Doubly Robust Inference in Causal Latent Factor Models

Alberto Abadie
MIT

Anish Agarwal
Columbia

Raaz Dwivedi
Cornell Tech

Abhin Shah
MIT

October 22, 2024

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.

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.

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.

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 predetermined relative to the platform's treatments (advertisements, discounts, webpage 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 for estimation.

To our knowledge, this is the first article that leverages latent structures in both the assignment and the outcome processes to obtain a doubly-robust estimator of average treatment effects in the presence of unobserved confounding. 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), and Xiong and Pelger (2023) propose estimators that apply matrix completion techniques to impute potential outcomes. Although these studies utilize low-rank restrictions in the outcome process, they do not investigate the possibility of similar latent structures in the treatment assignment process. Our article addresses this question, and demonstrate substantial benefits from incorporating knowledge about the structure of the assignment mechanism.

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. The matrices $\Theta^{(0)} \triangleq \{\theta_{i,j}^{(0)}\}_{i \in [N], j \in [M]}$, $\Theta^{(1)} \triangleq \{\theta_{i,j}^{(1)}\}_{i \in [N], j \in [M]}$, and $P \triangleq \{p_{i,j}\}_{i \in [N], j \in [M]}$ collect mean potential outcomes and assignment probabilities. Then, the matrices $E^{(0)} \triangleq \{\varepsilon_{i,j}^{(0)}\}_{i \in [N], j \in [M]}, E^{(1)} \triangleq \{\varepsilon_{i,j}^{(1)}\}_{i \in [N], j \in [M]}$, 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,

$$\text{ATE}_{\cdot,j} \triangleq \mu_{\cdot,j}^{(1)} - \mu_{\cdot,j}^{(0)} \tag{5}$$

where

$$\mu_{\cdot,j}^{(a)} \triangleq \frac{1}{N} \sum_{i \in [N]} \theta_{i,j}^{(a)}.$$

$\text{ATE}_{\cdot,j}$ 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

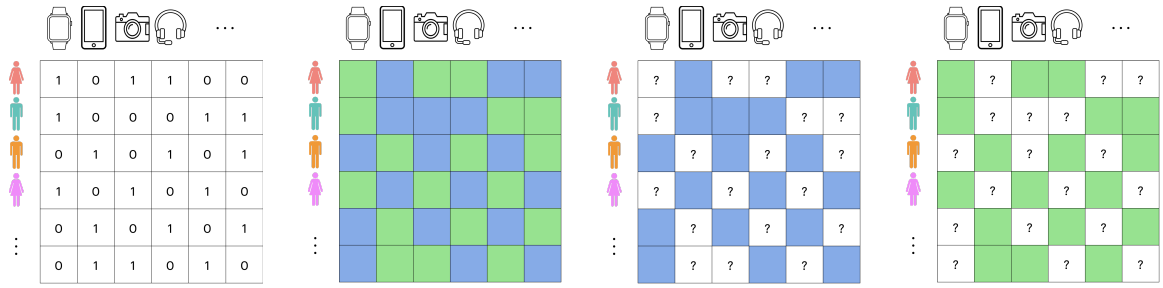
Consider a matrix of parameters $T \in \mathbb{R}^{N \times M}$. While T is unobserved, we observe the matrix $S \in \{\mathbb{R} \cup \{\,?\}\}^{N \times M}$ where ? denotes a missing value. The relationship between S and T is given by

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

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 $\texttt{MC}$, takes the S as its input, and returns an estimate of T, which we denote by $\widehat{T}$ or $\texttt{MC}(S)$. In other words, $\texttt{MC}$ 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 $\texttt{MC}$ that provide error guarantees, namely bounds on $\|\texttt{MC}(S) - T\|$ 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, $\texttt{MC}(S)$ subtracts the effects of H from S, i.e.,



(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.

it performs matrix denoising. Nguyen et al. (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 the following 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 satifies 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 $\widehat{\theta}^{(a)}_{i,j}$ 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:

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

where

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

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 the OI estimator 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:

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

where

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

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:

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

where

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

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 than 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 Section 6.

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*

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

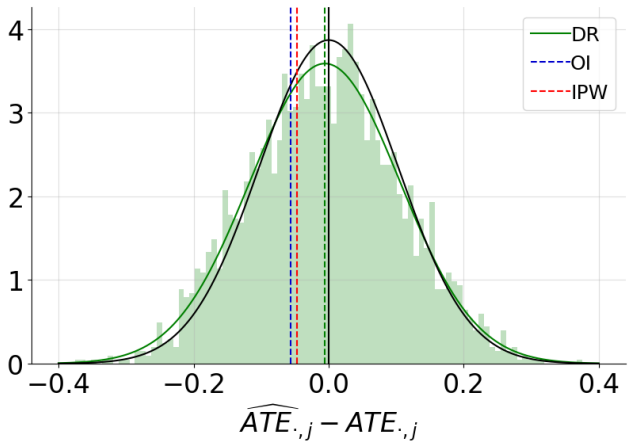


Figure 2: Simulation evidence of the convergence of the error of the doubly-robust (DR) estimator to a mean-zero Gaussian distribution. The histogram represents $\widehat{\text{ATE}}^{\text{DR}}_{\cdot,j} - \text{ATE}_{\cdot,j}$, the green curve represents the (best) fitted Gaussian distribution, and 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 6 reports complete simulation results.

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* $j \in [M]$*,* $(\varepsilon^{(0)}_{i,j}, \varepsilon^{(1)}_{i,j}) \perp\!\!\!\perp \eta_{i,j}$*; moreover, the distribution of* $(\varepsilon^{(0)}_{i,j}, \varepsilon^{(1)}_{i,j})$ *depends on* $(\Theta^{(0)}, \Theta^{(1)}, P)$ *only through* $(\theta^{(0)}_{i,j}, \theta^{(1)}_{i,j})$*, 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 disposed of 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*

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

and

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

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:

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

A variety of matrix completion algorithms deliver $\mathcal{E}\big(\widehat{P}\big) = O_p(\min\{N,M\}^{-\alpha})$ and $\mathcal{E}\big(\widehat{\Theta}\big) = O_p(\min\{N,M\}^{-\beta})$, where $0 < \alpha, \beta \leq 1/2$. The conditions in this section track dependence on N only. We say that the normalized errors $\mathcal{E}\big(\widehat{P}\big)$ and $\mathcal{E}\big(\widehat{\Theta}\big)$ 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 $\mathcal{E}\big(\widehat{P}\big)$ and $\mathcal{E}\big(\widehat{\Theta}\big)$ 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 $\mathcal{E}\big(\widehat{P}\big)$ and $\mathcal{E}\big(\widehat{\Theta}\big)$ 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*

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

where

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

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

The proof of Theorem 1 is given in Appendix A1. Eqs. (17) and (18) bound the absolute error of the DR estimator by the rate of $\mathcal{E}\big(\widehat{\Theta}\big)(\mathcal{E}\big(\widehat{P}\big) + N^{-0.5}) + N^{-0.5}$. When

$\mathcal{E}\big(\widehat{P}\big)$ is lower bounded at the parametric rate of $N^{-0.5}$, $\mathrm{Err}^{\mathrm{DR}}_{N,\delta}$ has the same rate as $\mathcal{E}\big(\widehat{P}\big)\mathcal{E}\big(\widehat{\Theta}\big) + N^{-0.5}$.

Doubly-robust behavior of $\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,j}$. The error rate of $\mathcal{E}\big(\widehat{P}\big)\mathcal{E}\big(\widehat{\Theta}\big) + N^{-0.5}$ 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, $\mathcal{E}\big(\widehat{P}\big)\mathcal{E}\big(\widehat{\Theta}\big)$, 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, $\mathrm{Err}^{\mathrm{DR}}_{N,\delta}$ decays to zero as long as either of $\mathcal{E}\big(\widehat{P}\big)$ or $\mathcal{E}\big(\widehat{\Theta}\big)$ 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 $\widehat{\mathrm{ATE}}^{\mathrm{OI}}_{\cdot,j}$ in terms of $\mathcal{E}\big(\widehat{\Theta}\big)$ and of $\widehat{\mathrm{ATE}}^{\mathrm{IPW}}_{\cdot,j}$ in terms of $\mathcal{E}\big(\widehat{P}\big)$.

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

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

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

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

where

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

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 S3 and S4). 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 $\mathcal{E}\big(\widehat{\Theta}\big)$ and $\mathcal{E}\big(\widehat{P}\big)$ are $O_p(N^{-0.5})$, respectively. The next corollary, proven in the supplementary appendix (Section S2), 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*

$\mathcal{E}\big(\widehat{P}\big) = O_p(N^{-\alpha})$ *and* $\mathcal{E}\big(\widehat{\Theta}\big) = O_p(N^{-\beta})$ *for* $0 \le \alpha \le 0.5$ *and* $0 \le \beta \le 0.5$*, then*

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

and

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

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 S2) as a corollary of Theorem 1, provides conditions on $\mathcal{E}\big(\widehat{P}\big)$ and $\mathcal{E}\big(\widehat{\Theta}\big)$ 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)* $\mathcal{E}\big(\widehat{P}\big) = o_p(1)$*,* $\mathcal{E}\big(\widehat{\Theta}\big) = O_p(1)$*, or (ii)* $\mathcal{E}\big(\widehat{\Theta}\big) = o_p(1)$*,* $\mathcal{E}\big(\widehat{P}\big) = O_p(1)$*, 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 Appendix A2, establishes a Gaussian approximation for $\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,j}$ under mild conditions on error rates $\mathcal{E}\big(\widehat{P}\big)$ and $\mathcal{E}\big(\widehat{\Theta}\big)$.

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

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

(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* $\sigma^{(0)}_{i,j}$ *and* $\sigma^{(1)}_{i,j}$ *be the standard deviations of* $\varepsilon^{(0)}_{i,j}$ *and* $\varepsilon^{(1)}_{i,j}$*, respectively. The sequence*

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

is bounded away from zero as N increases.

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

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

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 $\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,j}$ exhibits an asymptotic Gaussian distribution centered at $\mathrm{ATE}_{\cdot,j}$. 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 the 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, in that context, 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. We emphasize that 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, $\mathcal{E}\big(\widehat{\Theta}\big)$ and $\mathcal{E}\big(\widehat{P}\big)$, which are typically lower-bouded at the parametric rate of $N^{-0.5}$.

The next result, proven in Appendix A2.3, 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*

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

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

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

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 S6) 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 depends 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 $\{\widehat{p}_{i,j}\}_{i \in \mathcal{R}_s} \perp\!\!\!\perp \{\eta_{i,j}\}_{i \in \mathcal{R}_s}$ 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 $\{\widehat{p}_{i,j}\}_{i \in \mathcal{R}_s} \not\perp\!\!\!\perp \{\eta_{i,j}\}_{i \in \mathcal{R}_s}$, 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 $\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s} \perp\!\!\!\perp \{\eta_{i,j}\}_{i\in\mathcal{R}_s}$ requirement in Assumption 4. Likewise, the requirement $\{\widehat{\theta}^{(a)}_{i,j}\}_{i\in\mathcal{R}_s} \perp\!\!\!\perp \{\eta_{i,j}\}_{i\in\mathcal{R}_s}$ in Assumption 4 can be violated, because $\widehat{\Theta}^{(0)}$ and $\widehat{\Theta}^{(1)}$ depend respectively on $Y^{(0),\text{obs}}$ and $Y^{(1),\text{obs}}$, which themselves depend on the entire matrix 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 partitions $(\mathcal{C}_0, \mathcal{C}_1)$ 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 $k' \in \{0,1,\ldots,m\} \setminus \{k\}$ with $\{\eta_{i,j}\}_{(i,j)\in\mathcal{R}_s\times\mathcal{C}_k} \perp\!\!\!\perp \{\eta_{i,j}, \varepsilon^{(a)}_{i,j}\}_{(i,j)\in\mathcal{R}_{1-s}\times\mathcal{C}_{k'}}$ and $\{\eta_{i,j}\}_{(i,j)\in\mathcal{R}_{1-s}\times\mathcal{C}_{k'}} \perp\!\!\!\perp \{\eta_{i,j}, \varepsilon^{(a)}_{i,j}\}_{(i,j)\in\mathcal{R}_s\times\mathcal{C}_k}$. 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 S5.1), 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*

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

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

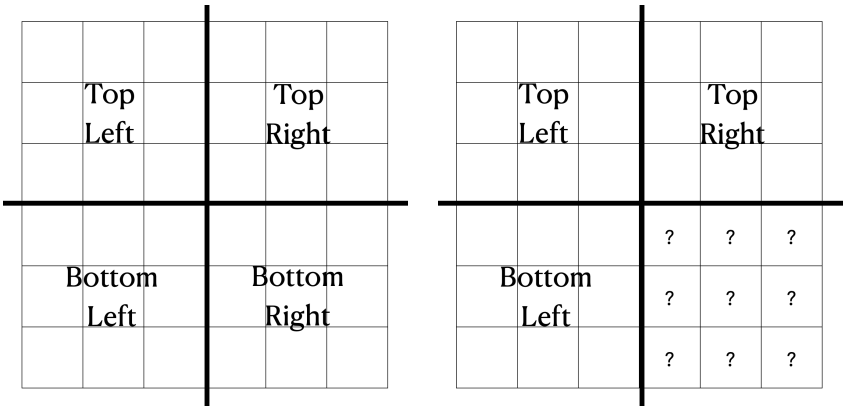
$$\widehat{P} = \texttt{Cross-Fitted-MC}(\texttt{MC}, A, \mathcal{P}), \tag{32}$$

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



(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 ?.

in one block are deterministically missing, as in Figure 3(b). A recent strand of research on the interplay 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 5.2.1, and then use it to describe `Cross-Fitted-SVD` in Section 5.2.2.

5.2.1. The `TW` algorithm of Bai and Ng (2021).

Bai and Ng (2021) propose `TW` to impute missing values in matrices with a set of rows and a set of columns without missing entries. More concretely, for any matrix

$S \in \{\mathbb{R} \cup \{?\}\}^{N \times M}$, let $\mathcal{R}_{\text{obs}} \subseteq [N]$ and $\mathcal{C}_{\text{obs}} \subseteq [M]$ denote the set of all rows and all columns, respectively, with all entries observed. Then, all missing entries of S belong to the block $\mathcal{I} = \mathcal{R}_{\text{miss}} \times \mathcal{C}_{\text{miss}}$, where $\mathcal{R}_{\text{miss}} \triangleq [N] \setminus \mathcal{R}_{\text{obs}}$ and $\mathcal{C}_{\text{miss}} \triangleq [M] \setminus \mathcal{C}_{\text{obs}}$.

Given a rank hyper-parameter $r \in [\min\{|\mathcal{R}_{\text{obs}}|, |\mathcal{C}_{\text{obs}}|\}]$, $\texttt{TW}_r$ produces an estimate of T as follows:

1. Run $\texttt{SVD}$ separately on $S^{(\text{tall})} \triangleq S_{[N] \times \mathcal{C}_{\text{obs}}}$ and $S^{(\text{wide})} \triangleq S_{\mathcal{R}_{\text{obs}} \times [M]}$, i.e.,

$$\texttt{SVD}(S^{(\text{tall})}) = (U^{(\text{tall})} \in \mathbb{R}^{N \times \overline{r}_N}, \Sigma^{(\text{tall})} \in \mathbb{R}^{\overline{r}_N \times \overline{r}_N}, V^{(\text{tall})} \in \mathbb{R}^{|\mathcal{C}_{\text{obs}}| \times \overline{r}_N})$$

and

$$\texttt{SVD}(S^{(\text{wide})}) = (U^{(\text{wide})} \in \mathbb{R}^{|\mathcal{R}_{\text{obs}}| \times \overline{r}_M}, \Sigma^{(\text{wide})} \in \mathbb{R}^{\overline{r}_M \times \overline{r}_M}, V^{(\text{wide})} \in \mathbb{R}^{M \times \overline{r}_M})$$

where $\overline{r}_N \triangleq \min\{N, |\mathcal{C}_{\text{obs}}|\}$ and $\overline{r}_M \triangleq \min\{|\mathcal{R}_{\text{obs}}|, M\}$. The columns of $U^{(\text{tall})}$ and $U^{(\text{wide})}$ are the left singular vectors of $S^{(\text{tall})}$ and $S^{(\text{wide})}$, respectively, and the columns of $V^{(\text{tall})}$ and $V^{(\text{wide})}$ are the right singular vectors of $S^{(\text{tall})}$ and $S^{(\text{wide})}$, respectively. The diagonal entries of $\Sigma^{(\text{tall})}$ and $\Sigma^{(\text{wide})}$ are the singular values of $S^{(\text{tall})}$ and $S^{(\text{wide})}$, respectively, and the off-diagonal entries are zeros. This step of $\texttt{TW}$ requires the existence of the fully observed blocks $S^{(\text{tall})}$ and $S^{(\text{wide})}$, i.e., $\mathcal{R}_{\text{obs}}$ and $\mathcal{C}_{\text{obs}}$ cannot be empty.

2. Let $\widetilde{V}^{(\text{tall})} \in \mathbb{R}^{|\mathcal{C}_{\text{obs}}| \times r}$ be the sub-matrix of $V^{(\text{tall})}$ that keeps the columns corresponding to the r largest singular values only. Let $\widetilde{V}^{(\text{wide})} \in \mathbb{R}^{|\mathcal{C}_{\text{obs}}| \times r}$ be the sub-matrix of $V^{(\text{wide})}$ that keeps the columns corresponding to the r largest singular values only and the rows corresponding to the indices in $\mathcal{C}_{\text{obs}}$ only. Obtain a rotation matrix $R \in \mathbb{R}^{r \times r}$ as follows:

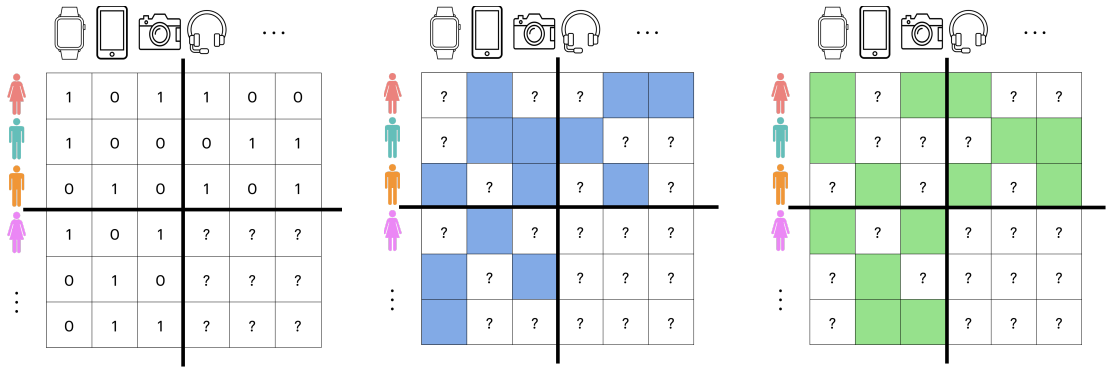
$$R \triangleq \widetilde{V}^{(\text{tall})\top} \widetilde{V}^{(\text{wide})} \big(\widetilde{V}^{(\text{wide})\top} \widetilde{V}^{(\text{wide})}\big)^{-1}.$$

That is, R is obtained by regressing $\widetilde{V}^{(\text{tall})}$ on $\widetilde{V}^{(\text{wide})}$. In essence, R aligns the right singular vectors of $S^{(\text{tall})}$ and $S^{(\text{wide})}$ using the entries that are common between these two matrices, i.e., the entries corresponding to indices $\mathcal{R}_{\text{obs}} \times \mathcal{C}_{\text{obs}}$. The formal guarantees of the $\texttt{TW}$ algorithm remains unchanged if one alternatively regresses $\widetilde{V}^{(\text{wide})}$ on $\widetilde{V}^{(\text{tall})}$, or uses the left singular vectors of $S^{(\text{tall})}$ and $S^{(\text{wide})}$ for alignment.

3. Let $\overline{\Sigma}^{(\text{tall})} \in \mathbb{R}^{\overline{r}_N \times r}$ be the sub-matrix of $\Sigma^{(\text{tall})}$ that keeps the columns corresponding to the r largest singular values only. Let $\overline{V}^{(\text{wide})} \in \mathbb{R}^{M \times r}$ be the sub-matrix of $V^{(\text{wide})}$ that keeps the columns corresponding to the r largest singular values only. Return $\widehat{T} \triangleq U^{(\text{tall})} \overline{\Sigma}^{(\text{tall})} R \overline{V}^{(\text{wide})\top}$ as an estimate for T.

5.2.2. `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,



(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), there exists rows and columns with all entries observed in Panel (a). 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 (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 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 $\widehat{P}$.

5.3. Theoretical guarantees for `Cross-Fitted-SVD`

To establish theoretical guarantees for `Cross-Fitted-SVD`, we adopt three assumptions from Bai and Ng (2021). The first assumption imposes a low-rank structure on the matrices P, $\Theta^{(0)}$, and $\Theta^{(1)}$, namely that their entries are given by an inner product of latent factors.

Assumption 6 (Linear latent factor model on the confounders)**.** *There exist constants* $r_p, r_{\theta_0}, r_{\theta_1} \in [\min\{N, M\}]$ *and a collection of latent factors*

$$U \in \mathbb{R}^{N \times r_p}, \quad V \in \mathbb{R}^{M \times r_p}, \quad U^{(a)} \in \mathbb{R}^{N \times r_{\theta_a}}, \quad \textit{and} \quad V^{(a)} \in \mathbb{R}^{M \times r_{\theta_a}} \quad \textit{for} \quad a \in \{0, 1\},$$

such that the unobserved confounders $(\Theta^{(0)}, \Theta^{(1)}, P)$ *satisfy the following factorization:*

$$P = UV^\top \quad \textit{and} \quad \Theta^{(a)} = U^{(a)} V^{(a)\top} \quad \textit{for} \quad a \in \{0, 1\}. \tag{34}$$

Assumption 6 decomposes each of the unobserved confounders (P, $\Theta^{(0)}$, and $\Theta^{(0)}$) into low-dimensional unit-dependent latent factors (U, $U^{(0)}$, and $U^{(1)}$) and measurement-dependent latent factors (V, $V^{(0)}$, and $V^{(1)}$). In particular, every unit $i \in [N]$ is associated with three low-dimensional factors: (i) $U_i \in \mathbb{R}^{r_p}$, (ii) $U_i^{(0)} \in \mathbb{R}^{r_{\theta_0}}$, and (iii) $U_i^{(1)} \in \mathbb{R}^{r_{\theta_1}}$. Similarly, every measurement $j \in [M]$ is associated with three factors: (i) $V_j \in \mathbb{R}^{r_p}$, (ii) $V_j^{(0)} \in \mathbb{R}^{r_{\theta_0}}$, and (iii) $V_j^{(1)} \in \mathbb{R}^{r_{\theta_1}}$. Low-rank assumptions are standard in the matrix completion literature.

The second assumption requires that the factors that determine P, $\Theta^{(0)} \odot (\mathbf{1} - P)$, and $\Theta^{(1)} \odot P$ explain a sufficiently large amount of the variation in the data. This assumption is made on the factors of $\Theta^{(0)} \odot (\mathbf{1} - P)$ and $\Theta^{(1)} \odot P$ instead of $\Theta^{(0)}$ and $\Theta^{(1)}$ as the `TW` algorithm is applied on $Y^{(0),\text{full}} = Y \odot (\mathbf{1} - A)$ and $Y^{(1),\text{full}} = Y \odot A$,

instead of $Y^{(0),\text{obs}}$ and $Y^{(1),\text{obs}}$ (see steps 4 and 5 of `Cross-Fitted-SVD`). To determine the factors of $\Theta^{(0)} \odot (\mathbf{1} - P)$ and $\Theta^{(1)} \odot P$, let

$$\overline{U} \triangleq [\mathbf{1}_N, -U] \in \mathbb{R}^{N \times (r_p+1)} \quad \text{and} \quad \overline{V} \triangleq [\mathbf{1}_M, V] \in \mathbb{R}^{M \times (r_p+1)},$$

where $\mathbf{1}_N \in \mathbb{R}^N$ and $\mathbf{1}_M \in \mathbb{R}^M$ are vectors of all 1's. Then,

$$\Theta^{(0)} \odot (\mathbf{1} - P) = \overline{U}^{(0)} \overline{V}^{(0)\top} \quad \text{and} \quad \Theta^{(1)} \odot P = \overline{U}^{(1)} \overline{V}^{(1)\top}, \tag{35}$$

where $\overline{U}^{(0)} \triangleq \overline{U} * U^{(0)} \in \mathbb{R}^{N \times r_{\theta_0}(r_p+1)}$, $\overline{V}^{(0)} \triangleq \overline{V} * V^{(0)} \in \mathbb{R}^{M \times r_{\theta_0}(r_p+1)}$, $\overline{U}^{(1)} \triangleq U * U^{(1)} \in \mathbb{R}^{N \times r_{\theta_1} r_p}$, and $\overline{V}^{(1)} \triangleq V * V^{(1)} \in \mathbb{R}^{M \times r_{\theta_1} r_p}$, with the operator $*$ denoting the Khatri-Rao product (see Section 1). We provide details of the derivation of these factors in the supplementary appendix (Section S5.2.3).

Assumption 7 (Strong factors)**.** *There exists a positive constant c such that*

$$\|U\|_{2,\infty} \leq c, \quad \|V\|_{2,\infty} \leq c, \quad \|U^{(a)}\|_{2,\infty} \leq c, \quad \textit{and} \quad \|V^{(a)}\|_{2,\infty} \leq c \quad \textit{for} \quad a \in \{0,1\}.$$

Further, the matrices defined below exist and are positive definite:

$$\lim_{N\to\infty} \frac{U^\top U}{N}, \quad \lim_{M\to\infty} \frac{V^\top V}{M}, \quad \lim_{N\to\infty} \frac{\overline{U}^{(a)\top} \overline{U}^{(a)}}{N}, \quad \textit{and} \quad \lim_{M\to\infty} \frac{\overline{V}^{(a)\top} \overline{V}^{(a)}}{M} \quad \textit{for} \quad a \in \{0,1\}.$$

Assumption 7, a classic assumption in the literature on latent factor models, ensures that the factor structure is strong. Specifically, it ensures that each eigenvector of P, $\Theta^{(0)} \odot (\mathbf{1} - P)$, and $\Theta^{(1)} \odot P$ carries sufficiently large signal.

The third assumption requires a strong factor structure on the sub-matrices of P, $\Theta^{(0)} \odot (\mathbf{1} - P)$, and $\Theta^{(1)} \odot P$ corresponding to every block $\mathcal{I}$ in the block partition $\mathcal{P}$ from Assumption 5. Further, it also requires that the size $\mathcal{I}$ grows linearly in N and M.

Assumption 8 (Strong block factors)**.** *Consider the block partition $\mathcal{P} \triangleq \{\mathcal{R}_s \times \mathcal{C}_k : s,k \in \{0,1\}\}$ from Assumption 5. For every $s \in \{0,1\}$, let $U_{(s)} \in \mathbb{R}^{|\mathcal{R}_s| \times r_p}$, $\overline{U}^{(0)}_{(s)} \in \mathbb{R}^{|\mathcal{R}_s| \times r_{\theta_0}(r_p+1)}$, and $\overline{U}^{(1)}_{(s)} \in \mathbb{R}^{|\mathcal{R}_s| \times r_{\theta_1} r_p}$ be the sub-matrices of U, $\overline{U}^{(0)}$, and $\overline{U}^{(1)}$, respectively, that keeps the rows corresponding to the indices in $\mathcal{R}_s$. For every $k \in \{0,1\}$, let $V_{(k)} \in \mathbb{R}^{|\mathcal{C}_k| \times r_p}$, $\overline{V}^{(0)}_{(k)} \in \mathbb{R}^{|\mathcal{C}_k| \times r_{\theta_0}(r_p+1)}$, and $\overline{V}^{(1)}_{(k)} \in \mathbb{R}^{|\mathcal{C}_k| \times r_{\theta_1} r_p}$ be the sub-matrices of V, $\overline{V}^{(0)}$, and $\overline{V}^{(1)}$, respectively, that keeps the rows corresponding to the indices in $\mathcal{C}_k$. Then, for every $s,k \in \{0,1\}$, the matrices defined below exist and are positive definite:*

$$\lim_{N\to\infty} \frac{U_{(s)}^\top U_{(s)}}{|\mathcal{R}_s|}, \quad \lim_{M\to\infty} \frac{V_{(k)}^\top V_{(k)}}{|\mathcal{C}_k|}, \quad \lim_{N\to\infty} \frac{\overline{U}^{(a)\top}_{(s)} \overline{U}^{(a)}_{(s)}}{|\mathcal{R}_s|}, \quad \textit{and} \quad \lim_{M\to\infty} \frac{\overline{V}^{(a)\top}_{(k)} \overline{V}^{(a)}_{(k)}}{|\mathcal{C}_k|} \quad \textit{for } a \in \{0,1\}.$$

Further, for every $s, k \in \{0,1\}$, $|\mathcal{R}_s| = \Omega(N)$ *and* $|\mathcal{C}_k| = \Omega(M)$.

