Causality: Philosophy, Math and Machine Learning

Renjie Yang

COMPHI LAB for Social Data Science

June 2019

Outline

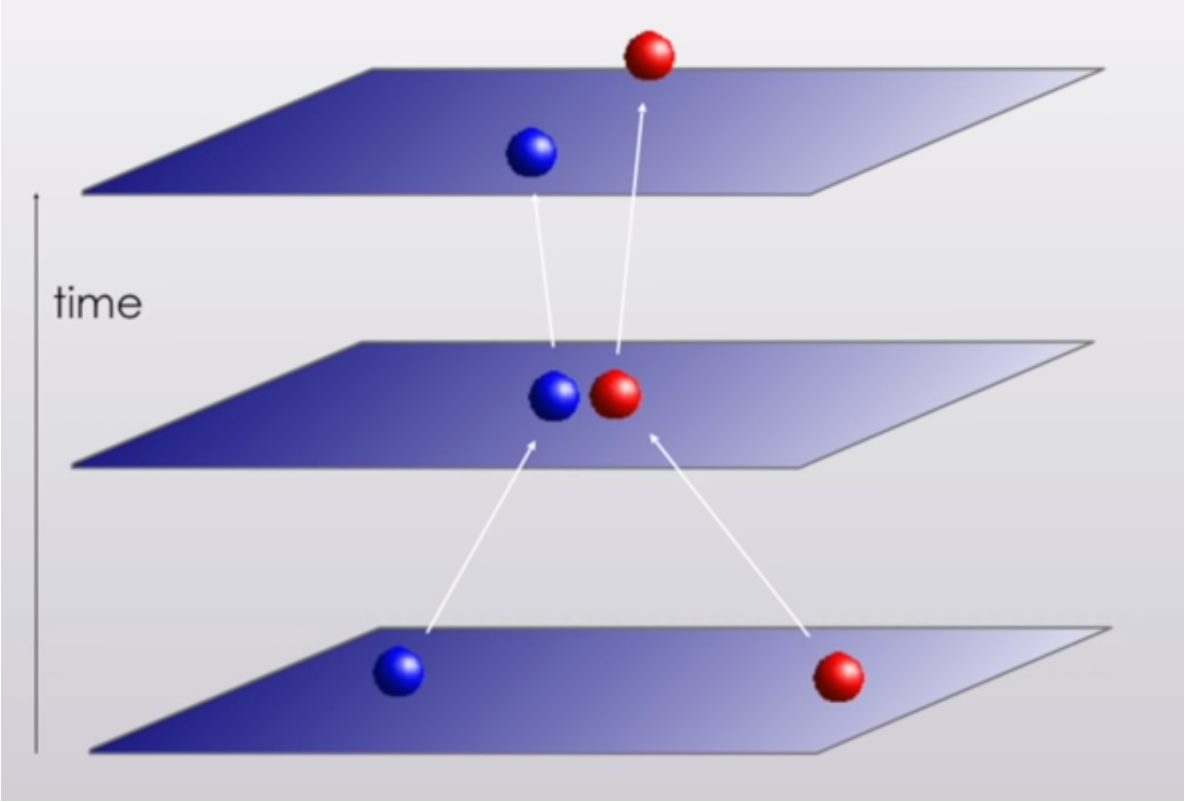
1. The Concept of Causality
2. Potential Outcomes and Counterfactuals
3. Causal Machine Learning: The Graphical Modeling Framework
4. The Limitation of Causal Discovery

Principle of Causality

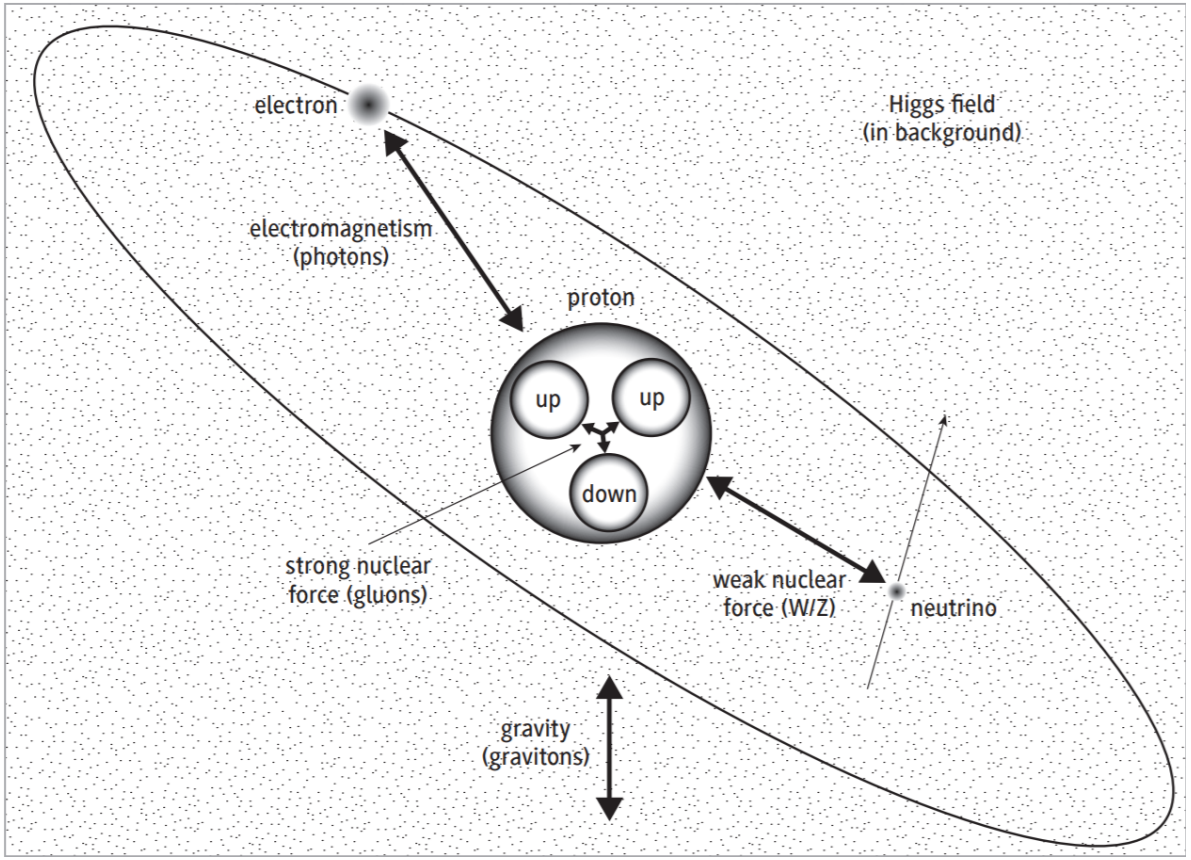
The laws that are used in the explanations of mature sciences cannot be interpreted as causal laws.

