PUATE: Efficient ATE Estimation from Treated (Positive) and Unlabeled Units

Masahiro Kato, Fumiaki Kozai, Ryo Inokuchi

Mizuho-DL Financial Technology

MIZUHO

1. Contributions

- ✓ Novel problem setting: PUATE.
- Estimation of average treatment effects (ATEs).
- Treatment indicators are missing values.
- ✓ Two data-generating process: censoring and case-control settings.
- ✓ For each setting, we develop asymptotically efficient estimators.

2. Setup

Variables

- **Binary treatment** 1 (treatment) and 0 (control).
- **Covariates** $X_i \in \mathbb{R}^k$.
- **Potential outcome** $Y_i(1), Y_i(0) \in \mathbb{R}$

PUATE

- Goal: estimation of the average treatment effect $\tau_0 \coloneqq \mathbb{E}[Y(1) - Y(0)]$.
- We can only observe outcomes of the treated units and outcomes of the mixture of treated and controlled units.
- We cannot observe the treatment indicator directly.

➢ We consider the following two sampling regimes in observations.
Under different sampling schemes, we can obtain different theoretical properties.

Censoring setting

- Observe the following dataset:

$$\mathcal{D} \coloneqq \{X_i, O_i, Y_i\}_{i=1}^n,$$

- $O_i \in \{1,0\}$: observation indicator.
- $Y_i \coloneqq O_i Y_i(1) + (1 - O_i)\tilde{Y}_i$.
- $\tilde{Y}_i \coloneqq \tilde{D}_i(1) + (1 - \tilde{D}_i)Y_i(0)$.
- $\tilde{D}_i$ is treatment indicator.

➢ Treatment indicator is unobserved.

- If $O_i = 1$, we can observe outcomes of treated units.
- If $O_i = 0$, we can only observe a outcomes of mixed treatments.
- Observation is a random event.

Case-control setting

- Observe the two datasets.

$$\mathcal{D}_{\mathrm{T}} \coloneqq \{X_{\mathrm{T},j}, Y_j(1)\}_{j=1}^m,$$

$$\mathcal{D}_{\mathrm{U}} \coloneqq \{X_k, Y_{\mathrm{U},k}\}_{k=1}^l.$$

- m and l are fixed sample sizes of each dataset.
- $Y_{\mathrm{U},k} \coloneqq D_i Y_k(1) + (1 - D_k)Y_k(0)$.
- $D_i \in \{1,0\}$ is a treatment indicator.

➢ Treatment indicator is unobserved.

- In $\mathcal{D}_{\mathrm{T}}$, all units are treated.
- In $\mathcal{D}_{\mathrm{U}}$, we can only observe outcomes of mixed treatments.
- Observation is a non-random event.

