Structure of causal DAGs

ISTA 410 / INFO 510 - Bayesian Modeling and Inference

University of Arizona School of Information

September 15, 2021

Outline

Last time:

- Multiple regression
- Causal DAGs

Today: more DAGs

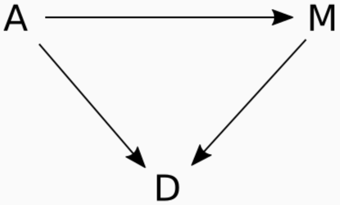
Chapter 5/6 of *Rethinking*

Recap

- Southern states, Waffle house, and divorce
  - Age at marriage and total marriage rate two potential predictors of divorce
  - Both associated with divorce; conditioning on age in a multiple regression suppressed the association between marriage rate and divorce rate
- Fox weight, food supply, and group size
  - Food supply not associated with fox weight in the full population
  - Conditioning on group size uncovered an association between food and weight
  - Association was masked by competing association between group size and weight

Our first DAG

We used this DAG to represent hypothetical causal associations between age, marriage rate, and divorce rate:



In this model, total causal effect of A on D:

1. $A \rightarrow D$ – direct causal effect
2. $A \rightarrow M \rightarrow D$ – indirect effect

Structure of DAGs

What is a DAG?

What is a DAG?

- Directed acyclic graph
- Nodes are variables
- Directed arrows are causal associations

What can we use DAGs for? Probabilistic models, on two levels:

- probabilistic model for associations between variables
- metadata that guides choice of variables for inference

Associations and information flow

In the DAG, we have paths from one variable to another:

- Paths go along the edges connecting variables
- Paths can go in the direction of the arrows or against them
    - With the arrows: causal association
    - Against the arrows: non-causal association
- Paths can be "blocked" or "opened" by conditioning on[1] variables along those paths

[1] sometimes called adjusting or controlling for

Three technical slides

Probabilistic model of a DAG

The probabilistic nature of a DAG is implied *conditional independence*.

Say we have n variables $X_1, \ldots, X_n$. We can always write

$$p(x_1, \ldots, x_n) = \prod_i p(x_i | x_1, x_2, \ldots, x_{i-1})$$

(the chain rule). We are interested in the case where each x_j is dependent on only some of the other variables:

$$p(x_i | x_1, \ldots, x_{i-1}) = p(x_i | pa_i)$$

where PA_i is a subset of the remaining variables, called the "parents" of X_i.

Graphical example

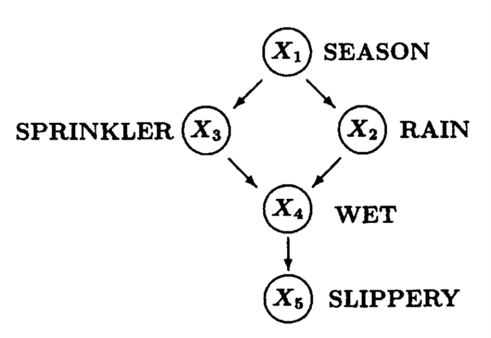


Figure from *Causality* by Judea Pearl

Graphical example

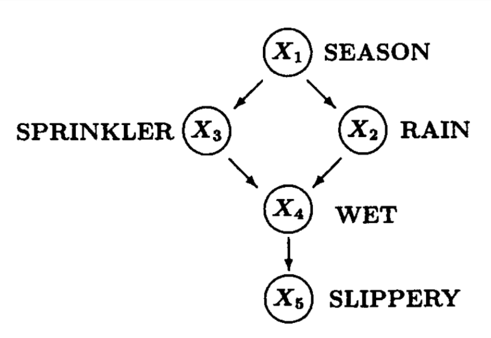


Figure from *Causality* by Judea Pearl

$$P(x_1, \ldots, x_5) = P(x_1)P(x_2|x_1)P(x_3|x_1)P(x_4|x_2, x_3)P(x_5|x_4)$$

Functional model of a DAG

Functional model: each variable X_i satisfies an equation in the graph:

$$x_i = f_i((pa_i), u_i)$$

where

- f_i is a function of the parent variables
- pa_i refers to the parent nodes of X_i in the graph
- u_i represents the unobserved and/or random components of the model

