

A brief introduction to causal inference

Qingsong Wang

Oct 28, 2022

Outline

- 1 What is causal inference?
- 2 Potential outcome framework
- 3 Causal graphs
- 4 Causal inference and machine learning

Why causal inference?

Why association/correlation alone is not good enough?

Example (Simpson's paradox)

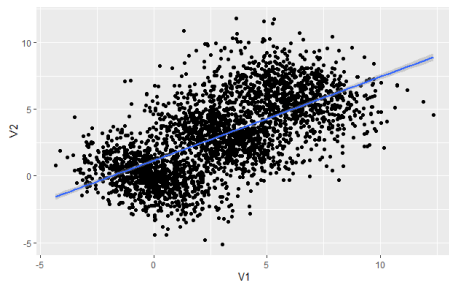
	Full Population, N = 52			Men (M), N = 20			Women (\neg M), N = 32		
	Success (S)	Failure (\neg S)	Success Rate	Success	Failure	Success Rate	Success	Failure	Success Rate
Treatment (T)	20	20	50%	8	5	$\approx 61\%$	12	15	$\approx 44\%$
Control (\neg T)	6	6	50%	4	3	$\approx 57\%$	2	3	$\approx 40\%$

TABLE 1: Simpson's Paradox: the type of association at the population level (positive, negative, independent) changes at the level of subpopulations. Numbers taken from Simpson's original example (1951).

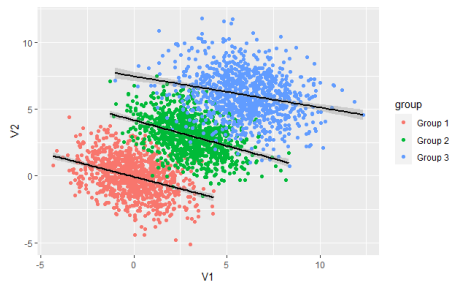
Question: Is the treatment effective?

Simpson's paradox in regression

Hypothetically, consider the following data where the x axis is the placement of Ads on the Google search page, y -axis is the number of clicks, and the data is grouped by different device types.

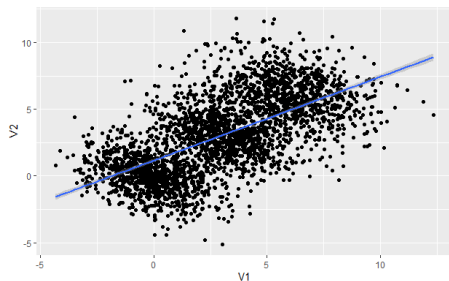


(a) **Positive** correlation among full population

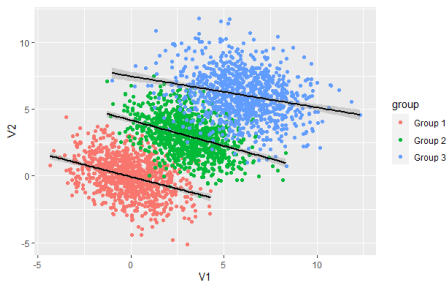


(b) **Negative** correlation within each subgroup

Simpson's paradox in regression



(a) **Positive** correlation among full population



(b) **Negative** correlation within each subgroup

What are the possible explanations for Simpson's paradox?

- 1 Not enough data?
- 2 Imbalanced data distribution? (**Randomization**)
- 3 Some unknown effects? (**Confounders**: a third variable that influences both the exposure and outcome)

General picture of causal inference

Causal inference is about¹

- 1 Build a framework and define causal effects under general scenarios
- 2 Specify assumptions under which one can declare/identify causation from association
- 3 Assess the sensitivity to the causal assumptions and find ways to mitigate

In this brief introduction, we will see some aspects of items 1 and 2.

¹Fan Li. *STA 640: Causal Inference*. 2022. URL:

<https://www2.stat.duke.edu/~fl35/CausalInferenceClass.html>

Languages for causality

We will talk about two languages for causal inference with an emphasis on the first one, and give motivations for the second one.

