Supplementary material for "A unified inference framework for multiple imputation using martingales"

Qian Guan and Shu Yang

Department of Statistics, North Carolina State University

S1 Common assumptions for ACE estimation

We make the following assumptions that are common in the causal inference literature.

Assumption 1 (Treatment ignorability)**.** $A \perp\!\!\!\perp \{Y(0), Y(1)\} \mid X$.

Assumption 1 rules out latent confounding between the treatment and outcome. It holds by the design of a completely randomized experiment, where the treatment is independent of all the potential values and covariates. It also holds by the design of a stratified experiment based on a discrete X, where the treatment is independent of the potential outcomes within each stratum of X. In observational studies, its plausibility relies on whether or not the observed covariates X include all the confounders that affect the treatment as well as the outcome. Under Assumption 1, $\mu_a(X)$ is identified by $\mathbb{E}(Y \mid A = a, X)$.

Assumption 2 (Overlap)**.** $0 < c_1 \leq e(X) \leq c_2 < 1$ almost surely for some c_1 and c_2.

Assumption 2 implies a sufficient overlap of the covariate distribution between the treatment groups. Yang and Ding (2018) suggested trimming the sample when it is violated.

S2 Influence functions of common estimators

Likelihood-based or equivalently score-based methods can be used to estimate the model parameters. For the outcome model, let $S_a(A, X, Y; \beta_a)$ be the estimating function for β_a^*, e.g.,

$$S_a(A, X, Y; \beta_a) = \frac{\partial \mu_a(X; \beta_a)}{\partial \beta_a}\{Y - \mu_a(X; \beta_a)\},$$

for $a = 0, 1$, which is a standard choice for the conditional mean model. For the propensity score model, let $S(A, X; \alpha)$ be the estimating function for α, e.g.,

$$S(A, X; \alpha) = \frac{A - e(X; \alpha)}{e(X; \alpha)\{1 - e(X; \alpha)\}} \frac{\partial e(X; \alpha)}{\partial \alpha},$$

which is the score function from the likelihood of a binary response model. Moreover, let

$$\Sigma_{\alpha\alpha} = \mathbb{E}\left\{S^{\otimes 2}(A, X; \alpha)\right\} = \mathbb{E}\left[\frac{1}{e(X; \alpha^*)\{1 - e(X; \alpha^*)\}}\left\{\frac{\partial e(X; \alpha^*)}{\partial \alpha}\right\}^{\otimes 2}\right]$$

be the Fisher information matrix for α in the propensity score model, where $S^{\otimes 2} = SS^T$. In addition, let $\hat{\beta}_a$ $(a = 0, 1)$ and $\hat{\alpha}$ be the estimators solving the corresponding empirical estimating equations, with probability limits β_a^* $(a = 0, 1)$ and α^*, respectively.

Below we review influence functions for the ACE estimators presented in Examples 1–4. We first consider the outcome regression, IPW, and AIPW estimators which belong to the class of RAL estimators. Their influence functions are not simply $\tau_{\text{reg},i}$, $\tau_{\text{IPW},i}$, and $\tau_{\text{AIPW},i}$ in (2.3)–(2.5) because they depend on the estimated model parameters in the outcome and/or propensity score model. To derive the influence functions, one can use the standard Taylor expansion technique. Lunceford and Davidian (2004) derived the results in Lemmas 1 and 2 below.

Lemma 1 (Outcome regression)**.** *Under Assumption 1, $\hat{\tau}_{n,\text{reg}}$ has the influence function*

$$\begin{aligned}
&\psi_{\text{reg}}(A_i, X_i, Y_i) = \mu_1(X_i; \beta_1^*) - \mu_0(X_i; \beta_0^*) - \tau \\
&-\mathbb{E}\left\{\frac{\partial \mu_1(X; \beta_1^*)}{\partial \beta_1^{\text{T}}}\right\} \mathbb{E}\left\{\frac{\partial S_1(A, X, Y; \beta_1^*)}{\partial \beta_1^{\text{T}}}\right\}^{-1} S_1(A_i, X_i, Y_i; \beta_1^*) \\
&+\mathbb{E}\left\{\frac{\partial \mu_0(X; \beta_0^*)}{\partial \beta_0^{\text{T}}}\right\} \mathbb{E}\left\{\frac{\partial S_0(A, X, Y; \beta_0^*)}{\partial \beta_0^{\text{T}}}\right\}^{-1} S_0(A_i, X_i, Y_i; \beta_0^*).
\end{aligned}$$

Lemma 2 (Inverse probability weighting)**.** *Under Assumption 2, $\hat{\tau}_{n,\text{IPW}}$*

has the influence function

$$\psi_{\mathrm{IPW}}(A_i, X_i, Y_i) = \frac{A_i Y_i}{e(X_i;\alpha^*)} - \frac{(1-A_i)Y_i}{1-e(X_i;\alpha^*)} - \tau - H_{\mathrm{IPW}}\Sigma_{\alpha\alpha}^{-1} S(A_i, X_i;\alpha^*),$$

where

$$H_{\mathrm{IPW}} = \mathbb{E}\left(\left[\frac{AY}{e(X;\alpha^*)^2} - \frac{(1-A)Y}{\{1-e(X;\alpha^*)\}^2}\right]\frac{\partial e(X;\alpha^*)}{\partial \alpha}\right).$$

Lemma 3 (Augmented inverse probability weighting)**.** *Under Assumption 1 or 2, $\hat{\tau}_{n,\mathrm{AIPW}}$ has the influence function*

