Pathwise Fairness

Sheridan Grant

University of Washington

slgstats@uw.edu

February 16, 2021

What is unfair?

Causal Inference Basics

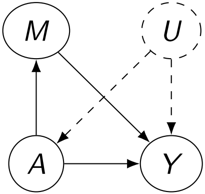

Figure: Mediation with an unobserved confounder

- Arrows represent direct causal effects
- Y is the outcome
- M *mediates* the effect of A on Y
- U is an unobserved *confounder*

Causal Inference Basics

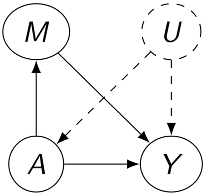

Figure: Mediation with an unobserved confounder

$Y(a)$: the outcome had A been intervened upon to take value a. *A may have taken on value a naturally, anyway.* Let a' denote the "control" level, a the "treatment" (or level of interest). E.g. when assessing racial discrimination, often a' represents white people and a represents Black people.

Causal Inference Basics

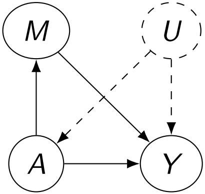

Figure: Mediation with an unobserved confounder

- Average treatment effect (ATE): $\mathrm{E}[Y(a) - Y(a')]$.
- Average treatment effect *on the treated* (ATT): $\mathrm{E}[Y(a) - Y(a')|A = a]$.
- If A is randomized, then ATE $=$ ATT.

Types of causal effects

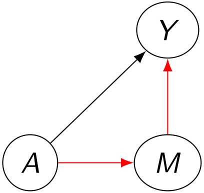

Figure: Mediation with no confounders

This paper is concerned with more interesting/unusual causal effects.

- Ignore issues of "on the treated" for this paper
- Direct effect: $E[Y(a, M(a')) - Y(a')]$
- Indirect (mediation) effect: $E[Y(a) - Y(a, M(a'))]$
- Total effect (ATE or ATT): sum of direct and indirect effects

Types of causal effects

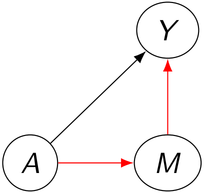

Figure: Mediation with no confounders

- Fit $Y = \beta_0 + \beta_A A + \beta M + \epsilon$; β_A is direct effect
- Fit $Y = \beta'_0 + \beta'_A A + \epsilon$; β'_A is total effect
- $\beta'_A - \beta_A$ is indirect effect

Types of causal effects

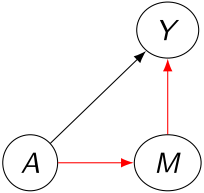

Figure: Mediation with no confounders

- All of this generalizes to complex diagrams, multiple mediators/paths, confounders, etc.
- Modern causal (often semiparametric) inference studies this
- Nabi and Shpitser 2017 points you to a lot of these semiparametric papers

When do associative metrics fail?

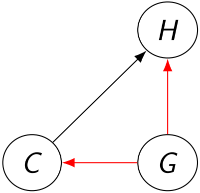

Figure: Prior conviction C, hiring H, and gender G

p(H=1| G,C)	G value	C value	p(C=1 |G)
0.06	1	1	0.99
0.01	0	1	0.01
0.2	1	0	
0.05	0	0	

Figure: Rates of hiring H for different genders G and prior conviction status C.

This distribution actually displays equality of opportunity! (Hardt, Price, and Srebro 2016)

How do we make it fair?

Hypothetically Fair Worlds

Causal models seek to reconstruct a hypothetical world in which the treatment was randomly assigned. Nabi and Shpitser 2017 do this with fairness: estimate a "fair" world that is KL-close to the observed world.

- Assume linearity, standardized variables for now
- "fair": PSE strengths restricted to $[\epsilon_l, \epsilon_u]$
- Divide covariates into X and Z, and condition on the Z covariates—that is, assume they come from a "fair world."
- Estimate parameters of p^* subject to PSE constraints.
- For future predictions: 1) use $\tilde{X}_i \equiv \mathrm{E}^*[X|Z_i]$ in place of X_i, 2) use $p^*(Y_i, \tilde{X}_i, Z_i)$ to make predictions
- Example: BART

COMPAS Results

Use BART (Chipman, George, and McCulloch 2010) as outcome model, but in MCMC reject any step yielding a PSE outside constrained range.

Model	Accuracy	NDE (1 = fair)
Unconstrained	67.8%	1.3
Constrained	66.4%	1.05
Race-unaware	64%	2.1

Table: Accuracies and race NDE for various BART models of COMPAS data.

Challenges for Future Work

- In general, constraining PSEs introduces nonconvex constraints: assuming a linear SEM, a 1-length path needs only convex constraints, but a 2-length path (e.g. $A \rightarrow M \rightarrow Y$) require a nonconvex constraint ($\epsilon_l < \beta_{A\rightarrow M} \cdot \beta_{M\rightarrow Y} < \epsilon_u$). This is clearly a serious problem and one of the main gaps in the paper.
- Choice of X and Z. Authors discuss "tradeoffs" but it appears to me that the more variables in Z the better (judging from the developments in "Fair Inference From Finite Samples," the authors seem to agree).

References I

Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. "BART: Bayesian additive regression trees". EN. In: The Annals of Applied Statistics 4.1 (Mar. 2010), pp. 266–298. ISSN: 1932-6157, 1941-7330. DOI: 10.1214/09-AOAS285. URL: https://projecteuclid.org/euclid.aoas/1273584455 (visited on 03/20/2019).

Moritz Hardt, Eric Price, and Nathan Srebro. "Equality of Opportunity in Supervised Learning". In: arXiv:1610.02413 [cs] (Oct. 2016). arXiv: 1610.02413. URL: http://arxiv.org/abs/1610.02413 (visited on 10/16/2018).

References II

Razieh Nabi and Ilya Shpitser. "Fair Inference On Outcomes". In: arXiv:1705.10378 [stat] (May 2017). arXiv: 1705.10378. URL: http://arxiv.org/abs/1705.10378 (visited on 08/21/2018).