A Semiparametric Approach to Data-Integrated Causal Inference

2024 SSC Annual Meeting

Archer Gong Zhang

Prof. Nancy Reid

Prof. Qiang Sun

Department of Statistical Sciences
University of Toronto

Outline

- Data-integrated causal inference
- A semiparametric model: density ratio model
- Inference procedure: empirical likelihood
- Simulation

Data-integrated causal inference

Causal inference with multi-source data

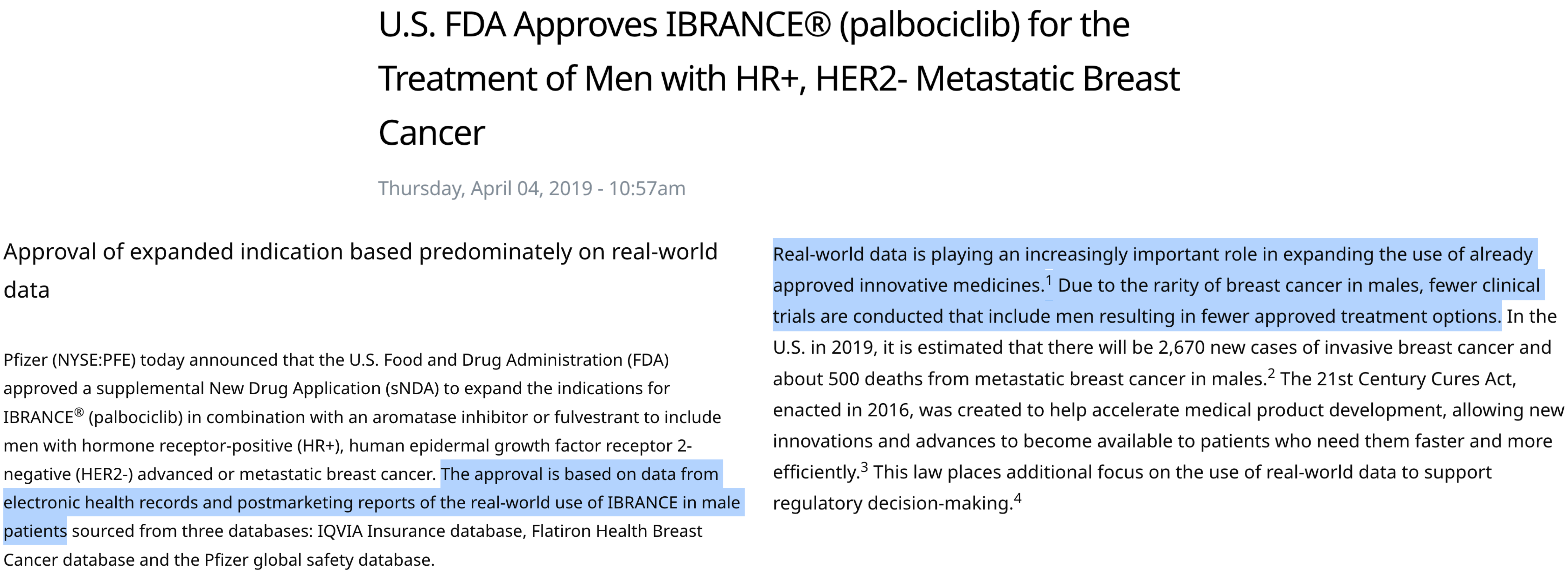
- Goal: estimate the causal effects on a **target population.**
- Data: often collected from several experimental (RCT) and observational studies.

	Experimental data	**Observational data**
Confounding	No	Inevitable
Representative of the target population	No	Yes
Size	Small	Large
Cost	High	Low
Disadvantage	Lack of external validity	Lack of internal validity

- Q: How to take advantage of both data with complementary features?

Use RCT and Obs data to generalize the treatment effect in a target population

A real-world example



U.S. FDA Approves IBRANCE® (palbociclib) for the Treatment of Men with HR+, HER2- Metastatic Breast Cancer

Thursday, April 04, 2019 - 10:57am

Approval of expanded indication based predominately on real-world data

Pfizer (NYSE:PFE) today announced that the U.S. Food and Drug Administration (FDA) approved a supplemental New Drug Application (sNDA) to expand the indications for IBRANCE® (palbociclib) in combination with an aromatase inhibitor or fulvestrant to include men with hormone receptor-positive (HR+), human epidermal growth factor receptor 2-negative (HER2-) advanced or metastatic breast cancer. The approval is based on data from electronic health records and postmarketing reports of the real-world use of IBRANCE in male patients sourced from three databases: IQVIA Insurance database, Flatiron Health Breast Cancer database and the Pfizer global safety database.

