Section 14

Lecture 4

Plan for today

- Look at examples of causal graphs.
  - Explain that an association may or may not be interpreted as a causal effect.
  - Illustrate that the structural equations tell us something about the law of counterfactual variables.
- Define SWIGs and study properties of SWIGs.
  - factorization and modularity.
  - Use SWIGs for identification.
  - Use SWIGs for identification in the presence of hidden variables.

Causal DAG

Definition (Robins EPI 207)

A causal model associated with a DAG satisfies:

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).

Absence of common causes in the NPSEM-IE and 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 exert direct effects 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, $U_k \not\perp\!\!\!\perp U_j$.

Factorization of the NPSEM-IE (point 4)

Argument for Markov factorization 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

We have seen from Slide 107 that the only restriction imposed on the observed law is the factorization

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

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

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

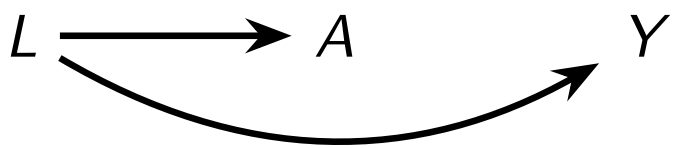
and we have not put any restrictions on the marginal density of U_{v_j}. □

So we have an algorithm for creating causal graphs

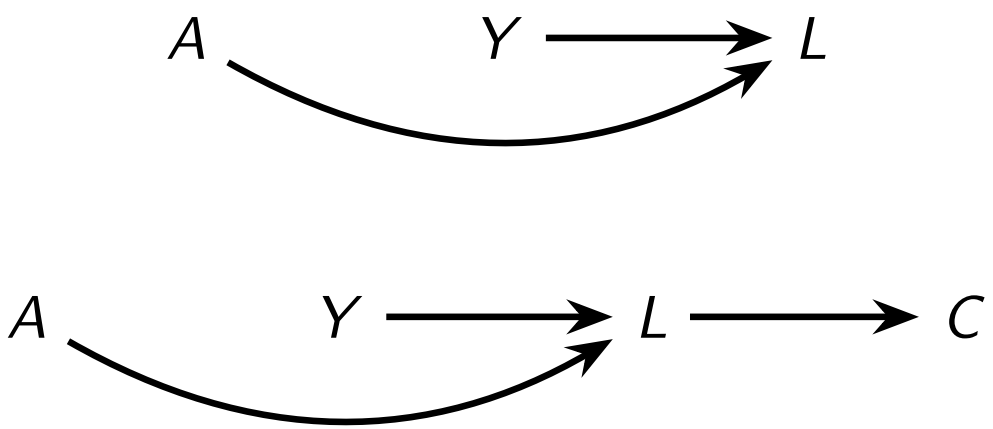
We can create a causal DAG by:

1. Draw nodes for the exposure A and the outcome Y of interest.
   - Draw an arrow from A to Y.
2. If there exists a common cause C of A and Y, write C in the graph.
   - Draw arrows from C to A and from C to Y.
     These common causes must be drawn, even if they are unmeasured.
3. If there exists a common cause C' of any pair $W, W' \in (C, A, Y)$, write C' in the graph.
   - Draw arrows from C' to W and from C' to W'.
4. Continue in this way until there are no common causes...

Carrying a lighter A and the risk of lung cancer Y



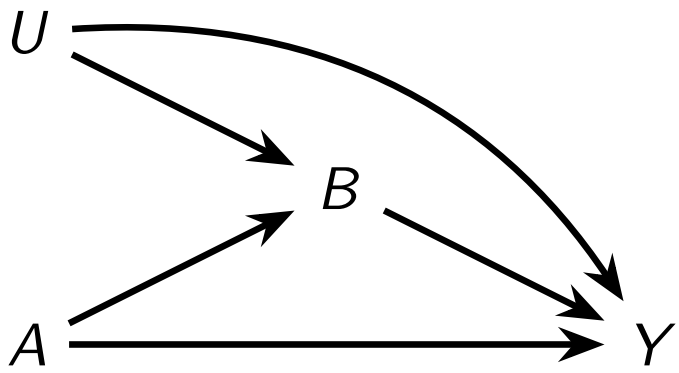
A gene A that causes heart disease L but not smoking Y, where C is taking aspirin (A cardiovascular drug)