Controlling “flows”

When we’re trying to estimate the effect of one variable on another:

- Control “flow” of information along paths
- Information flows along or against arrows
- Including a variable in the regression (conditioning on a variable) can either “block” or “open” paths

Whether a path is opened or blocked depends on the direction of the arrows connected to each variable

Three basic paths

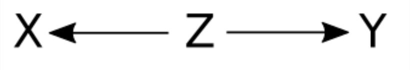
Three basic paths

In a DAG, information flows along paths (both with and against the arrows)

- A path from X to Y can be a direct path – an arrow between X and Y. Or it can be an indirect path $X \leftrightarrow Z \leftrightarrow Y$ (or a concatenation of several of these).
- Indirect paths can lead to confounding / spurious associations; to deal with this, we need to classify the different types of indirect paths
- All such paths can be built out of three-node paths, so we just need to examine the different types of these we could write

The "fork" path

The *fork* is the form most commonly thought of as "confounding": X and Y are confounded by their common cause, Z:



A statistical association exists between X and Y because they are both influenced by Z.

Example: X is ice cream sales; Y is drowning deaths; Z is temperature

The "fork" path

The *fork* is the form most commonly thought of as "confounding": X and Y are confounded by their common cause, Z:

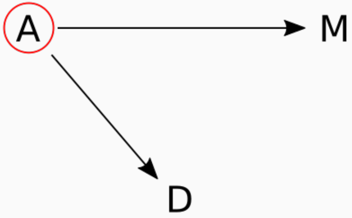
$$X \longleftarrow Z \longrightarrow Y$$

Conditional independence:

- DAG property means: conditional on Z, X and Y are independent.
- So, condition/stratify/control on Z to block the path and estimate effect of X on Y

Fork path in the divorce example

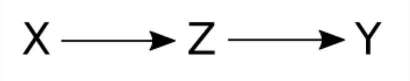
In our example of divorce rates:



- A sits on a fork path joining M and D
- If we *don't* condition on A, the path is open and there is an observed association between M and D
- If we do, the path is blocked and the observed association disappears

The "chain" path

The *chain* is a similar-looking form, where Z sits in the middle of a causal path:



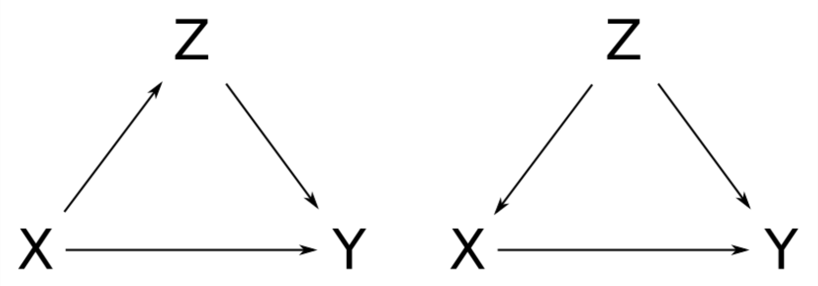
Typical case: Z is an effect of X that mediates the effect on Y

Example: X is pesticide application; Z is the pest population; Y is crop yield.

Controlling for Z blocks information flow along the path.

When the data can't tell you

Multiple paths: should you include the variable Z or not?



