Doubly Robust Inference in Causal Latent Factor Models

Alberto Abadie
MIT

Anish Agarwal
Columbia

Raaz Dwivedi
Cornell Tech

Abhin Shah
MIT

April 15, 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 practical 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. Modern data-rich environments are characterized by repeated measurements of outcomes, such as clinical metrics or purchase history, across a substantial number of units—be it patients in medical contexts or customers in online retail. As an example, consider 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. Given an observational dataset capturing such interactions, our objective is to infer the causal effect of offering the discount on consumer purchase behavior. More specifically, we aim to infer two kinds of treatment effects: (a) tailored to product categories, the average impact of the discount on a product across consumers, and (b) tailored to consumers, the average impact of the discount on a consumer across product categories. This task is challenging due to unobserved confounding that may cause spurious associations between discount allocation and product purchase.

Abadie: abadie@mit.edu; Agarwal: aa5194@columbia.edu; Dwivedi: dwivedi@cornell.edu; Shah: abhin@mit.edu

There are two widely used approaches for treatment effect estimation: outcome-based methods and assignment-based methods. 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). Assignment-based methods predict the probability with which a consumer is offered the discount on a product category, and inversely weight the observed outcomes by these estimated probabilities.

A substantial and influential 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, and the existing literature devoted to this problem is relatively small. In this context, a commonly adopted framework is the latent factor framework (Bai and Ng, 2002; Bai, 2009), wherein each element of the large-dimensional outcome vector is influenced by the same low-dimensional vector of unobserved confounders. A closely related approach is the technique of matrix completion (see, e.g., Chatterjee, 2015; Athey et al., 2021; Bai and Ng, 2021; Agarwal et al., 2023a; Dwivedi et al., 2022a) which has found widespread applications in recommendation systems and panel data models.

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.

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 $n \geq n_0$. For a sequence of random variables x_n and a sequence of positive constants b_n, $x_n = O_p(b_n)$ means that the sequence $|x_n/b_n|$ is stochastically bounded, i.e., for every $\epsilon > 0$, there exists a finite $\delta > 0$ and a finite $n_0 > 0$ such that $\mathbb{P}\big(|x_n/b_n| > \delta\big) < \epsilon$ for all $n \geq n_0$. Similarly, $x_n = o_p(b_n)$ means that the sequence $|x_n/b_n|$ converges to zero in probability, i.e., for every $\epsilon > 0$ and $\delta > 0$, there exists a finite $n_0 > 0$ such that $\mathbb{P}\big(|x_n/b_n| > \delta\big) < \epsilon$ for all $n \geq n_0$.

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 $\mathcal{I} \triangleq \mathcal{R} \times \mathcal{C}$. 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 $L_{1,2}$ 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 $L_{2,\infty}$ 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 Khatri-Rao product (or column-wise 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 $y_{i,j} \in \mathbb{R}$ under the treatment assignment. For the ease of exposition, we focus on binary treatments. However, our framework can be easily generalized to multi-ary 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 $y_{i,j}^{(a)} \in \mathbb{R}$. Here, it is implicitly assumed that the potential outcome for any unit i and measurement j does not depend on the treatment assignment for any other unit-measurement pair, i.e., there are no spillover effects across units or measurements. 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. The observed outcomes depend on the potential outcomes and the 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 $j \in [M]$.

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 all 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 identification restrictions made 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

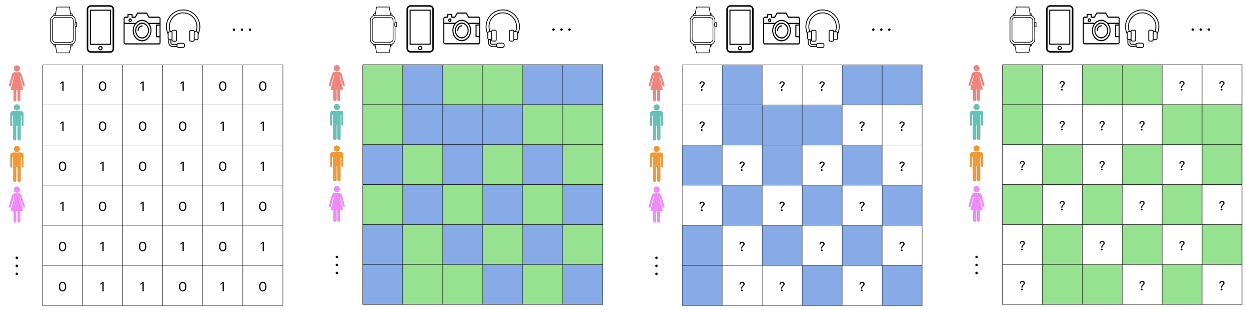
$$\mu_{\cdot,j}^{(a)} \triangleq \frac{1}{N} \sum_{i\in[N]} \theta_{i,j}^{(a)}.$$

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 an estimator that uses the treatment assignment matrix A and the observed outcomes matrix Y to estimate the target causal estimand $\{\text{ATE}_{\cdot,j}\}_{j\in[M]}$, where

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



(*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.

Our estimator leverages matrix completion as a key subroutine. We start with a brief overview of matrix completion below.

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 $S \in \{\mathbb{R}, ?\}^{N\times M}$ where ? denotes a missing value. The relationship between S and T is given by

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

where $H \in \mathbb{R}^{N\times M}$ represents a matrix of noise, $F \in \{0,1\}^{N\times M}$ is a masking matrix, and the operator $\otimes$ is as defined in Section 1. A matrix completion algorithm, denoted by $\texttt{MC}$, takes the matrix S as its input, and returns an estimate for the matrix 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., 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 $Y^{(0),\text{obs}} = Y \otimes (\mathbf{1} - A) \in \{\mathbb{R}, ?\}^{N \times M}$ 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 $Y^{(1),\text{obs}} = Y \otimes A \in \{\mathbb{R}, ?\}^{N \times M}$ 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 $\{y_{i,j}^{(0)}\}_{i\in[N],j\in[M]}$ and $\{y_{i,j}^{(1)}\}_{i\in[N],j\in[M]}$, 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

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

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

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

where

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

That is, the OI estimator is obtained by taking the difference of the average value of the j-th column of the estimates $\widehat{\Theta}^{(0)}$ and $\widehat{\Theta}^{(1)}$. The quality of 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:

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

where

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

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

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

3.3. Doubly robust (DR) estimator

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

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

where

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

and

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

In Section 4, we prove that $\widehat{\text{ATE}}_{\cdot,j}^{\text{DR}}$ 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 demonstrates a mean-zero Gaussian distribution. We provide a detailed description of the simulation setup in Section 6.

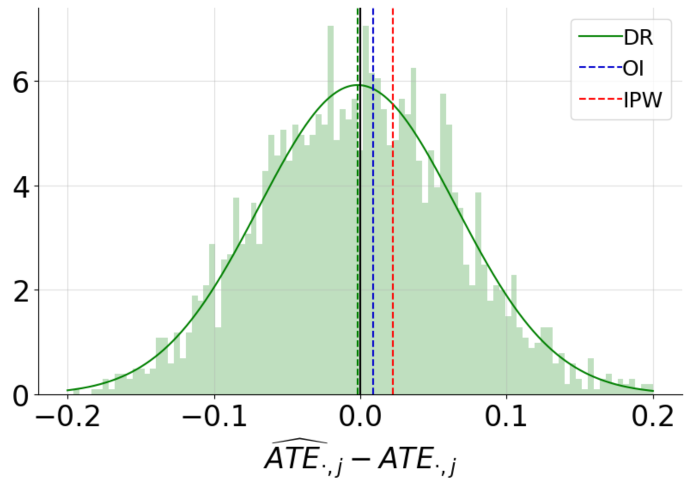


Figure 2: Empirical illustration of the convergence of the error of the doubly robust (DR) estimator to a mean-zero Gaussian distribution. The histogram represents $\widehat{\mathrm{ATE}}_{\cdot,j}^{\mathrm{DR}} - \mathrm{ATE}_{\cdot,j}$ and the curve represents the (best) fitted Gaussian distribution. 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. We provide details of the simulations, including the data-generating process, 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 $\widehat{\mathrm{ATE}}_{\cdot,j}^{\mathrm{DR}}$.

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). *The unknown assignment probability matrix P is such that*

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

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

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 that assume observed confounding. For simplicity of exposition and to avoid notational clutter, Assumption 1 requires Eq. (13) for all outcomes, $j \in [M]$. However, it is only necessary that Eq. (13) holds for the outcomes of interest, j, for which $\mathrm{ATE}_{\cdot,j}$ is estimated. Our framework leverages the availability of a large number of outcomes to control for the confounding effect of latent variables. In practical applications, however, $\mathrm{ATE}_{\cdot,j}$ 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 $\mathrm{ATE}_{\cdot,j}$ 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) $\{(\varepsilon_{i,j}^{(0)}, \varepsilon_{i,j}^{(1)}, \eta_{i,j}) : i \in [N]\}$ are mean zero and independent (across i),

(b) $(\varepsilon_{i,j}^{(0)}, \varepsilon_{i,j}^{(1)}) \perp\!\!\!\perp \eta_{i,j}$ for every $i \in [N]$, 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 the means of the potential outcomes and treatment assignment 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, for every unit and every measurement. This assumption is crucial for identification. 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 $\varepsilon_{i,j}^{(a)}$. Note that Assumption 2 does not restrict the dependence between $\varepsilon_{i,j}^{(0)}$ and $\varepsilon_{i,j}^{(1)}$ for any unit $i \in [N]$.

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

Assumption 3. *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 is achieved by truncating entries of $\widehat{P}$ to the range $[\bar{\lambda}, 1 - \bar{\lambda}]$. 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. *Fix any $j \in [M]$. There exists a non-empty partition $(\mathcal{R}_0, \mathcal{R}_1)$ of the units $[N]$ such that*

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

and

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

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. Analogous conditions appear in the literature on doubly robust estimation under observed confounding (e.g., Definition 3.1 in Chernozhukov et al., 2018). Specifically, in that context, Chernozhukov et al. (2018) split the available data into K-folds ($K \geq 2$), and require estimates of propensities and outcomes in each fold to be independent of the noise in that fold. Section 5 provides a way to ensure Assumption 4 holds for any `MC` algorithm using a cross-fitting procedure, with additional assumptions on the noise variables.

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

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

where

$$\mathcal{E}\big(\widehat{\Theta}^{(a)}\big) = \frac{\|\widehat{\Theta}^{(a)} - \Theta^{(a)}\|_{1,2}}{\sqrt{N}}.$$

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 $0 < \alpha, \beta \leq 1/2$. Throughout, our notation primarily tracks dependence on N. We say that these normalized errors achieve the parametric rate when they have the same rate as $O_p(N^{-1/2})$. Section 5 explicitly characterizes α and β under low-rank assumptions on $(\Theta^{(0)}, \Theta^{(1)})$ and P for a particular matrix completion algorithm.

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*

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

where

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

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

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

Doubly robust behavior of $\widehat{\mathrm{ATE}}_{\cdot,j}^{\mathrm{DR}}$. The error rate of $\mathcal{E}(\widehat{P})\mathcal{E}(\widehat{\Theta}) + N^{-0.5}$ 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 $\mathrm{Err}_N^{\mathrm{DR}}$ decays at a parametric rate of $O_p(N^{-0.5})$ as long as the product of error rates, $\mathcal{E}(\widehat{P})\mathcal{E}(\widehat{\Theta})$, decays as $O_p(N^{-0.5})$. As a result, $\widehat{\mathrm{ATE}}_{\cdot,j}^{\mathrm{DR}}$ can exhibit a parametric error rate even when neither the mean potential outcomes nor the assignment probabilities are estimated at a parametric rate. Second, $\mathrm{Err}_N^{\mathrm{DR}}$ decays to zero as long as either of $\mathcal{E}(\widehat{P})$ or $\mathcal{E}(\widehat{\Theta})$ decays to 0.

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

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

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

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

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

where

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

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

The proofs of Eq. (19) and Eq. (20) are given in Appendices E and F, respectively. 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 $\mathcal{E}(\widehat{\Theta})$ and $\mathcal{E}(\widehat{P})$ are $O_p(N^{-0.5})$, respectively. The next corollary, proven in Appendix C, 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 $\mathcal{E}(\widehat{P}) = O_p(N^{-\alpha})$ and $\mathcal{E}(\widehat{\Theta}) = O_p(N^{-\beta})$ for $0 \leq \alpha, \beta \leq 0.5$, then*

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

and

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

Corollary 1 demonstrates 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 $\mathrm{Err}_N^{\mathrm{DR}}$ scales strictly faster than both $\mathrm{Err}_N^{\mathrm{OI}}$ and $\mathrm{Err}_N^{\mathrm{IPW}}$ 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 Appendix C 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 $\widehat{\mathrm{ATE}}_{\cdot,j}^{\mathrm{DR}}$.

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*

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

for all $j \in [M]$.

