Singular Propensity Scores: Reducing Variance in Weighted Estimators

Joint Statistical Meeting 2024, Portland, USA
August 5th, 2024

Kosuke Morikawa[1] and Keisuke Yano[2]

Graduate School of Engineering Science, Osaka University, Japan[1]
Institute of Statistical Mathematics, Japan[2]

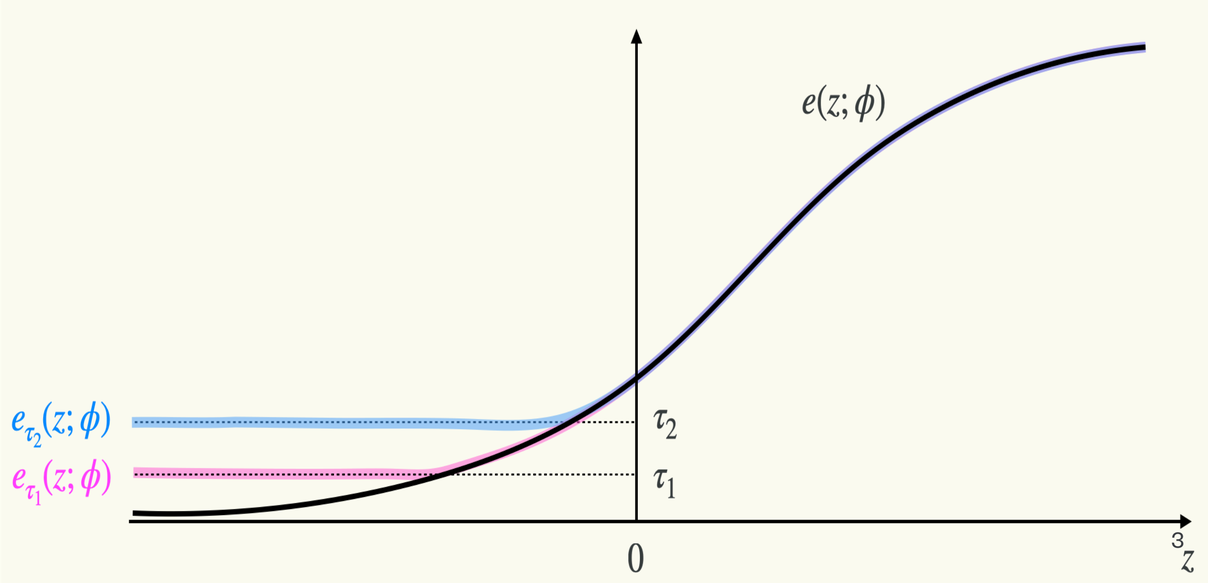
Introduction

Brief Summary

- Inverse weighted estimators $\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{e_i(\phi)}$ have been widely used in many fields, including **causal inference**, **missing data analysis**, and **survey sampling**.

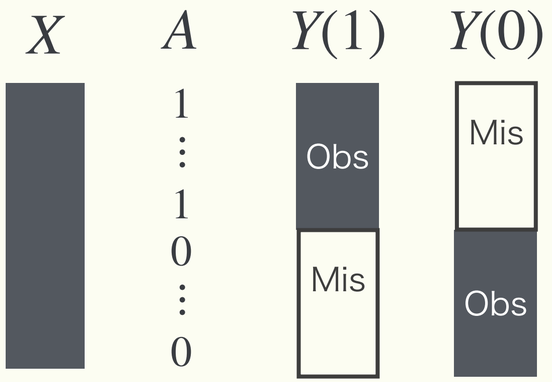
Propensity Score

- However, inverse weights $1/e_i(\phi)$ can diverge to infinity
- Our contribution is…
  1. to introduce a trimmed propensity score $e_{\tau,i}(\phi)$ with an appropriate threshold τ
  2. to propose an information criterion to select the best threshold τ^*



Settings: Average Treatment Effect

- $A_i \in \{0,1\}$: **Treatment**
- $Y_i = A_i Y_i(1) + (1 - A_i) Y_i(0)$: **Observed outcome**
  - $Y_i(a)\,(a = 0,1)$: **Potential outcomes**
- X_i: **Pre-treatment covariates**
- $Z_i = (X_i, Y_i)$
- $e(z; \phi) = P(A = 1 \mid Z = z; \phi)$: **Propensity score** (including MNAR case)
- θ: **Interesting parameter**
  - If $\theta = E\{Y(1)\}$, an estimator for θ is $\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} \frac{A_i}{e(Z_i)} Y_i$ ← We focus on this estimator
  - If $\theta = E\{Y(0)\}$, $\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} \frac{1 - A_i}{1 - e(Z_i)} Y_i$



✔ **Strongly Ignorability (or MAR).**

$A \perp \{Y(0), Y(1)\} \mid X$

Estimating Functions & Evaluation Metric — Regular Setup —

- $s_\phi(\phi; Z_i, A_i)$: Scoring function for ϕ
  - e.g., log-likelihood under MAR:
    $s_\phi(\phi; Z_i, A_i) = A_i \log e(\phi; X_i) + (1 - A_i)\log\{1 - e(\phi; X_i)\}$

- $s_\theta(\theta; \phi; Z_i, A_i)$: Scoring function for θ
  - e.g., IPW estimator $-\dfrac{A_i}{e(Z_i; \hat{\phi})}(\theta - Y_i)^2$

- $\nu(\hat{\theta}; \phi^*; Z_i, A_i)$: Evaluation metric for $\hat{\theta}$
  - e.g., $\dfrac{A_i}{e(Z_i; \phi^*)}(\hat{\theta} - Y_i)^2 \rightarrow$ Objective function $E_{\hat{\theta}, \tilde{Y}}\{(\hat{\theta} - \tilde{Y})^2\}$

True (fixed, for the time being…)

Train

Test

$$\sum_{i=1}^{n} s_\phi(\phi; Z_i, A_i) \xrightarrow{\text{maximizer}} \hat{\phi}$$

$$\sum_{i=1}^{n} s_\theta(\theta; \hat{\phi}; Z_i, A_i) \xrightarrow{\text{maximizer}} \hat{\theta}$$

$$E_{\tilde{Z},\tilde{A}}\left[E_{\hat{\theta}}\left[\nu(\hat{\theta}; \phi^*; \tilde{Z}, \tilde{A})\right]\right] \xrightarrow{\text{minimizer}} \text{Good model (w.r.t. MSE)}$$

Generalization Error (GE)

Estimating Functions & Evaluation Metric — Regular Setup —

- $s_\phi(\phi; Z_i, A_i)$: Scoring function for ϕ
  - e.g., log-likelihood under MAR:
    $$s_\phi(\phi; Z_i, A_i) = A_i \log e(\phi; X_i) + (1 - A_i)\log\{1 - e(\phi; X_i)\}$$

- $s_\theta(\theta; \phi; Z_i, A_i)$: Scoring function for θ
  - e.g., IPW estimator $-\frac{A_i}{e(Z_i; \hat{\phi})}(\theta - Y_i)^2$ → **unstable**

$$\sum_{i=1}^{n} s_\phi(\phi; Z_i, A_i) \xrightarrow{\text{maximizer}} \hat{\phi}$$

$$\sum_{i=1}^{n} s_\theta(\theta; \hat{\phi}; Z_i, A_i) \xrightarrow{\text{maximizer}} \hat{\theta}$$

