BIOST 578: Special Topics
Causal inference in biomedical studies

Richard Guo

Lecture # 5: Causal DAGs and their variants

April 8, 2024


ZOO

Overview

A zoo of graphical models (non-causal or causal) and a myriad of acronyms:

- ADMG
- PAG
- MAG
- **DAG**
- chain graph
- CPDAG
- UG
- ancestral graph
- factor graph
- path diagram

...

Overview

For this quick intro, we shall focus on DAG and its variant ADMG (aka DAG with latents).

1. DAG as a probability model
2. DAG as a causal model
3. DAG as a tool for practitioners

DAG as a probability model

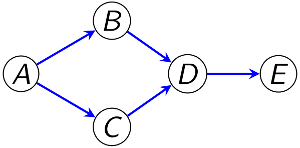
DAG

A graph $\mathcal{G}$ that consists of

- vertices $\boldsymbol{V}$,
- directed edges $\boldsymbol{E}$

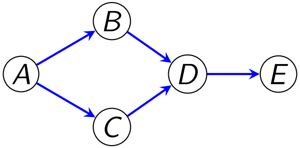
such that there is no directed cycle.

DAG



- $\mathrm{Pa}(D) = \{B, C\}$
- $\mathrm{Ch}(A) = \{B, C\}$
- $A \to B \to D \to E$ is a directed path $\quad A \in \mathrm{An}(E)$ and $E \in \mathrm{De}(A)$
- A and B are adjacent

DAG



- $\mathrm{Pa}(D) = \{B, C\}$
- $\mathrm{Ch}(A) = \{B, C\}$
- $A \to B \to D \to E$ is a directed path $\quad A \in \mathrm{An}(E)$ and $E \in \mathrm{De}(A)$
- A and B are adjacent

- Topological ordering: $A \prec B \prec C \prec D \prec E$ (not unique) such that

$$i \text{ and } j \text{ are adjacent with } i \prec j \implies i \to j.$$

Probability model

☞ Associate every vertex with a random variable. ▶ State space can be $\{0, 1\}$, $\mathbb{R}$ or anything

Probability model

☞ Associate every vertex with a random variable. ▶ State space can be $\{0, 1\}$, $\mathbb{R}$ or anything

Then a DAG $\mathcal{G} = (\boldsymbol{V}, \boldsymbol{E})$ is associated with

$$\begin{aligned} \mathcal{M}_{\mathcal{G}} &:= \{P : p(\boldsymbol{V}) \text{ factorizes according to } \mathcal{G}\} \\ &= \left\{P : p(\boldsymbol{V}) = \prod_{v \in \boldsymbol{V}} p(v \mid \mathrm{Pa}(v))\right\}. \end{aligned}$$

▶ Bayesian network. ▶ semiparametric model

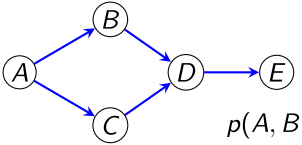
Probability model

☞ Associate every vertex with a random variable. ▶ State space can be $\{0, 1\}$, $\mathbb{R}$ or anything

Then a DAG $\mathcal{G} = (\boldsymbol{V}, \boldsymbol{E})$ is associated with

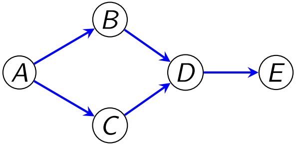
$$\begin{aligned}\mathcal{M}_{\mathcal{G}} &:= \{P : p(\boldsymbol{V}) \text{ factorizes according to } \mathcal{G}\} \\ &= \left\{P : p(\boldsymbol{V}) = \prod_{v \in \boldsymbol{V}} p(v \mid \mathrm{Pa}(v))\right\}.\end{aligned}$$

▶ Bayesian network. ▶ semiparametric model



$$p(A, B, C, D, E) = p(A)\, p(B \mid A)\, p(C \mid A)\, p(D \mid B, C)\, p(E \mid D)$$

Equivalent description: NPSEM-IE



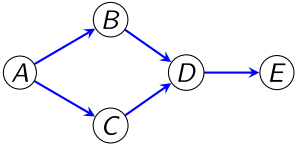
$$p(A, B, C, D, E) = p(A)\, p(B \mid A)\, p(C \mid A)\, p(D \mid B, C)\, p(E \mid D).$$

is equivalent to positing a nonparametric structural equation model with independent errors (NPSEM-IE):

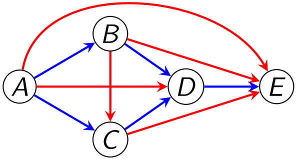
$$\begin{aligned}
\varepsilon_a, \varepsilon_b, \varepsilon_c, \varepsilon_d, \varepsilon_e &\overset{\text{iid}}{\sim} \mathrm{unif}(0, 1) \\
A &= f_a(\varepsilon_a) \\
B &= f_b(A, \varepsilon_b) \\
C &= f_c(A, \varepsilon_c) \\
D &= f_d(B, C, \varepsilon_d) \\
E &= f_e(D, \varepsilon_e)
\end{aligned}$$

Constraints: missing edges

Topological ordering: $A \prec B \prec C \prec D \prec E$



$$p(A)\, p(B \mid A)\, p(C \mid A)\, p(D \mid B, C)\, p(E \mid D)$$



$$p(A)\, p(B \mid A)\, p(C \mid A, {\color{red}B})\, p(D \mid B, C, {\color{red}A})\, p(E \mid D, {\color{red}A, B, C})$$

- The full DAG represents any P
- the **nonparametric** model.

Conditional independence

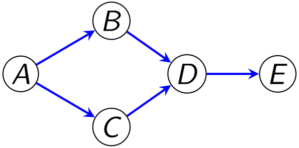
A DAG $\mathcal{G}$, as a probability model $\mathcal{M}_{\mathcal{G}}$, posits

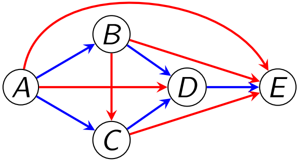
$$\boxed{\text{missing edges} \implies \text{conditional independence.}}$$

Conditional independence

A DAG $\mathcal{G}$, as a probability model $\mathcal{M}_{\mathcal{G}}$, posits

$$\boxed{\text{missing edges} \implies \text{conditional independence.}}$$



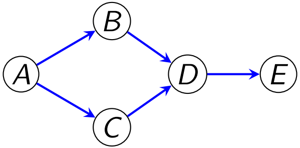


The missing '$B \to C$' posits

$$P(C \mid A, B) = P(C \mid A) \iff \boxed{B \perp\!\!\!\perp C \mid A} \iff P(B, C \mid A) = P(B \mid A)P(C \mid A).$$

Conditional independence

The graph



also implies, e.g.,

$$A, B, C \perp\!\!\!\perp E \mid D, \quad A, C \perp\!\!\!\perp E \mid B, D, \quad \ldots$$

☞ How we read off all the CIs a DAG implies ?

Dependence: mechanisms

Let A, B be the two fair coins.