The subsequent assumption introduces additional conditions on the noise variables in Bai and Ng (2021) than those specified in Assumptions 2 and 5.

Assumption 9 (Weak dependence in noise across measurements and independence in noise across units)**.**

(a) $\sum_{j' \in [M]} \left|\mathbb{E}[\eta_{i,j}\eta_{i,j'}]\right| \leq c$ *for every* $i \in [N]$ *and* $j \in [M]$,

(b) $\sum_{j' \in [M]} \left|\mathbb{E}[\bar{\varepsilon}^{(a)}_{i,j}\bar{\varepsilon}^{(a)}_{i,j'}]\right| \leq c$ *for every* $i \in [N]$, $j \in [M]$, *and* $a \in \{0,1\}$, *where* $\bar{\varepsilon}^{(a)}_{i,j} \triangleq \theta_{i,j}\eta_{i,j} + \varepsilon^{(a)}_{i,j}p_{i,j} + \varepsilon^{(a)}_{i,j}\eta_{i,j}$, *and*

(c) The elements of $\{(E^{(a)}_{i,\cdot}, W_{i,\cdot}) : i \in [N]\}$ *are mutually independent (across* i*) for* $a \in \{0,1\}$.

Assumption 9(a) and Assumption 9(b) requires the noise variables to exhibit only weak dependency across measurements. Still, these assumptions allow the existence of pairs of perfectly correlated outcomes (e.g., $j, j' \in [M]$ such that $a_{i,j} = a_{i,j'}$). Assumption 9(c) requires the noise $(E^{(a)}, W)$ to be jointly independent across units, for every $a \in \{0,1\}$. We are now ready to provide guarantees on the estimates produced by `Cross-Fitted-SVD`. The proof can be found in the supplementary appendix (Section S5.2).

Proposition 4 (**Guarantees for** `Cross-Fitted-SVD`)**.** *Suppose Assumptions 1, 2, and 6 to 9 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* $N, M \to \infty$,

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

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 $\mathcal{E}\big(\widehat{P}\big)$ and $\mathcal{E}\big(\widehat{\Theta}\big)$ achieve the parametric rate whenever $N/M = O(1)$.

5.4. 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 S7) 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

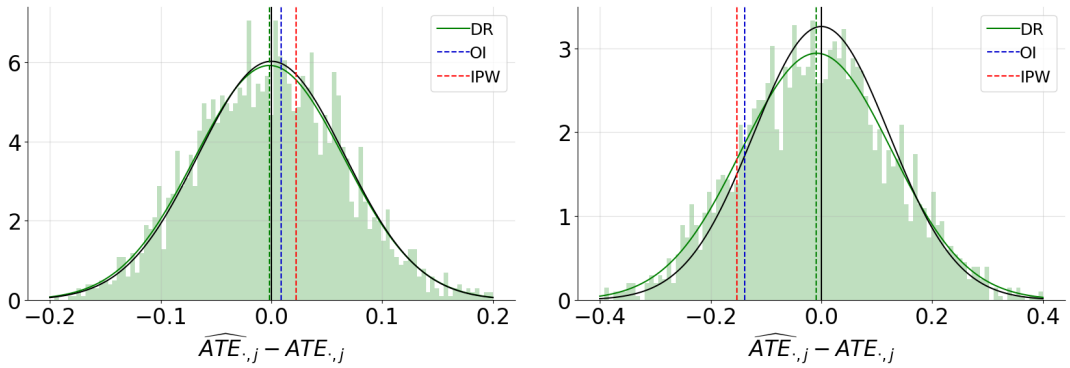
data setting M measurements correspond to T time periods, and t denotes the time index. Then, the supplementary appendix considers a setting where a unit remains under control for some period of time, after which it deterministically remains under treatment. In other words, for every unit $i \in [N]$, 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. Simulations

This section reports simulation results on the performance of the DR estimator of Eq. (11) and the OI and IPW estimators of Eqs. (9) and (10), respectively.

Data Generating Process (DGP). We now briefly describe the DGP for our simulations; Appendix A3 provides details. All simulations set $N = M$. To generate, P, $\Theta^{(0)}$, and $\Theta^{(1)}$, we use the latent factor model given in Eq. (34). To introduce unobserved confounding, we set the unit-specific latent factors to be the same across P, $\Theta^{(0)}$, and $\Theta^{(1)}$, i.e., $U = U^{(0)} = U^{(1)}$. The entries of U and the measurement-specific latent factors, $V, V^{(0)}, V^{(1)}$ are each sampled independently from a uniform distribution, with hyperparameter r_p equal to the dimension of U and V, and hyperparameter r_p equal to the dimension of $U^{(a)}$ and $V^{(a)}$ for $a = 0, 1$. Further, the entries of the noise matrices $E^{(0)}$ and $E^{(1)}$ are sampled independently from a normal distribution, and the entries of W are sampled independently as in Eq. (4). Then, $y^{(a)}_{i,j}$, $a_{i,j}$, and $y_{i,j}$ are determined from Eqs. (1) to (3), respectively. The simulation generates P, $\Theta^{(0)}$, and $\Theta^{(1)}$ once. Given the fixed values of P, $\Theta^{(0)}$, and $\Theta^{(1)}$, the simulation generates 2500 realizations of (Y, A)—that is, only the noise matrices $E^{(0)}, E^{(1)}, W$ are resampled for each of the 2500 realizations. For each simulation realization, we apply the `Cross-Fitted-SVD` algorithm with hyper-parameters as in Proposition 4 and $\bar{\lambda} = \lambda = 0.05$ to obtain $\widehat{P}$, $\widehat{\Theta}^{(0)}$, and $\widehat{\Theta}^{(1)}$, and compute $\text{ATE}_{\cdot,j}$ from Eq. (5), and $\widehat{\text{ATE}}^{\text{OI}}_{\cdot,j}$, $\widehat{\text{ATE}}^{\text{IPW}}_{\cdot,j}$ and $\widehat{\text{ATE}}^{\text{DR}}_{\cdot,j}$ from Eqs. (9) to (11).

Results. Figure 5 reports simulation results for $N = 1000$, with $r_p = 3$, $r_\theta = 3$ in Panel (a), and $r_p = 5$, $r_\theta = 3$ in Panel (b). Figure 2 in Section 3 reports simulation results for $r_p = 3$, $r_\theta = 5$. In each case, the figure shows a histogram of the distribution of $\widehat{\text{ATE}}^{\text{DR}}_{\cdot,j} - \text{ATE}_{\cdot,j}$ across 2500 simulation instances for a fixed j, along with the best fitting Gaussian distribution (green curve). The histogram counts are normalized so that the area under the histogram integrates to one. Figure 5 plots the Gaussian distribution in the result of Theorem 2 (black curve). The dashed blue, red and green lines in Figures 2 and 5 indicate the values of the means of the OI, IPW, and DR error, respectively, across simulation instances. For reference, we place a black solid line at zero. The DR estimator has minimal bias and a close-to-Gaussian distribution.



(a) $r_p = 3$, $r_\theta = 3$ (b) $r_p = 5$, $r_\theta = 3$

Figure 5: Empirical illustration of the asymptotic performance of DR as in Theorem 2. The histogram corresponds to the errors of 2500 independent instances of DR estimates, the green curve represents the (best) fitted Gaussian distribution, and the black curve represents the Gaussian approximation from Theorem 2. The dashed green, blue, and red lines represent the biases of DR, OI, and IPW estimators.

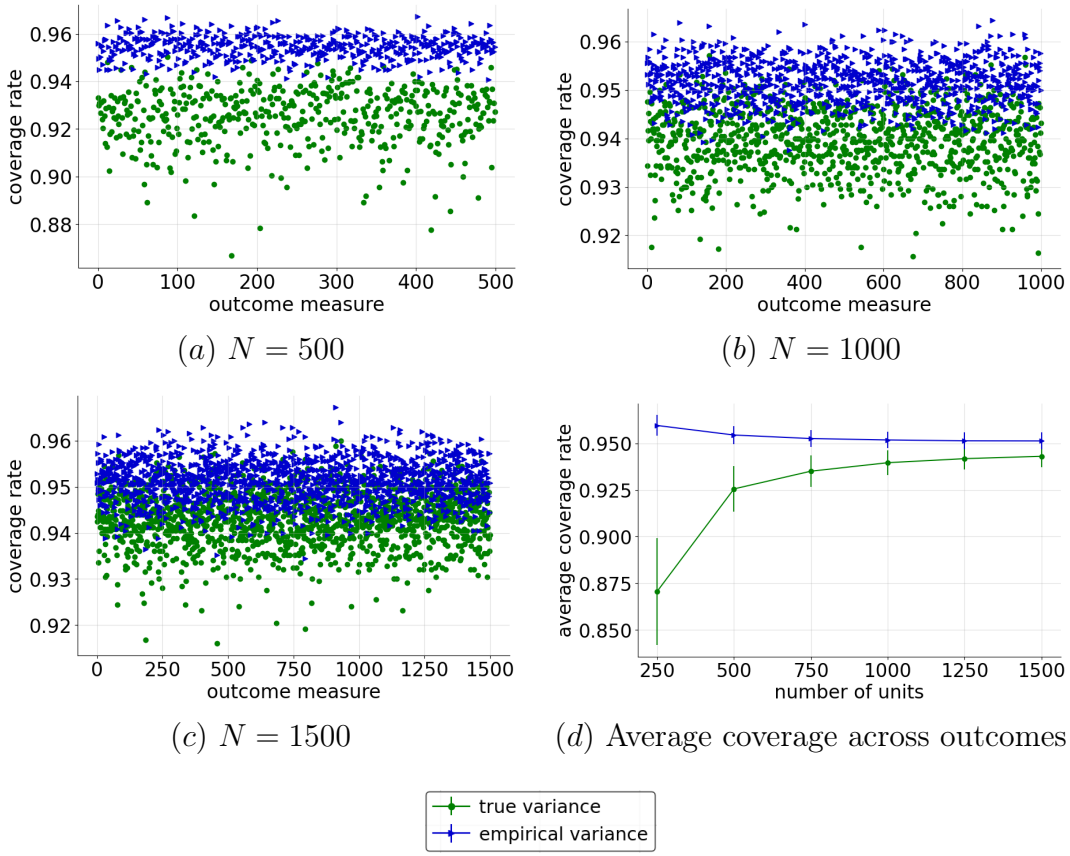
The biases of OI and IPW are non-negligible. In Appendix A3, we compare the biases and the standard deviations of OI, IPW, and DR across many j.

Panels (a), (b), and (c) of Figure 6 report coverage rates over the 2500 simulations for $\widehat{\text{ATE}}^{\text{DR}}_{\cdot,j}$-centered nominal 95% confidence intervals with $N = 500$, $N = 1000$, and $N = 1500$, respectively, all with $M = N$ and $r_p = r_\theta = 3$. For every $j \in [M]$, panels (a), (b) and (c) show $\widehat{\text{c}}_j$, the percentage of times $[\widehat{\text{ATE}}^{\text{DR}}_{\cdot,j} \pm 1.96\widehat{\sigma}_j/\sqrt{N}]$ covers $\text{ATE}_{\cdot,j}$ (in blue), and c_j, the percentage of times $[\widehat{\text{ATE}}^{\text{DR}}_{\cdot,j} \pm 1.96\sigma_j/\sqrt{N}]$ covers $\text{ATE}_{\cdot,j}$ (in green). Panel (d) shows the means and standard deviations of $\{\widehat{\text{c}}_j\}_{j\in[M]}$ and $\{\text{c}_j\}_{j\in[M]}$ for different values of N. Confidence intervals based on the large-sample approximation results of Section 4 exhibit small size distortion even for fairly small values of N.

7. Conclusion

This article introduces a new framework to estimate treatment effects in the presence unobserved confounding. We consider modern data-rich environments, where there are many units, and outcomes of interest per unit. 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.

Our proposed estimator is doubly-robust, combining outcome imputation and inverse probability weighting with matrix completion. Analytical tractability of its distribution is gained through a novel cross-fitting procedure for causal matrix completion. We study the properties of the doubly-robust estimator, along with the outcome imputation and inverse probability weighting-based estimators under black-box matrix completion error rates. We show that the decay rate of the error of the



(a) $N = 500$ (b) $N = 1000$

(c) $N = 1500$ (d) Average coverage across outcomes

Figure 6: Panels (a), (b), and (c) report coverage rates for nominal 95% confidence intervals constructed using the estimated variance from Eq. (25) (in blue) and the true variance from Eq. (22) (in green) for $N \in \{500, 1000, 1500\}$ and $M = N$. Panel (d) shows the means and standard deviations of coverage rates across outcomes for different values of N.

doubly-robust estimator dominates those of the outcome imputation and the inverse probability weighting estimators. Moreover, we establish a Gaussian approximation to the distribution of the doubly-robust estimator. Simulation results demonstrate the practical relevance of the formal properties of the doubly-robust estimator.

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 arXiv: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.

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.

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 arXiv: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 arXiv:2211.14297*.

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.

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.

Appendices

A1. Proof of Theorem 1: Finite Sample Guarantees for DR

Fix any $j \in [M]$. Recall the definitions of the parameter $\text{ATE}_{\cdot,j}$ and corresponding doubly-robust estimate $\widehat{\text{ATE}}^{\text{DR}}_{\cdot,j}$ from Eqs. (5) and (11), respectively. The error $\Delta\text{ATE}^{\text{DR}}_{\cdot,j} = \widehat{\text{ATE}}^{\text{DR}}_{\cdot,j} - \text{ATE}_{\cdot,j}$ can be re-expressed as

$$\Delta\text{ATE}^{\text{DR}}_{\cdot,j} = \frac{1}{N}\sum_{i\in[N]}\left(\widehat{\theta}^{(1,\text{DR})}_{i,j} - \widehat{\theta}^{(0,\text{DR})}_{i,j}\right) - \frac{1}{N}\sum_{i\in[N]}\left(\theta^{(1)}_{i,j} - \theta^{(0)}_{i,j}\right)$$

$$= \frac{1}{N}\sum_{i\in[N]}\Big(\big(\widehat{\theta}_{i,j}^{(1,\mathrm{DR})} - \theta_{i,j}^{(1)}\big) - \big(\widehat{\theta}_{i,j}^{(0,\mathrm{DR})} - \theta_{i,j}^{(0)}\big)\Big)$$
$$\overset{(a)}{=} \frac{1}{N}\sum_{i\in[N]}\big(\mathbb{T}_{i,j}^{(1,\mathrm{DR})} + \mathbb{T}_{i,j}^{(0,\mathrm{DR})}\big), \tag{A.1}$$

where (a) follows after defining $\mathbb{T}_{i,j}^{(1,\mathrm{DR})} \triangleq \big(\widehat{\theta}_{i,j}^{(1,\mathrm{DR})} - \theta_{i,j}^{(1)}\big)$ and $\mathbb{T}_{i,j}^{(0,\mathrm{DR})} \triangleq -\big(\widehat{\theta}_{i,j}^{(0,\mathrm{DR})} - \theta_{i,j}^{(0)}\big)$ for every $(i,j) \in [N]\times[M]$. Then, we have

$$\begin{aligned}
\mathbb{T}_{i,j}^{(1,\mathrm{DR})} &= \widehat{\theta}_{i,j}^{(1,\mathrm{DR})} - \theta_{i,j}^{(1)} \\
&\overset{(a)}{=} \widehat{\theta}_{i,j}^{(1)} + \big(y_{i,j} - \widehat{\theta}_{i,j}^{(1)}\big)\frac{a_{i,j}}{\widehat{p}_{i,j}} - \theta_{i,j}^{(1)} \\
&\overset{(b)}{=} \widehat{\theta}_{i,j}^{(1)} + \big(\theta_{i,j}^{(1)} + \varepsilon_{i,j}^{(1)} - \widehat{\theta}_{i,j}^{(1)}\big)\frac{p_{i,j}+\eta_{i,j}}{\widehat{p}_{i,j}} - \theta_{i,j}^{(1)} \qquad \text{(A.2)} \\
&= (\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)})\Big(1 - \frac{p_{i,j}+\eta_{i,j}}{\widehat{p}_{i,j}}\Big) + \varepsilon_{i,j}^{(1)}\Big(\frac{p_{i,j}+\eta_{i,j}}{\widehat{p}_{i,j}}\Big) \\
&= \frac{(\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)})(\widehat{p}_{i,j} - p_{i,j})}{\widehat{p}_{i,j}} - \frac{(\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)})\eta_{i,j}}{\widehat{p}_{i,j}} + \frac{\varepsilon_{i,j}^{(1)} p_{i,j}}{\widehat{p}_{i,j}} + \frac{\varepsilon_{i,j}^{(1)}\eta_{i,j}}{\widehat{p}_{i,j}}, \qquad \text{(A.3)}
\end{aligned}$$

where (a) follows from Eq. (12), and (b) follows from Eqs. (1) to (3). A similar derivation for $a = 0$ implies that

$$\begin{aligned}
\mathbb{T}_{i,j}^{(0,\mathrm{DR})} &= -\frac{(\widehat{\theta}_{i,j}^{(0)} - \theta_{i,j}^{(0)})(1 - \widehat{p}_{i,j} - (1 - p_{i,j}))}{1-\widehat{p}_{i,j}} + \frac{(\widehat{\theta}_{i,j}^{(0)} - \theta_{i,j}^{(0)})(-\eta_{i,j})}{1-\widehat{p}_{i,j}} - \frac{\varepsilon_{i,j}^{(0)}(1-p_{i,j})}{1-\widehat{p}_{i,j}} \\
&\quad - \frac{\varepsilon_{i,j}^{(0)}(-\eta_{i,j})}{1-\widehat{p}_{i,j}} \\
&= \frac{(\widehat{\theta}_{i,j}^{(0)} - \theta_{i,j}^{(0)})(\widehat{p}_{i,j} - p_{i,j})}{1-\widehat{p}_{i,j}} - \frac{(\widehat{\theta}_{i,j}^{(0)} - \theta_{i,j}^{(0)})\eta_{i,j}}{1-\widehat{p}_{i,j}} - \frac{\varepsilon_{i,j}^{(0)}(1-p_{i,j})}{1-\widehat{p}_{i,j}} + \frac{\varepsilon_{i,j}^{(0)}\eta_{i,j}}{1-\widehat{p}_{i,j}}. \qquad \text{(A.4)}
\end{aligned}$$

Consider any $a \in \{0,1\}$ and any $\delta \in (0,1)$. We claim that, with probability at least $1 - 6\delta$,

$$\frac{1}{N}\Big|\sum_{i\in[N]}\mathbb{T}_{i,j}^{(\mathrm{a},\mathrm{DR})}\Big| \le \frac{2}{\bar{\lambda}}\mathcal{E}\big(\widehat{\Theta}^{(a)}\big)\mathcal{E}\big(\widehat{P}\big) + \frac{2\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{\ell_1 N}}\mathcal{E}\big(\widehat{\Theta}^{(a)}\big) + \frac{2\overline{\sigma}\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{N}} + \frac{2\overline{\sigma}m(c\ell_\delta)}{\bar{\lambda}\sqrt{\ell_1 N}}, \tag{A.5}$$

where recall that $m(c\ell_\delta) = \max\big(c\ell_\delta, \sqrt{c\ell_\delta}\big)$. We provide a proof of this claim at the end of this section. Applying triangle inequality in Eq. (A.1) and using Eq. (A.5) with a union bound, we obtain that

$$\big|\Delta\mathrm{ATE}_{\cdot,j}^{\mathrm{DR}}\big| \le \frac{2}{\bar{\lambda}}\mathcal{E}\big(\widehat{\Theta}\big)\mathcal{E}\big(\widehat{P}\big) + \frac{2\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{\ell_1 N}}\mathcal{E}\big(\widehat{\Theta}\big) + \frac{4\overline{\sigma}\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{N}} + \frac{4\overline{\sigma}m(c\ell_\delta)}{\bar{\lambda}\sqrt{\ell_1 N}},$$

with probability at least $1 - 12\delta$. The claim in Eq. (18) follows by re-parameterizing δ.

Proof of bound Eq. (A.5). Recall the partitioning of the units $[N]$ into $\mathcal{R}_0$ and $\mathcal{R}_1$ from Assumption 4. Now, to enable the application of concentration bounds, we split the summation over $i \in [N]$ in the left hand side of Eq. (A.5) into two parts—one over $i \in \mathcal{R}_0$ and the other over $i \in \mathcal{R}_1$—such that the noise terms are independent of the estimates of $\Theta^{(0)}, \Theta^{(1)}, P$ in each of these parts as in Eqs. (14) and (15).

Fix $a = 1$ and note that $|\sum_{i\in[N]} \mathbb{T}^{(1,\mathrm{DR})}_{i,j}| \leq |\sum_{i\in\mathcal{R}_0} \mathbb{T}^{(1,\mathrm{DR})}_{i,j}| + |\sum_{i\in\mathcal{R}_1} \mathbb{T}^{(1,\mathrm{DR})}_{i,j}|$. Fix any $s \in \{0, 1\}$. Then, Eq. (A.3) and triangle inequality imply

$$\begin{aligned}\Big|\sum_{i\in\mathcal{R}_s} \mathbb{T}^{(1,\mathrm{DR})}_{i,j}\Big| \leq \Big|\sum_{i\in\mathcal{R}_s} \frac{\big(\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}\big)\big(\widehat{p}_{i,j} - p_{i,j}\big)}{\widehat{p}_{i,j}}\Big| + \Big|\sum_{i\in\mathcal{R}_s} \frac{\big(\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}\big)\eta_{i,j}}{\widehat{p}_{i,j}}\Big| \\ + \Big|\sum_{i\in\mathcal{R}_s} \frac{\varepsilon^{(1)}_{i,j} p_{i,j}}{\widehat{p}_{i,j}}\Big| + \Big|\sum_{i\in\mathcal{R}_s} \frac{\varepsilon^{(1)}_{i,j}\eta_{i,j}}{\widehat{p}_{i,j}}\Big|. \qquad \text{(A.6)}\end{aligned}$$

Applying the Cauchy-Schwarz inequality to bound the first term yields that

$$\begin{aligned}\Big|\sum_{i\in\mathcal{R}_s} \frac{\big(\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}\big)\big(\widehat{p}_{i,j} - p_{i,j}\big)}{\widehat{p}_{i,j}}\Big| &\leq \sqrt{\sum_{i\in\mathcal{R}_s} \Big(\frac{\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}}{\widehat{p}_{i,j}}\Big)^2 \sum_{i\in\mathcal{R}_s} \big(\widehat{p}_{i,j} - p_{i,j}\big)^2} \\ &\leq \big\|\big(\widehat{\Theta}^{(1)}_{\cdot,j} - \Theta^{(1)}_{\cdot,j}\big) \oslash \widehat{P}_{\cdot,j}\big\|_2 \big\|\widehat{P}_{\cdot,j} - P_{\cdot,j}\big\|_2. \qquad \text{(A.7)}\end{aligned}$$

To bound the second term in Eq. (A.6), note that $\eta_{i,j}$ is subGaussian$(1/\sqrt{\ell_1})$ (see Example 2.5.8 in Vershynin (2018)) as well as zero-mean and independent across all $i \in [N]$ due to Assumption 2(a). By Assumption 4, $\{(\widehat{p}_{i,j}, \widehat{\theta}^{(1)}_{i,j})\}_{i\in\mathcal{R}_s} \perp\!\!\!\perp \{\eta_{i,j}\}_{i\in\mathcal{R}_s}$. The subGaussian concentration result in Corollary S1 yields

$$\Big|\sum_{i\in\mathcal{R}_s} \frac{\big(\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}\big)\eta_{i,j}}{\widehat{p}_{i,j}}\Big| \leq \frac{\sqrt{c\ell_\delta}}{\sqrt{\ell_1}} \sqrt{\sum_{i\in\mathcal{R}_s} \Big(\frac{\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}}{\widehat{p}_{i,j}}\Big)^2} \leq \frac{\sqrt{c\ell_\delta}}{\sqrt{\ell_1}} \big\|\big(\widehat{\Theta}^{(1)}_{\cdot,j} - \Theta^{(1)}_{\cdot,j}\big) \oslash \widehat{P}_{\cdot,j}\big\|_2, \quad \text{(A.8)}$$

with probability at least $1 - \delta$.

To bound the third term in Eq. (A.6), note that $\varepsilon^{(1)}_{i,j}$ is subGaussian$(\overline{\sigma})$, zero-mean, and independent across all $i \in [N]$ due to Assumption 2. By Assumption Assumption 4, $\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s} \perp\!\!\!\perp \{\varepsilon^{(1)}_{i,j}\}_{i\in\mathcal{R}_s}$. The subGaussian concentration result in Corollary S1 yields

$$\Big|\sum_{i\in\mathcal{R}_s} \frac{\varepsilon^{(1)}_{i,j} p_{i,j}}{\widehat{p}_{i,j}}\Big| \leq \overline{\sigma}\sqrt{c\ell_\delta} \sqrt{\sum_{i\in\mathcal{R}_s} \Big(\frac{p_{i,j}}{\widehat{p}_{i,j}}\Big)^2} \leq \overline{\sigma}\sqrt{c\ell_\delta} \big\|P_{\cdot,j} \oslash \widehat{P}_{\cdot,j}\big\|_2, \qquad \text{(A.9)}$$

with probability at least $1 - \delta$.