- $\nu(\hat{\theta}; \phi^*; Z_i, A_i)$: Evaluation metric for $\hat{\theta}$
  - e.g., $\frac{A_i}{e(Z_i; \phi^*)}(\hat{\theta} - Y_i)^2$

True (fixed, for the time being…)

$$E_{\tilde{Z},\tilde{A}}\left[E_{\hat{\theta}}\left[\nu(\hat{\theta}; \phi^*; \tilde{Z}, \tilde{A})\right]\right] \xrightarrow{\text{minimizer}} \text{Good model (w.r.t. MSE)}$$

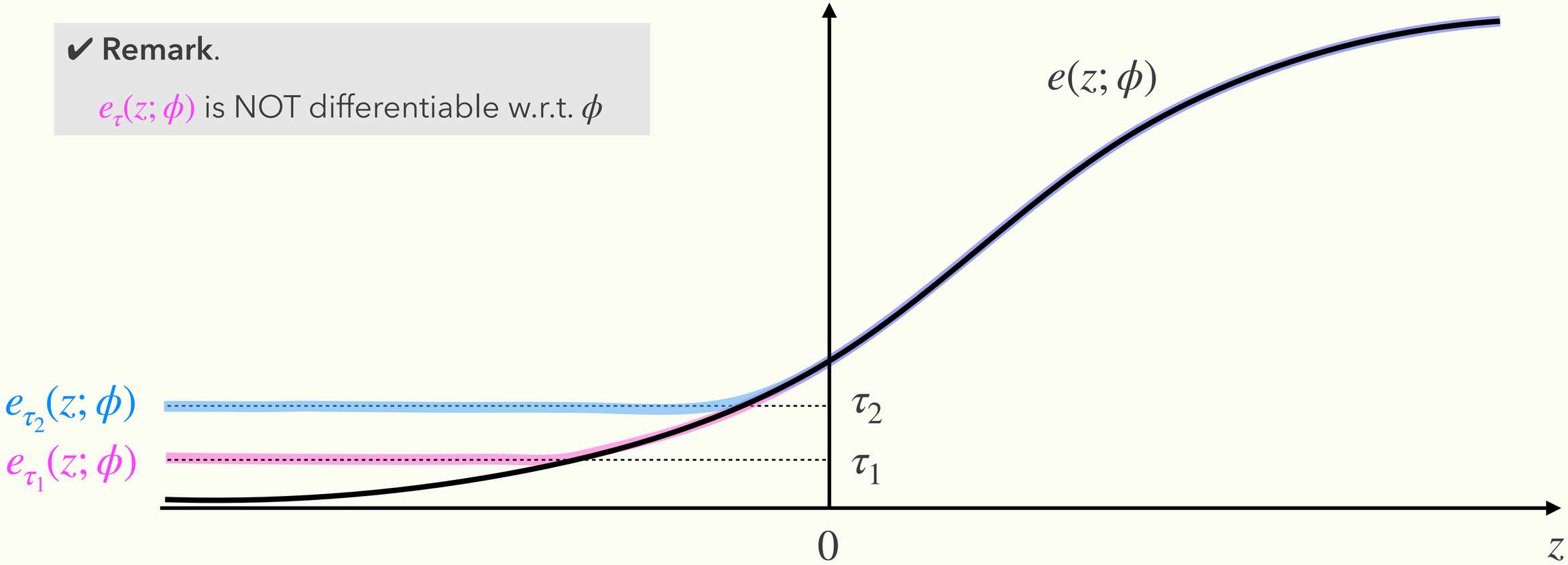
Generalization Error (GE)

Trimmed propensity score

$$e_\tau(Z_i;\phi) = \tau I\{e(Z_i;\phi) \le \tau\} + e(Z_i;\phi) I\{e(Z_i;\phi) > \tau\}$$

✔ **Remark**.

$e_\tau(z;\phi)$ is NOT differentiable w.r.t. ϕ



Estimating Functions & Evaluation Metric — Singular Setup —

- $s_\phi(\phi; Z_i, A_i)$: Scoring function for ϕ
- $s_\theta(\theta; \phi; Z_i, A_i)$: Scoring function for θ
  - e.g., IPW estimator $\frac{A_i}{\boxed{e_\tau(Z_i; \phi)}}(\theta - Y_i)^2$ ②

Trimmed propensity

- $\nu(\hat{\theta}[\phi]; \phi^*; Z_i, A_i)$: Evaluation metric for $\hat{\theta}[\phi]$
  - e.g., $\frac{A_i}{e(Z_i; \phi^*)}(\hat{\theta}[\phi] - Y_i)^2$

① **Bayesian Approach** ~~$\hat{\phi} = \arg\max_\phi \exp\left(\sum_{i=1}^n s_\phi(A_i, Z_i)\right)$~~

Pseudo likelihood Prior

$$\phi \sim \exp\left(\sum_{i=1}^n s_\phi(\phi; A_i, Z_i)\right)\pi(\phi)$$

$$\sum_{i=1}^n s_\theta(\theta; \phi; Z_i, A_i) \xrightarrow{\text{maximizer}} \hat{\theta}[\phi]$$

minimizer

$$E_{\tilde{Z},\tilde{A}}\left[E_{\text{pos},\phi}\left[\nu(\hat{\theta}[\phi]; \phi^*; \tilde{Z}, \tilde{A}) \,\middle|\, \mathcal{S}\right]\right] \longrightarrow \text{Good model (w.r.t. MSE)}$$

(Pseudo) Posterior

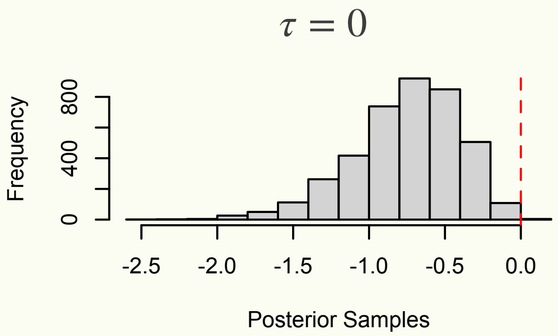
Expectation of ϕ

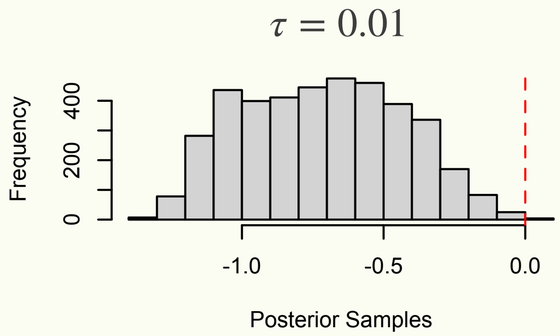
$\{A_i, X_i, Y_i\}_{i=1}^n$ (dataset)

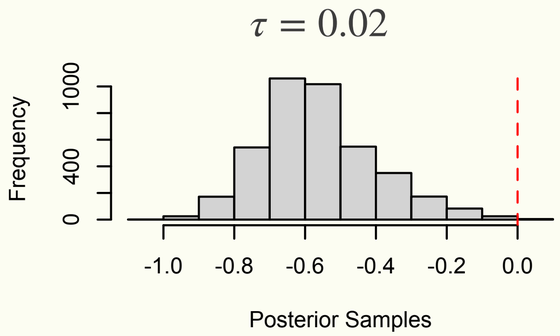
Posterior Dist. of IPW $\hat{\theta}[\phi]$