Real-world data is playing an increasingly important role in expanding the use of already approved innovative medicines.[1] Due to the rarity of breast cancer in males, fewer clinical trials are conducted that include men resulting in fewer approved treatment options. In the U.S. in 2019, it is estimated that there will be 2,670 new cases of invasive breast cancer and about 500 deaths from metastatic breast cancer in males.[2] The 21st Century Cures Act, enacted in 2016, was created to help accelerate medical product development, allowing new innovations and advances to become available to patients who need them faster and more efficiently.[3] This law places additional focus on the use of real-world data to support regulatory decision-making.[4]

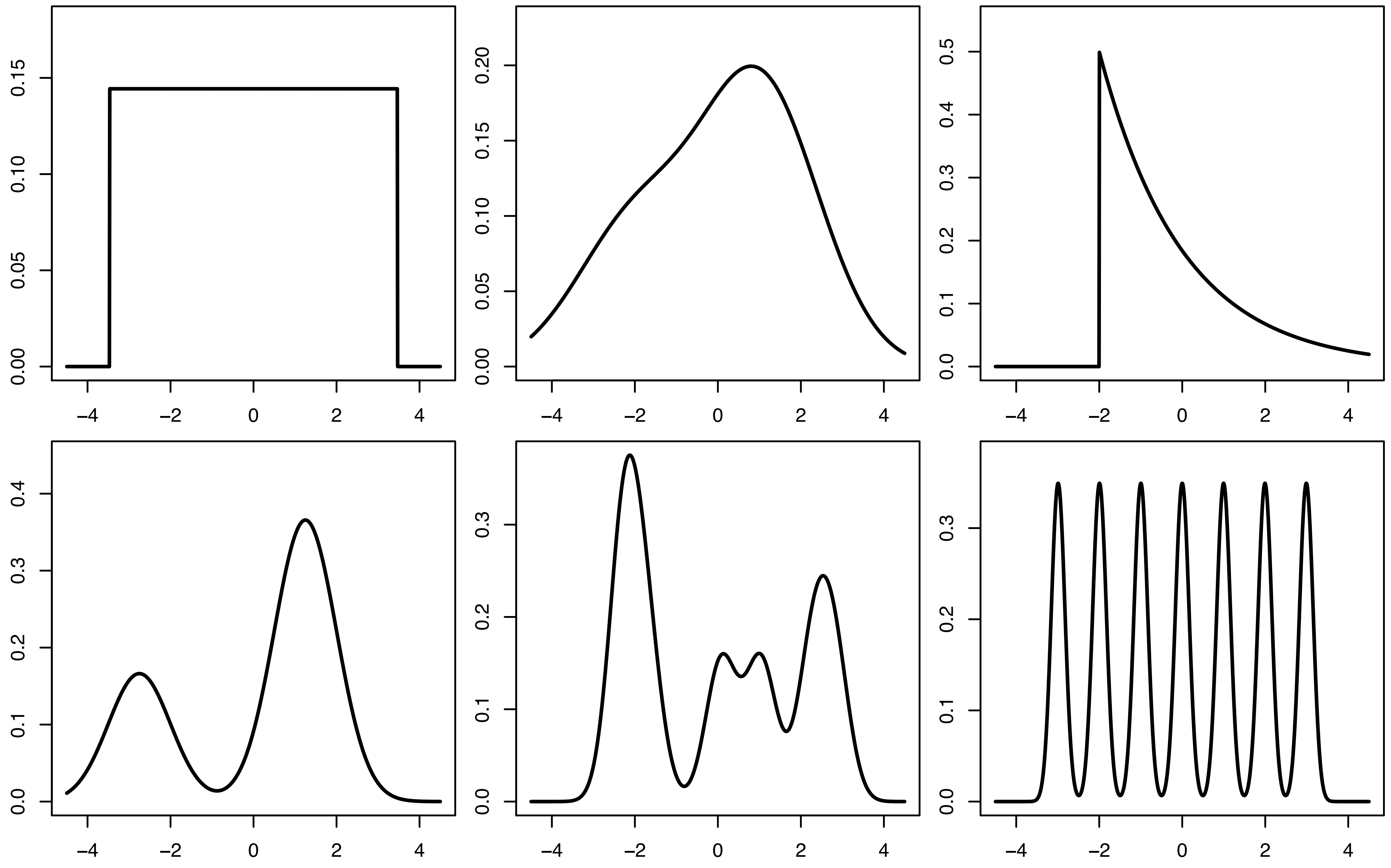
Clinical trials performed for authorization were mainly performed on the female population.

