Inference and learning in causal models: Lecture 3

Tsinghua Logic Summer School

Tadeg Quillien

University of Edinburgh

2026

Learning causal structure

Learning causal structure from data

- Suppose we have a system with two variables A, B, and no other variables (apart from noise variables).
- Can we learn the causal structure (i.e. the DAG) of the model?
- I.e. is it the case that $A \rightarrow B$, that $A \leftarrow B$, or that the two variables are not causally connected?

If we can perform experiments

Experimental data allow us to estimate the interventional distributions:

- $P(\mathbf{A}|\mathrm{Do}(\mathbf{B}))$,
- $P(\mathbf{B}|\mathrm{Do}(\mathbf{A}))$.

And this gives us what we need:

- If $P(\mathbf{B}|\mathrm{Do}(\mathbf{A})) \neq P(\mathbf{B})$ then $A \rightarrow B$,
- If $P(\mathbf{A}|\mathrm{Do}(\mathbf{B})) \neq P(\mathbf{A})$ then $A \leftarrow B$,
- If $P(\mathbf{B}|\mathrm{Do}(\mathbf{A})) = P(\mathbf{B})$ and $P(\mathbf{A}|\mathrm{Do}(\mathbf{B})) = P(\mathbf{A})$ then the variables are disconnected.

If we can collect observations but can't perform experiments

Then we can estimate the joint distribution $P(\mathbf{A}, \mathbf{B})$, but we don't have access to the interventional distributions.

Our inferences are now more limited:

- If $A \perp B$ then the variables are disconnected,
- If $A \not\perp B$, then either $A \rightarrow B$ or $A \leftarrow B$.

Two lessons here:

- Even observational data contain some information about causal structure.
- But some causal structures cannot be differentiated on the basis of observational data alone.
  - We will say that these causal structures are **Markov-equivalent**.

Markov equivalence

Definition

Two DAGs G_1 and G_2 are **Markov-equivalent** if:

$$\mathcal{D}(G_1) = \mathcal{D}(G_2),$$

where $\mathcal{D}(G)$ denotes the family of distributions that factorize according to G.

$\rightarrow$ Observational data can allow you to rule out some DAGs, but cannot decide between Markov-equivalent DAGs.

A slightly more complex example

Example

Suppose we don't know which of the following DAGs is the right one:

- G_1: $A \rightarrow B \rightarrow C$
- G_2: $A \rightarrow C \rightarrow B$

We cannot use patterns of unconditional dependence to tell them apart:

- Both DAGs predict that all variables depend on each other.

But we can use patterns of **conditional** dependence!

- G_1 predicts that $A \perp C|B$,
- G_2 predicts that $A \not\perp C|B$.

Remark

There are still some DAGs we can't tell apart, e.g.:

- $A \rightarrow B \rightarrow C$
- $A \leftarrow B \leftarrow C$

Important Remark

Throughout these lectures we make an idealized assumption:

- You always have enough data to accurately estimate observational probability distributions,
- e.g. you always have enough data to establish for sure whether X and Y are (conditionally or unconditionally) independent.

Of course in the real world this is rarely the case!

- Entire statistics books are written about statistical estimation from limited data.
- Our point is that even if you have solved these problems you still have some work to do to make correct causal conclusions.

Learning causal models (in general)

A general Bayesian perspective on learning causal models

We could 'simply' use Bayes' rule to compute the probability of different causal models $\mathcal{M}$ given the collected data d:

$$Pr(\mathcal{M}|d) = \frac{Pr(d|\mathcal{M})Pr(\mathcal{M})}{\sum_i Pr(d|\mathcal{M}_i)Pr(\mathcal{M}_i)},$$

where the data d can include observations and / or interventions.
Problem with this approach:

- There are very many possible causal models, $\mathcal{M}_i$,
- Computing $P(d|\mathcal{M}_i)$ **for all of them** is in general intractable.

A general Bayesian perspective on learning causal models

- There are very many possible causal models, $\mathcal{M}_i$,
- Computing $P(d|\mathcal{M}_i)$ **for all of them** is in general intractable.

This raises the questions:

- How do we even **represent** the space of all possible hypotheses about the correct causal model?
- We must use an approximation for computing $P(\mathcal{M}|d)$. Can we use sampling for this approximation? What sampling algorithm can we use?

See next lectures!