SUPPLEMENTARY MATERIAL

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

S1 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 $\widehat{\mathrm{ATE}}_{\cdot,j}^{\mathrm{DR}}$ from Eqs. (5) and (11), respectively. The error $\Delta\mathrm{ATE}_{\cdot,j}^{\mathrm{DR}} = \widehat{\mathrm{ATE}}_{\cdot,j}^{\mathrm{DR}} - \mathrm{ATE}_{\cdot,j}$ can be re-expressed as

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

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

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

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

$$\mathbb{T}_{i,j}^{(1,\mathrm{DR})} = \widehat{\theta}_{i,j}^{(1,\mathrm{DR})} - \theta_{i,j}^{(1)}$$

$$\overset{(a)}{=} \widehat{\theta}_{i,j}^{(1)} + \big(y_{i,j} - \widehat{\theta}_{i,j}^{(1)} \big) \frac{a_{i,j}}{\widehat{p}_{i,j}} - \theta_{i,j}^{(1)}$$

$$\overset{(b)}{=} \widehat{\theta}_{i,j}^{(1)} + \big(\theta_{i,j}^{(1)} + \varepsilon_{i,j}^{(1)} - \widehat{\theta}_{i,j}^{(1)} \big) \frac{p_{i,j} + \eta_{i,j}}{\widehat{p}_{i,j}} - \theta_{i,j}^{(1)} \tag{S.2}$$

$$= (\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)}) \Big(1 - \frac{p_{i,j} + \eta_{i,j}}{\widehat{p}_{i,j}} \Big) + \varepsilon_{i,j}^{(1)} \Big(\frac{p_{i,j} + \eta_{i,j}}{\widehat{p}_{i,j}} \Big)$$

$$= \frac{(\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)})(\widehat{p}_{i,j} - p_{i,j})}{\widehat{p}_{i,j}} - \frac{(\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)})\eta_{i,j}}{\widehat{p}_{i,j}} + \frac{\varepsilon_{i,j}^{(1)} p_{i,j}}{\widehat{p}_{i,j}} + \frac{\varepsilon_{i,j}^{(1)} \eta_{i,j}}{\widehat{p}_{i,j}}, \tag{S.3}$$

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

$$\mathbb{T}_{i,j}^{(0,\mathrm{DR})} = -\frac{(\widehat{\theta}_{i,j}^{(0)} - \theta_{i,j}^{(0)})(1 - \widehat{p}_{i,j} - (1 - p_{i,j}))}{1 - \widehat{p}_{i,j}} + \frac{(\widehat{\theta}_{i,j}^{(0)} - \theta_{i,j}^{(0)})(-\eta_{i,j})}{1 - \widehat{p}_{i,j}} - \frac{\varepsilon_{i,j}^{(0)}(1 - p_{i,j})}{1 - \widehat{p}_{i,j}}$$

$$- \frac{\varepsilon_{i,j}^{(0)}(-\eta_{i,j})}{1 - \widehat{p}_{i,j}}$$

$$= \frac{(\widehat{\theta}_{i,j}^{(0)} - \theta_{i,j}^{(0)})(\widehat{p}_{i,j} - p_{i,j})}{1 - \widehat{p}_{i,j}} - \frac{(\widehat{\theta}_{i,j}^{(0)} - \theta_{i,j}^{(0)})\eta_{i,j}}{1 - \widehat{p}_{i,j}} - \frac{\varepsilon_{i,j}^{(0)}(1 - p_{i,j})}{1 - \widehat{p}_{i,j}} + \frac{\varepsilon_{i,j}^{(0)} \eta_{i,j}}{1 - \widehat{p}_{i,j}}. \tag{S.4}$$

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

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

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

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

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

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

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

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

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

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

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

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

with probability at least $1-\delta$.

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

$$\Big|\sum_{i\in\mathcal{R}_s}\frac{\varepsilon_{i,j}^{(1)}p_{i,j}}{\widehat{p}_{i,j}}\Big| \le \overline{\sigma}\sqrt{c\ell_\delta}\sqrt{\sum_{i\in\mathcal{R}_s}\Big(\frac{p_{i,j}}{\widehat{p}_{i,j}}\Big)^2} \le \overline{\sigma}\sqrt{c\ell_\delta}\Big\|P_{\cdot,j}\oslash\widehat{P}_{\cdot,j}\Big\|_2, \tag{S.9}$$

with probability at least $1-\delta$.

To bound the fourth term in Eq. (S.6), note that $\varepsilon_{i,j}^{(1)}\eta_{i,j}$ is subExponential($\overline{\sigma}/\sqrt{\ell_1}$) because of Lemma S6 as well as zero-mean and independent across all $i \in [N]$ due to Assumption 2. By Assumption 4, $\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s} \perp\!\!\!\perp \{(\eta_{i,j},\varepsilon_{i,j}^{(1)})\}_{i\in\mathcal{R}_s}$. The subExponential concentration result in Corollary S2 yields that

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

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

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

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

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

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

S2 Proof of Theorem 2: Asymptotic Normality for DR

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

$$\mathbb{X}_{i,j}^{(1,\text{DR})} \triangleq \mathbb{T}_{i,j}^{(1,\text{DR})} - \varepsilon_{i,j}^{(1)} - \frac{\varepsilon_{i,j}^{(1)} \eta_{i,j}}{p_{i,j}} \tag{S.13}$$
$$\mathbb{X}_{i,j}^{(0,\text{DR})} \triangleq \mathbb{T}_{i,j}^{(0,\text{DR})} + \varepsilon_{i,j}^{(0)} - \frac{\varepsilon_{i,j}^{(0)} \eta_{i,j}}{1 - p_{i,j}},$$

and

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

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

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

We obtain the following convergence results.

Lemma S1 (Convergence of $\mathbb{X}_j^{\text{DR}}$)**.** *Fix any $j \in [M]$. Suppose Assumptions 1 to 4 and conditions (C1) to (C3) in Theorem 2 hold. Then,*

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

Lemma S2 (Convergence of $\mathbb{Z}_j^{\text{DR}}$)**.** *Fix any $j \in [M]$. Suppose Assumptions 1 and 2 hold and condition (C3) in Theorem 2 hold. Then,*

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

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

S2.1 Proof of Lemma S1

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

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

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

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

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

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

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

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

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

$$\frac{1}{\sqrt{N}}\Big|\sum_{i \in \mathcal{R}_s} \mathbb{X}_{i,j}^{(1,\mathrm{DR})}\Big| \leq \frac{1}{\sqrt{N}}\Big|\sum_{i \in \mathcal{R}_s} \frac{\big(\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)}\big)\big(\widehat{p}_{i,j} - p_{i,j}\big)}{\widehat{p}_{i,j}}\Big| + \frac{1}{\sqrt{N}}\Big|\sum_{i \in \mathcal{R}_s} \frac{\big(\widehat{\theta}_{i,j}^{(1)} - \theta_{i,j}^{(1)}\big)\eta_{i,j}}{\widehat{p}_{i,j}}\Big|$$

$$+\frac{1}{\sqrt{N}}\Big|\sum_{i\in\mathcal{R}_s}\frac{\varepsilon_{i,j}^{(1)}\big(\widehat{p}_{i,j}-p_{i,j}\big)}{\widehat{p}_{i,j}}\Big|+\frac{1}{\sqrt{N}}\Big|\sum_{i\in\mathcal{R}_s}\frac{\varepsilon_{i,j}^{(1)}\eta_{i,j}\big(\widehat{p}_{i,j}-p_{i,j}\big)}{\widehat{p}_{i,j}p_{i,j}}\Big|. \tag{S.16}$$

To control the first term in Eq. (S.16), we use the Cauchy-Schwarz inequality and Assumption 3 as in supplementary appendix S1 (see Eqs. (S.7), (S.11), and (S.12)).

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

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

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

To control the third term in Eq. (S.16), we condition on $\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s}$. Then, Assumption 4 (i.e., Eq. (15)) provides that $\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s}\perp\!\!\!\perp\{\varepsilon_{i,j}^{(1)}\}_{i\in\mathcal{R}_s}$. As a result, $\sum_{i\in\mathcal{R}_s}\varepsilon_{i,j}^{(1)}\big(\widehat{p}_{i,j}-p_{i,j}\big)/\widehat{p}_{i,j}$ is subGaussian$\Big(\bar{\sigma}\Big[\sum_{i\in\mathcal{R}_s}\big(\widehat{p}_{i,j}-p_{i,j}\big)^2/\big(\widehat{p}_{i,j}\big)^2\Big]^{1/2}\Big)$ because $\varepsilon_{i,j}^{(1)}$ is subGaussian$(\bar{\sigma})$, zero-mean, and independent across all $i\in[N]$ due to Assumption 2. Then, we have

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

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

To control the fourth term in Eq. (S.16), we condition on $\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s}$. Then, Assumption 4 (i.e., Eq. (15)) provides that $\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s} \perp\!\!\!\perp \{(\eta_{i,j}, \varepsilon^{(1)}_{i,j})\}_{i\in\mathcal{R}_s}$. 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 S6 as well as zero-mean and independent across all $i \in [N]$ due to Assumption 2. Then, we have

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

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

Putting together Eqs. (S.16) to (S.19) using Lemma S9, we have

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

Then, the claim in Eq. (S.15) 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. (S.15) for $a = 0$ follows similarly.

S2.2 Proof of Lemma S2

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

Lemma S3 (Lyapunov CLT, see Theorem 27.3 of Billingsley (2017))**.** *Consider a sequence $x_1, x_2, \cdots$ of mean-zero independent random variables such that the moments $\mathbb{E}[|x_i|^{2+\omega}]$ are finite for some $\omega > 0$. Moreover, assume that the Lyapunov's condition is satisfied, i.e.,*

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

as $N \to \infty$. *Then,*

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

as $N \to \infty$.

Fix any $j \in [M]$. We apply Lyapunov CLT in Lemma S3 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. (S.14). Note that this sequence is zero-mean from Assumption 2(a) and Assumption 2(b), and independent from Assumption 2(b). First, we show in supplementary appendix S2.2.1 that

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

for each $i \in [N]$. Next, we show in supplementary appendix S2.2.2 that Lyapunov's condition Eq. (S.20) holds for the sequence $\mathbb{Z}_{1,j}^{\mathrm{DR}}, \mathbb{Z}_{2,j}^{\mathrm{DR}}, \cdots$ with $\omega = 1$. Finally, applying Lemma S3 and using the definition of $\overline{\sigma}_j$ from Eq. (22) yields Lemma S2.

S2.2.1 Proof of Eq. (S.21)

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

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

We claim the following:

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

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

and

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

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

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

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

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

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

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 $\mathbb{E}[\eta_{i,j}^2] = \mathbb{V}\mathrm{ar}(\eta_{i,j}) = p_{i,j}(1-p_{i,j})$.

S2.2.2 Proof of Lyapunov's condition with $\omega = 1$

We have

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

where (a) follows by putting together Eqs. (S.21) and (22), (b) follows because $\overline{\sigma}_j \geq c_1 > 0$ as per condition (C3), (c) follows because the absolute third moments of subExponential random variables are bounded, after noting that $\mathbb{Z}_{i,j}^{\mathrm{DR}}$ is a subExponential random variable. Then, condition Eq. (S.20) holds for $\omega = 1$ as the right hand side of Eq. (S.26) goes to zero as $N \to \infty$.

S2.3 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

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

where (a) follows after defining

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

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

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

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

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

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

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

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

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

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

S2.3.0.1 Proof of bound Eq. (S.28). This proof follows a very similar road map to that used for establishing the inequality in Eq. (S.15). Recall the partitioning of the units $[N]$ into $\mathcal{R}_0$ and $\mathcal{R}_1$ from Assumption 4. Now, to enable the application of concentration bounds, we split the summation over $i \in [N]$ in the left hand side of Eq. (S.28) into two parts—one over $i \in \mathcal{R}_0$ and the other over $i \in \mathcal{R}_1$—such that the noise terms are independent of the estimates of $\Theta^{(0)}, \Theta^{(1)}, P$ in each of these parts as in Eqs. (14) and (15).

