Efficient Multiply Robust Estimation Under Informative Sampling

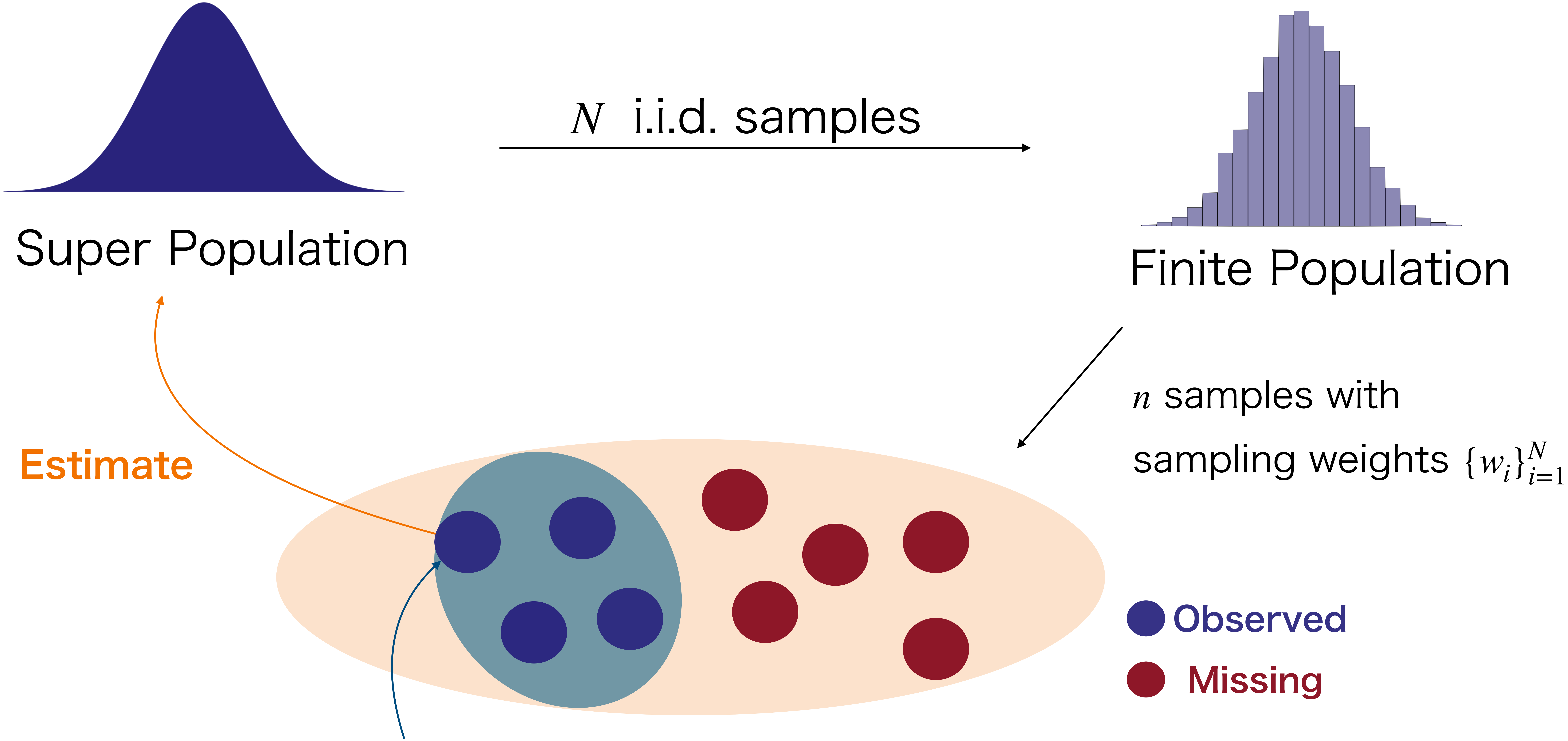
(https://arxiv.org/abs/2311.06719)

August 6th, 2024 Kenji Beppu[1]
e-mail: beppu@sigmath.es.osaka-u.ac.jp

This is joint work with Wataru Aida[1] and Kosuke Morikawa[1].

[1] **Graduate School of Engineering Science, Osaka University**

Strategy for Estimation in Survey Sampling



N i.i.d. samples

Super Population

Finite Population

n samples with

sampling weights $\{w_i\}_{i=1}^{N}$

Estimate

● **Observed**

● **Missing**

Obserbed data

Brief Summary

- In survey sampling, effective use of sampling weights enable us to remove the selection bias in the sample.
- We propose a semiparametric efficient estimator by using the Empirical Likelihood (EL) method that does not require assumptions on a specific distribution.
- The proposed EL method is easily extendable to cases where external information is available.

Overview

- Introduction
- Efficient estimator
- Multiple robust EL estimator
- Numerical experiment
- Conclusion

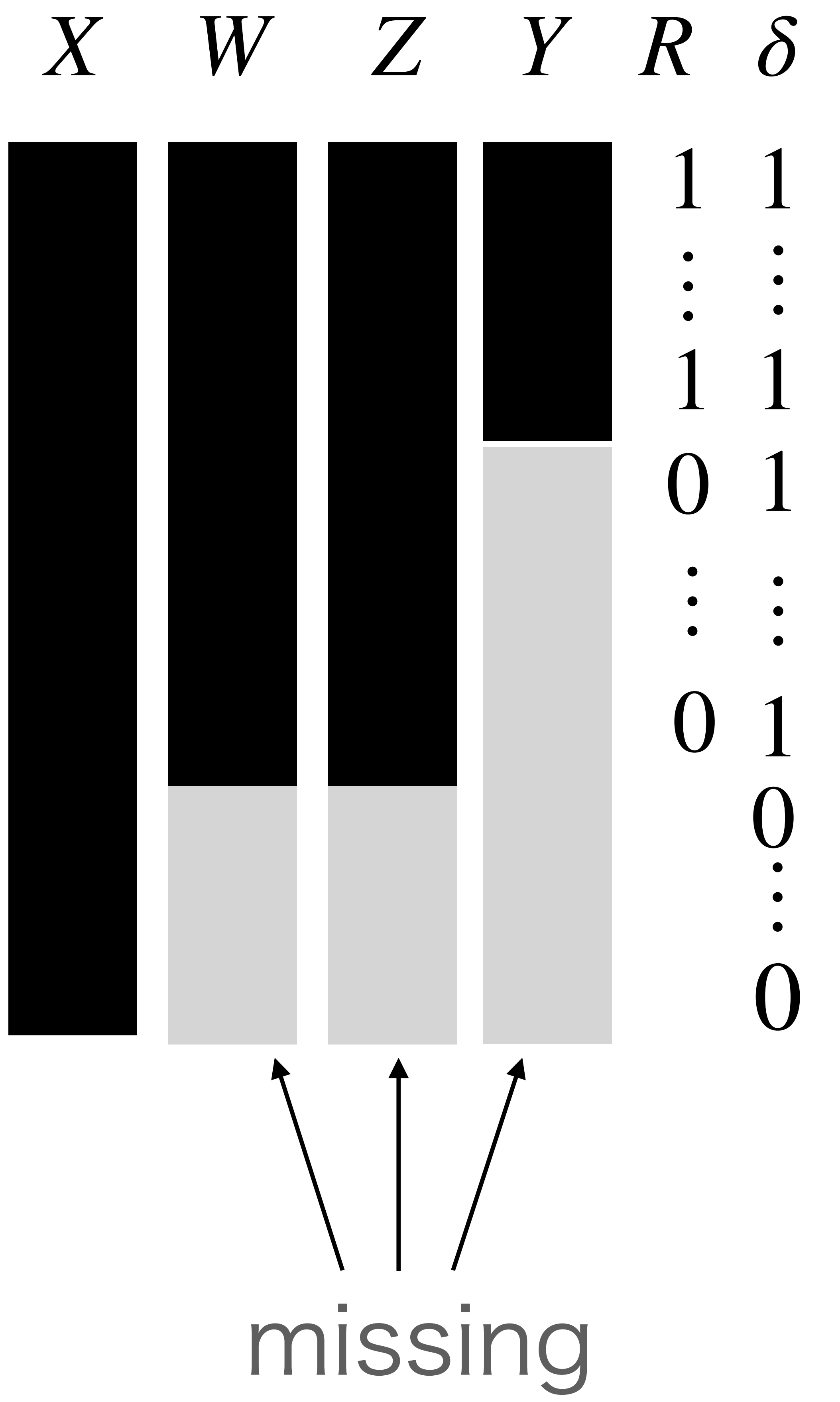
Introduction

Setup

$(X_i, Y_i, Z_i, W_i, \delta_i, R_i)_{i=1,\ldots,N} \overset{i.i.d.}{\sim} F,$

- Y : response variable
- X : covariate of interest
- Z : other covariates
- W : inverse of inclusion probability, i.e., $W := P(\delta = 1 \mid X, Y, Z, W)^{-1}$
- δ : sampling indicator
- $n := \sum_{i=1}^{N} \delta_i$: size of the sampled dataset
- R : response indicator of Y

Setting



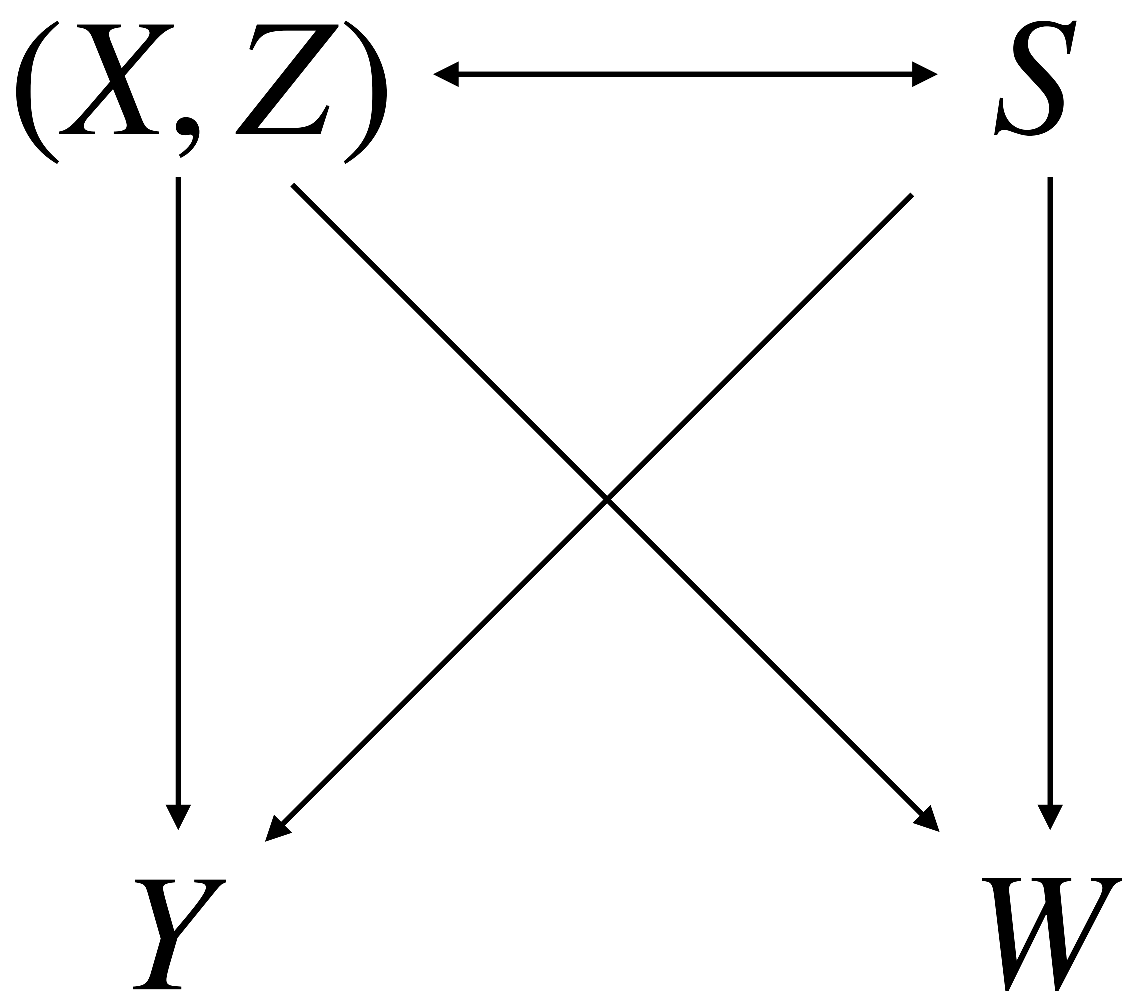
Informative sampling: $W \not\perp Y \mid (X, Z)$

Sample missing at random (Pfefferman, 1993, ISR): $Y \perp R \mid (\delta = 1, X, Z, W)$

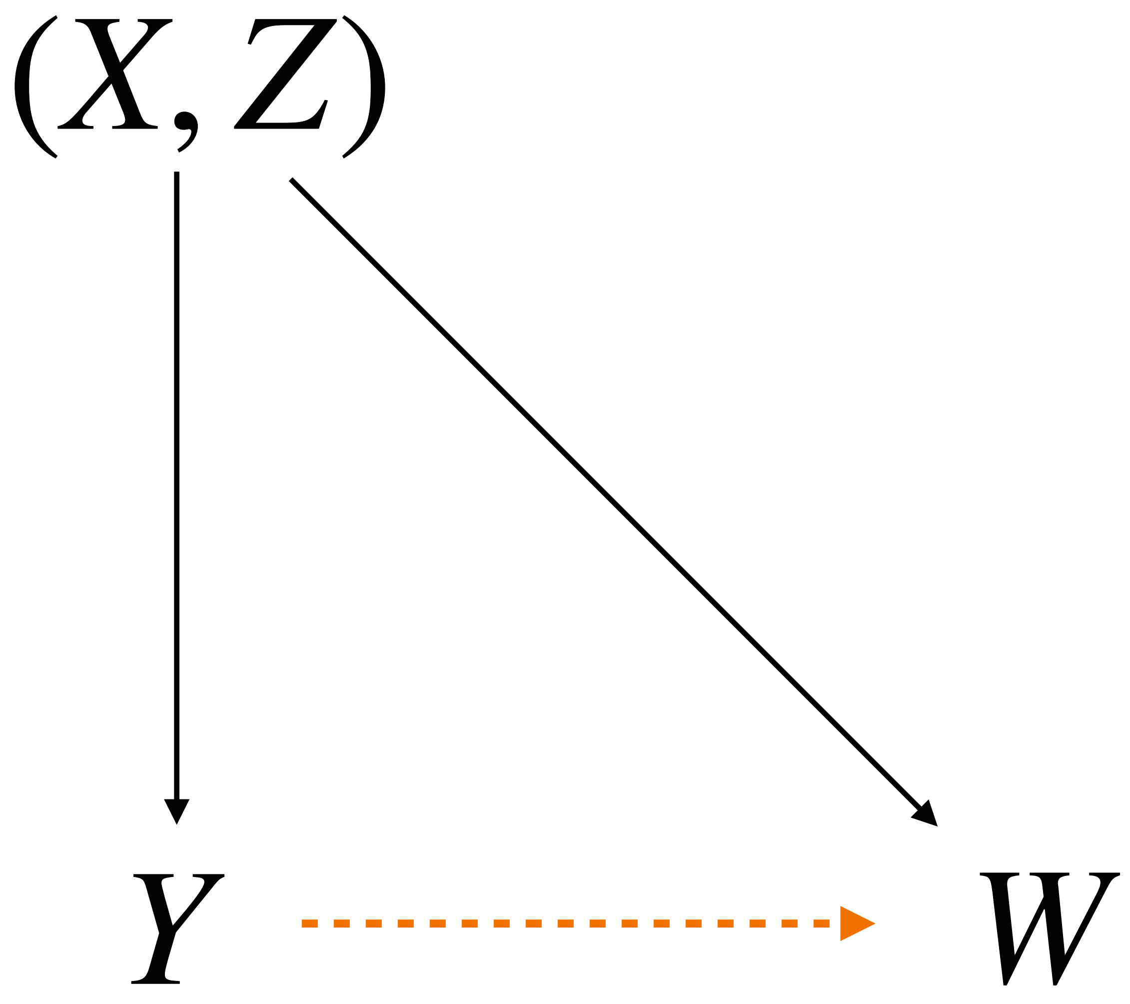
Example of Informative Sampling

Informative sampling: $W \not\perp Y \mid (X, Z)$

Data accessible to survey weight designers



Data accessible to data analysts



→ $W \not\perp Y \mid (X, Z)$

Estimator from previous studies

- Target: Z-estimator θ : the unique solution to

$$E\{U(X, Y; \theta)\} = 0$$

e.g. Mean of the response variable $\theta = E(Y) \Rightarrow U(X, Y; \theta) = \theta - Y$

- Doubly Robust (DR) estimator (Kim and Haziza, 2014): the solution to

$$\sum_{i=1}^{N} \delta_i W_i \left[\frac{R_i}{\pi_i} U(X_i, Y_i; \theta) + \left(1 - \frac{R_i}{\pi_i}\right) E\left\{U(X_i, Y; \theta) \mid X_i, Z_i, W_i\right\} \right] = 0$$

- DR estimator has double robustness and asymptotic normality but does NOT have semiparametric efficiency.

Our aim and contribution

- Deriving the semiparametric efficient score function.
- Proposing a double robust semiparametric efficient estimator.
- Constructing a multiple robust EL estimator.

Efficient estimator

What is semiparametric efficiency?

- The semiparametric efficient estimator gives the smallest asymptotic variance in the class of Regular and Asymptotic Linear (RAL) estimators.
  - $\hat{\theta}_{\mathrm{eff}}$ is the semiparametric efficient estimator (Bickel et al., 1993; Tsiatis, 2006)

$$\Leftrightarrow \sqrt{N}(\hat{\theta}_{\mathrm{eff}} - \theta_0) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \varphi_{\mathrm{eff}}(X_i, Y_i, Z_i, W_i, \delta_i, R_i) + o_p(1) \text{ and}$$

$$\mathrm{Var}(\varphi_{\mathrm{eff}}) \leq \mathrm{Var}(\varphi)$$

$$\left({}^{\forall}\varphi \in \left\{ \varphi \in \mathscr{H};\ \hat{\theta} \text{ is RAL estimartor, } \sqrt{N}(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \varphi(X_i, Y_i, Z_i, W_i, \delta_i, R_i) + o_p(1) \right\} \right)$$

- φ_{eff} is called the “efficient influence function”.
- $\varphi_{\theta,\,\mathrm{Eff}} = E\{S_{\theta,\,\mathrm{Eff}}^{\otimes 2}\}^{-1} S_{\theta,\,\mathrm{Eff}}$, $S_{\theta,\,\mathrm{Eff}}$ is called the “efficient score function”.
- φ_{eff} and $S_{\theta,\,\mathrm{Eff}}$ have the one-to-one correspondence.

