EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Causal Machine Learning

Causal Inference basis

Michael Knaus

WiSe 25/26

I ask you for some more patience

Without causal inference, the outputs of causal ML methods are just numbers

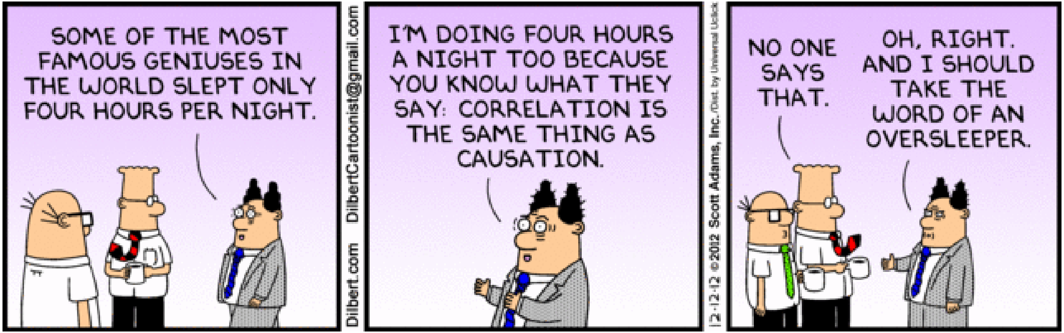
Plan for today

Crash course/recap in causal inference

1. Correlation vs. causation
2. We need a framework ... or two
3. Potential outcomes
4. Structural causal models and directed acyclic graphs
5. Finding (conditional) (in)dependencies
6. Single World Intervention Graphs
7. Identification in RCTs
8. Valid adjustment sets
9. Wrapping up

Correlation vs. causation

We are the guys on the left


SOME OF THE MOST FAMOUS GENIUSES IN THE WORLD SLEPT ONLY FOUR HOURS PER NIGHT.
I'M DOING FOUR HOURS A NIGHT TOO BECAUSE YOU KNOW WHAT THEY SAY: CORRELATION IS THE SAME THING AS CAUSATION.
NO ONE SAYS THAT.
OH, RIGHT. AND I SHOULD TAKE THE WORD OF AN OVERSLEEPER.

Dilbert.com DilbertCartoonist@gmail.com

12-12-12 ©2012 Scott Adams, Inc./Dist. by Universal Uclick

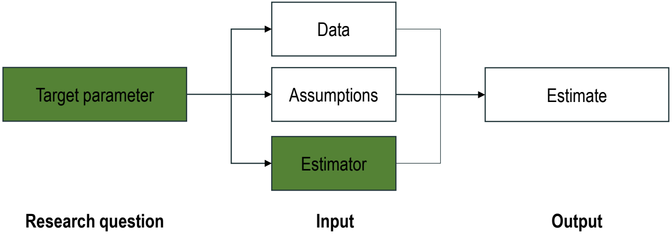
What's the deal?

Intuition by picture

Source

Recall from first week

I told you we will focus on extending the menu of parameters and estimators



BUT, we still need to understand the assumptions that are required for credible parameter estimates such that we can critically assess them in analyses

Only pushing buttons to produce numbers and stories around these numbers can be misleading

We need a framework … or two

Frameworks

There are two prominent frameworks for causal inference and a bridge between them:

- Potential Outcomes (POs): Prevalent in economics and statistics
- Structural Causal Models (SCM) often represented as Directed Acyclic Graphs (DAGs): Prevalent in computer science and industry
- Single World Intervention Graphs (SWIGs): A very useful bridge between the two

This slide deck introduces all three in a condensed form

For the deep dive, please see the slides of the Causal Inference course from last semester provided on ILIAS and the references therein

Causal inference pipeline on a high level

Regardless of the framework, three steps lead to an estimate of a causal effect:

1. **Definition:** Define the target parameter
2. **Identification:** Impose assumptions to express target parameter in terms of observable statistical quantities
3. **Estimation:** Select and apply a suitable estimator to estimate target parameter in a sample of the observable distribution

The general mantra is that **identification always comes before estimation**

This holds in general, but is important to emphasize in this course to avoid tempting traps like ”I use a causal forest, so I am estimating a causal effect”

Potential outcomes

Potential outcomes (dream) world

- Binary treatment: $W \in \{0, 1\}$
- Potential outcome under treatment w: $Y(w)$
- Individual Treatment Effect (ITE): $\Delta = Y(1) - Y(0)$

i	$Y_i(1)$	$Y_i(0)$	Δ_i
1	0	1	-1
2	3	2	1
3	1	1	0
4	2	1	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$

