NEURAL INFORMATION PROCESSING SYSTEMS

Double Machine Learning Density Estimation for Local Treatment Effects with Instruments

Yonghan Jung

PURDUE UNIVERSITY

Jin Tian

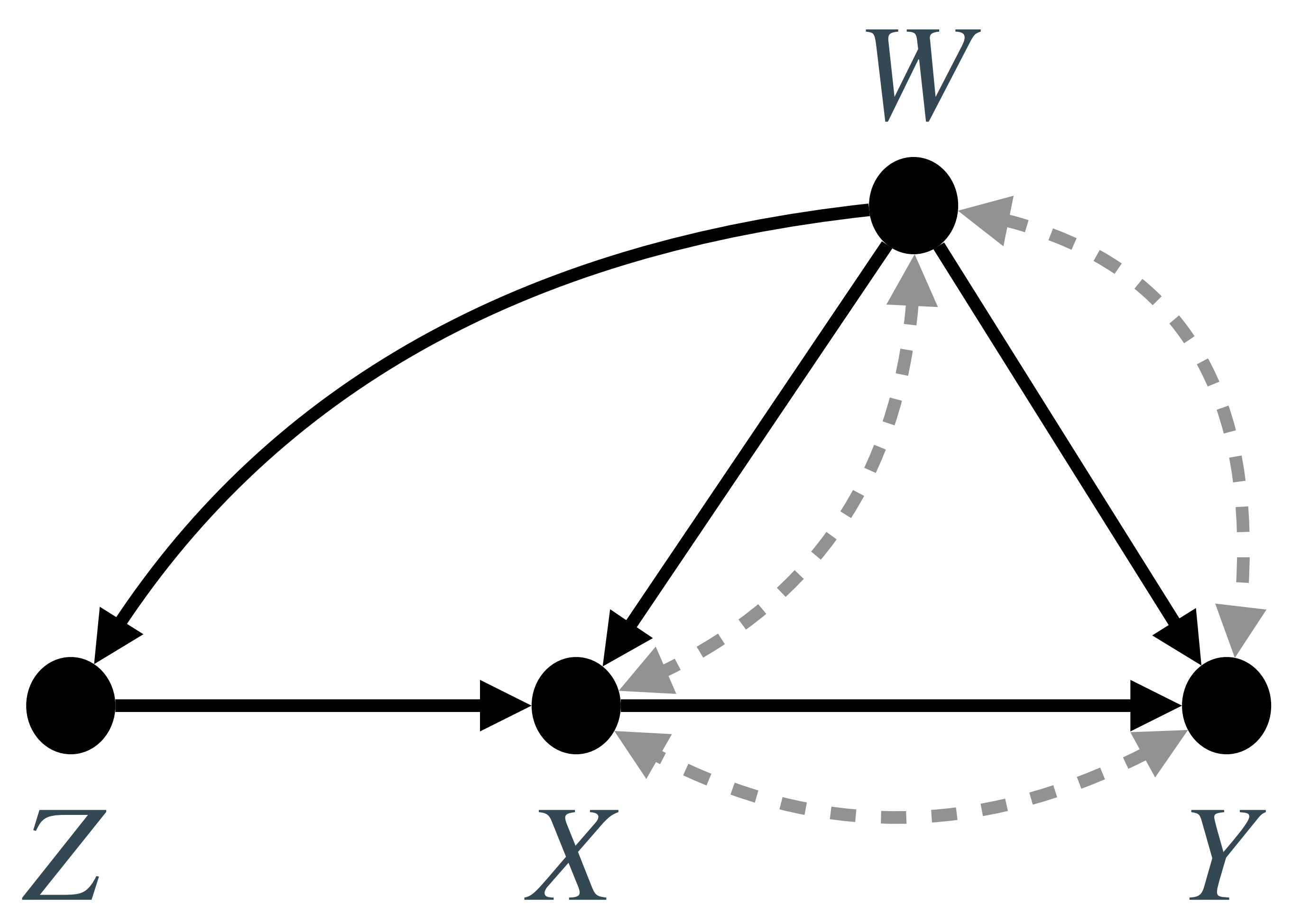
IOWA STATE UNIVERSITY

Elias Bareinboim

COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

Example: 401(k) on the financial asset

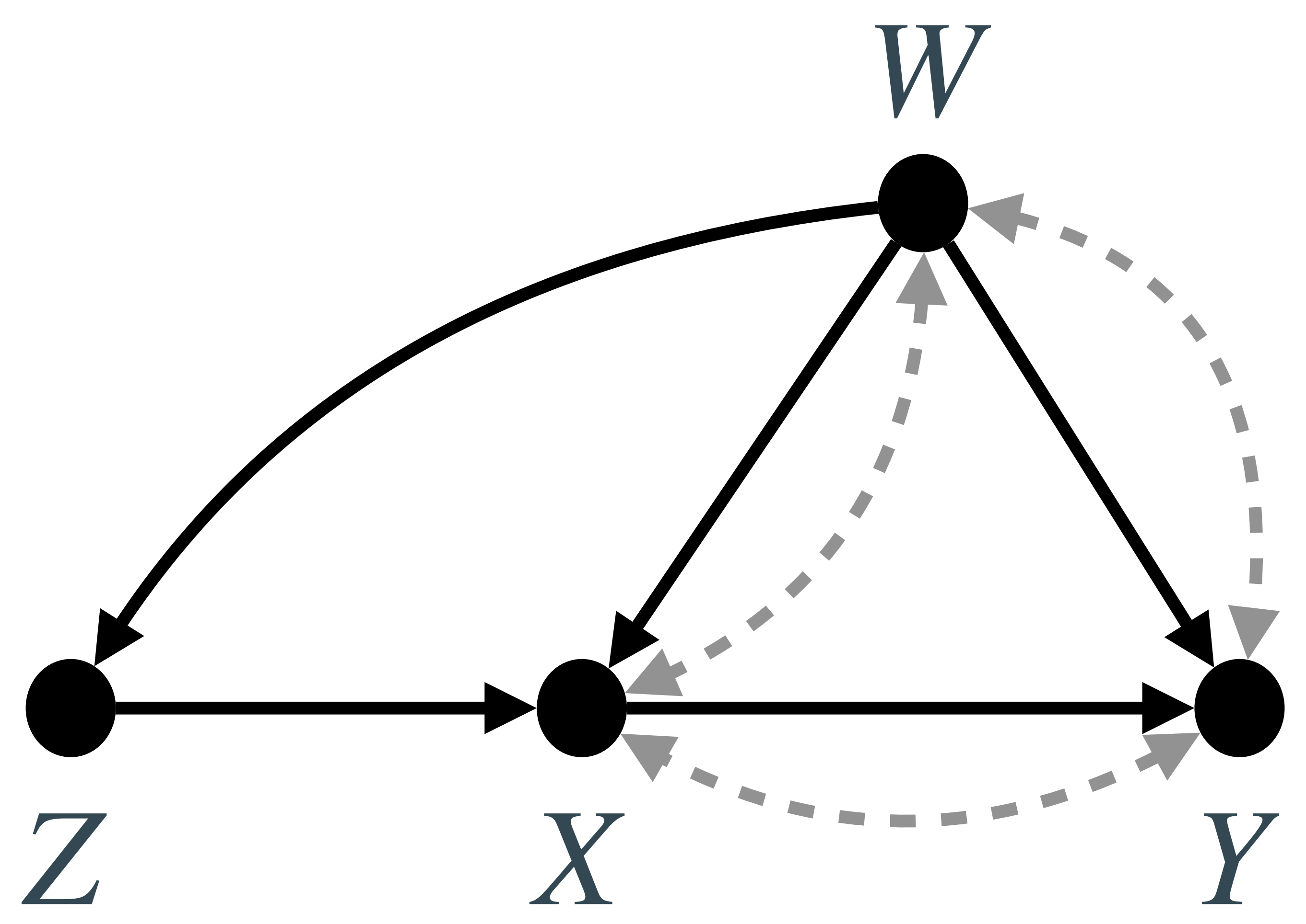
Query: Effect of the participation of 401(k) to the net financial asset



X	401(k) participation	$X = 1$ for participation
Z	401(k) eligibility	$Z = 1$ for being eligible
Y	Net financial asset	
W	Observed covariates (e.g., income, gender, family size, etc.)	
$\leftrightarrow$	Presence of unmeasured confounders	

Rewritten query: A causal effect of $X = x$ to $Y = y$; i.e., $P(y \mid do(x))$.

