A Semiparametric Approach to Data-Integrated Causal Inference

2024 Department of Statistical Sciences Postdoc Day

Archer Gong Zhang

Prof. Nancy Reid

Prof. Qiang Sun

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?

Combine RCT and Obs data: an 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.

Trend of data integration…

One Community: Informing Decisions and Driving Discovery
TORONTO
Ontario, Canada • August 5–10
JSM2023

Action	Time	Title	Type
Sunday, August 6, 2023			
View	2:00 PM - 3:50 PM	Advances in Joint Modeling and Data Integration	Contributed Papers
Back To Top			
Monday, August 7, 2023			
View	10:30 AM - 12:20 PM	Advances of Statistical Methodologies in Biomedical Data Integration	Invited Paper Session
View	10:30 AM - 12:20 PM	Frontiers and Challenges in Data Integration Analysis with Multiple Outcomes	Topic-Contributed Paper Session
View	2:00 PM - 3:50 PM	Integrating Information from Different Data Sources: Some New Developments	Invited Paper Session
Back To Top			
Tuesday, August 8, 2023			
View	8:30 AM - 10:20 AM	When Data Integration Meets Causal Inference	Invited Paper Session
View	10:30 AM - 12:20 PM	Making the case for data quality	Topic-Contributed Paper Session
View	2:00 PM - 3:50 PM	Novel statistical methods for high-dimensional metagenomics and multi-omics data analysis	Topic-Contributed Paper Session
Back To Top			
Wednesday, August 9, 2023			
View	8:30 AM - 10:20 AM	Model Transportation, Distribution Shift, and Data Integration	Invited Paper Session
View	8:30 AM - 10:20 AM	Our Healthcare Data Community: Statistical Challenges and Discoveries using EHRs and Beyond	Invited Paper Session
View	8:30 AM - 10:20 AM	Recent advances in high-dimensional data integration methods and applications	Invited Paper Session
View	10:30 AM - 12:20 PM	Distributed, adaptive and efficient inference for modern biomedical data in the post covid world.	Topic-Contributed Paper Session
View	10:30 AM - 12:20 PM	Harnessing multiple data sources to improve generalizability of findings from clinical trials	Invited Paper Session
View	10:30 AM - 12:20 PM	Optimal Transport and Applications to Statistics	Invited Paper Session
Back To Top			
Thursday, August 10, 2023			
View	8:30 AM - 10:20 AM	Contributions to Inference from Survey Samples: In Honor of Professor Joe Sedransk	Invited Paper Session
View	8:30 AM - 10:20 AM	Methods for large multi-cohort data integration in presence of missing and imbalanced covariates	Invited Paper Session

JSM PORTLAND OREGON | AUGUST 3–8, 2024

STATISTICS AND DATA SCIENCE: INFORMING POLICY AND COUNTERING MISINFORMATION

58 Matches found.

Sunday, August 4, 2024 | Sunday, August 4, 2024 | Monday, August 5, 2024 | Tuesday, August 6, 2024 | Wednesday, August 7, 2024 | Thursday, August 8, 2024

