4-2023

Adjustment with many regressors under covariate-adaptive randomizations

Liang JIANG

Liyao LI

Ke MIAO

Yichong ZHANG

Singapore Management University, yczhang@smu.edu.sg

Adjustment with Many Regressors Under Covariate-Adaptive Randomizations*

Liang Jiang† Liyao Li‡ Ke Miao§ Yichong Zhang¶

Abstract

Our paper identifies a trade-off when using regression adjustments (RAs) in causal inference under covariate-adaptive randomizations (CARs). On one hand, RAs can improve the efficiency of causal estimators by incorporating information from covariates that are not used in the randomization. On the other hand, RAs can degrade estimation efficiency due to their estimation errors, which are not asymptotically negligible when the number of regressors is of the same order as the sample size. Failure to account for the cost of RAs can result in over-rejection of causal inference under the null hypothesis. To address this issue, we develop a unified inference theory for the regression-adjusted average treatment effect (ATE) estimator under CARs. Our theory has two key features: (1) it ensures the exact asymptotic size under the null hypothesis, regardless of whether the number of covariates is fixed or diverges at most at the rate of the sample size, and (2) it guarantees weak efficiency improvement over the ATE estimator with no adjustments.

Keywords: Covariate-adaptive randomization, many regressors, regression adjustment.

JEL codes: C14, C21, D14, G21

*Yichong Zhang acknowledges the financial support from the Lee Kong Chian fellowship and the NSFC under the grant No. 72133002. Any and all errors are our own.

†Fudan University. E-mail address: jiangliang@fudan.edu.cn.

‡East China Normal University. E-mail address: lyli@fem.ecnu.edu.cn.

§Fudan University. E-mail address: miaoke@fudan.edu.cn.

¶Singapore Management University. E-mail address: yczhang@smu.edu.sg.

1 Introduction

Regression adjustments (RAs) are often used in randomized controlled trials to improve the efficiency of causal inference. However, Freedman (2008a,b) showed that OLS regression with covariates, as in the analysis of covariance (ANCOVA), can actually decrease the precision of the average treatment effect (ATE) estimator. Lin (2013) further examined Freedman's critique and found that for a linear RA to be "no-harm" (i.e., guarantee efficiency improvement over the unadjusted estimator), it must include a full set of interactions between treatment status and covariates. Since then, many "no-harm" RAs have been developed for various randomization schemes and causal parameters under both finite- and super-population asymptotics, including Shao, Yu, and Zhong (2010); Bloniarz, Liu, Zhang, Sekhon, and Yu (2016); Fogarty (2018); Liu, Tu, and Ma (2020); Li and Ding (2020); Ma, Tu, and Liu (2022); Cohen and Fogarty (2020); Lei and Ding (2021); Jiang, Phillips, Tao, and Zhang (2022); Jiang, Linton, Tang, and Zhang (2022); Reluga, Ye, and Zhao (2022); Ye, Shao, Yi, and Zhao (2022); Ye, Yi, and Shao (2022); Bai, Jiang, Romano, Shaikh, and Zhang (2023); Cytrynbaum (2023); Chiang, Matsushita, and Otsu (2023), among others.[1] When the regressors are chosen as sieve bases with a growing dimension, causal estimators with such "no-harm" RAs can potentially achieve the semiparametric efficiency bound, as shown in Jiang et al. (2022) and Bai et al. (2023).

This paper focuses on linear RAs for estimating the average treatment effect (ATE) under covariate adaptive randomizations (CARs).[2] The paper identifies a new efficiency trade-off of using "no-harm" RAs. On one hand, these adjustments can improve the estimation efficiency by incorporating information from covariates that are not used in the randomization; on the other hand, they can degrade estimation efficiency due to their estimation errors, which are not asymptotically negligible when the number of regressors is of the same order of the sample size. Ignoring the cost of RAs in asymptotic variance estimators can lead to over-rejection under the null. Furthermore, a consistent variance estimator that accounts for the cost of RAs can potentially be larger than that of the simple estimator without using RAs, even asymptotically, if the cost outweights the benefit. Therefore, using a "no-harm" RA that incurs such costs can be harmful.

Under the asymptotic framework where the number of regressors diverges to infinity at most at the rate of the sample size, we derive the joint asymptotic distribution for both estimators with and without RAs, as well as a consistent estimator of the asymptotic covariance matrix, which takes into account both the benefit and cost of RAs. Based on this covariance matrix, we construct a new estimator of ATE by optimally linearly combining the adjusted and unadjusted estimators. We

[1] We apologize for any unintentional omissions.

[2] Under CARs, units are first stratified using some baseline covariates, and then, within each stratum, the treatment status is assigned (independent of covariates) to achieve the balance between the numbers of treated and control units. Such randomization schemes have been widely used in economic research. See, for example, Chong, Cohen, Field, Nakasone, and Torero (2016); Greaney, Kaboski, and Van Leemput (2016); Jakiela and Ozier (2016); Burchardi, Gulesci, Lerva, and Sulaiman (2019); Anderson and McKenzie (2021), among others.

show that the corresponding Wald test (1) achieves the exact asymptotic size under the null and (2) is weakly more efficient than both the adjusted and unadjusted estimators. We also consider an alternative asymptotic framework under which the dimension of regressors is fixed and provide mild regularity conditions under which the same optimal estimator with the same covariance matrix estimator satisfies the above two properties. This implies that empirical researchers can use our inference method without explicitly specifying which asymptotic framework their analyses belong to.

The contributions of this paper relate to other recent research. First, Liu et al. (2020), Ma et al. (2022), and Ye et al. (2022) considered linear RAs for ATE under CARs when the (effective) number of covariates[3] is less than the square root of the sample size. At this rate, the cost of RAs is asymptotically negligible. Under a *finite-population* framework with the *complete random sampling*, Li and Ding (2020) and Chiang et al. (2023) considered bias-correction for the ATE allowing the number of regressors to grow at a rate that is *slower* than the sample size. In contrast, we consider a *super-population* framework with *CARs* as established by Bugni, Canay, and Shaikh (2018), in which the number of regressors can grow at the *same* rate as the sample size.

Second, the present paper complements the seminal work of Cattaneo, Jansson, and Newey (2018) in three ways. First, while Cattaneo et al. (2018) considered the inference of linear coefficients in OLS regression with many regressors under the assumption of independent observations or clusters, we consider observations generated by CARs, which violate this assumption by having cross-sectional dependence among treatment assignments and observed outcomes. Second, our ATE estimator involves a linear combination of intercepts from multiple OLS regressions with many regressors, and those intercepts are not asymptotically independent. The linear combination step will introduce additional estimation error which is not asymptotically negligible. Due to the two-step nature, the asymptotic distribution of our ATE estimator is not a direct consequence of results established by Cattaneo et al. (2018). Instead, we combine techniques from Bugni et al. (2018) to deal with cross-sectional dependence, Yurinskii's coupling, and an anti-concentration inequality from Chernozhukov, Chetverikov, and Kato (2014) to establish distributional theory in our setting. Third, we introduce a new optimal estimator and discuss efficiency improvement under the many regressors framework, which are new to the literature.

Third, while Cattaneo, Jansson, and Ma (2019) studied the two-step estimation when the number of regressors is of the same order of the square root of the sample size, our setting has the intercept estimators enter the second step in a linear manner, resulting in a knife-edge rate that is the order of the sample size.

Fourth, we propose a new estimator of the covariance matrix based on cross-fit estimators of both the variance of regression coefficients and the "variance component" with many coefficients, as proposed by Jochmans (2022) and Kline, Saggio, and Sølvsten (2020), respectively. Moreover, in

[3] They consider the ultra-high dimension that the number of regressors can be higher than the sample size and use Lasso to select the effective regressors, which is assumed to be less than the sample size.

order to take into account the dependence between the first-stage intercept estimators, we complement Kline et al. (2020) by proposing a new estimator for the "covariance component" with many coefficients.

Notation. For any positive integer m, let 1_m and I_m be the $m \times 1$ vector of ones and the $m \times m$ identity matrix, respectively. Let $||\cdot||_2$ and $||\cdot||_{op}$ denote the ℓ_2-norm for a vector and operator norm for a matrix, respectively. For a symmetric and positive semi-definite matrix Υ, $\lambda_{\max}(\Upsilon)$ denotes the maximum eigenvalue of Υ. We define $W^{(n)}$ as the sample of W's, i.e., $W^{(n)} = \{W_i\}_{i \in [n]}$, where $[n] = 1, \cdots, n$. We write $U \overset{d}{=} V$ for two random variables U and V if they share the same distribution.

2 Setup

Potential outcomes for treated and control groups are denoted by $Y(1)$ and $Y(0)$, respectively. Treatment status is denoted by A, with $A = 1$ indicating treated and $A = 0$ untreated. The stratum indicator is denoted by S, based on which the researcher implements the covariate-adaptive randomization. The support of S is denoted by $\mathcal{S}$, a finite set. After randomization, the researcher can observe the data $\{Y_i, S_i, A_i, X_i\}_{i \in [n]}$ where $Y_i = Y_i(1)A_i + Y_i(0)(1 - A_i)$ is the observed outcome, and X_i contains covariates besides S_i in the dataset. The support of X is denoted $\text{Supp}(X)$. In this paper, we allow X_i and S_i to be dependent. For $i \in [n]$, let $n_s = \sum_{i \in [n]} 1\{S_i = s\}$, $n_{1,s} = \sum_{i \in [n]} A_i 1\{S_i = s\}$, and $n_{0,s} = n_s - n_{1,s}$. Let $\aleph_{a,s} = \{i \in [n] : A_i = a, S_i = s\}$ denote the set of individuals in stratum s with treatment status a and $\aleph_s = \aleph_{1,s} \cup \aleph_{0,s}$. Our parameter of interest is the average treatment effect defined as $\tau = \mathbb{E}(Y(1) - Y(0))$.

We make the following assumptions on the data generating process (DGP) and the treatment assignment rule.

Assumption 1. *(i)* $\{Y_i(1), Y_i(0), S_i, X_i\}_{i \in [n]}$ *is i.i.d.*

(ii) $\{Y_i(1), Y_i(0), X_i\}_{i \in [n]} \perp\!\!\!\perp \{A_i\}_{i \in [n]} | \{S_i\}_{i \in [n]}$.

(iii) Suppose $p_s = \mathbb{P}(S_i = s)$ *is fixed with respect to (w.r.t.)* n *and is positive for every* $s \in \mathcal{S}$.

(iv) Let π_s *denote the target fraction of treatment for stratum* s. *Then,* $c < \min_{s \in \mathcal{S}} \pi_s \leq \max_{s \in \mathcal{S}} \pi_s < 1 - c$ *for some constant* $c \in (0, 0.5)$ *and* $\frac{D_{n,s}}{n_s} = o_P(1)$ *for* $s \in \mathcal{S}$, *where* $D_{n,s} = \sum_{i \in [n]} (A_i - \pi_s) 1\{S_i = s\}$.

Remark 1. *Several remarks are in order. First, Assumption 1(i) assumes* $(Y(1), Y(0), S, X)^{(n)}$ *are independent but allows for* $A^{(n)}$, *and thus,* $Y^{(n)}$ *to be cross-sectionally dependent, which will be the case under CARs. The identical distribution assumption can be relaxed by using a set of more complex notation. Second, Assumption 1(ii) implies that the treatment assignment* $A^{(n)}$ *are generated only based on strata indicators. Third, Assumption 1(iii) imposes that the number of* *strata is bounded and the strata sizes are approximately balanced. Fourth, Bugni et al. (2018) show that Assumption 1(iv) holds under several specific covariate-adaptive treatment assignment rules such as simple random sampling (SRS), biased-coin design (BCD), adaptive biased-coin design (WEI) and stratified block randomization (SBR). For completeness, we provide brief descriptions below. Note that the requirement $D_{n,s}/n_s = o_P(1)$ is weaker than the assumption imposed by Bugni et al. (2018), but it is the same as that imposed by Bugni, Canay, and Shaikh (2019) and Zhang and Zheng (2020).*

Example 1 (SRS)**.** *Let $\{A_i\}_{i=1}^n$ be drawn independently across i and of $\{S_i\}_{i=1}^n$ as Bernoulli random variables with success rate π_s, i.e., for $k = 1, \ldots, n$,*

$$\mathbb{P}\left(A_k = 1 \middle| \{S_i\}_{i=1}^n, \{A_j\}_{j=1}^{k-1}\right) = \mathbb{P}(A_k = 1|S_k) = \pi_{S_k}.$$

Example 2 (WEI)**.** *This design was first proposed by Wei (1978). Let $n_{k-1}(S_k) = \sum_{i=1}^{k-1} 1\{S_i = S_k\}$, $B_{k-1}(S_k) = \sum_{i=1}^{k-1}\left(A_i - \frac{1}{2}\right) 1\{S_i = S_k\}$, and*

$$\mathbb{P}\left(A_k = 1 \middle| \{S_i\}_{i=1}^k, \{A_i\}_{i=1}^{k-1}\right) = f\left(\frac{2B_{k-1}(S_k)}{n_{k-1}(S_k)}\right),$$

where $f(\cdot) : [-1, 1] \mapsto [0, 1]$ is a pre-specified non-increasing function satisfying $f(-x) = 1 - f(x)$. Here, $\frac{B_0(S_1)}{n_0(S_1)}$ and $B_0(S_1)$ are understood to be zero.

Example 3 (BCD)**.** *The treatment status is determined sequentially for $1 \leq k \leq n$ as*

$$\mathbb{P}\left(A_k = 1|\{S_i\}_{i=1}^k, \{A_i\}_{i=1}^{k-1}\right) = \begin{cases} \frac{1}{2} & \text{if } B_{k-1}(S_k) = 0 \\ \lambda & \text{if } B_{k-1}(S_k) < 0 \\ 1-\lambda & \text{if } B_{k-1}(S_k) > 0, \end{cases}$$

where $B_{k-1}(s)$ is defined as above and $\frac{1}{2} < \lambda \leq 1$.

Example 4 (SBR)**.** *For each stratum, $\lfloor \pi_s n_s \rfloor$ units are assigned to treatment and the rest are assigned to control.*

Remark 2. *We note that SRS and SBR allow for the target faction of treatment to be different for different strata. For WEI and BCD, we have $\pi_s = 1/2$ for $s \in \mathcal{S}$. Bugni et al. (2018) further shows BCD and SBR achieve strong balance in the sense that $D_{n,s} = o_P(n^{1/2})$.*

The fully saturated regression adjusted estimator is denoted as $\hat{\tau}^{adj}$ and defined as

$$\hat{\tau}^{adj} = \sum_{s \in \mathcal{S}} \hat{p}_s(\hat{\tau}_{1,s} - \hat{\tau}_{0,s}),$$

where $\hat{\tau}_{a,s}$ is compute as the intercept in the OLS regression of

$$Y_i \sim 1 + \breve{X}_i, \tag{2.1}$$

using all observations in $\aleph_{a,s}$, $\hat{p}_s = n_s/n$, and $\breve{X}_i = X_i - \overline{X}_{S_i}$ with $\overline{X}_s = \frac{1}{n_s}\sum_{i\in\aleph_s} X_i$.

Remark 3. *We consider this adjusted estimator for two reasons. First, it is a natural extension to the fully saturated estimator proposed by Bugni et al. (2019). Ye et al. (2022) have also considered this same estimator (their $\hat{\theta}_A$) and showed that it is "no-harm" when the dimension of X_i is fixed. Second, we can write $\hat{\tau}^{adj}$ as an augmented inverse propensity score weighted (AIPW) estimator:*

$$\hat{\tau}^{adj} = \frac{1}{n}\sum_{i\in[n]} \frac{A_i(Y_i - X_i^\top \hat{\beta}_{1,S_i})}{\hat{\pi}_{S_i}} - \frac{1}{n}\sum_{i\in[n]} \frac{(1-A_i)(Y_i - X_i^\top \hat{\beta}_{0,S_i})}{1-\hat{\pi}_{S_i}} + \frac{1}{n}\sum_{i\in[n]} X_i^\top(\hat{\beta}_{1,S_i} - \hat{\beta}_{0,S_i}),$$

where $\hat{\pi}_s = n_{1,s}/n_s$. It is important to note that, if an individual i belongs to stratum s, then $\breve{X}_i$ is defined as X_i demeaned by the stratum-level mean $\frac{1}{n_s}\sum_{i\in\aleph_s} X_i$, not the stratum-treatment level mean $\frac{1}{n_{a,s}}\sum_{i\in\aleph_{a,s}} X_i$. This is the key for the AIPW interpretation to hold.

We also consider the unadjusted IPW estimator

$$\hat{\tau}^{unadj} = \frac{1}{n}\sum_{i\in[n]} \frac{A_i Y_i}{\hat{\pi}_{S_i}} - \frac{1}{n}\sum_{i\in[n]} \frac{(1-A_i)Y_i}{1-\hat{\pi}_{S_i}}.$$

Remark 4. *It is natural to consider $\hat{\tau}^{unadj}$ in our context. First, it is simply the AIPW estimator with an empty set of covariates. Second, it is the polar opposite of $\hat{\tau}^{adj}$ in that it does not benefit from RAs but also does not suffer from the costs of RAs.*

Denote the dimension of X as k. In Sections 3 and 4 below, respectively, we study the asymptotic properties of $(\hat{\tau}^{adj}, \hat{\tau}^{unadj})$ in two scenarios: (1) $k = k_n$ which diverges with sample size n and (2) k is fixed. No matter which case we are in, we use the same covariance estimator $\hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_{1,1} & \hat{\Sigma}_{1,2} \\ \hat{\Sigma}_{1,2} & \hat{\Sigma}_{2,2} \end{pmatrix}$ defined in Section 3.2 below.

Then, our final estimator $\hat{\tau}^* = \hat{w}\hat{\tau}^{adj} + (1-\hat{w})\hat{\tau}^{unadj}$ is a linear combination of $\hat{\tau}^{adj}$ and $\hat{\tau}^{unadj}$ where the weight $\hat{\omega}$ is chosen to minimize the corresponding asymptotic variance. Given $\hat{\tau}^{adj}$ and $\hat{\tau}^{unadj}$ are not asymptotically equivalent, we have $p\lim_{n\to\infty}(\hat{\Sigma}_{1,1} + \hat{\Sigma}_{2,2} - 2\hat{\Sigma}_{1,2}) > 0$, which implies the optimal weight is

$$\hat{w} = \frac{\hat{\Sigma}_{2,2} - \hat{\Sigma}_{1,2}}{\hat{\Sigma}_{1,1} + \hat{\Sigma}_{2,2} - 2\hat{\Sigma}_{1,2}},$$

and the corresponding estimator of the asymptotic variance for $\hat{\tau}^*$ is $\hat{\sigma}_*^2 = (\hat{w}, 1-\hat{w})\hat{\Sigma}(\hat{w}, 1-\hat{w})^\top$.

In both diverging and fixed k cases mentioned above, we give regularity conditions under which (1) $\sqrt{n}(\hat{\tau}^* - \tau)/\ \hat{\sigma}_* \rightsquigarrow \mathcal{N}(0,1)$ so that the corresponding Wald test has the exact asymptotic size under the null and (2) $\hat{\tau}^*$ is weakly more efficient than both $\hat{\tau}^{adj}$ and $\hat{\tau}^{unadj}$.

3 Many Regressors

Recall k is the dimension of X. In this section, we consider the case that $k = k_n$ increases at most at the rate of the sample size n. To clearly state our assumptions, we need to introduce more notation. Let $\breve{X}_{\aleph_{a,s}}$ be a $n_{a,s} \times k_n$ matrix which is constructed by stacking $\breve{X}_i^\top$ for $i \in \aleph_{a,s}$. We define $A_{\aleph_{a,s}}$ similarly. Then, let $P_{a,s} = \breve{X}_{\aleph_{a,s}}(\breve{X}_{\aleph_{a,s}}^\top \breve{X}_{\aleph_{a,s}})^{-1}\breve{X}_{\aleph_{a,s}}^\top$ be the projection matrix of $\breve{X}_{\aleph_{a,s}}$, $M_{a,s} = I_{n_{a,s}} - P_{a,s}$,

$$\begin{aligned}
&\varepsilon_i(a) = Y_i(a) - \mathbb{E}(Y_i(a)|X_i, S_i),\\
&\gamma_{a,s,n} = (1_{n_{a,s}}^\top M_{a,s} 1_{n_{a,s}})/n_s(a),\\
&\sigma^2_{a,s,n} = \frac{1}{n_{a,s}} \sum_{i \in \aleph_{a,s}} \left[\left(\sum_{j \in \aleph_{a,s}} M_{a,s,i,j} \right)^2 \mathbb{E}(\varepsilon_i^2(a)|X_i, S_i = s) \right], \quad \text{and}\\
&\rho_{a,s,n} = \frac{1}{n_{a,s}} \sum_{i \in \aleph_{a,s}} \left[\left(\sum_{j \in \aleph_{a,s}} M_{a,s,i,j} \right) \mathbb{E}(\varepsilon_i^2(a)|X_i, S_i = s) \right].
\end{aligned}$$

Assumption 2. *(i) For $a = 0, 1$ and $s \in \mathcal{S}$, we have*

$$\mathbb{E}(Y_i(a)|X_i, S_i = s) = \alpha_{a,s} + X_i^\top \beta_{a,s} + e_{i,s}(a),$$

such that $\mathbb{E}(e^2_{i,s}(a)|S_i = s) = o(n^{-1})$.

(ii) $\max_{i \in [n]} \mathbb{E}\left[\varepsilon_i^4(a)|X_i, S_i\right] = O_P(1)$.

(iii) There exists a constant $b > 0$ such that $\min_{a=0,1,s \in \mathcal{S}, i \in [n]} \mathbb{E}\left[\varepsilon_i^2(a)|X_i, S_i = s\right] \geq b$.

(iv) $k_n/n_{a,s} \to \kappa_{a,s} \in [0,1)$ *and* $\limsup_{n \to \infty} \max_{a=0,1,s \in \mathcal{S}} P_{a,s,i,i} \leq 1 - \delta$ *for some constant* $\delta \in (0,1)$.

(v) $\max_{a=0,1,s \in \mathcal{S}} \max_{i \in \aleph_{a,s}} \left| \sum_{j \in \aleph_{a,s}} M_{a,s,i,j} \right| = o_P(n^{1/2})$.

Remark 5. *Assumption 2(i) assumes that $\mathbb{E}(Y_i(a)|X_i, S_i = s)$ is approximately linear with asymptotically negligible approximation error $e_{i,s}(a)$. This condition is commonly assumed in linear regression analyses with many regressors, as seen in Cattaneo et al. (2018, Assumption 3) and Jochmans (2022, Assumption 3). Kline et al. (2020) further require Assumption 2(i) holds with no approximation error, i.e., $e_{i,s}(a) = 0$. This condition is reasonable because we allow the dimension k_n to diverge to infinity. For example, if the fixed dimensional baseline covariatesï¼Œ denoted* *as Z_i, are continuous, Assumption 2(i) holds when X_i contains sieve basis functions of Z_i and $\mathbb{E}(Y_i(a)|Z_i = z, S_i = s)$ is sufficiently smooth in z.*[4] *When all baseline covariates are categorical (discrete), Assumption 2(i) holds when X_i contains the fully saturated dummies for all categories. When k_n is fixed and does not diverge to infinity, Assumption 2(i) may not hold. However, in Section 4, we show that, in the case with a fixed k, our estimator and inference method are still valid even the linear regression is misspecified. Assumption 2(ii) and 2(iii) are mild regularity conditions. Assumption 2(iv) implies we allow the number of covariates to diverge at the rate of the sample size n. Because we run the stratum-treatment level regression with an effective sample size of $n_{a,s}$, We require $\kappa_{a,s} < 1$ to avoid multicollinearity. Assumption 2(v) is the same as Jochmans (2022, Assumption 3).*[5] *For more discussion on this condition, we refer readers to Jochmans (2022).*

Assumption 3. *Suppose $\gamma_{a,s,n} \xrightarrow{p} \gamma_{a,s,\infty} > 0$, $\gamma_{a,s,n}^{-2}\sigma_{a,s,n}^2 \xrightarrow{p} \omega_{a,s,\infty}^2 > 0$, and $\gamma_{a,s,n}^{-1}\rho_{a,s,n} \xrightarrow{p} \varpi_{a,s,\infty}$ for some deterministic constants $(\gamma_{a,s,\infty}, \omega_{a,s,\infty}, \varpi_{a,s,\infty})$.*

Remark 6. *If $k_n \log k_n = o(n)$ (which means $\kappa_{a,s} = 0$ for all (a,s)), then under general conditions on the distribution of X_i, it is possible to show that Assumption 3 holds with $\gamma_{a,s,\infty} = 1$, $\omega_{a,s,\infty}^2 = \varpi_{a,s,\infty} = \mathbb{E}(\varepsilon_i^2(a)|S_i = s)$. See Belloni, Chernozhukov, Chetverikov, and Kato (2015) and Cattaneo, Farrell, and Feng (2020) for more discussion and examples. If $\kappa_{a,s} > 0$, then in Section A in the Online Supplement, we provide detailed calculation of $\gamma_{a,s,\infty}$ when the $k_n \times 1$ vector X_i given $S_i = s$ is jointly Gaussian with a covariance matrix $\Sigma_{s,n}$. Specifically, we show that*

$$\gamma_{a,s,\infty} = \frac{1}{1 + (a(1-\pi_s) + (1-a)\pi_s)\zeta_{a,s}} < 1,$$

where $\zeta_{a,s} = \int_{\lambda_-}^{\lambda_+} \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{2\pi\lambda^2} d\lambda$ and $\lambda_\pm = (1 \pm \sqrt{\kappa_{a,s}})^2$ are due to the Marčenko-Pastur theorem (Marčenko and Pastur, 1967). If we further assume homoskedasticity in the sense that $\mathbb{E}(\varepsilon_i^2(a)|X_i, S_i = s) = \mathbb{E}(\varepsilon_i^2(a)|S_i = s)$, then, we have

$$\omega_{a,s,\infty}^2 = \varpi_{a,s,\infty} = \gamma_{a,s,\infty}^{-1}\mathbb{E}(\varepsilon_i^2(a)|S_i = s) > \mathbb{E}(\varepsilon_i^2(a)|S_i = s).$$

We believe that Assumption 3 holds for non-Gaussian distributions of X_i and heteroskedastic errors. However, establishing general preliminary conditions for this assumption requires the use of advanced tools from random matrix theory[6]*. Such a discussion is beyond the scope of this paper. Finally, we want to emphasize that researchers do not need to know the exact values of $(\gamma_{a,s,\infty}, \omega_{a,s,\infty}^2, \varpi_{a,s,\infty})$ to apply our inference method.*

[4] In this case, we have $\mathbb{E}(Y_i(a)|Z_i, S_i = s) = \alpha_{a,s} + X_i^\top \beta_{a,s} + \tilde{e}_{i,s}(a)$ such that $\mathbb{E}(\tilde{e}_{i,s}^2(a)|S_i = s) = o(n^{-1})$. Then, we have $\mathbb{E}(Y_i(a)|X_i, S_i = s) = \alpha_{a,s} + X_i^\top \beta_{a,s} + e_{i,s}(a)$ where $e_{i,s}(a) = \mathbb{E}(\tilde{e}_{i,s}(a)|X_i, S_i = s)$. Then, by Jansen's inequality, we have $\mathbb{E}(e_{i,s}^2(a)|S_i = s) \leq \mathbb{E}(\tilde{e}_{i,s}^2(a)|S_i = s) = o(n^{-1})$.

[5] In fact, Jochmans's (2022) $\hat{v}_{i,n}$ is just $\sum_{j \in \aleph_{a,s}} M_{a,s,i,j}$.

[6] See, for example, Bai (2008) for a survey.