Efficient score

Theorem 1.

The efficient score function in this setting is

$$S_{\theta,\mathrm{Eff}} = \underbrace{\delta W \left[\frac{R}{\pi} U(X,Y;\theta) + \left(1 - \frac{R}{\pi}\right) g_\theta(X,Z,W) \right]}_{S_{\theta,\ \mathrm{DR}}} + \underbrace{(1-\delta W) C(X;\theta)}_{S_{\theta,\ \mathrm{Aug}}}.$$

where $\pi := \pi(x,z,w) = P(R=1 \mid \delta = 1, x, z, w)$, $g_\theta(X,Z,W) = E\left\{U(X,Y;\theta) \mid X,Z,W\right\}$,

$$C(X;\theta) = \frac{E\left[(W-1)U(X,Y;\theta) \mid X\right]}{E(W-1 \mid X)}.$$

- $S_{\theta,DR}$: double robust estimator (e.g. Kim and Haziza, 2014)
- $S_{\theta,\ \mathrm{Aug}}$; augmentation term

Adaptive estimator

- In practice, the functions $\pi(x, z, w)$, $g_\theta(x, z, w)$, $C(X; \theta)$ are generally unknown

  → some working models are required

- The method of moments estimator $\hat{\theta}_{\text{MM}}$ is the solution to

$$\sum_{i=1}^{N}\left[\delta_i W_i\left\{\frac{R_i U(X_i, Y_i; \theta)}{\hat{\pi}(X_i, Z_i, W_i)} + \left(1 - \frac{R_i}{\hat{\pi}(X_i, Z_i, W_i)}\right)\hat{g}_\theta(X_i, Z_i, W_i)\right\} + (1 - \delta_i W_i)\hat{C}(X_i; \theta)\right] = 0.$$

Theorem 2.

Under some regularity conditions,

- (Efficiency) if all the working models are correct, the estimator $\hat{\theta}_{\text{MM}}$ achieves the efficiency bound;
- (Double robustness) if either $\pi(X, Z, W)$ or $g_\theta(X, Z, W)$ is correct, the estimator $\hat{\theta}_{\text{MM}}$ has consistency and asymptotic normality.

Multiple robust EL estimator

Empirical Likelihood (EL) estimator

Advantages of the EL method for constructing estimators :

- Easy derivation of multiple robust estimators
- Possible integration of previous studies and external information

For multiple robust estimators, consider the following candidate models

$\mathscr{P} = \{\pi^{[j]};\ j = 1, \ldots, J\}$: multiple models for π

$\mathscr{G} = \{g_\theta^{[k]};\ k = 1, \ldots, K\}$: multiple models for g_θ

$\mathscr{C} = \{C_\theta^{[l]};\ l = 1, \ldots, L\}$: multiple models for $C(X; \theta)$

Proposed EL estimator

The proposed EL estimator $\hat{\theta}_{\mathrm{EL}}$: the solution to

$$\sum_{i=1}^{N} \hat{p}_i^{(1)} \hat{p}_i^{(2)} \delta_i R_i W_i U(X_i, Y_i; \theta) = 0$$

First step

$$\hat{\theta}_1 = \arg\max_{\theta} \arg\max_{p_i^{(1)}} \underline{\sum_{i=1}^{n_1} \log p_i^{(1)}(\theta)},$$

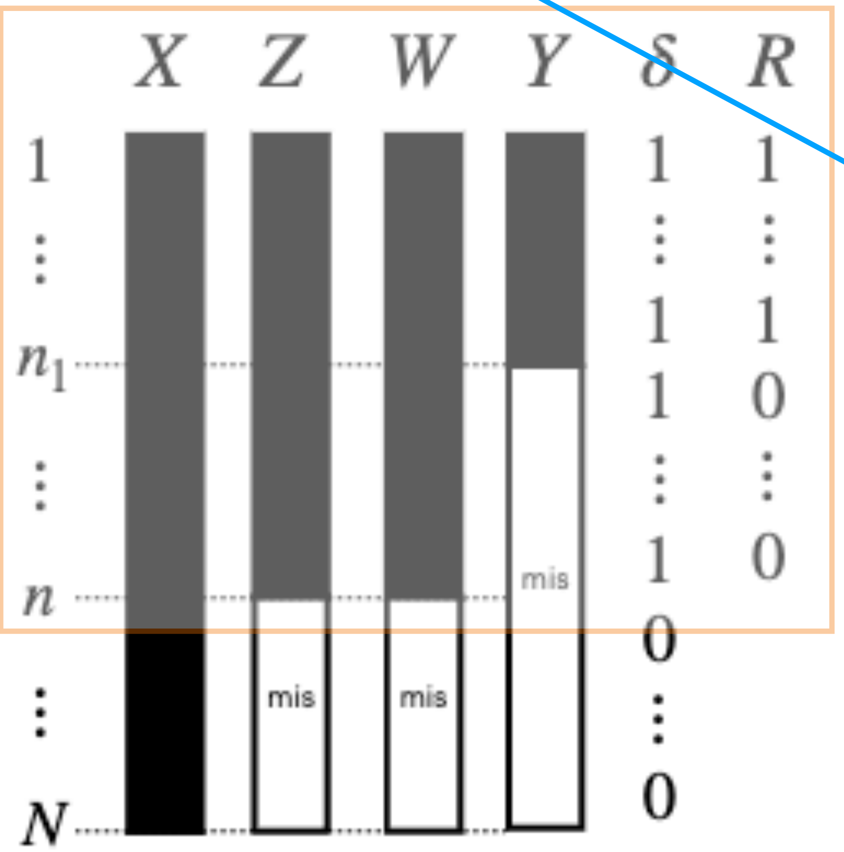
subject to $\sum_{i=1}^{n_1} p_i^{(1)} = 1,$

$$\sum_{i=1}^{n_1} p_i^{(1)} \left\{ \hat{\pi}^{[j]}(X_i, Z_i, W_i) - \bar{\pi}_n^{[j]} \right\} = 0 \quad (j = 1, \ldots, J),$$

$$\sum_{i=1}^{n_1} p_i^{(1)} \left\{ W_i \hat{g}_\theta^{[k]}(X_i, Z_i, W_i) - \bar{g}_\theta^{w[k]} \right\} = 0 \quad (k = 1, \ldots, K),$$

where

$$\bar{\pi}_n^{[j]} = \frac{1}{n} \sum_{i=1}^{n} \hat{\pi}^{[j]}(X_i, Z_i, W_i), \quad \bar{g}_\theta^{w[k]} = \frac{1}{n} \sum_{i=1}^{n} W_i \hat{g}_\theta^{[k]}(X_i, Z_i, W_i).$$

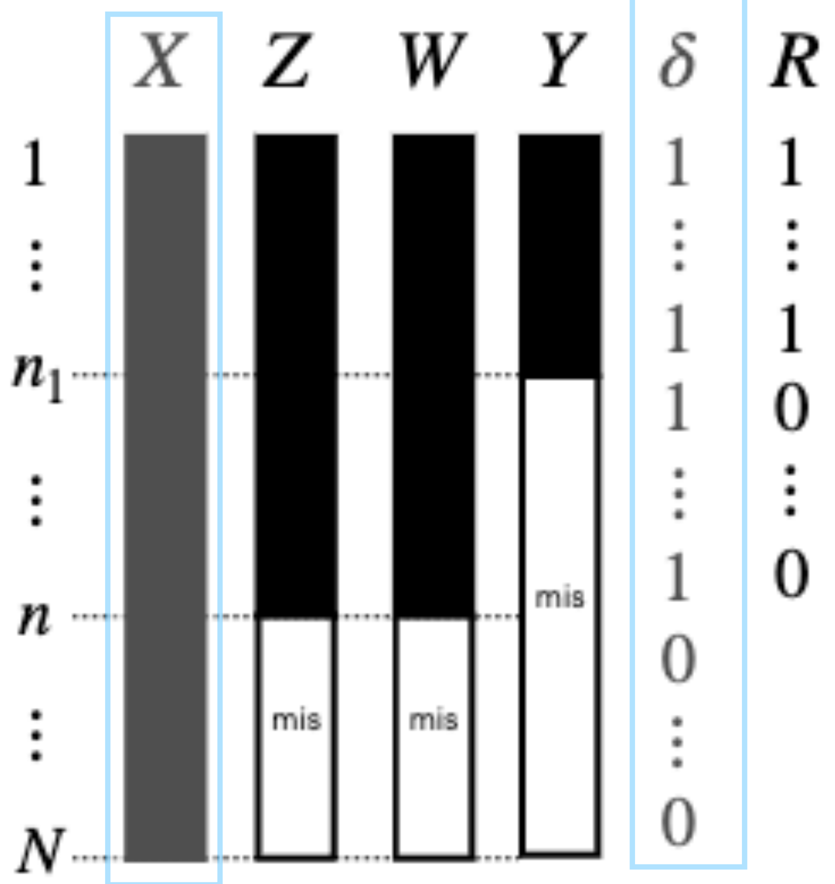


Second step

$$\hat{\theta}_2 = \arg\max_{\theta} \arg\max_{p_i^{(2)}} \underline{\sum_{i=1}^{N} \log p_i^{(2)}},$$

subject to $\sum_{i=1}^{n_1} p_i^{(2)} = 1,$

$$\sum_{i=1}^{N} p_i^{(2)} (1 - \delta_i W_i) \hat{C}_\theta^{[l]}(X_i) = 0 \quad (l = 1, \ldots, L).$$



Theoretical result of EL estimator

$\mathscr{P} = \{\pi^{[j]};\ j = 1,\ldots,J\}$: multiple models for π

$\mathscr{G} = \{g_\theta^{[k]};\ k = 1,\ldots,K\}$: multiple models for g_θ

$\mathscr{C} = \{C_\theta^{[l]};\ l = 1,\ldots,L\}$: multiple models for $C(X;\theta)$

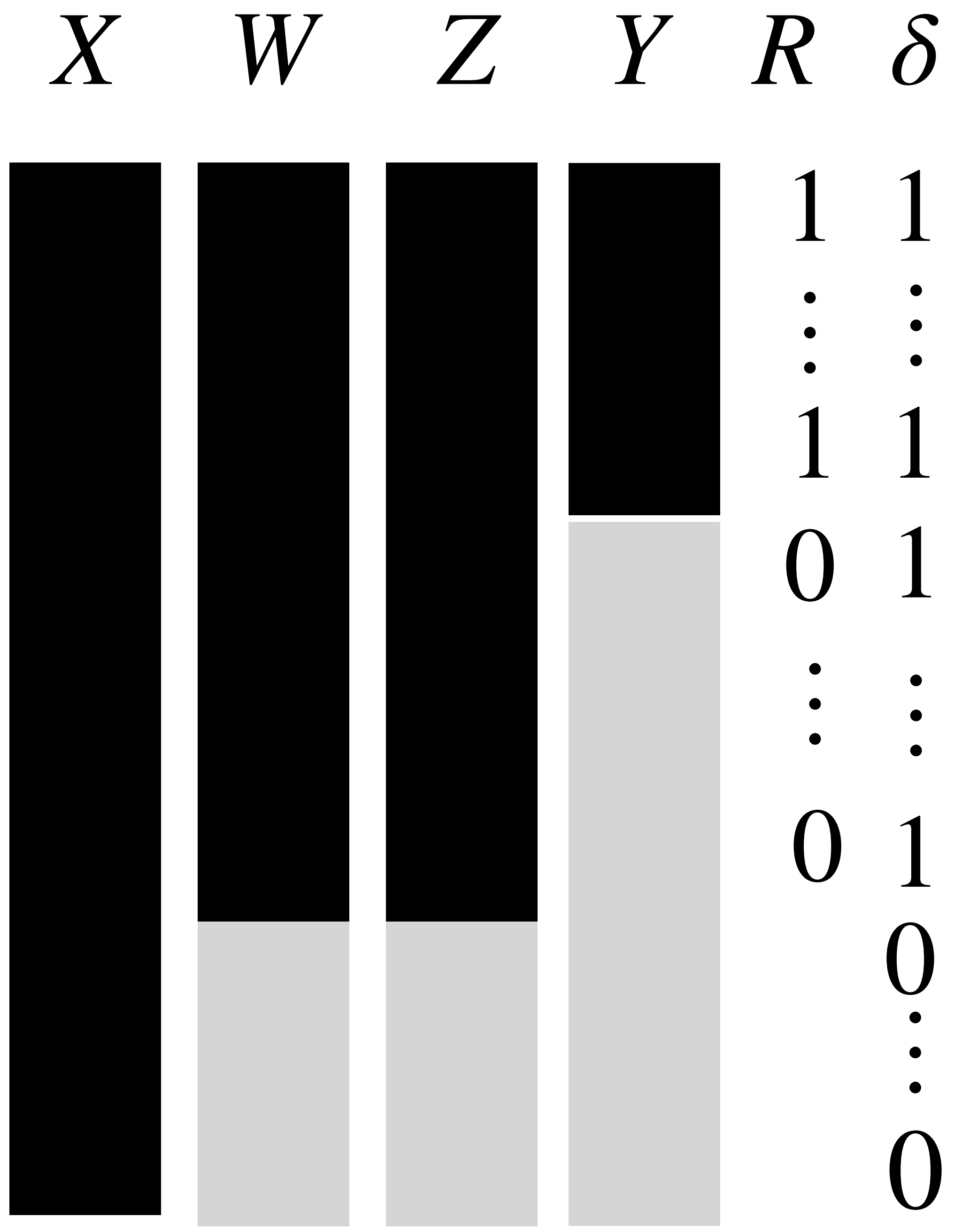
Theorem 3.

Under some regularity conditions,

- (Multiple robustness) When $J+K$ models have at least one correct model in $\mathscr{P}$ and $\mathscr{G}$, the proposed estimators $\hat{\theta}_{\mathrm{EL}}$ have consistency.
- (Semiparametric Efficiency) Moreover, $\hat{\theta}_{\mathrm{EL}}$ are efficient if each $\mathscr{P}$, $\mathscr{G}$, $\mathscr{C}$ contain at least one correct model.

EL method with external information

Available data are
- Individual data from an internal study
- Summary statistics from relevant external studies.



+

External data

$\hat{\beta}$: summary statistics of X

e.g. $\bar{X}$

How can we use the information of $\hat{\beta}$?

- $\hat{\beta}$ is an estimator of some functional β that is the unique solution to $E\{\psi(X;\beta)\} = 0$
- Σ_1: The asymptotic variance of $\hat{\beta}$
- A consistent estimator $\hat{\Sigma}_1$ is also available

Proposed EL estimator

The proposed EL estimator $\hat{\theta}_{\text{EL,ext}}$: the solution to

$$\sum_{i=1}^{N} \hat{p}_i^{(1)} \hat{p}_i^{(2)} \delta_i R_i W_i U(X_i, Y_i; \theta) = 0$$

First step $p_i^{(1)}$ is the same.

Second step (Zhang et al., 2020)

$$\hat{\theta}_2 = \arg\max_{\theta,\beta} \arg\max_{p_i^{(2)}} \sum_{i=1}^{N} \log p_i^{(2)} - \frac{m}{2}(\hat{\beta} - \beta)^\top \hat{\Sigma}_1^{-1} (\hat{\beta} - \beta),$$

subject to $\sum_{i=1}^{n_1} p_i^{(2)} = 1,\ \sum_{i=1}^{N} p_i^{(2)} (1 - \delta_i W_i) \hat{C}_\theta^{[l]}(X_i) = 0 \quad (l = 1, \ldots, L),\ \sum_{i=1}^{N} p_i^{(2)} \psi(X_i; \beta) = 0$.

Numerical experiment

HT and KH estimators

We compared Horvitz-Thompson (HT) and Kim and Haziza (KH) estimators with our proposed multiply-robust empirical likelihood estimators.