- 1 **Potential outcome:** Easy to incorporate additional assumptions; Convenient statistical inference; Not as convenient if the system is complex.
- 2 **Causal graphs:** Easy to visualize the causal assumptions; Difficult for statistical inference because the model is nonparametric. ²
- 3 **Structural equations:** Bridge between graphs and potential outcomes

²Qingyuan Zhao. *Lecture Notes on Causal Inference*. May 2022. URL:

http://www.statslab.cam.ac.uk/~qz280/teaching/causal-2022/notes_2021.pdf

Some basic notions for studying treatment effects

Treatment assignment mechanism

- Suppose there are n units. $(1, 2, \dots, n.)$
- For the i -th unit, some covariates X_i is observed prior to treatment assignments. (Let $A_{[n]} := (A_1, A_2, \dots, A_n)$)
- Binary treatments: $A_i \in \{0, 1\}$ ($A_i = 1$ is treated, $A_i = 0$ is control).
- The assignment mechanism is the conditional distribution:

$$P(A_{[n]} = a_{[n]} \mid X_{[n]} = x_{[n]}) = \pi(a_{[n]} \mid x_{[n]}),$$

where the function $\pi(a_{[n]} \mid x_{[n]})$ is prespecified.

Example: Bernoulli trial with covariates

$\pi(a_{[n]} \mid x_{[n]}) = \prod_{i=1}^n \mu(x_i)^{a_i} (1 - \mu(x_i))^{1-a_i}$ where $0 < \mu(x) < 1$ is a function in covariate x .

Some basic notions for studying treatment effects

After treatment assignment, an outcome variable Y_i is measured for each unit i .

Usual statistical estimation/ machine learning prediction

- 1 Compare the conditional expectations $E[Y | A = 0]$ with $E[Y | A = 1]$.
- 2 Compare the conditional distributions $P(Y \leq y | A = 0)$ with $P(Y \leq y | A = 1)$.
- 3 Further condition on X and compare $E[Y | A = 0, X = x]$ with $E[Y | A = 1, X = x]$ or the conditional distributions.

Difficult to obtain reasonable causal inference especially when the treatment assignment is not i.i.d.

How to correctly estimate the treatment effect?

Ideal scenario

For the same unit (patient), ideally, we want to measure the treatment effect as the difference in his responses between receiving and not receiving the treatment.

Remark

- 1 Causal effects are defined by counterfactual contrasts
- 2 In practice, the individual treatment effect is hardly accessible

How to correctly estimate the treatment effect?

Ideal scenario

For the same unit (patient), ideally, we want to measure the treatment effect as the difference in his responses between receiving and not receiving the treatment.

Remark

- 1 Causal effects are defined by counterfactual contrasts
- 2 In practice, the individual treatment effect is hardly accessible

Next, we will introduce the Neyman-Rubin's potential outcome causal model.

Potential outcomes (or counterfactuals)

Let $a_{[n]}$ be a treatment assignment (for all units). The *potential outcome* of unit i under $a_{[n]}$ is given as $Y_i(a_{[n]})$.

The potential outcome is linked with the *observed outcome* Y_i via the following **assumptions**.

Consistency assumption

Let $A_{[n]}$ be the observed treatment assignment. Then $Y_i = Y_i(A_{[n]})$ for all $i \in [n]$

No inference assumption

$Y_i(a_{[n]}) = Y_i(a_i)$ for all $i \in [n]$ and $a_{[n]} \in \{0, 1\}^n$.

The above two assumptions together are commonly denoted as *stable unit treatment value assumption* (SUTVA).

Stable unit treatment value assumption (SUTVA)

Under the *stable unit treatment value assumption* (SUTVA), each unit only has two potential outcomes $Y_i(1)$, $Y_i(0)$ and

- If $A_i = 1$ then $Y_i = Y_i(1)$
- If $A_i = 0$ then $Y_i = Y_i(0)$
- or equivalently: $Y_i = A_i Y_i(1) + (1 - A_i) Y_i(0)$.

Remark

The SUTVA is not always satisfied in practice, for example, one's exposure risk to Covid is influenced by other people.

Fundamental problem of causal inference

Some more notations

- ① individual treatment effect: $Y_i(1) - Y_i(0)$
- ② population average treatment effect (PATE): $E_{Y_i}(Y_i(1) - Y_i(0))$

Fundamental problem of causal inference

We most of the time can observe at most one of the potential outcomes for each unit, the other(s) are missing/counterfactual

Fundamental problem of causal inference

Some more notations

- 1 individual treatment effect: $Y_i(1) - Y_i(0)$
- 2 population average treatment effect (PATE): $E_{Y_i}(Y_i(1) - Y_i(0))$

Fundamental problem of causal inference

We most of the time can observe at most one of the potential outcomes for each unit, the other(s) are missing/counterfactual

Under the potential outcome framework, causal inference is a *missing data problem*.

Question

Can we provide a reasonable estimate of PATE even though not all potential outcomes are observed?

