Section 12

Lecture 4

Plan for lecture 4

- Target trials (briefly)
- Structural Equation Models
- Causal graphs
  - Bayesian networks
  - Link to structural equations
  - D-separation
  - Examples

Section 13

Target trial

The target trial

- We have argued that contrast between average counterfactual outcomes under different treatments are often of substantial interest.
- We have also clarified that conducting an experiment guarantees identification of a causal effect. However, conducting an experiment is not always feasible.
- For each causal effect of interest, we can conceptualize a (hypothetical) randomised experiment to quantify it. This hypothetical randomised experiment is called the **target experiment** or **target trial**.
- Being explicit about specifying the target trial forces us to be explicit about the causal question of interest. We ask the question: "What randomised experiment are you trying to emulate?"

Specification of the target trial

To make a causal question practically interesting and useful, it is important to clarify the following, which is part of the specification of the target trial:

- Target population (eligibility criteria).
- Interventions (the treatment strategies).
- Outcome (what is the outcome and when will the outcome be measured)
- Statistical analysis (application of estimators and their statistical properties).

Also clarifies how the claims made can be falsified in the future (in principle), by conducting the target trial. This fits with a positivist (Popperian) view of science.

Motivation

- You have seen that conditional dependencies are hard to interpret.
  - Death penalty example
  - GRE example (you will see this again today)
- We have also seen that (average) causal effects are identified by design in experiments, but also can be identified under assumptions (exchangeability, consistency and positivity) in an observational study. However, reasoning about *counterfactual* (in)dependencies is at least as hard as observed (in)dependencies.
- We will now introduce graphs to clarify when:
  1. Observed (in)dependencies can be interpreted causally.
  2. Counterfactual independencies are plausible, which can allow identification of causal effects.
- Importantly, graphs allow us to study much more complex and realistic settings than those we have considered so far.

Section 14

Structural equations

Structural equation model

Definition

A structural equation model (SEM) is a model that describes how values are assigned to each variable in a system

Think about nature (God) assigning values to each variable in the system. This describes a generative story of how the data came to be. Or think about each equation representing a physical mechanism that determines the value of the variable on the left (output) from values of the variable on the right (inputs)

We motivate structural equation models (SEMs) with an example (more general theory follows in later slides)

Consider

$$
\begin{aligned}
L &= f_L(U_L) \\
A &= f_A(L, U_A) \\
Y &= f_Y(A, L, U_Y) = Y^{a=A, l=L} = \sum_{a,l} I(A = a, L = l) Y^{a,l}
\end{aligned} \tag{1}
$$

Here U_L, U_A, U_Y are external unmeasured factors that are mutually independent. Here, the generative story is as follows:

- The value of L is determined as a function of the value of U_L as given by the function f_L.
- The value of A is determined as a function of the value of L, U_A as given by the function f_A.
- The value of Y is determined as a function of the value of L, A, U_Y as given by the function f_Y.

We will accompany the structural equations with a picture

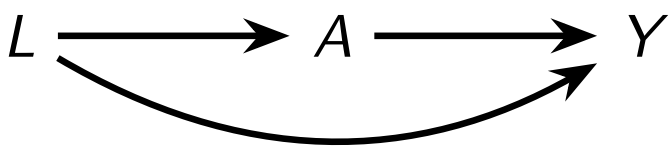
Structural equation models are typically accompanied with a corresponding picture known as a path diagram (as above): that is, a graph which makes explicit the directionality of the underlying process.
For a more compact representation, unmeasured factors that do not determine two or more variables in the system can be left out of the graph (I will repeat this point in later slides, and make the notion more formal).

SEM example (continued)

Consider the SEM $\mathcal{M}$

$$\begin{aligned} L &= f_L(U_L) \\ A &= f_A(L, U_A) \\ Y &= f_Y(A, L, U_Y) \end{aligned} \tag{2}$$

and the graph $\mathcal{G}$,



How does $\mathcal{M}$ induce an observed data distribution over $P(L = l, A = a, Y = y)$ and can this distribution be fully described in some way by simply looking at the graph $\mathcal{G}$?
And how about the distributions under interventions on A, that is, $P(L = l, A = a, Y^a = y)$?

Section 15

Graphs

What is a graph?

Definition (Graph)

A graph $\mathcal{G}$ is a collection of

- Nodes (vertices), $V = \{V_1, V_2, \cdots, V_m\}$.
- Edges $(V_i V_j)$ connecting some of the vertices.