$$\begin{aligned}
&\psi_{\mathrm{AIPW}}(A_i, X_i, Y_i) = \frac{A_i Y_i}{e(X_i;\alpha^*)} + \left\{1 - \frac{A_i}{e(X_i;\alpha^*)}\right\}\mu_1(X_i;\beta_1^*) \\
&-\frac{(1-A_i)Y_i}{1-e(X_i;\alpha^*)} - \left\{1 - \frac{1-A_i}{1-e(X_i;\alpha^*)}\right\}\mu_0(X_i;\beta_0^*) - \tau + H_{\mathrm{AIPW}}^{\mathrm{T}}\Sigma_{\alpha\alpha}^{-1} S(A_i, X_i;\alpha^*) \\
&+\mathbb{E}\left\{\frac{A - e(X;\alpha^*)}{1-e(X;\alpha^*)}\frac{\partial \mu_0(X;\beta_0^*)}{\partial \beta_0^{\mathrm{T}}}\right\}\mathbb{E}\left\{\frac{\partial S_0(A,X,Y;\beta_0^*)}{\partial \beta_0^{\mathrm{T}}}\right\}^{-1} S_0(A_i, X_i, Y_i;\beta_0^*) \quad \text{(S2.1)} \\
&-\mathbb{E}\left\{\frac{e(X;\alpha^*) - A}{e(X;\alpha^*)}\frac{\partial \mu_1(X;\beta_1^*)}{\partial \beta_1^{\mathrm{T}}}\right\}\mathbb{E}\left\{\frac{\partial S_1(A,X,Y;\beta_1^*)}{\partial \beta_1^{\mathrm{T}}}\right\}^{-1} S_1(A_i, X_i, Y_i;\beta_1^*), \quad \text{(S2.2)}
\end{aligned}$$

where

$$H_{\mathrm{AIPW}} = \mathbb{E}\left(\left[\frac{A\{Y-\mu_1(X;\beta_1^*)\}}{e(X;\alpha^*)^2} - \frac{(1-A)\{Y-\mu_0(X;\beta_0^*)\}}{\{1-e(X;\alpha^*)\}^2}\right]\frac{\partial e(X;\alpha^*)}{\partial \alpha}\right).$$

The derivation of Lemma 3 is provided in the supplementary material. Lunceford and Davidian (2004) suggested using the efficient influence function without (S2.1) and (S2.2) for $\psi_{\mathrm{AIPW}}(A_i, X_i, Y_i)$, which, however, works only when both Assumptions 1 and 2 hold. The influence function $\psi_{\mathrm{AIPW}}(A_i, X_i, Y_i)$ in Lemma 3 is agnostic about whether Assumption 1 or 2 holds.

We then consider the matching estimators which belong to the class of non-RAL estimators.

Lemma 4. *Continuing with Example 4, the bias-corrected matching estimator $\hat{\tau}_{n,\mathrm{mat}}$ can be expressed in a linear form (Abadie and Imbens, 2006) as*

$$\hat{\tau}_{n,\mathrm{mat}} - \tau = \frac{1}{n}\sum_{i=1}^{n} \psi_{\mathrm{mat},i} + o_{\mathbb{P}}(1), \tag{S2.3}$$

where

$$\psi_{\mathrm{mat},i} = \mu_1(X_i) - \mu_0(X_i) - \tau + (2A_i - 1)\left(1 + M^{-1}K_{X,i}\right)\left\{Y_i - \mu_{A_i}(X_i)\right\}. \tag{S2.4}$$

Although the matching estimators can be expressed in an asymptotically linear form, they are not regular because the functional forms are not smooth due to the fixed numbers of matches (Abadie and Imbens, 2008). Also, the number of times unit i is being matched to other units $K_{X,i}$ depends on the full sample, and therefore $\{\psi_{\mathrm{mat},i} : i = 1, \ldots, n\}$ are not i.i.d.

S3 Proof of Lemma 3

We derive the influence function of the AIPW estimator $\hat{\tau}_{n,\mathrm{AIPW}}$ under Assumption 1 or Assumption 2. We write $\hat{\tau}_{n,\mathrm{AIPW}} = \hat{\tau}_{n,\mathrm{AIPW}}(\hat{\alpha}, \hat{\beta}_0, \hat{\beta}_1)$ to emphasize its dependence on the parameter estimates $(\hat{\alpha}, \hat{\beta}_0, \hat{\beta}_1)$. By the

Taylor expansion,

$$
\begin{aligned}
& \hat{\tau}_{n,\mathrm{AIPW}}(\hat{\alpha}, \hat{\beta}_0, \hat{\beta}_1) \\
\cong\ & \hat{\tau}_{n,\mathrm{AIPW}}(\alpha^*, \beta_0^*, \beta_1^*) + \mathbb{E}\left\{\frac{\partial \hat{\tau}_{n,\mathrm{AIPW}}(\alpha^*, \beta_0^*, \beta_1^*)}{\partial \alpha^{\mathrm{T}}}\right\}(\hat{\alpha} - \alpha^*) \\
& + \mathbb{E}\left\{\frac{\partial \hat{\tau}_{n,\mathrm{AIPW}}(\alpha^*, \beta_0^*, \beta_1^*)}{\partial \beta_0^{\mathrm{T}}}\right\}(\hat{\beta}_0 - \beta_0^*) + \mathbb{E}\left\{\frac{\partial \hat{\tau}_{n,\mathrm{AIPW}}(\alpha^*, \beta_0^*, \beta_1^*)}{\partial \beta_1^{\mathrm{T}}}\right\}(\hat{\beta}_1 - \beta_1^*) \\
\cong\ & \hat{\tau}_{n,\mathrm{AIPW}}(\alpha^*, \beta_0^*, \beta_1^*) \\
& + n^{-1}\sum_{i=1}^{n} \mathbb{E}\left\{\frac{\partial \hat{\tau}_{n,\mathrm{AIPW}}(\alpha^*, \beta_0^*, \beta_1^*)}{\partial \alpha^{\mathrm{T}}}\right\} \mathbb{E}\left\{\frac{\partial S(A, X; \alpha^*)}{\partial \alpha^{\mathrm{T}}}\right\}^{-1} S(A_i, X_i; \alpha^*) \quad \text{(S3.5)} \\
& - n^{-1}\sum_{i=1}^{n} \mathbb{E}\left\{\frac{\partial \hat{\tau}_{n,\mathrm{AIPW}}(\alpha^*, \beta_0^*, \beta_1^*)}{\partial \beta_0^{\mathrm{T}}}\right\} \mathbb{E}\left\{\frac{\partial S_0(A, X, Y; \beta_0^*)}{\partial \beta_0^{\mathrm{T}}}\right\}^{-1} S_0(A_i, X_i, Y_i; \beta_0^*) \\
& - n^{-1}\sum_{i=1}^{n} \mathbb{E}\left\{\frac{\partial \hat{\tau}_{n,\mathrm{AIPW}}(\alpha^*, \beta_0^*, \beta_1^*)}{\partial \beta_1^{\mathrm{T}}}\right\} \mathbb{E}\left\{\frac{\partial S_1(A, X, Y; \beta_1^*)}{\partial \beta_1^{\mathrm{T}}}\right\}^{-1} S_1(A_i, X_i, Y_i; \beta_1^*).
\end{aligned}
$$

