Causal Inference Methods in Data Science
Lecture 2: More on Causal Identifications, DAGs, SWIGs, Sensitivity Analysis, Mediation

Lin Liu

July 17, 2025

two nice online seminars

online causal inference seminar (ocis):
https://sites.google.com/view/ocis/

the Gary Chamberlain online seminar in econometrics:
https://www.chamberlainseminar.org/

Definition of a confounder

A variable is a confounder, shall it satisfy the following axioms:

- controlling for all "confounders" suffices to control for "confounding"

Definition of a confounder

A variable is a confounder, shall it satisfy the following axioms:

- controlling for all "confounders" suffices to control for "confounding"
- each confounder should help eliminate or reduce "confounding bias" in some context

Definition of a confounder

A variable is a confounder, shall it satisfy the following axioms:

- controlling for all "confounders" suffices to control for "confounding"
- each confounder should help eliminate or reduce "confounding bias" in some context

REF: VanderWeele, Shpitser. On the definition of a confounder. AoS (2013).

Definition of a confounder

A variable is a confounder, shall it satisfy the following axioms:

- controlling for all "confounders" suffices to control for "confounding"
- each confounder should help eliminate or reduce "confounding bias" in some context

REF: VanderWeele, Shpitser. On the definition of a confounder. AoS (2013).

C is a confounder relative to the effect of A on Y if there exists a set of pre-treatment (A) covariates X such that

(1) $Y(a) \perp\!\!\!\perp A|X, C$

(2) But there is no proper subset $Z \subsetneq (X, C)$ such that $Y(a) \perp\!\!\!\perp A|Z$

Rigorous derivation of ATE identification

We need the following ID (identification) conditions

1. Consistency: $Y = AY(1) + (1 - A)Y(0)$
2. Positivity: $0 < \Pr(A = a|X) < 1$ almost surely
3. No unmeasured confounding/ignorability: $Y(a) \perp\!\!\!\perp A|X, \ \forall a \in \{0, 1\}$

Static g-formula of $\mathbb{E}[Y(1)]$ in observational studies under ignorability

$$\begin{aligned}\mathbb{E}[Y(1)] &= \mathbb{E}[\mathbb{E}[Y(1)|X]] \\ &\overset{2,3}{=} \mathbb{E}[\mathbb{E}[Y(1)|X, A = 1]] \\ &\overset{1}{=} \mathbb{E}[\mathbb{E}[Y|X, A = 1]]\end{aligned}$$

Rigorous derivation of ATE identification

We need the following ID (identification) conditions

1. Consistency: $Y = AY(1) + (1 - A)Y(0)$
2. Positivity: $0 < \Pr(A = a|X) < 1$ almost surely
3. No unmeasured confounding/ignorability: $Y(a) \perp\!\!\!\perp A|X, \ \forall a \in \{0, 1\}$

IPW of $\mathbb{E}[Y(1)]$ in observational studies under ignorability

$$
\begin{aligned}
\mathbb{E}[Y(1)] &= \mathbb{E}[\mathbb{E}[Y|X, A = 1]] \\
&= \int \mathbb{E}[Y|X = x, A = 1]p(x)\mathrm{d}x \\
&\overset{2}{=} \int \frac{y}{p(A = 1|X = x)} p(y|X = x, A = 1)p(A = 1|X = x)p(x)\mathrm{d}y\mathrm{d}x \\
&= \int \frac{ay}{p(A = 1|X = x)} p(y|X = x, A = 1)p(A = a|X = x)p(x)\mathrm{d}y\mathrm{d}a\mathrm{d}x \\
&= \mathbb{E}\left[\frac{AY}{p(A = 1|X)}\right]
\end{aligned}
$$

A question to think about

Can you derive the IPW formula without using g-formula?

Causal Identification

- As we have mentioned, identification is the core that distinguishes causal inference from other branches of statistics or data sciences

Causal Identification

- As we have mentioned, identification is the core that distinguishes causal inference from other branches of statistics or data sciences
- Identification connects data to science

Causal Identification

- As we have mentioned, identification is the core that distinguishes causal inference from other branches of statistics or data sciences
- Identification connects data to science
- High-level description:
  - What are we trying to learn from data? Write down the quantity of interest in terms of counterfactuals
  - Looking for a set of conditions (identification conditions) such that
  - Counterfactuals can be written as functionals/parameters of the distribution of observables

Causal Identification

- As we have mentioned, identification is the core that distinguishes causal inference from other branches of statistics or data sciences
- Identification connects data to science
- High-level description:
  - What are we trying to learn from data? Write down the quantity of interest in terms of counterfactuals
  - Looking for a set of conditions (identification conditions) such that
  - Counterfactuals can be written as functionals/parameters of the distribution of observables
  - The rest is statistics and computing (optimization)

outline

- We will take a deeper dive into counterfactuals and DAGs (SEM will not be our focus)

outline

- We will take a deeper dive into counterfactuals and DAGs (SEM will not be our focus)
- We will learn some basic knowledge of DAGs

outline

- We will take a deeper dive into counterfactuals and DAGs (SEM will not be our focus)
- We will learn some basic knowledge of DAGs
- Then we will show how to unify counterfactual and DAG into a single entity: SWIG

outline

- We will take a deeper dive into counterfactuals and DAGs (SEM will not be our focus)
- We will learn some basic knowledge of DAGs
- Then we will show how to unify counterfactual and DAG into a single entity: SWIG
- We will also learn some philosophical difference between Robins and Pearl, manifested by the difference between Robins' SWIG and Pearl's NPSEM-IE causal model

outline

- We will take a deeper dive into counterfactuals and DAGs (SEM will not be our focus)
- We will learn some basic knowledge of DAGs
- Then we will show how to unify counterfactual and DAG into a single entity: SWIG
- We will also learn some philosophical difference between Robins and Pearl, manifested by the difference between Robins' SWIG and Pearl's NPSEM-IE causal model
- Sensitivity analysis

outline

- We will take a deeper dive into counterfactuals and DAGs (SEM will not be our focus)
- We will learn some basic knowledge of DAGs
- Then we will show how to unify counterfactual and DAG into a single entity: SWIG
- We will also learn some philosophical difference between Robins and Pearl, manifested by the difference between Robins' SWIG and Pearl's NPSEM-IE causal model
- Sensitivity analysis
- Mediation analysis

outline

- We will take a deeper dive into counterfactuals and DAGs (SEM will not be our focus)
- We will learn some basic knowledge of DAGs
- Then we will show how to unify counterfactual and DAG into a single entity: SWIG
- We will also learn some philosophical difference between Robins and Pearl, manifested by the difference between Robins' SWIG and Pearl's NPSEM-IE causal model
- Sensitivity analysis
- Mediation analysis
- Next chapter is on dynamic causal inference, and mediation analysis is also a sort of dynamic causal inference

Directed Acyclic Graphs (DAGs)

- (Re-)invented by Judea Pearl in 1980's and percolated into statistics around 1990's after his paper in Biometrika

Directed Acyclic Graphs (DAGs)

- (Re-)invented by Judea Pearl in 1980's and percolated into statistics around 1990's after his paper in Biometrika

- Actual origin: geneticist Sewall Wright in 1918, 1934

Directed Acyclic Graphs (DAGs)

- (Re-)invented by Judea Pearl in 1980's and percolated into statistics around 1990's after his paper in Biometrika
- Actual origin: geneticist Sewall Wright in 1918, 1934
- DAGs are very convenient to represent background causal knowledge
- A causes Y: in word, it means

$$\begin{cases} \text{i)} & A \text{ precedes } Y \text{ in time ordering} \\ \text{ii)} & \text{a change in } Y \text{ is only due to a change in } A \end{cases}$$

Represent it in graph:

Directed Acyclic Graphs (DAGs)

- (Re-)invented by Judea Pearl in 1980's and percolated into statistics around 1990's after his paper in Biometrika

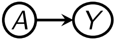
- Actual origin: geneticist Sewall Wright in 1918, 1934

- DAGs are very convenient to represent background causal knowledge

- A causes Y: in word, it means

$$\begin{cases} \text{i)} & A \text{ precedes } Y \text{ in time ordering} \\ \text{ii)} & \text{a change in } Y \text{ is only due to a change in } A \end{cases}$$

Represent it in graph:



A is called a parent of Y w.r.t. $\mathcal{G}$, denoted as $\text{pa}_{\mathcal{G}}(Y) \equiv \text{pa}(Y)$; A precedes Y in topological ordering in graphical language

Directed Acyclic Graphs (DAGs)

- (Re-)invented by Judea Pearl in 1980's and percolated into statistics around 1990's after his paper in Biometrika

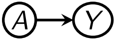
- Actual origin: geneticist Sewall Wright in 1918, 1934

- DAGs are very convenient to represent background causal knowledge

- A causes Y: in word, it means

$$\begin{cases} \text{i)} & A \text{ precedes } Y \text{ in time ordering} \\ \text{ii)} & \text{a change in } Y \text{ is only due to a change in } A \end{cases}$$

  Represent it in graph:



  A is called a parent of Y w.r.t. $\mathcal{G}$, denoted as $\text{pa}_{\mathcal{G}}(Y) \equiv \text{pa}(Y)$; A precedes Y in topological ordering in graphical language
  A precedes Y in time in human language

Definition of DAGs: Terminologies

- Graph: $\mathcal{G} = (\boldsymbol{V}, \boldsymbol{E})$, a set of vertices/nodes ($\boldsymbol{V}$) and a set of edges connecting the nodes ($\boldsymbol{E}$)

Definition of DAGs: Terminologies

- Graph: $\mathcal{G} = (\boldsymbol{V}, \boldsymbol{E})$, a set of vertices/nodes ($\boldsymbol{V}$) and a set of edges connecting the nodes ($\boldsymbol{E}$)
- Edge: either undirected – or directed $\rightarrow, \leftarrow$

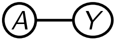


Figure: Either A causes Y or Y causes A

Definition of DAGs: Terminologies

- Graph: $\mathcal{G} = (\boldsymbol{V}, \boldsymbol{E})$, a set of vertices/nodes ($\boldsymbol{V}$) and a set of edges connecting the nodes ($\boldsymbol{E}$)

- Edge: either undirected – or directed $\rightarrow, \leftarrow$

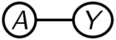


Figure: Either A causes Y or Y causes A

- Graphs with only (un)directed edges: (un)directed graphs

- Graphs with both directed and undirected edges: partially directed graphs

Definition of DAGs: More terminologies (paths family)

- *path* between nodes X and Y on $\mathcal{G}$
  any sequence of **distinct nodes** $(X, V_1, \cdots, V_k, Y)$, $k \geq 0$, such that any two successive nodes are connected by an edge
  e.g. $X \rightarrow V_1 - V_2 \leftarrow \cdots \leftarrow V_k - Y$

Definition of DAGs: More terminologies (paths family)

- *path* between nodes X and Y on $\mathcal{G}$
  any sequence of **distinct nodes** $(X, V_1, \cdots, V_k, Y)$, $k \geq 0$, such that any two successive nodes are connected by an edge
  e.g. $X \rightarrow V_1 - V_2 \leftarrow \cdots \leftarrow V_k - Y$

- *causal path* between X and Y: $X \rightarrow \cdots \rightarrow Y$

Definition of DAGs: More terminologies (paths family)

- *path* between nodes X and Y on $\mathcal{G}$
  any sequence of **distinct nodes** $(X, V_1, \cdots, V_k, Y)$, $k \geq 0$, such that any two successive nodes are connected by an edge
  e.g. $X \rightarrow V_1 - V_2 \leftarrow \cdots \leftarrow V_k - Y$

- *causal path* between X and Y: $X \rightarrow \cdots \rightarrow Y$

- Causal path between X and Y and an edge $Y \rightarrow X$ form a *directed cycle*

Definition of DAGs: More terminologies (paths family)

- *path* between nodes X and Y on $\mathcal{G}$
  any sequence of **distinct nodes** $(X, V_1, \cdots, V_k, Y)$, $k \geq 0$, such that any two successive nodes are connected by an edge
  e.g. $X \to V_1 - V_2 \leftarrow \cdots \leftarrow V_k - Y$

- *causal path* between X and Y: $X \to \cdots \to Y$

- Causal path between X and Y and an edge $Y \to X$ form a *directed cycle*

- Directed graphs without directed cycles are DAGs

Definition of DAGs: More terminologies (paths family)

- *path* between nodes X and Y on $\mathcal{G}$
  any sequence of **distinct nodes** $(X, V_1, \cdots, V_k, Y)$, $k \geq 0$, such that any two successive nodes are connected by an edge
  e.g. $X \rightarrow V_1 - V_2 \leftarrow \cdots \leftarrow V_k - Y$

- *causal path* between X and Y: $X \rightarrow \cdots \rightarrow Y$

- Causal path between X and Y and an edge $Y \rightarrow X$ form a *directed cycle*

- Directed graphs without directed cycles are DAGs

- $X = \text{de}_{\mathcal{G}}(Y) \equiv \text{de}(Y)$ if there is any causal path from Y to X

Definition of DAGs: More terminologies (paths family)

- *path* between nodes X and Y on $\mathcal{G}$
  any sequence of **distinct nodes** $(X, V_1, \cdots, V_k, Y)$, $k \geq 0$, such that any two successive nodes are connected by an edge
  e.g. $X \rightarrow V_1 - V_2 \leftarrow \cdots \leftarrow V_k - Y$

- *causal path* between X and Y: $X \rightarrow \cdots \rightarrow Y$

- Causal path between X and Y and an edge $Y \rightarrow X$ form a *directed cycle*

- Directed graphs without directed cycles are DAGs

- $X = \text{de}_{\mathcal{G}}(Y) \equiv \text{de}(Y)$ if there is any causal path from Y to X

- $X = \text{an}_{\mathcal{G}}(Y) \equiv \text{an}(Y)$ if there is any causal path from X to Y

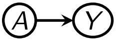
Definition of DAGs: More terminologies (paths family)

- *path* between nodes X and Y on $\mathcal{G}$
  any sequence of **distinct nodes** $(X, V_1, \cdots, V_k, Y)$, $k \geq 0$, such that any two successive nodes are connected by an edge
  e.g. $X \rightarrow V_1 - V_2 \leftarrow \cdots \leftarrow V_k - Y$
- *causal path* between X and Y: $X \rightarrow \cdots \rightarrow Y$
- Causal path between X and Y and an edge $Y \rightarrow X$ form a *directed cycle*
- Directed graphs without directed cycles are DAGs
- $X = \text{de}_{\mathcal{G}}(Y) \equiv \text{de}(Y)$ if there is any causal path from Y to X
- $X = \text{an}_{\mathcal{G}}(Y) \equiv \text{an}(Y)$ if there is any causal path from X to Y
- On a path, if $X_i \rightarrow X_j \leftarrow X_k$, then X_j is a collider

A quick quiz: is it possible $X = \mathrm{an}(Y)$ and $X = \mathrm{de}(Y)$ in a DAG?