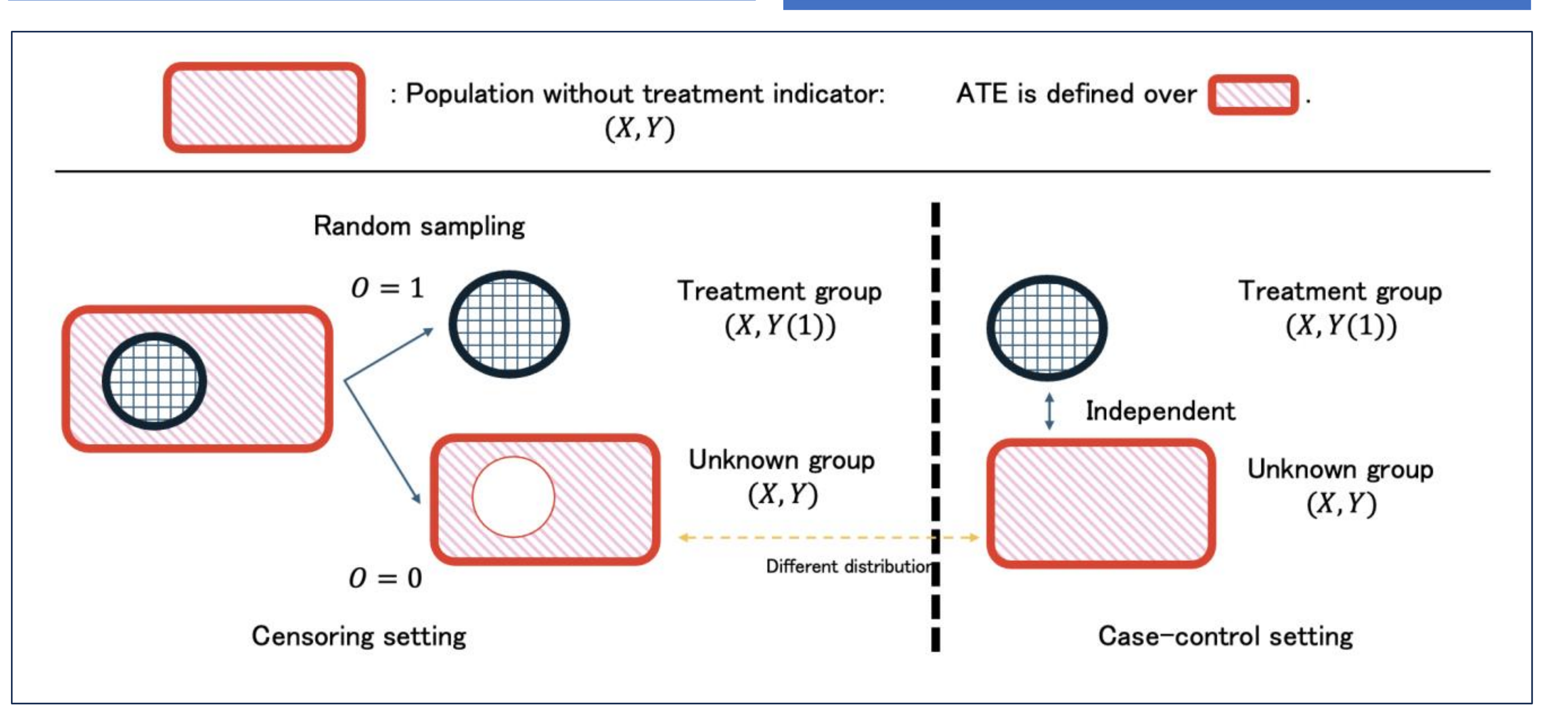


Illustration of the censoring and case-control settings

3. Censoring setting

What kind of estimators are desirable?
→ **Asymptotically efficient estimators**.
= Estimators whose asymptotic variance (mean squared error) is minimum.

Efficiency bound

- **Lower bound for the asymptotic variance.**
  - Generalization of the Cramer-Rao lower bounds.
- For any regular estimators, the smallest asymptotic variance is given as

$$V^{\mathrm{cens}} \coloneqq \mathbb{E}\left[\left(S^{\mathrm{cens}}(X, O, Y; \mu_{\mathrm{T},0}, v_0, \pi_0, g_0) - \tau_0\right)^2\right]$$
$$= \mathbb{E}\left[\left(1 - \frac{g_0(1|X)}{g_0(0|X)}\right)^2 \frac{\mathrm{Var}(Y(1)|X)}{\pi(1|X)} + \frac{\mathrm{Var}(\tilde{Y}|X)}{g_0(0|X)^2 \pi(1|X)} + (\tau_0(X) - \tau_0)\right].$$

- $S^{\mathrm{cens}}(X, O, Y) - \tau_0$ is the **efficient score function** defined as

$$S^{\mathrm{cens}}(X, O, Y; \mu_{\mathrm{T},0}, v_0, \pi_0, g_0) \coloneqq \frac{\mathbb{I}[O = 1]\left(Y - \mu_{\mathrm{T},0}(X)\right)}{\pi_0(1|X)} - \frac{\mathbb{I}[O = 0](Y - v_0(X))}{g_0(0|X)\pi_0(0|X)}$$
$$+ \frac{g_0(1|X)\mathbb{I}[O = 1]\left(Y - \mu_{\mathrm{T},0}(X)\right)}{g_0(0|X)\pi_0(1|X)} - \mu_{\mathrm{T},0}(X) - \frac{1}{g_0(0|X)} v(X) + \frac{g_0(1|X)}{g_0(0|X)} \mu_{\mathrm{T},0}(X).$$

- $\tau_0(X) \coloneqq \mathbb{E}[Y(1) - Y(0)|X]$: Conditional ATE.
- $\mu_{T,0}(X) \coloneqq \mathbb{E}[Y(1)|X], v_0(X) \coloneqq \mathbb{E}[Y(1)|X]$: Conditional mean outcomes.
- $\pi(o|X) \coloneqq \mathbb{P}(O = o|X)$: Conditional probability of observation $O = 1$.
- $g_0(d|X) \coloneqq \mathbb{P}(\tilde{D} = d|X, O = 0)$: Conditional probability of treatment $\tilde{D} = 1$ given X and $O = 0$.