To bound the fourth term in Eq. (A.6), note that $\varepsilon_{i,j}^{(1)}\eta_{i,j}$ is subExponential$(\overline{\sigma}/\sqrt{\ell_1})$ because of Lemma S3 as well as zero-mean and independent across all $i \in [N]$ due to Assumption 2. By Assumption 4, $\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s} \perp\!\!\!\perp \{(\eta_{i,j}, \varepsilon_{i,j}^{(1)})\}_{i\in\mathcal{R}_s}$. The subExponential concentration result in Corollary S2 yields that

$$\Big|\sum_{i\in\mathcal{R}_s} \frac{\varepsilon_{i,j}^{(1)}\eta_{i,j}}{\widehat{p}_{i,j}}\Big| \le \frac{\overline{\sigma}m(c\ell_\delta)}{\sqrt{\ell_1}}\|\mathbf{1}_N \oslash \widehat{P}_{\cdot,j}\|_2, \tag{A.10}$$

with probability at least $1-\delta$. Putting together Eqs. (A.6) to (A.10), we conclude that, with probability at least $1-3\delta$,

$$\begin{aligned}\frac{1}{N}\Big|\sum_{i\in\mathcal{R}_s}\mathbb{T}_{i,j}^{(1,\mathrm{DR})}\Big| \le \frac{1}{N}\big\|\big(\widehat{\Theta}_{\cdot,j}^{(1)}-\Theta_{\cdot,j}^{(1)}\big)\oslash\widehat{P}_{\cdot,j}\big\|_2\big\|\widehat{P}_{\cdot,j}-P_{\cdot,j}\big\|_2 + \frac{\sqrt{c\ell_\delta}}{\sqrt{\ell_1}N}\big\|\big(\widehat{\Theta}_{\cdot,j}^{(1)}-\Theta_{\cdot,j}^{(1)}\big)\oslash\widehat{P}_{\cdot,j}\big\|_2 \\ + \frac{\overline{\sigma}\sqrt{c\ell_\delta}}{N}\big\|P_{\cdot,j}\oslash\widehat{P}_{\cdot,j}\big\|_2 + \frac{\overline{\sigma}m(c\ell_\delta)}{\sqrt{\ell_1}N}\big\|\mathbf{1}_N\oslash\widehat{P}_{\cdot,j}\big\|_2.\end{aligned}$$

Then, noting that $1/\widehat{p}_{i,j} \le 1/\bar{\lambda}$ for every $i \in [N]$ and $j \in [M]$ from Assumption 3, and consequently that $\|B_{\cdot,j}\oslash\widehat{P}_{\cdot,j}\|_2 \le \|B\|_{1,2}/\bar{\lambda}$ for any matrix B and every $j \in [M]$, we obtain the following bound, with probability at least $1-3\delta$,

$$\begin{aligned}\frac{1}{N}\Big|\sum_{i\in\mathcal{R}_s}\mathbb{T}_{i,j}^{(1,\mathrm{DR})}\Big| &\le \frac{1}{\bar{\lambda}N}\|\widehat{\Theta}^{(1)}-\Theta^{(1)}\|_{1,2}\|\widehat{P}-P\|_{1,2} + \frac{\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{\ell_1}N}\|\widehat{\Theta}^{(1)}-\Theta^{(1)}\|_{1,2} \\ &\quad + \frac{\overline{\sigma}\sqrt{c\ell_\delta}}{\bar{\lambda}N}\|P\|_{1,2} + \frac{\overline{\sigma}m(c\ell_\delta)}{\bar{\lambda}\sqrt{\ell_1}N}\|\mathbf{1}\|_{1,2} \qquad (A.11)\\ &\overset{(a)}{\le} \frac{1}{\bar{\lambda}}\mathcal{E}\big(\widehat{\Theta}^{(1)}\big)\mathcal{E}\big(\widehat{P}\big) + \frac{\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{\ell_1}N}\mathcal{E}\big(\widehat{\Theta}^{(1)}\big) + \frac{\overline{\sigma}\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{N}} + \frac{\overline{\sigma}m(c\ell_\delta)}{\bar{\lambda}\sqrt{\ell_1}N}, \qquad (A.12)\end{aligned}$$

where (a) follows from Eq. (16) and because $\|P\|_{1,2} \le \sqrt{N}$ and $\|\mathbf{1}\|_{1,2} = \sqrt{N}$. Then, the claim in Eq. (A.5) follows for $a = 1$ by using Eq. (A.12) and applying a union bound over $s \in \{0,1\}$. The proof of Eq. (A.5) for $a = 0$ follows similarly.

A2. Proof of Theorem 2: Asymptotic Normality for DR

For every $(i,j) \in [N]\times[M]$, recall the definitions of $\mathbb{T}_{i,j}^{(1,\mathrm{DR})}$ and $\mathbb{T}_{i,j}^{(0,\mathrm{DR})}$ from Eq. (A.3) and Eq. (A.4), respectively. Then, define

$$\begin{aligned}\mathbb{X}_{i,j}^{(1,\mathrm{DR})} &\triangleq \mathbb{T}_{i,j}^{(1,\mathrm{DR})} - \varepsilon_{i,j}^{(1)} - \frac{\varepsilon_{i,j}^{(1)}\eta_{i,j}}{p_{i,j}} \qquad (A.13)\\ \mathbb{X}_{i,j}^{(0,\mathrm{DR})} &\triangleq \mathbb{T}_{i,j}^{(0,\mathrm{DR})} + \varepsilon_{i,j}^{(0)} - \frac{\varepsilon_{i,j}^{(0)}\eta_{i,j}}{1-p_{i,j}},\end{aligned}$$

and

$$\mathbb{Z}_{i,j}^{\mathrm{DR}} \triangleq \varepsilon_{i,j}^{(1)} + \frac{\varepsilon_{i,j}^{(1)}\eta_{i,j}}{p_{i,j}} - \varepsilon_{i,j}^{(0)} + \frac{\varepsilon_{i,j}^{(0)}\eta_{i,j}}{1-p_{i,j}}. \tag{A.14}$$

Then, $\Delta\mathrm{ATE}_{\cdot,j}^{\mathrm{DR}}$ in Eq. (A.1) can be expressed as

$$\Delta\mathrm{ATE}_{\cdot,j}^{\mathrm{DR}} = \frac{1}{N}\sum_{i\in[N]}\Big(\mathbb{X}_{i,j}^{(1,\mathrm{DR})} + \mathbb{X}_{i,j}^{(0,\mathrm{DR})} + \mathbb{Z}_{i,j}^{\mathrm{DR}}\Big).$$

We obtain the following convergence results.

Lemma A1 (Convergence of $\mathbb{X}_j^{\mathrm{DR}}$)**.** *Fix any $j \in [M]$. Suppose Assumptions 1 to 4 and conditions (C1) to (C3) in Theorem 2 hold. Then,*

$$\frac{1}{\overline{\sigma}_j\sqrt{N}}\sum_{i\in[N]}\Big(\mathbb{X}_{i,j}^{(1,\mathrm{DR})} + \mathbb{X}_{i,j}^{(0,\mathrm{DR})}\Big) = o_p(1).$$

Lemma A2 (Convergence of $\mathbb{Z}_j^{\mathrm{DR}}$)**.** *Fix any $j \in [M]$. Suppose Assumptions 1 and 2 hold and condition (C3) in Theorem 2 hold. Then,*

$$\frac{1}{\overline{\sigma}_j\sqrt{N}}\sum_{i\in[N]}\mathbb{Z}_{i,j}^{\mathrm{DR}} \xrightarrow{d} \mathcal{N}(0,1).$$

Now, the result in Theorem 2 follows from Slutsky's theorem.

A2.1. Proof of Lemma A1

Fix any $j \in [M]$. Consider any $a \in \{0,1\}$. We claim that

$$\frac{1}{\sqrt{N}}\sum_{i\in[N]}\mathbb{X}_{i,j}^{(a,\mathrm{DR})} \le O\Big(\sqrt{N}\mathcal{E}\big(\widehat{\Theta}^{(a)}\big)\mathcal{E}\big(\widehat{P}\big)\Big) + o_p(1). \tag{A.15}$$

We provide a proof of this claim at the end of this section. Then, using Eq. (A.15) and the fact that $\overline{\sigma}_j \ge c > 0$ as per condition (C3), we obtain the following,

$$\begin{aligned}\frac{1}{\overline{\sigma}_j\sqrt{N}}\sum_{i\in[N]}\Big(\mathbb{X}_{i,j}^{(1,\mathrm{DR})} + \mathbb{X}_{i,j}^{(0,\mathrm{DR})}\Big) &\le \frac{1}{c}\Big(O\Big(\sqrt{N}\mathcal{E}\big(\widehat{\Theta}\big)\mathcal{E}\big(\widehat{P}\big)\Big) + o_p(1)\Big)\\ &\overset{(a)}{=} \frac{1}{c}\Big(\sqrt{N}o_p(N^{-1/2}) + o_p(1)\Big) \overset{(b)}{=} o_p(1),\end{aligned}$$

where (a) follows from (C2), and (b) follows because $o_p(1) + o_p(1) = o_p(1)$.

Proof of Eq. (A.15) Recall the partitioning of the units $[N]$ into $\mathcal{R}_0$ and $\mathcal{R}_1$ from Assumption 4. Now, to enable the application of concentration bounds, we split the

summation over $i \in [N]$ in the left hand side of Eq. (A.15) into two parts—one over $i \in \mathcal{R}_0$ and the other over $i \in \mathcal{R}_1$—such that the noise terms are independent of the estimates of $\Theta^{(0)}, \Theta^{(1)}, P$ in each of these parts as in Eqs. (14) and (15).

Fix $a = 1$. Then, Eqs. (A.3) and (A.13) imply that

$$\begin{aligned}\mathbb{X}_{i,j}^{(1,\mathrm{DR})} &= \frac{\big(\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)}\big)\big(\widehat{p}_{i,j} - p_{i,j}\big)}{\widehat{p}_{i,j}} - \frac{\big(\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)}\big)\eta_{i,j}}{\widehat{p}_{i,j}} + \frac{\varepsilon_{i,j}^{(1)} p_{i,j}}{\widehat{p}_{i,j}} + \frac{\varepsilon_{i,j}^{(1)}\eta_{i,j}}{\widehat{p}_{i,j}} - \varepsilon_{i,j}^{(1)} - \frac{\varepsilon_{i,j}^{(1)}\eta_{i,j}}{p_{i,j}} \\ &= \frac{\big(\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)}\big)\big(\widehat{p}_{i,j} - p_{i,j}\big)}{\widehat{p}_{i,j}} - \frac{\big(\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)}\big)\eta_{i,j}}{\widehat{p}_{i,j}} - \frac{\varepsilon_{i,j}^{(1)}\big(\widehat{p}_{i,j} - p_{i,j}\big)}{\widehat{p}_{i,j}} - \frac{\varepsilon_{i,j}^{(1)}\eta_{i,j}\big(\widehat{p}_{i,j} - p_{i,j}\big)}{\widehat{p}_{i,j} p_{i,j}}.\end{aligned}$$

Now, note that $|\sum_{i\in[N]} \mathbb{X}_{i,j}^{(1,\mathrm{DR})}| \le |\sum_{i\in\mathcal{R}_0} \mathbb{X}_{i,j}^{(1,\mathrm{DR})}| + |\sum_{i\in\mathcal{R}_1} \mathbb{X}_{i,j}^{(1,\mathrm{DR})}|$. Fix any $s \in \{0,1\}$. Then, triangle inequality implies that

$$\begin{aligned}\frac{1}{\sqrt{N}}\Big|\sum_{i\in\mathcal{R}_s} \mathbb{X}_{i,j}^{(1,\mathrm{DR})}\Big| &\le \frac{1}{\sqrt{N}}\Big|\sum_{i\in\mathcal{R}_s} \frac{\big(\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)}\big)\big(\widehat{p}_{i,j} - p_{i,j}\big)}{\widehat{p}_{i,j}}\Big| + \frac{1}{\sqrt{N}}\Big|\sum_{i\in\mathcal{R}_s} \frac{\big(\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)}\big)\eta_{i,j}}{\widehat{p}_{i,j}}\Big| \\ &\quad + \frac{1}{\sqrt{N}}\Big|\sum_{i\in\mathcal{R}_s} \frac{\varepsilon_{i,j}^{(1)}\big(\widehat{p}_{i,j} - p_{i,j}\big)}{\widehat{p}_{i,j}}\Big| + \frac{1}{\sqrt{N}}\Big|\sum_{i\in\mathcal{R}_s} \frac{\varepsilon_{i,j}^{(1)}\eta_{i,j}\big(\widehat{p}_{i,j} - p_{i,j}\big)}{\widehat{p}_{i,j} p_{i,j}}\Big|.\end{aligned} \tag{A.16}$$

To control the first term in Eq. (A.16), we use the Cauchy-Schwarz inequality and Assumption 3 as in Appendix A1 (see Eqs. (A.7), (A.11), and (A.12)).

To control the second term in Eq. (A.16), we condition on $\{(\widehat{p}_{i,j}, \widehat{\theta}_{i,j}^{(1)})\}_{i\in\mathcal{R}_s}$. Then, Assumption 4 (i.e., Eq. (14)) provides that $\{(\widehat{p}_{i,j}, \widehat{\theta}_{i,j}^{(1)})\}_{i\in\mathcal{R}_s} \perp\!\!\!\perp \{\eta_{i,j}\}_{i\in\mathcal{R}_s}$. As a result, $\sum_{i\in\mathcal{R}_s}\big(\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)}\big)\eta_{i,j}/\widehat{p}_{i,j}$ is $\mathrm{subGaussian}\big(\big[\sum_{i\in\mathcal{R}_s}\big(\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)}\big)^2/\big(\widehat{p}_{i,j}\big)^2\big]^{1/2}/\sqrt{\ell_1}\big)$ because $\eta_{i,j}$ is $\mathrm{subGaussian}(1/\sqrt{\ell_1})$ (see Example 2.5.8 in Vershynin (2018)) as well as zero-mean and independent across all $i \in [N]$ due to Assumption 2(a). Then, we have

$$\begin{aligned}\frac{1}{\sqrt{N}}\mathbb{E}\Big[\Big|\sum_{i\in\mathcal{R}_s}\frac{\big(\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)}\big)\eta_{i,j}}{\widehat{p}_{i,j}}\Big|\Big|\{(\widehat{p}_{i,j}, \widehat{\theta}_{i,j}^{(1)})\}_{i\in\mathcal{R}_s}\Big] &\overset{(a)}{\le} \frac{c}{\sqrt{N}}\sqrt{\sum_{i\in\mathcal{R}_s}\left(\frac{\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)}}{\widehat{p}_{i,j}}\right)^2} \\ &\le \frac{c}{\sqrt{N}}\big\|\big(\widehat{\Theta}_{\cdot,j}^{(1)} - \Theta_{\cdot,j}^{(1)}\big) \oslash \widehat{P}_{\cdot,j}\big\|_2 \\ &\overset{(b)}{\le} \frac{c}{\bar{\lambda}}\mathcal{E}\big(\widehat{\Theta}^{(1)}\big) \le \frac{c}{\bar{\lambda}}\mathcal{E}\big(\widehat{\Theta}\big) \overset{(c)}{=} o_p(1),\end{aligned} \tag{A.17}$$

where (a) follows as the first moment of $\mathrm{subGaussian}(\sigma)$ is $O(\sigma)$, (b) follows from Assumption 3 and Eq. (16), and (c) follows from (C1).

To control the third term in Eq. (A.16), we condition on $\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s}$. Then, Assumption 4 (i.e., Eq. (15)) provides that $\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s} \perp\!\!\!\perp \{\varepsilon_{i,j}^{(1)}\}_{i\in\mathcal{R}_s}$. As a result,

$\sum_{i\in\mathcal{R}_s}\varepsilon_{i,j}^{(1)}\big(\widehat{p}_{i,j}-p_{i,j}\big)/\widehat{p}_{i,j}$ is subGaussian$\big(\overline{\sigma}\big[\sum_{i\in\mathcal{R}_s}\big(\widehat{p}_{i,j}-p_{i,j}\big)^2/\big(\widehat{p}_{i,j}\big)^2\big]^{1/2}\big)$ because $\varepsilon_{i,j}^{(1)}$ is subGaussian$(\overline{\sigma})$, zero-mean, and independent across all $i\in[N]$ due to Assumption 2. Then, we have

$$\begin{aligned}\frac{1}{\sqrt{N}}\mathbb{E}\bigg[\Big|\sum_{i\in\mathcal{R}_s}\frac{\varepsilon_{i,j}^{(1)}\big(\widehat{p}_{i,j}-p_{i,j}\big)}{\widehat{p}_{i,j}}\Big|\Big|\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s}\bigg] &\overset{(a)}{\leq} \frac{c\overline{\sigma}}{\sqrt{N}}\sqrt{\sum_{i\in\mathcal{R}_s}\bigg(\frac{\widehat{p}_{i,j}-p_{i,j}}{\widehat{p}_{i,j}}\bigg)^2}\\ &\leq \frac{c\overline{\sigma}}{\sqrt{N}}\big\|\big(\widehat{P}_{\cdot,j}-P_{\cdot,j}\big)\oslash\widehat{P}_{\cdot,j}\big\|_2\\ &\overset{(b)}{\leq}\frac{c\overline{\sigma}}{\underline{\lambda}}\mathcal{E}\big(\widehat{P}\big)\overset{(c)}{=}o_p(1),\end{aligned}\tag{A.18}$$

where (a) follows as the first moment of subGaussian(σ) is $O(\sigma)$, (b) follows from Assumption 3 and Eq. (16), and (c) follows from (C1).

To control the fourth term in Eq. (A.16), we condition on $\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s}$. Then, Assumption 4 (i.e., Eq. (15)) provides that $\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s}\perp\!\!\!\perp\{(\eta_{i,j},\varepsilon_{i,j}^{(1)})\}_{i\in\mathcal{R}_s}$. As a result, $\sum_{i\in\mathcal{R}_s}\varepsilon_{i,j}^{(1)}\eta_{i,j}\big(\widehat{p}_{i,j}-p_{i,j}\big)/\widehat{p}_{i,j}p_{i,j}$ is subExponential$\big(\overline{\sigma}\big[\sum_{i\in\mathcal{R}_s}\big(\widehat{p}_{i,j}-p_{i,j}\big)^2/\big(\widehat{p}_{i,j}p_{i,j}\big)^2\big]^{1/2}/\sqrt{\ell_1}\big)$ because $\varepsilon_{i,j}^{(1)}\eta_{i,j}$ is subExponential$(\overline{\sigma}/\sqrt{\ell_1})$ due to Lemma S3 as well as zero-mean and independent across all $i\in[N]$ due to Assumption 2. Then, we have

$$\begin{aligned}\frac{1}{\sqrt{N}}\mathbb{E}\bigg[\Big|\sum_{i\in\mathcal{R}_s}\frac{\varepsilon_{i,j}^{(1)}\eta_{i,j}\big(\widehat{p}_{i,j}-p_{i,j}\big)}{\widehat{p}_{i,j}p_{i,j}}\Big|\Big|\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s}\bigg] &\overset{(a)}{\leq} \frac{c\overline{\sigma}}{\sqrt{N}}\sqrt{\sum_{i\in\mathcal{R}_s}\bigg(\frac{\widehat{p}_{i,j}-p_{i,j}}{\widehat{p}_{i,j}p_{i,j}}\bigg)^2}\\ &\leq \frac{c\overline{\sigma}}{\sqrt{N}}\big\|\big(\widehat{P}_{\cdot,j}-P_{\cdot,j}\big)\oslash\big(\widehat{P}_{\cdot,j}\odot P_{\cdot,j}\big)\big\|_2\\ &\overset{(b)}{\leq}\frac{c\overline{\sigma}}{\underline{\lambda}\lambda}\mathcal{E}\big(\widehat{P}\big)\overset{(c)}{=}o_p(1),\end{aligned}\tag{A.19}$$

where (a) follows as the first moment of subExponential(σ) is $O(\sigma)$, (b) follows from Assumption 3 and Eq. (16), and (c) follows from (C1).

Putting together Eqs. (A.16) to (A.19) using Lemma S6, we have

$$\frac{1}{\sqrt{N}}\Big|\sum_{i\in\mathcal{R}_s}\mathbb{X}_{i,j}^{(1,\mathrm{DR})}\Big|\leq O\Big(\sqrt{N}\mathcal{E}\big(\widehat{\Theta}^{(1)}\big)\mathcal{E}\big(\widehat{P}\big)\Big)+o_p(1).$$

Then, the claim in Eq. (A.15) follows for $a=1$ by using $|\sum_{i\in[N]}\mathbb{X}_{i,j}^{(1,\mathrm{DR})}|\leq|\sum_{i\in\mathcal{R}_0}\mathbb{X}_{i,j}^{(1,\mathrm{DR})}|+|\sum_{i\in\mathcal{R}_1}\mathbb{X}_{i,j}^{(1,\mathrm{DR})}|$. The proof of Eq. (A.15) for $a=0$ follows similarly.

A2.2. Proof of Lemma A2

To prove this result, we invoke Lyapunov central limit theorem (CLT).

Lemma A3 (Lyapunov CLT, see Theorem 27.3 of Billingsley (2017))**.** *Consider a sequence $x_1, x_2, \cdots$ of mean-zero independent random variables such that the moments $\mathbb{E}[|x_i|^{2+\omega}]$ are finite for some $\omega > 0$. Moreover, assume that the Lyapunov's condition is satisfied, i.e.,*

$$\sum_{i=1}^{N} \mathbb{E}[|x_i|^{2+\omega}] \Big/ \Big(\sum_{i=1}^{N} \mathbb{E}[x_i^2] \Big)^{\frac{2+\omega}{2}} \longrightarrow 0, \tag{A.20}$$

as $N \to \infty$. Then,

$$\sum_{i=1}^{N} x_i \Big/ \Big(\sum_{i=1}^{N} \mathbb{E}[x_i^2] \Big)^{\frac{1}{2}} \xrightarrow{d} \mathcal{N}(0,1),$$

as $N \to \infty$.

Fix any $j \in [M]$. We apply Lyapunov CLT in Lemma A3 on the sequence $\mathbb{Z}_{1,j}^{\mathrm{DR}}, \mathbb{Z}_{2,j}^{\mathrm{DR}}, \cdots$ where $\mathbb{Z}_{i,j}^{\mathrm{DR}}$ is as defined in Eq. (A.14). Note that this sequence is zero-mean from Assumption 2(a) and Assumption 2(b), and independent from Assumption 2(b). First, we show in Appendix A2.2.1 that

$$\mathbb{V}\mathrm{ar}(\mathbb{Z}_{i,j}^{\mathrm{DR}}) = \frac{(\sigma_{i,j}^{(1)})^2}{p_{i,j}} + \frac{(\sigma_{i,j}^{(0)})^2}{1-p_{i,j}}, \tag{A.21}$$

for each $i \in [N]$. Next, we show in Appendix A2.2.2 that Lyapunov's condition Eq. (A.20) holds for the sequence $\mathbb{Z}_{1,j}^{\mathrm{DR}}, \mathbb{Z}_{2,j}^{\mathrm{DR}}, \cdots$ with $\omega = 1$. Finally, applying Lemma A3 and using the definition of $\overline{\sigma}_j$ from Eq. (22) yields Lemma A2.

A2.2.1. Proof of Eq. (A.21)
Fix any $i \in [N]$ and consider $\mathbb{V}\mathrm{ar}(\mathbb{Z}_{i,j}^{\mathrm{DR}})$. We have

$$\mathbb{V}\mathrm{ar}\Big(\mathbb{Z}_{i,j}^{\mathrm{DR}}\Big) = \mathbb{V}\mathrm{ar}\left(\varepsilon_{i,j}^{(1)}\Big(1 + \frac{\eta_{i,j}}{p_{i,j}}\Big) - \varepsilon_{i,j}^{(0)}\Big(1 - \frac{\eta_{i,j}}{1-p_{i,j}}\Big)\right). \tag{A.22}$$

We claim the following:

$$\mathbb{V}\mathrm{ar}\left(\varepsilon_{i,j}^{(1)}\Big(1 + \frac{\eta_{i,j}}{p_{i,j}}\Big)\right) = \frac{(\sigma_{i,j}^{(1)})^2}{p_{i,j}}, \tag{A.23}$$

$$\mathbb{V}\mathrm{ar}\left(\varepsilon_{i,j}^{(0)}\Big(1 - \frac{\eta_{i,j}}{1-p_{i,j}}\Big)\right) = \frac{(\sigma_{i,j}^{(0)})^2}{1-p_{i,j}}, \tag{A.24}$$

and

$$\mathbb{C}\mathrm{ov}\left(\varepsilon_{i,j}^{(1)}\Big(1 + \frac{\eta_{i,j}}{p_{i,j}}\Big), \varepsilon_{i,j}^{(0)}\Big(1 - \frac{\eta_{i,j}}{1-p_{i,j}}\Big)\right) = 0, \tag{A.25}$$

with Eq. (A.21) following from Eqs. (A.22) to (A.25).

To establish Eq. (A.23), notice that Assumption 2(a) and (b) imply $\varepsilon_{i,j}^{(1)} \perp\!\!\!\perp \eta_{i,j}$ and $\mathbb{E}[\varepsilon_{i,j}^{(1)}] = \mathbb{E}[\eta_{i,j}] = 0$ so that $\mathbb{E}[\varepsilon_{i,j}^{(1)}(1 + \eta_{i,j}/p_{i,j})] = 0$. Then,

$$\begin{aligned}
\mathbb{V}\mathrm{ar}\left(\varepsilon_{i,j}^{(1)}\left(1+\frac{\eta_{i,j}}{p_{i,j}}\right)\right) &= \mathbb{E}\left[\left(\varepsilon_{i,j}^{(1)}\left(1+\frac{\eta_{i,j}}{p_{i,j}}\right)\right)^2\right] = \mathbb{E}\left[\left(\varepsilon_{i,j}^{(1)}\right)^2\right]\mathbb{E}\left[\left(1+\frac{\eta_{i,j}}{p_{i,j}}\right)^2\right] \\
&= \mathbb{E}\left[\left(\varepsilon_{i,j}^{(1)}\right)^2\right]\left[1+\mathbb{E}\left[\frac{\eta_{i,j}^2}{p_{i,j}^2}\right]\right] \overset{(a)}{=} (\sigma_{i,j}^{(1)})^2\left[1+\frac{p_{i,j}(1-p_{i,j})}{p_{i,j}^2}\right] \\
&= \frac{(\sigma_{i,j}^{(1)})^2}{p_{i,j}},
\end{aligned}$$

where (a) follows because $\mathbb{E}[\eta_{i,j}^2] = \mathbb{V}\mathrm{ar}(\eta_{i,j}) = p_{i,j}(1-p_{i,j})$ from Eq. (3), and $\mathbb{E}\big[(\varepsilon_{i,j}^{(1)})^2\big] = \mathbb{V}\mathrm{ar}(\varepsilon_{i,j}^{(1)}) = (\sigma_{i,j}^{(1)})^2$ from condition (C3). A similar argument establishes Eq. (A.24). Eq. (A.25) follows from,

$$\begin{aligned}
\mathbb{C}\mathrm{ov}\left(\varepsilon_{i,j}^{(1)}\left(1+\frac{\eta_{i,j}}{p_{i,j}}\right), \varepsilon_{i,j}^{(0)}\left(1-\frac{\eta_{i,j}}{1-p_{i,j}}\right)\right) &= \mathbb{E}\left[\varepsilon_{i,j}^{(1)}\left(1+\frac{\eta_{i,j}}{p_{i,j}}\right)\varepsilon_{i,j}^{(0)}\left(1-\frac{\eta_{i,j}}{1-p_{i,j}}\right)\right] \\
&\overset{(a)}{=} \mathbb{E}\left[\left(1+\frac{\eta_{i,j}}{p_{i,j}}\right)\left(1-\frac{\eta_{i,j}}{1-p_{i,j}}\right)\right]\mathbb{E}[\varepsilon_{i,j}^{(1)}\varepsilon_{i,j}^{(0)}] \\
&= \left(1-\mathbb{E}\left[\frac{\eta_{i,j}^2}{p_{i,j}(1-p_{i,j})}\right]\right)\mathbb{E}[\varepsilon_{i,j}^{(1)}\varepsilon_{i,j}^{(0)}] \\
&\overset{(b)}{=} 0 \cdot \mathbb{E}[\varepsilon_{i,j}^{(1)}\varepsilon_{i,j}^{(0)}] = 0,
\end{aligned}$$

where (a) follows because $(\varepsilon_{i,j}^{(0)}, \varepsilon_{i,j}^{(1)}) \perp\!\!\!\perp \eta_{i,j}$ from Assumption 2(b) and (b) follows because $\mathbb{E}[\eta_{i,j}^2] = \mathbb{V}\mathrm{ar}(\eta_{i,j}) = p_{i,j}(1-p_{i,j})$.

A2.2.2. Proof of Lyapunov's condition with $\omega = 1$

We have

$$\begin{aligned}
\frac{\sum_{i\in[N]}\mathbb{E}\big[|\mathbb{Z}_{i,j}^{\mathrm{DR}}|^3\big]}{\big(\sum_{i\in[N]}\mathbb{V}\mathrm{ar}(\mathbb{Z}_{i,j}^{\mathrm{DR}})\big)^{3/2}} &= \frac{1}{N^{3/2}}\frac{\sum_{i\in[N]}\mathbb{E}\big[|\mathbb{Z}_{i,j}^{\mathrm{DR}}|^3\big]}{\big(\frac{1}{N}\sum_{i\in[N]}\mathbb{V}\mathrm{ar}(\mathbb{Z}_{i,j}^{\mathrm{DR}})\big)^{3/2}} \\
&\overset{(a)}{=} \frac{1}{N^{3/2}}\frac{\sum_{i\in[N]}\mathbb{E}\big[|\mathbb{Z}_{i,j}^{\mathrm{DR}}|^3\big]}{\big(\overline{\sigma}_j\big)^{3/2}} \\
&\overset{(b)}{\leq} \frac{1}{N^{3/2}}\frac{\sum_{i\in[N]}\mathbb{E}\big[|\mathbb{Z}_{i,j}^{\mathrm{DR}}|^3\big]}{c_1^{3/2}} \overset{(c)}{\leq} \frac{1}{N^{1/2}}\frac{c_2}{c_1^{3/2}},
\end{aligned} \tag{A.26}$$

where (a) follows by putting together Eqs. (A.21) and (22), (b) follows because $\overline{\sigma}_j \geq c_1 > 0$ as per condition (C3), (c) follows because the absolute third moments of subExponential random variables are bounded, after noting that $\mathbb{Z}_{i,j}^{\mathrm{DR}}$ is a

subExponential random variable. Then, condition Eq. (A.20) holds for $\omega = 1$ as the right hand side of Eq. (A.26) goes to zero as $N \to \infty$.

