Fair New World

for the seminar: *Fairness in Machine Learning*,
organized by M. Hardt, Fall 2017, UC Berkeley

Nima Hejazi

Division of Biostatistics
University of California, Berkeley
stat.berkeley.edu/~nhejazi

nimahejazi.org
twitter/@nshejazi
github/nhejazi

slides: goo.gl/8RWEy5

This slide deck is for a reading group presentation on the manuscript "Fair Inference on Outcomes" (Rabi & Shpitser, 2017), for the seminar on "Fairness in Machine Learning", organized in Fall 2017 by Moritz Hardt, at the University of California, Berkeley.

Source: https://github.com/nhejazi/talk_fair-outcomes
Slides: https://goo.gl/i3CxL9
With notes: https://goo.gl/8RWEy5

Preview: Summary

- Mediation analysis provides a framework under which intuitive definitions of fairness may be expressed.
- "Fair inference" is analogous to causal inference, except in that the counterfactuals explored refer to a "fair" world (n. b., intentionally vague).
- Fairness may be characterized as the absence (or dampening) of a **path-specific effect (PSE)**.
- Restriction of a PSE is easily expressed as a likelihood maximization problem that features contraining the magnitude of the undesirable PSE.
- This approach to fairness avoids throwing away information (i.e., "fairness through unawareness") but leaves the definition of fairness to the analyst.

We'll go over this summary again at the end of the talk. Hopefully, it will make more sense then.

Preliminaries: Notation

- Data $\mathcal{D} = (Y, \boldsymbol{X})$; outcome Y and feature vector $\boldsymbol{X}$.
- Sensitive features: $S \in \boldsymbol{X}$, where inference on Y using S *might* result in discrimination.
- Treatment variable: $A \in \boldsymbol{X}$.
- Mediator variables: $M \in \boldsymbol{X}$ or $\boldsymbol{M} \subseteq \boldsymbol{X}$.
- Potential outcome: $Y(a)$, realization of Y under $A = a$.

Preliminaries: Mediation Analysis

- **Goal:** understand the mechanism by which A influences Y.
- Decompose the **ACE** into *direct* and *indirect* effects mediated by a variable M.
- Partition feature space $\boldsymbol{X}$ into A (treatment), M (mediator), and $\mathbf{C} = \mathbf{X} \setminus \{A, M\}$ (baseline factors).
- Counterfactual contrasts are expressed via *nested* potential outcomes (i.e., $Y(a, M(a'))$).

- Nested potential outcomes read as "the outcome Y if A were set to a while M were set to whatever value it would have attained had A been set to a'".

The Average Causal Effect (ACE)

- ACE $= \mathbb{E}[Y(a)] - \mathbb{E}[Y(a')]$
- Not computed via $\mathbb{E}[Y \mid A]$, as associations between A and Y may be "partly causal" or spurious.
- Decomposition: ACE $=$ NDE $+$ NIE, where **NDE** is the *Natural Direct Effect* and **NIE** is the *Natural Indirect Effect*.

$$\begin{aligned}\text{ACE} &= \mathbb{E}[Y(a)] - \mathbb{E}[Y(a')] \\ &= \mathbb{E}[Y(a)] - \mathbb{E}[Y(a, M(a')] \\ &+ \mathbb{E}[(Y(a, M(a')] - \mathbb{E}[Y(a')]\end{aligned}$$

- Decomposition of the ACE gives us a way to express undesirable PSEs using mediators
- The decomposition is just by way of a telescoping sum arguemnt.

The Natural *Direct* Effect (NDE)

- Comparison of the mean outcome under only the part of the treatment that directly affects it ($A = a$) and the placebo treatment (i.e., $A = a'$).
- Note that the *indirect* effect of the treatment (through the mediator M) is "turned off" (i.e., $M(A = a')$).

Definition

Natural **Direct** Effect

$$\text{NDE} = \mathbb{E}[(Y(a, M(a')] - \mathbb{E}[Y(a')]$$

With additional causal assumptions, the NDE is identified as the

Definition

Mediation formula

$$\sum_{\boldsymbol{C},M} (\mathbb{E}[Y \mid a, M, \boldsymbol{C}] - \mathbb{E}[Y \mid a', M, \boldsymbol{C}])p(M \mid a', \boldsymbol{C})p(\boldsymbol{C})$$

- Estimation may be performed using plug-in estimators.

The Natural *Indirect* Effect (NIE)

- Comparison of the outcome affected by all treatment (both direct and indirect) and the outcome where the effect through the mediator (M) is "turned off" (i.e., $M(A = a')$).
- Although in a roundabout manner, this quantity gets at the effect of the path-specific effect through the mediator on the outcome.