3.1 Asymptotic Properties

Theorem 3.1. *Suppose Assumptions 1–3 hold. Then, we have*

$$\sqrt{n}\begin{pmatrix}\hat{\tau}^{adj}-\tau\\ \hat{\tau}^{unadj}-\tau\end{pmatrix} \rightsquigarrow \begin{pmatrix}\mathcal{U}^{adj}\\ \mathcal{U}^{unadj}\end{pmatrix}+\begin{pmatrix}\mathcal{V}^{adj}\\ \mathcal{V}^{unadj}\end{pmatrix}+\begin{pmatrix}\mathcal{W}\\ \mathcal{W}\end{pmatrix},$$

where

$$\mathcal{U}=\begin{pmatrix}\mathcal{U}^{adj}\\ \mathcal{U}^{unadj}\end{pmatrix}\stackrel{d}{=}\mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix},\Sigma_{\mathcal{U}}\right),\ \Sigma_{\mathcal{U}}=\begin{pmatrix}\mathbb{E}\left(\frac{\omega^2_{1,S_i,\infty}}{\pi_{S_i}}+\frac{\omega^2_{0,S_i,\infty}}{1-\pi_{S_i}}\right) & \mathbb{E}\left(\frac{\varpi_{1,S_i,\infty}}{\pi_{S_i}}+\frac{\varpi_{0,S_i,\infty}}{1-\pi_{S_i}}\right)\\ \mathbb{E}\left(\frac{\varpi_{1,S_i,\infty}}{\pi_{S_i}}+\frac{\varpi_{0,S_i,\infty}}{1-\pi_{S_i}}\right) & \mathbb{E}\left(\frac{\mathbb{E}(\varepsilon_i^2(1)|S_i)}{\pi_{S_i}}+\frac{\mathbb{E}(\varepsilon_i^2(0)|S_i)}{1-\pi_{S_i}}\right)\end{pmatrix},$$

$$\mathcal{V}=\begin{pmatrix}\mathcal{V}^{adj}\\ \mathcal{V}^{unadj}\end{pmatrix}\stackrel{d}{=}\mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix},\Sigma_{\mathcal{V}}\right),\ \Sigma_{\mathcal{V}}=\begin{pmatrix}\mathbb{E}var(\phi_i(1)-\phi_i(0)|S_i) & \mathbb{E}var(\phi_i(1)-\phi_i(0)|S_i)\\ \mathbb{E}var(\phi_i(1)-\phi_i(0)|S_i) & \mathbb{E}\left[\frac{var(\phi_i(1)|S_i)}{\pi_{S_i}}+\frac{var(\phi_i(0)|S_i)}{(1-\pi_{S_i})}\right]\end{pmatrix},$$

$$\mathcal{W}\stackrel{d}{=}\mathcal{N}(0,\Sigma_{\mathcal{W}}),\ \Sigma_{\mathcal{W}}=var(\mathbb{E}(Y_i(1)-Y_i(0)|S_i)),$$

$\phi_i(a)=\mathbb{E}(Y_i(a)|X_i,S_i)-\mathbb{E}(Y_i(a)|S_i)$ *for* $a=0,1$, *and* (U,V,W) *are independent.*

Remark 7. *The variance decomposition shows that* $\Sigma_{\mathcal{U}}$, $\Sigma_{\mathcal{V}}$, *and* $\Sigma_{\mathcal{W}}$ *capture different sources of variability in the outcome variable* $Y(a)$ *given* X_i *and* S_i, X_i *given* S_i, *and* S_i, *respectively. In particular,* $\Sigma_{\mathcal{U}}$ *represents the estimation error of the RA, and its* $(1,1)$ *and* $(2,2)$ *elements are* $Var(\mathcal{U}^{adj})$ *and* $Var(\mathcal{U}^{unadj})$, *respectively. Following Remark 6, we can show that when* $k_n\log k_n = o(n)$, $Var(\mathcal{U}^{adj})=Var(\mathcal{U}^{unadj})$, *i.e., the estimation error of the RA is asymptotically negligible. Previous literature establishes the same result under a more stringent condition that* $k_n=o(\sqrt{n})$. *However, when* k_n *is of the same order of* n, $Var(\mathcal{U}^{adj})$ *can be larger than* $Var(\mathcal{U}^{unadj})$. *In fact, under the Gaussian covariates example mentioned in Remark 6, we have*

$$\text{Variance Inflate Factor}\equiv\frac{\omega^2_{1,s,\infty}-var(\varepsilon_i^2(1)|S_i=s)}{var(\varepsilon_i^2(a)|S_i=s)}=(a((1-\pi_s))+(1-a)\pi_s)\zeta_{a,s}>0.$$

When $\pi_s=1/2$, *the variance inflate factor (VIF) is just* $\zeta_{a,s}/2$. *Figure 1 plots the values of the VIF as a function of* $\kappa_{a,s}$. *We can see that the VIF is more than* 12.5%, 25%, 50%, *and* 100% *if* $\kappa_{a,s}$ *is greater than* 1/5, 1/3, 1/2, 2/3, *respectively. Also note that the* $(1,1)$ *and* $(2,2)$ *elements of* $\Sigma_{\mathcal{V}}$ *are* $Var(\mathcal{V}^{adj})$ *and* $Var(\mathcal{V}^{unadj})$, *respectively. It can be shown that* $Var(\mathcal{V}^{adj})\leq Var(\mathcal{V}^{unadj})$, *which indicates the benefit of RA by incorporating information from* X_i. *Therefore, the difference between the asymptotic variances of* $\hat{\tau}^{adj}$ *and* $\hat{\tau}^{unadj}$ *is*

$$(var(\mathcal{U}^{adj})-var(\mathcal{U}^{unadj}))+(var(\mathcal{V}^{adj})-var(\mathcal{V}^{unadj})),$$

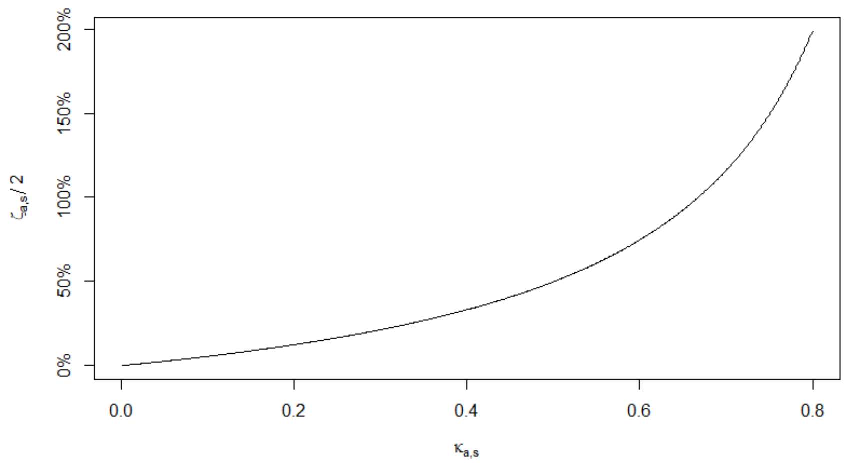


Figure 1: The Variance Inflation Factor

where the second term is always non-positive but the first term can be positive. In some scenarios, it is possible for the first term to dominate the second, resulting in the estimator with the theoretically "no-harm" adjustment to be even less efficient than the unadjusted estimator. This defeats the purpose of using RAs. See model 1 in Section 5 for a numerical illustration.

3.2 Variance Estimator and Further Improvement

In this section, we propose a consistent estimator $\hat{\Sigma}$ of $\Sigma = \Sigma_{\mathcal{U}} + \Sigma_{\mathcal{V}} + \Sigma_{\mathcal{W}} 1_2 1_2^\top$. We then construct our final estimator $\hat{\tau}^*$ based on $\hat{\Sigma}$ and show it is consistent, asymptotically normal, and weakly more efficient than both $\hat{\tau}^{adj}$ and $\hat{\tau}^{unadj}$.

To construct $\hat{\Sigma}$, we first note that

$$\Sigma_{2,2} = \mathbb{E}\left[\frac{var(Y_i(1)|S_i)}{\pi_{S_i}} + \frac{var(Y_i(0)|S_i)}{1-\pi_{S_i}}\right] + \Sigma_{\mathcal{W}}$$

is the asymptotic variance of the unadjusted estimator and does not depend on X_i. Let

$$\begin{aligned}\hat{\Sigma}_{2,2} =& \frac{1}{n}\sum_{s\in\mathcal{S}}\left[\sum_{i\in\aleph_{1,s}}\left(\frac{Y_i}{\hat{\pi}_s} - \frac{1}{n_{1,s}}\sum_{i\in\aleph_{1,s}}\frac{Y_i}{\hat{\pi}_s}\right)^2 + \sum_{i\in\aleph_{0,s}}\left(\frac{Y_i}{1-\hat{\pi}_s} - \frac{1}{n_{0,s}}\sum_{i\in\aleph_{0,s}}\frac{Y_i}{1-\hat{\pi}_s}\right)^2\right] \\ &+ \sum_{s\in\mathcal{S}}\hat{p}_s\left(\frac{1}{n_{1,s}}\sum_{i\in\aleph_{1,s}}Y_i - \frac{1}{n_{0,s}}\sum_{i\in\aleph_{0,s}}Y_i - \hat{\tau}^{unadj}\right)^2.\end{aligned}$$

This estimator is consistent as long as Assumption 1 holds. See, for example, Bugni et al. (2019).

Therefore, it suffices to construct consistent estimators for the $(1,1)$ and $(1,2)$ elements of $\Sigma_{\mathcal{U}}$ and $\Sigma_{\mathcal{V}}$, and $\Sigma_{\mathcal{W}}$. To this end, denote $\hat{\beta}_{a,s}$ as the coefficient of $\breve{X}_i$ in the linear regression (2.1) and recall the intercept is $\hat{\tau}_{a,s}$. The residual of the regression and its leave-one-out version are denoted as $\hat{\varepsilon}_{a,s,i} = Y_i - \hat{\tau}_{a,s} - \breve{X}_i'\hat{\beta}_{a,s}$ and $\acute{\varepsilon}_{a,s,i} = \hat{\varepsilon}_{a,s,i}/M_{a,s,i,i}$ for $i \in \aleph_{a,s}$, respectively. Then, we construct our estimator $\hat{\Sigma}_{\mathcal{U}}$ following the cross-fit approach developed by Jochmans (2022). It is also possible to construct the estimator based on the sample splitting approach proposed by Cattaneo et al. (2018). The cross-fit estimator, as shown below, is asymptotic valid as long as $\kappa_{a,s} < 1$, but is not guaranteed to be positive definite in finite sample. In contrast, the sample-splitting estimator is guaranteed to be positive semidefinite but requires $\kappa_{a,s} < 1/2$. When both of them are valid, they are asymptotically equivalent. We refer interested readers to Jochmans (2022) for more discussions. For the rest of the paper, we focus on cross-fit estimators of the $(1,1)$ and $(1,2)$ elements of $\hat{\Sigma}_{\mathcal{U}}$, which is defined as

$$\hat{\Sigma}_{\mathcal{U}} = \begin{pmatrix} \sum_{s\in\mathcal{S}}\sum_{a=0,1}\frac{n_s^2}{nn_{a,s}}\hat{\omega}_{a,s}^2 & \sum_{s\in\mathcal{S}}\sum_{a=0,1}\frac{n_s^2}{nn_{a,s}}\hat{\varpi}_{a,s} \\ \sum_{s\in\mathcal{S}}\sum_{a=0,1}\frac{n_s^2}{nn_{a,s}}\hat{\varpi}_{a,s} & \bullet \end{pmatrix},$$

where

$$\hat{\omega}_{a,s}^2 = \frac{1}{n_{a,s}}\gamma_{a,s,n}^{-2}\sum_{i\in\aleph_{a,s}}\left(\sum_{j\in\aleph_{a,s}}M_{a,s,i,j}\right)^2 Y_i\acute{\varepsilon}_{a,s,i},$$

$$\hat{\varpi}_{a,s} = \frac{1}{n_{a,s}}\gamma_{a,s,n}^{-1}\sum_{i\in\aleph_{a,s}}\left(\sum_{j\in\aleph_{a,s}}M_{a,s,i,j}\right) Y_i\acute{\varepsilon}_{a,s,i}.$$

To define $\hat{\Sigma}_{\mathcal{V}}$, we note that $\mathbb{E}(Y_i(a)|X_i,S_i) - \mathbb{E}(Y_i(a)|S_i)$ can be approximated by $\breve{X}_i^\top\beta_{a,s}$, where $\beta_{a,s}$ is defined in Assumption 2(i). Kline et al. (2020) consider the estimation and inference for quadratic functions of the coefficient in one linear regression with many regressors. However, in our setting, $\Sigma_{\mathcal{V}}$ depends on multiple linear coefficients $\beta_{a,s}$ which are estimated from separate strata of observations and each regression model may only be approximately linear. Therefore, the estimator $\hat{\Sigma}_{\mathcal{V}}$ includes both the quadratic and cross-product terms of the estimated coefficients $\hat{\beta}_{a,s}$. Denote $\Gamma_{a,s} = \sum_{i\in\aleph_{a,s}}\breve{X}_i\breve{X}_i^\top$ and $\Gamma_s = \sum_{i\in\aleph_s}\breve{X}_i\breve{X}_i^\top$. Following the spirit of Kline et al. (2020), our estimator of the $(1,1)$ and $(1,2)$ elements of $\hat{\Sigma}_{\mathcal{V}}$ is defined as

$$\hat{\Sigma}_{\mathcal{V}} = \begin{pmatrix} \hat{\Sigma}_{\mathcal{V}}^{adj} & \hat{\Sigma}_{\mathcal{V}}^{adj} \\ \hat{\Sigma}_{\mathcal{V}}^{adj} & \bullet \end{pmatrix},$$

where

$$\hat{\Sigma}_{\mathcal{V}}^{adj} = \sum_{s\in\mathcal{S}} \hat{p}_s \left[\sum_{a=0,1} \left(\frac{1}{n_{a,s}} \hat{\beta}_{a,s}^\top \Gamma_{a,s} \hat{\beta}_{a,s} - \frac{1}{n_{a,s}} \sum_{i\in\aleph_{a,s}} P_{a,s,i,i} Y_i \acute{\varepsilon}_{a,s,i} \right) - \frac{2}{n_s} \hat{\beta}_{1,s}^\top \Gamma_s \hat{\beta}_{0,s} \right].$$

The estimator of $\Sigma_{\mathcal{W}}$ is standard:

$$\hat{\Sigma}_{\mathcal{W}} = \sum_{s\in\mathcal{S}} \hat{p}_s \left(\hat{\tau}_{1,s} - \tau_{0,s} - \hat{\tau}^{adj} \right)^2.$$

Assumption 4. *Suppose* $\max_{a=0,1,s\in\mathcal{S}} \left\| \Gamma_s^{1/2} \Gamma_{a,s}^{-1} \Gamma_s^{1/2} \right\|_{op} = o_P(n)$.

Remark 8. *By the random matrix theory, under general conditions, one can show that all the eigenvalues of* Γ_s/n_s *and* $\Gamma_{a,s}/n_{a,s}$ *are bounded and bounded away from zero, even if* $\kappa_{a,s} > 0$. *This then implies* $\max_{a=0,1,s\in\mathcal{S}} \left\| \Gamma_s^{1/2} \Gamma_{a,s}^{-1} \Gamma_s^{1/2} \right\|_{op} = O_P(1)$, *and thus, Assumption 4 holds.*

Theorem 3.2. *Suppose Assumptions 1–4 hold. Let*

$$\hat{\Sigma} = \begin{pmatrix} \sum_{s\in\mathcal{S}} \sum_{a=0,1} \frac{n_s^2}{n n_{a,s}} \hat{\omega}_{a,s}^2 + \hat{\Sigma}_{\mathcal{V}}^{adj} + \hat{\Sigma}_{\mathcal{W}} & \sum_{s\in\mathcal{S}} \sum_{a=0,1} \frac{n_s^2}{n n_{a,s}} \hat{\varpi}_{a,s} + \hat{\Sigma}_{\mathcal{V}}^{adj} + \hat{\Sigma}_{\mathcal{W}} \\ \sum_{s\in\mathcal{S}} \sum_{a=0,1} \frac{n_s^2}{n n_{a,s}} \hat{\varpi}_{a,s} + \hat{\Sigma}_{\mathcal{V}}^{adj} + \hat{\Sigma}_{\mathcal{W}} & \hat{\Sigma}_{2,2} \end{pmatrix}.$$

Then, we have

$$\hat{\Sigma} \xrightarrow{p} \Sigma.$$

Theorem 3.3. *Suppose Assumptions 1–4 hold. Let* $\Sigma = \begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{1,2} & \Sigma_{2,2} \end{pmatrix}$. *If* $\Sigma_{1,1} - 2\Sigma_{1,2} + \Sigma_{2,2} > 0$, *then we have*

$$\sqrt{n} \left[(\hat{w}, 1-\hat{w}) \hat{\Sigma} (\hat{w}, 1-\hat{w})^\top \right]^{-1/2} (\hat{\tau}^* - \tau) \rightsquigarrow \mathcal{N}(0,1).$$

In this case, $\hat{\tau}^*$ *is asymptotically weakly more efficient than both* $\hat{\tau}^{adj}$ *and* $\hat{\tau}^{unadj}$.

Remark 9. *When* $\Sigma_{1,1} - 2\Sigma_{1,2} + \Sigma_{2,2} > 0$, *the adjusted estimator* $\hat{\tau}^{adj}$ *and the unadjusted estimator* $\hat{\tau}^{unadj}$ *are not asymptotically equivalent. In this case,* $(w, 1-w)\Sigma(w, 1-w)^\top$ *has a unique minimizer* $w^* = \frac{\Sigma_{2,2} - \Sigma_{1,2}}{\Sigma_{1,1} - 2\Sigma_{1,2} + \Sigma_{2,2}}$, *and our* $\hat{w}$ *is a consistent estimator of* w^*. *If the estimation error of the RA is asymptotically negligible and* X_i *does not contain useful information of* $(Y_i(1), Y_i(0))$ *in the sense that*

$$\mathbb{E}(Y_i(a)|X_i, S_i) = \mathbb{E}(Y_i(a)|S_i),$$

then we have $\Sigma_{1,1}-2\Sigma_{1,2}+\Sigma_{2,2}=0$ (we know $\Sigma_{1,1}-2\Sigma_{1,2}+\Sigma_{2,2}$ is nonnegative by Cauchy-Schwartz's inequality). In this case, any linear combinations of $\hat{\tau}^{adj}$ and $\hat{\tau}^{unadj}$ are asymptotically equivalent. If this case is a realistic concern, then we can always safeguard against the zero denominator by introducing a small positive constant λ and constructing the final estimator as $\tilde{\tau}=\tilde{w}\hat{\tau}^{adj}+(1-\tilde{w})\hat{\tau}^{unadj}$ with $\tilde{w}=\frac{\hat{\Sigma}_{2,2}-\hat{\Sigma}_{1,2}}{\hat{\Sigma}_{1,1}-2\hat{\Sigma}_{1,2}+\hat{\Sigma}_{2,2}+\lambda}$. It is possible to show that $\tilde{\tau}$ is always weakly more efficient than $\hat{\tau}^{unadj}$ as long as $\lambda>0$, regardless of whether $\Sigma_{1,1}-2\Sigma_{1,2}+\Sigma_{2,2}=0$ or $\Sigma_{1,1}-2\Sigma_{1,2}+\Sigma_{2,2}>0$. Finally, we note that the condition $\Sigma_{1,1}-2\Sigma_{1,2}+\Sigma_{2,2}>0$ is testable.

4 Fixed Number of Regressors

In this section, we consider the CARs specified in Assumption 1 in which the dimension of X_i (k) is fixed w.r.t. n. We then give another set of regularity conditions in replacement of Assumptions 2–4 under which the Wald test constructed based on the same estimator $\hat{\tau}^*$ defined in Section 2 (which depends on the same $\hat{\tau}^{adj}$ and $\hat{\tau}^{unadj}$) and the same $\hat{\Sigma}$ defined in Section 3.2 has the exact asymptotic size under the null and is weakly more powerful than the Wald test based on the unadjusted estimator $\hat{\tau}^{unadj}$. Most importantly, under such a fixed dimension framework, we can establish above results without assuming the RAs are approximately correctly specified.

Assumption 5. *Denote the dimension of X_i as k which is fixed w.r.t. the sample size. In addition, there exist constants c, C such that*

$$0<c\leq\lambda_{\min}\left(\frac{1}{n_{a,s}}\sum_{i\in\aleph_{a,s}}\breve{X}_i\breve{X}_i^\top\right)\leq\lambda_{\max}\left(\frac{1}{n_{a,s}}\sum_{i\in\aleph_{a,s}}\breve{X}_i\breve{X}_i^\top\right)\leq C<\infty$$

and $\max_{i\in[n]}||\breve{X}_i||_2=o_P(n^{1/2})$.

Remark 10. *Assumption 5 is standard when the dimension of $\breve{X}_i$ is fixed.*

Theorem 4.1. *Suppose Assumptions 1 and 5 holds. Then, we have*

$$\sqrt{n}\begin{pmatrix}\hat{\tau}^{adj}-\tau\\ \hat{\tau}^{unadj}-\tau\end{pmatrix}\rightsquigarrow\mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix},\Omega\right),$$

$$\Omega=\left\{\mathbb{E}\left[\frac{Var(Y_i(1)|X_i,S_i)}{\pi_{S_i}}+\frac{Var(Y_i(0)|X_i,S_i)}{1-\pi_{S_i}}\right]+\mathbb{E}(m_1(X_i,S_i)-m_0(X_i,S_i)-\tau)^2\right\}1_2 1_2^\top$$
$$+\sum_{s\in\mathcal{S}}\frac{p_s}{\pi_s(1-\pi_s)}\begin{pmatrix}V_s & V_s\\ V_s & V_s'\end{pmatrix},$$

and

$$\hat{\Sigma} \xrightarrow{p} \Omega,$$

where $m_a(x,s) = \mathbb{E}(Y_i(s)|X_i = x, S_i = s)$*,* $\overline{\beta}^*_s = (1-\pi_s)\beta^*_{1,s} + \pi_s\beta^*_{0,s}$*,*

$$\beta^*_{a,s} = Var(X_i \mid S_i = s)^{-1}Cov(X_i, Y_i(a)|S_i = s),$$
$$V_s = Var((1-\pi_s)m_1(X_i,s) + \pi_s m_0(X_i,s) - X_i^\top\overline{\beta}^*_s \mid S_i = s),$$
$$V'_s = Var((1-\pi_s)m_1(X_i,s) + \pi_s m_0(X_i,s) \mid S_i = s),$$

If we further assume $\pi_s\beta^*_{0,s} + (1-\pi_s)\beta_{1,s} \neq 0$*, then*

$$\sqrt{n}\left[(\hat{w}, 1-\hat{w})\hat{\Sigma}(\hat{w}, 1-\hat{w})^\top\right]^{-1/2}(\hat{\tau}^* - \tau) \rightsquigarrow \mathcal{N}(0,1).$$

In this case, $\hat{\tau}^*$ *is asymptotically equivalent to* $\hat{\tau}^{adj}$ *in sense that* $\hat{\tau}^* = \hat{\tau}^{adj} + o_P(n^{-1/2})$ *and weakly more efficient than* $\hat{\tau}^{unadj}$*.*

Remark 11. *The limit distribution of* $\hat{\tau}^{adj}$ *has already been derived by Ye et al. (2022). Here, our main contributions are (1) deriving the joint distribution of* $\hat{\tau}^{adj}$ *and* $\hat{\tau}^{unadj}$ *and (2) establishing the consistency of our cross-fit covariance matrix estimator* $\hat{\Sigma}$*. Theorem 4.1 shows that the Wald test based on our final estimator* $\hat{\tau}^*$ *proposed in Section 2 and the covariance matrix estimator* $\hat{\Sigma}$ *proposed in Section 3.2 still controls asymptotic size under the null and has the same power as* $\hat{\tau}^{adj}$ *under alternatives when the dimension of* X_i *is fixed. In fact, our* $\hat{\tau}^*$ *(and equivalently* $\hat{\tau}^{adj}$*) is the optimally linearly adjusted estimator in the sense that it achieves the minimum asymptotic variance among a class of estimators which are adjusted by linear functions of* X_i*. See, for example, Liu et al. (2020), Ma et al. (2022), and Ye et al. (2022).*

We can test the two-sided hypothesis $\tau = \tau_0$ v.s. $\tau \neq \tau_0$ using the usual Wald test

$$\mathbb{W}_n = 1\left\{n\left[(\hat{w}, 1-\hat{w})\hat{\Sigma}(\hat{w}, 1-\hat{w})^\top\right]^{-1}(\hat{\tau}^* - \tau_0)^2 \geq \mathcal{C}_\alpha\right\},$$

where $\alpha \in (0,1)$ is the significance level and $\mathcal{C}_\alpha$ is the $(1-\alpha)$ quantile of the chi-squared distribution with one degree of freedom. Combining Theorems 3.1 and 4.1, we can show that the Wald test $\mathbb{W}_n$ has the uniform asymptotic size control under the null over a wide range of DGPs that have both diverging and fixed dimensions of covariates.

We denote a DGP with sample size n as ψ_n which belongs to a class of DGPs Ψ_n. Then, we consider the sequence of classes of DGPs ($\{\Psi_n\}_{n\geq 1}$) that satisfies the following condition:

$$\left(\begin{array}{c} \text{for any subsequence } \{n_l\} \text{ of } \{n\}, \text{ there exists a subsequence } \{n_{l'}\} \text{ of } \{n_l\} \\ \text{such that along } \{\psi_{n_{l'}}\} \text{ for } \psi_{n_{l'}} \in \Psi_{n_{l'}}, \\ \text{either } k_{n_{l'}} \to \infty, \text{ Assumptions 1–4 hold, and } \liminf_{l'\to\infty} \Sigma_{1,1} - 2\Sigma_{1,2} + \Sigma_{2,2} > 0 \\ \text{or } k_{n_{l'}} = k \text{ is fixed, Assumptions 1 and 5 hold, and } \liminf_{l'\to\infty} |\pi_s \beta^*_{0,s} + (1-\pi_s)\beta_{1,s}| > 0 \end{array}\right). \tag{4.1}$$

Theorem 4.2. *Suppose $\{\Psi_n\}_{n\geq 1}$ satisfies (4.1), and we are under the null hypothesis that $\tau_0 = \tau$. Then, we have*

$$\liminf_{n\to\infty} \inf_{\psi_n \in \Psi_n} \mathbb{E}_{\psi_n}(\mathbb{W}_n) = \limsup_{n\to\infty} \sup_{\psi_n \in \Psi_n} \mathbb{E}_{\psi_n}(\mathbb{W}_n) = \alpha.$$

5 Simulations

We conduct a simulation study to evaluate the finite sample performance of our proposed inference methods. The nominal size of the tests is $\alpha = 5\%$ for all cases. For comparison, we also compare our methods with two existing methods introduced by Bugni et al. (2018) and Ye et al. (2022), respectively. For $a \in \{0, 1\}$, we generate potential outcomes according to the equation

$$Y_i(a) = \mu_a + m_a(Z_i) + \sigma_a(Z_i)\varepsilon_{a,i}, \tag{5.1}$$

where μ_a, and $m_a(Z_i)$ and $\sigma_a(Z_i)$are specified as follows. In each of the following specifications, $\{Z_i, \varepsilon_{1i}, \varepsilon_{0i}\}$ are independent and identically distributed (IID). We have considered two models following the designs by Cattaneo et al. (2018):

Model 1 (Linear model with many dummy variables)

The dimension of Z_i is set to $d_n = 0.2n/|\mathcal{S}|$. The first entry of Z_i is uniformly distributed in $[-1, 1]$, i.e. $Z_{1i} \sim U[-1, 1]$. The other entries of Z_i are dummies: $[Z_{2i}, Z_{3i}, \cdots, Z_{d_n,i}]^\top = \mathbf{1}(\mathbf{v}_i \geq \Phi^{-1}(0.8))$ with $\mathbf{v}_i \sim \mathcal{N}(0, \mathbf{I}_{d_n-1})$. $m_a(\cdot)$'s are linear

$$m_1(Z_i) = m_0(Z_i) = Z_{1i} + \sum_{j=2}^{d_n} Z_{ji}/\sqrt{k_n - 1},$$

and

$$\sigma_0(Z_i) = \sigma_1(Z_i) = c_\varepsilon \left[1 + \left(Z_{1i} + \sum_{j=2}^{d_n} Z_{ji}/\sqrt{k_n - 1}\right)^2\right]^{1/2}.$$

$(\varepsilon_{1,i}, \varepsilon_{0,i}) \sim \mathcal{N}(0, \mathbf{I}_2)$ is independent of Z_i.