Source: https://www.pfizer.com/news/press-release/press-release-detail/u_s_fda_approves_ibrance_palbociclib_for_the_treatment_of_men_with_hr_her2_metastatic_breast_cancer.

Distribution-centric causal inference is needed

- Many studies focus on mean estimation: e.g., average treatment effect (ATE) and conditional ATE (CATE).

- Kennedy et al. (2023): “Causal effects are often characterized with averages, which can give an incomplete picture of the underlying counterfactual distributions.”



Six distributions that all have the same mean and variance.

- It is more sensible to understand and study causal effects from a distributional viewpoint.

E. H. Kennedy, S. Balakrishnan, and L. Wasserman. Semiparametric counterfactual density estimation. *Biometrika*, 110(4):875–896, 2023.

Setup

- Potential outcome[1]: $Y(a)$ with treatment $a = 0,\ldots,K$, whose distribution is called a counterfactual distribution
- Data: $\{(X_i, A_i, Y_i, S_i) : i\}$
  - $Y = Y(A)$ is the observed actual outcome
  - $S_i = 1$ if $i \in$ RCT; $S_i = 0$ if $i \in$ Obs
- Goal: estimate the distribution of $Y(a)$ in the target population represented by the Obs population.
- Assumptions for *identifiable* causal inference:
  1. Internal validity of RCT: $Y(a) \perp A \,|\, X, S = 1$, for all $a = 0,\ldots,K$
  2. Transportability: $Y(a) \perp S \,|\, X$, for all $a = 0,\ldots,K$
- Strategy:
  1. Estimate the conditional distribution of $Y(a)\,|\,X$, which is identified by $Y\,|\,A = a, X, S = 1$
  2. Marginalize $Y\,|\,A = a, X, S = 1$ over X with $S = 0$

[1] D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

A semiparametric approach: density ratio model

Density ratio model (DRM) (Anderson, 1979)

- Let $G(y \mid x, a, s)$ be the distribution of $Y \mid X = x, A = a, S = s$.

- Model assumption: for all $a = 0, \ldots, K; s = 0,1,$

vector-valued function

$$\mathrm{d}G(y \mid x, a, s) = \exp\{\alpha(x, a, s) + \beta^{\top}(x, a, s) q(y)\} \mathrm{d}G_0(y).$$

“normalizing constant”

a baseline distribution

- Why DRM?
  - Flexible: G_0 is unspecified and users can specify $\beta(x, a, s), q(y)$ as they wish — it can be seen as a generalization of the GLM.
  - Interpretable: provides a structured framework for modelling distribution shifts caused by treatments a and populations s.

J. Anderson. Multivariate logistic compounds. *Biometrika*, 66(1):17–26, 1979.

Choices of $\beta(x, a, s), q(y)$

DRM: $\mathrm{d}G(y \mid x, a, s) = \exp\{\alpha(x, a, s) + \beta^\top(x, a, s) q(y)\}\mathrm{d}G_0(y),$

- We pre-specify $q(y)$ and delegate the inference of the DRM to $\beta(x, a, s)$.
- Choice of $q(y)$ has been explored in the literature under a marginal DRM for Y alone:
  - Exploratory data analysis
  - To ensure a sufficiently rich DRM: $q(y) = (\lvert y \rvert^{1/2}, y, y^2, \log \lvert y \rvert)^\top$
  - Data-adaptive $q(y)$ by Zhang and Chen (2022) using functional principal component analysis
- We allow a user-specified parametric form for $\beta(x, a, s) = \beta(x; \theta_{a,s})$ and estimate $\theta_{a,s}$:
  - e.g., $\beta(x; \theta_{a,s}) = x^\top \theta_{a,s}$ or also include higher-order terms, or splines
  - Without a known parametric form, estimating the infinite-dimensional $\beta(x, a, s)$ for each x becomes challenging, particularly in the absence of repeated x values in the data.

A. G. Zhang and J. Chen. Density ratio model with data-adaptive basis function. *Journal of Multivariate Analysis*, page 105043, 2022.