$\Rightarrow$ Each individual has as many POs as there are treatment states (two in our case)

Target parameters

Potential outcomes enable us to define parameter classics as (un-)conditional expectations of (differences) of POs

- Average potential outcome (APO): $\gamma_w := \mathbb{E}[Y(w)]$
  - What is the expected outcome if everybody receives treatment w?
- Average Treatment Effect (ATE): $\tau_{ATE} := \mathbb{E}[Y(1) - Y(0)] = \gamma_1 - \gamma_0$
  - What is the expected treatment effect in the population?
- Average Treatment Effect on the Treated (ATT): $\tau_{ATT} := \mathbb{E}[Y(1) - Y(0) \mid W = 1]$
  - What is the expected treatment effect in the subpopulation actually receiving the treatment?
- Conditional Average Treatment Effect (CATE): $\tau(x) := \mathbb{E}[Y(1) - Y(0) \mid X = x]$
  - What is the expected treatment effect for somebody with characteristics x?

Reality check

The parameters are defined with respect to hypothetical potential outcome distributions that exist in our imagination

In reality, we only observe Y ⇒ no (w) ⇒ no potential outcome (yet)

So, how can the observed outcomes help us to learn something about the unobservable causal effects? 🤔

The consistency assumption links potential outcomes and observable world

Under consistency, we can write

$$Y = Y(W) \text{ or } Y = (1 - W)Y(0) + WY(1) \tag{1}$$

⇒ We observe at least one potential outcome 🥳

⇒ Offers a glimpse into the dream world

(Hard) Reality

Only one potential outcome is observable:

	Partly observed		Unobserved	Observed	
i	$Y_i(1)$	$Y_i(0)$	Δ_i	W_i	Y_i
1	0	1	-1	0	1
2	3	2	1	1	3
3	1	1	0	0	1
4	2	1	1	1	2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

$\Rightarrow$ The counterfactual potential outcome is missing $\Rightarrow$ ITE is never observed $\Rightarrow$ ”fundamental problem of causal inference” (Holland, 1986)

Debunk (Naive) Group Comparisons

POs allow to unpack why comparing means of treated ($W = 1$) and control ($W = 0$) group may not provide a causal effect (correlation vs. causation reloaded/formal)

First see how consistency allows us to link observed and hypothetical distributions

$$
\begin{aligned}
&\mathbb{E}[Y \mid W = 1] - \mathbb{E}[Y \mid W = 0] \\
&\overset{(1)}{=} \mathbb{E}[\overbrace{(1 - W)}^{=0} Y(0) + \overbrace{W}^{=1} Y(1) \mid W = 1] - \mathbb{E}[\overbrace{(1 - W)}^{=1} Y(0) + \overbrace{W}^{=0} Y(1) \mid W = 0] \\
&= \mathbb{E}[Y(1) \mid W = 1] - \mathbb{E}[Y(0) \mid W = 0]
\end{aligned}
$$

Debunk (Naive) Group Comparisons

Without further assumptions, we can decompose the mean comparison as

$$
\begin{aligned}
&\mathbb{E}[Y \mid W = 1] - \mathbb{E}[Y \mid W = 0] \\
&\quad = \mathbb{E}[Y(1) \mid W = 1] - \mathbb{E}[Y(0) \mid W = 0] \\
&\quad = \mathbb{E}[Y(1) \mid W = 1] - \mathbb{E}[Y(0) \mid W = 0] \pm \mathbb{E}[Y(0) \mid W = 1] \\
&\quad = \underbrace{\overbrace{\mathbb{E}[Y(1) \mid W = 1]}^{\textit{factual}} - \overbrace{\mathbb{E}[Y(0) \mid W = 1]}^{\textit{counterfactual}}}_{\text{ATT}} + \underbrace{\overbrace{\mathbb{E}[Y(0) \mid W = 1]}^{\textit{counterfactual}} - \overbrace{\mathbb{E}[Y(0) \mid W = 0]}^{\textit{factual}}}_{\text{confounding bias}}
\end{aligned}
$$

$\Rightarrow$ The pure causal effect is contaminated if the expected potential outcomes under the control condition are different between treated and control subgroups