Unidentifiability of Causal Effects

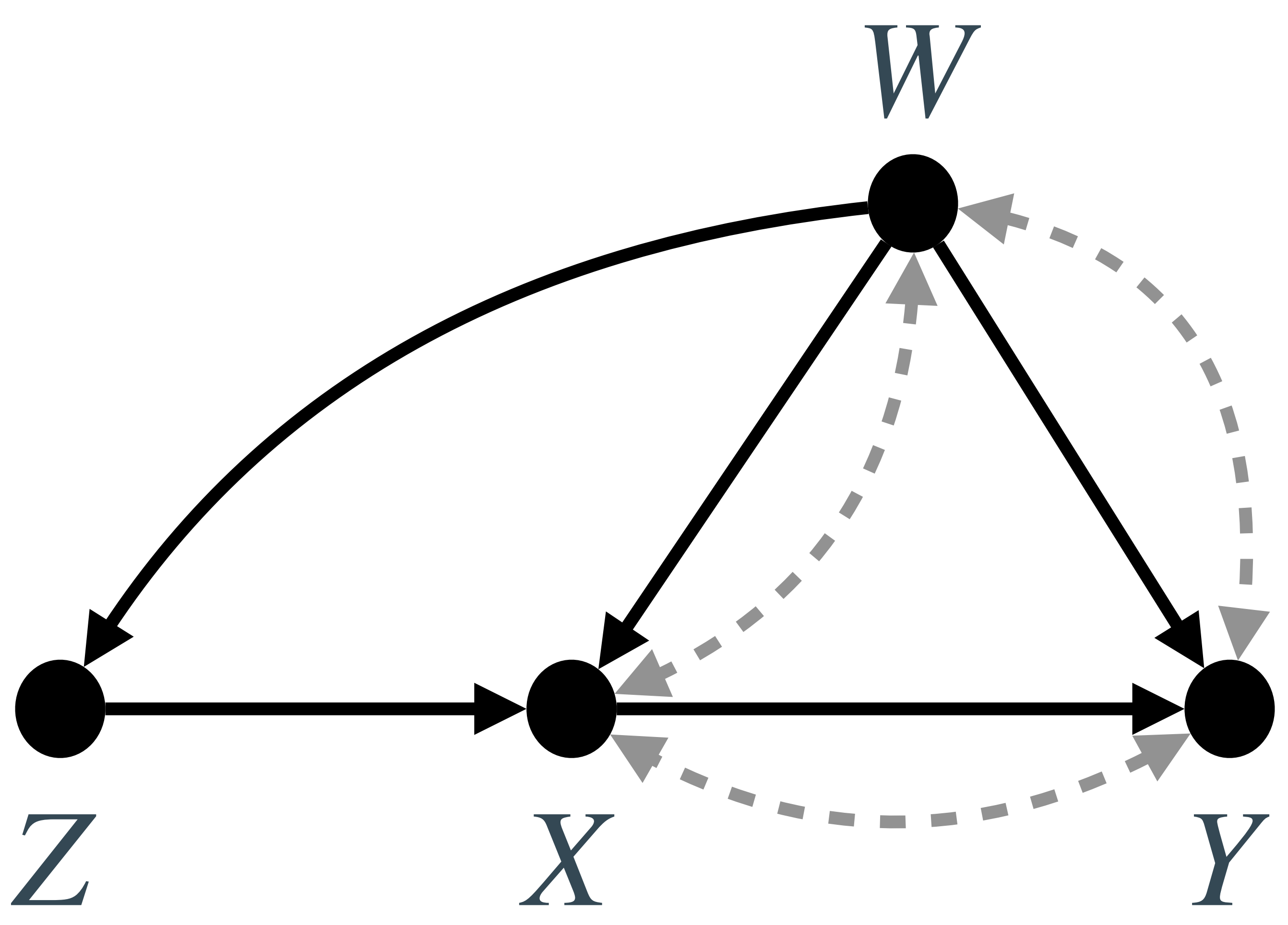


- $P(y \mid do(x))$ is unidentifiable

  (i.e., no unique representation of $P(y \mid do(x))$ w.r.t. the observational distribution P exists)

- Therefore, $P(y \mid do(x))$ is not computable from the observational data.

Exclusion of Defiers



- If a sample is eligible, then the sample either participates or not.

- No eligibility ⇒ No participation; i.e., $Z = 0 \nRightarrow X = 1$.

- This excludes a portion of samples called ‘*defiers*’.

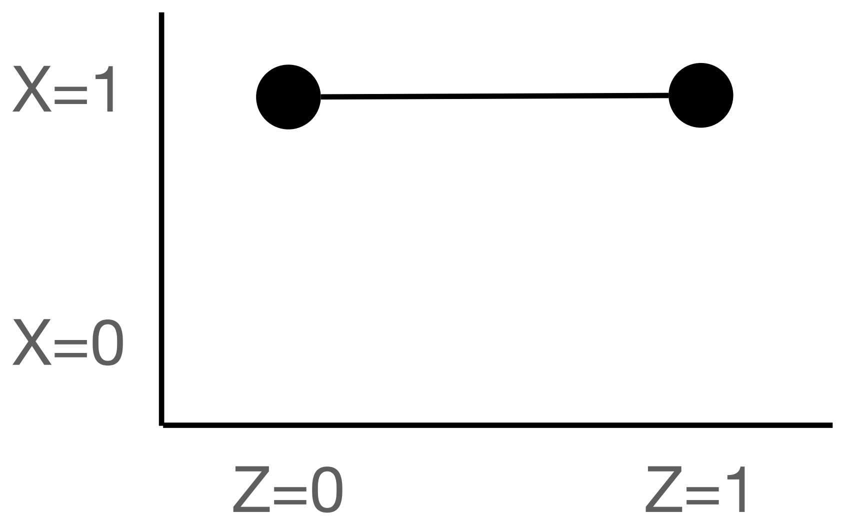
Four types of samples

Samples behaviors on $\{Z, X\}$ are always falling into one of the four types:

Always-taker

$$Z = 0 \Rightarrow X = 1$$

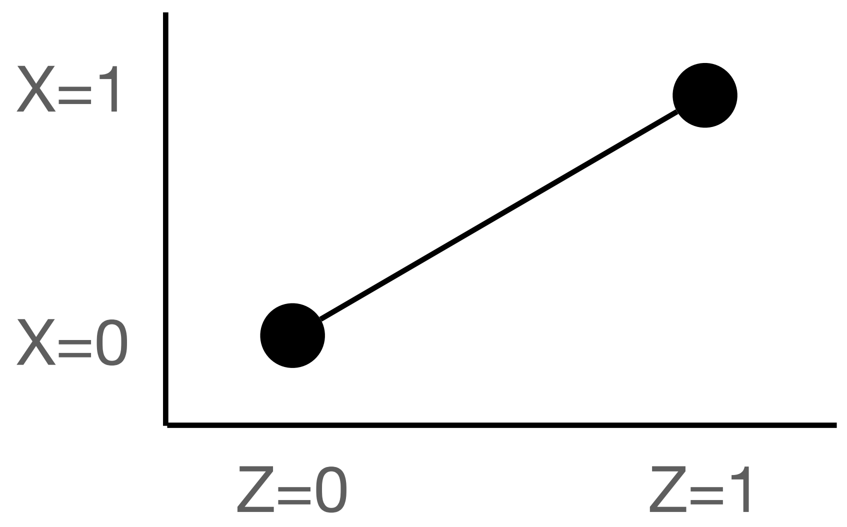
$$Z = 1 \Rightarrow X = 1$$



Compliers

$$Z = 0 \Rightarrow X = 0$$

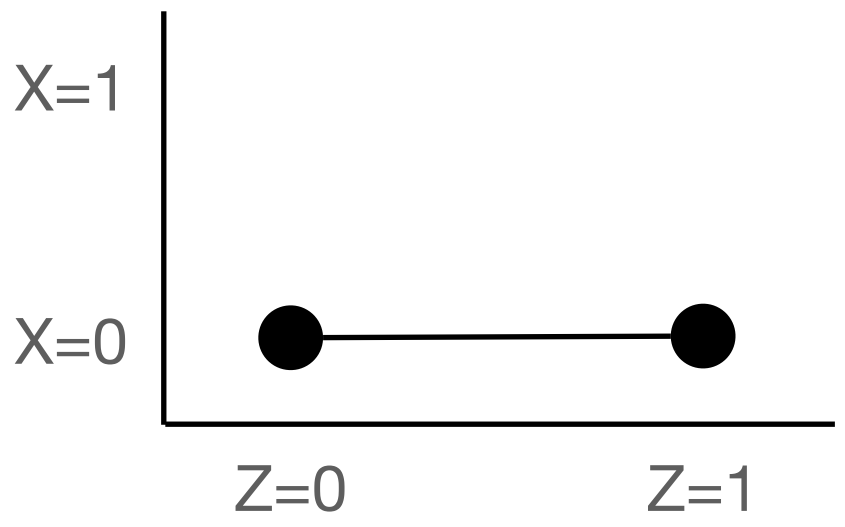
$$Z = 1 \Rightarrow X = 1$$



Never-taker

$$Z = 0 \Rightarrow X = 0$$

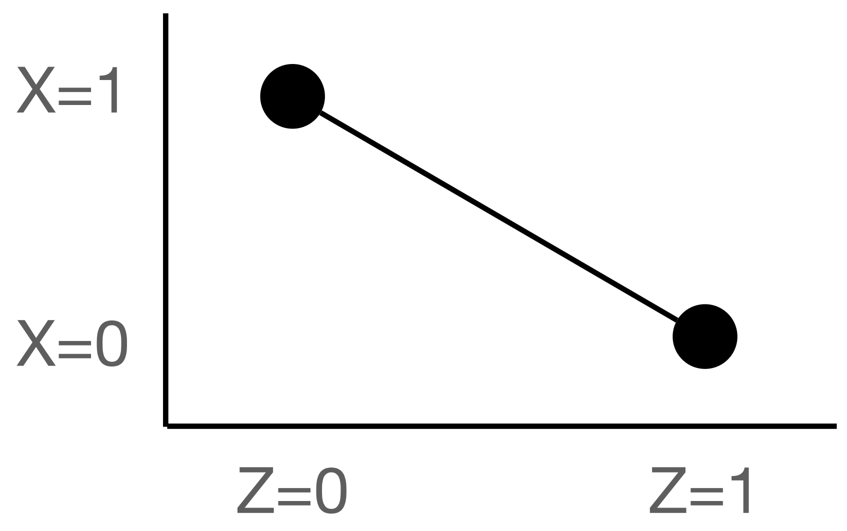
$$Z = 1 \Rightarrow X = 0$$



Defiers

$$Z = 0 \Rightarrow X = 1$$

$$Z = 1 \Rightarrow X = 0$$



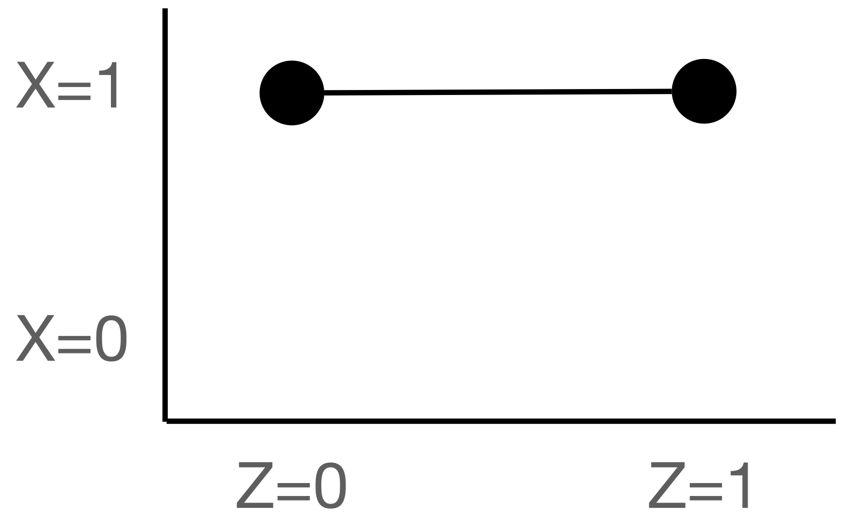
Monotonicity: No defiers

Suppose no defiers.