Inference procedures: empirical likelihood

Inference for the unspecified baseline $G_0(y)$

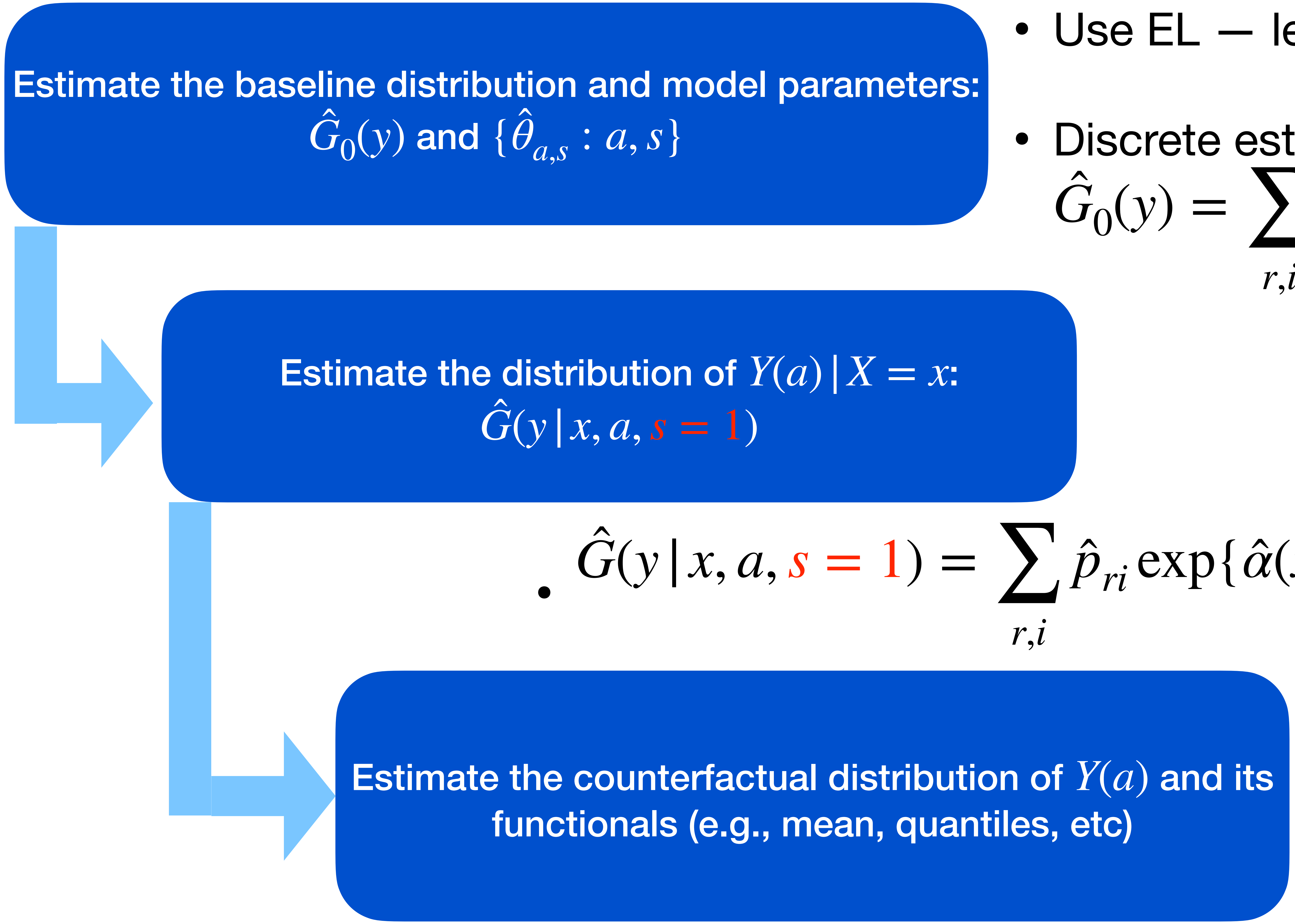
- If assigning a parametric form to G_0, DRM would reduce to a fully parametric model.
- Use a nonparametric inference method: empirical likelihood (EL; Owen, 1988).

Art B. Owen

Owen (2001): "EL keeps the effectiveness of likelihood methods and does not impose a known family distribution on the data."

EL-DRM framework enables utilization of the **entire data** to estimate each distributions, rather than **data only from themselves**.

Inference of the counterfactual distributions



Estimate the baseline distribution and model parameters: $\hat{G}_0(y)$ and $\{\hat{\theta}_{a,s} : a, s\}$

Estimate the distribution of $Y(a) \,|\, X = x$: $\hat{G}(y \,|\, x, a, s = 1)$

Estimate the counterfactual distribution of $Y(a)$ and its functionals (e.g., mean, quantiles, etc)