Definition

Natural **Indirect** Effect

$$\text{NIE} = \mathbb{E}[Y(a)] - \mathbb{E}[(Y(a, M(a')]$$

Example: *Thank You for Smoking*

For a better intuition of $Y(a, M(a'))$, consider the following:

- Let Y be a health outcome (e.g., survival probability), A be a treatment (e.g., smoking).
- Consider a decomposition of the effect of A on Y — that is, let M be a mediator (e.g., cancer).
- A affects Y directly (nicotine exposure) and indirectly (inducing lung cancer, through M).
- Here, $Y(a, M(a'))$ correponds to "the response of Y to an intervention that sets the nicotine exposure (direct effect) to what it would be in smokers, and the smoke exposure (indirect effect) to what it would be in non-smokers" (e.g., nicotine patch).

Path-Specific Effects

- A more general idea than the NDE and NIE — such effects are easily formulated as nested counterfactuals.
- *Intuition*: along a path of interest, all nodes behave as if the active rule were imposed (i.e., $A = a$) while, along all other paths, nodes behave as though the alternative were the case (i.e., $A = a'$).

Definition

Path-Specific Effect (PSE)

(Along a path, say $A \to W \to Y$)

$$\mathbb{E}[Y(a', W(M(a'), a), M(a'))] - \mathbb{E}[Y(a')]$$

With more complex assumptions, we get the

Definition

Edge G-formula

$$\sum_{\boldsymbol{C},M,W} \mathbb{E}[Y \mid a', W, M, \boldsymbol{C}]p(W \mid a, M, \boldsymbol{C})p(M \mid a', \boldsymbol{C})p(\boldsymbol{C})$$

- Estimation may be performed using plug-in estimators.

Finding Fairness

- Much work has focused on defining fairness via associative relationships (including equalized odds). Such criteria provided unintuive results when the sensitive feature is not randomly assigned.
- Here, an approach that ought to provide intuitive results (wrt fairness), even when the sensitive attribute is associated with the outcome (perhaps by way of an unobserved feature), is proposed.
- Associative fairness metrics fail to properly model sources of confounding (between S and Y).
- Generally, this failure is rooted in the fact that "counterfactual probabilities are complex functions of the observed data, no just conditional densities."

To see the problem with associative fairness criteria, consider the following example (see section 4, paragraph 3 of paper for details):

- Letting $p(H \mid C, G)$ be a hiring rule, we would assess fairness as follows: $p(H = 1 \mid C = 1) = p(H = 1 \mid C = 0) \approx 0.022$, using associative fairness criteria.
- Intuitively, fairness of a hiring rule would lead to equal hiring probabilities for cases and controls in a hypothetical randomized trial (RCT) — i.e., $p(H(C = 1)) = p(H(C = 0))$.
- In our example, application of the adjustment formula would yield $p(H(C = 1)) = 0.025$ and $p(H(C = 0)) = 0.25$, a rather striking difference.
- The difference between $p(H(C = 0))$ and $p(H \mid C = 0)$ is driven by extreme values of $p(C \mid G)$.
- This is basically Simpson's paradox, or rather close to it.

In Pursuit of "Fair Inference"

- Fairness is, at its core, rooted in counterfactuals. Thus, we can see *"fair inference"* as a branch of causal inference wherein the counterfactuals to be considered are with respect to a "fair" world.
- *Discrimination* may be expressed as the presence of a particular PSE, with choice of the specific PSE left as a domain-specific issue.
- Thus, minimization of specific PSEs corresponds to minimizing discrimination and is a problem of constrained inference on statistical models.

From the legal literature:
"The central question in any employment-discrimination case is whether the employer would have taken the same action had the employee been of a different race (age, sex, religion, national origin, etc.) and everything else had been the same."

Fairness as PSE Minimization

- Let $p(Y, \mathbf{X})$ be a statistical model, assumed to be induced by a *causal model*.
- Discrimination (wrt Y based on $S \in \boldsymbol{X}$) in this model is a PSE, identified as the functional $f(p(Y, \boldsymbol{X}))$.
- Let (ϵ_l, ϵ_u) be lower and upper bounds on the PSE, giving the degree of unfairness considered tolerable (n.b., the PSE is removed in the special case $\epsilon_l = \epsilon_u$).
- **Proposal**: transform $p(Y, \mathbf{X})$ into $p^*(Y, \boldsymbol{X})$ under the constraint that the PSE of interest lies within (ϵ_l, ϵ_u), where the two distributions are close in the sense of KL-divergence.

Definition

Kullback-Leibler Divergence

$$D(P \parallel Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx$$

- Why do we restrict ourselves to KL-divergence?
- Well, as we'll see shortly, our approach amounts to a constrained likelihood maximization problem. This makes KL-divergence a natural choice.
- Were we to consider formulating our fairness metric differently, we could use other distance metrics (e.g., Hellinger distance, total variation distance).

