Causal Directed Acylic Graphs

introduction

Wouter van Amsterdam

2024-08-06

Table of contents I

U

Causal inference frameworks

What are they for?

Mathematical language to

- define *causal* quantities
- express *assumptions*
- derive how to *estimate* causal effects

Causal inference frameworks

Why learn more than one?

- On day 1 we learned about the Potential Outcomes framework
  - Defines causal effects in terms of (averages of) *individual potential outcomes*
  - Estimation requires assumptions of (conditional) exchangeability and positivity / overlap and consistency
- There isn't only 1 way to think about causality, find one that '*clicks*'
- Now we will learn another framework: *Structural Causal Models* and *causal graphs*
  - causal relations and manipulations of *variables*
  - Developed by different people initially - Judea Pearl, Peter Spirtes, Clark Glymour
  - SCM approach is broader in that it can define more different types of causal questions
- Equivalence: given the same data and assumptions, get the same estimates

lecture 1 topics

- why use DAGs
-
- what are DAGs and where do come from
- SCMs as computer programs
  - intervention as change in program
- causal identifyability with DAGs
  - variable types: confounders, mediators, colliders (tinder: hot and intelligent / single ;on tinder)
    - confounders; storks; randomized, breaking arrows
  - d-separation
  - back-door criterion
  - follow rules of do calculus
- DAG

1. what is the world
2. what can we estimate and how?
3. confounders / colliders
4. backdoor

lecture 2 topics

- perfect versus soft intervention
- critique of scm / cross-world assumptions

practical 1

- drawing and using dags (what to condition on); daggity
- same data, different dags, different answers

lecture 3 topics

- bonus queries:
  - counterfactuals
  - probability of necessity, probability of sufficiency
  - actual causality (Joe Halpern)
- Pearl Causal Hierarchy
- other uses of DAGs: missing data, selection
- reflections on DAGs, limitations

practical 2

- causal ladder: what Q is this?
- give data of hierarchy and answer the Q
- give data of 2 treatments + SCM (treatment 3 which can be extrapolated from)

Why are DAGs useful

Example task: are hospital deliveries good for babies?

Example task: are hospital deliveries good for babies?

- You're a data scientist in the Wilhelmina Kinderziekenhuis (WKZ)
- Have data on
  - delivery location (home or hospital)
  - neonatal outcomes (good or bad)
  - pregnancy risk (high or low)
- Question: do hospital deliveries result in better outcomes for babies?

Example task: are hospital deliveries good for babies?

- You're a data scientist in the Wilhelmina Kinderziekenhuis (WKZ)
- Have data on
  - delivery location (home or hospital)
  - neonatal outcomes (good or bad)
  - pregnancy risk (high or low)
- Question: do hospital deliveries result in better outcomes for babies?

Example task: are hospital deliveries good for babies?

- You're a data scientist in the Wilhelmina Kinderziekenhuis (WKZ)
- Have data on
  - delivery location (home or hospital)
  - neonatal outcomes (good or bad)
  - pregnancy risk (high or low)
- Question: do hospital deliveries result in better outcomes for babies?

Observed data

		location	
		home	hospital
risk	low	648 / 720 = 90%	19 / 20 = 95%
	high	40 / 80 = 50%	144 / 180 = 80%

- better outcomes for babies delivered in the hospital for *both risk groups*

Observed data

		location	
		home	hospital
risk	low	$648 / 720 = 90\%$	$19 / 20 = 95\%$
	high	$40 / 80 = 50\%$	$144 / 180 = 80\%$
	marginal	$688 / 800 = 86\%$	$163 / 200 = 81.5\%$

- better outcomes for babies delivered in the hospital for *both risk groups*
- but not better *marginal* (‘overall’)
- how is this possible? (a.k.a. *simpsons paradox*)
- what is the correct way to estimate the effect of delivery location?

New question: hernia

- for a patient with a hernia, will they be able to walk sooner when recovering at home or when recovering in a hospital?

Observed data 2

		location	
		home	hospital
bedrest	no	648 / 720 = 90%	19 / 20 = 95%
	yes	40 / 80 = 50%	144 / 180 = 80%
	marginal	688 / 800 = 86%	163 / 200 = 81.5%

- more bed rest in hospital
- what is the correct way to estimate the effect of location?

How to unravel this?