- The same causes always have the same effects.
- Russell (1912): "Given any event E_1, there is an event E_2 and a time-interval τ such that, whenever E_1 occurs, E_2 follows after an interval τ"

Fundamental Physics vs. Causality



Fundamental Physics vs. Causality(Sean Carroll, 2016)



Functional Law vs. Causality(Sean Carroll, 2016)

quantum mechanics spacetime gravity

$$W = \int_{k<\Lambda} [Dg][DA][D\psi][D\Phi] \exp\left\{ i \int d^4x \sqrt{-g} \left[\frac{m_p^2}{2} R \right.\right.$$

$$\left.\left. -\frac{1}{4} F^a_{\mu\nu} F^{a\mu\nu} + i\bar{\psi}^i \gamma^\mu D_\mu \psi^i + \left(\bar{\psi}^i_L V_{ij} \Phi \psi^j_R + \text{h.c.} \right) - |D_\mu \Phi|^2 - V(\Phi) \right] \right\}$$

other forces matter Higgs

Philosophical Analysis of Causality I

- Counterfactual analysis: for any two events C and E that have actually occurred, C causes E if and only if it is true that: if C had not occurred, E would not have occurred.
- The process analysis: an event C causes another event E if and only if there is a physical process of transmission between C and E.

Philosophical Analysis of Causality II

- The probabilistic analysis: factor C exercises a causal influence on factor E if and only if an event of type C raises the probability of an event of type E.
- The manipulability analysis: there is a causal relation between two variables C and E if and only if interventions modifying the value of C modify the value of E.

Counterfactual Analysis (Lewis, 1986)

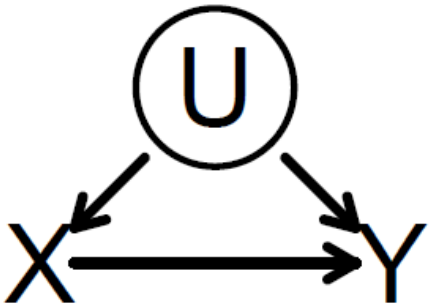
- "If C were the case, E would be the case" is true in a world w if and only if (1) C is not true in any possible world or (2) if some world in which both C and E are true is closer to w than all possible worlds in which C is true but E false.
- C is a cause of E if and only if there is a finite chain of intermediate events $E_1, E_2, ..., E_k$, between C and E, such that the nth link depends causally on the preceding $(n-1)$th link.

Process Analysis (Salmon, 1984)

- $pV = nRT$
- A causal process is a process that (1) has structure or qualities that are either permanent or only changing continuously and (2) is capable of transmitting a mark.

Probabilistic Analysis (Cartwright, 1979)

- C causes E if and only if $P(E|C\&D_i) > P(E|non-C\&D_i)$ for all D_i, where D_i are causes of E that are not caused by C.

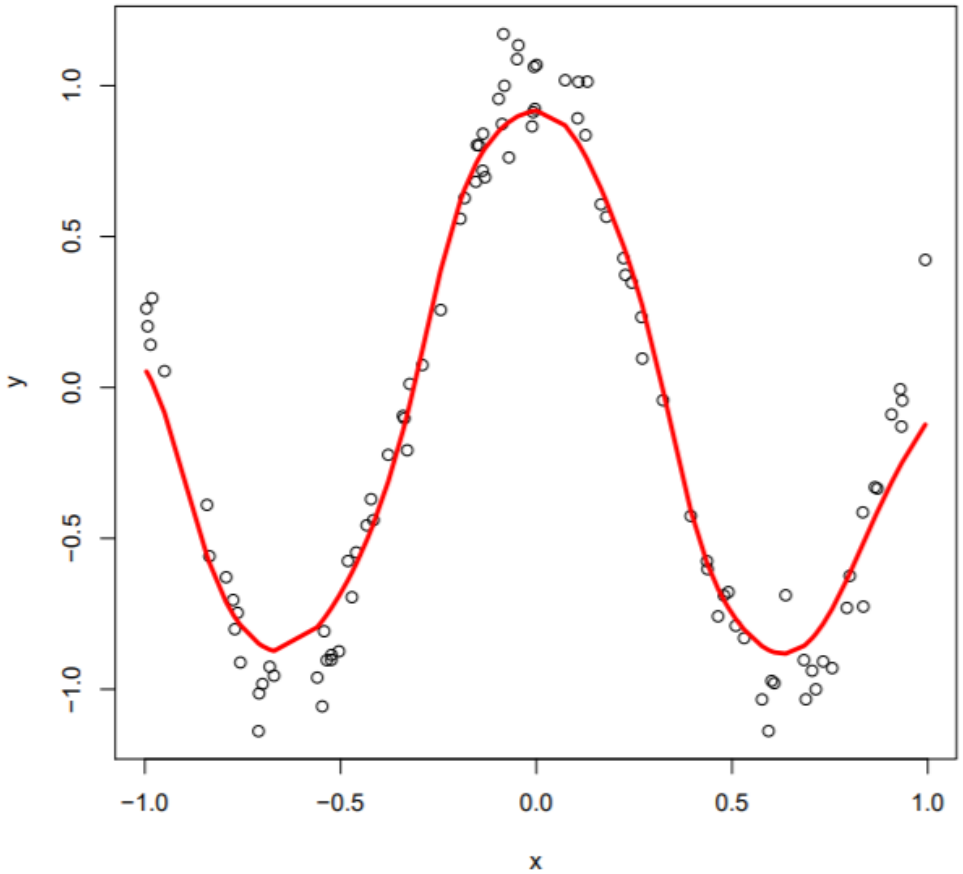


Manipulability Analysis

- A cause C of an effect E is an action that would give a human agent a means to obtain E if she decided to make C happen.

Probablistic Prediction

We observe training data $Z_1, ..., Z_n \sim P$ where $Z_i = (X_i, Y_i)$, $X_i \in R^d$. Given a new pair $Z = (X, Y)$ we want to predict Y from X.



Causal Questions

Prediction and causation are very different.