Corollary 2 states that $\widehat{\mathrm{ATE}}_{\cdot,j}^{\mathrm{DR}}$ 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 D, establishes a Gaussian approximation for $\widehat{\mathrm{ATE}}_{\cdot,j}^{\mathrm{DR}}$ 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 $\sigma_{i,j}^{(0)}$ and $\sigma_{i,j}^{(1)}$ be the standard deviations of $\varepsilon_{i,j}^{(0)}$ and $\varepsilon_{i,j}^{(1)}$, respectively. The sequence

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

is bounded away from zero as N increases.

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

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

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 $\widehat{\mathrm{ATE}}_{\cdot,j}^{\mathrm{DR}}$ 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 Theorem 2 applies to any matrix completion algorithm `MC` as long as conditions (C1) and (C2) are satisfied This property arises because the bias is dominated by the product of $\mathcal{E}\big(\widehat{P}\big)$ and $\mathcal{E}\big(\widehat{\Theta}\big)$, which can be shown to be $o_p\big(N^{-1/2}\big)$ for a broad class of `MC` algorithms under mild assumptions on $(P, \Theta^{(0)}, \Theta^{(1)})$. On the other hand, achieving such black-box asymptotic normality results for OI or IPW estimates is challenging, as their bias scales with individual error rates $\mathcal{E}\big(\widehat{\Theta}\big)$ and $\mathcal{E}\big(\widehat{P}\big)$, respectively, which are typically lower bounded at the parametric rate of $N^{-0.5}$. Our simulations in Section 6 corroborate these theoretical findings.

The next result, proven in Appendix D.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*

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

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

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

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. 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 $Y^{(0),\mathrm{obs}}$ and $Y^{(1),\mathrm{obs}}$, 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 approach that modifies any `MC` algorithm to produce $(\widehat{\Theta}^{(0)}, \widehat{\Theta}^{(1)}, \widehat{P})$ that satisfy Assumption 4 under the following additional assumption on the noise variables.

Assumption 5. *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 $\mathcal{I}$.

Recall the setup from Section 3.1: Given an observation matrix $S \in \{\mathbb{R}, ?\}^{N\times M}$, a matrix completion algorithm `MC` produces an estimate $\widehat{T} = \mathtt{MC}(S) \in \mathbb{R}^{N\times M}$ 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 $S \in \{\mathbb{R}, ?\}^{N\times M}$, 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 Appendix G.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*

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

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

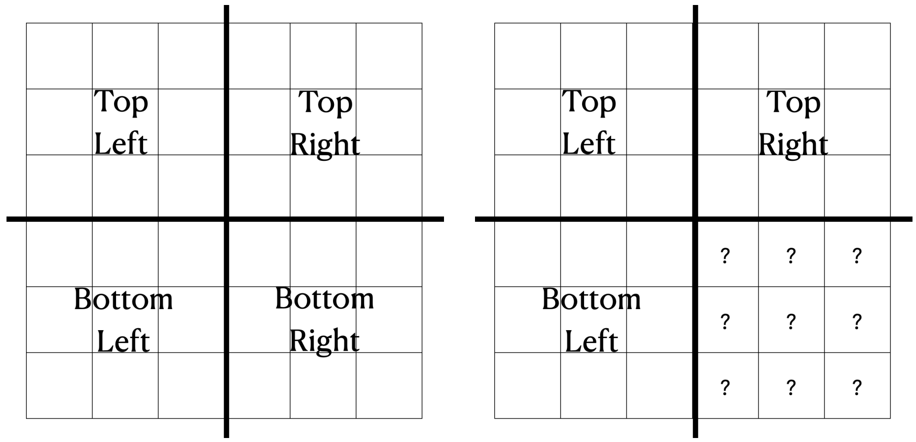
$$\widehat{P} = \texttt{Cross-Fitted-MC}(\texttt{MC}, A, \mathcal{P}), \tag{32}$$

where $Y^{(0),\mathrm{obs}}$ *and* $Y^{(1),\mathrm{obs}}$ *are defined in Eq.* (7)*. Then, Assumption 4 holds. Further, suppose*

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

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

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



(a) S $\qquad$ (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 ?.

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 any `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 those matrices where there exists a set of rows and a set of columns without missing entries. More concretely, for any matrix $S \in \{\mathbb{R}, ?\}^{N \times M}$, let $\mathcal{R}_{\text{obs}} \subseteq [N]$ and $\mathcal{C}_{\text{obs}} \subseteq [M]$ denote the set of rows and columns, respectively, with all entries observed. Then, 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 $\mathcal{C}_{\text{miss}} \triangleq [M] \setminus \mathcal{C}_{\text{obs}}$, is such that all the missing entries in S are a subset of it.

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.,

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

and

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

where $\bar{r}_N \triangleq \min\{N, |\mathcal{C}_{\text{obs}}|\}$ and $\bar{r}_M \triangleq \min\{|\mathcal{R}_{\text{obs}}|, M\}$. 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:

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

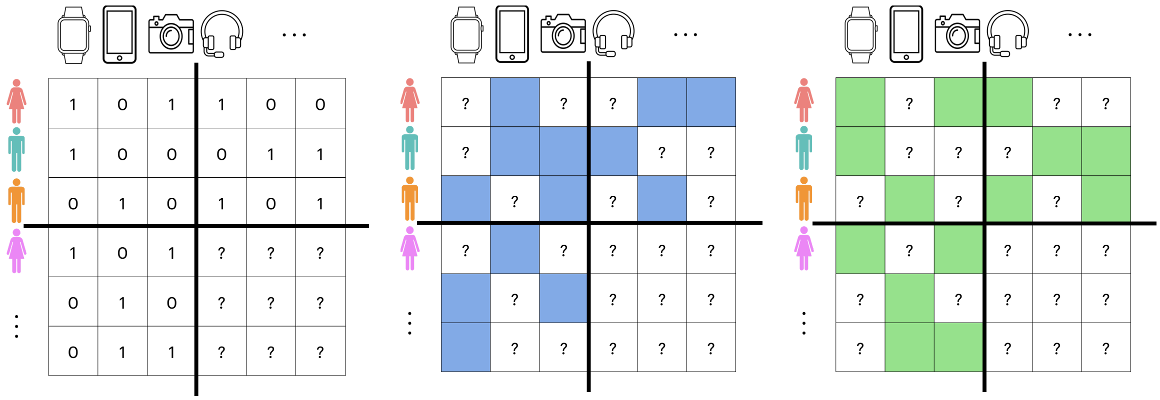
That is, R is obtained by regressing $\widetilde{V}^{(\text{tall})}$ on $\widetilde{V}^{(\text{wide})}$. In essence, R aligns the right singular vectors of $S^{(\text{tall})}$ and $S^{(\text{wide})}$ using the entries that are common between these two matrices, i.e., the entries corresponding to indices $\mathcal{R}_{\text{obs}} \times \mathcal{C}_{\text{obs}}$. 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 $\overline{\Sigma}^{(\text{tall})} \in \mathbb{R}^{\bar{r}_N\times r}$ 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) $Y^{(a),\text{obs}} \in \{\mathbb{R}, ?\}^{N\times M}$ for $a \in \{0,1\}$, (iii) a block partition $\mathcal{P}$ of the set $[N]\times[M]$ into four blocks as in Assumption 5, and (iv) hyper-parameters r_1, r_2, r_3, and $\bar{\lambda}$ such that $r_1, r_2, r_3 \in [\min\{N, M\}]$ and $0 < \bar{\lambda} \leq 1/2$.

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



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

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.

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, $\mathtt{MC}(Y^{(0),\text{full}})$ and $\mathtt{MC}(Y^{(1),\text{full}})$ provide estimates of $\Theta^{(0)} \odot (\mathbf{1} - P)$ and $\Theta^{(1)} \odot P$, respectively—recall the discussion in Section 3.1. To estimate $\Theta^{(0)}$ and $\Theta^{(1)}$, we divide the entries of $\mathtt{MC}(Y^{(0),\text{full}})$ and $\mathtt{MC}(Y^{(1),\text{full}})$ by the entries of $(\mathbf{1} - \widehat{P})$ and $\widehat{P}$, respectively. Adjustments of this type for heterogeneous missingness probabilities have been previously explored in Ma and Chen (2019); Bhattacharya and Chatterjee (2022).

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) $U_{i,\cdot} \in \mathbb{R}^{r_p}$, (ii) $U^{(0)}_{i,\cdot} \in \mathbb{R}^{r_{\theta_0}}$, and (iii) $U^{(1)}_{i,\cdot} \in \mathbb{R}^{r_{\theta_1}}$. Similarly, every measurement $j \in [M]$ is associated with three factors: (i) $V_{i,\cdot} \in \mathbb{R}^{r_p}$, (ii) $V^{(0)}_{i,\cdot} \in \mathbb{R}^{r_{\theta_0}}$, and (iii) $V^{(1)}_{i,\cdot} \in \mathbb{R}^{r_{\theta_1}}$. Such 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 row-wise Khatri-Rao product (see Section 1). We provide details of the derivation of these factors in Appendix G.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 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 $P \triangleq \{\mathcal{R}_s \times \mathcal{C}_k : s, k \in \{0,1\}\}$ from Assumption 5. For every $s \in \{0,1\}$, let $U_{(s)} \in \mathbb{R}^{|R_i| \times r_p}$, $\overline{U}^{(0)}_{(s)} \in \mathbb{R}^{|R_i| \times r_{\theta_0}(r_p+1)}$, and $\overline{U}^{(1)}_{(s)} \in \mathbb{R}^{|R_i| \times r_{\theta_1} r_p}$ 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 $V_{(k)} \in \mathbb{R}^{|C_j| \times r_p}$, $\overline{V}^{(0)}_{(k)} \in \mathbb{R}^{|C_j| \times r_{\theta_0}(r_p+1)}$, and $\overline{V}^{(1)}_{(k)} \in \mathbb{R}^{|C_j| \times r_{\theta_1} r_p}$ 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 $i, j \in \{0,1\}$, the matrices defined below are positive definite:*

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

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 across measurements and independence 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) $\{(E^{(a)}_{i,\cdot}, W_{i,\cdot}) : i \in [N]\}$ *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. 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 Appendix G.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 $N, M \to \infty$,*

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

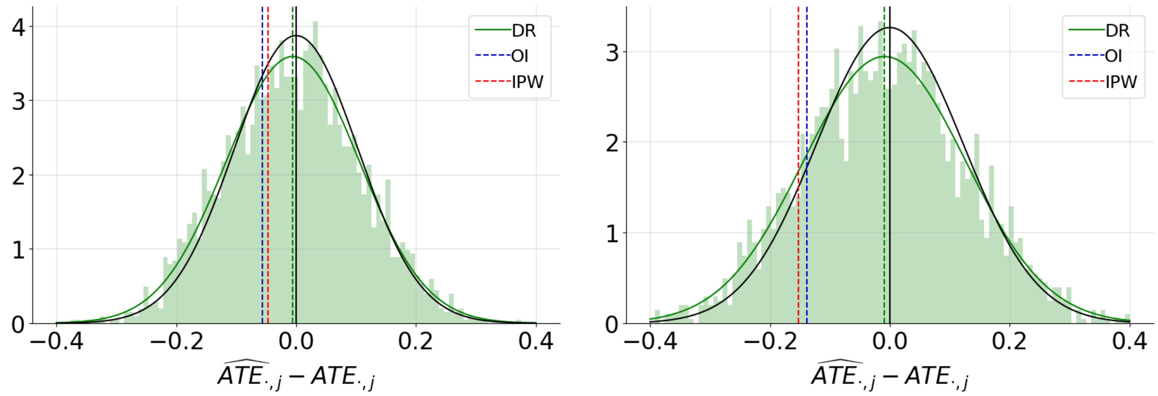
Proposition 4 implies that the conditions (C1) and (C2) in Theorem 2 hold whenever $N^{1/2}/M = o(1)$. Then, the DR estimator from Eq. (11) constructed using the estimates $\widehat{\Theta}^{(0)}$, $\widehat{\Theta}^{(1)}$, and $\widehat{P}$ returned by `Cross-Fitted-SVD` exhibits an asymptotic Gaussian 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 $N/M = o(1)$. In summary, Propositions 3 and 4 ensure that the estimates adhere to Assumption 4 and satisfy conditions (C1) and (C2) in Theorem 2.

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. For convenience, we let $N = M$.

Data Generating Process (DGP). We now briefly describe the DGP for our simulations; details can be found in Appendix H. 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. 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 per Eq. (4). Then, $y_{i,j}^{(a)}$, $a_{i,j}$, and $y_{i,j}$ are determined from Eqs. (1) to (3), respectively. The simulation generates P, $\Theta^{(0)}$, and $\Theta^{(1)}$ once. Then, given the fixed values of P, $\Theta^{(0)}$, and $\Theta^{(1)}$, the simulation generates Q realizations of (Y, A)—that is, only the noise matrices $E^{(0)}, E^{(1)}, W$ are resampled for each of the Q realizations. For each of these Q instances of the simulation, $\widehat{P}$, $\widehat{\Theta}^{(0)}$, and $\widehat{\Theta}^{(1)}$ are obtained by applying the `Cross-Fitted-SVD` algorithm to the corresponding A and Y with the choice of hyper-parameters as in Proposition 4 and $\bar{\lambda} = \lambda = 0.05$. For each of the instances of the simulation we compute $\mathrm{ATE}_{\cdot,j}$ from Eq. (5), and $\widehat{\mathrm{ATE}}_{\cdot,j}^{\mathrm{OI}}$, $\widehat{\mathrm{ATE}}_{\cdot,j}^{\mathrm{IPW}}$ and $\widehat{\mathrm{ATE}}_{\cdot,j}^{\mathrm{DR}}$ from Eqs. (9) to (11). We set $Q = 2500$.