- Use EL — leads to consistent estimators

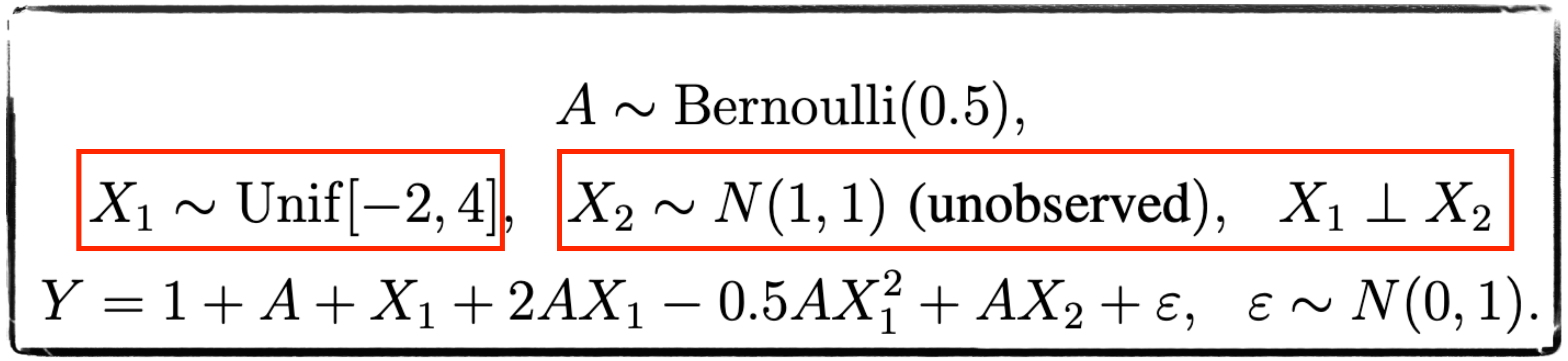
- Discrete estimator of baseline distribution:
  $$\hat{G}_0(y) = \sum_{r,i} \hat{p}_{ri} 1(y_{ri} \le y)$$

- $$\hat{G}(y \,|\, x, a, s = 1) = \sum_{r,i} \hat{p}_{ri} \exp\{\hat{\alpha}(x, a, 1) + \beta^\top(x; \hat{\theta}_{a,1}) q(y_{ri})\} 1(y_{ri} \le y)$$

- Marginalizing $\hat{G}(y \,|\, x, a, s = 1)$ over the observed x in **observational data**

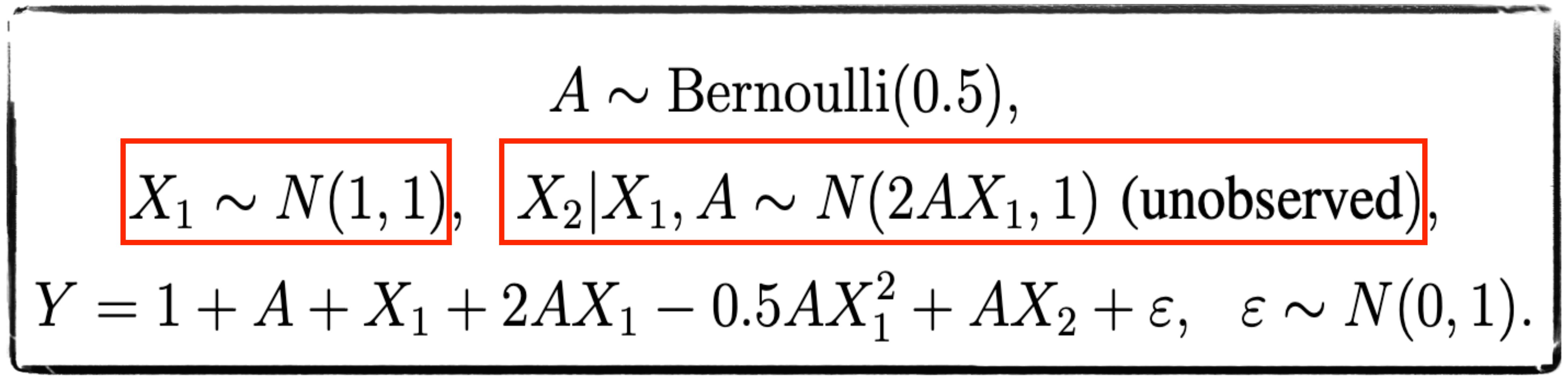
Simulation

Simulated data



$$A \sim \mathrm{Bernoulli}(0.5),$$
$$\boxed{X_1 \sim \mathrm{Unif}[-2,4]}, \quad \boxed{X_2 \sim N(1,1)\ \textbf{(unobserved)}, \quad X_1 \perp X_2}$$
$$Y = 1 + A + X_1 + 2AX_1 - 0.5AX_1^2 + AX_2 + \varepsilon, \quad \varepsilon \sim N(0,1).$$