We write $(V_i V_j)$ to denote an edge that connects V_i and V_j.
A **path** is a sequence of edges of the form
$\langle (V_1, V_2), (V_2, V_3), \cdots, (V_{k-1}, V_k) \rangle$,

What is a directed graph?

Definition (Directed Graph)

A directed graph is a graph with a set of nodes and *arrows* connecting some of the nodes. A graph $\mathcal{G}$ is a collection of

- Nodes (vertices) $V = \{V_1, V_2, \cdots, V_k\}$.
- **Directed** edges connecting some of the nodes.

We write $(V_i V_j)_{\rightarrow}$ to denote a directed edge from V_i to V_j.
It is directed, because the graphs A **directed path** is a sequence of edges of the form
$\langle (V_1, V_2)_{\rightarrow}, (V_2, V_3)_{\rightarrow}, \cdots, (V_{k-1}, V_k)_{\rightarrow} \rangle$,
A directed graph has a **cycle** if there exists a path
$\langle (V_1, V_2)_{\rightarrow}, (V_2, V_3)_{\rightarrow}, \cdots, (V_{k-1}, V_k)_{\rightarrow}, (V_k, V_1)_{\rightarrow} \rangle$.
A Directed Acyclic Graph is a directed graph with no cycles.
PS: Now the subscript does not longer indicate an individual. V_1 is now a random variable. From now on, I will use $V_1(\omega)$ when I talk about the value for a particular individual.

Basic definitions

In a DAG $\mathcal{G}$ we define the following sets (parents, children, ancestors and descendants):

- $\mathbf{pa}_G(V_i) \equiv \{V_t : V_t \to V_i \text{ exists in } \mathcal{G}\}$.
- $\mathbf{ch}_G(V_i) \equiv \{V_t : V_i \to V_t \text{ exists in } \mathcal{G}\}$.
- $\mathbf{an}_G(V_i) \equiv \{V_t : V_t \to V_a \to \cdots \to V_j \to V_i \text{ exists in } \mathcal{G}\} \cup V_i$.
- $\mathbf{de}_G(V_i) \equiv \{V_t : V_i \to V_a \to \cdots \to V_j \to V_t \text{ exists in } \mathcal{G}\}$.

Further terminology:

- A path where $V_a \to V_i \leftarrow V_b$ is called a collider path, and here V_i is a collider.
- A path where $V_a \leftarrow V_i \to V_b$ is called a fork.
- A path is *blocked* if it contains a collider. Otherwise it is *open*.
- A DAG is complete if there is an arrow between every pair of nodes.

Topological order with respect to a graph

Definition (Topological order of a DAG)

The nodes $V_1, V_2, \ldots$ follow a topological order relative to a DAG $\mathcal{G}$, if V_j is not ancestor of V_i whenever $j > i$.

Note that topological orders are not necessarily unique, but in the DAG in Figure 129 the only possible topological order is $\langle L, A, Y \rangle$.

Some preliminaries

- Consider a study population Ω.
- Let ω be an element (i.e. unit or individual) in Ω.
- Note that we used subscript i to denote an individual in the first lecture, but now the subscript just indicates a particular random variable, and we write $V_i(\omega)$ when we consider the value for individual ω.
- Consider a discrete random variable V_j.
- Let $V_j(\omega)$ be the value of V_j in ω.
- Let $\mathcal{G}$ be a DAG with nodes $V = \{V_1, V_2, \cdots, V_m\}$.
- We use overlines to denote histories of variables, e.g. $\overline{v}_j \equiv (v_1, v_2, \ldots, v_j) \in \mathcal{V}_1 \times \mathcal{V}_2 \times \cdots \times \mathcal{V}_m$.
- Let $PA_k = \{V_j : V_j \in \mathbf{pa}_G(V_k)\}$. A random vector
- Let $pa_k = \{v_j : V_j \in \mathbf{pa}_G(V_k)\}$ for a $\overline{v} \equiv (v_1, v_2, \ldots, v_m) \in \mathcal{V}_1 \times \mathcal{V}_2 \times \cdots \times \mathcal{V}_m$ A realisation of PA_k.
- From now on I will use $p(v_i \mid v_j)$ to denote conditional densities $P(V_i = v_i \mid V_j = v_j)$.