- Prediction: Predict Y after **observing** $X = x$
- Causation: Predict Y after **setting** $X = x$

Causal Questions

- Prediction: Predict health given that a person takes vitamin C
- Causation: Predict health if I give a person vitamin C

"If I place this ad on a web page, will people click on it?" and "If I recommend a product will people buy it?" are causal questions, not predictive questions.

Outline

1. The Concept of Causation
2. **Potential Outcomes and Counterfactuals**
3. Causal Machine Learning: The Graphical Modeling Framework
4. The Limitation of Causal Discovery

The Role of Causation in Social Science

Jon Elster(2015):

- "The main task of the social sciences is to explain social phenomena."
- "I argue that all explanation is causal."
- "To explain a phenomenon (an explanandum) is to cite an earlier phenomenon (the explanans) that caused it."

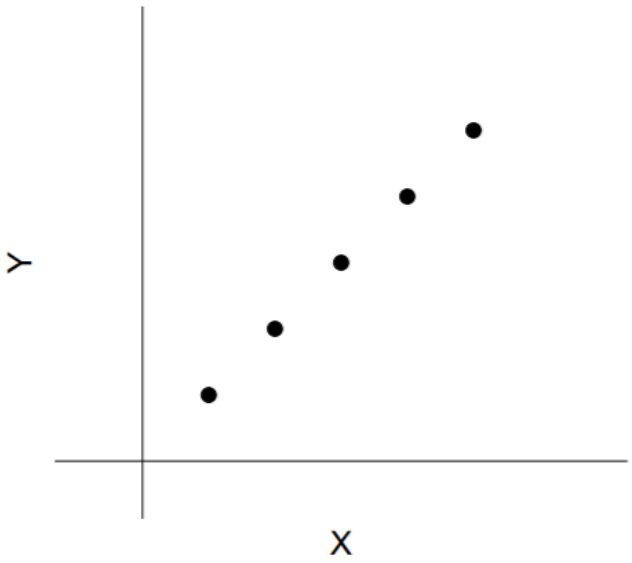
Two types of causal questions

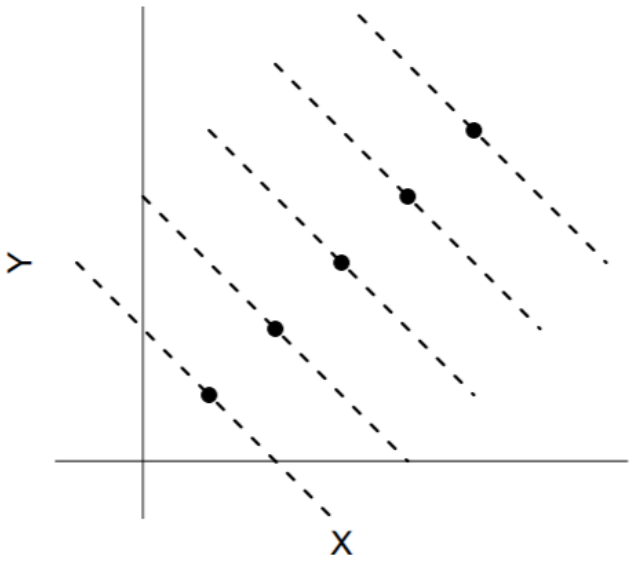
- Causal Inference(Causal Effect): How do cell phones cause brain cancer?
- Causal Discovery: Given a set of variables, determine the causal relationship between the variables.

Counterfactuals

$$Y = \begin{cases} Y_1 & \text{if } X = 1 \\ Y_0 & \text{if } X = 0. \end{cases}$$

Counterfactuals





Potential Outcomes

X	Y	Y_0	Y_1
1	1	*	1
1	1	*	1
1	0	*	0
1	1	*	1
0	1	1	*
0	0	0	*
0	1	1	*
0	1	1	*

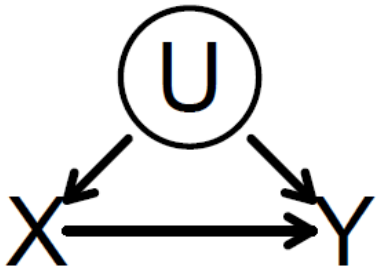
Causal Effect

$$\theta = E(Y_1) - E(Y_0) = E(Y|setX = 1) - E(Y|setX = 0)$$
$$\alpha = E(Y_1) - E(Y_0) = E(Y|X = 1) - E(Y|X = 0)$$

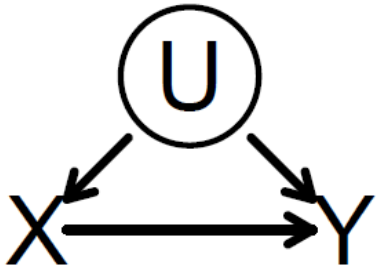
- There are uniform consistent estimators for α. But α and β are not equal!

$$\sup_{P \in \mathcal{P}} P(|\widehat{\theta}_n - \theta(P)| > \epsilon) \to 0$$

Two Ways to Make θ Estimable

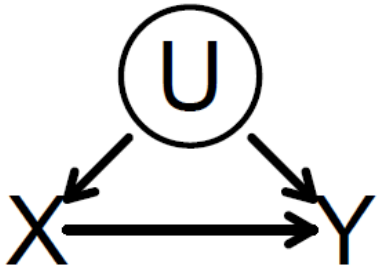


Randomized Experiments



1. **Randomization**: Suppose that we randomly assign X. Then X will be independent of $(Y_0; Y_1)$.
2. If X is randomly assigned then correlation = causation.