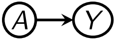
Examples of DAGs

- A causes Y

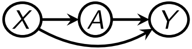


Examples of DAGs

- A causes Y

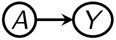


- A causes Y and X is a common cause (confounder) of A and Y

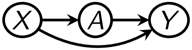


Examples of DAGs

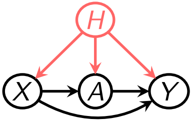
- A causes Y



- A causes Y and X is a common cause (confounder) of A and Y



- Observational studies with unmeasured/latent confounders:



One potential disadvantage of DAG: unless we are crystally clear about which important confounders are unmeasured, it is impossible to draw such a DAG

From DAGs without latent variables to probability distribution (data)

Markov factorization of DAGs:

Definition 1

A probability density function f over the variables $\boldsymbol{V}$ is consistent with DAG $\mathcal{G} = (\boldsymbol{V}, \boldsymbol{E})$ ($\boldsymbol{E}$ is the set of all directed edges) if it factorizes according to the following rule:

$$f(\boldsymbol{v}) = \prod_{V_i \in \boldsymbol{V}} f\left(v_i | \mathrm{pa}_{\mathcal{G}}(v_i)\right) \tag{1}$$

IPW, interventional distribution and do calculus

Derive the IPW formula without using g-formula

IPW, interventional distribution and do calculus

Derive the IPW formula without using g-formula

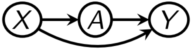
What is the “physical” meaning of $Y(1)$ or $Y|\mathrm{do}(A = 1)$?

IPW, interventional distribution and do calculus

Derive the IPW formula without using g-formula

What is the "physical" meaning of $Y(1)$ or $Y|\mathrm{do}(A=1)$?

Observational study satisfying "no unmeasured confounding"

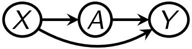


IPW, interventional distribution and do calculus

Derive the IPW formula without using g-formula

What is the “physical” meaning of $Y(1)$ or $Y|\mathrm{do}(A=1)$?

Observational study satisfying “no unmeasured confounding”



By Markovian factorization of DAG,

$$p(X, A, Y) = p(Y|X, A)p(A|X)p(X)$$

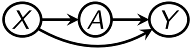
so when data is drawn from the above DAG, A is drawn from a distribution with pdf/pmf $p(A|X)$

IPW, interventional distribution and do calculus

Derive the IPW formula without using g-formula

What is the "physical" meaning of $Y(1)$ or $Y|\mathrm{do}(A=1)$?

Observational study satisfying "no unmeasured confounding"

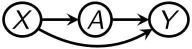


IPW, interventional distribution and do calculus

Derive the IPW formula without using g-formula

What is the "physical" meaning of $Y(1)$ or $Y|\mathrm{do}(A=1)$?

Observational study satisfying "no unmeasured confounding"



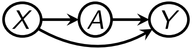
$\mathrm{do}(A=1)$ means intervening everybody's A to 1; the distribution of $Y(1)$ or $Y|\mathrm{do}(A=1)$ is the distribution Y under this interventional distribution!

IPW, interventional distribution and do calculus

Derive the IPW formula without using g-formula

What is the "physical" meaning of $Y(1)$ or $Y|\mathrm{do}(A=1)$?

Observational study satisfying "no unmeasured confounding"



$\mathrm{do}(A=1)$ means intervening everybody's A to 1; the distribution of $Y(1)$ or $Y|\mathrm{do}(A=1)$ is the distribution Y under this interventional distribution!

The interventional distribution:

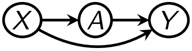
$$p(X, Y(1)) \equiv p(X, \mathrm{do}(A=1), Y) = p(Y|X, \mathrm{do}(A=1))\mathbb{1}\{A=1\}p(X)$$

IPW, interventional distribution and do calculus

Derive the IPW formula without using g-formula

What is the "physical" meaning of $Y(1)$ or $Y|\mathrm{do}(A=1)$?

Observational study satisfying "no unmeasured confounding"

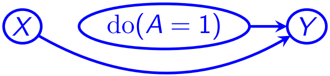


$\mathrm{do}(A=1)$ means intervening everybody's A to 1; the distribution of $Y(1)$ or $Y|\mathrm{do}(A=1)$ is the distribution Y under this interventional distribution!

The interventional distribution:

$$p(X, Y(1)) \equiv p(X, \mathrm{do}(A=1), Y) = p(Y|X, \mathrm{do}(A=1))\mathbb{1}\{A=1\}p(X)$$

The DAG:



IPW, interventional distribution and do calculus

From the observed data distribution to the interventional distribution?

$$p(X, A, Y) = p(Y|X, A)p(A|X)p(X) \text{ vs.}$$
$$p(X, \mathrm{do}(A = 1), Y) = p(Y|X, \mathrm{do}(A = 1))\mathbb{1}\{A = 1\}p(X)$$

IPW, interventional distribution and do calculus

From the observed data distribution to the interventional distribution?

$$p(X, A, Y) = p(Y|X, A)p(A|X)p(X) \text{ vs.}$$
$$p(X, \mathrm{do}(A = 1), Y) = p(Y|X, \mathrm{do}(A = 1))\mathbb{1}\{A = 1\}p(X)$$

Change of probability distribution

$$p(X, \mathrm{do}(A = 1), Y) = \frac{\mathbb{1}\{A = 1\}}{p(A|X)} p(X, A, Y) = \frac{A}{p(A|X)} p(X, A, Y)$$

IPW, interventional distribution and do calculus

From the observed data distribution to the interventional distribution?

$$p(X, A, Y) = p(Y|X, A)p(A|X)p(X) \text{ vs.}$$
$$p(X, \mathrm{do}(A = 1), Y) = p(Y|X, \mathrm{do}(A = 1))\mathbb{1}\{A = 1\}p(X)$$

Change of probability distribution

$$p(X, \mathrm{do}(A = 1), Y) = \frac{\mathbb{1}\{A = 1\}}{p(A|X)} p(X, A, Y) = \frac{A}{p(A|X)} p(X, A, Y)$$

So

$$\mathbb{E}[Y(1)] \equiv \mathbb{E}[Y|\mathrm{do}(A = 1)] = \mathbb{E}\left[\frac{A}{p(A|X)} Y\right]$$

d-separation

- To read off (conditional) independence constraints implied by the Markovian factorization, we need to introduce the following *d*-separation **graphical rule**:

d-separation

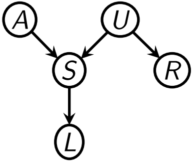
- To read off (conditional) independence constraints implied by the Markovian factorization, we need to introduce the following d-separation **graphical rule**:
  1. Between any two nodes X and Y, a set of nodes $\boldsymbol{Z}$ *not* containing X and Y, a path p from X to Y is blocked by $\boldsymbol{Z}$ if (1) there exists non-collider $W \in \boldsymbol{Z}$ also on p and (2) there exists a collider V but no member of de(V) is in $\boldsymbol{Z}$; otherwise, $\boldsymbol{Z}$ does not block the path p

d-separation

- To read off (conditional) independence constraints implied by the Markovian factorization, we need to introduce the following d-separation **graphical rule**:
  1. Between any two nodes X and Y, a set of nodes $\boldsymbol{Z}$ *not* containing X and Y, a path p from X to Y is blocked by $\boldsymbol{Z}$ if (1) there exists non-collider $W \in \boldsymbol{Z}$ also on p and (2) there exists a collider V but no member of $\mathrm{de}(V)$ is in $\boldsymbol{Z}$; otherwise, $\boldsymbol{Z}$ does not block the path p
  2. If $\boldsymbol{Z}$ blocks all paths between some disjoint subsets $\boldsymbol{X}$ and $\boldsymbol{Y}$, then $\boldsymbol{X}$ is d-separated from $\boldsymbol{Y}$ given $\boldsymbol{Z}$, denoted as $\boldsymbol{X} \perp_{\mathcal{G}} \boldsymbol{Y} | \boldsymbol{Z}$

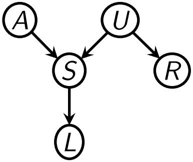
d-separation

- To read off (conditional) independence constraints implied by the Markovian factorization, we need to introduce the following *d*-separation **graphical rule**:
  1. Between any two nodes X and Y, a set of nodes $\boldsymbol{Z}$ *not* containing X and Y, a path p from X to Y is blocked by $\boldsymbol{Z}$ if (1) there exists non-collider $W \in \boldsymbol{Z}$ also on p and (2) there exists a collider V but no member of $\mathrm{de}(V)$ is in $\boldsymbol{Z}$; otherwise, $\boldsymbol{Z}$ does not block the path p
  2. If $\boldsymbol{Z}$ blocks all paths between some disjoint subsets $\boldsymbol{X}$ and $\boldsymbol{Y}$, then $\boldsymbol{X}$ is *d*-separated from $\boldsymbol{Y}$ given $\boldsymbol{Z}$, denoted as $\boldsymbol{X} \perp_{\mathcal{G}} \boldsymbol{Y} | \boldsymbol{Z}$
- Examples:



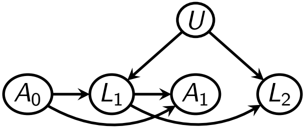
d-separation

- To read off (conditional) independence constraints implied by the Markovian factorization, we need to introduce the following d-separation **graphical rule**:
  1. Between any two nodes X and Y, a set of nodes $\boldsymbol{Z}$ *not* containing X and Y, a path p from X to Y is blocked by $\boldsymbol{Z}$ if (1) there exists non-collider $W \in \boldsymbol{Z}$ also on p and (2) there exists a collider V but no member of $\text{de}(V)$ is in $\boldsymbol{Z}$; otherwise, $\boldsymbol{Z}$ does not block the path p
  2. If $\boldsymbol{Z}$ blocks all paths between some disjoint subsets $\boldsymbol{X}$ and $\boldsymbol{Y}$, then $\boldsymbol{X}$ is d-separated from $\boldsymbol{Y}$ given $\boldsymbol{Z}$, denoted as $\boldsymbol{X} \perp_{\mathcal{G}} \boldsymbol{Y} | \boldsymbol{Z}$
- Examples:



$S \perp_{\mathcal{G}} R | U$, $A \not\perp_{\mathcal{G}} U | S$, $A \not\perp_{\mathcal{G}} U | L$

Another example: due to Verma



quiz: $\boldsymbol{X} = (A_0, A_1)$, $\boldsymbol{Y} = L_2$. Are $\boldsymbol{X}$ and $\boldsymbol{Y}$ d-separated given L_1?

connection to science: can you think of a scientific story for Verma's constraint?

d-separation and statistical independence

- When $\boldsymbol{X} \perp_{\mathcal{G}} \boldsymbol{Y}|\boldsymbol{Z}$, then for every distribution Markov factorized according to $\mathcal{G}$, $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{Z}$, where $\perp\!\!\!\perp$ stands for "statistical independence"
  – Soundness

d-separation and statistical independence

- When $\boldsymbol{X} \perp_{\mathcal{G}} \boldsymbol{Y} | \boldsymbol{Z}$, then for every distribution Markov factorized according to $\mathcal{G}$, $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} | \boldsymbol{Z}$, where $\perp\!\!\!\perp$ stands for "statistical independence"
  – Soundness

- When $\boldsymbol{X} \not\perp_{\mathcal{G}} \boldsymbol{Y} | \boldsymbol{Z}$, then there exists a distribution Markov factorized according to $\mathcal{G}$, such that $\boldsymbol{X} \not\perp\!\!\!\perp \boldsymbol{Y} | \boldsymbol{Z}$
  – Completeness