Results. Figure 5 reports simulation results for $N = 1000$, with $r_p = 3$, $r_\theta = 5$ in Panel (a), and $r_p = 5$, $r_\theta = 3$ in Panel (b). Figure 2 in Section 3 reports simulation results for $r_p = r_\theta = 3$. In each case, the figure shows a histogram of the distribution of $\widehat{\mathrm{ATE}}_{\cdot,j}^{\mathrm{DR}} - \mathrm{ATE}_{\cdot,j}$ 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



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

Figure 5: Empirical illustration of the asymptotic performance of DR as in Theorem 2. The histogram corresponds to the errors of 1000 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.

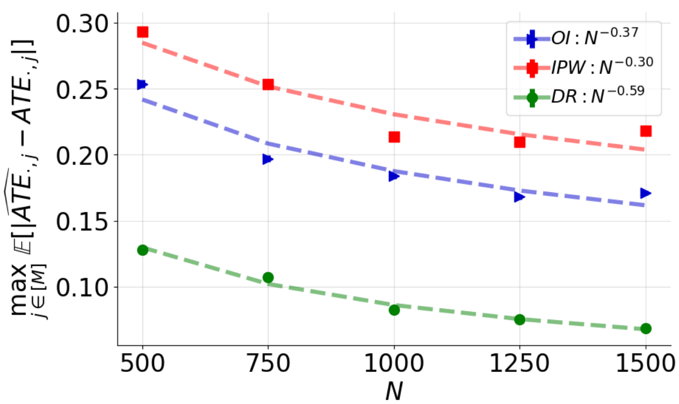


Figure 6: Comparison of OI, IPW, and DR in terms of finite sample performance as in Proposition 1. The estimates $\widehat{\mathrm{ATE}}^{\mathrm{OI}}_{\cdot,j}$, $\widehat{\mathrm{ATE}}^{\mathrm{IPW}}_{\cdot,j}$, and $\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,j}$ are obtained by taking an average over 2500 independent instances.

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

To further illustrate the different bias performance of the three estimators, Figure 6 reports the maximum over $j \in [M]$ of their respective mean absolute error estimates. For each j, the estimate of the mean absolute error of OI, IPW, and DR is the average of $|\widehat{\mathrm{ATE}}^{\mathrm{OI}}_{\cdot,j} - \mathrm{ATE}_{\cdot,j}|$, $|\widehat{\mathrm{ATE}}^{\mathrm{IPW}}_{\cdot,j} - \mathrm{ATE}_{\cdot,j}|$ and $|\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,j} - \mathrm{ATE}_{\cdot,j}|$ across the Q simulation instances, respectively. We set $r_p = r_\theta = 3$ and vary $N \in \{500, 750, 1000, 1250, 1500\}$. To make the scaling clear, we use least squares to produce the best $N^{-\rho}$ fit to the maximum bias as N varies. We state the empirical decay rates in the legend, e.g., for DR, we report an empirical rate of $N^{-0.62}$. The DR estimator consistently outperforms the OI and IPW estimators.

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 matrix completion to estimate the treatment assignment probabilities and mean potential outcomes. 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 mean absolute error for the 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.

Appendices

A. 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.

Lemma A.1 (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 $t \geq 0$,*

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

The following corollary expresses the bound in Lemma A.1 in a convenient form.

Corollary A.1 (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 $1-\delta$,*

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

Proof. The proof follows from Lemma A.1 by choosing $\delta \triangleq 2\exp(-ct^2/\sigma^2\|b\|_2^2)$. □

Lemma A.2 (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 $t \geq 0$,*

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

The following corollary expresses the bound in Lemma A.2 in a convenient form.

Corollary A.2 (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 $1-\delta$,*

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

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

Proof. Choosing $t = t_0\sigma\|b\|_2$ in Lemma A.2, 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 A.3 (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 1 (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. We note that subGaussian and subExponential random variables are subWeibull random variable with indices 2 and 1, respectively.

Lemma A.4 (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*

$$\sigma = \Pi_{i\in[d]}\sigma_i \quad \text{and} \quad \rho = 1/\sum_{i\in[d]} 1/\rho_i.$$

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

Lemma A.5. *Let X_n and $\overline{X}_n$ be sequences of random variables. Let $\delta_n = o(1)$ be a deterministic sequence such that $0 \leq \delta_n \leq 1$. 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 $n > \overline{n}$. Fix any $\epsilon > 0$. Because δ_n converges to zero, there exists a finite n_0 such that $\delta_n < \epsilon/2$, for all $n > n_0$. Because X_n is converges to zero in probability, there exist finite n_1, such that $\mathbb{P}(|X_n| > \delta) < \epsilon/2$ for all $n > n_1$. 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 $n > \overline{n} = \max\{n_0, n_1\}$. Therefore, $\overline{X}_n = o_p(1)$.

□

Lemma A.6. *Let X_n and $\overline{X}_n$ be sequences of random variables. Suppose $\mathbb{E}\big[|X_n|\big|\overline{X}_n\big]$ exists and is non-negative. Then, $X_n = O_p\big(\mathbb{E}\big[|X_n|\big|\overline{X}_n\big]\big)$.*

Proof. Consider any $\epsilon > 0$. Then, we have

$$\begin{aligned}\mathbb{P}\bigg(\frac{|X_n|}{\mathbb{E}\big[|X_n|\big|\overline{X}_n\big]} \geq \frac{1}{\epsilon}\bigg) &\overset{(a)}{=} \mathbb{E}\bigg[\mathbb{P}\bigg(\frac{|X_n|}{\mathbb{E}\big[|X_n|\big|\overline{X}_n\big]} \geq \frac{1}{\epsilon}\Big|\overline{X}_n\bigg)\bigg] \\ &\overset{(b)}{\leq} \mathbb{E}\bigg[\mathbb{E}\Big[\frac{|X_n|}{\mathbb{E}\big[|X_n|\big|\overline{X}_n\big]}\Big|\overline{X}_n\Big]\epsilon\bigg] = \epsilon.\end{aligned}$$

where (a) follows from the law of total expectation and (b) follows from Markov's inequality. Therefore, for every $\epsilon > 0$, we have $\delta = 1/\epsilon$ such that $\mathbb{P}\big(|X_n|/\mathbb{E}\big[|X_n|\big|\overline{X}_n\big] > \delta\big) < \epsilon$ for all n. □

Lemma A.7. *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 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 $\overline{\delta}$ and $\overline{n}$, such that

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

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

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

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

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

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

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

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

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

$$
\begin{aligned}
\mathbb{T}^{(0,\mathrm{DR})}_{i,j} &= -\widehat{\theta}^{(0,\mathrm{DR})}_{i,j} + \theta^{(0)}_{i,j}\\
&= -\frac{(\widehat{\theta}^{(0)}_{i,j} - \theta^{(0)}_{i,j})(1-\widehat{p}_{i,j} - (1-p_{i,j}))}{1-\widehat{p}_{i,j}} + \frac{(\widehat{\theta}^{(0)}_{i,j} - \theta^{(0)}_{i,j})(-\eta_{i,j})}{1-\widehat{p}_{i,j}} - \frac{\varepsilon^{(0)}_{i,j}(1-p_{i,j})}{1-\widehat{p}_{i,j}} - \frac{\varepsilon^{(0)}_{i,j}(-\eta_{i,j})}{1-\widehat{p}_{i,j}}\\
&= \frac{(\widehat{\theta}^{(0)}_{i,j} - \theta^{(0)}_{i,j})(\widehat{p}_{i,j} - p_{i,j})}{1-\widehat{p}_{i,j}} - \frac{(\widehat{\theta}^{(0)}_{i,j} - \theta^{(0)}_{i,j})\eta_{i,j}}{1-\widehat{p}_{i,j}} - \frac{\varepsilon^{(0)}_{i,j}(1-p_{i,j})}{1-\widehat{p}_{i,j}} + \frac{\varepsilon^{(0)}_{i,j}\eta_{i,j}}{1-\widehat{p}_{i,j}}.
\end{aligned} \tag{A.3}
$$

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

$$
\frac{1}{N}\Big|\sum_{i\in[N]}\mathbb{T}^{(\mathrm{a,DR})}_{i,j}\Big| \leq \frac{2}{\bar{\lambda}}\mathcal{E}\big(\widehat{\Theta}^{(a)}\big)\mathcal{E}\big(\widehat{P}\big) + \frac{2\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{\ell_1 N}}\mathcal{E}\big(\widehat{\Theta}^{(a)}\big) + \frac{2\overline{\sigma}\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{N}} + \frac{2\overline{\sigma}m(c\ell_\delta)}{\bar{\lambda}\sqrt{\ell_1 N}}, \tag{A.4}
$$

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.4) with a union bound, we obtain that

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

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

Proof of bound (A.4)**.** 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.4) 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 $|\sum_{i\in[N]} \mathbb{T}_{i,j}^{(1,\mathrm{DR})}| \leq |\sum_{i\in\mathcal{R}_0} \mathbb{T}_{i,j}^{(1,\mathrm{DR})}| + |\sum_{i\in\mathcal{R}_1} \mathbb{T}_{i,j}^{(1,\mathrm{DR})}|$. Fix any $s \in \{0, 1\}$. Then, Eq. (A.2) and triangle inequality imply

$$\begin{aligned}\Big|\sum_{i\in\mathcal{R}_s} \mathbb{T}_{i,j}^{(1,\mathrm{DR})}\Big| \leq \Big|\sum_{i\in\mathcal{R}_s} \frac{\big(\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)}\big)\big(\widehat{p}_{i,j} - p_{i,j}\big)}{\widehat{p}_{i,j}}\Big| + \Big|\sum_{i\in\mathcal{R}_s} \frac{\big(\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)}\big)\eta_{i,j}}{\widehat{p}_{i,j}}\Big| \\ + \Big|\sum_{i\in\mathcal{R}_s} \frac{\varepsilon_{i,j}^{(1)} p_{i,j}}{\widehat{p}_{i,j}}\Big| + \Big|\sum_{i\in\mathcal{R}_s} \frac{\varepsilon_{i,j}^{(1)} \eta_{i,j}}{\widehat{p}_{i,j}}\Big|. \qquad \text{(A.5)}\end{aligned}$$

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

$$\begin{aligned}\Big|\sum_{i\in\mathcal{R}_s} \frac{\big(\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)}\big)\big(\widehat{p}_{i,j} - p_{i,j}\big)}{\widehat{p}_{i,j}}\Big| &\leq \sqrt{\sum_{i\in\mathcal{R}_s} \Big(\frac{\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)}}{\widehat{p}_{i,j}}\Big)^2 \sum_{i\in\mathcal{R}_s} \big(\widehat{p}_{i,j} - p_{i,j}\big)^2} \\ &\leq \big\|\big(\widehat{\Theta}_{\cdot,j}^{(1)} - \Theta_{\cdot,j}^{(1)}\big) \oslash \widehat{P}_{\cdot,j}\big\|_2 \big\|\widehat{P}_{\cdot,j} - P_{\cdot,j}\big\|_2. \qquad \text{(A.6)}\end{aligned}$$

To bound the second term in Eq. (A.5), 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). Moreover, Assumption 4 (i.e., Eq. (14)) provides that $\{(\widehat{p}_{i,j}, \widehat{\theta}_{i,j}^{(1)})\}_{i\in\mathcal{R}_s} \perp\!\!\!\perp \{\eta_{i,j}\}_{i\in\mathcal{R}_s}$. Hence, applying the subGaussian concentration (Corollary A.1) for $\{\eta_{i,j}\}_{i\in\mathcal{R}_s}$ yields that

$$\Big|\sum_{i\in\mathcal{R}_s} \frac{\big(\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)}\big)\eta_{i,j}}{\widehat{p}_{i,j}}\Big| \leq \frac{\sqrt{c\ell_\delta}}{\sqrt{\ell_1}} \sqrt{\sum_{i\in\mathcal{R}_s} \Big(\frac{\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)}}{\widehat{p}_{i,j}}\Big)^2} \leq \frac{\sqrt{c\ell_\delta}}{\sqrt{\ell_1}} \big\|\big(\widehat{\Theta}_{\cdot,j}^{(1)} - \Theta_{\cdot,j}^{(1)}\big) \oslash \widehat{P}_{\cdot,j}\big\|_2, \qquad \text{(A.7)}$$

with probability at least $1 - \delta$.