Debunk (Naive) Group Comparisons

The mean comparison can alternatively be decomposed as

$$
\begin{aligned}
&\mathbb{E}[Y \mid W = 1] - \mathbb{E}[Y \mid W = 0] \\
&\quad = \mathbb{E}[Y(1) \mid W = 1] - \mathbb{E}[Y(0) \mid W = 0] \\
&\quad = \underbrace{\mathbb{E}[Y(1) - Y(0)]}_{\text{ATE}} \\
&\qquad + \underbrace{\mathbb{E}[Y(0) \mid W = 1] - \mathbb{E}[Y(0) \mid W = 0]}_{\text{confounding bias}} \\
&\qquad + \underbrace{(\mathbb{E}[Y(1) - Y(0) \mid W = 1] - \mathbb{E}[Y(1) - Y(0)])}_{\text{heterogeneous effect bias}}
\end{aligned}
$$

How?

$$\mathbb{E}[Y \mid W = 1] - \mathbb{E}[Y \mid W = 0] \overset{(1)}{=} \mathbb{E}[\underbrace{(1 - W)}_{=0} Y(0) + \underbrace{W}_{=1} Y(1) \mid W = 1] - \mathbb{E}[\underbrace{(1 - W)}_{=1} Y(0) + \underbrace{W}_{=0} Y(1) \mid W = 0]$$

$$= \mathbb{E}[Y(1) \mid W = 1] - \mathbb{E}[Y(0) \mid W = 0] + \tau_{ATE} - \tau_{ATE} + \mathbb{E}[Y(0) \mid W = 1] - \mathbb{E}[Y(0) \mid W = 1]$$

$$= \tau_{ATE} + \mathbb{E}[Y(0) \mid W = 1] - \mathbb{E}[Y(0) \mid W = 0] + \underbrace{\mathbb{E}[Y(1) \mid W = 1] - \mathbb{E}[Y(0) \mid W = 1]}_{\tau_{ATT}} - \tau_{ATE}$$

$$= \tau_{ATE} + \mathbb{E}[Y(0) \mid W = 1] - \mathbb{E}[Y(0) \mid W = 0] + \tau_{ATT} - \tau_{ATE}$$ 🤓

Note that The Mixtape provides a different but equivalent decomposition (nice exercise to show why)

$$\tau_{ATE} + \underbrace{\mathbb{E}[Y(0) \mid W = 1] - \mathbb{E}[Y(0) \mid W = 0]}_{\text{confounding bias}} + \underbrace{(1 - \mathbb{E}[W])(\tau_{ATT} - \tau_{ATU})}_{\text{heterogeneous effect bias}}$$

where $\tau_{ATU} := \mathbb{E}[Y(1) - Y(0) \mid W = 0]$ is the ATE ON THE UNTREATED (ATU)

Discussion of potential outcomes

Pros:

- Very powerful to define causal target parameters
- Very powerful to reverse engineer the causal content of statistical quantities

Cons:

- Some may feel uncomfortable with taking POs as primitives (where do they come from?)
- Managing complex causal structures can be cumbersome

Structural causal models and directed acyclic graphs

Characterization

Structural Causal Models (SCM) a.k.a. Nonparametric Structural Equation Models (NPSEM) formalize causal dependencies between RVs

They consist of three components:

- U: a set of exogenous variables that we assume to be jointly independent
- V: a set of endogenous variables
- F: a set of functions defining the variables in V using variables in U and V as inputs (define ”who listens to whom”)

Every SCM is associated with a *graphical causal model*

In our settings these will be the famous Directed Acyclic Graphs (DAGs)

Definition of causation in SCMs/DAGs

Every argument in the function defining an endogenous variable is a *direct cause* of this variable

We say: Every parent variable is a direct cause of its child variable(s)

We write every endogenous variable V_j as function of its parents and a j-specific exogenous variable:

$$V_j := f_j(PA_j, U_j)$$

$\Rightarrow$ We can use SCMs and DAGs to encode the causal assumptions we are willing to make in our setting

It is completely nonparametric, i.e exogenous variables and functions creating the endogenous variables do not take any specific, e.g. linear, form

Example

One classic SCM is the one displaying simple confounding of the treatment W

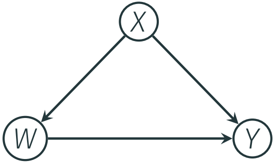
$$U = \{U_X, U_W, U_Y\}, \quad V = \{X, W, Y\}, \quad F = \{f_X, f_W, f_Y\}$$

$$X := f_X(U_X)$$

$$W := f_W(X, U_W)$$

$$Y := f_Y(X, W, U_Y)$$

This SCM has the associated DAG



where we follow the convention to suppress the exogenous variables

Finding (conditional) (in)dependencies

Conditional) (in)dependencies in DAGs

One powerful feature of DAGs is that they encode (conditional) (in)dependencies between their RVs