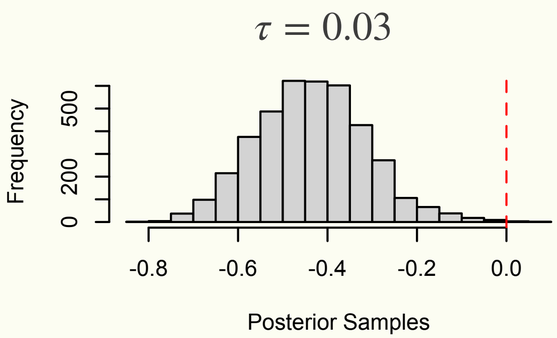
Heavily biased

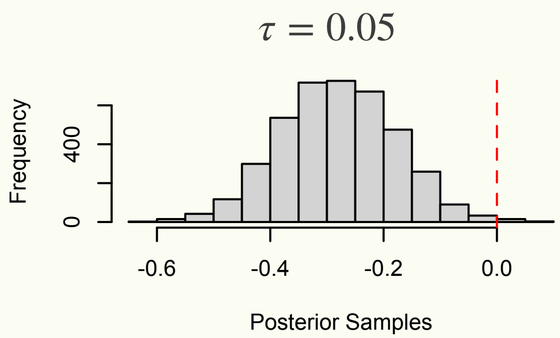
Non-normal distribution

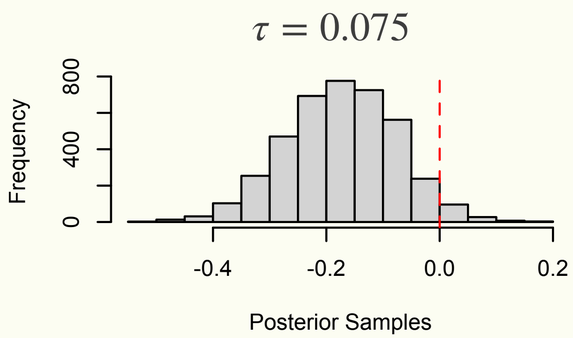




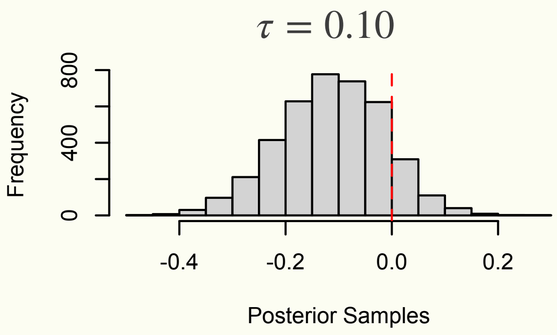


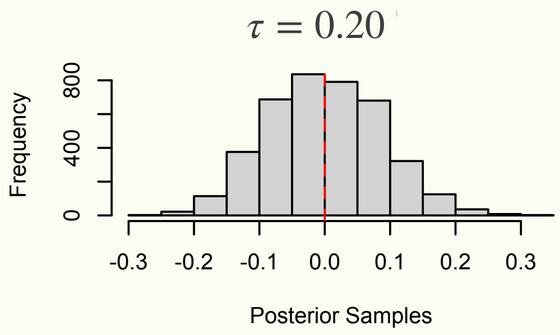


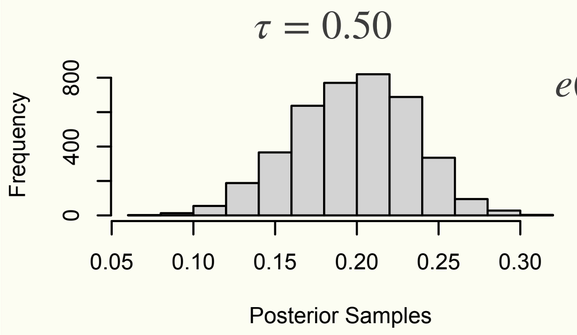




Can we choose this ideal threshold?







Outcome model

$$Y = X_1 + X_2 - X_3 - X_4 + \varepsilon$$

$$X_1, X_2, X_3, X_4, \varepsilon \sim N(0, 1)$$

Propensity score

$$e(X) = P(A = 1 \mid X)$$

$$= \frac{1}{1 + \exp\{(1 + X_1 + X_2 + X_3 - X_4)/2\}}$$

Estimating Functions & Evaluation Metric — Singular Setup —

- $s_\phi(\phi; Z_i, A_i)$: Scoring function for ϕ
- $s_\theta(\theta; \phi; Z_i, A_i)$: Scoring function for θ
  - e.g., IPW estimator $\frac{A_i}{②\ \boxed{e_\tau(Z_i; \phi)}}(\theta - Y_i)^2$ Trimmed propensity
- $\nu(\hat{\theta}[\phi]; \phi^*; Z_i, A_i)$: Evaluation metric for $\hat{\theta}[\phi]$
  - e.g., $\frac{A_i}{e(Z_i; \phi^*)}(\hat{\theta}[\phi] - Y_i)^2$

① Bayesian Approach ~~$\hat{\phi} = \arg\max_\phi \exp\left(\sum_{i=1}^n s_\phi(A_i, Z_i)\right)$~~

Pseudo likelihood Prior

$$\phi \sim \exp\left(\sum_{i=1}^n s_\phi(\phi; A_i, Z_i)\right)\pi(\phi)$$

$$\sum_{i=1}^n s_\theta(\theta; \phi; Z_i, A_i) \xrightarrow{\text{maximizer}} \hat{\theta}[\phi]$$

minimizer

$$E_{\tilde{Z},\tilde{A}}\left[E_{\text{pos},\phi}\left[\nu(\hat{\theta}[\phi]; \phi^*; \tilde{Z}, \tilde{A}) \,\middle|\, \mathcal{S}\right]\right] \longrightarrow \text{Good model (w.r.t. MSE)}$$

(Pseudo) Posterior

Expectation of ϕ

$\{A_i, X_i, Y_i\}_{i=1}^n$ (dataset)

Problem: estimation of the Generalization Error (GE)

$$E_{\tilde{Z},\tilde{A}}\left[E_{\mathrm{pos},\phi}\left[\nu(\hat{\theta}[\phi];\phi^*;\tilde{Z},\tilde{A})\,\middle|\,\mathcal{S}\right]\right] \approx \frac{1}{n}\sum_{i=1}^{n} E_{\mathrm{pos},\phi}\left[\nu(\hat{\theta}[\phi];\phi^*;Z_i,A_i)\,\middle|\,\mathcal{S}\right]$$

Pseudo posterior

Dual use of the same dataset in ESTIMATION and EVALUATION ⇒ Bias

Leave-One-Out-Cross-Validation Criterion

$$\mathrm{CV} := \frac{1}{n}\sum_{i=1}^{n} E_{\mathrm{pos},\phi^{-i}}\left[\nu(\hat{\theta}[\phi^{-i}]; \phi^*; Z_i, A_i) \,\middle|\, \mathcal{S}^{-i}\right]$$

Dataset excluding i-th data

- We start from the Cross-validation type criterion
  - 😀 r.v. between “ESTIMATION” and “EVALUATION” is independent
  - 😞 Computationally very intensive (posterior sampling $\times\, n$)

⇒ We derive an information criterion that is asymptotically unbiased as **CV**, by using only “one-set posterior samples”.