Fundamental problem of causal inference

Some more notations

- 1 individual treatment effect: $Y_i(1) - Y_i(0)$
- 2 population average treatment effect (PATE): $E_{Y_i}(Y_i(1) - Y_i(0))$

Fundamental problem of causal inference

We most of the time can observe at most one of the potential outcomes for each unit, the other(s) are missing/counterfactual

Under the potential outcome framework, causal inference is a *missing data problem*.

Question

Can we provide a reasonable estimate of PATE even though not all potential outcomes are observed?

One must make more assumptions.

Fundamental problem of causal inference

Question

Can we provide a reasonable estimate of PATE even though not all potential outcomes are observed?

One approach: Close substitutes

- 1 the same unit took two different treatments at different times (and no inference involved)
- 2 finds two identical units that take different treatments.

Fundamental problem of causal inference

Question

Can we provide a reasonable estimate of PATE even though not all potential outcomes are observed?

One approach: Close substitutes

- 1 the same unit took two different treatments at different times (and no inference involved)
- 2 finds two identical units that take different treatments.

The existence of close substitutes requires strong assumptions. In the next few slides, we will see that PATE can be estimated if the treatment is randomized.

The role of randomization

Randomization (ignorability) assumption

$$A_{[n]} \perp\!\!\!\perp Y_{[n]}(a_{[n]}) \mid X_{[n]} \text{ for } a_{[n]} \in \{0, 1\}^n.$$

That means when conditioned on the covariate $X_{[n]}$, the assignment of the treatment is independent of the **potential outcome**. (The treatment assignment is random within subpopulation conditioned by values of observed covariates)

- 1 The key is to rule out unobserved confounders.
- 2 Well-designed randomized control trials (RCT) should satisfy this assumption
- 3 Untestable in most observational studies, e.g. a patient may have a preference for a certain treatment based on his expected potential treatment effectiveness which is not captured in $X_{[n]}$

Positivity assumption

To identify causal effects in randomized experiments, we also need the following assumption:

Overlap (or positivity) assumption

$$P(A = a|X = x) > 0, \forall a \in A, x \in X$$

Remark

- 1 the positivity assumption requires, for all possible values of the covariate, there are both treated and control units.
- 2 the positivity assumption can be directly checked from data (unlike the ignorability assumption)
- 3 ignorability and positivity jointly define the "strong ignorability"

Causal identification

Causal identification under “strong ignorability”

Under the “strong ignorability” assumption, one has:

$$PATE := E_{Y_i}[Y_i(1) - Y_i(0)] = E_X\{E_I[Y_i | A = 1, X] - E[Y_i | A = 0, X]\}$$

Proof:

For any $y \in R$, $a \in \{0, 1\}$ and x , the strong ignorability implies

$$\begin{aligned} P(Y_i(a) \leq y | X = x) &= P(Y_i(a) \leq y | X = x, A = a) \quad (\text{strong ignorability}) \\ &= P(Y_i \leq y | X = x, A = a) \quad (\text{Consistency}) \end{aligned}$$

Then averaging Y_i over all units and taking the expectation over X .

More on potential outcome framework

If we back to Simpson's paradox, we know that the data does not resemble a nicely conducted randomized control trial. If we assume the strong ignorability holds, then certain weighted average among subgroups will give a better estimate of the treatment effect.

More on potential outcome framework

If we back to Simpson's paradox, we know that the data does not resemble a nicely conducted randomized control trial. If we assume the strong ignorability holds, then certain weighted average among subgroups will give a better estimate of the treatment effect.

- ① In general one needs to balance covariates (both measured and unmeasured confounders)
 - Measured covariate can be summarized into **propensity score**:

$$e(X) = P(A = 1 | X)$$

Therefore, instead of balancing all covariates, one can balance the propensity score.

- for unmeasured confounders, one approach is to use an **instrumental variable** ((i.e. IV) that influences treatment assignment but is independent of unmeasured confounders and has no direct effect on the outcome except through its effect on treatment)

A diagram for instrumental variable

The instrumental variable ((i.e. IV) is a variable that influences treatment assignment but is independent of unmeasured confounders and has no direct effect on the outcome except through its effect on treatment)

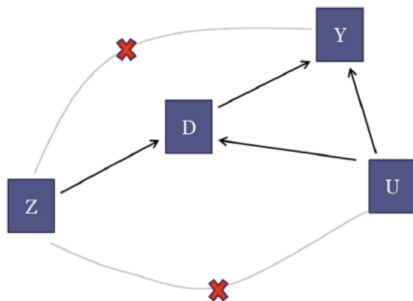


Figure 1. Directed acyclic graph for the relationship between an instrumental variable Z , a treatment D , unmeasured confounders U , and an outcome Y .