Always-taker

$$Z = 0 \Rightarrow X = 1$$

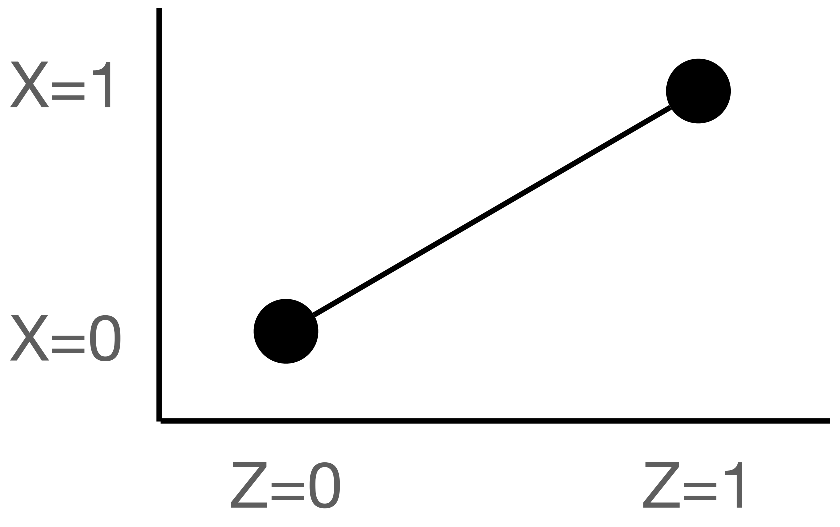
$$Z = 1 \Rightarrow X = 1$$



Compliers

$$Z = 0 \Rightarrow X = 0$$

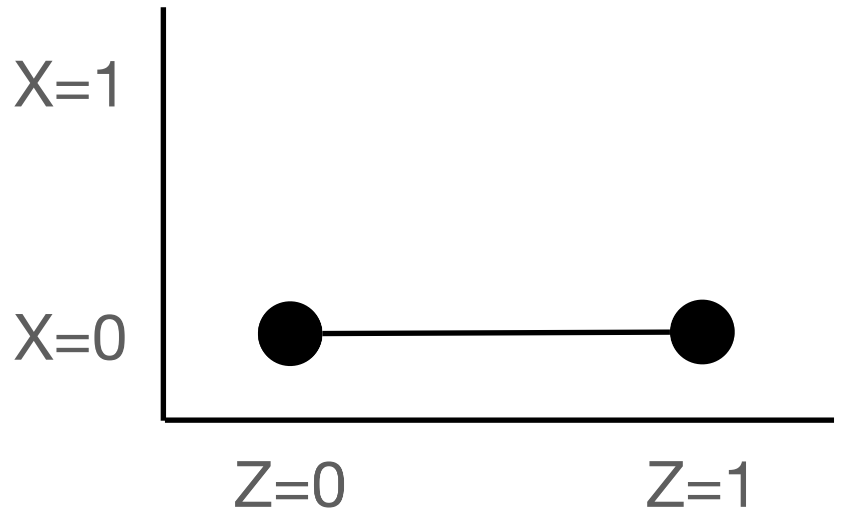
$$Z = 1 \Rightarrow X = 1$$



Never-taker

$$Z = 0 \Rightarrow X = 0$$

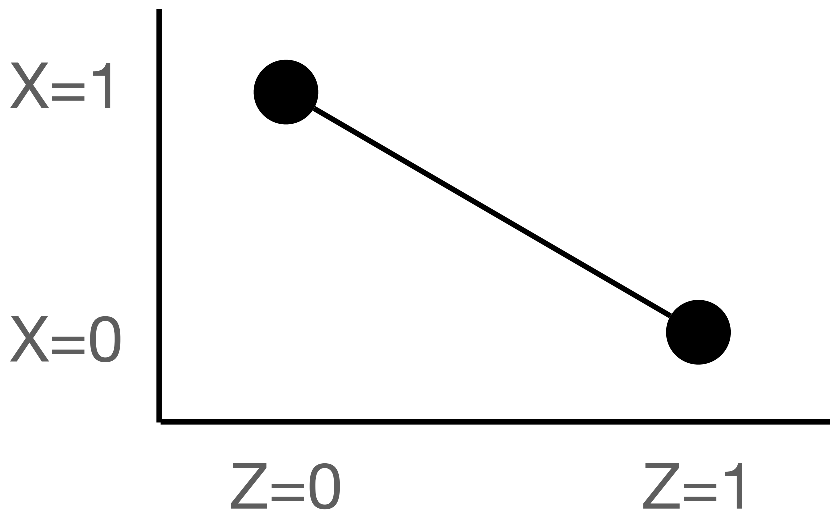
$$Z = 1 \Rightarrow X = 0$$



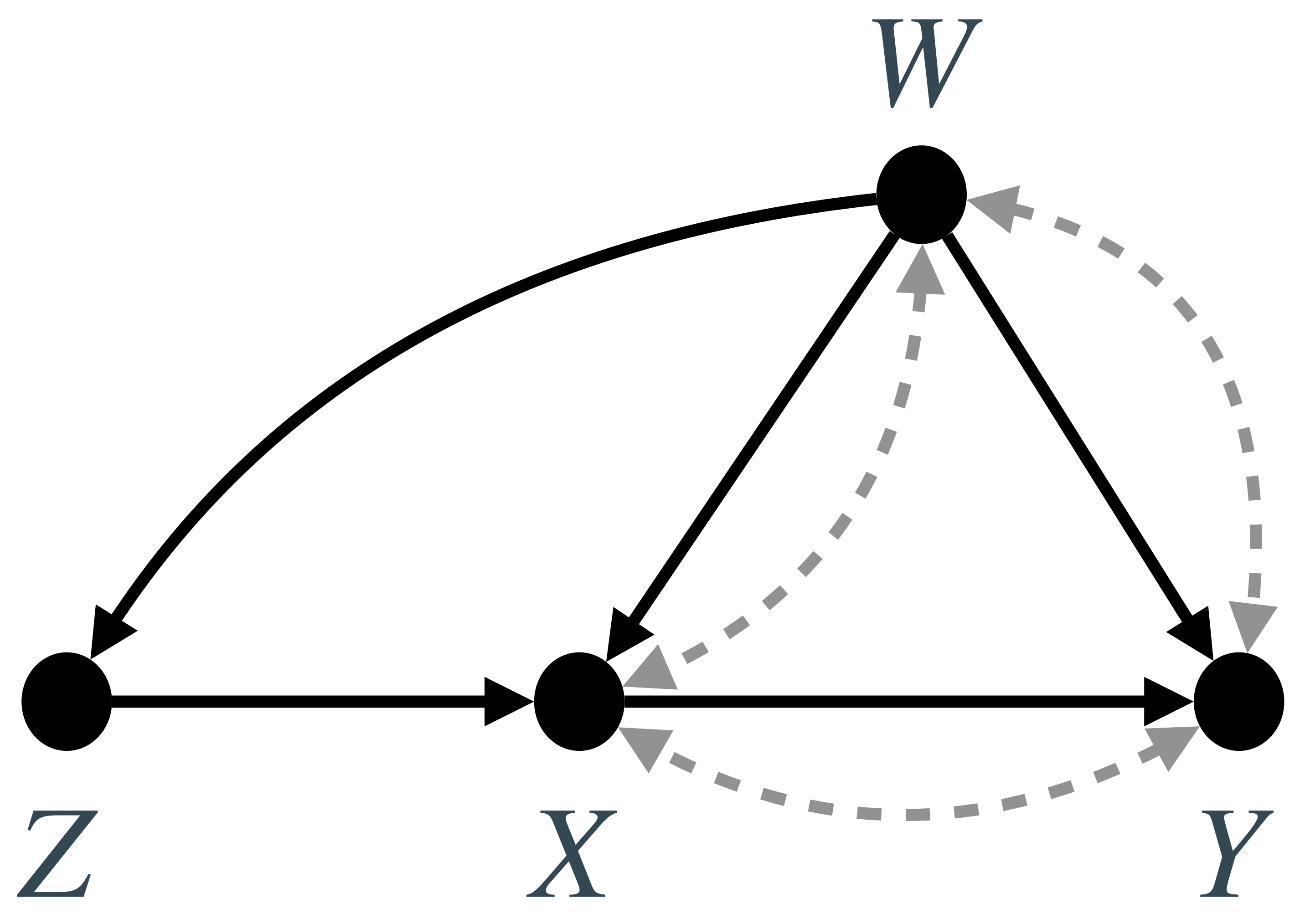
Defiers

$$Z = 0 \Rightarrow X = 1$$

$$Z = 1 \Rightarrow X = 0$$



Local Average Treatment Effect



Let C denote the complier. Under the *monotonicity*, the causal effect for compliers is identifiable and given as follow [🏆Imben and Angrist, 1994; Abadie, 2003]

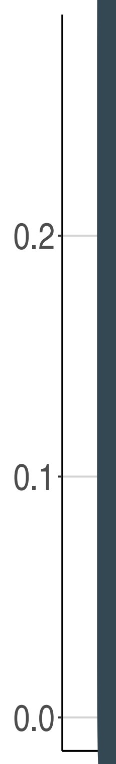
$$\mathbb{E}[f(Y) \mid do(x), C] = \frac{\mathbb{E}[\mathbb{E}[f(Y)I_x(X) \mid Z = x, W] - \mathbb{E}[f(Y)I_x(X) \mid Z = 1 - x, W]]}{\mathbb{E}[P(X = 1 \mid Z = 1, W) - P(X = 1 \mid Z = 0, W)]}$$

where f is a function of Y.

Insufficiency of LATE

- Most work considered $\mathbb{E}[Y \mid do(x), C]$ (with $f(Y) = Y$) called 'local average treatment effect (LATE)'.

-



We need the density estimator!

LATE is insufficient to understand the treatment effect.

Challenges in estimating the density

$$\mathbb{E}[f(Y) \mid do(x), C] = \frac{\mathbb{E}[\mathbb{E}[f(Y)I_x(X) \mid Z = x, W] - \mathbb{E}[f(Y)I_x(X) \mid Z = 1 - x, W]]}{\mathbb{E}[P(X = 1 \mid Z = 1, W) - P(X = 1 \mid Z = 0, W)]}$$

😊 We can leverage the existing method for estimating LATE.

😔 The existing methods cannot be directly applied for estimating the density (non-regular estimand).

Toward estimating the density

We propose two methods for estimating the density $p(y_x \mid C)$.

