DAGs and the backdoor criterion

ISTA 410 / INFO 510: Bayesian Modeling and Inference

U. of Arizona School of Information

September 14, 2020

Outline

- DAGs as probabilistic models
- How associations “flow” through DAGs
- Causal inference and paradoxes
    - Aside: multicollinearity

DAGs as probabilistic models

Recap: what is a DAG?

What is a DAG?

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

What are we using DAGs for? Probabilistic models, on two levels:

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

Reference

BDA mentions causal inference and gives some details, but doesn't use DAGs

Main reference: Judea Pearl, *Causality* (available online through UofA library)
Chapter/section references:

- DAGs as probabilistic models: Chapter 1
- The backdoor criterion: Section 3.3
- Simpson's paradox and confounding: Chapter 6

Four technical slides

Probabilistic model of a DAG

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 (Markovian) parents of X_i.

Probabilistic model of a DAG

Why Markovian? Because the restriction that

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

is a sort of Markov property: the distribution of X_i depends only on its immediate parents.

If the joint probability distribution of all the variables obeys this Markov property with respect to the parent relationships described by the graph:

$$p(x_1, x_2, \ldots, x_n) = \prod_i p(x_i|pa_i)$$

then the probability distribution is said to be *compatible* with the graph.

Graphical example

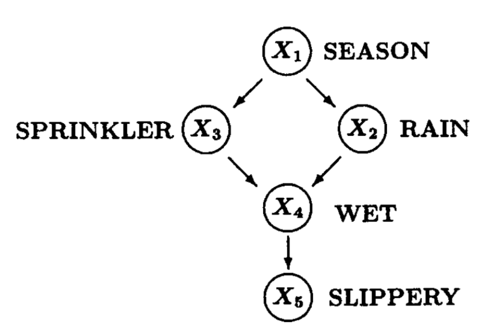


Figure from *Causality*

Graphical example

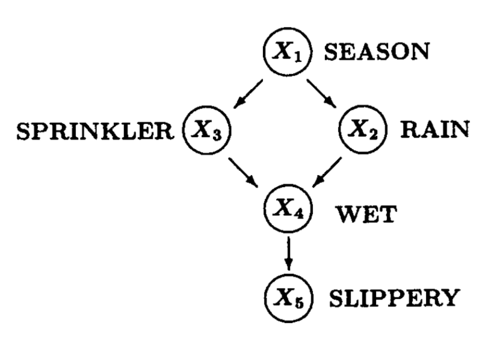


Figure from *Causality*

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

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

Special case: linear structural equation models:

$$x_i = \sum_j \alpha_j x_j + u_i$$

The hierarchical models we'll look at next week are restricted versions of this idea.

Three elemental paths

Three basic paths

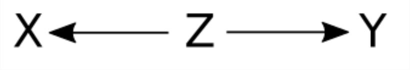
In a Bayesian network, information flows along paths (both with and against the arrows).

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 lead to confounding / spurious associations; to deal with this, we need to classify the different types of indirect paths.

The "fork" path

The *fork* is the form most students learn as the sole definition of "confounding" in introductory classes: X and Y are confounded by their common cause, Z:

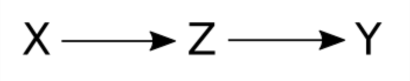


A statistical association exists between X and Y because they are both influenced by Z; if there is no arrow from X to Y, this association will be eliminated by controlling for Z. That is, controlling for Z blocks information flow along the path.

To estimate the causal effect of X on Y, control/stratify for Z.

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.

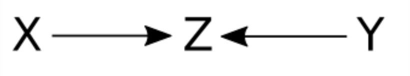
Controlling for Z blocks information flow along the path.

The "collider" path

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

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: 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

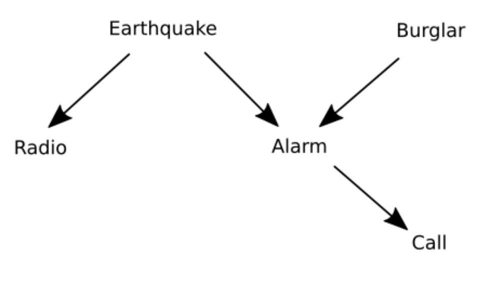
Explaining away: the burglar alarm

From Pearl by way of 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

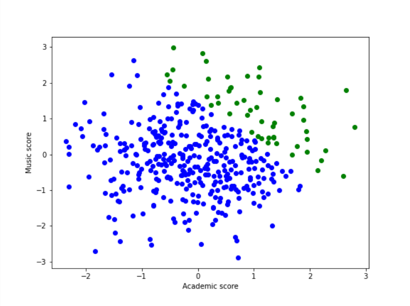
A DAG for the burglar alarm problem, showing the collider:



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*

The backdoor criterion

d-separation

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

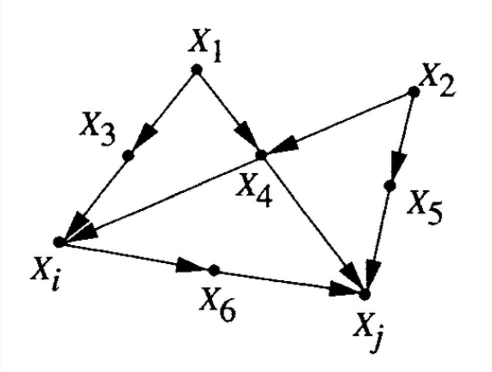
A related definition:

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.*

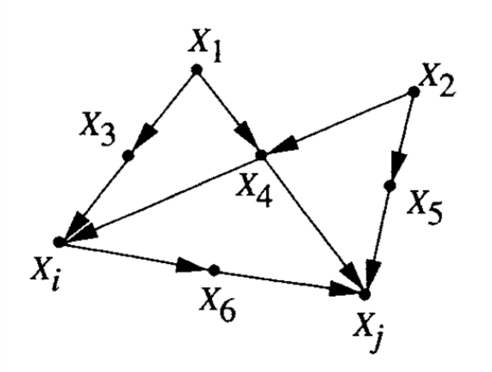
To estimate the causal effect of X on Y, condition on a set of variables satisfying the backdoor criterion with respect to (X, Y).