- we got two questions with exactly the same data
- in one example, ‘stratified analysis’ seemed best
- in the other example, ‘marginal analysis’ seemed best
- with *Directed Acyclic Graphs* we can make our decision

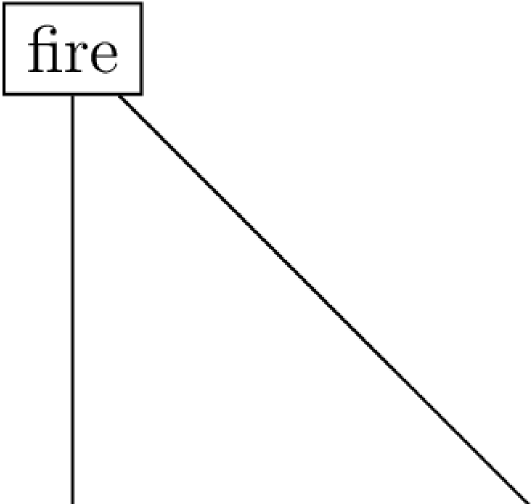
Assumption parlance

- necessary assumption:
  - A **must** hold for B to be true
- sufficient assumption:
  - B is always true when A holds
- strong assumption:
  - requires *strong* evidence, we rather not make these
- weak assumption:
  - requires *weak* evidence
- strong vs weak assumption are judged on relative terms
  - if assumption A is sufficient for B, B cannot be a stronger assumption that A

Causal Directed Acyclic Graphs

diagram that represents our assumptions on causal relations

1. nodes are variables
2. arrows (directed edges) point from cause to effect



DAGs convey two types of assumptions:

causal direction and conditional independence

1. causal direction: what causes what?

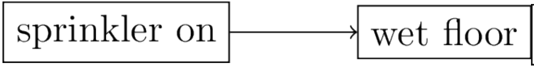


Figure 2: DAG 1

(a) DAG 2

- read Figure 2 as
  - `sprinkler on` **may** (or may not) cause `wet floor`
  - `wet floor` **cannot** cause `sprinkler on`

DAGs convey two types of assumptions:

causal direction and conditional independence

1. conditional indepence (e.g. exclusion of influence / information)

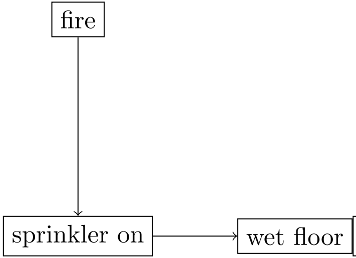


Figure 4: DAG 1

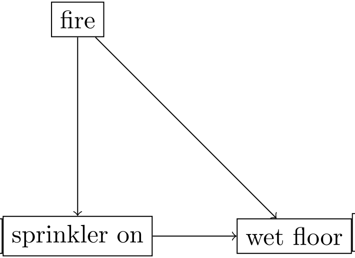


Figure 5: DAG 2

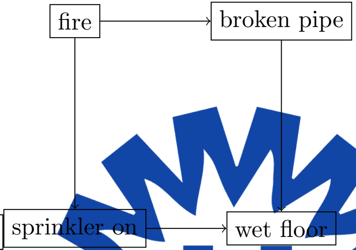
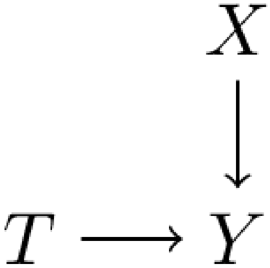


Figure 6: DAG 3

- Figure 4 says `fire` can **only** cause `wet floor` through `sprinkler on`
- Figure 5 says *there may be other ways through which `fire` causes `wet floor`*
    - Figure 5 is thus a *weaker* assumption than Figure 4
- Figure 6 is also compatible with Figure 5

DAGs are ‘non-parametric’

They relay what variable ‘listens’ to what, but not in what way



(a) DAG

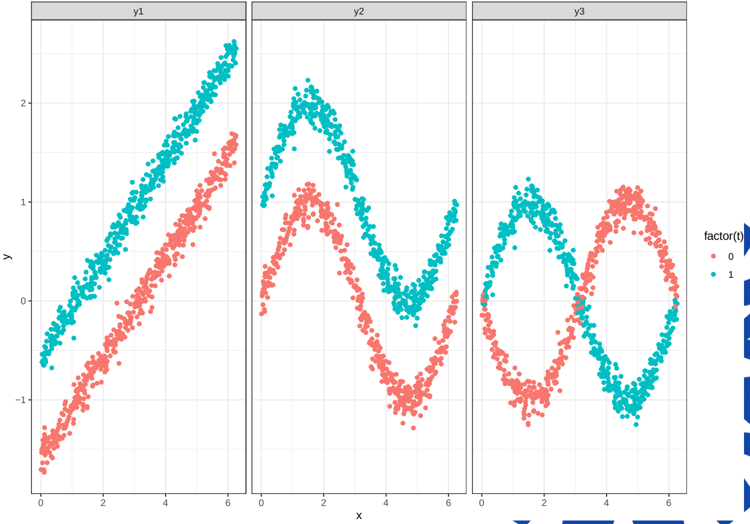


Figure 8: Three datasets with the same DAG

DAGs are 'non-parametric'