Example: Birth weight paradox

- Birth weight predicts infant mortality.
- Investigators often stratify on birth weight when evaluating the effect of maternal smoking on infant mortality.
- Among infants with low birth weight, the mortality rate ratio for smoke exposed infants versus non-exposed infants is 0.79 (95% CI: 0.76, 0.82).
- This birth weight paradox has been a controversy for decades.
- One suggestion is that the effect of maternal smoking is modified by birth weight in such a way that smoking is beneficial for LBW babies.
- Is this indeed the likely explanation?

Example: Birth weight paradox

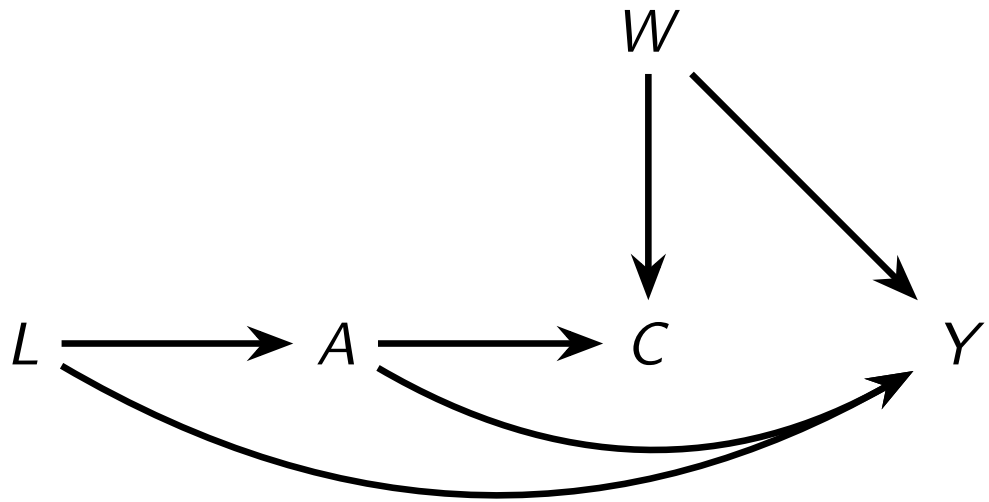


- A Smoking status of the mother
- B Birth weight
- U Unknown factor (e.g. genetic) causing low birth weight
- Y Infant mortality

PS: for this graph to be more plausible, we should also add common causes of A and Y.

Example: Randomised study with loss to follow-up

- Consider our running, conceptual example on a conditional randomised study of heart transplant.
- Suppose that some *young* people who were transplanted were lost to follow-up because they moved to another population. The fact that they were young also made them, of course, less likely to die.
- Can we use the observed data to make causal inference?



Factorization on the counterfactual law

Theorem (Factorization of a counterfactual law)

When positivity holds under a regime g defined in a counterfactual causal model (in particular the NPSEM-IE), the counterfactual law factorizes as

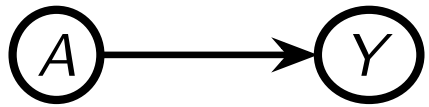
$$p^g(v) = \prod_{j:j \notin \{j_1,\ldots,j_t\}} p(v_j \mid pa_j) \times \prod_{j:j \in \{j_1,\ldots,j_t\}} I(g_j(\overline{v}_{j-1}) = v_j)$$

We will not show this result, but it follows from a similar argument as given for the graph with (L, A, Y).
This factorization is sometimes called the truncation formula.
Thus, when positivity holds for a regime g that fixes V_j to v_j for j in $j \in \{j_1, \ldots, j_t\}$ under a counterfactual causal model, we have that

$$\frac{p^g(v)}{p(v)} = \frac{\prod_{j:j \in \{j_1,\ldots,j_t\}} I(g_j(\overline{v}_{j-1}) = v_j)}{\prod_{j:j \in \{j_1,\ldots,j_t\}} p(v_j \mid pa_j)}.$$

Elephant in the room...