Section 16

The next slides on Non-Parametric Structural Equation Models give some more details

Non-parametric structural equation model (NSPEM) with respect to a DAG

There exist unknown functions $f_1, \ldots, f_m$ such that the observed ("factual") variables $V_1, \ldots, V_m$ satisfy

$$
\begin{aligned}
V_1 &= f_1(U_1) \\
V_2 &= f_2(PA_2, U_2) \\
V_3 &= f_3(PA_3, U_3) \\
&\vdots \\
V_m &= f_m(PA_m, U_m)
\end{aligned} \tag{3}
$$

where:

- $f_0, f_1, \ldots$ are unknown deterministic functions.
- PA_i is the set of random variables that are parents of V_i.
- $U_0, U_1, \ldots$ are random variables ("*disturbances*" *or*" *errorterms*") (not drawn in the graph). Sometimes called exogenous variables.

NPSEM continues

For any treatment regime $g = (g_{j_1}, \ldots, g_{j_t})$, the counterfactual variables under g are generated by replacing the functions $(f_{j_1}, \ldots, f_{j_t})$ with the functions $(g_{j_1}, \ldots, g_{j_t})$, where $t \leq m$. Below is an illustration. This is called performing recursive substitution.

$$
\begin{aligned}
V_1^g &= f_1(U_1) \\
V_2^g &= f_2(PA_2^g, U_2) \\
&\vdots \\
V_{j_1}^{g+} &= g_{j_1}(PA_{j_1}^g, U_{j_1}) \\
&\vdots \\
V_m^g &= f_m(PA_m^g, U_m) \qquad (4)
\end{aligned}
$$

The superscript "g" indicates that V_i^g is a counterfactual variable (in other words, potential outcome variable). The superscript "$g+$" is given to the variables on which we intervene. A NPSEM requires (3) and (4) to hold.

Some remarks

- Structural: f_k not only generates observed (factual variables), but also variables in other counterfactual worlds where we have done interventions.
- Counterfactual: The variable $V_j^g, j \in \{0, \ldots, m\}$ are called counterfactual variables under treatment regime g.
- A cause (intuitively): A variable A is a cause of a variable Y if an intervention that specifically changes A can lead to a change in Y.

Example: Point intervention

Let the regime g be defined by the intervention that sets V_2 to a.

$$\begin{aligned} V_1^a &= f_1(U_1) \\ V_1^{a+} &= a \\ V_3^a &= f_3(PA_3^a, U_3) \\ &\vdots \\ V_m^a &= f_m(PA_m^a, U_m) \end{aligned} \tag{5}$$

The superscript "a" indicates that V_i^a is a counterfactual variable (or potential outcome variable)
where we have intervened to set a variable, here V_1 (now with a superscript $a+$) to a.

Let's interpret this model, specifically

- Only the arguments to the structural equation determine the value of a node.
  That is, the value of $V_j(\omega)$ does not depend on any other unit ω' in the population.
  (**No interference**)
- Suppose that a unit ω has $PA_k(\omega) = pa_k$. Then, under any intervention g that fixes $PA_k^g = pa_k$ we have that $V_k(\omega) = V_k^g(\omega)$.
  (**Consistency**)
- When PA_k is known, the value of other variables $\overline{V} \setminus PA_k$ do not determine V_k. (**Exclusion restriction**).

The causal inference part is an assumption about the errors!

We must say something about the dependencies between the U's to encode causal relations.

Definition (Independent error model)

A NPSEM wrt. a DAG $\mathcal{G}$ such that $U_0, \ldots, U_M$ are mutually independent.

`This is Pearl's NPSEM-IE`[18]`.`
`"IE" stands for independent errors.`
`NB: The independent error assumption is not really needed, and can be relaxed in the more general FFRCISTG model`[19]
The U_ks represent all other variables that are used by nature, the decision maker or anyone else to determine the value of V_k.

[18] Judea Pearl. *Causality: Models, Reasoning and Inference 2nd Edition*. Cambridge University Press, 2000.

[19] Robins, "A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect".

