}\right)$ is lower bounded at the parametric rate of $N^{-0.5}$, $\mathrm{Err}^{\mathrm{DR}}_{N,\delta}$ has the same rate as $\mathcal{E}\left(\widehat{P}\right)\mathcal{E}\left(\widehat{\Theta}\right) + N^{-0.5}$.

Doubly-robust behavior of $\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,j}$. The error rate of $\mathcal{E}\left(\widehat{P}\right)\mathcal{E}\left(\widehat{\Theta}\right) + N^{-0.5}$ immediately reveals that the DR estimate is doubly-robust with respect to the error in estimating the mean potential outcomes $(\Theta^{(0)}, \Theta^{(1)})$ and the assignment probabilities P. First, the error $\mathrm{Err}^{\mathrm{DR}}_{N,\delta}$ decays at a parametric rate of $O_p(N^{-0.5})$ as long as the product of error rates, $\mathcal{E}\left(\widehat{P}\right)\mathcal{E}\left(\widehat{\Theta}\right)$, decays as $O_p(N^{-0.5})$. As a result, $\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,j}$ can exhibit a parametric error rate even when neither the mean potential outcomes nor the assignment probabilities are estimated at a parametric rate. Second, $\mathrm{Err}_{N,\delta}^{\mathrm{DR}}$ decays to zero as long as either of $\mathcal{E}\left(\widehat{P}\right)$ or $\mathcal{E}\left(\widehat{\Theta}\right)$ decays to zero, provided both errors are $O_p(1)$.

We next compare the performance of DR estimator with the OI and IPW estimators from Eqs. (9) and (10), respectively. Towards this goal, we characterize the $\mathrm{ATE}_{\cdot,j}$ estimation error of $\widehat{\mathrm{ATE}}_{\cdot,j}^{\mathrm{OI}}$ in terms of $\mathcal{E}\left(\widehat{\Theta}\right)$ and of $\widehat{\mathrm{ATE}}_{\cdot,j}^{\mathrm{IPW}}$ in terms of $\mathcal{E}\left(\widehat{P}\right)$.

Proposition 1 (**Finite Sample Guarantees for OI and IPW**). *Fix any $j \in [M]$. For OI, we have*

$$\left|\widehat{\mathrm{ATE}}_{\cdot,j}^{\mathrm{OI}} - \mathrm{ATE}_{\cdot,j}\right| \leq \mathrm{Err}_N^{\mathrm{OI}} \triangleq \mathcal{E}\left(\widehat{\Theta}\right). \tag{19}$$

For IPW, suppose Assumptions 1 to 4 hold. Define $\theta_{\max} \triangleq \sum_{a\in\{0,1\}} \|\Theta^{(a)}\|_{\max}$, and fix any $\delta \in (0,1)$. Then, with probability at least $1-\delta$, we have

$$\left|\widehat{\mathrm{ATE}}_{\cdot,j}^{\mathrm{IPW}} - \mathrm{ATE}_{\cdot,j}\right| \leq \mathrm{Err}_{N,\delta}^{\mathrm{IPW}}, \tag{20}$$

where

$$\mathrm{Err}_{N,\delta}^{\mathrm{IPW}} \triangleq \frac{2}{\bar{\lambda}}\left[\theta_{\max}\,\mathcal{E}\left(\widehat{P}\right) + \left(\frac{\sqrt{c\ell_{\delta/12}}}{\sqrt{\ell_1}}\theta_{\max} + 2\bar{\sigma}\sqrt{c\ell_{\delta/12}} + \frac{2\bar{\sigma}m(c\ell_{\delta/12})}{\sqrt{\ell_1}}\right)\frac{1}{\sqrt{N}}\right],$$

for $m(c)$ and ℓ_c as defined in Section 1.

The proofs of Eq. (19) and Eq. (20) are given in the supplementary appendix (Sections S6 and S7). Proposition 1 implies that in an asymptotic sequence with bounded $\theta_{\max}$, OI and IPW attain the parametric rate $O_p(N^{-0.5})$ provided $\mathcal{E}\left(\widehat{\Theta}\right)$ and $\mathcal{E}\left(\widehat{P}\right)$ are $O_p(N^{-0.5})$, respectively. The next corollary, proven in the supplementary appendix (Section S5), compares these error rates with those obtained for the DR estimator in Theorem 1.

Corollary 1 (**Gains of DR over OI and IPW**). *Suppose Assumptions 1 to 4 hold. Fix any $j \in [M]$. Consider an asymptotic sequence such that $\theta_{\max}$ is bounded. If $\mathcal{E}\left(\widehat{P}\right) = O_p(N^{-\alpha})$ and $\mathcal{E}\left(\widehat{\Theta}\right) = O_p(N^{-\beta})$ for $0 \leq \alpha \leq 0.5$ and $0 \leq \beta \leq 0.5$, then*

$$\left|\widehat{\mathrm{ATE}}_{\cdot,j}^{\mathrm{OI}} - \mathrm{ATE}_{\cdot,j}\right| = O_p(N^{-\beta}), \qquad \left|\widehat{\mathrm{ATE}}_{\cdot,j}^{\mathrm{IPW}} - \mathrm{ATE}_{\cdot,j}\right| = O_p(N^{-\alpha}),$$

and

$$\left|\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,j} - \mathrm{ATE}_{\cdot,j}\right| = O_p(N^{-\min\{\alpha+\beta,0.5\}}).$$

Corollary 1 shows that the DR estimate's error decay rate is consistently superior to that of the OI and IPW estimates across a variety of regimes for α, β. Specifically, the error $\mathrm{Err}^{\mathrm{DR}}_{N,\delta}$ scales strictly faster than both $\mathrm{Err}^{\mathrm{OI}}_{N}$ and $\mathrm{Err}^{\mathrm{IPW}}_{N,\delta}$ if the estimation errors of $\widehat{\Theta}^{(0)}$, $\widehat{\Theta}^{(1)}$, and $\widehat{P}$ converge slower than at the parametric rate $O_p(N^{-1/2})$. When the estimation errors of $\widehat{\Theta}^{(0)}$, $\widehat{\Theta}^{(1)}$, and $\widehat{P}$ all decay at a parametric rate, OI, IPW, and DR estimation errors decay also at a parametric rate.