Measure Confounding Variables



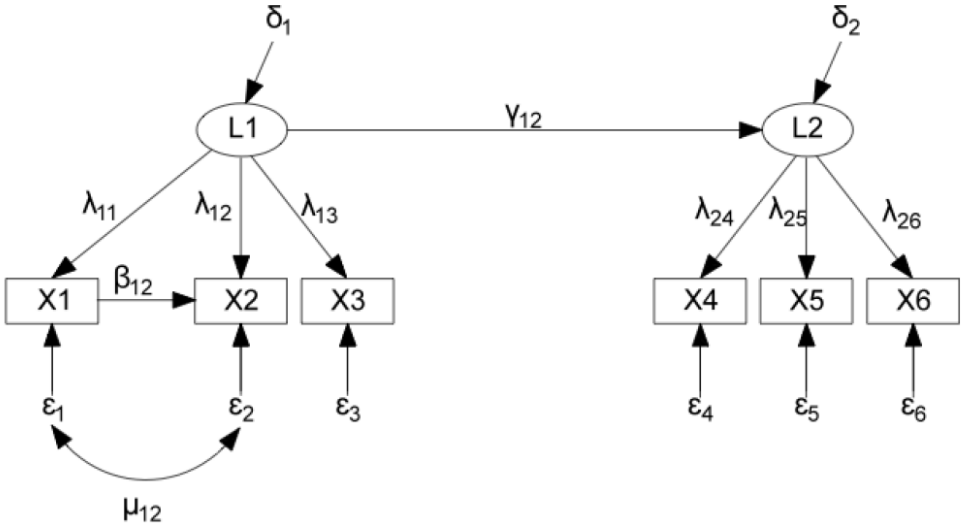
1. If we measure enough confounding variables $U = (U_1; \ldots; U_k)$, then, perhaps the treated and untreated groups will be comparable, conditional on Z.
2. Causal inference in observational studies is not possible without subject matter knowledge.

Outline

1. The Concept of Causation
2. Potential Outcomes and Counterfactuals
3. **Causal Machine Learning: The Graphical Modeling Framework**
4. The Limitation of Causal Discovery

Representation: SEM

$$X_1 := \lambda_{11} L_1 + \epsilon_1$$
$$X_2 := \lambda_{12} L_1 + \beta_{12} X_1 \epsilon_2$$
$$X_3 := \lambda_{13} L_1 + \epsilon_3$$
$$X_4 := \lambda_{24} L_2 + \epsilon_4$$
$$X_5 := \lambda_{25} L_2 + \epsilon_5$$
$$X_6 := \lambda_{26} L_2 + \epsilon_6$$
$$L_1 := \delta_1$$
$$L_2 := \gamma_{12} L_1 + \delta_2$$
$$cov(\epsilon_1, \epsilon_2) = \mu_{12}$$



Structural Equation Models: Definition

A *structural equation model* is a quadruple $S = \langle \mathbf{V}, \mathbf{E}, \mathbf{F}, P(\mathbf{E}) \rangle$, where

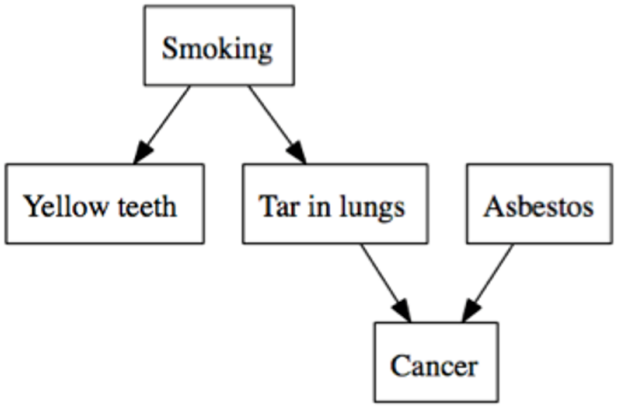
- $\mathbf{V}$ is a finite set of random variables $V_i, i \in \{1, 2, \ldots, n\}, n \in N$;
- $\mathbf{E}$ is another set of random variables $E_i, i \in \{1, 2, \ldots, n\}, n \in N$, where each element E_i in $\mathbf{E}$ corresponds to an element V_i in $\mathbf{V}$;
- $\mathbf{F}$ is a set of measurable functions $f_i, i \in \{1, 2, \ldots, n\}, n \in N$ such that each function f_i is a mapping from $\mathbf{V} \cup \mathbf{E} \backslash V_i$ to V_i;
- $P(\mathbf{E})$ is a joint probability distribution over the set of random variables $\mathbf{E}$.

(Pearl 2000)

Causal Inferences: Assumptions (Spirtes, 2010)

Weak Causal Markov Assumption: For a causally sufficient set of variables $\mathbf{V}$ in a population, if no variable in set $\mathbf{X} \subset \mathbf{V}$ causes any variable in set $\mathbf{Y} \subset \mathbf{V}$, and no variable in $\mathbf{Y} \subset \mathbf{V}$ causes any variable in $\mathbf{X} \subset \mathbf{V}$, then $\mathbf{X}$ and $\mathbf{Y}$ are independent, or in other words, any variable in $\mathbf{X}$ is independent with any variable in $\mathbf{Y}$.

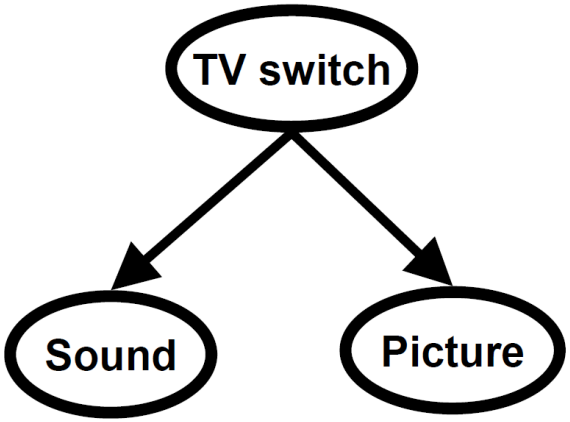
Causal Faithfullness Assumption: For a causally sufficient set of variables $\mathbf{V}$ in a population, the population distribution f is faithful to the causal graph G for the population in the sense that every conditional independence relation that is true in f is entail by G based on the Causal Markov Assumption.

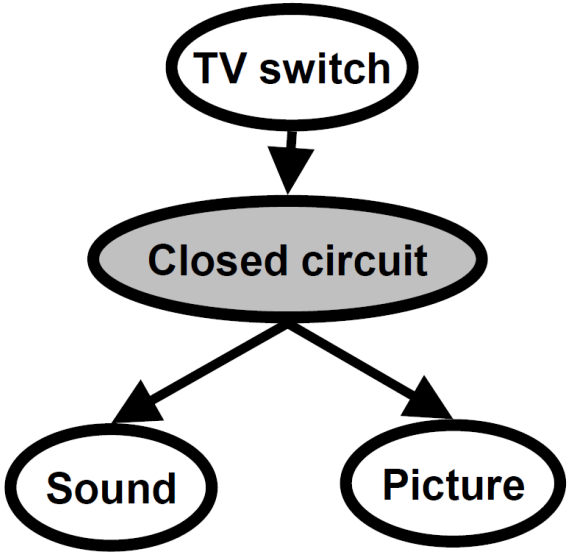


Causal Markov Assumption(Spirtes, 2000)