Data Integration for Heterogeneous Data

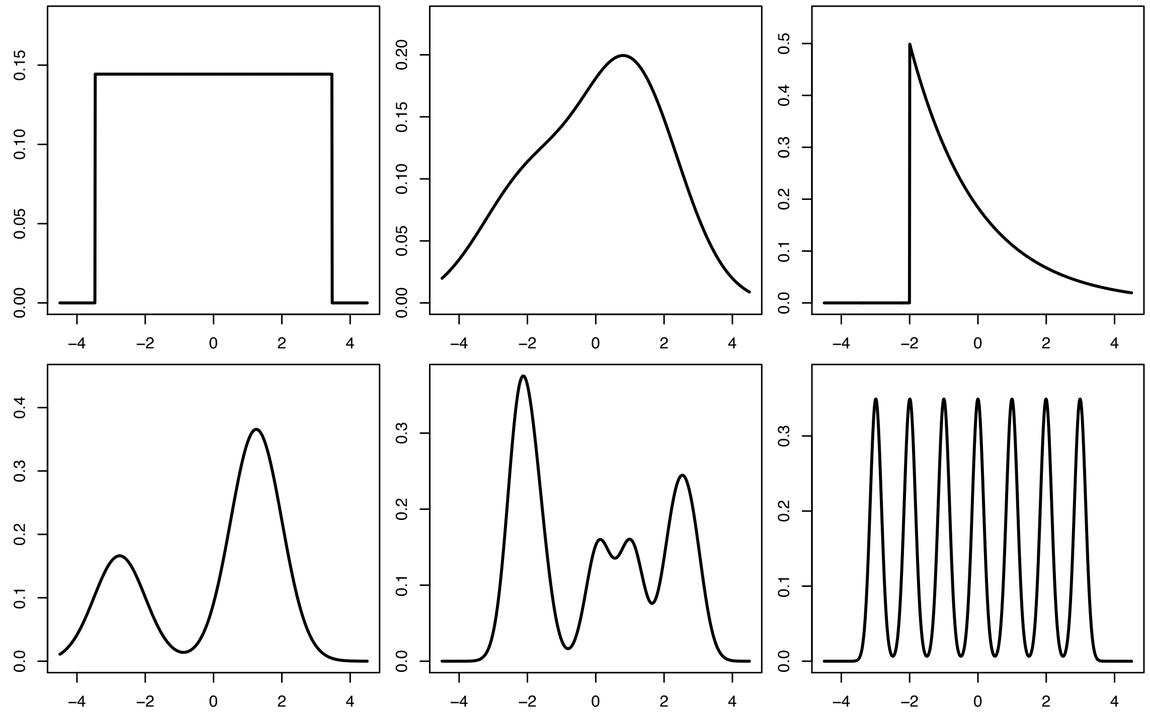
Presented During: IMS Medallion Lecture I
Annie Qu Speaker
University of California At Irvine

Fusion Learning: Combining Inferences from Diverse Data Sources

Presented During: COPSS Elizabeth L. Scott Lecture
Regina Liu Speaker
Rutgers University

Distribution-centric causal inference

- 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 $a = 0, \ldots, K$.
- Data: $\{(X_i, A_i, Y_i, S_i) : i\}$, where $S_i = 1$ if $i \in$ RCT and $S_i = 0$ if $i \in$ Obs.
- Goal: infer the **distribution of $Y(a)$** in the target population represented by the Obs.
- 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 $S = 0$ (from Obs).

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)*

- Let $G(y \mid x, a, s)$ be the distribution of $Y \mid X = x, A = a, S = s$.
- Model: for all $a = 0, \ldots, K;\ \ s = 0, 1,$

vector-valued functions

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

"normalizing constant"

a baseline distribution

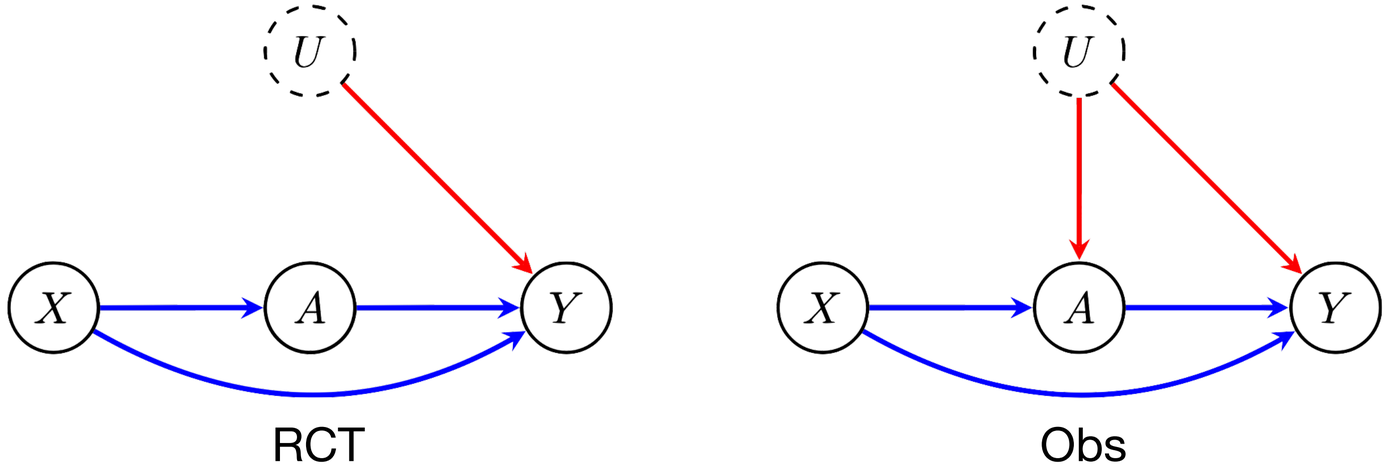
- Flexible: G_0 is unspecified and users can specify $\beta(x, a, s)$ & $q(y)$ as they wish — it can be seen as a generalization of the GLM.
- Interpretable: provides a structured framework for modelling distribution shifts caused by treatments a and populations s.