List of Relevalent Information Criteria (IC)

IC	Target	Evaluation Metric	Propensity Score (PS)	PS Estimation	Singularity
AIC (Akaike, 1973)	Density function	KL divergence		NA	
WAIC (Watanabe, 2010, JMLR)	Density function	KL divergence		NA	✔
wCp (Baba et al., 2017, Biometrika)	Regression	MSE	✔	△	
PCIC (Iba and Yano, 2023, Neural Comp.)	General	General	✔		✔
Proposed	General	General	✔	✔	✔

Proposed Information Criterion

Proposed Information Criterion With Known Outer Propensity

Theorem 1 (Morikawa and Yano, 2024+)

Under some regularity conditions,

$$E(\mathrm{IC}) = E(\mathrm{CV}) + o(n^{-1})$$

$$\mathrm{IC} = \frac{1}{n}\sum_{i=1}^{n} E_{\mathrm{pos},\phi}\left[\nu(\hat{\theta}[\phi];\phi^*;Z_i,A_i)\,\middle|\,\mathcal{S}\right] + \underbrace{I_1 + I_2}_{=\text{Bias corrected term}},$$

$$I_1 = -\frac{1}{n}\sum_{i=1}^{n} \mathrm{Cov}_{\mathrm{pos},\phi}\left[\nu(\hat{\theta}[\phi];\phi^*;Z_i,A_i),\quad s_\phi(\hat{\theta}[\phi];\phi;Z_i,A_i)\mid\mathcal{S}\right],$$

$$I_2 = -\frac{1}{n}\sum_{i=1}^{n} E_{\mathrm{pos},\phi}\left[\nabla_\theta\nu(\hat{\theta}[\phi];\phi^*;Z_i,A_i)\left\{-\sum_{j=1}^{n}\nabla_\theta^2 s_\theta(\hat{\theta}[\phi];\phi;Z_j,A_j)\right\}^{-1}\nabla_\theta s_\theta(\hat{\theta}[\phi];\phi;Z_i,A_i)\,\middle|\,\mathcal{S}\right].$$

Effects of Estimation of Outer Propensity Score

$$\nu(\hat{\theta}[\phi];\psi;\tilde{Z},\tilde{A}) = \frac{\tilde{A}}{e(\psi;\tilde{Z})}\left(\tilde{Y} - \underbrace{\frac{\sum_{i=1}^{n} A_i Y_i / e_\tau(\phi;Z_i)}{\sum_{i=1}^{n} A_i / e_\tau(\phi;Z_i)}}_{=\hat{\theta}[\phi]}\right)^2$$

Inner propensity

Outer propensity

- We do NOT allow trimming for outer propensity in this study
  - Previous researches have NOT cared about effects of $\hat{\psi}$ (e.g. MLE)

- In this talk, we also derive the asymptotic bias caused by replacement of ϕ^* with $\hat{\psi}$

Difficulty in Deriving Asymptotic Bias With Unknown Outer Propensity

By Taylor Expansion,

$$\frac{1}{n}\sum_{i=1}^{n} E_{\text{pos},\phi}\left[\nu(\hat{\theta}[\phi];\hat{\psi};Z_i,A_i)\,\middle|\,\mathcal{S}\right]$$

$$\approx \frac{1}{n}\sum_{i=1}^{n} E_{\text{pos},\phi}\left[\nu(\hat{\theta}[\phi];\psi^*;Z_i,A_i)\,\middle|\,\mathcal{S}\right] + \frac{1}{n}\sum_{i=1}^{n} E_{\text{pos},\phi}\left[\nabla_{\psi}\nu(\hat{\theta}[\phi];\psi^*;Z_i,A_i)(\hat{\psi}-\psi^*)\,\middle|\,\mathcal{S}\right]$$

$$= O_p(n^{-1/2})$$

$$+\frac{1}{2n}\sum_{i=1}^{n} E_{\text{pos},\phi}\left[\nabla^2_{\psi}\nu(\hat{\theta}[\phi];\psi^*;Z_i,A_i)(\hat{\psi}-\psi^*)^2\,\middle|\,\mathcal{S}\right]$$

↓

Higher order expansion is required…

Key Idea: Data Splitting

$$\frac{1}{n}\sum_{i=1}^{n} E_{\mathrm{pos},\phi}\left[\nabla_{\psi}\nu(\hat{\theta}[\phi];\psi^{*};Z_i,A_i)(\hat{\psi}-\psi^{*}) \;\middle|\; \mathcal{S}\right]$$

Dual use of the same dataset $\Rightarrow$ Data splitting

- Data split: $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2$
  - $\mathcal{S}_1$: Computation of the evaluation metric
  - $\mathcal{S}_2$: Estimation of $\psi \rightarrow \hat{\psi}$
- Data-Splitting CV criterion

$$\mathrm{CV}_1 := \frac{1}{n_1}\sum_{i\in\mathcal{S}_1} E_{\mathrm{pos},\phi^{-i}}\left[\nu(\hat{\theta}[\phi^{-i}];\hat{\psi};Z_i,A_i) \;\middle|\; \mathcal{S}_1^{-i}\right], \quad \text{where } n_1 = \#\{\mathcal{S}_1\}$$

($\mathcal{S}_2$ is used to obtain $\hat{\psi}$)

  - We can consider CV_2 by changing the roles of $\mathcal{S}_1$ and $\mathcal{S}_2$
  - Our final criterion is $\mathrm{CV}_1 + \mathrm{CV}_2$

Proposed Information Criterion With Unknown Outer Propensity

$$\mathrm{IC}_1 = \frac{1}{n_1}\sum_{i\in\mathcal{S}_1} E_{\mathrm{pos},\phi}\left[\nu(\hat{\theta}[\phi];\hat{\psi};Z_i,A_i)\,\middle|\,\mathcal{S}_1\right] + \underbrace{I_1 + I_2 + I_3 + I_4}_{=\text{Bias corrected term}},$$

$$I_1 = -\frac{1}{n_1}\sum_{i\in\mathcal{S}_1} \mathrm{Cov}_{\mathrm{pos},\phi}\left[\nu(\hat{\theta}[\phi];\hat{\psi};Z_i,A_i),\quad s_\phi(\phi;Z_i,A_i)\mid\mathcal{S}_1\right],$$

$$I_2 = -\frac{1}{n_1}\sum_{i\in\mathcal{S}_1} E_{\mathrm{pos},\phi}\left[\nabla_\theta\nu(\hat{\theta}[\phi];\hat{\psi};Z_i,A_i)\left\{-\sum_{j\in\mathcal{S}_1}\nabla_\theta^2 s_\theta(\hat{\theta}[\phi];\phi;Z_j,A_j)\right\}^{-1}\nabla_\theta s_\theta(\hat{\theta}[\phi];\phi;Z_i,A_i)\,\middle|\,\mathcal{S}_1\right],$$

$$I_3 = -\frac{1}{n_1}\sum_{i\in\mathcal{S}_1} E_{\mathrm{pos},\phi}\left[\nabla_\psi\nu(\hat{\theta}[\phi];\hat{\psi};Z_i,A_i)\mid\mathcal{S}_1\right]\underbrace{E(\hat{\psi}-\phi^*)}_{=O_p(n^{-1})},$$

$$I_4 = -\frac{1}{2n_1}\sum_{i\in\mathcal{S}_1} E_{\mathrm{pos},\phi}\left[\nabla_\psi^2\nu(\hat{\theta}[\phi];\hat{\psi};Z_i,A_i)\mid\mathcal{S}_1\right]\underbrace{Var(\hat{\psi})}_{=O_p(n^{-1})},$$