We have the following calculations:

$$
\begin{aligned}
\mathbb{E}\left\{\frac{\partial \hat{\tau}_{n,\mathrm{AIPW}}(\alpha^*, \beta_0^*, \beta_1^*)}{\partial \alpha}\right\} &= H_{\mathrm{AIPW}}, \\
\mathbb{E}\left\{\frac{\partial \hat{\tau}_{n,\mathrm{AIPW}}(\alpha^*, \beta_0^*, \beta_1^*)}{\partial \beta_0}\right\} &= -\mathbb{E}\left[\left\{1 - \frac{1-A}{1-e(X;\alpha^*)}\right\}\frac{\partial \mu_0(X;\beta_0^*)}{\partial \beta_0}\right], \\
\mathbb{E}\left\{\frac{\partial \hat{\tau}_{n,\mathrm{AIPW}}(\alpha^*, \beta_0^*, \beta_1^*)}{\partial \beta_1}\right\} &= \mathbb{E}\left[\left\{1 - \frac{A}{e(X;\alpha^*)}\right\}\frac{\partial \mu_1(X;\beta_1^*)}{\partial \beta_1}\right].
\end{aligned}
$$

Under Assumption 1, $H_{\mathrm{AIPW}} = 0$. Under Assumption 2,

$$
\mathbb{E}\left\{\frac{\partial S(A, X; \alpha^*)}{\partial \alpha^{\mathrm{T}}}\right\} = \mathbb{E}\left\{S^{\otimes 2}(A, X; \alpha^*)\right\} = \Sigma_{\alpha\alpha}.
$$

Therefore, we can always replace $\mathbb{E}\left\{\partial S(A, X; \alpha^*)/\partial \alpha^{\mathrm{T}}\right\}$ by $\Sigma_{\alpha\alpha}$ in expression (S3.5) if either Assumption 1 or 2 holds. Thus, we can derive the influence function for the AIPW estimator in Lemma 3.

S4 Regularity conditions for the matching estimator

We review the assumptions for the matching estimators, which can also be found in Abadie and Imbens (2006).

Assumption 3 (Population distributions)**.** (i) X is continuously distributed on a compact and convex support. The density of X is bounded and bounded away from zero on its support.

(ii) For $a = 0, 1$, $\mu_a(x)$ and $\sigma_a^2(x)$ are Lipschitz, $\sigma_a^2(x)$ is bounded away from zero, and $\mathbb{E}(Y^4 \mid A = a, X = x)$ is bounded uniformly over its support.

Assumption 3 (i) can be relaxed by allowing X to have discrete components. We only need to obtain results on each level of discrete covariates and derive the same result. Assumption 3 (ii) requires the conditional mean and variance functions to be bounded and satisfy certain smoothness conditions, which are rather mild.

Assumption 4 (Estimators of mean functions)**.** For $a = 0, 1$, the estimator $\hat{\mu}_a(x)$ satisfies the following asymptotic condition: $|\hat{\mu}_a(x) - \mu_a(x)| = o_{\mathbb{P}}\left\{n_2^{-1/2+1/\dim(x)}\right\}$.

If $\hat{\mu}_a(x)$ is obtained under correctly specified parametric models, then Assumption 4 holds. If $\hat{\mu}_a(x)$ is obtained using nonparametric methods, such as power series regression (Newey, 1997) or kernel regression (Fan

and Gijbels, 1996) estimators, we need to select their tuning parameters properly to ensure Assumption 4 holds. Assumption 4 is needed so that the bias correction terms achieve fast convergence; e.g., $n^{1/2}(\hat{C}_n - C_n) \to 0$ in probability, as $n \to \infty$.

S5 Proof of Theorem 1

For simplicity of the notation, we use $\xrightarrow{\mathbb{P}}$ and $\xrightarrow{d}$ to denote "converge in probability as $n \to \infty$" and "converge in distribution as $n \to \infty$" respectively.

Firstly, using the law of large numbers,

$$\begin{aligned}\sum_{k=1}^{n} \xi_{n,k}^2 &= \frac{1}{n}\sum_{i=1}^{n}\left[\mathbb{E}\{\psi(L_i) \mid \mathbf{Z}_{\text{obs}}, \theta_0\} + \Gamma\mathcal{I}_{\text{obs}}^{-1}\bar{S}(\theta_0; Z_{\text{obs},i})\right]^2 \\ &\xrightarrow{\mathbb{P}} \mathbb{V}\left[\mathbb{E}\{\psi(L_i) \mid \mathbf{Z}_{\text{obs}}, \theta_0\} + \Gamma\mathcal{I}_{\text{obs}}^{-1}\bar{S}(\theta_0; Z_{\text{obs},i})\right],\end{aligned}$$

and

$$\begin{aligned}\sum_{k=n+1}^{n(1+m)} \xi_{n,k}^2 &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m^2}\sum_{j=1}^{m}\left[\psi(L_i^{*(j)}) - \mathbb{E}\{\psi(L_i) \mid \mathbf{Z}_{\text{obs}}, \hat{\theta}\}\right]^2 \\ &\xrightarrow{\mathbb{P}} \frac{1}{m}\mathbb{E}\left[\mathbb{V}\left\{\psi(L_i^{*(j)}) \mid \mathbf{Z}_{\text{obs}}\right\}\right].\end{aligned}$$

Therefore,

$$\sum_{k=1}^{n(1+m)} \xi_{n,k}^2 \xrightarrow{\mathbb{P}} \sigma^2, \tag{S5.6}$$

where

$$\sigma^2 = \mathbb{V}\left[\mathbb{E}\{\psi(L_i) \mid \mathbf{Z}_{\text{obs}}, \theta_0\} + \Gamma\mathcal{I}_{\text{obs}}^{-1}\bar{S}(\theta_0; Z_{\text{obs},i})\right] + m^{-1}\mathbb{E}\left[\mathbb{V}\left\{\psi(L_i^{*(j)}) \mid \mathbf{Z}_{\text{obs}}\right\}\right]. \tag{S5.7}$$