Directed acyclic graphs (DAG)

Path and directed path

Let G be a directed graph.

- 1 A **path** on G between vertices i and j are sequences of *distinct vertices* $k_0 = i, k_1, k_2, \dots, k_m = j$ such that the consecutive vertices are adjacent, that is (k_{l-1}, k_l) or (k_l, k_{l-1}) is an edge for all $l = 1, 2, \dots, m$.
- 2 A **directed path** on G is a path where all arrows are going “forward”.
- 3 A **cycle** is a **directed path** such that the first and the last vertices are the same.

Definition: DAG

A directed acyclic graph is a directed graph with no cycles.

- 1 Because causality implies ordering in time from cause to effect, cycles are not possible

Analysis causal effects using DAGs

Causal diagrams

A *causal diagram* is a DAG where the vertices are random variables and edges represent the direct causes.

- 1 If there is a directed path from vertex v_i to vertex v_j , we say v_i is a cause of v_j .

Analysis causal effects using DAGs

Causal diagrams

A *causal diagram* is a DAG where the vertices are random variables and edges represent the direct causes.

- 1 If there is a directed path from vertex v_i to vertex v_j , we say v_i is a cause of v_j .

The causal diagram is linked with data by the following key assumption.

Causal Markov assumption (in terms of pdf)

Let G be a causal graph with m vertices $\{v_j\}_{j=1}^m$. Then for any variable v_i , when conditioned on its parents $pa(v_i)$, v_i is independent of any other variable for which it is not a cause.

Causal Markov assumption

Causal Markov assumption

The causal Markov assumption implies the following decomposition of the density function $pdf(v_1, \dots, v_m)$ of the joint distribution:

$$pdf(v_1, \dots, v_m) = \prod_{j=1}^m pdf(v_j \mid pa(v_j))$$

where $pdf(v_j \mid pa(v_j))$ is the density function of v_j when conditioned on its parent variables given by the graph G .

Benefits of causal DAGs

- 1 Use causal DAGs to help understand possible bias when making causal inferences but not meant to have an exact, accurate representation of the world.
- 2 Every assumption in potential outcomes can be depicted using a graph.

An example: Estrogens and Uterine Cancer

We will use an example to show the effectiveness of using a causal graph. ³

³[https:](https://www.edx.org/course/causal-diagrams-draw-your-assumptions-before-your)

[//www.edx.org/course/causal-diagrams-draw-your-assumptions-before-your](https://www.edx.org/course/causal-diagrams-draw-your-assumptions-before-your),
adapted from a course by Miguel Hernán

An example: Estrogens and Uterine Cancer

We will use an example to show the effectiveness of using a causal graph. ³

In the 1970s, people were concerned about the possible link between women *receiving estrogens* and an increased risk of *getting cancer*. There were two explanations.

- 1 Estrogens cause cancer
- 2 Estrogens accelerate the diagnosis silent cancer due to uterine bleeding

³[https:](https://www.edx.org/course/causal-diagrams-draw-your-assumptions-before-your)

[//www.edx.org/course/causal-diagrams-draw-your-assumptions-before-your](https://www.edx.org/course/causal-diagrams-draw-your-assumptions-before-your),
adapted from a course by Miguel Hernán

An example: Estrogens and Uterine Cancer

- 1 Estrogens cause cancer
- 2 Estrogens accelerate the diagnosis of silent cancer due to uterine bleeding