A2.3. Proof of Proposition 2: Consistent variance estimation

Fix any $j \in [M]$ and recall the definitions of $\overline{\sigma}_j^2$ and $\widehat{\sigma}_j^2$ from Eqs. (22) and (25), respectively. The error $\Delta_j = \widehat{\sigma}_j^2 - \overline{\sigma}_j^2$ can be expressed as

$$\begin{aligned}
\Delta_j &= \frac{1}{N}\sum_{i\in[N]}\left(\frac{\big(\widehat{\theta}_{i,j}^{(1)} - y_{i,j}\big)^2 a_{i,j}}{\big(\widehat{p}_{i,j}\big)^2} + \frac{\big(\widehat{\theta}_{i,j}^{(0)} - y_{i,j}\big)^2 (1-a_{i,j})}{\big(1-\widehat{p}_{i,j}\big)^2}\right) - \left(\frac{(\sigma_{i,j}^{(1)})^2}{p_{i,j}} + \frac{(\sigma_{i,j}^{(0)})^2}{1-p_{i,j}}\right) \\
&= \frac{1}{N}\sum_{i\in[N]}\left(\frac{\big(\widehat{\theta}_{i,j}^{(1)} - y_{i,j}\big)^2 a_{i,j}}{\big(\widehat{p}_{i,j}\big)^2} - \frac{(\sigma_{i,j}^{(1)})^2}{p_{i,j}}\right) + \left(\frac{\big(\widehat{\theta}_{i,j}^{(0)} - y_{i,j}\big)^2 (1-a_{i,j})}{\big(1-\widehat{p}_{i,j}\big)^2} - \frac{(\sigma_{i,j}^{(0)})^2}{1-p_{i,j}}\right) \\
&\overset{(a)}{=} \frac{1}{N}\sum_{i\in[N]}\Big(\mathbb{T}_{i,j}^{(1)} + \mathbb{T}_{i,j}^{(0)}\Big),
\end{aligned} \tag{A.27}$$

where (a) follows after defining

$$\mathbb{T}_{i,j}^{(1)} \triangleq \frac{\big(\widehat{\theta}_{i,j}^{(1)} - y_{i,j}\big)^2 a_{i,j}}{\big(\widehat{p}_{i,j}\big)^2} - \frac{(\sigma_{i,j}^{(1)})^2}{p_{i,j}} \quad \text{and} \quad \mathbb{T}_{i,j}^{(0)} \triangleq \frac{\big(\widehat{\theta}_{i,j}^{(0)} - y_{i,j}\big)^2 (1-a_{i,j})}{\big(1-\widehat{p}_{i,j}\big)^2} - \frac{(\sigma_{i,j}^{(0)})^2}{1-p_{i,j}}.$$

for every $(i,j) \in [N] \times [M]$. Then, we have

$$\begin{aligned}
\mathbb{T}_{i,j}^{(1)} &\overset{(a)}{=} \frac{\big(\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)} - \varepsilon_{i,j}^{(1)}\big)^2 \big(p_{i,j} + \eta_{i,j}\big)}{\big(\widehat{p}_{i,j}\big)^2} - \frac{(\sigma_{i,j}^{(1)})^2}{p_{i,j}} \\
&= \frac{\big(\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)}\big)^2 a_{i,j}}{\big(\widehat{p}_{i,j}\big)^2} - \frac{2\varepsilon_{i,j}^{(1)} p_{i,j}\big(\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)}\big)}{\big(\widehat{p}_{i,j}\big)^2} - \frac{2\varepsilon_{i,j}^{(1)}\eta_{i,j}\big(\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)}\big)}{\big(\widehat{p}_{i,j}\big)^2} \\
&\quad + \frac{\big(\varepsilon_{i,j}^{(1)}\big)^2 p_{i,j}}{\big(\widehat{p}_{i,j}\big)^2} + \frac{\big(\varepsilon_{i,j}^{(1)}\big)^2 \eta_{i,j}}{\big(\widehat{p}_{i,j}\big)^2} - \frac{(\sigma_{i,j}^{(1)})^2}{p_{i,j}},
\end{aligned}$$

where (a) follows from Eqs. (1) to (3). A similar derivation for $a = 0$ implies that

$$\begin{aligned}
\mathbb{T}_{i,j}^{(0)} &= \frac{\big(\widehat{\theta}_{i,j}^{(0)} - \theta_{i,j}^{(0)} - \varepsilon_{i,j}^{(0)}\big)^2 (1 - p_{i,j} - \eta_{i,j})}{\big(1-\widehat{p}_{i,j}\big)^2} - \frac{(\sigma_{i,j}^{(0)})^2}{1-p_{i,j}} \\
&= \frac{\big(\widehat{\theta}_{i,j}^{(0)} - \theta_{i,j}^{(0)}\big)^2 \big(1-a_{i,j}\big)}{\big(1-\widehat{p}_{i,j}\big)^2} - \frac{2\varepsilon_{i,j}^{(0)}\big(1-p_{i,j}\big)\big(\widehat{\theta}_{i,j}^{(0)} - \theta_{i,j}^{(0)}\big)}{\big(1-\widehat{p}_{i,j}\big)^2} + \frac{2\varepsilon_{i,j}^{(0)}\eta_{i,j}\big(\widehat{\theta}_{i,j}^{(0)} - \theta_{i,j}^{(0)}\big)}{\big(1-\widehat{p}_{i,j}\big)^2} \\
&\quad + \frac{\big(\varepsilon_{i,j}^{(0)}\big)^2\big(1-p_{i,j}\big)}{\big(1-\widehat{p}_{i,j}\big)^2} - \frac{\big(\varepsilon_{i,j}^{(0)}\big)^2 \eta_{i,j}}{\big(1-\widehat{p}_{i,j}\big)^2} - \frac{(\sigma_{i,j}^{(0)})^2}{1-p_{i,j}}.
\end{aligned}$$

Consider any $a \in \{0,1\}$. We claim that

$$\frac{1}{N}\Big|\sum_{i\in[N]} \mathbb{T}^{(\mathrm{a})}_{i,j}\Big| = o_p(1). \tag{A.28}$$

We provide a proof of this claim at the end of this section. Then, applying triangle inequality in Eq. (A.27), we obtain the following

$$\Delta_j \leq o_p(1) + o_p(1) \overset{(a)}{=} o_p(1),$$

where (a) follows because $o_p(1) + o_p(1) = o_p(1)$.

Proof of bound Eq. (A.28). This proof follows a very similar road map to that used for establishing the inequality in Eq. (A.15). Recall the partitioning of the units $[N]$ into $\mathcal{R}_0$ and $\mathcal{R}_1$ from Assumption 4. Now, to enable the application of concentration bounds, we split the summation over $i \in [N]$ in the left hand side of Eq. (A.28) into two parts—one over $i \in \mathcal{R}_0$ and the other over $i \in \mathcal{R}_1$—such that the noise terms are independent of the estimates of $\Theta^{(0)}, \Theta^{(1)}, P$ in each of these parts as in Eqs. (14) and (15).

Fix $a = 1$. Now, note that $|\sum_{i\in[N]} \mathbb{T}^{(1)}_{i,j}| \leq |\sum_{i\in\mathcal{R}_0} \mathbb{T}^{(1)}_{i,j}| + |\sum_{i\in\mathcal{R}_1} \mathbb{T}^{(1)}_{i,j}|$. Fix any $s \in \{0,1\}$. Then, triangle inequality implies that

$$\begin{aligned}\frac{1}{N}\Big|\sum_{i\in\mathcal{R}_s} \mathbb{T}^{(1)}_{i,j}\Big| \leq{}& \frac{1}{N}\Big|\sum_{i\in\mathcal{R}_s} \frac{\big(\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}\big)^2 a_{i,j}}{\big(\widehat{p}_{i,j}\big)^2}\Big| + \frac{1}{N}\Big|\sum_{i\in\mathcal{R}_s} \frac{2\varepsilon^{(1)}_{i,j} p_{i,j}\big(\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}\big)}{\big(\widehat{p}_{i,j}\big)^2}\Big| \\ &+ \frac{1}{N}\Big|\sum_{i\in\mathcal{R}_s} \frac{2\varepsilon^{(1)}_{i,j}\eta_{i,j}\big(\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}\big)}{\big(\widehat{p}_{i,j}\big)^2}\Big| + \frac{1}{N}\Big|\sum_{i\in\mathcal{R}_s} \frac{\big(\varepsilon^{(1)}_{i,j}\big)^2\eta_{i,j}}{\big(\widehat{p}_{i,j}\big)^2}\Big| + \frac{1}{N}\Big|\sum_{i\in\mathcal{R}_s} \frac{\big(\varepsilon^{(1)}_{i,j}\big)^2 p_{i,j}}{\big(\widehat{p}_{i,j}\big)^2} - \frac{(\sigma^{(1)}_{i,j})^2}{p_{i,j}}\Big|.\end{aligned} \tag{A.29}$$

To bound the first term in Eq. (A.29), we have

$$\begin{aligned}\frac{1}{N}\Big|\sum_{i\in\mathcal{R}_s} \frac{\big(\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}\big)^2 a_{i,j}}{\big(\widehat{p}_{i,j}\big)^2}\Big| &\overset{(a)}{\leq} \frac{1}{N}\Big|\sum_{i\in\mathcal{R}_s} \frac{\big(\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}\big)^2}{\big(\widehat{p}_{i,j}\big)^2}\Big| \\ &\overset{(b)}{\leq} \frac{1}{\overline{\lambda}^2 N}\big\|\widehat{\Theta}^{(1)}_{\cdot,j} - \Theta^{(1)}_{\cdot,j}\big\|_2^2 \\ &\overset{(c)}{=} \frac{1}{\overline{\lambda}^2}\Big[\mathcal{E}\big(\widehat{\Theta}^{(1)}\big)\Big]^2 \leq \frac{1}{\overline{\lambda}^2}\Big[\mathcal{E}\big(\widehat{\Theta}\big)\Big]^2 \overset{(d)}{=} o_p(1)o_p(1) \overset{(e)}{=} o_p(1),\end{aligned} \tag{A.30}$$

where (a) follows as $a_{i,j} \in \{0,1\}$, (b) follows from Assumption 3, (c) follows from Eq. (16), (d) follows from (C1), and (e) follows because $o_p(1)o_p(1) = o_p(1)$.

To control second term in Eq. (A.29), we condition on $\{\big(\widehat{p}_{i,j}, \widehat{\theta}^{(1)}_{i,j}\big)\}_{i\in\mathcal{R}_s}$. Then,

Eq. (24) provides that $\{(\widehat{p}_{i,j}, \widehat{\theta}^{(1)}_{i,j})\}_{i\in\mathcal{R}_s} \perp\!\!\!\perp \{\varepsilon^{(1)}_{i,j}\}_{i\in\mathcal{R}_s}$. As a result, $\sum_{i\in\mathcal{R}_s}\varepsilon^{(1)}_{i,j}p_{i,j}\big(\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}\big)/\big(\widehat{p}_{i,j}\big)^2$ is subGaussian$\big(\overline{\sigma}\big[\sum_{i\in\mathcal{R}_s}\big(p_{i,j}\big)^2\big(\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}\big)^2/\big(\widehat{p}_{i,j}\big)^4\big]^{1/2}\big)$ because $\varepsilon^{(1)}_{i,j}$ is subGaussian$(\overline{\sigma})$, zero-mean and independent across all $i \in [N]$ due to Assumption 2. Then, we have

$$\begin{aligned}
&\frac{1}{N}\mathbb{E}\bigg[\bigg|\sum_{i\in\mathcal{R}_s}\frac{2\varepsilon^{(1)}_{i,j}p_{i,j}\big(\widehat{\theta}^{(1)}_{i,j}-\theta^{(1)}_{i,j}\big)}{\big(\widehat{p}_{i,j}\big)^2}\bigg|\bigg|\{(\widehat{p}_{i,j},\widehat{\theta}^{(1)}_{i,j})\}_{i\in\mathcal{R}_s}\bigg]\\
&\overset{(a)}{\leq}\frac{c\overline{\sigma}}{N}\sqrt{\sum_{i\in\mathcal{R}_s}\bigg(\frac{p_{i,j}\big(\widehat{\theta}^{(1)}_{i,j}-\theta^{(1)}_{i,j}\big)}{\big(\widehat{p}_{i,j}\big)^2}\bigg)^2}\\
&\overset{(b)}{\leq}\frac{c\overline{\sigma}}{\overline{\lambda}^2N}\big\|\widehat{\Theta}^{(1)}_{\cdot,j}-\Theta^{(1)}_{\cdot,j}\big\|_2\overset{(c)}{=}\frac{c\overline{\sigma}}{\overline{\lambda}^2}\frac{\mathcal{E}\big(\widehat{\Theta}^{(1)}\big)}{\sqrt{N}}\leq\frac{c\overline{\sigma}}{\overline{\lambda}^2}\frac{\mathcal{E}\big(\widehat{\Theta}\big)}{\sqrt{N}}\overset{(d)}{=}o_p(1),
\end{aligned}\tag{A.31}$$

where (a) follows as the first moment of subGaussian(σ) is $O(\sigma)$, (b) follows from Assumptions 1 and 3, (c) follows from Eq. (16), and (d) follows from (C1).

To control third term in Eq. (A.29), we condition on $\{(\widehat{p}_{i,j}, \widehat{\theta}^{(1)}_{i,j})\}_{i\in\mathcal{R}_s}$. Then, Eq. (24) provides that $\{(\widehat{p}_{i,j}, \widehat{\theta}^{(1)}_{i,j})\}_{i\in\mathcal{R}_s} \perp\!\!\!\perp \{(\eta_{i,j}, \varepsilon^{(1)}_{i,j})\}_{i\in\mathcal{R}_s}$. As a result, $\sum_{i\in\mathcal{R}_s}\varepsilon^{(1)}_{i,j}\eta_{i,j}\big(\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}\big)/\big(\widehat{p}_{i,j}\big)^2$ is subExponential$\big(\overline{\sigma}\big[\sum_{i\in\mathcal{R}_s}\big(\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}\big)^2/\big(\widehat{p}_{i,j}\big)^4\big]^{1/2}/\sqrt{\ell_1}\big)$ because $\varepsilon^{(1)}_{i,j}\eta_{i,j}$ is subExponential$(\overline{\sigma}/\sqrt{\ell_1})$ due to Lemma S3 as well as zero-mean and independent across all $i \in [N]$ due to Assumption 2. Then, we have

$$\begin{aligned}
&\frac{1}{N}\mathbb{E}\bigg[\bigg|\sum_{i\in\mathcal{R}_s}\frac{2\varepsilon^{(1)}_{i,j}\eta_{i,j}\big(\widehat{\theta}^{(1)}_{i,j}-\theta^{(1)}_{i,j}\big)}{\big(\widehat{p}_{i,j}\big)^2}\bigg|\bigg|\{(\widehat{p}_{i,j},\widehat{\theta}^{(1)}_{i,j})\}_{i\in\mathcal{R}_s}\bigg]\\
&\overset{(a)}{\leq}\frac{c\overline{\sigma}}{N}\sqrt{\sum_{i\in\mathcal{R}_s}\bigg(\frac{\widehat{\theta}^{(1)}_{i,j}-\theta^{(1)}_{i,j}}{\big(\widehat{p}_{i,j}\big)^2}\bigg)^2}\\
&\overset{(b)}{\leq}\frac{c\overline{\sigma}}{\overline{\lambda}^2N}\big\|\widehat{\Theta}^{(1)}_{\cdot,j}-\Theta^{(1)}_{\cdot,j}\big\|_2\overset{(c)}{=}\frac{c\overline{\sigma}}{\overline{\lambda}^2}\frac{\mathcal{E}\big(\widehat{\Theta}^{(1)}\big)}{\sqrt{N}}\leq\frac{c\overline{\sigma}}{\overline{\lambda}^2}\frac{\mathcal{E}\big(\widehat{\Theta}\big)}{\sqrt{N}}\overset{(d)}{=}o_p(1),
\end{aligned}\tag{A.32}$$

where (a) follows as the first moment of subExponential(σ) is $O(\sigma)$ (Zhang and Wei, 2022, Corollary 3), (b) follows from Assumption 3, (c) follows from Eq. (16), and (d) follows from (C1).

To control fourth term in Eq. (A.29), we condition on $\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s}$. Then, Eq. (24) provides that $\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s} \perp\!\!\!\perp \{(\eta_{i,j}, \varepsilon^{(1)}_{i,j})\}_{i\in\mathcal{R}_s}$. As a result, $\sum_{i\in\mathcal{R}_s}\big(\varepsilon^{(1)}_{i,j}\big)^2\eta_{i,j}/\big(\widehat{p}_{i,j}\big)^2$ is subWeibull$_{2/3}\big(\overline{\sigma}^2\big[\sum_{i\in\mathcal{R}_s}1/\big(\widehat{p}_{i,j}\big)^4\big]^{1/2}/\sqrt{\ell_1}\big)$ because $(\varepsilon^{(1)}_{i,j})^2\eta_{i,j}$ is subWeibull$_{2/3}(\overline{\sigma}^2/\sqrt{\ell_1})$ due to Lemma S4 as well as zero-mean and independent across all $i \in [N]$ due to

Assumption 2. Then, we have

$$\frac{1}{N}\mathbb{E}\Big[\Big|\sum_{i\in\mathcal{R}_s}\frac{\big(\varepsilon_{i,j}^{(1)}\big)^2\eta_{i,j}}{\big(\widehat{p}_{i,j}\big)^2}\Big|\Big|\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s}\Big]\overset{(a)}{\leq}\frac{c\overline{\sigma}^2}{N}\sqrt{\sum_{i\in\mathcal{R}_s}\frac{1}{\big(\widehat{p}_{i,j}\big)^4}}\overset{(b)}{\leq}\frac{c\overline{\sigma}^2}{\bar{\lambda}^2\sqrt{N}}=o_p(1),\quad(\text{A.33})$$

where (a) follows as the first moment of $\text{subWeibull}_{2/3}(\sigma)$ is $O(\sigma)$ (Zhang and Wei, 2022, Corollary 3) and (b) follows from Assumption 3.

To control fifth term in Eq. (A.29), we have

$$\begin{aligned}\Big|\sum_{i\in\mathcal{R}_s}\Big(\frac{\big(\varepsilon_{i,j}^{(1)}\big)^2p_{i,j}}{\big(\widehat{p}_{i,j}\big)^2}-\frac{(\sigma_{i,j}^{(1)})^2}{p_{i,j}}\Big)\Big|&=\Big|\sum_{i\in\mathcal{R}_s}\Big(\frac{\big(\varepsilon_{i,j}^{(1)}\big)^2p_{i,j}}{\big(\widehat{p}_{i,j}\big)^2}-\frac{\big(\sigma_{i,j}^{(1)}\big)^2p_{i,j}}{\big(\widehat{p}_{i,j}\big)^2}+\frac{\big(\sigma_{i,j}^{(1)}\big)^2p_{i,j}}{\big(\widehat{p}_{i,j}\big)^2}-\frac{(\sigma_{i,j}^{(1)})^2}{p_{i,j}}\Big)\Big|\\&\overset{(a)}{\leq}\Big|\sum_{i\in\mathcal{R}_s}\Big(\frac{\big[\big(\varepsilon_{i,j}^{(1)}\big)^2-\big(\sigma_{i,j}^{(1)}\big)^2\big]p_{i,j}}{\big(\widehat{p}_{i,j}\big)^2}\Big)\Big|+\Big|\sum_{i\in\mathcal{R}_s}\Big(\frac{\big(\sigma_{i,j}^{(1)}\big)^2p_{i,j}}{\big(\widehat{p}_{i,j}\big)^2}-\frac{(\sigma_{i,j}^{(1)})^2}{p_{i,j}}\Big)\Big|,\end{aligned}\quad(\text{A.34})$$

where (a) follows from the triangle inequality. To control the first term in Eq. (A.34), we condition on $\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s}$. Then, Eq. (24) provides that $\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s}\perp\!\!\!\perp\{\varepsilon_{i,j}^{(1)}\}_{i\in\mathcal{R}_s}$. Further, $\mathbb{E}[(\varepsilon_{i,j}^{(1)})^2-(\sigma_{i,j}^{(1)})^2]=0$ due to (C3) and Assumption 2. As a result, $\sum_{i\in\mathcal{R}_s}\big[\big(\varepsilon_{i,j}^{(1)}\big)^2-\big(\sigma_{i,j}^{(1)}\big)^2\big]p_{i,j}/\big(\widehat{p}_{i,j}\big)^2$ is $\text{subExponential}\big(\overline{\sigma}^2\big[\sum_{i\in\mathcal{R}_s}\big(p_{i,j}\big)^2/\big(\widehat{p}_{i,j}\big)^4\big]^{1/2}\big)$ because $(\varepsilon_{i,j}^{(1)})^2-(\sigma_{i,j}^{(1)})^2$ is $\text{subExponential}(\overline{\sigma}^2)$ and independent across all $i\in[N]$ due to Lemma S3. Then, we have

$$\frac{1}{N}\mathbb{E}\Big[\Big|\sum_{i\in\mathcal{R}_s}\frac{\big[\big(\varepsilon_{i,j}^{(1)}\big)^2-\big(\sigma_{i,j}^{(1)}\big)^2\big]p_{i,j}}{\big(\widehat{p}_{i,j}\big)^2}\Big|\Big|\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s}\Big]\overset{(a)}{\leq}\frac{c\overline{\sigma}^2}{N}\sqrt{\sum_{i\in\mathcal{R}_s}\Big(\frac{p_{i,j}}{\big(\widehat{p}_{i,j}\big)^2}\Big)^2}\overset{(b)}{\leq}\frac{c\overline{\sigma}^2}{\bar{\lambda}^2\sqrt{N}}=o_p(1),\quad(\text{A.35})$$

where (a) follows as the first moment of $\text{subExponential}(\sigma)$ is $O(\sigma)$ and (b) follows from Assumption 3. To bound the second term in Eq. (A.34), applying the Cauchy-Schwarz inequality yields that

$$\begin{aligned}\frac{1}{N}\Big|\sum_{i\in\mathcal{R}_s}\Big(\frac{\big(\sigma_{i,j}^{(1)}\big)^2p_{i,j}}{\big(\widehat{p}_{i,j}\big)^2}-\frac{(\sigma_{i,j}^{(1)})^2}{p_{i,j}}\Big)\Big|&=\frac{1}{N}\Big|\sum_{i\in\mathcal{R}_s}\frac{\big(\sigma_{i,j}^{(1)}\big)^2\big(\big(p_{i,j}\big)^2-\big(\widehat{p}_{i,j}\big)^2\big)}{\big(\widehat{p}_{i,j}\big)^2p_{i,j}}\Big|\\&\overset{(a)}{\leq}\frac{2}{N}\sum_{i\in\mathcal{R}_s}\frac{\big(\sigma_{i,j}^{(1)}\big)^2\big|p_{i,j}-\widehat{p}_{i,j}\big|}{\big(\widehat{p}_{i,j}\big)^2p_{i,j}}\\&\overset{(b)}{\leq}\frac{2\overline{\sigma}^2}{\underline{\lambda}\bar{\lambda}^2N}\sum_{i\in\mathcal{R}_s}\big|p_{i,j}-\widehat{p}_{i,j}\big|\end{aligned}$$

$$\overset{(c)}{\leq} \frac{2\overline{\sigma}^2}{\lambda\bar{\lambda}^2\sqrt{N}}\big\|P_{\cdot,j}-\widehat{P}_{\cdot,j}\big\|_2 \overset{(d)}{=} \frac{2\overline{\sigma}^2}{\lambda\bar{\lambda}^2}\mathcal{E}\big(\widehat{P}\big) \overset{(e)}{=} o_p(1), \quad \text{(A.36)}$$

where (a) follows by using $\big(p_{i,j}\big)^2 - \big(\widehat{p}_{i,j}\big)^2 = (p_{i,j}+\widehat{p}_{i,j})(p_{i,j}-\widehat{p}_{i,j}) \leq 2|p_{i,j}-\widehat{p}_{i,j}|$, (b) follows from Assumptions 1 and 3, and because the variance of a subGaussian random variable is upper bounded by the square of its subGaussian norm, (c) follows by the relationship between ℓ_1 and ℓ_2 norms of a vector, (d) follows from Eq. (16), and (e) follows from (C1).

Putting together Eqs. (A.29) to (A.36) using Lemma S6,

$$\frac{1}{N}\Big|\sum_{i\in\mathcal{R}_s}\mathbb{T}^{(1)}_{i,j}\Big| = o_p(1).$$

Then, the claim in Eq. (A.28) follows for $a=1$ by using $|\sum_{i\in[N]}\mathbb{T}^{(1)}_{i,j}| \leq |\sum_{i\in\mathcal{R}_0}\mathbb{T}^{(1)}_{i,j}| + |\sum_{i\in\mathcal{R}_1}\mathbb{T}^{(1)}_{i,j}|$. The proof of Eq. (A.28) for $a=0$ follows similarly.

A3. Data generating process for the simulations

The inputs of the data generating process (DGP) are: the probability bound λ; two positive constants $c^{(0)}$ and $c^{(1)}$; and the standard deviations $\sigma^{(a)}_{i,j}$ for every $i\in[N], j\in[M], a\in\{0,1\}$. The DGP is:

1. For positive integers r_p, r_θ and $r=\max\{r_p,r_\theta\}$, generate a proxy for the common unit-level latent factors $U^{\text{shared}}\in\mathbb{R}^{N\times r}$, such that, for all $i\in[N]$ and $j\in[r]$, $u^{\text{shared}}_{i,j}$ is independently sampled from a $\texttt{Uniform}(\sqrt{\lambda},\sqrt{1-\lambda})$ distribution, with $\lambda\in(0,1)$.

2. Generate proxies for the measurement-level latent factors $V, V^{(0)}, V^{(1)}\in\mathbb{R}^{M\times r}$, such that, for all $i\in[M]$ and $j\in[r]$, $v_{i,j}, v^{(0)}_{i,j}, v^{(1)}_{i,j}$ are independently sampled from a $\texttt{Uniform}(\sqrt{\lambda},\sqrt{1-\lambda})$ distribution.