Secondly, we show

$$\max_{1\le k\le n(1+m)} |\xi_{n,k}| \xrightarrow{\mathbb{P}} 0. \tag{S5.8}$$

For any $\epsilon > 0$,

$$\begin{aligned}\mathbb{P}\left(\max_{1\le k\le n}|\xi_{n,k}| > \epsilon\right) &\le n\mathbb{P}\left(|\xi_{n,k}| > \epsilon\right) = n\mathbb{P}\left(|\xi_{n,k}^4| > \epsilon^4\right)\\ &\le \frac{1}{n\epsilon^4}\mathbb{E}\left[\mathbb{E}\{\psi(L_i) \mid \mathbf{Z}_{\text{obs}}, \theta_0\} + \Gamma\mathcal{I}_{\text{obs}}^{-1}\bar{S}(\theta_0; Z_{\text{obs},i})\right]^4 \to 0,\end{aligned}$$

where the second inequality follows from the Markov inequality, and the convergence follows by Assumption 6 and that $\hat{\theta}$ is consistent for θ_0 by Assumption 4. Similarly,

$$\mathbb{P}\left(\max_{n+1\le k\le n(1+m)}|\xi_{n,k}| > \epsilon\right) \le \frac{1}{nm^3\epsilon^4}\mathbb{E}\left[\psi(L_i^{*(j)}) - \mathbb{E}\{\psi(L_i) \mid \mathbf{Z}_{\text{obs}}, \hat{\theta}\}\right]^4 \to 0.$$

Therefore, $\mathbb{P}\left(\max_{1\le k\le n(1+m)}|\xi_{n,k}| > \epsilon\right) \to 0$, and then (S5.8) holds.

Next we show

$$\sup_n \mathbb{E}\left(\max_{1\le k\le n(1+m)} \xi_{n,k}^2\right) < \infty. \tag{S5.9}$$

For any n,

$$\begin{aligned}\mathbb{E}\left(\max_{1\le k\le n}\xi_{n,k}^2\right) &\le \mathbb{E}(n\xi_{n,k}^2)\\ &= \mathbb{E}\left[\mathbb{E}\{\psi(L_i) \mid \mathbf{Z}_{\text{obs}}, \theta_0\} - \tau + \Gamma\mathcal{I}_{\text{obs}}^{-1}\bar{S}(\theta_0; Z_{\text{obs},i})\right]^2 < \infty,\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\left(\max_{n+1\le k\le n(1+m)}\xi_{n,k}^2\right) &\le \mathbb{E}(nm\xi_{n,k}^2)\\ &= \frac{1}{m}\mathbb{E}\left[\psi(L_i^{*(j)})-\mathbb{E}\{\psi(L_i)\mid \mathbf{Z}_{\text{obs}},\hat{\theta}\}\right]^2<\infty.\end{aligned}$$

Therefore, $\mathbb{E}\left(\max_{1\le k\le n(1+m)}\xi_{n,k}^2\right)\le \mathbb{E}\left(\max_{1\le k\le n}\xi_{n,k}^2\right)+\mathbb{E}\left(\max_{n+1\le k\le n(1+m)}\xi_{n,k}^2\right)<\infty$, and (S5.9) holds.

Given conditions (S5.6) and (S5.8), the martingale central limit theorem implies that

$$\sum_{k=1}^{n(1+m)}\xi_{n,k}\xrightarrow{d}\mathcal{N}(0,\sigma^2),$$

where σ^2 is given in (S5.7). Also, under Assumption 4, $n^{1/2}(\hat{\tau}_{\text{MI}}-\tau)=\sum_{k=1}^{n+nm}\xi_{n,k}+o_{\mathbb{P}}(1)$, so we obtain

$$n^{1/2}(\hat{\tau}_{\text{MI}}-\tau)\xrightarrow{d}\mathcal{N}(0,\sigma^2). \tag{S5.10}$$

Given conditions (S5.6), (S5.8) and (S5.9), we apply Theorem 2.1 in Pauly et al. (2011), leading to

$$\sup_r\left|\mathbb{P}\left\{\{n(1+m)\}^{1/2}\sum_{k=1}^{n(1+m)}\frac{u_k}{\{n(1+m)\}^{1/2}}\xi_{n,k}\le r\mid \mathbf{Z}_{\text{obs}}\right\}-\Phi\left(\frac{r}{\sigma}\right)\right|\xrightarrow{\mathbb{P}}0, \tag{S5.11}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. Let $T_n^*=n^{-1/2}\sum_{k=1}^{n+nm}\xi_{n,k}u_k$. By (S5.10) and (S5.11), we can get

$$\sup_r\left|\mathbb{P}(n^{1/2}T_n^*\le r\mid \mathbf{Z}_{\text{obs}})-\mathbb{P}\{n^{1/2}(\hat{\tau}_{\text{MI}}-\tau)\le r\}\right|\xrightarrow{\mathbb{P}}0. \tag{S5.12}$$

Lastly, to prove Theorem 1, it remains to show that

$$\mathbb{P}\left\{n^{1/2}(T_n^* - T^*) \mid \mathbf{Z}_{\text{obs}}\right\} \xrightarrow{\mathbb{P}} 0. \tag{S5.13}$$

The difference between T_n^* and T^* can be decomposed into three parts,

$$n^{1/2}(T_n^* - T^*) = \sum_{k=1}^{n(1+m)} n^{-1/2} u_k (n^{1/2}\hat{\xi}_{n,k} - n^{1/2}\xi_{n,k}) = R_{1n} + R_{2n} + R_{3n},$$

where

$$\begin{aligned}
R_{1n} &= \sum_{i=1}^{n} n^{-1/2} u_i (\hat{\tau} - \tau), \\
R_{2n} &= \sum_{i=1}^{n} n^{-1/2} u_i \left[\mathbb{E}\{\psi(L_i) \mid \mathbf{Z}_{\text{obs}}, \hat{\theta}\} - \mathbb{E}\{\psi(L_i) \mid \mathbf{Z}_{\text{obs}}, \theta_0\}\right], \\
R_{3n} &= \sum_{i=1}^{n} n^{-1/2} u_i \left\{\hat{\Gamma}\hat{\mathcal{I}}_{\text{obs}}^{-1}\bar{S}(\hat{\theta}; Z_{\text{obs},i}) - \Gamma\mathcal{I}_{\text{obs}}^{-1}\bar{S}(\theta_0; Z_{\text{obs},i})\right\}.
\end{aligned}$$