RCT data



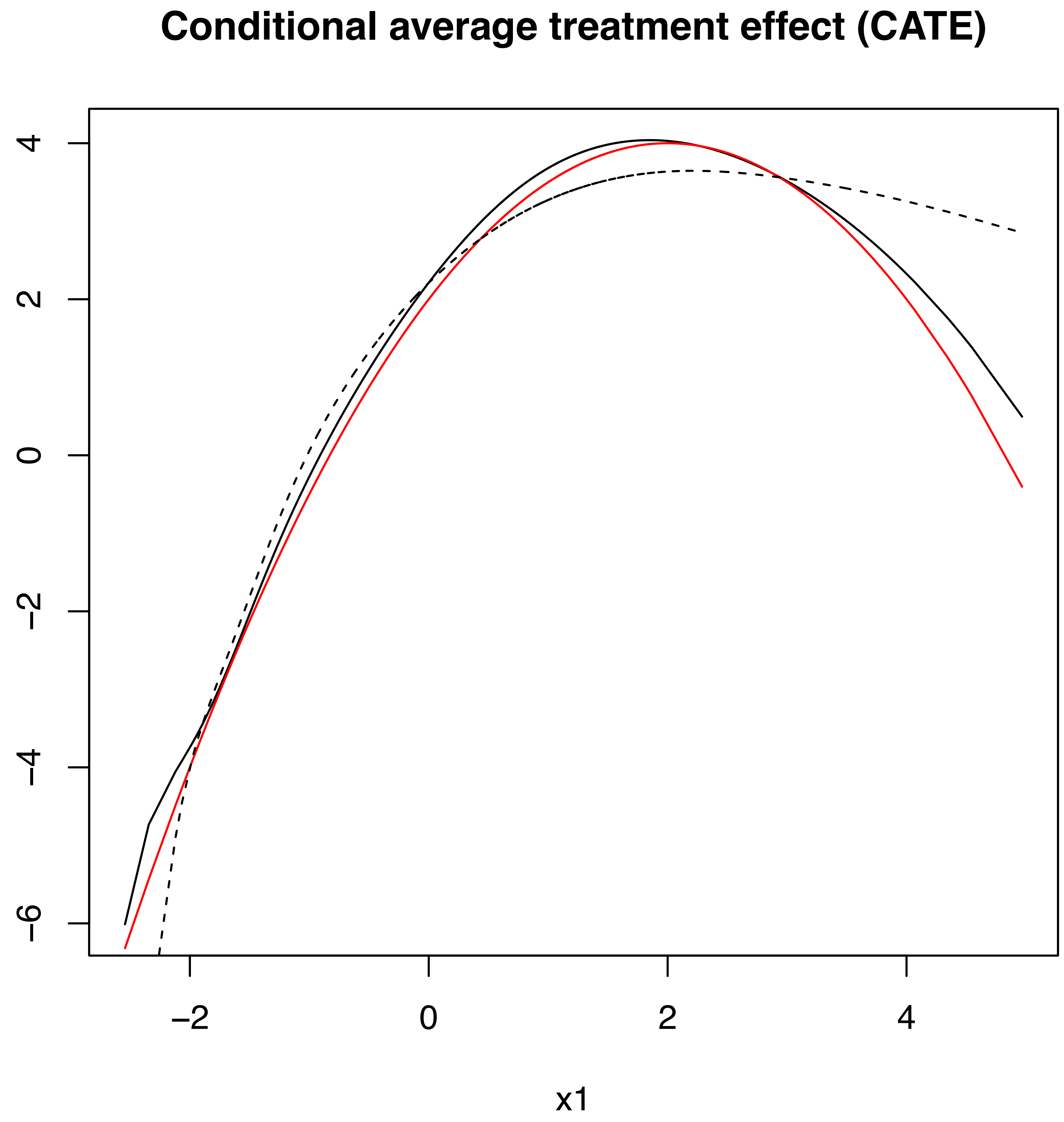
$$A \sim \mathrm{Bernoulli}(0.5),$$
$$\boxed{X_1 \sim N(1,1)}, \quad \boxed{X_2 | X_1, A \sim N(2AX_1, 1)\ \textbf{(unobserved)}},$$
$$Y = 1 + A + X_1 + 2AX_1 - 0.5AX_1^2 + AX_2 + \varepsilon, \quad \varepsilon \sim N(0,1).$$

Observational data

- X_1 has a larger support for observational data: mimics the real-world scenario.
- X_2 is not a confounder for RCT but is for Obs.
- Correctly specified DRM: $q(y) = (y, y^2)^\top$ and $\beta_{\mathrm{cor}}(x, a, s) = (x_1, x_1^2)^\top \theta_{a,s}$.
- To account for possible model misspecification, we also use $\beta_{\mathrm{mis}}(x, a, s) = x_1^\top \theta_{a,s}$.
- RCT sample size = 500; Obs sample size = 5000; 1000 simulation repetitions.