For regressors, we may choose $X_i = Z_i$ or include first a few terms in Z_i.

Model 2 Same as model 1 but Z_{ji}'s are $U(-1, 1)$.

In each case, strata are determined by dividing the support of Z_{1i} into $|\mathcal{S}|$ intervals of equal length. For our models 1-2, when $Z_{1i} \sim U[-1, 1]$, we have $S_i = \sum_{j=1}^{|\mathcal{S}|} 1\{Z_{1i} \leq g_j\}$, where $g_j = 2j/|\mathcal{S}| - 1$. For all strata, we set $\pi_s = 1/2$. The treatment status is determined according to one of the following four covariate-adaptive randomization schemes:

1. SRS: Treatment assignment is generated as in Example 1.
2. BCD: Treatment assignment is generated as in Example 2 with $\lambda = 0.75$.
3. WEI: Treatment assignment is generated as in Example 3 with $\phi(x) = (1 - x)/2$.
4. SBR: Treatment assignment is generated as in Example 4.

5.1 Results

Rejection probabilities are computed using $10,000$ replications. The methods considered are as follows.

$\hat{\tau}^{adj}$: The fully saturated regression adjusted estimator introduced in Section 2. The variance estimator is $\hat{\Sigma}_{1,1}$ as in Theorem 3.2.

$\hat{\tau}^{*}$: The efficient estimator introduced at the end of Section 2. The inference is based on Theorem 3.3.

YYS: The inference method introduced by Ye et al. (2022). The estimator for τ is exactly the same as $\hat{\tau}^*$. However, the variance estimator does not take into account the issue of many covariates.

BCS: The method introduced by Bugni et al. (2019). The estimator for τ is exactly the unadjusted IPW estimator $\hat{\tau}^u$.

Tables 1 reports the results under the null hypothesis $\mu_1 = \mu_0 = 0$, for $n = 400$ and 800. The number of strata is set to $S = 2$ for all cases. We include all elements of Z_i, i.e., $X_i = Z_i$. On average, each regression in (2.1) has the effective sample size $n_{a,s} \approx n/(2S)$. Therefore, the number of covariates included in each regression is approximately **40**% of the effective sample size (i.e., $\kappa_{a,s} = 0.4$). Both $\hat{\tau}^{adj}$ and $\hat{\tau}^*$ performed well in this many covariates scenario. We see that they both have rejection probability close to the nominal level under four randomization schemes. From $n = 400$ to $n = 800$, the size control for our methods has remarkable improvement. The method

Table 1: Rejection rate (in percent) under H_0

		n=800, $k_n = 80$				n=400, $k_n = 40$			
model	method	SRS	BCD	WEI	SBR	SRS	BCD	WEI	SBR
1	$\hat{\tau}^{adj}$	5.43	5.32	5.46	5.89	5.70	5.79	5.60	5.90
1	$\hat{\tau}^{*}$	5.38	5.41	5.29	5.51	5.56	6.01	5.76	5.78
1	YYS	11.73	11.53	11.53	12.27	12.05	12.26	12.59	12.18
1	BCS	5.12	5.28	5.19	5.05	4.94	5.42	5.19	5.03
2	$\hat{\tau}^{adj}$	5.08	5.72	5.32	5.47	5.36	5.41	5.28	5.44
2	$\hat{\tau}^{*}$	5.10	5.35	5.33	5.27	5.39	5.28	4.95	5.75
2	YYS	11.60	11.65	11.34	11.98	11.59	11.57	11.55	11.53
2	BCS	4.99	4.93	5.08	5.16	4.92	5.02	5.12	5.47

Remark: Under H_0, we set $\mu_1 - \mu_0 = 0$.

Table 2: Rejection rate (in percent) under $H_1 : \delta = 0.2$

		n=800, $k_n = 80$				n=400, $k_n = 40$			
model	method	SRS	BCD	WEI	SBR	SRS	BCD	WEI	SBR
1	$\hat{\tau}^{adj}$	68.45	69.12	68.29	68.82	41.41	41.52	41.86	41.36
1	$\hat{\tau}^{*}$	75.58	75.56	75.55	75.41	46.71	47.20	47.72	47.08
1	YYS	80.33	80.85	80.70	80.40	55.64	55.90	56.72	55.40
1	BCS	71.59	71.67	71.38	71.88	42.68	43.76	43.41	42.53
2	$\hat{\tau}^{adj}$	68.28	69.03	68.68	68.69	41.31	41.69	41.82	42.05
2	$\hat{\tau}^{*}$	73.35	73.92	73.81	73.97	45.77	45.60	46.05	45.42
2	YYS	79.86	80.35	80.72	79.99	55.79	55.58	55.13	55.70
2	BCS	64.78	66.35	65.19	65.22	38.90	38.68	38.61	38.97

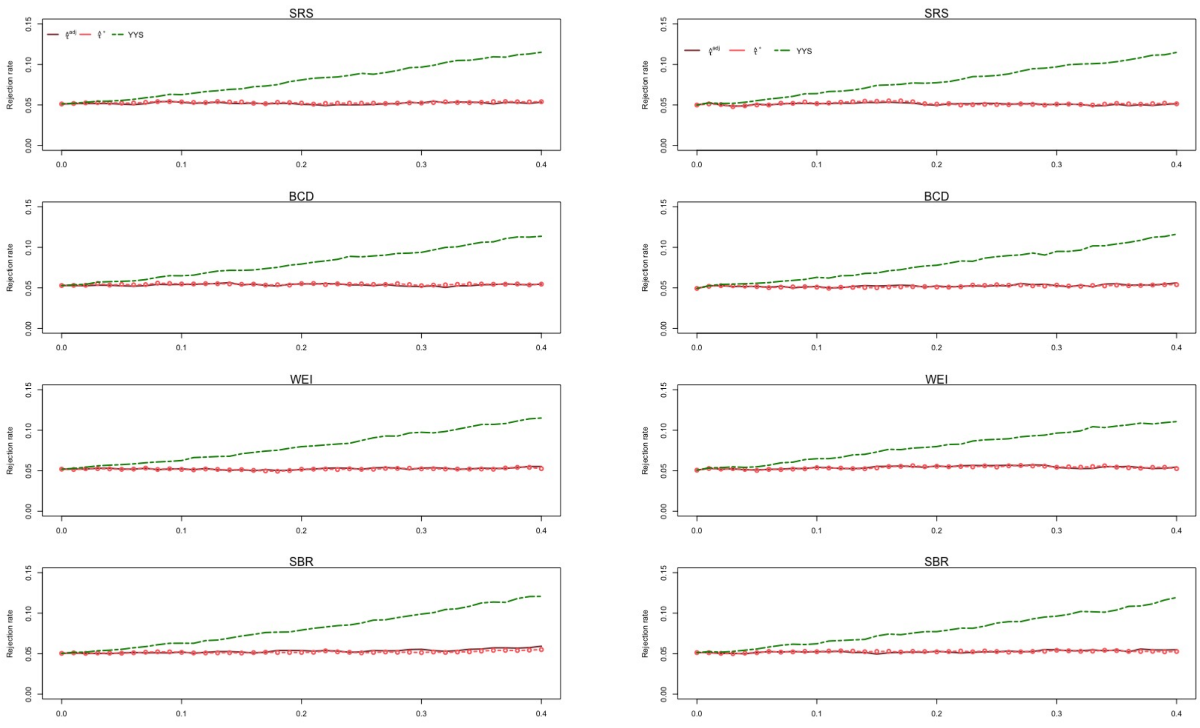
Remark: Under H_1, we set $\mu_1 - \mu_0 = 0.2$.

by YYS has rejection probabilities over 10% for all cases. BCS does not include any regressors and has a rejection probability close to the nominal level.

Table 2 report the results under the alternative hypothesis $\mu_1 - \mu_0 = 0.2$, for $n = 800$ and 400. The power of $\hat{\tau}^{adj}$ is slightly smaller than BCS for model 1. This observation reflects that the "no-harm" regression adjustment can actually degrade estimation efficiency because the cost of including many regressors outweights the benefit. The power of $\hat{\tau}^{*}$, on the other hand, always dominates $\hat{\tau}^{adj}$ and BCS, which is consistent with out theory. YYS has the highest power but its variance estimator is incorrect in this case.

Figures 2 and 3 display the relationship between the numbers of included regressors and the rejection probabilities of $\hat{\tau}^{adj}$, $\hat{\tau}^{*}$, and YYS under the null and alternative hypotheses, respectively. For both figures, we generate observations in models 1 and 2 using 80 regressors. We set $n = 800$ and the number of included regressors $k_n = 0, 2, \cdots, 80$, corresponding to $\kappa_{a,s} = 0, 0.02, \cdots, 0.4$. The x-axis represents $\kappa_{a,s}$. Figure 2 shows our inference methods (i.e., $\hat{\tau}^{adj}$ and $\hat{\tau}^{*}$) have uniform size control over $\kappa_{a,s}$, which is consistent with Theorem 4.2. On the other hand, we see that the rejection probabilities of YYS grow linearly in $\kappa_{a.s}$.

Figure 3 shows the power of the four inference methods under the alternative hypothesis $\mu_1 - \mu_0 = 0.2$. BCS' power, which is our benchmark, does not vary with $\kappa_{a,s}$ because it does not use any regressors. For both models, $\hat{\tau}^*$ is always more powerful than BCS, and, in most of the time, more powerful than $\hat{\tau}^{adj}$, which is consistent with our theory. As the number of regressors increases, the misspecification error decreases, and thus, $\hat{\tau}^*$ becomes more powerful. On the other hand, for model 1, the power of $\hat{\tau}^{adj}$ eventually drop below the benchmark of the unadjusted estimator as k_n increases. This reflects that including new regressors does not always increase the estimation accuracy in many covariates scenarios.



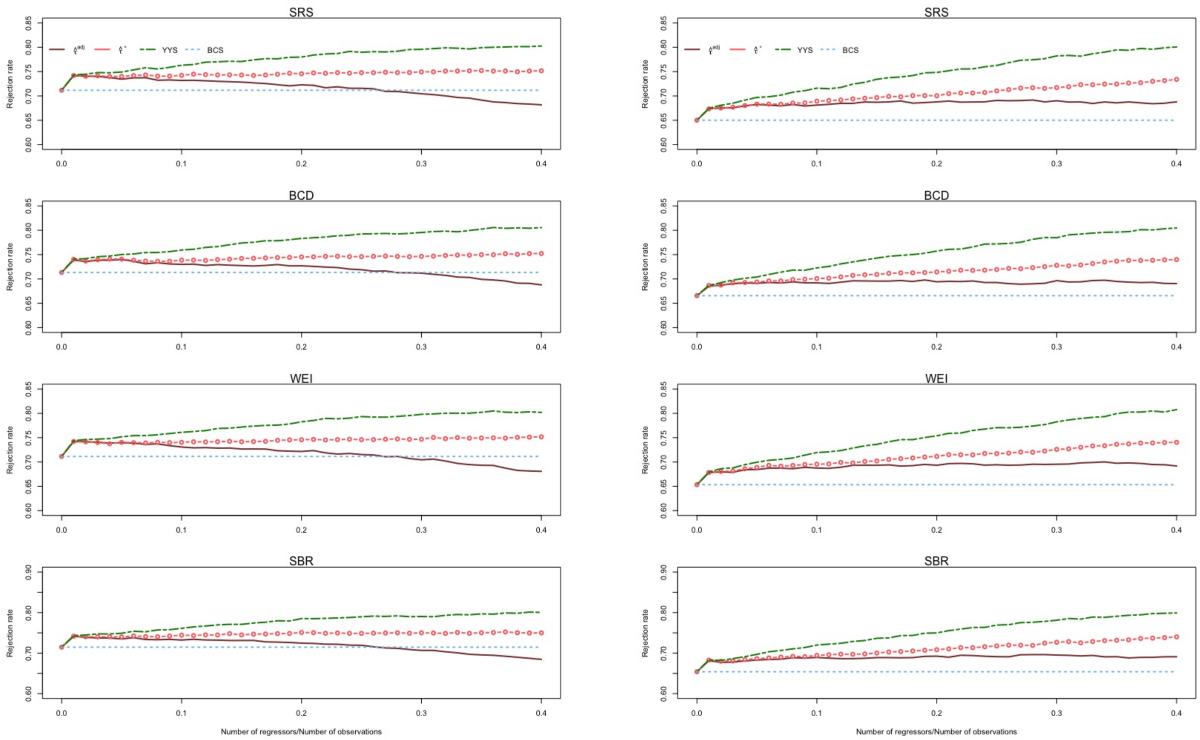
(a) Model 1

(b) Model 2

Figure 2: Rejection Probabilities under H_0

6 Conclusion

In this paper, we show that the estimation error of the "no-harm" regression adjustments can contaminate the ATE estimator and degrade estimation efficiency when the number of regressors is of the same order of the sample size. We then propose a new estimator for the ATE which is guaranteed to be more efficient than the unadjusted estimator. Last, we propose a consistent estimator of its asymptotic variance, construct the corresponding Wald-statistic, and show that it



(a) Model 1

(b) Model 2

Figure 3: Rejection Probabilities under H_1

has a uniform size control over two asymptotic regimes: (1) the dimension of the regressors is fixed and the regression adjustments are arbitrarily misspecified and (2) the dimension of the regressors is diverging at most at the rate of the sample size and the regression adjustments are approximately correctly specified.

A Verifying Assumption 3 for Gaussian Covariates

For $i \in \aleph_s$, we suppose $X_i = \Sigma_{s,n}^{1/2} Z_i$ where $Z_i = (Z_{i,1}, \cdots, Z_{i,k_n})$ are k_n independent standard normal random variables and the covariance matrix $\Sigma_{s,n} \in \Re^{k_n \times k_n}$ is symmetric and positive definite. The assumption that Z_i has mean zero is without loss of generality because the means will be cancelled in the definition of $\breve{X}_i$. By an abuse of notation, we assume $\aleph_{1,s} = \{1, \cdots, n_{1,s}\}$ and $\aleph_{0,s} = \{1 + n_{1,s}, \cdots, n_s\}$ and denote $Z_{\aleph_{a,s}} = (Z_1, \cdots, Z_{n_{1,s}})^\top \in \Re^{n_{1,s} \times k_n}$ and $Z_{\aleph_s} = (Z_1, \cdots, Z_{n_s})^\top \in \Re^{n_s \times k_n}$. Then, we have

$$\breve{X}_{\aleph_{1,s}} = X_{\aleph_{1,s}} - 1_{n_{1,s}} 1_{n_s}^\top X_{\aleph_s}/n_s \stackrel{d}{=} (Z_{\aleph_{1,s}} - 1_{n_{a,s}} 1_{n_s}^\top Z_{\aleph_s}/n_s)\Sigma_{s,n}^{1/2} = \Gamma_n Z_{\aleph_s} \Sigma_{s,n}^{1/2}$$

and

$$M_{a,s} = \breve{X}_{\aleph_{a,s}} \left[\breve{X}_{\aleph_{a,s}}^\top \breve{X}_{\aleph_{a,s}}\right]^{-1} \breve{X}_{\aleph_{a,s}}^\top \overset{d}{=} \Gamma_n Z_{\aleph_s} \left[Z_{\aleph_s}^\top \Gamma_n^\top \Gamma_n Z_{\aleph_s}\right]^{-1} Z_{\aleph_s}^\top \Gamma_n^\top,$$

where $\Gamma_n = (I_{n_{a,s}}, 0_{n_{a,s}, n_{1-a,s}}) - 1_{n_{a,s}} 1_{n_s}^\top / n_s$. We focus on the case with $a = 1$. For the l-th column of $\Gamma_n Z_{\aleph_s}$ denoted as $\Gamma_n Z_{\aleph_s, l}$, we have

$$\Gamma_n Z_{\aleph_s, l} \overset{d}{=} W_n^{1/2} Z_{\aleph_{1,s}, l},$$

where $Z_{\aleph_{1,s}, l}$ is an $n_{1,s} \times 1$ standard normal vector and W_n is a $n_{1,s} \times n_{1,s}$ matrix with $1 - 1/n_s$ and $-1/n_s$ being the diagonal and off-diagonal elements, respectively. Then, we have

$$\begin{aligned}
1 - \gamma_{1,s,n} &\overset{d}{=} \frac{1_{n_{1,s}}^\top \Gamma_n Z_{\aleph_s}}{n_{1,s}} \left[\frac{Z_{\aleph_s}^\top \Gamma_n^\top \Gamma_n Z_{\aleph_s}}{n_{1,s}}\right]^{-1} \frac{Z_{\aleph_s}^\top \Gamma_n^\top 1_{n_{1,s}}}{n_{1,s}} \\
&\overset{d}{=} \frac{1_{n_{1,s}}^\top W_n^{1/2} Z_{\aleph_{1,s}}}{n_{1,s}} \left[\frac{Z_{\aleph_{1,s}}^\top W_n Z_{\aleph_{1,s}}}{n_{1,s}}\right]^{-1} \frac{Z_{\aleph_{1,s}}^\top W_n^{1/2} 1_{n_{1,s}}}{n_{1,s}} \\
&\overset{d}{=} \frac{1_{n_{1,s}}^\top U_n D_n^{1/2} Z_{\aleph_{1,s}}}{n_{1,s}} \left[\frac{Z_{\aleph_{1,s}}^\top D_n Z_{\aleph_{1,s}}}{n_{1,s}}\right]^{-1} \frac{Z_{\aleph_{1,s}}^\top D_n^{1/2} U_n^\top 1_{n_{1,s}}}{n_{1,s}} \\
&\overset{d}{=} \frac{(1 - u_{1,s}/n_s)}{n_{1,s}} Z_{n_{1,s}}^\top \left[\frac{1}{n_{1,s}} \sum_{i \in [n_{1,s}-1]} Z_i Z_i^\top + \left(\frac{1}{n_{1,s}} - \frac{1}{n_s}\right) Z_{n_{1,s}} Z_{n_{1,s}}^\top\right]^{-1} Z_{n_{1,s}} \\
&\overset{d}{=} \frac{(1 - u_{1,s}/n_s)}{n_{1,s}} Z_{n_{1,s}}^\top \left[\mathcal{A}_n^{-1} - \frac{\mathcal{A}_n^{-1}\left(\left(\frac{1}{n_{1,s}} - \frac{1}{n_s}\right) Z_{n_{1,s}} Z_{n_{1,s}}^\top\right)\mathcal{A}_n^{-1}}{1 + \left(\frac{1}{n_{1,s}} - \frac{1}{n_s}\right) Z_{n_{1,s}}^\top \mathcal{A}_n^{-1} Z_{n_{1,s}}}\right] Z_{n_{1,s}},
\end{aligned}$$

where $W_n = U_n D_n U_n^\top$ is the eigenvalue decomposition of W_n, $D_n = \text{diag}(1, \cdots, 1, 1 - n_{1,s}/n_s)$ is the eigenvalues, U_n is the corresponding eigenvectors, $\mathcal{A}_n = \frac{1}{n_{1,s}} \sum_{i \in [n_{1,s}-1]} Z_i Z_i^\top$, the fourth equality is by $1_{n_{1,s}}^\top U_n = (0, \cdots, 0, \sqrt{n_{1,s}})$, and the last equality follows the Sherman-Morrison formula. Consider the eigenvalue decomposition of $\frac{n_{1,s}}{n_{1,s}-1} \mathcal{A}_n = \mathcal{U}_n \Lambda_n \mathcal{U}_n^\top$ where $\Lambda_n = \text{diag}(\lambda_1, \cdots, \lambda_{k_n})$. We note that $Z_{n_{1,s}}$ and $\mathcal{A}_n$, and thus, $\mathcal{U}_n$ and Λ_n, are independent by construction. Then, for any bounded and smooth function $f : \Re^{k_n} \mapsto \Re$, we have

$$\mathbb{E}(f(\mathcal{U}_n^\top Z_{n_{1,s}} \mid \Lambda_n) = \mathbb{E}[\mathbb{E}(f(\mathcal{U}_n^\top Z_{n_{1,s}} \mid \mathcal{U}_n, \Lambda_n) \mid \Lambda_n] = \mathbb{E} f(Z_{n_{1,s}}) = \mathbb{E} f(\mathcal{U}_n^\top Z_{n_{1,s}}),$$

where we use the fact that the distribution of $Z_{n_{1,s}}$ ($k_n \times 1$ standard normal random vector) is invariant to rotations. This implies $\mathcal{U}_n^\top Z_{n_{1,s}}$ and Λ_n are independent. Denote $\mathcal{U}_n^\top Z_{n_{1,s}}$ as $\mathcal{G} =$

$(g_1, \cdots, g_{K_n})$ which are independent of Λ_n and suppose $k_n/n_{1,s} \to \kappa_{1,s}$. Then, we have

$$\begin{aligned}
\frac{1}{n_{1,s}} Z_{n_{1,s}}^\top \left(\frac{n_{1,s}}{n_{1,s}-1} \mathcal{A}_n \right)^{-1} Z_{n_{1,s}} &\overset{d}{=} \frac{1}{n_{1,s}} \sum_{l=1}^{k_n} \lambda_l^{-1} g_l^2 \\
&= \frac{1}{n_{1,s}} \sum_{l=1}^{k_n} \lambda_l^{-1} + \frac{1}{n_{1,s}} \sum_{l=1}^{k_n} \lambda_l^{-1} (g_l^2 - 1) \\
&= \frac{k_n}{n_{1,s}} \frac{1}{k_n} \sum_{l=1}^{k_n} \lambda_l^{-1} + o_P(1) \\
&\xrightarrow{p} \int_{\lambda_-}^{\lambda_+} \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{2\pi\lambda^2} d\lambda,
\end{aligned}$$

where $\lambda_\pm = (1 \pm \sqrt{\kappa_{1,s}})^2$, the third equality holds because $\lambda_{k_n} \xrightarrow{p} 1 - \sqrt{\kappa_{1,s}}$ which is bounded away from zero,

$$\mathbb{E}\left(\frac{1}{n_{1,s}} \sum_{l=1}^{k_n} \lambda_l^{-1} (g_l^2 - 1) \mid \Lambda_n \right) = 0,$$

and

$$Var\left(\frac{1}{n_{1,s}} \sum_{l=1}^{k_n} \lambda_l^{-1} (g_l^2 - 1) \mid \Lambda_n \right) \le \frac{k_n}{n_{1,s}^2} \lambda_{k_n}^2 \to 0,$$

and the last equality holds because by Marčenko-Pastur theorem (see, for example, Bai (2008)). Denote $\int_{\lambda_-}^{\lambda_+} \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{2\pi\lambda^2} d\lambda$ as $\zeta_{1,s}$, then we have

$$\gamma_{1,s,n} \xrightarrow{p} \frac{1}{1 + (1 - \pi_s)\zeta_{1,s}} \equiv \gamma_{1,s,\infty}.$$

B Proof of Theorem 3.1

Denote $\hat{\tau}_s = \hat{\tau}_{1,s} - \hat{\tau}_{0,s}$, $\tau_{a,s} = \mathbb{E}(Y(a)|S = s)$ for $a = 0, 1$, and $\tau_s = \mathbb{E}(Y(1) - Y(0)|S = s)$. Note that

$$\begin{aligned}
Y_i(1) &= \alpha_{1,S_i} + X_i^\top \beta_{1,S_i} + e_{i,S_i}(1) + \varepsilon_i(1) \\
&= \alpha_{1,S_i} + \overline{X}_{S_i}^\top \beta_{1,S_i} + \breve{X}_i^\top \beta_{1,S_i} + e_{i,S_i}(1) + \varepsilon_i(1),
\end{aligned}$$

which implies

$$\hat{\tau}_{1,s} = (1_{n_{1,s}}^\top M_{1,s} 1_{n_{1,s}})^{-1} (1_{n_{1,s}}^\top M_{1,s} Y_{\aleph_{1,s}}(1))$$

$$
\begin{aligned}
&= \alpha_{1,s} + \overline{X}_s^\top \beta_{1,s} + (1_{n_{1,s}}^\top M_{1,s} 1_{n_{1,s}})^{-1}(1_{n_{1,s}}^\top M_{1,s}(e_{\aleph_{1,s}}(1) + \varepsilon_{\aleph_{1,s}}(1))) \\
&= \frac{1}{n_s} \sum_{i \in [n]} 1\{S_i = s\} \left(\mathbb{E}(Y_i(1)|X_i, S_i = s) - e_{i,s}(1)\right) \\
&+ (1_{n_{1,s}}^\top M_{1,s} 1_{n_{1,s}})^{-1}(1_{n_{1,s}}^\top M_{1,s}(e_{\aleph_{1,s}}(1) + \varepsilon_{\aleph_{1,s}}(1))).
\end{aligned}
$$

By Assumption 2(iii), we have

$$
\left| \frac{1}{n_s} \sum_{i \in [n]} 1\{S_i = s\} e_{i,s}(1) \right| \leq \left(\frac{1}{n_s} \sum_{i \in [n]} 1\{S_i = s\} e_{i,s}^2(1) \right)^{1/2} = o_P(n^{-1/2})
$$

and

$$
\begin{aligned}
\left| n_{1,s}^{-1}(1_{n_{1,s}}^\top M_{1,s} e_{\aleph_{1,s}}(1)) \right| &\leq n_{1,s}^{-1} ||M_{1,s}||_{op} ||1_{n_{1,s}}||_2 ||e_{\aleph_{1,s}}(1)||_2 \\
&\leq \left(\frac{1}{n_s} \sum_{i \in [n]} 1\{S_i = s\} e_{i,s}^2(1) \right)^{1/2} = o_P(n^{-1/2}).
\end{aligned}
$$

This implies the linear expansion

$$
\begin{aligned}
\hat{\tau}_{1,s} - \tau_{1,s} &= \frac{1}{n_s} \sum_{i \in [n]} 1\{S_i = s\} \left(\mathbb{E}(Y_i(1)|X_i, S_i = s) - \mathbb{E}(Y_i(1)|S_i = s)\right) \\
&\quad + \gamma_{1,s,n}^{-1} n_{1,s}^{-1} (1_{n_{1,s}}^\top M_{1,s} \varepsilon_{\aleph_{1,s}}(1)) + o_P(n^{-1/2}).
\end{aligned}
$$

Similarly, we have

$$
\begin{aligned}
\hat{\tau}_{0,s} - \tau_{0,s} &= \frac{1}{n_s} \sum_{i \in [n]} 1\{S_i = s\} \left(\mathbb{E}(Y_i(0)|X_i, S_i = s) - \mathbb{E}(Y_i(0)|S_i = s)\right) \\
&\quad + \gamma_{0,s,n}^{-1} n_{0,s}^{-1} (1_{n_{0,s}}^\top M_{0,s} \varepsilon_{\aleph_{0,s}}(0)) + o_P(n^{-1/2}),
\end{aligned}
$$

and thus

$$
\begin{aligned}
\hat{\tau}_s - \tau_s &= \frac{1}{n_s} \sum_{i \in [n]} 1\{S_i = s\} \left[\mathbb{E}(Y_i(1)|X_i, S_i = s) - \mathbb{E}(Y_i(0)|X_i, S_i = s) - \tau_s\right] \\
&\quad + \gamma_{1,s,n}^{-1} n_{1,s}^{-1} (1_{n_{1,s}}^\top M_{1,s} \varepsilon_{\aleph_{1,s}}(1)) - \gamma_{0,s,n}^{-1} n_{0,s}^{-1} (1_{n_{0,s}}^\top M_{0,s} \varepsilon_{\aleph_{0,s}}(0)) + o_P(n^{-1/2}).
\end{aligned}
$$