3. Generate the treatment assignment probability matrix P

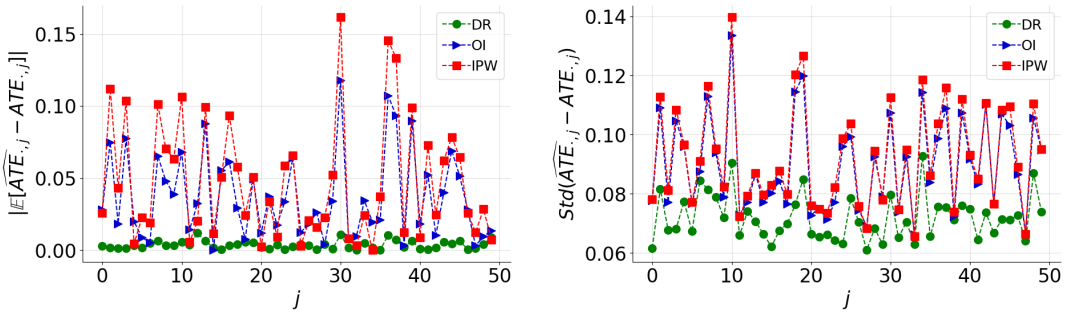
$$P = \frac{1}{r_p}U^{\text{shared}}_{[N]\times[r_p]}V^{\top}_{[M]\times[r_p]}.$$

4. For $a\in\{0,1\}$, run $\texttt{SVD}$ on $U^{\text{shared}}V^{(a)\top}$, i.e.,

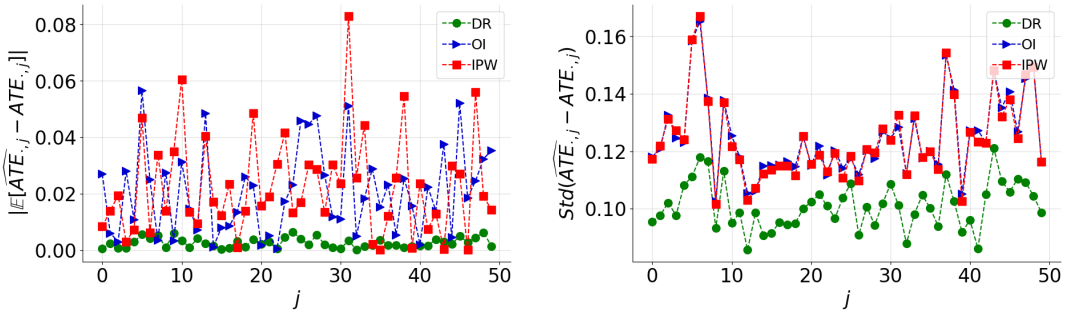
$$\texttt{SVD}(U^{\text{shared}}V^{(a)\top}) = (U^{(a)},\Sigma^{(a)},W^{(a)}).$$

Then, generate the mean potential outcome matrices $\Theta^{(0)}$ and $\Theta^{(1)}$:

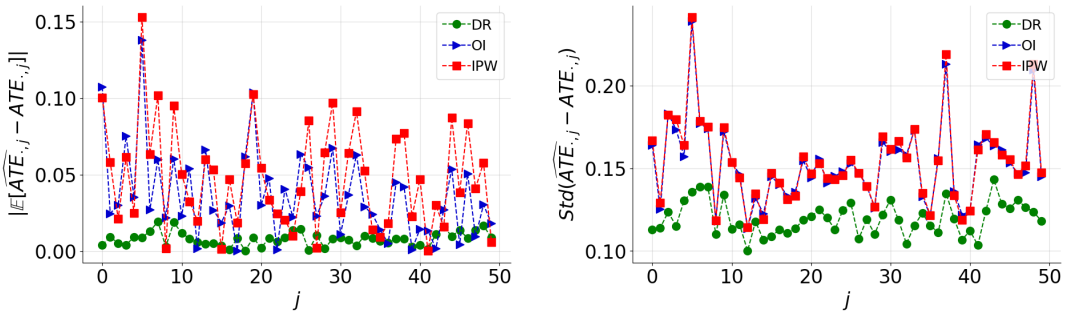
$$\Theta^{(a)} = \frac{c^{(a)}\texttt{Sum}(\Sigma^{(a)})}{r_\theta}U^{(a)}_{[N]\times[r_\theta]}W^{(a)\top}_{[M]\times[r_\theta]},$$



(a) $r_p = 3$, $r_\theta = 3$



(b) $r_p = 3$, $r_\theta = 5$



(c) $r_p = 5$, $r_\theta = 3$

Figure 7: Empirical illustration of the biases and the standard deviations of DR, OI, and IPW estimators for different j, and for different $r_p = 5$ and r_θ.

where $\texttt{Sum}(\Sigma^{(a)})$ denotes the sum of all entries of $\Sigma^{(a)}$.

5. Generate the noise matrices $E^{(0)}$ and $E^{(1)}$, such that, for all $i \in [N], j \in [M], a \in \{0, 1\}$, $\varepsilon_{i,j}^{(a)}$ is independently sampled from a $\mathcal{N}(0, (\sigma_{i,j}^{(a)})^2)$ distribution. Then, determine $y_{i,j}^{(a)}$ from Eq. (2).

6. Generate the noise matrix W, such that, for all $i \in [N], j \in [M]$, $\eta_{i,j}$ is independently sampled as per Eq. (4). Then, determine $a_{i,j}$ and $y_{i,j}$ from Eq. (3) and Eq. (1), respectively.

In our simulations, we set $\lambda = 0.05$, $c^{(0)} = 1$ and $c^{(1)} = 2$. In practice, instead of

choosing the values of $\sigma_{i,j}^{(a)}$ as ex-ante inputs, we make them equal to the standard deviation of all the entries in $\Theta^{(a)}$ for every i and j, separately for $a \in \{0, 1\}$.

In Figure 7, we compare the absolute biases and the standard deviations of OI, IPW, and DR across the first 50 values of j for $N = 1000$, with $r_p = 3$, $r_\theta = 3$ in Panel (a), $r_p = 3$, $r_\theta = 5$ in Panel (b), and $r_p = 5$, $r_\theta = 3$ in Panel (c). For each j, the estimate of the biases of OI, IPW, and DR is the average of $\widehat{\text{ATE}}_{\cdot,j}^{\text{OI}} - \text{ATE}_{\cdot,j}$, $\widehat{\text{ATE}}_{\cdot,j}^{\text{IPW}} - \text{ATE}_{\cdot,j}$ and $\widehat{\text{ATE}}_{\cdot,j}^{\text{DR}} - \text{ATE}_{\cdot,j}$ across the Q simulation instances. Likewise, the estimate of the standard deviation of OI, IPW, and DR is the standard deviation of $\widehat{\text{ATE}}_{\cdot,j}^{\text{OI}} - \text{ATE}_{\cdot,j}$, $\widehat{\text{ATE}}_{\cdot,j}^{\text{IPW}} - \text{ATE}_{\cdot,j}$ and $\widehat{\text{ATE}}_{\cdot,j}^{\text{DR}} - \text{ATE}_{\cdot,j}$ across the Q simulation instances. The DR estimator consistently outperforms the OI and IPW estimators in reducing both absolute biases and standard deviations.

Supplement to "Doubly Robust Inference in Causal Latent Factor Models"

S1. Supporting Concentration and Convergence Results

This section presents known results on subGaussian, subExponential, and subWeibull random variables (defined below), along with few basic results on convergence of random variables.

We use subGaussian(σ) to represent a subGaussian random variable, where σ is a bound on the subGaussian norm; and subExponential(σ) to represent a subExponential random variable, where σ is a bound on the subExponential norm. Recall the definitions of the norms from Section 1 of the main article.

Lemma S1 (subGaussian concentration: Theorem 2.6.3 of Vershynin (2018))**.** *Let $x \in \mathbb{R}^n$ be a random vector whose entries are independent, zero-mean, subGaussian(σ) random variables. Then, for any $b \in \mathbb{R}^n$ and $t \geq 0$,*

$$\mathbb{P}\Big\{\big|b^\top x\big| \geq t\Big\} \leq 2\exp\Big(\frac{-ct^2}{\sigma^2\|b\|_2^2}\Big).$$

The following corollary expresses the bound in Lemma S1 in a convenient form.

Corollary S1 (subGaussian concentration)**.** *Let $x \in \mathbb{R}^n$ be a random vector whose entries are independent, zero-mean, subGaussian(σ) random variables. Then, for any $b \in \mathbb{R}^n$ and any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\big|b^\top x\big| \leq \sigma\sqrt{c\ell_\delta} \cdot \|b\|_2.$$

Proof. The proof follows from Lemma S1 by choosing $\delta \triangleq 2\exp(-ct^2/\sigma^2\|b\|_2^2)$. □

Lemma S2 (subExponential concentration: Theorem 2.8.2 of Vershynin (2018))**.** *Let $x \in \mathbb{R}^n$ be a random vector whose entries are independent, zero-mean, subExponential(σ)*

random variables. Then, for any $b \in \mathbb{R}^n$ *and* $t \geq 0$,

$$\mathbb{P}\Big\{\big|b^\top x\big| \geq t\Big\} \leq 2\exp\bigg(-c\min\Big(\frac{t^2}{\sigma^2\|b\|_2^2}, \frac{t}{\sigma\|b\|_\infty}\Big)\bigg).$$

The following corollary expresses the bound in Lemma S2 in a convenient form.

Corollary S2 (subExponential concentration)**.** *Let* $x \in \mathbb{R}^n$ *be a random vector whose entries are independent, zero-mean, subExponential*(σ) *random variables. Then, for any* $b \in \mathbb{R}^n$ *and any* $\delta \in (0,1)$*, with probability at least* $1-\delta$,

$$\big|b^\top x\big| \leq \sigma m(c\ell_\delta)\cdot\|b\|_2,$$

where recall that $m(c\ell_\delta) = \max\big(c\ell_\delta, \sqrt{c\ell_\delta}\big)$.

Proof. Choosing $t = t_0\sigma\|b\|_2$ in Lemma S2, we have

$$\mathbb{P}\Big\{\big|b^\top x\big| \geq t_0\sigma\|b\|_2\Big\} \leq 2\exp\Big(-ct_0\min\Big(t_0, \frac{\|b\|_2}{\|b\|_\infty}\Big)\Big) \leq 2\exp\Big(-ct_0\min\big(t_0,1\big)\Big),$$

where the second inequality follows from $\min\{t_0, c\} \geq \min\{t_0, 1\}$ for any $c \geq 1$ and $\|b\|_2 \geq \|b\|_\infty$. Then, the proof follows by choosing $\delta \triangleq 2\exp\big(-ct_0\min\big(t_0,1\big)\big)$ which fixes $t_0 = \max\{\sqrt{c\ell_\delta}, c\ell_\delta\} = m(c\ell_\delta)$.

□

Lemma S3 (Product of subGaussians is subExponential: Lemma. 2.7.7 of Vershynin (2018))**.** *Let* x_1 *and* x_2 *be subGaussian*(σ_1) *and subGaussian*(σ_2) *random variables, respectively. Then,* x_1x_2 *is subExponential*$(\sigma_1\sigma_2)$ *random variable.*

Next, we provide the definition of a subWeibull random variable.

Definition S1 (subWeibull random variable: Definition 1 of Zhang and Wei (2022))**.** *For* $\rho > 0$*, a random variable* x *is subWeibull with index* ρ *if it has a bounded subWeibull norm defined as follows:*

$$\|x\|_{\psi_\rho} \triangleq \inf\{t > 0 : \mathbb{E}[\exp(|x|^\rho/t^\rho)] \leq 2\}.$$

We use subWeibull$_\rho(\sigma)$ to represent a subWeibull random variable with index ρ, where σ is a bound on the subWeibull norm. Note that subGaussian and subExponential random variables are subWeibull random variable with indices 2 and 1, respectively.

Lemma S4 (Product of subWeibulls is subWeibull: Proposition 2 of Zhang and Wei (2022))**.** *For* $i \in [d]$*, let* x_i *be a subWeibull*$_{\rho_i}(\sigma_i)$ *random variable. Then,* $\Pi_{i\in[d]}x_i$ *is subWeibull*$_\rho(\sigma)$ *random variable where*

$$\sigma = \Pi_{i\in[d]}\sigma_i \quad \textit{and} \quad \rho = \left(\sum_{i\in[d]} 1/\rho_i\right)^{-1}.$$

Next set of lemmas provide useful intermediate results on stochastic convergence.

Lemma S5. *Let X_n and $\overline{X}_n$ be sequences of random variables. Let δ_n be a deterministic sequence such that $0 \leq \delta_n \leq 1$ and $\delta_n \to 0$. Suppose $X_n = o_p(1)$ and $\mathbb{P}(|\overline{X}_n| \leq |X_n|) \geq 1 - \delta_n$. Then, $\overline{X}_n = o_p(1)$.*

Proof. We need to show that for any $\epsilon > 0$ and $\delta > 0$, there exist finite $\overline{n}$, such that

$$\mathbb{P}(|\overline{X}_n| > \delta) < \epsilon$$

for all $n \geq \overline{n}$. Fix any $\epsilon > 0$. As δ_n converges to zero, there exists a finite n_0 such that $\delta_n < \epsilon/2$, for all $n \geq n_0$. As X_n is converges to zero in probability, there exists finite n_1, such that $\mathbb{P}(|X_n| > \delta) < \epsilon/2$ for all $n \geq n_1$. Now, the event $\{|\overline{X}_n| > \delta\}$ belongs to the union of $\{|\overline{X}_n| > |X_n|\}$ and $\{|X_n| > \delta\}$. As a result, we obtain

$$\mathbb{P}(|\overline{X}_n| > \delta) \leq \mathbb{P}(|\overline{X}_n| > |X_n|) + \mathbb{P}(|X_n| > \delta) \leq \delta_n + \mathbb{P}(|X_n| > \delta) < \epsilon,$$

for $n \geq \overline{n} = \max\{n_0, n_1\}$. Therefore, $\overline{X}_n = o_p(1)$. □

Lemma S6. *Let X_n and $\overline{X}_n$ be sequences of random variables. Suppose $\mathbb{E}\big[|X_n|\big|\overline{X}_n\big] = o_p(1)$. Then, $X_n = o_p(1)$.*

Proof. Fix any $\delta > 0$. Markov's inequality implies

$$\mathbb{P}\Big(|X_n| \geq \delta\Big|\overline{X}_n\Big) \leq \frac{1}{\delta}\mathbb{E}\Big[|X_n|\Big|\overline{X}_n\Big] = o_p(1).$$

The law of total probability and the boundedness of conditional probabilities yield

$$\mathbb{P}\Big(|X_n| \geq \delta\Big) = \mathbb{E}\Big[\mathbb{P}\Big(|X_n| \geq \delta\Big|\overline{X}_n\Big)\Big] \longrightarrow 0.$$

□

Lemma S7. *Let X_n and $\overline{X}_n$ be sequences of random variables. Suppose $X_n = O_p(1)$ and $\mathbb{P}\big(|\overline{X}_n| \geq |X_n| + f(\epsilon)\big) < \epsilon$ for some positive function f and every $\epsilon \in (0, 1)$. Then, $\overline{X}_n = O_p(1)$.*

Proof. We need to show that for any $\epsilon > 0$, there exist finite $\overline{\delta} > 0$ and $\overline{n} > 0$, such that

$$\mathbb{P}(|\overline{X}_n| > \overline{\delta}) < \epsilon$$

for all $n \geq \overline{n}$. Fix any $\epsilon > 0$. Because X_n is bounded in probability, there exist finite δ and n_0, such that $\mathbb{P}(|X_n| > \delta) < \epsilon/2$ for all $n \geq n_0$. Further, we have $\mathbb{P}\big(|\overline{X}_n| \geq |X_n| + f(\epsilon/2)\big) < \epsilon/2$. Now, the event $\{|\overline{X}_n| > \delta + f(\epsilon/2)\}$ belongs to the union of $\{|\overline{X}_n| > |X_n| + f(\epsilon/2)\}$ and $\{|X_n| > \delta\}$. As a result, we obtain

$$\mathbb{P}\big(|\overline{X}_n| > \delta + f(\epsilon/2)\big) \leq \mathbb{P}\big(|\overline{X}_n| > |X_n| + f(\epsilon/2)\big) + \mathbb{P}\big(|X_n| > \delta\big) < \epsilon.$$

for all $n \geq n_0$. In other words, $\mathbb{P}(|\overline{X}_n| > \overline{\delta}) < \epsilon$ for all $n \geq \overline{n}$, where $\overline{\delta} = \delta + f(\epsilon/2) > 0$ and $\overline{n} = n_0$. Therefore, $\overline{X}_n = O_p(1)$. □

S2. Proofs of Corollaries 1 and 2

S2.1. Proof of Corollary 1: Gains of DR over OI and IPW

Fix any $j \in [M]$ and any $\delta \in (0,1)$. First, consider IPW. Take any $\alpha \in [0, 1/2]$. From Eq. (20), with probability at least $1-\delta$,

$$N^{\alpha}\big|\widehat{\mathrm{ATE}}_{\cdot,j}^{\mathrm{IPW}} - \mathrm{ATE}_{\cdot,j}\big| \leq \frac{2\theta_{\max}}{\bar{\lambda}} N^{\alpha}\mathcal{E}\big(\widehat{P}\big) + f_1(\delta)N^{\alpha-1/2} \leq \frac{2\theta_{\max}}{\bar{\lambda}} N^{\alpha}\mathcal{E}\big(\widehat{P}\big) + f_1(\delta),$$

where

$$f_1(\delta) \triangleq \frac{2}{\bar{\lambda}}\bigg(\frac{\sqrt{c\ell_{\delta/12}}}{\sqrt{\ell_1}}\theta_{\max} + 2\overline{\sigma}\sqrt{c\ell_{\delta/12}} + \frac{2\overline{\sigma}m(c\ell_{\delta/12})}{\sqrt{\ell_1}}\bigg),$$

for $m(c)$ and ℓ_c as defined in Section 1 of the main article. Then, if $\mathcal{E}\big(\widehat{P}\big) = O_p\big(N^{-\alpha}\big)$, Lemma S7 implies

$$\big|\widehat{\mathrm{ATE}}_{\cdot,j}^{\mathrm{IPW}} - \mathrm{ATE}_{\cdot,j}\big| = O_p\big(N^{-\alpha}\big).$$

Next, consider DR. From Eq. (17), with probability at least $1-\delta$,

$$\big|\widehat{\mathrm{ATE}}_{\cdot,j}^{\mathrm{DR}} - \mathrm{ATE}_{\cdot,j}\big| \leq \frac{2}{\bar{\lambda}}\mathcal{E}\big(\widehat{\Theta}\big)\mathcal{E}\big(\widehat{P}\big) + f_2(\delta)N^{-1/2},$$

where

$$f_2(\delta) \triangleq \frac{2}{\bar{\lambda}}\bigg(\frac{\sqrt{c\ell_{\delta/12}}}{\sqrt{\ell_1}}\mathcal{E}\big(\widehat{\Theta}\big) + 2\overline{\sigma}\sqrt{c\ell_{\delta/12}} + \frac{2\overline{\sigma}m(c\ell_{\delta/12})}{\sqrt{\ell_1}}\bigg).$$

Suppose $\mathcal{E}\big(\widehat{P}\big) = O_p\big(N^{-\alpha}\big)$ and $\mathcal{E}\big(\widehat{\Theta}\big) = O_p\big(N^{-\beta}\big)$. Consider two cases. First, suppose $\alpha + \beta \leq 0.5$. Then, with probability at least $1-\delta$,

$$\begin{aligned} N^{\alpha+\beta}\big|\widehat{\mathrm{ATE}}_{\cdot,j}^{\mathrm{DR}} - \mathrm{ATE}_{\cdot,j}\big| &\leq \frac{2}{\bar{\lambda}}N^{\alpha+\beta}\mathcal{E}\big(\widehat{\Theta}\big)\mathcal{E}\big(\widehat{P}\big) + f_2(\delta)N^{\alpha+\beta-1/2} \\ &\leq \frac{2}{\bar{\lambda}}N^{\alpha+\beta}\mathcal{E}\big(\widehat{\Theta}\big)\mathcal{E}\big(\widehat{P}\big) + f_2(\delta). \end{aligned}$$

Lemma S7 implies $\big|\widehat{\mathrm{ATE}}_{\cdot,j}^{\mathrm{DR}} - \mathrm{ATE}_{\cdot,j}\big| = O_p(N^{-(\alpha+\beta)})$. Next, suppose $\alpha + \beta > 0.5$. With probability at least $1-\delta$,

$$N^{1/2}\big|\widehat{\mathrm{ATE}}_{\cdot,j}^{\mathrm{DR}} - \mathrm{ATE}_{\cdot,j}\big| \leq \frac{2}{\bar{\lambda}}N^{1/2}\mathcal{E}\big(\widehat{\Theta}\big)\mathcal{E}\big(\widehat{P}\big) + f_2(\delta) \leq \frac{2}{\bar{\lambda}}N^{\alpha+\beta}\mathcal{E}\big(\widehat{\Theta}\big)\mathcal{E}\big(\widehat{P}\big) + f_2(\delta).$$

Lemma S7 implies $\big|\widehat{\mathrm{ATE}}_{\cdot,j}^{\mathrm{DR}} - \mathrm{ATE}_{\cdot,j}\big| = O_p(N^{-1/2})$.

S2.2. Proof of Corollary 2: Consistency for DR

Fix any $j \in [M]$. Then, choose $\delta = 1/N$ in Eq. (18) and note that every term in the right hand side of Eq. (18) is $o_p(1)$ under the conditions on $\mathcal{E}\big(\widehat{\Theta}\big)$ and $\mathcal{E}\big(\widehat{P}\big)$. Then, Eq. (21) follows from Lemma S5.

S3. Proof of Proposition 1 (19): Finite Sample Guarantees for OI

Fix any $j \in [M]$. Recall the definitions of the parameter $\mathrm{ATE}_{\cdot,j}$ and corresponding outcome imputation estimate $\widehat{\mathrm{ATE}}_{\cdot,j}^{\mathrm{OI}}$ from Eqs. (5) and (9), respectively. The error $\Delta\mathrm{ATE}_{\cdot,j}^{\mathrm{OI}} = \widehat{\mathrm{ATE}}_{\cdot,j}^{\mathrm{OI}} - \mathrm{ATE}_{\cdot,j}$ can be re-expressed as

$$\Delta\mathrm{ATE}_{\cdot,j}^{\mathrm{OI}} = \frac{1}{N}\sum_{i\in[N]}\Big(\widehat{\theta}_{i,j}^{(1)} - \widehat{\theta}_{i,j}^{(0)}\Big) - \frac{1}{N}\sum_{i\in[N]}\Big(\theta_{i,j}^{(1)} - \theta_{i,j}^{(0)}\Big) = \frac{1}{N}\sum_{i\in[N]}\Big(\big(\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)}\big) - \big(\widehat{\theta}_{i,j}^{(0)} - \theta_{i,j}^{(0)}\big)\Big).$$

Using the triangle inequality, we have

$$\big|\Delta\mathrm{ATE}_{\cdot,j}^{\mathrm{OI}}\big| \le \frac{1}{N}\Big|\sum_{i\in[N]}\big(\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)}\big)\Big| + \frac{1}{N}\Big|\sum_{i\in[N]}\big(\widehat{\theta}_{i,j}^{(0)} - \theta_{i,j}^{(0)}\big)\Big|. \tag{S.1}$$

Consider any $a \in \{0,1\}$. We claim that

$$\frac{1}{N}\Big|\sum_{i\in[N]}\big(\widehat{\theta}_{i,j}^{(a)} - \theta_{i,j}^{(a)}\big)\Big| \le \mathcal{E}\big(\widehat{\Theta}^{(a)}\big). \tag{S.2}$$

The proof is complete by putting together Eqs. (S.1) and (S.2).

Proof of Eq. (S.2) Fix any $a \in \{0,1\}$. Using the Cauchy-Schwarz inequality, we have

$$\frac{1}{N}\Big|\sum_{i\in[N]}\big(\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)}\big)\Big| \le \frac{1}{N}\|\mathbf{1}_N\|_2\|\widehat{\Theta}_{\cdot,j}^{(1)} - \Theta_{\cdot,j}^{(1)}\|_2 \le \frac{1}{\sqrt{N}}\|\widehat{\Theta}^{(1)} - \Theta^{(1)}\|_{1,2}.$$

S4. Proof of Proposition 1 (20): Finite Sample Guarantees for IPW

Fix any $j \in [M]$. Recall the definitions of the parameter $\mathrm{ATE}_{\cdot,j}$ and corresponding inverse probability weighting estimate $\widehat{\mathrm{ATE}}_{\cdot,j}^{\mathrm{IPW}}$ from Eqs. (5) and (10), respectively.

The error $\Delta\mathrm{ATE}_{\cdot,j}^{\mathrm{IPW}} = \widehat{\mathrm{ATE}}_{\cdot,j}^{\mathrm{IPW}} - \mathrm{ATE}_{\cdot,j}$ can be re-expressed as

$$\begin{aligned}
\Delta\mathrm{ATE}_{\cdot,j}^{\mathrm{IPW}} &= \frac{1}{N}\sum_{i\in[N]}\Big(\frac{y_{i,j}a_{i,j}}{\widehat{p}_{i,j}} - \frac{y_{i,j}(1-a_{i,j})}{1-\widehat{p}_{i,j}}\Big) - \frac{1}{N}\sum_{i\in[N]}\Big(\theta_{i,j}^{(1)} - \theta_{i,j}^{(0)}\Big) \\
&= \frac{1}{N}\sum_{i\in[N]}\Big(\Big(\frac{y_{i,j}a_{i,j}}{\widehat{p}_{i,j}} - \theta_{i,j}^{(1)}\Big) - \Big(\frac{y_{i,j}(1-a_{i,j})}{1-\widehat{p}_{i,j}} - \theta_{i,j}^{(0)}\Big)\Big) \\
&\overset{(a)}{=} \frac{1}{N}\sum_{i\in[N]}\Big(\mathbb{T}_{i,j}^{(1,\mathrm{IPW})} + \mathbb{T}_{i,j}^{(0,\mathrm{IPW})}\Big),
\end{aligned} \tag{S.3}$$

where (a) follows after defining $\mathbb{T}_{i,j}^{(1,\mathrm{IPW})} \triangleq y_{i,j}a_{i,j}/\widehat{p}_{i,j} - \theta_{i,j}^{(1)}$ and $\mathbb{T}_{i,j}^{(0,\mathrm{IPW})} \triangleq \theta_{i,j}^{(0)} - y_{i,j}(1-a_{i,j})/(1-\widehat{p}_{i,j})$. Then, we have

$$\begin{aligned}
\mathbb{T}_{i,j}^{(1,\mathrm{IPW})} &= \frac{y_{i,j}a_{i,j}}{\widehat{p}_{i,j}} - \theta_{i,j}^{(1)} \\
&\overset{(a)}{=} \frac{\big(\theta_{i,j}^{(1)} + \varepsilon_{i,j}^{(1)}\big)\big(p_{i,j} + \eta_{i,j}\big)}{\widehat{p}_{i,j}} - \theta_{i,j}^{(1)} \\
&= \theta_{i,j}^{(1)}\Big(\frac{p_{i,j}+\eta_{i,j}}{\widehat{p}_{i,j}} - 1\Big) + \varepsilon_{i,j}^{(1)}\Big(\frac{p_{i,j}+\eta_{i,j}}{\widehat{p}_{i,j}}\Big) \\
&= \frac{\theta_{i,j}^{(1)}\big(p_{i,j} - \widehat{p}_{i,j}\big)}{\widehat{p}_{i,j}} + \frac{\theta_{i,j}^{(1)}\eta_{i,j}}{\widehat{p}_{i,j}} + \frac{\varepsilon_{i,j}^{(1)}p_{i,j}}{\widehat{p}_{i,j}} + \frac{\varepsilon_{i,j}^{(1)}\eta_{i,j}}{\widehat{p}_{i,j}},
\end{aligned} \tag{S.4}$$

where (a) follows from Eqs. (1) to (3). A similar derivation for $a = 0$ implies that

$$\begin{aligned}
\mathbb{T}_{i,j}^{(0,\mathrm{IPW})} &= \theta_{i,j}^{(0)} - \frac{y_{i,j}(1-a_{i,j})}{1-\widehat{p}_{i,j}} \\
&= -\frac{\theta_{i,j}^{(0)}\big(1-p_{i,j}-\big(1-\widehat{p}_{i,j}\big)\big)}{1-\widehat{p}_{i,j}} - \frac{\theta_{i,j}^{(0)}(-\eta_{i,j})}{1-\widehat{p}_{i,j}} - \frac{\varepsilon_{i,j}^{(0)}\big(1-p_{i,j}\big)}{1-\widehat{p}_{i,j}} - \frac{\varepsilon_{i,j}^{(0)}(-\eta_{i,j})}{1-\widehat{p}_{i,j}} \\
&= \frac{\theta_{i,j}^{(0)}\big(p_{i,j} - \widehat{p}_{i,j}\big)}{1-\widehat{p}_{i,j}} + \frac{\theta_{i,j}^{(0)}\eta_{i,j}}{1-\widehat{p}_{i,j}} - \frac{\varepsilon_{i,j}^{(0)}\big(1-p_{i,j}\big)}{1-\widehat{p}_{i,j}} + \frac{\varepsilon_{i,j}^{(0)}\eta_{i,j}}{1-\widehat{p}_{i,j}}.
\end{aligned}$$

Consider any $a \in \{0,1\}$ and $\delta \in (0,1)$. We claim that, with probability at least $1-6\delta$,

$$\frac{1}{N}\Big|\sum_{i\in[N]}\mathbb{T}_{i,j}^{(\mathrm{a,IPW})}\Big| \le \frac{2}{\overline{\lambda}}\|\Theta^{(a)}\|_{\max}\mathcal{E}\big(\widehat{P}\big) + \frac{2\sqrt{c\ell_\delta}}{\overline{\lambda}\sqrt{\ell_1 N}}\|\Theta^{(a)}\|_{\max} + \frac{2\overline{\sigma}\sqrt{c\ell_\delta}}{\overline{\lambda}\sqrt{N}} + \frac{2\overline{\sigma}m(c\ell_\delta)}{\overline{\lambda}\sqrt{\ell_1 N}}. \tag{S.5}$$

where recall that $m(c\ell_\delta) = \max\big(c\ell_\delta, \sqrt{c\ell_\delta}\big)$. We provide a proof of this claim at the end of this section. Applying triangle inequality in Eq. (S.3) and using Eq. (S.5) with

a union bound, we obtain that

$$\left|\Delta\mathrm{ATE}^{\mathrm{IPW}}_{\cdot,j}\right| \leq \frac{2}{\bar{\lambda}}\theta_{\max}\,\mathcal{E}\big(\widehat{P}\big) + \frac{2\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{\ell_1 N}}\theta_{\max} + \frac{4\overline{\sigma}\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{N}} + \frac{4\overline{\sigma}m(c\ell_\delta)}{\bar{\lambda}\sqrt{\ell_1 N}},$$

with probability at least $1-12\delta$. The claim in Eq. (20) follows by re-parameterizing δ.