4.3 Asymptotic guarantees

The next result, proven in the supplementary appendix (Section S5) as a corollary of Theorem 1, provides conditions on $\mathcal{E}\left(\widehat{P}\right)$ and $\mathcal{E}\left(\widehat{\Theta}\right)$ for consistency of $\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,j}$.

Corollary 2 (**Consistency for DR**). *Suppose Assumptions 1 to 4 hold. As $N \to \infty$, if either (i) $\mathcal{E}\left(\widehat{P}\right) = o_p(1)$, $\mathcal{E}\left(\widehat{\Theta}\right) = O_p(1)$, or (ii) $\mathcal{E}\left(\widehat{\Theta}\right) = o_p(1)$, $\mathcal{E}\left(\widehat{P}\right) = O_p(1)$, it holds that*

$$\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,j} - \mathrm{ATE}_{\cdot,j} \xrightarrow{p} 0, \tag{21}$$

for all $j \in [M]$.

Corollary 2 states that $\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,j}$ is a consistent estimator for $\mathrm{ATE}_{\cdot,j}$ as long as either the mean potential outcomes or the assignment probabilities are estimated consistently.

The next theorem, proven in supplementary appendix S2, establishes a Gaussian approximation for $\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,j}$ under mild conditions on error rates $\mathcal{E}\left(\widehat{P}\right)$ and $\mathcal{E}\left(\widehat{\Theta}\right)$.

Theorem 2 (**Asymptotic Normality for DR**). *Suppose Assumptions 1 to 4 and the following conditions hold,*

(C1) $\mathcal{E}\left(\widehat{P}\right) = o_p(1)$ and $\mathcal{E}\left(\widehat{\Theta}\right) = o_p(1)$.

(C2) $\mathcal{E}\Big(\widehat{P}\Big)\mathcal{E}\Big(\widehat{\Theta}\Big) = o_p\Big(N^{-1/2}\Big)$.

(C3) For every $i \in [N]$ *and* $j \in [M]$, *let* $\sigma_{i,j}^{(0)}$ *and* $\sigma_{i,j}^{(1)}$ *be the standard deviations of* $\varepsilon_{i,j}^{(0)}$ *and* $\varepsilon_{i,j}^{(1)}$, *respectively. The sequence*

$$\overline{\sigma}_j^2 \triangleq \frac{1}{N}\sum_{i\in[N]}\frac{(\sigma_{i,j}^{(1)})^2}{p_{i,j}} + \frac{1}{N}\sum_{i\in[N]}\frac{(\sigma_{i,j}^{(0)})^2}{1-p_{i,j}}, \tag{22}$$

is bounded away from zero as N *increases.*

Then, for all $j \in [M]$,

$$\sqrt{N}\Big(\widehat{\text{ATE}}_{\cdot,j}^{\text{DR}} - \text{ATE}_{\cdot,j}\Big)/\overline{\sigma}_j \xrightarrow{d} \mathcal{N}\Big(0,1\Big), \tag{23}$$

as $N \to \infty$.

Theorem 2 describes two simple requirements on the estimated matrices $\widehat{P}$ and $(\widehat{\Theta}^{(0)}, \widehat{\Theta}^{(1)})$, under which $\widehat{\text{ATE}}_{\cdot,j}^{\text{DR}}$ exhibits an asymptotic Gaussian distribution centered at $\text{ATE}_{\cdot,j}$. Condition (C1) requires that the estimation errors of $\widehat{P}$ and $(\widehat{\Theta}^{(0)}, \widehat{\Theta}^{(1)})$ converge to zero in probability. Condition (C2) requires that the product of errors decays sufficiently fast, at a rate $o_p\Big(N^{-1/2}\Big)$, ensuring that the bias of the normalized estimator in Eq. (23) converges to zero. Condition (C2) is similar to conditions in the literature on doubly-robust estimation of average treatment effects under observed confounding (e.g., Assumption 5.1 in Chernozhukov et al., 2018). Specifically, Chernozhukov et al. (2018) assume that the product of propensity estimation error and outcome regression error decays faster than $N^{-1/2}$.

Black-box asymptotic normality. Theorem 2 applies to any matrix completion algorithm `MC`, provided conditions (C1) and (C2) hold. This level of generality is useful because the product of $\mathcal{E}\Big(\widehat{P}\Big)$ and $\mathcal{E}\Big(\widehat{\Theta}\Big)$ is $o_p\Big(N^{-1/2}\Big)$ for a wide range of `MC` algorithms, under mild assumptions on $(\Theta^{(0)}, \Theta^{(1)}, P)$. In contrast, achieving such black-box asymptotic normality for OI or IPW estimates is challenging. Their biases are tied to the individual error rates,

$\mathcal{E}\left(\widehat{\Theta}\right)$ and $\mathcal{E}\left(\widehat{P}\right)$, which are typically lower-bounded at the parametric rate of $N^{-0.5}$.