Therefore, we have

$$
\hat{\tau}^{adj} - \tau = \sum_{s \in \mathcal{S}} \hat{p}_s \left[\tilde{\gamma}_{1,s}^{-1} n_{1,s}^{-1} (1_{n_{1,s}}^\top M_{1,s} \varepsilon_{\aleph_{1,s}}(1)) - \tilde{\gamma}_{1,s}^{-1} n_{0,s}^{-1} (1_{n_{0,s}}^\top M_{0,s} \varepsilon_{\aleph_{0,s}}(0)) \right]
$$

$$+\sum_{s\in\mathcal{S}}\frac{\hat{p}_s}{n_s}\sum_{i\in[n]}1\{S_i=s\}\left[\mathbb{E}(Y_i(1)|X_i,S_i=s)-\mathbb{E}(Y_i(1)|S_i=s)\right]$$
$$-\sum_{s\in\mathcal{S}}\frac{\hat{p}_s}{n_s}\sum_{i\in[n]}1\{S_i=s\}\left[\mathbb{E}(Y_i(0)|X_i,S_i=s)-\mathbb{E}(Y_i(0)|S_i=s)\right]$$
$$+\frac{1}{n}\sum_{i\in[n]}\left[\mathbb{E}(Y_i(1)|S_i)-\mathbb{E}(Y_i(0)|S_i)-\tau\right]+o_P(n^{-1/2}).$$

In addition, we have

$$\begin{aligned}\hat{\tau}^{unadj}-\tau=&\sum_{s\in\mathcal{S}}\frac{\hat{p}_s}{n_{1,s}}\sum_{i\in[n]}A_i1\{S_i=s\}\varepsilon_i(1)-\sum_{s\in\mathcal{S}}\frac{\hat{p}_s}{n_{0,s}}\sum_{i\in[n]}(1-A_i)1\{S_i=s\}\varepsilon_i(0)\\&+\sum_{s\in\mathcal{S}}\frac{\hat{p}_s}{n_{1,s}}\sum_{i\in[n]}A_i1\{S_i=s\}(\mathbb{E}(Y_i(1)|X_i,S_i=s)-\mathbb{E}(Y_i(1)|S_i=s))\\&-\sum_{s\in\mathcal{S}}\frac{\hat{p}_s}{n_{0,s}}\sum_{i\in[n]}(1-A_i)1\{S_i=s\}(\mathbb{E}(Y_i(0)|X_i,S_i=s)-\mathbb{E}(Y_i(0)|S_i=s))\\&+\frac{1}{n}\sum_{i\in[n]}(\mathbb{E}(Y_i(1)|S_i)-\mathbb{E}(Y_i(0)|S_i)-\tau)\end{aligned}$$

Therefore, we have

$$\sqrt{n}\begin{pmatrix}\hat{\tau}^{adj}-\tau\\\hat{\tau}^{unadj}-\tau\end{pmatrix}=\mathcal{U}_n+\mathcal{V}_n+\begin{pmatrix}\mathcal{W}_n\\\mathcal{W}_n\end{pmatrix}+o_P(1),$$

where

$$\mathcal{U}_n=\sum_{s\in\mathcal{S}}\hat{p}_s\sqrt{n}\begin{pmatrix}\gamma_{1,s,n}^{-1}n_{1,s}^{-1}(1_{n_{1,s}}^\top M_{1,s}\varepsilon_{\aleph_{1,s}}(1))-\gamma_{0,s,n}^{-1}n_{0,s}^{-1}(1_{n_{0,s}}^\top M_{0,s}\varepsilon_{\aleph_{0,s}}(0))\\n_{1,s}^{-1}(1_{n_{1,s}}^\top\varepsilon_{\aleph_{1,s}}(1))-n_{0,s}^{-1}(1_{n_{0,s}}^\top\varepsilon_{\aleph_{0,s}}(0))\end{pmatrix}, \tag{B.1}$$

$$\mathcal{V}_n=\sum_{s\in\mathcal{S}}\hat{p}_s\sqrt{n}\begin{pmatrix}\frac{1}{n_s}\sum_{i\in[n]}1\{S_i=s\}\left(\phi_i(1)-\phi_i(0)\right)\\\frac{1}{n_{1,s}}\sum_{i\in\aleph_{1,s}}\phi_i(1)-\frac{1}{n_{0,s}}\sum_{i\in\aleph_{0,s}}\phi_i(0)\end{pmatrix}, \tag{B.2}$$

$$\mathcal{W}_n=\frac{1}{\sqrt{n}}\sum_{i\in[n]}(\mathbb{E}(Y_i(1)|S_i)-\mathbb{E}(Y_i(0)|S_i)-\tau), \tag{B.3}$$

and $\phi_i(a)=\mathbb{E}(Y_i(a)|X_i,S_i)-\mathbb{E}(Y_i(a)|S_i)$. Then, the desired result holds by Lemma G.1.

C Proof of Theorem 3.2

We derive the limits of $\hat{\Sigma}_{\mathcal{U}}$, $\hat{\Sigma}_{\mathcal{V}}$, and $\hat{\Sigma}_{\mathcal{W}}$ in the following three steps.

Step 1: Limit of $\hat{\Sigma}_{\mathcal{U}}$. Following Bugni et al. (2018), we define $\{(X_i^s, \varepsilon_i^s(1), \varepsilon_i^s(0)) : 1 \le i \le n\}$ as a sequence of i.i.d. random variables with marginal distributions equal to the distribution of $(X_i, \varepsilon_i(1), \varepsilon_i(0))|S_i = s$ and $N_s = \sum_{i=1}^n 1\{S_i < s\}$. Then, we have $Y_i(a) \mid S_i = s \overset{d}{=} Y_i^s(a)$ and $\tilde{e}_i(a) \mid S_i = s \overset{d}{=} e_{i,s}(a)$ where

$$Y_i^s(a) = \alpha_{a,s} + (X_i^s)^\top \beta_{a,s} + \tilde{e}_i^s(a) + \varepsilon_i^s(a) \quad \text{and} \quad \tilde{e}_i^s(a) = \mathbb{E}(Y_i(a)|X_i = x, S_i = s) - \alpha_{a,s} - (X_i^s)^\top \beta_{a,s}.$$

We further define $\tilde{M}_{a,s,i,j}$ as the (i,j)th entry of the $n_{a,s} \times n_{a,s}$ matrix, $\tilde{M}_{a,s} = \Xi_{a,s}(\Xi_{a,s}^\top \Xi_{a,s})^{-1}\Xi_{a,s}^\top$, $\Xi_{1,s} = ((\breve{X}_{N_s+1}^s)^\top, \cdots, (\breve{X}_{N_s+n_{1,s}}^s)^\top)^\top$ is an $n_{1,s} \times k_n$ matrix, $\Xi_{0,s} = ((\breve{X}_{N_s+1+n_{1,s}}^s)^\top, \cdots, (\breve{X}_{N_s+n_s}^s)^\top)^\top$ is an $n_{0,s} \times k_n$ matrix,

$$\breve{X}_i^s = X_i^s - \frac{1}{n_s} \sum_{j=N_s+1}^{N_s+n_s} X_j^s, \quad \text{if} \quad N_s + 1 \le i \le N_s + n_s.$$

$\tilde{\gamma}_{a,s,n} = 1_{n_{a,s}}^\top \tilde{M}_{a,s} 1_{n_{a,s}}$. and $\phi_i^s(a) = \mathbb{E}(Y_i(a)|X_i, S_i = s) - \mathbb{E}(Y_i(a)|S_i = s)$.

Step 1.1: Limit of $\hat{\omega}_{a,s}^2$. We consider the case with $a = 1$. The result for $a = 0$ can be established in the same manner. Given the above definitions, we see that, conditionally on $(A^{(n)}, S^{(n)})$

$$\hat{\omega}_{1,s}^2 \overset{d}{=} \frac{1}{n_{1,s}} \tilde{\gamma}_{1,s,n}^{-2} \sum_{i=N_s+1}^{N_s+n_{1,s}} \left(\sum_{j=N_s+1}^{N_s+n_{1,s}} \tilde{M}_{1,s,i,j} \right)^2 Y_i^s(1) \tilde{\tilde{\varepsilon}}_{1,s,i},$$

where $\tilde{\tilde{\varepsilon}}_{1,s,i} = \tilde{\varepsilon}_{1,s,i}/\tilde{M}_{1,s,i,i}$ and $\tilde{\varepsilon}_{1,s,i}$ is the residual from the linear regression of $Y_i^s(1)$ on $(1, \breve{X}_i^s)$ with observations $N_s + 1 \le i \le N_s + n_{1,s}$. Because $\{(X_i^s, Y_i^s(1)) : 1 \le i \le n\}$ are i.i.d. and independent of $(A^{(n)}, S^{(n)})$ we can directly apply Jochmans (2022, Theorem 1) and conclude that

$$\hat{\omega}_{1,s}^2 - \tilde{\gamma}_{1,s,n}^{-2} \sigma_{1,s,n}^2 = o_P(1).$$

Then, because $\tilde{\gamma}_{a,s,n} \overset{d}{=} \gamma_{a,s,n}$ and Assumption 3, we have

$$\hat{\omega}_{1,s}^2 \xrightarrow{p} \omega_{1,s,\infty}^2.$$

Step 1.2: Limit of $\hat{\varpi}_{a,s}$. Again, we focus on $\hat{\varpi}_{1,s}$. Following the argument in **Step 1.1**, we have

$$\hat{\varpi}_{1,s} \stackrel{d}{=} \frac{1}{n_{1,s}} \tilde{\gamma}_{1,s,n}^{-1} \sum_{i=N_s+1}^{N_s+n_{1,s}} \left(\sum_{j=N_s+1}^{N_s+n_{1,s}} \tilde{M}_{1,s,i,j} \right) Y_i^s(1) \tilde{\check{\varepsilon}}_{1,s,i}.$$

Note that

$$\begin{aligned}
\tilde{\check{\varepsilon}}_{1,s,i} &= \frac{\sum_{k=N_s+1}^{N_s+n_{1,s}} \tilde{M}_{1,s,i,k}(Y_k^s(1) - \tilde{\tau}_{1,s})}{\tilde{M}_{1,s,i,i}} \\
&= \frac{\sum_{k=N_s+1}^{N_s+n_{1,s}} M_{1,s,i,k}(\tilde{e}_k^s(1) + \alpha_{1,s} + \widetilde{X}_s^\top \beta_{1,s} - \tilde{\tau}_{1,s}) + \varepsilon_k^s(1)}{\tilde{M}_{1,s,i,i}},
\end{aligned}$$

where $\tilde{\tau}_{1,s}$ is the intercept of the OLS regression of $Y_i^s(1)$ on $(1, \breve{X}_i^s)$ using observations $N_s+1 \leq i \leq N_s+n_{1,s}$ and $\widetilde{X}_s = \frac{1}{n_s} \sum_{i=N_s+1}^{N_s+n_s} X_i^s$. By construction, we have $\tilde{\tau}_{1,s} \stackrel{d}{=} \hat{\tau}_{1,s}$. Let $\theta_i = \sum_{k=N_s+1}^{N_s+n_{1,s}} \tilde{M}_{1,s,i,k}$. Therefore, we have

$$\begin{aligned}
&\frac{1}{n_{1,s}} \tilde{\gamma}_{1,s,n}^{-1} \sum_{i=N_s+1}^{N_s+n_{1,s}} \left(\sum_{j=N_s+1}^{N_s+n_{1,s}} \tilde{M}_{1,s,i,j} \right) Y_i^s(1) \tilde{\check{\varepsilon}}_{1,s,i} \\
&= \frac{1}{n_{1,s}} \tilde{\gamma}_{1,s,n}^{-1} \sum_{i=N_s+1}^{N_s+n_{1,s}} \sum_{k=N_s+1}^{N_s+n_{1,s}} \frac{\theta_i Y_i^s(1) M_{1,s,i,k} \varepsilon_k^s(1)}{\tilde{M}_{1,s,i,i}} + \frac{1}{n_{1,s}} \tilde{\gamma}_{1,s,n}^{-1} \sum_{i=N_s+1}^{N_s+n_{1,s}} \sum_{k=N_s+1}^{N_s+n_{1,s}} \frac{\theta_i Y_i^s(1) M_{1,s,i,k} \tilde{e}_k^s(1)}{\tilde{M}_{1,s,i,i}} \\
&+ \frac{1}{n_{1,s}} \tilde{\gamma}_{1,s,n}^{-1} \sum_{i=N_s+1}^{N_s+n_{1,s}} \frac{\theta_i^2 Y_i^s(1)}{\tilde{M}_{1,s,i,i}} (\alpha_{1,s} + \widetilde{X}_s^\top \beta_{1,s} - \tilde{\tau}_{1,s})) \\
&\equiv T_1 + T_2 + T_3.
\end{aligned}$$

For T_2, we have

$$\begin{aligned}
|T_2| &\leq \tilde{\gamma}_{1,s,n}^{-1} \left[\frac{1}{n_{1,s}} \sum_{i=N_s+1}^{N_s+n_{1,s}} \frac{\theta_i^2 (Y_i^s(1))^2}{\tilde{M}_{1,s,i,i}^2} \right]^{1/2} \left[\frac{1}{n_{1,s}} \sum_{i=N_s+1}^{N_s+n_{1,s}} (\tilde{e}_i^s(1))^2 \right]^{1/2} \\
&\leq \gamma_{1,s,n}^{-1} \frac{\max_{N_s+1 \leq i \leq N_s+n_{1,s}} |\theta_i|}{\min_{N_s+1 \leq i \leq N_s+n_{1,s}} \tilde{M}_{1,s,i,i}} \left[\frac{1}{n_{1,s}} \sum_{i=N_s+1}^{N_s+n_{1,s}} (Y_i^s(1))^2 \right]^{1/2} \left[\frac{1}{n_{1,s}} \sum_{i=N_s+1}^{N_s+n_{1,s}} (\tilde{e}_i^s(1))^2 \right]^{1/2} \\
&= o_P(1),
\end{aligned}$$

where the last equality holds because by Assumption 2

$$\min_{N_s+1 \leq i \leq N_s+n_{1,s}} \tilde{M}_{1,s,i,i} \stackrel{d}{=} \min_{N_s+1 \leq i \in \aleph_{1,s}} M_{1,s,i,i} \geq \delta > 0,$$

$$\frac{1}{n_{1,s}}\sum_{i=N_s+1}^{N_s+n_{1,s}}(\tilde{e}_i^s(1))^2 \stackrel{d}{=} \frac{1}{n_{1,s}}\sum_{i\in\aleph_{1,s}} e_{i,s}^2(1) = o_P(n^{-1}), \quad \text{and}$$

$$\max_{N_s+1\le i\le N_s+n_{1,s}}|\theta_i| \stackrel{d}{=} \max_{i\in\aleph_{1,s}}|\sum_{j\in\aleph_{1,s}} M_{1,s,i,j}| = o_P(n^{1/2}).$$

For T_3, we have

$$\begin{aligned}
|T_3| &\le \tilde{\gamma}_{1,s,n}^{-1}\left[\frac{1}{n_{1,s}}\sum_{i=N_s+1}^{N_s+n_{1,s}}\frac{\theta_i^4}{\widetilde{M}_{1,s,i,i}^2}\right]^{1/2}\left[\frac{1}{n_{1,s}}\sum_{i=N_s+1}^{N_s+n_{1,s}}(Y_i^s(1))^2\right]^{1/2}|\alpha_{1,s}+\widetilde{X}_s^\top\beta_{1,s}-\tilde{\tau}_{1,s})| \\
&\le \gamma_{1,s,n}^{-1}\frac{\max_{N_s+1\le i\le N_s+n_{1,s}}|\theta_i|}{\min_{N_s+1\le i\le N_s+n_{1,s}}\tilde{M}_{1,s,i,i}}\left[\frac{1}{n_{1,s}}\sum_{i=N_s+1}^{N_s+n_{1,s}}\theta_i^2\right]^{1/2} \\
&\times\left[\frac{1}{n_{1,s}}\sum_{i=N_s+1}^{N_s+n_{1,s}}(Y_i^s(1))^2\right]^{1/2}|\alpha_{1,s}+\widetilde{X}_s^\top\beta_{1,s}-\tilde{\tau}_{1,s})| \\
&= o_P(1),
\end{aligned}$$

where the last equality holds because

$$\frac{1}{n_{1,s}}\sum_{i=N_s+1}^{N_s+n_{1,s}}\theta_i^2 = \frac{1}{n_{1,s}}\sum_{i=N_s+1}^{N_s+n_{1,s}}\sum_{j=N_s+1}^{N_s+n_{1,s}}\tilde{M}_{1,s,i,j} \le 1$$

and

$$\begin{aligned}
|\alpha_{1,s}+\widetilde{X}_s^\top\beta_{1,s}-\tilde{\tau}_{1,s})| \stackrel{d}{=} |\alpha_{1,s}+\overline{X}_s^\top\beta_{1,s}-\hat{\tau}_{1,s})| &= |(1_{n_{1,s}}^\top M_{1,s}1_{n_{1,s}})^{-1}(1_{n_{1,s}}^\top M_{1,s}\varepsilon_{\aleph_{1,s}}(1))| + o_P(n^{-1/2}) \\
&= O_P(n^{-1/2}).
\end{aligned}$$

Next, we focus on T_1. Denote $\mu_i^s(1) = \mathbb{E}(Y_i^s(1)|X_i^s) \stackrel{d}{=} \mathbb{E}(Y_i(1)|X_i, S_i = s)$. Then, for T_1, we have

$$\begin{aligned}
T_1 &= \frac{1}{n_{1,s}}\tilde{\gamma}_{1,s,n}^{-1}\sum_{i=N_s+1}^{N_s+n_{1,s}}\sum_{k=N_s+1}^{N_s+n_{1,s}}\frac{\theta_i\mu_i^s(1)M_{1,s,i,k}\varepsilon_k^s(1)}{\tilde{M}_{1,s,i,i}} \\
&+ \frac{1}{n_{1,s}}\tilde{\gamma}_{1,s,n}^{-1}\sum_{i=N_s+1}^{N_s+n_{1,s}}\sum_{N_s+1\le K\le N_s+n_{1,s},k\ne i}\frac{\theta_i\varepsilon_i^s(1)M_{1,s,i,k}\varepsilon_k^s(1)}{\widetilde{M}_{1,s,i,i}} \\
&+ \frac{1}{n_{1,s}}\tilde{\gamma}_{1,s,n}^{-1}\sum_{i=N_s+1}^{N_s+n_{1,s}}\theta_i[(\varepsilon_i^s(1))^2-\tilde{V}_{1,s,i}^2] + \frac{1}{n_{1,s}}\tilde{\gamma}_{1,s,n}^{-1}\sum_{i=N_s+1}^{N_s+n_{1,s}}\theta_i\tilde{V}_{1,s,i}^2 \\
&\equiv T_{1,1}+T_{1,2}+T_{1,3}+T_{1,4},
\end{aligned}$$

where $\tilde{V}^2_{1,s,i} = \mathbb{E}[(\varepsilon^s_i(1))^2 | X^s_i] \overset{d}{=} \mathbb{E}(\varepsilon^2_i(1)|X_i, S_i = s)$. Recall $\Xi_{1,s} = ((\breve{X}^s_{N_s+1})^\top, \cdots, (\breve{X}^s_{N_s+n_{1,s}})^\top)^\top$. We have $\mathbb{E}(T_{1,1} \mid \Xi_{1,s}) = 0$ and

$$\begin{aligned}
Var(T_{1,1} \mid \Xi_{1,s}) &= \frac{1}{n^2_{1,s}\tilde{\gamma}^2_{1,s,n}} \sum_{k=N_s+1}^{N_s+n_{1,s}} \tilde{V}^2_{1,s,k} \left[\sum_{i=N_s+1}^{N_s+n_{1,s}} \frac{\theta_i \mu^s_i(1) M_{1,s,i,k}}{\tilde{M}_{1,s,i,i}}\right]^2 \\
&\leq \frac{\max_{i\in[n]} \tilde{V}^2_{1,s,i}}{n^2_{1,s}\tilde{\gamma}^2_{1,s,n}} \sum_{k=N_s+1}^{N_s+n_{1,s}} \left[\sum_{i=N_s+1}^{N_s+n_{1,s}} \frac{\theta_i \mu^s_i(1) M_{1,s,i,k}}{\tilde{M}_{1,s,i,i}}\right] \left[\sum_{j=N_s+1}^{N_s+n_{1,s}} \frac{\tilde{v}_j \mu^s_j(1) M_{1,s,j,k}}{\tilde{M}_{1,s,j,j}}\right] \\
&= \frac{\max_{i\in[n]} \tilde{V}^2_{1,s,i}}{n^2_{1,s}\tilde{\gamma}^2_{1,s,n}} \sum_{i=N_s+1}^{N_s+n_{1,s}} \sum_{j=N_s+1}^{N_s+n_{1,s}} \left[\frac{\theta_i \mu^s_i(1) \tilde{v}_j \mu^s_j(1) M_{1,s,i,j}}{\tilde{M}_{1,s,i,i}\tilde{M}_{1,s,j,j}}\right] \\
&\leq \frac{\max_{i\in[n]} \tilde{V}^2_{1,s,i}}{n^2_{1,s}\tilde{\gamma}^2_{1,s,n}} \sum_{i=N_s+1}^{N_s+n_{1,s}} \left[\frac{\theta_i \mu^s_i(1)}{\tilde{M}_{1,s,i,i}}\right]^2 \\
&\leq \left[\frac{\max_{i\in[n]} \tilde{V}^2_{1,s,i}}{n_{1,s}\tilde{\gamma}^2_{1,s,n} \min_{N_s+1\leq i\leq N_s+n_{1,s}} \tilde{M}_{1,s,i,i}} \sum_{i=N_s+1}^{N_s+n_{1,s}} (\mu^s_i(1))^2\right] \left[\frac{\max_{N_s+1\leq i\leq N_s+n_{1,s}} \theta^2_i}{n_{1,s}}\right] \\
&= o_P(1),
\end{aligned}$$

where the second equality is by the matrix $[\tilde{M}_{1,s,i,j}]_{N_s+1\leq i,j\leq N_s+n_{1,s}}$ is idempotent, the second inequality is because the operator norm of the matrix $[\tilde{M}_{1,s,i,j}]_{N_s+1\leq i,j\leq N_s+n_{1,s}}$ is at most 1, and the last equality is because by Assumption 2,

$$\begin{aligned}
&\min_{N_s+1\leq i\leq N_s+n_{1,s}} \tilde{M}_{1,s,i,i} \overset{d}{=} \min_{N_s+1\leq i\in\aleph_{1,s}} M_{1,s,i,i} \geq \delta > 0, \quad \max_{i\in[n]} \tilde{V}^2_{1,s,i} = O_P(1), \quad \text{and} \\
&\max_{N_s+1\leq i\leq N_s+n_{1,s}} |\theta_i| \overset{d}{=} \max_{i\in\aleph_{1,s}} |\sum_{j\in\aleph_{1,s}} M_{1,s,i,j}| = o_P(n^{1/2})
\end{aligned}$$

This implies $T_{1,1} = o_P(1)$.

For $T_{1,2}$, we have $\mathbb{E}(T_{1,2} \mid \Xi_{1,s}) = 0$ and

$$\begin{aligned}
Var(T_{1,2} \mid \Xi_{1,s}) &= \frac{1}{n^2_{1,s}}\tilde{\gamma}^{-2}_{1,s,n} \sum_{i=N_s+1}^{N_s+n_{1,s}} \sum_{N_s+1\leq K\leq N_s+n_{1,s}, k\neq i} \frac{\theta^4_i M^2_{1,s,i,k} \tilde{V}^2_{1,s,i} \tilde{V}^2_{1,s,k}}{\tilde{M}^2_{1,s,i,i}} \\
&+ \frac{1}{n^2_{1,s}}\tilde{\gamma}^{-2}_{1,s,n} \sum_{i=N_s+1}^{N_s+n_{1,s}} \sum_{N_s+1\leq K\leq N_s+n_{1,s}, k\neq i} \frac{\theta^2_i \tilde{v}^2_j M^2_{1,s,i,k} \tilde{V}^2_{1,s,i} \tilde{V}^2_{1,s,k}}{\tilde{M}^2_{1,s,i,i}} \\
&\leq \frac{\max_{i\in[n]} \tilde{V}^4_{1,s,i}}{n^2_{1,s}\tilde{\gamma}^2_{1,s,n}} \sum_{i=N_s+1}^{N_s+n_{1,s}} \frac{\theta^4_i \sum_{k=N_s+1}^{N_s+n_{1,s}} \tilde{M}^2_{1,s,i,k}}{\tilde{M}^2_{1,s,i,i}}
\end{aligned}$$

$$+\frac{\max_{i\in[n]}\tilde{V}^4_{1,s,i}}{n^2_{1,s}\tilde{\gamma}^2_{1,s,n}}\sum_{i=N_s+1}^{N_s+n_{1,s}}\sum_{k=N_s+1}^{N_s+n_{1,s}}\frac{\theta_i^2\tilde{v}_k^2\tilde{M}^2_{1,s,i,k}}{\tilde{M}^2_{1,s,i,i}}$$

$$\leq 2\left[\frac{\max_{N_s+1\leq i\leq N_s+n_{1,s}}\tilde{V}^4_{1,s,i}\max_{N_s+1\leq i\leq N_s+n_{1,s}}\theta_i^2}{n_{1,s}\tilde{\gamma}^2_{1,s,n}\min_{N_s+1\leq i\leq N_s+n_{1,s}}\tilde{M}_{1,s,i,i}}\right]\left[\frac{1}{n_{1,s}}\sum_{i=N_s+1}^{N_s+n_{1,s}}\theta_i^2\right]$$

$$= o_P(1).$$

because $\max_{N_s+1\leq i\leq N_s+n_{1,s}}\theta_i^2 = o_P(n)$ and $\frac{1}{n_{1,s}}\sum_{i=N_s+1}^{N_s+n_{1,s}}\theta_i^2 \leq 1$.

For $T_{1,3}$, we have $\mathbb{E}(T_{1,3}\mid\Xi_{1,s}) = 0$ and

$$Var(T_{1,3}\mid\Xi_{1,s}) = \frac{\tilde{\gamma}^{-2}_{1,s,n}}{n^2_{1,s}}\sum_{i=N_s+1}^{N_s+n_{1,s}}\theta_i^2\mathbb{E}((\varepsilon_i^s(1))^2-\tilde{V}^2_{1,s,i}\mid\Xi_{1,s})$$
$$\leq \frac{\tilde{\gamma}^{-2}_{1,s,n}\max_{N_s+1\leq i\leq N_s+n_{1,s}}\mathbb{E}((\varepsilon_i^s(1))^2-\tilde{V}^2_{1,s,i}\mid\Xi_{1,s})}{n^2_{1,s}}\sum_{i=N_s+1}^{N_s+n_{1,s}}\theta_i^2 = O_P(n^{-1}).$$

This implies $T_{1,3} = o_P(1)$.