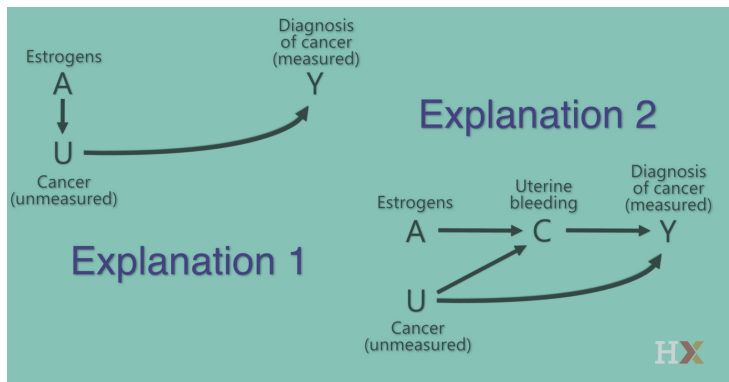
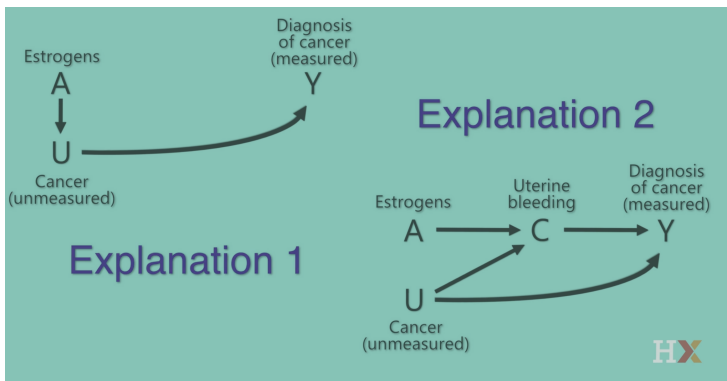


Figure: Causal graphs of two explanations

An example: Estrogens and Uterine Cancer

- To test the correctness of explanation 2, two investigators from Yale proposed to **restrict the data analysis to women who bleed, whether they are taking estrogens or not.**

Claim: If the use of estrogen still correlates with the diagnosis of cancer, then estrogen must have been more directly causing cancer.



An example: Estrogens and Uterine Cancer

- 1 However, scholars from Boston and Harvard disagree and argue that an **association** between estrogen use and the cancer diagnosis can arise in analyses **restricted to women who bleed**, even if estrogens don't cause cancer.

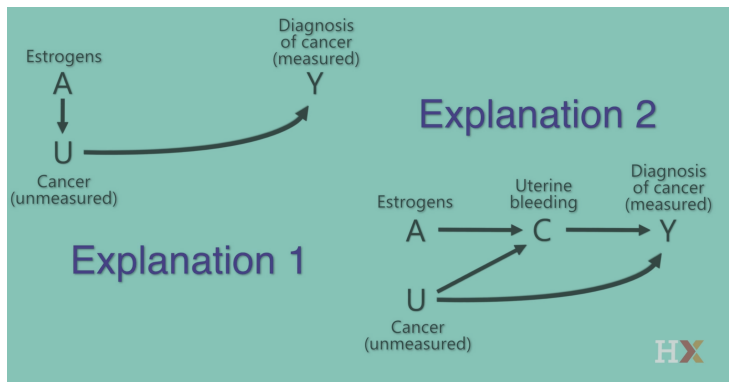


Figure: Causal graphs of two explanations

Three basic (local) causal graphs

Chain

- 1 A affects both B and Y .



Three basic (local) causal graphs

Chain

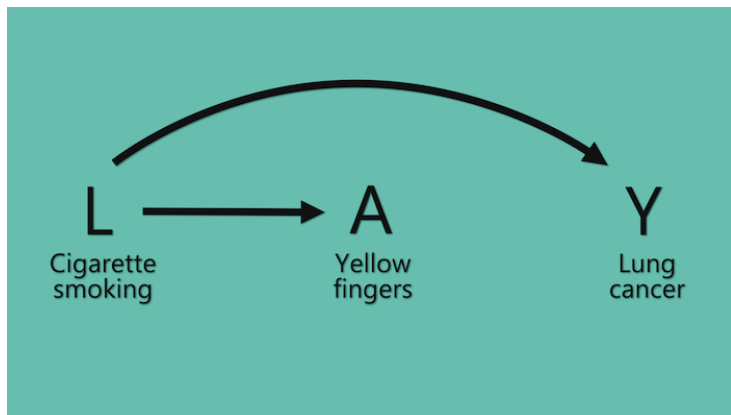
- 1 A and Y are independent when conditioned on B (path is blocked)



Three basic (local) causal graphs

Fork/confounder

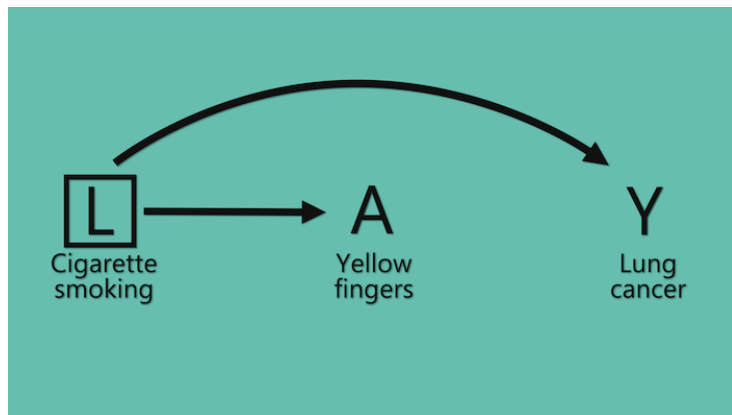
- 1 A is dependent on Y due to the common cause