1. **Kernel-smoothing based approach**

   In $\mathbb{E}[f(Y) \mid do(x), C]$, *choose f such that approximates* $\mathbb{E}[f(Y) \mid do(x), C] \approx p(y \mid do(x), C)$.

   Then, we can use the existing methods designed for estimating the LATE $\mathbb{E}[Y \mid do(x), C]$.

2. **Model-based approach**

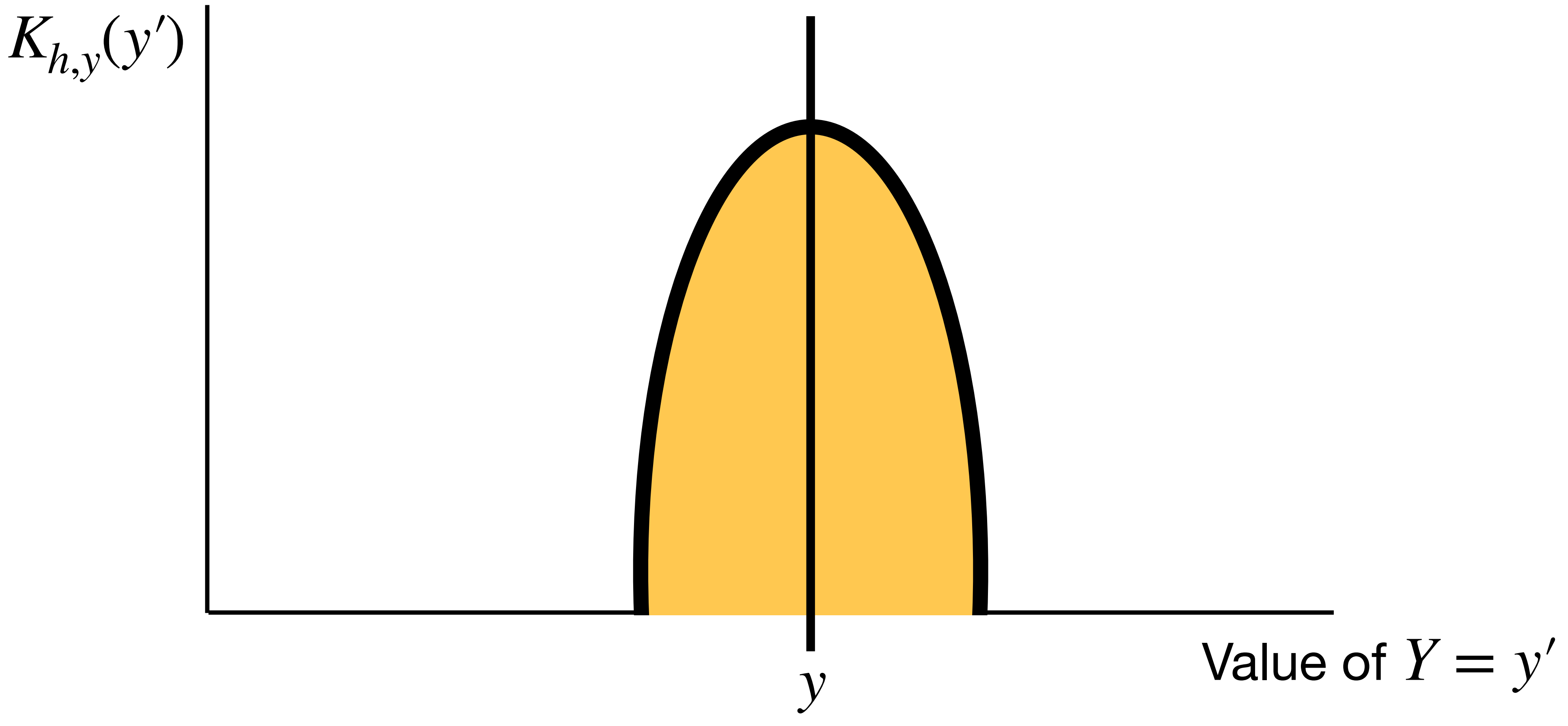
   Assume that the density can be modeled by a parametric model $g(y; \beta)$ (e.g., $p(y \mid do(x), C)$ can be modeled as a normal distribution $g(y; \beta = \{\mu, \sigma\})$).

   Then, we estimate β.

Approach 1 — Kernel-smooth based

In $\mathbb{E}[f(Y) \mid do(x), C]$, choose f for approximating $\mathbb{E}[f(Y) \mid do(x), C] \approx p(y \mid do(x), C)$.

We used $f(Y)$ as a Kernel $K_{h,y}(\cdot)$, which is a smooth approximation for denoting $Y = y$.



For example, Gaussian kernel is given as

$$K_{h,y}(y') \equiv \frac{1}{h\sqrt{2\pi}} \exp\left(-\frac{(y'-y)^2}{2h^2}\right)$$

Approach 1 — Kernel-smooth based

We propose a *double/debiased machine learning* (DML, Chernozhukov et al., 2018) based estimator which is a function of the followings; $\widehat{\psi}_h(y) \equiv f(\widehat{\pi}_z, \widehat{\xi}_x, \widehat{\theta}[K_{h,y}(Y)])$.

$$\pi_z(w) \equiv P(z \mid w); \ \xi_x(z,w) \equiv P(x \mid z, w); \ \theta(x,z,w)[f(Y)] \equiv \mathbb{E}[f(Y)I_x(X) \mid x, w].$$

The proposed estimator exhibits robustness property of DML:

Thm. 2. Debiasedness property of $\widehat{\psi}_h$

$\widehat{\psi}_h$ converges to $p(y \mid do(x), C)$ in $\sqrt{nh}$-rate if

- $\widehat{\pi}_z, \widehat{\xi}_x, \widehat{\theta}$ converges in much slower $n^{-1/4}$ rate; or
- $\widehat{\pi}_z$ or $\{\widehat{\xi}_x, \widehat{\theta}\}$ are correctly specified.

Approach 2 — Model-based

Goal — Learn β that minimizes the divergence between $p(y \mid do(x), C)$ and $g(y; \beta)$.

We propose a *DML* based estimator using the following parameters:

$$\pi_z(w) \equiv P(z \mid w); \; \xi_x(z, w) \equiv P(x \mid z, w); \; \theta(x, z, w)[f(Y)] \equiv \mathbb{E}[f(Y) I_x(X) \mid x, w].$$

For example, suppose $g(y; \beta = \{\mu, \sigma\})$ is a normal distribution. Then, the DML estimates of μ, σ are given as a function $\widehat{\mu} = g_{\mu}(\widehat{\pi}_z, \widehat{\xi}_x, \widehat{\theta})$ and $\widehat{\sigma} = g_{\sigma}(\widehat{\pi}_z, \widehat{\xi}_x, \widehat{\theta})$.