From DAG to causal DAG

- Causal DAG: DAG that contains all variables, observed or unobserved
- If all variables on a causal DAG are observed: “causal sufficiency”

From DAG to causal DAG

- Causal DAG: DAG that contains all variables, observed or unobserved
- If all variables on a causal DAG are observed: "causal sufficiency"
- Lack of an arrow from node X to node Y: absence of a **direct causal effect** from X to Y

From DAG to causal DAG

- Causal DAG: DAG that contains all variables, observed or unobserved
- If all variables on a causal DAG are observed: "causal sufficiency"
- Lack of an arrow from node X to node Y: absence of a **direct causal effect** from X to Y
- Any node is a cause of all its descendant; any node is caused by all its ancestors

From DAG to causal DAG

- Causal DAG: DAG that contains all variables, observed or unobserved
- If all variables on a causal DAG are observed: "causal sufficiency"
- Lack of an arrow from node X to node Y: absence of a **direct causal effect** from X to Y
- Any node is a cause of all its descendant; any node is caused by all its ancestors
- Markov factorization implies:

Conditional on its direct causes/parents, node X is independent of any node it does not cause (any non-descendant of X)

Faithfulness

Recall:

- When $\boldsymbol{X} \perp_{\mathcal{G}} \boldsymbol{Y} | \boldsymbol{Z}$, then for every distribution Markov factorized according to $\mathcal{G}$, $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} | \boldsymbol{Z}$, where $\perp\!\!\!\perp$ stands for "statistical independence"
  – Soundness

Faithfulness

Recall:

- When $\boldsymbol{X} \perp_{\mathcal{G}} \boldsymbol{Y} | \boldsymbol{Z}$, then for every distribution Markov factorized according to $\mathcal{G}$, $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} | \boldsymbol{Z}$, where $\perp\!\!\!\perp$ stands for "statistical independence"
  – Soundness

- When $\boldsymbol{X} \not\perp_{\mathcal{G}} \boldsymbol{Y} | \boldsymbol{Z}$, then there exists a distribution Markov factorized according to $\mathcal{G}$, such that $\boldsymbol{X} \not\perp\!\!\!\perp \boldsymbol{Y} | \boldsymbol{Z}$
  – Completeness

Faithfulness

Recall:

- When $\boldsymbol{X} \perp_{\mathcal{G}} \boldsymbol{Y}|\boldsymbol{Z}$, then for every distribution Markov factorized according to $\mathcal{G}$, $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{Z}$, where $\perp\!\!\!\perp$ stands for "statistical independence"
  – Soundness

- When $\boldsymbol{X} \not\perp_{\mathcal{G}} \boldsymbol{Y}|\boldsymbol{Z}$, then there exists a distribution Markov factorized according to $\mathcal{G}$, such that $\boldsymbol{X} \not\perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{Z}$
  – Completeness

But in some application, one might hope $\boldsymbol{X} \not\perp_{\mathcal{G}} \boldsymbol{Y}|\boldsymbol{Z} \Rightarrow \boldsymbol{X} \not\perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{Z}$

Faithfulness

Recall:

- When $\boldsymbol{X} \perp_{\mathcal{G}} \boldsymbol{Y}|\boldsymbol{Z}$, then for every distribution Markov factorized according to $\mathcal{G}$, $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{Z}$, where $\perp\!\!\!\perp$ stands for "statistical independence"
  – Soundness

- When $\boldsymbol{X} \not\perp_{\mathcal{G}} \boldsymbol{Y}|\boldsymbol{Z}$, then there exists a distribution Markov factorized according to $\mathcal{G}$, such that $\boldsymbol{X} \not\perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{Z}$
  – Completeness

But in some application, one might hope $\boldsymbol{X} \not\perp_{\mathcal{G}} \boldsymbol{Y}|\boldsymbol{Z} \Rightarrow \boldsymbol{X} \not\perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{Z}$

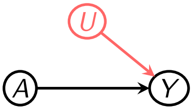
Does not always hold, how to proceed? Making such an assumption – called "faithfulness"; we will come back to this when we talk about causal discovery/structure learning (a lot of works on linear models done by the statistics group from ETH Zürich)

Causal DAG for RCT

Can you draw a causal DAG for RCT?

Causal DAG for RCT

Can you draw a causal DAG for RCT?

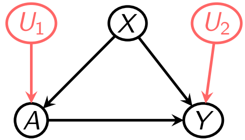


Causal DAG for observational studies under ignorability

Can you draw a causal DAG for observational studies under ignorability?

Causal DAG for observational studies under ignorability

Can you draw a causal DAG for observational studies under ignorability?



From causal graph to identification

- From last lecture, we have seen that identification is the process of turning parameters defined via counterfactuals into parameters defined via observables

From causal graph to identification

- From last lecture, we have seen that identification is the process of turning parameters defined via counterfactuals into parameters defined via observables
- There are a set of rules that can assist you to derive if the average causal effect of some node X to some other node Y can be identified by speculating the causal DAG

From causal graph to identification

- From last lecture, we have seen that identification is the process of turning parameters defined via counterfactuals into parameters defined via observables
- There are a set of rules that can assist you to derive if the average causal effect of some node X to some other node Y can be identified by speculating the causal DAG
- But they are not very convenient to use because there are no counterfactuals on DAG!

From causal graph to identification

- From last lecture, we have seen that identification is the process of turning parameters defined via counterfactuals into parameters defined via observables
- There are a set of rules that can assist you to derive if the average causal effect of some node X to some other node Y can be identified by speculating the causal DAG
- But they are not very convenient to use because there are no counterfactuals on DAG!
  - When causal graph theory was first introduced into the statistics community, this is the very reason why Donald Rubin and Guido Imbens (economist) are emphatically against the adoption of DAG

Backdoor criterion: identification rule 1 using DAG

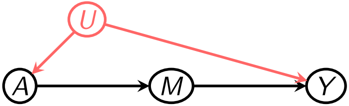
A set of covariates X satisfies the backdoor criterion relative to A and Y if all backdoor paths between A and Y are blocked given X and X does not include descendants of A

X satisfies backdoor criterion $\Leftrightarrow$ $Y(a) \perp\!\!\!\perp A|X$

Backdoor between A and Y: a path that starts with $A \leftarrow \cdots$

Front door criterion: identification rule 2 using DAG

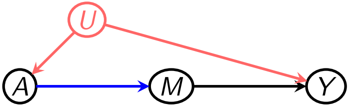
- Can we identify $\mathbb{E}[Y(a)]$ in the DAG below?



Yes! Front door criterion.

Front door criterion: identification rule 2 using DAG

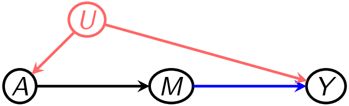
- Can we identify $\mathbb{E}[Y(a)]$ in the DAG below?



$p(M(a)) = p(M|A = a)$

Front door criterion: identification rule 2 using DAG

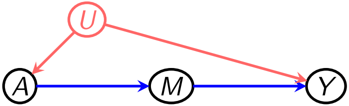
- Can we identify $\mathbb{E}[Y(a)]$ in the DAG below?



$p(Y(m)) = \int_a p(Y|M = m, A = a)p(a)\mathrm{d}a$

Front door criterion: identification rule 2 using DAG

- Can we identify $\mathbb{E}[Y(a)]$ in the DAG below?



$\mathbb{E}[Y(a)] = \int_y \int_m y p(Y(m) = y) p(M(a) = m) \mathrm{d}m \mathrm{d}y$ but can you see why?

SWIGs: putting counterfactuals on DAGs

Recall the three identification conditions

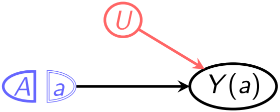
1. Consistency: $Y = \sum_a Y(a)\mathbb{1}\{A = a\}$
2. Positivity/Overlap: $0 < \Pr(A = a|X) < 1 \ \forall a$
3. Ignorability/Randomization: $Y(a) \perp\!\!\!\perp A|X \ \forall a$

SWIGs: putting counterfactuals on DAGs

Recall the three identification conditions

1. Consistency: $Y = \sum_a Y(a)\mathbb{1}\{A = a\}$
2. Positivity/Overlap: $0 < \Pr(A = a|X) < 1 \ \forall a$
3. Ignorability/Randomization: $Y(a) \perp\!\!\!\perp A|X \ \forall a$

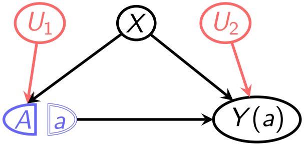
SWIG for RCT: splitting treatment node, then all descendants become counterfactuals



Reading independence: it is immediate $Y(a) \perp\!\!\!\perp A$ because A and $Y(a)$ are d-separated

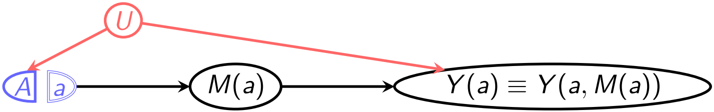
SWIGs: putting counterfactuals on DAGs

SWIG for observational studies under ignorability



Reading conditional independence (CI): it is immediate $Y(a) \perp\!\!\!\perp A|X$ because $Y(a) \perp_{\mathcal{G}_{swig}} A|X$

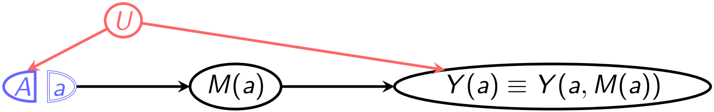
Front door with SWIG



so we have $(\star)$ $M(a) \perp\!\!\!\perp A$

$$\mathbb{E}[Y(a)] = \mathbb{E}[\mathbb{E}[Y(a)|M(a)]] = \int_m \mathbb{E}[Y(a)|M(a) = m]f(M(a) = m)\mathrm{d}m$$

Front door with SWIG

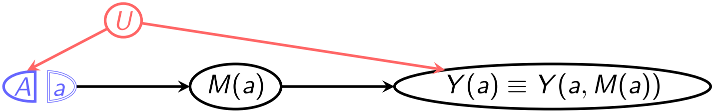


so we have $(\star)$ $M(a) \perp\!\!\!\perp A$

$$\mathbb{E}[Y(a)] = \mathbb{E}[\mathbb{E}[Y(a)|M(a)]] = \int_m \mathbb{E}[Y(a)|M(a) = m] f(M(a) = m)\mathrm{d}m$$

$$\overset{(\star)}{=} \int_m \mathbb{E}[Y(a)|M(a) = m] f(M = m|A = a)\mathrm{d}m$$

Front door with SWIG



so we have $(\star)$ $M(a) \perp\!\!\!\perp A$

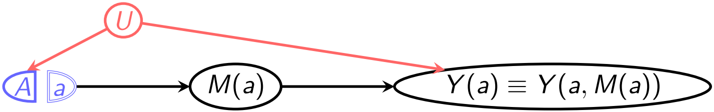
$$\mathbb{E}[Y(a)] = \mathbb{E}[\mathbb{E}[Y(a)|M(a)]] = \int_m \mathbb{E}[Y(a)|M(a) = m]f(M(a) = m)\mathrm{d}m$$

$$\stackrel{(\star)}{=} \int_m \mathbb{E}[Y(a)|M(a) = m]f(M = m|A = a)\mathrm{d}m$$

$$= \int_m \mathbb{E}[Y(a, M(a))|M(a) = m]f(M = m|A = a)\mathrm{d}m$$

consistency: by $Y(a, M(a)) = Y(a, m)$ given $M(a) = m$

Front door with SWIG



so we have $(\star)$ $M(a) \perp\!\!\!\perp A$

$$\mathbb{E}[Y(a)] = \mathbb{E}[\mathbb{E}[Y(a)|M(a)]] = \int_m \mathbb{E}[Y(a)|M(a) = m] f(M(a) = m)\mathrm{d}m$$

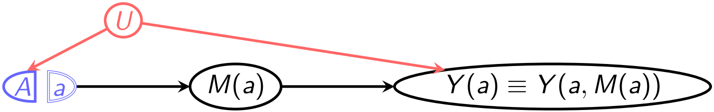
$$\overset{(\star)}{=} \int_m \mathbb{E}[Y(a)|M(a) = m] f(M = m|A = a)\mathrm{d}m$$

$$= \int_m \mathbb{E}[Y(a, M(a))|M(a) = m] f(M = m|A = a)\mathrm{d}m$$

consistency: by $Y(a, M(a)) = Y(a, m)$ given $M(a) = m$

$$= \int_m \mathbb{E}[Y(a, m)|M(a) = m] f(M = m|A = a)\mathrm{d}m$$

Front door with SWIG



so we have $(\star)$ $M(a) \perp\!\!\!\perp A$

$$\mathbb{E}[Y(a)] = \mathbb{E}[\mathbb{E}[Y(a)|M(a)]] = \int_m \mathbb{E}[Y(a)|M(a) = m] f(M(a) = m) \mathrm{d}m$$

$$\overset{(\star)}{=} \int_m \mathbb{E}[Y(a)|M(a) = m] f(M = m|A = a) \mathrm{d}m$$