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

Where the distribution shifts come from

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

- Causal models for the two studies & data:



- If 1) U is a hidden confounder for Obs; 2) $Y(a) \perp A \mid X, U$; and 3) $Y(a) \perp S \mid X, U$,

$$Y \mid X = x, A = a, S = s \sim G(y \mid x, a, s) = \int \left[Y(a) \mid X, U\right] \times \left[U \mid X, A, S\right] \mathrm{d}U\,.$$

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

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 $q(y)$ in the literature under a marginal DRM for Y alone:
  - Exploratory data analysis.
  - To ensure a sufficiently rich DRM: $q(y) = (|y|^{1/2}, y, y^2, \log|y|)^\top$.
  - Data-adaptive by Z. and Chen (2022) via 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., $\beta(x; \theta_{a,s}) = \theta_{a,s}^\top x$, or also include higher-order terms, or splines.
  - Without this, estimating the infinite-dimensional $\beta(x, a, s)$ becomes challenging.

A. G. Zhang and J. Chen. Density ratio model with data-adaptive basis function. *Journal of Multivariate Analysis*, 191:105043, 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, 2001).

Art B. Owen

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

Archer (today): “EL-DRM framework enables utilization of the **entire data** to estimate each distribution.” 😆

A. B. Owen. *Empirical Likelihood*. Chapman & Hall/CRC, New York, 2001.

Inference of the distribution of $Y(a)$

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

- EL: utilizing the **entire data** to estimate $G_0(y)$.
- Discrete estimator of baseline distribution: $\hat{G}_0(y) = \sum_{r,i} \hat{p}_{ri} 1(y_{ri} \le y)$.

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

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

Inference on the distribution of $Y(a)$ and its functionals (e.g., mean, CDF, quantiles, etc)

- Marginalize $\hat{G}(y \,|\, x, a, s = 1)$ over the observed x in *Obs data*.
- Consistent estimators.
- Confidence regions.
- Hypothesis tests: Wald test & LRT.

Simulation

Simulation

$$A \sim \text{Bernoulli}(0.5),$$
$$X \sim \text{Unif}[-2, 4], \quad U \sim N(1, 1) \text{ (unobserved)}, \quad X \perp U$$
$$Y = 1 + A + X + 2AX - 0.5AX^2 + AU + \varepsilon, \quad \varepsilon \sim N(0, 1).$$

RCT data

$$A \sim \text{Bernoulli}(0.5),$$
$$X \sim N(1, 1), \quad U|X, A \sim N(2AX, 1) \text{ (unobserved)},$$
$$Y = 1 + A + X + 2AX - 0.5AX^2 + AU + \varepsilon, \quad \varepsilon \sim N(0, 1).$$