To bound the third term in Eq. (A.5), note that $\varepsilon_{i,j}^{(1)}$ is subGaussian$(\overline{\sigma})$, zero-mean, and independent across all $i \in [N]$ due to Assumption 2. Moreover, Assumption 4 (i.e., Eq. (15)) provides that $\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s} \perp\!\!\!\perp \{\varepsilon_{i,j}^{(1)}\}_{i\in\mathcal{R}_s}$. Hence, applying the subGaussian concentration (Corollary A.1) for $\{\varepsilon_{i,j}^{(1)}\}_{i\in\mathcal{R}_s}$ yields that

$$\Big|\sum_{i\in\mathcal{R}_s} \frac{\varepsilon_{i,j}^{(1)} p_{i,j}}{\widehat{p}_{i,j}}\Big| \leq \overline{\sigma}\sqrt{c\ell_\delta} \sqrt{\sum_{i\in\mathcal{R}_s} \Big(\frac{p_{i,j}}{\widehat{p}_{i,j}}\Big)^2} \leq \overline{\sigma}\sqrt{c\ell_\delta} \big\|P_{\cdot,j} \oslash \widehat{P}_{\cdot,j}\big\|_2, \qquad \text{(A.8)}$$

with probability at least $1 - \delta$.

Finally, to bound the fourth term in Eq. (A.5), note that $\varepsilon_{i,j}^{(1)} \eta_{i,j}$ is subExponential$(\overline{\sigma}/\sqrt{\ell_1})$ due to Lemma A.3 as well as zero-mean and independent across all $i \in [N]$ due to Assumption 2.

Moreover, Assumption 4 (i.e., Eq. (15)) provides that $\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s} \perp\!\!\!\perp \{(\eta_{i,j}, \varepsilon^{(1)}_{i,j})\}_{i\in\mathcal{R}_s}$. Hence, applying the subExponential concentration (Corollary A.2) for $\{\varepsilon^{(1)}_{i,j}\eta_{i,j}\}_{i\in\mathcal{R}_s}$ yields that

$$\Big|\sum_{i\in\mathcal{R}_s}\frac{\varepsilon^{(1)}_{i,j}\eta_{i,j}}{\widehat{p}_{i,j}}\Big| \leq \frac{\overline{\sigma}m(c\ell_\delta)}{\sqrt{\ell_1}}\|\mathbf{1}_N \oslash \widehat{P}_{\cdot,j}\|_2, \tag{A.9}$$

with probability at least $1-\delta$. Putting together Eqs. (A.5) to (A.9), we conclude that, with probability at least $1-3\delta$,

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

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

$$\frac{1}{N}\Big|\sum_{i\in\mathcal{R}_s}\mathbb{T}^{(1,\mathrm{DR})}_{i,j}\Big| \leq \frac{1}{\bar{\lambda}N}\|\widehat{\Theta}^{(1)}-\Theta^{(1)}\|_{1,2}\|\widehat{P}-P\|_{1,2} + \frac{\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{\ell_1}N}\|\widehat{\Theta}^{(1)}-\Theta^{(1)}\|_{1,2}$$
$$+\frac{\overline{\sigma}\sqrt{c\ell_\delta}}{\bar{\lambda}N}\|P\|_{1,2} + \frac{\overline{\sigma}m(c\ell_\delta)}{\bar{\lambda}\sqrt{\ell_1}N}\|\mathbf{1}\|_{1,2} \tag{A.10}$$
$$\stackrel{(a)}{\leq} \frac{1}{\bar{\lambda}}\mathcal{E}\big(\widehat{\Theta}^{(1)}\big)\mathcal{E}\big(\widehat{P}\big) + \frac{\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{\ell_1 N}}\mathcal{E}\big(\widehat{\Theta}^{(1)}\big) + \frac{\overline{\sigma}\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{N}} + \frac{\overline{\sigma}m(c\ell_\delta)}{\bar{\lambda}\sqrt{\ell_1 N}}, \tag{A.11}$$

where (a) follows from Eq. (16) and because $\|P\|_{1,2}\leq\sqrt{N}$ and $\|\mathbf{1}\|_{1,2}=\sqrt{N}$. Then, the claim in Eq. (A.4) follows for $a=1$ by using Eq. (A.11) and applying a union bound over $s\in\{0,1\}$. The proof of Eq. (A.4) for $a=0$ follows similarly.

C. Proofs of Corollaries 1 and 2

C.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. From Eq. (20), with probability at least $1-\delta$,

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

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. Then, if $\mathcal{E}\big(\widehat{P}\big) = O_p\big(N^{-\alpha}\big)$, with probability at least $1-\delta$,

$$\big|\widehat{\mathrm{ATE}}^{\mathrm{IPW}}_{\cdot,j} - \mathrm{ATE}_{\cdot,j}\big| \leq O_p\big(N^{-\alpha}\big) + f_1(\delta)N^{-1/2} = N^{-\alpha}\Big(O_p\big(1\big) + f_1(\delta)N^{\alpha-1/2}\Big) \stackrel{(a)}{\leq} N^{-\alpha}\Big(O_p\big(1\big) + f_1(\delta)\Big),$$

where (a) follows because $0 \leq \alpha \leq 0.5$. Then, from Lemma A.7, we have

$$\big|\widehat{\mathrm{ATE}}^{\mathrm{IPW}}_{\cdot,j} - \mathrm{ATE}_{\cdot,j}\big| \leq N^{-\alpha}O_p\big(1\big) = O_p\big(N^{-\alpha}\big).$$

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

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

where

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

Then, if $\mathcal{E}\big(\widehat{P}\big) = O_p\big(N^{-\alpha}\big)$ and $\mathcal{E}\big(\widehat{\Theta}\big) = O_p\big(N^{-\beta}\big)$, with probability at least $1-\delta$,

$$\big|\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,j} - \mathrm{ATE}_{\cdot,j}\big| \leq O_p\big(N^{-(\alpha+\beta)}\big) + f_2(\delta)N^{-1/2}. \tag{A.12}$$

Consider two cases. First, suppose $\alpha+\beta \leq 0.5$. Then, we can rewrite Eq. (A.12) as follows

$$\big|\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,j} - \mathrm{ATE}_{\cdot,j}\big| \leq N^{-(\alpha+\beta)}\Big(O_p\big(1\big) + f_2(\delta)N^{\alpha+\beta-1/2}\Big) \stackrel{(a)}{\leq} N^{-(\alpha+\beta)}\Big(O_p\big(1\big) + f_2(\delta)\Big), \tag{A.13}$$

where (a) follows because $\alpha+\beta \leq 0.5$. Next, suppose $\alpha+\beta > 0.5$. Then, we can rewrite Eq. (A.12) as follows

$$\big|\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,j} - \mathrm{ATE}_{\cdot,j}\big| \leq N^{-0.5}\Big(O_p\big(N^{0.5-\alpha-\beta}\big) + f_2(\delta)\Big) \stackrel{(a)}{\leq} N^{-0.5}\Big(O_p\big(1\big) + f_2(\delta)\Big), \tag{A.14}$$

where (a) follows because $\alpha+\beta > 0.5$. Combining Eqs. (A.13) and (A.14), we have

$$\big|\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,j} - \mathrm{ATE}_{\cdot,j}\big| \leq N^{-\min\{\alpha+\beta,0.5\}}\Big(O_p\big(1\big) + f_2(\delta)\Big).$$

Then, from Lemma A.7, we have

$$\big|\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,j} - \mathrm{ATE}_{\cdot,j}\big| \leq N^{-\min\{\alpha+\beta,0.5\}}O_p\big(1\big) = O_p\big(N^{-\min\{\alpha+\beta,0.5\}}\big).$$

C.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 A.5.

D. 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.2) and Eq. (A.3), respectively. Then, define

$$\mathbb{X}_{i,j}^{(1,\mathrm{DR})} \triangleq \mathbb{T}_{i,j}^{(1,\mathrm{DR})} - \varepsilon_{i,j}^{(1)} - \frac{\varepsilon_{i,j}^{(1)}\eta_{i,j}}{p_{i,j}} \tag{A.15}$$

$$\mathbb{X}_{i,j}^{(0,\mathrm{DR})} \triangleq \mathbb{T}_{i,j}^{(0,\mathrm{DR})} + \varepsilon_{i,j}^{(0)} - \frac{\varepsilon_{i,j}^{(0)}\eta_{i,j}}{1-p_{i,j}},$$

and

$$\mathbb{Z}_{i,j}^{\mathrm{DR}} \triangleq \varepsilon_{i,j}^{(1)} + \frac{\varepsilon_{i,j}^{(1)}\eta_{i,j}}{p_{i,j}} - \varepsilon_{i,j}^{(0)} + \frac{\varepsilon_{i,j}^{(0)}\eta_{i,j}}{1-p_{i,j}}. \tag{A.16}$$

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 D.1 (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 D.2 (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.

D.1. Proof of Lemma D.1

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

$$\frac{1}{\sqrt{N}}\sum_{i\in[N]}\mathbb{X}_{i,j}^{(a,\mathrm{DR})} \le O\Big(\sqrt{N}\mathcal{E}\big(\widehat{\Theta}^{(a)}\big)\mathcal{E}\big(\widehat{P}\big)\Big) + O_p\Big(\mathcal{E}\big(\widehat{\Theta}^{(a)}\big)\Big) + O_p\Big(\mathcal{E}\big(\widehat{P}\big)\Big). \tag{A.17}$$

We provide a proof of this claim at the end of this section. Then, using Eq. (A.17) and the fact that $\overline{\sigma}_j \geq c > 0$ as per condition (C3), we obtain the following,

$$\frac{1}{\overline{\sigma}_j\sqrt{N}}\sum_{i\in[N]}\Big(\mathbb{X}^{(1,\mathrm{DR})}_{i,j}+\mathbb{X}^{(0,\mathrm{DR})}_{i,j}\Big)\leq\frac{1}{c}\Big(O\Big(\sqrt{N}\mathcal{E}(\widehat{\Theta})\mathcal{E}(\widehat{P})\Big)+O_p\Big(\mathcal{E}(\widehat{\Theta})\Big)+O_p\Big(\mathcal{E}(\widehat{P})\Big)\Big)$$

$$\overset{(a)}{=}\frac{1}{c}\Big(\sqrt{N}o_p(N^{-1/2})+O_p\Big(o_p(1)\Big)+O_p\Big(o_p(1)\Big)\Big)\overset{(b)}{=}o_p(1),$$

where (a) follows from (C1) and (C2), and (b) follows because $O_p\Big(o_p(1)\Big)=o_p(1)$ and $o_p(1)+o_p(1)=o_p(1)$.

Proof of Eq. (A.17) 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.17) 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.2) and (A.15) imply that

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

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

$$\begin{aligned}\frac{1}{\sqrt{N}}\Big|\sum_{i\in\mathcal{R}_s}\mathbb{X}^{(1,\mathrm{DR})}_{i,j}\Big|&\leq\frac{1}{\sqrt{N}}\Big|\sum_{i\in\mathcal{R}_s}\frac{\big(\widehat{\theta}^{(1)}_{i,j}-\theta^{(1)}_{i,j}\big)\big(\widehat{p}_{i,j}-p_{i,j}\big)}{\widehat{p}_{i,j}}\Big|+\frac{1}{\sqrt{N}}\Big|\sum_{i\in\mathcal{R}_s}\frac{\big(\widehat{\theta}^{(1)}_{i,j}-\theta^{(1)}_{i,j}\big)\eta_{i,j}}{\widehat{p}_{i,j}}\Big|\\&\quad+\frac{1}{\sqrt{N}}\Big|\sum_{i\in\mathcal{R}_s}\frac{\varepsilon^{(1)}_{i,j}\big(\widehat{p}_{i,j}-p_{i,j}\big)}{\widehat{p}_{i,j}}\Big|+\frac{1}{\sqrt{N}}\Big|\sum_{i\in\mathcal{R}_s}\frac{\varepsilon^{(1)}_{i,j}\eta_{i,j}\big(\widehat{p}_{i,j}-p_{i,j}\big)}{\widehat{p}_{i,j}p_{i,j}}\Big|.\end{aligned}\tag{A.18}$$

To control the first term in Eq. (A.18), we use the Cauchy-Schwarz inequality and Assumption 3 as in Appendix B (see Eqs. (A.6), (A.10), and (A.11)).