HT :
$$\sum_{i=1}^{N} \frac{\delta_i R_i W_i}{\hat{\pi}_i}(Y_i - \hat{\theta}_{\mathrm{HT}}) = 0 \Leftrightarrow \hat{\theta}_{\mathrm{HT}} := \frac{\sum_{i=1}^{N} \delta_i R_i W_i Y_i / \hat{\pi}_i}{\sum_{i=1}^{N} \delta_i R_i W_i / \hat{\pi}_i}.$$

KH :
$$\hat{\theta}_{\mathrm{KH}} := \frac{1}{N} \sum_{i=1}^{N} \delta_i W_i \left\{ \frac{R_i Y_i}{\hat{\pi}_i} + \left(1 - \frac{R_i}{\hat{\pi}_i}\right) \hat{E}\left(Y \mid X_i, Z_i, W_i\right) \right\}.$$

Working models for π and g

We prepared three models for each π, m :

$$\text{(Correct)}\ \pi^{(1)}(x, z, w; \phi^{(1)}) = \text{expit}(\phi_0^{(1)} + \phi_1^{(1)}x + \phi_2^{(1)}z + \phi_3^{(1)}w),$$

$$\pi^{(2)}(x, z, w; \phi^{(2)}) = \text{expit}(\phi_0^{(2)} + \phi_1^{(2)}x + \phi_2^{(2)}z + \phi_3^{(2)}xz),$$

$$\pi^{(3)}(x, z, w; \phi^{(3)}) = \text{expit}\{s_1(x; \phi^{(3)}) + s_2(z; \phi^{(3)})\},$$

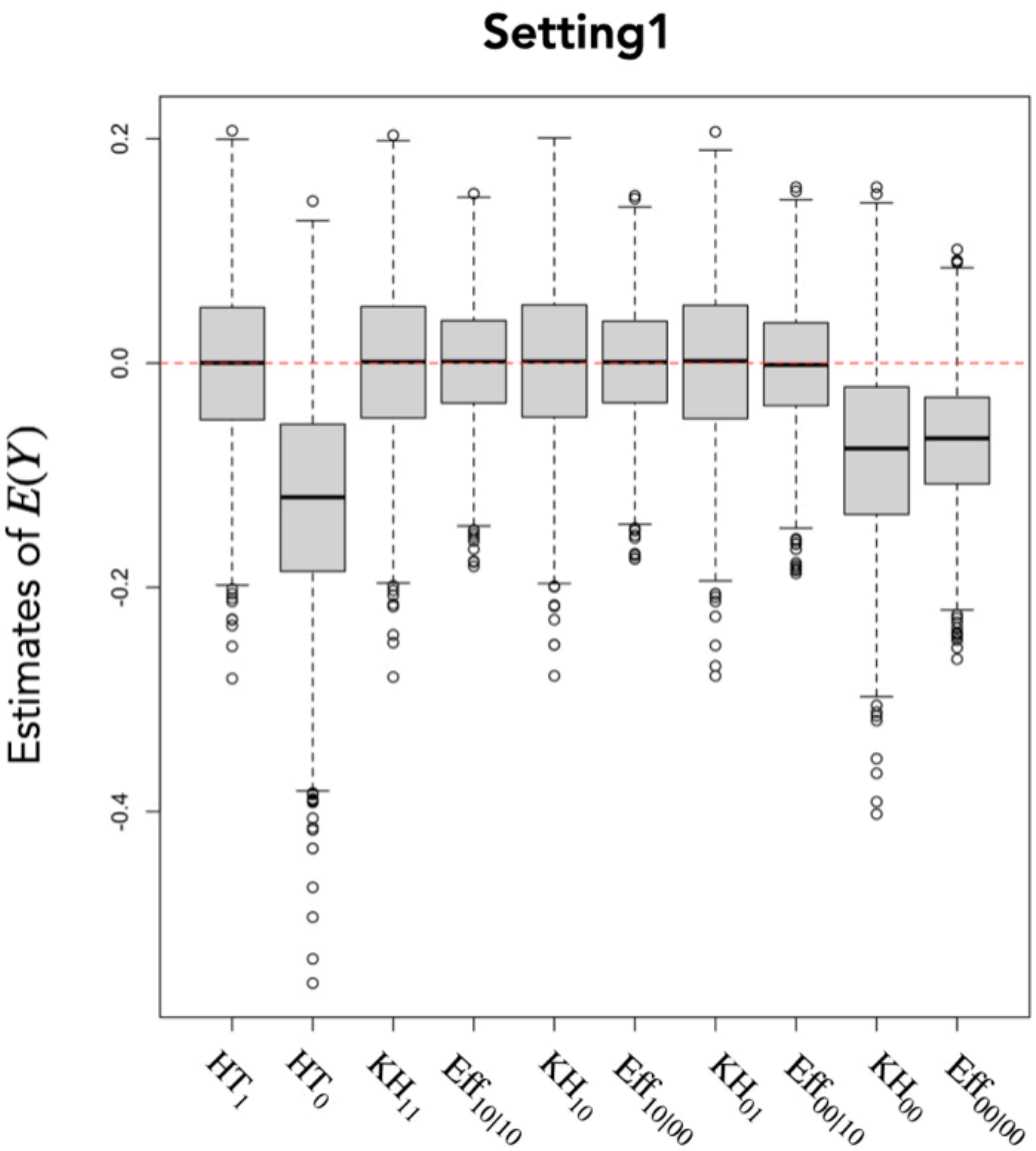
$$\text{(Correct)}\ g^{(1)}(x, z, w; \xi^{(1)}) = \xi_0^{(1)} + \xi_1^{(1)}x + \xi_2^{(1)}z + \xi_3^{(1)}\log(w-1),$$

$$g^{(2)}(x, z, w; \xi^{(2)}) = \xi_0^{(2)} + \xi_1^{(2)}x + \xi_2^{(2)}z + \xi_3^{(2)}xz,$$

$$g^{(3)}(x, z, w; \xi^{(3)}) = \tilde{s}_1(x; \phi^{(3)}) + \tilde{s}_2(z; \phi^{(3)}),$$

. $s_j, \tilde{s}_j$: smoothing splines with three degrees of freedom.

Result


Setting1

We prepared four models for our multiply-robust estimators:

$\mathrm{Eff}_{10|10}$: combination of $\pi^{(1)}, \pi^{(2)}, g^{(1)}, g^{(2)}$;

$\mathrm{Eff}_{10|00}$: combination of $\pi^{(1)}, \pi^{(2)}, g^{(2)}, g^{(3)}$;

$\mathrm{Eff}_{00|10}$: combination of $\pi^{(2)}, \pi^{(3)}, g^{(1)}, g^{(2)}$;

$\mathrm{Eff}_{00|00}$: combination of $\pi^{(2)}, \pi^{(3)}, g^{(2)}, g^{(3)}$.

.HT_i , KH_{ij} , $\mathrm{Eff}_{ij|kl}$:

$i, j, k, l = 1 \rightarrow$ correct working model

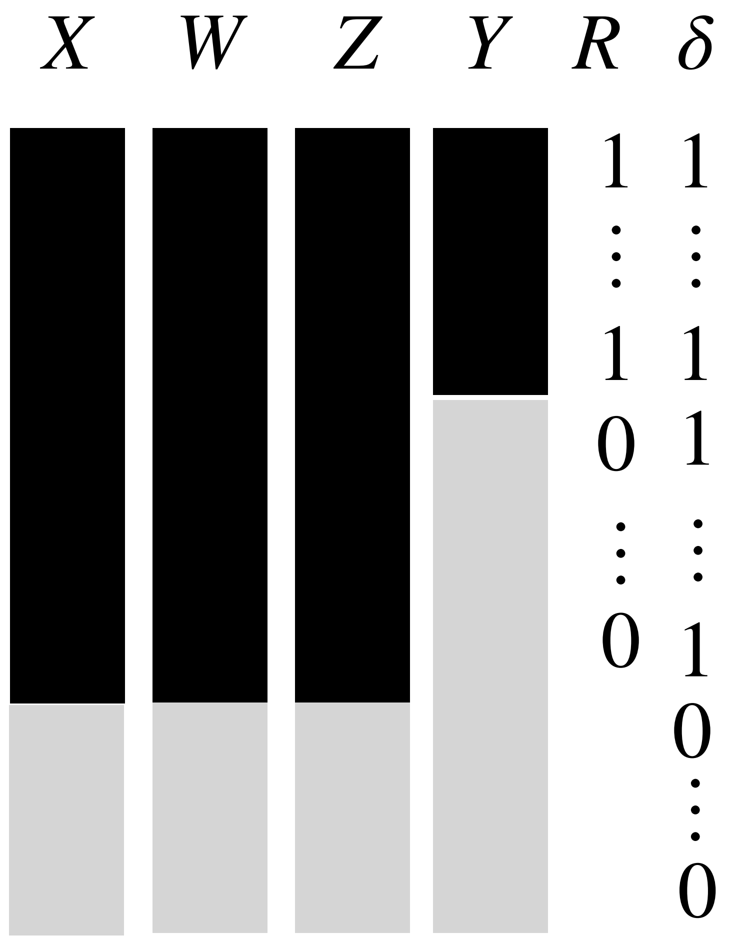
・KH: double robustness but less efficient

Conclusion

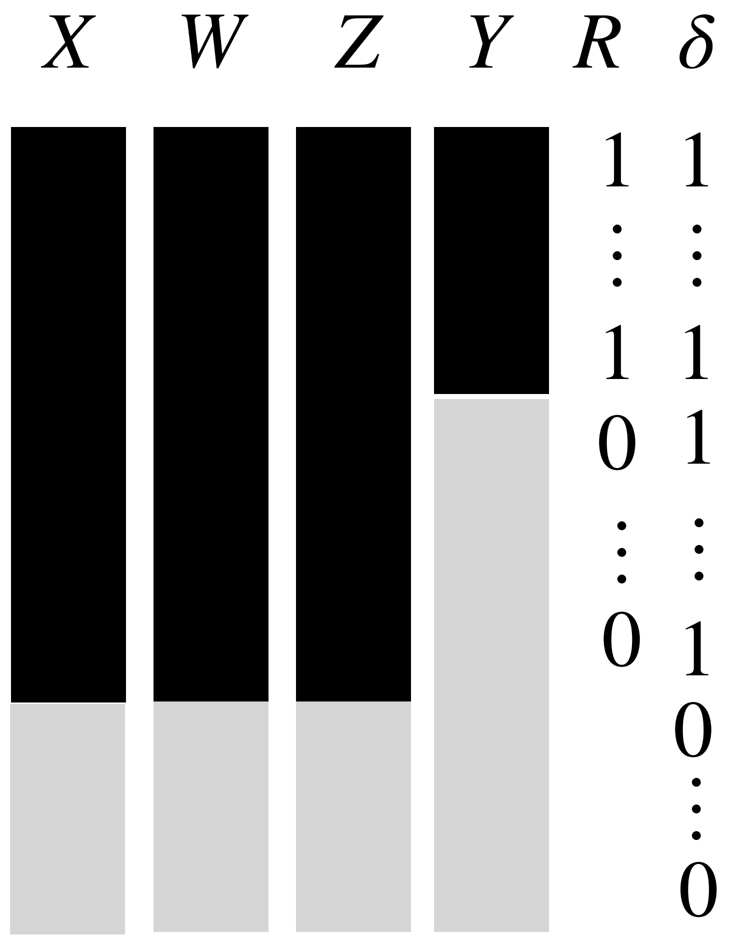
Contribution

- We derived the semiparametric efficient score function and proposed a double robust estimator.
- By using the two-step EL method, we proposed the EL estimator with multiple robustness.
- We provided a similar argument in Setting 2. As Setting 3, we also proposed an EL estimator when an external summary statistic is available. See Morikawa, Beppu, and Aida (2023, arXiv: 2311.06719) for details.

Setting 2



Setting 3



+

External summary statistics e.g. $\bar{X}$

Thank you !

← arXiv link

References

- Bickel, P. J., Klaassen, C. A., Ritov, Y. A. and Wellner, J. A. (1993). Efficient and adaptive estimation for semiparametric models. Springer.
- Kim, J. K. and Haziza, D. (2014). Doubly robust inference with missing data in survey sampling. *Statistica Sinica*, **24**(1), 375-394.
- Morikawa, K., Beppu, K. and Aida, W. (2023). Efficient Multiply Robust Estimation Under Informative Sampling. *arXiv:2311.06719*.
- Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**(2), 237-249.
- Pfeffermann, D. (1993) The role of sampling weights when modeling survey data. *International Statistical Review,* **61**, 317–337.
- Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *Annals of Statistics*, **22**(1), 300-325.
- Tsiatis, A. A. (2006). Semiparametric theory and missing data. Springer.

Appendix

Setup

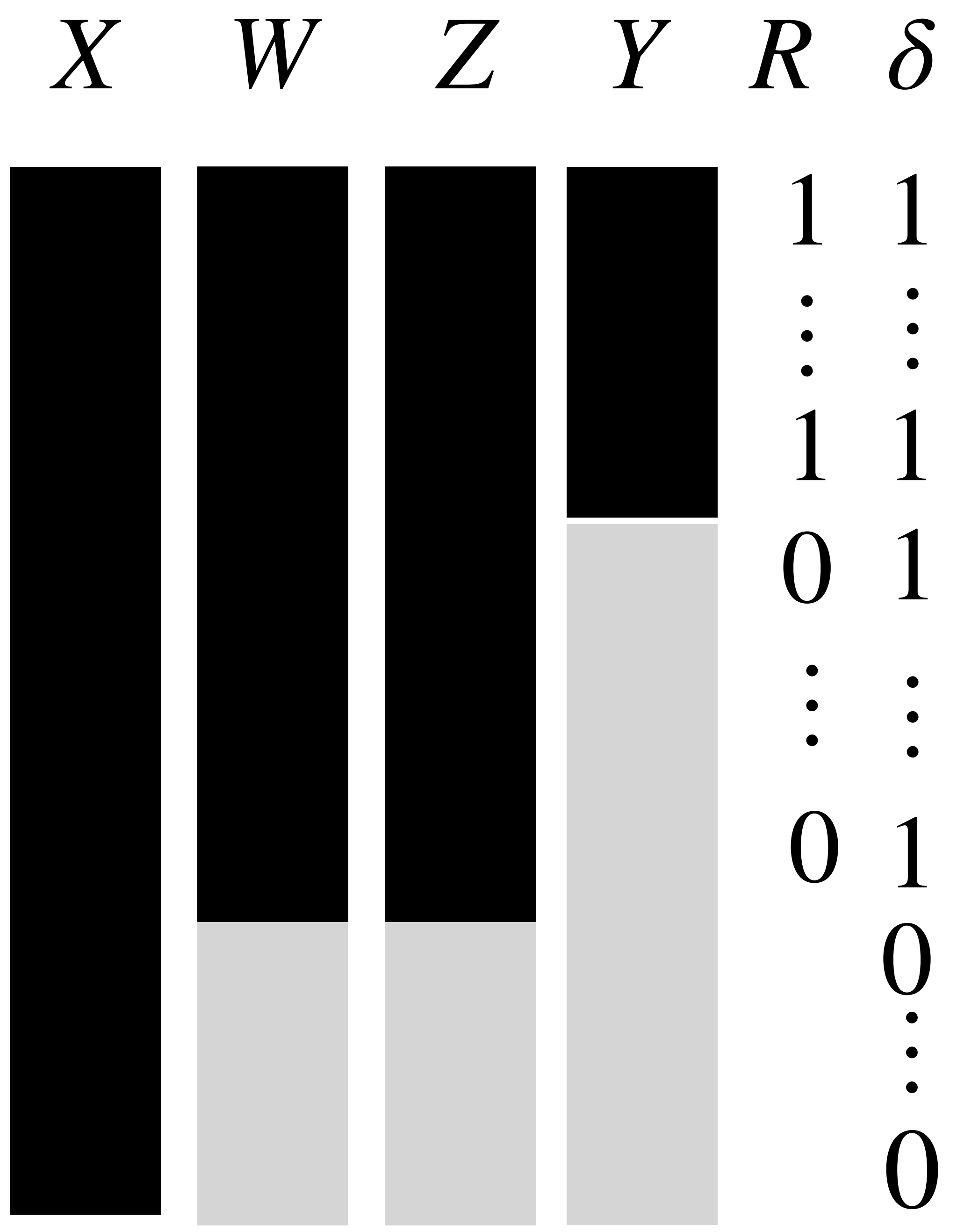
$$(X_i, Y_i, Z_i, W_i, \delta_i, R_i)_{i=1,\ldots,N} \overset{i.i.d.}{\sim} F,$$

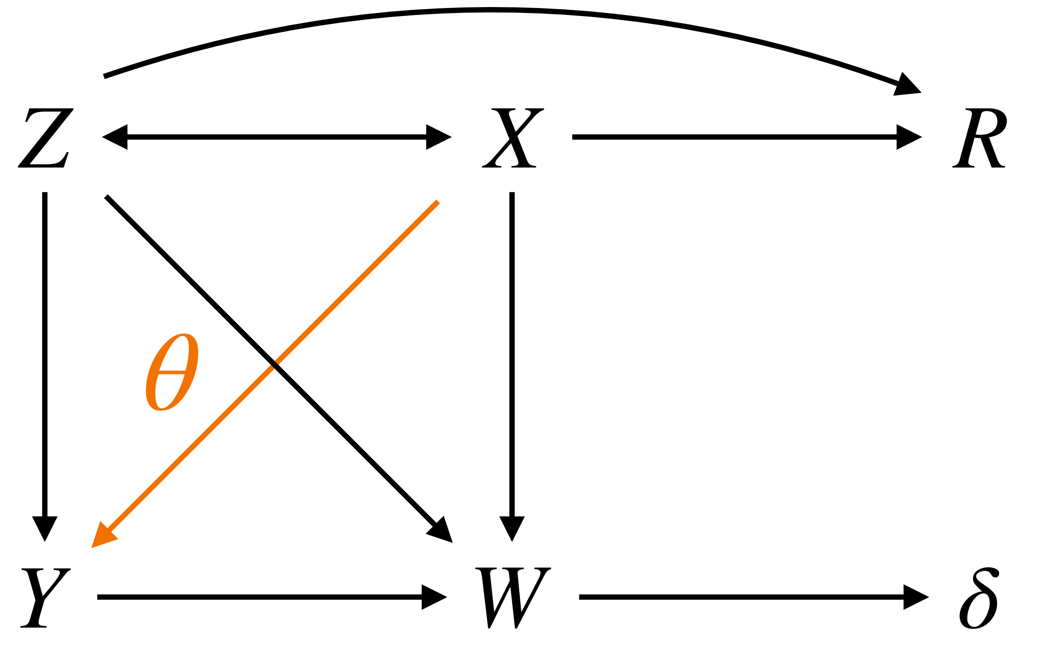
- Y : response variable
- X : covariate of interest
- Z : other covariates
- W : inverse of inclusion probability, i.e., $W := P(\delta = 1 \mid X, Y, Z, W)^{-1}$
- δ : sampling indicator
- $n := \sum_{i=1}^{N} \delta_i$: size of the sampled dataset
- R : response indicator of Y

Informative sampling: $W \not\perp Y \mid (X, Z)$

Sample missing at random (Pfefferman, 1993, ISR): $Y \perp R \mid (\delta = 1, X, Z, W)$

Setting





Adaptive estimator for $E(Y)$

- For example, we consider the estimation of $E(Y)$ (i.e. $U(X, Y, ; \theta) = Y - \theta$) the method of moments. The estimator is

$$\hat{\theta} = \frac{1}{N}\sum_{i=1}^{N}\left[\delta_i W_i \left\{ \frac{R_i Y_i}{\hat{\pi}(X_i, Z_i, W_i)} + \left(1 - \frac{R_i}{\hat{\pi}(X_i, Z_i, W_i)}\right)\hat{g}(X_i, Z_i, W_i) \right\} + (1 - \delta_i W_i)\hat{\kappa}(X_i)\right],$$

where $\hat{\kappa}(x)$ is an estimated parametric working model for

$$\kappa(x) = \frac{E\{(W-1)Y \mid x\}}{E(W-1 \mid x)}$$

Empirical Likelihood (EL)