Additional penalty due to the estimation of $\hat{\psi}$

Proposed Information Criterion With Unknown Outer Propensity

Theorem 2 (Morikawa and Yano, 2024+)

Under some regularity conditions,

$$E(\tilde{\mathrm{IC}}) = E(\tilde{\mathrm{CV}}) + o(n^{-1}),$$

where $\tilde{\mathrm{IC}} = \mathrm{IC}_1 + \mathrm{IC}_2$, $\tilde{\mathrm{CV}} = \mathrm{CV}_1 + \mathrm{CV}_2$

- In practice, to reduce the randomness of data splitting, we can compute the proposed IC multiple times $\tilde{\mathrm{IC}}^{(b)}$ $(b = 1, \ldots, B)$
- Then, we can take a vote among minimizer of each $\tilde{\mathrm{IC}}^{(b)}$ $(b = 1, \ldots, B)$

Numerical Study

Numerical Study (Toy Example)

1. **Outcome model**

$$Y = X_1 + X_2 - X_3 - X_4 + \varepsilon,$$

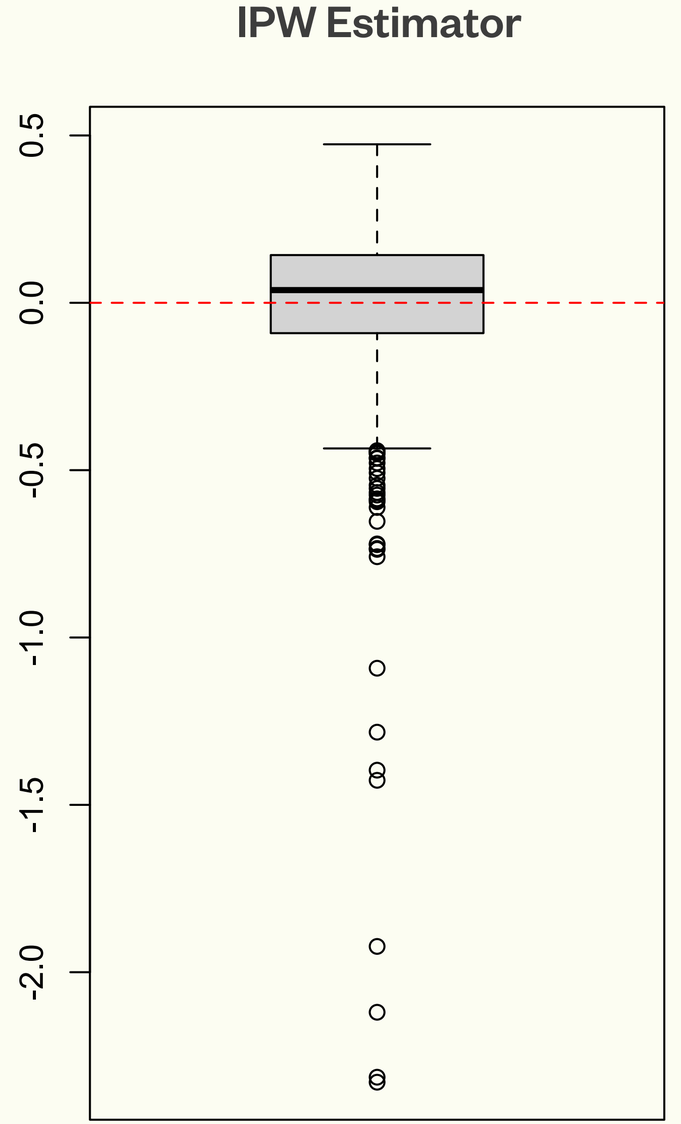
$$X_1, X_2, X_3, X_4, \varepsilon \sim N(0, 1)$$

2. **Propensity Score**

$$e(X) = P(A = 1 \mid X)$$

$$= \frac{1}{1 + \exp\{(1 + X_1 + X_2 + X_3 - X_4)/2\}}$$

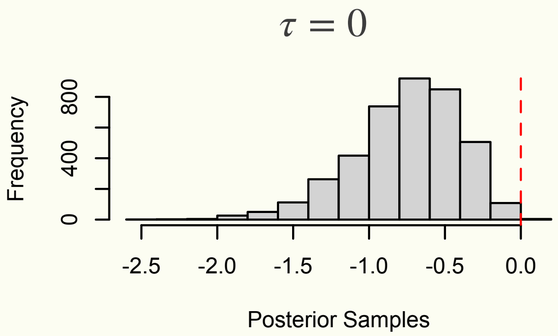
- **Target :** $E(Y)$
  - s_ϕ: log-likelihood; s_θ: IPW, $\nu(\cdot)$: IPW L_2-metric
- $n = 400,$ #iteration $= 100$

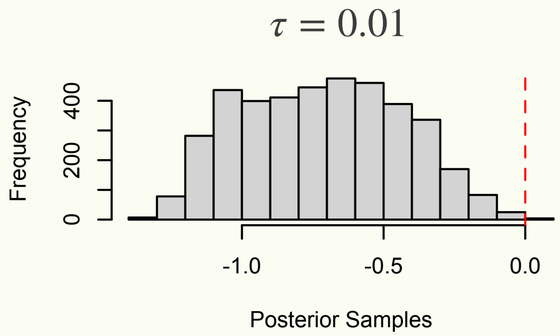


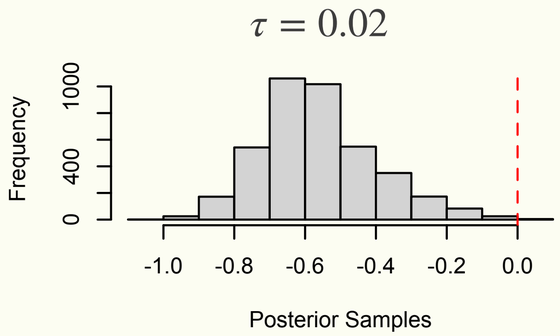
Posterior Dist. of an IPW Estimator (cont'd)