- The data cannot tell you the difference between these, because they imply the same set of conditional independences
- Choice of DAG is a model assumption – based on prior information, domain expertise
  - Inferences are conditional on the model you use

Simpson's paradox

Very famous phenomenon: an observed association reverses direction after conditioning on another variable

Often framed as: population-wide association is reversed after stratification on sub-populations

- Kidney stones: treatment A succeeds more often than treatment B, but treatment B performs better on large stones and on small stones
- Graduate admissions: men admitted to graduate programs at a higher rate, but women more successful in admission to every individual department

Simpson's paradox: example

Fake data about a drug:

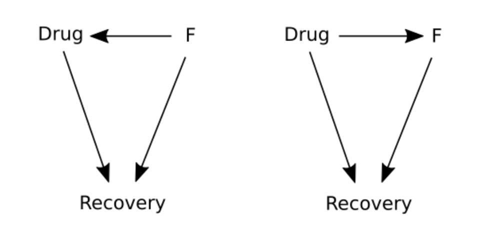
Combined	Recovered	Not recovered	% Recovery
Drug	20	20	50%
No drug	16	24	40%
$F = 1$	Recovered	Not recovered	% Recovery
Drug	18	12	60%
No drug	7	3	70%
$F = 0$	Recovered	Not recovered	% Recovery
Drug	2	8	20%
No drug	9	21	30%

The variable F is a potential confound; this data displays Simpson's paradox.

Question: does the drug help people recover?

Two DAGs

The data from the previous slide could be generated by processes represented by either of the following causal DAGs:



But the inference we should make about the effectiveness of the drug is very different in each case!

Situation 1: sex and compliance

Situation 1: F is a fork variable, influencing both recovery and treatment

Example:

- F is sex (male or female)
- the drug negatively influences recovery
- men are both less likely to recover *and* less likely to take the treatment, so a positive association between treatment and recovery is observed in the pooled data

Action: to estimate causal effect of treatment, condition on the fork; conclude the treatment is bad

Situation 2: post-treatment effect

Situation 2: F is a treatment effect that mediates the recovery (a chain)

Example:

- F is blood pressure (high or low)
- Mechanism by which the drug works is by reducing blood pressure
- Controlling for post-treatment effect masks influence of the drug

Action: to estimate causal effect of treatment, don't condition on the post-treatment effect; conclude the treatment is good

The "collider" path

The third form is the *collider* or inverted fork, and it behaves quite differently:

$$X \longrightarrow Z \longleftarrow Y$$

In contrast to the fork or chain, information flows through the collider only when it *is* observed / controlled; controlling *unblocks* the path.

Heuristic example

$$X \longrightarrow Z \longleftarrow Y$$

X: switch state on/off Z: light bulb on/off Y: power working/not working

The presence of power and the state of the switch are independent; but,

- turn on the switch and observe the light: it's off
- is the power working?

The explaining-away effect

This property of colliders is responsible for a sometimes counterintuitive effect:

- "explaining away": observing one of the common causes
- Berkson's paradox: conditioning on a variable can introduce a spurious association

They're really the same effect; explaining away common in AI/ML; Berkson's paradox in statistics

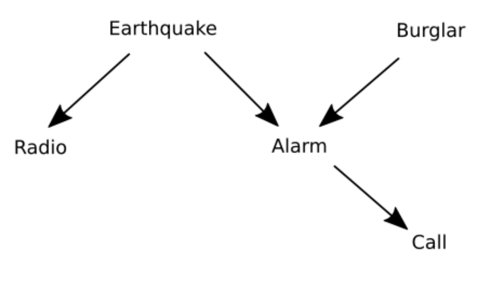
Explaining away: the burglar alarm

From Judea Pearl by way of David MacKay:

> *Fred lives in Los Angeles and commutes 60 miles to work. Whilst at work, he receives a phone-call from his neighbour saying that Fred's burglar alarm is ringing. What is the probability that there was a burglar in his house today? While driving home to investigate, Fred hears on the radio that there was a small earthquake that day near his home. 'Oh', he says, feeling relieved, 'it was probably the earthquake that set off the alarm'. What is the probability that there was a burglar in his house?*

Explaining away: the burglar alarm

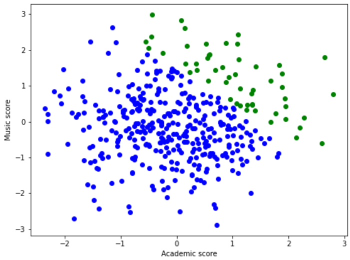
A DAG for the burglar alarm problem, showing the collider:



The alarm sits at a collider.