The next result, proven in supplementary appendix S2.3, provides a consistent estimator for the asymptotic variance $\overline{\sigma}_j^2$ from Theorem 2.

Proposition 2 (**Consistent variance estimation**). *Suppose Assumptions 1 to 3 and condition (C1) in Theorem 2 holds. Suppose the partition $(\mathcal{R}_0, \mathcal{R}_1)$ of the units $[N]$ from Assumption 4 is such that*

$$\{\left(\widehat{p}_{i,j}, \widehat{\theta}_{i,j}^{(a)}\right)\}_{i\in\mathcal{R}_s} \perp\!\!\!\perp \{(\eta_{i,j}, \varepsilon_{i,j}^{(a)})\}_{i\in\mathcal{R}_s}, \tag{24}$$

for every $j \in [M]$, $a \in \{0,1\}$ and $s \in \{0,1\}$. Then, for all $j \in [M]$, $\widehat{\sigma}_j^2 - \overline{\sigma}_j^2 \xrightarrow{p} 0$, where

$$\widehat{\sigma}_j^2 \triangleq \frac{1}{N}\sum_{i\in[N]} \frac{\left(y_{i,j} - \widehat{\theta}_{i,j}^{(1)}\right)^2 a_{i,j}}{\left(\widehat{p}_{i,j}\right)^2} + \frac{1}{N}\sum_{i\in[N]} \frac{\left(y_{i,j} - \widehat{\theta}_{i,j}^{(0)}\right)^2 (1-a_{i,j})}{\left(1-\widehat{p}_{i,j}\right)^2}. \tag{25}$$

4.4 Application to panel data with lagged treatment effects

Sections 4.2 and 4.3 considered a model where the outcome $y_{i,j}$ for unit $i \in [N]$ and measurement $j \in [M]$ depends on treatment assignment only for unit i and measurement j, i.e., $a_{i,j}$. The supplementary appendix (Section S10) discusses how to extend the results of this section to a setting of panel data with lagged treatment effects. In a panel data setting, the M measurements correspond to T time periods, and t denotes the time index. Then, the supplementary appendix considers an auto-regressive setting, where the potential outcomes at time t depend on the treatment assignment at time t and the realized outcome at time $t-1$, i.e., for all $i \in [N], t \in [T]$, and $a \in \{0,1\}$,

$$y_{i,t}^{(a|y_{i,t-1})} = \alpha^{(a)} y_{i,t-1} + \theta_{i,t}^{(a)} + \varepsilon_{i,t}^{(a)},$$

and observed outcomes satisfy

$$y_{i,t} = y_{i,t}^{(0|y_{i,t-1})}(1-a_{i,t}) + y_{i,t}^{(1|y_{i,t-1})} a_{i,t}.$$

The presence of lagged treatment effects in this model makes it crucial to define causal estimands for entire sequences of treatments. The supplementary appendix describes how the proposed doubly-robust estimation can be extended to treatment sequences and derives a generalization of Theorem 1.

5 Matrix Completion with Cross-Fitting

In this section, we introduce a novel algorithm designed to construct estimates $(\widehat{\Theta}^{(0)}, \widehat{\Theta}^{(1)}, \widehat{P})$ that adhere to Assumption 4 and satisfy conditions (C1) and (C2) in Theorem 2. We first explain why traditional matrix completion algorithms fail to deliver the properties required by Assumption 4. We then present `Cross-Fitted-MC`, a meta-algorithm that takes any matrix completion algorithm and uses it to construct $(\widehat{\Theta}^{(0)}, \widehat{\Theta}^{(1)}, \widehat{P})$ that satisfy Assumption 4, and the stronger independence condition in Proposition 2. Finally, we describe `Cross-Fitted-SVD`, an end-to-end algorithm obtained by combining `Cross-Fitted-MC` with the singular value decomposition (`SVD`)-based algorithm of Bai and Ng (2021), and establish that it also satisfies conditions (C1) and (C2) in Theorem 2.

Traditional matrix completion. Estimates $(\widehat{\Theta}^{(0)}, \widehat{\Theta}^{(1)}, \widehat{P})$ obtained from existing matrix completion algorithms need not satisfy Assumption 4. In particular, using the entire assignment matrix A to estimate each element of P typically results in a violation of $\left\{\widehat{p}_{i,j}\right\}_{i\in\mathcal{R}_s} \perp\!\!\!\perp \left\{\eta_{i,j}\right\}_{i\in\mathcal{R}_s}$ in Assumption 4, as each entry of $\widehat{P}$ is allowed to depend on the entire noise matrix W. For example, in spectral methods (e.g., Nguyen et al., 2019), $\widehat{P}$ is a function of the `SVD` of the entire matrix A, and

$$\widehat{p}_{i,j} \not\perp\!\!\!\perp a_{i',j'}, \tag{26}$$

for all $(i,j),(i',j') \in [N]\times[M]$ in general, which implies $\left\{\widehat{p}_{i,j}\right\}_{i\in\mathcal{R}_s} \not\perp\!\!\!\perp \left\{\eta_{i,j}\right\}_{i\in\mathcal{R}_s}$, for every $\mathcal{R}_s \subset [N]$. Similarly, in matching methods such as nearest neighbors (Li et al., 2019), $\widehat{P}$