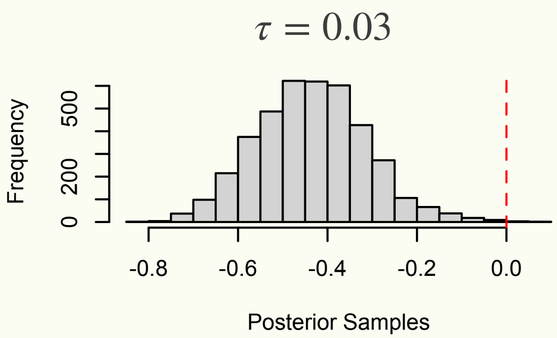
Heavily biased

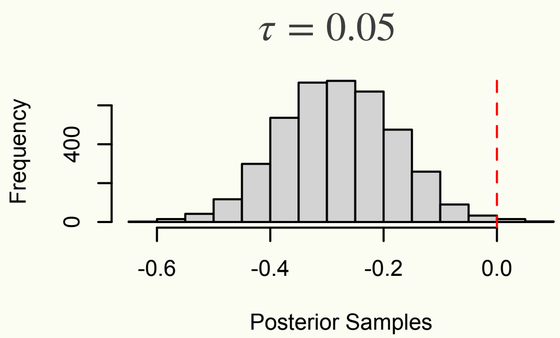
Non-normal distribution

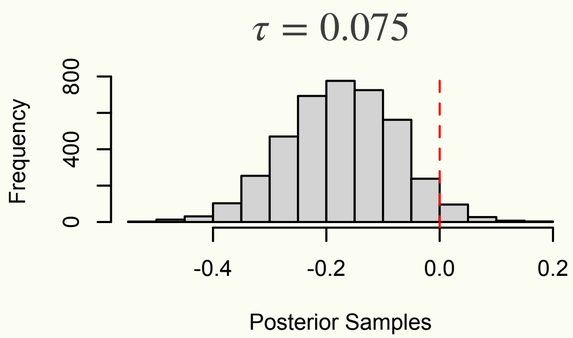




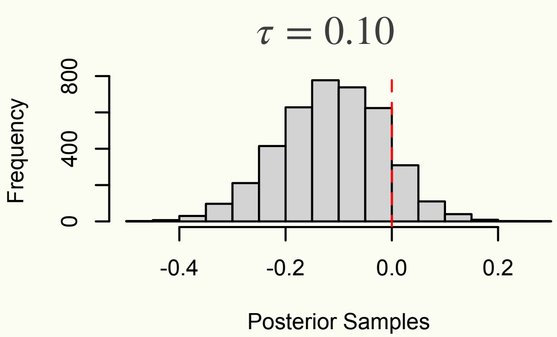


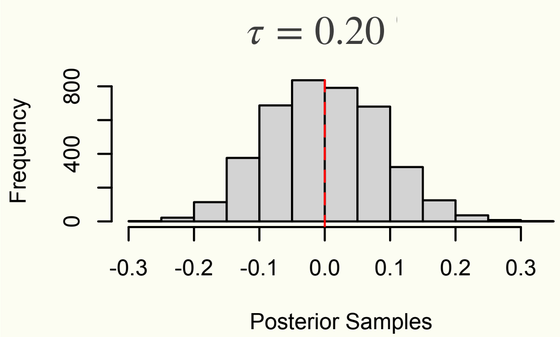


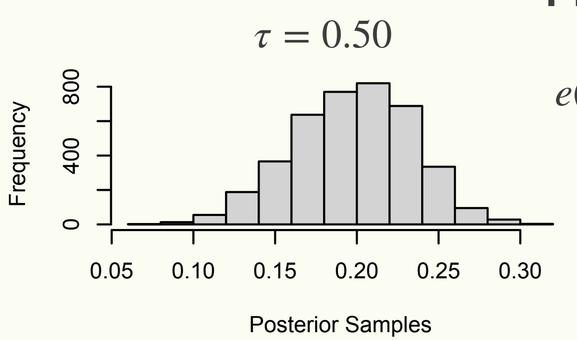




Can we choose this ideal threshold?