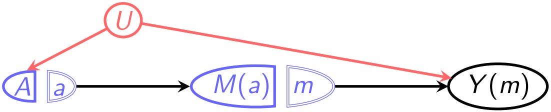
$$= \int_m \mathbb{E}[Y(a, M(a))|M(a) = m] f(M = m|A = a) \mathrm{d}m$$

consistency: by $Y(a, M(a)) = Y(a, m)$ given $M(a) = m$

$$= \int_m \mathbb{E}[Y(a, m)|M(a) = m] f(M = m|A = a) \mathrm{d}m$$

now we need a SWIG by intervening both A and M

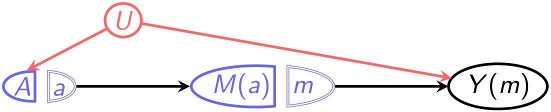
Front door with SWIG



so we have $M(a) \perp\!\!\!\perp A$ and $Y(m) \perp\!\!\!\perp M(a)$

$$\mathbb{E}[Y(a)] = \int_m \mathbb{E}[Y(a,m)|M(a) = m] f(M = m|A = a) \mathrm{d}m$$

Front door with SWIG

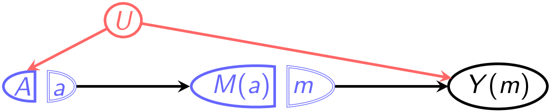


so we have $M(a) \perp\!\!\!\perp A$ and $Y(m) \perp\!\!\!\perp M(a)$

$$\mathbb{E}[Y(a)] = \int_m \mathbb{E}[Y(a,m)|M(a) = m] f(M = m|A = a)\mathrm{d}m$$
$$= \int_m \mathbb{E}[Y(m)|M(a) = m] f(M = m|A = a)\mathrm{d}m$$

by no direct causal effect of A on Y

Front door with SWIG



so we have $M(a) \perp\!\!\!\perp A$ and $Y(m) \perp\!\!\!\perp M(a)$

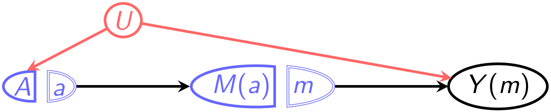
$$\mathbb{E}[Y(a)] = \int_m \mathbb{E}[Y(a,m)|M(a) = m] f(M = m|A = a) \mathrm{d}m$$

$$= \int_m \mathbb{E}[Y(m)|M(a) = m] f(M = m|A = a) \mathrm{d}m$$

by no direct causal effect of A on Y

$$= \int_m \mathbb{E}[Y(m)] f(M = m|A = a) \mathrm{d}m$$

Front door with SWIG



so we have $M(a) \perp\!\!\!\perp A$ and $Y(m) \perp\!\!\!\perp M(a)$

$$\mathbb{E}[Y(a)] = \int_m \mathbb{E}[Y(a,m)|M(a) = m] f(M = m|A = a)\mathrm{d}m$$

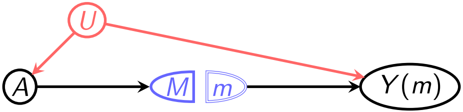
$$= \int_m \mathbb{E}[Y(m)|M(a) = m] f(M = m|A = a)\mathrm{d}m$$

by no direct causal effect of A on Y

$$= \int_m \mathbb{E}[Y(m)] f(M = m|A = a)\mathrm{d}m$$

now we need a world by intervening M alone

Front door with SWIG

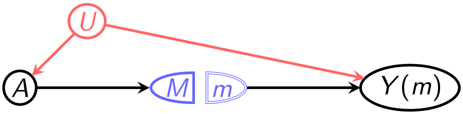


so we have $(\star)$ $Y(m) \perp\!\!\!\perp M | A$

$$\mathbb{E}[Y(a)]$$

$$= \int_m \mathbb{E}[Y(m)] f(M = m | A = a) \mathrm{d}m$$

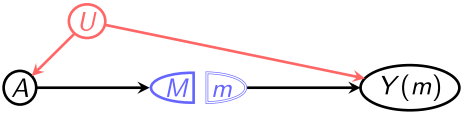
Front door with SWIG



so we have (⋆) $Y(m) \perp\!\!\!\perp M|A$

$$
\begin{aligned}
&\mathbb{E}[Y(a)] \\
&= \int_m \mathbb{E}[Y(m)] f(M=m|A=a)\mathrm{d}m \\
&\overset{(\star)}{=} \int_m \mathbb{E}_A[\mathbb{E}_{Y(m)}[Y(m)|A, M=m]] f(M=m|A=a)\mathrm{d}m \\
&= \int_m \mathbb{E}_A[\mathbb{E}_Y[Y|A, M=m]] f(M=m|A=a)\mathrm{d}m
\end{aligned}
$$

Front door with SWIG



so we have $(\star)$ $Y(m) \perp\!\!\!\perp M|A$

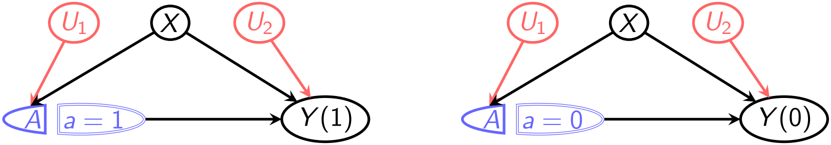
$$
\begin{aligned}
&\mathbb{E}[Y(a)] \\
&= \int_m \mathbb{E}[Y(m)] f(M=m|A=a)\mathrm{d}m \\
&\overset{(\star)}{=} \int_m \mathbb{E}_A[\mathbb{E}_{Y(m)}[Y(m)|A, M=m]] f(M=m|A=a)\mathrm{d}m \\
&= \int_m \mathbb{E}_A[\mathbb{E}_Y[Y|A, M=m]] f(M=m|A=a)\mathrm{d}m \\
&= \int_m \left\{ \int_{a'} \int_y y f(Y=y|A=a', M=m) f(A=a')\mathrm{d}a'\mathrm{d}y \right\} f(M=m|A=a)\mathrm{d}m
\end{aligned}
$$

What the hell is “single world”-ness?

- By node splitting, A becomes $A||a$

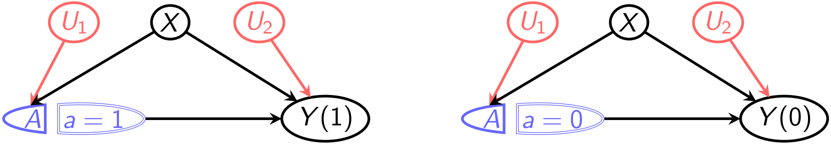
What the hell is “single world”-ness?

- By node splitting, A becomes $A||a$
- Say $A \in \{0, 1\}$, then if one is interested in two cases $a = 0, a = 1$, need to draw two separate graphs



What the hell is "single world"-ness?

- By node splitting, A becomes $A||a$
- Say $A \in \{0, 1\}$, then if one is interested in two cases $a = 0, a = 1$, need to draw two separate graphs



- $Y(1)$ and $Y(0)$ are never on the same graph! So the theory of SWIG cannot be used to identify the joint distribution of $(Y(1), Y(0))$

No cross-world identification – feature or bug?

- What do you think? Not allowing cross-world identification, is it an advantage or a disadvantage?

No cross-world identification – feature or bug?

- What do you think? Not allowing cross-world identification, is it an advantage or a disadvantage?
- Discuss among yourselves

No cross-world identification – feature or bug?

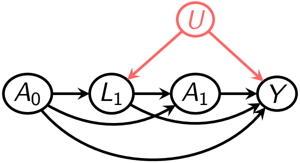
- What do you think? Not allowing cross-world identification, is it an advantage or a disadvantage?
- Discuss among yourselves
- Pros: For causal inference in reality, we can never observe $Y(1)$ and $Y(0)$ for one person so there is no way to estimate their joint distribution without making assumptions

No cross-world identification – feature or bug?

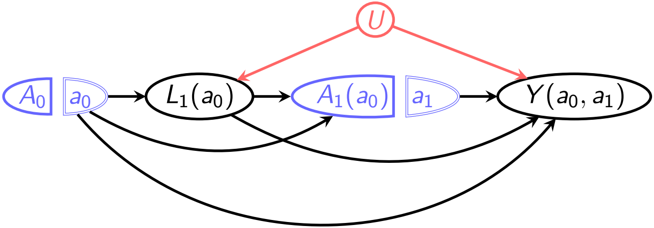
- What do you think? Not allowing cross-world identification, is it an advantage or a disadvantage?
- Discuss among yourselves
- Pros: For causal inference in reality, we can never observe $Y(1)$ and $Y(0)$ for one person so there is no way to estimate their joint distribution without making assumptions
- Cons: For those taking causality as a much higher-level meta-physical type of human inquiry, this is definitely limiting our "imagination power"

One more example of SWIG: Verma

Can we identify $\mathbb{E}[Y(a_0, a_1)]$ in the following complete Verma's graph?

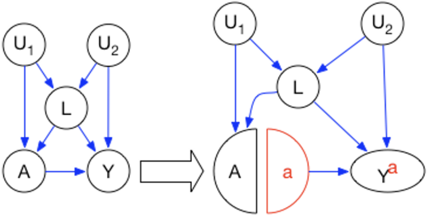


SWIG?



Reading CI: $Y(a_0, a_1) \perp\!\!\!\perp A_0$ and $Y(a_0, a_1) \perp\!\!\!\perp A_1(a_0) | L_1(a_0), A_0$. We will come back to this on Friday.

More Examples

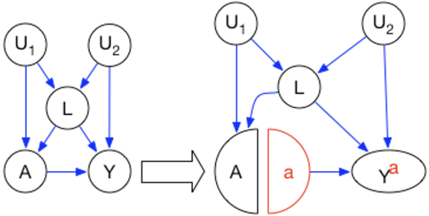


- We can read directly from the template that:

  $A \not\perp\!\!\!\perp Y^{a} \mid L.$

  – Conditioning on L opens the path $A \leftarrow U_1 \rightarrow L \leftarrow U_2 \rightarrow Y^{a}$.

More Examples



- We can also directly read that: $A \not\perp\!\!\!\perp Y^a$.
  - $A \leftarrow U_1 \rightarrow L \rightarrow Y^a$ and $A \leftarrow L \leftarrow U_2 \rightarrow Y^a$ are both open.
- So we have intractable confounding.

Non-Homework :-)

Read the following papers:
Richardson & Robins, Single world intervention graphs (148 pages);

Shpitser, Richardson & Robins, Multivariate Counterfactual Systems And Causal Graphical Models (34 pages)

Sensitivity Analysis

Sensitivity analysis

Sensitivity analysis

At the end of this lecture, let's discuss what sensitivity analysis is

Sensitivity analysis

At the end of this lecture, let's discuss what sensitivity analysis is

Sensitivity analysis is a technique very commonly used in applied math, computational physics, or any area mathematical modeling is used: check how our results change by perturbing some parameters in the model; it is important for science

Sensitivity analysis

At the end of this lecture, let's discuss what sensitivity analysis is

Sensitivity analysis is a technique very commonly used in applied math, computational physics, or any area mathematical modeling is used: check how our results change by perturbing some parameters in the model; it is important for science

Have you done any sensitivity analysis yourself in your research?

Basic idea of sensitivity analysis

What does sensitivity analysis protect against in causal inference?

Basic idea of sensitivity analysis

What does sensitivity analysis protect against in causal inference?

Potential violation of identification conditions. Which one? Discuss

Basic idea of sensitivity analysis

What does sensitivity analysis protect against in causal inference?

Potential violation of identification conditions. Which one? Discuss

Randomization/Ignorability/No unmeasured confounding
REF: Cornfield et al. Smoking and lung cancer: recent evidence and a discussion of some questions. JNCI (1959)

Basic idea of sensitivity analysis

What does sensitivity analysis protect against in causal inference?

Potential violation of identification conditions. Which one? Discuss

Randomization/Ignorability/No unmeasured confounding
REF: Cornfield et al. Smoking and lung cancer: recent evidence and a discussion of some questions. JNCI (1959)

Smoking-Lung Cancer story

Smoking-Lung Cancer Story

To make the story easier, we consider binary outcome (lung caner or not), so $Y \in \{0, 1\}$

Smoking-Lung Cancer Story

To make the story easier, we consider binary outcome (lung caner or not), so $Y \in \{0, 1\}$

We further consider a new "causal measure" risk ratio (RR): $RR = \mathbb{E}[Y(1)]/\mathbb{E}[Y(0)]$ so RR $= 1$ means no causal effect

Smoking-Lung Cancer Story

To make the story easier, we consider binary outcome (lung caner or not), so $Y \in \{0, 1\}$

We further consider a new "causal measure" risk ratio (RR): $RR = \mathbb{E}[Y(1)]/\mathbb{E}[Y(0)]$ so RR = 1 means no causal effect

Doll and Hill (1950 BMJ): smoking-lung cancer RR under ignorability $RR^{obs}_{AY} = \frac{\Pr(Y=1|A=1)}{\Pr(Y=1|A=0)} \approx 9$

Figure: Sir Austin Bradford Hill (1897-1991)

Fisher's dispute

Sir R.A. Fisher (founding father of statistics, also Donald B. Rubin's academic grand father) disagreed with Doll and Hill (Fisher 1957 BMJ)

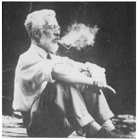


Figure: Sir Ronald Fisher (1890-1962)

Fisher's dispute

Sir R.A. Fisher (founding father of statistics, also Donald B. Rubin's academic grand father) disagreed with Doll and Hill (Fisher 1957 BMJ)

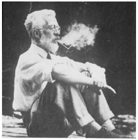


Figure: Sir Ronald Fisher (1890-1962)

Fisher: "... cigarette-smoking and lung cancer, though not mutually causative, are both influenced by a common cause U, in this case the individual genotype."