Proof of Eq. (S.5). This proof follows a very similar road map to that used for establishing the inequality in Eq. (A.5). Recall the partitioning of the units $[N]$ into $\mathcal{R}_0$ and $\mathcal{R}_1$ from Assumption 4. Now, to enable the application of concentration bounds, we split the summation over $i \in [N]$ in the left hand side of Eq. (S.5) into two parts—one over $i \in \mathcal{R}_0$ and the other over $i \in \mathcal{R}_1$—such that the noise terms are independent of the estimates of $\Theta^{(0)}, \Theta^{(1)}, P$ in each of these parts as in Eqs. (14) and (15).

Fix $a = 1$ and note that $|\sum_{i\in[N]} \mathbb{T}^{(1,\mathrm{IPW})}_{i,j}| \leq |\sum_{i\in\mathcal{R}_0} \mathbb{T}^{(1,\mathrm{IPW})}_{i,j}| + |\sum_{i\in\mathcal{R}_1} \mathbb{T}^{(1,\mathrm{IPW})}_{i,j}|$. Fix any $s \in \{0,1\}$. Then, Eq. (S.4) and triangle inequality imply that

$$\Big|\sum_{i\in\mathcal{R}_s} \mathbb{T}^{(1,\mathrm{IPW})}_{i,j}\Big| \leq \Big|\sum_{i\in\mathcal{R}_s} \frac{\theta^{(1)}_{i,j}\big(p_{i,j}-\widehat{p}_{i,j}\big)}{\widehat{p}_{i,j}}\Big| + \Big|\sum_{i\in\mathcal{R}_s} \frac{\theta^{(1)}_{i,j}\eta_{i,j}}{\widehat{p}_{i,j}}\Big| + \Big|\sum_{i\in\mathcal{R}_s} \frac{\varepsilon^{(1)}_{i,j}p_{i,j}}{\widehat{p}_{i,j}}\Big| + \Big|\sum_{i\in\mathcal{R}_s} \frac{\varepsilon^{(1)}_{i,j}\eta_{i,j}}{\widehat{p}_{i,j}}\Big|. \tag{S.6}$$

Next, note that the decomposition in Eq. (S.6) is identical to the one in Eq. (A.6), except for the fact when compared to Eq. (A.6), the first two terms in Eq. (S.6) have a factor of $\theta^{(1)}_{i,j}$ instead of $\big(\widetilde{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}\big)$. As a result, mimicking steps used to derive Eq. (A.11), we obtain the following bound, with probability at least $1-3\delta$,

$$\begin{aligned}
\frac{1}{N}\Big|\sum_{i\in\mathcal{R}_s} \mathbb{T}^{(1,\mathrm{IPW})}_{i,j}\Big| &\leq \frac{1}{\bar{\lambda}N}\|\Theta^{(1)}\|_{1,2}\|\widehat{P}-P\|_{1,2} + \frac{\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{\ell_1}N}\|\Theta^{(1)}\|_{1,2} + \frac{\overline{\sigma}\sqrt{c\ell_\delta}}{\bar{\lambda}N}\|P\|_{1,2} + \frac{\overline{\sigma}m(c\ell_\delta)}{\bar{\lambda}\sqrt{\ell_1}N}\|\mathbf{1}\|_{1,2} \\
&\overset{(a)}{\leq} \frac{1}{\bar{\lambda}\sqrt{N}}\|\Theta^{(1)}\|_{\max}\|\widehat{P}-P\|_{1,2} + \frac{\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{\ell_1 N}}\|\Theta^{(1)}\|_{\max} + \frac{\overline{\sigma}\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{N}} + \frac{\overline{\sigma}m(c\ell_\delta)}{\bar{\lambda}\sqrt{\ell_1 N}} \\
&\overset{(b)}{\leq} \frac{1}{\bar{\lambda}}\|\Theta^{(1)}\|_{\max}\,\mathcal{E}\big(\widehat{P}\big) + \frac{\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{\ell_1 N}}\|\Theta^{(1)}\|_{\max} + \frac{\overline{\sigma}\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{N}} + \frac{\overline{\sigma}m(c\ell_\delta)}{\bar{\lambda}\sqrt{\ell_1 N}},
\end{aligned} \tag{S.7}$$

where (a) follows because $\|\Theta^{(1)}\|_{1,2} \leq \sqrt{N}\|\Theta^{(1)}\|_{\max}$, $\|P\|_{1,2} \leq \sqrt{N}$ and $\|\mathbf{1}\|_{1,2} = \sqrt{N}$, and (b) follows from Eq. (16). Then, the claim in Eq. (S.5) follows for $a = 1$ by using Eq. (S.7) and applying a union bound over $s \in \{0,1\}$. The proof of Eq. (S.5) for $a = 0$ follows similarly.

S5. Proofs of Propositions 3 and 4

In Section S5.1, we prove Proposition 3, i.e., we show that the estimates of P, $\Theta^{(0)}$, and $\Theta^{(1)}$ generated by `Cross-Fitted-MC` satisfy Assumption 4. Next, we prove

Proposition 4 implying that the estimates of P, $\Theta^{(0)}$, and $\Theta^{(1)}$ generated by `Cross-Fitted-SVD` satisfy the condition (C2) in Theorem 2 as long as $\sqrt{N}/M = o(1)$.

S5.1. Proof of Proposition 3: Guarantees for `Cross-Fitted-MC`

Consider any matrix completion algorithm `MC`. We show that

$$\widehat{P}_{\mathcal{I}}, \widehat{\Theta}^{(a)}_{\mathcal{I}} \perp\!\!\!\perp W_{\mathcal{I}} \tag{S.8}$$

and

$$\widehat{P}_{\mathcal{I}} \perp\!\!\!\perp W_{\mathcal{I}}, E^{(a)}_{\mathcal{I}}, \tag{S.9}$$

for every $\mathcal{I} \in \mathcal{P}$ and $a \in \{0,1\}$, where $\mathcal{P}$ is the block partition of $[N] \times [M]$ into four blocks from Assumption 5. Then, Eqs. (14) and (15) in Assumption 4 follow from Eqs. (S.8) and (S.9), respectively.

Consider $\widehat{\Theta}^{(0)}, \widehat{\Theta}^{(1)}$, and $\widehat{P}$ as in Eqs. (30) to (32). Fix any $a \in \{0,1\}$. From Eq. (29), note that $\widehat{P}_{\mathcal{I}}$ depends only on $A \otimes \mathbf{1}^{-\mathcal{I}}$ and $\widehat{\Theta}^{(a)}_{\mathcal{I}}$ depends on $Y^{(a),\text{obs}} \otimes \mathbf{1}^{-\mathcal{I}}$. In other words, the randomness in $\big(\widehat{P}_{\mathcal{I}}, \widehat{\Theta}^{(a)}_{\mathcal{I}}\big)$ stems from the randomness in $\big(A_{-\mathcal{I}}, Y^{(a),\text{obs}}_{-\mathcal{I}}\big)$ which in turn stems from the randomness in $\big(W_{-\mathcal{I}}, E^{(a)}_{-\mathcal{I}}\big)$. Then, Eq. (S.8) follows from Eq. (27). Likewise, the randomness in $\widehat{P}_{\mathcal{I}}$ stems from the randomness in $A_{-\mathcal{I}}$ which in turn stems from the randomness in $W_{-\mathcal{I}}$. Then, Eq. (S.9) follows from Eq. (28).

To prove Eq. (24), we show that

$$\widehat{P}_{\mathcal{I}}, \widehat{\Theta}^{(a)}_{\mathcal{I}} \perp\!\!\!\perp W_{\mathcal{I}}, E^{(a)}_{\mathcal{I}}, \tag{S.10}$$

for every $\mathcal{I} \in \mathcal{P}$ and $a \in \{0,1\}$. As mentioned above, the randomness in $\big(\widehat{P}_{\mathcal{I}}, \widehat{\Theta}^{(a)}_{\mathcal{I}}\big)$ stems from the randomness in $\big(A_{-\mathcal{I}}, Y^{(a),\text{obs}}_{-\mathcal{I}}\big)$ which in turn stems from the randomness in $\big(W_{-\mathcal{I}}, E^{(a)}_{-\mathcal{I}}\big)$. Then, Eq. (S.10) follows from Eq. (33).

S5.2. Proof of Proposition 4: Guarantees for `Cross-Fitted-SVD`

To prove this result, we first derive a corollary of Lemma A.1 in Bai and Ng (2021) for a generic matrix of interest T, such that $S = (T + H) \otimes F$, and apply it to P, $\Theta^{(0)} \odot (\mathbf{1} - P)$, and $\Theta^{(1)} \odot P$. We impose the following restrictions on T, H, and F.

Assumption S1 (Strong linear latent factors)**.** *There exist a constant $r_T \in [\min\{N, M\}]$ and a collection of latent factors*

$$\widetilde{U} \in \mathbb{R}^{N \times r_T} \quad \textit{and} \quad \widetilde{V} \in \mathbb{R}^{M \times r_T},$$

such that,

(a) T satisfies the factorization: $T = \widetilde{U}\widetilde{V}^\top$,

(b) $\|\widetilde{U}\|_{2,\infty} \leq c$ and $\|\widetilde{V}\|_{2,\infty} \leq c$ for some positive constant c, and

(c) The matrices defined below exist and are positive definite:

$$\lim_{N\to\infty} \frac{\widetilde{U}^\top \widetilde{U}}{N} \quad \textit{and} \quad \lim_{M\to\infty} \frac{\widetilde{V}^\top \widetilde{V}}{M}.$$

Assumption S2 (Zero-mean, weakly dependent, and subExponential noise)**.** *The noise matrix H is such that,*

(a) $\{h_{i,j} : i \in [N], j \in [M]\}$ are zero-mean subExponential with the subExponential norm bounded by a constant $\overline{\sigma}$,

(b) $\sum_{j' \in [M]} \left|\mathbb{E}[h_{i,j} h_{i,j'}]\right| \leq c$ for every $i \in [N]$ and $j \in [M]$, and

(c) The elements of $\{H_{i,\cdot} : i \in [N]\}$ are mutually independent (across i).

Assumption S3 (Strong block factors)**.** *Consider the latent factors $\widetilde{U} \in \mathbb{R}^{N \times r_T}$ and $\widetilde{V} \in \mathbb{R}^{M \times r_T}$ from Assumption S1. Let $\mathcal{R}_{\text{obs}} \subseteq [N]$ and $\mathcal{C}_{\text{obs}} \subseteq [M]$ denote the set of rows and columns of S, respectively, with all entries observed, and $\mathcal{R}_{\text{miss}} \triangleq [N] \setminus \mathcal{R}_{\text{obs}}$ and $\mathcal{C}_{\text{miss}} \triangleq [M] \setminus \mathcal{C}_{\text{obs}}$. Let $\widetilde{U}^{\text{obs}} \in \mathbb{R}^{|\mathcal{R}_{\text{obs}}| \times r_T}$ and $\widetilde{U}^{\text{miss}} \in \mathbb{R}^{|\mathcal{R}_{\text{miss}}| \times r_T}$ be the sub-matrices of $\widetilde{U}$ that keeps the rows corresponding to the indices in $\mathcal{R}_{\text{obs}}$ and $\mathcal{R}_{\text{miss}}$, respectively. Let $\widetilde{V}^{\text{obs}} \in \mathbb{R}^{|\mathcal{C}_{\text{obs}}| \times r_T}$ and $\widetilde{V}^{\text{miss}} \in \mathbb{R}^{|\mathcal{C}_{\text{miss}}| \times r_T}$ be the sub-matrices of $\widetilde{V}$ that keeps the rows corresponding to the indices in $\mathcal{C}_{\text{obs}}$ and $\mathcal{C}_{\text{miss}}$, respectively. Then, the matrices defined below exist and are positive definite:*

$$\lim_{N\to\infty} \frac{\widetilde{U}^{\text{obs}\top} \widetilde{U}^{\text{obs}}}{|\mathcal{R}_{\text{obs}}|}, \ \lim_{M\to\infty} \frac{\widetilde{U}^{\text{miss}\top} \widetilde{U}^{\text{miss}}}{|\mathcal{R}_{\text{miss}}|}, \ \lim_{N\to\infty} \frac{\widetilde{V}^{\text{obs}\top} \widetilde{V}^{\text{obs}}}{|\mathcal{C}_{\text{obs}}|}, \ \textit{and} \ \lim_{M\to\infty} \frac{\widetilde{V}^{\text{miss}\top} \widetilde{V}^{\text{miss}}}{|\mathcal{C}_{\text{miss}}|}. \quad \text{(S.11)}$$

Further, the mask matrix F is such that

$$|\mathcal{R}_{\text{obs}}| = \Omega(N), \quad |\mathcal{R}_{\text{miss}}| = \Omega(N), \quad |\mathcal{C}_{\text{obs}}| = \Omega(M), \quad \textit{and} \quad |\mathcal{C}_{\text{miss}}| = \Omega(M). \quad \text{(S.12)}$$

The next result characterizes the entry-wise error in recovering the missing entries of a matrix where all entries in one block are deterministically missing (see the discussion in Section 5.1 of the main article) using the `TW` algorithm (summarized in Section 5.2.1 of the main article). Its proof, essentially established as a corollary of Bai and Ng (2021, Lemma A.1), is provided in Section S5.3.

Corollary S3. *Consider a matrix of interest T, a noise matrix H, and a mask matrix F such that that Assumptions S1 to S3 hold. Let $S \in \{\mathbb{R} \cup \{?\}\}^{N \times M}$ be the observed matrix as in Eq. (6). Let $\mathcal{R}_{\text{obs}} \subseteq [N]$ and $\mathcal{C}_{\text{obs}} \subseteq [M]$ denote the set of rows and columns of S, respectively, with all entries observed. Let $\mathcal{I} = \mathcal{R}_{\text{miss}} \times \mathcal{C}_{\text{miss}}$ where $\mathcal{R}_{\text{miss}} \triangleq [N] \setminus \mathcal{R}_{\text{obs}}$ and $\mathcal{C}_{\text{miss}} \triangleq [M] \setminus \mathcal{C}_{\text{obs}}$. Then, $\texttt{TW}_{r_T}$ produces an estimate $\widehat{T}_{\mathcal{I}}$ of $T_{\mathcal{I}}$ such that*

$$\|\widehat{T}_{\mathcal{I}} - T_{\mathcal{I}}\|_{\max} = O_p\left(\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{M}}\right),$$

as $N, M \to \infty$.

Given this corollary, we now complete the proof of Proposition 4. Consider the partition $\mathcal{P}$ from Assumption 5 and fix any $\mathcal{I} \in \mathcal{P}$. Recall that `Cross-Fitted-SVD` applies `TW` on $P \otimes \mathbf{1}^{-\mathcal{I}}$, $Y^{(0),\text{full}} \otimes \mathbf{1}^{-\mathcal{I}}$, and $Y^{(1),\text{full}} \otimes \mathbf{1}^{-\mathcal{I}}$, and note that the mask matrix $\mathbf{1}^{-\mathcal{I}}$ satisfies the requirement in Assumption S3, i.e., Eq. (S.12) under Assumption 8.

S5.2.1. Estimating P.
Consider estimating P using `Cross-Fitted-SVD`. To apply Corollary S3, we use Assumptions 6 and 7 to note that P satisfies Assumption S1 with rank parameter r_p. Then, we use Eq. (4), Assumption 2, and Assumption 9 to note that W satisfies Assumption S2. Finally, we use Assumption 8 to note that Assumption S3 holds. Step 2 of `Cross-Fitted-SVD` can be rewritten as $\widehat{P} = \texttt{Proj}_{\bar{\lambda}}(\overline{P})$ and $\overline{P} = \texttt{Cross-Fitted-MC}(\texttt{TW}_{r_1}, A, \mathcal{P})$ where $r_1 = r_p$. Then,

$$\|\widehat{P}_{\mathcal{I}} - P_{\mathcal{I}}\|_{\max} \overset{(a)}{\leq} \|\overline{P}_{\mathcal{I}} - P_{\mathcal{I}}\|_{\max} \overset{(b)}{=} O_p\bigg(\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{M}}\bigg),$$

where (a) follows from Assumption 1, the choice of $\bar{\lambda}$, and the definition of $\texttt{Proj}_{\bar{\lambda}}(\cdot)$, and (b) follows from Corollary S3. Applying a union bound over all $\mathcal{I} \in \mathcal{P}$, we have

$$\mathcal{E}\big(\widehat{P}\big) \overset{(a)}{\leq} \|\widehat{P} - P\|_{\max} = O_p\bigg(\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{M}}\bigg), \tag{S.13}$$

where (a) follows from the definition of $(1,2)$ operator norm.

S5.2.2. Estimating $\Theta^{(0)}$ *and* $\Theta^{(1)}$.
For every $a \in \{0,1\}$, we show that

$$\mathcal{E}\big(\widehat{\Theta}^{(a)}\big) = O_p\bigg(\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{M}}\bigg). \tag{S.14}$$

We focus on $a = 1$ noting that the proof for $a = 0$ is analogous. We split the proof in two cases: (i) $\|\big(\widehat{\Theta}^{(1)} - \Theta^{(1)}\big) \odot \widehat{P}\|_{\max} \leq \|\Theta^{(1)} \odot \big(\widehat{P} - P\big)\|_{\max}$ and (ii) $\|\big(\widehat{\Theta}^{(1)} - \Theta^{(1)}\big) \odot \widehat{P}\|_{\max} \geq \|\Theta^{(1)} \odot \big(\widehat{P} - P\big)\|_{\max}$.

In the first case, we have

$$\bar{\lambda}\|\widehat{\Theta}^{(1)} - \Theta^{(1)}\|_{\max} \overset{(a)}{\leq} \|\big(\widehat{\Theta}^{(1)} - \Theta^{(1)}\big) \odot \widehat{P}\|_{\max} \leq \|\Theta^{(1)} \odot \big(\widehat{P} - P\big)\|_{\max} \overset{(b)}{\leq} \|\Theta^{(1)}\|_{\max}\|\widehat{P} - P\|_{\max}, \tag{S.15}$$

where (a) follows from Assumption 3 and (b) follows from the definition of $\|\Theta^{(1)}\|_{\max}$.

Then,

$$\mathcal{E}\big(\widehat{\Theta}^{(1)}\big) \overset{(a)}{\leq} \|\widehat{\Theta}^{(1)} - \Theta^{(1)}\|_{\max} \overset{(b)}{\leq} \frac{\|\Theta^{(1)}\|_{\max}}{\bar{\lambda}} \|\widehat{P} - P\|_{\max} \overset{(c)}{=} \frac{\|\Theta^{(1)}\|_{\max}}{\bar{\lambda}} O_p\bigg(\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{M}}\bigg),$$

where (a) follows from the definition of $(1,2)$ operator norm, (b) follows from Eq. (S.15), and (c) follows from Eq. (S.13). Then, Eq. (S.14) follows as $1/\bar{\lambda}$ and $\|\Theta^{(1)}\|_{\max}$ are assumed to be bounded.

In the second case, using Eqs. (2) and (3) to expand $Y^{(1),\text{full}}$, we have

$$Y^{(1),\text{full}} = \Theta^{(1)} \odot P + \Theta^{(1)} \odot W + E^{(1)} \odot P + E^{(1)} \odot W.$$

Next, we utilize two claims proven in Sections S5.2.3 and S5.2.4 respectively: $\Theta^{(1)} \odot P$ satisfies Assumption S1 with rank parameter $r_{\theta_1} r_p$ and

$$\overline{\varepsilon}^{(1)} \triangleq \Theta^{(1)} \odot W + E^{(1)} \odot P + E^{(1)} \odot W,$$

satisfies Assumption S2. Finally, Assumption 8 in Section 5 of the main article implies that Assumption S3 holds.

Now, note that step 5 of `Cross-Fitted-SVD` can be rewritten as $\widehat{\Theta}^{(1)} = \overline{\Theta}^{(1)} \oslash \widehat{P}$ and $\overline{\Theta}^{(1)} = \texttt{Cross-Fitted-MC}(\texttt{TW}_{r_3}, Y^{(1),\text{full}}, \mathcal{P})$ where $r_3 = r_{\theta_1} r_p$. Then, from Corollary S3,

$$\|\overline{\Theta}^{(1)}_{\mathcal{I}} - \Theta^{(1)}_{\mathcal{I}} \odot P_{\mathcal{I}}\|_{\max} = O_p\bigg(\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{M}}\bigg).$$

Applying a union bound over all $\mathcal{I} \in \mathcal{P}$ and noting that $\overline{\Theta}^{(1)} = \widehat{\Theta}^{(1)} \odot \widehat{P}$, we have

$$\|\widehat{\Theta}^{(1)} \odot \widehat{P} - \Theta^{(1)} \odot P\|_{\max} = O_p\bigg(\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{M}}\bigg). \tag{S.16}$$

The left hand side of Eq. (S.16) can be written as,

$$\begin{aligned}
\|\widehat{\Theta}^{(1)} \odot \widehat{P} - \Theta^{(1)} \odot P\|_{\max} &= \|\widehat{\Theta}^{(1)} \odot \widehat{P} - \Theta^{(1)} \odot \widehat{P} + \Theta^{(1)} \odot \widehat{P} - \Theta^{(1)} \odot P\|_{\max} \\
&\overset{(a)}{\geq} \|\big(\widehat{\Theta}^{(1)} - \Theta^{(1)}\big) \odot \widehat{P}\|_{\max} - \|\Theta^{(1)} \odot \big(\widehat{P} - P\big)\|_{\max} \\
&\overset{(b)}{\geq} \bar{\lambda} \|\widehat{\Theta}^{(1)} - \Theta^{(1)}\|_{\max} - \|\Theta^{(1)}\|_{\max} \|\widehat{P} - P\|_{\max}, \qquad \text{(S.17)}
\end{aligned}$$

where (a) follows from triangle inequality as $\|\big(\widehat{\Theta}^{(1)} - \Theta^{(1)}\big) \odot \widehat{P}\|_{\max} \geq \|\Theta^{(1)} \odot \big(\widehat{P} - P\big)\|_{\max}$ and (b) follows from the choice of $\bar{\lambda}$ and the definition of $\|\Theta^{(1)}\|_{\max}$. Then,

$$\mathcal{E}\big(\widehat{\Theta}^{(1)}\big) \overset{(a)}{\leq} \|\widehat{\Theta}^{(1)} - \Theta^{(1)}\|_{\max} \overset{(b)}{\leq} \frac{1}{\bar{\lambda}} \|\widehat{\Theta}^{(1)} \odot \widehat{P} - \Theta^{(1)} \odot P\|_{\max} + \frac{\|\Theta^{(1)}\|_{\max}}{\bar{\lambda}} \|\widehat{P} - P\|_{\max}$$

$$\overset{(c)}{=} \frac{1}{\bar{\lambda}} O_p\Big(\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{M}}\Big) + \frac{\|\Theta^{(1)}\|_{\max}}{\bar{\lambda}} O_p\Big(\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{M}}\Big),$$

where (a) follows from the definition of $L_{1,2}$ norm, (b) follows from Eq. (S.17), and (c) follows from Eqs. (S.13) and (S.16). Then, Eq. (S.14) follows as $1/\bar{\lambda}$ and $\|\Theta^{(1)}\|_{\max}$ are assumed to be bounded.

S5.2.3. Proof that $\Theta^{(0)} \odot (\mathbf{1} - P)$ and $\Theta^{(1)} \odot P$ satisfy Assumption S1.
First, we show that $\overline{U}^{(0)} \in \mathbb{R}^{N \times r_{\theta_0}(r_p+1)}$ and $\overline{V}^{(0)} \in \mathbb{R}^{N \times r_{\theta_0}(r_p+1)}$ are factors of $\Theta^{(0)} \odot (\mathbf{1} - P)$, and $\overline{U}^{(1)} \in \mathbb{R}^{N \times r_{\theta_1} r_p}$ and $\overline{V}^{(1)} \in \mathbb{R}^{N \times r_{\theta_1}}$ are factors of $\Theta^{(1)} \odot P$ as claimed in Eq. (35). We have

$$\Theta^{(1)} \odot P = \Big(\sum_{i \in [r_{\theta_1}]} U^{(1)}_{i,\cdot} V^{(1)\top}_{i,\cdot}\Big) \odot \Big(\sum_{j \in [r_p]} U_{j,\cdot} V^{\top}_{j,\cdot}\Big) = \sum_{i \in [r_{\theta_1}]} \sum_{j \in [r_p]} \Big(U^{(1)}_{i,\cdot} \odot U_{j,\cdot}\Big)\Big(V^{(1)}_{i,\cdot} \odot V_{j,\cdot}\Big)^{\top}$$
$$\overset{(a)}{=} \big(U * U^{(1)}\big)\big(V * V^{(1)}\big)^{\top} \overset{(b)}{=} \overline{U}^{(1)} \overline{V}^{(1)\top},$$

where (a) follows from the definition of Khatri-Rao product (see Section 1 of the main article) and (b) follows from the definitions of $\overline{U}^{(1)}$ and $\overline{V}^{(1)}$. The proof for $\Theta^{(0)} \odot (\mathbf{1} - P)$ follows similarly. Then, Assumption S1(a) holds from Eq. (35). Next, we note that

$$\|\overline{U}^{(1)}\|_{2,\infty} = \|U * U^{(1)}\|_{2,\infty} \overset{(a)}{=} \max_{i \in [N]} \sqrt{\sum_{j \in [r_p]} u^2_{i,j} \sum_{j' \in [r_{\theta_1}]} (u^{(1)}_{i,j'})^2} \leq \|U\|_{2,\infty} \|U^{(1)}\|_{2,\infty} \overset{(b)}{\leq} c,$$

where (a) follows from the definition of Khatri-Rao product (see Section 1 of the main article), and (b) follows from Assumption 7. Then, $\Theta^{(1)} \odot P$ satisfies Assumption S1(b) by using similar arguments on $\overline{V}^{(1)}$. Further, $\Theta^{(0)} \odot (\mathbf{1} - P)$ satisfies Assumption S1(b) by noting that $\|\overline{U}\|_{2,\infty}$ and $\|\overline{V}\|_{2,\infty}$ are bounded whenever $\|U\|_{2,\infty}$ and $\|V\|_{2,\infty}$ are bounded, respectively. Finally, Assumption S1(c) holds from Assumption 7.

S5.2.4. Proof that $\overline{\varepsilon}^{(1)}$ satisfies Assumption S2
Recall that $\overline{\varepsilon}^{(1)} \triangleq \Theta^{(1)} \odot W + E^{(1)} \odot P + E^{(1)} \odot W$. Then, Assumption S2(a) holds as $\overline{\varepsilon}^{(1)}_{i,j}$ is zero-mean from Assumption 2 and Eq. (3), and $\overline{\varepsilon}^{(1)}_{i,j}$ is subExponential because $\varepsilon^{(1)}_{i,j} \eta_{i,j}$ is a subExponential random variable Lemma S3, every subGaussian random variable is subExponential random variable, and sum of subExponential random variables is a subExponential random variable. Finally, Assumption S2(b) and Assumption S2(b) hold from Assumption 9(b) and Assumption 9(c), respectively.