Last, by Assumption 3, we have

$$T_{1,4}\overset{d}{=}\frac{1}{n_{1,s}}\gamma^{-1}_{1,s,n}\sum_{i\in\aleph_{1,s}}(\sum_{j\in\aleph_{1,s}}M_{1,s,i,j})\mathbb{E}(\varepsilon_i^2(1)|X_i,S_i=s)\xrightarrow{p}\varpi_{a,s,\infty}.$$

This concludes the proof of **Step 1.2.**

Step 1.3: Consistency of $\hat{\Sigma}_{\mathcal{U}}$. We note that $n_s^2/(nn_{1,s})\xrightarrow{p}p_s/\pi_s$ and $n_s^2/(nn_{0,s})\xrightarrow{p}p_s/(1-\pi_s)$. Therefore, we have

$$\hat{\Sigma}_{\mathcal{U}}\xrightarrow{p}\sum_{s\in\mathcal{S}}p_s\begin{pmatrix}\frac{\omega^2_{1,s,\infty}}{\pi_s}+\frac{\omega^2_{0,s,\infty}}{1-\pi_s} & \frac{\omega^2_{1,s,\infty}}{\pi_s}+\frac{\omega^2_{0,s,\infty}}{1-\pi_s}\\ \frac{\omega^2_{1,s,\infty}}{\pi_s}+\frac{\omega^2_{0,s,\infty}}{1-\pi_s} & \cdot\end{pmatrix}.$$

Step 2: Limit of $\hat{\Sigma}^{adj}_{\mathcal{V}}$. It suffices to consider the limits of $\hat{\beta}^\top_{a,s}\Gamma_{a,s}\hat{\beta}_{a,s}-\sum_{i\in\aleph_{a,s}}P_{a,s,i,i}Y_i\acute{\varepsilon}_{a,s,i}$ and $\hat{\beta}^\top_{1,s}\Gamma_s\hat{\beta}_{0,s}$, which are derived in the following two steps.

Step 2.1: Limit of $(\hat{\beta}^\top_{a,s}\Gamma_{a,s}\hat{\beta}_{a,s}-\sum_{i\in\aleph_{a,s}}P_{a,s,i,i}Y_i\acute{\varepsilon}_{a,s,i})/n_{a,s}$. We note that

$$\hat{\beta}_{a,s} = \Gamma^{-1}_{a,s}\sum_{i\in\aleph_{a,s}}\breve{X}_i(Y_i-\hat{\tau}_a(s))$$
$$= \Gamma^{-1}_{a,s}\sum_{i\in\aleph_{a,s}}\breve{X}_i\left[(\alpha_{a,s}+\overline{X}_s^\top\beta_{a,s}-\hat{\tau}_a(s))+\breve{X}_i^\top\beta_{a,s}+e_{i,s}(a)+\varepsilon_i(a)\right]$$
$$= \beta_{a,s}+\Gamma^{-1}_{a,s}\sum_{i\in\aleph_{a,s}}\breve{X}_i\varepsilon_i(a)+\Gamma^{-1}_{a,s}\sum_{i\in\aleph_{a,s}}\breve{X}_i\left[(\alpha_{a,s}+\overline{X}_s^\top\beta_{a,s}-\hat{\tau}_a(s))+e_{i,s}(a)\right]. \quad \text{(C.1)}$$

By (C.1), we have

$$
\begin{aligned}
&\hat{\beta}_{a,s}^\top \Gamma_{a,s} \hat{\beta}_{a,s} - \sum_{i \in \aleph_{a,s}} P_{a,s,i,i} Y_i \acute{\varepsilon}_{a,s,i} \\
&= \beta_{a,s}^\top \Gamma_{a,s} \beta_{a,s} + \left(\varepsilon_{\aleph_{a,s}}^\top(a) P_{a,s} \varepsilon_{\aleph_{a,s}}(a) - \sum_{i \in \aleph_{a,s}} P_{a,s,i,i} Y_i \acute{\varepsilon}_{a,s,i} \right) + 2 \sum_{i \in \aleph_{a,s}} \breve{X}_i^\top \beta_{a,s} \varepsilon_i(a) \\
&+ 2 \sum_{i \in \aleph_{a,s}} \breve{X}_i^\top \beta_{a,s} (\alpha_{a,s} + \overline{X}_s^\top \beta_{a,s} - \hat{\tau}_a(s)) + \varepsilon_{\aleph_{a,s}}^\top(a) P_{a,s} \left[1_{n_{a,s}} (\alpha_{a,s} + \overline{X}_s^\top \beta_{a,s} - \hat{\tau}_a(s)) + e_{\aleph_{a,s}}(a) \right] \\
&+ \left[1_{n_{a,s}} (\alpha_{a,s} + \overline{X}_s^\top \beta_{a,s} - \hat{\tau}_a(s)) + e_{\aleph_{a,s}}(a) \right]^\top P_{a,s} \left[1_{n_{a,s}} (\alpha_{a,s} + \overline{X}_s^\top \beta_{a,s} - \hat{\tau}_a(s)) + e_{\aleph_{a,s}}(a) \right] \\
&\equiv \beta_{a,s}^\top \Gamma_{a,s} \beta_{a,s} + \sum_{l \in [5]} I_l, \qquad \text{(C.2)}
\end{aligned}
$$

where $e_{\aleph_{a,s}}(a)$ and $\varepsilon_{\aleph_{a,s}}(a)$ are $n_{a,s} \times 1$ vectors of $\{e_{i,s}(a)\}_{i \in \aleph_{a,s}}$ and $\{\varepsilon_{i,s}(a)\}_{i \in \aleph_{a,s}}$, respectively. We note that

$$
\begin{aligned}
\sum_{i \in \aleph_{a,s}} (\breve{X}_i^\top \beta_{a,s})^2 &= \sum_{i \in \aleph_{a,s}} \left[\mathbb{E}(Y_i(a)|X_i, S_i = s) - e_{i,s}(a) - \frac{1}{n_s} \sum_{j \in \aleph_s} (\mathbb{E}(Y_j(a)|X_j, S_j = s) - e_{j,s}(a)) \right]^2 \\
&\lesssim \sum_{i \in \aleph_{a,s}} [\mathbb{E}(Y_i(a)|X_i, S_i = s) - e_{i,s}(a)]^2 + \frac{n_{a,s}}{n_s} \sum_{j \in \aleph_s} [(\mathbb{E}(Y_j(a)|X_j, S_j = s) - e_{j,s}(a))]^2 \\
&= O_P(n),
\end{aligned}
$$

$$
\begin{aligned}
&\left| \left[\frac{1}{n_{a,s}} \beta_{a,s}^\top \Gamma_{a,s} \beta_{a,s} \right]^{1/2} - \left[\frac{1}{n_{a,s}} \sum_{i \in \aleph_{a,s}} (\mathbb{E}(Y_i(a)|X_i, S_i = s) - \mathbb{E}(Y_i(a)|S_i = s))^2 \right]^{1/2} \right| \\
&= \left| \left[\frac{1}{n_{a,s}} \sum_{i \in \aleph_{a,s}} (\breve{X}_i^\top \beta_{a,s})^2 \right]^{1/2} - \left[\frac{1}{n_{a,s}} \sum_{i \in \aleph_{a,s}} (\mathbb{E}(Y_i(a)|X_i, S_i = s) - \mathbb{E}(Y_i(a)|S_i = s))^2 \right]^{1/2} \right| \\
&\leq \left[\frac{1}{n_{a,s}} \sum_{i \in \aleph_{a,s}} \left(e_{i,s}(a) - \frac{1}{n_s} \sum_{j \in \aleph_s} e_{j,s}(a) \right)^2 \right]^{1/2} + \left| \frac{1}{n_s} \sum_{j \in \aleph_s} \mathbb{E}(Y_j(a)|X_j, S_j = s) - \mathbb{E}(Y_i(a)|S_i = s) \right| \\
&= o_P(1),
\end{aligned}
$$

and

$$
\frac{1}{n_{a,s}} \sum_{i \in \aleph_{a,s}} (\mathbb{E}(Y_i(a)|X_i, S_i = s) - \mathbb{E}(Y_i(a)|S_i = s))^2 \xrightarrow{p} var(\phi_i(a)|S_i = s),
$$

which implies

$$\frac{1}{n_{a,s}}\beta_{a,s}^{\top}\Gamma_{a,s}\beta_{a,s} \xrightarrow{p} var(\phi_i(a)|S_i = s).$$

Next, we show $I_1, \cdots, I_5$ are $o_P(n)$. Lemma G.2 shows $I_1 = o_P(n)$. For I_2, we note that, conditional on $(A^{(n)}, S^{(n)}, X^{(n)})$, $\{\varepsilon_i(a)\}_{i\in\aleph_{a,s}}$ is independent across i and mean zero, which implies the conditional variance of $\sum_{i\in\aleph_{a,s}}\breve{X}_i^{\top}\beta_{a,s}\varepsilon_i(a)$ is $\sum_{i\in\aleph_{a,s}}(\breve{X}_i^{\top}\beta_{a,s})^2 = O_P(n)$ as shown below, and thus, $I_2 = O_P(n^{1/2}) = o_P(n)$. Because $\alpha_{a,s} + \overline{X}_s^{\top}\beta_{a,s} - \hat{\tau}_a(s) \xrightarrow{p} o_P(1)$ by the proof of Theorem 3.1, $\sum_{i\in\aleph_{a,s}} e_{i,s}^2(a) = o(n)$, and $\lambda_{\max}(P_{a,s}) \leq 1$, we have $I_l = o_P(n)$ for $l = 3, 4, 5$. This means

$$\frac{1}{n_{a,s}}\hat{\beta}_{a,s}^{\top}\Gamma_{a,s}\hat{\beta}_{a,s} - \sum_{i\in\aleph_{a,s}} P_{a,s,i,i}Y_i\acute{\varepsilon}_{a,s,i} \xrightarrow{p} var(\phi_i(a)|S_i = s).$$

Step 2.2: Limit of $\hat{\beta}_{1,s}^{\top}\Gamma_s\hat{\beta}_{0,s}$. By (C.1), we have

$$\begin{aligned}
\hat{\beta}_{1,s}^{\top}\Gamma_s\hat{\beta}_{0,s} &= \beta_{1,s}^{\top}\Gamma_s\beta_{0,s} + \beta_{1,s}^{\top}\Gamma_s\Gamma_{0,s}^{-1}\sum_{i\in\aleph_{0,s}}\breve{X}_i\varepsilon_i(0) \\
&+ \beta_{1,s}^{\top}\Gamma_s\Gamma_{0,s}^{-1}\left[\sum_{i\in\aleph_{0,s}}\breve{X}_i\left(\alpha_{0,s} + \overline{X}_s^{\top}\beta_{0,s} - \hat{\tau}_0(s) + e_{i,s}(0)\right)\right] \\
&+ \left[\sum_{i\in\aleph_{1,s}}\breve{X}_i\varepsilon_i(1)\right]^{\top}\Gamma_{1,s}^{-1}\Gamma_s\beta_{0,s} + \left[\sum_{i\in\aleph_{1,s}}\breve{X}_i\varepsilon_i(1)\right]^{\top}\Gamma_{1,s}^{-1}\Gamma_s\Gamma_{0,s}^{-1}\sum_{i\in\aleph_{0,s}}\breve{X}_i\varepsilon_i(0) \\
&+ \left[\sum_{i\in\aleph_{1,s}}\breve{X}_i\varepsilon_i(1)\right]^{\top}\Gamma_{1,s}^{-1}\Gamma_s\Gamma_{0,s}^{-1}\left[\sum_{i\in\aleph_{0,s}}\breve{X}_i\left(\alpha_{0,s} + \overline{X}_s^{\top}\beta_{0,s} - \hat{\tau}_0(s) + e_{i,s}(0)\right)\right] \\
&+ \left[\sum_{i\in\aleph_{1,s}}\breve{X}_i\left(\alpha_{1,s} + \overline{X}_s^{\top}\beta_{1,s} - \hat{\tau}_1(s) + e_{i,s}(1)\right)\right]^{\top}\Gamma_{1,s}^{-1}\Gamma_s\beta_{0,s} \\
&+ \left[\sum_{i\in\aleph_{1,s}}\breve{X}_i\left(\alpha_{1,s} + \overline{X}_s^{\top}\beta_{1,s} - \hat{\tau}_1(s) + e_{i,s}(1)\right)\right]^{\top}\Gamma_{1,s}^{-1}\Gamma_s\Gamma_{0,s}^{-1}\sum_{i\in\aleph_{0,s}}\breve{X}_i\varepsilon_i(0) \\
&+ \left[\sum_{i\in\aleph_{1,s}}\breve{X}_i\left(\alpha_{1,s} + \overline{X}_s^{\top}\beta_{1,s} - \hat{\tau}_1(s) + e_{i,s}(1)\right)\right]^{\top}\Gamma_{1,s}^{-1}\Gamma_s\Gamma_{0,s}^{-1} \\
&\times\left[\sum_{i\in\aleph_{0,s}}\breve{X}_i\left(\alpha_{0,s} + \overline{X}_s^{\top}\beta_{0,s} - \hat{\tau}_0(s) + e_{i,s}(0)\right)\right] \\
&\equiv \sum_{l\in[9]} I_l.
\end{aligned}$$

Following the previous argument, we can show that

$$\frac{I_1}{n} = \frac{1}{n}\sum_{i\in\aleph_s}(\mathbb{E}(Y_i(1)|X_i,S_i=s)-\mathbb{E}(Y_i(1)|S_i=s))(\mathbb{E}(Y_i(0)|X_i,S_i=s)-\mathbb{E}(Y_i(0)|S_i=s))+o_P(1)$$
$$\xrightarrow{p} cov(\phi_i(1),\phi_i(0)|S_i=s).$$

For I_2, we have

$$\mathbb{E}(I_2|A^{(n)},S^{(n)},X^{(n)})=0$$

and

$$var(I_2|A^{(n)},S^{(n)},X^{(n)}) \lesssim \beta_{1,s}^\top\Gamma_s\Gamma_{0,s}^{-1}\Gamma_s\beta_{1,s}\max_{i\in\aleph_{0,s}}\mathbb{E}(\varepsilon_i^2(0)|A^{(n)},S^{(n)},X^{(n)})=o_P(n^2),$$

where we use the fact that

$$\beta_{1,s}^\top\Gamma_s\Gamma_{0,s}^{-1}\Gamma_s\beta_{1,s} \le \left\|\Gamma_s^{1/2}\Gamma_{0,s}^{-1}\Gamma_s^{1/2}\right\|_{op}||\breve{X}_{\aleph_s}\beta_{1,s}||_2^2 = o_P(n^2). \tag{C.3}$$

For I_3, we have

$$|I_3| \le \left\|\beta_{1,s}^\top\Gamma_s\Gamma_{0,s}^{-1}\breve{X}_{\aleph_{0,s}}^\top\right\|_2 \times O_P(1) = o_P(n),$$

where the first inequality holds by Assumption 2(i) and the facts that

$$\alpha_{0,s}+\overline{X}_s^\top\beta_{0,s}-\hat{\tau}_0(s)=O_P(n^{-1/2}), \quad \sum_{i\in\aleph_{0,s}}e_{i,s}^2(a)=o_P(1)$$

and the last equality holds by Assumption 4, (C.3), and the fact that

$$\left\|\beta_{1,s}^\top\Gamma_s\Gamma_{0,s}^{-1}\breve{X}_{\aleph_{0,s}}^\top\right\|_2 = \left(\beta_{1,s}^\top\Gamma_s\Gamma_{0,s}^{-1}\Gamma_s\beta_{1,s}\right)^{1/2} = o_P(n).$$

We can show $I_4 = o_P(n)$ following the same argument in bounding I_2.

For I_5, we note that the two index sets $\aleph_{0,s}$ and $\aleph_{1,s}$ do not overlap. Therefore, we have

$$\mathbb{E}(I_5|A^{(n)},S^{(n)},X^{(n)})=0$$

and

$$var(I_5|A^{(n)},S^{(n)},X^{(n)}) \le \sum_{i\in\aleph_{1,s}}\sum_{j\in\aleph_{0,s}}B_{i,j}^2\left[\max_{a=0,1}\max_{i\in\aleph_{a,s}}\mathbb{E}(\varepsilon_i^2(0)|A^{(n)},S^{(n)},X^{(n)})\right]^2 = o_P(n^2),$$

where we have $B_{i,j} = \breve{X}_i^\top \Gamma_{1,s}^{-1}\Gamma_s\Gamma_{0,s}^{-1}\breve{X}_j$ and

$$\begin{aligned}\sum_{i\in\aleph_{1,s}}\sum_{j\in\aleph_{0,s}} B_{i,j}^2 &= \sum_{i\in\aleph_{1,s}} \mathrm{trace}(\breve{X}_i^\top \Gamma_{1,s}^{-1}\Gamma_s\Gamma_{0,s}^{-1}\Gamma_s\Gamma_{1,s}^{-1}\breve{X}_i)\\ &= \mathrm{trace}(\Gamma_s^{1/2}\Gamma_{0,s}^{-1}\Gamma_s\Gamma_{1,s}^{-1}\Gamma_s^{1/2})\\ &= \mathrm{trace}(\Gamma_s^{1/2}\Gamma_{0,s}^{-1}\Gamma_s^{1/2}) + \mathrm{trace}(\Gamma_s^{1/2}\Gamma_{1,s}^{-1}\Gamma_s^{1/2}) = o_P(n^2),\end{aligned}$$

where we use the fact that $\Gamma_s = \Gamma_{0,s} + \Gamma_{1,s}$.

For I_6, we have

$$\begin{aligned}|I_6| &\le \left|\left[\sum_{i\in\aleph_{1,s}} \breve{X}_i\varepsilon_i(1)\right]^\top \Gamma_{1,s}^{-1}\Gamma_s\Gamma_{0,s}^{-1}\left[\sum_{i\in\aleph_{0,s}}\breve{X}_i\right]\right| |\alpha_{0,s} + \overline{X}_s^\top\beta_{0,s} - \hat{\tau}_0(s)|\\ &\quad + \left|\left[\sum_{i\in\aleph_{1,s}} \breve{X}_i\varepsilon_i(1)\right]^\top \Gamma_{1,s}^{-1}\Gamma_s\Gamma_{0,s}^{-1}\left[\sum_{i\in\aleph_{0,s}}\breve{X}_i e_{i,s}(0)\right]\right|\\ &\equiv |I_{6,1}||\alpha_{0,s} + \overline{X}_s^\top\beta_{0,s} - \hat{\tau}_0(s)| + |I_{6,2}|.\end{aligned}$$

We note that

$$\mathbb{E}(I_{6,1}|A^{(n)}, S^{(n)}, X^{(n)}) = 0$$

and

$$\begin{aligned}&var(I_{6,1}|A^{(n)}, S^{(n)}, X^{(n)})\\ &\le \left[\sum_{i\in\aleph_{0,s}}\breve{X}_i\right]^\top \Gamma_{0,s}^{-1}\Gamma_s\Gamma_{1,s}^{-1}\Gamma_s S_0^{-1}\left[\sum_{i\in\aleph_{0,s}}\breve{X}_i\right]\left[\max_{i\in\aleph_{a,s}}\mathbb{E}(\varepsilon_i^2(1)|A^{(n)}, S^{(n)}, X^{(n)})\right]\\ &= o_P(n^3).\end{aligned}$$

where the second last equality holds because

$$\begin{aligned}\left[\sum_{i\in\aleph_{0,s}}\breve{X}_i\right]^\top \Gamma_{0,s}^{-1}\Gamma_s\Gamma_{1,s}^{-1}\Gamma_s S_0^{-1}\left[\sum_{i\in\aleph_{0,s}}\breve{X}_i\right] &\le n_{0,s}\left\|\breve{X}_{\aleph_{0,s}}\Gamma_{0,s}^{-1}\Gamma_s\Gamma_{1,s}^{-1}\Gamma_s S_0^{-1}\breve{X}_{\aleph_{0,s}}\right\|_{op}\\ &\le n_{0,s}\left\|\Gamma_s^{1/2}\Gamma_{1,s}^{-1}\Gamma_s^{1/2}\right\|_{op}\left\|\breve{X}_{\aleph_{0,s}}\Gamma_{0,s}^{-1}\Gamma_s^{1/2}\right\|_{op}^2\\ &\le n_{0,s}\left\|\Gamma_s^{1/2}\Gamma_{1,s}^{-1}\Gamma_s^{1/2}\right\|_{op}\left\|\Gamma_s^{1/2}\Gamma_{0,s}^{-1}\Gamma_s^{1/2}\right\|_{op} = o_P(n^3).\end{aligned}$$

This implies $|I_{6,1}||\alpha_{0,s}+\overline{X}_s^\top\beta_{0,s}-\hat{\tau}_0(s)|=o_P(n)$. Similarly, for $I_{6,2}$, we have

$$\mathbb{E}(I_{6,2}|A^{(n)},S^{(n)},X^{(n)})=0$$

and

$$\begin{aligned}
&var(I_{6,2}|A^{(n)},S^{(n)},X^{(n)})\\
&\le\left[\sum_{i\in\aleph_{0,s}}\breve{X}_ie_{i,s}(0)e_{i,s}(0)\right]^\top\Gamma_{0,s}^{-1}\Gamma_s\Gamma_{1,s}^{-1}\Gamma_sS_0^{-1}\left[\sum_{i\in\aleph_{0,s}}\breve{X}_ie_{i,s}(0)\right]\left[\max_{i\in\aleph_{a,s}}\mathbb{E}(\varepsilon_i^2(1)|A^{(n)},S^{(n)},X^{(n)})\right]\\
&\le\left\|\breve{X}_{\aleph_{0,s}}\Gamma_{0,s}^{-1}\Gamma_s\Gamma_{1,s}^{-1}\Gamma_sS_0^{-1}\breve{X}_{\aleph_{0,s}}\right\|_{op}\left\|e_{\aleph_{0,s}}(0)\right\|_2^2\left[\max_{i\in\aleph_{a,s}}\mathbb{E}(\varepsilon_i^2(1)|A^{(n)},S^{(n)},X^{(n)})\right]=o_P(n^2),
\end{aligned}$$

where we use the fact that $\left\|e_{\aleph_{0,s}}(0)\right\|_2^2=O_P(1)$.

We can show $I_7=o_P(n)$ and $I_8=o_P(n)$ following the same argument used to bound I_3 and I_6, respectively.

For I_9, we have

$$\begin{aligned}
|I_9|\le&\left\|\breve{X}_{\aleph_{1,s}}\Gamma_{1,s}^{-1}\Gamma_s\Gamma_{0,s}^{-1}\breve{X}_{\aleph_{0,s}}\right\|_{op}\left\|1_{n_{1,s}}(\alpha_{1,s}+\overline{X}_s^\top-\hat{\tau}_1(s))+e_{\aleph_{1,s}}(1)\right\|_2\\
&\times\left\|1_{n_{0,s}}(\alpha_{0,s}+\overline{X}_s^\top-\hat{\tau}_0(s))+e_{\aleph_{0,s}}(0)\right\|_2=o_P(n),
\end{aligned}$$

where we use the fact that

$$\begin{aligned}
\left\|\breve{X}_{\aleph_{1,s}}\Gamma_{1,s}^{-1}\Gamma_s\Gamma_{0,s}^{-1}\breve{X}_{\aleph_{0,s}}\right\|_{op}&=\left\|\breve{X}_{\aleph_{1,s}}\Gamma_{1,s}^{-1}\Gamma_s\Gamma_{0,s}^{-1}\Gamma_s\Gamma_{1,s}^{-1}\breve{X}_{\aleph_{1,s}}^\top\right\|_{op}^{1/2}\\
&\le\left[\left\|\Gamma_s^{1/2}\Gamma_{0,s}^{-1}\Gamma_s^{1/2}\right\|_{op}\left\|\Gamma_s^{1/2}\Gamma_{1,s}^{-1}\breve{X}_{\aleph_{1,s}}\right\|_{op}^2\right]^{1/2}\\
&=\left[\left\|\Gamma_s^{1/2}\Gamma_{0,s}^{-1}\Gamma_s^{1/2}\right\|_{op}\left\|\Gamma_s^{1/2}\Gamma_{1,s}^{-1}\Gamma_s^{1/2}\right\|_{op}\right]^{1/2}=o_P(n)
\end{aligned}$$

and

$$\left\|1_{n_{0,s}}(\alpha_{0,s}+\overline{X}_s^\top-\hat{\tau}_0(s))+e_{\aleph_{0,s}}(0)\right\|_2=O_P(1).$$

This leads to the desired result in this step that $\hat{\beta}_{1,s}^\top\Gamma_s\hat{\beta}_{0,s}\xrightarrow{p}cov(\phi_i(1),\phi_i(0)|S_i=s)$.

Step 3: Limit of $\hat{\Sigma}_{\mathcal{W}}$. For $\hat{\Sigma}_{\mathcal{W}}$, by the proof of Theorem 3.1, we have $\hat{\tau}_{1,s}-\hat{\tau}_{0,s}\xrightarrow{p}\mathbb{E}(Y(1)-Y(0)|S=s)$ and $\hat{\tau}^{adj}\xrightarrow{p}\tau$. By Assumption 1, we also have $\hat{p}_s\xrightarrow{p}p_s$, which implies the desired result.