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

$$\frac{1}{N}\Big|\sum_{i\in\mathcal{R}_s} \mathbb{T}^{(1)}_{i,j}\Big| \le \frac{1}{N}\Big|\sum_{i\in\mathcal{R}_s} \frac{\Big(\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}\Big)^2 a_{i,j}}{\Big(\widehat{p}_{i,j}\Big)^2}\Big| + \frac{1}{N}\Big|\sum_{i\in\mathcal{R}_s} \frac{2\varepsilon^{(1)}_{i,j}p_{i,j}\Big(\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}\Big)}{\Big(\widehat{p}_{i,j}\Big)^2}\Big|$$

$$+\frac{1}{N}\Big|\sum_{i\in\mathcal{R}_s}\frac{2\varepsilon_{i,j}^{(1)}\eta_{i,j}\Big(\widehat{\theta}_{i,j}^{(1)}-\theta_{i,j}^{(1)}\Big)}{\Big(\widehat{p}_{i,j}\Big)^2}\Big|+\frac{1}{N}\Big|\sum_{i\in\mathcal{R}_s}\frac{\Big(\varepsilon_{i,j}^{(1)}\Big)^2\eta_{i,j}}{\Big(\widehat{p}_{i,j}\Big)^2}\Big|+\frac{1}{N}\Big|\sum_{i\in\mathcal{R}_s}\frac{\Big(\varepsilon_{i,j}^{(1)}\Big)^2 p_{i,j}}{\Big(\widehat{p}_{i,j}\Big)^2}-\frac{(\sigma_{i,j}^{(1)})^2}{p_{i,j}}\Big|. \tag{S.29}$$

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

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

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

To control second term in Eq. (S.29), we condition on $\{\big(\widehat{p}_{i,j},\widehat{\theta}_{i,j}^{(1)}\big)\}_{i\in\mathcal{R}_s}$. Then, Eq. (24) provides that $\{\big(\widehat{p}_{i,j},\widehat{\theta}_{i,j}^{(1)}\big)\}_{i\in\mathcal{R}_s}\perp\!\!\!\perp\{\varepsilon_{i,j}^{(1)}\}_{i\in\mathcal{R}_s}$. As a result, $\sum_{i\in\mathcal{R}_s}\varepsilon_{i,j}^{(1)}p_{i,j}\big(\widehat{\theta}_{i,j}^{(1)}-\theta_{i,j}^{(1)}\big)/\big(\widehat{p}_{i,j}\big)^2$ is subGaussian$\Big(\overline{\sigma}\Big[\sum_{i\in\mathcal{R}_s}\big(p_{i,j}\big)^2\big(\widehat{\theta}_{i,j}^{(1)}-\theta_{i,j}^{(1)}\big)^2/\big(\widehat{p}_{i,j}\big)^4\Big]^{1/2}\Big)$ because $\varepsilon_{i,j}^{(1)}$ is subGaussian$(\overline{\sigma})$, zero-mean and independent across all $i\in[N]$ due to Assumption 2. Then, we have

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

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

To control third term in Eq. (S.29), we condition on $\{\big(\widehat{p}_{i,j},\widehat{\theta}_{i,j}^{(1)}\big)\}_{i\in\mathcal{R}_s}$. Then, Eq. (24) provides that $\{\big(\widehat{p}_{i,j},\widehat{\theta}_{i,j}^{(1)}\big)\}_{i\in\mathcal{R}_s}\perp\!\!\!\perp\{(\eta_{i,j},\varepsilon_{i,j}^{(1)})\}_{i\in\mathcal{R}_s}$. As a result, $\sum_{i\in\mathcal{R}_s}\varepsilon_{i,j}^{(1)}\eta_{i,j}\big(\widehat{\theta}_{i,j}^{(1)}-\theta_{i,j}^{(1)}\big)/\big(\widehat{p}_{i,j}\big)^2$ is subExponential$\Big(\overline{\sigma}\Big[\sum_{i\in\mathcal{R}_s}\big(\widehat{\theta}_{i,j}^{(1)}-\theta_{i,j}^{(1)}\big)^2/\big(\widehat{p}_{i,j}\big)^4\Big]^{1/2}/\sqrt{\ell_1}\Big)$ because $\varepsilon_{i,j}^{(1)}\eta_{i,j}$ is subExponential$(\overline{\sigma}/\sqrt{\ell_1})$ due to Lemma S6 as well as zero-mean and independent across all

$i \in [N]$ due to Assumption 2. Then, we have

$$\frac{1}{N}\mathbb{E}\left[\left|\sum_{i\in\mathcal{R}_s}\frac{2\varepsilon_{i,j}^{(1)}\eta_{i,j}\left(\widehat{\theta}_{i,j}^{(1)}-\theta_{i,j}^{(1)}\right)}{\left(\widehat{p}_{i,j}\right)^2}\right|\left|\{(\widehat{p}_{i,j},\widehat{\theta}_{i,j}^{(1)})\}_{i\in\mathcal{R}_s}\right.\right]$$
$$\overset{(a)}{\le} \frac{c\overline{\sigma}}{N}\sqrt{\sum_{i\in\mathcal{R}_s}\left(\frac{\widehat{\theta}_{i,j}^{(1)}-\theta_{i,j}^{(1)}}{\left(\widehat{p}_{i,j}\right)^2}\right)^2}$$
$$\overset{(b)}{\le} \frac{c\overline{\sigma}}{\bar{\lambda}^2 N}\left\|\widehat{\Theta}_{\cdot,j}^{(1)}-\Theta_{\cdot,j}^{(1)}\right\|_2 \overset{(c)}{=} \frac{c\overline{\sigma}}{\bar{\lambda}^2}\frac{\mathcal{E}\left(\widehat{\Theta}^{(1)}\right)}{\sqrt{N}} \le \frac{c\overline{\sigma}}{\bar{\lambda}^2}\frac{\mathcal{E}\left(\widehat{\Theta}\right)}{\sqrt{N}} \overset{(d)}{=} o_p(1), \tag{S.32}$$

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

To control fourth term in Eq. (S.29), we condition on $\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s}$. Then, Eq. (24) provides that $\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s} \perp\!\!\!\perp \{(\eta_{i,j},\varepsilon_{i,j}^{(1)})\}_{i\in\mathcal{R}_s}$. As a result, $\sum_{i\in\mathcal{R}_s}\left(\varepsilon_{i,j}^{(1)}\right)^2\eta_{i,j}/\left(\widehat{p}_{i,j}\right)^2$ is subWeibull$_{2/3}$ $\left(\overline{\sigma}^2\left[\sum_{i\in\mathcal{R}_s}1/\left(\widehat{p}_{i,j}\right)^4\right]^{1/2}/\sqrt{\ell_1}\right)$ because $(\varepsilon_{i,j}^{(1)})^2\eta_{i,j}$ is subWeibull$_{2/3}(\overline{\sigma}^2/\sqrt{\ell_1})$ due to Lemma S7 as well as zero-mean and independent across all $i \in [N]$ due to Assumption 2. Then, we have

$$\frac{1}{N}\mathbb{E}\left[\left|\sum_{i\in\mathcal{R}_s}\frac{\left(\varepsilon_{i,j}^{(1)}\right)^2\eta_{i,j}}{\left(\widehat{p}_{i,j}\right)^2}\right|\left|\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s}\right.\right] \overset{(a)}{\le} \frac{c\overline{\sigma}^2}{N}\sqrt{\sum_{i\in\mathcal{R}_s}\frac{1}{\left(\widehat{p}_{i,j}\right)^4}} \overset{(b)}{\le} \frac{c\overline{\sigma}^2}{\bar{\lambda}^2\sqrt{N}} = o_p(1), \tag{S.33}$$

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. (S.29), we have

$$\begin{aligned}\left|\sum_{i\in\mathcal{R}_s}\left(\frac{\left(\varepsilon_{i,j}^{(1)}\right)^2 p_{i,j}}{\left(\widehat{p}_{i,j}\right)^2}-\frac{(\sigma_{i,j}^{(1)})^2}{p_{i,j}}\right)\right| &= \left|\sum_{i\in\mathcal{R}_s}\left(\frac{\left(\varepsilon_{i,j}^{(1)}\right)^2 p_{i,j}}{\left(\widehat{p}_{i,j}\right)^2}-\frac{\left(\sigma_{i,j}^{(1)}\right)^2 p_{i,j}}{\left(\widehat{p}_{i,j}\right)^2}+\frac{\left(\sigma_{i,j}^{(1)}\right)^2 p_{i,j}}{\left(\widehat{p}_{i,j}\right)^2}-\frac{(\sigma_{i,j}^{(1)})^2}{p_{i,j}}\right)\right| \\ &\overset{(a)}{\le} \left|\sum_{i\in\mathcal{R}_s}\left(\frac{\left[\left(\varepsilon_{i,j}^{(1)}\right)^2-\left(\sigma_{i,j}^{(1)}\right)^2\right]p_{i,j}}{\left(\widehat{p}_{i,j}\right)^2}\right)\right|+\left|\sum_{i\in\mathcal{R}_s}\left(\frac{\left(\sigma_{i,j}^{(1)}\right)^2 p_{i,j}}{\left(\widehat{p}_{i,j}\right)^2}-\frac{(\sigma_{i,j}^{(1)})^2}{p_{i,j}}\right)\right|,\end{aligned} \tag{S.34}$$

where (a) follows from the triangle inequality. To control the first term in Eq. (S.34), we condition on $\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s}$. Then, Eq. (24) provides that $\{\widehat{p}_{i,j}\}_{i\in\mathcal{R}_s} \perp\!\!\!\perp \{\varepsilon_{i,j}^{(1)}\}_{i\in\mathcal{R}_s}$. Further,

$\mathbb{E}[(\varepsilon_{i,j}^{(1)})^2 - (\sigma_{i,j}^{(1)})^2] = 0$ due to (C3) and Assumption 2. As a result, $\sum_{i\in\mathcal{R}_s}\Big[\big(\varepsilon_{i,j}^{(1)}\big)^2 - \big(\sigma_{i,j}^{(1)}\big)^2\Big]p_{i,j}/\big(\widehat{p}_{i,j}\big)^2$ is 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 S6. Then, we have

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

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. (S.34), applying the Cauchy-Schwarz inequality yields that

$$\begin{aligned}
\frac{1}{N}\Bigg|\sum_{i\in\mathcal{R}_s}\Bigg(\frac{\big(\sigma_{i,j}^{(1)}\big)^2 p_{i,j}}{\big(\widehat{p}_{i,j}\big)^2}-\frac{(\sigma_{i,j}^{(1)})^2}{p_{i,j}}\Bigg)\Bigg| &= \frac{1}{N}\Bigg|\sum_{i\in\mathcal{R}_s}\frac{\big(\sigma_{i,j}^{(1)}\big)^2\Big(\big(p_{i,j}\big)^2-\big(\widehat{p}_{i,j}\big)^2\Big)}{\big(\widehat{p}_{i,j}\big)^2 p_{i,j}}\Bigg| \\
&\overset{(a)}{\leq} \frac{2}{N}\sum_{i\in\mathcal{R}_s}\frac{\big(\sigma_{i,j}^{(1)}\big)^2\big|p_{i,j}-\widehat{p}_{i,j}\big|}{\big(\widehat{p}_{i,j}\big)^2 p_{i,j}} \\
&\overset{(b)}{\leq} \frac{2\overline{\sigma}^2}{\lambda\bar{\lambda}^2 N}\sum_{i\in\mathcal{R}_s}\big|p_{i,j}-\widehat{p}_{i,j}\big| \\
&\overset{(c)}{\leq} \frac{2\overline{\sigma}^2}{\lambda\bar{\lambda}^2\sqrt{N}}\big\|P_{\cdot,j}-\widehat{P}_{\cdot,j}\big\|_2 \overset{(d)}{=} \frac{2\overline{\sigma}^2}{\lambda\bar{\lambda}^2}\mathcal{E}\big(\widehat{P}\big) \overset{(e)}{=} o_p(1),
\end{aligned} \tag{S.36}$$

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

Putting together Eqs. (S.29) to (S.36) using Lemma S9,

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

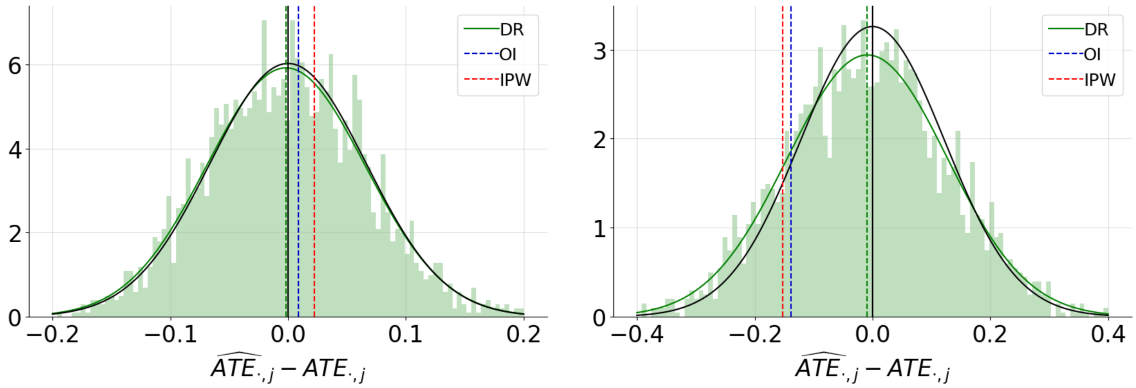
Then, the claim in Eq. (S.28) follows for $a = 1$ by using $|\sum_{i\in[N]}\mathbb{T}_{i,j}^{(1)}| \leq |\sum_{i\in\mathcal{R}_0}\mathbb{T}_{i,j}^{(1)}| + |\sum_{i\in\mathcal{R}_1}\mathbb{T}_{i,j}^{(1)}|$. The proof of Eq. (S.28) for $a = 0$ follows similarly.