Dependence: mechanisms

Let A, B be the two fair coins.

$$\text{HH, TT, HT, TH with equal prob.} \iff A \perp\!\!\!\perp B$$

Dependence: mechanisms

Let A, B be the two fair coins.

$$\text{HH, TT, HT, TH with equal prob.} \iff A \perp\!\!\!\perp B$$

A B $\qquad A \perp\!\!\!\perp B$

Mechanisms of inducing dependence

Let A, B be the two fair coins.

$$\text{only HH and TT} \implies A \not\perp B$$

Mechanisms of inducing dependence

Let A, B be the two fair coins.

$$\text{only HH and TT} \implies A \not\perp B$$

(1) Causal relations

$$A \rightarrow B \qquad A \leftarrow B \qquad A \not\perp B$$

Mechanisms of inducing dependence

Let A, B be the two fair coins.

$$\text{only HH and TT} \implies A \not\perp\!\!\!\perp B$$

(1) Causal relations

$$A \rightarrow B \qquad A \leftarrow B \qquad A \not\perp\!\!\!\perp B$$

(2) Common cause (unconditionally)

$$\begin{tikzcd} & C \arrow[dl] \arrow[dr] & \\ A & & B \end{tikzcd} \qquad A \not\perp\!\!\!\perp B$$

Mechanisms of inducing dependence

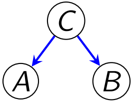
Let A, B be the two fair coins.

$$\text{only HH and TT} \implies A \not\perp\!\!\!\perp B$$

(1) Causal relations

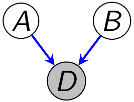
$A \not\perp\!\!\!\perp B$

(2) Common cause (unconditionally)



$A \not\perp\!\!\!\perp B$

(3) Conditioning on a common effect



$A \not\perp\!\!\!\perp B \mid D$

d-connecting path

- A **path** between A and B: a sequence of distinct, adjacent vertices

$$A \to \circ \to \circ \leftarrow \cdots \to B,$$

where every non-endpoint vertex is either a **collider** ($\to \circ \leftarrow$) or a **non-collider** ($\to \circ \to$, $\leftarrow \circ \leftarrow$, $\leftarrow \circ \to$)

A path is **d-connecting given $\boldsymbol{C}$** if

1. every non-collider $\notin \boldsymbol{C}$, and
2. every collider is $\in \boldsymbol{C}$ or is an ancestor of $\boldsymbol{C}$.

d-separation

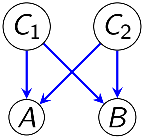
Vertex A and vertex B are d-separated by vertex set $\boldsymbol{C}$, written as $A \perp\!\!\!\perp_{\mathcal{G}} B \mid \boldsymbol{C}$, if there is no d-connecting path between A and B given $\boldsymbol{C}$.

- Extended to $\boldsymbol{A} \perp\!\!\!\perp_{\mathcal{G}} \boldsymbol{B} \mid C$ for disjoint vertex sets $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$.

d-separation

Vertex A and vertex B are d-separated by vertex set $\boldsymbol{C}$, written as $A \perp\!\!\!\perp_{\mathcal{G}} B \mid \boldsymbol{C}$, if there is no d-connecting path between A and B given $\boldsymbol{C}$.

- Extended to $\boldsymbol{A} \perp\!\!\!\perp_{\mathcal{G}} \boldsymbol{B} \mid C$ for disjoint vertex sets $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$.



Global Markov property

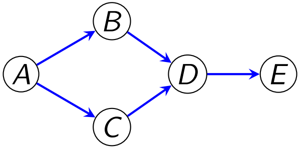
Global Markov property For disjoint vertex sets $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$, it holds that

$$\boldsymbol{A} \perp\!\!\!\perp_{\mathcal{G}} \boldsymbol{B} \mid \boldsymbol{C} \implies \boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B} \mid \boldsymbol{C} \; [P], \quad P \in \mathcal{M}_{\mathcal{G}}.$$

Global Markov property

Global Markov property For disjoint vertex sets $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$, it holds that

$$\boldsymbol{A} \perp\!\!\!\perp_{\mathcal{G}} \boldsymbol{B} \mid \boldsymbol{C} \implies \boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B} \mid \boldsymbol{C} \; [P], \quad P \in \mathcal{M}_{\mathcal{G}}.$$



DAG as a CI model

- The global Markov property also holds **reversely**. If P satisfies

$$\boldsymbol{A} \perp\!\!\!\perp_{\mathcal{G}} \boldsymbol{B} \mid \boldsymbol{C} \implies \boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B} \mid \boldsymbol{C} \; [P],$$

then $P \in \mathcal{M}_{\mathcal{G}}$.

DAG as a CI model

- The global Markov property also holds **reversely**. If P satisfies

$$\boldsymbol{A} \perp\!\!\!\perp_{\mathcal{G}} \boldsymbol{B} \mid \boldsymbol{C} \implies \boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B} \mid \boldsymbol{C} \; [P],$$

then $P \in \mathcal{M}_{\mathcal{G}}$.

Theorem Factorization $\iff$ Global Markov $\iff$ Local Markov.

- Local Markov: $P \in \mathcal{M}_{\mathcal{G}} \implies A \perp\!\!\!\perp \text{non-descendants of } A \mid \text{Pa}(A)$

DAG as a CI model

- The global Markov property also holds **reversely**. If P satisfies

$$\boldsymbol{A} \perp\!\!\!\perp_{\mathcal{G}} \boldsymbol{B} \mid \boldsymbol{C} \implies \boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B} \mid \boldsymbol{C}\ [P],$$

then $P \in \mathcal{M}_{\mathcal{G}}$.

Theorem Factorization $\iff$ Global Markov $\iff$ Local Markov.

- Local Markov: $P \in \mathcal{M}_{\mathcal{G}} \implies A \perp\!\!\!\perp \text{non-descendants of } A \mid \text{Pa}(A)$

☞ That is, the model defined as $\mathcal{M}_{\mathcal{G}} := \{P : P \text{ factorizes according to } \mathcal{G}\}$ can be viewed as a **CI model**

$$\{P : \boldsymbol{A} \perp\!\!\!\perp_{\mathcal{G}} \boldsymbol{B} \mid \boldsymbol{C} \implies \boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B} \mid \boldsymbol{C}\ [P]\},$$

i.e.,

$$\{P : P \text{ satisfies CIs that are encoded as d-separations in } \mathcal{G}\}.$$

Graphoid axioms

- From a set of CIs, new CIs may be derived, e.g., with applications of ‘graphoid axioms’:
  1. Symmetry: $A \perp\!\!\!\perp B \mid \boldsymbol{C} \implies B \perp\!\!\!\perp A \mid \boldsymbol{C}$
  2. Decomposition: $A \perp\!\!\!\perp B, D \mid \boldsymbol{C} \implies A \perp\!\!\!\perp B \mid \boldsymbol{C}$ and $A \perp\!\!\!\perp D \mid \boldsymbol{C}$
  3. Weak union: $A \perp\!\!\!\perp B, D \mid \boldsymbol{C} \implies A \perp\!\!\!\perp B \mid D, \boldsymbol{C}$
  4. Contraction: $A \perp\!\!\!\perp B \mid \boldsymbol{C}$ and $A \perp\!\!\!\perp D \mid B, \boldsymbol{C} \implies A \perp\!\!\!\perp B, D \mid \boldsymbol{C}$

Graphoid axioms

- From a set of CIs, new CIs may be derived, e.g., with applications of 'graphoid axioms':
  1. Symmetry: $A \perp\!\!\!\perp B \mid \boldsymbol{C} \implies B \perp\!\!\!\perp A \mid \boldsymbol{C}$
  2. Decomposition: $A \perp\!\!\!\perp B, D \mid \boldsymbol{C} \implies A \perp\!\!\!\perp B \mid \boldsymbol{C}$ and $A \perp\!\!\!\perp D \mid \boldsymbol{C}$
  3. Weak union: $A \perp\!\!\!\perp B, D \mid \boldsymbol{C} \implies A \perp\!\!\!\perp B \mid D, \boldsymbol{C}$
  4. Contraction: $A \perp\!\!\!\perp B \mid \boldsymbol{C}$ and $A \perp\!\!\!\perp D \mid B, \boldsymbol{C} \implies A \perp\!\!\!\perp B, D \mid \boldsymbol{C}$