S5.3. Proof of Corollary S3

Corollary S3 is a direct application of Bai and Ng (2021, Lemma A.1), specialized to our setting. Notably, Bai and Ng (2021) make three assumptions numbered A, B, and C in their paper to establish the corresponding result. It remains to establish that the conditions assumed in Corollary S3 imply the necessary conditions used in the proof of Bai and Ng (2021, Lemma A.1). First, note that certain assumptions in Bai and Ng (2021) are not actually used in their proof of Lemma A.1 (or in the proof of other results used in that proof), namely, the distinct eigenvalue condition in Assumption A(a)(iii), the asymptotic normality conditions in Assumption A(c) and the asymptotic normality conditions in Assumption C. Next, Eq. (S.12) in Assumption S3 implies Assumption B and Eq. (S.11) in Assumption S3 is equivalent to the remaining conditions in Assumption C.

It remains to show how Assumptions S1 and S2 imply the remainder of conditions in Bai and Ng (2021, Assumptions A). For completeness, these conditions are collected in the following assumption.

Assumption S4. *The noise matrix H is such that,*

(a) $\max_{j\in[M]} \frac{1}{N} \sum_{j'\in[M]} \left| \sum_{i\in[N]} \mathbb{E}[h_{i,j}h_{i,j'}] \right| \leq c,$

(b) $\max_{j\in[M]} \left|\mathbb{E}[h_{i,j}h_{i',j}]\right| \leq c_{i,i'}$ *and* $\max_{i\in[N]} \sum_{i'\in[N]} c_{i,i'} \leq c,$

(c) $\frac{1}{NM} \sum_{i,i'\in[N]} \sum_{j,j'\in[M]} \left|\mathbb{E}[h_{i,j}h_{i',j'}]\right| \leq c,$ *and*

(d) $\max_{j,j'\in[M]} \frac{1}{N^2} \mathbb{E}\big[\big| \sum_{i\in[N]} \big(h_{i,j}h_{i,j'} - \mathbb{E}[h_{i,j}h_{i,j'}]\big) \big|^4 \big].$

Assumption S4 is a restatement of the subset of conditions from Bai and Ng (2021, Assumption A) necessary in Bai and Ng (2021, proof of Lemma A.1) and it essentially requires weak dependence in the noise across measurements and across units. In particular, Assumption S4(a), (b), (c), and (d) correspond to Assumption A(b)(ii), (iii), (iv), (v), respectively, of Bai and Ng (2021). For the other conditions in Bai and Ng (2021, Assumption A), note that Assumption S1 above is equivalent to their Assumption A(a)(i) and (ii) of Bai and Ng (2021) when the factors are non-random as in this work. Similarly, Assumption S2(a) above is analogous to Assumption A(b)(i) of Bai and Ng (2021). Assumption A(b)(vi) of Bai and Ng (2021) is implied by their other Assumptions for non-random factors as stated in Bai (2003).

To establish Corollary S3, it remains to establish that Assumption S4 holds, which is done in Section S5.3.1 below.

S5.3.1. Assumption S4 holds

First, Assumption S4(a) holds as follows,

$$\max_{j\in[M]} \frac{1}{N} \sum_{j'\in[M]} \Big| \sum_{i\in[N]} \mathbb{E}\big[h_{i,j}h_{i,j'}\big] \Big| \overset{(a)}{\leq} \max_{j\in[M]} \frac{1}{N} \sum_{i\in[N]} \sum_{j'\in[M]} \Big| \mathbb{E}\big[h_{i,j}h_{i,j'}\big] \Big| \overset{(b)}{\leq} \max_{j\in[M]} \frac{1}{N} \sum_{i\in[N]} c = c,$$

where (a) follows from triangle inequality and (b) follows from Assumption S2(b). Next, from Assumption S2(a) and Assumption S2(c), we have

$$\max_{j\in[M]} \left|\mathbb{E}[h_{i,j}h_{i',j}]\right| = \begin{cases} 0 & \text{if } i \neq i' \\ \max_{j\in[M]} \left|\mathbb{E}[h_{i,j}^2]\right| \leq c & \text{if } i = i' \end{cases}$$

Then, Assumption S4(b) holds as $\max_{i\in[N]} \max_{j\in[M]} \sum_{i'\in[N]} \left|\mathbb{E}[h_{i,j}h_{i',j}]\right| \leq c$. Next, Assumption S4(c) holds as follows,

$$\frac{1}{NM} \sum_{i,i'\in[N]} \sum_{j,j'\in[M]} \left|\mathbb{E}[h_{i,j}h_{i',j'}]\right| \overset{(a)}{=} \frac{1}{NM} \sum_{i\in[N]} \sum_{j,j'\in[M]} \left|\mathbb{E}[h_{i,j}h_{i,j'}]\right| \overset{(b)}{\leq} \frac{1}{NM} \sum_{i\in[N]} \sum_{j\in[M]} c = c,$$

where (a) follows from Assumption S2(c) and (b) follows from Assumption S2(b). Next, let $\gamma_{i,j,j'} \triangleq h_{i,j}h_{i,j'} - \mathbb{E}[h_{i,j}h_{i,j'}]$ and fix any $j, j' \in [M]$. Then, Assumption S4(d) holds as follows,

$$\begin{aligned} \frac{1}{N^2}\mathbb{E}\Big[\Big(\sum_{i\in[N]} \gamma_{i,j,j'}\Big)^4\Big] &= \frac{1}{N^2}\mathbb{E}\Big[\Big(\sum_{i_1\in[N]} \gamma_{i_1,j,j'}\Big)\Big(\sum_{i_2\in[N]} \gamma_{i_2,j,j'}\Big)\Big(\sum_{i_3\in[N]} \gamma_{i_3,j,j'}\Big)\Big(\sum_{i_4\in[N]} \gamma_{i_4,j,j'}\Big)\Big] \\ &\overset{(a)}{=} \frac{1}{N^2} \sum_{i\in[N]} \mathbb{E}\Big[\gamma_{i,j,j'}^4\Big] + \frac{3}{N^2} \sum_{i\neq i'\in[N]} \mathbb{E}\Big[\gamma_{i,j,j'}^2\gamma_{i',j,j'}^2\Big] \leq c, \end{aligned}$$

where (a) follows from linearity of expectation and Assumption S2(c) after by noting that $\mathbb{E}[\gamma_{i,j,j'}] = 0$ for all $i, j, j' \in [N] \times [M] \times [M]$ and (b) follows because $\gamma_{i,j,j'}$ has bounded moments due to Assumption S2(a).

S6. Doubly-robust estimation in panel data with lagged effects

This section describes how the doubly-robust framework of this article can be generalized to a panel data setting with lagged treatment effects. We highlight that, as is the convention in a panel data setting, t denotes the column (time) index and T denotes the total number of columns (time periods).

S6.1. Setup

As described in Section 4.4 of the main article, potential outcomes are generated as follows: for all $i \in [N], t \in [T]$, and $a \in \{0, 1\}$,

$$y_{i,t}^{(a|y_{i,t-1})} = \alpha^{(a)} y_{i,t-1} + \theta_{i,t}^{(a)} + \varepsilon_{i,t}^{(a)}, \tag{S.18}$$

where $y_{i,t}^{(a|y_{i,t-1})}$ is the potential outcome for unit i at time t given treatment $a \in \{0, 1\}$ and lagged outcome $y_{i,t-1}$. This model combines unobserved confounding and lagged treatment effects, where the lagged effect is carried over via the auto-regressive term,

$\alpha^{(a)} y_{i,t-1}$, with $\alpha^{(a)}$ being the auto-regressive parameter for treatment $a \in \{0,1\}$. The treatment possibly starts at $t = 1$, and $y_{i,0}$ is assumed to not be affected by any future exposure to the treatment. Treatment assignments are continually assumed to be generated via Eq. (3). As in Eq. (1), realized outcomes, $y_{i,t}$, depend on potential outcomes and treatment assignments,

$$y_{i,t} = y_{i,t}^{(0|y_{i,t-1})}(1 - a_{i,t}) + y_{i,t}^{(1|y_{i,t-1})} a_{i,t}, \tag{S.19}$$

for all $i \in [N]$ and $t \in [T]$.

S6.2. Target causal estimand

The lagged effects in Eq. (S.18) imply that the treatment effects need to be defined for sequences of treatments. For concreteness, consider the effect at time T for an always-treat policy, i.e., $a_{i,t} = 1$, versus never-treat, i.e., $a_{i,t} = 0$, for $i \in [N]$ and $j \in [T]$. Let $y_{i,T}^{[1]}$ be the potential outcome for unit i at time T under always-treat and $y_{i,T}^{[0]}$ be the potential outcome for unit i at time T under never-treat. We aim to estimate the difference in the expected potential outcomes under these two treatment policies averaged over all units,

$$\text{ATE}_{\cdot,T} \triangleq \mu_{\cdot,T}^{[1]} - \mu_{\cdot,T}^{[0]}, \quad \text{where} \quad \mu_{\cdot,T}^{[a]} \triangleq \frac{1}{N} \sum_{i \in [N]} \mathbb{E}[y_{i,T}^{[a]}],$$

with the expectation taken over the distribution of $\{\varepsilon_{i,t}^{(a)}\}_{i \in [N], t \in [T]}$, conditioned on the initial outcomes $\{y_{i,0}\}_{i \in [N]}$. We make the following assumption about the noise in potential outcomes.

Assumption S5 (Zero-mean noise conditioned on the initial outcomes)**.** $\{\varepsilon_{i,t}^{(a)} : i \in [N], t \in [T], a \in \{0,1\}\}$ *are mean zero conditioned on* $\{y_{i,0}\}_{i \in [N]}$.

Assumption S5 holds whenever Assumption 2(a) holds conditioned on the initial outcomes $\{y_{i,0}\}_{i \in [N]}$. Another sufficient condition for Assumption S5 is that $(\varepsilon_{i,t}^{(0)}, \varepsilon_{i,t}^{(1)})$ are independent in time. Given this, the time dependence in the expected potential outcome $\mathbb{E}[y_{i,T}^{[a]}]$ is captured as follows: for $a \in \{0,1\}$

$$\mathbb{E}[y_{i,T}^{[a]}] = (\alpha^{(a)})^T y_{i,0} + \sum_{s=0}^{T-1} (\alpha^{(a)})^s \theta_{i,T-s}^{(a)}. \tag{S.20}$$

Eq. (S.20) forms the basis of our doubly-robust estimator of $\text{ATE}_{\cdot,T}$.

We chose the contrast between always-treat and never-treat for concreteness. However, the framework and the results in this section can be generalized in a straightforward manner to contrast any two pre-specified sequences of treatments, where the treatment can also be chosen stochastically with pre-specified probabilities.

For the remainder of this section, we condition on the initial outcomes $\{y_{i,0}\}_{i\in[N]}$ but omit it from our notation for brevity.

S6.3. Doubly-robust estimator

The DR estimator of $\text{ATE}_{\cdot,T}$ combines the estimates of $(\alpha^{(0)}, \alpha^{(1)})$, $(\Theta^{(0)}, \Theta^{(1)})$, and P. First, we obtain the estimates $(\widehat{\alpha}^{(0)}, \widehat{\alpha}^{(1)})$. These estimates can be computed using the likelihood approach of Bai (2024) whenever there exists some units such that they all have treatment a for some consecutive time points, for $a \in \{0, 1\}$.

Next, we define the residual matrices $\widetilde{Y}^{(0),\text{obs}}$ and $\widetilde{Y}^{(1),\text{obs}}$. Let $\widetilde{Y}^{(0),\text{obs}} \in \{\mathbb{R} \cup \{\,?\}\}^{N\times T}$ be a matrix with (i,t)-th entry equal to $y_{i,t} - \widehat{\alpha}^{(0)} y_{i,t-1}$ if $a_{i,t} = 0$, and equal to ? otherwise. Analogously, let $\widetilde{Y}^{(1),\text{obs}} \in \{\mathbb{R} \cup \{\,?\}\}^{N\times T}$ be a matrix with (i,t)-th entry equal to $y_{i,t} - \widehat{\alpha}^{(1)} y_{i,t-1}$ if $a_{i,t} = 1$, and equal to ? otherwise. Then, similar to Eq. (8), the application of matrix completion yields the following estimates:

$$\widehat{\Theta}^{(0)} = \texttt{MC}(\widetilde{Y}^{(0),\text{obs}}), \quad \widehat{\Theta}^{(1)} = \texttt{MC}(\widetilde{Y}^{(1),\text{obs}}), \quad \text{and} \quad \widehat{P} = \texttt{MC}(A). \tag{S.21}$$

Then, the DR estimate is defined as follows:

$$\widehat{\text{ATE}}^{\text{DR}}_{\cdot,T,J} \triangleq \widehat{\mu}^{[1,\text{DR}]}_{\cdot,T,J} - \widehat{\mu}^{[0,\text{DR}]}_{\cdot,T,J} \quad \text{where} \quad \widehat{\mu}^{[a,\text{DR}]}_{\cdot,T,J} = \frac{1}{N}\sum_{i\in[N]}\left[(\widehat{\alpha}^{(a)})^T y_{i,0} + \sum_{s=0}^{J-1}(\widehat{\alpha}^{(a)})^s \widehat{\theta}^{[a,\text{DR}]}_{i,T-s}\right], \tag{S.22}$$

where

$$\widehat{\theta}^{[0,\text{DR}]}_{i,T-s} \triangleq \widehat{\theta}^{(0)}_{i,T-s} + \big(y_{i,T-s} - \widehat{\alpha}^{(0)} y_{i,T-s-1} - \widehat{\theta}^{(0)}_{i,T-s}\big)\frac{1 - a_{i,T-s}}{1 - \widehat{p}_{i,T-s}},$$

and

$$\widehat{\theta}^{[1,\text{DR}]}_{i,T-s} \triangleq \widehat{\theta}^{(1)}_{i,T-s} + \big(y_{i,T-s} - \widehat{\alpha}^{(1)} y_{i,T-s-1} - \widehat{\theta}^{(1)}_{i,T-s}\big)\frac{a_{i,T-s}}{\widehat{p}_{i,T-s}}$$

The estimator is parameterized by an integer J, which denotes the contiguous number of time periods preceding time T that are used to estimate the expectations at time T (see the summation in Eq. (S.20)). Notably, using preceding J terms instead of $T-1$ terms allows us to adapt cross-fitting for the setting with lagged treatment effects. Let us briefly elaborate: suppose $(\widehat{\alpha}^{(0)}, \widehat{\alpha}^{(1)})$ are estimated from entries of Y in $[N] \times [L]$ for some $L < T - J$. Consider the column partitions $\mathcal{C}_0 = \{L+1, \ldots, T-J\}$ and $\mathcal{C}_1 = \{T-J+1, \ldots, T\}$ of times $[T] \setminus [L]$. Suppose Eqs. (27) and (28) in Assumption 5 hold for $\mathcal{I} = \mathcal{R}_0 \times \mathcal{C}_1$ and $\mathcal{I} = \mathcal{R}_1 \times \mathcal{C}_1$ for some row partitions $\mathcal{R}_0$ and $\mathcal{R}_1$ of units $[N]$. Then, applying `Cross-Fitted-MC` on the residual matrices $\widetilde{Y}^{(0),\text{obs}}$ and $\widetilde{Y}^{(1),\text{obs}}$ with row partitions $(\mathcal{R}_0, \mathcal{R}_1)$ and column partitions $(\mathcal{C}_0, \mathcal{C}_1)$ ensures that Assumption 4 holds for every column in $\mathcal{C}_1$ with row partitions $(\mathcal{R}_0, \mathcal{R}_1)$.

S6.4. Non-asymptotic guarantees

Recall the notation for $\mathcal{E}\big(\widehat{\Theta}\big)$ and $\mathcal{E}\big(\widehat{P}\big)$ from Eq. (16) and define

$$\mathcal{E}(\widehat{\alpha}) \triangleq \sum_{a\in\{0,1\}} \mathcal{E}(\widehat{\alpha}^{(a)}) \quad \text{where} \quad \mathcal{E}(\widehat{\alpha}^{(a)}) \triangleq |\widehat{\alpha}^{(a)} - \alpha^{(a)}|. \tag{S.23}$$

Our analysis makes two additional assumptions to state a non-asymptotic error bound for $\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,T,J} - \mathrm{ATE}_{\cdot,T}$.

Assumption S6 (Bounded auto-regressive parameters and estimates)**.** *The auto-regressive parameters and their estimates are such that $|\alpha^{(a)}| \leq \overline{\alpha}$ and $|\widehat{\alpha}^{(a)}| \leq \overline{\alpha}$, for all $a \in \{0,1\}$, where $\overline{\alpha} \in [0,1)$.*

Assumption S6 requires the regression parameters to be bounded by a fixed constant less than 1. This condition is standard for auto-regressive models, as it implies stability of the outcome process in Eq. (S.18). The analogous condition on the estimated parameters can be ensured by truncating the estimates to $[0, \overline{\alpha}]$.

Assumption S7 (Bounded observed outcomes, mean potential outcomes, and estimated mean potential outcomes)**.** *The observed outcomes, the mean potential outcomes, and the estimates of the mean potential outcomes are such that $|y_{i,t}| \leq C_1$, $|\theta^{(a)}_{i,t}| \leq C_2$, and $|\widehat{\theta}^{(a)}_{i,t}| \leq C_3$, for all $i \in [N]$, $j \in [M]$, and $a \in \{0,1\}$, where C_1, C_2, and C_3 are universal constants.*

Assumption S7 requires the observed outcomes, the mean potential outcomes, and the estimates of the mean potential outcomes to be bounded to simplify our proof. With a more delicate analysis, Assumption S7 can be relaxed to require the average observed outcomes over $i \in [N]$, the average mean potential outcomes over $i \in [N]$, and the average estimated mean potential outcomes over $i \in [N]$ to be bounded.

Theorem S1 (Finite Sample Guarantees for DR with lagged effects)**.** *Consider the panel data model with lagged effects defined via Eqs.* (S.18) *and* (S.19)*. Suppose Assumptions 1 to 3, S6, and S7 hold and Assumption 4 holds for $t \in \{T-J+1, \ldots, T\}$ for some integer $J \in [T]$. Fix $\delta \in (0,1)$. Then, with probability at least $1-\delta$, we have*

$$\big|\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,T,J} - \mathrm{ATE}_{\cdot,T}\big| \leq \frac{\mathrm{Err}^{\mathrm{DR}}_{N,\delta/J}}{1-\overline{\alpha}} + C\left[\frac{\overline{\alpha}^J}{1-\overline{\alpha}} + \mathcal{E}(\widehat{\alpha})\Big(T\overline{\alpha}^{T-1} + \frac{1}{1-\overline{\alpha}}\Big)\right], \tag{S.24}$$

for $\mathrm{Err}^{\mathrm{DR}}_{N,\delta}$ as defined in Eq. (18) *in Theorem 1 and a universal constant C.*

The proof of Theorem S1 is given in Section S6.5. For brevity, the finite sample guarantees above uses $\mathcal{E}\big(\widehat{\Theta}\big)$ and $\mathcal{E}\big(\widehat{P}\big)$ as defined in Eq. (16), but the proof can be easily modified to replace the $\max_{j\in[T]}$ appearing in the definition of $\|\cdot\|_{1,2}$ in Eq. (16) with $\max_{j\in\{T-J+1,\cdots,T\}}$.

Next, we remark that Theorem S1 is a strict generalization of Theorem 1. To this end, note that when $\alpha^{(a)} = 0$ for all $a \in \{0, 1\}$, the model considered in Theorem S1 simplifies to the model considered in Theorem 1. For this setting, the assumptions in Theorem 1 imply that the assumptions in Theorem S1 hold with $J = 1$. First, Assumption S6 holds with $\overline{\alpha} = 0$ when $\alpha^{(a)} = 0$ for all $a \in \{0, 1\}$. Second, the proof of Theorem S1 can be easily modified to drop the requirement of Assumption S7 when $J = 1$ and $\overline{\alpha} = 0$. Substituting $\overline{\alpha} = 0$, $\mathcal{E}(\widehat{\alpha}) = 0$ (i.e., the auto-regressive parameters are known to be 0), and $J = 1$ in Eq. (S.24) recovers the guarantee stated in Theorem 1.

Doubly-robust behavior of $\widehat{\text{ATE}}^{\text{DR}}_{\cdot,T,J}$. When $\overline{\alpha} \neq 0$ and bounded away from one, Eq. (S.24) bounds the absolute error of the DR estimator by the rate of

$$\mathcal{E}\big(\widehat{\Theta}\big)\Big(\mathcal{E}\big(\widehat{P}\big) + \sqrt{\frac{\log J}{N}}\Big) + \frac{1}{\sqrt{N}} + \overline{\alpha}^{J} + \mathcal{E}(\widehat{\alpha}).$$

Then, if the conditions of Theorem S1 are satisfied for some J such that $C \log N \geq J \geq \log N/(2\log(1/\overline{\alpha}))$, the error rate of the DR estimator is bounded by

$$\mathcal{E}\big(\widehat{\Theta}\big)\Big(\mathcal{E}\big(\widehat{P}\big) + \sqrt{\frac{\log\log N}{N}}\Big) + \frac{1}{\sqrt{N}} + \mathcal{E}(\widehat{\alpha}),$$

which decays a parametric rate of $O_p(N^{-0.5})$ as long as

$$\mathcal{E}\big(\widehat{\Theta}\big)\mathcal{E}\big(\widehat{P}\big) = O_p\Big(\frac{1}{\sqrt{N}}\Big), \quad \mathcal{E}\big(\widehat{\Theta}\big) = O_p\Big(\frac{1}{\sqrt{\log\log N}}\Big), \quad \text{and} \quad \mathcal{E}(\widehat{\alpha}) = O_p\Big(\frac{1}{\sqrt{N}}\Big).$$

Note that Proposition 4 still implies that `Cross-Fitted-SVD` achieves $\mathcal{E}\big(\widehat{P}\big) = O_p(N^{-0.5} + T^{-0.5})$ under suitable conditions. To estimate the auto-regressive parameter $\alpha^{(a)}$ for $a \in \{0, 1\}$, Bai (2024, Section 5) shows that whenever there exist K units such that they all have treatment a for L consecutive time points, a full information maximum likelihood estimator provides $|\alpha^{(a)} - \widehat{\alpha}^{(a)}| = O_p((KL)^{-0.5})$. Next, establishing a matrix completion guarantee for the mean potential outcomes by residualizing as in Eq. (S.21) can be reduced to deriving a matrix completion guarantee for an approximately low-rank matrix. To this end, Agarwal and Singh (2024, Theorem 5) suggests that, up to logarithmic factors, an error rate of $N^{-0.5} + T^{-0.5} + \mathcal{E}(\widehat{\alpha})$ is plausible for $\mathcal{E}\big(\widehat{\Theta}\big)$ for our setting. A complete derivation of error guarantees for $\mathcal{E}(\widehat{\alpha})$ and $\mathcal{E}\big(\widehat{\Theta}\big)$ in the dynamic model is an interesting venue for future work.

S6.5. Proof of Theorem S1: Finite Sample Guarantees for DR with lagged effects

The error $\Delta\mathrm{ATE}^{\mathrm{DR}}_{\cdot,T} = \widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,T,J} - \mathrm{ATE}_{\cdot,T}$ can be re-expressed as

$$\Delta\mathrm{ATE}^{\mathrm{DR}}_{\cdot,T} = \big(\widehat{\mu}^{[1,\mathrm{DR}]}_{\cdot,T,J} - \widehat{\mu}^{[0,\mathrm{DR}]}_{\cdot,T,J}\big) - \big(\mu^{[1]}_{\cdot,T} - \mu^{[0]}_{\cdot,T}\big) = \big(\widehat{\mu}^{[1,\mathrm{DR}]}_{\cdot,T,J} - \mu^{[1]}_{\cdot,T}\big) - \big(\widehat{\mu}^{[0,\mathrm{DR}]}_{\cdot,T,J} - \mu^{[0]}_{\cdot,T}\big). \quad \text{(S.25)}$$

We claim that, with probability at least $1-\delta$,

$$\Big|\widehat{\mu}^{[1,\mathrm{DR}]}_{\cdot,T,J} - \mu^{[1]}_{\cdot,T}\Big| \le C\left[\frac{|\alpha^{(1)}|^J - |\alpha^{(1)}|^T}{1-|\alpha^{(1)}|} + \mathcal{E}(\widehat{\alpha}^{(1)})\Big(T\overline{\alpha}^{T-1} + \frac{1-|\alpha^{(1)}|^J}{1-|\alpha^{(1)}|} + \frac{1}{(1-|\alpha^{(1)}|)^2}\Big)\right]$$
$$+\frac{2}{(1-|\alpha^{(1)}|)\bar{\lambda}}\left[\mathcal{E}(\widehat{\Theta}^{(1)})\mathcal{E}\big(\widehat{P}\big) + \frac{1}{\sqrt{N}}\Big(\frac{\sqrt{c\ell_{\delta/(12J)}}}{\sqrt{\ell_1}}\mathcal{E}(\widehat{\Theta}^{(1)}) + 2\overline{\sigma}\sqrt{c\ell_{\delta/(12J)}} + \frac{2\overline{\sigma}m(c\ell_{\delta/(12J)}}{\sqrt{\ell_1}}\Big)\right], \quad \text{(S.26)}$$

and

$$\Big|\widehat{\mu}^{[0,\mathrm{DR}]}_{\cdot,T,J} - \mu^{[0]}_{\cdot,T}\Big| \le C\left[\frac{|\alpha^{(0)}|^J - |\alpha^{(0)}|^T}{1-|\alpha^{(0)}|} + \mathcal{E}(\widehat{\alpha}^{(0)})\Big(T\overline{\alpha}^{T-1} + \frac{1-|\alpha^{(0)}|^J}{1-|\alpha^{(0)}|} + \frac{1}{(1-|\alpha^{(0)}|)^2}\Big)\right]$$
$$+\frac{2}{(1-|\alpha^{(0)}|)\bar{\lambda}}\left[\mathcal{E}(\widehat{\Theta}^{(0)})\mathcal{E}\big(\widehat{P}\big) + \frac{1}{\sqrt{N}}\Big(\frac{\sqrt{c\ell_{\delta/(12J)}}}{\sqrt{\ell_1}}\mathcal{E}(\widehat{\Theta}^{(0)}) + 2\overline{\sigma}\sqrt{c\ell_{\delta/(12J)}} + \frac{2\overline{\sigma}m(c\ell_{\delta/(12J)}}{\sqrt{\ell_1}}\Big)\right]. \quad \text{(S.27)}$$

Then, the claim in Eq. (S.24) follows by applying triangle inequality in Eq. (S.25) and using Assumption S6. We prove the bound (S.26) in Section S6.5.1, and also provide an expression for C. The proof of Eq. (S.27) follows similarly.