A ⫫ {All of its (non-parental) non-descendants} | A's parents

"The Interactive Fork"

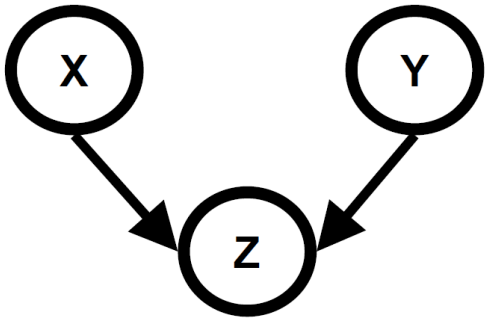


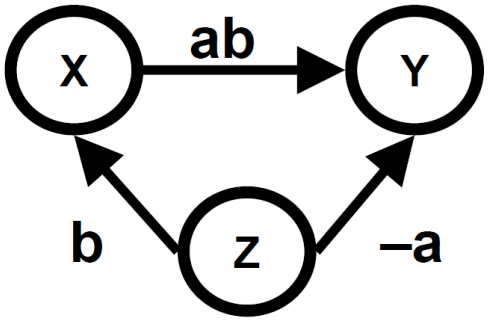


P(Picture | Switch) < P(Picture | Switch, Sound)

Where did the independence go?

Causal Faithfulness Assumption(Pearl, 2000)

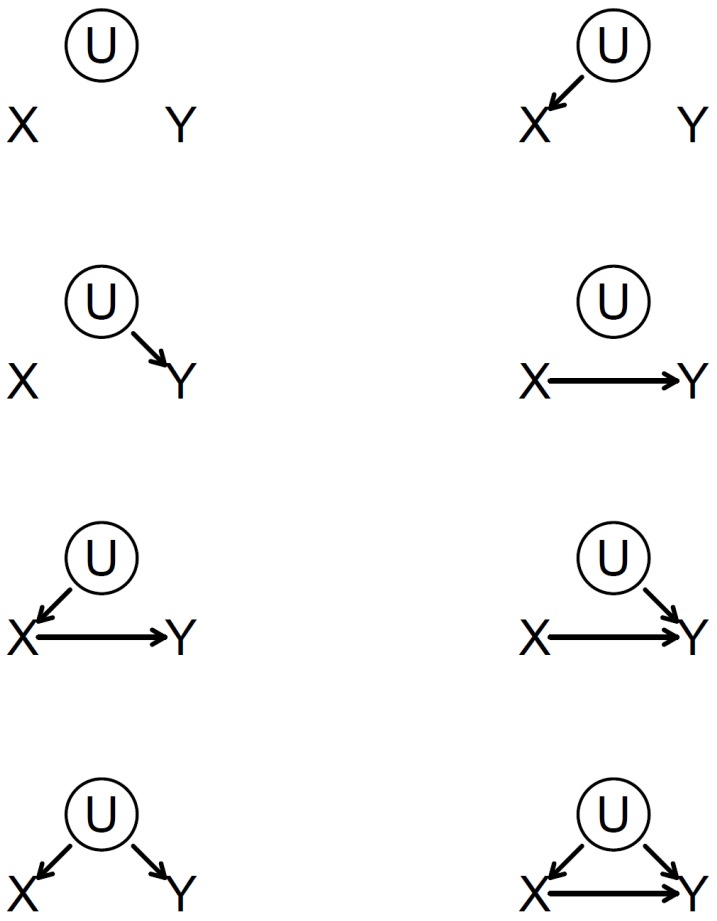




Graph	Distribution
$X \perp\!\!\!\perp Y$	$X \perp\!\!\!\perp Y$
$X \not\perp\!\!\!\perp Y \mid Z$	$X \not\perp\!\!\!\perp Y \mid Z$

Graph	Distribution
$X \not\perp\!\!\!\perp Y$	$X \perp\!\!\!\perp Y$
$X \not\perp\!\!\!\perp Y \mid Z$	$X \not\perp\!\!\!\perp Y \mid Z$

Causal Discovery

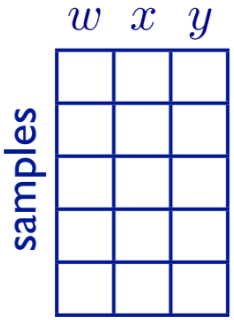


Causal Discovery

$$\begin{aligned} \alpha \neq 0 \quad &\Longrightarrow \quad \theta \text{ can be } 0 \text{ or nonzero (no conclusion)} \\ \alpha = 0 \text{ and faithfulness} \quad &\Longrightarrow \quad \theta = 0 \text{ (no causal effect).} \end{aligned}$$

Causal Discovery

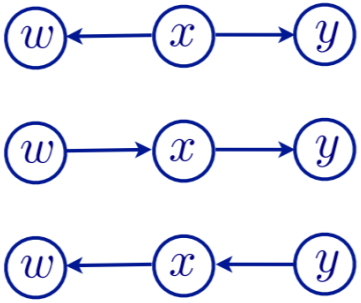
data sample



Assumptions
- causal Markov
- causal faithfulness
- causal sufficiency
- acyclicity

- independence based algorithms (e.g. PC and its variants)
- greedy Bayesian algorithms (e.g. GES)
- exact Bayesian algorithms

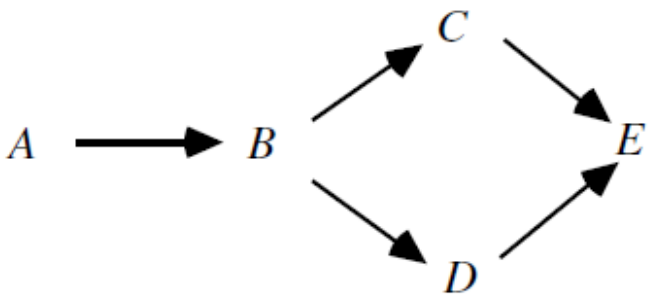
Markov equivalence class



The PC Algorithm (Spirtes, 2000)