☞ **Example** Given

$$A \perp\!\!\!\perp B, \quad A \perp\!\!\!\perp C \mid B,$$

we can derive

$$A \perp\!\!\!\perp B, \quad A \perp\!\!\!\perp C \mid B \implies A \perp\!\!\!\perp B, C \implies A \perp\!\!\!\perp C.$$

Completeness of d-separation

- **Question**: From the list CIs encoded by d-separations, can we derive a new CI (e.g., with graphoid axioms) that holds for **every** $P \in \mathcal{M}_{\mathcal{G}}$ but does not correspond to any d-separation in the graph? **NO!**

Completeness of d-separation

- **Question**: From the list CIs encoded by d-separations, can we derive a new CI (e.g., with graphoid axioms) that holds for **every** $P \in \mathcal{M}_{\mathcal{G}}$ but does not correspond to any d-separation in the graph? **NO!**

Theorem For every $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ such that $\boldsymbol{A}$ and $\boldsymbol{B}$ are d-connected given $\boldsymbol{C}$ on $\mathcal{G}$, there exists $P \in \mathcal{M}_{\mathcal{G}}$ such that

$$\boldsymbol{A} \not\perp\!\!\!\perp \boldsymbol{B} \mid \boldsymbol{C} \; [P].$$

Completeness of d-separation

- Why is this important?

Completeness of d-separation

- Why is this important?

Milan Studenỳ (1992) showed that CIs cannot be axiomatized by a finite set of rules. That is, one cannot deduce all the consequences of an arbitrary set $\{\mathrm{CI}_1, \mathrm{CI}_2, \ldots, \mathrm{CI}_k\}$ using a finite number of rules (e.g. graphoid axioms).
☞ **Graphoid axioms are incomplete and cannot be completed**, if one is free to specify the list of CIs.

Completeness of d-separation

- Why is this important?

Milan Studený (1992) showed that CIs cannot be axiomatized by a finite set of rules. That is, one cannot deduce all the consequences of an arbitrary set $\{\mathrm{CI}_1, \mathrm{CI}_2, \ldots, \mathrm{CI}_k\}$ using a finite number of rules (e.g. graphoid axioms).
☞ **Graphoid axioms are incomplete and cannot be completed**, if one is free to specify the list of CIs.

☞ However, DAG models are a class of nice CI models by confining the set of CIs (reducing complexity).

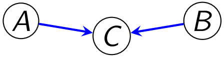
Examples over three variables

- $\boldsymbol{V} = \{A, B, C\}$.

Examples over three variables

- $\boldsymbol{V} = \{A, B, C\}$.

$$A \perp\!\!\!\perp B$$



Examples over three variables

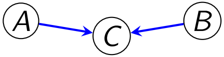
- $\boldsymbol{V} = \{A, B, C\}$.

$A \perp\!\!\!\perp B$	$A \rightarrow C \leftarrow B$
$A \perp\!\!\!\perp B \mid C$	$A \rightarrow C \rightarrow B$ $A \leftarrow C \leftarrow B$ $A \leftarrow C \rightarrow B$

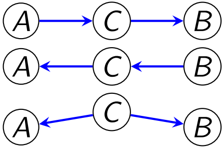
Examples over three variables

- $\boldsymbol{V} = \{A, B, C\}$.

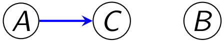
$A \perp\!\!\!\perp B$



$A \perp\!\!\!\perp B \mid C$



$A \perp\!\!\!\perp B$
$A \perp\!\!\!\perp B \mid C$
$B \perp\!\!\!\perp C$
$B \perp\!\!\!\perp C \mid A$



Markov equivalence

- $\mathcal{G}$ and $\mathcal{G}'$ are called '**Markov equivalent**', written as $\mathcal{G} \sim \mathcal{G}'$, if they define the same model.
  - i.e., they encode the same CIs.

Markov equivalence

- $\mathcal{G}$ and $\mathcal{G}'$ are called '**Markov equivalent**', written as $\mathcal{G} \sim \mathcal{G}'$, if they define the same model.
  - i.e., they encode the same CIs.

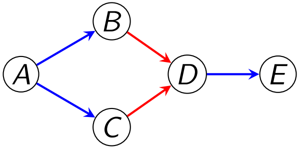
Theorem Two DAGs over the same set of vertices are **Markov equivalent** iff they share the same **adjacencies** and **unshielded colliders**.

Markov equivalence

- $\mathcal{G}$ and $\mathcal{G}'$ are called '**Markov equivalent**', written as $\mathcal{G} \sim \mathcal{G}'$, if they define the same model.
  - i.e., they encode the same CIs.

Theorem Two DAGs over the same set of vertices are **Markov equivalent** iff they share the same **adjacencies** and **unshielded colliders**.

- Unshielded collider: $B \to D \leftarrow C$ but B, C are not adjacent



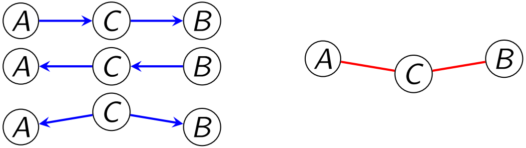
Markov equivalence class

- A Markov equivalence class can be represented by an essential graph / CPDAG.

(Without extra assumptions, DAGs can only be learned from data up to Markov equivalence.)

$A \perp\!\!\!\perp B$

$A \perp\!\!\!\perp B \mid C$



$A \perp\!\!\!\perp B$

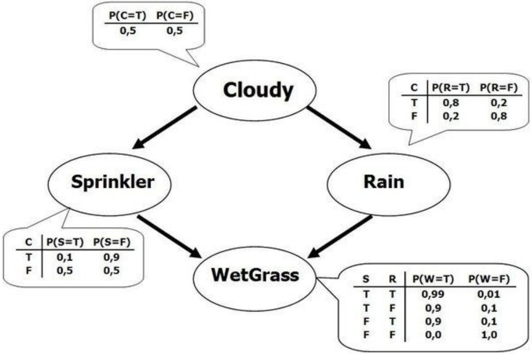
$A \perp\!\!\!\perp B \mid C$

$B \perp\!\!\!\perp C$

$B \perp\!\!\!\perp C \mid A$

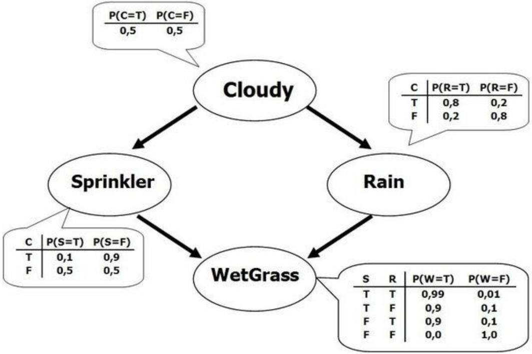
Parametric case: finite state space

- When every variable only takes finitely many levels, the model can be parametrized in terms of conditional probability tables $\{p(A \mid \mathrm{Pa}(A)) : A \in \boldsymbol{V}\}$.



Parametric case: finite state space

- When every variable only takes finitely many levels, the model can be parametrized in terms of conditional probability tables $\{p(A \mid \mathrm{Pa}(A)) : A \in \boldsymbol{V}\}$.



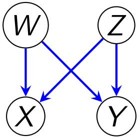
- Efficient algorithms exist for **marginalization** and computing **posterior probabilities**

Parametric case: linear SEM

Each linear equation posits that

$$V_i = \beta_i^{\mathsf{T}} \operatorname{Pa}(V_i) + \varepsilon_i,$$

where ε_i is exogenous error (drawn independently).