Three basic (local) causal graphs

Fork/confounder

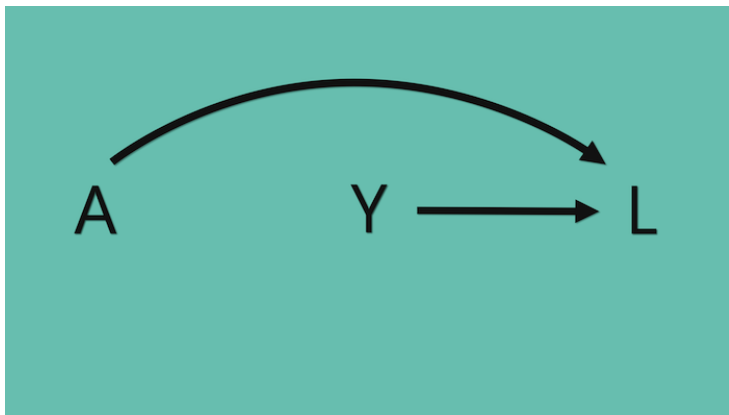
- 1 A and Y are independent when conditioned on L



Three basic (local) causal graphs

Collider

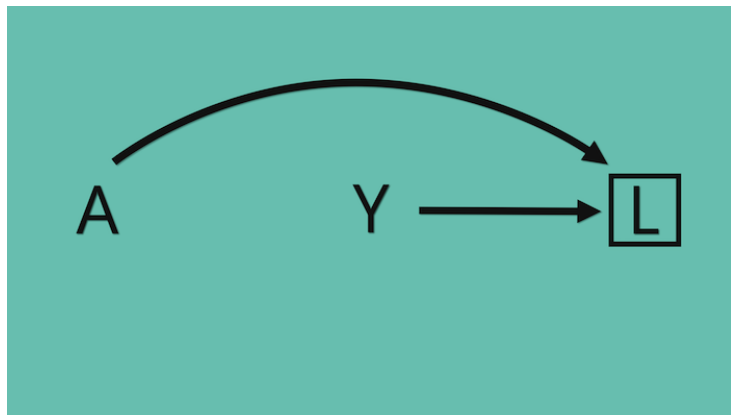
- 1 A and L are independent.



Three basic (local) causal graphs

Collider

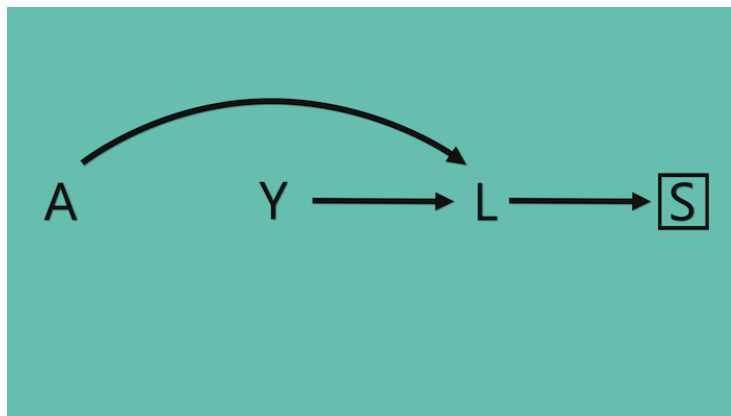
- 1 A and Y are **dependent** when conditioned on L . (think A : generic, Y environmental covariates, L is cancer)



Three basic (local) causal graphs

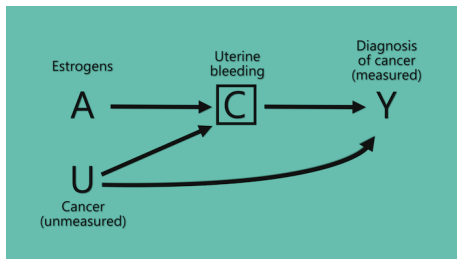
Collider

- 1 A and Y are **dependent** when conditioned on common effects.