In the following we will learn a general recipe to check whether two RVs in a DAG are

- independent,
- independent conditionally on a set of RVs,
- likely dependent or
- likely dependent conditionally on a set of RVs

This process is called *d*-separation

(Conditional) (in)dependence of RVs

Two variables X and Y can be

- *d*-separated w/o conditioning on any variable $\Rightarrow X \perp\!\!\!\perp Y$
- *d*-separated conditional on (a set of nodes) $Z \Rightarrow X \perp\!\!\!\perp Y \mid Z$
- *d*-connected w/o conditioning on any variable $\Rightarrow X \not\!\perp\!\!\!\perp Y$
- *d*-connected conditional on (a set of nodes) $Z \Rightarrow X \not\!\perp\!\!\!\perp Y \mid Z$

A pair of RVs can be *d*-separated by ”blocking all paths” between them

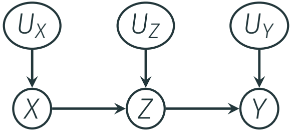
But how can we block paths?

Three useful patterns

For blocking paths we introduce three useful patterns that populate also more complex graphs:

- *Chains:* e.g. $X \to Y \to Z$
- *Forks:* e.g. $X \leftarrow Y \to Z$
- *Colliders:* e.g. $X \to Y \leftarrow Z$

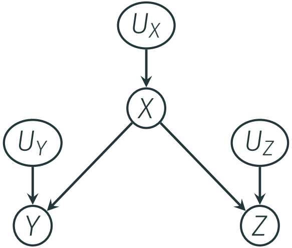
Chains



This pattern is called a *chain* and implies the following (in)dependencies:

1. X and Z are likely dependent: $X \not\perp\!\!\!\perp Z$
2. Z and Y are likely dependent: $Z \not\perp\!\!\!\perp Y$
3. X and Y are likely dependent: $X \not\perp\!\!\!\perp Y$
4. X and Y are independent, conditional on Z: $X \perp\!\!\!\perp Y \mid Z$

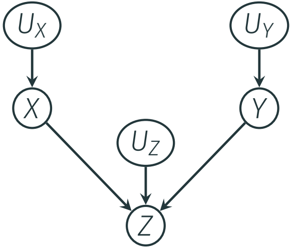
Forks



This pattern is called a *fork* and implies the following (in)dependencies:

1. Z and Y are likely dependent: $Z \not\perp\!\!\!\perp Y$
2. Y and X are likely dependent: $Y \not\perp\!\!\!\perp X$
3. Z and X are likely dependent: $Z \not\perp\!\!\!\perp X$
4. Z and Y are independent, conditional on X: $Z \perp\!\!\!\perp Y \mid X$

Collider



This pattern is called a *collider* and implies the following (in)dependencies:

1. X and Z are likely dependent: $X \not\perp\!\!\!\perp Z$
2. Y and Z are likely dependent: $Y \not\perp\!\!\!\perp Z$
3. X and Y are independent: $X \perp\!\!\!\perp Y$
4. X and Y are likely dependent, conditional on Z: $X \not\perp\!\!\!\perp Y \mid Z$

Three rules

Rule 1 (Conditional Independence in Chains)

Two variables, X and Y, are conditionally independent given Z, if there is only one unidirectional path between X and Y, and Z is any set of variables that intercepts that path.

Rule 2 (Conditional Independence in Forks)

If a variable X is a common cause of variables Y and Z, and there is only one path between Y and Z, then Y and Z are independent conditional on X.

Rule 3 (Unconditional Independence in Colliders)

If a variable Z is the collision node between two variables X and Y, and there is only one path between X and Y, then X and Y are unconditionally independent but are dependent conditional on Z and any descendants of Z.

d-separation - definition

Definition (d-separation)

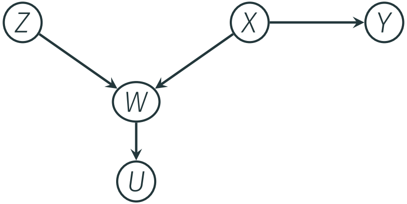
A path p is blocked by a set of nodes Z if and only if

1. p contains a chain of nodes $A \rightarrow B \rightarrow C$ or a fork $A \leftarrow B \rightarrow C$ such that the middle node B is in Z (i.e., B is conditioned on), or
2. p contains a collider $A \rightarrow B \leftarrow C$ such that the collision node B is not in Z, and no descendant of B is in Z.