They relay what variable 'listens' to what, but not in what way

$$\begin{tikzcd} & X \arrow[d] \\ T \arrow[r] & Y \end{tikzcd}$$

(a) DAG

- this DAG says Y is a function of X, T, or:
- $Y = f_Y(X, T)$
- in the next lecture we'll talk more about these 'structural equations'

Making DAGs for our examples:
The pregnancy DAG

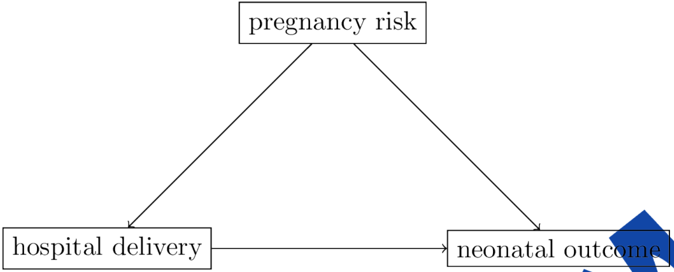


Figure 10

- assumptions:
  - women with high risk of bad neonatal outcomes (`pregnancy risk`) are referred to the hospital for delivery
  - hospital deliveries lead to better outcomes for babies as more emergency treatments possible
  - both `pregnancy risk` and `hospital delivery` cause `neonatal outcome`
- the *other variable* `pregnancy risk` is a common cause of the treatment (`hospital delivery`) and the outcome (this is what's called a confounder)

Making DAGs for our examples:

The hernia DAG

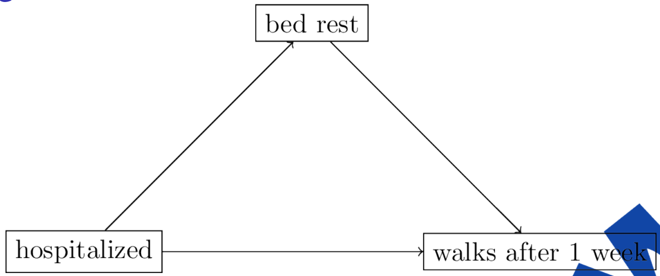


Figure 11

- assumptions:
  - patients admitted to the hospital keep more `bed rest` than those who remain at home
  - `bed rest` leads to lower recovery times thus less walking patients after 1 week
- the *other variable* `bed rest` is a *mediator* between the treatment (`hospitalized`) and the outcome

DAG definitions

The DAG definition of an intervention

assume this is our DAG for a situation:

- then intervening on variable T means removing all incoming arrows

$$\begin{tikzcd} W \arrow[d] & Z \arrow[dl] \arrow[d] \\ T \arrow[r] & Y \end{tikzcd}$$

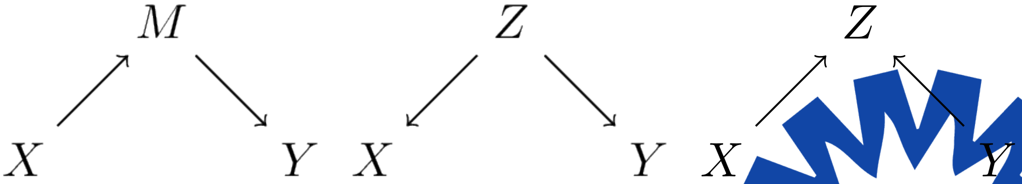
(a) observational data

$$\begin{tikzcd} W & Z \arrow[d] \\ T \arrow[r] & Y \end{tikzcd}$$

(b) intervened DAG

- which means T does not *listen* to other variables anymore, but is set at a particular value, like in an experiment

Basic DAG types



(a) chain (b) fork (c) collider

next steps

conclusion 1: seeing is not doing

to follow-up

- DAG to picture the game
- doing = mutilating DAG
- hotel 2: the Randtz
- identifyability: more SCMs with same marginals
- another hidden variable: are the guests brittish
- SCM = know rules of the game
- DAG = know who listens to what
- why are DAGs useful? know what you can compute

References