S3 Simulations

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

Data Generating Process (DGP). We now briefly describe the DGP for our simulations; supplementary appendix S3.1 provides details. All simulations set $N = M$. To generate, P, $\Theta^{(0)}$, and $\Theta^{(1)}$, we use the latent factor model given in Eq. (S.47). To introduce unobserved confounding, we set the unit-specific latent factors to be the same across P, $\Theta^{(0)}$, and $\Theta^{(1)}$, i.e., $U = U^{(0)} = U^{(1)}$. The entries of U and the measurement-specific latent factors, $V, V^{(0)}, V^{(1)}$ are each sampled independently from a uniform distribution, with hyperparameter r_p equal to the dimension of U and V, and hyperparameter r_θ equal to the dimension of $U^{(a)}$ and $V^{(a)}$ for $a = 0, 1$. Further, the entries of the noise matrices $E^{(0)}$ and $E^{(1)}$ are sampled independently from a normal distribution, and the entries of W are sampled independently as in Eq. (4). Then, $y_{i,j}^{(a)}$, $a_{i,j}$, and $y_{i,j}$ are determined from Eqs. (1) to (3), respectively. The simulation generates P, $\Theta^{(0)}$, and $\Theta^{(1)}$ once. Given the fixed values of P, $\Theta^{(0)}$, and $\Theta^{(1)}$, the simulation generates 2500 realizations of (Y, A)—that is, only the noise matrices $E^{(0)}, E^{(1)}, W$ are resampled for each of the 2500 realizations. For each simulation realization, we apply the `Cross-Fitted-SVD` algorithm with hyper-parameters as in Proposition 4 and $\bar{\lambda} = \lambda = 0.05$ to obtain $\widehat{P}$, $\widehat{\Theta}^{(0)}$, and $\widehat{\Theta}^{(1)}$, and compute $\text{ATE}_{\cdot,j}$ from Eq. (5), and $\widehat{\text{ATE}}_{\cdot,j}^{\text{OI}}$, $\widehat{\text{ATE}}_{\cdot,j}^{\text{IPW}}$ and $\widehat{\text{ATE}}_{\cdot,j}^{\text{DR}}$ from Eqs. (9) to (11).

Results. Figure 5 reports simulation results for $N = 1000$, with $r_p = 3$, $r_\theta = 3$ in Panel (a), and $r_p = 5$, $r_\theta = 3$ in Panel (b). Figure 2 in Section 3 reports simulation results for $r_p = 3$, $r_\theta = 5$. In each case, the figure shows a histogram of the distribution of $\widehat{\text{ATE}}_{\cdot,j}^{\text{DR}} - \text{ATE}_{\cdot,j}$ across 2500 simulation instances for a fixed j, along with the best fitting Gaussian distribution (green curve). The histogram counts are normalized so that the area under the histogram integrates to one. Figure 5 plots the Gaussian distribution in



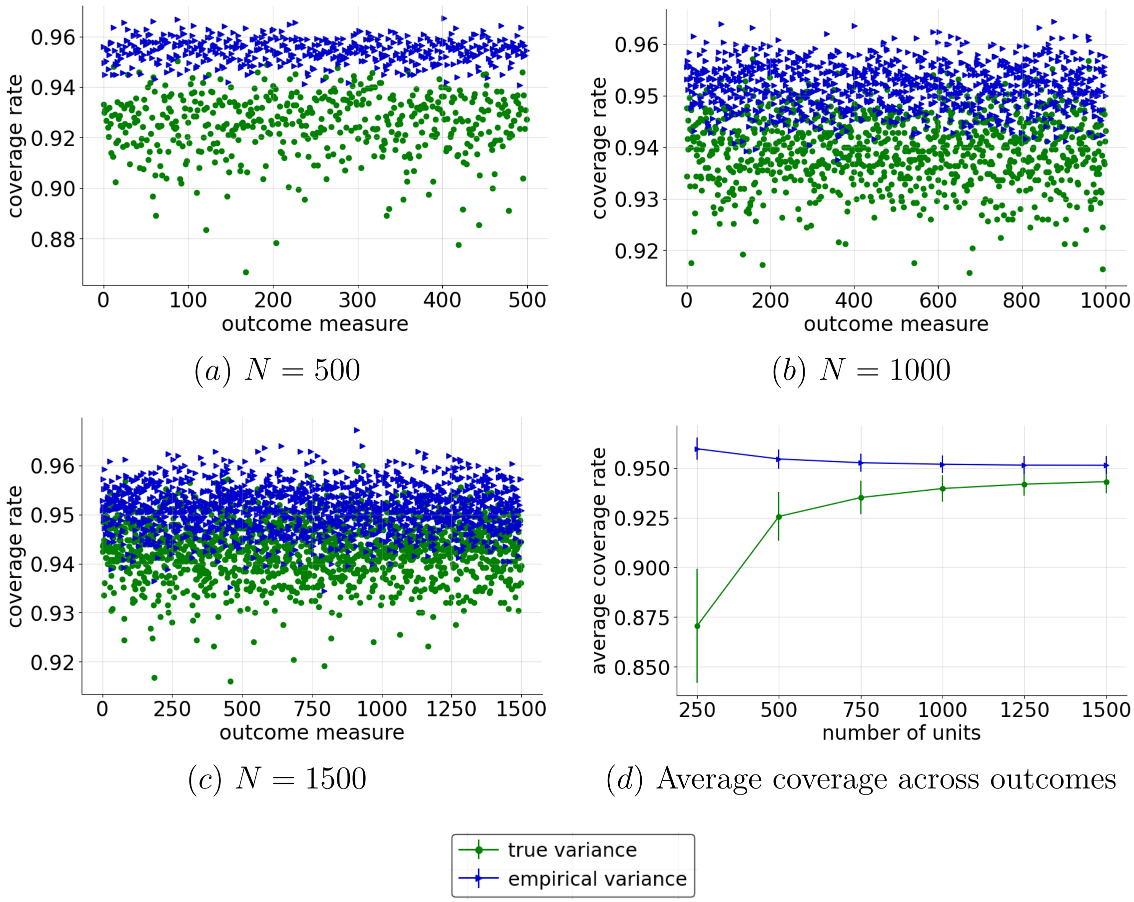
(a) $r_p = 3$, $r_\theta = 3$ (b) $r_p = 5$, $r_\theta = 3$

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

the 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 supplementary appendix S3.1, we compare the biases and the standard deviations of OI, IPW, and DR across many j.

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

In Figure 7, we compare the absolute biases and the standard deviations of OI, IPW,



(a) $N = 500$

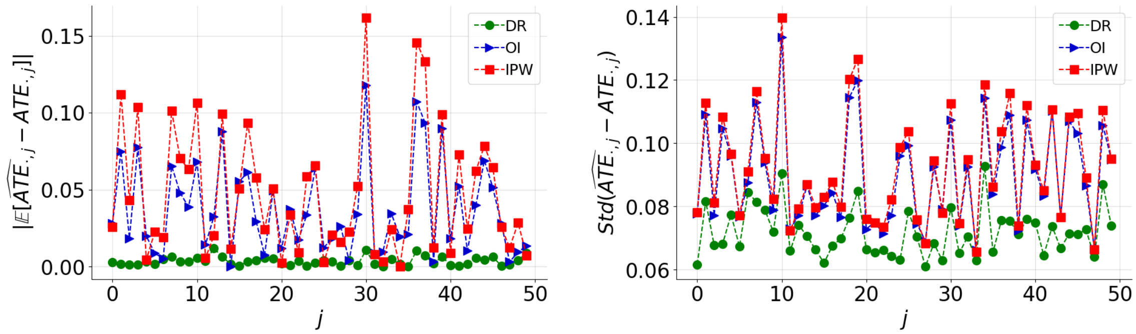
(b) $N = 1000$

(c) $N = 1500$

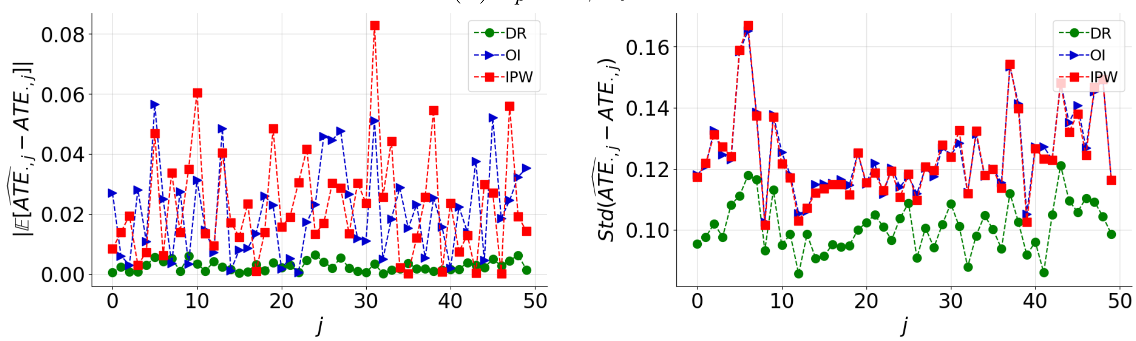
(d) Average coverage across outcomes

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

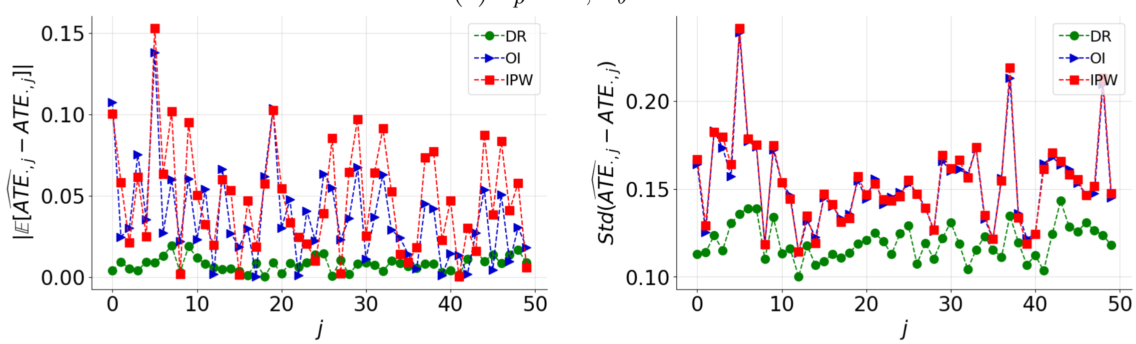
and DR across the first 50 values of j for $N = 1000$, with $r_p = 3$, $r_\theta = 3$ in Panel (a), $r_p = 3$, $r_\theta = 5$ in Panel (b), and $r_p = 5$, $r_\theta = 3$ in Panel (c). For each j, the estimate of the biases of OI, IPW, and DR is the average of $\widehat{\text{ATE}}_{\cdot,j}^{\text{OI}}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



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



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



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

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 and r_θ.

S3.1 Details for the data generating process

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 $\sigma_{i,j}^{(a)}$ 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 $j \in [r]$, $u_{i,j}^{\text{shared}}$

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 $j \in [r]$, $v_{i,j}, v_{i,j}^{(0)}, v_{i,j}^{(1)}$ 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 $\Sigma^{(a)}$.

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

S4 Supporting Concentration and Convergence Results

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

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

Lemma S4 (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 $t \geq 0$,*

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

The following corollary expresses the bound in Lemma S4 in a convenient form.

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

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

Proof. The proof follows from Lemma S4 by choosing $\delta \triangleq 2\exp(-ct^2/\sigma^2\|b\|_2^2)$. □

Lemma S5 (subExponential concentration: Theorem 2.8.2 of Vershynin (2018))**.** *Let $x \in \mathbb{R}^n$ be a random vector whose entries are independent, zero-mean, subExponential(σ)*

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

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

The following corollary expresses the bound in Lemma S5 in a convenient form.