In a randomised study, the following graph is a causal DAG:



And we know that $Y^a \perp\!\!\!\perp A$ for $a \in \{0, 1\}$.
But the counterfactual independence cannot be read off from the graph!
This raises some questions:

- Can we construct graphs to read off such counterfactual independencies?
- Can we read off factorizations of *counterfactual* laws from graphs?

D-separation allows us to read off whether an association is causal

- We can graphically check – using d-separation – whether an observed association between two variables A and B conditional on C is (solely) due to a causal effect (i.e. that the association is unconfounded).
- However, we also want to use graph to evaluate if we can identify functionals of *counterfactual* variables, for example $\mathbb{E}(Y^a)$. Now, the elephant in the room is that there are no counterfactual variables on the DAG! And we did want to reason about counterfactual independencies. Thus, whereas we can evaluate independencies between *factual* variables in a DAG, we cannot study *counterfactual* independencies.
- Here we will study a recent and elegant[26] transformation of the DAG – the so-called Single World Intervention Graph (SWIG) – that does allow us to read off independencies between factual and counterfactual variables.

[26] Thomas S Richardson and James M Robins. "Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality". In: *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper* 128.30 (2013).

Section 15

Single World Intervention Graphs (SWIGs)

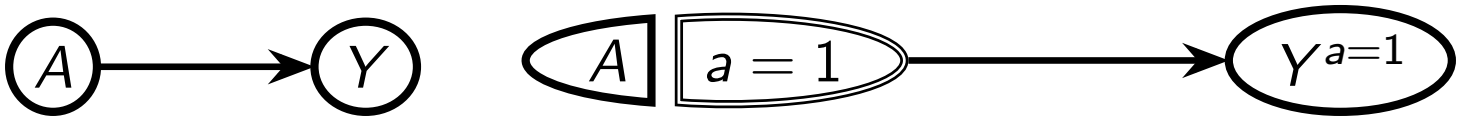
Creation of SWIGs

The SWIG $\mathcal{G}(a)$ is created as follows (it can be conceived as a function that transforms the original DAG into a new graph, which is still a DAG):

1. Draw the DAG that represents the causal model.
2. Split treatment variables into two nodes (indicated by semi-circles), left and right.
   - The left node encodes the random variable treatment that would have been observed in the absence of an intervention. This is called the *natural value of treatment* node. Natural value of treatment nodes should be treated as nodes of an ordinary DAG, i.e., ordinary random variables.
   - The right node encodes the value of treatment under the intervention. These nodes should be treated as constants, i.e. fixed nodes.
3. Re-label every non-manipulated descendant of an intervention node with superscript: the superscripts indicate the counterfactual.
   - Use consistency to obtain graphs with minimal labelling, i.e. the minimal set of counterfactuals in the superscript.

Example: SWIG in a simple randomised trial

SWIG under treatment $a = 1$:



We can read off the independence $Y^{a=1} \perp\!\!\!\perp A$.
We also associate the new **factorization**:

$$P(A = a', Y^{a=1} = y) = P(A = a')P(Y^{a=1} = y),$$

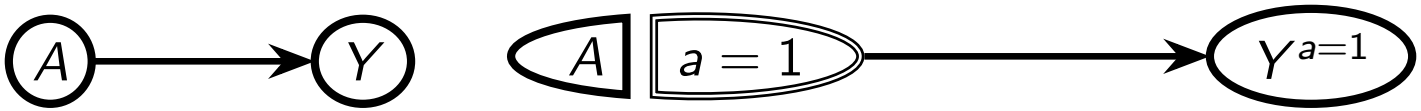
where we omit the fixed nodes from the conditioning set. Furthermore, we make a **modularity** assumption (which would be implied by the NPSEM-IE)

$$P(Y^{a=1} = y) = P(Y = y \mid A = 1),$$

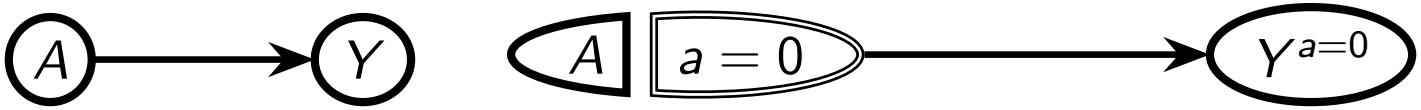
which links the original factorization to the original DAG factorization.