is a function of the matches/neighbors estimated from the entire matrix A. Dependence structures such as $\widehat{p}_{i,j} \not\perp\!\!\!\perp a_{i,j}$ for any $i, j \in [N] \times [M]$—which is weaker than Eq. (26)—are enough to violate the $\left\{\widehat{p}_{i,j}\right\}_{i\in\mathcal{R}_s} \perp\!\!\!\perp \left\{\eta_{i,j}\right\}_{i\in\mathcal{R}_s}$ requirement in Assumption 4. Likewise, the requirement $\left\{\widehat{\theta}^{(a)}_{i,j}\right\}_{i\in\mathcal{R}_s} \perp\!\!\!\perp \left\{\eta_{i,j}\right\}_{i\in\mathcal{R}_s}$ in Assumption 4 can be violated, because $\widehat{\Theta}^{(0)}$ and $\widehat{\Theta}^{(1)}$ depend respectively on $Y^{(0),\mathrm{obs}}$ and $Y^{(1),\mathrm{obs}}$, which themselves depend on A.

5.1 `Cross-Fitted-MC`: A meta-cross-fitting algorithm for matrix completion

We now introduce `Cross-Fitted-MC`, a cross-fitting procedure that modifies any `MC` algorithm to produce $(\widehat{\Theta}^{(0)}, \widehat{\Theta}^{(1)}, \widehat{P})$ that satisfy Assumption 4. We employ the following assumption on the noise variables.

Assumption 5 (Block independence between noise)**.** *Let* $(\mathcal{R}_0, \mathcal{R}_1)$ *denote the partition of the units* $[N]$ *from Assumption 4. There exists a partition* $(\mathcal{C}_0, \mathcal{C}_1)$ *of the measurements* $[M]$*, such that for each block* $\mathcal{I} \in \mathcal{P} \triangleq \{\mathcal{R}_s \times \mathcal{C}_k : s, k \in \{0, 1\}\}$,

$$W_{\mathcal{I}} \perp\!\!\!\perp W_{-\mathcal{I}}, E^{(a)}_{-\mathcal{I}} \tag{27}$$

and

$$W_{-\mathcal{I}} \perp\!\!\!\perp W_{\mathcal{I}}, E^{(a)}_{\mathcal{I}}. \tag{28}$$

for every $a \in \{0, 1\}$.

For a given block $\mathcal{I}$, Eq. (27) requires the noise in the treatment assignments corresponding to $\mathcal{I}$ to be independent jointly of the noise in the treatment assignments and the potential outcomes corresponding to the remaining three blocks. Likewise, Eq. (28) requires the noise in the treatment assignments corresponding to the remaining three blocks to be independent jointly of the noise in the treatment assignments and the potential outcomes corresponding to $\mathcal{I}$. Assumption 5 leaves unrestricted the dependence of the noise variables across outcomes that belong to the same block.

For notational simplicity, Assumption 5 imposes independence conditions across blocks of outcomes in a partition of $[M]$ into two blocks only. It is important to note, however, that the results in this section hold under more general dependence patterns. In particular, at the cost of additional notational complexity, it is straightforward to extend the result in this section to partitions of outcomes $(\mathcal{C}_0, \mathcal{C}_1, \ldots, \mathcal{C}_m)$ such that for each $k \in \{0, 1, \ldots, m\}$, $s \in \{0, 1\}$ and $a \in \{0, 1\}$, there exists $k' \in \{0, 1, \ldots, m\} \backslash \{k\}$ with $\{\eta_{i,j}\}_{(i,j) \in \mathcal{R}_s \times \mathcal{C}_k} \perp\!\!\!\perp \{\eta_{i,j}, \varepsilon_{i,j}^{(a)}\}_{(i,j) \in \mathcal{R}_{1-s} \times \mathcal{C}_{k'}}$ and $\{\eta_{i,j}\}_{(i,j) \in \mathcal{R}_{1-s} \times \mathcal{C}_{k'}} \perp\!\!\!\perp \{\eta_{i,j}, \varepsilon_{i,j}^{(a)}\}_{(i,j) \in \mathcal{R}_s \times \mathcal{C}_k}$. This allows for rather general patterns of dependence across outcomes while preserving independence across specific sets of outcomes (e.g., certain product categories in the retail example of Section 1).

Recall the setup from Section 3.1: Given an observation matrix $S \in \{\mathbb{R} \cup \{\,?\}\}^{N \times M}$, a matrix completion algorithm MC produces an estimate $\widehat{T} = \texttt{MC}(S) \in \mathbb{R}^{N \times M}$ of a matrix of interest T, where S and T are related via Eq. (6). With this background, we now describe the `Cross-Fitted-MC` meta-algorithm.

1. The inputs are (i) a matrix completion algorithm MC, (ii) an observation matrix $S \in \{\mathbb{R} \cup \{\,?\}\}^{N \times M}$, and (iii) a block partition $\mathcal{P}$ of the set $[N] \times [M]$ into four blocks as in Assumption 5.

2. For each block $\mathcal{I} \in \mathcal{P}$, construct $\widehat{T}_{\mathcal{I}}$ by applying MC on $S \otimes \mathbf{1}^{-\mathcal{I}}$ where $\mathbf{1}^{-\mathcal{I}} \in \mathbb{R}^{N \times M}$ denotes a masking matrix with (i, j)-th entry equal to 0 if $(i, j) \in \mathcal{I}$ and 1 otherwise, and the operator $\otimes$ is as defined in Section 1. In other words,

$$\widehat{T}_{\mathcal{I}} = \overline{T}_{\mathcal{I}} \quad \text{where} \quad \overline{T} = \texttt{MC}(S \otimes \mathbf{1}^{-\mathcal{I}}). \tag{29}$$

3. Return $\widehat{T} \in \mathbb{R}^{N \times M}$ obtained by collecting together $\{\widehat{T}_{\mathcal{I}}\}_{\mathcal{I} \in \mathcal{P}}$, with each entry in its original position.