Estimator

- Based on the efficient score (influence) function, we construct the following estimator for the ATE τ_0.

$$\hat{\tau}_n^{\mathrm{cens-eff}} \coloneqq \frac{1}{n}\sum_{i=1}^n S^{\mathrm{cens}}(X_i, O_i, Y_i; \hat{\mu}_{\mathrm{T},n}, \hat{v}_n, \hat{\pi}_n, \hat{g}_n)$$

- $\hat{\mu}_{\mathrm{T},n}$, $\hat{v}_n$, $\hat{\pi}_n$, and $\hat{g}_n$ are the estimators of $\mu_{\mathrm{T},0}$, v_0, π_0, and g_0.
- The quality of the estimation of the nuisance parameters $\mu_{T,0}$, v_0, π_0, and g_0 affects the root-n consistency of $\hat{\tau}_n^{\mathrm{cens-eff}}$. See below.

Asymptotic normality

- Propensity score g_0 is known ($\hat{g}_n = g_0$).
- Assum the following convergence rate conditions hold as $n \to \infty$:
  $\|\pi_0(d|X) - \hat{\pi}_n(d|X)\|_2 = o_p(1)$, $\left\|\mu_{\mathrm{T},0}(X) - \hat{\mu}_{\mathrm{T},n}(X)\right\|_2 = o_p(1)$,
  $\|v_0(X) - \hat{v}_n(X)\|_2 = o_p(1)$,
  $\|\pi_0(d|X) - \hat{\pi}_n(d|X)\|_2 \left\|\mu_{\mathrm{T},0}(X) - \hat{\mu}_{T,n}(X)\right\|_2 = o_p(n^{-1/2})$,
  $\|\pi_0(d|X) - \hat{\pi}_n(d|X)\|_2 \|v_0(X) - \hat{v}_n(X)\|_2 = o_p(n^{-1/2})$
- $\hat{\mu}_{T,n}$, $\hat{v}_n$, and $\hat{\pi}_n$ are constructed via cross fitting or satisfy Donsker condition.
- Then, the asymptotic normality holds:

$$\sqrt{n}\left(\hat{\tau}_n^{\mathrm{cens-eff}} - \tau_0\right) \xrightarrow{\mathrm{d}} \mathcal{N}(0, V^{\mathrm{cens}}). \quad (n \to \infty)$$

- This estimator is asymptotically efficient because the asymptotic variance V^{cens} matches the efficiency bound.

4. Case-control setting

- Let $n = m + l$, where $m = \alpha n$ and $l = (1 - \alpha)n$ for $\alpha \in (0,1)$.
- We consider asymptotic efficiency for $n \to \infty$.

Estimator

- We use the following ATE estimator.

$$\hat{\tau}^{\mathrm{cc-eff}} \coloneqq \frac{1}{m}\sum_{j=1}^m \left(1 - \frac{\hat{e}_n(1 \mid X_j)}{\hat{e}_n(0 \mid X_j)}\right)\left(Y_j(1) - \hat{\mu}_{\mathrm{T},n}(X_j)\right)\hat{r}_n(X_j)$$
$$+ \frac{1}{l}\sum_{k=1}^l \left(\frac{Y_{\mathrm{U},k} - \hat{\mu}_{\mathrm{U},n}(X_k)}{\hat{e}_n(0 \mid X_k)} + \hat{\mu}_{\mathrm{T},n}(X_k) - \frac{\hat{\mu}_{\mathrm{U},n}(X_k)}{\hat{e}_n(0 \mid X_k)} + \frac{\hat{e}_n(1 \mid X_k)\hat{\mu}_{\mathrm{T},n}(X_k)}{\hat{e}_n(0 \mid X_k)}\right).$$

- $\hat{e}_n$, $\hat{r}_n(X_j)$, $\hat{\mu}_{\mathrm{T},n}$, and $\hat{\mu}_{\mathrm{U},n}$ are estimators of e_0, r_0, $\mu_{\mathrm{T},0}$, and $\mu_{\mathrm{U},0}$.
- $\mu_{\mathrm{U},0}(X) \coloneqq \mathbb{E}[Y_{\mathrm{U}}|X]$: Conditional mean outcomes.
- $e_0(d|X) \coloneqq \mathbb{P}(D = d|X)$: Conditional probability of treatment $D = 1$.
- $r_0(X) \coloneqq \frac{p_{\mathrm{U},0}(X)}{p_{\mathrm{T},0}(X)}$: Density ratio between covariate densities of $\mathcal{D}_{\mathrm{T}}$ and $\mathcal{D}_{\mathrm{U}}$.

