Causality and Graphical Models

Chris Williams

Probabilistic Modelling and Reasoning (INFR11134)
School of Informatics, The University of Edinburgh

Spring Semester 2024

The Causal Hierarchy

Pearl (2000)

- **Association** (seeing)
  Example: What does a symptom tell me about a disease?
- **Intervention** (doing)
  Example: If I take an aspirin, will my headache be cured?
- **Counterfactuals** (imagining)
  Example: was it the aspirin that stopped my headache?

[This is a short introduction to causality. There is a whole course *Methods for Causal Inference* if you want to learn more.]

Outline

- Structural Causal Models
- Interventions
- Causal effects
- Confounding
- Adjustment for direct causes
- Causal identifiability
- Counterfactuals

Structural Causal Models

- A **structural causal model (SCM)** M is given by a set of variables $X_1, \ldots, X_d$ and corresponding assignments of the form
$$X_i := f_i(Pa_i, U_i) \qquad \text{for } i = 1, \ldots, d$$
where Pa_i is the parents of X_i, and the U's are jointly independent noise variables (aka exogenous factors). The f_is are deterministic functions
- The DAG corresponding to the model has one node for each X_i. This is termed the **causal graph** corresponding to the structural causal model
- SCM goes beyond the causal graphical model (CGM) with factorization
$$p(X_1, \ldots, X_d) = \prod_{i=1}^{d} p(X_i | Pa_i)$$

- A causal graphical model is a DGM in which each arc is interpreted as a direct causal influence between a parent node and a child node, relative to the other nodes in the network.
- Both SCMs and CGMs can handle interventions, but SCMs are needed to handle counterfactuals
- Different SCMs can give rise to the same CGM

Interventions

- Given a SCM M we can take any assignment

$$X := f(Pa, U)$$

and replace it by

$$X := x.$$

- We denote this as $M' = M[X := x]$ or $M' = M; do(X := x)$
- Graphically, the operation eliminates all incoming edges into X; this is called the *modified* graphical model
- The assignment operator is called the **do-operator**
- After applying the do-operator, we obtain probabilities for an event E in the new graph M' as $p_{M[X:=x]}(E)$
- Can also write $p(E|do(X := x))$
  but the do-operator is fundamentally different from conditioning

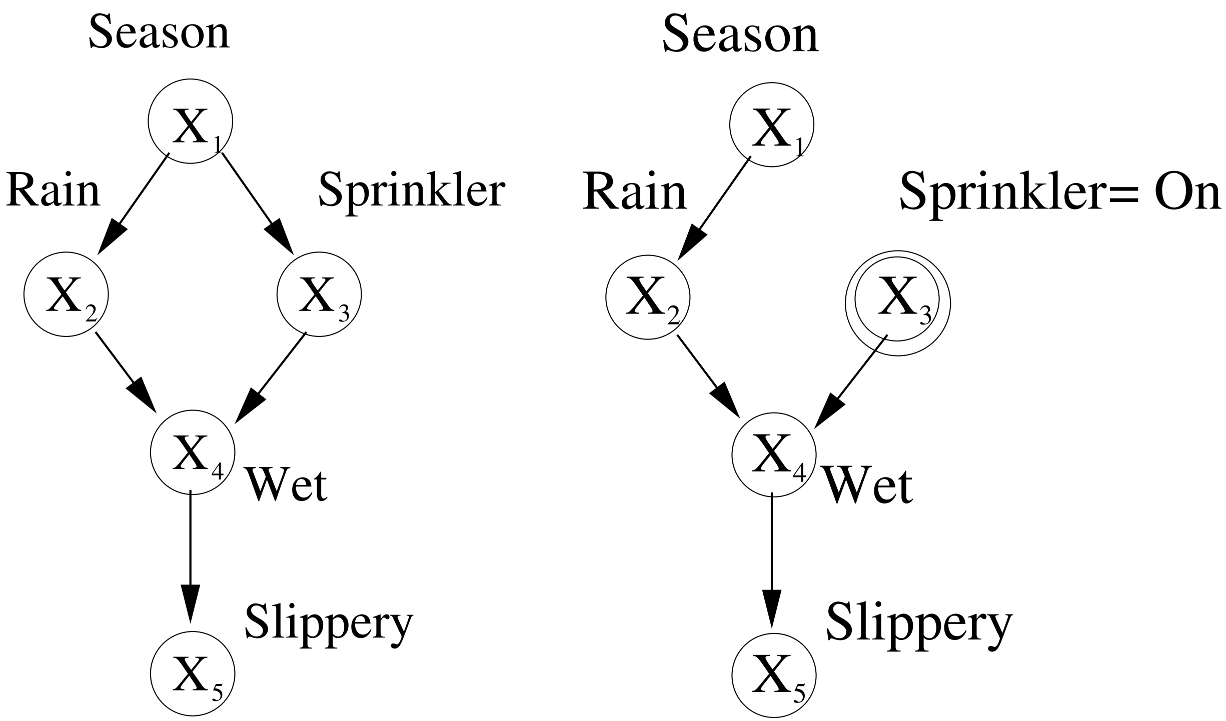
- $p_{M[X_i=x_i]}(x_1, \ldots, x_d)$ is defined by the *truncated factorization formula*