Given the property of the bootstrap weight that $\mathbb{E}(u_k^2 \mid \mathbf{Z}_{\text{obs}}) = 1$, we can obtain that

$$\mathbb{E}(R_{1n}^2 | \mathbf{Z}_{\text{obs}}) = \frac{1}{n} n \mathbb{E}(u_i^2)(\hat{\tau} - \tau)^2 = (\hat{\tau} - \tau)^2 \xrightarrow{\mathbb{P}} 0,$$

and

$$\begin{aligned}
\mathbb{E}(R_{2n}^2 | \mathbf{Z}_{\text{obs}}) &= \frac{1}{n}\sum_{i=1}^{n}\left[\mathbb{E}\{\psi(L_i) \mid \mathbf{Z}_{\text{obs}}, \hat{\theta}\} - \mathbb{E}\{\psi(L_i) \mid \mathbf{Z}_{\text{obs}}, \theta_0\}\right]^2 \\
&\xrightarrow{d} \mathbb{P}\left\{\bar{\psi}(\hat{\theta}; \mathbf{Z}_{\text{obs}}) - \bar{\psi}(\theta_0; \mathbf{Z}_{\text{obs}})\right\}^2 \xrightarrow{\mathbb{P}} 0,
\end{aligned}$$

where $\mathbb{P}\{f(Z)\} = \int f(z)\mathrm{d}\mathbb{P}$ denotes the expectation of $f(Z)$. The first convergence follows by Assumption 7(Kennedy, 2016), and the second convergence follows by Assumption 5 and the fact $\hat{\theta}$ is $n^{1/2}$-consistent for θ_0

under conditions in Assumption 4. Similarly,

$$\mathbb{E}(R_{3n}^2|\mathbf{Z}_{\text{obs}}) \;=\; \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{\Gamma}\hat{\mathcal{I}}_{\text{obs}}^{-1}\bar{S}(\hat{\theta};Z_{\text{obs},i}) - \Gamma\mathcal{I}_{\text{obs}}^{-1}\bar{S}(\theta_0;Z_{\text{obs},i})\right\}^2 \xrightarrow{\mathbb{P}} 0.$$

Therefore, for any $\epsilon > 0$,

$$\mathbb{P}\{|R_{1n}| > \epsilon \mid \mathbf{Z}_{\text{obs}}\} \xrightarrow{\mathbb{P}} 0, \quad \mathbb{P}\{|R_{2n}| > \epsilon \mid \mathbf{Z}_{\text{obs}}\} \xrightarrow{\mathbb{P}} 0, \quad ,\mathbb{P}\{|R_{3n}| > \epsilon \mid \mathbf{Z}_{\text{obs}}\} \xrightarrow{\mathbb{P}} 0,$$

and we obtain (S5.13). The conclusion of Theorem 1 follows.

S6 Posterior sampling details

In this section, we describe the priors and MCMC details for both the simulation study and application.

S6.1 Simulation study

Here are the covariate distribution, outcome model and the propoensity score model we assume for the scenario (a) in the simulation study ($X_{[2]}$ is missing at random).

$$\begin{aligned} X_i &\sim \text{Normal}(\mu_X, \Sigma_X) \\ Y_i(0) &\sim \text{Normal}(X_i\beta_0, \sigma_0^2) \\ Y_i(1) &\sim \text{Normal}(X_i\beta_1, \sigma_1^2) \end{aligned}$$

$$A_i = \begin{cases} 1, & \text{if } A_i^\star >= 0, \\ 0, & \text{otherwise,} \end{cases} \quad \text{where } A_i^\star \sim \text{Normal}(X_i\alpha, 1)$$

We assume non-informative priors for the parameters: $\mu_X \sim \text{Normal}(0, I_p)$, $\Sigma_X^{-1} \sim \text{Wishsart}(p{+}1, (p{+}1)I_p)$ $\beta_0 \sim \text{Normal}(0, b_0^2 I_p)$, $\sigma_0^{-2} \sim \text{Gamma}(c_0, d_0)$, $\beta_1 \sim \text{Normal}(0, b_1^2 I_p)$, $\sigma_1^{-2} \sim \text{Gamma}(c_1, d_1)$, $\alpha \sim \text{Normal}(0, b_\alpha^2 I_p)$ where p is the dimension of the covariate and we assume $b_0 = b_1 = b_\alpha = 100$ and $c_0 = d_0 = c_1 = d_1 = 0.01$ in the simulation.

Let $X_{R_X,i}$ and $X_{\overline{R}_X,i}$ represent the observed and missing parts of X_i respectively, and β_{0R_X}, β_{1R_X} and $\beta_{0\overline{R}_X}$, $\beta_{1\overline{R}_X}$ are their corresponding coefficients in the outcome model, α_{R_X} and $\alpha_{\overline{R}_X}$ are their corresponding coefficients in the propensity score model. Let $\mu_{X_{R_X}|X_{\overline{R}_X}}$ and $\Sigma_{X_{R_X}|X_{\overline{R}_X}}$ be the conditional mean and variance of the missing part of X conditional on the observed part of X, which can be derived from μ_X and Σ_X.

We get the posterior distribution of the parameters and missing values using Gibbs sampling. The algorithm begins with setting all parameters and missing values to initial values and drawing the parameters from their full conditional distributions in sequence. The full conditional distributions are given below:

$$\beta_0|\text{rest} \sim \text{Normal}\Bigg(\Big(\sigma_0^{-2}\sum_{i:A_i=0} X_i^T X_i + b_0^{-2} I_p\Big)^{-1}\Big(\sigma_0^{-2}\sum_{i:A_i=0} X_i^T Y_i\Big), \Big(\sigma_0^{-2}\sum_{i:A_i=0} X_i^T X_i + b_0^{-2} I_p\Big)^{-1}\Bigg)$$

$$\sigma_0|\text{rest} \sim \text{Gamma}\Bigg(\frac{1}{2}\sum_{i=1}^{n} I(A_i=0) + c_0, \frac{1}{2}\sum_{i:A_i=0}(Y_i - X_i\beta_0)^2 + d_0\Bigg)$$

$$\beta_1|\text{rest} \sim \text{Normal}\Bigg(\Big(\sigma_1^{-2}\sum_{i:A_i=1} X_i^T X_i + b_1^{-2} I_p\Big)^{-1}\Big(\sigma_1^{-2}\sum_{i:A_i=1} X_i^T Y_i\Big), \Big(\sigma_1^{-2}\sum_{i:A_i=1} X_i^T X_i + b_1^{-2} I_p\Big)^{-1}\Bigg)$$