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

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

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

Proof. Choosing $t = t_0\sigma\|b\|_2$ in Lemma S5, 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 S6 (Product of subGaussians is subExponential: Lemma. 2.7.7 of Vershynin (2018))**.** *Let* x_1 *and* x_2 *be subGaussian*(σ_1) *and subGaussian*(σ_2) *random variables, respectively. Then,* x_1x_2 *is subExponential*($\sigma_1\sigma_2$) *random variable.*

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

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

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

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

Lemma S7 (Product of subWeibulls is subWeibull: Proposition 2 of Zhang and Wei (2022))**.** *For $i \in [d]$, let x_i be a subWeibull$_{\rho_i}(\sigma_i)$ random variable. Then, $\Pi_{i\in[d]} x_i$ is subWeibull$_\rho(\sigma)$ random variable where*

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

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

Lemma S8. *Let X_n and $\overline{X}_n$ be sequences of random variables. Let δ_n be a deterministic sequence such that $0 \le \delta_n \le 1$ and $\delta_n \to 0$. Suppose $X_n = o_p(1)$ and $\mathbb{P}(|\overline{X}_n| \le |X_n|) \ge 1-\delta_n$. 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 $n \ge \overline{n}$. Fix any $\epsilon > 0$. As δ_n converges to zero, there exists a finite n_0 such that $\delta_n < \epsilon/2$, for all $n \ge n_0$. As X_n is converges to zero in probability, there exists finite n_1, such that $\mathbb{P}(|X_n| > \delta) < \epsilon/2$ for all $n \ge n_1$. 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

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

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

Lemma S9. *Let X_n and $\overline{X}_n$ be sequences of random variables. Suppose $\mathbb{E}\Big[|X_n|\Big|\overline{X}_n\Big] = o_p(1)$. Then, $X_n = o_p(1)$.*

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

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

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

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

□

Lemma S10. *Let X_n and $\overline{X}_n$ be sequences of random variables. Suppose $X_n = O_p(1)$ and $\mathbb{P}\Big(|\overline{X}_n| \geq |X_n| + f(\epsilon)\Big) < \epsilon$ for some positive function f and every $\epsilon \in (0, 1)$. Then, $\overline{X}_n = O_p(1)$.*

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

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

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

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

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

S5 Proofs of Corollaries 1 and 2

S5.1 Proof of Corollary 1: Gains of DR over OI and IPW

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

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

where

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

for $m(c)$ and ℓ_c as defined in Section 1 of the main article. Then, if $\mathcal{E}\left(\widehat{P}\right) = O_p\left(N^{-\alpha}\right)$, Lemma S10 implies

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

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

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

where

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

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

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

Lemma S10 implies $\left|\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,j} - \mathrm{ATE}_{\cdot,j}\right| = O_p(N^{-(\alpha+\beta)})$. Next, suppose $\alpha+\beta > 0.5$. With probability at least $1-\delta$,

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

Lemma S10 implies $\left|\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,j} - \mathrm{ATE}_{\cdot,j}\right| = O_p(N^{-1/2})$.

S5.2 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 S8.

S6 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 $\widehat{\mathrm{ATE}}^{\mathrm{OI}}_{\cdot,j}$ from Eqs. (5) and (9), respectively. The error $\Delta\mathrm{ATE}^{\mathrm{OI}}_{\cdot,j} = \widehat{\mathrm{ATE}}^{\mathrm{OI}}_{\cdot,j} - \mathrm{ATE}_{\cdot,j}$ can be re-expressed as

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

Using the triangle inequality, we have

$$\left|\Delta\mathrm{ATE}^{\mathrm{OI}}_{\cdot,j}\right| \leq \frac{1}{N}\left|\sum_{i\in[N]}\Big(\widehat{\theta}^{(1)}_{i,j} - \theta^{(1)}_{i,j}\Big)\right| + \frac{1}{N}\left|\sum_{i\in[N]}\Big(\widehat{\theta}^{(0)}_{i,j} - \theta^{(0)}_{i,j}\Big)\right|. \tag{S.37}$$

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

$$\frac{1}{N}\left|\sum_{i\in[N]}\Big(\widehat{\theta}^{(a)}_{i,j} - \theta^{(a)}_{i,j}\Big)\right| \leq \mathcal{E}\Big(\widehat{\Theta}^{(a)}\Big). \tag{S.38}$$

The proof is complete by putting together Eqs. (S.37) and (S.38).

Proof of Eq. (S.38) Fix any $a \in \{0, 1\}$. Using the Cauchy-Schwarz inequality, we have

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

S7 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 $\widehat{\mathrm{ATE}}^{\mathrm{IPW}}_{\cdot,j}$ from Eqs. (5) and (10), respectively. The error $\Delta\mathrm{ATE}^{\mathrm{IPW}}_{\cdot,j} = \widehat{\mathrm{ATE}}^{\mathrm{IPW}}_{\cdot,j} - \mathrm{ATE}_{\cdot,j}$ can be re-expressed as

$$\begin{aligned}\Delta\mathrm{ATE}^{\mathrm{IPW}}_{\cdot,j} &= \frac{1}{N}\sum_{i\in[N]}\Big(\frac{y_{i,j}a_{i,j}}{\widehat{p}_{i,j}} - \frac{y_{i,j}(1-a_{i,j})}{1-\widehat{p}_{i,j}}\Big) - \frac{1}{N}\sum_{i\in[N]}\Big(\theta^{(1)}_{i,j} - \theta^{(0)}_{i,j}\Big)\\ &= \frac{1}{N}\sum_{i\in[N]}\bigg(\Big(\frac{y_{i,j}a_{i,j}}{\widehat{p}_{i,j}} - \theta^{(1)}_{i,j}\Big) - \Big(\frac{y_{i,j}(1-a_{i,j})}{1-\widehat{p}_{i,j}} - \theta^{(0)}_{i,j}\Big)\bigg)\\ &\overset{(a)}{=} \frac{1}{N}\sum_{i\in[N]}\Big(\mathbb{T}^{(1,\mathrm{IPW})}_{i,j} + \mathbb{T}^{(0,\mathrm{IPW})}_{i,j}\Big),\end{aligned} \tag{S.39}$$

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

$$\begin{aligned}\mathbb{T}^{(1,\mathrm{IPW})}_{i,j} &= \frac{y_{i,j}a_{i,j}}{\widehat{p}_{i,j}} - \theta^{(1)}_{i,j}\\ &\overset{(a)}{=} \frac{\Big(\theta^{(1)}_{i,j} + \varepsilon^{(1)}_{i,j}\Big)\Big(p_{i,j} + \eta_{i,j}\Big)}{\widehat{p}_{i,j}} - \theta^{(1)}_{i,j}\\ &= \theta^{(1)}_{i,j}\Big(\frac{p_{i,j} + \eta_{i,j}}{\widehat{p}_{i,j}} - 1\Big) + \varepsilon^{(1)}_{i,j}\Big(\frac{p_{i,j} + \eta_{i,j}}{\widehat{p}_{i,j}}\Big)\\ &= \frac{\theta^{(1)}_{i,j}\Big(p_{i,j} - \widehat{p}_{i,j}\Big)}{\widehat{p}_{i,j}} + \frac{\theta^{(1)}_{i,j}\eta_{i,j}}{\widehat{p}_{i,j}} + \frac{\varepsilon^{(1)}_{i,j}p_{i,j}}{\widehat{p}_{i,j}} + \frac{\varepsilon^{(1)}_{i,j}\eta_{i,j}}{\widehat{p}_{i,j}},\end{aligned} \tag{S.40}$$

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

$$\mathbb{T}^{(0,\mathrm{IPW})}_{i,j} = \theta^{(0)}_{i,j} - \frac{y_{i,j}(1-a_{i,j})}{1-\widehat{p}_{i,j}}$$

$$= -\frac{\theta^{(0)}_{i,j}\Big(1-p_{i,j}-\Big(1-\widehat{p}_{i,j}\Big)\Big)}{1-\widehat{p}_{i,j}} - \frac{\theta^{(0)}_{i,j}(-\eta_{i,j})}{1-\widehat{p}_{i,j}} - \frac{\varepsilon^{(0)}_{i,j}\Big(1-p_{i,j}\Big)}{1-\widehat{p}_{i,j}} - \frac{\varepsilon^{(0)}_{i,j}(-\eta_{i,j})}{1-\widehat{p}_{i,j}}$$
$$= \frac{\theta^{(0)}_{i,j}\Big(p_{i,j}-\widehat{p}_{i,j}\Big)}{1-\widehat{p}_{i,j}} + \frac{\theta^{(0)}_{i,j}\eta_{i,j}}{1-\widehat{p}_{i,j}} - \frac{\varepsilon^{(0)}_{i,j}\Big(1-p_{i,j}\Big)}{1-\widehat{p}_{i,j}} + \frac{\varepsilon^{(0)}_{i,j}\eta_{i,j}}{1-\widehat{p}_{i,j}}.$$

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

$$\frac{1}{N}\Big|\sum_{i\in[N]}\mathbb{T}^{(\mathrm{a,IPW})}_{i,j}\Big| \leq \frac{2}{\bar{\lambda}}\|\Theta^{(a)}\|_{\max}\,\mathcal{E}\Big(\widehat{P}\Big) + \frac{2\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{\ell_1 N}}\|\Theta^{(a)}\|_{\max} + \frac{2\overline{\sigma}\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{N}} + \frac{2\overline{\sigma}m(c\ell_\delta)}{\bar{\lambda}\sqrt{\ell_1 N}}. \tag{S.41}$$

where recall that $m(c\ell_\delta) = \max\Big(c\ell_\delta, \sqrt{c\ell_\delta}\Big)$. We provide a proof of this claim at the end of this section. Applying triangle inequality in Eq. (S.39) and using Eq. (S.41) with a union bound, we obtain that

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

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

Proof of Eq. (S.41). This proof follows a very similar road map to that used for establishing the inequality in Eq. (S.5). Recall the partitioning of the units $[N]$ into $\mathcal{R}_0$ and $\mathcal{R}_1$ from Assumption 4. Now, to enable the application of concentration bounds, we split the summation over $i \in [N]$ in the left hand side of Eq. (S.41) into two parts—one over $i \in \mathcal{R}_0$ and the other over $i \in \mathcal{R}_1$—such that the noise terms are independent of the estimates of $\Theta^{(0)}, \Theta^{(1)}, P$ in each of these parts as in Eqs. (14) and (15).

Fix $a = 1$ and note that $|\sum_{i\in[N]}\mathbb{T}^{(1,\mathrm{IPW})}_{i,j}| \leq |\sum_{i\in\mathcal{R}_0}\mathbb{T}^{(1,\mathrm{IPW})}_{i,j}| + |\sum_{i\in\mathcal{R}_1}\mathbb{T}^{(1,\mathrm{IPW})}_{i,j}|$. Fix any $s \in \{0,1\}$. Then, Eq. (S.40) and triangle inequality imply that

$$\Big|\sum_{i\in\mathcal{R}_s}\mathbb{T}^{(1,\mathrm{IPW})}_{i,j}\Big| \leq \Big|\sum_{i\in\mathcal{R}_s}\frac{\theta^{(1)}_{i,j}\Big(p_{i,j}-\widehat{p}_{i,j}\Big)}{\widehat{p}_{i,j}}\Big| + \Big|\sum_{i\in\mathcal{R}_s}\frac{\theta^{(1)}_{i,j}\eta_{i,j}}{\widehat{p}_{i,j}}\Big| + \Big|\sum_{i\in\mathcal{R}_s}\frac{\varepsilon^{(1)}_{i,j}p_{i,j}}{\widehat{p}_{i,j}}\Big| + \Big|\sum_{i\in\mathcal{R}_s}\frac{\varepsilon^{(1)}_{i,j}\eta_{i,j}}{\widehat{p}_{i,j}}\Big|. \tag{S.42}$$

Next, note that the decomposition in Eq. (S.42) is identical to the one in Eq. (S.6), except for the fact when compared to Eq. (S.6), the first two terms in Eq. (S.42) 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. (S.11), we obtain the following bound, with probability at least $1-3\delta$,

$$\frac{1}{N}\Big|\sum_{i\in\mathcal{R}_s}\mathbb{T}_{i,j}^{(1,\mathrm{IPW})}\Big| \leq \frac{1}{\bar{\lambda}N}\|\Theta^{(1)}\|_{1,2}\|\widehat{P}-P\|_{1,2}+\frac{\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{\ell_1}N}\|\Theta^{(1)}\|_{1,2}+\frac{\overline{\sigma}\sqrt{c\ell_\delta}}{\bar{\lambda}N}\|P\|_{1,2}+\frac{\overline{\sigma}m(c\ell_\delta)}{\bar{\lambda}\sqrt{\ell_1}N}\|\mathbf{1}\|_{1,2}$$
$$\overset{(a)}{\leq} \frac{1}{\bar{\lambda}\sqrt{N}}\|\Theta^{(1)}\|_{\max}\|\widehat{P}-P\|_{1,2}+\frac{\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{\ell_1N}}\|\Theta^{(1)}\|_{\max}+\frac{\overline{\sigma}\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{N}}+\frac{\overline{\sigma}m(c\ell_\delta)}{\bar{\lambda}\sqrt{\ell_1N}}$$
$$\overset{(b)}{\leq} \frac{1}{\bar{\lambda}}\|\Theta^{(1)}\|_{\max}\,\mathcal{E}\big(\widehat{P}\big) + \frac{\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{\ell_1N}}\|\Theta^{(1)}\|_{\max} + \frac{\overline{\sigma}\sqrt{c\ell_\delta}}{\bar{\lambda}\sqrt{N}} + \frac{\overline{\sigma}m(c\ell_\delta)}{\bar{\lambda}\sqrt{\ell_1N}}, \tag{S.43}$$

where (a) follows because $\|\Theta^{(1)}\|_{1,2}\leq\sqrt{N}\|\Theta^{(1)}\|_{\max}$, $\|P\|_{1,2}\leq\sqrt{N}$ and $\|\mathbf{1}\|_{1,2}=\sqrt{N}$, and (b) follows from Eq. (16). Then, the claim in Eq. (S.41) follows for $a=1$ by using Eq. (S.43) and applying a union bound over $s\in\{0,1\}$. The proof of Eq. (S.41) for $a=0$ follows similarly.