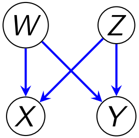
$$\begin{aligned} W &= && && \varepsilon_w \\ Z &= && && \varepsilon_z \\ X &= \beta_{wx} W && + \beta_{zx} Z + && \varepsilon_x \\ Y &= \beta_{wy} W && + \beta_{zy} Z + && \varepsilon_y \end{aligned}$$

Parametric case: linear SEM

Each linear equation posits that

$$V_i = \beta_i^\intercal \, \mathrm{Pa}(V_i) + \varepsilon_i,$$

where ε_i is exogenous error (drawn independently).



$$\begin{aligned} W &= & & & \varepsilon_w \\ Z &= & & & \varepsilon_z \\ X &= \beta_{wx} W & +\beta_{zx} Z + & & \varepsilon_x \\ Y &= \beta_{wy} W & +\beta_{zy} Z + & & \varepsilon_y \end{aligned}$$

- Because of acyclicity, it admits a unique solution:

$$\boldsymbol{V} = \boldsymbol{B}^\intercal \boldsymbol{V} + \boldsymbol{\varepsilon} \iff \boldsymbol{V} = (\boldsymbol{I} - \boldsymbol{B})^{-\intercal} \boldsymbol{\varepsilon}.$$

Limitation of DAGs

DAGs are not closed (in general) under marginalization and selection.

Limitation of DAGs

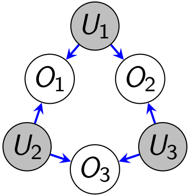
DAGs are not closed (in general) under marginalization and selection.

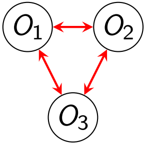
- Marginalization: Consider $\boldsymbol{V} = \boldsymbol{O} \cup \boldsymbol{U}$ for $\boldsymbol{O} \cap \boldsymbol{U} = \emptyset$, where we only get to observe $P(\boldsymbol{O})$.
  - $\boldsymbol{U}$ are **latent variables**

Limitation of DAGs

DAGs are not closed (in general) under marginalization and selection.

- Marginalization: Consider $\boldsymbol{V} = \boldsymbol{O} \cup \boldsymbol{U}$ for $\boldsymbol{O} \cap \boldsymbol{U} = \emptyset$, where we only get to observe $P(\boldsymbol{O})$.
- $\boldsymbol{U}$ are **latent variables**



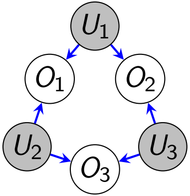


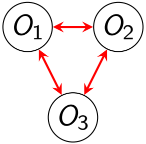
$$
\begin{aligned}
O_1 &\perp\!\!\!\perp O_2, \quad & O_1 &\not\!\perp\!\!\!\perp O_2 \mid O_3 \\
O_1 &\perp\!\!\!\perp O_3, \quad & O_1 &\not\!\perp\!\!\!\perp O_3 \mid O_2 \\
O_2 &\perp\!\!\!\perp O_3, \quad & O_2 &\not\!\perp\!\!\!\perp O_3 \mid O_1
\end{aligned}
$$

Limitation of DAGs

DAGs are not closed (in general) under marginalization and selection.

- Marginalization: Consider $\boldsymbol{V} = \boldsymbol{O} \cup \boldsymbol{U}$ for $\boldsymbol{O} \cap \boldsymbol{U} = \emptyset$, where we only get to observe $P(\boldsymbol{O})$.
  - $\boldsymbol{U}$ are **latent variables**





$$O_1 \perp\!\!\!\perp O_2, \quad O_1 \not\perp\!\!\!\perp O_2 \mid O_3$$
$$O_1 \perp\!\!\!\perp O_3, \quad O_1 \not\perp\!\!\!\perp O_3 \mid O_2$$
$$O_2 \perp\!\!\!\perp O_3, \quad O_2 \not\perp\!\!\!\perp O_3 \mid O_1$$

☞ Such a CI model does not correspond to any DAG over $\boldsymbol{O}$.

DAGs with latent variables

For a DAG over $\boldsymbol{V} = \boldsymbol{O} \cup \boldsymbol{U}$,

$$\text{constraints in } P(\boldsymbol{O}) = \underbrace{\overbrace{\text{CIs}}^{\text{ancestral graphs}} + \text{`Verma' constraints}}_{\text{equalities (nested Markov models)}} + \text{inequalities}$$

DAGs with latent variables

For a DAG over $\boldsymbol{V} = \boldsymbol{O} \cup \boldsymbol{U}$,

$$\text{constraints in } P(\boldsymbol{O}) = \underbrace{\overbrace{\text{CIs}}^{\text{ancestral graphs}} + \text{`Verma' constraints}}_{\text{equalities (nested Markov models)}} + \text{inequalities}$$

☞ See also Richardson and Spirtes (2002), Richardson (2003), Robin J Evans (2016), and Richardson, Robin J. Evans, et al. (2023).

DAG as a causal model

What makes it causal?

We have already seen that a DAG is a probability model as it defines a set of probability distributions $\mathcal{M}_{\mathcal{G}}$.

- $P \in \mathcal{M}_{\mathcal{G}}$ is an observed distribution over factual random variables.

What makes it causal?

We have already seen that a DAG is a probability model as it defines a set of probability distributions $\mathcal{M}_{\mathcal{G}}$.

- $P \in \mathcal{M}_{\mathcal{G}}$ is an observed distribution over factual random variables.

☞ What makes it a causal model?

What makes it causal?

We have already seen that a DAG is a probability model as it defines a set of probability distributions $\mathcal{M}_{\mathcal{G}}$.

- $P \in \mathcal{M}_{\mathcal{G}}$ is an observed distribution over factual random variables.

☞ What makes it a causal model?

- It must be augmented with extra semantics that

What makes it causal?

We have already seen that a DAG is a probability model as it defines a set of probability distributions $\mathcal{M}_{\mathcal{G}}$.

- $P \in \mathcal{M}_{\mathcal{G}}$ is an observed distribution over factual random variables.

☞ What makes it a causal model?

- It must be augmented with extra semantics that
  1. posits the existence of counterfactuals (i.e., potential outcomes),

What makes it causal?

We have already seen that a DAG is a probability model as it defines a set of probability distributions $\mathcal{M}_{\mathcal{G}}$.

- $P \in \mathcal{M}_{\mathcal{G}}$ is an observed distribution over factual random variables.

☞ What makes it a causal model?

- It must be augmented with extra semantics that
  1. posits the existence of counterfactuals (i.e., potential outcomes),
  2. makes assumptions about factual (e.g., Y) and counterfactual (e.g., $Y(a)$) variables, and

What makes it causal?

We have already seen that a DAG is a probability model as it defines a set of probability distributions $\mathcal{M}_{\mathcal{G}}$.

- $P \in \mathcal{M}_{\mathcal{G}}$ is an observed distribution over factual random variables.

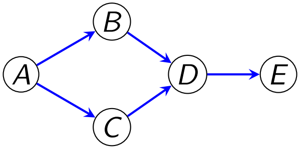
☞ What makes it a causal model?

- It must be augmented with extra semantics that
  1. posits the existence of counterfactuals (i.e., potential outcomes),
  2. makes assumptions about factual (e.g., Y) and counterfactual (e.g., $Y(a)$) variables, and
  3. connects the counterfactual distributions with the observed distribution.