To control the second term in Eq. (A.18), we condition on $\{(\widehat{p}_{i,j},\widehat{\theta}^{(1)}_{i,j})\}_{i\in\mathcal{R}_s}$. Then, Assumption 4 (i.e., Eq. (14)) provides that $\{(\widehat{p}_{i,j},\widehat{\theta}^{(1)}_{i,j})\}_{i\in\mathcal{R}_s}\perp\!\!\!\perp\{\eta_{i,j}\}_{i\in\mathcal{R}_s}$. As a result, $\sum_{i\in\mathcal{R}_s}\big(\widehat{\theta}^{(1)}_{i,j}-\theta^{(1)}_{i,j}\big)\eta_{i,j}/\widehat{p}_{i,j}$ is subGaussian$\big(\big[\sum_{i\in\mathcal{R}_s}\big(\widehat{\theta}^{(1)}_{i,j}-\theta^{(1)}_{i,j}\big)^2/\big(\widehat{p}_{i,j}\big)^2\big]^{1/2}/\sqrt{\ell_1}\big)$ because $\eta_{i,j}$ 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). Then, we have

$$\frac{1}{\sqrt{N}}\mathbb{E}\Big[\Big|\sum_{i\in\mathcal{R}_s}\frac{\big(\widehat{\theta}^{(1)}_{i,j}-\theta^{(1)}_{i,j}\big)\eta_{i,j}}{\widehat{p}_{i,j}}\Big|\Big|\{(\widehat{p}_{i,j},\widehat{\theta}^{(1)}_{i,j})\}_{i\in\mathcal{R}_s}\Big]\overset{(a)}{\leq}\frac{c}{\sqrt{N}}\sqrt{\sum_{i\in\mathcal{R}_s}\Big(\frac{\widehat{\theta}^{(1)}_{i,j}-\theta^{(1)}_{i,j}}{\widehat{p}_{i,j}}\Big)^2}$$

$$
\begin{aligned}
&\leq \frac{c}{\sqrt{N}}\big\|\big(\widehat{\Theta}^{(1)}_{\cdot,j}-\Theta^{(1)}_{\cdot,j}\big) \oslash \widehat{P}_{\cdot,j}\big\|_2 \\
&\overset{(b)}{\leq} \frac{c}{\bar{\lambda}}\mathcal{E}\big(\widehat{\Theta}^{(1)}\big),
\end{aligned} \tag{A.19}
$$

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

To control the third term in Eq. (A.18), we condition on $\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s}$. Then, Assumption 4 (i.e., Eq. (15)) provides that $\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s} \perp\!\!\!\perp \{\varepsilon^{(1)}_{i,j}\}_{i\in\mathcal{R}_s}$. As a result, $\sum_{i\in\mathcal{R}_s}\varepsilon^{(1)}_{i,j}\big(\widehat{p}_{i,j}-p_{i,j}\big)/\widehat{p}_{i,j}$ 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 $\varepsilon^{(1)}_{i,j}$ is subGaussian$(\overline{\sigma})$, zero-mean, and independent across all $i \in [N]$ due to Assumption 2. Then, we have

$$
\begin{aligned}
\frac{1}{\sqrt{N}}\mathbb{E}\Big[\Big|\sum_{i\in\mathcal{R}_s}\frac{\varepsilon^{(1)}_{i,j}\big(\widehat{p}_{i,j}-p_{i,j}\big)}{\widehat{p}_{i,j}}\Big|\Big|\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s}\Big] &\overset{(a)}{\leq} \frac{c\overline{\sigma}}{\sqrt{N}}\sqrt{\sum_{i\in\mathcal{R}_s}\Big(\frac{\widehat{p}_{i,j}-p_{i,j}}{\widehat{p}_{i,j}}\Big)^2} \\
&\leq \frac{c\overline{\sigma}}{\sqrt{N}}\big\|\big(\widehat{P}_{\cdot,j}-P_{\cdot,j}\big) \oslash \widehat{P}_{\cdot,j}\big\|_2 \\
&\overset{(b)}{\leq} \frac{c\overline{\sigma}}{\bar{\lambda}}\mathcal{E}\big(\widehat{P}\big),
\end{aligned} \tag{A.20}
$$

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

To control the fourth term in Eq. (A.18), we condition on $\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s}$. Then, Assumption 4 (i.e., Eq. (15)) provides that $\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s} \perp\!\!\!\perp \{(\eta_{i,j},\varepsilon^{(1)}_{i,j})\}_{i\in\mathcal{R}_s}$. As a result, $\sum_{i\in\mathcal{R}_s}\varepsilon^{(1)}_{i,j}\eta_{i,j}\big(\widehat{p}_{i,j} - p_{i,j}\big)/\widehat{p}_{i,j}p_{i,j}$ 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 $\varepsilon^{(1)}_{i,j}\eta_{i,j}$ is subExponential$(\overline{\sigma}/\sqrt{\ell_1})$ due to Lemma A.3 as well as zero-mean and independent across all $i \in [N]$ due to Assumption 2. Then, we have

$$
\begin{aligned}
\frac{1}{\sqrt{N}}\mathbb{E}\Big[\Big|\sum_{i\in\mathcal{R}_s}\frac{\varepsilon^{(1)}_{i,j}\eta_{i,j}\big(\widehat{p}_{i,j}-p_{i,j}\big)}{\widehat{p}_{i,j}p_{i,j}}\Big|\Big|\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s}\Big] &\overset{(a)}{\leq} \frac{c\overline{\sigma}}{\sqrt{N}}\sqrt{\sum_{i\in\mathcal{R}_s}\Big(\frac{\widehat{p}_{i,j}-p_{i,j}}{\widehat{p}_{i,j}p_{i,j}}\Big)^2} \\
&\leq \frac{c\overline{\sigma}}{\sqrt{N}}\big\|\big(\widehat{P}_{\cdot,j}-P_{\cdot,j}\big) \oslash \big(\widehat{P}_{\cdot,j} \odot P_{\cdot,j}\big)\big\|_2 \\
&\overset{(b)}{\leq} \frac{c\overline{\sigma}}{\bar{\lambda}\lambda}\mathcal{E}\big(\widehat{P}\big),
\end{aligned} \tag{A.21}
$$

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

Putting together Eqs. (A.18) to (A.21) using Lemma A.6, we have

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

Then, the claim in Eq. (A.17) follows for $a = 1$ by using $|\sum_{i\in[N]}\mathbb{X}^{(1,\mathrm{DR})}_{i,j}| \leq |\sum_{i\in\mathcal{R}_0}\mathbb{X}^{(1,\mathrm{DR})}_{i,j}| + |\sum_{i\in\mathcal{R}_1}\mathbb{X}^{(1,\mathrm{DR})}_{i,j}|$. The proof of Eq. (A.17) for $a = 0$ follows similarly.

D.2. Proof of Lemma D.2

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

Lemma D.3 (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.,*

$$\frac{\sum_{i=1}^{N} \mathbb{E}[|x_i|^{2+\omega}]}{(\sum_{i=1}^{N} \mathbb{E}[x_i^2])^{\frac{2+\omega}{2}}} \longrightarrow 0, \tag{A.22}$$

as $N \to \infty$. Then,

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

as $N \to \infty$.

Fix any $j \in [M]$. We apply Lyapunov CLT in Lemma D.3 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.16). Note that Assumption 2(a) implies $\mathbb{E}[\mathbb{Z}_{i,j}^{\mathrm{DR}}] = 0$ for all $i \in [N]$, and Assumption 2(b) implies that $\mathbb{Z}_{i,j}^{\mathrm{DR}} \perp\!\!\!\perp \mathbb{Z}_{i',j}^{\mathrm{DR}}$ for all $i \neq i' \in [N]$. First, we show in Appendix D.2.1 that

$$\mathbb{V}\mathrm{ar}(\mathbb{Z}_{i,j}^{\mathrm{DR}}) = \frac{(\sigma_{i,j}^{(1)})^2}{p_{i,j}} + \frac{(\sigma_{i,j}^{(0)})^2}{1 - p_{i,j}}, \tag{A.23}$$

for each $i \in [N]$. Next, we show in Appendix D.2.2 that Lyapunov's condition (A.22) holds for the sequence $\mathbb{Z}_{1,j}^{\mathrm{DR}}, \mathbb{Z}_{2,j}^{\mathrm{DR}}, \cdots$ with $\omega = 1$. Finally, applying Lemma D.3 and using the definition of $\overline{\sigma}_j$ from Eq. (22) yields Lemma D.2.

D.2.1. Proof of Eq. (A.23)

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

$$\mathbb{V}\mathrm{ar}\Big(\mathbb{Z}_{i,j}^{\mathrm{DR}}\Big) = \mathbb{V}\mathrm{ar}\bigg(\varepsilon_{i,j}^{(1)}\Big(1 + \frac{\eta_{i,j}}{p_{i,j}}\Big) - \varepsilon_{i,j}^{(0)}\Big(1 - \frac{\eta_{i,j}}{1 - p_{i,j}}\Big)\bigg). \tag{A.24}$$

We claim the following:

$$\mathbb{V}\mathrm{ar}\bigg(\varepsilon_{i,j}^{(1)}\Big(1 + \frac{\eta_{i,j}}{p_{i,j}}\Big)\bigg) = \frac{(\sigma_{i,j}^{(1)})^2}{p_{i,j}}, \tag{A.25}$$

$$\mathbb{V}\mathrm{ar}\bigg(\varepsilon_{i,j}^{(0)}\Big(1 - \frac{\eta_{i,j}}{1 - p_{i,j}}\Big)\bigg) = \frac{(\sigma_{i,j}^{(0)})^2}{1 - p_{i,j}}, \tag{A.26}$$

and

$$\mathbb{C}\mathrm{ov}\bigg(\varepsilon_{i,j}^{(1)}\Big(1 + \frac{\eta_{i,j}}{p_{i,j}}\Big), \varepsilon_{i,j}^{(0)}\Big(1 - \frac{\eta_{i,j}}{1 - p_{i,j}}\Big)\bigg) = 0, \tag{A.27}$$

with Eq. (A.23) following from Eqs. (A.24) to (A.27).

To establish Eq. (A.25), 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.26). Eq. (A.27) 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 $\mathbb{E}[\eta_{i,j}^2] = \mathbb{V}\mathrm{ar}(\eta_{i,j}) = p_{i,j}(1-p_{i,j})$.

D.2.2. Proof of Lyapunov's condition with $\omega = 1$
We have

$$\begin{aligned}
\frac{\sum_{i\in[N]}\mathbb{E}\big[|\mathbb{Z}_{i,j}^{\mathrm{DR}}|^3\big]}{\big(\sum_{i\in[N]}\mathbb{V}\mathrm{ar}(\mathbb{Z}_{i,j}^{\mathrm{DR}})\big)^{3/2}} &= \frac{1}{N^{3/2}}\frac{\sum_{i\in[N]}\mathbb{E}\big[|\mathbb{Z}_{i,j}^{\mathrm{DR}}|^3\big]}{\big(\frac{1}{N}\sum_{i\in[N]}\mathbb{V}\mathrm{ar}(\mathbb{Z}_{i,j}^{\mathrm{DR}})\big)^{3/2}} \overset{(a)}{=} \frac{1}{N^{3/2}}\frac{\sum_{i\in[N]}\mathbb{E}\big[|\mathbb{Z}_{i,j}^{\mathrm{DR}}|^3\big]}{\big(\overline{\sigma}_j\big)^{3/2}} \\
&\overset{(b)}{\leq} \frac{1}{N^{3/2}}\frac{\sum_{i\in[N]}\mathbb{E}\big[|\mathbb{Z}_{i,j}^{\mathrm{DR}}|^3\big]}{c_1^{3/2}} \overset{(c)}{\leq} \frac{1}{N^{1/2}}\frac{c_2}{c_1^{3/2}},
\end{aligned} \tag{A.28}$$

where (a) follows by putting together Eqs. (22) and (A.23), (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 (A.22) holds for $\omega = 1$ as the right hand side of Eq. (A.28) goes to zero as $N \to \infty$.

D.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

$$\begin{aligned}
\Delta_j &= \frac{1}{N}\sum_{i\in[N]}\bigg(\frac{\big(\widehat{\theta}_{i,j}^{(1)} - y_{i,j}\big)^2 a_{i,j}}{\big(\widehat{p}_{i,j}\big)^2} + \frac{\big(\widehat{\theta}_{i,j}^{(0)} - y_{i,j}\big)^2(1-a_{i,j})}{\big(1-\widehat{p}_{i,j}\big)^2}\bigg) - \bigg(\frac{(\sigma_{i,j}^{(1)})^2}{p_{i,j}} + \frac{(\sigma_{i,j}^{(0)})^2}{1-p_{i,j}}\bigg)\\
&= \frac{1}{N}\sum_{i\in[N]}\bigg(\frac{\big(\widehat{\theta}_{i,j}^{(1)} - y_{i,j}\big)^2 a_{i,j}}{\big(\widehat{p}_{i,j}\big)^2} - \frac{(\sigma_{i,j}^{(1)})^2}{p_{i,j}}\bigg) + \bigg(\frac{\big(\widehat{\theta}_{i,j}^{(0)} - y_{i,j}\big)^2(1-a_{i,j})}{\big(1-\widehat{p}_{i,j}\big)^2} - \frac{(\sigma_{i,j}^{(0)})^2}{1-p_{i,j}}\bigg)\\
&\overset{(a)}{=} \frac{1}{N}\sum_{i\in[N]}\Big(\mathbb{T}_{i,j}^{(1)} + \mathbb{T}_{i,j}^{(0)}\Big),
\end{aligned} \tag{A.29}$$

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

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

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

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

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

$$\frac{1}{N}\Big|\sum_{i\in[N]}\mathbb{T}_{i,j}^{(\mathrm{a})}\Big| \le O\Big(\mathcal{E}\big(\widehat{\Theta}^{(a)}\big)\Big)^2 + O_p\bigg(\frac{\mathcal{E}\big(\widehat{\Theta}^{(a)}\big)}{\sqrt{N}}\bigg) + O_p\bigg(\frac{1}{\sqrt{N}}\bigg) + O\Big(\mathcal{E}\big(\widehat{P}\big)\Big). \tag{A.30}$$

We provide a proof of this claim at the end of this section. Applying triangle inequality in Eq. (A.29), we obtain the following

$$\begin{aligned}\Delta_j &\leq O\Big(\mathcal{E}\big(\widehat{\Theta}\big)\Big)^2 + O_p\bigg(\frac{\mathcal{E}\big(\widehat{\Theta}\big)}{\sqrt{N}}\bigg) + O_p\bigg(\frac{1}{\sqrt{N}}\bigg) + O\Big(\mathcal{E}\big(\widehat{P}\big)\Big) \\ &\overset{(a)}{=} O\Big(o_p(1)\Big)^2 + O_p\bigg(\frac{o_p(1)}{\sqrt{N}}\bigg) + O_p\bigg(\frac{1}{\sqrt{N}}\bigg) + O\Big(o_p(1)\Big) \overset{(b)}{=} o_p(1),\end{aligned}$$

where (a) follows from (C1), and (b) follows because $O_p\Big(o_p(1)\Big) = o_p(1)$, $o_p(1)o_p(1) = o_p(1)$, and $o_p(1) + o_p(1) = o_p(1)$.