We represent this meta-algorithm succinctly as below:

$$\widehat{T} = \texttt{Cross-Fitted-MC}(\texttt{MC}, S, \mathcal{P}).$$

In summary, `Cross-Fitted-MC` produces an estimate $\widehat{T}$ such that for each block $\mathcal{I} \in \mathcal{P}$, the sub-matrix $\widehat{T}_{\mathcal{I}}$ is constructed only using the entries of S corresponding to the remaining three blocks of $\mathcal{P}$. Figure 3(a) provides a schematic of the block partition $\mathcal{P}$ for $\mathcal{R}_0 = [\lfloor N/2 \rfloor]$ and $\mathcal{C}_0 = [\lfloor M/2 \rfloor]$. See Figure 3($b$) for a visualization of $S \otimes \mathbf{1}^{-\mathcal{I}}$. The following result, proven in the supplementary appendix (Section S8.1), establishes $(\widehat{\Theta}^{(0)}, \widehat{\Theta}^{(1)}, \widehat{P})$ generated by `Cross-Fitted-MC` satisfy Assumption 4.

Proposition 3 (**Guarantees for `Cross-Fitted-MC`**)**.** *Suppose Assumptions 2 and 5 hold. Let `MC` be any matrix completion algorithm and $\mathcal{P}$ be the block partition of the set $[N] \times [M]$ into four blocks from Assumption 5. Let*

$$\widehat{\Theta}^{(0)} = \textit{Cross-Fitted-MC}(\textit{MC}, Y^{(0),\text{obs}}, \mathcal{P}), \tag{30}$$

$$\widehat{\Theta}^{(1)} = \textit{Cross-Fitted-MC}(\textit{MC}, Y^{(1),\text{obs}}, \mathcal{P}), \tag{31}$$

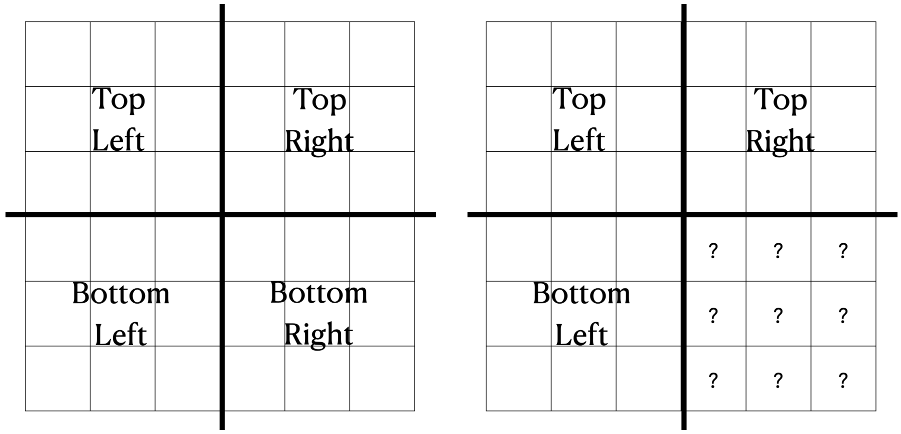
$$\widehat{P} = \textit{Cross-Fitted-MC}(\textit{MC}, A, \mathcal{P}), \tag{32}$$

where $Y^{(0),\text{obs}}$ and $Y^{(1),\text{obs}}$ are defined in Eq. (7). Then, Assumption 4 holds for all $j \in [M]$. Further, suppose

$$W_{\mathcal{I}}, E^{(a)}_{\mathcal{I}} \perp\!\!\!\perp W_{-\mathcal{I}}, E^{(a)}_{-\mathcal{I}}, \tag{33}$$

for every block $\mathcal{I} \in \mathcal{P}$ and $a \in \{0, 1\}$. Then, Eq. (24) holds too.

A host of `MC` algorithms are designed to de-noise and impute missing entries of matrices under random patterns of missingness; the most common missingness pattern studied is where each entry has the same probability of being missing, independent of everything else. In contrast, `Cross-Fitted-MC` generates patterns where all entries in one block are deterministically missing, as in Figure 3(b). A recent strand of research on the interplay



(a) S (b) $S \otimes \mathbf{1}^{-\text{Bottom Right}}$

Figure 3: Panel (a): A matrix S partitioned into four blocks when $\mathcal{R}_0 = [N/2]$ and $\mathcal{C}_0 = [M/2]$ in Assumption 5, i.e., $\mathcal{P} = \{\text{Top Left, Top Right, Bottom Left, Bottom Right}\}$. Panel (b): The matrix $S \otimes \mathbf{1}^{-\text{Bottom Right}}$ obtained from the matrix S by masking the entries corresponding to the Bottom Right block with ?.

between matrix completion methods and causal inference models—specifically, within the synthetic controls framework—has contributed matrix completion algorithms that allow for block missingness (see, e.g., Athey et al., 2021; Agarwal et al., 2021; Bai and Ng, 2021; Agarwal et al., 2023b; Arkhangelsky et al., 2021; Agarwal et al., 2023a; Dwivedi et al., 2022a,b). However, it is a challenge to apply known theoretical guarantees for these methods to the setting in this article because of: (i) the use of cross-fitting—which creates blocks where all observations are missing—and (ii) outside of the completely-missing blocks, there can still be missing observations with heterogeneous probabilities of missingness. In the next section, we show how to modify an `MC` algorithm designed for block missingness patterns so that it can be applied to our setting with cross-fitting and heterogeneous probabilities of missingness outside the folds. For concreteness, we work with the Tall-Wide matrix completion algorithm of Bai and Ng (2021).