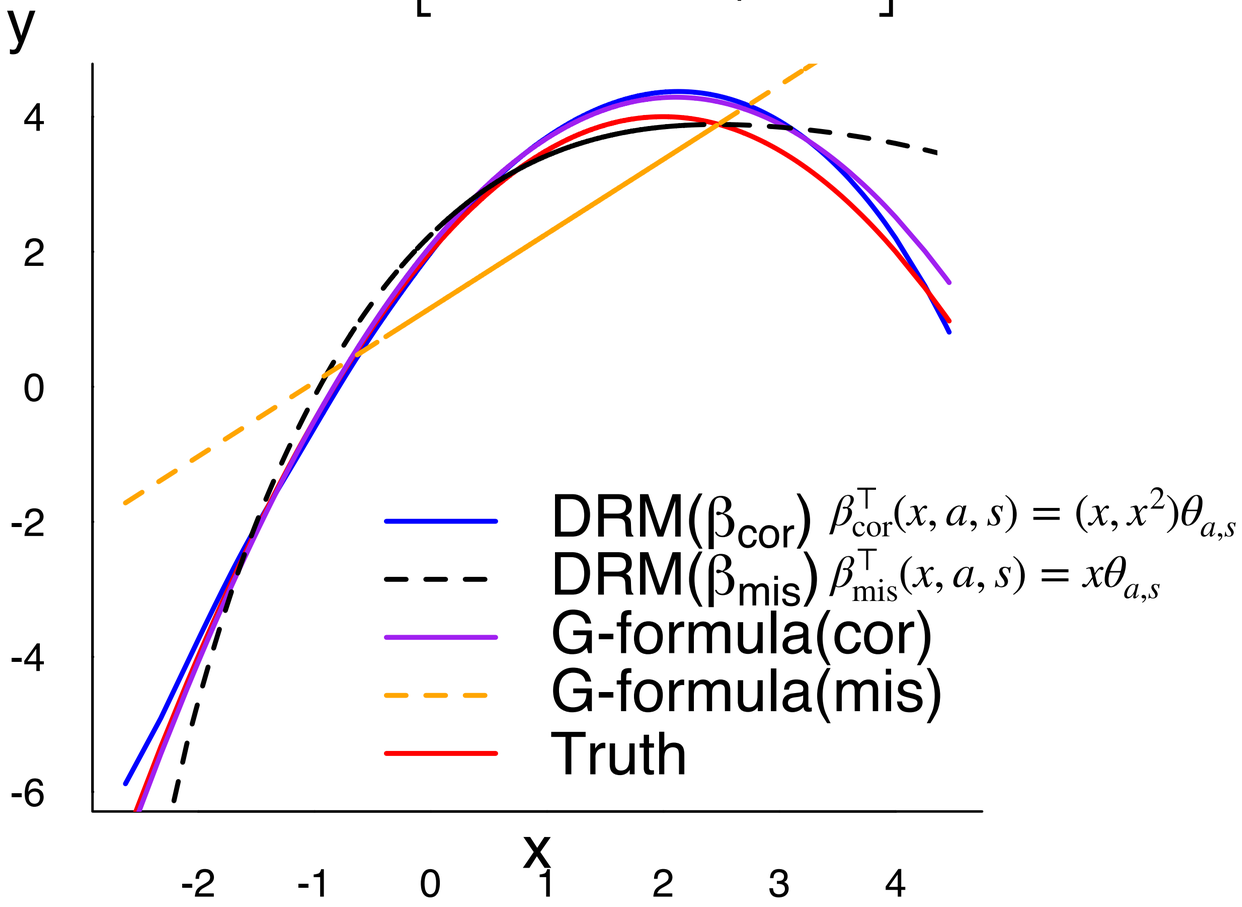
Observational data

- X from RCT & Obs have different distributions: mimics the real-world scenario.
- U is a hidden confounder for Obs.
- Correctly specified DRM: $q^\top(y) = (y, y^2)$ and $\beta_{\text{cor}}^\top(x, a, s) = (x, x^2)\theta_{a,s}$.
- To account for possible model misspecification, we also use $\beta_{\text{mis}}^\top(x, a, s) = x\theta_{a,s}$.
- RCT sample size = 500; Obs sample size = 5000; 1000 simulation repetitions.