Outcome model

$$Y = X_1 + X_2 - X_3 - X_4 + \varepsilon$$

$$X_1, X_2, X_3, X_4, \varepsilon \sim N(0, 1)$$

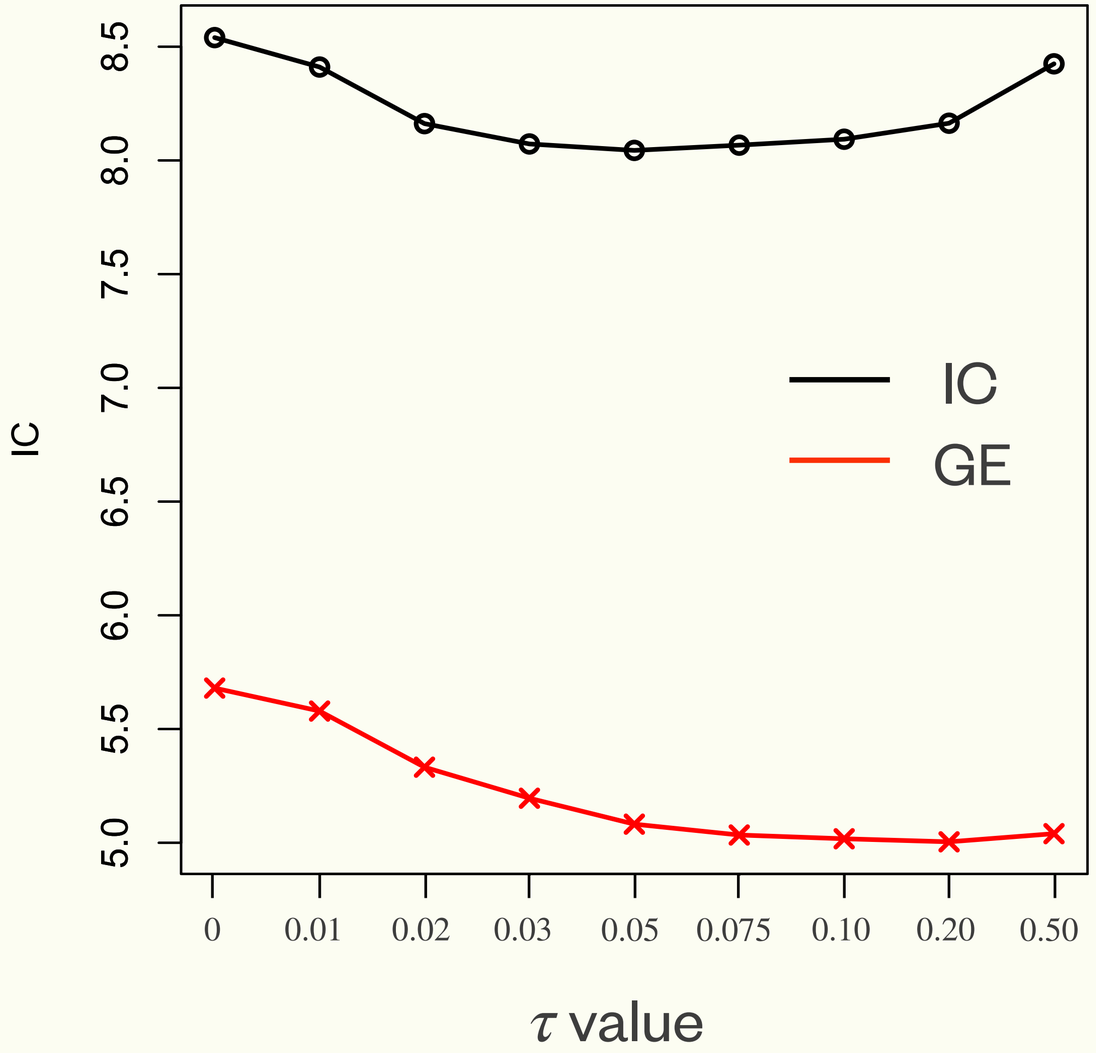
Propensity score

$$e(X) = P(A = 1 \mid X)$$

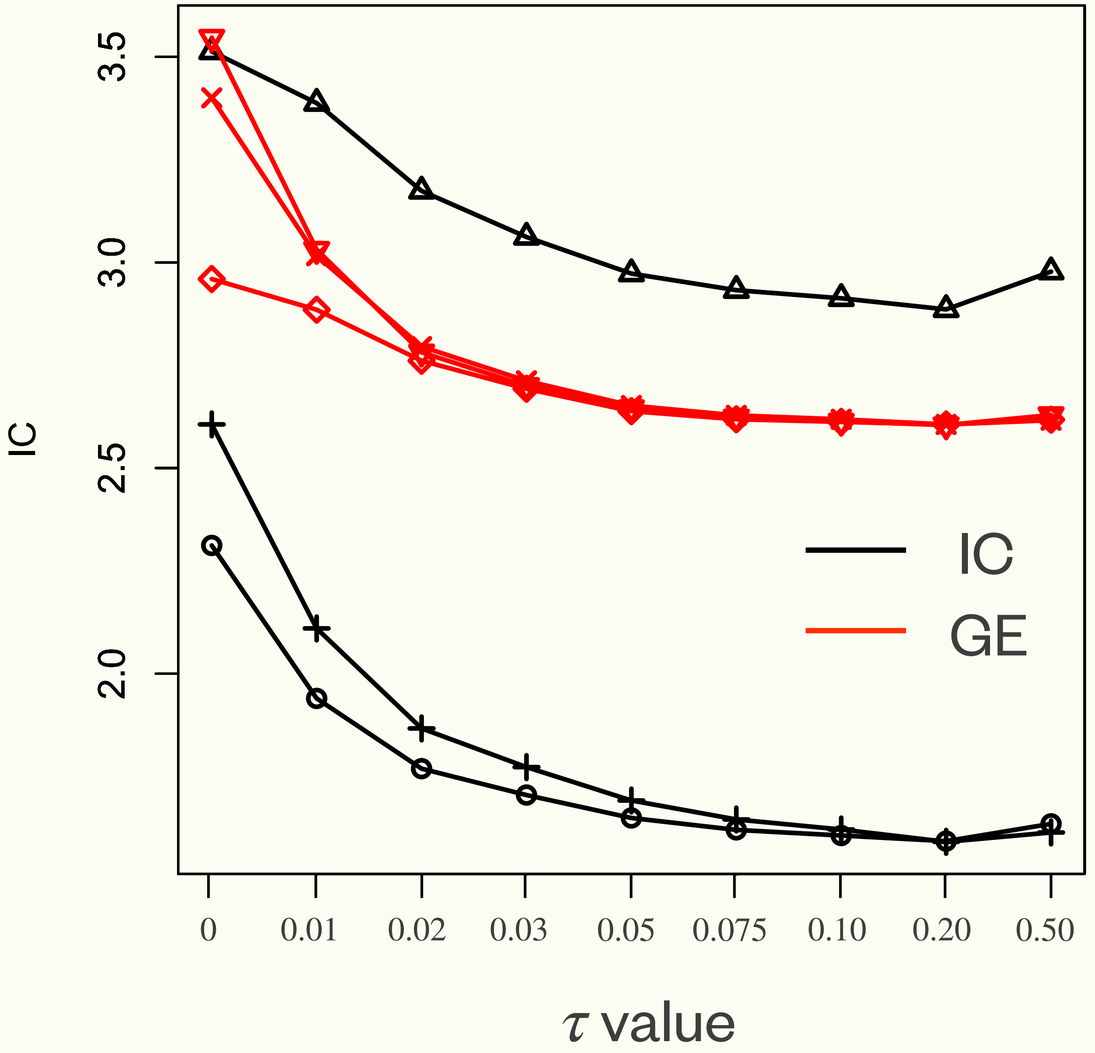
$$= \frac{1}{1 + \exp\{(1 + X_1 + X_2 + X_3 - X_4)/2\}}$$

An Application of Proposed IC

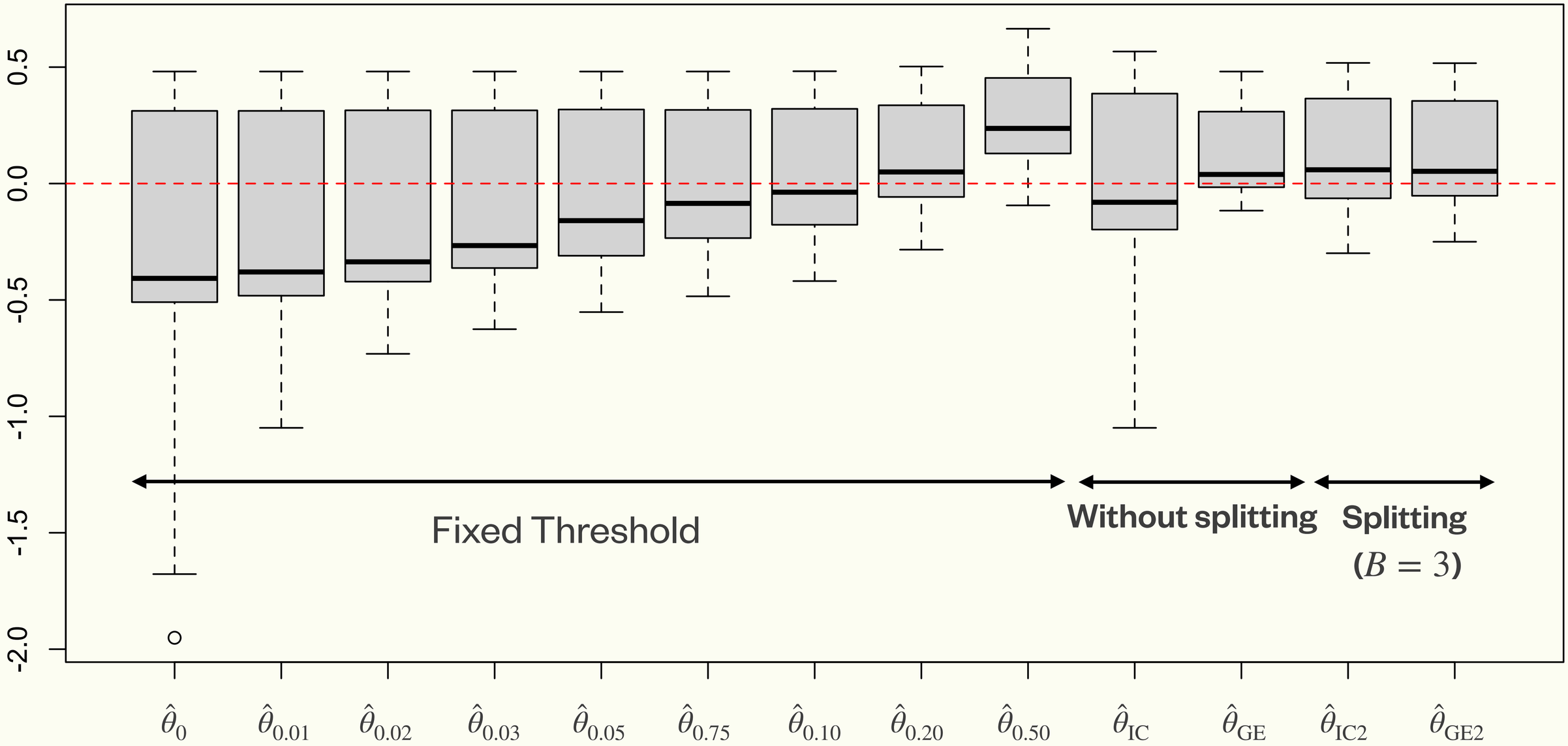
Without Splitting



Random Splittings ($B = 3$)



Boxplots of the Estimators



Conclusion

Discussion

- We proposed a statistical framework to use the **trimmed propensity score**

$$\nu(\hat{\theta}[\phi];\psi;\tilde{Z},\tilde{A}) = \frac{\tilde{A}}{e(\psi;\tilde{Z})}\left(\tilde{Y} - \underbrace{\frac{\sum_{i=1}^{n} A_i Y_i / e_\tau(\phi;Z_i)}{\sum_{i=1}^{n} A_i / e_\tau(\phi;Z_i)}}_{=\hat{\theta}[\phi]}\right)^2$$