If Z blocks every path between two nodes X and Y, then X and Y are d-separated, conditional on Z, and thus are independent conditional on Z.

Note that $Z = \{\}$, i.e. an empty set, is a special case of this where only 2. can block paths

Example

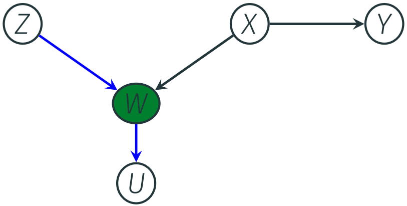


We omit the exogenous variables for compactness

Every two non-adjacent variables span a path that can be investigated:

$$\{Z, U\}, \{Z, X\}, \{Z, Y\}, \{W, Y\}, \{X, U\}, \{Y, U\}$$

$\{Z, U\}$ unconditional



RV pair $\{Z, U\}$ is *d*-connected as the path is not blocked by a collider

$$\Rightarrow Z \not\perp\!\!\!\perp U$$

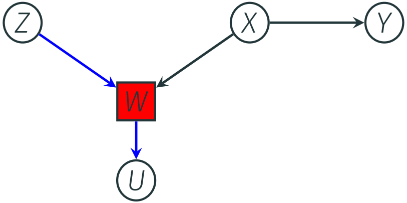
I highlight the path under consideration in blue

The green node on the path means dependence can flow through this node 🚦

$\{Z, U\}$ conditional on $\{W\}$

A square node means we condition on that node

A red node means that dependence flow is blocked 🚦

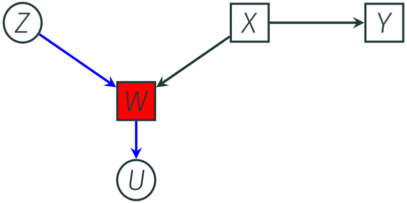


RV pair $\{Z, U\}$ can be *d*-separated by conditioning on $\{W\}$

It is the middle variable of the chain $Z \to W \to U$

$$\Rightarrow Z \perp\!\!\!\perp U \mid W$$

$\{Z, U\}$ conditional on W and more?



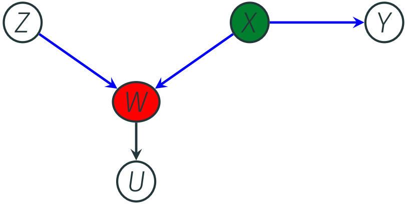
Neither conditioning on X nor on Y can block the path, simply because they do not lie on the path

However, conditioning does not open a path either because they are no descendants of a collider (see 2. in d-separation definition on slide 27)

$\Rightarrow$ Also conditioning on $\{W, X\}$, $\{W, Y\}$, and $\{W, X, Y\}$ d-separates Z and U

$\Rightarrow$ We can condition on them, or not 🤷

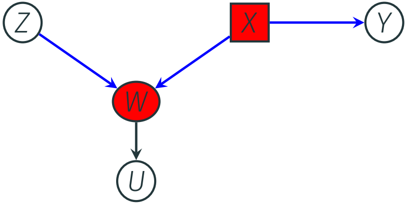
$\{Z, Y\}$ unconditional



RV pair $\{Z, Y\}$ is unconditionally *d*-separated because the collider W blocks the path

$$\Rightarrow \; Z \perp\!\!\!\perp Y$$

$\{Z, Y\}$ conditional on $\{X\}$

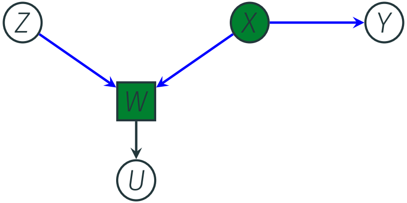


RV pair $\{Z, Y\}$ is d-separated conditional on $\{X\}$ because it is in the middle of fork $W \leftarrow X \rightarrow Y$ and therefore blocks the path additional to the collider

$$\Rightarrow Z \perp\!\!\!\perp Y \mid X$$

Take-away: While one block suffices to d-separate, we can also block a path multiple times

$\{Z, Y\}$ conditional on $\{W\}$



RV pair $\{Z, Y\}$ is *d*-connected conditional on $\{W\}$ because conditioning on a collider opens/unblocks the path