D Proof of Theorem 3.3

We note that

$$\hat{w} \xrightarrow{p} \frac{\Sigma_{2,2} - \Sigma_{1,2}}{\Sigma_{1,1} + \Sigma_{2,2} - 2\Sigma_{1,2}} = \arg\min_{w}(w, 1-w)\Sigma(w, 1-w)^\top.$$

This implies $\hat{\tau}^*$ is weakly more efficient than both $\hat{\tau}^{adj}$ and $\hat{\tau}^{adj}$ which correspond to $\hat{w} = 1$ and 0, respectively. Then, by the continuous mapping theorem, we have

$$\sqrt{n}\left[(\hat{w}, 1-\hat{w})\hat{\Sigma}(\hat{w}, 1-\hat{w})^\top\right]^{-1/2}(\hat{\tau}^* - \tau) \rightsquigarrow \mathcal{N}(0,1).$$

E Proof of Theorem 4.1

Step 1: Asymptotic normality and the expression of Ω. The asymptotic normality of $\hat{\tau}^{adj}$ has already been established by Ye et al. (2022) (our $\hat{\tau}^{adj}$ is just their $\hat{\theta}_A$). In the following, we just sketch the proof of the joint distribution of $(\hat{\tau}^{adj}, \hat{\tau}^{unadj})$. Following Remark 3, we have

$$\begin{aligned}
\hat{\tau}^{adj} &= \frac{1}{n}\sum_{i\in[n]}\frac{A_i(Y_i - X_i^\top\hat{\beta}_{1,S_i})}{\hat{\pi}_{S_i}} - \frac{1}{n}\sum_{i\in[n]}\frac{(1-A_i)(Y_i - X_i^\top\hat{\beta}_{0,S_i})}{1-\hat{\pi}_{S_i}} + \frac{1}{n}\sum_{i\in[n]}X_i^\top(\hat{\beta}_{1,S_i} - \hat{\beta}_{0,S_i}) \\
&= \frac{1}{n}\sum_{i\in[n]}\frac{A_i(Y_i(1) - X_i^\top\beta^*_{1,S_i})}{\hat{\pi}_{S_i}} - \frac{1}{n}\sum_{i\in[n]}\frac{(1-A_i)(Y_i(0) - X_i^\top\beta^*_{0,S_i})}{1-\hat{\pi}_{S_i}} + \frac{1}{n}\sum_{i\in[n]}X_i^\top(\beta^*_{1,S_i} - \beta^*_{0,S_i}) \\
&+ \sum_{s\in\mathcal{S}}\hat{p}_s\left(\frac{1}{n_{1,s}}\sum_{i\in\aleph_{1,s}}X_i - \frac{1}{n_s}\sum_{i\in\aleph_s}X_i\right)^\top(\beta^*_{1,s} - \hat{\beta}_{1,s}) \\
&- \sum_{s\in\mathcal{S}}\hat{p}_s\left(\frac{1}{n_{0,s}}\sum_{i\in\aleph_{0,s}}X_i - \frac{1}{n_s}\sum_{i\in\aleph_s}X_i\right)^\top(\beta^*_{0,s} - \hat{\beta}_{0,s}) \\
&= \frac{1}{n}\sum_{i\in[n]}\frac{A_i(Y_i(1) - X_i^\top\beta^*_{1,S_i})}{\hat{\pi}_{S_i}} - \frac{1}{n}\sum_{i\in[n]}\frac{(1-A_i)(Y_i(0) - X_i^\top\beta^*_{0,S_i})}{1-\hat{\pi}_{S_i}} \\
&+ \frac{1}{n}\sum_{i\in[n]}X_i^\top(\beta^*_{1,S_i} - \beta^*_{0,S_i}) + o_P(n^{-1/2}),
\end{aligned}$$

where the last equality holds because

$$\max_{a=0,1,s\in\mathcal{S}}\left\|\frac{1}{n_{a,s}}\sum_{i\in\aleph_{a,s}}X_i - \frac{1}{n_s}\sum_{i\in\aleph_s}X_i\right\|_2 = O_P(n^{-1/2}) \quad \text{and} \quad \max_{a=0,1,s\in\mathcal{S}}||\hat{\beta}_{a,s} - \beta^*_{a,s}||_2 = o_P(1).$$

Let $\delta_{Y,a,s} = Y_i(a) - \mathbb{E}(Y_i(a)|S_i = s)$ and $\delta_{X,s,i} = X_i - \mathbb{E}(X_i|S_i = s)$. Then, we have

$$\begin{aligned}
\begin{pmatrix} \sqrt{n}(\hat{\tau}^{adj} - \tau) \\ \sqrt{n}(\hat{\tau}^{unadj} - \tau) \end{pmatrix} &= \sum_{s\in\mathcal{S}} \hat{p}_s \sqrt{n} \begin{pmatrix} \frac{1}{n_{1,s}} \sum_{i\in\aleph_{1,s}} (\delta_{Y,1,s,i} - (1-\hat{\pi}_s)\delta_{X,s,i}^\top \beta_{1,s}^* - \hat{\pi}_s \delta_{X,s,i}^\top \beta_{0,s}^*) \\ \frac{1}{n_{1,s}} \sum_{i\in\aleph_{1,s}} \delta_{Y,1,s,i} \end{pmatrix} \\
&- \sum_{s\in\mathcal{S}} \hat{p}_s \sqrt{n} \begin{pmatrix} \frac{1}{n_{0,s}} \sum_{i\in\aleph_{0,s}} (\delta_{Y,0,s} - \hat{\pi}_s \delta_{X,s,i}^\top \beta_{0,s}^* - (1-\hat{\pi}_s)\delta_{X,s,i}^\top \beta_{1,s}^*) \\ \frac{1}{n_{1,s}} \sum_{i\in\aleph_{0,s}} \delta_{Y,0,s} \end{pmatrix} \\
&+ \sqrt{n} \sum_{s\in\mathcal{S}} \hat{p}_s (\mathbb{E}(Y_i(1)|S_i = s) - \mathbb{E}(Y_i(0)|S_i = s) - \tau) 1_2 + o_P(1) \\
&= \sum_{s\in\mathcal{S}} \hat{p}_s \sqrt{n} \begin{pmatrix} \frac{1}{n_{1,s}} \sum_{i\in\aleph_{1,s}} (\delta_{Y,1,s,i} - (1-\pi_s)\delta_{X,s,i}^\top \beta_{1,s}^* - \pi_s \delta_{X,s,i}^\top \beta_{0,s}^*) \\ \frac{1}{n_{1,s}} \sum_{i\in\aleph_{1,s}} \delta_{Y,1,s,i} \end{pmatrix} \\
&- \sum_{s\in\mathcal{S}} \hat{p}_s \sqrt{n} \begin{pmatrix} \frac{1}{n_{0,s}} \sum_{i\in\aleph_{0,s}} (\delta_{Y,0,s} - \pi_s \delta_{X,s,i}^\top \beta_{0,s}^* - (1-\pi_s)\delta_{X,s,i}^\top \beta_{1,s}^*) \\ \frac{1}{n_{1,s}} \sum_{i\in\aleph_{0,s}} \delta_{Y,0,s} \end{pmatrix} \\
&+ \frac{1}{\sqrt{n}} \sum_{i\in[n]} (\mathbb{E}(Y_i(1)|S_i) - \mathbb{E}(Y_i(0)|S_i) - \tau) 1_2 + o_P(1) \\
&\equiv \tilde{U}_{n,1} - \tilde{U}_{n,0} + W_n 1_2,
\end{aligned}$$

where we denote

$$\begin{aligned}
\tilde{U}_{n,1} &= \sum_{s\in\mathcal{S}} \hat{p}_s \sqrt{n} \begin{pmatrix} \frac{1}{n_{1,s}} \sum_{i\in\aleph_{1,s}} (\delta_{Y,1,s,i} - (1-\pi_s)\delta_{X,s,i}^\top \beta_{1,s}^* - \pi_s \delta_{X,s,i}^\top \beta_{0,s}^*) \\ \frac{1}{n_{1,s}} \sum_{i\in\aleph_{1,s}} \delta_{Y,1,s,i} \end{pmatrix} \\
\tilde{U}_{n,0} &= \sum_{s\in\mathcal{S}} \hat{p}_s \sqrt{n} \begin{pmatrix} \frac{1}{n_{0,s}} \sum_{i\in\aleph_{0,s}} (\delta_{Y,0,s} - \pi_s \delta_{X,s,i}^\top \beta_{0,s}^* - (1-\pi_s)\delta_{X,s,i}^\top \beta_{1,s}^*) \\ \frac{1}{n_{1,s}} \sum_{i\in\aleph_{0,s}} \delta_{Y,0,s} \end{pmatrix},
\end{aligned}$$

W_n is defined in (B.3), and the second equality is by the facts that

$$\hat{\pi}_s - \pi_s = \frac{\sum_{i\in[n]} (A_i - \pi_s) 1\{S_i = s\}}{n_s} = o_P(1), \quad \left\| \frac{1}{n_{a,s}} \sum_{i\in\aleph_{a,s}} \delta_{X,s,i} \right\|_2 = O_P(n^{-1/2}), \quad \text{and thus,}$$

$$\max_{a=0,1,s\in\mathcal{S}} \left| (\hat{\pi}_s - \pi_s) \left(\frac{1}{n_{a,s}} \sum_{i\in\aleph_{a,s}} \delta_{X,s,i}^\top (\beta_{1,s}^* - \beta_{0,s}^*) \right) \right| = o_P(n^{-1/2}).$$

Following the same argument in the proof of Lemma G.1,[7] we have $(\tilde{U}_{n,1}, \tilde{U}_{n,0}, W_n) \xrightarrow{p} (\tilde{U}_1, \tilde{U}_0, W)$,

[7] Also see Lemma N.3 in Jiang et al. (2022) for a similar argument for the regression adjusted quantile treatment effect estimator.

where $(\tilde{U}_1, \tilde{U}_0, W)$ are independent, $m_a(x,s) = \mathbb{E}(Y_i(a)|X_i = x, S_i = s)$, $\overline{\beta}^*_s = (1-\pi_s)\beta_{1,s} + \pi_s \beta^*_{0,s}$,

$$\tilde{U}_1 \stackrel{d}{=} \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \Omega_{U,1}\right), \quad \tilde{U}_0 \stackrel{d}{=} \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \Omega_{U,0}\right), \quad W \stackrel{d}{=} \mathcal{N}(0, var(\mathbb{E}(Y_i(1)|S_i) - \mathbb{E}(Y_i(0)|S_i))),$$

$$\begin{aligned}
\Omega_{U,1} &= \mathbb{E}\frac{Var(Y_i(1)|X_i,S_i)}{\pi_{S_i}} 1_2 1_2^\top \\
&+ \sum_{s\in\mathcal{S}} \frac{p_s}{\pi_s} \begin{pmatrix} Var(m_1(X_i,s) - X_i^\top \overline{\beta}^*_s \mid S_i = s) & cov(m_1(X_i,s) - X_i^\top \overline{\beta}^*_s, m_1(X_i,s) \mid S_i = s) \\ cov(m_1(X_i,s) - X_i^\top \overline{\beta}^*_s, m_1(X_i,s) \mid S_i = s) & Var(m_1(X_i,s)) \end{pmatrix} \\
\Omega_{U,0} &= \mathbb{E}\frac{Var(Y_i(0)|X_i,S_i)}{1-\Pi_{S_i}} 1_2 1_2^\top \\
&+ \sum_{s\in\mathcal{S}} \frac{p_s}{1-\pi_s} \begin{pmatrix} Var(m_0(X_i,s) - X_i^\top \overline{\beta}^*_s \mid S_i = s) & cov(m_0(X_i,s) - X_i^\top \overline{\beta}^*_s, m_0(X_i,s) \mid S_i = s) \\ cov(m_0(X_i,s) - X_i^\top \overline{\beta}^*_s, m_0(X_i,s) \mid S_i = s) & Var(m_0(X_i,s)) \end{pmatrix}.
\end{aligned}$$

We further note that

$$\begin{aligned}
\Omega &= \Omega_{U,1} + \Omega_{U,0} + var(\mathbb{E}(Y_i(1)|S_i) - \mathbb{E}(Y_i(0)|S_i)) 1_2 1_2^\top \\
&= \left\{ \mathbb{E}\left[\frac{Var(Y_i(1)|X_i,S_i)}{\pi_{S_i}} + \frac{Var(Y_i(0)|X_i,S_i)}{1-\pi_{S_i}}\right] + \mathbb{E}(m_1(X_i,S_i) - m_0(X_i,S_i) - \tau)^2 \right\} 1_2 1_2^\top \\
&+ \sum_{s\in\mathcal{S}} \frac{p_s}{\pi_s(1-\pi_s)} \begin{pmatrix} V_s & V_s \\ V_s & V'_s \end{pmatrix}
\end{aligned}$$

where

$$\begin{aligned}
V_s &= Var((1-\pi_s)m_1(X_i,s) + \pi_s m_0(X_i,s) - X_i^\top \overline{\beta}^*_s \mid S_i = s) \\
V'_s &= Var((1-\pi_s)m_1(X_i,s) + \pi_s m_0(X_i,s) \mid S_i = s).
\end{aligned}$$

Then, by the definition of $\beta^*_{a,s}$, we see that $V'_s \geq V_s$ for $s \in \mathcal{S}$ and thus, $\Omega_{1,1} = \Omega_{1,2} \leq \Omega_{2,2}$.

Step 2: Limit of $\hat{\Sigma}_{\mathcal{U}}$. We have

$$\gamma_{a,s,n} = 1 - \left(\frac{1}{n_{a,s}} \sum_{i\in\aleph_{a,s}} \breve{X}_i^\top\right) \left(\frac{1}{n_{a,s}} \sum_{i\in\aleph_{a,s}} \breve{X}_i \breve{X}_i^\top\right) \left(\frac{1}{n_{a,s}} \sum_{i\in\aleph_{a,s}} \breve{X}_i\right) \xrightarrow{p} 1.$$

In addition, we have $\sum_{j\in\aleph_{a,s}} M_{a,s,i,j} = 1 - R_{a,s,i}$ where $R_i = \breve{X}_i^\top \left(\frac{1}{n_{a,s}} \sum_{i\in\aleph_{a,s}} \breve{X}_i \breve{X}_i^\top\right)^{-1} \left(\frac{1}{n_{a,s}} \sum_{i\in\aleph_{a,s}} \breve{X}_i\right)$ so that

$$\max_{i\in\aleph_{a,s}} ||R_{a,s,i}||_2 \leq \max_{i\in[n]} ||\breve{X}_i||_2 \lambda_{\min} \left(\frac{1}{n_{a,s}} \sum_{i\in\aleph_{a,s}} \breve{X}_i \breve{X}_i^\top\right)^{-1} \left\|\frac{1}{n_{a,s}} \sum_{i\in\aleph_{a,s}} \breve{X}_i\right\|_2 = o_P(1)$$

and

$$\left|\frac{1}{n_{a,s}}\sum_{i\in\aleph_{a,s}}\left[\left(\sum_{j\in\aleph_{a,s}}M_{a,s,i,j}\right)^2-1\right]Y_i\acute{\varepsilon}_{a,s,i}\right|\leq 2\max_{i\in\aleph_{a,s}}\left(|R_{a,s,i}|+R_{a,s,i}^2\right)\frac{1}{n_{a,s}}\sum_{i\in\aleph_{a,s}}|Y_i\acute{\varepsilon}_{a,s,i}|=o_P(1)$$

where we use the facts that

$$\max_{i\in\aleph_{a,s}}|1-M_{a,s,i,i}|=\max_{i\in\aleph_{a,s}}P_{a,s,i,i}\leq\frac{1}{n_{a,s}}\lambda_{\max}\left(\frac{1}{n_{a,s}}\breve{X}_i\breve{X}_i^\top\right)\max_{i\in\aleph_{a,s}}||X_i||_2^2=o_P(1)$$

and

$$\frac{1}{n_{a,s}}\sum_{i\in\aleph_{a,s}}|Y_i\acute{\varepsilon}_{a,s,i}|\leq\frac{1}{n_{a,s}\min_{i\in\aleph_{a,s}}M_{a,s,i,i}}|Y_i||Y_i-\hat{\tau}_{a,s}-\breve{X}_i^\top\hat{\beta}_{a,s}|=O_P(1).$$

Next, recall $\hat{\tau}_{a,s}$ and $\hat{\beta}_{a,s}$ are the intercept and slope of the OLS regression of Y_i on $(1,\breve{X}_i)$ using observations $i\in\aleph_{a,s}$. Therefore, we have $\hat{\beta}_{a,s}\xrightarrow{p}\beta^*_{a,s}$,

$$\hat{\tau}_{a,s}=\frac{1}{n_{a,s}}\sum_{i\in\aleph_{a,s}}(Y_i-\breve{X}_i^\top\hat{\beta}_{a,s})\xrightarrow{p}\mathbb{E}(Y_i(1)|S_i=s),$$

$$\left|\frac{1}{n_{a,s}}\sum_{i\in\aleph_{a,s}}Y_i(\acute{\varepsilon}_{a,s,i}-\hat{\varepsilon}_{a,s,i})\right|\leq\left(\max_{i\in\aleph_{a,s}}P_{a,s,i,i}\right)\left(\frac{1}{n_{a,s}}\sum_{i\in\aleph_{a,s}}|Y_i\acute{\varepsilon}_{a,s,i}|\right)=o_P(1),$$

and

$$\begin{aligned}\frac{1}{n_{a,s}}\sum_{i\in\aleph_{a,s}}Y_i\hat{\varepsilon}_{a,s,i}&=\frac{1}{n_{a,s}}\sum_{i\in\aleph_{a,s}}Y_i(Y_i-\hat{\tau}_{a,s}-\breve{X}_i^\top\hat{\beta}_{a,s})\\&=\frac{1}{n_{a,s}}\sum_{i\in\aleph_{a,s}}Y_i^2-\left(\frac{1}{n_{a,s}}\sum_{i\in\aleph_{a,s}}Y_i\right)\hat{\tau}_{a,s}\\&-\left[\frac{1}{n_{a,s}}\sum_{i\in\aleph_{a,s}}Y_iX_i-\left(\frac{1}{n_{a,s}}\sum_{i\in\aleph_{a,s}}Y_i\right)\left(\frac{1}{n_s}\sum_{i\in\aleph_s}X_i\right)\right]^\top\hat{\beta}_{a,s}\\&\xrightarrow{p}Var(Y_i(a)|S_i=s)-cov(Y_i(a),X_i^\top\beta^*_{a,s}|S_i=s)=Var(Y_i(a)-X_i^\top\beta^*_{a,s}|S_i=s).\end{aligned}$$

Therefore, we have $\hat{\omega}^2_{a,s}\xrightarrow{p}Var(Y_i(a)-X_i^\top\beta^*_{a,s}|S_i=s)$. By a similar argument, we have

$$\hat{\varpi}_{a,s}\xrightarrow{p}Var(Y_i(a)-X_i^\top\beta^*_{a,s}|S_i=s).$$

This implies

$$\sum_{s\in\mathcal{S}}\sum_{a=0,1}\frac{n_s^2}{nn_{a,s}}\hat{\omega}_{a,s}^2 \xrightarrow{p} \sum_{s\in\mathcal{S}}p_s\left[\frac{Var(Y_i(1)-X_i^\top\beta_{1,s}|S_i=s)}{\pi_s}+\frac{Var(Y_i(0)-X_i^\top\beta_{0,s}|S_i=s)}{1-\pi_s}\right]$$

and

$$\sum_{s\in\mathcal{S}}\sum_{a=0,1}\frac{n_s^2}{nn_{a,s}}\hat{\varpi}_{a,s} \xrightarrow{p} \sum_{s\in\mathcal{S}}p_s\left[\frac{Var(Y_i(1)-X_i^\top\beta_{1,s}|S_i=s)}{\pi_s}+\frac{Var(Y_i(0)-X_i^\top\beta_{0,s}|S_i=s)}{1-\pi_s}\right].$$

Step 3: Limit of $\hat{\Sigma}_{\mathcal{V}}^{adj}$. We note that

$$\left|\frac{1}{n_{a,s}}\sum_{i\in\aleph_{a,s}}P_{a,s,i,i}Y_i\acute{\varepsilon}_{a,s,i}\right| \le \left(\max_{i\in\aleph_{a,s}}P_{a,s,i,i}\right)\left(\frac{1}{n_{a,s}}\sum_{i\in\aleph_{a,s}}|Y_i\acute{\varepsilon}_{a,s,i}|\right)=o_P(1),$$

$$\frac{1}{n_{a,s}}\hat{\beta}_{a,s}^\top\Gamma_{a,s}\hat{\beta}_{a,s}\xrightarrow{p}\mathbb{E}\left[(X_i-\mathbb{E}(X_i|S_i=s))^\top\beta_{a,s}^*\right]^2,$$

and

$$\frac{1}{n_s}\hat{\beta}_{1,s}^\top\Gamma_s\hat{\beta}_{0,s}\xrightarrow{p}\mathbb{E}\left[(X_i-\mathbb{E}(X_i|S_i=s))^\top\beta_{1,s}^*\right]\left[(X_i-\mathbb{E}(X_i|S_i=s))^\top\beta_{0,s}^*\right].$$

There imply that

$$\hat{\Sigma}_{\mathcal{V}}^{adj}\xrightarrow{p}\mathbb{E}Var(X_i^\top(\beta_{1,s}^*-\beta_{0,s}^*)|S_i).$$

Step 4: $\hat{\Sigma}\xrightarrow{p}\Omega$. We we have already shown in Step 1, we have $\hat{\tau}_{a,s}\xrightarrow{p}\mathbb{E}(Y_i(a)|S_i=s)$. This implies $\hat{\Sigma}_{\mathcal{W}}\xrightarrow{p}var(\mathbb{E}(Y_i(1)|S_i)-\mathbb{E}(Y_i(0)|S_i))$. Therefore, combining results in Steps 2 and 3, we have

$$\begin{aligned}
\hat{\Sigma}\xrightarrow{p}&\left\{\sum_{s\in\mathcal{S}}p_s\left[\frac{Var(Y_i(1)-X_i^\top\beta_{1,s}|S_i=s)}{\pi_s}+\frac{Var(Y_i(0)-X_i^\top\beta_{0,s}|S_i=s)}{1-\pi_s}+\right]\right\}1_21_2^\top\\
&+\begin{pmatrix}\mathbb{E}Var(X_i^\top(\beta_{1,s}^*-\beta_{0,s}^*)|S_i) & \mathbb{E}Var(X_i^\top(\beta_{1,s}^*-\beta_{0,s}^*)|S_i)\\ \mathbb{E}Var(X_i^\top(\beta_{1,s}^*-\beta_{0,s}^*)|S_i) & \mathbb{E}\left[\frac{Var(X_i^\top\beta_{1,s}^*|S_i)}{\pi_{S_i}}+\frac{Var(X_i^\top\beta_{0,s}^*|S_i)}{1-\pi_{S_i}}\right]\end{pmatrix}\\
&+var(\mathbb{E}(Y_i(1)|S_i)-\mathbb{E}(Y_i(0)|S_i))1_21_2^\top\\
&=\sum_{s\in\mathcal{S}}p_s\left\{\frac{\mathbb{E}[Var(Y_i(1)|X_i,S_i)|S_i=s]}{\pi_s}+\frac{\mathbb{E}[Var(Y_i(0)|X_i,S_i)|S_i=s]}{1-\pi_s}\right\}1_21_2^\top\\
&+\sum_{s\in\mathcal{S}}\frac{p_s}{(1-\pi_s)\pi_s}\left\{Var((1-\pi_s)(m_1(X_i,S_i)-X_i^\top\beta_{1,s}^*)|S_i=s)\right\}1_21_2^\top
\end{aligned}$$

$$
\begin{aligned}
&+\sum_{s\in\mathcal{S}}\frac{p_s}{(1-\pi_s)\pi_s}\left\{Var(\pi_s(m_0(X_i,S_i)-X_i^\top\beta_{0,s}^*)|S_i=s)\right\}1_2 1_2^\top\\
&+\sum_{s\in\mathcal{S}}p_s\left\{Var(m_1(X_i,S_i)|S_i=s)+Var(m_0(X_i,S_i)|S_i=s)-2cov(X_i^\top\beta_{1,s}^*,X_i^\top\beta_{1,s}^*\mid S_i=s)\right\}1_2 1_2^\top\\
&+\begin{pmatrix}0 & 0\\ 0 & \mathbb{E}\left[\frac{Var(X_i^\top\beta_{1,s}^*|S_i)}{\pi_{S_i}}+\frac{Var(X_i^\top\beta_{0,s}^*|S_i)}{1-\pi_{S_i}}-Var(X_i^\top(\beta_{1,s}^*-\beta_{0,s}^*|S_i))\right]\end{pmatrix}\\
&+var(\mathbb{E}(Y_i(1)|S_i)-\mathbb{E}(Y_i(0)|S_i))1_2 1_2^\top\\
&=\sum_{s\in\mathcal{S}}p_s\left\{\frac{\mathbb{E}[Var(Y_i(1)|X_i,S_i)|S_i=s]}{\pi_s}+\frac{\mathbb{E}[Var(Y_i(0)|X_i,S_i)|S_i=s]}{1-\pi_s}\right\}1_2 1_2^\top\\
&+\mathbb{E}(m_1(X_i,S_i)-m_1(X_i,S_i)-\tau)^2 1_2 1_2^\top\\
&+\sum_{s\in\mathcal{S}}\frac{p_s}{(1-\pi_s)\pi_s}\left\{Var((1-\pi_s)m_1(X_i,S_i)+\pi_s m_0(X_i,S_i)-X_i^\top\overline{\beta}_s^*)|S_i=s)\right\}1_2 1_2^\top\\
&+\begin{pmatrix}0 & 0\\ 0 & \mathbb{E}\left[\frac{Var(X_i^\top\beta_{1,s}^*|S_i)}{\pi_{S_i}}+\frac{Var(X_i^\top\beta_{0,s}^*|S_i)}{1-\pi_{S_i}}-Var(X_i^\top(\beta_{1,s}^*-\beta_{0,s}^*|S_i))\right]\end{pmatrix}\\
&=\Omega,
\end{aligned}
$$

where the first equality holds by the facts that

$$
\begin{aligned}
&\frac{Var(Y_i(1)-X_i^\top\beta_{1,s}^*|S_i=s)}{\pi_s}\\
&=\frac{\mathbb{E}[Var(Y_i(1)|X_i,S_i)|S_i=s]}{\pi_s}+\frac{Var(m_1(X_i,S_i)-X_i^\top\beta_{1,s}^*|S_i=s)}{\pi_s}\\
&=\frac{\mathbb{E}[Var(Y_i(1)|X_i,S_i)|S_i=s]}{\pi_s}+\frac{Var((1-\pi_s)(m_1(X_i,S_i)-X_i^\top\beta_{1,s}^*)|S_i=s)}{(1-\pi_s)\pi_s}\\
&+Var(m_1(X_i,S_i)|S_i=s)-Var(X_i^\top\beta_{1,s}^*|S_i=s),
\end{aligned}
$$

$$
\begin{aligned}
&\frac{Var(Y_i(0)-X_i^\top\beta_{0,s}^*|S_i=s)}{1-\pi_s}\\
&=\frac{\mathbb{E}[Var(Y_i(0)|X_i,S_i)|S_i=s]}{1-\pi_s}+\frac{Var(\pi_s(m_0(X_i,S_i)-X_i^\top\beta_{0,s}^*)|S_i=s)}{(1-\pi_s)\pi_s}\\
&+Var(m_0(X_i,S_i)|S_i=s)-Var(X_i^\top\beta_{0,s}^*|S_i=s)
\end{aligned}
$$

and

$$
Var(X_i^\top(\beta_{1,s}^*-\beta_{0,s}^*)|S_i=s)-Var(X_i^\top(\beta_{1,s}^*)|S_i=s)-Var(X_i^\top(\beta_{0,s}^*)|S_i=s)
$$

$$
\begin{aligned}
&= -2cov(X_i^\top \beta_{1,s}^*, X_i^\top \beta_{0,s}^* | S_i = s) \\
&= 2cov(m_1(X_i, S_i) - X_i^\top \beta_{1,s}^*, m_0(X_i, S_i) - X_i^\top \beta_{0,s}^* | S_i = s) \\
&- 2cov(m_1(X_i, S_i), m_0(X_i, S_i) | S_i = s),
\end{aligned}
$$

the second equality holds by the fact that

$$
\begin{aligned}
&\sum_{s \in \mathcal{S}} \frac{p_s}{(1-\pi_s)\pi_s} \left\{ Var((1-\pi_s)m_1(X_i, S_i) + \pi_s m_0(X_i, S_i) - X_i^\top \overline{\beta}_s^*) | S_i = s) \right\} \\
&= \sum_{s \in \mathcal{S}} \frac{p_s}{(1-\pi_s)\pi_s} \left\{ Var((1-\pi_s)(m_1(X_i, S_i) - X_i^\top \beta_{1,s}^*) | S_i = s) \right\} \\
&+ \sum_{s \in \mathcal{S}} \frac{p_s}{(1-\pi_s)\pi_s} \left\{ Var(\pi_s(m_0(X_i, S_i) - X_i^\top \beta_{0,s}^*) | S_i = s) \right\} \\
&+ 2cov(m_1(X_i, S_i) - X_i^\top \beta_{1,s}^*, m_0(X_i, S_i) - X_i^\top \beta_{0,s}^* | S_i = s),
\end{aligned}
$$

and the last equality holds by the fact that

$$
\begin{aligned}
&\sum_{s \in \mathcal{S}} \frac{p_s}{(1-\pi_s)\pi_s} \left\{ Var((1-\pi_s)m_1(X_i, S_i) + \pi_s m_0(X_i, S_i) | S_i = s) \right\} \\
&= \mathbb{E}\left[\frac{Var(X_i^\top \beta_{1,s}^* | S_i)}{\pi_{S_i}} + \frac{Var(X_i^\top \beta_{0,s}^* | S_i)}{1-\pi_{S_i}} - Var(X_i^\top (\beta_{1,s}^* - \beta_{0,s}^* | S_i)) \right] \\
&+ \sum_{s \in \mathcal{S}} \frac{p_s}{(1-\pi_s)\pi_s} \left\{ Var((1-\pi_s)m_1(X_i, S_i) + \pi_s m_0(X_i, S_i) - X_i^\top \overline{\beta}_s^*) | S_i = s) \right\}.
\end{aligned}
$$

This concludes the proof of $\hat{\Sigma} \xrightarrow{p} \Omega$.