Performance of CATE estimator

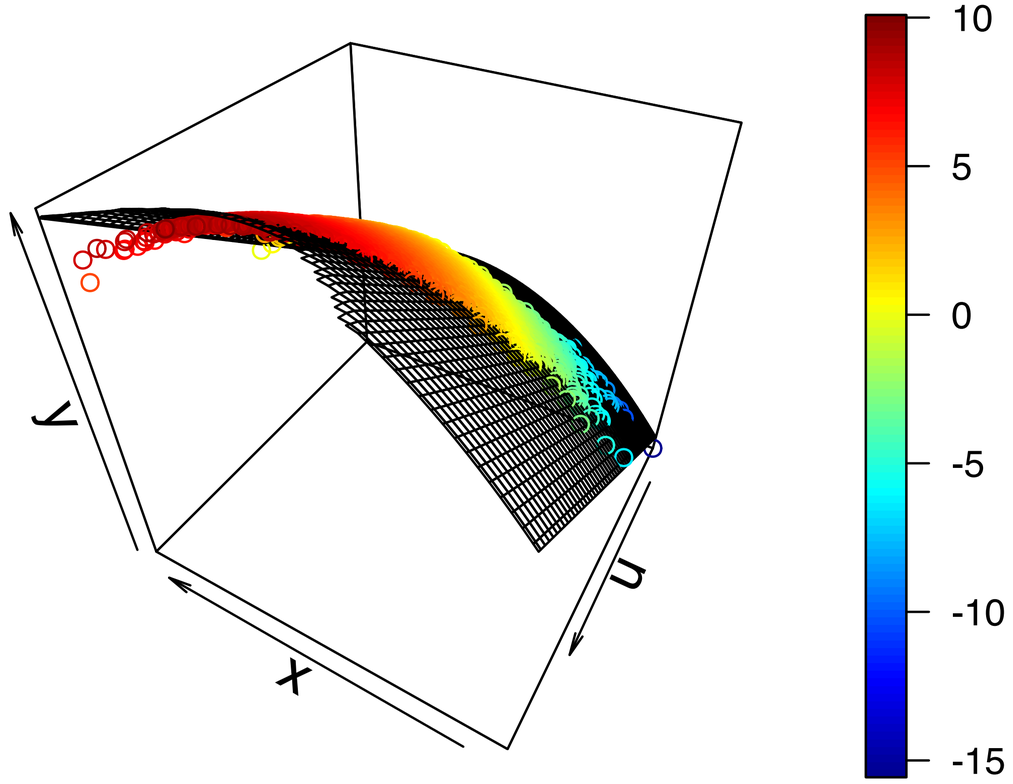
Based on one simulation repetition. All DRM use $q^\top(y) = (y, y^2)$.

$$\mathrm{CATE}: \mathrm{E}\left[\mathrm{Y}(1) - \mathrm{Y}(0) \middle| \mathrm{X} = \mathrm{x}\right]$$



Our proposed method is relatively robust to model misspecification!

$$\mathrm{CATE}: \mathrm{E}\left[\mathrm{Y}(1) - \mathrm{Y}(0) \middle| \mathrm{X} = \mathrm{x}, \mathrm{U} = \mathrm{u}\right]$$



Our proposed method is also applicable to situations of no hidden confounding!