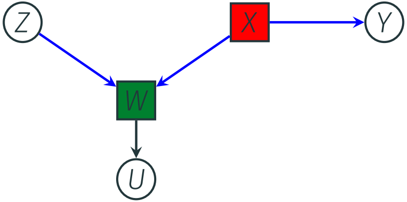
$$\Rightarrow \; Z \not\perp\!\!\!\perp Y \mid W$$

So we can never condition on a collider if we want to block a path? 🤔

$\{Z, Y\}$ conditional on $\{W, X\}$

No, it suffices to block the path once

In the example conditioning additionally on X blocks the opened path again because it is in the middle of fork $W \leftarrow X \rightarrow Y$



$\Rightarrow$ RV pair $\{Z, Y\}$ is *d*-separated conditional on $\{W, X\}$

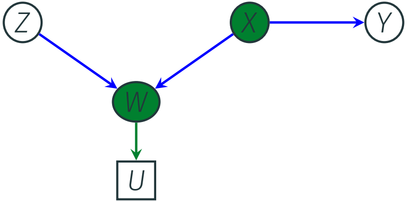
$$\Rightarrow \; Z \perp\!\!\!\perp Y \mid W, X$$

Take-away: We can condition on colliders if the path is blocked by other nodes

$\{Z, Y\}$ conditional on $\{U\}$

Conditioning on $\{U\}$ should not hurt as it is not on the path, right?

No, because it is a descendant of a collider and they unblock the path (2nd part of 2. in definition on slide 27)



$\Rightarrow$ RV pair $\{Z, Y\}$ is d-connected conditional on $\{U\}$

$$\Rightarrow \; Z \not\perp\!\!\!\perp Y \mid U$$

Again conditioning additionally on X would block the path

Discussion of structural causal models

Pros:

- Very powerful to express complex causal systems and to derive their testable implications via *d*-separation
- Comes with the very powerful *do*-calculus to show identification of causal quantities

Cons:

- *do*-calculus non-trivial to learn and apply (beyond the scope of this course)
- Definition of target parameters can be cumbersome

Single World Intervention Graphs

Unifying DAGs and POs

Single World Intervention Graphs (SWIGs) unify POs and SCMs

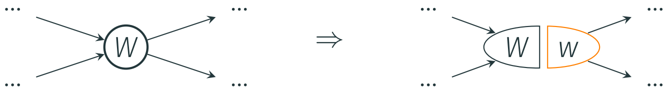
They provide a recipe to show which POs live in a DAG and to derive (conditional) independencies with respect to *potential* instead of *observed* outcomes

These independencies allow us then to compactly show identification of causal effects in settings in which Causal ML estimators are most prominent

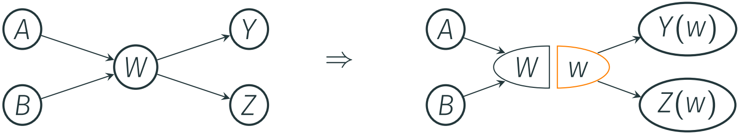
The recipe

Single World Intervention Graphs (SWIGs) transform a DAG in two steps:

1. *Node splitting:* Split the treatment node(s) into
   - a random node inheriting all incoming edges (capital letter)
   - a fixed node inheriting all outgoing edges (small letter)

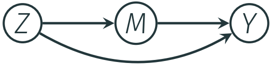


2. *Relabeling:* Descendants of fixed nodes become potential outcomes



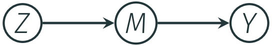
Working with SWIGs

In the following we will consider two examples that introduce and illustrate how to work with SWIGs using DAGs $\mathcal{G}$ as starting point



Implying the SCM

- $Z := f_Z(U_Z)$
- $M := M(Z) = f_M(Z, U_M)$
- $Y := Y(Z, M(Z)) = f_Y(Z, M, U_Y)$



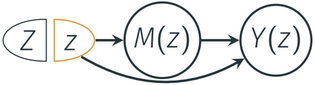
Implying the SCM

- $Z := f_Z(U_Z)$
- $M := M(Z) = f_M(Z, U_M)$
- $Y := Y(M(Z)) = f_Y(M, U_Y)$

POs in the SCMs for intervention on Z

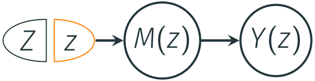
By fixing the endogenous variables in the SCM, we can define different POs

We start with fixing $Z = z$ to create the graphs $\mathcal{G}(z)$



Encoding the POs

- $M(\textcolor{orange}{z}) = f_M(\textcolor{orange}{z}, U_M)$
- $Y(\textcolor{orange}{z}) = f_Y(\textcolor{orange}{z}, M(\textcolor{orange}{z}), U_Y)$