$$\sigma_1|\text{rest} \sim \text{Gamma}\Bigg(\frac{1}{2}\sum_{i=1}^{n} I(A_i=1) + c_1, \frac{1}{2}\sum_{i:A_i=1}(Y_i - X_i\beta_1)^2 + d_1\Bigg)$$

$$A_i^\star|\text{rest} \sim \begin{cases} \text{rtNormal}(X_i\alpha, 1, -\infty, 0), & \text{if } A_i = 0, \\ \text{rtNormal}(X_i\alpha, 1, 0, \infty), & \text{if } A_i = 1. \end{cases}$$

$$\alpha|\text{rest} \sim \text{Normal}\Bigg(\Big(\sum_{i=1}^{n} X_i^T X_i + b_\alpha I_p\Big)^{-1}\Big(\sum_{i=1}^{n} X_i^T A_i^\star\Big), \Big(\sum_{i=1}^{n} X_i^T X_i + b_\alpha I_p\Big)^{-1}\Bigg)$$

$$\mu_X|\text{rest} \sim \text{Normal}\Bigg(\Big(n\Sigma_X^{-1} + I_p\Big)^{-1}\Big(\Sigma_X^{-1}\sum_{i=1}^{n} X_i^T\Big), \Big(n\Sigma_X^{-1} + I_p\Big)^{-1}\Bigg)$$

$$\Sigma_X^{-1}|\text{rest} \sim \text{Wishart}\Bigg(n+p+1, \Big[\sum_{i=1}^{n}(X_i^T - \mu_X)(X_i^T - \mu_X)^T + (p+1)I_p\Big]^{-1}\Bigg)$$

$$X_{\overline{R}_X,i}|\text{rest} \sim \text{Normal}\Bigg(\frac{1}{\alpha_{\overline{R}_X}^2 + \beta_{0\overline{R}_X}^2\sigma_0^{-2}(1-A_i) + \beta_{1\overline{R}_X}^2\sigma_1^{-2}A_i + \Sigma^{-1}_{X_{R_X}|X_{\overline{R}_X}}}\Big[\alpha_{\overline{R}_X}(A_i^\star - X_{R_X,i}\alpha_{R_X})$$

$$+\beta_{0\overline{R}_X}\sigma_0^{-2}(Y_i - X_{R_X,i}\beta_{0R_X})(1-A_i) + \beta_{1\overline{R}_X}\sigma_1^{-2}(Y_i - X_{R_X,i}\beta_{1R_X})A_i + \mu_{X_{R_X}|X_{\overline{R}_X}}\Sigma^{-1}_{X_{R_X}|X_{\overline{R}_X}}\Big],$$

$$\frac{1}{\alpha_{\overline{R}_X}^2 + \beta_{0\overline{R}_X}^2\sigma_0^{-2}(1-A_i) + \beta_{1\overline{R}_X}^2\sigma_1^{-2}A_i + \Sigma^{-1}_{X_{R_X}|X_{\overline{R}_X}}}\Bigg)$$

For the scenario (b) in the simulation study where we assume $X_{[2]}$ is missing not at random, we need to introduce the model for the missing indicator in order to get the posterior distribution of missing values:

$$R_i = \begin{cases} 1, & \text{if } R_i^\star >= 0, \\ 0, & \text{otherwise,} \end{cases} \quad \text{where } R_i^\star \sim \text{Normal}((X_i, A_i)\gamma, 1)$$

with prior $\gamma \sim \text{Normal}(0, b_\gamma^2 I_{p+1})$.

The full conditional distribution are slightly different for missing covaraites but are the same for all other parameters as the scenario (a).

$$R_i^\star|\text{rest} \sim \begin{cases} \text{rtNormal}((X_i, A_i)\gamma, 1, -\infty, 0), & \text{if } R_i = 0, \\ \text{rtNormal}((X_i, A_i)\gamma, 1, 0, \infty), & \text{if } R_i = 1. \end{cases}$$

$$\gamma|\text{rest} \sim \text{Normal}\Bigg(\Big(\sum_{i=1}^n (X_i, A_i)^T (X_i, A_i) + b_\gamma I_{p+1}\Big)^{-1} \Big(\sum_{i=1}^n (X_i, A_i)^T R_i^\star\Big), \Big(\sum_{i=1}^n (X_i, A_i)^T (X_i, A_i) + b_\gamma I_{p+1}\Big)^{-}$$

$$X_{\overline{R}_X,i}|\text{rest} \sim \text{Normal}\Bigg(\frac{1}{\alpha_{\overline{R}_X}^2 + \beta_{0\overline{R}_X}^2 \sigma_0^{-2}(1-A_i) + \beta_{1\overline{R}_X}^2 \sigma_1^{-2} A_i + \Sigma_{X_{R_X}|X_{\overline{R}_X}}^{-1} + \gamma_{\overline{R}_X}^2} \Big[\alpha_{\overline{R}_X}(A_i^\star - X_{R_X,i}\alpha_{R_X})$$

$$+\beta_{0\overline{R}_X}\sigma_0^{-2}(Y_i - X_{R_X,i}\beta_{0R_X})(1-A_i) + \beta_{1\overline{R}_X}\sigma_1^{-2}(Y_i - X_{R_X,i}\beta_{1R_X})A_i + \mu_{X_{R_X}|X_{\overline{R}_X}}\Sigma_{X_{R_X}|X_{\overline{R}_X}}^{-1} + \gamma_{\overline{R}_X}(R_i^\star - (X_{R_X,i}, A$$

$$\frac{1}{\alpha_{\overline{R}_X}^2 + \beta_{0\overline{R}_X}^2 \sigma_0^{-2}(1-A_i) + \beta_{1\overline{R}_X}^2 \sigma_1^{-2} A_i + \Sigma_{X_{R_X}|X_{\overline{R}_X}}^{-1} + \gamma_{\overline{R}_X}^2}\Bigg)$$

For the scenario (d) in the simulation study where we assume both $X_{[2]}$ and Y are missing not at random, we need to introduce another model for

the missing indicator R_Y for Y on top of scenario (b):

$$R_{Yi} = \begin{cases} 1, & \text{if } R^{\star}_{Yi} >= 0, \\ 0, & \text{otherwise,} \end{cases} \quad \text{where } R^{\star}_{Yi} \sim \text{Normal}((X_i, A_i)\gamma_Y, 1)$$

with prior $\gamma_Y \sim \text{Normal}(0, b^2_{\gamma_Y} I_{p+1})$.

