Causal Inference

Vikram Mullachery

Jan 2019

References

- **Elements of Causal Inference: Foundations and Learning Algorithms** by Jonas Peters, Dominik Janzing, Bernhard Schölkopf (2017)
- **Causality: Models, Reasoning and Inference** by Judea Pearl (2000)

Thus far

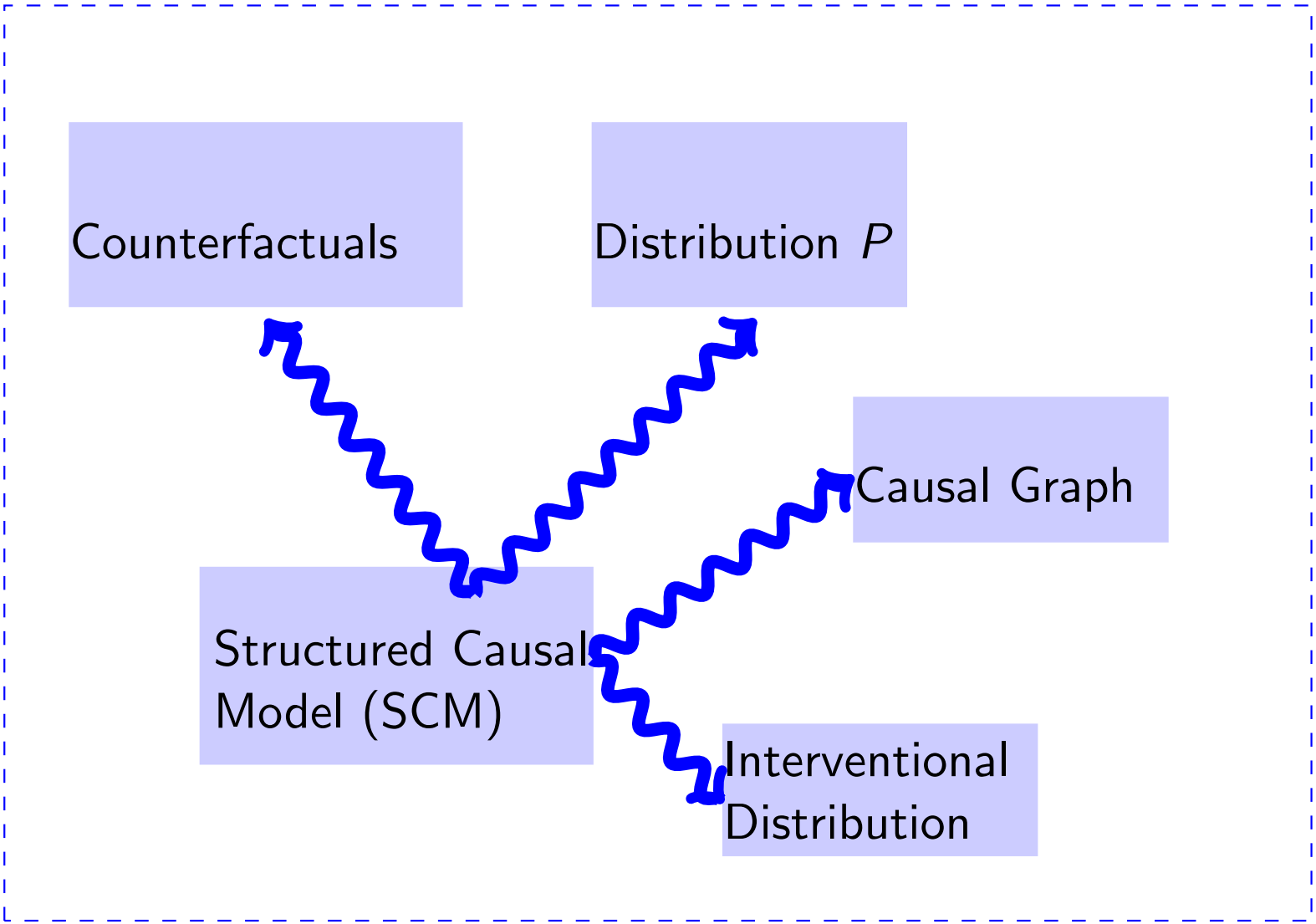
Challenges to Learning from data

- Independent Identically Distributed (iid) sample assumption
- Observational data vs. Experimental data vs. Interventional data
- Challenges to randomized experiments
- ATE, ITE
- Counterfactuals, bias, propensity score

Causal Language and Reasoning

- Counterfactuals
- Distributions (Probability)
- Causal Graphs
- Structured Causal Models (SCM)
- Interventional distributions