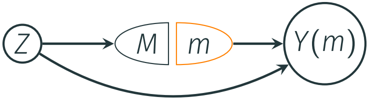


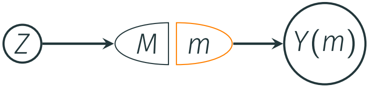
Encoding the POs

- $M(\textcolor{orange}{z}) = f_M(\textcolor{orange}{z}, U_M)$
- $Y(\textcolor{orange}{z}) = f_Y(M(\textcolor{orange}{z}), U_Y)$

POs in the SCMs for intervention on M

Now we fix $M = m$ to create the graphs $\mathcal{G}(m)$





Encoding the PO

- $Y(m) = f_Y(Z, m, U_Y)$

Encoding the PO

- $Y(m) = f_Y(m, U_Y)$

What to do with SWIGs?

SWIGs are populated by observed and potential RVs

The nice thing is that we can use standard *d*-separation to read off (conditional) independencies between them

The only additional rule is that a split node also blocks a path

DAGs and SWIGs are therefore complementary

DAGs encode independencies that can be used to test the causal model (not scope of this course, see e.g. Peters, Janzing and Schölkopf book)

SWIGs encode independencies that can be used to show identification of causal target parameters in a lean manner (crucial for us)

Identification in RCTs

Randomized Controlled Trials

Whether you call it randomized controlled trial (RCT) or A/B testing, randomizing the treatment is arguably the cleanest way to obtain causal effect estimates

Why does randomization work so well?

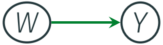
Randomness does not care about anything $\Rightarrow$ W can not be predicted

Most importantly, W is independent of other variables

In particular it is independent of potential outcomes as we can now derive using SWIGs

DAG and SWIG of RCT

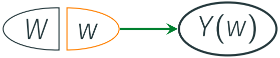
The DAG



implies only $W \not\perp\!\!\!\perp Y$ because of the directed path

$\Rightarrow$ Not very surprising or useful

The created SWIG



implies $Y(w) \perp\!\!\!\perp W$ because the path between the potential outcome and the factual treatment is blocked by the split node

SWIGs + *d*-separation allow us to derive in a principled way what intuition tells us:

$$Y(w) \perp\!\!\!\perp W \tag{2}$$

Identification of APO

How can we use this independence for identification?

In few steps, we can identify the average potential outcome (APO)

$$\begin{aligned} \mathbb{E}[Y(w)] &= \mathbb{E}[Y(w) \mid W = w] && (2) \\ &= \mathbb{E}[Y \mid W = w] && (\textit{consistency}) \end{aligned}$$

We expressed an unobservable quantity in terms of observable variables

$\Rightarrow$ identification of APO as simple conditional expectation ✅

Identification of ATE in RCT

This means that also the average treatment effect (ATE) is identified in an RCT

$$\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[Y \mid W = 1] - \mathbb{E}[Y \mid W = 0]$$

Note that also the effects of the treated and untreated are identified in the same way because under (2)

$$\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[Y(1) - Y(0) \mid W = 1] = \mathbb{E}[Y(1) - Y(0) \mid W = 0]$$

Estimation of ATE in RCT

The next step after identification is now estimation

The easiest way is to apply a simple mean comparison:

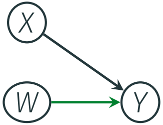
$$\hat{\tau}_{ATE} = \frac{1}{\sum_i W_i} \sum_i W_i Y_i - \frac{1}{\sum_i (1 - W_i)} \sum_i (1 - W_i) Y_i$$

or equivalently run an OLS regression of the form

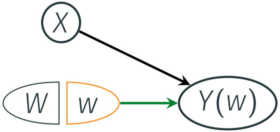
$$Y = \alpha + \tau W + \varepsilon_{Y \sim W}$$

Identification of CATE in RCT (1/2)

The DAG with heterogeneity variables/moderators X



The created SWIG



As conditioning on X neither blocks nor opens a path in the SWIG, it keeps W and $Y(w)$ also conditionally d-separated and therefore

$$Y(w) \perp\!\!\!\perp W \mid X \tag{3}$$

can be established from the SWIG

Identification of CATE in RCT (2/2)