Step 5: Asymptotic normality of $\hat{\tau}^*$. We see that $\Omega_{1,1} = \Omega_{1,2}$ and

$$
\Omega_{2,2} - \Omega_{1,1} = \sum_{s \in \mathcal{S}} \frac{p_s}{\pi_s(1-\pi_s)} (V_s' - V_s) = \sum_{s \in \mathcal{S}} \frac{p_s}{\pi_s(1-\pi_s)} Var(X_i^\top \overline{\beta}_s^*) > 0.
$$

This implies $\hat{w} \xrightarrow{p} 1$, which leads to the fact that

$$
|\sqrt{n}(\hat{\tau}^* - \hat{\tau}^{adj})| \leq |1-\hat{w}| \left[|\sqrt{n}(\hat{\tau}^{adj} - \tau)| + |\sqrt{n}(\hat{\tau}^{unadj} - \tau)| \right] = o_P(1).
$$

Therefore, we have

$$
\sqrt{n} \left[(\hat{w}, 1-\hat{w}) \hat{\Sigma} (\hat{w}, 1-\hat{w})^\top \right]^{-1/2} (\hat{\tau}^* - \tau) \rightsquigarrow \mathcal{N}(0,1)
$$

and $(\hat{w}, 1-\hat{w}) \hat{\Sigma} (\hat{w}, 1-\hat{w})^\top \xrightarrow{p} \Omega_{1,1} < \Omega_{2,2}$.

F Proof of Theorem 4.2

We prove the result that $\limsup_{n\to\infty}\sup_{\psi_n\in\Psi_n}\mathbb{E}_{\psi_n}(\mathbb{W}_n)=\alpha$. The other one can be proved in the same manner. Throughout the proof, we are under the null, i.e., $\tau_0=\tau$. Suppose the claim does not hold, then there exists a constant $c>0$ and a subsequence $\{n_l\}$ of $\{n\}$ such that

$$\left|\mathbb{E}_{\psi_{n_l}}(\mathbb{W}_{n_l})-\alpha\right|\geq c.$$

By the definition of Λ_n, we can further find a subsequence $\{n_{l'}\}$ of $\{n_l\}$. If for such a subsequence, we have $k_{n'_l}\to\infty$, then Assumptions 1–4 holds and $\Sigma_{1,1}-2\Sigma_{1,2}+\Sigma_{2,2}>0$. Then, by Theorem 3.1, we have

$$\lim_{l'\to\infty}\mathbb{E}_{\psi_{n_{l'}}}(\mathbb{W}_{n_{l'}})=\alpha,$$

which is a contradiction. On the other hand, if there does not exist such a subsequence $\{n_{l'}\}$ with $k_{n_{l'}}\to\infty$, that means we can find a subsequence $\{n_{l'}\}$ such that $k_{n_{l'}}=k$ is fixed. Then, (4.1) implies along $\psi_{n_{l'}}$, Assumptions 1 and 5 hold, and $\pi_s\beta^*_{0,s}+(1-\pi_s)\beta_{1,s}\neq 0$. Then, Theorem 4.1 implies

$$\lim_{l'\to\infty}\mathbb{E}_{\psi_{n_{l'}}}(\mathbb{W}_{n_{l'}})=\alpha,$$

which is also a contradiction.

G Technical Lemmas

Lemma G.1. *Recall the definitions of $\mathcal{U}_n$, $\mathcal{V}_n$, and $\mathcal{W}_n$ in (B.1), (B.2), and (B.3), respectively. Suppose Assumptions 1–3 holds. Then, we have*

$$\mathcal{U}_n\rightsquigarrow\mathcal{U}\stackrel{d}{=}\mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}\mathbb{E}\left(\frac{\omega^2_{1,S_i,\infty}}{\pi_{S_i}}+\frac{\omega^2_{0,S_i,\infty}}{1-\pi_{S_i}}\right) & \mathbb{E}\left(\frac{\varpi_{1,S_i,\infty}}{\pi_{S_i}}+\frac{\varpi_{0,S_i,\infty}}{1-\pi_{S_i}}\right)\\ \mathbb{E}\left(\frac{\varpi_{1,S_i,\infty}}{\pi_{S_i}}+\frac{\varpi_{0,S_i,\infty}}{1-\pi_{S_i}}\right) & \mathbb{E}\left(\frac{\mathbb{E}(\varepsilon_i^2(1)|S_i)}{\pi_{S_i}}+\frac{\mathbb{E}(\varepsilon_i^2(0)|S_i)}{1-\pi_{S_i}}\right)\end{pmatrix}\right),$$

$$\mathcal{V}_n\rightsquigarrow\mathcal{V}\stackrel{d}{=}\mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}\mathbb{E}var(\phi_i(1)-\phi_i(0)|S_i) & \mathbb{E}var(\phi_i(1)-\phi_i(0)|S_i)\\ \mathbb{E}var(\phi_i(1)-\phi_i(0)|S_i) & \mathbb{E}\left[\frac{var(\phi_i(1)|S_i)}{\pi_{S_i}}+\frac{var(\phi_i(0)|S_i)}{(1-\pi_{S_i})}\right]\end{pmatrix}\right),$$

and

$$\mathcal{W}_n\rightsquigarrow\mathcal{W}\stackrel{d}{=}\mathcal{N}(0,var(\mathbb{E}(Y_i(1)|S_i)-\mathbb{E}(Y_i(0)|S_i))).$$

In addition, $(\mathcal{U}, \mathcal{V}, \mathcal{W})$ are independent.

Proof. Following Bugni et al. (2018), we define $\{(X_i^s, \varepsilon_i^s(1), \varepsilon_i^s(0)) : 1 \leq i \leq n\}$ as a sequence of i.i.d. random variables with marginal distributions equal to the distribution of $(X_i, \varepsilon_i(1), \varepsilon_i(0))|S_i = s$ and $N_s = \sum_{i=1}^n 1\{S_i < s\}$. Then, we have $Y_i(a) \mid S_i = s \stackrel{d}{=} Y_i^s(a)$ where

$$Y_i^s(a) = \mu_a(X_i^s, s) + \varepsilon_i^s(a) \quad \text{and} \quad \mu_a(x, s) = \mathbb{E}(Y_i(a)|X_i = x, S_i = s).$$

We further define $\tilde{M}_{a,s,i,j}$ as the (i,j)th entry of the $n_{a,s} \times n_{a,s}$ matrix, $\tilde{M}_{a,s} = \Xi_{a,s}(\Xi_{a,s}^\top \Xi_{a,s})^{-1}\Xi_{a,s}^\top$, $\Xi_{1,s} = ((\breve{X}_{N_s+1}^s)^\top, \cdots, (\breve{X}_{N_s+n_{1,s}}^s)^\top)^\top$ is an $n_{1,s} \times k_n$ matrix, $\Xi_{0,s} = ((\breve{X}_{N_s+1+n_{1,s}}^s)^\top, \cdots, (\breve{X}_{N_s+n_s}^s)^\top)^\top$ is an $n_{0,s} \times k_n$ matrix,

$$\breve{X}_i^s = X_i^s - \frac{1}{n_s}\sum_{j=N_s+1}^{N_s+n_s} X_j^s, \quad \text{if} \quad N_s + 1 \leq i \leq N_s + n_s.$$

$\tilde{\gamma}_{a,s,n} = 1_{n_{a,s}}^\top \tilde{M}_{a,s} 1_{n_{a,s}}$. and $\phi_i^s(a) = \mathbb{E}(Y_i(a)|X_i, S_i = s) - \mathbb{E}(Y_i(a)|S_i = s)$.

Conditional on $(A^{(n)}, S^{(n)})$, the joint distribution of $\mathcal{U}_n$ and $\mathcal{V}_n$ are the same as the counterpart with units ordered by strata and then ordered by $A_i = 1$ first and $A_i = 0$ second within each stratum, i.e.,

$$(\mathcal{U}_n, \mathcal{V}_n)|(A^{(n)}, S^{(n)}) \stackrel{d}{=} (\tilde{\mathcal{U}}_n, \tilde{\mathcal{V}}_n)|(A^{(n)}, S^{(n)}).$$

To see the detailed definition of $\tilde{\mathcal{U}}_n$ and $\tilde{\mathcal{V}}_n$,

Then, we have

$$\tilde{\mathcal{U}}_n = \sum_{s \in \mathcal{S}} \hat{p}_s \sqrt{n} \begin{pmatrix} \tilde{\gamma}_{1,s,n}^{-1} n_{1,s}^{-1} \left[\sum_{i=N_s+1}^{N_s+n_{1,s}} \left(\sum_{j=N_s+1}^{N_s+n_{1,s}} \tilde{M}_{1,s,i,j}\right) \varepsilon_i^s(1)\right] \\ -\tilde{\gamma}_{0,s,n}^{-1} n_{0,s}^{-1} \left[\sum_{i=N_s+n_{1,s}+1}^{N_s+n_s} \left(\sum_{j=N_s+n_{1,s}+1}^{N_s+n_s} \tilde{M}_{0,s,i,j}\right) \varepsilon_i^s(0)\right] \\ n_{1,s}^{-1} \left[\sum_{i=N_s+1}^{N_s+n_{1,s}} \varepsilon_i^s(1)\right] - n_{0,s}^{-1} \left[\sum_{i=N_s+n_{1,s}+1}^{N_s+n_s} \varepsilon_i^s(0)\right] \end{pmatrix},$$

$$\tilde{\mathcal{V}}_n = \sum_{s \in \mathcal{S}} \hat{p}_s \sqrt{n} \begin{pmatrix} \frac{1}{n_s} \sum_{i=N_s+1}^{N_s+n_s} [\phi_i^s(1) - \phi_i^s(0)] \\ \frac{1}{n_{1,s}} \sum_{i=N_s+1}^{N_s+n_{1,s}} \phi_i^s(1) - \frac{1}{n_{0,s}} \sum_{i=N_s+n_{1,s}+1}^{N_s+n_s} \phi_i^s(0) \end{pmatrix},$$

In addition, because $\mathcal{W}_n$ only depends on $S^{(n)}$, it implies the joint distribution of $(\mathcal{U}_n, \mathcal{V}_n, \mathcal{W}_n)$ and $(\tilde{\mathcal{U}}_n, \tilde{\mathcal{V}}_n, \mathcal{W}_n)$ are the same, i.e.,

$$(\mathcal{U}_n, \mathcal{V}_n, \mathcal{W}_n) \stackrel{d}{=} (\tilde{\mathcal{U}}_n, \tilde{\mathcal{V}}_n, \mathcal{W}_n). \tag{G.1}$$

We aim to show that, for any $t_1 \in \Re^2$, $t_2 \in \Re^2$, and $t_3 \in \Re$,

$$|\mathbb{P}\left(\mathcal{U}_n \le t_1, \mathcal{V}_n \le t_2, \mathcal{W}_n \le t_3\right) - \mathbb{P}\left(\mathcal{U} \le t_1\right)\mathbb{P}\left(\mathcal{V} \le t_2\right)\mathbb{P}\left(\mathcal{W} \le t_3\right)| = o(1).$$

By the above definition, we have

$$\begin{aligned}
&|\mathbb{P}\left(\mathcal{U}_n \le t_1, \mathcal{V}_n \le t_2, \mathcal{W}_n \le t_3\right) - \mathbb{P}\left(\mathcal{U} \le t_1\right)\mathbb{P}\left(\mathcal{V} \le t_2\right)\mathbb{P}\left(\mathcal{W} \le t_3\right)| \\
&= \left|\mathbb{P}\left(\tilde{\mathcal{U}}_n \le t_1, \tilde{\mathcal{V}}_n \le t_2, \mathcal{W}_n \le t_3\right) - \mathbb{P}\left(\mathcal{U} \le t_1\right)\mathbb{P}\left(\mathcal{V} \le t_2\right)\mathbb{P}\left(\mathcal{W} \le t_3\right)\right| \\
&= \left|\mathbb{E}\mathbb{P}\left(\tilde{\mathcal{U}}_n \le t_1, \tilde{\mathcal{V}}_n \le t_2 \middle| A^{(n)}, S^{(n)}\right) 1\{\mathcal{W}_n \le t_3\} - \mathbb{P}\left(\mathcal{U} \le t_1\right)\mathbb{P}\left(\mathcal{V} \le t_2\right)\mathbb{P}\left(\mathcal{W} \le t_3\right)\right| \\
&\le \mathbb{E}\left|\mathbb{P}\left(\tilde{\mathcal{U}}_n \le t_1, \tilde{\mathcal{V}}_n \le t_2 \middle| A^{(n)}, S^{(n)}\right) - \mathbb{P}\left(\mathcal{U} \le t_1\right)\mathbb{P}\left(\mathcal{V} \le t_2\right)\right| + |\mathbb{P}(\mathcal{W}_n \le t_3) - \mathbb{P}(\mathcal{W} \le t_3)| \\
&\le \mathbb{E}\left|\mathbb{P}\left(\tilde{\mathcal{U}}_n \le t_1, \tilde{\mathcal{V}}_n \le t_2 \middle| A^{(n)}, S^{(n)}\right) - \mathbb{P}\left(\mathcal{U} \le t_1\right)\mathbb{P}\left(\mathcal{V} \le t_2\right)\right| + o(1),
\end{aligned}$$

where first equality is by (G.1) and the second inequality is because by the Lindeberg central limit theorem, $\mathcal{W}_n \rightsquigarrow \mathcal{W}$. Therefore, it suffices to show

$$\mathbb{P}\left(\tilde{\mathcal{U}}_n \le t_1, \tilde{\mathcal{V}}_n \le t_2 \middle| A^{(n)}, S^{(n)}\right) - \mathbb{P}\left(\mathcal{U} \le t_1\right)\mathbb{P}\left(\mathcal{V} \le t_2\right) \xrightarrow{p} 0.$$

Let $\Xi^{(n)} = \{X_i^s\}_{i\in[n], s\in\mathcal{S}}$. Then, $\tilde{\mathcal{V}}_n$ belongs to the sigma field generated by $(A^{(n)}, S^{(n)}, \Xi^{(n)})$. We have

$$\begin{aligned}
&\left|\mathbb{P}\left(\tilde{\mathcal{U}}_n \le t_1, \tilde{\mathcal{V}}_n \le t_2 \middle| A^{(n)}, S^{(n)}\right) - \mathbb{P}\left(\mathcal{U} \le t_1\right)\mathbb{P}\left(\mathcal{V} \le t_2\right)\right| \\
&\le \mathbb{E}\left[\left|\mathbb{P}\left(\tilde{\mathcal{U}}_n \le t_1 \middle| A^{(n)}, S^{(n)}, \Xi^{(n)}\right) - \mathbb{P}\left(\mathcal{U} \le t_1\right)\right| \middle| A^{(n)}, S^{(n)}\right] + \left|\mathbb{P}\left(\tilde{\mathcal{V}}_n \le t_2 \middle| A^{(n)}, S^{(n)}\right) - \mathbb{P}\left(\mathcal{V} \le t_2\right)\right|.
\end{aligned} \tag{G.2}$$

We show the two terms on the RHS of (G.2) vanish in probability in the following two steps.

Step 1: The first term on the RHS of (G.2). It suffices to show that

$$\mathbb{P}\left(\tilde{\mathcal{U}}_n \le t_1 \middle| A^{(n)}, S^{(n)}, \Xi^{(n)}\right) - \mathbb{P}\left(\mathcal{U} \le t_1\right) = o_P(1).$$

Let

$$Z_i = \begin{cases} \left((n_s/n_{1,s})\tilde{\gamma}_{1,s,n}^{-1}\theta_i\varepsilon_i^s(1), (n_s/n_{1,s})\varepsilon_i^s(1)\right)^\top & \text{if} \quad N_s + 1 \le i \le N_s + n_{1,s} \\ \left((n_s/n_{0,s})\tilde{\gamma}_{0,s,n}^{-1}\theta_i\varepsilon_i^s(0), (n_s/n_{0,s})\varepsilon_i^s(0)\right)^\top & \text{if} \quad N_s + n_{1,s} + 1 \le i \le N_s + n_s \end{cases},$$

where

$$\theta_i = \begin{cases} \sum_{j=N_s+1}^{N_s+n_{1,s}} \tilde{M}_{1,s,i,j} & \text{if} \quad N_s+1 \le i \le N_s+n_{1,s} \\ \sum_{j=N_s+n_{1,s}+1}^{N_s+n_s} \tilde{M}_{0,s,i,j} & \text{if} \quad N_s+n_{1,s}+1 \le i \le N_s+n_s \end{cases}.$$

Then, we have

$$\tilde{\mathcal{U}}_n = \frac{1}{\sqrt{n}} \sum_{i \in [n]} Z_i.$$

By construction,

$$\max_{N_s+1 \le i \le N_s+n_{1,s}} \left| \sum_{j=N_s+1}^{N_s+n_{1,s}} \tilde{M}_{1,s,i,j} \right| \quad \text{and} \quad \max_{i \in \aleph_{1,s}} \left| \sum_{j \in \aleph_{1,s}} M_{1,s,i,j} \right|$$

share the same distribution conditional on $(A^{(n)}, S^{(n)})$, and thus, unconditionally. Same for

$$\max_{N_s+n_{1,s}+1 \le i \le N_s+n_s} \left| \sum_{j=N_s+n_{1,s}+1}^{N_s+n_s} \tilde{M}_{0,s,i,j} \right| \quad \text{and} \quad \max_{i \in \aleph_{0,s}} \left| \sum_{j \in \aleph_{0,s}} M_{0,s,i,j} \right|.$$

Therefore, Assumption 2(vi) implies

$$\max_{i \in [n]} |\theta_i| = o_P(n^{1/2}) \tag{G.3}$$

In addition, we note that $(\hat{p}_s, n_{1,s}, n_{0,s}, n_s, \tilde{\gamma}_{1,s,n}, \tilde{\gamma}_{0,s,n}, \theta_i)$ belong to the sigma field generated by $(A^{(n)}, S^{(n)}, \Xi^{(n)})$ and $\{Z_i\}_{i \in [n]}$ are independent and mean-zero conditional on $(A^{(n)}, S^{(n)}, \Xi^{(n)})$. Define

$$\begin{aligned} H_n &= \frac{1}{n} \sum_{i \in [n]} \mathbb{E}\left(Z_i Z_i^\top \middle| A^{(n)}, S^{(n)}, \Xi^{(n)} \right) \\ &= \frac{1}{n} \sum_{s \in \mathcal{S}} \Bigg[\sum_{i=N_s+1}^{N_s+n_{1,s}} \begin{pmatrix} (n_s/n_{1,s})^2 \tilde{\gamma}_{1,s,n}^{-2} \theta_i^2 \eta_i^2 & (n_s/n_{1,s})^2 \tilde{\gamma}_{1,s,n}^{-1} \theta_i \eta_i^2 \\ (n_s/n_{1,s})^2 \tilde{\gamma}_{1,s,n}^{-1} \theta_i \eta_i^2 & (n_s/n_{1,s})^2 \eta_i^2 \end{pmatrix} \\ &+ \sum_{i=N_s+n_{1,s}+1}^{N_s+n_s} \begin{pmatrix} (n_s/n_{0,s})^2 \tilde{\gamma}_{0,s,n}^{-2} \theta_i^2 \eta_i^2 & (n_s/n_{0,s})^2 \tilde{\gamma}_{0,s,n}^{-1} \theta_i \eta_i^2 \\ (n_s/n_{0,s})^2 \tilde{\gamma}_{0,s,n}^{-1} \theta_i \eta_i^2 & (n_s/n_{0,s})^2 \eta_i^2 \end{pmatrix} \Bigg], \end{aligned}$$

where

$$\eta_i^2 = \begin{cases} \mathbb{E}\left[(\varepsilon_i^s(1))^2 \middle| A^{(n)}, S^{(n)}, \Xi^{(n)}\right] & \text{if} \quad N_s + 1 \le i \le N_s + n_{1,s} \\ \mathbb{E}\left[(\varepsilon_i^s(0))^2 \middle| A^{(n)}, S^{(n)}, \Xi^{(n)}\right] & \text{if} \quad N_s + n_{1,s} + 1 \le i \le N_s + n_s \end{cases}.$$

By construction, we have

$$\eta_i^2 \stackrel{d}{=} \begin{cases} \mathbb{E}(\varepsilon_i^2(1)|X_i, S_i = s) & \text{if} \quad N_s + 1 \le i \le N_s + n_{1,s} \\ \mathbb{E}(\varepsilon_i^2(0)|X_i, S_i = s) & \text{if} \quad N_s + n_{1,s} + 1 \le i \le N_s + n_s \end{cases}.$$

Further define $Z_i = (Z_{i,1}, Z_{i,2})^\top$ and

$$L_n = \max_{\ell=1,2} \frac{1}{n^{3/2}} \sum_{i \in [n]} \mathbb{E}\left(|Z_{i,\ell}^3| \mid A^{(n)}, S^{(n)}, \Xi^{(n)}\right).$$

Then, we have

$$\begin{aligned} L_n &\le \frac{\overline{\sigma}_n^3}{n^{3/2}} \sum_{s \in \mathcal{S}} (n/n_{1,s})^3 \left[\sum_{i=N_s+1}^{N_s+n_{1,s}} \tilde{\gamma}_{1,s,n}^{-3} |\theta_i|^3 + (n/n_{0,s})^3 \sum_{i=N_s+n_{1,s}+1}^{N_s+n_s} \gamma_{0,s,n}^{-3} |\theta_i|^3\right] \\ &\le \frac{\overline{\sigma}_n^3 n^3}{\min_{a=0,1,s\in\mathcal{S}} n_a^3(s) \min_{a=0,1,s\in\mathcal{S}} \tilde{\gamma}_{a,s,n}^2} \frac{\max_{i\in[n]} |\theta_i|}{\sqrt{n}} \\ &\xrightarrow{p} 0, \end{aligned}$$

where $\tilde{\gamma}_{a,s,n} \stackrel{d}{=} \gamma_{a,s,n}$,

$$\overline{\sigma}_n^3 = \max_{i\in[n],a=0,1,s\in\mathcal{S}} \mathbb{E}\left[|\varepsilon_i^s(a)|^3 \mid A^{(n)}, S^{(n)}, \Xi^{(n)}\right] \stackrel{d}{=} \max_{i\in[n],a=0,1,s\in\mathcal{S}} \mathbb{E}(|\varepsilon_i^3(a)| \mid X_i, S_i = s),$$

the second inequality holds by the facts that

$$\sum_{i=N_s+1}^{N_s+n_{1,s}} |\theta_i|^3 \le \max_{i\in[n]} |\theta_i| \sum_{i=N_s+1}^{N_s+n_{1,s}} |\theta_i|^2 = \max_{i\in[n]} |\theta_i| \tilde{\gamma}_{1,s,n}$$

and

$$\sum_{i=N_s+n_{1,s}+1}^{N_s+n_s} |\theta_i|^3 \le \max_{i\in[n]} |\theta_i| \sum_{i=N_s+n_{1,s}+1}^{N_s+n_s} |\theta_i|^2 = \max_{i\in[n]} |\theta_i| \tilde{\gamma}_{0,s,n},$$

and the last convergence is by (G.3).

By the Yurinskii's coupling (Pollard (2002, Theorem 10)), there exists a version of $\breve{\mathcal{U}}_n \stackrel{d}{=}$

$\mathcal{N}(0, H_n)$ and a universal constant C_0 such that

$$\mathbb{P}\left(\left\|\breve{\mathcal{U}}_n - \tilde{\mathcal{U}}_n\right\|_2 \geq 3\delta \middle| A^{(n)}, S^{(n)}, \Xi^{(n)}\right) \leq C_0 2 L_n \delta^{-3}\left(1 + \frac{|\log(1/(2L_n\delta^{-3}))|}{2}\right).$$

Because $L_n \xrightarrow{p} 0$, we have $C_0 2 L_n \delta^{-3}\left(1 + \frac{|\log(1/(2L_n\delta^{-3}))|}{2}\right) \xrightarrow{p} 0$ for any $\delta > 0$, which implies

$$\left\|\breve{\mathcal{U}}_n - \tilde{\mathcal{U}}_n\right\|_2 = o_P(1).$$

Furthermore, note that

$$\begin{aligned}
H_n &= \sum_{s\in\mathcal{S}} \frac{n^2(s)}{n n_{1,s}}\left[\frac{1}{n_{1,s}} \sum_{i=N_s+1}^{N_s+n_{1,s}} \begin{pmatrix} \tilde{\gamma}_{1,s,n}^{-2}\theta_i^2\eta_i^2 & \tilde{\gamma}_{1,s,n}^{-1}\theta_i\eta_i^2 \\ \tilde{\gamma}_{1,s,n}^{-1}\theta_i\eta_i^2 & \eta_i^2 \end{pmatrix}\right] \\
&+ \sum_{s\in\mathcal{S}} \frac{n^2(s)}{n n_{0,s}}\left[\frac{1}{n_{0,s}} \sum_{i=N_s+n_{1,s}+1}^{N_s+n_s} \begin{pmatrix} \tilde{\gamma}_{0,s,n}^{-2}\theta_i^2\eta_i^2 & \tilde{\gamma}_{0,s,n}^{-1}\theta_i\eta_i^2 \\ \tilde{\gamma}_{0,s,n}^{-1}\theta_i\eta_i^2 & \eta_i^2 \end{pmatrix}\right] \\
&\stackrel{d}{=} \sum_{s\in\mathcal{S}} \frac{n^2(s)}{n n_{1,s}}\left[\begin{pmatrix} \gamma_{1,s,n}^{-2}\sigma_{1,s,n}^2 & \gamma_{1,s,n}^{-1}\rho_{1,s,n} \\ \gamma_{1,s,n}^{-1}\rho_{1,s,n} & \tilde{\gamma}_{1,s,n}^{-1}\frac{1}{n_{1,s}}\sum_{i\in\aleph_{1,s}}\left[\mathbb{E}(\varepsilon_i^2(1)|X_i, S_i = s)\right] \end{pmatrix}\right] \\
&+ \sum_{s\in\mathcal{S}} \frac{n^2(s)}{n n_{0,s}}\left[\begin{pmatrix} \gamma_{0,s,n}^{-2}\sigma_{0,s,n}^2 & \tilde{\gamma}_{0,s,n}^{-1}\rho_{0,s,n} \\ \gamma_{0,s,n}^{-1}\rho_{0,s,n} & \frac{1}{n_{0,s}}\sum_{i\in\aleph_{0,s}}\left[\mathbb{E}(\varepsilon_i^2(0)|X_i, S_i = s)\right] \end{pmatrix}\right] \\
&\xrightarrow{p} \sum_{s\in\mathcal{S}}\left[\frac{p_s}{\pi_s}\begin{pmatrix} \omega_{1,s,n}^2 & \varpi_{1,s,n} \\ \varpi_{1,s,n} & \mathbb{E}(\varepsilon_i^2(1)|S_i = s) \end{pmatrix} + \frac{p_s}{1-\pi_s}\begin{pmatrix} \omega_{0,s,n}^2 & \varpi_{0,s,n} \\ \varpi_{0,s,n} & \mathbb{E}(\varepsilon_i^2(0)|S_i = s) \end{pmatrix}\right] \\
&= H,
\end{aligned}$$

where the second last line holds by Assumption 3.