Approach 2 — model-smooth based

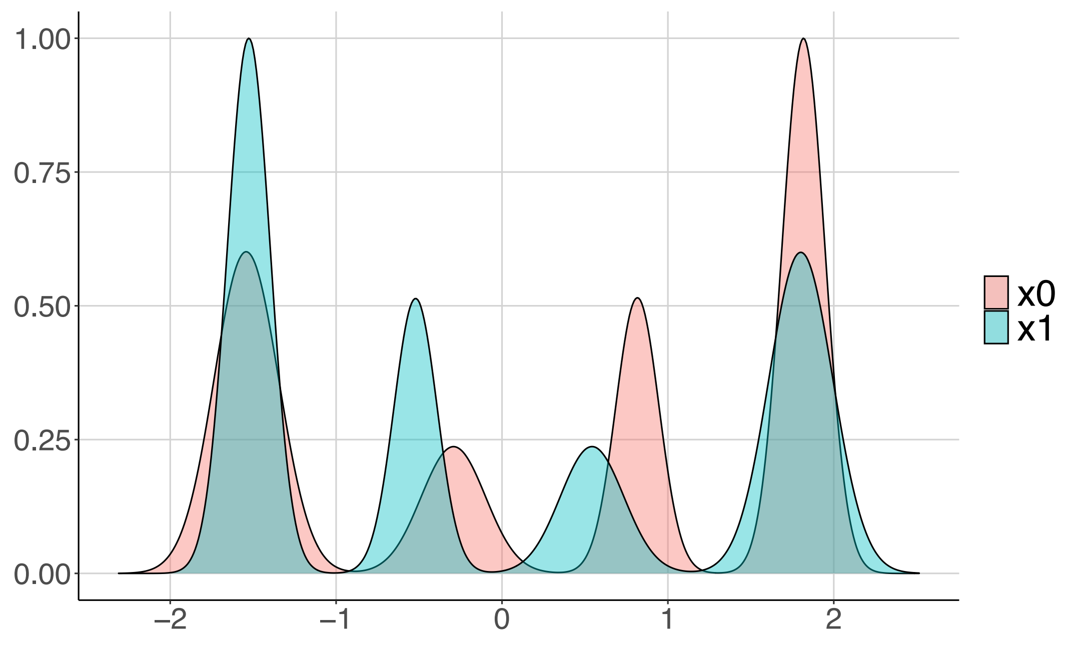
The proposed model-based DML estimators estimates (e.g., $\widehat{\mu} = g_\mu(\widehat{\pi}_z, \widehat{\xi}_x, \widehat{\theta})$ and $\widehat{\sigma} = g_\sigma(\widehat{\pi}_z, \widehat{\xi}_x, \widehat{\theta})$) exhibit robustness property of DML:

Thm. 2. Debiasedness of the model-based approach

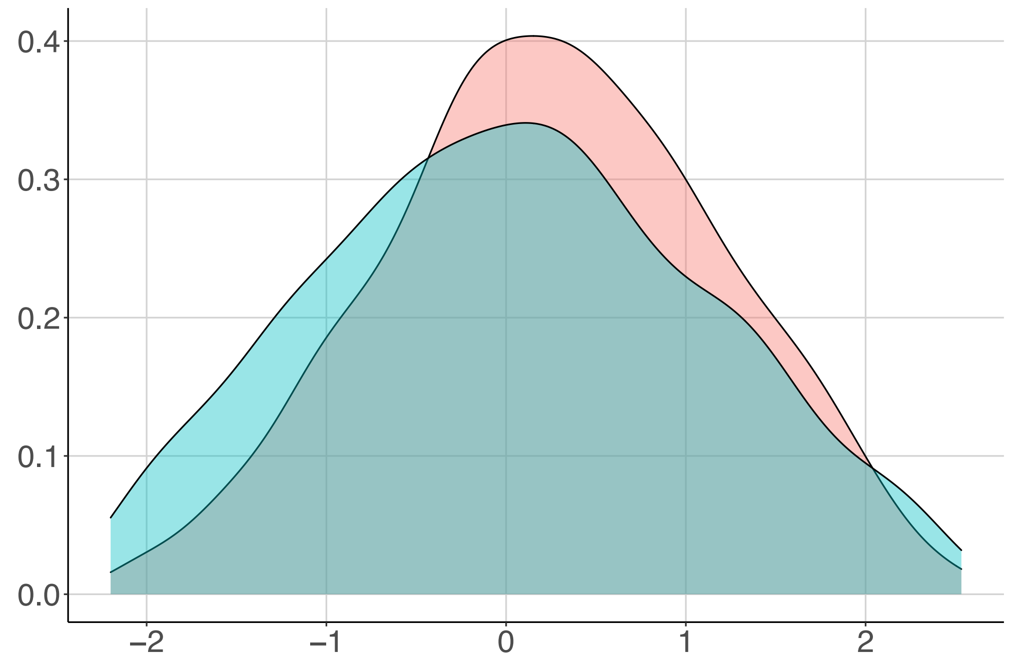
The proposed model-based DML estimator estimates $\widehat{\beta}$ (e.g., $\widehat{\beta} = \{\widehat{\mu}, \widehat{\sigma}\}$) in root-$n$ rate if

- $\widehat{\pi}_z, \widehat{\xi}_x, \widehat{\theta}$ converges in much slower $n^{-1/4}$ rate; or
- $\widehat{\pi}_z$ or $\{\widehat{\xi}_x, \widehat{\theta}\}$ are correctly specified.

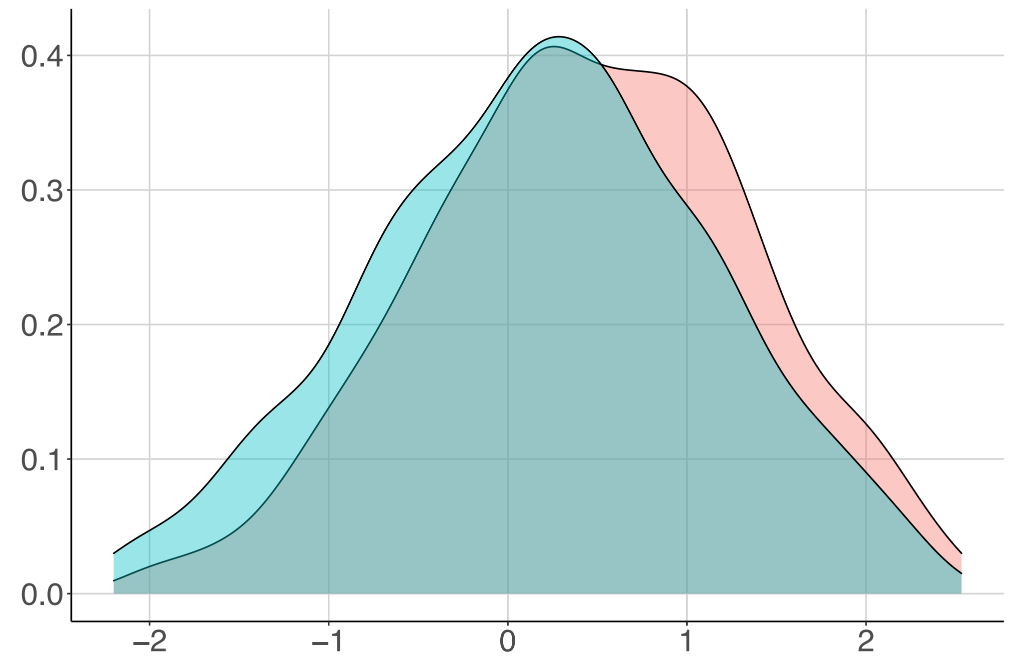
Results: Synthetic dataset



Ground-truth density.