Proof of bound Eq. (A.30). This proof follows a very similar road map to that used for establishing the inequality in Eq. (A.17). 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.30) 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

$$\begin{aligned}\frac{1}{N}\Big|\sum_{i\in\mathcal{R}_s} \mathbb{T}^{(1)}_{i,j}\Big| &\leq \frac{1}{N}\Big|\sum_{i\in\mathcal{R}_s} \frac{\big(\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}\big)^2 a_{i,j}}{\big(\widehat{p}_{i,j}\big)^2}\Big| + \frac{1}{N}\Big|\sum_{i\in\mathcal{R}_s} \frac{2\varepsilon^{(1)}_{i,j} p_{i,j}\big(\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}\big)}{\big(\widehat{p}_{i,j}\big)^2}\Big| \\ &+ \frac{1}{N}\Big|\sum_{i\in\mathcal{R}_s} \frac{2\varepsilon^{(1)}_{i,j}\eta_{i,j}\big(\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}\big)}{\big(\widehat{p}_{i,j}\big)^2}\Big| + \frac{1}{N}\Big|\sum_{i\in\mathcal{R}_s} \frac{\big(\varepsilon^{(1)}_{i,j}\big)^2\eta_{i,j}}{\big(\widehat{p}_{i,j}\big)^2}\Big| + \frac{1}{N}\Big|\sum_{i\in\mathcal{R}_s} \frac{\big(\varepsilon^{(1)}_{i,j}\big)^2 p_{i,j}}{\big(\widehat{p}_{i,j}\big)^2} - \frac{(\sigma^{(1)}_{i,j})^2}{p_{i,j}}\Big|.\end{aligned} \tag{A.31}$$

To bound the first term in Eq. (A.31), we have

$$\frac{1}{N}\Big|\sum_{i\in\mathcal{R}_s}\frac{\big(\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}\big)^2 a_{i,j}}{\big(\widehat{p}_{i,j}\big)^2}\Big| \overset{(a)}{\leq} \frac{1}{N}\Big|\sum_{i\in\mathcal{R}_s}\frac{\big(\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}\big)^2}{\big(\widehat{p}_{i,j}\big)^2}\Big| \overset{(b)}{\leq} \frac{1}{\bar{\lambda}^2 N}\|\widehat{\Theta}^{(1)}_{\cdot,j} - \Theta^{(1)}_{\cdot,j}\|_2^2 \overset{(c)}{=} \frac{1}{\bar{\lambda}^2}\Big[\mathcal{E}\big(\widehat{\Theta}^{(1)}\big)\Big]^2, \tag{A.32}$$

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

To control second term in Eq. (A.31), we condition on $\{\big(\widehat{p}_{i,j}, \widehat{\theta}^{(1)}_{i,j}\big)\}_{i\in\mathcal{R}_s}$. Then, Eq. (24) provides that $\{\big(\widehat{p}_{i,j}, \widehat{\theta}^{(1)}_{i,j}\big)\}_{i\in\mathcal{R}_s} \perp\!\!\!\perp \{\varepsilon^{(1)}_{i,j}\}_{i\in\mathcal{R}_s}$. 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

$$\frac{1}{N}\mathbb{E}\bigg[\Big|\sum_{i\in\mathcal{R}_s}\frac{2\varepsilon^{(1)}_{i,j} p_{i,j}\big(\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}\big)}{\big(\widehat{p}_{i,j}\big)^2}\Big|\Big|\{\big(\widehat{p}_{i,j}, \widehat{\theta}^{(1)}_{i,j}\big)\}_{i\in\mathcal{R}_s}\bigg] \overset{(a)}{\leq} \frac{c\overline{\sigma}}{N}\sqrt{\sum_{i\in\mathcal{R}_s}\bigg(\frac{p_{i,j}\big(\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}\big)}{\big(\widehat{p}_{i,j}\big)^2}\bigg)^2}$$

$$\overset{(b)}{\leq} \frac{c\overline{\sigma}}{\overline{\lambda}^2 N}\big\|\widehat{\Theta}^{(1)}_{\cdot,j} - \Theta^{(1)}_{\cdot,j}\big\|_2 \overset{(c)}{=} \frac{c\overline{\sigma}}{\overline{\lambda}^2}\frac{\mathcal{E}\big(\widehat{\Theta}^{(1)}\big)}{\sqrt{N}}, \quad (A.33)$$

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

To control third term in Eq. (A.31), we condition on $\{\big(\widehat{p}_{i,j}, \widehat{\theta}^{(1)}_{i,j}\big)\}_{i\in\mathcal{R}_s}$. Then, Eq. (24) provides that $\{\big(\widehat{p}_{i,j}, \widehat{\theta}^{(1)}_{i,j}\big)\}_{i\in\mathcal{R}_s} \perp\!\!\!\perp \{(\eta_{i,j}, \varepsilon^{(1)}_{i,j})\}_{i\in\mathcal{R}_s}$. 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 A.3 as well as zero-mean and independent across all $i \in [N]$ due to Assumption 2. Then, we have

$$\frac{1}{N}\mathbb{E}\Big[\Big|\sum_{i\in\mathcal{R}_s}\frac{2\varepsilon^{(1)}_{i,j}\eta_{i,j}\big(\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}\big)}{\big(\widehat{p}_{i,j}\big)^2}\Big|\Big|\{(\widehat{p}_{i,j}, \widehat{\theta}^{(1)}_{i,j})\}_{i\in\mathcal{R}_s}\Big] \overset{(a)}{\leq} \frac{c\overline{\sigma}}{N}\sqrt{\sum_{i\in\mathcal{R}_s}\bigg(\frac{\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}}{\big(\widehat{p}_{i,j}\big)^2}\bigg)^2}$$
$$\overset{(b)}{\leq} \frac{c\overline{\sigma}}{\overline{\lambda}^2 N}\big\|\widehat{\Theta}^{(1)}_{\cdot,j} - \Theta^{(1)}_{\cdot,j}\big\|_2 \overset{(c)}{=} \frac{c\overline{\sigma}}{\overline{\lambda}^2}\frac{\mathcal{E}\big(\widehat{\Theta}^{(1)}\big)}{\sqrt{N}}, \quad (A.34)$$

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

To control fourth term in Eq. (A.31), 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 A.4 as well as zero-mean and independent across all $i \in [N]$ due to Assumption 2. Then, we have

$$\frac{1}{N}\mathbb{E}\Big[\Big|\sum_{i\in\mathcal{R}_s}\frac{\big(\varepsilon^{(1)}_{i,j}\big)^2\eta_{i,j}}{\big(\widehat{p}_{i,j}\big)^2}\Big|\Big|\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s}\Big] \overset{(a)}{\leq} \frac{c\overline{\sigma}^2}{N}\sqrt{\sum_{i\in\mathcal{R}_s}\frac{1}{\big(\widehat{p}_{i,j}\big)^4}} \overset{(b)}{\leq} \frac{c\overline{\sigma}^2}{\overline{\lambda}^2\sqrt{N}}, \quad (A.35)$$

where (a) follows as the first moment of 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.31), we have

$$\Big|\sum_{i\in\mathcal{R}_s}\Big(\frac{\big(\varepsilon^{(1)}_{i,j}\big)^2 p_{i,j}}{\big(\widehat{p}_{i,j}\big)^2} - \frac{(\sigma^{(1)}_{i,j})^2}{p_{i,j}}\Big)\Big| = \Big|\sum_{i\in\mathcal{R}_s}\Big(\frac{\big(\varepsilon^{(1)}_{i,j}\big)^2 p_{i,j}}{\big(\widehat{p}_{i,j}\big)^2} - \frac{\big(\sigma^{(1)}_{i,j}\big)^2 p_{i,j}}{\big(\widehat{p}_{i,j}\big)^2} + \frac{\big(\sigma^{(1)}_{i,j}\big)^2 p_{i,j}}{\big(\widehat{p}_{i,j}\big)^2} - \frac{(\sigma^{(1)}_{i,j})^2}{p_{i,j}}\Big)\Big|$$
$$\overset{(a)}{\leq} \Big|\sum_{i\in\mathcal{R}_s}\Big(\frac{\big[\big(\varepsilon^{(1)}_{i,j}\big)^2 - \big(\sigma^{(1)}_{i,j}\big)^2\big]p_{i,j}}{\big(\widehat{p}_{i,j}\big)^2}\Big)\Big| + \Big|\sum_{i\in\mathcal{R}_s}\Big(\frac{\big(\sigma^{(1)}_{i,j}\big)^2 p_{i,j}}{\big(\widehat{p}_{i,j}\big)^2} - \frac{(\sigma^{(1)}_{i,j})^2}{p_{i,j}}\Big)\Big| \quad (A.36)$$

where (a) follows from the triangle inequality. To control the first term in Eq. (A.36), we condition on $\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s}$. Then, Eq. (24) provides that $\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s} \perp\!\!\!\perp \{\varepsilon^{(1)}_{i,j}\}_{i\in\mathcal{R}_s}$. Further, $\mathbb{E}[(\varepsilon^{(1)}_{i,j})^2 - (\sigma^{(1)}_{i,j})^2] = 0$ due to (C3) and Assumption 2. As a result, $\sum_{i\in\mathcal{R}_s}\big[\big(\varepsilon^{(1)}_{i,j}\big)^2 - \big(\sigma^{(1)}_{i,j}\big)^2\big]p_{i,j}/\big(\widehat{p}_{i,j}\big)^2$ is 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 subExponential$(\overline{\sigma}^2)$ and independent across all $i\in[N]$ due to Lemma A.3. Then, we have

$$\frac{1}{N}\mathbb{E}\Big[\Big|\sum_{i\in\mathcal{R}_s}\frac{\big[\big(\varepsilon_{i,j}^{(1)}\big)^2-\big(\sigma_{i,j}^{(1)}\big)^2\big]p_{i,j}}{\big(\widehat{p}_{i,j}\big)^2}\Big|\Big|\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s}\Big]\overset{(a)}{\leq}\frac{c\overline{\sigma}^2}{N}\sqrt{\sum_{i\in\mathcal{R}_s}\bigg(\frac{p_{i,j}}{\big(\widehat{p}_{i,j}\big)^2}\bigg)^2}\overset{(b)}{\leq}\frac{c\overline{\sigma}^2}{\bar{\lambda}^2\sqrt{N}},\quad\text{(A.37)}$$

where (a) follows as the first moment of subExponential(σ) is $O(\sigma)$ and (b) follows from Assumption 3. To bound the second term in Eq. (A.36), applying the Cauchy-Schwarz inequality yields that

$$\begin{aligned}\frac{1}{N}\Big|\sum_{i\in\mathcal{R}_s}\bigg(\frac{\big(\sigma_{i,j}^{(1)}\big)^2p_{i,j}}{\big(\widehat{p}_{i,j}\big)^2}-\frac{(\sigma_{i,j}^{(1)})^2}{p_{i,j}}\bigg)\Big|&=\frac{1}{N}\Big|\sum_{i\in\mathcal{R}_s}\frac{\big(\sigma_{i,j}^{(1)}\big)^2\big(\big(p_{i,j}\big)^2-\big(\widehat{p}_{i,j}\big)^2\big)}{\big(\widehat{p}_{i,j}\big)^2p_{i,j}}\Big|\\&\overset{(a)}{\leq}\frac{2}{N}\sum_{i\in\mathcal{R}_s}\frac{\big(\sigma_{i,j}^{(1)}\big)^2\big|p_{i,j}-\widehat{p}_{i,j}\big|}{\big(\widehat{p}_{i,j}\big)^2p_{i,j}}\\&\overset{(b)}{\leq}\frac{2\overline{\sigma}^2}{\lambda\bar{\lambda}^2N}\sum_{i\in\mathcal{R}_s}\big|p_{i,j}-\widehat{p}_{i,j}\big|\\&\overset{(c)}{\leq}\frac{2\overline{\sigma}^2}{\lambda\bar{\lambda}^2\sqrt{N}}\big\|P_{\cdot,j}-\widehat{P}_{\cdot,j}\big\|_2\overset{(d)}{=}\frac{2\overline{\sigma}^2}{\lambda\bar{\lambda}^2}\mathcal{E}\big(\widehat{P}\big),\end{aligned}\quad\text{(A.38)}$$

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, and (d) follows from Eq. (16).