$$p_{M[X_i:=x_i]}(x_1, \ldots, x_d) = \begin{cases} \prod_{j \neq i} p(x_j | pa_j) & \text{if } X_i = x_i \\ 0 & \text{if } X_i \neq x_i \end{cases}$$

- By marginalizing out the other variables, we can see that

$$\begin{aligned} p_{M'}(X_i = x_i) &= 1 \\ p_{M'}(X_i = x_i') &= 0 \qquad \text{if } x_i' \neq x_i \end{aligned}$$

Intervention as surgery on graphs



- Intervening on X_3 produces the modified graphical model on the right

Causal effects

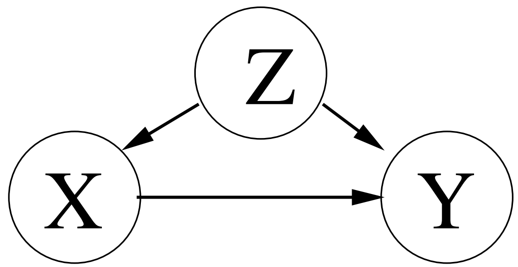
- The **causal effect** of an action $X := x$ on a variable Y refers to the distribution of the variable Y in the model $M' = M[X := x]$
- Suppose X denotes the presence or absence of an intervention of treatment (e.g., taking a drug or not)
- Assume Y takes on values of 0 and 1
- The **average treatment effect**

$$\mathrm{ATE} = \mathbb{E}_{M[X:=1]}[Y] - \mathbb{E}_{M[X:=0]}[Y]$$

- Causal effects are population quantities, relating to effects averaged over the whole population

Confounding

- In general the causal effect $p(Y|do(X := x))$ does not coincide with the conditional $p(Y|X = x)$
- The difference between interventional statements and conditional statements in known as **confounding**
- Classic setup, Z is a common cause of X and Y

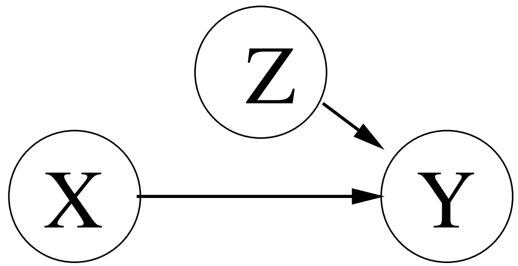


- Example: X is taking a drug or not, Y is recovery (or not), and Z is a patient's blood pressure. The blood pressure Z influences a patient being assigned to the drug, as well as their chances of recovery Y
- Confounders may be observed, or unobserved

Adjusting for Direct Causes

$$p(Y = y|do(X := x)) = \sum_z p(Y = y|X = x, Pa = z)p(Pa = z)$$

- This is called the **adjustment formula**
- Follows from the **modified** graphical model



- Contrast the adjustment formula with conditioning

$$\begin{aligned} p(Y = y|X = x) &= \sum_z p(Y = y, Pa = z|X = x) \\ &= \sum_z p(Y = y|X = x, Pa = z)p(Pa = z|X = x) \end{aligned}$$