5.2 The `Cross-Fitted-SVD` algorithm

`Cross-Fitted-SVD` is an end-to-end `MC` algorithm obtained by instantiating the `Cross-Fitted-MC` meta-algorithm with the Tall-Wide algorithm of Bai and Ng (2021), which we denote as `TW`. For completeness, we detail the `TW` algorithm in Section S8.2.

5.2.1 `Cross-Fitted-SVD` algorithm.

1. The inputs are (i) $A \in \mathbb{R}^{N\times M}$, (ii) $Y^{(a),\text{obs}} \in \{\mathbb{R} \cup \{\,?\}\}^{N\times M}$ for $a \in \{0,1\}$, (iii) a block partition $\mathcal{P}$ of the set $[N] \times [M]$ into four blocks as in Assumption 5, and (iv) hyper-parameters r_1, r_2, r_3, and $\bar{\lambda}$ such that $r_1, r_2, r_3 \in [\min\{N, M\}]$ and $0 < \bar{\lambda} \leq 1/2$.

2. Return $\widehat{P} = \texttt{Proj}_{\bar{\lambda}}\Big(\texttt{Cross-Fitted-MC}(\texttt{TW}_{r_1}, A, \mathcal{P})\Big)$ where $\texttt{Proj}_{\bar{\lambda}}(\cdot)$ projects each entry of its input to the interval $[\bar{\lambda}, 1-\bar{\lambda}]$.

3. Define $Y^{(0),\text{full}}$ as equal to $Y^{(0),\text{obs}}$, but with all missing entries in $Y^{(0),\text{obs}}$ set to zero. Define $Y^{(1),\text{full}}$ analogously with respect to $Y^{(1),\text{obs}}$.

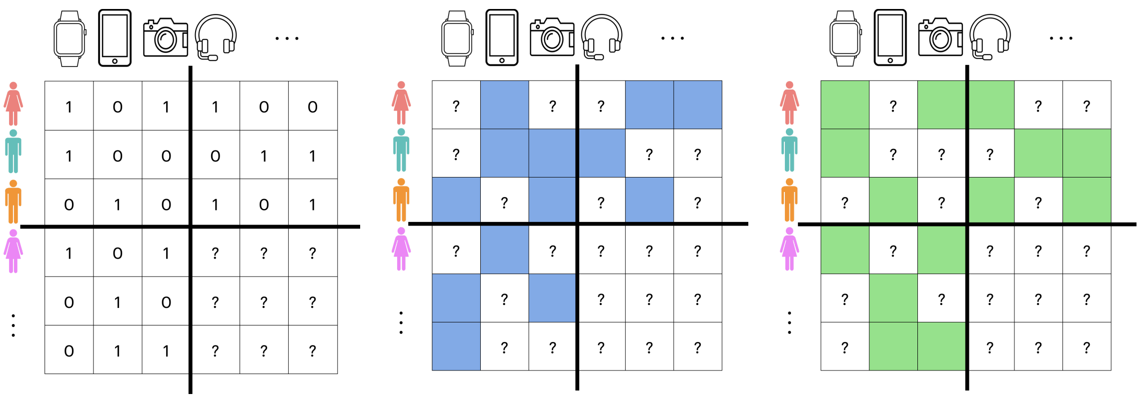
4. Return $\widehat{\Theta}^{(0)} = \texttt{Cross-Fitted-MC}(\texttt{TW}_{r_2}, Y^{(0),\text{full}}, \mathcal{P}) \oslash (\mathbf{1} - \widehat{P})$.

5. Return $\widehat{\Theta}^{(1)} = \texttt{Cross-Fitted-MC}(\texttt{TW}_{r_3}, Y^{(1),\text{full}}, \mathcal{P}) \oslash \widehat{P}$.

We provide intuition on the key steps of the `Cross-Fitted-SVD` algorithm next.

Computing $\widehat{P}$. The estimate $\widehat{P}$ comes from applying `Cross-Fitted-MC` with `TW` on A and truncating the entries of the resulting matrix to the range $[\bar{\lambda}, 1-\bar{\lambda}]$, in accordance with Assumption 3. The `TW` sub-routine is directly applicable to A, because for any block $\mathcal{I} = \mathcal{R}_s \times \mathcal{C}_k \in \mathcal{P}$ the masked matrix $A \otimes \mathbf{1}^{-\mathcal{I}}$ has $[N] \setminus \mathcal{R}_s$ fully observed rows and $[M] \setminus \mathcal{C}_k$ fully observed columns. See Figure 4(a) for a visualization of $A \otimes \mathbf{1}^{-\mathcal{I}}$.