Fisher's dispute

Sir R.A. Fisher (founding father of statistics, also Donald B. Rubin's academic grand father) disagreed with Doll and Hill (Fisher 1957 BMJ)

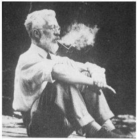


Figure: Sir Ronald Fisher (1890-1962)

Fisher: "... cigarette-smoking and lung cancer, though not mutually causative, are both influenced by a common cause U, in this case the individual genotype."

Can you draw Fisher's DAG?

Cornfield's dispute of Fisher's dispute

At the end of the world, two curmudgeons are still alive and they must be statisticians...

Cornfield's dispute of Fisher's dispute

At the end of the world, two curmudgeons are still alive and they must be statisticians...

Cornfield's back-of-envelope calculation: if Fisher is right, then we need

$$RR_{AU} \geq RR_{AY} \approx 9$$

Cornfield's dispute of Fisher's dispute

At the end of the world, two curmudgeons are still alive and they must be statisticians...

Cornfield's back-of-envelope calculation: if Fisher is right, then we need

$$RR_{AU} \geq RR_{AY} \approx 9$$

Geneticists: such smoking-genetic association is too strong to be realistic

Cornfield's dispute of Fisher's dispute

At the end of the world, two curmudgeons are still alive and they must be statisticians...

Cornfield's back-of-envelope calculation: if Fisher is right, then we need

$$RR_{AU} \geq RR_{AY} \approx 9$$

Geneticists: such smoking-genetic association is too strong to be realistic

Fisher is likely to be incorrect

How to get $RR_{AU} \geq RR_{AY}$?

Ignorability assumption ($\star$): $Y(a) \perp\!\!\!\perp A|U$, U unmeasured

How to get $RR_{AU} \geq RR_{AY}$?

Ignorability assumption ($\star$): $Y(a) \perp\!\!\!\perp A|U$, U unmeasured

Observed RR:

$$RR_{AY}^{obs} = \frac{\Pr(Y=1|A=1)}{\Pr(Y=1|A=0)} = \frac{\sum_{u=0,1}\Pr(Y=1|A=1,U=u)\Pr(U=u|A=1)}{\sum_{u=0,1}\Pr(Y=1|A=0,U=u)\Pr(U=u|A=0)}$$

How to get $RR_{AU} \geq RR_{AY}$?

Ignorability assumption ($\star$): $Y(a) \perp\!\!\!\perp A|U$, U unmeasured

Observed RR:

$$RR_{AY}^{obs} = \frac{\Pr(Y=1|A=1)}{\Pr(Y=1|A=0)} = \frac{\sum_{u=0,1}\Pr(Y=1|A=1,U=u)\Pr(U=u|A=1)}{\sum_{u=0,1}\Pr(Y=1|A=0,U=u)\Pr(U=u|A=0)}$$

Causal RR:

$$RR_{AY}^{true} = \frac{\Pr(Y(1)=1)}{\Pr(Y(0)=1)} = \frac{\sum_{u=0,1}\Pr(Y(1)=1|U=u)\Pr(U=u)}{\sum_{u=0,1}\Pr(Y(0)=1|U=u)\Pr(U=u)}$$

How to get $RR_{AU} \geq RR_{AY}$?

Ignorability assumption ($\star$): $Y(a) \perp\!\!\!\perp A|U$, U unmeasured

Observed RR:

$$RR_{AY}^{obs} = \frac{\Pr(Y=1|A=1)}{\Pr(Y=1|A=0)} = \frac{\sum_{u=0,1}\Pr(Y=1|A=1,U=u)\Pr(U=u|A=1)}{\sum_{u=0,1}\Pr(Y=1|A=0,U=u)\Pr(U=u|A=0)}$$

Causal RR:

$$RR_{AY}^{true} = \frac{\Pr(Y(1)=1)}{\Pr(Y(0)=1)} = \frac{\sum_{u=0,1}\Pr(Y(1)=1|U=u)\Pr(U=u)}{\sum_{u=0,1}\Pr(Y(0)=1|U=u)\Pr(U=u)}$$

$$\overset{(\star)}{=} \frac{\sum_{u=0,1}\Pr(Y=1|A=1,U=u)\Pr(U=u)}{\sum_{u=0,1}\Pr(Y=1|A=0,U=u)\Pr(U=u)}$$

Sensitivity analysis: can we bound RR_{AY}^{true} by RR_{AY}^{obs}?

Two definitions: $A - U$ association and $U - Y(a)$ association by RR

$$RR_{AU} := \frac{\Pr(U=1|A=1)}{\Pr(U=1|A=0)}$$

$$RR_{UY(a)} := \max\left\{\frac{\Pr(Y(a)=1|U=1)}{\Pr(Y(a)=1|U=0)}, \frac{\Pr(Y(a)=1|U=0)}{\Pr(Y(a)=1|U=1)}\right\}$$

$$RR_{UY} := \max\left\{RR_{UY(1)}, RR_{UY(0)}\right\}$$

so we reduce the sensitivity parameters to two measures RR_{AU} and RR_{UY} and they have nice interpretation

Sensitivity analysis: can we bound RR_{AY}^{true} by RR_{AY}^{obs}?

Two definitions: $A - U$ association and $U - Y(a)$ association by RR

$$RR_{AU} := \frac{\Pr(U=1|A=1)}{\Pr(U=1|A=0)}$$

$$RR_{UY(a)} := \max\left\{\frac{\Pr(Y(a)=1|U=1)}{\Pr(Y(a)=1|U=0)}, \frac{\Pr(Y(a)=1|U=0)}{\Pr(Y(a)=1|U=1)}\right\}$$

$$RR_{UY} := \max\left\{RR_{UY(1)}, RR_{UY(0)}\right\}$$

so we reduce the sensitivity parameters to two measures RR_{AU} and RR_{UY} and they have nice interpretation

BOUND (REF: Ding, VanderWeele. Sensitivity analysis without assumptions. Epidemiology (2017)): if $RR_{AY}^{obs} > 1$

$$\frac{RR_{AY}^{obs}}{RR_{AY}^{true}} \leq \frac{RR_{AU}RR_{UY}}{RR_{AU} + RR_{UY} - 1}$$

Proof sketch

Recall

$$RR_{AY}^{true} = \frac{\sum_{u=0,1} \Pr(Y=1|A=1,U=u)\Pr(U=u)}{\sum_{u=0,1} \Pr(Y=1|A=0,U=u)\Pr(U=u)}$$

Introduce two new notation

$$RR_{AY}^{true,+} = \frac{\sum_{u=0,1} \Pr(Y=1|A=1,U=u)\Pr(U=u|A=1)}{\sum_{u=0,1} \Pr(Y=1|A=0,U=u)\Pr(U=u|A=1)}$$

$$RR_{AY}^{true,-} = \frac{\sum_{u=0,1} \Pr(Y=1|A=1,U=u)\Pr(U=u|A=0)}{\sum_{u=0,1} \Pr(Y=1|A=0,U=u)\Pr(U=u|A=0)}$$

With some algebra, one can show

$$RR_{AY}^{true} = wRR_{AY}^{true,+} + (1-w)RR_{AY}^{true,-}$$

with

$$w = \frac{\sum_{u=0,1} \Pr(Y=1|A=0,U=u)\Pr(U=u,A=1)}{\sum_{u=0,1} \Pr(Y=1|A=0,U=u)\Pr(U=u,A=1)+\sum_{u=0,1} \Pr(Y=1|A=0,U=u)\Pr(U=u,A=0)}$$

Proof sketch

Similarly,

$$\left(\frac{RR_{AY}^{obs}}{RR_{AY}^{true}}\right)^{-1} = w\left(\frac{RR_{AY}^{obs}}{RR_{AY}^{true,+}}\right)^{-1} + (1-w)\left(\frac{RR_{AY}^{obs}}{RR_{AY}^{true,-}}\right)^{-1}$$

Now

$$\frac{RR_{AY}^{obs}}{RR_{AY}^{true,+}} = \frac{\frac{\sum_{u=0,1}\Pr(Y=1|A=1,U=u)\Pr(U=u|A=1)}{\sum_{u=0,1}\Pr(Y=1|A=0,U=u)\Pr(U=u|A=0)}}{\frac{\sum_{u=0,1}\Pr(Y=1|A=1,U=u)\Pr(U=u|A=1)}{\sum_{u=0,1}\Pr(Y=1|A=0,U=u)\Pr(U=u|A=1)}}$$

$$= \frac{\sum_{u=0,1}\Pr(Y=1|A=0,U=u)\Pr(U=u|A=1)}{\sum_{u=0,1}\Pr(Y=1|A=0,U=u)\Pr(U=u|A=0)}$$

$$= \frac{RR_{AU}\Pr(U=1|A=0)\left(\Pr(Y=1|A=0,U=1)-\Pr(Y=1|A=0,U=0)\right)+\Pr(Y=1|A=0,U=0)}{\Pr(U=1|A=0)\left(\Pr(Y=1|A=0,U=1)-\Pr(Y=1|A=0,U=0)\right)+\Pr(Y=1|A=0,U=0)}$$

$$= \frac{RR_{AU}\Pr(U=1|A=0)\left(RR_{UY(0)}-1\right)+1}{\Pr(U=1|A=0)\left(RR_{UY(0)}-1\right)+1}$$

$$\leq \frac{RR_{AU}RR_{UY(0)}}{RR_{AU}+RR_{UY(0)}-1}$$

in the last step, maximized at $RR_{AU}\Pr(U=1|A=0)=1$.

How to use such a bound?

$$\underbrace{\frac{RR_{AU}RR_{UY}}{RR_{AU} + RR_{UY} - 1}}_{\text{Bounding Factor (BF)}} \geq \frac{RR_{AY}^{obs}}{RR_{AY}^{true}}$$

How to use such a bound?

$$\underbrace{\frac{RR_{AU}RR_{UY}}{RR_{AU}+RR_{UY}-1}}_{\text{Bounding Factor (BF)}} \geq \frac{RR_{AY}^{obs}}{RR_{AY}^{true}}$$

To reduce the observed RR to the hypothetical true causal RR (e.g. no effect, taking value 1), BF needs to exceed RR_{AY}^{obs} to explain away the observed RR

How to use such a bound?

$$\underbrace{\frac{RR_{AU}RR_{UY}}{RR_{AU}+RR_{UY}-1}}_{\text{Bounding Factor (BF)}} \geq \frac{RR_{AY}^{obs}}{RR_{AY}^{true}}$$

To reduce the observed RR to the hypothetical true causal RR (e.g. no effect, taking value 1), BF needs to exceed RR_{AY}^{obs} to explain away the observed RR

Connect to Cornfield's analysis?

Fact 1

$$a \geq \frac{ab}{a+b-1} \text{ if } a \geq 1$$

so $RR_{AU} \geq BF$ and $RR_{UY} \geq BF$

BF table

bounding factor		RR_{UD} 1.3	1.5	1.8	2	2.5	3	3.5	4	5
	1.3	1.06	1.08	1.11	1.13	1.16	1.18	1.20	1.21	1.23
	1.5	1.08	1.12	1.17	1.20	1.25	1.29	1.31	1.33	1.36
	1.8	1.11	1.17	1.25	1.29	1.36	1.42	1.47	1.50	1.55
	2	1.13	1.20	1.29	1.33	1.43	1.50	1.56	1.60	1.67
RR_{EU}	2.5	1.16	1.25	1.36	1.43	1.56	1.67	1.75	1.82	1.92
	3	1.18	1.29	1.42	1.50	1.67	1.80	1.91	2.00	2.14
	3.5	1.20	1.31	1.47	1.56	1.75	1.91	2.04	2.15	2.33
	4	1.21	1.33	1.50	1.60	1.82	2.00	2.15	2.29	2.50
	5	1.23	1.36	1.55	1.67	1.92	2.14	2.33	2.50	2.78

E-value

VanderWeele, Ding. Sensitivity Analysis in Observational Research: Introducing the E-Value. Annals of Internal Medicine (2017)
!!Also read the response letter and the authors' rejoinder!!

E-value

VanderWeele, Ding. Sensitivity Analysis in Observational Research: Introducing the E-Value. Annals of Internal Medicine (2017)
!!Also read the response letter and the authors' rejoinder!!

E-value is nothing but imposing $RR_{AU} = RR_{UY} = \rho$ as we do not really know their value anyway

E-value

VanderWeele, Ding. Sensitivity Analysis in Observational Research: Introducing the E-Value. Annals of Internal Medicine (2017)
!!Also read the response letter and the authors' rejoinder!!