Back to estrogens example

What happens when we restrict to patients with uterine bleeding as suggested by scholars from Yale?

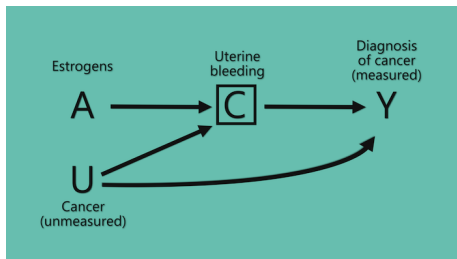


Question

Is it true that if we condition on C on the proposed causal graph (from explanation 2), A and Y will be independent?

Back to estrogens example

What happens when we restrict to patients with uterine bleeding as suggested by scholars from Yale?



Question

Is it true that if we condition on C on the proposed causal graph (from explanation 2), A and Y will be independent?

This is not true, as we know that C is a collider and condition on C will create an association between A and U and hence between A and Y .

Back to estrogens example

A causally correct approach: Instead of restricting on uterine bleeding, we should sample data screening for cancer regardless of bleeding or not.

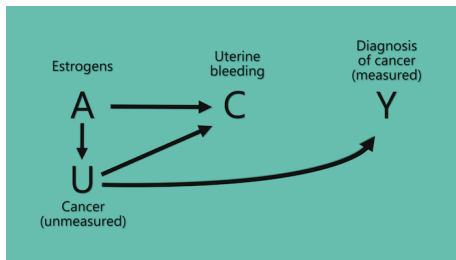


Figure: Randomized screening

d -separation

On a DAG, our previous knowledge about chain, fork, and collider can be generalized to more general paths.

Definition

Given a DAG G , a path is said to be blocked by $K \subseteq V$ if there exists a vertex k on the path such that either

- 1 k is not a collider on this path and $k \in K$; or
- 2 k is a collider on this path and $k \notin K$ and all its descendants are not in K ;

d -separation

On a DAG, our previous knowledge about chain, fork, and collider can be generalized to more general paths.

Definition

Given a DAG G , a path is said to be blocked by $K \subseteq V$ if there exists a vertex k on the path such that either

- 1 k is not a collider on this path and $k \in K$; or
- 2 k is a collider on this path and $k \notin K$ and all its descendants are not in K ;

For disjoint subsets of nodes $I, J, K \subset V$, we say **I and J are d -separated by K** , written as $I \perp\!\!\!\perp J \mid K[G]$, if all paths from a vertex in I to a vertex in J are blocked by K .

d -separation and causal discovery

With certain assumptions, we can infer the graphic model (d -separation) from the conditional independence in the observed data.

Faithfulness assumption

We say a distribution P of X that factorizes according to G is faithful to G if $I \perp\!\!\!\perp J | K [G] \iff X_I \perp\!\!\!\perp X_J | X_K$ for all disjoint $I, J, K \subset V$.

That is, sometimes the data cannot faithfully represent the causal effect. (Say A affects each unit in B but on average the effect is canceled.)

Next step: Causal **structure** model

So far in the causal graph,

- ① arrow only represents direct causal effect but not any specific data generation mechanism
- ② we haven't incorporated unobserved variables and counterfactuals into the graphs.

Next step: Causal **structure** model

So far in the causal graph,

- 1 arrow only represents direct causal effect but not any specific data generation mechanism
- 2 we haven't incorporated unobserved variables and counterfactuals into the graphs.

The definition of the causal structural model is lengthy and out of the scope of this introduction. (cf. Chapter 5 in Zhao, *Lecture Notes on Causal Inference*)

Causal inference and machine learning

Machine learning in general can make very good predictions. However, most ML models fall inside the statistical inference region and often require the training and testing data coming from the same i.i.d. distribution.

Benefits from including causal structure into ML models

- 1 Robustness
- 2 Learning reusable mechanism (adaptation to a new environment)
- 3 Correctly applies to data where requires causal perspective (Health data, economic data)

Causal inference and machine learning

Machine learning in general can make very good predictions. However, most ML models fall inside the statistical inference region and often require the training and testing data coming from the same i.i.d. distribution.

Benefits from including causal structure into ML models

- 1 Robustness
- 2 Learning reusable mechanism (adaptation to a new environment)
- 3 Correctly applies to data where requires causal perspective (Health data, economic data)

For more discussion: see [Bernhard Schölkopf et al. "Toward causal representation learning"](#). In: *Proceedings of the IEEE 109.5 (2021)*, pp. 612–634.