Performance of CATE estimator

Based on one simulation repetition. All DRM use $q(y) = (y, y^2)^\top$.



CATE: $\mathbb{E}[Y(1) - Y(0) \,|\, X = x]$.

Solid black —: $\beta_{\mathrm{cor}}(x, a, s) = (x_1, x_1^2)^\top \theta_{a,s}$

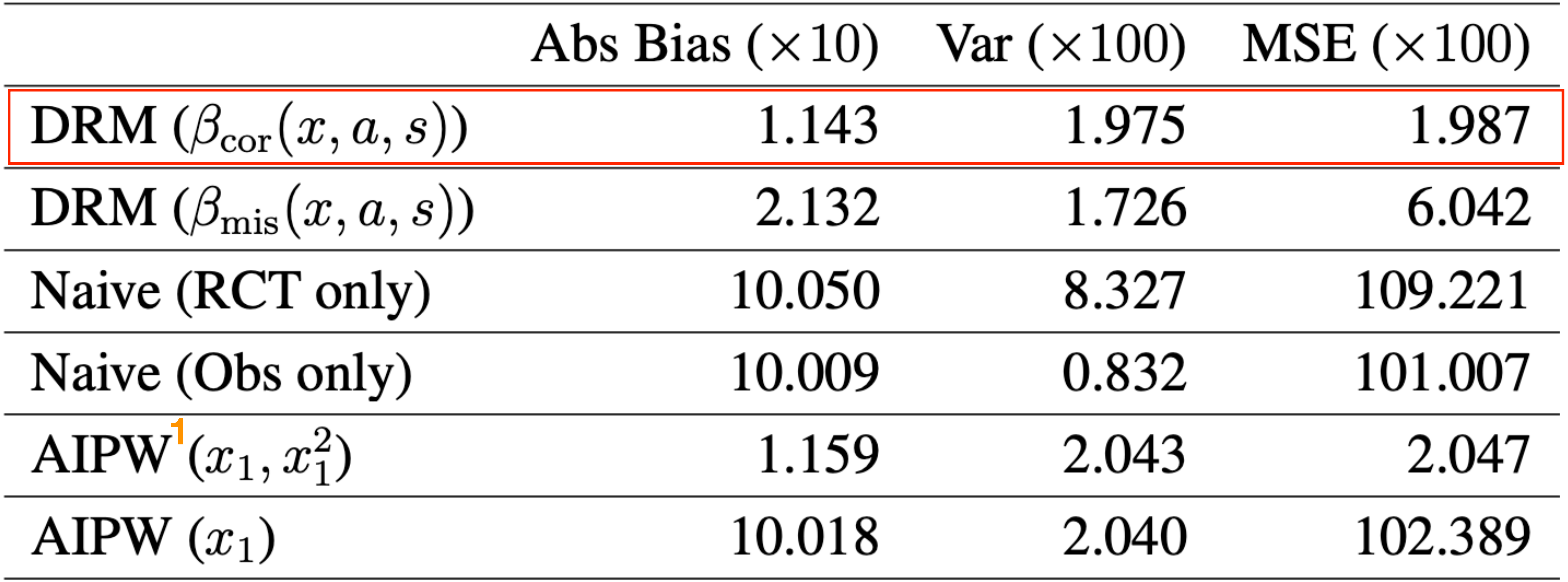
Dashed black - - -: $\beta_{\mathrm{mis}}(x, a, s) = x_1^\top \theta_{a,s}$

Solid red —: the truth

Performance of ATE estimators

ATE: $\mathbb{E}[Y(1) - Y(0)]$.

DRM: $q(y) = (y, y^2)^\top$ **and** $\beta_{\mathrm{cor}}(x, a, s) = (x_1, x_1^2)^\top \theta_{a,s}$ **or** $\beta_{\mathrm{mis}}(x, a, s) = x_1^\top \theta_{a,s}$