We can write $\breve{\mathcal{U}}_n = H_n^{1/2}\mathcal{G}_2$ and $\mathcal{U} = H^{1/2}\mathcal{G}_2$, where $\mathcal{G}_2$ is a two-dimensional standard Gaussian vector. Note that

$$||H_n^{1/2} - H^{1/2}||_{op} \leq ||H_n - H||_{op}^{1/2} = o_P(1),$$

where the inequality is by Bhatia (2013, Theorem X.1.1) with $f(u) = u^{1/2}$. Let $F_n = \{||H_n - H||_{op}^{1/2} \leq (\delta')^2\}$ for any $(\delta')^2 > 0$. Then, $\mathbb{P}(F_n) \to 1$ and F_n belongs to the sigma field generated by $(A^{(n)}, S^{(n)}, \Xi^{(n)})$. On F_n, we have

$$\begin{aligned}
&\mathbb{P}\left(\breve{\mathcal{U}}_n \leq t_1 \middle| A^{(n)}, S^{(n)}, \Xi^{(n)}\right) - \mathbb{P}\left(U \leq t_1\right) \\
&= \mathbb{P}\left(H_n^{1/2}\mathcal{G}_2 \leq t_1 \middle| A^{(n)}, S^{(n)}, \Xi^{(n)}\right) - \mathbb{P}\left(H^{1/2}\mathcal{G}_2 \leq t_1\right)
\end{aligned}$$

$$\leq \mathbb{P}\left(H^{1/2}\mathcal{G}_2 \leq t + 1_2 ||H_n^{1/2} - H^{1/2}||_{op}||\mathcal{G}_2||_2 \middle| A^{(n)}, S^{(n)}, \Xi^{(n)}\right) - \mathbb{P}(H^{1/2}\mathcal{G}_2 \leq t_1)$$
$$\leq \mathbb{P}\left(H^{1/2}\mathcal{G}_2 \leq t + 1_2\delta'\right) - \mathbb{P}(H^{1/2}\mathcal{G}_2 \leq t_1) + \mathbb{P}(||\mathcal{G}_2||_2 \geq 1/\delta').$$

By Assumption 2(iii), we see that $H_{1,1}$ and $H_{2,2}$ are bounded above from zero. Therefore, by Chernozhukov et al. (2014, Lemma A.1), we have

$$\mathbb{P}\left(H^{1/2}\mathcal{G}_2 \leq t + 1_2\delta'\right) - \mathbb{P}(H^{1/2}\mathcal{G}_2 \leq t_1) \leq C\delta',$$

which implies

$$\mathbb{P}\left(\breve{\mathcal{U}}_n \leq t_1 \middle| A^{(n)}, S^{(n)}, \Xi^{(n)}\right) - \mathbb{P}\left(\mathcal{U} \leq t_1\right) \leq C\delta' + \mathbb{P}(||\mathcal{G}_2||_2 \geq 1/\delta').$$

By a similar argument, we have

$$\mathbb{P}\left(\breve{\mathcal{U}}_n > t_1 \middle| A^{(n)}, S^{(n)}, \Xi^{(n)}\right) - \mathbb{P}\left(\mathcal{U} > t_1\right)$$
$$= \mathbb{P}\left(H_n^{1/2}\mathcal{G}_2 > t_1 \middle| A^{(n)}, S^{(n)}, \Xi^{(n)}\right) - \mathbb{P}\left(H^{1/2}\mathcal{G}_2 > t_1\right)$$
$$\leq \mathbb{P}\left(H^{1/2}\mathcal{G}_2 > t - 1_2 ||H_n^{1/2} - H^{1/2}||_{op}||\mathcal{G}_2||_2 \middle| A^{(n)}, S^{(n)}, \Xi^{(n)}\right) - \mathbb{P}(H^{1/2}\mathcal{G}_2 > t_1)$$
$$\leq \mathbb{P}\left(H^{1/2}\mathcal{G}_2 > t - 1_2\delta'\right) - \mathbb{P}(H^{1/2}\mathcal{G}_2 > t_1) + \mathbb{P}(||\mathcal{G}_2||_2 \geq 1/\delta')$$
$$\leq C\delta' + \mathbb{P}(||\mathcal{G}_2||_2 \geq 1/\delta').$$

Combining these two bounds, we have, on F_n,

$$\left|\mathbb{P}\left(\mathcal{U} \leq t_1\right) - \mathbb{P}\left(\breve{\mathcal{U}}_n \leq t_1 \middle| A^{(n)}, S^{(n)}, \Xi^{(n)}\right)\right| \leq C\delta' + \mathbb{P}(||\mathcal{G}_2||_2 \geq 1/\delta').$$

By letting $n \to \infty$ followed by $\delta' \downarrow 0$, we have the desired result that

$$\mathbb{P}\left(\mathcal{U} \leq t_1\right) - \mathbb{P}\left(\tilde{\mathcal{U}}_n \leq t_1 \middle| A^{(n)}, S^{(n)}, \Xi^{(n)}\right)$$
$$= \mathbb{P}\left(\mathcal{U} \leq t_1\right) - \mathbb{P}\left(\breve{\mathcal{U}}_n \leq t_1 \middle| A^{(n)}, S^{(n)}, \Xi^{(n)}\right)$$
$$+ \mathbb{P}\left(\breve{\mathcal{U}}_n \leq t_1 \middle| A^{(n)}, S^{(n)}, \Xi^{(n)}\right) - \mathbb{P}\left(\tilde{\mathcal{U}}_n \leq t_1 \middle| A^{(n)}, S^{(n)}, \Xi^{(n)}\right) = o_P(1).$$

Step 2: The second term on the RHS of (G.2). We first define

$$\tilde{V}_n^* = \sum_{s\in\mathcal{S}} p_s\sqrt{n}\begin{pmatrix} \frac{1}{np_s}\sum_{i=\lfloor n\mathbb{P}(S<s)\rfloor+1}^{\lfloor n\mathbb{P}(S\le s)\rfloor}[\phi_i^s(1)-\mathbb{E}(\phi_i^s(1))-(\phi_i^s(0)-\mathbb{E}(\phi_i^s(0)))] \\ \frac{1}{n\pi_s p_s}\sum_{i=\lfloor n\mathbb{P}(S<s)\rfloor+1}^{\lfloor n(\mathbb{P}(S<s)+\pi_s p_s)\rfloor}(\phi_i^s(1)-\mathbb{E}(\phi_i^s(1))) \\ -\frac{1}{n(1-\pi_s)p_s}\sum_{i=\lfloor n(\mathbb{P}(S<s)+\pi_s p_s)\rfloor+1}^{\lfloor n(\mathbb{P}(S\le s))\rfloor}(\phi_i^s(0)-\mathbb{E}(\phi_i^s(0))) \end{pmatrix},$$

we note that

$$N_s/n \xrightarrow{p} \mathbb{P}(S<s), \quad n_{1,s}/n \xrightarrow{p} \pi_s p_s, \quad \text{and} \quad n_{0,s}/n \xrightarrow{p} (1-\pi_s)p_s.$$

Because the partial sum process is stochastically equicontinuous, we have $\tilde{\mathcal{V}}_n = \tilde{\mathcal{V}}_n^* + o_P(1)$. By construction, we have $\tilde{\mathcal{V}}_n^* \perp\!\!\!\perp (A^{(n)}, X^{(n)})$, and by the Lindeberg CLT, $\tilde{\mathcal{V}}_n^* \rightsquigarrow V$. In particular, we see that the limit distribution of $\tilde{\mathcal{V}}_n^*$

$$\begin{aligned}&\mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \sum_{s\in\mathcal{S}} p^2(s)\begin{pmatrix}\frac{var(\phi_i^s(1)-\phi_i^s(0))}{p_s} & \frac{var(\phi_i^s(1)-\phi_i^s(0))}{p_s}\\ \frac{var(\phi_i^s(1)-\phi_i^s(0))}{p_s} & \frac{var(\phi_i^s(1))}{\pi_s p_s}+\frac{var(\phi_i^s(0))}{(1-\pi_s)p_s}\end{pmatrix}\right)\\ &\stackrel{d}{=}\mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\mathbb{E}var(\phi_i(1)-\phi_i(0)|S_i) & \mathbb{E}var(\phi_i(1)-\phi_i(0)|S_i)\\ \mathbb{E}var(\phi_i(1)-\phi_i(0)|S_i) & \mathbb{E}\left[\frac{var(\phi_i(1)|S_i)}{\pi_{S_i}}+\frac{var(\phi_i(0)|S_i)}{(1-\pi_{S_i})}\right]\end{pmatrix}\right).\end{aligned}$$

Therefore, we have

$$\begin{aligned}&\mathbb{P}\left(\tilde{\mathcal{V}}_n \le t_2 \middle| A^{(n)}, S^{(n)}\right)\\ &\le \mathbb{P}\left(\tilde{\mathcal{V}}_n^* \le t_2+\delta \middle| A^{(n)}, S^{(n)}\right) + \mathbb{P}\left(|\tilde{\mathcal{V}}_n - \tilde{\mathcal{V}}_n^*| \ge \delta \middle| A^{(n)}, S^{(n)}\right)\\ &= \mathbb{P}\left(\tilde{\mathcal{V}}_n^* \le t_2+\delta\right) + \mathbb{P}\left(|\tilde{\mathcal{V}}_n - \tilde{\mathcal{V}}_n^*| \ge \delta \middle| A^{(n)}, S^{(n)}\right)\\ &= \mathbb{P}\left(\tilde{\mathcal{V}}_n^* \le t_2+\delta\right) + o_P(1)\\ &= \mathbb{P}\left(\mathcal{V} \le t_2+\delta\right) + o_P(1),\end{aligned}$$

where the first equality is by the fact that $\tilde{\mathcal{V}}_n^*$ is independent of $(A^{(n)}, S^{(n)})$ and the second equality holds because by Markov's inequality, for any $\delta' > 0$,

$$\mathbb{P}\left(\mathbb{P}\left(|\tilde{\mathcal{V}}_n - \tilde{\mathcal{V}}_n^*| \ge \delta \middle| A^{(n)}, S^{(n)}\right) \ge \delta'\right) \le \frac{\mathbb{P}\left(|\tilde{\mathcal{V}}_n - \tilde{\mathcal{V}}_n^*| \ge \delta\right)}{\delta'} \to 0.$$

Similarly, we can show that

$$\mathbb{P}\left(\tilde{\mathcal{V}}_n > t_2 \middle| A^{(n)}, S^{(n)}\right) \le \mathbb{P}\left(\mathcal{V} > t_2-\delta\right) + o_P(1),$$

or equivalently,

$$\mathbb{P}\left(\tilde{\mathcal{V}}_n \leq t_2 \middle| A^{(n)}, S^{(n)}\right) \geq \mathbb{P}\left(\mathcal{V} \leq t_2 - \delta\right) + o_P(1),$$

By letting $n \to \infty$ followed by $\delta \downarrow 0$, we have

$$\left|\mathbb{P}\left(\tilde{\mathcal{V}}_n \leq t_2 \middle| A^{(n)}, S^{(n)}\right) - \mathbb{P}\left(\mathcal{V} \leq t_2\right)\right| \xrightarrow{p} 0.$$

This concludes the proof.

□

Lemma G.2. *Suppose Assumptions 1–4 hold and recall I_1 defined in (C.2). Then, we have*

$$\varepsilon_{\aleph_{a,s}}^\top(a) P_{a,s} \varepsilon_{\aleph_{a,s}}(a) - \sum_{i \in \aleph_{a,s}} P_{a,s,i,i} Y_i \acute{\varepsilon}_{a,s,i} = o_P(n).$$

Proof. Let $H_{a,s,i,j} = M_{a,s,i,j} - (\sum_{j \in \aleph_{a,s}} M_{a,s,i,j})\gamma_{a,s,\infty}^{-1}(\sum_{i \in \aleph_{a,s}} M_{a,s,i,j})$ and $H_{a,s}$ be the $n_{a,s} \times n_{a,s}$ matrix with its (i,j)th entry being $H_{a,s,i,j}$. We can see that $\sum_{j \in \aleph_{a,s}} H_{a,s,i,j} H_{a,s,j,k} = H_{a,s,i,k}$, which means $H_{a,s}$ is idempotent. In addition, we $\hat{\varepsilon}_{a,s,i} = \sum_{j \in \aleph_{a,s}} H_{a,s,i,j}(e_{j,s}(a) + \varepsilon_j(a))$. Further denote $\tilde{H}_{a,s,i,j} = H_{a,s,i,j}/M_{a,s,i,i}$ and $\mu_i(a) = \mathbb{E}(Y_i(a)|X_i, S_i)$.

This implies

$$\begin{aligned}
I_1 &= \sum_{i \in \aleph_{a,s}} \sum_{j \in \aleph_{a,s}, j \neq i} \varepsilon_i(a) P_{a,s,i,j} \varepsilon_j(a) + \sum_{i \in \aleph_{a,s}} P_{a,s,i,i}(\varepsilon_i^2(a) - Y_i \acute{\varepsilon}_{a,s,i}) \\
&= \left[\sum_{i \in \aleph_{a,s}} P_{a,s,i,i} \varepsilon_i^2(a)(1 - \tilde{H}_{a,s,i,i})\right] + \left[\sum_{i \in \aleph_{a,s}} \sum_{j \in \aleph_{a,s}, j \neq i} \varepsilon_i(a) P_{a,s,i,j} \varepsilon_i(a)\right] \\
&- \left[\sum_{i \in \aleph_{a,s}} \sum_{j \in \aleph_{a,s}, j \neq i} P_{a,s,i,i} \tilde{H}_{a,s,i,j} \varepsilon_i(a) \varepsilon_j(a)\right] - \left[\sum_{i \in \aleph_{a,s}} \sum_{j \in \aleph_{a,s}} P_{a,s,i,i} \tilde{H}_{a,s,i,j} \mu_i(a) e_{j,s}(a)\right] \\
&- \left[\sum_{i \in \aleph_{a,s}} \sum_{j \in \aleph_{a,s}} P_{a,s,i,i} \tilde{H}_{a,s,i,j} \mu_i(a) \varepsilon_j(a)\right] - \left[\sum_{i \in \aleph_{a,s}} \sum_{j \in \aleph_{a,s}} P_{a,s,i,i} \tilde{H}_{a,s,i,j} \varepsilon_i(a) e_{j,s}(a)\right] \\
&\equiv I_{1,1} + I_{1,2} - I_{1,3} - I_{1,4} - I_{1,5} - I_{1,6}.
\end{aligned}$$

Conditional on $(A^{(n)}, S^{(n)}, X^{(n)})$, $\aleph_{a,s}$, $P_{a,s}$, $\tilde{H}_{a,s}$, $\phi_i(a)$, and $e_{i,s}(a)$ are all deterministic and $\{\varepsilon_i(a)\}_{i \in \aleph_{a,s}}$ are independent across i, conditionally mean zero, and $\varepsilon_i(a)|(A^{(n)}, S^{(n)}, X^{(n)}) \stackrel{d}{=}$

$\varepsilon_i(a)|X_i, S_i = s$. For $I_{1,1}$, we have

$$|I_{1,1}| = \sum_{i \in \aleph_{a,s}} P_{a,s,i,i} \varepsilon_i^2(a) \frac{(\sum_{j \in \aleph_{a,s}} M_{a,s,i,j})^2}{1_{\aleph_{a,s}}^\top M_{a,s} 1_{\aleph_{a,s}} M_{a,s,i,i}}$$
$$\leq \left(\sum_{i \in \aleph_{a,s}} P_{a,s,i,i} \varepsilon_i^2(a) \right) \left(\frac{\max_{i \in \aleph_{a,s}} (\sum_{j \in \aleph_{a,s}} M_{a,s,i,j})^2}{n} \right) \left(\frac{n}{n_{a,s} \gamma_{a,s,n} \min_{i \in \aleph_{a,s}} M_{a,s,i,i}} \right) = o_P(n)$$

where we use Assumptions 2(vi), 3, 4, and the facts that $\min_{i \in \aleph_{a,s}} M_{a,s,i,i} \geq \delta > 0$ which is bounded away from zero, implying that $1/\min_{i \in \aleph_{a,s}} M_{a,s,i,i} = O_P(1)$ and

$$\left(\sum_{i \in \aleph_{a,s}} P_{a,s,i,i} \varepsilon_i^2(a) \right) = O_P(n).$$

For $I_{1,2}$, we have

$$\mathbb{E}(I_{1,2}|A^{(n)}, S^{(n)}, X^{(n)}) = 0$$

and

$$var(I_{1,2}|A^{(n)}, S^{(n)}, X^{(n)}) = \sum_{i \in \aleph_{a,s}} \sum_{j \in \aleph_{a,s}, j \neq i} \mathbb{E}(\varepsilon_i^2(a)|A^{(n)}, S^{(n)}, X^{(n)}) P_{a,s,i,j}^2 \mathbb{E}(\varepsilon_j^2(a)|A^{(n)}, S^{(n)}, X^{(n)})$$
$$\leq \sum_{i \in \aleph_{a,s}} \sum_{j \in \aleph_{a,s}} P_{a,s,i,j}^2 \max_{i \in [n]} \mathbb{E}(\varepsilon_i^2(a)|X_i, S_i) = O_P(n),$$

which implies

$$I_{1,2} = O_P(n^{1/2}) = o_P(n).$$

Similarly, we have

$$\mathbb{E}(I_{1,3}|A^{(n)}, S^{(n)}, X^{(n)}) = 0$$

and

$$var(I_{1,3}|A^{(n)}, S^{(n)}, X^{(n)}) \lesssim \sum_{i \in \aleph_{a,s}} \sum_{j \in \aleph_{a,s}} \tilde{H}_{a,s,i,j}^2 P_{a,s,i,i}^2 \max_{i \in [n]} \mathbb{E}(\varepsilon_i^2(a)|X_i, S_i)$$
$$\lesssim \sum_{i \in \aleph_{a,s}} \sum_{j \in \aleph_{a,s}} H_{a,s,i,j}^2 (\min_{i \in \aleph_{a,s}} M_{a,s,i,i})^{-2} \max_{i \in [n]} \mathbb{E}(\varepsilon_i^2(a)|X_i, S_i)$$
$$= O_P(n),$$

where the second inequality holds by the fact that $P^2_{a,s,i,i} \leq 1$. This implies $I_{1,3} = o_P(n)$.

For $I_{1,4}$, we have

$$\begin{aligned}
I^2_{1,4} &\leq \left(\sum_{i\in\aleph_{a,s}} P^2_{a,s,i,i}\mu^2_i(a)M^{-2}_{a,s,i,i}\right)\left(\sum_{i\in\aleph_{a,s}}\left(\sum_{j\in\aleph_{a,s}} H_{a,s,i,j}e_{j,s}(a)\right)^2\right) \\
&\leq \left(\sum_{i\in\aleph_{a,s}} P^2_{a,s,i,i}M^{-2}_{a,s,i,i}\mu^2_i(a)\right)\left(\sum_{i\in\aleph_{a,s}}\left(\sum_{j\in\aleph_{a,s}} H_{a,s,i,j}e_{j,s}(a)\right)^2\right) \\
&\leq \left(\max_{i\in\aleph_{a,s}} P^2_{a,s,i,i}M^{-2}_{a,s,i,i}\right)\left(\sum_{i\in\aleph_{a,s}}\mu^2_i(a)\right)\left(\sum_{j\in\aleph_{a,s}} e^2_{j,s}(a)\right) = o_P(n),
\end{aligned}$$

where the last inequality is by $||H_{a,s}||_{op} \leq 1$ and the last equality is by the facts that

$$\left(\sum_{i\in\aleph_{a,s}}\mu^2_i(a)\right) = O_P(n), \quad \max_{i\in\aleph_{a,s}} P^2_{a,s,i,i}M^{-2}_{a,s,i,i} = O_P(1), \quad \left(\sum_{j\in\aleph_{a,s}} e^2_{j,s}(a)\right) = o_P(1).$$

This implies $I_{1,4} = o_P(n^{1/2}) = o_P(n)$.

For $I_{1,5}$, we have

$$\mathbb{E}(I_{1,5}|A^{(n)}, S^{(n)}, X^{(n)}) = 0$$

and

$$\begin{aligned}
&var(I_{1,5}|A^{(n)}, S^{(n)}, X^{(n)}) \\
&= \sum_{j\in\aleph_{a,s}}\left[\sum_{i\in\aleph_{a,s}} P_{a,s,i,i}\tilde{H}_{a,s,i,j}\mu_i(a)\right]^2 \mathbb{E}(\varepsilon^2_j(a)|A^{(n)}, S^{(n)}, X^{(n)}) \\
&\leq \sum_{j\in\aleph_{a,s}}\sum_{i\in\aleph_{a,s}} \frac{P_{a,s,i,i}\mu_i(a)}{M_{a,s,i,i}} H_{a,s,i,j}\frac{P_{a,s,j,j}\mu_j(a)}{M_{a,s,j,j}}\left[\max_{j\in\aleph_{a,s}}\mathbb{E}(\varepsilon^2_j(a)|A^{(n)}, S^{(n)}, X^{(n)})\right] \\
&\leq \sum_{i\in\aleph_{a,s}} \frac{P^2_{a,s,i,i}\mu^2_i(a)}{M^2_{a,s,i,i}}\left[\max_{j\in\aleph_{a,s}}\mathbb{E}(\varepsilon^2_j(a)|A^{(n)}, S^{(n)}, X^{(n)})\right] \\
&\leq \left(\sum_{i\in\aleph_{a,s}}\mu^2_i(a)\right)\left(\max_{i\in\aleph_{a,s}}\frac{P^2_{a,s,i,i}}{M^2_{a,s,i,i}}\right)\left(\max_{j\in\aleph_{a,s}}\mathbb{E}(\varepsilon^2_j(a)|A^{(n)}, S^{(n)}, X^{(n)})\right) \\
&= O_P(n),
\end{aligned}$$

where the second inequality is by $||H_{a,s}||_{op} \leq 1$. This implies $I_{1,5} = O_P(n^{1/2}) = o_P(n)$. In the same manner, we can show that $I_{1,6} = o_P(n)$, which concludes the proof of this lemma. □

References

Anderson, S. J. and D. McKenzie (2021). Improving business practices and the boundary of the entrepreneur: a randomized experiment comparing training, consulting, insourcing and outsourcing. *Journal of Political Economy, 130*(1), 157–209.

Bai, Y., L. Jiang, J. P. Romano, A. M. Shaikh, and Y. Zhang (2023). Covariate adjustment in experiments with matched pairs. *arXiv preprint arXiv:2302.04380*.

Bai, Z. D. (2008). Methodologies in spectral analysis of large dimensional random matrices, a review. In *Advances in statistics*, pp. 174–240. World Scientific.

Belloni, A., V. Chernozhukov, D. Chetverikov, and K. Kato (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics 186*(2), 345–366.

Bhatia, R. (2013). *Matrix analysis*, Volume 169. Springer Science & Business Media.

Bloniarz, A., H. Liu, C.-H. Zhang, J. S. Sekhon, and B. Yu (2016). Lasso adjustments of treatment effect estimates in randomized experiments. *Proceedings of the National Academy of Sciences 113*(27), 7383–7390.

Bugni, F. A., I. A. Canay, and A. M. Shaikh (2018). Inference under covariate-adaptive randomization. *Journal of the American Statistical Association 113*(524), 1741–1768.

Bugni, F. A., I. A. Canay, and A. M. Shaikh (2019). Inference under covariate-adaptive randomization with multiple treatments. *Quantitative Economics 10*(4), 1747–1785.

Burchardi, K. B., S. Gulesci, B. Lerva, and M. Sulaiman (2019). Moral hazard: experimental evidence from tenancy contracts. *Quarterly Journal of Economics 134*(1), 281–347.

Cattaneo, M. D., M. H. Farrell, and Y. Feng (2020). Large sample properties of partitioning-based series estimators. *Annals of Statistics 48*(3), 1718–1741.

Cattaneo, M. D., M. Jansson, and X. Ma (2019). Two-step estimation and inference with possibly many included covariates. *The Review of Economic Studies 86*(3), 1095–1122.

Cattaneo, M. D., M. Jansson, and W. K. Newey (2018). Inference in linear regression models with many covariates and heteroscedasticity. *Journal of the American Statistical Association 113*(523), 1350–1361.

Chernozhukov, V., D. Chetverikov, and K. Kato (2014). Gaussian approximation of suprema of empirical processes. *The Annals of Statistics 42*(4), 1564–1597.

Chiang, H. D., Y. Matsushita, and T. Otsu (2023). Regression adjustment in randomized controlled trials with many covariates. *arXiv preprint arXiv:2302.00469*.

Chong, A., I. Cohen, E. Field, E. Nakasone, and M. Torero (2016). Iron deficiency and schooling attainment in Peru. *American Economic Journal: Applied Economics 8*(4), 222–55.

Cohen, P. L. and C. B. Fogarty (2020). No-harm calibration for generalized oaxaca-blinder estimators. *arXiv preprint arXiv:2012.09246*.

Cytrynbaum, M. (2023). Covariate adjustment in stratified experiments.

Fogarty, C. B. (2018). Regression-assisted inference for the average treatment effect in paired experiments. *Biometrika 105*(4), 994–1000.

Freedman, D. A. (2008a). On regression adjustments in experiments with several treatments. *Annals of Applied Statistics 2*(1), 176–196.

Freedman, D. A. (2008b). On regression adjustments to experimental data. *Advances in Applied Mathematics 40*(2), 180–193.

Greaney, B. P., J. P. Kaboski, and E. Van Leemput (2016). Can self-help groups really be "self-help"? *Review of Economic Studies 83*(4), 1614–1644.

Jakiela, P. and O. Ozier (2016). Does Africa need a rotten kin theorem? Experimental evidence from village economies. *Review of Economic Studies 83*(1), 231–268.

Jiang, L., O. B. Linton, H. Tang, and Y. Zhang (2022). Improving estimation efficiency via regression-adjustment in covariate-adaptive randomizations with imperfect compliance.

Jiang, L., P. C. Phillips, Y. Tao, and Y. Zhang (2022). Regression-adjusted estimation of quantile treatment effects under covariate-adaptive randomizations. *Journal of Econometrics, forthcoming*.

Jochmans, K. (2022). Heteroscedasticity-robust inference in linear regression models with many covariates. *Journal of the American Statistical Association 117*(538), 887–896.

Kline, P., R. Saggio, and M. Sølvsten (2020). Leave-out estimation of variance components. *Econometrica 88*(5), 1859–1898.

Lei, L. and P. Ding (2021). Regression adjustment in completely randomized experiments with a diverging number of covariates. *Biometrika 108*(4), 815–828.

Li, X. and P. Ding (2020). Rerandomization and regression adjustment. *Journal of the Royal Statistical Society Series B: Statistical Methodology 82*(1), 241–268.

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: reexamining Freedman's critique. *Annals of Applied Statistics 7*(1), 295–318.

Liu, H., F. Tu, and W. Ma (2020). A general theory of regression adjustment for covariate-adaptive randomization: OLS, Lasso, and beyond. *arXiv preprint arXiv: 2011.09734*.

Ma, W., F. Tu, and H. Liu (2022). Regression analysis for covariate-adaptive randomization: a robust and efficient inference perspective. *Statistics in Medicine 41*(29).

Marčenko, V. A. and L. A. Pastur (1967). Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik 1*(4), 457.

Pollard, D. (2002). *A user's guide to measure theoretic probability.* Number 8. Cambridge University Press.

Reluga, K., T. Ye, and Q. Zhao (2022). A unified analysis of regression adjustment in randomized experiments. *arXiv preprint arXiv:2210.04360*.

Shao, J., X. Yu, and B. Zhong (2010). A theory for testing hypotheses under covariate-adaptive randomization. *Biometrika 97*(2), 347–360.

Wei, L. (1978). An application of an urn model to the design of sequential controlled clinical trials. *Journal of the American Statistical Association 73*(363), 559–563.

Ye, T., J. Shao, Y. Yi, and Q. Zhao (2022). Toward better practice of covariate adjustment in analyzing randomized clinical trials. *Journal of the American Statistical Association, forthcoming*, 1–13.

Ye, T., Y. Yi, and J. Shao (2022). Inference on the average treatment effect under minimization and other covariate-adaptive randomization methods. *Biometrika 109*(1), 33–47.

Zhang, Y. and X. Zheng (2020). Quantile treatment effects and bootstrap inference under covariate-adaptive randomization. *Quantitative Economics 11*(3), 957–982.