We can therefore show also identification of CATEs using this conditional independence:

$$\begin{aligned} \mathbb{E}[Y(1) - Y(0) \mid X] &= \mathbb{E}[Y(1) \mid X] - \mathbb{E}[Y(0) \mid X] && (LOE) \\ &= \mathbb{E}[Y(1) \mid W = 1, X] - \mathbb{E}[Y(0) \mid W = 0, X] && (3) \\ &= \mathbb{E}[Y \mid W = 1, X] - \mathbb{E}[Y \mid W = 0, X] && (consistency) \end{aligned}$$

CATE is identified as the difference in two CEFs

We will have some fun to cleverly estimate this expression later

Identification notebook: DAG and SWIG for RCTs

Valid adjustment sets

Finding valid adjustment sets

Causal inference in RCTs is by design relatively straightforward and the technology introduced above is not really required

However, in many cases we either can and/or do not want to randomize the treatment

Then identification is more complicated

In some cases we know enough about the structure of our data that we can justify to control/adjust for variables to identify our target parameters

But how to find valid adjustment sets?

Many different adjustment criteria available

We focus on the imo most simple to proof one using SWIGs + d-separation

Counterfactual Adjustment Criterion

Richardson and Robins (2013) introduce the

Counterfactual Adjustment Criterion (CAC)

If $Y(w) \perp\!\!\!\perp W \mid X$ is implied by the SWIG $\mathcal{G}(w)$, then

$$\mathbb{E}[Y(w)] = \mathbb{E}[\mathbb{E}[Y \mid X, W = w]]$$

Therefore any set of variables that d-separates $Y(w)$ and W in the SWIG is a valid adjustment set

However, note that conditioning on other POs in the SWIG are ruled out b/c CAC does NOT include $Y(w) \perp\!\!\!\perp W \mid X(w)$

An easy to proof adjustment criterion (2/2)

The proof for the APO is basically a conditional version of the RCT identification:

$$\begin{aligned} \mathbb{E}[Y(w)] &= \mathbb{E}[\mathbb{E}[Y(w) \mid X]] && (LIE) \\ &= \mathbb{E}[\mathbb{E}[Y(w) \mid X, W = w]] && (Y(w) \perp\!\!\!\perp W \mid X) \\ &= \mathbb{E}[\mathbb{E}[Y \mid X, W = w]] && (consistency) \end{aligned}$$

Therefore the ATE is identified as

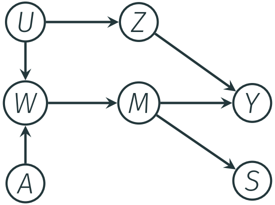
$$\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[\mathbb{E}[Y \mid X, W = 1]] - \mathbb{E}[\mathbb{E}[Y \mid X, W = 0]]$$

Estimation of this guy is covered in the upcoming lectures

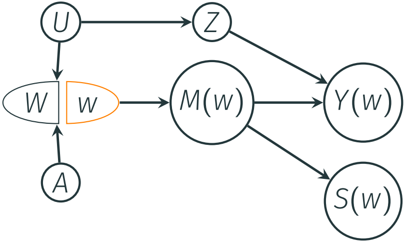
Similar derivations show also that CATE, ATT or ATU can be identified with a valid adjustment set

Example

DAG $\mathcal{G}$:



SWIG $\mathcal{G}(w)$:



Recipe to find valid adjustment sets to identify effect of W on Y:

1. Rule out all variables that are potential outcomes in the SWIG: M, S, and Y
2. Check which variable sets d-separate W and $Y(w)$ in the SWIG:
   $X \in \{\{U\}, \{Z\}, \{A, U\}, \{A, Z\}, \{U, Z\}, \{A, U, Z\}\}$

$\Rightarrow$ We have six valid adjustment sets according to the CAC

Identification notebook: DAG and SWIG with valid adjustment set

Wrapping up

Wrapping up

The punchline is that the combination DAGs + SWIGs + d-separation is very powerful to show identification of causal target parameters defined using POs

The crucial assumption is that the DAG properly represents the causal structure in our specific problem

Once a credible graph exists, valid adjustment sets can be found manually or computationally (check `dagitty`)

We do not study how to draw a credible DAG but assume throughout that it is given to us

Outlook

We cover identification with instrumental variables at a later stage

We will not look into difference-in-differences and regression discontinuity

Causal ML for them build on the same principles as the ones we introduce, but these strategies would require more investment on the causal inference side

If causal inference is new for you (and even if not), I highly recommend to read Chapters 3 & 4 of Causal Inference - The Mixtape by Scott Cunningham

Ceterum censeo a fancy method alone is not a credible identification strategy

$\Rightarrow$ separate identification and estimation