Sampling from a DAG

$$\begin{tikzcd}[row sep=small] & B \arrow[dr] & & \\ A \arrow[ur] \arrow[dr] & & D \arrow[r] & E \\ & C \arrow[ur] & & \end{tikzcd}$$

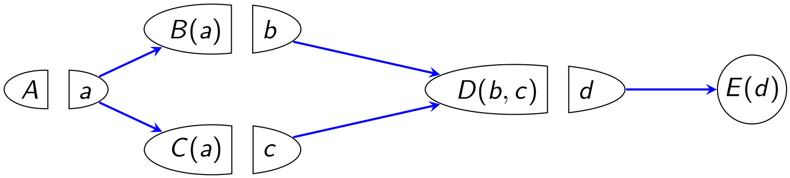
Sampling from a DAG



Following the topological ordering $A \prec B \prec C \prec D \prec E$,

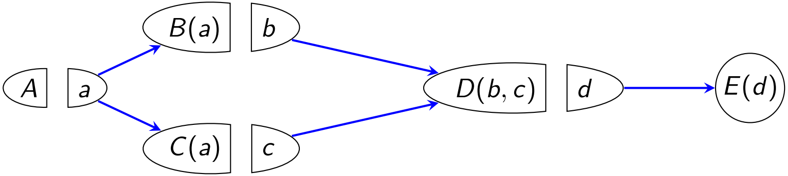
1. Draw $A \sim P(A)$
2. Draw $B \sim P(B \mid A)$, $C \sim P(C \mid A)$
3. Draw $D \sim P(D \mid B, C)$
4. Draw $E \sim P(E \mid D)$

Alternative sampling (I): one-step-ahead counterfactuals



- Single-World Intervention Graph (SWIG) (Richardson and J. M. Robins, 2013)

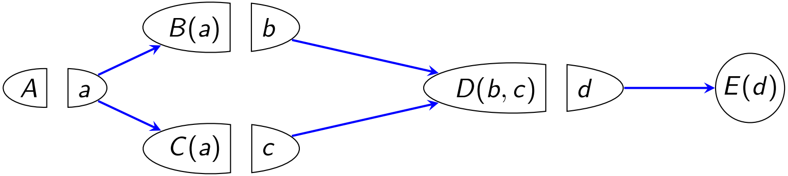
Alternative sampling (I): one-step-ahead counterfactuals



- Single-World Intervention Graph (SWIG) (Richardson and J. M. Robins, 2013)

1. Draw $A \sim P(A)$

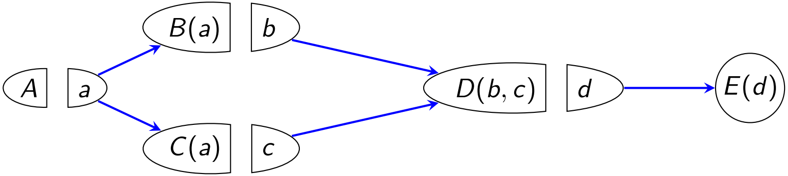
Alternative sampling (I): one-step-ahead counterfactuals



- Single-World Intervention Graph (SWIG) (Richardson and J. M. Robins, 2013)

1. Draw $A \sim P(A)$
2. For every potential a, draw $B(a) \sim P(B \mid A = a)$, $C(a) \sim P(A \mid A = a)$ independent of A

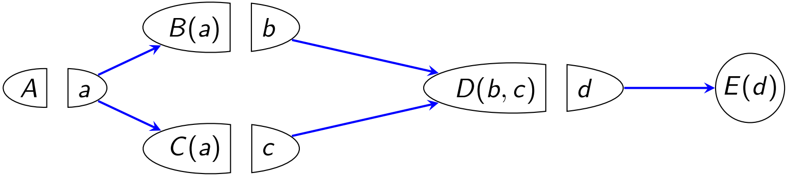
Alternative sampling (I): one-step-ahead counterfactuals



- Single-World Intervention Graph (SWIG) (Richardson and J. M. Robins, 2013)

1. Draw $A \sim P(A)$
2. For every potential a, draw $B(a) \sim P(B \mid A = a)$, $C(a) \sim P(A \mid A = a)$ independent of A
3. For every potential (b, c), draw $D(b, c) \sim P(D \mid B = b, C = c)$ independent of previously drawn.

Alternative sampling (I): one-step-ahead counterfactuals



- Single-World Intervention Graph (SWIG) (Richardson and J. M. Robins, 2013)

1. Draw $A \sim P(A)$
2. For every potential a, draw $B(a) \sim P(B \mid A = a)$, $C(a) \sim P(A \mid A = a)$ independent of A
3. For every potential (b, c), draw $D(b, c) \sim P(D \mid B = b, C = c)$ independent of previously drawn.
4. For every potential d, draw $E(d) \sim P(E \mid D = d)$ independent of previously drawn.

Alternative sampling (I): one-step-ahead counterfactuals

Alternative sampling (I): one-step-ahead counterfactuals

- This is called 'single-world' because we only posit that

$$A \perp\!\!\!\perp B(a), C(a) \quad \text{for every } a$$

and

$$B(a) \sim P(B \mid A = a), \quad C(a) \sim P(C \mid A = a) \quad \text{for every } a.$$

Alternative sampling (I): one-step-ahead counterfactuals

- This is called 'single-world' because we only posit that

$$A \perp\!\!\!\perp B(a), C(a) \quad \text{for every } a$$

and

$$B(a) \sim P(B \mid A = a), \quad C(a) \sim P(C \mid A = a) \quad \text{for every } a.$$

- Refrain from making 'cross-world' statements such as

$$A \perp\!\!\!\perp B(a), B(a'), B(a''), C(a), C(a'), C(a'')$$

because we will never see $B(a)$ and $B(a')$ together for $a \neq a'$.

☞ Cross-world assumptions cannot be empirically verified.

Alternative sampling (I): one-step-ahead counterfactuals

- This is called 'single-world' because we only posit that

$$A \perp\!\!\!\perp B(a), C(a) \quad \text{for every } a$$

and

$$B(a) \sim P(B \mid A = a), \quad C(a) \sim P(C \mid A = a) \quad \text{for every } a.$$

- Refrain from making 'cross-world' statements such as

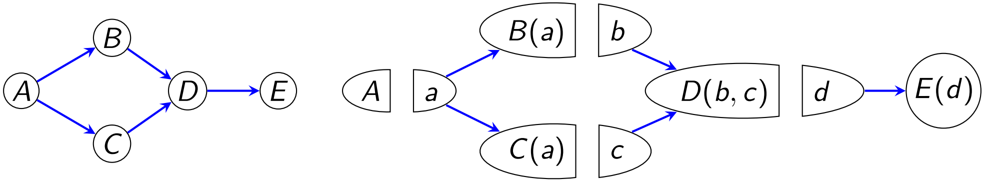
$$A \perp\!\!\!\perp B(a), B(a'), B(a''), C(a), C(a'), C(a'')$$

because we will never see $B(a)$ and $B(a')$ together for $a \neq a'$.

☞ Cross-world assumptions cannot be empirically verified.

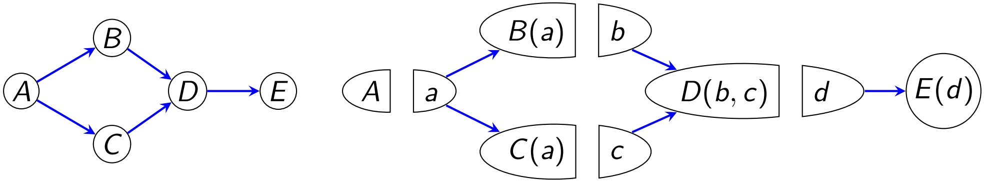
- Nevertheless, we can empirically examine $A \perp\!\!\!\perp B(a), C(a)$, if we can observe the naturally occurring value of A immediately before we intervene on it.

