Identification of Causal Effect in the Presence of Selection Bias

Juan D. Correa  Jin Tian  Elias Bareinboim

AAAI
Honolulu, 2019
Challenge 1: Confounding Bias

What’s the causal effect of Exercise on Cholesterol?
What about $P(\text{cholesterol} \mid \text{exercise})$?
Challenge 1: Confounding Bias

- Age 10
- Age 20
- Age 30
- Age 40
- Age 50
Challenge 1: Confounding Bias

This difference is called **Confounding Bias**
Challenge 2: Selection Bias

Variables in the system affect the inclusion of units in the sample.

- **Exercise**
- **Cholesterol**
- **Fitness**
- **S**

![Graph showing the relationship between exercise hours, cholesterol levels, and fitness with selection bias indicated by S=0 and S=1.]
Challenge 2: Selection Bias

Variables in the system affect the inclusion of units in the sample

\[ P(\text{age}, \text{ex}, \text{ch}, \text{fit}) \neq P(\text{age}, \text{ex}, \text{ch}, \text{fit} \mid S = 1) \]

This difference is due to Selection Bias
## Current literature

<table>
<thead>
<tr>
<th>No Confounding</th>
<th>Confounding</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Association = Causation</strong>&lt;br&gt;No control</td>
<td><strong>Complete Algorithms</strong>&lt;br&gt;[Tian and Pearl ’02; Huang and Valtorta ’06; Shpitser and Pearl ’06; Bareinboim and Pearl ’12]</td>
</tr>
<tr>
<td><strong>Controlling Selection Bias</strong>&lt;br&gt;[Bareinboim and Pearl ’12]&lt;br&gt;Recovering from Selection Bias in Causal and Statistical Inference&lt;br&gt;[Bareinboim, Tian, Pearl ’14]</td>
<td><strong>RCE</strong>&lt;br&gt;[Bareinboim, Tian, Pearl ’15]&lt;br&gt;<strong>Generalized Adjustment</strong>&lt;br&gt;[Correa, Tian, Bareinboim ’18]&lt;br&gt;<strong>IDSB</strong>&lt;br&gt;[Correa, Tian, Bareinboim ’19]</td>
</tr>
</tbody>
</table>
Problem I

Given:

Is there a function $f$ such that

$$P(y|do(x)) = f(P_1)$$
Result 1

Theorem 1:
Let $X, Y \subset V$ be two disjoint sets of variables and $G$ a causal diagram over $V$ and $S$. If $(Y \perp S)_{G_{XY}}^{pbd}$, then $P_x(y)$ is not recoverable from $P(v \mid S = 1)$ in $G$. 
Problem II

Given:

Is there a function $f$ such that

$$P(y|do(x)) = f(P_1, P_2)$$
Result II

Algorithm **IDSB**

Given a causal diagram, a selection-biased distribution and external data over a subset of the variables and the variables of interest \((X, Y)\); returns an expression for \(P_x(y)\) in terms of the input or failure.

Strictly more powerful than the best known algorithm that accepts both biased and unbiased data.
Decomposing the Problem

Intervention

\[ P_x(y) = \sum_{w_1, w_2, w_3} P_x(y, w_3, w_2, w_1) \]
Decomposing the Problem

C-Components

\[ P_x(y) = \sum_{w_1,w_2,w_3} P_x(y, w_3, w_2, w_1) = \sum_{w_1,w_2,w_3} P_{x,w_1,w_3}(y, w_2) P_{w_2,y}(w_3, w_1) \]
Summary

1. Complete characterization recoverable causal effects from the causal diagram and a selection-biased probability distribution.

2. Sufficient procedure to recover causal effects from a causal diagram, selection-biased distributions and auxiliary unbiased data which is strictly more powerful than state-of-the-art procedure.

Thanks!