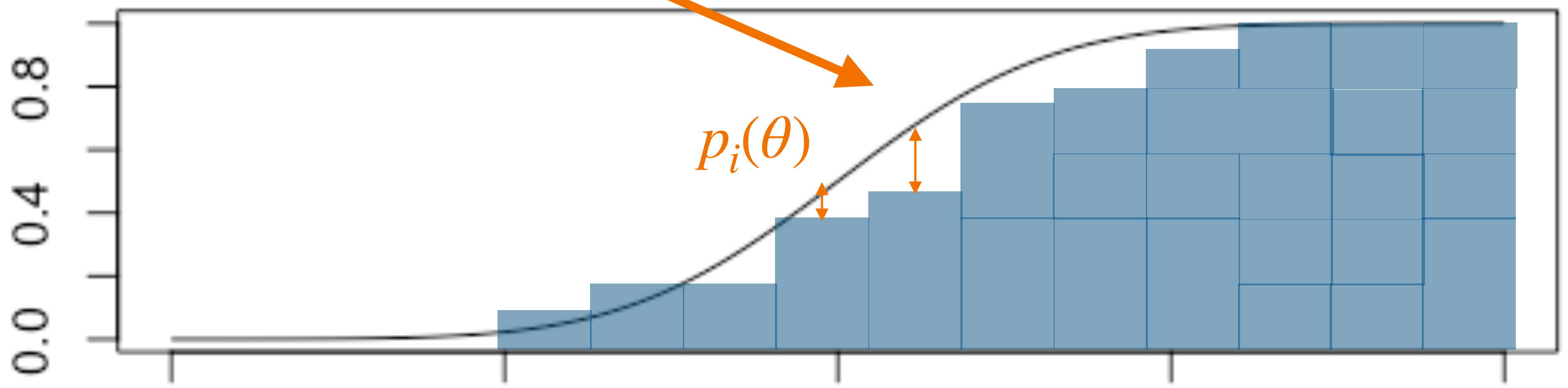
- EL is a non-semi parametric method commonly applied in survey sampling (Owen, 1988, Biometrika; Qin, 1994, AoS)
- Empirical weights p_i are estimated subject to constraints:

$$\arg\max_{\theta} \arg\max_{p_1,\ldots,p_n} \prod_{i=1}^{N} p_i$$

subject to $\sum_{i=1}^{N} p_i = 1,\ \sum_{i=1}^{N} p_i U(y_i;\theta) = 0,\ \cdots$

uneven step widths.

$$F_n(x) = \sum_{i=1}^{N} p_i(\theta) I(y_i \le x)$$



We propose the multiple robust semiparametric estimator by using EL

Model Candidates：

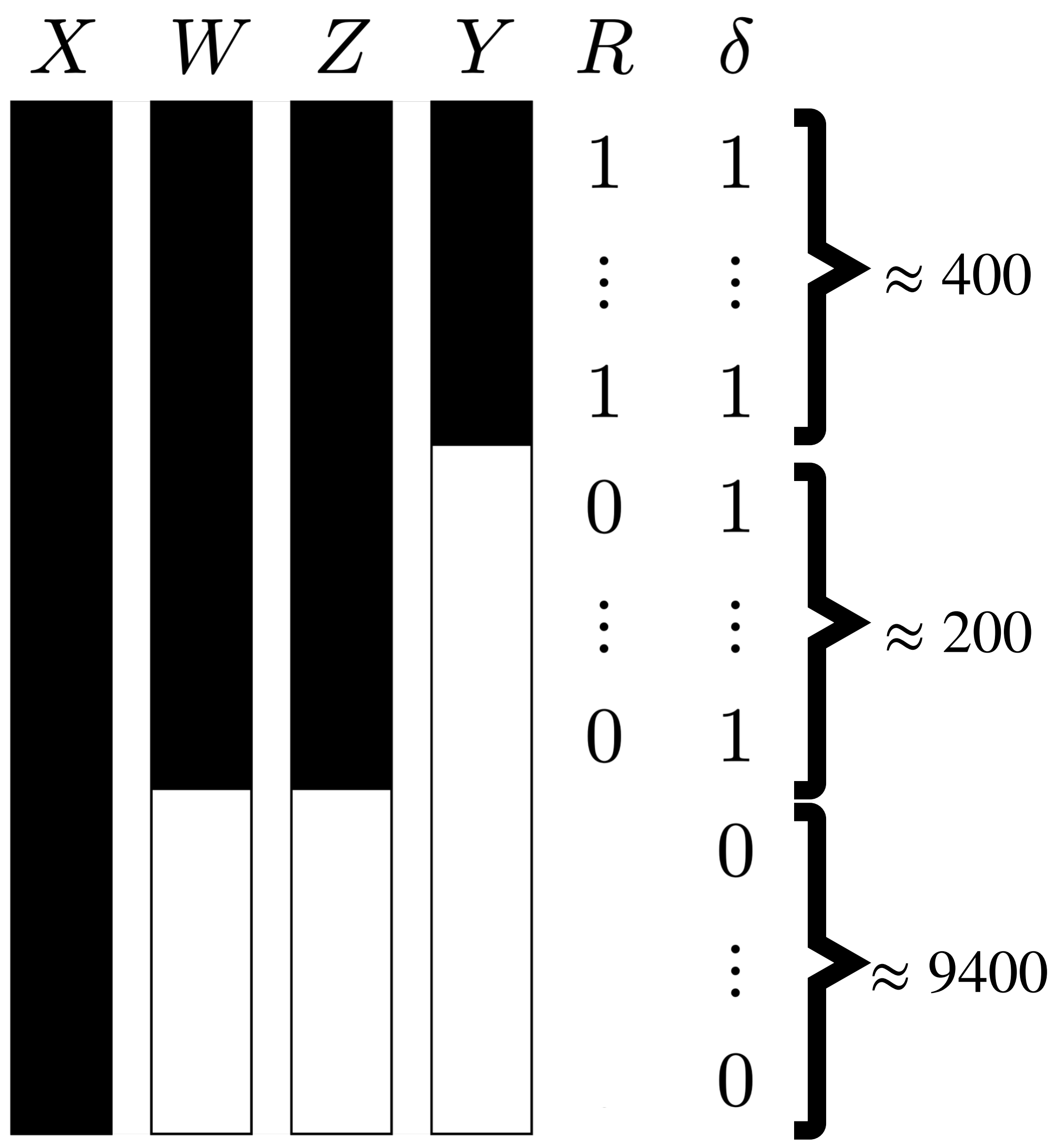
$\mathscr{P} = \{\pi^{[j]};\ j = 1,\ldots,J\}$: multiple models for π,

$\mathscr{G} = \{g_\theta^{[k]};\ k = 1,\ldots,K\}$: multiple models for g

$\mathscr{C} = \{C_\theta^{[l]};\ l = 1,\ldots,L\}$: multiple models for $C(X;\theta)$

Setup

- $X \sim N\left(0, \frac{1}{2}\right), Z \sim N\left(0, \frac{1}{2}\right), Y \mid x, z \sim N\left(x - z, \frac{1}{2}\right).$
- $\log(W - 1) \mid x, y, z \sim N\left(2.95 - 0.25x - 0.45y - 0.1z, \sqrt{0.05}^2\right)$
- $\delta \mid w \sim \text{Binom}(w^{-1})$
- $P(R = 1 \mid \delta = 1, x, z, w) = \dfrac{\exp(-0.3 + 0.75x - 0.5z + 0.05w)}{1 + \exp(-0.3 + 0.75x - 0.5z + 0.05w)}$
- $N = 10000$, iteration $= 2000.$
- Target: $\theta = E(Y)$



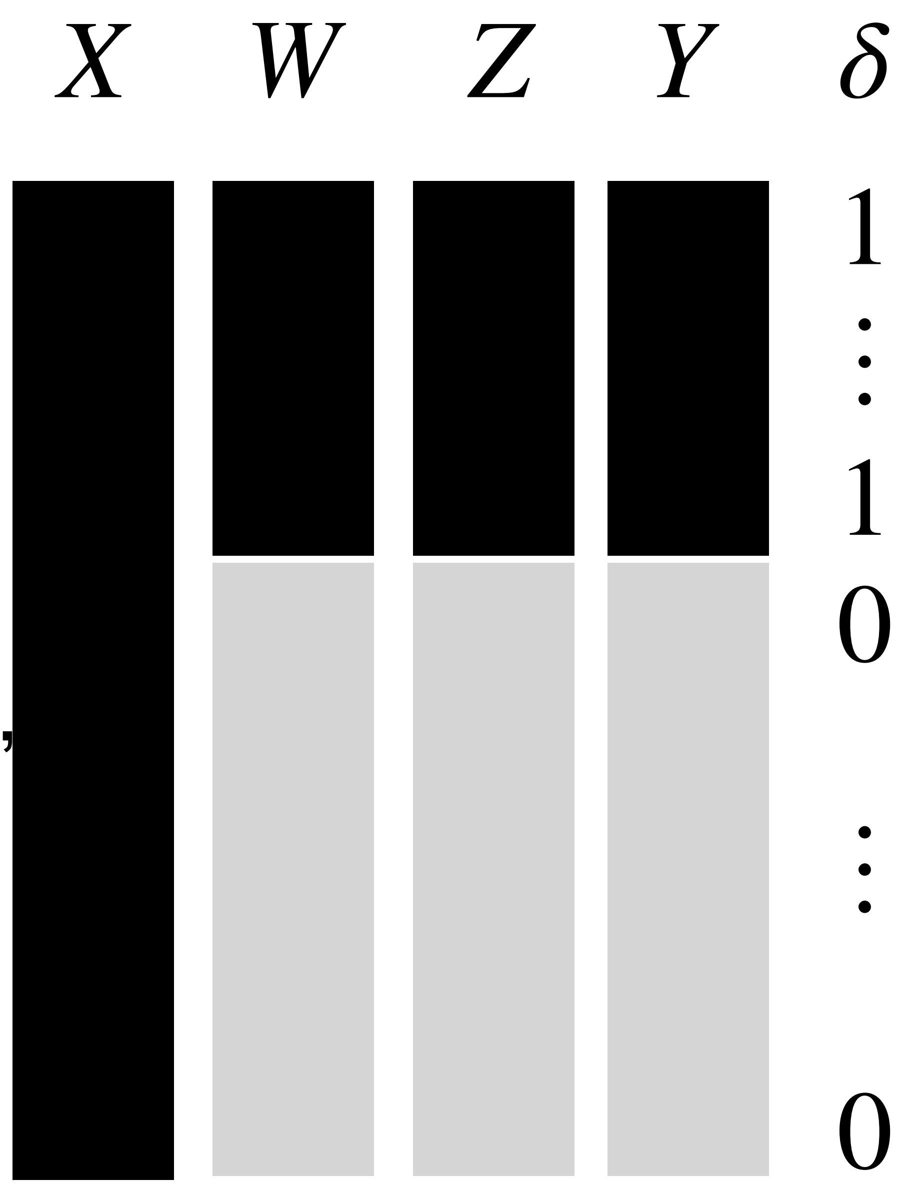
Setup

$(X_i, Y_i, Z_i, W_i, \delta_i)_{i=1,\dots,N} \sim F$, i.i.d.

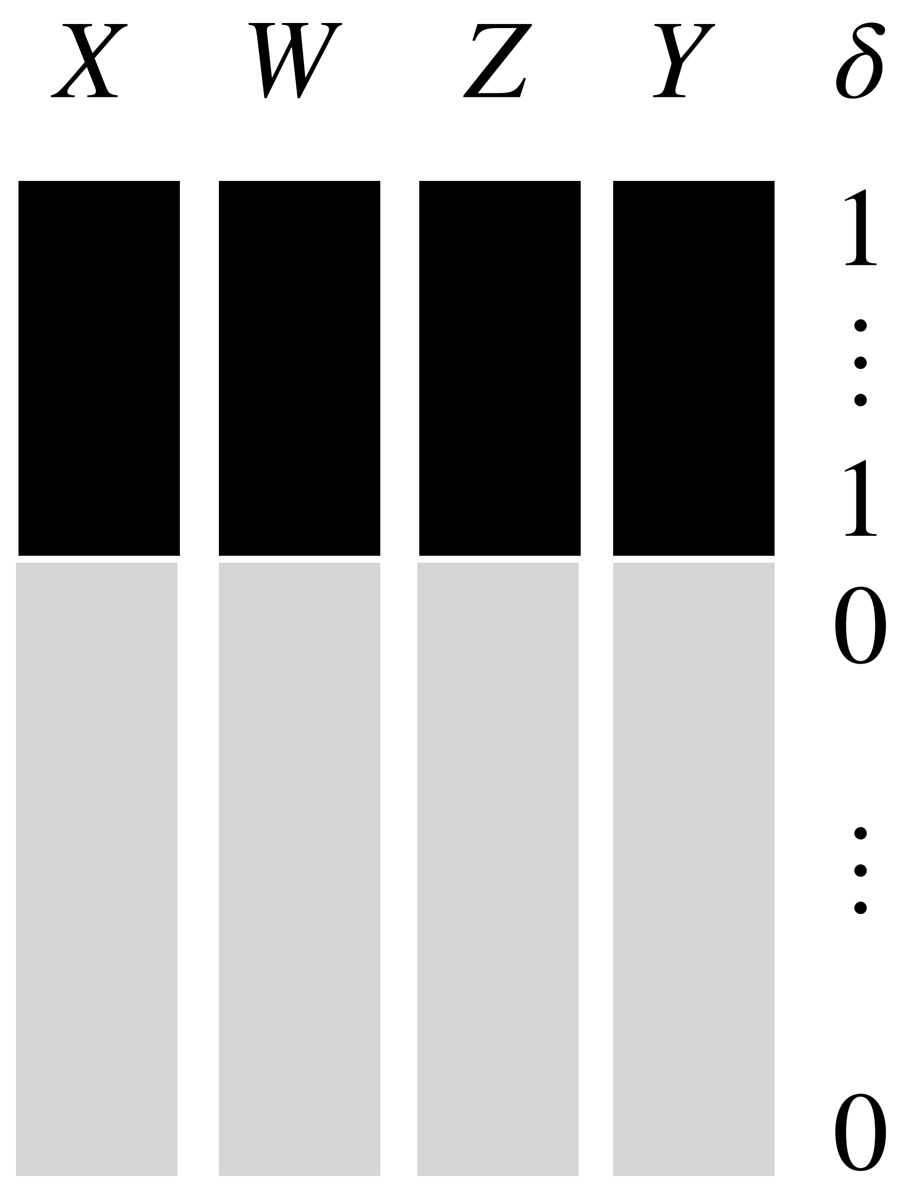
- Y : response variable
- X : covariate of interest
- Z : other covariates
- W : inverse of inclusion probability,

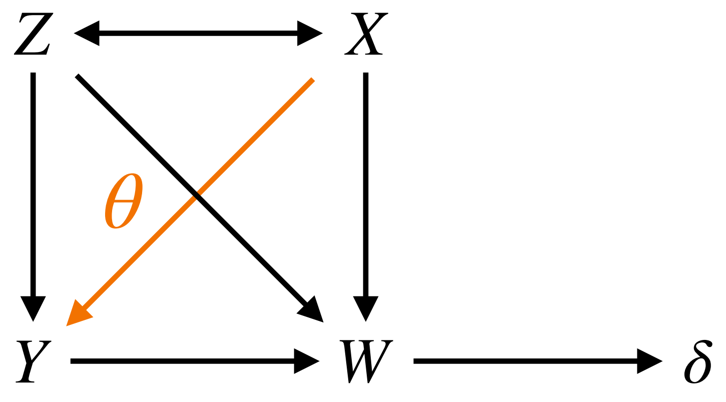
  i.e., $W := P(\delta = 1 \mid X, Y, Z, W)^{-1}$
- δ : sampling indicator
- $n := \sum_{i=1}^{N} \delta_i$: size of the sampled dataset

Setting 1



Setting 2



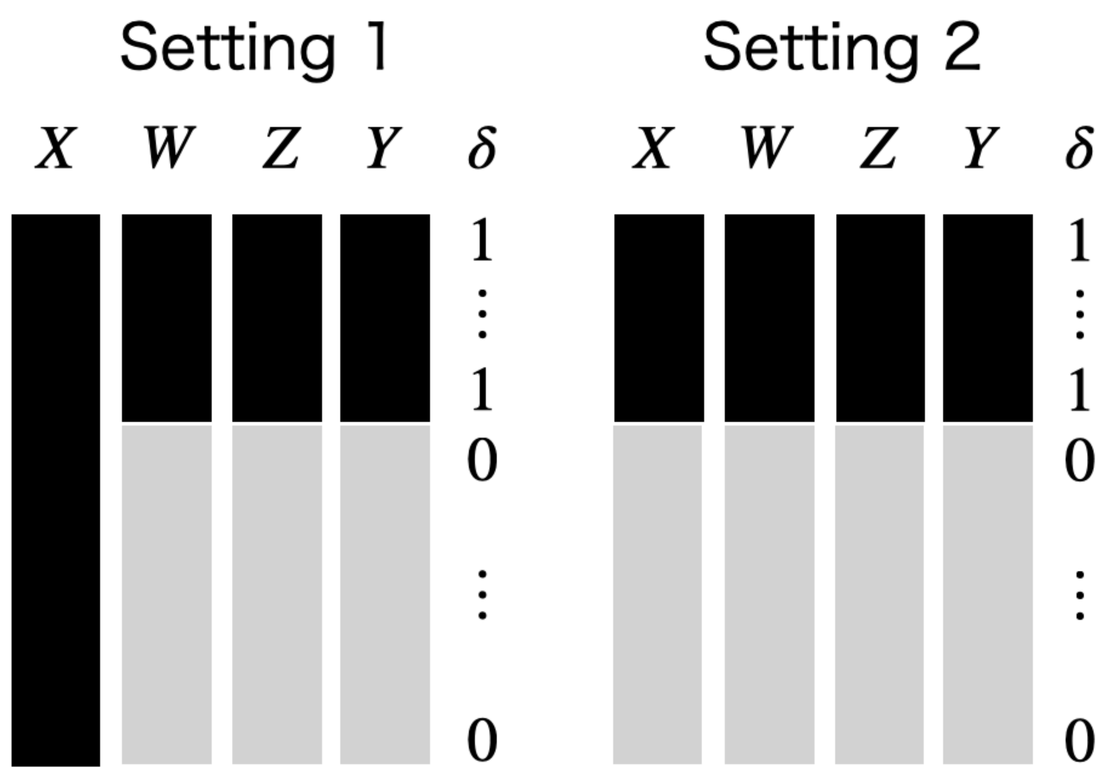


■ : available

░ : unavailable

Informative sampling: $W \not\perp Y \mid (X, Z)$

Efficient estimator in survey sampling



- Morikawa et al.(2022, arXiv) proposed the estimator that is solution to

$$\sum_{i=1}^{N}\left\{\delta_i W_i U(X_i, Y_i; \theta) + (1-\delta_i W_i) C(\theta; X_i)\right\} = 0,$$

where $C(\theta; X) = \dfrac{E\{(W-1)U(\theta) \mid X\}}{E(W-1 \mid X)}$ in Setting 1 and

$C(\theta; X) = C(\theta) = \dfrac{E\{(W-1)U(\theta)\}}{E(W-1)}$ in Setting 2.