Computing $\widehat{\Theta}^{(0)}$ and $\widehat{\Theta}^{(1)}$. The estimates $\widehat{\Theta}^{(0)}$ and $\widehat{\Theta}^{(1)}$ are constructed by applying `Cross-Fitted-MC` with `TW` on $Y^{(0),\text{full}}$ and $Y^{(1),\text{full}}$, which do not have missing entries. `TW` is not directly applicable on $Y^{(0),\text{obs}}$ and $Y^{(1),\text{obs}}$, as both matrices may not have any rows



(a) $A \otimes \mathbf{1}^{-\text{Bottom Right}}$ (b) $Y^{(0),\text{obs}} \otimes \mathbf{1}^{-\text{Bottom Right}}$ (c) $Y^{(1),\text{obs}} \otimes \mathbf{1}^{-\text{Bottom Right}}$

Figure 4: Panels (a), (b), and (c) illustrate the matrices $A \otimes \mathbf{1}^{-\mathcal{I}}$, $Y^{(0),\text{obs}} \otimes \mathbf{1}^{-\mathcal{I}}$, and $Y^{(1),\text{obs}} \otimes \mathbf{1}^{-\mathcal{I}}$ obtained from A, $Y^{(0),\text{obs}}$ and $Y^{(1),\text{obs}}$, respectively, for the block partition $\mathcal{P}$ in Figure 3(a) and the block $\mathcal{I}$ = Bottom Right. Unlike Panels (b) and (c), Panel (a) contains fully observed rows and columns. To enable the application of `TW` for Panels (b) and (c), we replace missing entries in blocks Top Left, Top Right, and Bottom Left with zeros.

and columns that are fully observed. See Figure 4(b) and Figure 4(c) for visualizations of $Y^{(0),\text{obs}} \otimes \mathbf{1}^{-\mathcal{I}}$ and $Y^{(1),\text{obs}} \otimes \mathbf{1}^{-\mathcal{I}}$, respectively. However, notice that, due to Assumption 2(a) and Assumption 2(b),

$$\mathbb{E}[Y^{(0),\text{full}}] = \mathbb{E}[Y \odot (\mathbf{1} - A)] = \Theta^{(0)} \odot (\mathbf{1} - P),$$

and

$$\mathbb{E}[Y^{(1),\text{full}}] = \mathbb{E}[Y \odot A] = \Theta^{(1)} \odot P.$$

As a result, $\texttt{MC}(Y^{(0),\text{full}})$ and $\texttt{MC}(Y^{(1),\text{full}})$ provide estimates of $\Theta^{(0)} \odot (\mathbf{1} - P)$ and $\Theta^{(1)} \odot P$, respectively—recall the discussion in Section 3.1. To construct $\widehat{\Theta}^{(0)}$ and $\widehat{\Theta}^{(1)}$, we divide the entries of $\texttt{MC}(Y^{(0),\text{full}})$ and $\texttt{MC}(Y^{(1),\text{full}})$ by the entries of $(\mathbf{1} - \widehat{P})$ and $\widehat{P}$, respectively, to adjust for heterogeneous probabilities of missingness (see, e.g., Jin et al., 2021; Bhattacharya and Chatterjee, 2022; Xiong and Pelger, 2023, for related procedures). This inverse probability of treatment weighting adjustment to estimate $\widehat{\Theta}^{(0)}$ and $\widehat{\Theta}^{(1)}$ is distinct and in addition to the augmented IPW procedure that generates $\widehat{\text{ATE}}^{\text{DR}}_{\cdot,j}$ from estimates $\widehat{\Theta}^{(0)}$, $\widehat{\Theta}^{(1)}$ and $\widehat{P}$.

5.2.2 Theoretical guarantees for `Cross-Fitted-SVD`

The following result, proven in supplementary appendix (Section S9.1), provides theoretical guarantees for `Cross-Fitted-SVD`.

Proposition 4 (Guarantees for `Cross-Fitted-SVD`). *Suppose Assumptions 1 and 2 and Assumptions S1 to S4 in supplementary appendix (Section S9) hold. Consider an asymptotic sequence such that $\theta_{\max}$ is bounded as both N and M increase. Let $\widehat{P}$, $\widehat{\Theta}^{(0)}$, and $\widehat{\Theta}^{(1)}$ be the estimates returned by `Cross-Fitted-SVD` with the block partition $\mathcal{P}$ from Assumption 5, $r_1 = r_p$, $r_2 = r_{\theta_0}(r_p + 1)$, $r_3 = r_{\theta_1} r_p$, and any $\bar{\lambda}$ such that $0 < \bar{\lambda} \leq \lambda$ with λ denoting the constant from Assumption 1. Then, as $N, M \to \infty$,*

$$\mathcal{E}\left(\widehat{P}\right) = O_p\left(\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{M}}\right) \quad \textit{and} \quad \mathcal{E}\left(\widehat{\Theta}\right) = O_p\left(\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{M}}\right).$$

Proposition 4 implies that the conditions (C1) and (C2) in Theorem 2 hold whenever $N^{1/2}/M = o(1)$. Then, the DR estimator from Eq. (11) constructed using `Cross-Fitted-SVD` estimates $\widehat{\Theta}^{(0)}$, $\widehat{\Theta}^{(1)}$, and $\widehat{P}$ exhibits an asymptotic Gaussian distribution centered at the target causal estimand. Further, Proposition 4 implies that the estimation errors $\mathcal{E}\left(\widehat{P}\right)$ and $\mathcal{E}\left(\widehat{\Theta}\right)$ achieve the parametric rate whenever $N/M = O(1)$.

5.3 Application to panel data with staggered adoption

Section 5.1 considered a setting with block independence between noise (formalized in Assumption 5). The supplementary appendix (Section S11) discusses how to extend the proposed doubly-robust framework to a setting of panel data with staggered adoption, where this assumption may not hold. Recall (from Section 4.4) that in the panel data setting M measurements correspond to T time periods, and t denotes the time index. Then, the supplementary appendix considers a setting where a unit remains under control for some period of time, after which it deterministically remains under treatment. In other words, for every unit $i \in [N]$, there exists a time point $t_i \in [T]$ such that $a_{i,t} = 0$ for $t \leq t_i$, and $a_{i,t} = 1$ for $t > t_i$. Such a treatment assignment pattern leads to a heavy dependence in the noise $\{\eta_{i,t}\}_{t\in[T]}$ for every unit $i \in [N]$. The supplementary appendix describes an alternative approach to the `Cross-Fitted-SVD` algorithm and shows that Assumption 4 still holds for a suitable staggered adoption model.