Causal inference and machine learning




Machine learning models can also help causal inference, for example, instead of being restricted to working with *average treatment effects*, the prediction power provided by the ML model can work with *individual treatment effects*.

Some Challenges for using ML models for causal inference

- 1 Cross-validation is difficult due to no access to the true causal effect
- 2 Good performance at predicting (fitting) the observed outcomes does not necessarily translate into good performance in causal estimation.

Useful links

- [Mihaela van der Schaar's Lab at Cambridge and UCLA](#)
- [Kun Zhang's group at CMU](#)

-  Li, Fan. *STA 640: Causal Inference*. 2022. URL: <https://www2.stat.duke.edu/~fl35/CausalInferenceClass.html>.
-  Schölkopf, Bernhard et al. “Toward causal representation learning”. In: *Proceedings of the IEEE* 109.5 (2021), pp. 612–634.
-  Zhao, Qingyuan. *Lecture Notes on Causal Inference*. May 2022. URL: http://www.statslab.cam.ac.uk/~qz280/teaching/causal-2022/notes_2021.pdf.

Sparse, Multivariate Functional Outcome Data

Given that,

- 1 $N_i(t)$: the assessment time process.
- 2 $Y_i(t)$: the repeated measure which follows a Gauss process $GP(\mu(t), C(s, t)) + \epsilon$.
- 3 $\lambda(t)$: the intensity for the recurrent event $N_i(t)$.

Assumption on intensity function

$$\lambda(t) = \lambda_0(t) \exp\{\alpha(t) + h(Y(t))\}$$

where $\lambda_0(t)$ is the baseline intensity function at time t , $\alpha(t)$ can be considered as the exponential mean effect.

Sparse, Multivariate Functional Outcome Data

Likelihood of intensity function approach

$$\begin{aligned} L_{N|Y}(\theta, N|Y) &= \prod_i \left\{ \left(\prod_i [\exp\{\alpha_i(t) + h(Y_i(t))\} \lambda_0(t)]^{\Delta N_i(t)} \right) \right. \\ &\quad \left. \times \exp \left[- \int_0^\tau G_i(t) \exp\{\alpha_i(t) + h(Y_i(t))\} dA_0(t) \right] \right\} \end{aligned}$$

Where, $A_0(t) = \int_0^t \lambda_0(u) du$ is the cumulative baseline intensity and $G_i(t)$ denotes whether subject i is available for assessment at time t ($G_i(t) = I(C_i \geq t)$).

Sparse, Multivariate Functional Outcome Data

Event rate function when Assume the event process if a Poisson process when given information on $Y_i(t)$.

$$r_i^*(t) = E[r_0(t) \exp\{\alpha(t) + h(Y(t))\}]$$

Maximum likelihood of rate function approach

$$\prod_{i=1}^n \prod_{j=1}^{m_i} \frac{r_i^*(t_{ij})}{\sum_{k=1}^n G_k(t_{ij}) r_k^*(t_{ij})}$$

Which is maximized at the solution of

$$\sum_{i=1}^n \sum_{j=1}^{m_i} \left(\frac{r_i^*(t_{ij})'}{r_i^*(t_{ij})} - \frac{\sum_{k=1}^n G_k(t_{ij}) r_k^*(t_{ij})'}{\sum_{k=1}^n G_k(t_{ij}) r_k^*(t_{ij})} \right) = 0$$

where $r_k^*(t_{ij})$ denotes the first derivative taken on subject k 's conditional rate function at time t_{ij}

Sparse, Multivariate Functional Outcome Data

Formula of $r_i^*(t)$

$$r_i^*(t) = r_0(t) \exp\{\alpha(t)\} E(\exp\{h(Y_i)\} \mid Y_i) = r_0(t) \exp\{\alpha(t)\} \exp\{h(Y_i)\}$$

Then we obtain

Conditional partial likelihood

$$\begin{aligned} & \prod_{i=1}^n \prod_{j=1}^{m_i} \frac{r_0(t_{ij}) \exp\{\alpha(t_{ij})\} \exp\{h(Y_i)\}}{\sum_{k=1}^n G_k(t_{ij}) r_0(t_{ij}) \exp\{\alpha(t_{ij})\} \exp\{h(Y_k)\}} \\ &= \prod_{i=1}^n \prod_{j=1}^{m_i} \frac{\exp\{h(Y_i)\}}{\sum_{k=1}^n G_k(t_{ij}) \exp\{h(Y_k)\}} \end{aligned}$$