- This estimator is a semiparametric efficient estimator, i.e., it attains the smallest asymptotic variance in the class of RAL estimators

Empirical Likelihood method

We will explain the Empirical Likelihood (EL) method in a simple case.

$\boldsymbol{y} = (y_1, y_2, \ldots, y_N)$: observed variable

$\boldsymbol{p} = (p_1, p_2, \ldots, p_N)$: event probability of $\boldsymbol{y}$, i.e. $\sum_{i=1}^{N} p_i = 1,\ 0 \leq p_i \leq 1$

$\prod_{i=1}^{N} p_i$: likelihood function

θ_0 : unique solution to $E\{U(Y;\theta)\} = 0$

Consider following maximization problem:

$$\arg\max_{\theta} \arg\max_{p_1,\ldots,p_n} \prod_{i=1}^{N} p_i$$

subject to $\sum_{i=1}^{N} p_i U(y_i;\theta) = 0,\ \sum_{i=1}^{N} p_i = 1.$

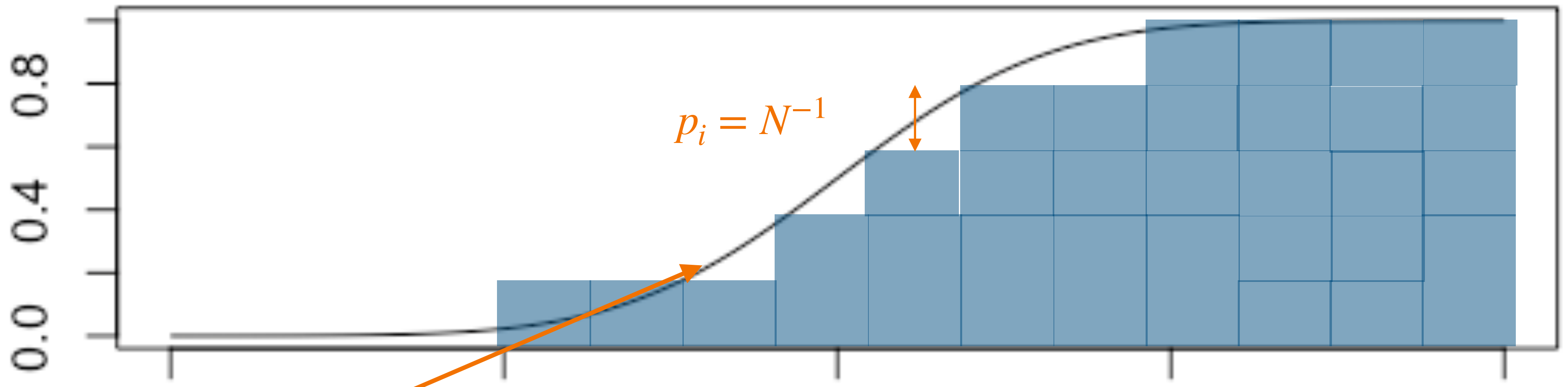
Impression of EL

Empirical distribution
based on EL weights

Constraints are $\sum_{i=1}^{N} p_i = 1$ only, i.e.,

$$\arg\max_{p_1,\ldots,p_n} \prod_{i=1}^{N} p_i \text{ subject to } \sum_{i=1}^{N} p_i = 1.$$

→ same weights $p_i = N^{-1}$



Equal width of steps

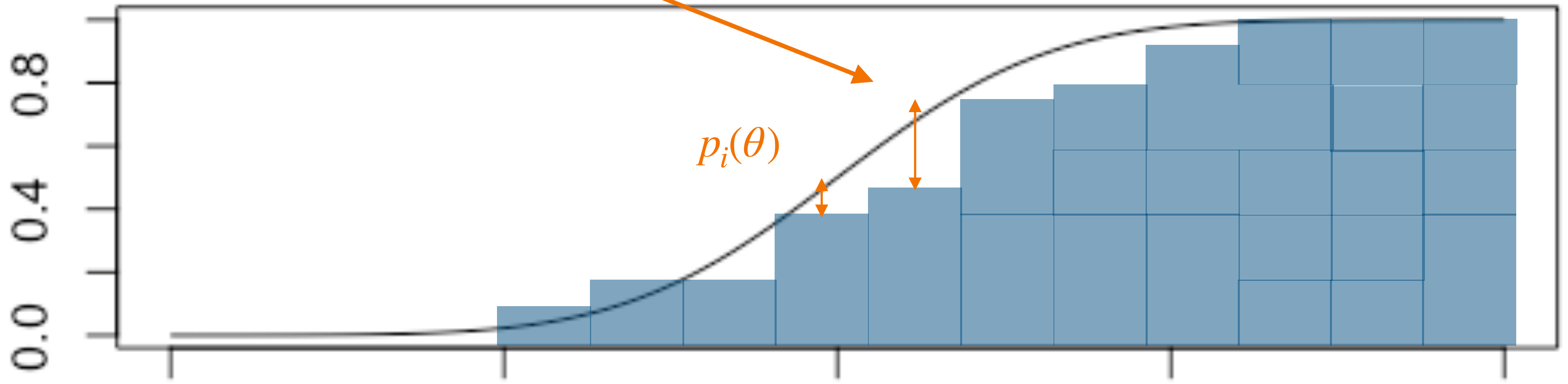
$$F_n(x) = \sum_{i=1}^{N} p_i I(y_i \le x) = N^{-1} \sum_{i=1}^{N} I(y_i \le x)$$

uneven step widths.

$$\arg\max_{\theta} \arg\max_{p_1,\ldots,p_n} \prod_{i=1}^{N} p_i$$

subject to $\sum_{i=1}^{N} p_i U(y_i;\theta) = 0,\ \sum_{i=1}^{N} p_i = 1.$

→ weights $p_i(\theta)$ depends on θ



$$F_n(x) = \sum_{i=1}^{N} p_i(\theta) I(y_i \le x)$$

How to calculate EL estimator

- By using the method of Lagrange multiplier, we maximize

$$H = \sum_{i=1}^{N} \log p_i - N\lambda \sum_{i=1}^{N} p_i U(y_i; \theta) - \gamma \left(\sum_{i=1}^{N} p_i - 1 \right)$$

with respect to $\boldsymbol{p}, \lambda, \gamma$.

- Simple calculation yields

$$\frac{\partial H}{\partial p_i} = \frac{1}{p_i} - N\lambda U(y; \theta) - \gamma$$

$$\sum_{i=1}^{N} p_i \frac{\partial H}{\partial p_i} = N - \gamma = 0 \iff \gamma = N$$

$$\Rightarrow \qquad p_i = \frac{1}{N} \frac{1}{1 + \lambda U(y_i; \theta)}$$

How to calculate EL estimator

- The likelihood function is

$$L = \sum_{i=1}^{N} \log p_i \propto - \sum_{i=1}^{N} \log\{1 + \hat{\lambda} U(y_i; \hat{\theta})\},$$

where $\hat{\theta}, \hat{\lambda}$ satisfies the following equation:

$$\frac{1}{N} \sum_{i=1}^{N} \frac{\hat{\lambda} \nabla_{\theta} U(y_i; \hat{\theta})}{1 + \hat{\lambda} U(y_i; \hat{\theta})} = 0,$$

$$\frac{1}{N} \sum_{i=1}^{N} \frac{U(y_i; \hat{\theta})}{1 + \hat{\lambda} U(y_i; \hat{\theta})} = 0.$$

- Under mild conditions, the maximum empirical likelihood estimator (MELE) $\hat{\theta}$ has consistency and asymptotic normality.
- See Qin and Lawless(1994) and Owen(1988) for more details.

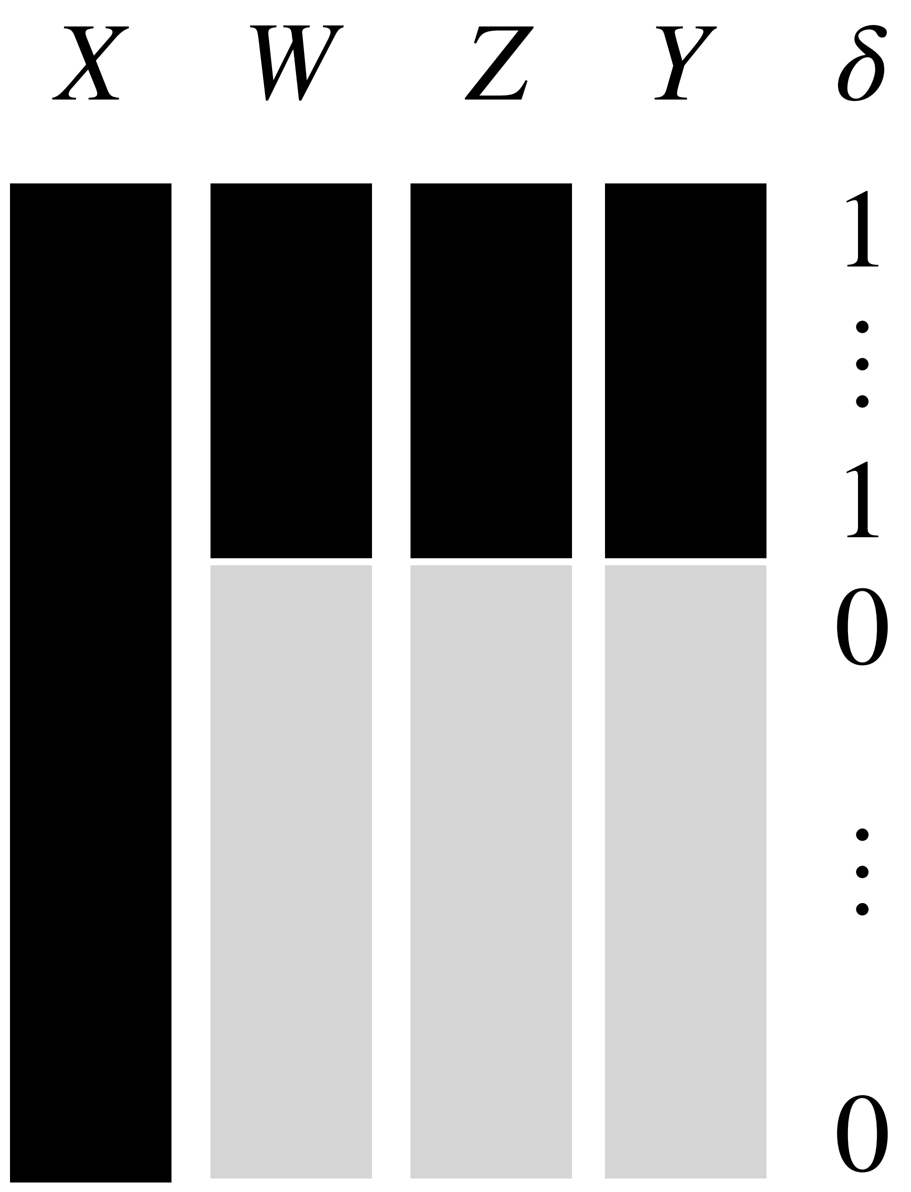
Setup of Setting 1

$(X_i, Y_i, Z_i, W_i, \delta_i, R_i)_{i=1,\ldots,N} \sim F$, i.i.d.

- Y : response variable
- X : covariate of interest
- Z : other covariates
- W : inverse of inclusion probability,

  i.e., $W := P(\delta = 1 \mid X, Y, Z, W)^{-1}$
- δ : sampling indicator

Setting 1



Semiparametric efficient estimator $\hat{\theta}$ is the solution to

$$\sum_{i=1}^{N} \left\{ \delta_i W_i U(X_i, Y_i; \theta) + (1 - \delta_i W_i) C(\theta; X_i) \right\} = 0,$$

where $$C(\theta; X) = \frac{E\{(W-1)U(\theta) \mid X\}}{E(W - 1 \mid X)}$$ (Morikawa et al., 2022)

Proposed EL estimator

- We propose the following EL estimator:

$$\hat{\theta}_{\mathrm{EL}} = \arg\max_{\theta} \arg\max_{p_1,\ldots,p_N} \sum_{i=1}^{N} \log p_i,$$

subject to

$$\sum_{i=1}^{N} p_i = 1, \quad \sum_{i=1}^{N} p_i \delta_i W_i U(X_i, Y_i; \theta) = 0, \quad \sum_{i=1}^{N} p_i (1 - \delta_i W_i) C(\theta; X_i) = 0,$$

Theorem 1 .

Under some regularity conditions,

- The proposed EL estimator has **consistency** and **asymptotic normality**.
- The asymptotic variance of $\hat{\theta}_{\mathrm{EL}}$ attains the **semiparametric efficient bound**.

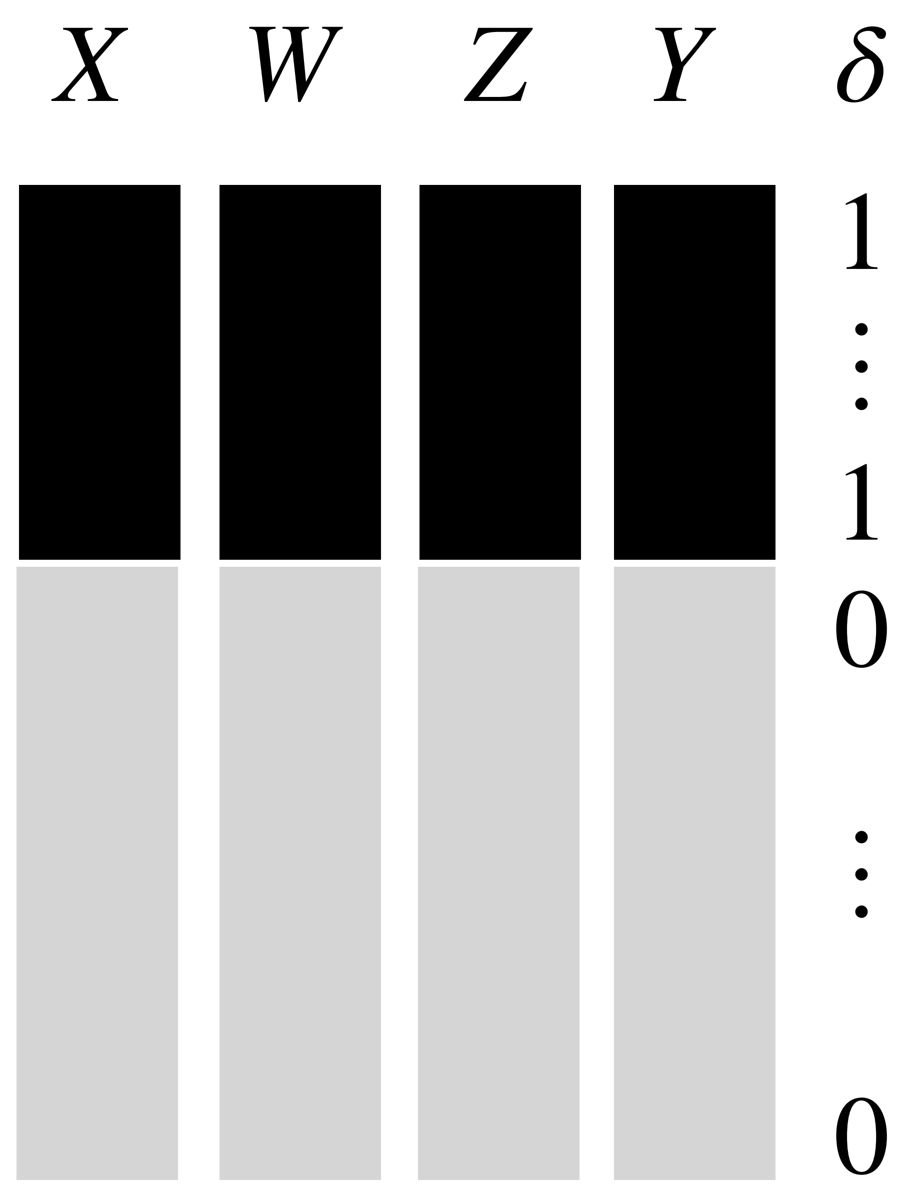
Setup of Setting 2

$(X_i, Y_i, Z_i, W_i, \delta_i, R_i)_{i=1,\ldots,N} \sim F$, i.i.d.

- Y : response variable
- X : covariate of interest
- Z : other covariates
- W : inverse of inclusion probability,

  i.e., $W := P(\delta = 1 \mid X, Y, Z, W)^{-1}$
- δ : sampling indicator

Setting 2



Semiparametric efficient estimator $\hat{\theta}$ is solution to

$$\sum_{i=1}^{N} \left\{ \delta_i W_i U(X_i, Y_i; \theta) + (1 - \delta_i W_i) C(\theta) \right\} = 0,$$

where $$C(\theta) = \frac{E\{(W-1)U(\theta)\}}{E(W-1)}$$ (Morikawa et al. 2022)

Likelihood function

- Likelihood function:

$$\prod_{i=1}^{N}\left\{P(\delta=1\mid X_i,Y_i,Z_i,W_i)f(X_i,Y_i,Z_i,W_i)\right\}^{\delta_i}\left\{\int\{1-P(\delta=1\mid x,y,z,w)\}f(x,y,z,w)dxdydzdw\right\}^{1-\delta_i}$$

$$\prod_{i=1}^{N}\left\{W_i^{-1}p_i\right\}^{\delta_i}\left\{1-\sum_{i=1}^{n}W_i^{-1}p_i\right\}^{1-\delta_i}$$

Use $W=P(\delta=1\mid X,Y,Z,W)^{-1}$ and replace $f(X_i,Y_i,Z_i,W_i)$ with p_i

- $\prod_{i=1}^{N}\left\{W_i^{-1}p_i\right\}^{\delta_i}\left\{1-\sum_{i=1}^{n}W_i^{-1}p_i\right\}^{1-\delta_i}\propto\prod_{i=1}^{N}\left\{p_i\right\}^{\delta_i}\left\{1-V\right\}^{1-\delta_i}$

where $V=\sum_{i=1}^{n}W_i^{-1}p_i\,.$

- $\log\prod_{i=1}^{N}\left\{p_i\right\}^{\delta_i}\left\{1-V\right\}^{1-\delta_i}=\sum_{i=1}^{n}\delta_i\log p_i+(N-n)\log(1-V),$

Setting 2

Proposed EL estimator

- We propose the following EL estimator:

$$(\hat{\theta}_{\mathrm{EL}}, \hat{V}) = \arg\max_{\theta, V} \arg\max_{p_1, \ldots, p_n} \sum_{i=1}^{n} \delta_i \log p_i + (N - n)\log(1 - V),$$

subject to

$$\sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i U(X_i, Y_i; \theta) = 0, \quad \sum_{i=1}^{n} p_i (W_i^{-1} - V) = 0$$

Theorem 2.