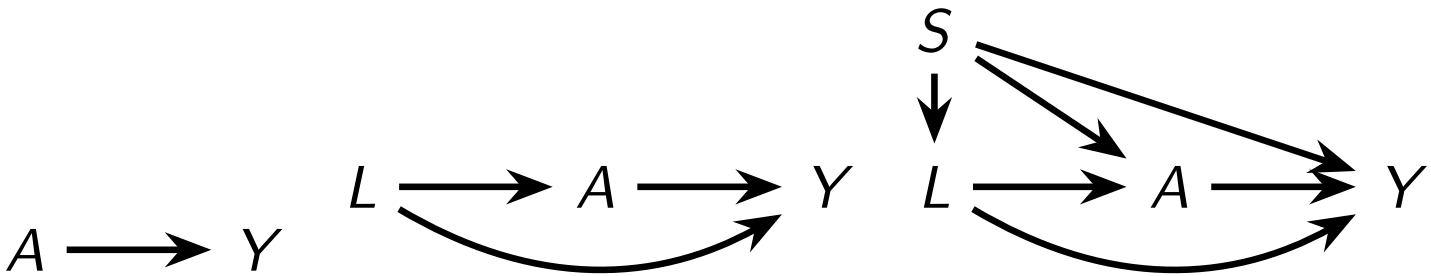
Section 17

Causal graphs

Let us start with some intuition

Suppose I were to explain what is going on in the experiment on heart transplant for my friend who studied literature. I will draw intuitive diagrams that can be formalised as causal graphs. We have previously discussed:

- Completely randomised experiment.
- Conditional randomised experiment.
- Observational study with smoking.



This way of building causal stories using diagrams can be formalised by graphs.

Next step

- In the previous slide, we just made these diagrams to encode qualitative subject matter knowledge.
- However, we shall see that the diagram can be formalised as a causal directed acyclic graph, DAG, which encodes information about causal and non-causal associations in a causal network: it allows us to represent both association and causation in the same graph.

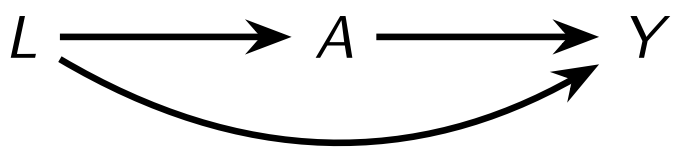
What is the role of causal graphs?

- Graphs help us to reason about independencies; that is, they help us reason about whether certain exchangeability assumptions (conditional independencies) hold.
- This agrees with the mantra: "draw your assumptions before your conclusions".[20]
- Graphs help us to conceptualize problems and have intuitive appeal, also for researchers who are illiterate in math.
- However, the intuitive graphical representations have a **mathematical justification**. Therefore you can translate the intuitive subject-matter expertise (from doctors, economists, social scientists) to precise mathematical statements.
- Graphs allow us to encode *causation* and *association*.

[20] Hernan and Robins, *Causal inference: What if?*

Example

We can now define the graph below as a causal DAG that describes the conditional randomised trial on heart transplants,



where $V_1 = L, V_2 = A, V_3 = Y$.
Here $\mathbf{pa}_G(Y) = (L, A)$.
The graph is complete because there is an arrow between every pair of nodes.

What is a model

Definition (Statistical model)

A statistical model $\mathcal{P}$ is a collection of laws, $\mathcal{P} = \{P_\eta : \eta \in \Gamma\}$.

Here Γ could be an infinite dimensional space. We will typically only restrict ourselves to the space of models with finite variance.

Bayesian network

Definition (Bayesian network)

A Bayesian Network with respect to a DAG $\mathcal{G}$ with nodes $V = (V_1, \ldots, V_m)$ is a statistical model for the random vector V specifying that V belongs to the collection of laws $\mathcal{B}$ satisfying the Markovian factorisation

$$p(v) = \prod_{j=1}^{m} p(v_j \mid pa_j)$$

Here, $p(x \mid y) \equiv P(X = x \mid Y = y)$.
We say that the DAG $\mathcal{G}$ represents the Bayesian Network $\mathcal{B}$.
For any law p in $\mathcal{B}$, we say that p factors according to $\mathcal{G}$, or that p is represented by $\mathcal{B}$.

Causal DAG

Definition (Robins EPI 207)

A causal model associated with a DAG has to satisfy the criteria below:

1. The lack of an arrow from node V_i to V_j can be interpreted as the absence of a direct causal effect of V_i on V_j (relative to the other variables on the graph).
2. Any variable is a cause of all its descendants. Equivalently, any variable is caused by all its ancestors.
3. All common causes, even if unmeasured, of any pair of variables on the graph, are themselves on the graph.
4. The Causal Markov Assumption (CMA): The causal DAG is a statistical DAG, i.e., the distribution of V factors.
5. Because of the causal meaning of parents and descendants on a causal DAG, the Causal Markov Assumption is equivalent to the statement:
    - Conditional on its direct causes (i.e., parents), a variable V_i is independent of any variable it does not cause (i.e., any nondescendant).

