MAS 420: Causal Inference Overview

Adam Rohde
Department of Statistics
University of California, Los Angeles

November 29, 2022

Recommended resources

Useful resources

1. Causality - Pearl (2009)
2. Adjustment Criterion - Shpitser, VanderWeele, Robins (2012) [link]

Example

Suppose we have a binary treatment (D), some outcome (Y), and some observed covariates (X). We might hope that we can use a "no unobserved confounding" argument after conditioning on X. So we might hope that we have a DAG something like this:

$$\begin{tikzcd} & X \arrow[dl] \arrow[dr] & \\ D \arrow[rr] & & Y \end{tikzcd}$$

Assuming this DAG captures the true data generating process (there are no unobserved confounders or anything else), why does conditioning on X allow us to identify causal effects?

Adjustment criterion

You might cite the adjustment (or back-door) criterion. And you'd be right! But why?

Definition (Adjustment criterion)

X satisfies the adjustment criterion relative to D and Y in a causal graph G if

1. No element in X lies on or is a descendant of a node on a causal path from D to Y.[a]
2. All non-causal paths from D to Y are blocked by X.

[a] Elements of X can be descendants of D if they are not on causal paths from D to Y.

Theorem

Assume the adjustment criterion holds for X and (D, Y) in G. Then $Y_d \perp\!\!\!\perp D|X$.[a]

[a] For every model inducing G.

Recall that $Y_d \perp\!\!\!\perp D|X$ is called **conditional ignorability**.

Identification

Recall the "adjustment formula" for the ATE where we use conditional ignorability. This is an identification result.[1] We **identify** an expression containing potential outcomes (i.e., the ATE) with an expression that does not contain potential outcomes.

$$
\begin{aligned}
\text{ATE} &= \mathbb{E}[Y_{1i} - Y_{0i}] = \mathbb{E}[Y_{1i}] - \mathbb{E}[Y_{0i}] \\
&= \sum_x \left(\mathbb{E}[Y_{1i}|X_i = x] - \mathbb{E}[Y_{0i}|X_i = x]\right) P(X_i = x) \\
&= \sum_x \left(\mathbb{E}[Y_{1i}|D_i = 1, X_i = x] - \mathbb{E}[Y_{0i}|D_i = 0, X_i = x]\right) P(X_i = x) \text{ by } Y_{di} \perp\!\!\!\perp D_i|X_i \\
&= \sum_x \left(\mathbb{E}[Y_i|D_i = 1, X_i = x] - \mathbb{E}[Y_i|D_i = 0, X_i = x]\right) P(X_i = x) \text{ by consistency}
\end{aligned}
$$

This seems nice but how do we connect $Y_{di} \perp\!\!\!\perp D_i|X_i$ with what we see in the DAG? That is, why does the adjustment criterion work?

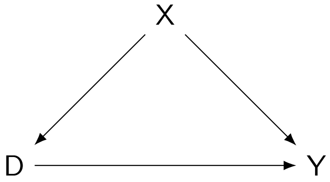
[1] Other identification results, like instrumental variables, use other assumptions in addition to assumptions like conditional ignorability.

Structrual Causal Models

DAGs are graphical representations of structural causal models.
The structural causal model for our DAG is

$$
\begin{aligned}
X_i &= f_X(U_{Xi}) \\
D_i &= f_D(X_i, U_{Di}) \\
Y_i &= f_Y(D_i, X_i, U_{Yi})
\end{aligned}
$$

This is a non-parametric (meaning no assumption on the distributions or relationships) representation of how the variables relate as functions of one another. It is easy to see that the DAG captures the same relationships (recall, we often do not draw the Us).



Sub-Models

Potential outcomes are the values for the outcome that a unit would have had, if their treatment had be $D_i = d$. So POs are actually solutions to the SCM, where we alter the equation for the treatment to take a particular value of interest. These altered SCMs are called **sub-models**. The sub-model in which we, perhaps counterfactually, set the value of the treatment to be $D_i = d$.

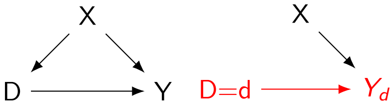
$$\begin{array}{ll} \text{SCM} & \text{Sub-Model} \\ X_i = f_X(U_{Xi}) & X_i = f_X(U_{Xi}) \\ D_i = f_D(X_i, U_{Di}) & D_i = d \\ Y_i = f_Y(D_i, X_i, U_{Yi}) & Y_{di} = f_Y(d, X_i, U_{Yi}) \end{array}$$

In the sub-model, Y_i becomes Y_{di}, the potential outcome, because we have intervened to set the value of D_i to be d. This means that the value of D_i that Y_i is listening to is now d and so this is the value that appears in $f_Y(d, X_i, U_{Yi})$.

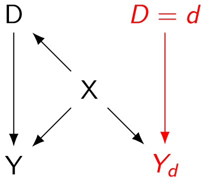
The SCM and the sub-model represent to alternative "worlds." One in which D_i is allowed to take on the value it would naturally and one in which we set $D_i = d$.

Twin Networks

It is possible to draw a graph for both the SCM and the sub-model. Importantly, since we are setting $D = d$, D no longer listens to any other variables in the sub-model.



X is the same as in the SCM and in the sub-model; so we combine the graphs into a single graph that shows us both the "pre-intervention" world and the "post-intervention" world.

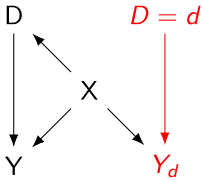


This is called a **twin-network**.

Twin Networks

In the twin network and in the SCM + sub-model, we can see how D relates to Y_d. They both listen to X. We easily see that they are d-separated (all paths are blocked between them) and are equivalently independent, when we condition on X.
Therefore, $Y_d \not\perp\!\!\!\perp D$ but $Y_d \perp\!\!\!\perp D|X$.

SCM	Sub-Model
$X_i = f_X(U_{Xi})$	$X_i = f_X(U_{Xi})$
$D_i = f_D(X_i, U_{Di})$	$D_i = d$
$Y_i = f_Y(D_i, X_i, U_{Yi})$	$Y_{di} = f_Y(d, X_i, U_{Yi})$



Summary

- We've seen how conditional ignorability (and possibly other assumptions) can provide us with identification results that express causal effects in terms of quantities that can be estimated from the data.
- We've also seen how conditional ignorability can be shown on a twin network.
- Twin networks (and DAGs) are graphical representations of SCMs (and sub-models).
- These points show us why and how we can use DAGs to justify assumptions of conditional ignorability. But you still need to defend your DAG.

We can then actually estimate the quantities from our identification result.

$$\text{ATE} = \sum_{x} \left(\mathbb{E}[Y_i | D_i = 1, X_i = x] - \mathbb{E}[Y_i | D_i = 0, X_i = x] \right) P(X_i = x)$$

Matching, IPW, propensity scores, regression, or more sophisticated approaches cab be used for estimation, depending on whether our Xs are discrete or continuous and how many Xs we have. We will also want to quantify the statistical uncertainty in our estimates and explore how sensitive our results are to unobserved confounders etc.

Any remaining time

questions / break