Under some regularity conditions,

- The proposed EL estimator has **consistency** and **asymptotic normality**.
- The asymptotic variance of $\hat{\theta}_{\mathrm{EL}}$ attains the **semiparametric efficient bound**.

EL method with missing value

Setup

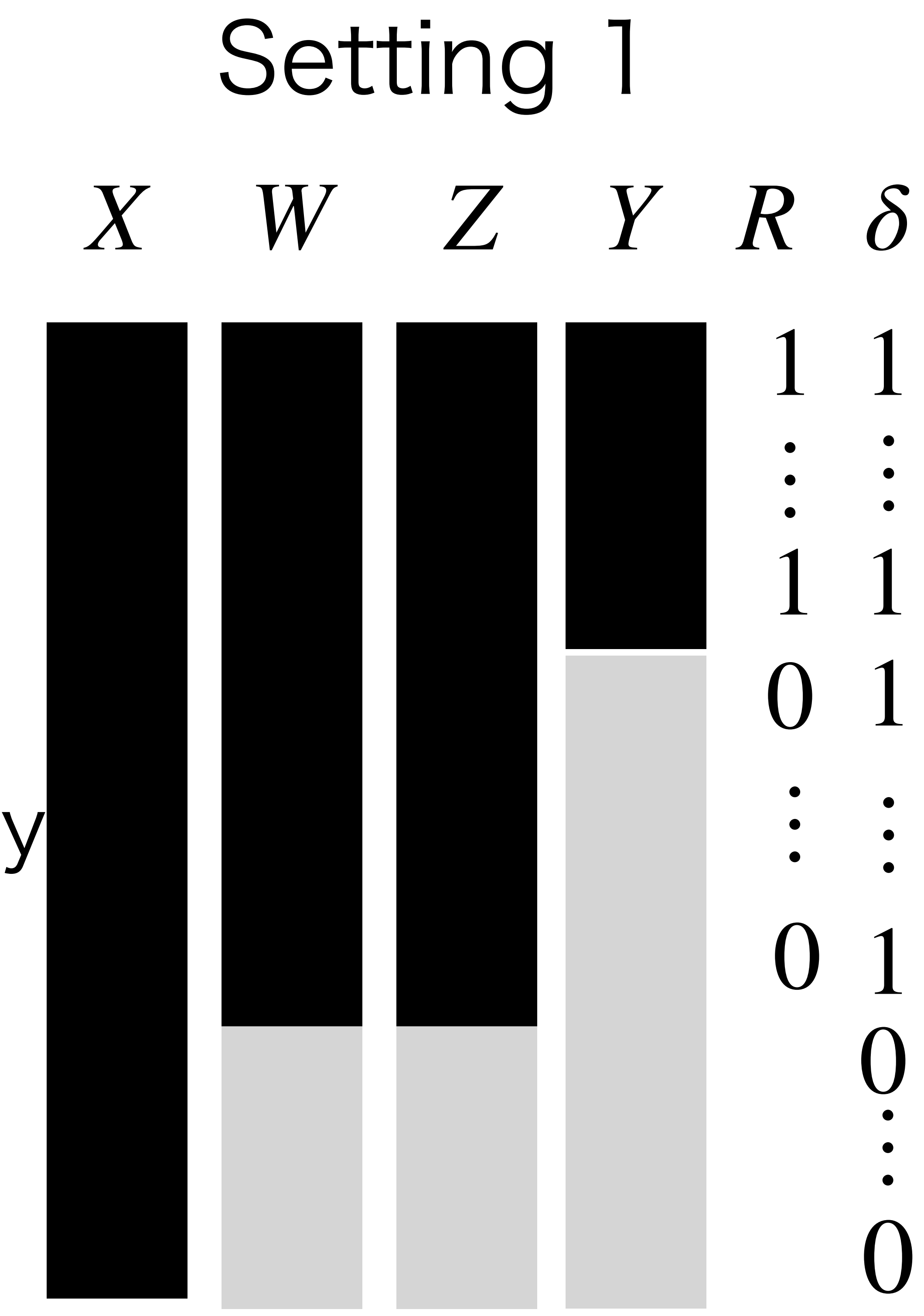
$(X_i, Y_i, Z_i, W_i, \delta_i, R_i)_{i=1,\dots,N} \sim F$, i.i.d.

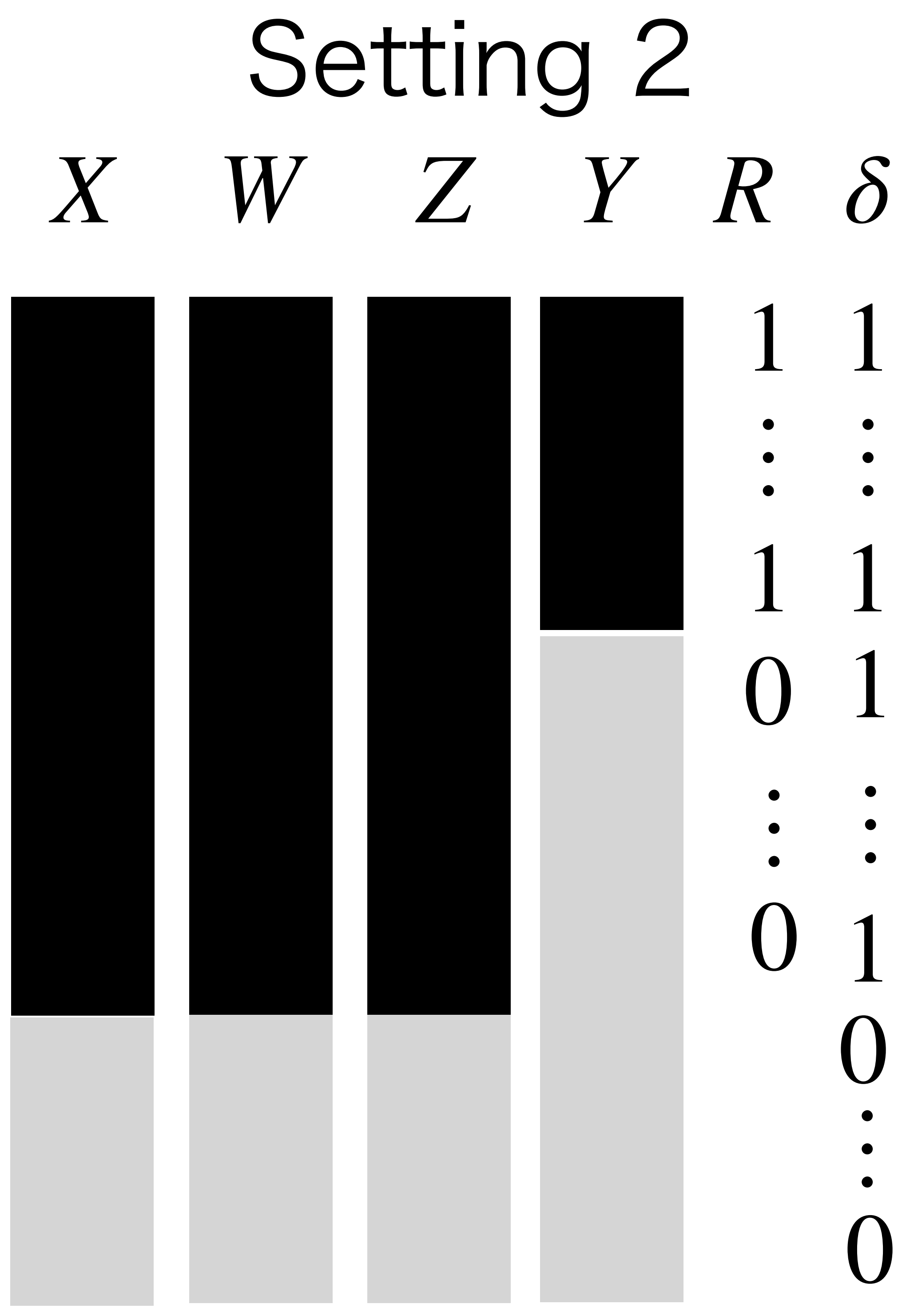
- Y : response variable
- X : covariate of interest
- Z : other covariates
- W : inverse of inclusion probability

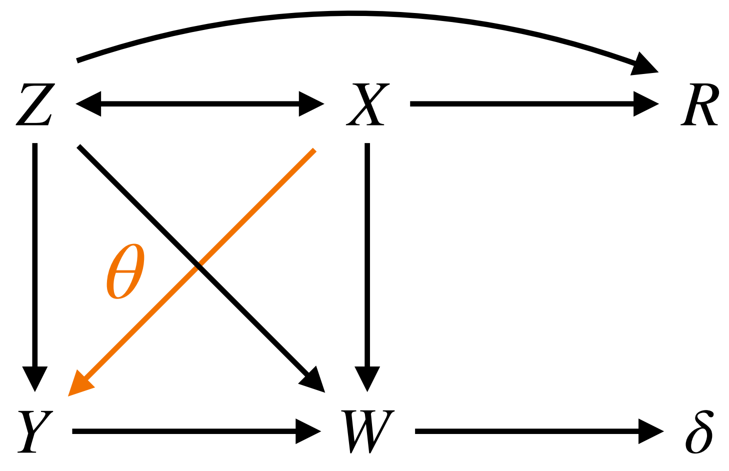
  i.e., $W := P(\delta = 1 \mid X, Y, Z, W)^{-1}$
- δ : sampling indicator
- $n := \sum_{i=1}^{N} \delta_i$: size of the sampled dataset
- R : response indicator of Y

Informative sampling: $W \perp \cancel{Y} \mid (X, Z)$

MAR missingness: $P(R = 1 \mid \delta = 1, x, y, z, w) = P(R = 1 \mid \delta = 1, x, z, w) =: \pi(x, z, w)$.







Efficient score in Setting 1,2

Theorem 3.

The efficient score function in Settings 1 and 2 are

$$S_{\theta,\,\mathrm{Eff}} = \underbrace{\delta W\left[\frac{R}{\pi}U(X,Y;\theta) + \left(1-\frac{R}{\pi}\right)E\left\{U(X,Y;\theta)\mid X,Z,W\right\}\right]}_{S_{\theta,\ \mathrm{DR}}} + \underbrace{(1-\delta W)C(X;\theta)}_{S_{\theta,\ \mathrm{Aug}}}.$$

where $\pi := \pi(x,z,w)$,

$C(X;\theta) = \dfrac{E\left[(W-1)U(X,Y;\theta)\mid X\right]}{E(W-1\mid X)}$ in Setting 1 and $C(X;\theta) = C(\theta) = \dfrac{E\left[(W-1)U(X,Y;\theta)\right]}{E(W-1)}$ in Setting 2.

- $S_{\theta,DR}$: double robust estimator (e.g. Kim and Haziza, 2014)
- $S_{\theta,\ \mathrm{Aug}}$; augmentation term

EL estimator with missing

- We derive the semiparametric efficient estimator via EL.
- One natural EL estimator based on efficient score is

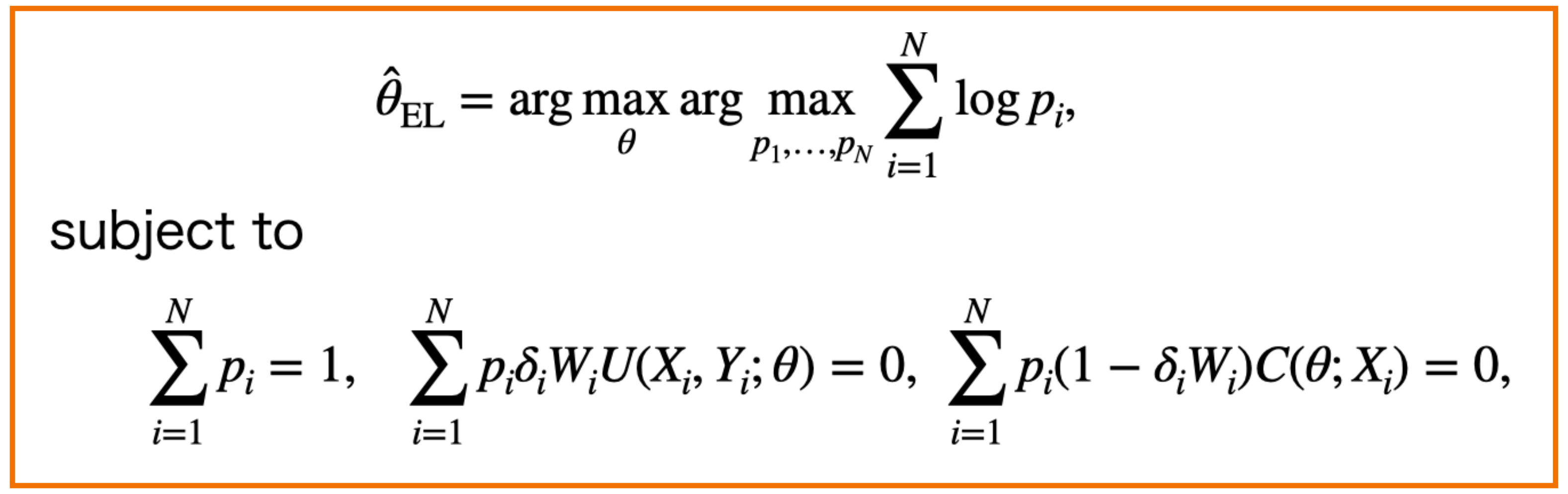
$$\hat{\theta}_{\mathrm{EL}} = \arg\max_{\theta} \arg\max_{p_1,\ldots,p_N} \sum_{i=1}^{N} \log p_i,$$

subject to

$$\sum_{i=1}^{N} p_i = 1, \quad \sum_{i=1}^{N} p_i (1 - \delta_i W_i) C(\theta; X_i) = 0,$$

$$\sum_{i=1}^{N} p_i \delta_i W_i \left\{ \frac{R_i U(X_i, Y_i; \theta)}{\hat{\pi}(X_i, Z_i, W_i)} + \left(1 - \frac{R_i}{\hat{\pi}(X_i, Z_i, W_i)} \right) \hat{m}(X_i, Z_i, W_i) \right\} = 0,$$

EL without missing



$$\hat{\theta}_{\mathrm{EL}} = \arg\max_{\theta} \arg\max_{p_1,\ldots,p_N} \sum_{i=1}^{N} \log p_i,$$

subject to

$$\sum_{i=1}^{N} p_i = 1, \quad \sum_{i=1}^{N} p_i \delta_i W_i U(X_i, Y_i; \theta) = 0, \quad \sum_{i=1}^{N} p_i (1 - \delta_i W_i) C(\theta; X_i) = 0,$$

Challenging of multiple robustness

- The constraint includes two working models $\hat{\pi}, \hat{m}$,

$$\sum_{i=1}^{N} p_i \delta_i W_i \left\{ \frac{R_i U(X_i, Y_i; \theta)}{\hat{\pi}(X_i, Z_i, W_i)} + \left(1 - \frac{R_i}{\hat{\pi}(X_i, Z_i, W_i)}\right) \hat{m}(X_i, Z_i, W_i) \right\} = 0,$$

<u>Reason for difficulty</u>

$\pi^{[j]}(x, z, w) \quad j = 1, \ldots, J$: candidate models for π

$m^{[k]}(x, z, w) \quad k = 1, \ldots, K$: candidate models for m

Need $J \times K$ constraints $\left\{ \sum_{i=1}^{N} p_i \delta_i W_i \left\{ \frac{R_i U}{\hat{\pi}^{[j]}} + \left(1 - \frac{R_i}{\hat{\pi}^{[j]}} \hat{m}^{[k]}\right) \right\} = 0; \quad j = 1, \ldots, J,\ k = 1, \ldots, K \right\}$

We want to reduce the number of constraints to $J + K$

Proposed EL estimator

The EL estimator is constructed in 2 steps.