Inner propensity

Outer propensity

- Crump et al. (2009, Biometrika) and Yang and Ding (2018, Biometrika) considered a statistical inference for trimmed estimand $E\{Y \cdot I(\alpha_1 \le e(X) \le \alpha_2)\}$
  - α_1 and α_2 are user-specified constants
  - Our estimand is $E(Y)$

- However, when the propensity is extremely close to zero or one , we also need to consider the trimmed estimand, because the outer propensity also diverges

Appendix

Selected Models: #iteration = 100

- Without splitting
  - **IC**: Proposed Information Criterion

ε	0	0.01	0.02	0.03	0.05	0.075	0.10	0.20	0.50
	1	3	1	1	8	10	35	27	14

  - **GE**: Generalization error with the use of the true propensity score for the outer propensity

ε	0	0.01	0.02	0.03	0.05	0.075	0.10	0.20	0.50
	15	1	2	1	6	8	13	35	19

- #Random splitting = 3
  - **IC2**: Proposed Information Criterion

ε	0	0.01	0.02	0.03	0.05	0.075	0.10	0.20	0.50
	0	0	0	1	0	0	21	47	31

  - **GE2**: Generalization error with the use of the true propensity score for the outer propensity

ε	0	0.01	0.02	0.03	0.05	0.075	0.10	0.20	0.50
	1	3	1	3	8	9	14	40	21

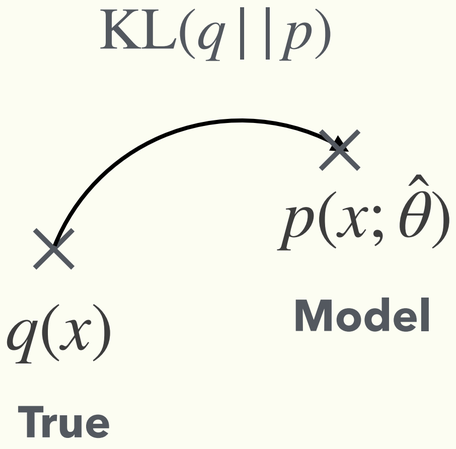
Classical Information Criterion: AIC $\simeq$ CV-type Criterion

- Estimation of Kullbuck-Leibler (KL) divergence:

Our Target!!

$$\mathrm{KL}(q||p) = \int q(x)\log\frac{q(x)}{p(x;\hat{\theta})}\mathrm{d}x \propto -\int q(x)\log p(x;\hat{\theta})dx =: \tilde{\mathrm{KL}}(q||p)$$

Prof. Akaike with a Google logo



Classical Information Criterion: AIC ≃ CV-type Criterion

- Estimation of Kullbuck-Leibler (KL) divergence:

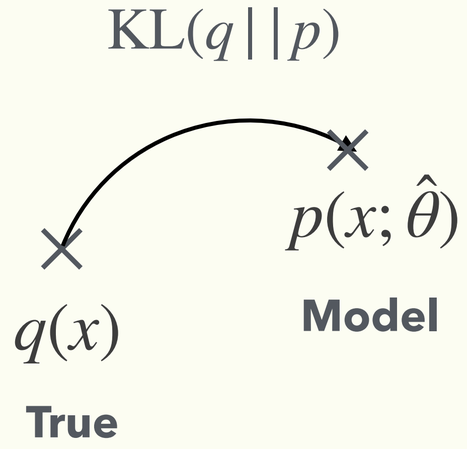
Our Target!!

$$\mathrm{KL}(q||p) = \int q(x)\log\frac{q(x)}{p(x;\hat{\theta})}\mathrm{d}x \propto -\int q(x)\log p(x;\hat{\theta})dx =: \tilde{\mathrm{KL}}(q||p)$$

Empirical Dist.

$$\approx -\frac{1}{n}\sum_{i=1}^{n}\log p(X_i;\hat{\theta}(X_1,\ldots,X_n))$$

Prof. Akaike (赤池) in Google logo



Classical Information Criterion: AIC ≃ CV-type Criterion

- **Estimation of Kullbuck-Leibler (KL) divergence**:

Our Target!!

$$\mathrm{KL}(q||p) = \int q(x)\log\frac{q(x)}{p(x;\hat{\theta})}\mathrm{d}x \propto -\int q(x)\log p(x;\hat{\theta})dx =: \tilde{\mathrm{KL}}(q||p)$$

Empirical Dist.

$$\approx -\frac{1}{n}\sum_{i=1}^{n}\log p(X_i;\hat{\theta}(X_1,\ldots,X_n))$$

Prof. Akaike (赤池) in Google logo

- **AIC** (Akaike, 1973)

$$-\frac{1}{n}\sum_{i=1}^{n}\log p(X_i;\hat{\theta}) + \frac{1}{n}\#\dim(\hat{\theta})$$

= Asymptotic Bias

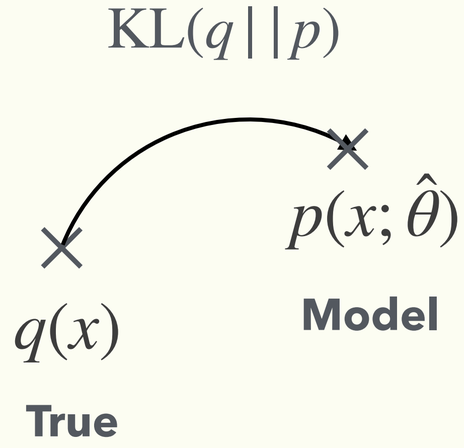
- **CV** (Stone, 1974, JRSSB)

$$-\frac{1}{n}\sum_{i=1}^{n}\log p(X_i;\hat{\theta}^{(-i)})$$

Estimator excluding i-th dataset

Remark.

$$E[\mathrm{CV}] = E[\mathrm{AIC}] + o(n^{-1})$$
$$= E[\tilde{\mathrm{KL}}(q||p)] + o(n^{-1})$$



Pseudo Posterior & Its Bias Correction

- **Pseudo Posterior** of ϕ

Pseudo likelihood

prior

$$\underbrace{\pi_{\text{pos},\phi}}_{\text{Pseudo posterior}}(\phi \mid \mathcal{S}) = \frac{\exp\left\{ \sum_{i=1}^{n} s_\phi(\phi; Z_i, A_i) \right\} \pi(\phi)}{\int \exp\left\{ \sum_{i=1}^{n} s_\phi(\bar{\phi}; Z_i, A_i) \right\} \pi(\bar{\phi}) d\bar{\phi}}$$

  - If $s_\phi(\phi; Z_i, A_i) = \log p(Z_i, A_i; \phi)$ (log-likelihood), $\pi_{\text{pos},\phi}(\phi \mid \mathcal{S})$ is the usual posterior distribution

- Problem: estimation of the Generalization Error (GE)

$$E_{\tilde{Z},\tilde{A}}\left[E_{\text{pos},\phi}\left[\nu(\hat{\theta}[\phi]; \phi^*; \tilde{Z}, \tilde{A}) \,\middle|\, \mathcal{S} \right] \right] \approx \frac{1}{n} \sum_{i=1}^{n} E_{\text{pos},\phi}\left[\nu(\hat{\theta}[\phi]; \phi^*; Z_i, A_i) \,\middle|\, \mathcal{S} \right]$$

Dual use of the same dataset in ESTIMATION and EVALUATION ⇒ Bias