S8 Proof of Proposition 3 and `TW` algorithm of Bai and Ng (2021)

In Section S8.1, 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 detail the `TW` algorithm in Section S8.2.

S8.1 Proof of Proposition 3: Guarantees for `Cross-Fitted-MC`

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

$$\widehat{P}_{\mathcal{I}}, \widehat{\Theta}_{\mathcal{I}}^{(a)} \perp\!\!\!\perp W_{\mathcal{I}} \tag{S.44}$$

and

$$\widehat{P}_{\mathcal{I}} \perp\!\!\!\perp W_{\mathcal{I}}, E_{\mathcal{I}}^{(a)}, \tag{S.45}$$

for every $\mathcal{I} \in \mathcal{P}$ and $a \in \{0, 1\}$, where $\mathcal{P}$ is the block partition of $[N] \times [M]$ into four blocks from Assumption 5. Then, Eqs. (14) and (15) in Assumption 4 follow from Eqs. (S.44) and (S.45), respectively.

Consider $\widehat{\Theta}^{(0)}, \widehat{\Theta}^{(1)}$, and $\widehat{P}$ as in Eqs. (30) to (32). Fix any $a \in \{0, 1\}$. From Eq. (29), note that $\widehat{P}_{\mathcal{I}}$ depends only on $A \otimes \mathbf{1}^{-\mathcal{I}}$ and $\widehat{\Theta}^{(a)}_{\mathcal{I}}$ depends on $Y^{(a),\text{obs}} \otimes \mathbf{1}^{-\mathcal{I}}$. In other words, the randomness in $\left(\widehat{P}_{\mathcal{I}}, \widehat{\Theta}^{(a)}_{\mathcal{I}}\right)$ stems from the randomness in $\left(A_{-\mathcal{I}}, Y^{(a),\text{obs}}_{-\mathcal{I}}\right)$ which in turn stems from the randomness in $\left(W_{-\mathcal{I}}, E^{(a)}_{-\mathcal{I}}\right)$. Then, Eq. (S.44) follows from Eq. (27). Likewise, the randomness in $\widehat{P}_{\mathcal{I}}$ stems from the randomness in $A_{-\mathcal{I}}$ which in turn stems from the randomness in $W_{-\mathcal{I}}$. Then, Eq. (S.45) follows from Eq. (28).

To prove Eq. (24), we show that

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

for every $\mathcal{I} \in \mathcal{P}$ and $a \in \{0, 1\}$. As mentioned above, the randomness in $\left(\widehat{P}_{\mathcal{I}}, \widehat{\Theta}^{(a)}_{\mathcal{I}}\right)$ stems from the randomness in $\left(A_{-\mathcal{I}}, Y^{(a),\text{obs}}_{-\mathcal{I}}\right)$ which in turn stems from the randomness in $\left(W_{-\mathcal{I}}, E^{(a)}_{-\mathcal{I}}\right)$. Then, Eq. (S.46) follows from Eq. (33).

S8.2 The TW algorithm of Bai and Ng (2021).

Bai and Ng (2021) propose `TW` to impute missing values in matrices with a set of rows and a set of columns without missing entries. More concretely, for any matrix $S \in \{\mathbb{R} \cup \{\,?\}\}^{N \times M}$, let $\mathcal{R}_{\text{obs}} \subseteq [N]$ and $\mathcal{C}_{\text{obs}} \subseteq [M]$ denote the set of all rows and all columns, respectively, with all entries observed. Then, all missing entries of S belong to the block $\mathcal{I} = \mathcal{R}_{\text{miss}} \times \mathcal{C}_{\text{miss}}$, where $\mathcal{R}_{\text{miss}} \triangleq [N] \setminus \mathcal{R}_{\text{obs}}$ and $\mathcal{C}_{\text{miss}} \triangleq [M] \setminus \mathcal{C}_{\text{obs}}$.

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

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

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

and

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

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

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

$$R \triangleq \widetilde{V}^{(\mathrm{tall})\top}\widetilde{V}^{(\mathrm{wide})}\left(\widetilde{V}^{(\mathrm{wide})\top}\widetilde{V}^{(\mathrm{wide})}\right)^{-1}.$$

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

3. Let $\overline{\Sigma}^{(\mathrm{tall})} \in \mathbb{R}^{\bar{r}_N\times r}$ be the sub-matrix of $\Sigma^{(\mathrm{tall})}$ 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.

S9 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 S1 (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:*

$$P = UV^\top \quad \textit{and} \quad \Theta^{(a)} = U^{(a)}V^{(a)\top} \quad \textit{for} \quad a \in \{0,1\}. \tag{S.47}$$

Assumption S1 decomposes each of the unobserved confounders (P, $\Theta^{(0)}$, and $\Theta^{(0)}$) into low-dimensional unit-dependent latent factors (U, $U^{(0)}$, and $U^{(1)}$) and measurement-dependent latent factors (V, $V^{(0)}$, and $V^{(1)}$). In particular, every unit $i \in [N]$ is associated with three low-dimensional factors: (i) $U_i \in \mathbb{R}^{r_p}$, (ii) $U_i^{(0)} \in \mathbb{R}^{r_{\theta_0}}$, and (iii) $U_i^{(1)} \in \mathbb{R}^{r_{\theta_1}}$. Similarly, every measurement $j \in [M]$ is associated with three factors: (i) $V_j \in \mathbb{R}^{r_p}$, (ii) $V_j^{(0)} \in \mathbb{R}^{r_{\theta_0}}$, and (iii) $V_j^{(1)} \in \mathbb{R}^{r_{\theta_1}}$. Low-rank assumptions are widespread 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,

$$\Theta^{(0)} \odot (\mathbf{1} - P) = \overline{U}^{(0)}\overline{V}^{(0)\top} \quad \text{and} \quad \Theta^{(1)} \odot P = \overline{U}^{(1)}\overline{V}^{(1)\top}, \tag{S.48}$$

where $\overline{U}^{(0)} \triangleq \overline{U} * U^{(0)} \in \mathbb{R}^{N\times r_{\theta_0}(r_p+1)}$, $\overline{V}^{(0)} \triangleq \overline{V} * V^{(0)} \in \mathbb{R}^{M\times r_{\theta_0}(r_p+1)}$, $\overline{U}^{(1)} \triangleq U * U^{(1)} \in \mathbb{R}^{N\times r_{\theta_1} r_p}$, and $\overline{V}^{(1)} \triangleq V * V^{(1)} \in \mathbb{R}^{M\times r_{\theta_1} r_p}$, with the operator $*$ denoting the Khatri-Rao product (see Section 1). We provide details of the derivation of these factors in the supplementary appendix (Section S9.1.3). The ranks r_p, r_{θ_0}, and r_{θ_1} can be consistently estimated using the full matrices A, $Y^{(0),\text{full}}$, and $Y^{(1),\text{full}}$, and the methods in Bai and Ng (2002) and Bai (2003). Hence, they are treated as known.

Assumption S2 (Strong factors)**.** *There exists a positive constant c such that*

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

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

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

Assumption S2, 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 S3 (Strong block factors)**.** *Consider the block partition* $\mathcal{P} \triangleq \{\mathcal{R}_s \times \mathcal{C}_k : s, k \in \{0,1\}\}$ *from Assumption 5. For every* $s \in \{0,1\}$*, let* $U_{(s)} \in \mathbb{R}^{|\mathcal{R}_s| \times r_p}$*,* $\overline{U}^{(0)}_{(s)} \in \mathbb{R}^{|\mathcal{R}_s| \times r_{\theta_0}(r_p+1)}$*, and* $\overline{U}^{(1)}_{(s)} \in \mathbb{R}^{|\mathcal{R}_s| \times r_{\theta_1} r_p}$ *be the sub-matrices of* U*,* $\overline{U}^{(0)}$*, and* $\overline{U}^{(1)}$*, respectively, that keeps the rows corresponding to the indices in* $\mathcal{R}_s$*. For every* $k \in \{0,1\}$*, let* $V_{(k)} \in \mathbb{R}^{|\mathcal{C}_k| \times r_p}$*,* $\overline{V}^{(0)}_{(k)} \in \mathbb{R}^{|\mathcal{C}_k| \times r_{\theta_0}(r_p+1)}$*, and* $\overline{V}^{(1)}_{(k)} \in \mathbb{R}^{|\mathcal{C}_k| \times r_{\theta_1} r_p}$ *be the sub-matrices of* V*,* $\overline{V}^{(0)}$*, and* $\overline{V}^{(1)}$*, respectively, that keeps the rows corresponding to the indices in* $\mathcal{C}_k$*. Then, for every* $s, k \in \{0,1\}$*, the matrices defined below exist and are positive definite:*

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

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

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

Assumption S4 (Weak dependence in noise across measurements and independence in noise across units)**.**

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

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

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

Assumption S4(a) and Assumption S4(b) requires the noise variables to exhibit only weak dependency across measurements. Still, these assumptions allow the existence of pairs of perfectly correlated outcomes (e.g., $j, j' \in [M]$ such that $a_{i,j} = a_{i,j'}$). Assumption S4(c) requires the noise $(E^{(a)}, W)$ to be jointly independent across units, for every $a \in \{0,1\}$. We are now ready to prove the guarantees on the estimates produced by `Cross-Fitted-SVD`.

S9.1 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 S5 (Strong linear latent factors). *There exist a constant $r_T \in [\min\{N, M\}]$ and a collection of latent factors*

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

such that,

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

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

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

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

Assumption S6 (Zero-mean, weakly dependent, and subExponential noise). *The noise matrix H is such that,*

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

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

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

Assumption S7 (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 S5. 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 $\mathcal{R}_{\text{obs}}$ and $\mathcal{R}_{\text{miss}}$, respectively. Let $\widetilde{V}^{\text{obs}} \in \mathbb{R}^{|\mathcal{C}_{\text{obs}}| \times r_T}$ and $\widetilde{V}^{\text{miss}} \in \mathbb{R}^{|\mathcal{C}_{\text{miss}}| \times r_T}$ be the sub-matrices of $\widetilde{V}$ that keeps the rows corresponding to the indices in $\mathcal{C}_{\text{obs}}$ and $\mathcal{C}_{\text{miss}}$, respectively. Then, the matrices defined below exist and are positive definite:

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

Further, the mask matrix F is such that

$$|\mathcal{R}_{\text{obs}}| = \Omega(N), \quad |\mathcal{R}_{\text{miss}}| = \Omega(N), \quad |\mathcal{C}_{\text{obs}}| = \Omega(M), \quad \text{and} \quad |\mathcal{C}_{\text{miss}}| = \Omega(M). \tag{S.50}$$

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 of the main article) using the `TW` algorithm (summarized in Section S8.2). Its proof, essentially established as a corollary of Bai and Ng (2021, Lemma A.1), is provided in Section S9.2.