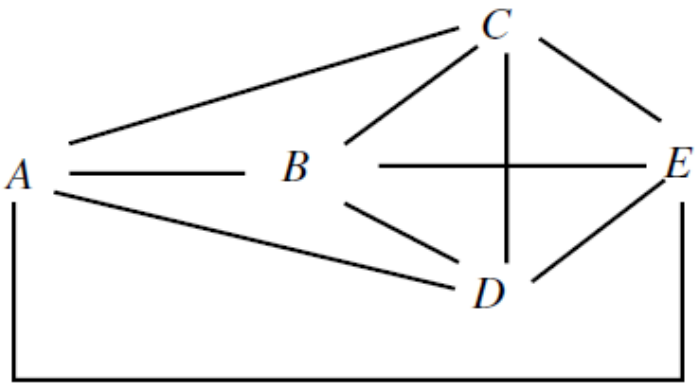
1. Start by testing marginal independencies
2. Next step: conditional independencies tests of size 1, 2, ..., until no tests of a given size pass
3. Orient edges according to which tests passed

The PC Algorithm (Spirtes, 2000)



True Graph

The PC Algorithm (Spirtes, 2000)



Complete Undirected Graph

The PC Algorithm (Spirtes, 2000)

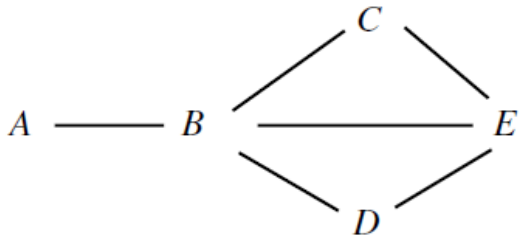
n = 0 No zero order independencies

n = 1 First order independencies

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

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

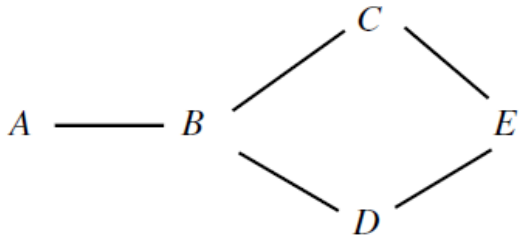
Resulting Adjacencies



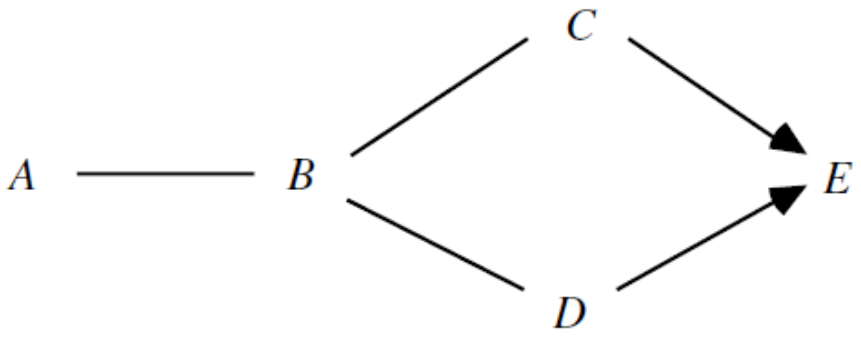
n = 2: Second order independencies

$B \perp\!\!\!\perp E \mid \{C,D\}$

Resulting Adjacencies



The PC Algorithm (Spirtes, 2000)



The PC Algorithm (Spirtes, 2000)

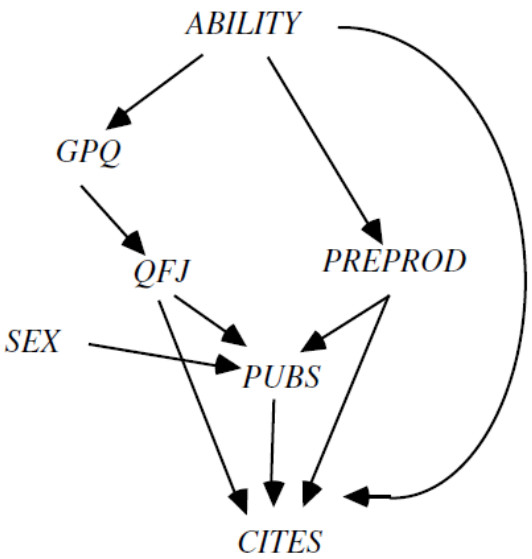
D.) repeat

If $A \rightarrow B$, B and C are adjacent, A and C are not adjacent, and there is no arrowhead at B, then orient B - C as $B \rightarrow C$.

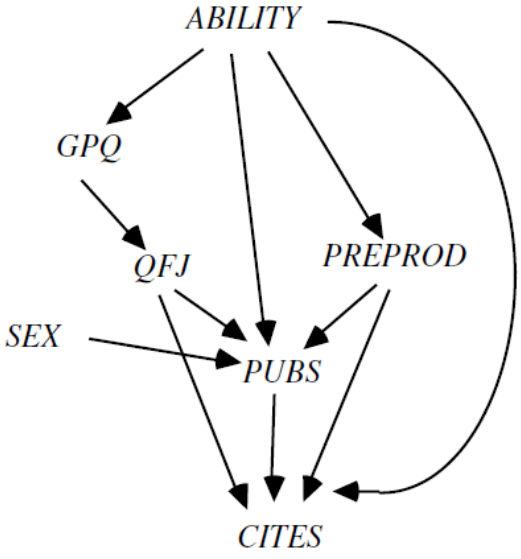
If there is a *directed* path from A to B, and an edge between A and B, then orient A - B as $A \rightarrow B$.

until no more edges can be oriented.

PC Algorithm Application (Spirtes, 2000)