Example



Which variables satisfy the backdoor criterion?

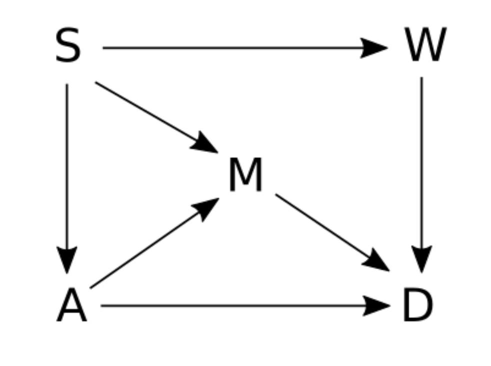
Example



Which variables satisfy the backdoor criterion?

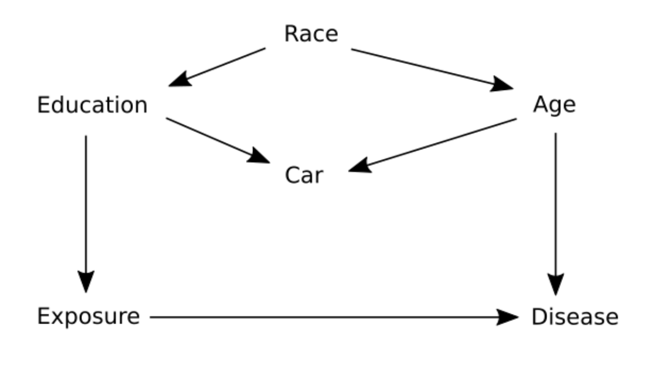
- $\{X_3, X_4\}$ or $\{X_4, X_5\}$
- Not $\{X_4\}$ (doesn't block every backdoor path), nor $\{X_6\}$ (descendent of X_i)

Return to the Waffle House



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

Example



What to condition on to estimate Exposure $\rightarrow$ Disease?

Aside: multicollinearity

Multicollinearity in regression

Multicollinearity is a problem that appears in linear regression when two or more predictors are not linearly independent – when they are tightly correlated with one another.

Multicollinearity is bad:

- models overly complex
- numerical problems in variable fitting
- uninformative inferences

Heuristic example

Problem: predict human height from leg length

- We could reasonably expect leg length to be a strong, but not perfect predictor
- Total height $\approx$ leg length $+$ torso length $+$ const.

What if we have more granular data: measurements on both left and right legs?

Heuristic example

Model based on artificial data:

Left leg only:

	mean	sd	hpd_3%	hpd_97%
betal	2.151	0.049	2.065	2.240
alpha	5.909	3.593	-0.921	12.224

Right leg only:

	mean	sd	hpd_3%	hpd_97%
betar	2.158	0.048	2.070	2.251
alpha	5.516	3.576	-1.820	11.498

Both legs:

	mean	sd	hpd_3%	hpd_97%
betal	-0.627	1.467	-3.279	2.232
betar	2.785	1.470	-0.007	5.509

Heuristic example

- We can precisely estimate the effect of left leg length on height
- We can precisely estimate the effect of left leg length on height
- If we try to estimate both at once, we lose all precision

The problem: once we control for one leg length, all that's left in the other observation is noise

Variable selection

To deal with multicollinearity, there are several tools for *model comparison*, to choose which variables should go in a model

- e.g., stepwise selection with information criteria (we'll cover these criteria later)

While DAGs can be seen partially as a tool for fighting multicollinearity, it's not exactly the same:

- Variable selection: numerical tools based on fit of predictions to data. Attempt to maximize out-of-sample prediction accuracy.
- DAGs: external causal models, related to but distinct from the data. Attempt to estimate causal effects.

More on confounding and Simpson's paradox

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 every sub-population

- 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:

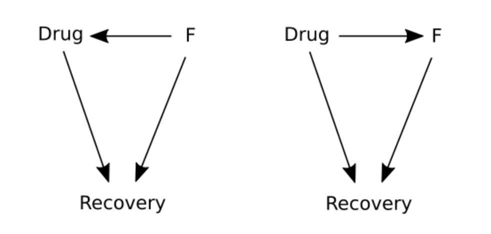
Combined	Recovered	Not recovered	% Recovery
Drug	20	20	50%
No drug	16	24	40%
$F = 1$	Recovered	Not recovered	% Recovery
Drug	18	12	60%
No drug	7	3	70%
$F = 0$	Recovered	Not recovered	% Recovery
Drug	2	8	20%
No drug	9	21	30%

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: gender and compliance

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

Example:

- F is gender
- 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)
- One 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

Reference

BDA mentions causal inference and gives some details, but doesn't use DAGs

Main reference: Judea Pearl, *Causality* (available online through UofA library)
Chapter/section references:

- DAGs as probabilistic models: Chapter 1
- The backdoor criterion: Section 3.3
- Simpson's paradox and confounding: Chapter 6

Summary

Summary:

- DAGs are probabilistic and/or functional models of dependency in multi-variable systems
- Confounding and statistical "paradoxes" can be modeled by information flow through the graph

Next time:

- More DAGs and backdoors
- Unobserved variables
- Example(s)