Conditioning on colliders creates confounding

The spurious-association effect of conditioning on a collider:



Berkson's paradox a.k.a. *selection bias*

Recent example

Recent example: risk factors for COVID-19

- Early studies of COVID-19 were based on observational studies
- Testing availability was low

This led to the potential for collider bias. Why?

- Any study of confirmed COVID-19 cases can only be applied to people who are tested (still true!)
- Data sets are implicitly conditional on having been tested

Examining the effect of smoking

Example study: does smoking protect against severe disease?

- early observational data suggested a negative association between smoking and probability of severe COVID-19
- this is a surprising finding!

Implicit collider: who is getting tested in early 2020?

Examining the effect of smoking

In the early stages of the pandemic, two groups of people were tested most commonly:

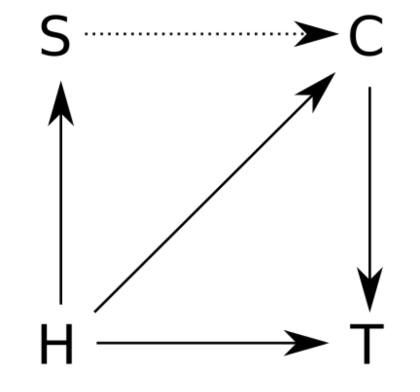
- people with severe disease
- healthcare workers

Conditioning on testing introduces an association between these two traits

Griffith et al., “Collider bias undermines our understanding of COVID-19 risk and severity” (Nature, 12 Nov 2020)

A DAG for the smoking confound

Here is a DAG:



The backdoor criterion

A (possibly undirected) path p through a DAG G is said to be *d-separated* or *blocked* by a set of nodes Z if:

1. p contains a chain $X_i \to M \to X_j$ or fork $X_i \leftarrow M \to X_j$ such that $M \in Z$; or,
2. p contains a collider $X_i \to M \leftarrow X_j$ such that $M \notin Z$ and no descendent of M is in Z.

(Why the descendant property? Look back at the burglar alarm.)

The d-separation (blocking) definition for paths leads to another definition, for sets of variables.

The backdoor criterion

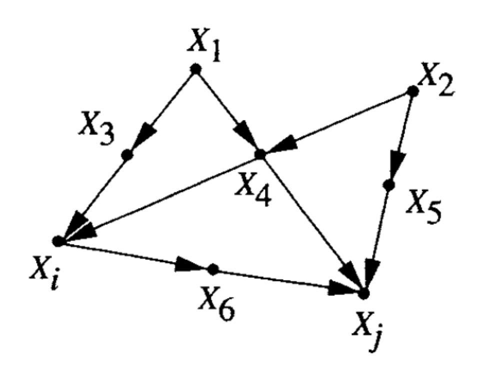
A related definition:

Definiton
A set of variables Z satisfies the backdoor criterion with respect to an ordered pair of variables (X_i, X_j) in G if:

1. *no node in Z is a descendent of X_i; and,*
2. *Z blocks every path from X_i to X_j that contains an arrow into X.*

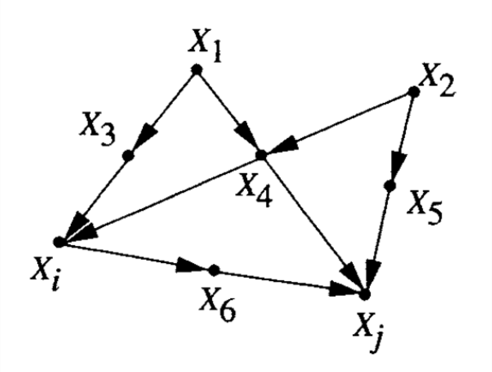
To estimate the causal effect of X on Y, condition on a set of variables satisfying the backdoor criterion with respect to (X, Y).