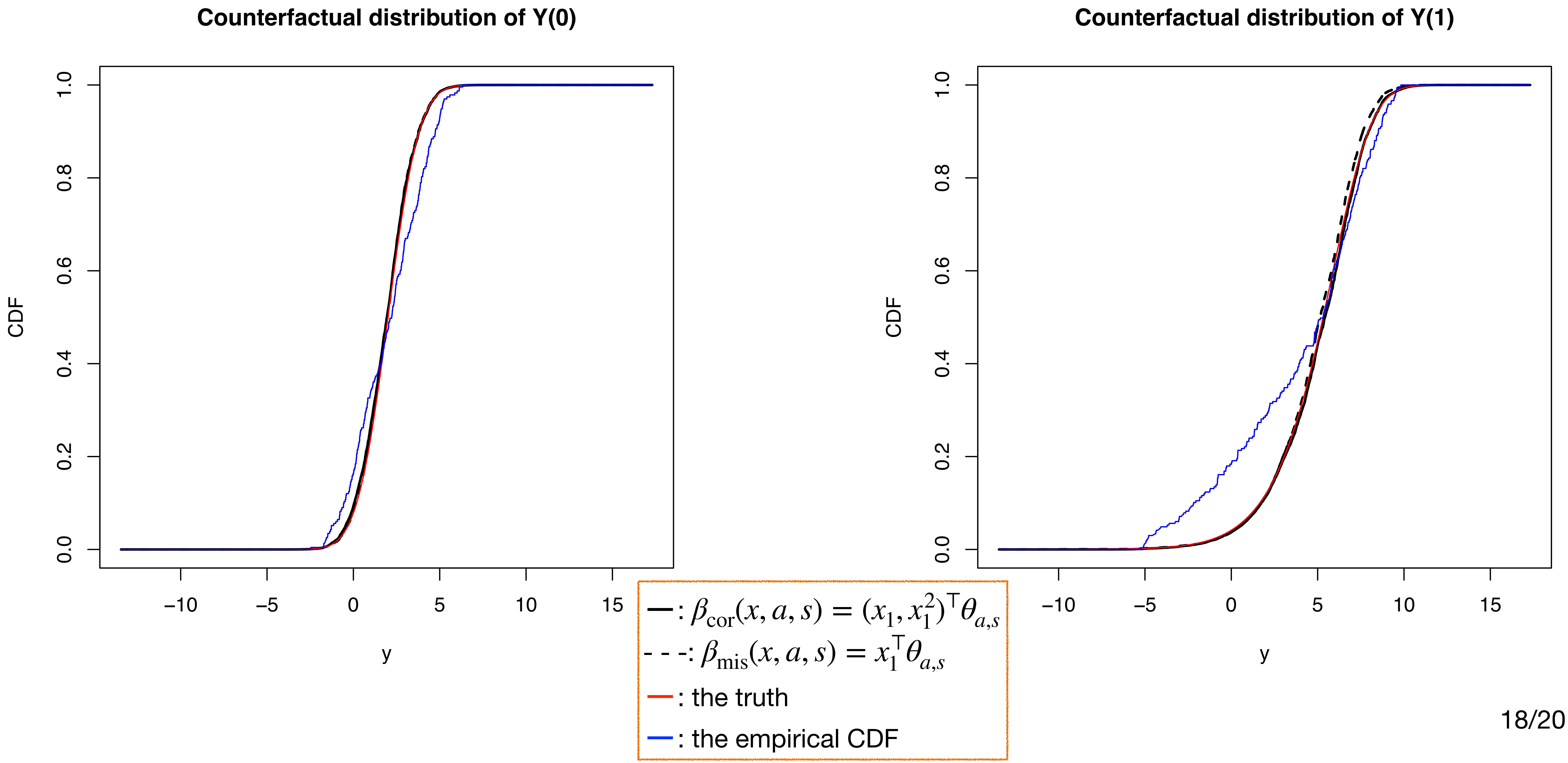
	Abs Bias ($\times$10)	Var ($\times$100)	MSE ($\times$100)
DRM ($\beta_{\mathrm{cor}}(x, a, s)$)	1.143	1.975	1.987
DRM ($\beta_{\mathrm{mis}}(x, a, s)$)	2.132	1.726	6.042
Naive (RCT only)	10.050	8.327	109.221
Naive (Obs only)	10.009	0.832	101.007
AIPW[1] (x_1, x_1^2)	1.159	2.043	2.047
AIPW (x_1)	10.018	2.040	102.389

lower is better

[1]: Colnet, Bénédicte, et al. "Causal inference methods for combining randomized trials and observational studies: a review." Statistical science 39.1 (2024): 165–191.

Performance of counterfactual distribution estimator

Based on one simulation repetition. All DRM use $q(y) = (y, y^2)^\top$.



Summary

- We propose a flexible and interpretable model for data-integrated causal inference.
    - Capture common latent structures across all counterfactual distributions:
        - 1) among treatments $a = 0, \ldots, K$
        - 2) observational versus experimental populations ($S = 0,1$)
    - Mild model assumption: the baseline distribution G_0 is unspecified.
    - Address the necessity of studying causal effects from a distributional perspective.
- Other inferences such as hypothesis testing and confidence interval is possible with our EL-DRM framework.

Thank you!

Questions & discussions are welcome! :-)