The full conditional distribution are slightly different for missing covaraites but are the same for all other parameters as the scenario (a). Additionally, we also need to derive the posterior distribution of missing Y.

$$R^{\star}_{Yi}|\text{rest} \sim \begin{cases} \text{rtNormal}((X_i, A_i)\gamma_Y, 1, -\infty, 0), & \text{if } R_{Yi} = 0, \\ \text{rtNormal}((X_i, A_i)\gamma_Y, 1, 0, \infty), & \text{if } R_{Yi} = 1. \end{cases}$$

$$\gamma_Y|\text{rest} \sim \text{Normal}\Bigg(\Big(\sum_{i=1}^{n}(X_i, A_i)^T(X_i, A_i)+b_{\gamma_Y} I_{p+1}\Big)^{-1}\Big(\sum_{i=1}^{n}(X_i, A_i)^T R^{\star}_{Yi}\Big), \Big(\sum_{i=1}^{n}(X_i, A_i)^T(X_i, A_i)+b_{\gamma_Y} I_{p}$$

$$X_{\overline{R}_X,i}|\text{rest} \sim \text{Normal}\Bigg(\frac{1}{\alpha^2_{\overline{R}_X} + \beta^2_{0\overline{R}_X}\sigma_0^{-2}(1-A_i) + \beta^2_{1\overline{R}_X}\sigma_1^{-2}A_i + \Sigma^{-1}_{X_{R_X}|X_{\overline{R}_X}} + \gamma^2_{\overline{R}_X} + \gamma^2_{Y\overline{R}_X}}\Big[\alpha_{\overline{R}_X}(A^{\star}_i - X_R$$

$$+\beta_{0\overline{R}_X}\sigma_0^{-2}(Y_i - X_{R_X,i}\beta_{0R_X})(1-A_i) + \beta_{1\overline{R}_X}\sigma_1^{-2}(Y_i - X_{R_X,i}\beta_{1R_X})A_i + \mu_{X_{R_X}|X_{\overline{R}_X}}\Sigma^{-1}_{X_{R_X}|X_{\overline{R}_X}} + \gamma_{\overline{R}_X}(R^{\star}_i - (X_{R_X,i}, A$$

$$+\gamma_{Y\overline{R}_X}(R^{\star}_{Yi} - (X_{R_X,i}, A_i)\gamma_{YR_X})\Big], \frac{1}{\alpha^2_{\overline{R}_X} + \beta^2_{0\overline{R}_X}\sigma_0^{-2}(1-A_i) + \beta^2_{1\overline{R}_X}\sigma_1^{-2}A_i + \Sigma^{-1}_{X_{R_X}|X_{\overline{R}_X}} + \gamma^2_{\overline{R}_X} + \gamma^2_{Y\overline{R}_X}}\Bigg)$$

$$Y_{\overline{R}_Y,i}|\text{rest} \sim \begin{cases} \text{Normal}(X_i\beta_0, \sigma_0^2) & \text{if } A_i = 0, \\ \text{Normal}(X_i\beta_1, \sigma_1^2) & \text{if } A_i = 1 \end{cases}$$

S6.2 Application

The setting in the application example is very similar to that in the scenario (b) in the simulation study where Y is fully observed and X are partially observed, and we assume the poverty ratio $X_{[1]}$ is missing not at random. More specifically as described in Application Section, we assume

$$Y = \begin{cases} 1 & \text{if } Y^* < 1, \\ [Y^*] & \text{if } 1 \leq Y^* \leq 5, \\ 5 & \text{if } Y^* > 5. \end{cases}$$

for the ordinary response variable general health satisfaction outcome and

$$X_{[1]} = \begin{cases} 0 & \text{if } X^*_{[1]} < 0, \\ X^*_{[1]} & \text{if } 0 \leq X^*_{[1]} \leq 5, \\ 5 & \text{if } X^*_{[1]} > 5. \end{cases}$$

for the bounded continuous variable family poverty ratio. Also, we assume we assume the latent outcome Y^* follows a linear regression model, i.e., $Y^*(a) = X^{*\mathrm{T}}\beta_a + \epsilon(a)$, where $\epsilon(a) \sim \mathcal{N}(0, \sigma_a^2)$ for $a = 0, 1$. The treatment

indicator follows Bernoulli$\{\pi_A(X^*)\}$ with $\pi_A(X^*) = \Phi(X^{*\mathrm{T}}\alpha)$. The missing indicator follows Bernoulli$\{\pi_R(X^*, A)\}$ with $\pi_R(X^*, A) = \Phi\{(X^*, A)^{\mathrm{T}}\gamma\}$. we assume the latent family poverty ratio follows a linear regression model with the other covariates, i.e., $X^*_{\overline{R}_X} = X_{R_X}\eta + \epsilon_X$, where $X^*_{\overline{R}_X} = X^*_{[1]}$ represents the latent family poverty ratio and X_R represents the other four covariates, $\epsilon_X \sim \mathcal{N}(0, \sigma_X^2)$.

We assume non-informative priors for η and σ_X^2: $\eta \sim \text{Normal}(0, b_\eta^2 I_{p-1})$ and $\sigma_\eta^{-2} \sim \text{Gamma}(c_\eta, d_\eta)$ where $b_\eta = 100$ and $c_\eta = d_\eta = 0.01$, and priors for all the other parameters are similarly to those in the simulation study.