Alternative sampling (II): recursive substitution



To generate the observed, factual variables,

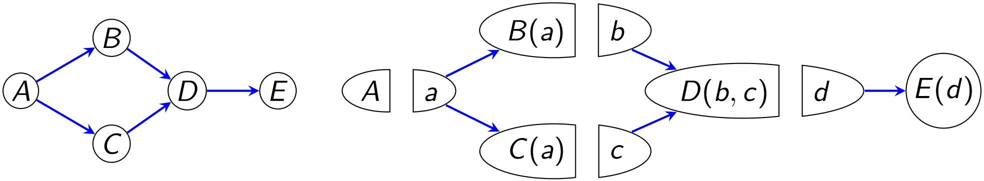
Alternative sampling (II): recursive substitution



To generate the observed, factual variables,

1. $A = A$,

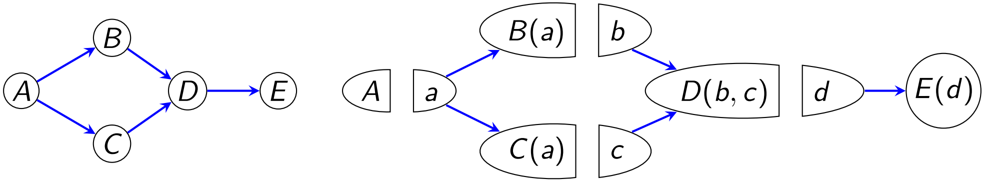
Alternative sampling (II): recursive substitution



To generate the observed, factual variables,

1. $A = A$,
2. $B = B(A)$, $C = C(A)$,

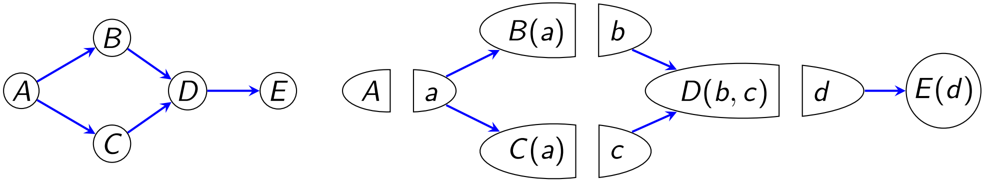
Alternative sampling (II): recursive substitution



To generate the observed, factual variables,

1. $A = A$,
2. $B = B(A)$, $C = C(A)$,
3. $D = D(B, C)$,

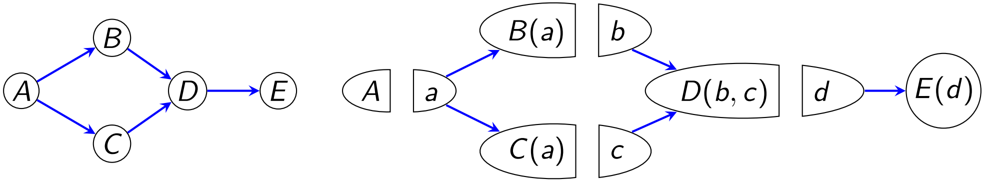
Alternative sampling (II): recursive substitution



To generate the observed, factual variables,

1. $A = A$,
2. $B = B(A)$, $C = C(A)$,
3. $D = D(B, C)$,
4. $E = E(D)$.

Alternative sampling (II): recursive substitution



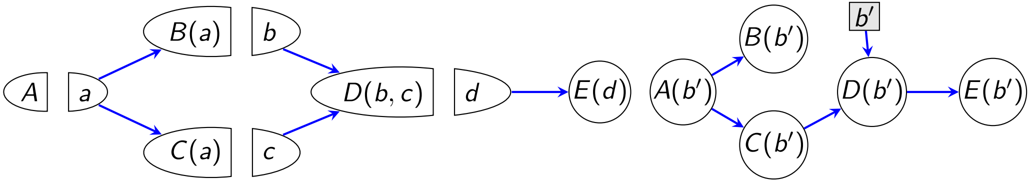
To generate the observed, factual variables,

1. $A = A$,
2. $B = B(A)$, $C = C(A)$,
3. $D = D(B, C)$,
4. $E = E(D)$.

- Apparently, $(A, B, C, D, E) \sim P$

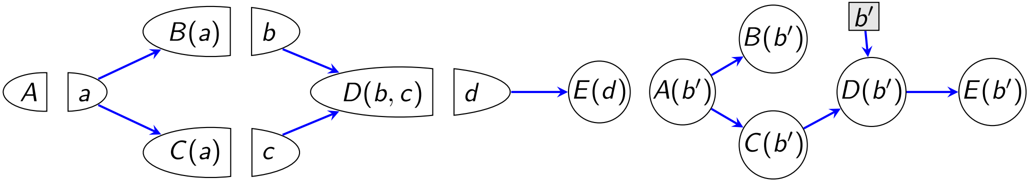
Alternative sampling (III): intervention

☞ Suppose that we **intervene on B and set it to b'** — imposes input to B's children.



Alternative sampling (III): intervention

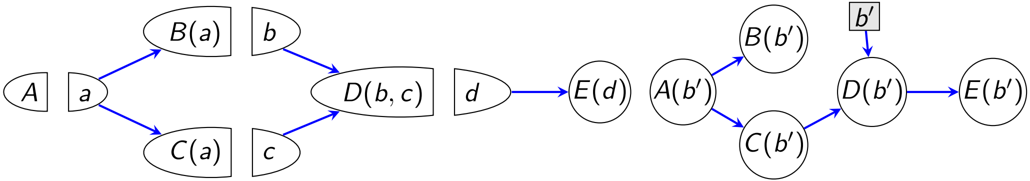
☞ Suppose that we **intervene on B and set it to b'** — imposes input to B's children.



1. $A(b') = A$,

Alternative sampling (III): intervention

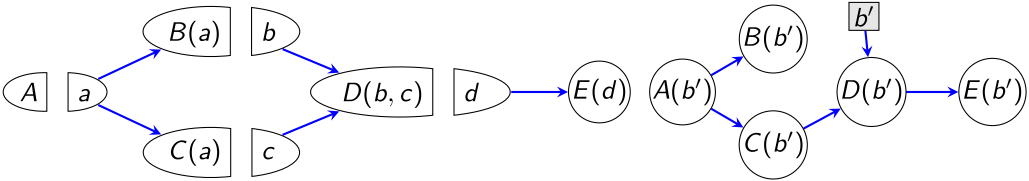
☞ Suppose that we **intervene on B and set it to b'** — imposes input to B's children.



1. $A(b') = A$,
2. $B(b') = B(A(b')) = B(A)$, $C(b') = C(A(b')) = C(A)$
   - $B(b')$ is the naturally occurring value of B immediately before it is intervened on

Alternative sampling (III): intervention

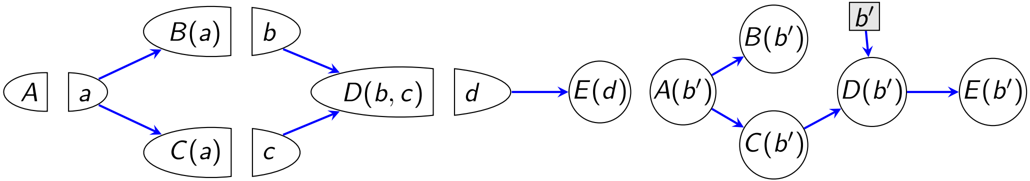
☞ Suppose that we **intervene on B and set it to b'** — imposes input to B's children.