Finding Fair Worlds I

- **Proposal**: We can make *any* function of p ***fair***, merely by computing it from p^* (instead of from p).
- To ensure fairness, we must make inference only in the "fair world", just as we only perform inference on counterfactuals in causal inference.
- To do this, map any x^i from p to a sensible version of it drawn from p^* — i.e., find a $g : x^i_p \mapsto x^i_{p^*}$.
- I want to be fair, so what exactly do I do?

- The intuition here is simply that p^* includes all the same information that was found in p, except that information that leads to discrimination (is used in the unfavorable PSE).
- We'll shortly see what it means to construct a fair world p^*.

Finding Fair Worlds II

- Consider the following general setup:
  - finite samples $\mathcal{D}$ drawn from $p(Y, \boldsymbol{X})$
  - a likelihood function $\mathcal{L}_{Y,\boldsymbol{X}}(\mathcal{D}; \alpha)$
  - a discriminative PSE $f(p(Y, \mathbf{X}))$ with bounds (ϵ_l, ϵ_u)
  - an estimator of the PSE $g(\mathcal{D})$.
- We obtain fairness by solving:

$$\hat{\alpha} = \arg\max_{\alpha} \mathcal{L}_{Y,\boldsymbol{X}}(\mathcal{D}; \alpha),$$

  subject to $\epsilon_l \leq g(\mathcal{D}) \leq \epsilon_u$.
- In this setup, fairness is achieved by constraining parts of $p(Y, \boldsymbol{X}, ; \alpha)$, with the choice of g determining exactly what is constrained.

- Note that knowing p and p^* exactly would make the problem rather easy, since we would merely use $x^i_{\boldsymbol{W}}$, where $\boldsymbol{W}$ is simply the largest subset of $\boldsymbol{X}$ that contrains the PSE in a desirable manner (i.e., where $p(\boldsymbol{W}) = p^*(\boldsymbol{W})$.
- Further, it's important to consider "large" statistical models — perhaps infinite-dimensional models? — in order to ensure that prediction (and inference, if so desired) can be optimized with respect to out-of-sample observations.
- In regard to both prediction and inference, it turns out that most estimators that rely on the outcome (Y) model are not robust to its misspecification. Thus, it is important to either use simpler (IPW) estimators or more modern approaches (e.g., triply robust estimators).

Fairness is (Partial?) (Un)Awareness

- Since using all of the information contained in p leads to unfairness, this approach amounts to discarding information that is exclusively in p, relative to p^*.
- The goal of this approach is to use the available information as well as possible, but only in so far as our inferences are drawn from the "fair world."
- In this approach, fairness is characterized as the *a priori* inadmissability of certain paths in the DAG of interest — that is, paths other than a single edge path might cause discrimination.

- Consider the experiment of the authors using the "adult" dataset: here, there is concern about the direct effect of gender on income class as well as the effect of gender on income class through marital status. Using the proposed constrained optimization approach, they are able to avoid dropping sensitive variables from the analysis while improving upon the efficacy of "blind" models.
- Specifically, they estimate the PSE in the unconstrained model to be 3.16, and then restric the PSE to lie within $(0.95, 1.00)$ in the "fair" (constrained) model. The "fair" model achives accuracy of 72% while the "blind" model only achieves 42% (the unconstrained model does better, with an accuracy of 82%).

Review: Summary

- Mediation analysis provides a framework under which intuitive definitions of fairness may be expressed.
- "Fair inference" is analogous to causal inference, except in that the counterfactuals explored refer to a "fair" world (n. b., intentionally vague).
- Fairness may be characterized as the absence (or dampening) of a **path-specific effect (PSE)**.
- Restriction of a PSE is easily expressed as a likelihood maximization problem that features contraining the magnitude of the undesirable PSE.
- This approach to fairness avoids throwing away information (i.e., "fairness through unawareness") but leaves the definition of fairness to the analyst.

It's always good to include a summary. You've seen this all before.

References I

Hardt, M., Price, E., Srebro, N., et al. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323.

Miles, C. H., Kanki, P., Meloni, S., and Tchetgen, E. J. T. (2015). On partial identification of the pure direct effect. *arXiv preprint arXiv:1509.01652*.

Nabi, R. and Shpitser, I. (2017). Fair Inference On Outcomes. *ArXiv e-prints*.

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*, pages 411–420. Morgan Kaufmann Publishers Inc.

References II

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Robins, J. M. (2000). Marginal structural models versus structural nested models as tools for causal inference. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 95–133. Springer.

Tchetgen, E. J. T. and Shpitser, I. (2012). Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of statistics*, 40(3):1816.

Thank you.

Slides: goo.gl/i3CxL9

Notes: goo.gl/8RWEy5

Source (repo): goo.gl/qJSoz6

`stat.berkeley.edu/~nhejazi`

`nimahejazi.org`

`twitter/@nshejazi`

`github/nhejazi`

Here's where you can find me, as well as the materials from this talk.