Example



Which variables satisfy the backdoor criterion?

Example



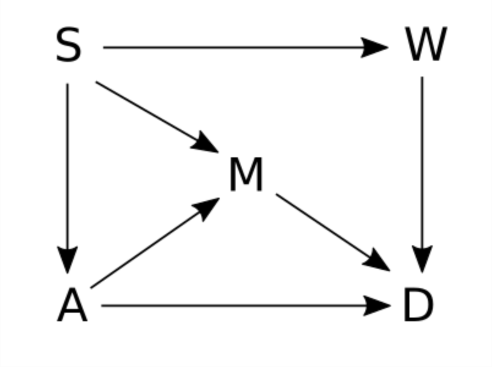
Which variables satisfy the backdoor criterion?

- $\{X_3, X_4\}$ or $\{X_4, X_5\}$
- Not $\{X_4\}$ (doesn't block every backdoor path), nor $\{X_6\}$ (descendent of X_i)

Return to the Waffle House

A bigger DAG from the Waffle House example, including:

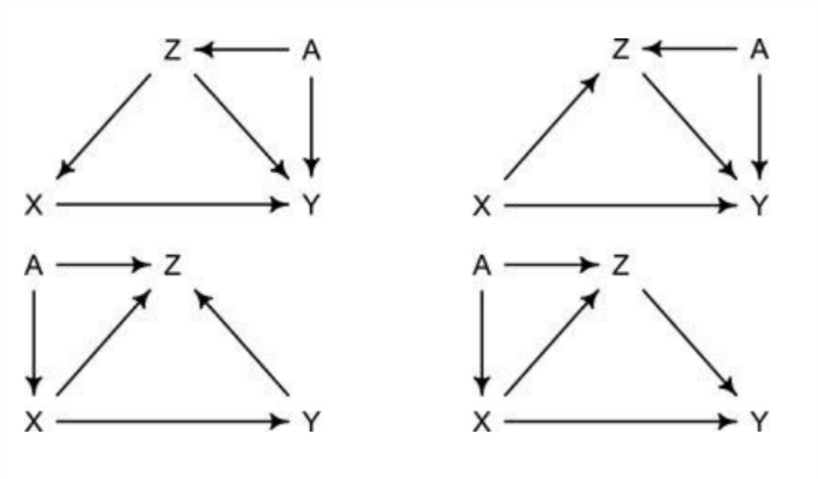
- W: number of Waffle Houses in the state
- S: Indicator variable for South



To estimate the direct effect of W on D, what do we condition on?

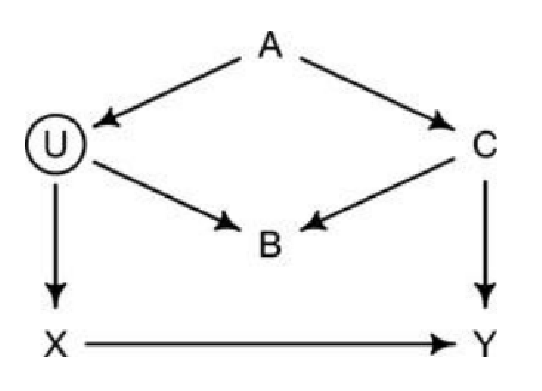
Group exercise

For each DAG, which variable should be conditioned on to estimate total causal influence of X on Y?



Demonstrative example

Another DAG:



We want to estimate $X \to Y$. What should we condition on? (U is unobserved; we can't use it.)

Fake data simulation

To demonstrate the effect, let's use a fake data simulation:

$$
\begin{aligned}
A &\sim \mathrm{Normal}(0, 1) \\
U &= A + \mathrm{noise} \\
C &= -2A + \mathrm{noise} \\
B &= -2C + 3U + \mathrm{noise} \\
X &= U + \mathrm{noise} \\
Y &= 1.5X + C + \mathrm{noise}
\end{aligned}
$$

In all cases, $\mathrm{noise} \sim \mathrm{Normal}(0, 0.1)$

What do we seek?