First step (same in Settings 1 and 2)

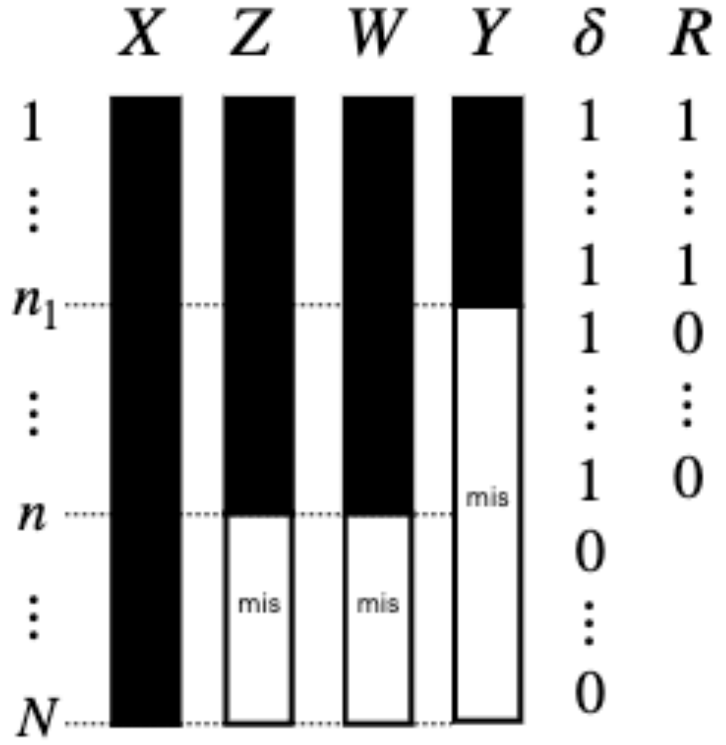
$$(\hat{p}_1^{(1)}, \dots, \hat{p}_{n_1}^{(1)}) = \arg\max_{p_i^{(1)}} \sum_{i=1}^{n_1} \log p_i^{(1)},$$

subject to $\sum_{i=1}^{n_1} p_i^{(1)} = 1,$

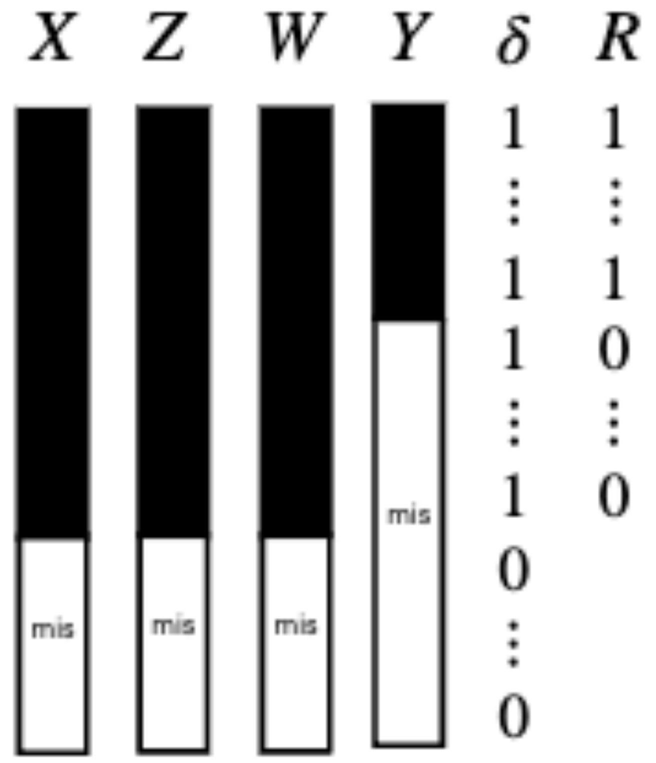
$$\underset{\text{constraints}}{J+K} \left\{ \begin{array}{l} \sum_{i=1}^{n_1} p_i^{(1)} \left\{ \hat{\pi}^{[j]}(X_i, Z_i, W_i) - \bar{\pi}_n^{[j]} \right\} = 0 \quad (j = 1, \dots, J), \\ \sum_{i=1}^{n_1} p_i^{(1)} \left\{ W_i \hat{m}^{[k]}(X_i, Z_i, W_i) - \bar{m}_n^{w[k]} \right\} = 0 \quad (k = 1, \dots, K), \end{array} \right.$$

where $\bar{\pi}_n^{[j]} = \frac{1}{n} \hat{\pi}^{[j]}(X_i, Z_i, W_i), \quad \bar{m}_n^{w[k]} = \frac{1}{n} W_i \hat{m}_n^{[k]}(X_i, Z_i, W_i).$

Setting 1

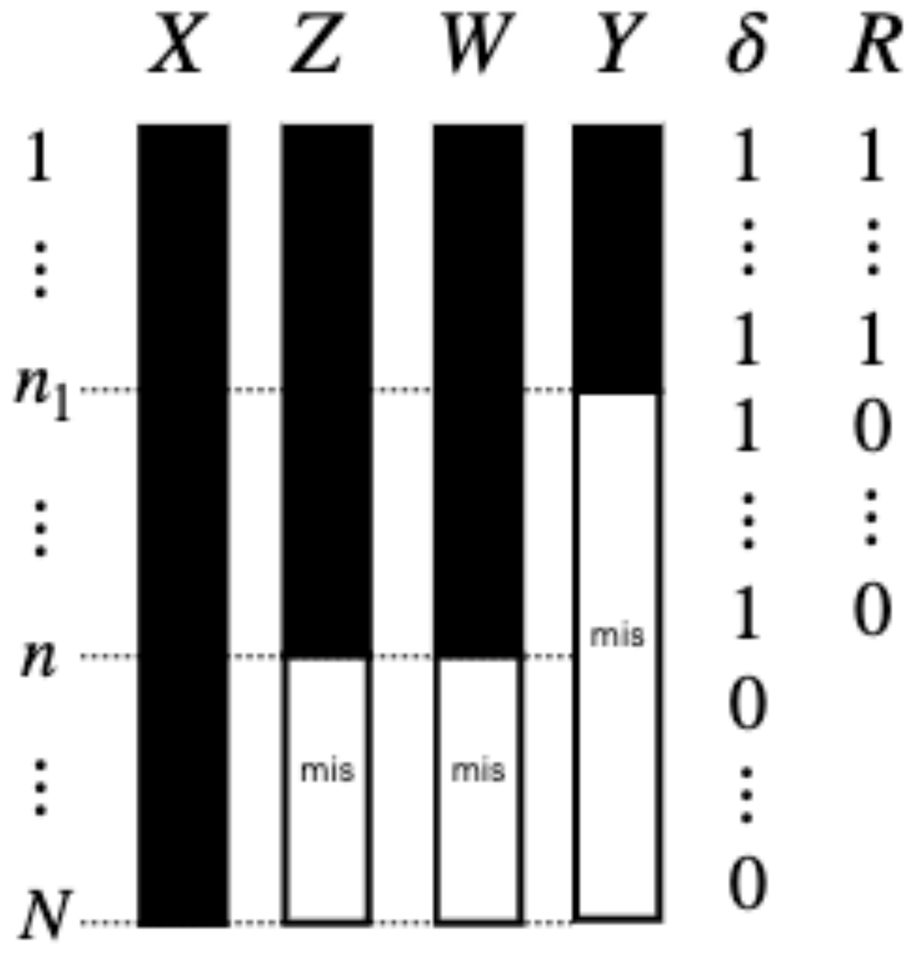


Setting 2



Proposed EL estimator in Setting 1

Second step in Setting 1



$$\hat{\theta}_{\mathrm{EL1}} = \arg\max_{\theta} \arg\max_{p_i^{(2)}} \sum_{i=1}^{N} \log p_i^{(2)},$$

subject to $\sum_{i=1}^{N} p_i^{(2)} = 1,$

$$\sum_{i=1}^{N} \hat{p}_i^{(1)} p_i^{(2)} \delta_i R_i W_i U(X_i, Y_i; \theta) = 0,$$

$$\sum_{i=1}^{N} p_i^{(2)} (1 - \delta_i W_i) \hat{C}_{\theta}^{[l]}(X_i) = 0 \quad (l = 1, \ldots, L),$$

Proposed EL estimator in Setting 2

Second step in Setting 2

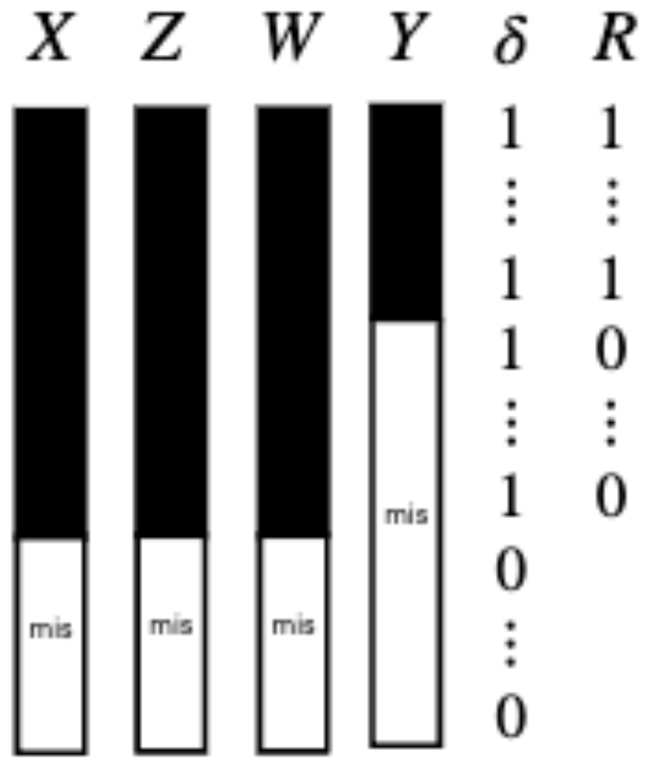
$$(\hat{\theta}_{\mathrm{EL2}}, \hat{V}) = \arg\max_{\theta, V} \arg \max_{p_1^{(2)}, \ldots, p_n^{(2)}} \sum_{i=1}^{n} \delta_i \log p_i + (N-n) \log(1-V),$$

subject to $\sum_{i=1}^{n} p_i^{(2)} = 1,$

$$\sum_{i=1}^{n} \hat{p}_i^{(1)} p_i^{(2)} \delta_i R_i U(X_i, Y_i; \theta) = 0,$$

$$\sum_{i=1}^{n} p_i^{(2)} (W_i^{-1} - V) = 0$$

Setting 2



Theoretical result of EL estimator

$\mathscr{P} = \{\pi^{[j]};\ j = 1,\ldots,J\}$: multiple models for π

$\mathscr{M} = \{m^{[k]};\ k = 1,\ldots,K\}$: multiple models for m

$\mathscr{C} = \{\hat{C}^{l}_{\theta};\ l = 1,\ldots,L\}$: multiple models for $C(X;\theta)$

Theorem 5.

Under some regularity conditions,

- (Multiple robustness) When $J+K$ models have at least one correct model in $\mathscr{P}$ and $\mathscr{M}$, the proposed estimators $\hat{\theta}_{\mathrm{EL1}}, \hat{\theta}_{\mathrm{EL2}}$ have consistency.
- (Efficiency) Moreover, $\hat{\theta}_{\mathrm{EL1}}, \hat{\theta}_{\mathrm{EL2}}$ are efficient if each $\mathscr{P}$, $\mathscr{M}$, $\mathscr{C}$ contain at least one correct model.

Remark

- When $W_i = 1$, the first step of the EL weights $p_i^{(1)}$ is the same as Han (2014, JASA).

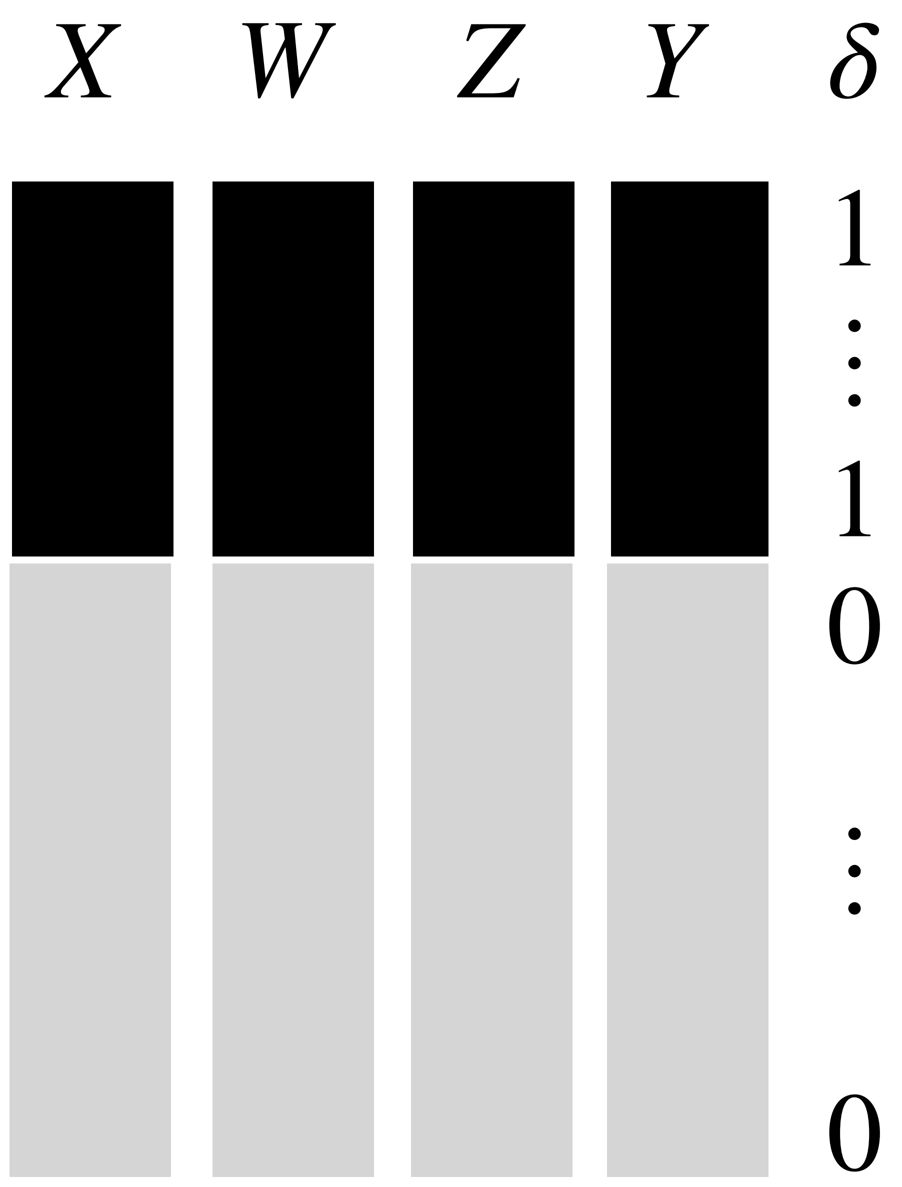
- Robustness against near-zero $\hat{\pi}(x, z, w)$ Proposed EL estimator does not use $1/\hat{\pi}(x, z, w)$, unlike the method of moments estimator
→ good finite-sample performance (Han, 2014, JASA).

EL method with external information

Setup

Available

- Individual data from an internal study
- Summary statistics from relevant external studies.



+

External data

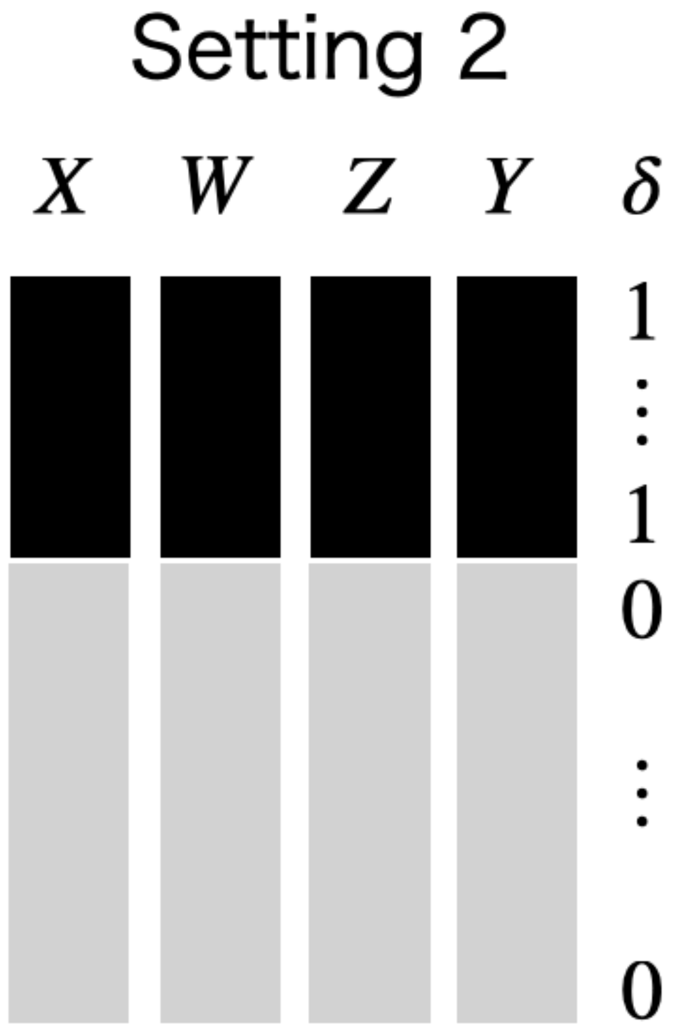
$\tilde{\beta}$: summary statistics of X

e.g. $\bar{X}$

How can we use the information of $\tilde{\beta}$?

Setup

- $\{(X_i, Z_i, W_i, Y_i, \delta_i)\}_{i=1}^{n}$: individual observations in Setting 2 from the internal distribution $P_0 \in \mathscr{P}_0$
- $\tilde{\beta}$: summary statistics based on individual observations from the external distribution $P_1 \in \mathscr{P}_1$
- $\tilde{\beta}$ is an estimator of some functional $\beta(P_1)$ that is the unique solution to $E_{P_1}\{\psi(X, Y; \beta)\} = 0$



Mathematical Assumption

- Hu et al. (2023, arXiv) assumed the following conditions:

Assumption

- The summary statistics $\tilde{\beta}$ is a RAL estimator of $\beta(P_1)$.
- Asymptotic normality $m^{1/2}\{\tilde{\beta} - \beta(P_1)\} \to N(0, \Sigma_1)$.
- A consistent estimator $\hat{\Sigma}_1$ and the sample size of external data m are available.
- Sample size ratios converge $m/N \to \rho \in (0, \infty)$.
- $\beta(P_0) = \beta(P_1)$

Proposed EL estimator

- Motivated by Zhang et al.(2020, Biometrika), we propose the following EL estimator:

$$(\hat{\theta}_{\mathrm{FEL}}, \hat{V}) = \arg\max_{\theta, V} \arg\max_{p_1,\ldots,p_n} \sum_{i=1}^{n} \delta_i \log p_i + (N-n)\log(1-V) \quad -m(\tilde{\beta}-\beta)^{\top}\hat{\Sigma}_1^{-1}(\tilde{\beta}-\beta)/2$$

subject to

$$\sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i U(X_i, Y_i; \theta) = 0, \quad \sum_{i=1}^{n} p_i (W_i^{-1} - V) = 0$$

$$\sum_{i=1}^{n} p_i \psi(X_i, Y_i; \beta) = 0$$

Proposed EL estimator

Theorem 6.