S6.5.1. Proof of Eq. (S.26)

We start by decomposing $\mu^{[1]}_{\cdot,T}$ as follows:

$$\mu^{[1]}_{\cdot,T} = \frac{1}{N}\left[\sum_{i\in[N]}(\alpha^{(1)})^T y_{i,0} + \sum_{s=0}^{T-1}(\alpha^{(1)})^s \sum_{i\in[N]}\theta^{(1)}_{i,T-s}\right] = \mathbb{T}^{(1)}_J + \mathbb{U}^{(1)}_J + \mathbb{V}^{(1)},$$

where

$$\mathbb{T}^{(1)}_J \triangleq \frac{1}{N}\sum_{s=0}^{J-1}(\alpha^{(1)})^s \sum_{i\in[N]}\theta^{(1)}_{i,T-s}, \quad \mathbb{U}^{(1)}_J \triangleq \frac{1}{N}\sum_{s=J}^{T-1}(\alpha^{(1)})^s \sum_{i\in[N]}\theta^{(1)}_{i,T-s}, \quad \text{(S.28)}$$

and

$$\mathbb{V}^{(1)} \triangleq (\alpha^{(1)})^T \frac{1}{N}\sum_{i\in[N]} y_{i,0}. \quad \text{(S.29)}$$

Next, we decompose $\widehat{\mu}^{[1,\mathrm{DR}]}_{\cdot,T,J}$ in Eq. (S.22) as $\widehat{\mu}^{[1,\mathrm{DR}]}_{\cdot,T,J} = \widehat{\mathbb{T}}^{(1)}_J + \widehat{\mathbb{V}}^{(1)}$, where

$$\widehat{\mathbb{T}}^{(1)}_J \triangleq \frac{1}{N}\sum_{s=0}^{J-1}(\widehat{\alpha}^{(1)})^s \sum_{i\in[N]} \widehat{\theta}^{[1,\mathrm{DR}]}_{i,T-s}, \quad \text{and} \quad \widehat{\mathbb{V}}^{(1)} \triangleq (\widehat{\alpha}^{(1)})^T \frac{1}{N}\sum_{i\in[N]} y_{i,0}. \tag{S.30}$$

Finally, we define

$$\widetilde{\mathbb{T}}^{(1)}_J \triangleq \frac{1}{N}\sum_{s=0}^{J-1}(\alpha^{(1)})^s \sum_{i\in[N]} \left[\widehat{\theta}^{(1)}_{i,T-s} + \big(y_{i,T-s} - \alpha^{(1)} y_{i,T-s-1} - \widehat{\theta}^{(1)}_{i,T-s}\big)\frac{a_{i,T-s}}{\widehat{p}_{i,T-s}}\right], \tag{S.31}$$

which is similar to $\widehat{\mathbb{T}}^{(1)}_J$ except that $\widehat{\alpha}^{(1)}$ is replaced by $\alpha^{(1)}$. The proof proceeds by bounding each term in the following fundamental decomposition:

$$\widehat{\mu}^{[1,\mathrm{DR}]}_{\cdot,T,J} - \mu^{[1]}_{\cdot,T} = (\widehat{\mathbb{V}}^{(1)} - \mathbb{V}^{(1)}) + (\widetilde{\mathbb{T}}^{(1)}_J - \mathbb{T}^{(1)}_J) + (\widehat{\mathbb{T}}^{(1)}_J - \widetilde{\mathbb{T}}^{(1)}_J) - \mathbb{U}^{(1)}_J. \tag{S.32}$$

With $C_0 \triangleq \max_{i\in[N]} |y_{i,0}|$ and $C_{\mathrm{DR}} \triangleq C_3 + (2C_1 + C_3)/\bar{\lambda}$, we claim that the bounds

$$\big|\widehat{\mathbb{V}}^{(1)} - \mathbb{V}^{(1)}\big| \le C_0 T \mathcal{E}(\widehat{\alpha}^{(1)})\overline{\alpha}^{T-1}, \qquad |\mathbb{U}^{(1)}_J| \le C_2 \frac{|\alpha^{(1)}|^J - |\alpha^{(1)}|^T}{1-|\alpha^{(1)}|}, \tag{S.33}$$

and

$$|\widehat{\mathbb{T}}^{(1)}_J - \widetilde{\mathbb{T}}^{(1)}_J| \le \mathcal{E}(\widehat{\alpha}^{(1)})\bigg(\frac{C_1}{\lambda}\frac{(1-|\alpha^{(1)}|^J)}{1-|\alpha^{(1)}|} + C_{\mathrm{DR}}\frac{1}{(1-|\alpha^{(1)}|)^2}\bigg), \tag{S.34}$$

hold deterministically (conditioned on $\widehat{\alpha}^{(1)}$), and that the bound

$$\begin{aligned}|\widetilde{\mathbb{T}}^{(1)}_J - \mathbb{T}^{(1)}_J| \le \frac{2}{(1-|\alpha^{(1)}|)\bar{\lambda}}\bigg[&\mathcal{E}(\widehat{\Theta}^{(1)})\mathcal{E}\big(\widehat{P}\big) \\ &+ \Big(\frac{\sqrt{c\ell_{\delta/(12J)}}}{\sqrt{\ell_1}}\mathcal{E}(\widehat{\Theta}^{(1)}) + 2\overline{\sigma}\sqrt{c\ell_{\delta/(12J)}} + \frac{2\overline{\sigma}m(c\ell_{\delta/(12J)}}{\sqrt{\ell_1}}\Big)\frac{1}{\sqrt{N}}\bigg],\end{aligned} \tag{S.35}$$

holds with probability at least $1-\delta/2$. The claim in Eq. (S.26) follows by applying triangle inequality in Eq. (S.32) and using the above bounds.

It remains to establish the intermediate claims Eqs. (S.33) to (S.35). Throughout the rest of the proof, we repeatedly use the inequality below that holds for all $s \in [T]$:

$$\begin{aligned}\big|(\widehat{\alpha}^{(1)})^s - (\alpha^{(1)})^s\big| = \Big|(\widehat{\alpha}^{(1)} - \alpha^{(1)})\big(\sum_{l\in[s]}(\widehat{\alpha}^{(1)})^{s-l}(\alpha^{(1)})^{l-1}\big)\Big| &\overset{(a)}{\le} s\big|(\widehat{\alpha}^{(1)} - \alpha^{(1)})\big|\overline{\alpha}^{s-1} \\ &\overset{(b)}{=} s\mathcal{E}(\widehat{\alpha}^{(1)})\overline{\alpha}^{s-1},\end{aligned} \tag{S.36}$$

where (a) follows from Assumption S6 and (b) follows from Eq. (S.23).

Proof of Eq. (S.33) First, from Eq. (S.28), we have

$$|\mathbb{U}_J^{(1)}| = \Big|\frac{1}{N}\sum_{s=J}^{T-1}(\alpha^{(1)})^s \sum_{i\in[N]}\theta_{i,T-s}^{(1)}\Big| \overset{(a)}{\leq} C_2\sum_{s=J}^{T-1}\big|\alpha^{(1)}\big|^s \overset{(b)}{=} C_2\frac{|\alpha^{(1)}|^J - |\alpha^{(1)}|^T}{1-|\alpha^{(1)}|},$$

where (a) follows from Assumption S7 and (b) follows from the sum of geometric series. Next, from Eqs. (S.29) and (S.30), we have

$$\big|\widehat{\mathbb{V}}^{(1)} - \mathbb{V}^{(1)}\big| = \Big|\Big((\widehat{\alpha}^{(1)})^T - (\alpha^{(1)})^T\Big)\frac{1}{N}\sum_{i\in[N]} y_{i,0}\Big| \overset{(a)}{\leq} C_0 T\mathcal{E}(\widehat{\alpha}^{(1)})\overline{\alpha}^{T-1},$$

where (a) follows from the definition of C_0 and Eq. (S.36).

Proof of Eq. (S.34) From Eqs. (S.30) and (S.31), and triangle inequality, we have

$$\begin{aligned}
\big|\widehat{\mathbb{T}}_J^{(1)} - \widetilde{\mathbb{T}}_J^{(1)}\big| &\leq \frac{1}{N}\sum_{i\in[N]}\sum_{s=0}^{J-1}\Bigg|(\widehat{\alpha}^{(1)})^s\Big(\widehat{\theta}_{i,T-s}^{(1)} + \big(y_{i,T-s} - \widehat{\alpha}^{(1)}y_{i,T-s-1} - \widehat{\theta}_{i,T-s}^{(1)}\big)\frac{a_{i,T-s}}{\widehat{p}_{i,T-s}}\Big) \\
&\qquad\qquad - (\alpha^{(1)})^s\left(\widehat{\theta}_{i,T-s}^{(1)} + \big(y_{i,T-s} - \alpha^{(1)}y_{i,T-s-1} - \widehat{\theta}_{i,T-s}^{(1)}\big)\frac{a_{i,T-s}}{\widehat{p}_{i,T-s}}\right)\Bigg| \\
&= \frac{1}{N}\sum_{i\in[N]}\sum_{s=0}^{J-1}\Bigg|(\alpha^{(1)})^s(\alpha^{(1)} - \widehat{\alpha}^{(1)})y_{i,T-s-1}\frac{a_{i,T-s}}{\widehat{p}_{i,T-s}} + \Big((\widehat{\alpha}^{(1)})^s - (\alpha^{(1)})^s\Big) \\
&\qquad\qquad \cdot\left(\widehat{\theta}_{i,T-s}^{(1)} + \big(y_{i,T-s} - \widehat{\alpha}^{(1)}y_{i,T-s-1} - \widehat{\theta}_{i,T-s}^{(1)}\big)\frac{a_{i,T-s}}{\widehat{p}_{i,T-s}}\right)\Bigg| \\
&\overset{(a)}{\leq} \frac{1}{N}\sum_{i\in[N]}\sum_{s=0}^{J-1}\Big|\frac{C_1}{\lambda}|\alpha^{(1)}|^s\mathcal{E}(\widehat{\alpha}^{(1)}) + C_{\mathrm{DR}}s\mathcal{E}(\widehat{\alpha}^{(1)})\overline{\alpha}^{s-1}\Big| \\
&= \mathcal{E}(\widehat{\alpha}^{(1)})\left(\frac{C_1}{\lambda}\frac{(1-|\alpha^{(1)}|^J)}{1-|\alpha^{(1)}|} + C_{\mathrm{DR}}\frac{1}{(1-|\alpha^{(1)}|)^2}\right),
\end{aligned}$$

where (a) follows from Eq. (S.23), Assumptions 3 and S7, and because $\max_{i\in[N],t\in[T]}\big|\widehat{\theta}_{i,t}^{[1,\mathrm{DR}]}\big| \leq C_{\mathrm{DR}}$ from Assumptions 3, S6, and S7, and (b) follows from the sum of geometric and arithmetico-geometric sequences.

Proof of Eq. (S.35) We start by defining

$$\widetilde{\theta}_{i,T-s}^{[1,\mathrm{DR}]} \triangleq \widehat{\theta}_{i,T-s}^{(1)} + \big(y_{i,T-s} - \alpha^{(1)}y_{i,T-s-1} - \widehat{\theta}_{i,T-s}^{(1)}\big)\frac{a_{i,T-s}}{\widehat{p}_{i,T-s}}.$$

Then, from Eqs. (S.28) and (S.31), we have

$$|\widetilde{\mathbb{T}}_J^{(1)} - \mathbb{T}_J^{(1)}| = \Big| \sum_{s=0}^{J-1} (\alpha^{(1)})^s \frac{1}{N} \sum_{i \in [N]} (\widetilde{\theta}_{i,T-s}^{[1,\mathrm{DR}]} - \theta_{i,T-s}^{(1)}) \Big| \overset{(a)}{\leq} \sum_{s=0}^{J-1} |\alpha^{(1)}|^s \frac{1}{N} \Big| \sum_{i \in [N]} (\widetilde{\theta}_{i,T-s}^{[1,\mathrm{DR}]} - \theta_{i,T-s}^{(1)}) \Big|,$$

where (a) follows from triangle inequality. From Eqs. (3) and (S.18), we have

$$\widetilde{\theta}_{i,T-s}^{[1,\mathrm{DR}]} - \theta_{i,T-s}^{(1)} = \widehat{\theta}_{i,T-s}^{(1)} + (\theta_{i,T-s}^{(1)} + \varepsilon_{i,T-s}^{(1)} - \widehat{\theta}_{i,T-s}^{(1)}) \frac{p_{i,T-s} + \eta_{i,T-s}}{\widehat{p}_{i,T-s}} - \theta_{i,T-s}^{(1)}.$$

Then, the term $\widetilde{\theta}_{i,T-s}^{[1,\mathrm{DR}]} - \theta_{i,T-s}^{(1)}$ is analogous to the display Eq. (A.2) in the proof of Theorem 1. Following similar algebra as in Appendix A1, we first obtain

$$\widetilde{\theta}_{i,T-s}^{[1,\mathrm{DR}]} - \theta_{i,T-s}^{(1)} = \frac{(\widehat{\theta}_{i,T-s}^{(1)} - \theta_{i,T-s}^{(1)})(\widehat{p}_{i,T-s} - p_{i,T-s})}{\widehat{p}_{i,T-s}} - \frac{(\widehat{\theta}_{i,T-s}^{(1)} - \theta_{i,T-s}^{(1)})\eta_{i,T-s}}{\widehat{p}_{i,T-s}} + \frac{\varepsilon_{i,T-s}^{(1)} p_{i,T-s}}{\widehat{p}_{i,T-s}} + \frac{\varepsilon_{i,T-s}^{(1)} \eta_{i,T-s}}{\widehat{p}_{i,T-s}}.$$

Now, note that Assumption 4 holds for $j = T - s$ for all $s \in \{0, \ldots, J-1\}$. Hence, for any such s and for any $\delta \in (0,1)$, mimicking the derivation of Eq. (A.5) from Appendix A1, we obtain, with probability at least $1 - \delta/(2J)$,

$$\frac{1}{N} \Big| \sum_{i \in [N]} (\widetilde{\theta}_{i,T-s}^{[1,\mathrm{DR}]} - \theta_{i,T-s}^{(1)}) \Big| \leq \frac{2}{\bar{\lambda}} \mathcal{E}\big(\widehat{\Theta}^{(1)}\big) \mathcal{E}\big(\widehat{P}\big) + \frac{2\sqrt{c\ell_{\delta/(12J)}}}{\bar{\lambda}\sqrt{\ell_1 N}} \mathcal{E}\big(\widehat{\Theta}^{(1)}\big) + \frac{2\overline{\sigma}\sqrt{c\ell_{\delta/(12J)}}}{\bar{\lambda}\sqrt{N}} + \frac{2\overline{\sigma} m(c\ell_{\delta/(12J)})}{\bar{\lambda}\sqrt{\ell_1 N}}. \tag{S.37}$$

Finally, multiplying both sides of Eq. (S.37) by $(\alpha^{(1)})^s$, summing it over $s \in \{0, \ldots, J-1\}$, and using a union bound argument yields that the bound in Eq. (S.35) holds with probability at least $1 - \delta/2$.

S7. Doubly-robust estimation in panel data with staggered adoption

This section shows how to extend the doubly-robust framework of this article to a setting with panel data and staggered adoption. Recall (from Section S6) that for panel data, t denotes the column (time) index and T denotes the total number of columns (time periods). In a staggered adoption setting, 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 $t > t_i$. This defines the observed treatment assignment matrix A. As mentioned in Section 5.4 of the main article and illustrated in the example below, a staggered treatment assignment leads

to a heavy time-series dependence in $\{\eta_{i,t}\}_{t\in[T]}$.

Example S1 (Single adoption time)**.** *Consider a panel data setting where all units remain in the control group until time T_0. At time T_0+1, each unit $i \in [N]$ receives treatment with probability p_i, and remains in treatment until time T. With probability $1-p_i$, each unit $i \in [N]$ stays in the control group until time T. In other words, for each unit $i \in [N]$*

$$p_{i,t} = 0 \quad \textit{for all} \quad t \le T_0 \quad \textit{and} \quad p_{i,t} = p_i \quad \textit{for all} \quad T_0 < t \le T.$$

Further, for units remaining in control,

$$\eta_{i,t} = 0 \quad \textit{for all} \quad t \le T_0 \quad \textit{and} \quad \eta_{i,t} = -p_i \quad \textit{for all} \quad T_0 < t \le T,$$

and for units receiving treatment,

$$\eta_{i,t} = 0 \quad \textit{for all} \quad t \le T_0 \quad \textit{and} \quad \eta_{i,t} = 1-p_i \quad \textit{for all} \quad T_0 < t \le T.$$

The strong time-series dependence in $\eta_{i,t}$ above implies that Assumption 8 or Assumption 9(a) do not hold, which in turn implies that the guarantees for `Cross-Fitted-SVD`, as in Proposition 4, may not hold. To see this, first note that to ensure Assumption 5, the set of column partitions $\{\mathcal{C}_0, \mathcal{C}_1\}$ must be equal to $\{[T_0], [T] \setminus [T_0]\}$ due to the dependence in the noise W. Now, for Assumption 8 to hold, we need $|\mathcal{C}_k| = \Omega(T)$ for every $k \in \{0,1\}$. However, for Assumption 9(a) to hold, we need $T - T_0$ to be a constant with respect to T as, for any $t \in [T] \setminus [T_0]$ and $i \in [N]$, $\sum_{t' \in [T]} \left|\mathbb{E}[\eta_{i,t}\eta_{i,t'}]\right| = (T-T_0)c_i$ where $c_i \in \{p_i^2, (1-p_i)^2\}$.

Moreover, in Example S1, $t_i = T_0$ for all treated units. This allows the choice of $\{[T_0], [T] \setminus [T_0]\}$ as the set of column partitions $\{\mathcal{C}_0, \mathcal{C}_1\}$ in Assumption 5. More generally, if treatment adoption times $\{t_i\}_{i\in[N]}$ differ across units, then it may not be feasible to obtain a partition of $[T]$ into $\{\mathcal{C}_0, \mathcal{C}_1\}$ such that Assumption 5 holds.

In this section, we propose an alternative approach to the `Cross-Fitted-SVD` algorithm such that Assumption 4 still holds for a suitable staggered adoption model.

Assumption S8 (Staggered adoption and common unit factors)**.** *We consider a panel data setting with staggered adoption where*

1. *all units remain under control till time T_0, i.e., for every unit $i \in [N]$, there exists a time point $t_i \ge T_0$ such that $a_{i,t} = 0$ for $t \le t_i$, and $a_{i,t} = 1$ for $t > t_i$, and*
2. *the unit-dependent latent factors corresponding to P, $\Theta^{(0)}$, and $\Theta^{(1)}$ are the same, i.e., $U = U^{(0)} = U^{(1)} \in \mathbb{R}^{N\times r}$. In other words, for every $i \in [N]$ and $t \in [T]$, $p_{i,t} = g(U_i, V_t)$, $\theta^{(0)}_{i,t} = \langle U_i, V^{(0)}_t\rangle$, and $\theta^{(1)}_{i,t} = \langle U_i, V^{(1)}_t\rangle$ for some known function $g : \mathbb{R}^r \times \mathbb{R}^r \to \mathbb{R}$, with $\langle\cdot,\cdot\rangle$ denoting the inner product.*

For Example S1, the function g corresponds to the inner product, the unit-dependent latent factors are 1-dimensional (i.e., $r = 1$) with $U_i = p_i$ for every $i \in [N]$,

and the time-dependent latent factors for the assignment probability are such that $V_t = 0$ for every $t \in [T_0]$ and $V_t = 1$ for every $t \in [T] \setminus [T_0]$. Consequently, Example S1 is consistent with Assumption S8 if $U_i^{(a)} = p_i$ for every $a \in \{0, 1\}$ and $i \in [N]$. Next, we provide a more flexible version of Example S1 that allows different adoption times for different units.

Example S2 (Different adoption times)**.** *Consider a panel data setting where all units remain in the control group until time T_0. At every time $t \in [T] \setminus [T_0]$, each unit $i \in [N]$ receives treatment with probability p_i, and remains in treatment until time T. Therefore, for $t \in [T] \setminus [T_0]$ and $i \in [N]$, $a_{i,t} = 1$ if the adoption time point $t_i \in \{T_0 + 1, \cdots, t\}$, which occurs with probability $\sum_{t' \in [t-T_0-1]} (1-p_i)^{t'-1} p_i$. In other words, for each unit $i \in [N]$,*

$$p_{i,t} = 0 \quad \textit{for all} \quad t \leq T_0 \quad \textit{and} \quad p_{i,t} = 1 - (1-p_i)^{t-T_0} \quad \textit{for all} \quad T_0 < t \leq T.$$

For Example S2, the unit-dependent latent factors are 1-dimensional (i.e., $r = 1$) with $U_i = p_i$ for every $i \in [N]$, and the time-dependent latent factors for the assignment probability are such that $V_t = 0$ for every $t \in [T_0]$ and $V_t = t - T_0$ for every $t \in [T] \setminus [T_0]$. Further the function g is such that $g(U_i, V_t) = 1 - (1-U_i)^{V_t}$. Consequently, Example S2 is consistent with Assumption S8 if $U_i^{(a)} = p_i$ for every $a \in \{0, 1\}$ and $i \in [N]$.

We now describe `Cross-Fitted-Regression`, an algorithm that generates estimates of $(\Theta^{(0)}, \Theta^{(1)}, P)$ for the staggered adoption model in Assumption S8 such that Assumption 4 holds.

1. The inputs are (i) $A \in \mathbb{R}^{N \times T}$, (ii) $Y^{(a),\text{obs}} \in \{\mathbb{R} \cup \{\,?\}\}^{N \times T}$ for $a \in \{0, 1\}$, (iii) the rank r of the unit-dependent latent factors, (iv) the time period T_0 until which all units remain under control, (v) the time period $t \in [T] \setminus [T_0]$ for which we want to estimate the average treatment effect, and (vi) the function g.

2. Let $Y^{(0),\text{pre}} \in \mathbb{R}^{N \times T_0}$ be the sub-matrix of $Y^{(0),\text{obs}}$ that keeps the first T_0 columns only. Run `SVD` on $Y^{(0),\text{pre}}$, i.e.,

$$\texttt{SVD}(Y^{(0),\text{pre}}) = (\widehat{U} \in \mathbb{R}^{N \times r}, \widehat{\Sigma} \in \mathbb{R}^{r \times r}, \widehat{V} \in \mathbb{R}^{|T_0| \times r}).$$

3. Let $\mathcal{R}^{(0)}$ and $\mathcal{R}^{(1)}$ be the set of units receiving control and treatment at time t, respectively. In other words, for every $a \in \{0, 1\}$, $\mathcal{R}^{(a)} \triangleq \{i \in [N] : a_{i,t} = a\}$. Next, randomly partition $\mathcal{R}^{(a)}$ into two nearly equal parts $\mathcal{R}_0^{(a)}$ and $\mathcal{R}_1^{(a)}$. For every $s \in \{0, 1\}$, define $\mathcal{R}_s = \mathcal{R}_s^{(0)} \cup \mathcal{R}_s^{(1)}$.

4. For every $s \in \{0, 1\}$, regress $\{a_{i,t}\}_{i \in \mathcal{R}_s}$ on $\{\widehat{U}_i\}_{i \in \mathcal{R}_s}$ using g to obtain $\widehat{V}_{1-s}$. For every $s \in \{0, 1\}$ and $i \in \mathcal{R}_s$, return $\widehat{p}_{i,t} = g(\widehat{U}_i, \widehat{V}_s)$.

5. For every $a \in \{0, 1\}$ and $s \in \{0, 1\}$, regress $\{y_{i,t}\}_{i \in \mathcal{R}_s^{(a)}}$ on $\{\widehat{U}_i\}_{i \in \mathcal{R}_s^{(a)}}$ to obtain $\widehat{V}_{1-s}^{(a)}$. For every $a \in \{0, 1\}$, $s \in \{0, 1\}$, and $i \in \mathcal{R}_s$, return $\widehat{\theta}_{i,t}^{(a)} = \widehat{U}_i \widehat{V}_s^{(a)\top}$.

In summary, `Cross-Fitted-Regression` estimates the shared unit-dependent latent factors using the observed outcomes for all units until time period T_0. Then, for every $s \in \{0,1\}$, the time-dependent latent factors $\widehat{V}_s$, $\widehat{V}_s^{(0)}$, and $\widehat{V}_s^{(1)}$ are estimated using the treatment assignments and the observed outcomes for units in $\mathcal{R}_{1-s}$.

To establish guarantees for `Cross-Fitted-Regression`, we adopt the subsequent assumption on the noise variables.

Assumption S9 (Independence across units and with respect to pre-adoption noise)**.**

(a) $\{(\eta_{i,t}, \varepsilon_{i,t}^{(a)}) : i \in [N]\}$ *are mutually independent (across* i*) given* $\{\varepsilon_{i,t}^{(0)}\}_{i\in[N],t\in[T_0]}$ *for every* $t \in [T] \setminus [T_0]$ *and* $a \in \{0,1\}$*.*

(b) $\{\varepsilon_{i,t}^{(0)}\}_{i\in[N],t\in[T_0]} \perp\!\!\!\perp \{\eta_{i,t}, \varepsilon_{i,t}^{(a)}\}_{i\in[N]}$ *for every* $t \in [T] \setminus [T_0]$ *and* $a \in \{0,1\}$*.*

Assumption S9(a) requires the noise $(E^{(a)}, W)$ corresponding to a time period $t \in T \setminus [T_0]$ to be jointly independent across units given the noise $E^{(0)}$ corresponding to time periods $[T_0]$, for every $a \in \{0,1\}$. Assumption S9(b) is satisfied if, for instance, the noise variables follow a moving average model of order $t - T_0 - 1$. The following result, proven in Section S7.1, establishes that the estimates generated by `Cross-Fitted-Regression` satisfy Assumption 4. Deriving error bounds, i.e., $\mathcal{E}\big(\widehat{P}\big)$ and $\mathcal{E}\big(\widehat{\Theta}\big)$, for the estimates generated by `Cross-Fitted-Regression` for the staggered adoption model is an interesting direction for future research.

Proposition S1 (Guarantees for `Cross-Fitted-Regression`)**.** *Consider the staggered adoption model in Assumption S8 and suppose Assumption S9 holds. Fix any* $t \in [T] \setminus [T_0]$*, and* $\{\widehat{\theta}_{i,t}^{(0)}, \widehat{\theta}_{i,t}^{(1)}, \widehat{p}_{i,t}\}_{i\in[N]}$ *be the estimates returned by* `Cross-Fitted-Regression`*. Then, Assumption 4 holds.*

S7.1. Proof of Proposition S1: Guarantees for `Cross-Fitted-Regression`

Fix any $s \in \{0,1\}$. Then, Assumption S9(a) and Assumption S9(b) imply that

$$\{\varepsilon_{i,t}^{(0)}\}_{i\in[N],t\in[T_0]} \cup \{\eta_{i,t}, \varepsilon_{i,t}^{(a)}\}_{i\in\overline{\mathcal{R}}_{1-s}} \perp\!\!\!\perp \{\eta_{i,t}, \varepsilon_{i,t}^{(a)}\}_{i\in\overline{\mathcal{R}}_s}, \tag{S.38}$$

for every partition $(\overline{\mathcal{R}}_0, \overline{\mathcal{R}}_1)$ of the units $[N]$.

`Cross-Fitted-Regression` estimates $\{\widehat{p}_{i,t}\}_{i\in\mathcal{R}_s}$ using $\{\widehat{U}_i\}_{i\in\mathcal{R}_s}$ and $\widehat{V}_s$, where $\widehat{V}_s$ is estimated using $\{\widehat{U}_i\}_{i\in\mathcal{R}_{1-s}}$ and $\{a_{i,t}\}_{i\in\mathcal{R}_{1-s}}$. Therefore, the randomness in $\{\widehat{p}_{i,t}\}_{i\in\mathcal{R}_s}$ stems from the randomness in $Y^{(0),\text{pre}}$ and $\{a_{i,t}\}_{i\in\mathcal{R}_{1-s}}$ which in turn stems from the randomness in $\{\varepsilon_{i,t}^{(0)}\}_{i\in[N],t\in[T_0]}$ and $\{\eta_{i,t}\}_{i\in\mathcal{R}_{1-s}}$. Then, Eq. (15) follows from Eq. (S.38).

Next, fix any $a \in \{0,1\}$. Then, `Cross-Fitted-Regression` estimates $\{\widehat{\theta}^{(a)}\}_{i\in\mathcal{R}_s}$ using $\{\widehat{U}_i\}_{i\in\mathcal{R}_s}$ and $\widehat{V}_s^{(a)}$, where $\widehat{V}_s^{(a)}$ is estimated using $\{\widehat{U}_i\}_{i\in\mathcal{R}_{1-s}^{(a)}}$ and $\{y_{i,t}\}_{i\in\mathcal{R}_{1-s}^{(a)}}$. Therefore, the randomness in $\{\widehat{\theta}^{(a)}\}_{i\in\mathcal{R}_s}$ stems from the randomness in $Y^{(0),\text{pre}}$ and $\{y_{i,t}\}_{i\in\mathcal{R}_{1-s}^{(a)}}$ which in turn stems from the randomness in $\{\varepsilon_{i,t}^{(0)}\}_{i\in[N],t\in[T_0]}$ and $\{\varepsilon_{i,t}^{(a)}\}_{i\in\mathcal{R}_{1-s}^{(a)}}$. Then, Eq. (14) follows from Eq. (S.38).