Before we run any regressions, what *should* we see?

- All paths $X \to Y$ except the direct one are non-causal (backdoor)
- We want to estimate the direct (causal) effect
- We know $Y = 1.5X +$ other effects
- An unconfounded estimate of $\hat{Y} = \beta_x X +$ others should have $\beta_x \approx 1.5$

Let's see some regression results

- Regression including only X:

	mean	sd	hdi_3%	hdi_97%
alpha	0.070	0.031	0.011	0.125
beta_x	-0.465	0.033	-0.525	-0.403
sigma	0.302	0.023	0.261	0.342

Let's see some regression results

- Regression including only X:

	mean	sd	hdi_3%	hdi_97%
alpha	0.070	0.031	0.011	0.125
beta_x	-0.465	0.033	-0.525	-0.403
sigma	0.302	0.023	0.261	0.342

- Condition on A:

	mean	sd	hdi_3%	hdi_97%
alpha	0.027	0.013	0.006	0.054
beta_x	1.461	0.093	1.290	1.636
beta_a	-1.948	0.093	-2.124	-1.782
sigma	0.126	0.009	0.109	0.145

Let's see some regression results

- Regression including only X:

	mean	sd	hdi_3%	hdi_97%
alpha	0.070	0.031	0.011	0.125
beta_x	-0.465	0.033	-0.525	-0.403
sigma	0.302	0.023	0.261	0.342

- Condition on A:

	mean	sd	hdi_3%	hdi_97%
alpha	0.027	0.013	0.006	0.054
beta_x	1.461	0.093	1.290	1.636
beta_a	-1.948	0.093	-2.124	-1.782
sigma	0.126	0.009	0.109	0.145

- Condition on A and B:

	mean	sd	hdi_3%	hdi_97%
alpha	0.020	0.011	-0.002	0.040
beta_x	1.893	0.110	1.687	2.107
beta_a	-1.044	0.190	-1.428	-0.710
beta_b	-0.192	0.036	-0.263	-0.129
sigma	0.112	0.008	0.097	0.127

Let's see some regression results

- Using X alone: badly confounded
- Using X, A: good estimate of $X \to Y$
- Using X, A, B: confounded again

What other options do we have?

More variables

- Condition on C:

	mean	sd	hdi_3%	hdi_97%
alpha	0.018	0.009	0.002	0.034
beta_x	1.539	0.065	1.410	1.650
beta_c	1.017	0.033	0.957	1.076
sigma	0.087	0.006	0.075	0.098

More variables

- Condition on C:

	mean	sd	hdi_3%	hdi_97%
alpha	0.018	0.009	0.002	0.034
beta_x	1.539	0.065	1.410	1.650
beta_c	1.017	0.033	0.957	1.076
sigma	0.087	0.006	0.075	0.098

- Use everything:

	mean	sd	hdi_3%	hdi_97%
alpha	0.019	0.009	0.003	0.036
beta_x	1.559	0.100	1.372	1.756
beta_a	-0.233	0.186	-0.593	0.102
beta_b	-0.002	0.038	-0.076	0.067
beta_c	0.905	0.121	0.686	1.138
sigma	0.087	0.007	0.075	0.099

Using everything works because the collider path is blocked at C; but note precision

Unobserved variables in a DAG

The previous example had an unobserved variable, U:

- If a variable is unobserved, then we can't stratify/adjust for it in the regression
- ...but that doesn't mean we are off the hook for thinking about it!
- Unobserved variables can confound estimates
- Unobserved variables can form colliders

Sometimes this means there is no way to make the estimates that you want!

Summary

Today:

- Structure of DAGs
- Blocking and opening paths
- Criteria for variable inclusion

Next time:

- Going further with multiple regression

HW2 up soon, due next Friday – first steps with PyMC3 and quap