Single world

We can read the independence $Y^{a=1} \perp\!\!\!\perp A$ from the SWIG for treatment $a = 1$:



We can read the independence $Y^{a=0} \perp\!\!\!\perp A$ from the SWIG for treatment $a = 0$:



Why do we need both graphs? These are two different graphs that represent the factorization of different margins: $P(A = a', Y^{a=1} = y)$ and $P(A = a', Y^{a=0} = y)$. None of the SWIGs encodes assumptions between counterfactuals from different worlds $Y^{a=0}$ and $Y^{a=1}$. This is a feature, not a bug.

It has to do with identification. Node splitting preserves identification. If I observe every node that I included in the original DAG, then the counterfactual laws defined by the node splittings are also going to be identified: If $P(A = a', Y = y)$ is identified, then $P(A = a', Y^{a=1} = y)$ is identified and so is $P(A = a', Y^{a=0} = y)$, but not $P(A = a', Y^{a=1} = y', Y^{a=0} = y)$.

SWIT in a simple randomised trial

A SWIT is a SWIG template[27], i.e. a graph valued function:

- It takes a specific value a as input.
- Returns a SWIG $G(a)$.
- SWIG $G(0)$ represents $p(A = a', Y^{a=0} = y)$.
- SWIG $G(1)$ represents $p(A = a', Y^{a=1} = y)$.

The SWIT represents both the SWIGs from the previous slide. Hereafter we will use SWITs for simplicity, most of the time.

[27] Note that I am sometimes sloppy and use the word SWIG when I formally talk about a SWIT.

Factorization

Definition (SWIG factorization)

The factorization associated with a SWIG is

$$P(V^{\overline{a}} = v) = \prod_{V_i \in V} P(V_i^{\overline{a}_i} = v_i \mid (PA_{\mathcal{G}(\overline{a}),i} \setminus \overline{a}) = q)$$

where $q \subseteq pa_i \subset v$ and $\overline{a}_i \subseteq \overline{a}$ ($\overline{a}_i$ are the elements of $\overline{a}$ that are ancestors of V_i).

Modularity

Definition (Modularity)

The DAG pair $(\mathcal{G}, p(v))$ and the SWIG pair $(\mathcal{G}(\overline{a}), p^{\overline{a}}(v))$ under an intervention that sets $\overline{A} = (A_0, \ldots, A_k)$ to $\overline{a} = (a_0, \ldots, a_k)$ satisfy modularity for every $V_i \in V$ if

$$
\begin{aligned}
&P(V_i^{\overline{a}_i} = v_i \mid (PA_{\mathcal{G}(\overline{a}),i} \setminus \overline{a}) = q) \\
=&P(V_i = v_i \mid (PA_{\mathcal{G},i} \setminus \overline{A}) = q, (PA_{\mathcal{G},i} \cap \overline{A}) = \overline{a}_{PA_{\mathcal{G},i} \cap \overline{A}})
\end{aligned}
$$

This definition looks like a mouthful, but it is conceptually quite easy to understand. It bridges counterfactual densities and observable densities.
It is implied by the independent error assumption of the NPSEM-IE, and it holds under a weaker causal model, the FFRCISTG[28] (I have not shown this).

[28] Richardson and Robins, "Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality".

Causal models, factorization and modularity

Theorem

A NPSEM-IE model (and the FFRCISTG model that includes the NPSEM-IE model as a strict submodel) obeys factorization and modularity.

We will not prove this result, but we will use it extensively.
In our saturated graph when we intervene to set $a = 1$, it implies that $P(Y^{a=1} = y) = P(Y = y \mid A = 1)$.

D separation of a path (minimal modification in SWIGs)

A slight twist of D-separation for SWIGs

Definition (d-separation of a path)

A path r is d-separated by a set of nodes Z iff

1. r contains a chain $V_i \to V_j \to V_k$ or a fork $V_i \leftarrow V_j \to V_k$ such that V_j is in Z, or
2. r contains a collider $V_i \to V_j \leftarrow V_k$ such that V_j *is not* in Z and such that no descendant of V_j is in Z.

If a path is not d-separated by Z and there is no fixed node on the path, then the path is d-connected given Z.