Putting together Eqs. (A.31) to (A.38) using Lemma A.6,

$$\frac{1}{N}\Big|\sum_{i\in\mathcal{R}_s}\mathbb{T}_{i,j}^{(1)}\Big|\leq O\Big(\mathcal{E}\big(\widehat{\Theta}^{(1)}\big)\Big)^2+O_p\bigg(\frac{\mathcal{E}\big(\widehat{\Theta}^{(1)}\big)}{\sqrt{N}}\bigg)+O_p\bigg(\frac{1}{\sqrt{N}}\bigg)+O\Big(\mathcal{E}\big(\widehat{P}\big)\Big).$$

Then, the claim in Eq. (A.30) follows for $a=1$ by using $|\sum_{i\in[N]}\mathbb{T}_{i,j}^{(1)}|\leq|\sum_{i\in\mathcal{R}_0}\mathbb{T}_{i,j}^{(1)}|+|\sum_{i\in\mathcal{R}_1}\mathbb{T}_{i,j}^{(1)}|$. The proof of Eq. (A.30) for $a=0$ follows similarly.

E. 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

$$\big| \Delta \mathrm{ATE}_{\cdot,j}^{\mathrm{OI}} \big| \leq \frac{1}{N} \Big| \sum_{i \in [N]} \big(\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)} \big) \Big| + \frac{1}{N} \Big| \sum_{i \in [N]} \big(\widehat{\theta}_{i,j}^{(0)} - \theta_{i,j}^{(0)} \big) \Big|. \tag{A.39}$$

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

$$\frac{1}{N} \Big| \sum_{i \in [N]} \big(\widehat{\theta}_{i,j}^{(a)} - \theta_{i,j}^{(a)} \big) \Big| \leq \mathcal{E}\big(\widehat{\Theta}^{(a)} \big). \tag{A.40}$$

The proof is complete by putting together Eqs. (A.39) and (A.40).

Proof of Eq. (A.40) Fix any $a \in \{0, 1\}$. Using the Cauchy-Schwarz inequality, we have

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

F. 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

$$\begin{aligned}
\Delta \mathrm{ATE}_{\cdot,j}^{\mathrm{IPW}} &= \frac{1}{N} \sum_{i \in [N]} \Big(\frac{y_{i,j} a_{i,j}}{\widehat{p}_{i,j}} - \frac{y_{i,j}(1 - a_{i,j})}{1 - \widehat{p}_{i,j}} \Big) - \frac{1}{N} \sum_{i \in [N]} \Big(\theta_{i,j}^{(1)} - \theta_{i,j}^{(0)} \Big) \\
&= \frac{1}{N} \sum_{i \in [N]} \bigg(\Big(\frac{y_{i,j} a_{i,j}}{\widehat{p}_{i,j}} - \theta_{i,j}^{(1)} \Big) - \Big(\frac{y_{i,j}(1 - a_{i,j})}{1 - \widehat{p}_{i,j}} - \theta_{i,j}^{(0)} \Big) \bigg) \\
&\overset{(a)}{=} \frac{1}{N} \sum_{i \in [N]} \Big(\mathbb{T}_{i,j}^{(1,\mathrm{IPW})} + \mathbb{T}_{i,j}^{(0,\mathrm{IPW})} \Big),
\end{aligned} \tag{A.41}$$

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

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

$$= \frac{\theta^{(1)}_{i,j}\big(p_{i,j} - \widehat{p}_{i,j}\big)}{\widehat{p}_{i,j}} + \frac{\theta^{(1)}_{i,j}\eta_{i,j}}{\widehat{p}_{i,j}} + \frac{\varepsilon^{(1)}_{i,j} p_{i,j}}{\widehat{p}_{i,j}} + \frac{\varepsilon^{(1)}_{i,j}\eta_{i,j}}{\widehat{p}_{i,j}}, \tag{A.42}$$

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

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

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

$$\frac{1}{N}\Big|\sum_{i\in[N]} \mathbb{T}^{(\mathrm{a},\mathrm{IPW})}_{i,j}\Big| \leq \frac{2}{\bar{\lambda}}\|\Theta^{(a)}\|_{\max}\,\mathcal{E}\big(\widehat{P}\big) + \frac{2\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{\ell_1 N}}\|\Theta^{(a)}\|_{\max} + \frac{2\bar{\sigma}\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{N}} + \frac{2\bar{\sigma}m(c\ell_\delta)}{\bar{\lambda}\sqrt{\ell_1 N}}. \tag{A.43}$$

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.41) and using Eq. (A.43) with a union bound, we obtain that

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

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

Proof of Eq. (A.43). This proof follows a very similar road map to that used for establishing the inequality in Eq. (A.4). 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.43) 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. (A.42) and triangle inequality imply that

$$\Big|\sum_{i\in\mathcal{R}_s} \mathbb{T}^{(1,\mathrm{IPW})}_{i,j}\Big| \leq \Big|\sum_{i\in\mathcal{R}_s} \frac{\theta^{(1)}_{i,j}\big(p_{i,j} - \widehat{p}_{i,j}\big)}{\widehat{p}_{i,j}}\Big| + \Big|\sum_{i\in\mathcal{R}_s} \frac{\theta^{(1)}_{i,j}\eta_{i,j}}{\widehat{p}_{i,j}}\Big| + \Big|\sum_{i\in\mathcal{R}_s} \frac{\varepsilon^{(1)}_{i,j} p_{i,j}}{\widehat{p}_{i,j}}\Big| + \Big|\sum_{i\in\mathcal{R}_s} \frac{\varepsilon^{(1)}_{i,j}\eta_{i,j}}{\widehat{p}_{i,j}}\Big|. \tag{A.44}$$

Next, note that the decomposition in Eq. (A.44) is identical to the one in Eq. (A.5), except for the fact when compared to Eq. (A.5), the first two terms in Eq. (A.44) have a factor of $\theta^{(1)}_{i,j}$ instead of $\big(\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}\big)$. As a result, mimicking steps used to derive Eq. (A.10), we obtain the following bound, with probability at least $1 - 3\delta$,

$$\frac{1}{N}\Big|\sum_{i\in\mathcal{R}_s} \mathbb{T}^{(1,\mathrm{IPW})}_{i,j}\Big| \leq \frac{1}{\bar{\lambda}N}\|\Theta^{(1)}\|_{1,2}\|\widehat{P} - P\|_{1,2} + \frac{\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{\ell_1}N}\|\Theta^{(1)}\|_{1,2} + \frac{\bar{\sigma}\sqrt{c\ell_\delta}}{\bar{\lambda}N}\|P\|_{1,2} + \frac{\bar{\sigma}m(c\ell_\delta)}{\bar{\lambda}\sqrt{\ell_1}N}\|\mathbf{1}\|_{1,2}$$

$$\overset{(a)}{\leq} \frac{1}{\bar{\lambda}\sqrt{N}} \|\Theta^{(1)}\|_{\max} \|\widehat{P} - P\|_{1,2} + \frac{\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{\ell_1 N}} \|\Theta^{(1)}\|_{\max} + \frac{\overline{\sigma}\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{N}} + \frac{\overline{\sigma} m(c\ell_\delta)}{\bar{\lambda}\sqrt{\ell_1 N}}$$

$$\overset{(b)}{\leq} \frac{1}{\bar{\lambda}} \|\Theta^{(1)}\|_{\max} \mathcal{E}\big(\widehat{P}\big) + \frac{\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{\ell_1 N}} \|\Theta^{(1)}\|_{\max} + \frac{\overline{\sigma}\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{N}} + \frac{\overline{\sigma} m(c\ell_\delta)}{\bar{\lambda}\sqrt{\ell_1 N}}, \tag{A.45}$$

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. (A.43) follows for $a = 1$ by using Eq. (A.45) and applying a union bound over $s \in \{0, 1\}$. The proof of Eq. (A.43) for $a = 0$ follows similarly.

G. Proofs of Propositions 3 and 4

In Appendix G.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 $\sqrt{N}/M = o(1)$.

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

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

$$\widehat{P}_{\mathcal{I}}, \widehat{\Theta}^{(a)}_{\mathcal{I}} \perp\!\!\!\perp W_{\mathcal{I}} \tag{A.46}$$

and

$$\widehat{P}_{\mathcal{I}} \perp\!\!\!\perp W_{\mathcal{I}}, E^{(a)}_{\mathcal{I}}, \tag{A.47}$$

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. (A.46) and (A.47), 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 $Y^{(a),\mathrm{obs}} \otimes \mathbf{1}^{-\mathcal{I}}$. In other words, the randomness in $\big(\widehat{P}_{\mathcal{I}}, \widehat{\Theta}^{(a)}_{\mathcal{I}}\big)$ stems from the randomness in $\big(A_{-\mathcal{I}}, Y^{(a),\mathrm{obs}}_{-\mathcal{I}}\big)$ which in turn stems from the randomness in $\big(W_{-\mathcal{I}}, E^{(a)}_{-\mathcal{I}}\big)$. Then, Eq. (A.46) 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. (A.47) follows from Eq. (28).

To prove Eq. (24), we show that

$$\widehat{P}_{\mathcal{I}}, \widehat{\Theta}^{(a)}_{\mathcal{I}} \perp\!\!\!\perp W_{\mathcal{I}}, E^{(a)}_{\mathcal{I}}, \tag{A.48}$$

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 $\big(A_{-\mathcal{I}}, Y^{(a),\mathrm{obs}}_{-\mathcal{I}}\big)$ which in turn stems from the randomness in $\big(W_{-\mathcal{I}}, E^{(a)}_{-\mathcal{I}}\big)$. Then, Eq. (A.48) follows from Eq. (33).

G.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 10 (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 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 11 (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) $\{H_{i,\cdot} : i \in [N]\}$ *are mutually independent (across* i*).*

Assumption 12 (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 10. Let* $\mathcal{R}_{\mathrm{obs}} \subseteq [N]$ *and* $\mathcal{C}_{\mathrm{obs}} \subseteq [M]$ *denote the set of rows and columns of* S*, respectively, with all entries observed, and* $\mathcal{R}_{\mathrm{miss}} \triangleq [N] \setminus \mathcal{R}_{\mathrm{obs}}$ *and* $\mathcal{C}_{\mathrm{miss}} \triangleq [M] \setminus \mathcal{C}_{\mathrm{obs}}$*. Let* $\widetilde{U}^{\mathrm{obs}} \in \mathbb{R}^{|\mathcal{R}_{\mathrm{obs}}| \times r_T}$ *and* $\widetilde{U}^{\mathrm{miss}} \in \mathbb{R}^{|\mathcal{R}_{\mathrm{miss}}| \times r_T}$ *be the sub-matrices of* $\widetilde{U}$ *that keeps the rows corresponding to the indices in* $\mathcal{R}_{\mathrm{obs}}$ *and* $\mathcal{R}_{\mathrm{miss}}$*, respectively. Let* $\widetilde{V}^{\mathrm{obs}} \in \mathbb{R}^{|\mathcal{C}_{\mathrm{obs}}| \times r_T}$ *and* $\widetilde{V}^{\mathrm{miss}} \in \mathbb{R}^{|\mathcal{C}_{\mathrm{miss}}| \times r_T}$ *be the sub-matrices of* $\widetilde{V}$ *that keeps the rows corresponding to the indices in* $\mathcal{C}_{\mathrm{obs}}$ *and* $\mathcal{C}_{\mathrm{miss}}$*, respectively. Then, the matrices defined below are positive definite:*

$$\lim_{N\to\infty} \frac{\widetilde{U}^{\mathrm{obs}\top} \widetilde{U}^{\mathrm{obs}}}{|\mathcal{R}_{\mathrm{obs}}|}, \quad \lim_{M\to\infty} \frac{\widetilde{U}^{\mathrm{miss}\top} \widetilde{U}^{\mathrm{miss}}}{|\mathcal{R}_{\mathrm{miss}}|}, \quad \lim_{N\to\infty} \frac{\widetilde{V}^{\mathrm{obs}\top} \widetilde{V}^{\mathrm{obs}}}{|\mathcal{C}_{\mathrm{obs}}|}, \quad \textit{and} \quad \lim_{M\to\infty} \frac{\widetilde{V}^{\mathrm{miss}\top} \widetilde{V}^{\mathrm{miss}}}{|\mathcal{C}_{\mathrm{miss}}|}. \tag{A.49}$$

Further, the mask matrix F *is such that*

$$|\mathcal{R}_{\mathrm{obs}}| = \Omega(N), \qquad |\mathcal{R}_{\mathrm{miss}}| = \Omega(N), \qquad |\mathcal{C}_{\mathrm{obs}}| = \Omega(M), \quad \textit{and} \quad |\mathcal{C}_{\mathrm{miss}}| = \Omega(M). \tag{A.50}$$

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) using the `TW` algorithm (summarized in Section 5.2.1). Its proof, essentially established as a corollary of Bai and Ng (2021, Lemma A.1), is provided in Appendix G.3.