Section 18

The next slides were not discussed explicitly in the lectures but give some more justification and background on graphs and NPSEM's

Absence of common causes in the DAG (point 3)

The arguments here are analogous to the motivating example for the simple graph with A, L, Y and smoking S.

- Remember that U_k represents all other variables that determines (causes) V_k except the parents PA_k.
- Suppose that there exists a variable C that is a direct determinant of V_k relative to the DAG (i.e. it does not only determine V_k through variables in the DAG).
- This means that $U_k = m_k(C, U_k^*)$ for some function m_k.
- Suppose that C is also a direct determinant of a node j (but C is still not in the DAG).
- Thus, $U_j = m_j(C, U_j^*)$ for some function m_j.
- Thus, $U_k \not\perp\!\!\!\perp U_j$.

Factorisation of the NPSEM-IE (point 4)

Argument for Markov factorisation of causal model wrt. a DAG

$$p(v) = \prod_{j=1}^{m} p(v_j \mid pa_j).$$

Proof.

Consider $p(v_j \mid \overline{v}_{j-1})$ for any $j \in \{0, \ldots, m\}$. Here pa_j are the parents of v_j.

$$\begin{aligned}
&p(v_j \mid \overline{v}_{j-1}) \\
&= p(f_{v_j}(PA_j, U_{v_j}) = v_j \mid \overline{V}_{j-1} = \overline{v}_{j-1}) \\
&= p(f_{v_j}(pa_j, U_{v_j}) = v_j \mid \overline{V}_{j-1} = \overline{v}_{j-1}) \\
&= p(f_{v_j}(pa_j, U_{v_j}) = v_j \mid f_{v_{j-1}}(pa_{j-1}, U_{v_{j-1}}) = v_{j-1}, \ldots, f_{v_1}(pa_1, U_{v_1}) = v_1) \\
&= p(f_{v_j}(PA_j, U_{v_j}) = v_j \mid PA_j = pa_j).
\end{aligned}$$

□

No restrictions on $p(v)$ imposed by the NPSEM-IE

The only restriction imposed on the *observed* law is the factorisation

$$p(v) = \prod_{j=1}^{m} p(v_j \mid pa_j).$$

Proof.

Any further restriction must be a restriction on the form of $p(v_j \mid pa_j)$ for any $j \in \{0, \ldots, m\}$. But

$$P(V_j = v_j \mid PA_j = pa_j) = P(f_{v_j}(pa_j, U_{v_j}) = v_j),$$

and we have not put any restrictions on the marginal density of U_{v_j}. □

Factorisation of the nodes V

Lemma

If V follows a NPSEM-IE, then for any $p(\overline{v}_{j-1})$ with $p(\overline{v}_{j-1}) > 0$ we have that $p(v_j \mid \overline{v}_{j-1}) = p(v_j \mid pa_j)$ and therefore the joint density factorizes as

$$p(v) = \prod_{j=1}^{m} p(v_j \mid pa_j).$$

This factorisation is the only restriction that the causal model implies on the law of the observed data.

Thus, in our example from slide 140, the observed law factorizes as

$$p(v) = p(l, a', y) = p(l)p(a' \mid l)p(y \mid a', l),$$

which means that here we put absolutely no restrictions on the law $p(v) \equiv P(V = v)$. You do not have to prove this.

Markov equivalence classes

Definition (Markov equivalence class)

A Markov equivalence class is a set of DAGs that encode the same set of conditional independencies.

Example of markov equivalent DAGs:

Implication: We cannot use data alone to distinguish between causal graphs.

Linear structural equation example

We have not imposed any parametric assumptions so far. However, just for the illustration, suppose we have a (partially) linear structural equation model with two variables satisfying

$$\begin{aligned} A &= f(U_A) \\ Y &= \alpha + \beta A + U_Y \end{aligned} \tag{6}$$

This structural equation model implies that the individual level causal effects is $Y^{a=1} - Y^{a=0} = \beta$!
We conclude that the linear equation model relies on extremely strong assumptions that usually will be implausible. In this course, we will not rely on such assumptions.

Modified non-parametric example

A different SEM $\mathcal{M}$

$$\begin{aligned} L &= f_L(U_L) \\ A &= f_A(L, U_A) \\ Y &= f_Y(A, U_Y) \end{aligned} \tag{7}$$

and the graph $\mathcal{G}$,

- Encodes that, changes in L leaves Y unchanged, provided that U_Y and A remain constant.
- Does this graph encode any restrictions on the distribution of (L, A, Y)?
  `We will formally study what kind of restrictions the structural models involve`