Under some regularity conditions,

- The proposed EL estimator $\hat{\theta}_{\text{FEL}}$ has consistency and asymptotic normality.
- The asymptotic variance is

$$E(\phi_{\text{eff}}^{\otimes 2}) + A\{\Sigma_1/\rho - E(\eta_{\text{eff}}^{\otimes 2})\}A^\top,$$

where $A = E(\phi_{\text{eff}}\eta_{\text{eff}}^\top)E(\eta_{\text{eff}}^{\otimes 2})^{-1}$,

ϕ_{eff} : efficient influence function only using internal data, and

η_{eff} : efficient influence function for β based on internal individual data.

Remark

The asymptotic variance of the proposed estimator is

$$\boxed{E(\phi_{\text{eff}}^{\otimes 2})} - A\{\Sigma_1/\rho + E(\eta_{\text{eff}}^{\otimes 2})\}^{-1}A^\top.$$

Efficiency bound only
using internal data

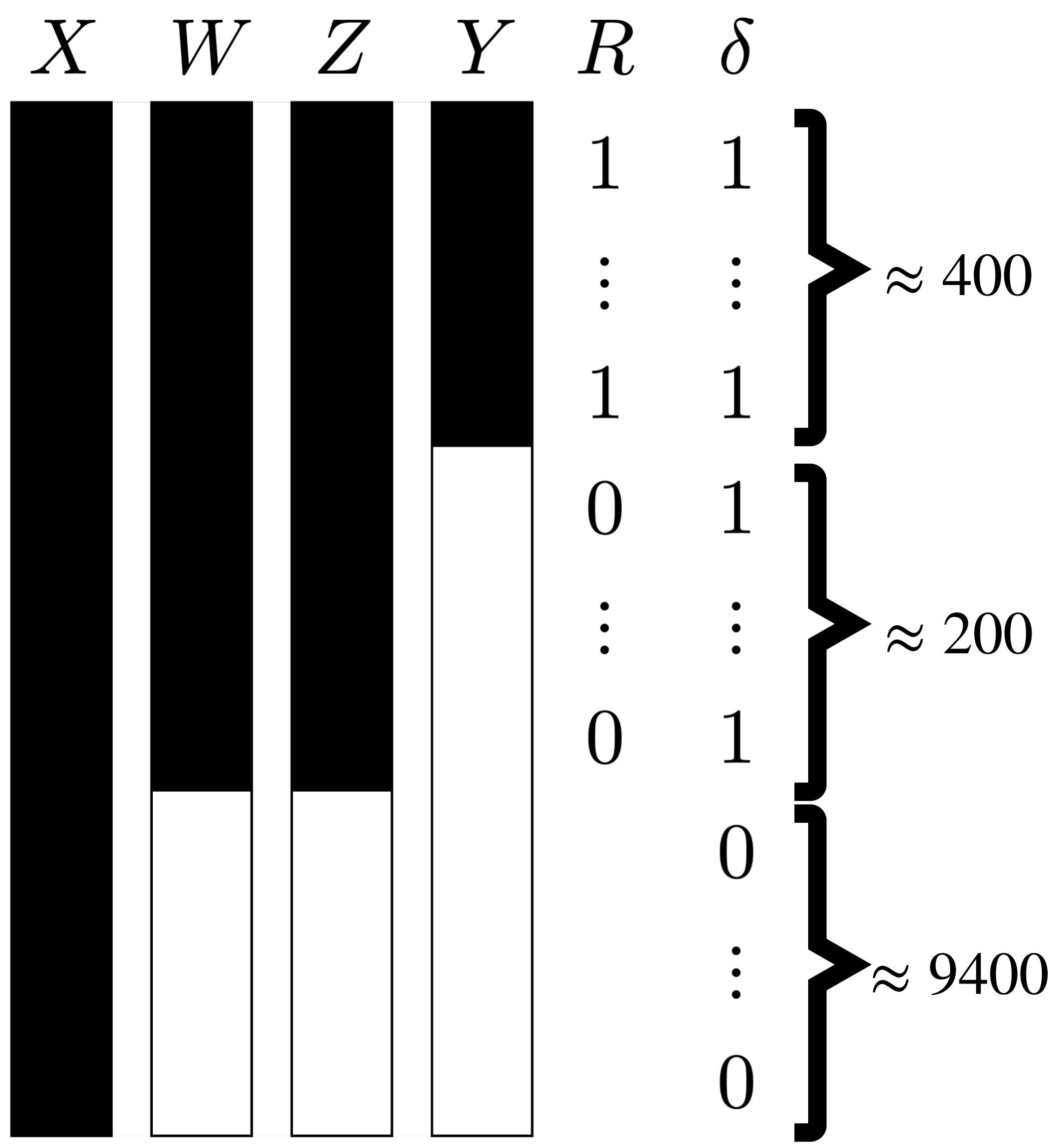
- When $\rho \to \infty$, i.e., the external sample size is larger than that of the internal data; thus $\hat{\theta}_{\text{FEL}}$ is more efficient than the estimators using only internal data.
- $\rho \to 0$ → Negligible impact of data consolidation.
- See Hu et al. (2023, arXiv) for more details.

Numerical experiment

Setup

- $X \sim N\left(0, \frac{1}{2}\right), Z \sim N\left(0, \frac{1}{2}\right), Y \mid x, z \sim N\left(x - z, \frac{1}{2}\right).$
- $\log(W - 1) \mid x, y, z \sim N\left(2.95 - 0.25x - 0.45y - 0.1z, \sqrt{0.05}^{2}\right)$
- $\delta \mid w \sim \text{Binom}(w^{-1})$
- $P(R = 1 \mid \delta = 1, x, z, w) = \dfrac{\exp(-0.3 + 0.75x - 0.5z + 0.05w)}{1 + \exp(-0.3 + 0.75x - 0.5z + 0.05w)}$

- $N = 10000$, iteration $= 2000$.
- Target: $\theta = E(Y)$



Setup

We compared Horvitz-Thompson (HT) and Kim and Haziza (KH) estimators with our proposed multiply-robust empirical likelihood estimators.

HT :

$$\sum_{i=1}^{N} \frac{\delta_i R_i W_i}{\hat{\pi}_i}(Y_i - \hat{\theta}_{\mathrm{HT}}) = 0 \Leftrightarrow \hat{\theta}_{\mathrm{HT}} := \frac{\sum_{i=1}^{N} \delta_i R_i W_i Y_i / \hat{\pi}_i}{\sum_{i=1}^{N} \delta_i R_i W_i / \hat{\pi}_i}.$$

KH :

$$\hat{\theta}_{\mathrm{KH}} := \frac{1}{N}\sum_{i=1}^{N} \delta_i W_i \left\{ \frac{R_i Y_i}{\hat{\pi}_i} + \left(1 - \frac{R_i}{\hat{\pi}_i}\right) \hat{E}\left(Y \mid X_i, Z_i, W_i\right) \right\}.$$

Setup

We prepared three models for each π, m :

$$\text{(Correct)}\ \pi^{(1)}(x, z, w; \phi^{(1)}) = \operatorname{expit}(\phi_0^{(1)} + \phi_1^{(1)}x + \phi_2^{(1)}z + \phi_3^{(1)}w),$$

$$\pi^{(2)}(x, z, w; \phi^{(2)}) = \operatorname{expit}(\phi_0^{(2)} + \phi_1^{(2)}x + \phi_2^{(2)}z + \phi_3^{(2)}xz),$$

$$\pi^{(3)}(x, z, w; \phi^{(3)}) = \operatorname{expit}\{g_1(x; \phi^{(3)}) + g_2(z; \phi^{(3)})\},$$

$$\text{(Correct)}\ m^{(1)}(x, z, w; \xi^{(1)}) = \xi_0^{(1)} + \xi_1^{(1)}x + \xi_2^{(1)}z + \xi_3^{(1)}\log(w - 1),$$

$$m^{(2)}(x, z, w; \xi^{(2)}) = \xi_0^{(2)} + \xi_1^{(2)}x + \xi_2^{(2)}z + \xi_3^{(2)}xz,$$

$$m^{(3)}(x, z, w; \xi^{(3)}) = \tilde{g}_1(x; \phi^{(3)}) + \tilde{g}_2(z; \phi^{(3)}),$$

- $g_j, \tilde{g}_j$: smoothing splines with three degrees of freedom.
- $\pi^{(3)}$ and $m^{(3)}$ are misspecified, but it is expected that show better performance b
own more complex models.

Setup

We prepared four models for our multiply-robust estimators:

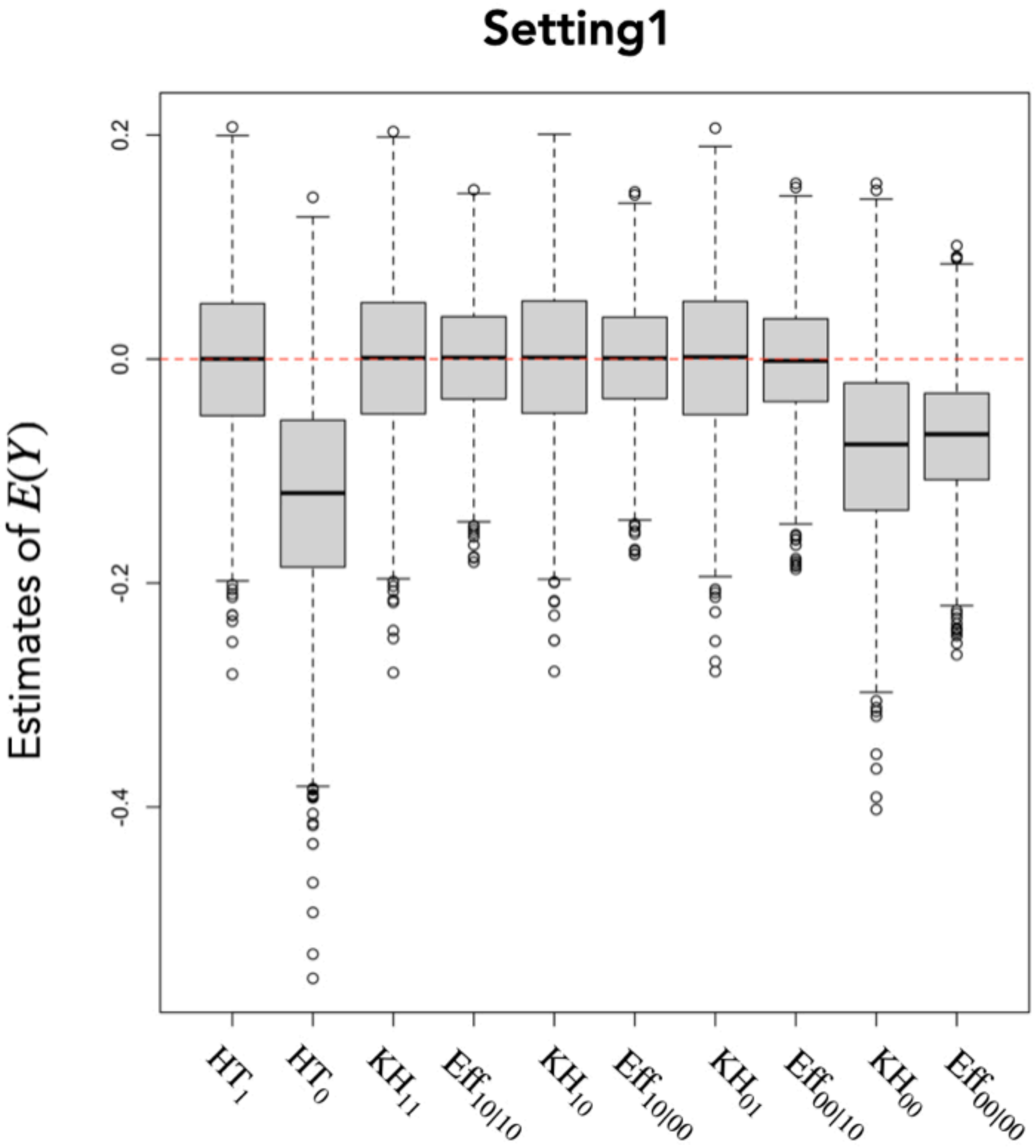
$\mathrm{Eff}_{10|10}$: combination of $\pi^{(1)}, \pi^{(2)}, m^{(1)}, m^{(2)}$;

$\mathrm{Eff}_{10|00}$: combination of $\pi^{(1)}, \pi^{(2)}, m^{(2)}, m^{(3)}$;

$\mathrm{Eff}_{00|10}$: combination of $\pi^{(2)}, \pi^{(3)}, m^{(1)}, m^{(2)}$;

$\mathrm{Eff}_{00|00}$: combination of $\pi^{(2)}, \pi^{(3)}, m^{(2)}, m^{(3)}$.

Result



HT_i , KH_{ij} , $\mathrm{Eff}_{ij|kl}$:

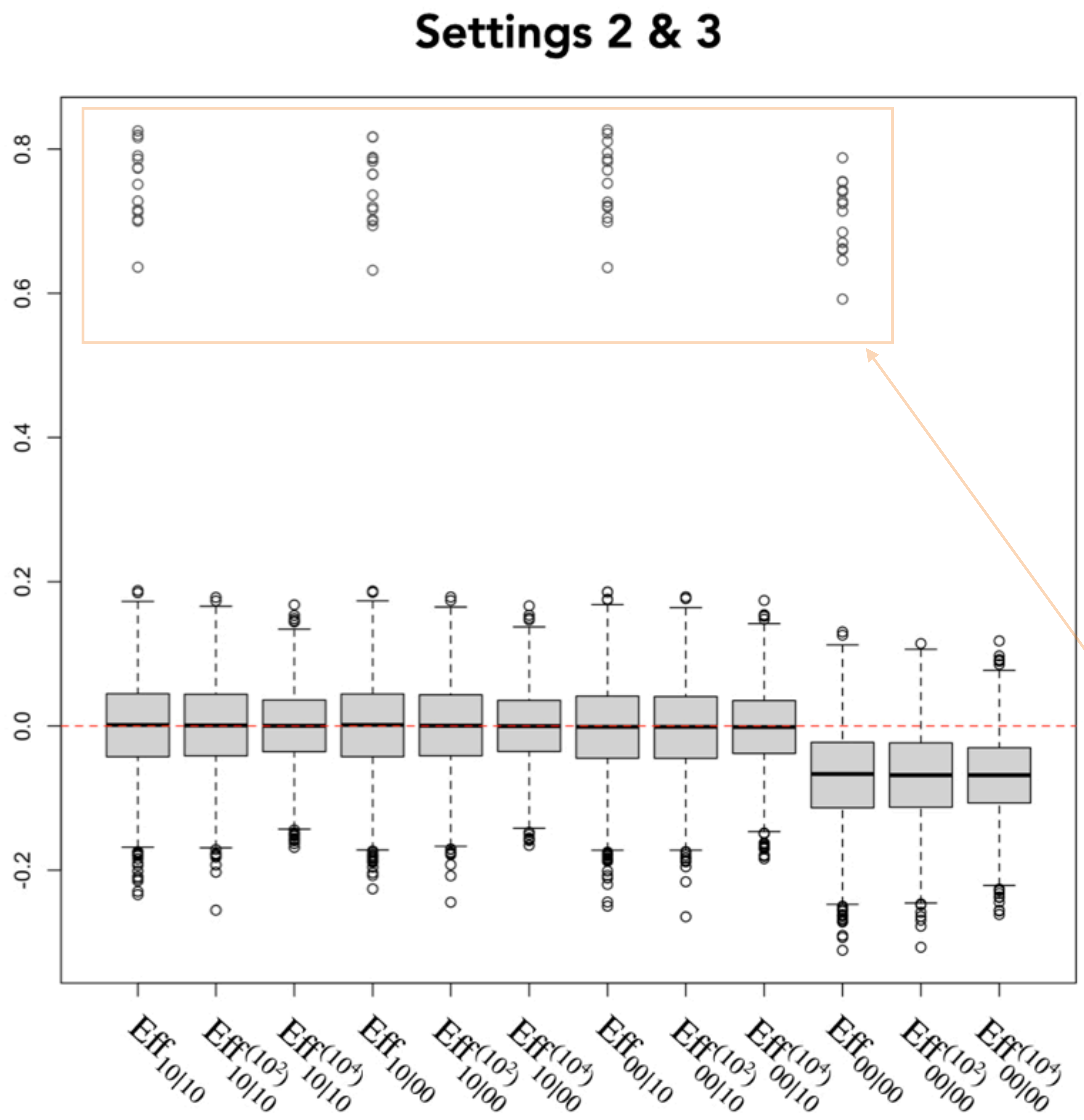
$i, j, k, l = 1$ → correct working model

- KH: double robustness
  but less efficient
- $\mathrm{Eff}_{10|10}$: semiparametric efficient

Proposed estimators perform well even if all the working models are misspecified

∵ better representation of the generalized additive models.

Result



Sample mean of $E(X)$, i.e. $M^{-1}\sum_{i=1}^{M} X_i$ is available from another external data source. $(M = 10^2 \text{ or } 10^4)$

$\mathrm{Eff}_{ij|kl}$ in Setting 2,
$\mathrm{Eff}^{(M)}_{ij|kl}$ in Setting 3.

Some iteration suffer from weak identification

References

- Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y. A., Klaassen, J., Wellner, J. A., & Ritov, Y. A. (1993). Efficient and adaptive estimation for semiparametric models. Springer.
- Han, P. (2014). Multiply robust estimation in regression analysis with missing data. *Journal of the American Statistical Association*, **109**(507), 1159-1173.
- Hu, W., Wang, R., Li, W., & Miao, W. (2023). Paradoxes and resolutions for semiparametric fusion of individual and summary data. *arXiv preprint arXiv:2210.00200*.
- Kim, J. K., & Haziza, D. (2014). Doubly robust inference with missing data in survey sampling. *Statistica Sinica*, **24**(1), 375-394.
- Morikawa, K., Beppu, K., and Aida, W. (2023). Efficient Multiply Robust Estimation Under Informative Sampling. *arXiv:2311.06719*.

References

- Morikawa, K., Terada, Y., & Kim, J. K. (2022). Semiparametric adaptive estimation under informative sampling. *arXiv preprint arXiv:2208.06039*.
- Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**(2), 237-249.
- Qin, J., & Lawless, J. (1994). Empirical likelihood and general estimating equations. *Annals of Statistics*, **22**(1), 300-325.
- Tsiatis, A. A. (2006). Semiparametric theory and missing data. Springer.
- Zhang, H., Deng, L., Schiffman, M., Qin, J., & Yu, K. (2020). Generalized integration model for improved statistical inference by leveraging external summary data. *Biometrika*, **107**(3), 689-703.