Performance of ATE estimators

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

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

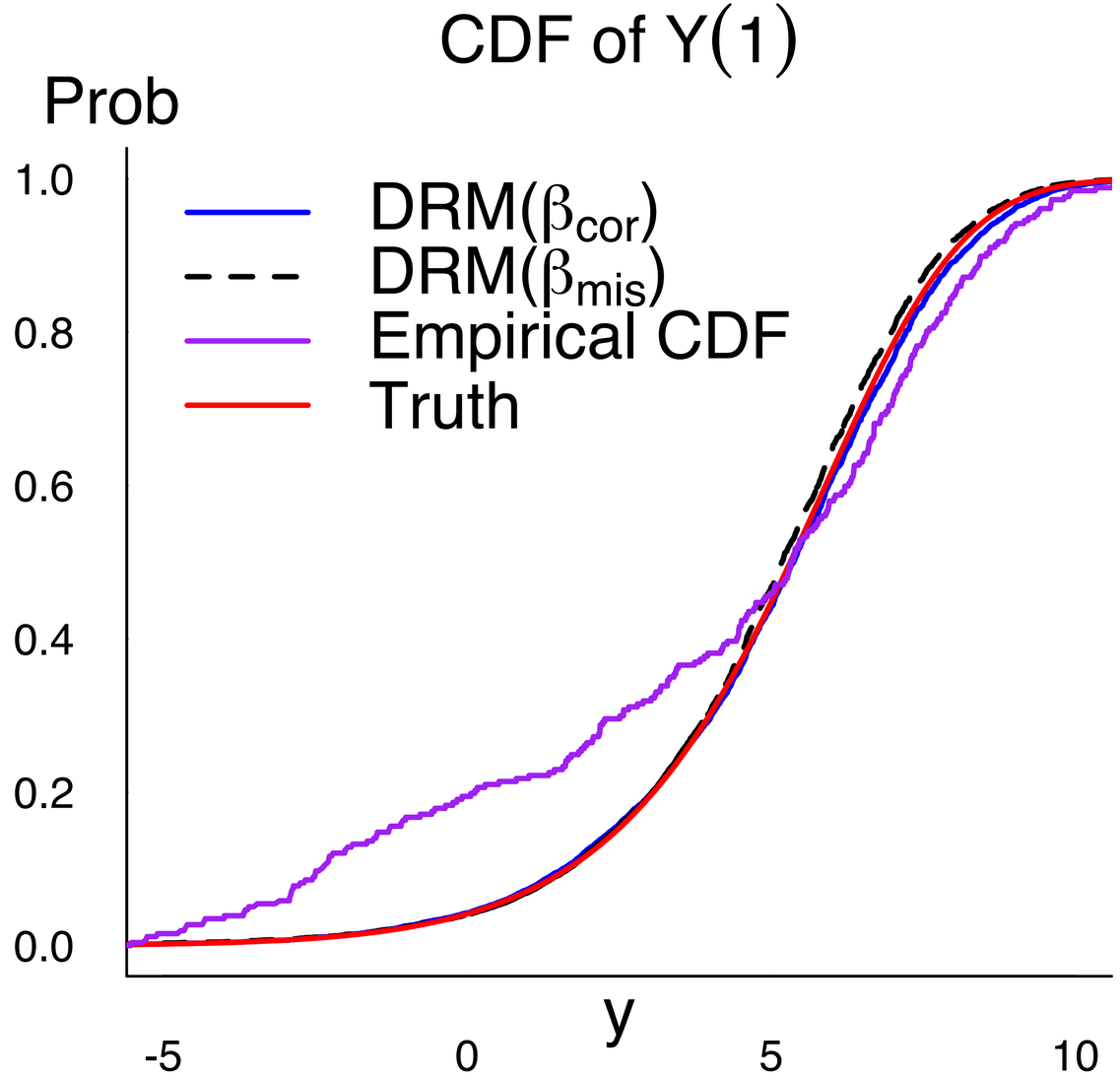
	Abs Bias ($\times 10$)	Var ($\times 100$)	MSE ($\times 100$)
DRM ($\beta_{\text{cor}}(x, a, s)$)	1.143	1.975	1.987
DRM ($\beta_{\text{mis}}(x, a, s)$)	2.132	1.726	6.042
Naive (RCT only)	10.050	8.327	109.221
Naive (Obs only)	10.009	0.832	101.007
AIPW[1] (x_1, x_1^2)	1.159	2.043	2.047
AIPW (x_1)	10.018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

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



$$\beta_{\text{cor}}^\top(x, a, s) = (x, x^2)\theta_{a,s}$$
$$\beta_{\text{mis}}^\top(x, a, s) = x\theta_{a,s}$$

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 within our EL-DRM framework.

Thank you! :-)

Questions & discussions are welcome!