1. $A(b') = A$,
2. $B(b') = B(A(b')) = B(A)$, $C(b') = C(A(b')) = C(A)$
   - $B(b')$ is the naturally occurring value of B immediately before it is intervened on
3. $D(b') = D(b', C(b'))$,

Alternative sampling (III): intervention

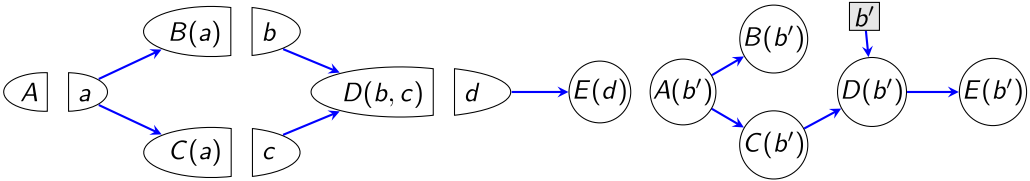
☞ Suppose that we **intervene on B and set it to b'** — imposes input to B's children.



1. $A(b') = A$,
2. $B(b') = B(A(b')) = B(A)$, $C(b') = C(A(b')) = C(A)$
   - $B(b')$ is the naturally occurring value of B immediately before it is intervened on
3. $D(b') = D(b', C(b'))$,
4. $E(b') = E(D(b'))$.

Alternative sampling (III): intervention

☞ Suppose that we **intervene on B and set it to b'** — imposes input to B's children.



1. $A(b') = A$,
2. $B(b') = B(A(b')) = B(A)$, $C(b') = C(A(b')) = C(A)$
   - $B(b')$ is the naturally occurring value of B immediately before it is intervened on
3. $D(b') = D(b', C(b'))$,
4. $E(b') = E(D(b'))$.

- This defines the distribution of $P((A,B,C,D,E)(b'))$, or $P(A,B,C,D,E \mid \text{do}(B=b'))$.

Alternative sampling: the causal model

☞ This set of semantics defines the FFRCISTG / SWIG causal model associated with a DAG $\mathcal{G}$.

- 'Finest Fully Randomized Causally Interpreted Structured Tree Graph' (J. Robins, 1986)

Alternative sampling: the causal model

☞ This set of semantics defines the FFRCISTG / SWIG causal model associated with a DAG $\mathcal{G}$.

- 'Finest Fully Randomized Causally Interpreted Structured Tree Graph' (J. Robins, 1986)

☞ It makes **weaker assumptions** than Pearl's NPSEM-IE (nonparametric structural equation model with independent errors) causal model.

- Even though NPSEM-IE and DAG define the same **probability model**!

g-formula

g-formula

With the semantics just described, the counterfactual distribution is

$$P\left(A(b') = a, B(b') = b, C(b') = c, D(b') = d, E(b') = e\right) =$$
$$P(A = a)\, P(B = b \mid A = a)\, P(C = c \mid A = a)\, P(D = d \mid B = b', C = c)\, P(E = e \mid D = e).$$

- This is **identified** from the observed distribution P because every **one-step-ahead conditional** is identified from P.

g-formula

With the semantics just described, the counterfactual distribution is

$$P\left(A(b') = a, B(b') = b, C(b') = c, D(b') = d, E(b') = e\right) =$$
$$P(A = a)\, P(B = b \mid A = a)\, P(C = c \mid A = a)\, P(D = d \mid B = b', C = c)\, P(E = e \mid D = e).$$

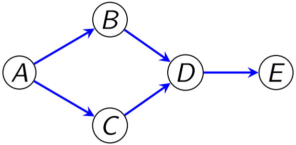
- This is **identified** from the observed distribution P because every **one-step-ahead conditional** is identified from P.

g-formula

$$P(\boldsymbol{V}(a) = \boldsymbol{v}) = \prod_{i=1}^{|V|} P(v_i \mid a_{\mathrm{Pa}(i)\cap A}, v_{\mathrm{Pa}(i)\setminus A})$$

- From this we can identify counterfactual means, etc.

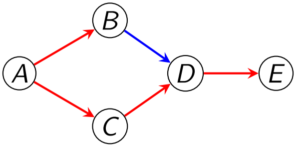
g-formula



- Example

$$
\begin{aligned}
\mathbb{E}\, E(b') &= \sum_{a,b,c,d,e} eP(A=a)\, P(B=b \mid A=a)\, P(C=c \mid A=a) \\
&\qquad \times P(D=d \mid B=b', C=c)\, P(E=e \mid D=d) \\
&= \sum_{d,c} \mathbb{E}[E \mid D=d] P(D=d \mid B=b', C=c) P(C=c) \\
&= \sum_{d,c} \mathbb{E}[E \mid D=d, B=b', C=c] P(D=d \mid B=b', C=c) P(C=c) \quad (why?)
\end{aligned}
$$

g-formula



$$
\begin{aligned}
\mathbb{E}\, E(b') &= \sum_c \sum_d \mathbb{E}[E \mid D = d, B = b', C = c] P(D = d \mid B = b', C = c) P(C = c) \\
&= \sum_c \mathbb{E}[E \mid B = b', C = c] P(C = c) \\
&= \mathbb{E}\left\{\mathbb{E}[E \mid B = b', C = c]\right\}.
\end{aligned}
$$

- Backdoor/adjustment formula that adjusts for C

Backdoor/adjustment

Suppose A is the treatment and Y is the outcome.

Backdoor/adjustment

Suppose A is the treatment and Y is the outcome.

- A set of variables $\boldsymbol{S} \subseteq \boldsymbol{V} \setminus \{A, Y\}$ is a valid adjustment set if

$$P(Y(a) = y) = \mathbb{E}\{P(Y = y \mid A = a, \boldsymbol{S})\}$$

under the SWIG causal model associated with $\mathcal{G}$.

Backdoor/adjustment

Suppose A is the treatment and Y is the outcome.

- A set of variables $\boldsymbol{S} \subseteq \boldsymbol{V} \setminus \{A, Y\}$ is a valid adjustment set if

$$P(Y(a) = y) = \mathbb{E}\{P(Y = y \mid A = a, \boldsymbol{S})\}$$

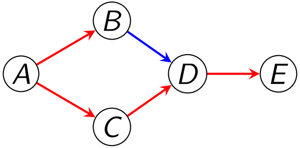
under the SWIG causal model associated with $\mathcal{G}$.

Theorem $\boldsymbol{S}$ is a valid adjustment set if it satisfies the backdoor criterion:

1. $\boldsymbol{S}$ contains no descendant of A,
2. No d-connecting path between A and Y given $\boldsymbol{S}$ with an arrowhead into A.

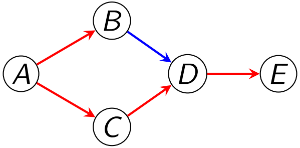
Backdoor/adjustment

- Identifying $P(E(b'))$.



Backdoor/adjustment

- Identifying $P(E(b'))$.

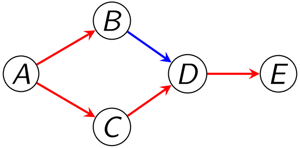


- Valid adjustment sets:

$$\{C\}, \quad \{A\}, \quad \{A, C\}.$$

Backdoor/adjustment

- Identifying $P(E(b'))$.



- Valid adjustment sets:

$$\{C\}, \quad \{A\}, \quad \{A, C\}.$$

☞ Adjusting for $\{C\}$ is the most efficient (Henckel et al., 2022; Rotnitzky and Smucler, 2020).

- A different, but more efficient estimator uses C, D (Guo, Perković, et al., 2023).

In the presence of latent variables

- For a DAG with latent variables $\boldsymbol{U}$, we can use latent projection to obtain an ADMG (acyclic directed mixed graph) over observed variables.