E-value is nothing but imposing $RR_{AU} = RR_{UY} = \rho$ as we do not really know their value anyway

To explain away the observed RR, we need $\frac{\rho^2}{2\rho-1} \geq RR_{AY}^{obs}$, which gives us the quadratic inequality: when $RR_{AY}^{obs} \geq 1$

$$\rho^2 - 2RR_{AY}^{obs}\rho + RR_{AY}^{obs} \geq 0$$
$$\Rightarrow \rho \geq RR_{AY}^{obs} + \sqrt{RR_{AY}^{obs}(RR_{AY}^{obs} - 1)}$$

E-value

VanderWeele, Ding. Sensitivity Analysis in Observational Research: Introducing the E-Value. Annals of Internal Medicine (2017)
!!Also read the response letter and the authors' rejoinder!!

E-value is nothing but imposing $RR_{AU} = RR_{UY} = \rho$ as we do not really know their value anyway

To explain away the observed RR, we need $\frac{\rho^2}{2\rho-1} \geq RR_{AY}^{obs}$, which gives us the quadratic inequality: when $RR_{AY}^{obs} \geq 1$

$$\rho^2 - 2RR_{AY}^{obs}\rho + RR_{AY}^{obs} \geq 0$$
$$\Rightarrow \rho \geq RR_{AY}^{obs} + \sqrt{RR_{AY}^{obs}(RR_{AY}^{obs} - 1)}$$

What if $RR_{AY}^{obs} \leq 1$?

Other sensitivity analysis strategies

Most other sensitivity analysis strategies rely on further untestable assumptions to argue against the untestable ignorability assumption

Other sensitivity analysis strategies

Most other sensitivity analysis strategies rely on further untestable assumptions to argue against the untestable ignorability assumption

Under ignorability: $p(Y(1) = y|X, A = 1) \equiv p(Y(1) = y|X, A = 0)$

Other sensitivity analysis strategies

Most other sensitivity analysis strategies rely on further untestable assumptions to argue against the untestable ignorability assumption

Under ignorability: $p(Y(1) = y|X, A = 1) \equiv p(Y(1) = y|X, A = 0)$

For example, one could postulate the following "exponential tilting model"

$$\frac{p(Y(a) = y|X, A = 1 - a)}{p(Y(a) = y|X, A = a)} = \frac{\exp\{\gamma_a f_a(y)\}}{\mathbb{E}\left[\exp\{\gamma_a f_a(Y)\}|X, A = a\right]}$$

Other sensitivity analysis strategies

Most other sensitivity analysis strategies rely on further untestable assumptions to argue against the untestable ignorability assumption

Under ignorability: $p(Y(1) = y|X, A = 1) \equiv p(Y(1) = y|X, A = 0)$

For example, one could postulate the following "exponential tilting model"

$$\frac{p(Y(a) = y|X, A = 1 - a)}{p(Y(a) = y|X, A = a)} = \frac{\exp\{\gamma_a f_a(y)\}}{\mathbb{E}\left[\exp\{\gamma_a f_a(Y)\}\,|X, A = a\right]}$$

With such a model, one immediately have

$$\mathbb{E}[Y(a)] = \int_x \left\{ \begin{array}{c} \mathbb{E}[Y|X = x, A = a]\Pr(A = a|X = x) \\ + \frac{\mathbb{E}[Y \exp\{\gamma_a f_a(Y)\}|X=x, A=a]}{\mathbb{E}[\exp\{\gamma_a f_a(Y)\}|X=x, A=a]}\Pr(A = 1 - a|X = x) \end{array} \right\} p(x)\mathrm{d}x$$

and hence the ATE can be identified as

$$\tau(\gamma_0, \gamma_1; f_0, f_1) = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$$

Floor discussion

Why do you think one postulate the sensitivity analysis model as

$$\frac{p(Y(a) = y|X, A = 1 - a)}{p(Y(a) = y|X, A = a)} = \frac{\exp\{\gamma_a f_a(y)\}}{\mathbb{E}\left[\exp\{\gamma_a f_a(Y)\}|X, A = a\right]}$$

Floor discussion

Why do you think one postulate the sensitivity analysis model as

$$\frac{p(Y(a) = y|X, A = 1 - a)}{p(Y(a) = y|X, A = a)} = \frac{\exp\{\gamma_a f_a(y)\}}{\mathbb{E}\left[\exp\{\gamma_a f_a(Y)\}\,|X, A = a\right]}$$

"Counterfactuals are the ultimate unmeasured confounder"
– REF: Robins et al. Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. Stats in Med (2004).

In summary

In general there are two different strategies to perform sensitivity analysis: one relatively more straightforward, one calling for deeper theoretical analysis

- Given estimated causal effect, obtain the strength of unmeasured confounding to explain away the effect (e.g. E-value type analysis, becoming standard in medical practice and getting popular in industry)

In summary

In general there are two different strategies to perform sensitivity analysis: one relatively more straightforward, one calling for deeper theoretical analysis

- Given estimated causal effect, obtain the strength of unmeasured confounding to explain away the effect (e.g. E-value type analysis, becoming standard in medical practice and getting popular in industry)

- Postulating a model that incorporates unmeasured confounding, then estimate the causal effect using the most advanced statistical methodology and see how the result changes with the sensitivity parameter γ
  If interested in further theory, take a look at:
  REF: Scharfstein et al. Semiparametric Sensitivity Analysis: Unmeasured Confounding in Observational Studies. 2021
  and see how they developed the theoretical results for sensitivity analysis


self-esteem
grades
happiness

Mediation analysis

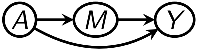
- In lecture 1, we mentioned it is often not the end of the story after we figure out if A causes Y

Mediation analysis

- In lecture 1, we mentioned it is often not the end of the story after we figure out if A causes Y
- In some applications, one might also be interested in the mechanism of A causing Y

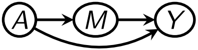
Mediation analysis

- In lecture 1, we mentioned it is often not the end of the story after we figure out if A causes Y
- In some applications, one might also be interested in the mechanism of A causing Y
- The simplest scenario is as follows



Mediation analysis

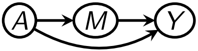
- In lecture 1, we mentioned it is often not the end of the story after we figure out if A causes Y
- In some applications, one might also be interested in the mechanism of A causing Y
- The simplest scenario is as follows



- Mechanistic question: does A directly cause Y or A causes Y through M or both?

Mediation analysis

- In lecture 1, we mentioned it is often not the end of the story after we figure out if A causes Y
- In some applications, one might also be interested in the mechanism of A causing Y
- The simplest scenario is as follows



- Mechanistic question: does A directly cause Y or A causes Y through M or both?
- Can you create a story based on this graph?

Examples

- Medicine: A chemotherapy, Y 5-year survival, M tumor resistance

Examples

- Medicine: A chemotherapy, Y 5-year survival, M tumor resistance
- Medicine: A blood pressure medication, Y heart attack, M blood pressure

Examples

- Medicine: A chemotherapy, Y 5-year survival, M tumor resistance
- Medicine: A blood pressure medication, Y heart attack, M blood pressure
- Machine learning fairness: A gender, Y college admission, M applying to department with lower admission rate

mediation analysis: motivations

Mediation questions:

- You have a theory for why the effect of a treatment/exposure on the outcome is mediated by > 1 variables
- You wish to frame your study in terms of causal questions, including hypothetical interventions

Non-mediation questions:

- Is it better to intervene on the treatment or the mediator (if you cannot do both)?
- What are the various effects of treatment?

How would you proceed?

- Floor discussion

How would you proceed?

- Floor discussion
- Jamie Robins in his 1986 paper (g-formula) has defined the concept direct and indirect effect

How would you proceed?

- Floor discussion
- Jamie Robins in his 1986 paper (g-formula) has defined the concept direct and indirect effect
- Around the same time, Judea Pearl also started to consider direct and indirect effect

Zoo of direct and indirect effects

- Treatment $A \in \{0, 1\}$, Mediator $M \in \{0, 1\}$, Outcome Y

Zoo of direct and indirect effects

- Treatment $A \in \{0, 1\}$, Mediator $M \in \{0, 1\}$, Outcome Y
- Total effect of A on Y:

$$\tau_{tot} := \mathbb{E}[Y(1) - Y(0)] \equiv \mathbb{E}[Y(1, M(1)) - Y(0, M(0))]$$

Zoo of direct and indirect effects

- Treatment $A \in \{0, 1\}$, Mediator $M \in \{0, 1\}$, Outcome Y

- Total effect of A on Y:

$$\tau_{tot} := \mathbb{E}[Y(1) - Y(0)] \equiv \mathbb{E}[Y(1, M(1)) - Y(0, M(0))]$$

- Controlled direct effect (CDE) of A on Y:

$$\tau_{cde}(0) := \mathbb{E}[Y(a = 1, m = 0) - Y(a = 0, m = 0)]$$
$$\tau_{cde}(1) := \mathbb{E}[Y(a = 1, m = 1) - Y(a = 0, m = 1)]$$

Zoo of direct and indirect effects

- Treatment $A \in \{0, 1\}$, Mediator $M \in \{0, 1\}$, Outcome Y
- Total effect of A on Y:

$$\tau_{tot} := \mathbb{E}[Y(1) - Y(0)] \equiv \mathbb{E}[Y(1, M(1)) - Y(0, M(0))]$$

- Controlled direct effect (CDE) of A on Y:

$$\tau_{cde}(0) := \mathbb{E}[Y(a = 1, m = 0) - Y(a = 0, m = 0)]$$
$$\tau_{cde}(1) := \mathbb{E}[Y(a = 1, m = 1) - Y(a = 0, m = 1)]$$

- Controlled direct effect (CDE) of M on Y:

$$\tau_{cie}(0) := \mathbb{E}[Y(a = 0, m = 1) - Y(a = 0, m = 0)]$$
$$\tau_{cie}(1) := \mathbb{E}[Y(a = 1, m = 1) - Y(a = 1, m = 0)]$$

Warning about CDE

- CDE is NOT for mediation analysis, as we have discussed

Warning about CDE

- CDE is NOT for mediation analysis, as we have discussed
- CDE is designed to answer questions like "the effect of intervening both A and M"

Zoo of direct and indirect effects

- Treatment $A \in \{0, 1\}$, Mediator $M \in \{0, 1\}$, Outcome Y

- Total effect of A on Y:

$$\tau_{tot} := \mathbb{E}[Y(1) - Y(0)] \equiv \mathbb{E}[Y(1, M(1)) - Y(0, M(0))]$$

Zoo of direct and indirect effects

- Treatment $A \in \{0,1\}$, Mediator $M \in \{0,1\}$, Outcome Y

- Total effect of A on Y:

$$\tau_{tot} := \mathbb{E}[Y(1) - Y(0)] \equiv \mathbb{E}[Y(1, M(1)) - Y(0, M(0))]$$

- Natural direct effect (NDE) of A on Y: the part of τ that does not go through M

$$\tau_{nde}(0) := \mathbb{E}[Y(1, M(0)) - Y(0, M(0))]$$
$$\tau_{nde}(1) := \mathbb{E}[Y(1, M(1)) - Y(0, M(1))]$$

Zoo of direct and indirect effects

- Treatment $A \in \{0,1\}$, Mediator $M \in \{0,1\}$, Outcome Y

- Total effect of A on Y:

$$\tau_{tot} := \mathbb{E}[Y(1) - Y(0)] \equiv \mathbb{E}[Y(1, M(1)) - Y(0, M(0))]$$

- Natural direct effect (NDE) of A on Y: the part of τ that does not go through M

$$\tau_{nde}(0) := \mathbb{E}[Y(1, M(0)) - Y(0, M(0))]$$
$$\tau_{nde}(1) := \mathbb{E}[Y(1, M(1)) - Y(0, M(1))]$$

- Natural indirect effect (NIE) of A on Y: the part of τ that does not go through M

$$\tau_{nie}(0) := \mathbb{E}[Y(0, M(1)) - Y(0, M(0))]$$
$$\tau_{nie}(1) := \mathbb{E}[Y(1, M(1)) - Y(1, M(0))]$$

Trivial decompositions

$$\tau_{tot} = \mathbb{E}[Y(1, M(1)) - Y(0, M(0))] = \tau_{nde}(1) + \tau_{nie}(0)$$
$$\tau_{tot} = \mathbb{E}[Y(1, M(1)) - Y(0, M(0))] = \tau_{nde}(0) + \tau_{nie}(1)$$

Not so trivial 4-way decomposition: connections with causal interactions (study on your own)

A less trivial but useful decomposition of the total effect τ_{tot}
REF: VanderWeele, A Unification of Mediation and Interaction: A 4-Way Decomposition. Epidemiology (2014)

Not so trivial 4-way decomposition: connections with causal interactions (study on your own)

A less trivial but useful decomposition of the total effect τ_{tot}

REF: VanderWeele, A Unification of Mediation and Interaction: A 4-Way Decomposition. Epidemiology (2014)