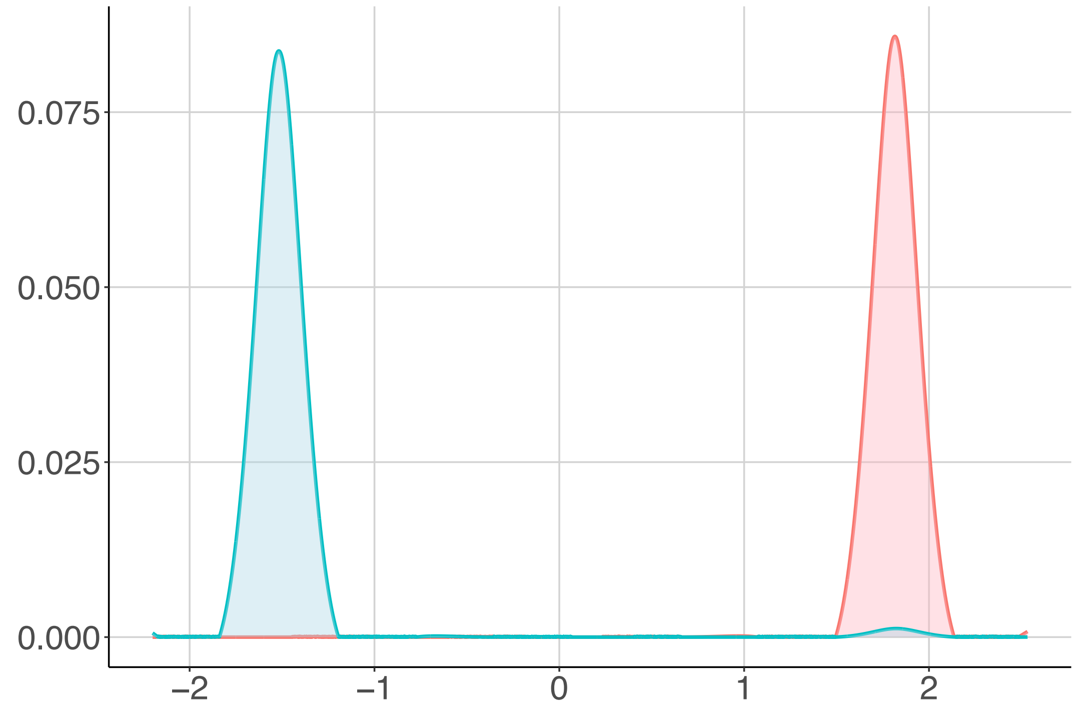


Model-based.
Non-DML

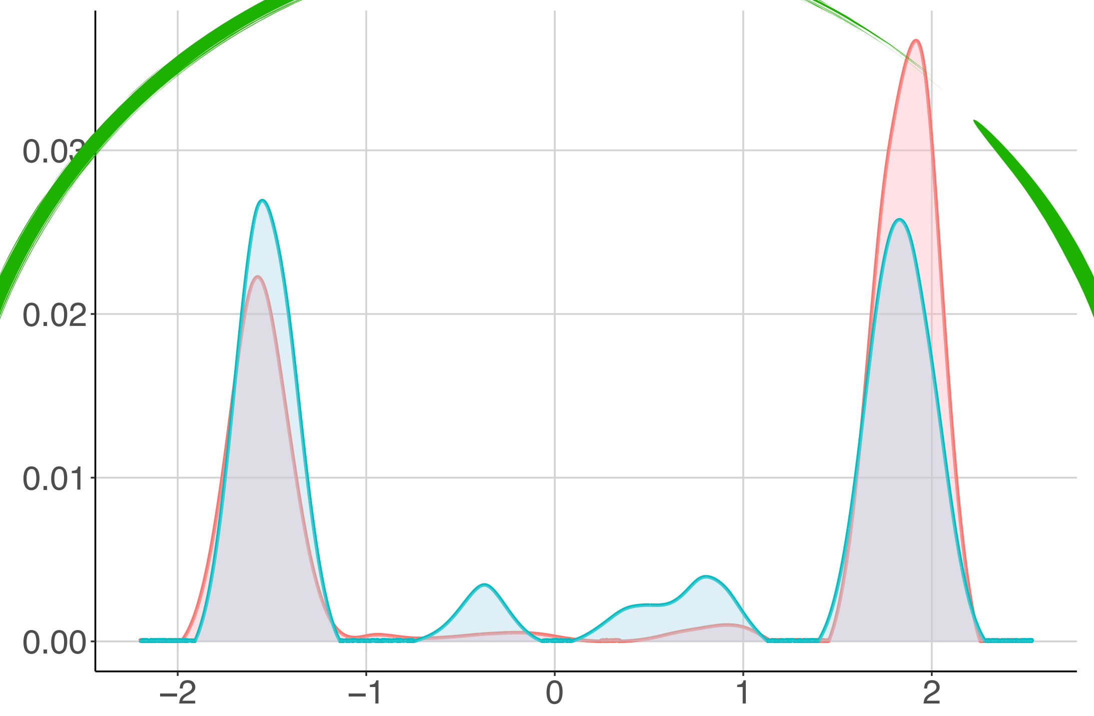


Model-based.
DML

Inaccurate result due to inconsistent assumption (Normal dist.) for the model



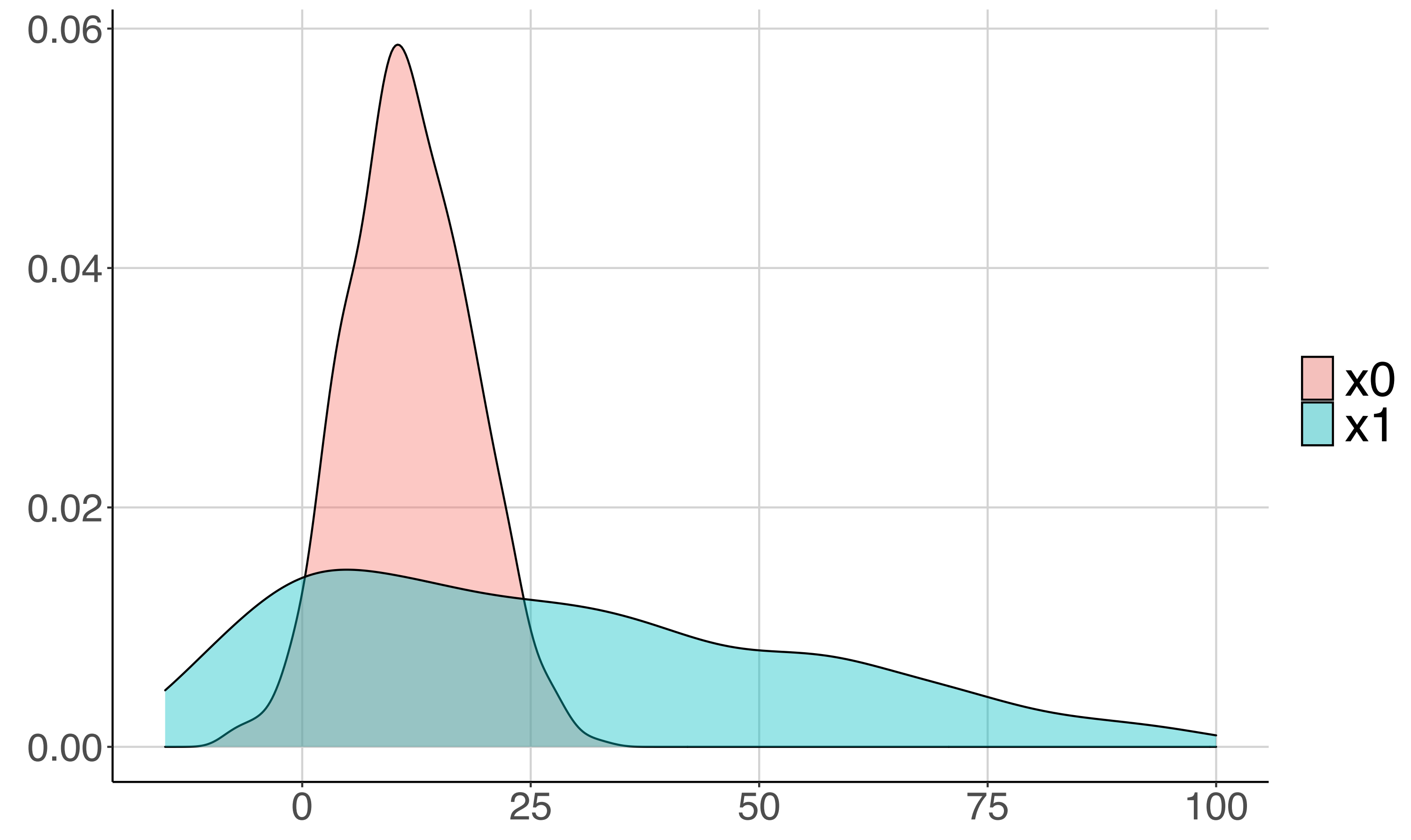
Kernel-based.
Non-DML



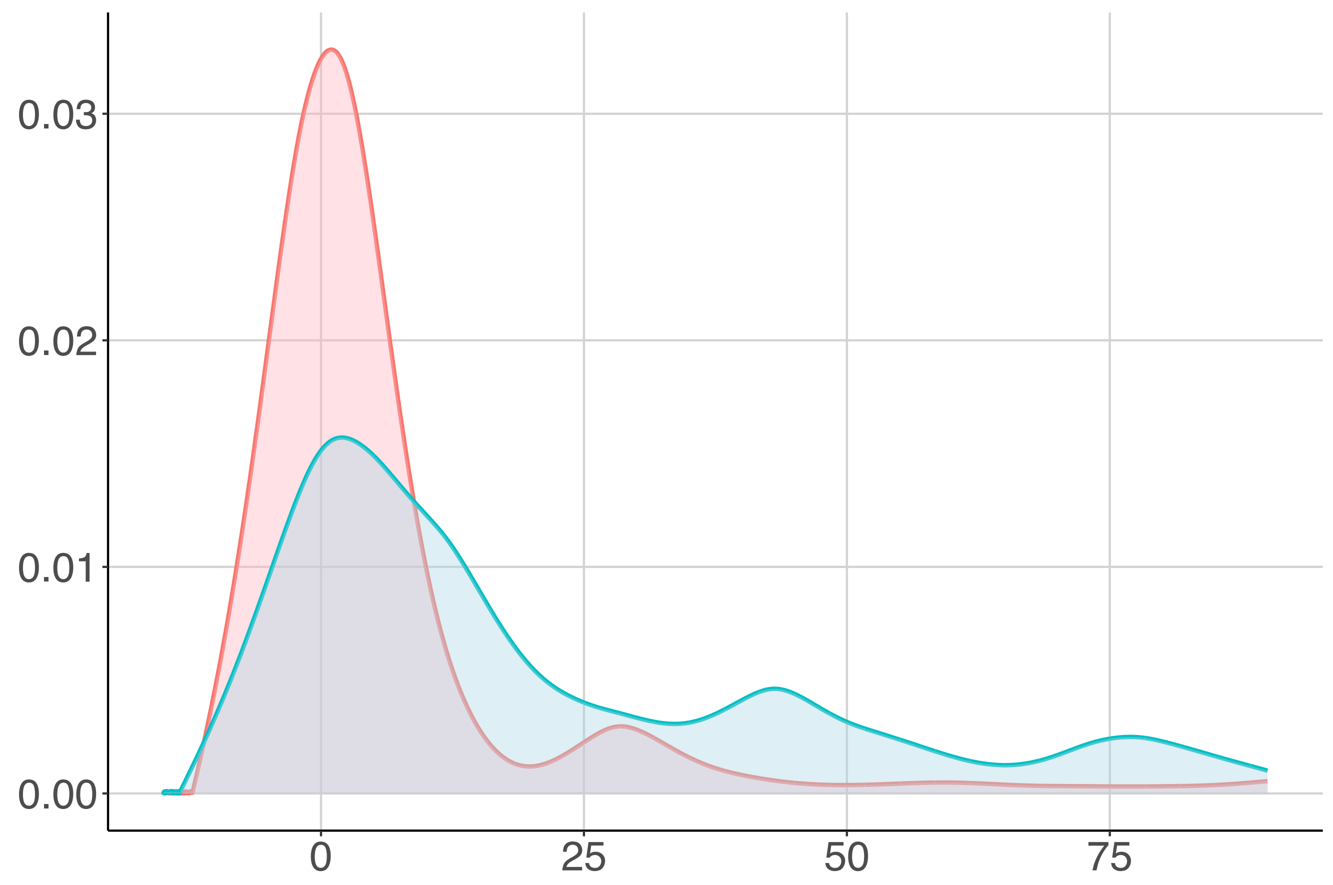
Kernel-based.
DML

Kernel-based DML estimator results in the most accurate result, capturing all modes.

Results: 401(k)



Model-based, DML

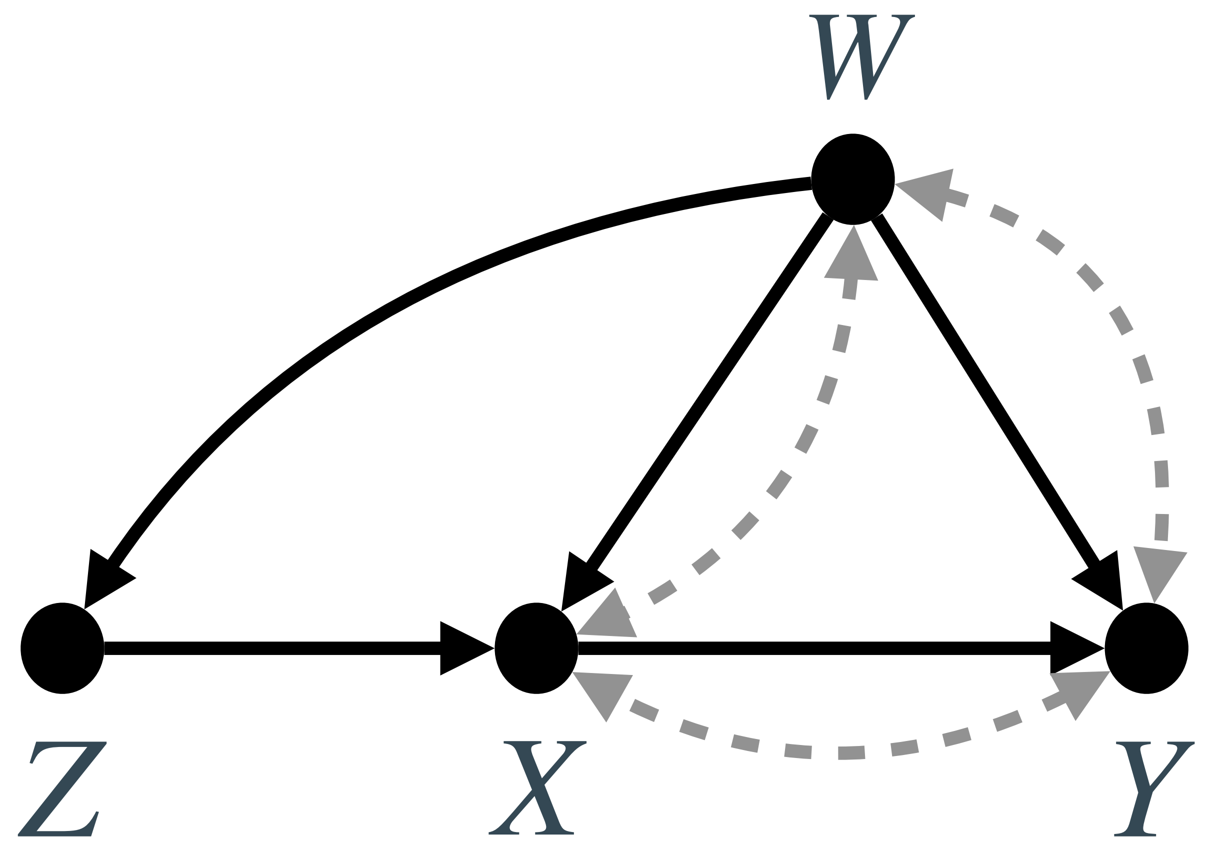


Kernel-based, DML

- Since this is a real-dataset, there is no ground-truth density.
- The implication $-$ $X = 1$ has a positive causal effect on $Y = y$ $-$ is well-known.
- For both of methods, our result is consistent with the known implication.

Conclusion

For a well-known instrumental variable setting,



- We provide the (1) Kernel-smoothing; and (2) model-based on the DML which exhibits robustness properties against the slow convergence of parameter estimates.

- We illustrated the proposed method to the synthetic and real-dataset.