Corollary S3. *Consider a matrix of interest T, a noise matrix H, and a mask matrix F such that that Assumptions S5 to S7 hold. Let $S \in \{\mathbb{R} \cup \{\,?\}\}^{N\times M}$ be the observed matrix as in Eq. (6). Let $\mathcal{R}_{\text{obs}} \subseteq [N]$ and $\mathcal{C}_{\text{obs}} \subseteq [M]$ denote the set of rows and columns of S, respectively, with all entries observed. Let $\mathcal{I} = \mathcal{R}_{\text{miss}} \times \mathcal{C}_{\text{miss}}$ where $\mathcal{R}_{\text{miss}} \triangleq [N] \setminus \mathcal{R}_{\text{obs}}$ and $\mathcal{C}_{\text{miss}} \triangleq [M] \setminus \mathcal{C}_{\text{obs}}$. Then, $\texttt{TW}_{r_T}$ produces an estimate $\widehat{T}_{\mathcal{I}}$ of $T_{\mathcal{I}}$ such that*

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

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 S7, i.e., Eq. (S.50) under Assumption S3.

S9.1.1 Estimating P.

Consider estimating P using `Cross-Fitted-SVD`. To apply Corollary S3, we use Assumptions S1 and S2 to note that P satisfies Assumption S5 with rank parameter r_p. Then, we use Eq. (4), Assumption 2, and Assumption S4 to note that W satisfies Assumption S6. Finally, we use Assumption S3 to note that Assumption S7 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,

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

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

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

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

S9.1.2 Estimating $\Theta^{(0)}$ and $\Theta^{(1)}$.

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

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

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

$$\bar{\lambda}\|\widehat{\Theta}^{(1)} - \Theta^{(1)}\|_{\max} \overset{(a)}{\leq} \|\big(\widehat{\Theta}^{(1)} - \Theta^{(1)}\big) \odot \widehat{P}\|_{\max} \leq \|\Theta^{(1)} \odot \big(\widehat{P} - P\big)\|_{\max} \overset{(b)}{\leq} \|\Theta^{(1)}\|_{\max}\|\widehat{P} - P\|_{\max}, \tag{S.53}$$

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 $(1, 2)$ operator norm, (b) follows from Eq. (S.53), and (c) follows from Eq. (S.51). Then, Eq. (S.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

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

Next, we utilize two claims proven in Sections S9.1.3 and S9.1.4 respectively: $\Theta^{(1)} \odot P$ satisfies Assumption S5 with rank parameter $r_{\theta_1} r_p$ and

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

satisfies Assumption S6. Finally, Assumption S3 implies that Assumption S7 holds.

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

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

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

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

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

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

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,

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

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

S9.1.3 Proof that $\Theta^{(0)} \odot (\mathbf{1} - P)$ and $\Theta^{(1)} \odot P$ satisfy Assumption S5.

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. (S.48). We have

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

where (a) follows from the definition of Khatri-Rao product (see Section 1 of the main article) and (b) follows from the definitions of $\overline{U}^{(1)}$ and $\overline{V}^{(1)}$. The proof for $\Theta^{(0)} \odot (\mathbf{1} - P)$ follows similarly. Then, Assumption S5(a) holds from Eq. (S.48). Next, we note that

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

where (a) follows from the definition of Khatri-Rao product (see Section 1 of the main article), and (b) follows from Assumption S2. Then, $\Theta^{(1)} \odot P$ satisfies Assumption S5(b) by using similar arguments on $\overline{V}^{(1)}$. Further, $\Theta^{(0)} \odot (\mathbf{1} - P)$ satisfies Assumption S5(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 S5(c) holds from Assumption S2.

S9.1.4 Proof that $\bar{\varepsilon}^{(1)}$ satisfies Assumption S6

Recall that $\bar{\varepsilon}^{(1)} \triangleq \Theta^{(1)} \odot W + E^{(1)} \odot P + E^{(1)} \odot W$. Then, Assumption S6(a) holds as $\bar{\varepsilon}^{(1)}_{i,j}$ is zero-mean from Assumption 2 and Eq. (3), and $\bar{\varepsilon}^{(1)}_{i,j}$ is subExponential because $\varepsilon^{(1)}_{i,j}\eta_{i,j}$ is a subExponential random variable Lemma S6, every subGaussian random variable is subExponential random variable, and sum of subExponential random variables is a subExponential random variable. Finally, Assumption S6(b) and Assumption S6(b) hold from Assumption S4(b) and Assumption S4(c), respectively.

S9.2 Proof of Corollary S3

Corollary S3 is a direct application of Bai and Ng (2021, Lemma A.1), specialized to our setting. Notably, Bai and Ng (2021) make three assumptions numbered A, B, and C in their paper to establish the corresponding result. It remains to establish that the conditions assumed in Corollary S3 imply the necessary conditions used in the proof of Bai and Ng (2021, Lemma A.1). First, note that certain assumptions in Bai and Ng (2021) are not actually used in their proof of Lemma A.1 (or in the proof of other results used in that proof), namely, the distinct eigenvalue condition in Assumption A(a)(iii), the asymptotic normality conditions in Assumption A(c) and the asymptotic normality conditions in Assumption C. Next, Eq. (S.50) in Assumption S7 implies Assumption B and Eq. (S.49) in Assumption S7 is equivalent to the remaining conditions in Assumption C.

It remains to show how Assumptions S5 and S6 imply the remainder of conditions in Bai and Ng (2021, Assumptions A). For completeness, these conditions are collected in the following assumption.

Assumption S8. *The noise matrix H is such that,*

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

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

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

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

Assumption S8 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 S8(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 S5 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 S6(a) above is analogous to Assumption A(b)(i) of Bai and Ng (2021). Assumption A(b)(vi) of Bai and Ng (2021) is implied by their other Assumptions for non-random factors as stated in Bai (2003).

To establish Corollary S3, it remains to establish that Assumption S8 holds, which is done in Section S9.2.1 below.

S9.2.1 Assumption S8 holds

First, Assumption S8(a) holds as follows,

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

where (a) follows from triangle inequality and (b) follows from Assumption S6(b). Next, from Assumption S6(a) and Assumption S6(c), we have

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

Then, Assumption S8(b) holds as $\max_{i\in[N]}\max_{j\in[M]}\sum_{i'\in[N]}\left|\mathbb{E}[h_{i,j}h_{i',j}]\right| \leq c$. Next, Assumption S8(c) holds as follows,

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

where (a) follows from Assumption S6(c) and (b) follows from Assumption S6(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 S8(d) holds as follows,

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

where (a) follows from linearity of expectation and Assumption S6(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 S6(a).

S10 Doubly-robust estimation in panel data with lagged effects

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

S10.1 Setup

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

$$y_{i,t}^{(a|y_{i,t-1})} = \alpha^{(a)} y_{i,t-1} + \theta_{i,t}^{(a)} + \varepsilon_{i,t}^{(a)}, \tag{S.56}$$

where $y_{i,t}^{(a|y_{i,t-1})}$ is the potential outcome for unit i at time t given treatment $a \in \{0, 1\}$ and lagged outcome $y_{i,t-1}$. This model combines unobserved confounding and lagged treatment effects, where the lagged effect is carried over via the auto-regressive term, $\alpha^{(a)} y_{i,t-1}$, with $\alpha^{(a)}$ being the auto-regressive parameter for treatment $a \in \{0, 1\}$. The treatment possibly starts at $t = 1$, and $y_{i,0}$ is assumed to not be affected by any future exposure to the treatment. Treatment assignments are continually assumed to be generated via Eq. (3). As in Eq. (1), realized outcomes, $y_{i,t}$, depend on potential outcomes and treatment assignments,

$$y_{i,t} = y_{i,t}^{(0|y_{i,t-1})}(1 - a_{i,t}) + y_{i,t}^{(1|y_{i,t-1})} a_{i,t}, \tag{S.57}$$

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

S10.2 Target causal estimand

The lagged effects in Eq. (S.56) imply that the treatment effects need to be defined for sequences of treatments. For concreteness, consider the effect at time T for an always-treat policy, i.e., $a_{i,t} = 1$, versus never-treat, i.e., $a_{i,t} = 0$, for $i \in [N]$ and $j \in [T]$. Let $y_{i,T}^{[1]}$ be the potential outcome for unit i at time T under always-treat and $y_{i,T}^{[0]}$ be the potential outcome for unit i at time T under never-treat. We aim to estimate the difference in the expected potential outcomes under these two treatment policies averaged over all units,

$$\text{ATE}_{\cdot,T} \triangleq \mu_{\cdot,T}^{[1]} - \mu_{\cdot,T}^{[0]},$$

where

$$\mu_{\cdot,T}^{[a]} \triangleq \frac{1}{N}\sum_{i\in[N]} \mathbb{E}[y_{i,T}^{[a]}],$$

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

Assumption S9 (Zero-mean noise conditioned on the initial outcomes)**.** $\{\varepsilon_{i,t}^{(a)} : i \in [N], t \in [T], a \in \{0,1\}\}$ *are mean zero conditioned on* $\{y_{i,0}\}_{i\in[N]}$.

Assumption S9 holds whenever Assumption 2(a) holds conditioned on the initial outcomes $\{y_{i,0}\}_{i\in[N]}$. Another sufficient condition for Assumption S9 is that $(\varepsilon_{i,t}^{(0)}, \varepsilon_{i,t}^{(1)})$ are independent in time. Given this, the time dependence in the expected potential outcome $\mathbb{E}[y_{i,T}^{[a]}]$ is captured as follows: for $a \in \{0,1\}$

$$\mathbb{E}[y_{i,T}^{[a]}] = (\alpha^{(a)})^T y_{i,0} + \sum_{s=0}^{T-1} (\alpha^{(a)})^s \theta_{i,T-s}^{(a)}. \tag{S.58}$$

Eq. (S.58) forms the basis of our doubly-robust estimator of $\mathrm{ATE}_{\cdot,T}$.

We chose the contrast between always-treat and never-treat for concreteness. However, the framework and the results in this section can be generalized in a straightforward manner to contrast any two pre-specified sequences of treatments, where the treatment can also be chosen stochastically with pre-specified probabilities. For the remainder of this section, we condition on the initial outcomes $\{y_{i,0}\}_{i\in[N]}$ but omit it from our notation for brevity.

S10.3 Doubly-robust estimator

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

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

$$\widehat{\Theta}^{(0)} = \mathtt{MC}(\widetilde{Y}^{(0),\mathrm{obs}}), \quad \widehat{\Theta}^{(1)} = \mathtt{MC}(\widetilde{Y}^{(1),\mathrm{obs}}), \quad \text{and} \quad \widehat{P} = \mathtt{MC}(A). \tag{S.59}$$

Then, the DR estimate is defined as follows:

$$\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,T,J} \triangleq \widehat{\mu}^{[1,\mathrm{DR}]}_{\cdot,T,J} - \widehat{\mu}^{[0,\mathrm{DR}]}_{\cdot,T,J} \quad \text{where} \quad \widehat{\mu}^{[a,\mathrm{DR}]}_{\cdot,T,J} = \frac{1}{N}\sum_{i\in[N]}\left[(\widehat{\alpha}^{(a)})^T y_{i,0} + \sum_{s=0}^{J-1}(\widehat{\alpha}^{(a)})^s \widehat{\theta}^{[a,\mathrm{DR}]}_{i,T-s}\right], \tag{S.60}$$

where

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

and

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

The estimator is parameterized by an integer J, which denotes the contiguous number of time periods preceding time T that are used to estimate the expectations at time T (see the summation in Eq. (S.58)). Notably, using preceding J terms instead of $T-1$ terms allows us to adapt cross-fitting for the setting with lagged treatment effects. Let us briefly elaborate: suppose $(\widehat{\alpha}^{(0)}, \widehat{\alpha}^{(1)})$ are estimated from entries of Y in $[N]\times[L]$ for some $L < T-J$. Consider the column partitions $\mathcal{C}_0 = \{L+1, \ldots, T-J\}$ and $\mathcal{C}_1 = \{T-J+1, \ldots, T\}$ of times $[T]\setminus[L]$. Suppose Eqs. (27) and (28) in Assumption 5 hold for $\mathcal{I} = \mathcal{R}_0 \times \mathcal{C}_1$ and $\mathcal{I} = \mathcal{R}_1 \times \mathcal{C}_1$ for some row partitions $\mathcal{R}_0$ and $\mathcal{R}_1$ of units $[N]$. Then, applying `Cross-Fitted-MC` on the residual matrices $\widetilde{Y}^{(0),\mathrm{obs}}$ and $\widetilde{Y}^{(1),\mathrm{obs}}$ with row partitions $(\mathcal{R}_0, \mathcal{R}_1)$ and column partitions $(\mathcal{C}_0, \mathcal{C}_1)$

ensures that Assumption 4 holds for every column in $\mathcal{C}_1$ with row partitions $(\mathcal{R}_0, \mathcal{R}_1)$.