Asymptotic normality

- Under appropriate conditions, we have

$$\sqrt{n}\left(\hat{\tau}^{\mathrm{cc-eff}} - \tau_0\right) \xrightarrow{\mathrm{d}} \mathcal{N}(0, V^{\mathrm{cc}}). \quad (n \to \infty)$$

- Here, V^{cc} is the efficiency bound defined as

$$V^{\mathrm{cc}} \coloneqq \frac{1}{\alpha}\mathbb{E}\left[\left(1 - \frac{e_0(1 \mid X_j)}{e_0(0 \mid X_j)}\right)\mathrm{Var}(Y(1) \mid X) r_0(X)\right]$$
$$+ \frac{1}{1 - \alpha}\mathbb{E}\left[\frac{\mathrm{Var}(Y_{\mathrm{U}} \mid X)}{e_0(0 \mid X_j)} + (\tau_0(X) - \tau_0)^2\right].$$

- The proposed ATE estimator is efficient.

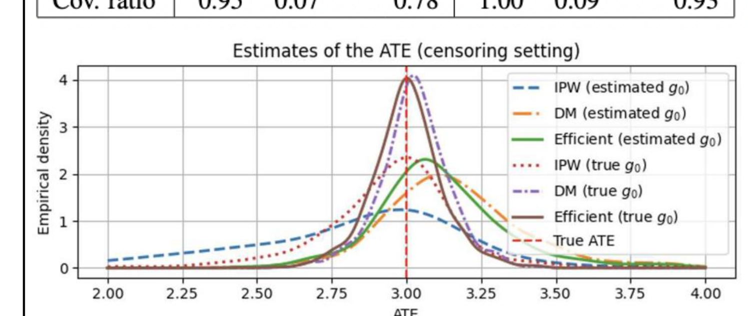
5. Simulation studies

- Simulation studies using synthetic datasets.

Censoring setting

- Set $\mathrm{k} = 3$.
- X is drawn from a multivariate normal dist. $\mathcal{N}(0, I_k)$
- D is drawn from a logistic model.
- O is drawn from a Bernoulli dist.
- $\tau_0 = 3$.
- $Y = X^\top \beta + \tau_0 \cdot D + 1.1 + \varepsilon$.

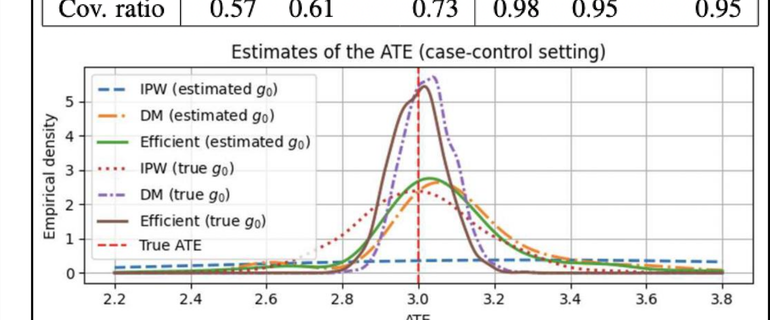
Censoring	IPW (estimated g_0)	DM (estimated g_0)	Efficient (estimated g_0)	IPW (true g_0)	DM (true g_0)	Efficient (true g_0)
MSE	0.31	0.08	0.06	0.06	0.01	0.01
Bias	-0.26	0.16	0.12	-0.06	0.03	0.00
Cov. ratio	0.95	0.07	0.78	1.00	0.09	0.93



Case-control setting

- Set $\mathrm{k} = 3$.
- X_{T} and X are drawn from different multivariate normal distributions $\mathcal{N}(0, I_k)$
- D is drawn from a logistic model.
- $\tau_0 = 3$.
- $Y = X^\top \beta + \tau_0 \cdot D + 1.1 + \varepsilon$.

Case-control	IPW (estimated e_0)	DM (estimated e_0)	Efficient (estimated e_0)	IPW (true e_0)	DM (true e_0)	Efficient (true e_0)
MSE	10.85	0.07	0.06	0.03	0.01	0.00
Bias	1.44	0.11	0.07	0.00	0.03	0.00
Cov. ratio	0.57	0.61	0.73	0.98	0.95	0.95



References

- Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal, 2018.
- Marthinus Christoffel du Plessis, Gang. Niu, and Masashi Sugiyama. Convex formulation for learning from positive and unlabeled data. In ICML, 2015.
- Guido W. Imbens and Tony Lancaster. Efficient estimation and stratified sampling. Journal of Econometrics, 74(2):289–318, 1996