PC Output

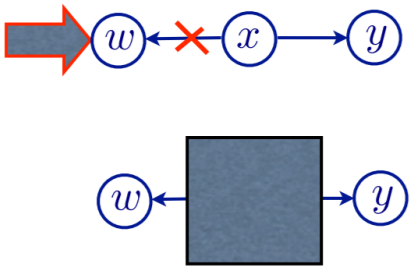


Rodgers and Maranto Graph

Causal Machine Learning

generalize

experimental / observational conditions



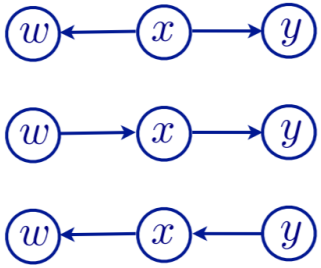
weaken

Assumptions

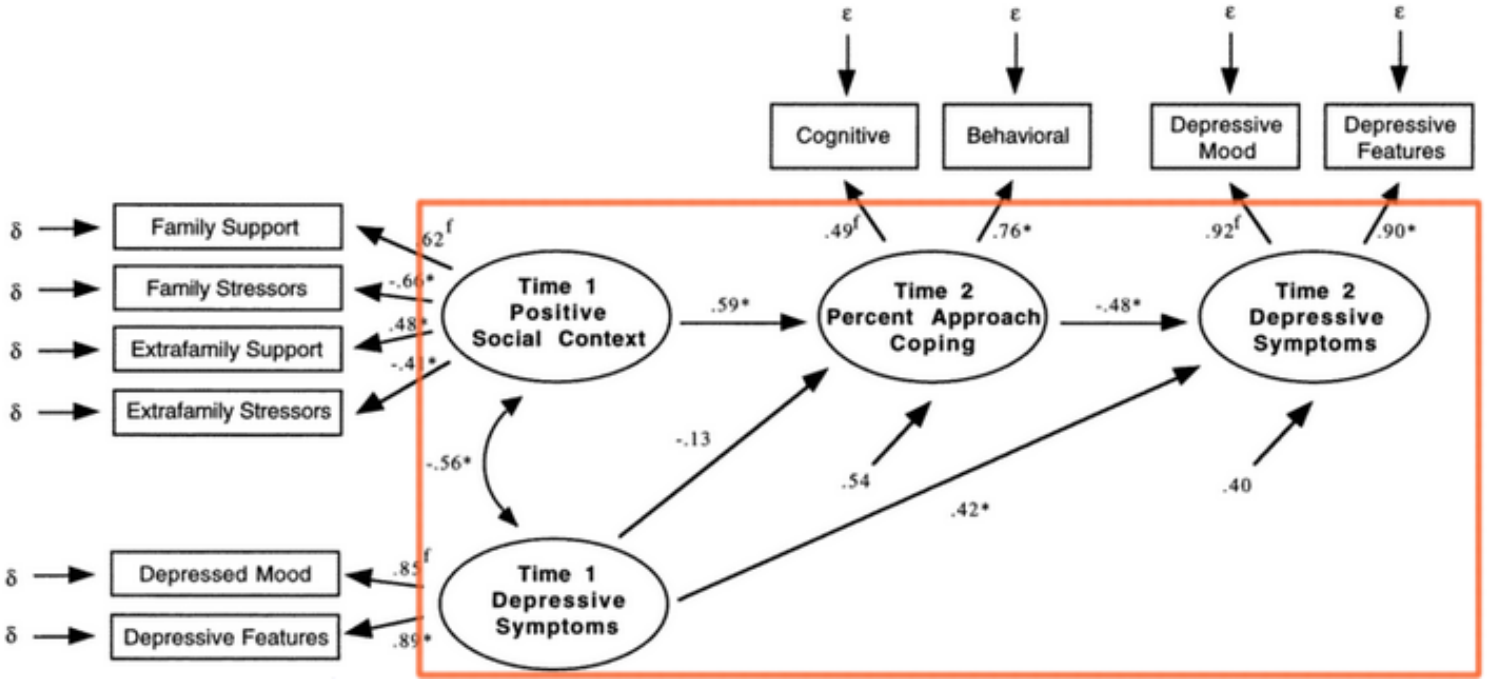
- causal Markov
- causal faithfulness
- causal sufficiency
- acyclicity

reduce

Markov equivalence class



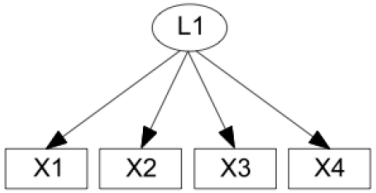
Causal Machine Learning for Social Science



Causal Machine Learning with Latent Constructs

Non-experimental Data

Path Diagram ← t-separation ← Rank Constraint



$\{\{L_1\};\phi\}$ t-separate $\{\{X_1, X_3\}, \{X_2, X_4\}\}$

$$\begin{array}{c c} & \begin{array}{cc} 2 & 4 \end{array} \\ \begin{array}{c} 1 \\ 3 \end{array} & \begin{vmatrix} \sigma_{12} & \sigma_{14} \\ \sigma_{23} & \sigma_{34} \end{vmatrix} = 0 \end{array}$$

Outline

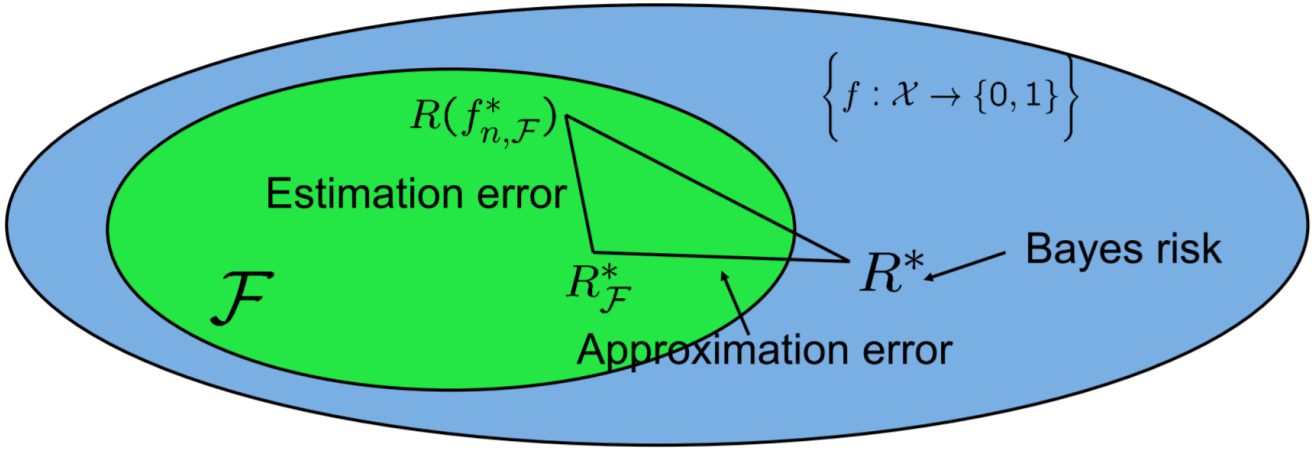
1. The Concept of Causation
2. Potential Outcomes and Counterfactuals
3. Causal Machine Learning: The Graphical Modeling Framework
4. **The Limitation of Causal Discovery**

Causal Effect

$$\theta = E(Y_1) - E(Y_0) = E(Y|setX = 1) - E(Y|setX = 0)$$
$$\alpha = E(Y_1) - E(Y_0) = E(Y|X = 1) - E(Y|X = 0)$$