The full conditional distributions in Gibbs sampling for most of the parameters are the similar to those in the simulation study with only difference in the posterior distribution for the new parameters and missing covaraite $X_{[1]}$ or $X_{\overline{R}_X}$:

$$\eta|\text{rest} \sim \text{Normal}\left(\left(\sigma_X^{-2}\sum_{i=1}^n X_{Ri}^T X_{Ri} + b_\eta^{-2} I_{p-1}\right)^{-1}\left(\sigma_X^{-2}\sum_{i=1}^n X_{Ri}^T X^\star_{\overline{R}_X,i}\right), \left(\sigma_X^{-2}\sum_{i=1}^n X_{Ri}^T X_{Ri} + b_\eta^{-2} I_{p-1}\right)^{-1}\right)$$

$$\sigma_X|\text{rest} \sim \text{Gamma}\left(\frac{1}{2}n + c_\eta, \frac{1}{2}\sum_{i=1}^n (X^\star_{\overline{R}_X,i} - X_{Ri}\eta)^2 + d_\eta\right)$$

where $X^\star_{\overline{R}_X,i}$ is generated from the truncated normal distribution based on $X_{\overline{R}_X,i}$.

$$X^{\star}_{\overline{R}_X,i}|\text{rest} \sim \text{Normal}\Bigg(\frac{1}{\alpha^2_{\overline{R}_X} + \beta^2_{0\overline{R}_X}\sigma_0^{-2}(1-A_i) + \beta^2_{1\overline{R}_X}\sigma_1^{-2}A_i + \sigma_X^{-2} + \gamma^2_{\overline{R}_X}}\Big[\alpha_{\overline{R}_X}(A^{\star}_i - X_{R_X,i}\alpha_{R_X})$$

$$+\beta_{0\overline{R}_X}\sigma_0^{-2}(Y^{\star}_i - X_{R_X,i}\beta_{0R_X})(1-A_i) + \beta_{1\overline{R}_X}\sigma_1^{-2}(Y^{\star}_i - X_{R_X,i}\beta_{1R_X})A_i + X_{R_X,i}\eta\sigma_X^{-2} + \gamma_{\overline{R}_X}(R^{\star}_{[1]i} - (X_{R_X,i}, A_i)\gamma_{R_X}$$

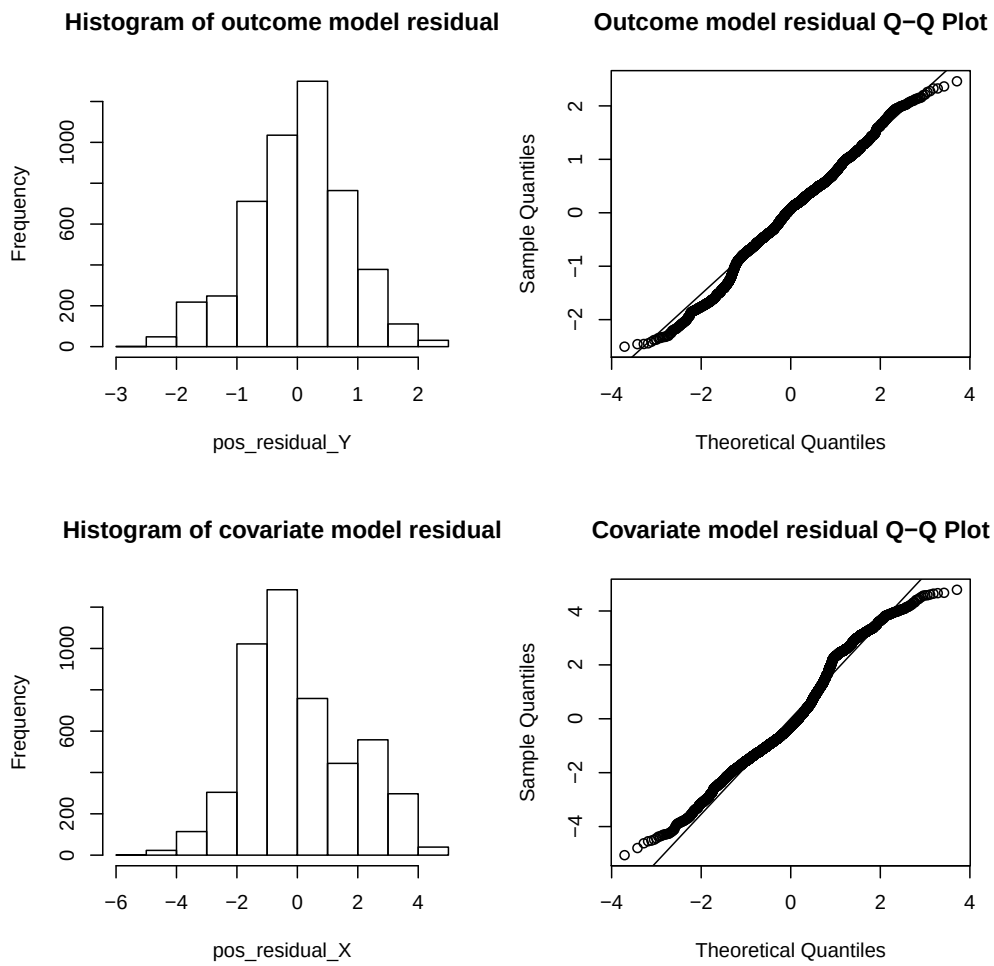
$$\frac{1}{\alpha^2_{\overline{R}_X} + \beta^2_{0\overline{R}_X}\sigma_0^{-2}(1-A_i) + \beta^2_{1\overline{R}_X}\sigma_1^{-2}A_i + \sigma_X^{-2} + \gamma^2_{\overline{R}_X}}\Bigg)$$

where $Y_i^{\star}$ is generated from the truncated normal distribution based on Y_i.

S7 Model diagnosis for the real data application

We provide model diagnoses for the real data application. We assume that the latent outcome Y^* follows a linear regression model and the family poverty ratio follows a linear regression model with other covariates. We assess these model assumptions by checking the distributions of the residuals after fitting the two models. Figure 1 displays the histograms and Q-Q plots of the residuals (top for the outcome model and bottom for the covariate model) and suggests that the models fit the data well and that the model assumptions are reasonable.

Figure 1: Diagnosis plots for the outcome model and the covariate model.



Bibliography

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

Abadie, A. and G. W. Imbens (2008). On the failure of the bootstrap for matching estimators. *Econometrica 76*, 1537–1557.

Fan, J. and I. Gijbels (1996). *Local Polynomial Modelling and Its Applications*. Chapman & Hall, London: CRC Press.

Kennedy, E. H. (2016). Semiparametric theory and empirical processes in causal inference. In *Statistical Causal Inferences and Their Applications in Public Health Research*, pp. 141–167. Springer.

Lunceford, J. K. and M. Davidian (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat. Med. 23*, 2937–2960.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *J. Econometrics 79*, 147–168.

Pauly, M. et al. (2011). Weighted resampling of martingale difference arrays with applications. *Electronic Journal of Statistics 5*, 41–52.

Yang, S. and P. Ding (2018). Asymptotic inference of causal effects

with observational studies trimmed by the estimated propensity scores. *Biometrika 105*, 487–493.