S10.4 Non-asymptotic guarantees

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

$$\mathcal{E}(\widehat{\alpha}) \triangleq \sum_{a\in\{0,1\}} \mathcal{E}(\widehat{\alpha}^{(a)}) \quad \text{where} \quad \mathcal{E}(\widehat{\alpha}^{(a)}) \triangleq |\widehat{\alpha}^{(a)} - \alpha^{(a)}|. \tag{S.61}$$

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

Assumption S10 (Bounded auto-regressive parameters and estimates)**.** *The auto-regressive parameters and their estimates are such that* $|\alpha^{(a)}| \leq \overline{\alpha}$ *and* $|\widehat{\alpha}^{(a)}| \leq \overline{\alpha}$*, for all* $a \in \{0, 1\}$*, where* $\overline{\alpha} \in [0, 1)$*.*

Assumption S10 requires the regression parameters to be bounded by a fixed constant less than 1. This condition is standard for auto-regressive models, as it implies stability of the outcome process in Eq. (S.56). The analogous condition on the estimated parameters can be ensured by truncating the estimates to $[0, \overline{\alpha}]$.

Assumption S11 (Bounded observed outcomes, mean potential outcomes, and estimated mean potential outcomes)**.** *The observed outcomes, the mean potential outcomes, and the estimates of the mean potential outcomes are such that* $|y_{i,t}| \leq C_1$*,* $|\theta^{(a)}_{i,t}| \leq C_2$*, and* $|\widehat{\theta}^{(a)}_{i,t}| \leq C_3$*, for all* $i \in [N]$*,* $j \in [M]$*, and* $a \in \{0, 1\}$*, where* C_1*,* C_2*, and* C_3 *are universal constants.*

Assumption S11 requires the observed outcomes, the mean potential outcomes, and the estimates of the mean potential outcomes to be bounded to simplify our proof. With a more delicate analysis, Assumption S11 can be relaxed to require the average observed outcomes over $i \in [N]$, the average mean potential outcomes over $i \in [N]$, and the average estimated mean potential outcomes over $i \in [N]$ to be bounded.

Theorem A.1 (Finite Sample Guarantees for DR with lagged effects)**.** *Consider the panel data model with lagged effects defined via Eqs.* (S.56) *and* (S.57)*. Suppose Assumptions 1 to 3, S10, and S11 hold and Assumption 4 holds for $t \in \{T-J+1, \ldots, T\}$ for some integer $J \in [T]$. Fix $\delta \in (0,1)$. Then, with probability at least $1-\delta$, we have*

$$\left|\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,T,J} - \mathrm{ATE}_{\cdot,T}\right| \leq \frac{\mathrm{Err}^{\mathrm{DR}}_{N,\delta/J}}{1-\overline{\alpha}} + C\left[\frac{\overline{\alpha}^J}{1-\overline{\alpha}} + \mathcal{E}(\widehat{\alpha})\Big(T\overline{\alpha}^{T-1} + \frac{1}{1-\overline{\alpha}}\Big)\right], \tag{S.62}$$

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

The proof of Theorem A.1 is given in Section S10.5. For brevity, the finite sample guarantees above use $\mathcal{E}\big(\widehat{\Theta}\big)$ and $\mathcal{E}\big(\widehat{P}\big)$ as defined in Eq. (16), but the proof can be easily modified to replace the $\max_{j\in[T]}$ appearing in the definition of $\|\cdot\|_{1,2}$ in Eq. (16) with $\max_{j\in\{T-J+1,\cdots,T\}}$.

Next, we remark that Theorem A.1 is a strict generalization of Theorem 1. To this end, note that when $\alpha^{(a)} = 0$ for all $a \in \{0,1\}$, the model considered in Theorem A.1 simplifies to the model considered in Theorem 1. For this setting, the assumptions in Theorem 1 imply that the assumptions in Theorem A.1 hold with $J=1$. First, Assumption S10 holds with $\overline{\alpha} = 0$ when $\alpha^{(a)} = 0$ for all $a \in \{0,1\}$. Second, the proof of Theorem A.1 can be easily modified to drop the requirement of Assumption S11 when $J=1$ and $\overline{\alpha}=0$. Substituting $\overline{\alpha}=0$, $\mathcal{E}(\widehat{\alpha}) = 0$ (i.e., the auto-regressive parameters are known to be 0), and $J=1$ in Eq. (S.62) recovers the guarantee stated in Theorem 1.

Doubly-robust behavior of $\widehat{\mathrm{ATE}}^{\mathrm{DR}}_{\cdot,T,J}$. When $\overline{\alpha} \neq 0$ and bounded away from one, Eq. (S.62) bounds the absolute error of the DR estimator by the rate of

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

Then, if the conditions of Theorem A.1 are satisfied for some J such that $C\log N \geq J \geq$

$\log N/(2\log(1/\overline{\alpha}))$, the error rate of the DR estimator is bounded by

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

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

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

Note that Proposition 4 still implies that `Cross-Fitted-SVD` achieves $\mathcal{E}\Big(\widehat{P}\Big)=O_p(N^{-0.5}+T^{-0.5})$ under suitable conditions. To estimate the auto-regressive parameter $\alpha^{(a)}$ for $a\in\{0,1\}$, Bai (2024, Section 5) shows that whenever there exist K units such that they all have treatment a for L consecutive time points, a full information maximum likelihood estimator provides $|\alpha^{(a)}-\widehat{\alpha}^{(a)}|=O_p((KL)^{-0.5})$. Next, establishing a matrix completion guarantee for the mean potential outcomes by residualizing as in Eq. (S.59) can be reduced to deriving a matrix completion guarantee for an approximately low-rank matrix. To this end, Agarwal and Singh (2024, Theorem 5) suggests that, up to logarithmic factors, an error rate of $N^{-0.5}+T^{-0.5}+\mathcal{E}(\widehat{\alpha})$ is plausible for $\mathcal{E}\Big(\widehat{\Theta}\Big)$ for our setting. A complete derivation of error guarantees for $\mathcal{E}(\widehat{\alpha})$ and $\mathcal{E}\Big(\widehat{\Theta}\Big)$ in the dynamic model is an interesting venue for future work.

S10.5 Proof of Theorem A.1: Finite Sample Guarantees for DR with lagged effects

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

$$\Delta\mathrm{ATE}^{\mathrm{DR}}_{\cdot,T}=\Big(\widehat{\mu}^{[1,\mathrm{DR}]}_{\cdot,T,J}-\widehat{\mu}^{[0,\mathrm{DR}]}_{\cdot,T,J}\Big)-\Big(\mu^{[1]}_{\cdot,T}-\mu^{[0]}_{\cdot,T}\Big)=\Big(\widehat{\mu}^{[1,\mathrm{DR}]}_{\cdot,T,J}-\mu^{[1]}_{\cdot,T}\Big)-\Big(\widehat{\mu}^{[0,\mathrm{DR}]}_{\cdot,T,J}-\mu^{[0]}_{\cdot,T}\Big).\quad\text{(S.63)}$$

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

$$\Big|\widehat{\mu}^{[1,\mathrm{DR}]}_{\cdot,T,J}-\mu^{[1]}_{\cdot,T}\Big|\le C\left[\frac{|\alpha^{(1)}|^J-|\alpha^{(1)}|^T}{1-|\alpha^{(1)}|}+\mathcal{E}(\widehat{\alpha}^{(1)})\Big(T\overline{\alpha}^{T-1}+\frac{1-|\alpha^{(1)}|^J}{1-|\alpha^{(1)}|}+\frac{1}{(1-|\alpha^{(1)}|)^2}\Big)\right]$$

$$+\frac{2}{(1-|\alpha^{(1)}|)\bar{\lambda}}\Big[\mathcal{E}(\widehat{\Theta}^{(1)})\mathcal{E}\big(\widehat{P}\big)+\frac{1}{\sqrt{N}}\Big(\frac{\sqrt{c\ell_{\delta/(12J)}}}{\sqrt{\ell_1}}\mathcal{E}(\widehat{\Theta}^{(1)})+2\overline{\sigma}\sqrt{c\ell_{\delta/(12J)}}+\frac{2\overline{\sigma}m(c\ell_{\delta/(12J)})}{\sqrt{\ell_1}}\Big)\Big],\tag{S.64}$$

and

$$\left|\widehat{\mu}^{[0,\mathrm{DR}]}_{\cdot,T,J}-\mu^{[0]}_{\cdot,T}\right|\leq C\left[\frac{|\alpha^{(0)}|^J-|\alpha^{(0)}|^T}{1-|\alpha^{(0)}|}+\mathcal{E}(\widehat{\alpha}^{(0)})\Big(T\overline{\alpha}^{T-1}+\frac{1-|\alpha^{(0)}|^J}{1-|\alpha^{(0)}|}+\frac{1}{(1-|\alpha^{(0)}|)^2}\Big)\right]$$
$$+\frac{2}{(1-|\alpha^{(0)}|)\bar{\lambda}}\Big[\mathcal{E}(\widehat{\Theta}^{(0)})\mathcal{E}\big(\widehat{P}\big)+\frac{1}{\sqrt{N}}\Big(\frac{\sqrt{c\ell_{\delta/(12J)}}}{\sqrt{\ell_1}}\mathcal{E}(\widehat{\Theta}^{(0)})+2\overline{\sigma}\sqrt{c\ell_{\delta/(12J)}}+\frac{2\overline{\sigma}m(c\ell_{\delta/(12J)})}{\sqrt{\ell_1}}\Big)\Big].\tag{S.65}$$

Then, the claim in Eq. (S.62) follows by applying triangle inequality in Eq. (S.63) and using Assumption S10. We prove the bound (S.64) in Section S10.5.1, and also provide an expression for C. The proof of Eq. (S.65) follows similarly.

S10.5.1 Proof of Eq. (S.64)

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

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

where

$$\mathbb{T}^{(1)}_J\triangleq\frac{1}{N}\sum_{s=0}^{J-1}(\alpha^{(1)})^s\sum_{i\in[N]}\theta^{(1)}_{i,T-s},\quad \mathbb{U}^{(1)}_J\triangleq\frac{1}{N}\sum_{s=J}^{T-1}(\alpha^{(1)})^s\sum_{i\in[N]}\theta^{(1)}_{i,T-s},\tag{S.66}$$

and

$$\mathbb{V}^{(1)}\triangleq(\alpha^{(1)})^T\frac{1}{N}\sum_{i\in[N]}y_{i,0}.\tag{S.67}$$

Next, we decompose $\widehat{\mu}^{[1,\mathrm{DR}]}_{\cdot,T,J}$ in Eq. (S.60) as $\widehat{\mu}^{[1,\mathrm{DR}]}_{\cdot,T,J}=\widehat{\mathbb{T}}^{(1)}_J+\widehat{\mathbb{V}}^{(1)}$, where

$$\widehat{\mathbb{T}}^{(1)}_J\triangleq\frac{1}{N}\sum_{s=0}^{J-1}(\widehat{\alpha}^{(1)})^s\sum_{i\in[N]}\widehat{\theta}^{[1,\mathrm{DR}]}_{i,T-s},\quad\text{and}\quad\widehat{\mathbb{V}}^{(1)}\triangleq(\widehat{\alpha}^{(1)})^T\frac{1}{N}\sum_{i\in[N]}y_{i,0}.\tag{S.68}$$

Finally, we define

$$\widetilde{\mathbb{T}}_J^{(1)} \triangleq \frac{1}{N}\sum_{s=0}^{J-1}(\alpha^{(1)})^s \sum_{i\in[N]}\left[\widehat{\theta}_{i,T-s}^{(1)} + \Big(y_{i,T-s} - \alpha^{(1)}y_{i,T-s-1} - \widehat{\theta}_{i,T-s}^{(1)}\Big)\frac{a_{i,T-s}}{\widehat{p}_{i,T-s}}\right], \tag{S.69}$$

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

$$\widehat{\mu}_{\cdot,T,J}^{[1,\mathrm{DR}]} - \mu_{\cdot,T}^{[1]} = (\widehat{\mathbb{V}}^{(1)} - \mathbb{V}^{(1)}) + (\widetilde{\mathbb{T}}_J^{(1)} - \mathbb{T}_J^{(1)}) + (\widehat{\mathbb{T}}_J^{(1)} - \widetilde{\mathbb{T}}_J^{(1)}) - \mathbb{U}_J^{(1)}. \tag{S.70}$$

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

$$\Big|\widehat{\mathbb{V}}^{(1)} - \mathbb{V}^{(1)}\Big| \leq C_0 T\mathcal{E}(\widehat{\alpha}^{(1)})\overline{\alpha}^{T-1}, \qquad |\mathbb{U}_J^{(1)}| \leq C_2\frac{|\alpha^{(1)}|^J - |\alpha^{(1)}|^T}{1-|\alpha^{(1)}|}, \tag{S.71}$$

and

$$|\widehat{\mathbb{T}}_J^{(1)} - \widetilde{\mathbb{T}}_J^{(1)}| \leq \mathcal{E}(\widehat{\alpha}^{(1)})\bigg(\frac{C_1}{\lambda}\frac{(1-|\alpha^{(1)}|^J)}{1-|\alpha^{(1)}|} + C_{\mathrm{DR}}\frac{1}{(1-|\alpha^{(1)}|)^2}\bigg), \tag{S.72}$$

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

$$\begin{aligned}|\widetilde{\mathbb{T}}_J^{(1)} - \mathbb{T}_J^{(1)}| \leq \frac{2}{(1-|\alpha^{(1)}|)\bar{\lambda}}\bigg[&\mathcal{E}(\widehat{\Theta}^{(1)})\mathcal{E}\Big(\widehat{P}\Big) \\ &+ \bigg(\frac{\sqrt{c\ell_{\delta/(12J)}}}{\sqrt{\ell_1}}\mathcal{E}(\widehat{\Theta}^{(1)}) + 2\overline{\sigma}\sqrt{c\ell_{\delta/(12J)}} + \frac{2\overline{\sigma}m(c\ell_{\delta/(12J)})}{\sqrt{\ell_1}}\bigg)\frac{1}{\sqrt{N}}\bigg],\end{aligned} \tag{S.73}$$

holds with probability at least $1-\delta/2$. The claim in Eq. (S.64) follows by applying triangle inequality in Eq. (S.70) and using the above bounds.