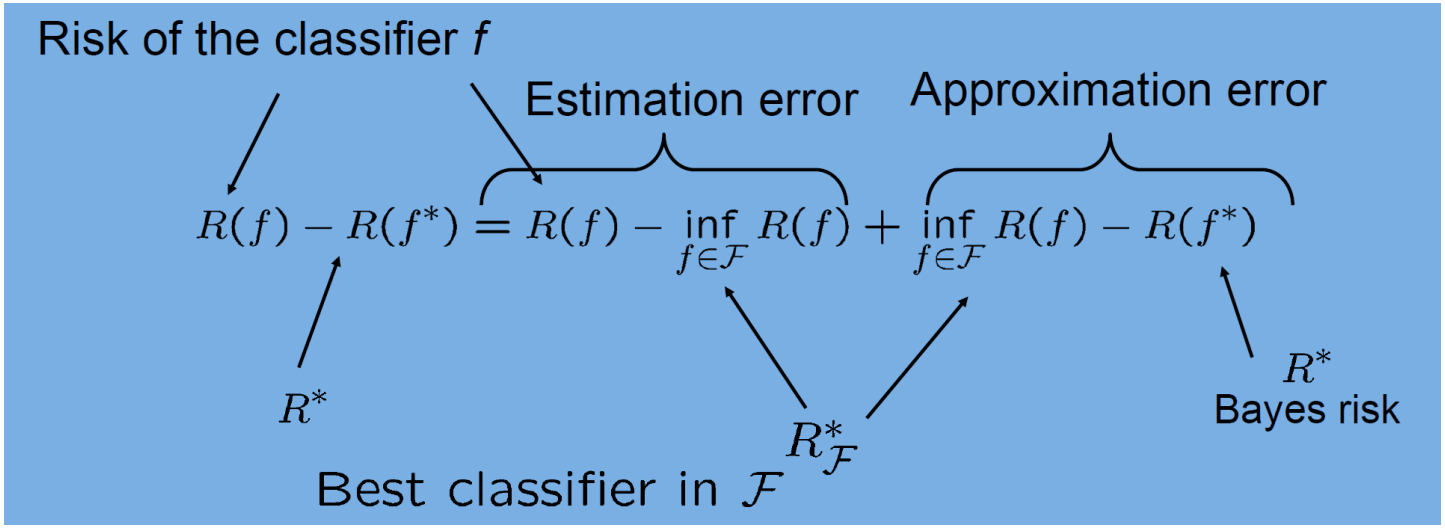
- There are uniform consistent estimators for α. But α and β are not equal!
- In general, there does not exist a uniformly consistent estimator of θ.

Statistical Learning Theory



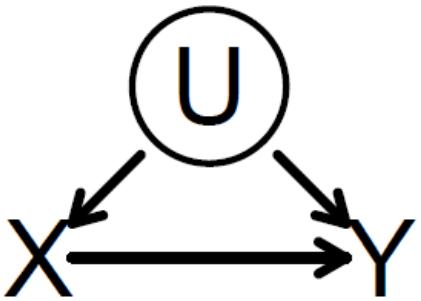
$$\sup_{P \in \mathcal{P}} P(|\widehat{\theta}_n - \theta(P)| > \epsilon) \to 0$$

Statistical Learning Theory



Test for Zero Causal Effect

Consider testing $H_0 : \theta = 0$ vs. $H_1 : \theta \neq 0$



Consistency for Causal Discovery(Robins et al, 2003)

Definition 2 . *A test ϕ is pointwise consistent if*

(i) for every $P \in \Omega_{\mathcal{G}0}$, $\lim_{n\to\infty} P^n\{\phi_n(O^n) = 1\} = 0$ and

(ii) for every $P \in \Omega_{\mathcal{G}1}$, $\lim_{n\to\infty} P^n\{\phi_n(O^n) = 0\} = 0$.

Definition 3 . *A test ϕ is uniformly consistent if*

(i) $\lim_{n\to\infty} \sup_{P\in\Omega_{\mathcal{G}0}} P^n\{\phi_n(O^n) = 1\} = 0$ and

(ii) for every $\delta > 0$, $\lim_{n\to\infty} \sup_{P\in\Omega_{\mathcal{G}\delta}} P^n(\phi_n(O^n) = 0) = 0$.

Consistency for Causal Discovery(Robins et al, 2003)

Theorem 1 . *Consider testing $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. Given the Markov assumption and full Markov support, if Z is observed then there exist pointwise and uniformly consistent tests of this hypothesis. Given the Markov assumption, the faithfulness assumption and full Markov support, if Z is not observed then there exist pointwise consistent tests but there is no uniformly consistent test.*

Theorem 4 . *Given the Markov assumption, the faithfulness assumption and full Markov support, if in Example 2 Z is unobserved then there does not exist a consistent, confidence map for θ. The same holds in Example 3, if R and S are unobserved.*

Causal Effect

$$\theta = E(Y_1) - E(Y_0) = E(Y|setX = 1) - E(Y|setX = 0)$$
$$\alpha = E(Y_1) - E(Y_0) = E(Y|X = 1) - E(Y|X = 0)$$

- There are uniform consistent estimators for α. But α and β are not equal.
- Even assuming unfaithfulness, there does not exist a uniformly consistent estimator of θ!

So what to do next?

Take Away Message

1. Causal inferences are typically much harder than statistical inferences.
2. Causal effects can be estimated consistently from randomized experiments.
3. It is difficult to estimate causal effects from observational data or non-randomized experiments.
4. Causal discovery is statistically impossible.

References

1. Barnabs Pczos' lecture Slides "Risk Minimization"
2. Bertrand Russell, "On the Notion of Cause"
3. David Lewis, "Causation", in *Philosophical Papers*
4. Frederick Eberhardt's lecture Slides "All of Causal Discovery"
5. Judea Pearl, *Causality*
6. Larry Wasserman's notes on Causation
7. Max Kistler, "Causality", in *The Philosophy of Science: A Companion*
8. Nancy Cartwright, *Causal Laws and Effective Strategies*
9. Ricardo Silva's slides "Causal Inference in Machine Learning"
10. Robins et al., "Uniform Consistency in Causal Inference"
11. Sean Carroll, *The Big Picture*
12. Spirtes et al., *Causation, Prediction and Search*
13. Wesley Salmon, *Scientific Explanation and the Causal Structure of the World*