Propensity score and inverse probability weighting

$$\begin{aligned} p(Y = y | do(X := x)) &= \sum_z p(Y = y | X = x, Pa = z) p(Pa = z) \\ &= \sum_z p(Y = y | X = x, Pa = z) \frac{p(X = x | Pa = z)}{p(X = x | Pa = z)} p(Pa = z) \\ &= \sum_z \frac{p(X = x, Y = y, Pa = z)}{p(X = x | Pa = z)} \end{aligned}$$

- The term $p(X = x | Pa = z)$ is known as the *propensity score*
- It is the propensity (probability) that a unit is assigned to a particular treatment, given $Pa = z$, in the observations
- Division by this term gives rise to the name "inverse probability weighting"

Example of Adjustment

Success rates for treatment of kidney stones

	Overall	Small stones	Large stones
Treatment a	78% (273/350)	93% (81/87)	73% (192/263)
Treatment b	83% (289/350)	87% (234/270)	69% (55/80)

- Overall, treatment b looks to be more effective, but when broken down for both small and large kidney stones, treatment a is more effective. What's going on?
- Note that treatment a tends to be assigned for cases of large stones, and treatment b for small stones.
- The possibility of higher risks with treatment a may mean that it is not always used
- This pattern is an example of "Simpson's paradox" (where a trend that holds in all subpopulations may not hold at the population level)

Example 6.37 in Peters, Janzing and Schölkopf, 2017

- Let X denote the treatment (a or b), Y the outcome (1 for success, 0 for failure) and Z the size of the stone
- Adjustment formula

$$\begin{aligned} p(Y=1|do(X:=a)) &= \sum_z p(Y=1|X=a, Z=z)p(Z=z) \\ &= 0.93\frac{(87+270)}{700} + 0.73\frac{(263+80)}{700} = 0.832 \\ p(Y=1|do(X:=b)) &= \sum_z p(Y=1|X=b, Z=z)p(Z=z) \\ &= 0.87\frac{(87+270)}{700} + 0.69\frac{(263+80)}{700} = 0.782 \end{aligned}$$

- Average treatment effect

$$\mathrm{ATE} = 0.832 - 0.782$$

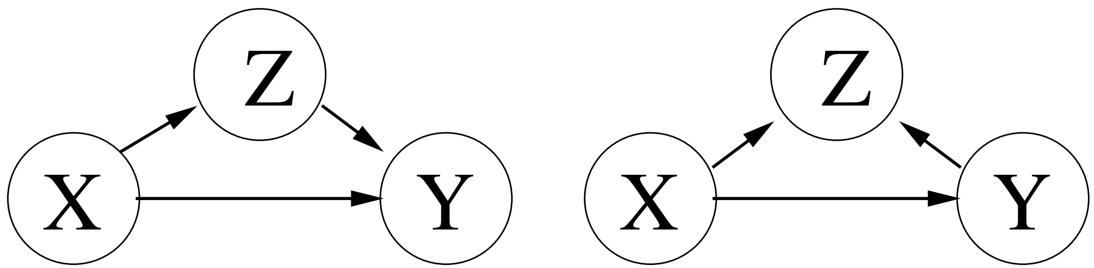
- In contrast the risk difference is $p(Y=1|X=a) - p(Y=1|X=b) = 0.780 - 0.826$

What is Adjustment?

- We wish to evaluate the effect of interventions on X on the target Y
- How do we take into account other variables Z, which may be called covariates, or confounders?
- *Adjustment* means partitioning the population into groups that are homogeneous relative to Z, assessing the effect of X on Y in each group, and then averaging the results (as per the adjustment formula)
- This is exactly what we did in the treatment of kidney stones example, where Z was the size of the stone
- "Adjust for" and "control for" are commonly used terms

Don't adjust for everything!