For simplicity, assume all variables are $\{0, 1\}$-valued

$$
\begin{aligned}
\tau_{tot} &= \mathbb{E}[Y(1, M(1)) - Y(0, M(1))] + \tau_{nie}(0) \\
&= \mathbb{E}[Y(1, M(0)) - Y(0, M(0))] + \mathbb{E}[Y(1, M(1)) - Y(0, M(1)) - Y(1, M(0)) + Y(0, M(0))] + \tau_{nie}(0) \\
&= \mathbb{E}[Y(1, 0) - Y(0, 0)] + \mathbb{E}[Y(1, M(0)) - Y(0, M(0)) - Y(1, 0) + Y(0, 0)] \\
&\quad + \mathbb{E}[Y(1, M(1)) - Y(0, M(1)) - Y(1, M(0)) + Y(0, M(0))] + \tau_{nie}(0) \\
&= \tau_{cde}(0) + \mathbb{E}[Y(1, M(0)) - Y(0, M(0)) - Y(1, 0) + Y(0, 0)] \\
&\quad + \mathbb{E}[Y(1, M(1)) - Y(0, M(1)) - Y(1, M(0)) + Y(0, M(0))] + \tau_{nie}(0).
\end{aligned}
$$

$$
\begin{aligned}
& \mathbb{E}[Y(1, M(0)) - Y(0, M(0)) - Y(1,0) + Y(0,0)] \\
= \; & \mathbb{E}\left[\sum_{m=0,1} \{Y(1,m) - Y(0,m)\} \mathbb{1}\{M(0) = m\} - Y(1,0) + Y(0,0)\right] \\
= \; & \mathbb{E}\left[\left\{\sum_{m=0,1} Y(1,m) - Y(0,m) - Y(1,0) + Y(0,0)\right\} \mathbb{1}\{M(0) = 1\}\right] \\
= \; & \underbrace{\mathbb{E}\left[\underbrace{\{Y(1,1) - Y(0,1) - Y(1,0) + Y(0,0)\}}_{\text{interaction between } a \text{ and } m} \mathbb{1}\{M(0) = 1\}\right]}_{\text{reference interaction}}
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{E}[Y(1, M(1)) - Y(0, M(1)) - Y(1, M(0)) + Y(0, M(0))] \\
&= \mathbb{E}\left[\sum_{m=0,1} \{Y(1, m) - Y(0, m)\}\{\mathbb{1}\{M(1) = m\} - \mathbb{1}\{M(0) = m\}\}\right] \\
&= \underbrace{\mathbb{E}\left[\underbrace{\{Y(1,1) - Y(0,1) - Y(1,0) + Y(0,0)\}}_{\text{interaction between } a \text{ and } m}\{\mathbb{1}\{M(1) = 1\} - \mathbb{1}\{M(0) = 1\}\}\right]}_{\text{mediated interaction}}
\end{aligned}
$$

To summarize:

$$
\tau_{tot} = \tau_{cde}(0) + \text{reference interaction} + \text{mediated interaction} + \tau_{nie}(0)
$$

How to identify natural direct/indirect effect?

Consider NDE: $\tau_{nde}(0) := \mathbb{E}[Y(1, M(0)) - Y(0, M(0))]$

How to identify natural direct/indirect effect?

Consider NDE: $\tau_{nde}(0) := \mathbb{E}[Y(1, M(0)) - Y(0, M(0))]$

$\mathbb{E}[Y(0, M(0))] \equiv \mathbb{E}[Y(0)]$ by definition, easy to handle

How to identify natural direct/indirect effect?

Consider NDE: $\tau_{nde}(0) := \mathbb{E}[Y(1, M(0)) - Y(0, M(0))]$

$\mathbb{E}[Y(0, M(0))] \equiv \mathbb{E}[Y(0)]$ by definition, easy to handle

But $\mathbb{E}[Y(1, M(0))]$ seems tricky to handle

How to identify natural direct/indirect effect?

Consider NDE: $\tau_{nde}(0) := \mathbb{E}[Y(1, M(0)) - Y(0, M(0))]$

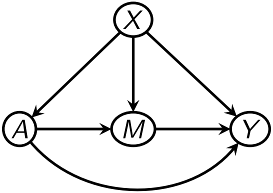
$\mathbb{E}[Y(0, M(0))] \equiv \mathbb{E}[Y(0)]$ by definition, easy to handle

But $\mathbb{E}[Y(1, M(0))]$ seems tricky to handle

Discuss why

How to identify $\mathbb{E}[Y(1, M(0))]$?

Mediation DAG with confounders



REF: Pearl, Direct and indirect effects, UAI (2001) consider the following ignorability conditions:

1. $Y(a,m) \perp\!\!\!\perp A|X$: no unmeasured treatment-outcome confounder
2. $Y(a,m) \perp\!\!\!\perp M|\{X,A\}$: no unmeasured mediator-outcome confounder
3. $M(a) \perp\!\!\!\perp A|X$: no unmeasured treatment-mediator confounder
4. $Y(a,m) \perp\!\!\!\perp M(a')|X$: will come back later

How to identify $\mathbb{E}[Y(1, M(0))]$?

REF: Pearl, Direct and indirect effects, UAI (2001)

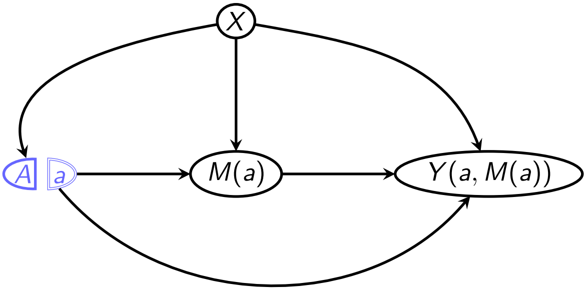
How to identify $\mathbb{E}[Y(1, M(0))]$?

REF: Pearl, Direct and indirect effects, UAI (2001)

$$
\begin{aligned}
&\mathbb{E}[Y(1, M(0))] \\
&= \mathbb{E}_X\left[\mathbb{E}_{Y(1,M(0))}[Y(1, M(0))|X, M(0)]\right] \\
&= \int_x \int_m \left\{ \int_y y f(Y(1, M(0)) = y|X = x, M(0) = m)\mathrm{d}y \right\} f(X = x, M(0) = m)\mathrm{d}m\mathrm{d}x \\
&= \int_x \int_m \left\{ \int_y y f(Y(1, m) = y|X = x, M(0) = m)\mathrm{d}y \right\} f(M(0) = m|X = x) f(X = x)\mathrm{d}m\mathrm{d}x \\
&\overset{4}{=} \int_x \int_m \left\{ \int_y y f(Y(1, m) = y|X = x)\mathrm{d}y \right\} f(M(0) = m|X = x) f(X = x)\mathrm{d}m\mathrm{d}x \\
&\overset{3}{=} \int_x \int_m \left\{ \int_y y f(Y(1, m) = y|X = x)\mathrm{d}y \right\} f(M(0) = m|X = x, A = 0) f(X = x)\mathrm{d}m\mathrm{d}x \\
&\overset{1}{=} \int_x \int_m \left\{ \int_y y f(Y(1, m) = y|X = x, A = 1)\mathrm{d}y \right\} f(M = m|X = x, A = 0) f(X = x)\mathrm{d}m\mathrm{d}x \\
&\overset{2}{=} \int_x \int_m \left\{ \int_y y f(Y(1, m) = y|X = x, A = 1, M = m)\mathrm{d}y \right\} f(M = m|X = x, A = 0) f(X = x)\mathrm{d}m\mathrm{d}x \\
&= \int_x \int_m \left\{ \int_y y f(Y = y|X = x, A = 1, M = m)\mathrm{d}y \right\} f(M = m|X = x, A = 0) f(X = x)\mathrm{d}m\mathrm{d}x \\
&= \int_x \int_m \mathbb{E}[Y|X = x, A = 1, M = m] f(M = m|X = x, A = 0) f(X = x)\mathrm{d}m\mathrm{d}x
\end{aligned}
$$

SWIG for mediation DAG

SWIG for mediation DAG



Issue: SWIG does not allow $Y(a, M(a'))$ on the graph for $a \neq a'$ because it is a cross-world counterfactual

Philosophical forkpath

Based on the causal model defined via SWIG, impossible to identify NDE or NIE because they are cross-world counterfactuals

Philosophical forkpath

Based on the causal model defined via SWIG, impossible to identify NDE or NIE because they are cross-world counterfactuals

SWIG-based causal model was defined in Robins 1986, called "Finest Fully Randomized Causally Interpretable Structured Tree Graphs (FFRCISTG)"

Philosophical forkpath

Based on the causal model defined via SWIG, impossible to identify NDE or NIE because they are cross-world counterfactuals

SWIG-based causal model was defined in Robins 1986, called "Finest Fully Randomized Causally Interpretable Structured Tree Graphs (FFRCISTG)"

But Pearl's theory says NDE and NIE are identified if one is willing to assume cross-world independence assumption $Y(a, m) \perp\!\!\!\perp M(a')|X$

Philosophical forkpath

Based on the causal model defined via SWIG, impossible to identify NDE or NIE because they are cross-world counterfactuals

SWIG-based causal model was defined in Robins 1986, called "Finest Fully Randomized Causally Interpretable Structured Tree Graphs (FFRCISTG)"

But Pearl's theory says NDE and NIE are identified if one is willing to assume cross-world independence assumption $Y(a, m) \perp\!\!\!\perp M(a')|X$

Pearl's causal model is called "Non-Parametric Structural Equation Models with Independent Errors (NPSEM-IE)"

Philosophical forkpath

Based on the causal model defined via SWIG, impossible to identify NDE or NIE because they are cross-world counterfactuals

SWIG-based causal model was defined in Robins 1986, called "Finest Fully Randomized Causally Interpretable Structured Tree Graphs (FFRCISTG)"

But Pearl's theory says NDE and NIE are identified if one is willing to assume cross-world independence assumption $Y(a, m) \perp\!\!\!\perp M(a')|X$

Pearl's causal model is called "Non-Parametric Structural Equation Models with Independent Errors (NPSEM-IE)"

Therefore NPSEM $\subset$ FFRCISTG

Abstract difference between FFRCISTG vs. NPSEM-IE

Let's look at the key difference between FFRCISTG vs. NPSEM-IE

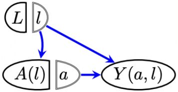
Abstract difference between FFRCISTG vs. NPSEM-IE

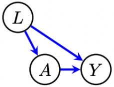
Let's look at the key difference between FFRCISTG vs. NPSEM-IE

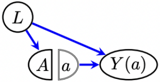
FFRCISTG/SWIG

	One-Step Ahead Counterfactuals		Passive Observation		Experimental Intervention on A

Graph:







	One-Step Ahead Counterfactuals		Passive Observation		Experimental Intervention on A
Variables:	L $A(l)$ $Y(a,l)$	$\Rightarrow$	L $A \equiv A(L)$ $Y \equiv Y(A,L)$	$\Rightarrow$	L $A \equiv A(L)$ $Y(a) \equiv Y(a,L)$
Interpretation:	Counterfactuals when A and L are intervened on;		Observed System;		Prior to intervention: L and A; after intervention: $Y(a)$.
Meaning of A:	(A does not appear)		Natural value of A;		Natural value of A (observed prior to intervention on A).

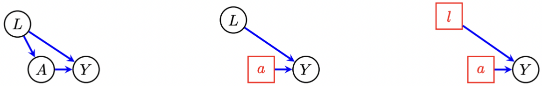
Single World No Confounding Assumption: for each pair a, l: $L \perp\!\!\!\perp A(l) \perp\!\!\!\perp Y(a,l)$

Abstract difference between FFRCISTG vs. NPSEM-IE

Let's look at the key difference between FFRCISTG vs. NPSEM-IE

NPSEM-IE

	Passive Observation		Experimental Intervention on A		Experimental Intervention on A and L
Graph:					



	Passive Observation		Experimental Intervention on A		Experimental Intervention on A and L
Variables:	$L = f_L(\varepsilon_L)$ $A = f_A(L, \varepsilon_A)$ $Y = f_Y(A, L, \varepsilon_Y)$	$\Rightarrow$	$L = f_L(\varepsilon_L)$ $A = a$ $Y = f_Y(A, L, \varepsilon_Y)$	$\Rightarrow$	$L = l$ $A = a$ $Y = f_Y(A, L, \varepsilon_Y)$
Interpretation:	Observed system;		Variables in system in which A is set to a;		L and A after each is intervened on; Y after both interventions.
Meaning of A:	Natural value of A;		Value of A after intervention on A;		Value of A after intervention on A (and L).

