A Semiparametric Approach to Data-Integrated Causal Inference

Archer Gong Zhang Nancy Reid Qiang Sun

Department of Statistical Sciences, University of Toronto

Statistical Sciences
UNIVERSITY OF TORONTO

Data-Integrated Causal Inference

- Goal: estimate the causal effects on a target population.
- Multi-source data: collected from experimental (RCT) and observational studies (Obs).

	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

- Question: how to take advantage of both data with complementary features?
- Example: in 2019, U.S. FDA approved IBRANCE® for the treatment of men with breast cancer.
  - Clinical trials performed for authorization were mainly performed on the female population.
  - The approval was based on data from EHRs and postmarketing reports of the real use of drug in male patients.

Importance of Distribution-Centric Causal Inference

- Many studies focus on mean: 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."

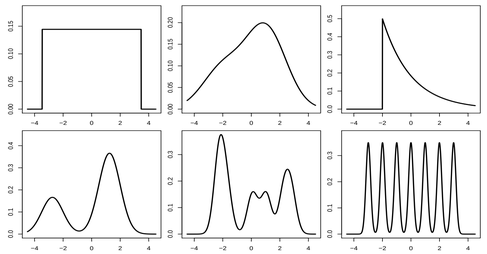


Figure 1. Six distributions that all have the same mean and variance (from Kennedy et al. (2023)).

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

Setup

- Potential outcome (Rubin, 1974): $Y(a)$ with treatment level $a = 0, \ldots, K$.
- Data: $\{(X_i, A_i, Y_i, S_i) : i\}$ with $Y = Y(A)$ is the observed actual outcome and $S_i = \mathbb{1}(i \in \text{RCT})$.
- Goal: estimate 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.
  2. Transportability/Generalizability: $Y(a)|X, S = 1 \overset{d}{=} Y(a)|X$ for all a.

A Semiparametric Approach: Density Ratio Model

- 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 $S = 0$.
- Density ratio model (DRM) (Anderson, 1979): let $G(y|x, a, s)$ be the distribution of $Y|X, A, S$.

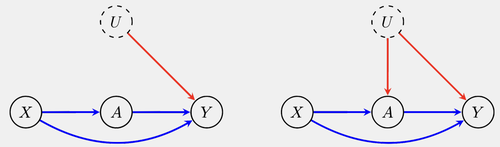
vector-valued function

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

"normalizing constant" baseline distribution

- DRM is flexible as G_0 is unspecified: it can be seen as a generalization of the GLM.
- DRM is interpretable as it provides a structured framework for modelling distribution shift:

Causal models for
- Left: RCT data
- Right: Obs data

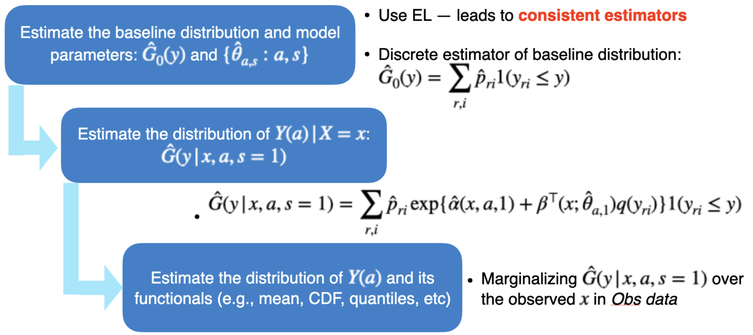


If U is a hidden confounder for Obs and 1) $Y(a) \perp A|X, U$ and 2) $Y(a) \perp S|X, U$. Then,

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

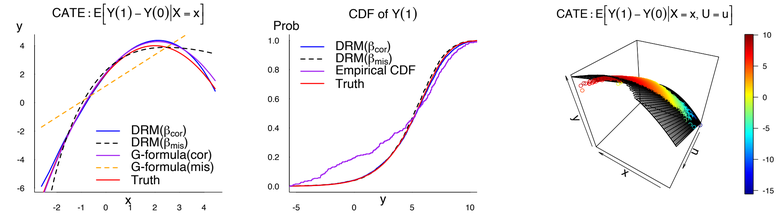
Inference Procedure: Empirical Likelihood

- Empirical likelihood (EL) (Owen, 2001): a nonparametric likelihood-based inference method.
- EL-DRM framework enables utilization of the ***entire data*** to estimate ***each distribution***.



Simulation

- RCT data: $A \sim \text{Bernoulli}(0.5)$, $X \sim \text{Unif}[-2, 4]$, $U \sim N(1, 1)$ (unobserved), $X \perp U$.
  Obs data: $A \sim \text{Bernoulli}(0.5)$, $X \sim N(1, 1)$, $U|X, A \sim N(2AX, 1)$ (unobserved).
  Observed outcome: $Y = 1 + A + X + 2AX - 0.5AX^2 + AU + \varepsilon$, $\varepsilon \sim N(0, 1)$.
- Correctly specified DRM: $q(y) = (y, y^2)^\top$ and $\beta_{\text{cor}}(x; \theta_{a,s}) = (x, x^2)^\top \theta_{a,s}$.
- To consider possible model misspecification, also use $\beta_{\text{mis}}(x; \theta_{a,s}) = x^\top \theta_{a,s}$.
- Estimation results with RCT sample size = 500 and Obs sample size = 5000:

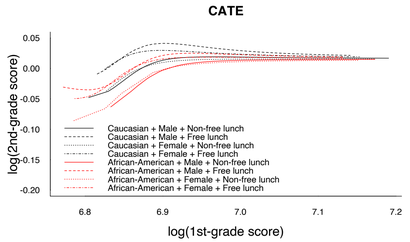


Application to Tennessee Student/Teacher Achievement Ratio (STAR)

- An experiment for the effect of class size on students' test scores (Achilles et al., 2008).
- Treatment A: students' class-type (small and regular) at their second grade.
- Outcome Y: log sum of the students' test scores in reading and math at their second grade.
- Covariates X: 1) log(first-grade score), 2) race, 3) gender, and 4) whether qualified for free lunch.
- Fit the DRM with a data-adaptive $q(y)$ (Zhang and Chen, 2022) and a linear $\beta(x; \theta_{a,s})$.

Generate RCT data:
- Control: sample 20% from the students with first-grade scores in the lower half of the data.
- Treatment: sample 10% from all those treated.

Generate Obs data:
- Control: all the controls not included in the RCT.
- Treatment: all those whose outcomes were in the upper half of the outcomes among the treated students not included in the RCT.



References

C. Achilles, H. P. Bain, F. Bellott, J. Boyd-Zaharias, J. Finn, J. Folger, J. Johnston, and E. Word. Tennessee's Student Teacher Achievement Ratio (STAR) project, 2008. URL https://doi.org/10.7910/DVN/SIWH9F.

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

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

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

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

A. G. Zhang and J. Chen. Density ratio model with data-adaptive basis function. *Journal of Multivariate Analysis*, 191:105043, 2022.