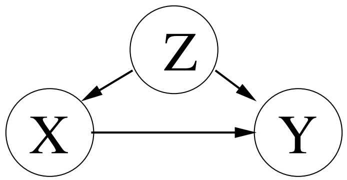
- In the adjustment formula above, we adjust for *the parents* of X
- It is also possible to use other valid adjustment sets, e.g, Pearl's "backdoor" and "frontdoor" criteria (details not required)
- But we **should not** control for all variables in the graph, e.g.



mediator collider

- Note that Z is not a parent of X in these two configurations

Causal Identifiability

- An intervention distribution $p(Y|M; do(X := x))$ is **identifiable** if it can be computed from the observational distribution and the graph structure
- Pearl's **do-calculus** determines the identifiability for a given graph and a set of observed variables
- Example: the confounder structure 

is identifiable if we observe X, Y and Z (adjustment formula)
- However, if Z is not observed, it is an unobserved confounder, and $p(Y|M; do(X := x))$ is not identifiable

Randomized Trials

- A do-operation does not have to be a fixed assignment
- In a randomized trial we have the operation $do(X := U_X)$
- E.g. in a drug trial, one might have 3 states: no medication, placebo, drug of interest, and U_X randomly chooses between these with (say) equal probability
- The randomization over X removes the influence of any other variable on X, and thus there cannot be any hidden common cause between X and Y
- This is an *experimental* manipulation, in contrast to only using observed data

Counterfactuals

Example from Hardt and Recht (2022, ch 9)

- We wish to drive to work, and can choose two routes $X = 0$ and $X = 1$. We decide randomly, i.e. $X := U_X \sim B(1/2)$
- On bad traffic days ($U = 1$), both routes are bad
- On good traffic days ($U = 0$) the traffic on either route is good unless there is an accident on the same route
- Accidents occur independently on either route with probability 1/2, so that $U_0,\ U_1 \sim B(1/2)$
- Our outcome variable is whether the traffic is good ($Y = 0$) or bad ($Y = 1$) on the chosen route
- Outcome Y is determined as

$$Y := X \cdot \max(U, U_1) + (1 - X) \max(U, U_0)$$

- Decoding the equation: say $X := 1$, then $Y = 0$ only if both U and U_1 are 0, otherwise $Y = 1$

- **Counterfactual question**: suppose we have $X := 1$ and observe bad traffic $Y = 1$. Would we have been better off taking the alternative route this morning?
- Notation $p(Y = 0|X = 1, Y = 1, do(X := 0))$
- To answer this, we need to compute $p(U, U_0, U_1|X = 1, Y = 1)$
- As $X = 1$, we cannot find out anything about U_0, thus this retains its prior distribution $U_0 \sim B(1/2)$
- As $Y = 1$, it must be that at least one of U and U_1 is equal to 1, so the posterior for $(U, U_1) = \{(1,0),(0,1),(1,1)\}$, each with probability $1/3$.
- Hence the posterior prob that $U = 1$ is $2/3$
- For the counterfactual query, $Y = 0$ if both U_0 and U are zero. This occurs with probability $\frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}$
- Interpretation: the evidence made it more likely to be a bad traffic day ($U = 1$), and this drops the probability from $1/4$ ($p(Y = 0|do(X := 0))$) to $1/6$

General recipe

Given a SCM M, observations $E = e$, action $X := x$ and a target variable Y, the counterfactual $p(Y = y | E = e, do(X := x))$ is defined by the three-step procedure

1. **Abduction:** Condition the joint distribution of the exogenous variables $U = (U_1, \ldots, U_d)$ on the event $E = e$ to obtain $p(U | E = e)$
2. **Action:** Perform the do-intervention $X := x$ in M resulting in the model $M' = M[X := x]$ and the modified graph
3. **Prediction:** Compute the target counterfactual using the noise distribution $p(U | E = e)$ in M'

This procedure defines what a counterfactual is in a SCM

What we are not covering

- Backdoor and frontdoor criteria
- Causal inference in practice
- Potential outcomes framework
- Causal discovery
- and lots more ...

Summary

- Structural Causal Models
- Interventions
- Causal effects
- Confounding
- Adjustment for direct causes
- Causal identifiability
- Counterfactuals

Further Reading

The material in these slides is covered largely by chapter 9 of

- *Patterns, Predictions, and Actions*, Moritz Hardt and Benjamin Recht, Princeton University Press (2022) [available free online]

Other more advanced texts include

- *Causal Inference in Statistics: A Primer*, Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell, Wiley (2016)
- *Causality*, Judea Pearl, Cambridge University Press (2000). Second edition in 2009.
- *Elements of Causal Inference*, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf, MIT Press (2017)