Corollary G.1. *Consider a matrix of interest T, a noise matrix H, and a mask matrix F such that that Assumptions 10 to 12 hold. Let $S \in \{\mathbb{R}, ?\}^{N\times M}$ 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*

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

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 12, i.e., Eq. (A.50) under Assumption 8.

G.2.1. Estimating P.

Consider estimating P using `Cross-Fitted-SVD`. To apply Corollary G.1, we use Assumptions 6 and 7 to note that P satisfies Assumption 10 with rank parameter r_p. Then, we use Eq. (4), Assumption 2, and Assumption 9 to note that W satisfies Assumption 11. Finally, we use Assumption 8 to note that Assumption 12 holds. Step 2 of `Cross-Fitted-SVD` can be rewritten as $\widehat{P} = \texttt{Proj}_{\bar{\lambda}}\big(\overline{P}\big)$ and $\overline{P} = \texttt{Cross-Fitted-MC}(\texttt{TW}_{r_1}, A, \mathcal{P})$ where $r_1 = r_p$. Then,

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

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 G.1. Applying a union bound over all $\mathcal{I} \in \mathcal{P}$, we have

$$\mathcal{E}\big(\widehat{P}\big) \overset{(a)}{\leq} \|\widehat{P} - P\|_{\max} = O_p\Big(\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{M}}\Big), \tag{A.51}$$

where (a) follows from the definition of $L_{1,2}$ norm.

G.2.2. Estimating $\Theta^{(0)}$ and $\Theta^{(1)}$.

For every $a \in \{0, 1\}$, we show that

$$\mathcal{E}\big(\widehat{\Theta}^{(a)}\big) = O_p\Big(\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{M}}\Big). \tag{A.52}$$

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

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

$$\overset{(b)}{\leq} \|\Theta^{(1)}\|_{\max}\|\widehat{P} - P\|_{\max}, \tag{A.53}$$

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

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

where (a) follows from the definition of $L_{1,2}$ norm, (b) follows from Eq. (A.53), and (c) follows from Eq. (A.51). Then, Eq. (A.52) 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

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

Next, we utilize two claims proven in Appendices G.2.3 and G.2.4 respectively: $\Theta^{(1)} \odot P$ satisfies Assumption 10 with rank parameter $r_{\theta_1} r_p$ and

$$\bar{\varepsilon}^{(1)} \triangleq \Theta^{(1)} \odot \eta + \varepsilon^{(1)} \odot P + \varepsilon^{(1)} \odot \eta,$$

satisfies Assumption 11. Finally, we use Assumption 8 to note that Assumption 12 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 G.1,

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

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

$$\|\widehat{\Theta}^{(1)} \odot \widehat{P} - \Theta^{(1)} \odot P\|_{\max} = O_p\Big(\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{M}}\Big). \tag{A.54}$$

The left hand side of Eq. (A.54) can be written as,

$$\begin{aligned}
\|\widehat{\Theta}^{(1)} \odot \widehat{P} - \Theta^{(1)} \odot P\|_{\max} &= \|\widehat{\Theta}^{(1)} \odot \widehat{P} - \Theta^{(1)} \odot \widehat{P} + \Theta^{(1)} \odot \widehat{P} - \Theta^{(1)} \odot P\|_{\max} \\
&\overset{(a)}{\geq} \|\big(\widehat{\Theta}^{(1)} - \Theta^{(1)}\big) \odot \widehat{P}\|_{\max} - \|\Theta^{(1)} \odot \big(\widehat{P} - P\big)\|_{\max} \\
&\overset{(b)}{\geq} \bar{\lambda}\|\widehat{\Theta}^{(1)} - \Theta^{(1)}\|_{\max} - \|\Theta^{(1)}\|_{\max}\|\widehat{P} - P\|_{\max},
\end{aligned} \tag{A.55}$$

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,

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

where (a) follows from the definition of $L_{1,2}$ norm, (b) follows from Eq. (A.55), and (c) follows from Eqs. (A.51) and (A.54). Then, Eq. (A.52) follows as $1/\bar{\lambda}$ and $\|\Theta^{(1)}\|_{\max}$ are assumed to be bounded.

G.2.3. Proof that $\Theta^{(0)} \odot (\mathbf{1} - P)$ and $\Theta^{(1)} \odot P$ satisfy Assumption 10.
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

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

where (a) follows from the definition of Khatri-Rao product 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 10(a) holds from Eq. (35). Next, we note that

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

where (a) follows from the definition of Khatri-Rao product (see Section 1), and (b) follows from Assumption 7. Then, $\Theta^{(1)} \odot P$ satisfies Assumption 10(b) by using similar arguments on $\overline{V}^{(1)}$. Further, $\Theta^{(0)} \odot (\mathbf{1} - P)$ satisfies Assumption 10(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 10(c) holds from Assumption 7.

G.2.4. Proof that $\overline{\varepsilon}^{(1)}$ satisfies Assumption 11
Recall that $\overline{\varepsilon}^{(1)} \triangleq \Theta^{(1)} \odot \eta + \varepsilon^{(1)} \odot P + \varepsilon^{(1)} \odot \eta$. Then, Assumption 11(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 A.3, every subGaussian random variable is subExponential random variable, and sum of subExponential random variables is a subExponential random variable. Finally, Assumption 11(b) and Assumption 11(b) hold from Assumption 9(b) and Assumption 9(c), respectively.

G.3. Proof of Corollary G.1

Corollary G.1 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 G.1 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. (A.50) in Assumption 12 implies Assumption B and Eq. (A.49) in Assumption 12 is equivalent to the remaining conditions in Assumption C.

It remains to show how Assumptions 10 and 11 imply the remainder of conditions in Bai and Ng (2021, Assumptions A). For completeness, these conditions are collected in the following assumption.

Assumption 13. *The noise matrix H is such that,*

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

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

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

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

Assumption 13 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 13(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 10 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 11(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 G.1, it remains to establish that Assumption 13 holds, which is done in Appendix G.3.1 below.

G.3.1. Assumption 13 holds

First, Assumption 13(a) holds as follows,

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

where (a) follows from triangle inequality and (b) follows from Assumption 11(b). Next, from Assumption 11(a) and Assumption 11(c), we have

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

Then, Assumption 13(b) holds as follows,

$$\max_{i\in[N]}\max_{j\in[M]}\sum_{i'\in[N]}\left|\mathbb{E}[h_{i,j}h_{i',j}]\right| \le c.$$

Next, Assumption 13(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 11(c) and (b) follows from Assumption 11(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 13(d) holds as follows,

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

where (a) follows from linearity of expectation and Assumption 11(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 11(a).

H. 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

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

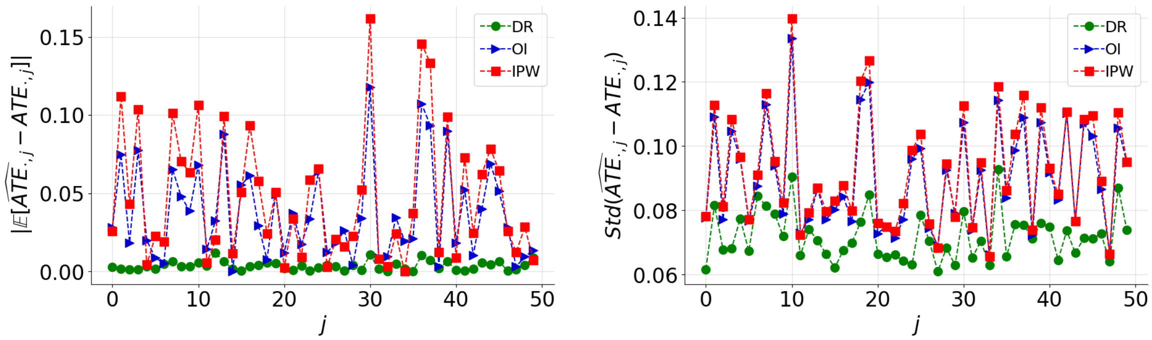
4. For $a \in \{0,1\}$, run $\texttt{SVD}$ on $U^{\text{shared}}V^{(a)\top}$, i.e.,

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

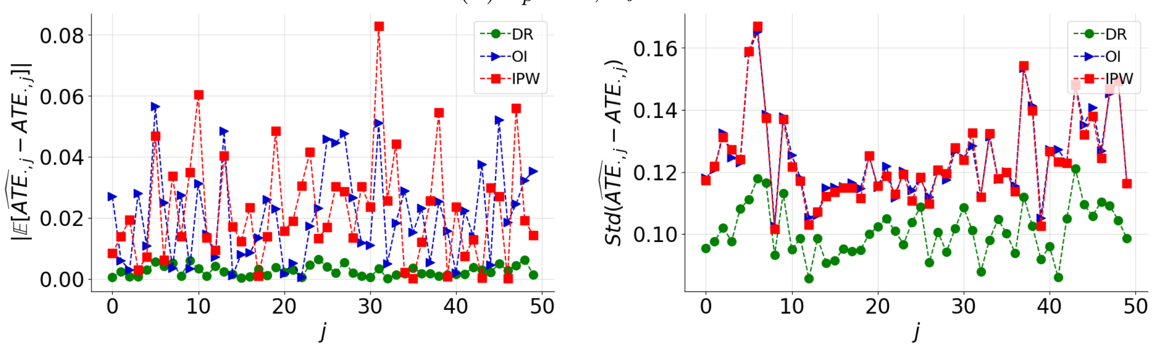
   Then, generate the mean potential outcome matrices $\Theta^{(0)}$ and $\Theta^{(1)}$:

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

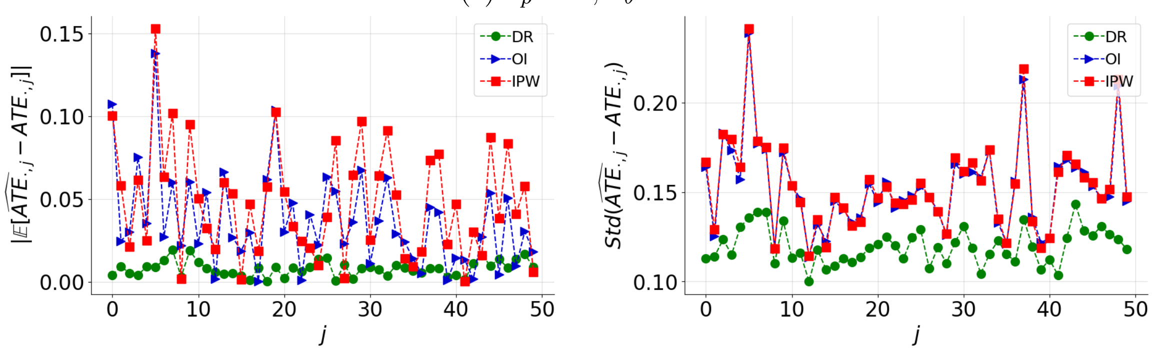
   where $\texttt{Sum}(\Sigma^{(a)})$ denotes the sum of all entries of $\Sigma^{(a)}$.



(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.

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 $\widehat{\text{ATE}}^{\text{OI}}_{\cdot,j} - \text{ATE}_{\cdot,j}$, $\widehat{\text{ATE}}^{\text{IPW}}_{\cdot,j} - \text{ATE}_{\cdot,j}$ and $\widehat{\text{ATE}}^{\text{DR}}_{\cdot,j} - \text{ATE}_{\cdot,j}$ across the Q simulation instances. Likewise, the estimate of the standard deviation of OI, IPW, and DR is the standard deviation of $\widehat{\text{ATE}}^{\text{OI}}_{\cdot,j} - \text{ATE}_{\cdot,j}$, $\widehat{\text{ATE}}^{\text{IPW}}_{\cdot,j} - \text{ATE}_{\cdot,j}$ and $\widehat{\text{ATE}}^{\text{DR}}_{\cdot,j} - \text{ATE}_{\cdot,j}$ across the Q simulation instances. The DR estimator consistently outperforms the OI and IPW estimators in reducing both absolute biases and standard deviations.

References

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.

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.

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. 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.

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*.

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

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.

Ma, W. and Chen, G. H. (2019). Missing not at random in matrix completion: The effectiveness of estimating missingness probabilities under a low nuclear norm assumption. *Advances in neural information processing systems*, 32.

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

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.

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

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