It remains to establish the intermediate claims Eqs. (S.71) to (S.73). Throughout the rest of the proof, we repeatedly use the inequality below that holds for all $s \in [T]$:

$$\begin{aligned}\Big|(\widehat{\alpha}^{(1)})^s - (\alpha^{(1)})^s\Big| = \Big|(\widehat{\alpha}^{(1)} - \alpha^{(1)})\Big(\sum_{l\in[s]}(\widehat{\alpha}^{(1)})^{s-l}(\alpha^{(1)})^{l-1}\Big)\Big| &\overset{(a)}{\leq} s\Big|(\widehat{\alpha}^{(1)} - \alpha^{(1)})\Big|\overline{\alpha}^{s-1} \\ &\overset{(b)}{=} s\mathcal{E}(\widehat{\alpha}^{(1)})\overline{\alpha}^{s-1},\end{aligned} \tag{S.74}$$

where (a) follows from Assumption S10 and (b) follows from Eq. (S.61).

Proof of Eq. (S.71) First, from Eq. (S.66), we have

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

where (a) follows from Assumption S11 and (b) follows from the sum of geometric series.

Next, from Eqs. (S.67) and (S.68), we have

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

where (a) follows from the definition of C_0 and Eq. (S.74).

Proof of Eq. (S.72) From Eqs. (S.68) and (S.69), and the triangle inequality, we have

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

where (a) follows from Eq. (S.61), Assumptions 3 and S11, and because $\max_{i\in[N],t\in[T]}\left|\widehat{\theta}_{i,t}^{[1,\mathrm{DR}]}\right| \le C_{\mathrm{DR}}$ from Assumptions 3, S10, and S11, and (b) follows from the sum of geometric and arithmetico-geometric sequences.

Proof of Eq. (S.73) We start by defining

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

Then, from Eqs. (S.66) and (S.69), we have

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

where (a) follows from triangle inequality. From Eqs. (3) and (S.56), we have

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

Then, the term $\widetilde{\theta}_{i,T-s}^{[1,\mathrm{DR}]} - \theta_{i,T-s}^{(1)}$ is analogous to the display Eq. (S.2) in the proof of Theorem 1. Following similar algebra as in Section S1, we first obtain

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

Now, note that Assumption 4 holds for $j = T - s$ for all $s \in \{0, \ldots, J-1\}$. Hence, for any such s and for any $\delta \in (0,1)$, mimicking the derivation of Eq. (S.5) from Section S1, we obtain, with probability at least $1 - \delta/(2J)$,

$$\frac{1}{N}\Big|\sum_{i\in[N]}(\widetilde{\theta}_{i,T-s}^{[1,\mathrm{DR}]} - \theta_{i,T-s}^{(1)})\Big| \leq \frac{2}{\bar{\lambda}}\mathcal{E}\big(\widehat{\Theta}^{(1)}\big)\mathcal{E}\big(\widehat{P}\big) + \frac{2\sqrt{c\ell_{\delta/(12J)}}}{\bar{\lambda}\sqrt{\ell_1 N}}\mathcal{E}\big(\widehat{\Theta}^{(1)}\big) + \frac{2\overline{\sigma}\sqrt{c\ell_{\delta/(12J)}}}{\bar{\lambda}\sqrt{N}} + \frac{2\overline{\sigma}m(c\ell_{\delta/(12J)})}{\bar{\lambda}\sqrt{\ell_1 N}}. \tag{S.75}$$

Finally, multiplying both sides of Eq. (S.75) by $(\alpha^{(1)})^s$, summing it over $s \in \{0, \ldots, J-1\}$, and using a union bound argument yields that the bound in Eq. (S.73) holds with probability at least $1 - \delta/2$.

S11 Doubly-robust estimation in panel data with staggered adoption

This section shows how to extend the doubly-robust framework of this article to a setting with panel data and staggered adoption. Recall (from Section S10) that for panel data, t denotes the column (time) index and T denotes the total number of columns (time periods). In a staggered adoption setting, for every unit $i \in [N]$, there exists a time point $t_i \in [T]$ such that $a_{i,t} = 0$ for $t \leq t_i$, and $a_{i,t} = 1$ for $t > t_i$. This defines the observed treatment assignment matrix A. As mentioned in Section 5.3 of the main article and illustrated in the example below, a staggered treatment assignment leads to a heavy time-series dependence in $\{\eta_{i,t}\}_{t\in[T]}$.

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

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

Further, for units remaining in control,

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

and for units receiving treatment,

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

The strong time-series dependence in $\eta_{i,t}$ above implies that Assumption S3 or Assumption S4(a) do not hold, which in turn implies that the guarantees for `Cross-Fitted-SVD`, as in Proposition 4, may not hold. To see this, first note that to ensure Assumption 5, the set of column partitions $\{\mathcal{C}_0, \mathcal{C}_1\}$ must be equal to $\{[T_0], [T] \setminus [T_0]\}$ due to the dependence in the noise W. Now, for Assumption S3 to hold, we need $|\mathcal{C}_k| = \Omega(T)$ for every $k \in \{0, 1\}$. However, for Assumption S4(a) to hold, we need $T - T_0$ to be a constant with respect to T as, for any $t \in [T] \setminus [T_0]$ and $i \in [N]$, $\sum_{t' \in [T]} \left|\mathbb{E}[\eta_{i,t}\eta_{i,t'}]\right| = (T - T_0)c_i$ where $c_i \in \{p_i^2, (1 - p_i)^2\}$.

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

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

Assumption S12 (Staggered adoption and common unit factors)**.** *We consider a panel data setting with staggered adoption where*

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

For Example S1, the function g corresponds to the inner product, the unit-dependent latent factors are 1-dimensional (i.e., $r = 1$) with $U_i = p_i$ for every $i \in [N]$, and the time-dependent latent factors for the assignment probability are such that $V_t = 0$ for every $t \in [T_0]$ and $V_t = 1$ for every $t \in [T] \setminus [T_0]$. Consequently, Example S1 is consistent with Assumption S12 if $U_i^{(a)} = p_i$ for every $a \in \{0, 1\}$ and $i \in [N]$. Next, we provide a more flexible version of Example S1 that allows different adoption times for different units.

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

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

For Example S2, the unit-dependent latent factors are 1-dimensional (i.e., $r = 1$) with $U_i = p_i$ for every $i \in [N]$, and the time-dependent latent factors for the assignment probability are such that $V_t = 0$ for every $t \in [T_0]$ and $V_t = t - T_0$ for every $t \in [T] \setminus [T_0]$. Further the function g is such that $g(U_i, V_t) = 1 - (1 - U_i)^{V_t}$. Consequently, Example S2 is consistent with Assumption S12 if $U_i^{(a)} = p_i$ for every $a \in \{0, 1\}$ and $i \in [N]$.

We now describe `Cross-Fitted-Regression`, an algorithm that generates estimates of $(\Theta^{(0)}, \Theta^{(1)}, P)$ for the staggered adoption model in Assumption S12 such that Assumption 4 holds.

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

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

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

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

$s \in \{0,1\}$, define $\mathcal{R}_s = \mathcal{R}_s^{(0)} \cup \mathcal{R}_s^{(1)}$.

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

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

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

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

Assumption S13 (Independence across units and with respect to pre-adoption noise)**.**

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

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

Assumption S13(a) requires the noise $(E^{(a)}, W)$ corresponding to a time period $t \in T \setminus [T_0]$ to be jointly independent across units given the noise $E^{(0)}$ corresponding to time periods $[T_0]$, for every $a \in \{0,1\}$. Assumption S13(b) is satisfied if, for instance, the noise variables follow a moving average model of order $t - T_0 - 1$. The following result, proven in Section S11.1, establishes that the estimates generated by `Cross-Fitted-Regression` satisfy Assumption 4. Deriving error bounds, i.e., $\mathcal{E}\left(\widehat{P}\right)$ and $\mathcal{E}\left(\widehat{\Theta}\right)$, for the estimates generated by `Cross-Fitted-Regression` for the staggered adoption model is an interesting direction for future research.

Proposition S1 (Guarantees for `Cross-Fitted-Regression`)**.** *Consider the staggered adoption model in Assumption S12 and suppose Assumption S13 holds. Fix any* $t \in [T] \setminus [T_0]$, *and* $\{\widehat{\theta}^{(0)}_{i,t}, \widehat{\theta}^{(1)}_{i,t}, \widehat{p}_{i,t}\}_{i\in[N]}$ *be the estimates returned by* `Cross-Fitted-Regression`*. Then, Assumption 4 holds.*

S11.1 Proof of Proposition S1: Guarantees for `Cross-Fitted-Regression`

Fix any $s \in \{0, 1\}$. Then, Assumption S13(a) and Assumption S13(b) imply that

$$\{\varepsilon^{(0)}_{i,t}\}_{i\in[N],t\in[T_0]} \cup \{\eta_{i,t}, \varepsilon^{(a)}_{i,t}\}_{i\in\overline{\mathcal{R}}_{1-s}} \perp\!\!\!\perp \{\eta_{i,t}, \varepsilon^{(a)}_{i,t}\}_{i\in\overline{\mathcal{R}}_s}, \tag{S.76}$$

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

`Cross-Fitted-Regression` estimates $\{\widehat{p}_{i,t}\}_{i\in\mathcal{R}_s}$ using $\{\widehat{U}_i\}_{i\in\mathcal{R}_s}$ and $\widehat{V}_s$, where $\widehat{V}_s$ is estimated using $\{\widehat{U}_i\}_{i\in\mathcal{R}_{1-s}}$ and $\{a_{i,t}\}_{i\in\mathcal{R}_{1-s}}$. Therefore, the randomness in $\{\widehat{p}_{i,t}\}_{i\in\mathcal{R}_s}$ stems from the randomness in $Y^{(0),\text{pre}}$ and $\{a_{i,t}\}_{i\in\mathcal{R}_{1-s}}$ which in turn stems from the randomness in $\{\varepsilon^{(0)}_{i,t}\}_{i\in[N],t\in[T_0]}$ and $\{\eta_{i,t}\}_{i\in\mathcal{R}_{1-s}}$. Then, Eq. (15) follows from Eq. (S.76).

Next, fix any $a \in \{0, 1\}$. Then, `Cross-Fitted-Regression` estimates $\{\widehat{\theta}^{(a)}\}_{i\in\mathcal{R}_s}$ using $\{\widehat{U}_i\}_{i\in\mathcal{R}_s}$ and $\widehat{V}^{(a)}_s$, where $\widehat{V}^{(a)}_s$ is estimated using $\{\widehat{U}_i\}_{i\in\mathcal{R}^{(a)}_{1-s}}$ and $\{y_{i,t}\}_{i\in\mathcal{R}^{(a)}_{1-s}}$. Therefore, the randomness in $\{\widehat{\theta}^{(a)}\}_{i\in\mathcal{R}_s}$ stems from the randomness in $Y^{(0),\text{pre}}$ and $\{y_{i,t}\}_{i\in\mathcal{R}^{(a)}_{1-s}}$ which in turn stems from the randomness in $\{\varepsilon^{(0)}_{i,t}\}_{i\in[N],t\in[T_0]}$ and $\{\varepsilon^{(a)}_{i,t}\}_{i\in\mathcal{R}^{(a)}_{1-s}}$. Then, Eq. (14) follows from Eq. (S.76).