Independent Errors No Confounding Assumption: $\varepsilon_L \perp\!\!\!\perp \varepsilon_A \perp\!\!\!\perp \varepsilon_Y$

Relationships to Counterfactuals:

Error terms:	$\varepsilon_L = L$;	$\varepsilon_A = \{A(l) \text{ for all } l\}$;	$\varepsilon_Y = \{Y(a, l) \text{ for all } l, a\}$
Structural equations:	$L = f_L(\varepsilon_L)$	$A(l) = f_A(l, \varepsilon_A)$	$Y(a, l) = f_Y(a, l, \varepsilon_Y)$

Difference presented in math

- Pearl's NPSEM-IE: for all variables $V_1, \cdots, V_N$ on a causal DAG $\mathcal{G}$:

$$V_1 = f_1(\text{pa}_1; \varepsilon_1)$$
$$\vdots$$
$$V_N = f_N(\text{pa}_N; \varepsilon_N)$$

s.t. $\{V_1\} \perp\!\!\!\perp \{V_2(x_{\text{pa}_2}); \forall x_{\text{pa}_2}\} \perp\!\!\!\perp \cdots \perp\!\!\!\perp \{V_N(x_{\text{pa}_N}); \forall x_{\text{pa}_N}\}$

Difference presented in math

· Pearl's NPSEM-IE: for all variables $V_1, \cdots, V_N$ on a causal DAG $\mathcal{G}$:

$$\begin{aligned} V_1 &= f_1(\text{pa}_1; \varepsilon_1) \\ &\vdots \\ V_N &= f_N(\text{pa}_N; \varepsilon_N) \end{aligned}$$

s.t. $\{V_1\} \perp\!\!\!\perp \{V_2(x_{\text{pa}_2}); \forall x_{\text{pa}_2}\} \perp\!\!\!\perp \cdots \perp\!\!\!\perp \{V_N(x_{\text{pa}_N}); \forall x_{\text{pa}_N}\}$

· Robins' FFRCISTG/SWIG: for all variables $V_1, \cdots, V_N$ on a causal DAG $\mathcal{G}$:
for each x_V: $V_1 \perp\!\!\!\perp V_2(x_{\text{pa}_2}) \perp\!\!\!\perp \cdots \perp\!\!\!\perp V_N(x_{\text{pa}_N})$

What to do without cross-world independence assumption?

When one cannot make further progress with the current definition, then change the definition

What to do without cross-world independence assumption?

When one cannot make further progress with the current definition, then change the definition

Here comes "Interventional Direct/Indirect Effect (IDE/IIE)":
REF: VanderWeele, Vansteelandt, Robins. Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. Epidemiology (2014).

Definition of IDE/IIE

How do we change the definition(s)? Let's look at the derivation of $\mathbb{E}[Y(1, M(0))]$, where we use the cross-world independence condition (4)

$$
\begin{aligned}
&\mathbb{E}[Y(1, M(0))] \\
&= \mathbb{E}_X\left[\mathbb{E}_{Y(1,M(0))}[Y(1, M(0))|X, M(0)]\right] \\
&= \int_x \int_m \left\{ \int_y y f(Y(1, M(0)) = y|X = x, M(0) = m)\mathrm{d}y \right\} f(X = x, M(0) = m)\mathrm{d}m\mathrm{d}x \\
&= \int_x \int_m \left\{ \int_y y f(Y(1, m) = y|X = x, M(0) = m)\mathrm{d}y \right\} f(M(0) = m|X = x) f(X = x)\mathrm{d}m\mathrm{d}x \\
&\overset{4}{=} \int_x \int_m \left\{ \int_y y f(Y(1, m) = y|X = x)\mathrm{d}y \right\} f(M(0) = m|X = x) f(X = x)\mathrm{d}m\mathrm{d}x \\
&\overset{3}{=} \int_x \int_m \left\{ \int_y y f(Y(1, m) = y|X = x)\mathrm{d}y \right\} f(M(0) = m|X = x, A = 0) f(X = x)\mathrm{d}m\mathrm{d}x \\
&\overset{1}{=} \int_x \int_m \left\{ \int_y y f(Y(1, m) = y|X = x, A = 1)\mathrm{d}y \right\} f(M = m|X = x, A = 0) f(X = x)\mathrm{d}m\mathrm{d}x \\
&\overset{2}{=} \int_x \int_m \left\{ \int_y y f(Y(1, m) = y|X = x, A = 1, M = m)\mathrm{d}y \right\} f(M = m|X = x, A = 0) f(X = x)\mathrm{d}m\mathrm{d}x \\
&= \int_x \int_m \left\{ \int_y y f(Y = y|X = x, A = 1, M = m)\mathrm{d}y \right\} f(M = m|X = x, A = 0) f(X = x)\mathrm{d}m\mathrm{d}x \\
&= \int_x \int_m \mathbb{E}[Y|X = x, A = 1, M = m] f(M = m|X = x, A = 0) f(X = x)\mathrm{d}m\mathrm{d}x
\end{aligned}
$$

Definition of IDE/IIE

Cross-world assumption is used, from $\mathbb{E}[Y(1, M(0))]$ to

$$\int_x \int_m \left\{ \int_y y f(Y(1,m) = y | X = x) \mathrm{d}y \right\} f(M(0) = m | X = x) f(X = x) \mathrm{d}m \mathrm{d}x$$

Definition of IDE/IIE

Cross-world assumption is used, from $\mathbb{E}[Y(1, M(0))]$ to

$$\int_x \int_m \left\{ \int_y y f(Y(1,m) = y | X = x) \mathrm{d}y \right\} f(M(0) = m | X = x) f(X = x) \mathrm{d}m \mathrm{d}x$$

Let's just define the latter as the parameter that we are interested in and figure out its interpretation

Definition of IDE/IIE

Cross-world assumption is used, from $\mathbb{E}[Y(1, M(0))]$ to

$$\int_x \int_m \left\{ \int_y y f(Y(1,m) = y | X = x) \mathrm{d}y \right\} f(M(0) = m | X = x) f(X = x) \mathrm{d}m \mathrm{d}x$$

Let's just define the latter as the parameter that we are interested in and figure out its interpretation

Can you give a meaning to this quantity?

Definition of IDE/IIE

Cross-world assumption is used, from $\mathbb{E}[Y(1, M(0))]$ to

$$\int_x \int_m \left\{ \int_y y f(Y(1, m) = y | X = x) \mathrm{d}y \right\} f(M(0) = m | X = x) f(X = x) \mathrm{d}m \mathrm{d}x$$

Let's just define the latter as the parameter that we are interested in and figure out its interpretation

Can you give a meaning to this quantity?

$$\mathbb{E}\left[Y(1, \widetilde{M}_{0|X}) \right]$$

where $\widetilde{M}_{0|X}$ is a random draw from the probability distribution of $M(0)|X$ independent of everything else

Definition of IDE/IIE

So we eventually define interventional direct and indirect effects as:

$$\tau_{ide}(0) \coloneqq \mathbb{E}\left[Y(1, \widetilde{M}_{0|X})\right] - \mathbb{E}\left[Y(0, \widetilde{M}_{0|X})\right]$$
$$\tau_{ide}(1) \coloneqq \mathbb{E}\left[Y(1, \widetilde{M}_{1|X})\right] - \mathbb{E}\left[Y(0, \widetilde{M}_{1|X})\right]$$
$$\tau_{iie}(0) \coloneqq \mathbb{E}\left[Y(0, \widetilde{M}_{1|X})\right] - \mathbb{E}\left[Y(0, \widetilde{M}_{0|X})\right]$$
$$\tau_{iie}(1) \coloneqq \mathbb{E}\left[Y(1, \widetilde{M}_{1|X})\right] - \mathbb{E}\left[Y(1, \widetilde{M}_{0|X})\right]$$

Definition of IDE/IIE

So we eventually define interventional direct and indirect effects as:

$$\tau_{ide}(0) \coloneqq \mathbb{E}\left[Y(1, \widetilde{M}_{0|X})\right] - \mathbb{E}\left[Y(0, \widetilde{M}_{0|X})\right]$$
$$\tau_{ide}(1) \coloneqq \mathbb{E}\left[Y(1, \widetilde{M}_{1|X})\right] - \mathbb{E}\left[Y(0, \widetilde{M}_{1|X})\right]$$
$$\tau_{iie}(0) \coloneqq \mathbb{E}\left[Y(0, \widetilde{M}_{1|X})\right] - \mathbb{E}\left[Y(0, \widetilde{M}_{0|X})\right]$$
$$\tau_{iie}(1) \coloneqq \mathbb{E}\left[Y(1, \widetilde{M}_{1|X})\right] - \mathbb{E}\left[Y(1, \widetilde{M}_{0|X})\right]$$

What do you think of this level of rigor when it comes to impose scientific meaning to a parameter?

Other related definitions of direct/indirect effects

Other than IDE/IIE, people have developed other definitions of direct/indirect effects, but they all have similar spirit to IDE/IIE

Other related definitions of direct/indirect effects

Other than IDE/IIE, people have developed other definitions of direct/indirect effects, but they all have similar spirit to IDE/IIE

Organic direct/indirect effect
REF: Lok, Bosch. Causal organic indirect and direct effects: closer to Baron and Kenny. Epidemiology (2021).

Other related definitions of direct/indirect effects

Other than IDE/IIE, people have developed other definitions of direct/indirect effects, but they all have similar spirit to IDE/IIE

Organic direct/indirect effect
REF: Lok, Bosch. Causal organic indirect and direct effects: closer to Baron and Kenny. Epidemiology (2021).

Mediators are often difficult to even imagine an intervention (e.g. BMI) so $Y(a, m)$ is ill-defined
Organic DE/IE: hypothesize an organic intervention I on mediator that does not have a direct effect on Y but $M(0, I = 1)|X \sim M(1)|X$. Then, for $a = 0, 1$

$$\tau_{ode}(a) := \mathbb{E}[Y(1, I = a) - Y(0, I = a)]$$
$$\tau_{oie}(a) := \mathbb{E}[Y(a, I = 1) - Y(a, I = 0)]$$

Interpretation of ODE/OIE

Example? A blood pressure drug, M blood pressure, Y heart attack
What is I?

Interpretation of ODE/OIE

Example? A blood pressure drug, M blood pressure, Y heart attack
What is I?

I: Reduction in salt intake in diet; salt should only cause Y through M, so no direct effect on Y

Others: learn on your own

Separable effects:
REF: Robins, Richardson, Shpitser. An Interventionist Approach to Mediation Analysis.
REF: Robins, Richardson. Alternative Graphical Causal Models and the Identification of Direct Effects.
REF: Didelez. Defining causal mediation with a longitudinal mediator and a survival outcome. Life Time Data Analysis (2019).

Others: learn on your own

Separable effects:
REF: Robins, Richardson, Shpitser. An Interventionist Approach to Mediation Analysis.
REF: Robins, Richardson. Alternative Graphical Causal Models and the Identification of Direct Effects.
REF: Didelez. Defining causal mediation with a longitudinal mediator and a survival outcome. Life Time Data Analysis (2019).

Population direct/indirect effect:
REF: Fulcher, Shpitser, Marealle, Tchetgen Tchetgen. Robust inference on population indirect causal effects: The generalized front-door criterion.

Path-specific effects

Mediation analysis is about delineating the mechanism of a causal effect

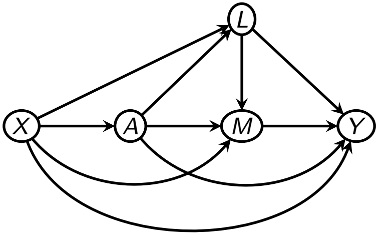
Path-specific effects

Mediation analysis is about delineating the mechanism of a causal effect
It can be further generalized: path-specific effects (identified by **edge g-formula**)

REF: Avin, Shpitser, Pearl. Identifiability of path-specific effects. (2005)
REF: Shpitser, Pearl. Complete identification methods for the causal hierarchy. JMLR (2008)
REF: Shpitser, Tchetgen Tchetgen. Causal inference with a graphical hierarchy of interventions. AoS (2016).

Path-specific effects

Mediation analysis is about delineating the mechanism of a causal effect
It can be further generalized: path-specific effects (identified by **edge g-formula**)
REF: Avin, Shpitser, Pearl. Identifiability of path-specific effects. (2005)
REF: Shpitser, Pearl. Complete identification methods for the causal hierarchy. JMLR (2008)
REF: Shpitser, Tchetgen Tchetgen. Causal inference with a graphical hierarchy of interventions. AoS (2016).



Total effect τ_{tot} of A on Y can be decomposed into

$$\tau_{tot} = \tau_{A\to Y} + \tau_{A\to M\to Y} + \tau_{A\to L\to Y} + \tau_{A\to L\to M\to Y}$$

Next time

Time-varying causal inference, (Optimal) dynamic treatment regimes, dynamic regime SWIG

Next time

Time-varying causal inference, (Optimal) dynamic treatment regimes, dynamic regime SWIG

Here dynamic treatment does not necessarily mean treatment at multiple points; it is a term opposite to "static treatment" such as $A = 1$; An example of dynamic treatment is
$A = \mathbb{1}\{\text{blood pressure } > 200, \text{age } < 60\}$

Next time

Time-varying causal inference, (Optimal) dynamic treatment regimes, dynamic regime SWIG

Here dynamic treatment does not necessarily mean treatment at multiple points; it is a term opposite to "static treatment" such as $A = 1$; An example of dynamic treatment is
$A = \mathbb{1}\{\text{blood pressure } > 200, \text{age } < 60\}$

Both of the above treatment examples are loosely called "hard intervention" according to Pearl and many other people

Next time

Time-varying causal inference, (Optimal) dynamic treatment regimes, dynamic regime SWIG

Here dynamic treatment does not necessarily mean treatment at multiple points; it is a term opposite to "static treatment" such as $A = 1$; An example of dynamic treatment is
$A = \mathbb{1}\{\text{blood pressure } > 200, \text{age } < 60\}$

Both of the above treatment examples are loosely called "hard intervention" according to Pearl and many other people

"Soft intervention" loosely means stochastic treatment regimes, e.g.
$A \sim \mathrm{Bernoulli}(\text{softmax (blood pressure)})$

Any Questions?