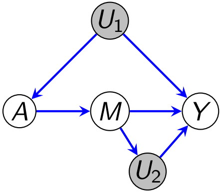
1. Whenever there is a path of the form $w \to u_1 \to \cdots \to u_2 \to v$ add $w \to v$ (if not already present).
2. Whenever there is a path of the form $w \leftarrow u_1 \leftarrow \cdots \to u_2 \to v$ add $w \leftrightarrow v$.

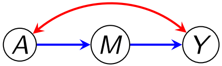
In the presence of latent variables

- For a DAG with latent variables $\boldsymbol{U}$, we can use latent projection to obtain an ADMG (acyclic directed mixed graph) over observed variables.

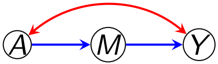
> 1. Whenever there is a path of the form $w \to u_1 \to \cdots \to u_2 \to v$ add $w \to v$ (if not already present).
> 2. Whenever there is a path of the form $w \leftarrow u_1 \leftarrow \cdots \to u_2 \to v$ add $w \leftrightarrow v$.

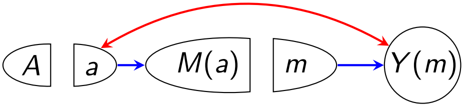
- Example





Front-door formula

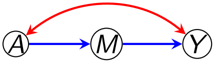


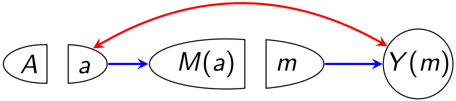


Front-door formula

$$\mathbb{E}\,Y(a) = \sum_m \overbrace{\mathbb{E}\{\mathbb{E}[Y \mid M = m, A]\}}^{\mathbb{E}\,Y(m)} \times P(M = m \mid A = a).$$

Front-door formula





$$
\begin{aligned}
P(Y(a)=y) &= P(Y(M(a))=y) \\
&= \sum_m P(Y(M(a))=y \mid M(a)=m)\, P(M(a)=m) \quad \text{(why?)} \\
&= \sum_m P(Y(m)=y \mid M(a)=m)\, P(M(a)=m) \\
&= \sum_m P(Y(m)=y)\, P(M(a)=m) \quad \text{(why?)} \\
&= \sum_m \left\{ \sum_{a'} P(Y=y \mid M=m, A=a') P(A=a') \right\} P(M(a)=m \mid A=a) \quad \text{(why?)} \\
&= \sum_m \left\{ \sum_{a'} P(Y=y \mid M=m, A=a') P(A=a') \right\} P(M=m \mid A=a).
\end{aligned}
$$

Identification in the presence of latent variables

- Not all (single-world) counterfactual quantities are identified.
- The ID algorithm due to Jin Tian provides a **complete solution**.
  - See also Shpitser and Pearl (2006), Richardson, Robin J. Evans, et al. (2023, §4.3).

DAG as a tool for practitioners

Use of DAGs in practice

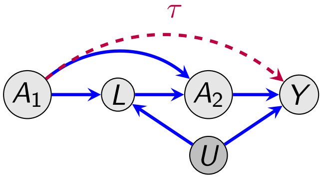
Practitioners can use DAGs/ADMGs to

Use of DAGs in practice

Practitioners can use DAGs/ADMGs to

1. Identify and communicate **biases** (confounding, selection). (yes!)

e.g., when there is a time-varying confounder

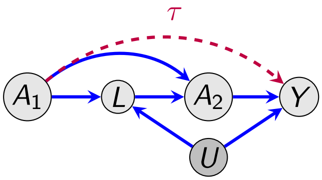


Use of DAGs in practice

Practitioners can use DAGs/ADMGs to

1. Identify and communicate **biases** (confounding, selection). (yes!)

   e.g., when there is a time-varying confounder



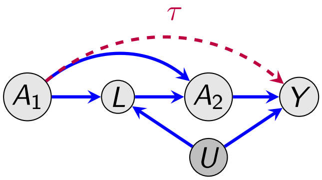
2. **Identify** a causal quantity of interest and consider its **estimation**. (to some extent)

Use of DAGs in practice

Practitioners can use DAGs/ADMGs to

1. Identify and communicate **biases** (confounding, selection). (yes!)

   e.g., when there is a time-varying confounder



2. **Identify** a causal quantity of interest and consider its **estimation**. (to some extent)
3. **Design** an observational study. (largely open)

Challenges

Challenges

1. Results strongly rely on the DAG model being correctly specified.

Challenges

1. Results strongly rely on the DAG model being correctly specified.
2. Often difficult to confidently specify (or learn) a DAG for applications.

Challenges

1. Results strongly rely on the DAG model being correctly specified.
2. Often difficult to confidently specify (or learn) a DAG for applications.
3. Sophisticated forms of nonparametric identification require detailed assumptions that are often difficult to justify in practice.

Challenges

1. Results strongly rely on the DAG model being correctly specified.
2. Often difficult to confidently specify (or learn) a DAG for applications.
3. Sophisticated forms of nonparametric identification require detailed assumptions that are often difficult to justify in practice.
4. Drawing a DAG can be a bad idea: unreliable and unnecessary.

Challenges

1. Results strongly rely on the DAG model being correctly specified.
2. Often difficult to confidently specify (or learn) a DAG for applications.
3. Sophisticated forms of nonparametric identification require detailed assumptions that are often difficult to justify in practice.
4. Drawing a DAG can be a bad idea: unreliable and unnecessary.
5. Model elicitation, robust methods, sensitivity analysis.
    - See Guo and Zhao (2023) for an interactive protocol of eliciting an adjustment set.

References I

Studenỳ, Milan (1992). "Conditional independence relations have no finite complete characterization". In: *Information Theory, Statistical Decision Functions and Random Processes. Transactions of the 11th Prague Conference vol. B*, pp. 377–396.

Richardson, Thomas and Peter Spirtes (2002). "Ancestral graph Markov models". In: *The Annals of Statistics* 30.4, pp. 962–1030.

Richardson, Thomas (2003). "Markov properties for acyclic directed mixed graphs". In: *Scandinavian Journal of Statistics* 30.1, pp. 145–157.

Evans, Robin J (2016). "Graphs for margins of Bayesian networks". In: *Scandinavian Journal of Statistics* 43.3, pp. 625–648.

References II

Richardson, Thomas, Robin J. Evans, et al. (2023). "Nested Markov properties for acyclic directed mixed graphs". In: *The Annals of Statistics* 51.1, pp. 334–361. DOI: 10.1214/22-AOS2253. URL: https://doi.org/10.1214/22-AOS2253.

Richardson, Thomas and James M Robins (2013). "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, p. 2013.

Robins, James (1986). "A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect". In: *Mathematical modelling* 7.9-12, pp. 1393–1512.

References III

- Henckel, Leonard et al. (2022). "Graphical criteria for efficient total effect estimation via adjustment in causal linear models". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 84.2, pp. 579–599.
- Rotnitzky, Andrea and Ezequiel Smucler (2020). "Efficient adjustment sets for population average causal treatment effect estimation in graphical models". In: *Journal of Machine Learning Research* 21.188, pp. 1–86.
- Guo, F Richard, Emilija Perković, et al. (2023). "Variable elimination, graph reduction and the efficient g-formula". In: *Biometrika* 110.3, pp. 739–761.
- Shpitser, Ilya and Judea Pearl (2006). "Identification of joint interventional distributions in recursive semi-Markovian causal models". In: *AAAI*, pp. 1219–1226.
- Guo, F Richard and Qingyuan Zhao (2023). "Confounder selection via iterative graph expansion". In: *arXiv preprint arXiv:2309.06053*.