6 Conclusion

This article contributes to a rapidly growing literature on causal factor models and treatment effect estimation in non-linear panel data models, where the availability of multiple outcomes offers a way to overcome the challenge of non-additive unobserved confounding. We show it is possible to control for the confounding effects of a set of latent variables when this set is low-dimensional relative to the number of observed treatments and outcomes. Future research could extend our methods to alternative treatment assignment mechanisms and develop primitive conditions for matrix completion approaches beyond the one used in Section 5.2. Especially promising directions include methods that accommodate weak factor structures (Armstrong et al., 2025) and those that require only a fixed number of outcomes (Lei and Ross, 2024).

References

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, 113(2):231–263.

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

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(490):493–505.

Abadie, A. and Gardeazabal, J. (2003). The economic costs of conflict: A case study of the basque country. *American Economic Review*, 93(1):113–132.

Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267.

Agarwal, A., Dahleh, M., Shah, D., and Shen, D. (2023a). Causal matrix completion. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 3821–3826. PMLR.

Agarwal, A., Shah, D., and Shen, D. (2023b). Synthetic interventions.

Agarwal, A., Shah, D., Shen, D., and Song, D. (2021). On robustness of principal component regression. *Journal of the American Statistical Association*, pages 1–34.

Agarwal, A. and Singh, R. (2024). Causal inference with corrupted data: Measurement error, missing values, discretization, and differential privacy. *arXiv preprint: 2107.02780*.

Angrist, J. D. (1998). Estimating the labor market impact of voluntary military service using social security data on military applicants. *Econometrica*, 66(2):249–288.

Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., and Wager, S. (2021). Synthetic difference-in-differences. *American Economic Review*, 111(12):4088–4118.

Arkhangelsky, D. and Imbens, G. W. (2022). Doubly robust identification for causal panel data models. *The Econometrics Journal*, 25(3):649–674.

Armstrong, T. B., Weidner, M., and Zeleneev, A. (2025). Robust estimation and inference in panels with interactive fixed effects. *arxiv preprint: 2210.06639v4*.

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

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

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

Bai, J. (2024). Likelihood approach to dynamic panel models with interactive effects. *Journal of Econometrics*, 240(1):105636.

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

Bai, J. and Ng, S. (2021). Matrix completion, counterfactuals, and factor analysis of missing data. *Journal of the American Statistical Association*, 116(536):1746–1763.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–972.

Bhattacharya, S. and Chatterjee, S. (2022). Matrix completion with data-dependent missingness probabilities. *IEEE Transactions on Information Theory*, 68(10):6762–6773.

Billingsley, P. (2017). *Probability and measure*. John Wiley & Sons.

Cattaneo, M. D., Feng, Y., and Titiunik, R. (2021). Prediction intervals for synthetic control methods. *Journal of the American Statistical Association*, 116(536):1865–1880.

Chatterjee, S. (2015). Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177 – 214.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Choi, J., Kwon, H., and Liao, Y. (2024). Inference for low-rank completion without sample splitting with application to treatment effect estimation. *Journal of Econometrics*, 240(1):105682.

Choi, J. and Yuan, M. (2024). Matrix completion when missing is not at random and its applications in causal panel data models. *Journal of the American Statistical Association*, pages 1–15.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, pages 295–313.

Dwivedi, R., Tian, K., Tomkins, S., Klasnja, P., Murphy, S., and Shah, D. (2022a). Counterfactual inference for sequential experiments. *arXiv preprint: 2202.06891*.

Dwivedi, R., Tian, K., Tomkins, S., Klasnja, P., Murphy, S., and Shah, D. (2022b). Doubly robust nearest neighbors in factor models. *arXiv preprint: 2211.14297*.

Feng, Y. (2021). Causal inference in possibly nonlinear factor models. *arxiv preprint: 2008.13651*.

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

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

Jin, S., Miao, K., and Su, L. (2021). On factor models with random missing: EM estimation, inference, and cross validation. *Journal of Econometrics*, 222(1):745–777.

Lei, L. and Ross, B. (2024). Estimating counterfactual matrix means with short panel data. *arxiv preprint: 2312.07520*.

Li, Y., Shah, D., Song, D., and Yu, C. L. (2019). Nearest neighbors for matrix estimation interpreted as blind regression for latent variable model. *IEEE Transactions on Information Theory*, 66(3):1760–1784.

Nguyen, L. T., Kim, J., and Shim, B. (2019). Low-rank matrix completion: A contemporary survey. *IEEE Access*, 7:94215–94237.

Robins, J. M., Hernan, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866.

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

Sloczynski, T., Uysal, S. D., and Wooldridge, J. M. (2024). Abadie's kappa and weighting estimators of the local average treatment effect. *Journal of Business & Economic Statistics*. Forthcoming.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press.

Wooldridge, J. M. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of econometrics*, 141(2):1281–1301.

Xiong, R. and Pelger, M. (2023). Large dimensional latent factor modeling with missing observations and applications to causal inference. *Journal of Econometrics*, 233(1):271–301.

Zhang, H. and Wei, H. (2022). Sharper sub-Weibull concentrations. *Mathematics*, 10(13):2252.