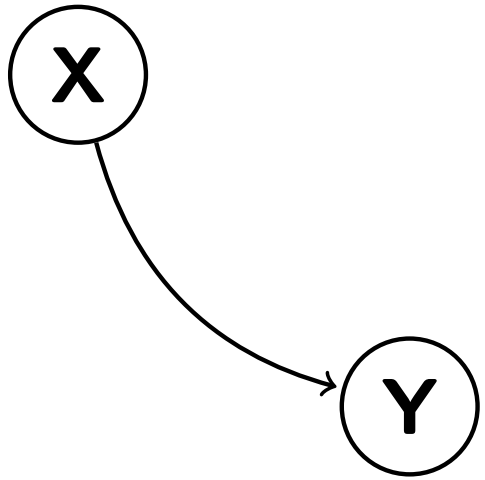
Causal Language and Reasoning



Structured Causal Model

SCM ($\boldsymbol{S}$, $\boldsymbol{P}^N$) models observational distributions

$$X := N_x$$
$$Y := \alpha X + N_y$$
$$N_x, N_y \overset{iid}{\sim} \mathcal{N}(0,1)$$

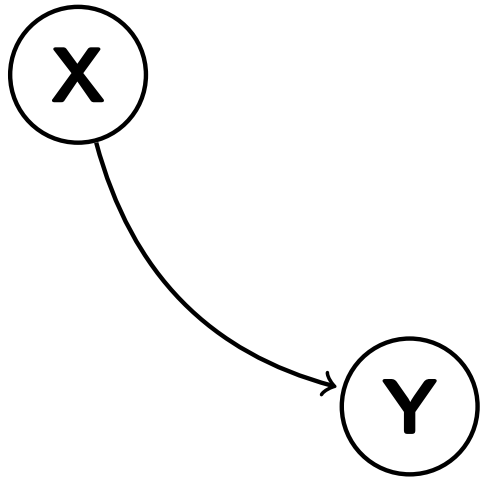


$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \alpha \\ \alpha & \alpha^2 + 1^2 \end{pmatrix} \right)$$

Modeling Interventions

SCM ($\boldsymbol{S}$, $\boldsymbol{P}^N$) models observational distributions

$$X := N_x \qquad X := 7$$
$$Y := \alpha X + N_y$$
$$N_x, N_y \overset{iid}{\sim} \mathcal{N}(0, 1)$$

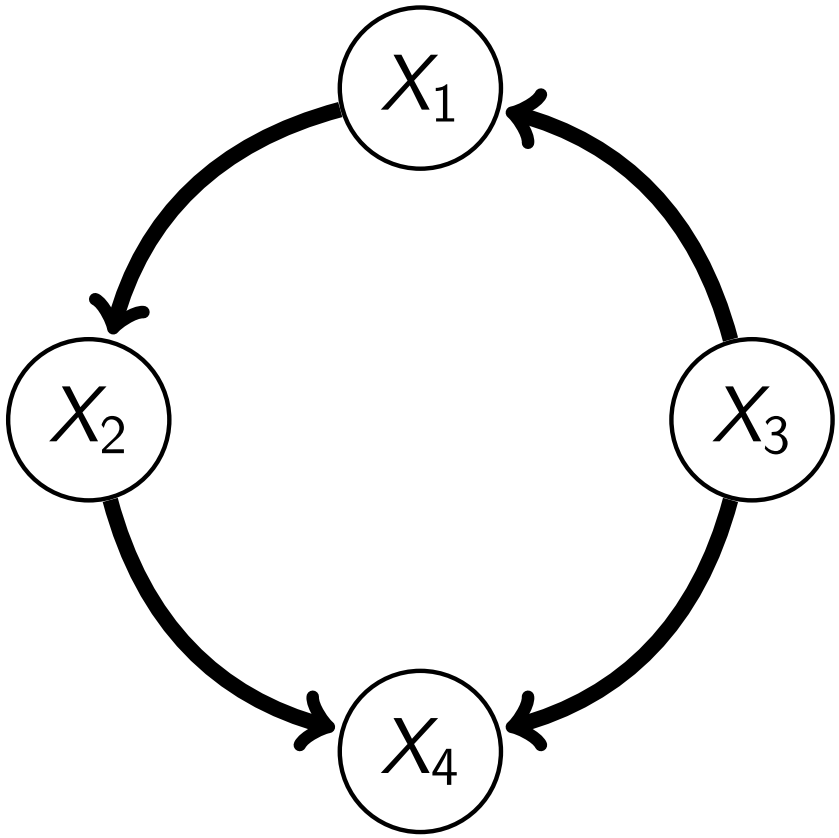


$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 7 \\ 7\alpha \end{pmatrix}, \begin{pmatrix} 0 & 49\alpha \\ 49\alpha & 1 \end{pmatrix} \right)$$

Structural Equations with Noise Distribution

SCM $(\boldsymbol{S}, \boldsymbol{P}^N)$

$$
\begin{aligned}
X_1 &:= f_1(X_3, N_1) \\
X_2 &:= f_2(X_1, N_2) \\
X_3 &:= f_3(N_3) \\
X_4 &:= f_4(X_2, X_3, N_4)
\end{aligned}
$$



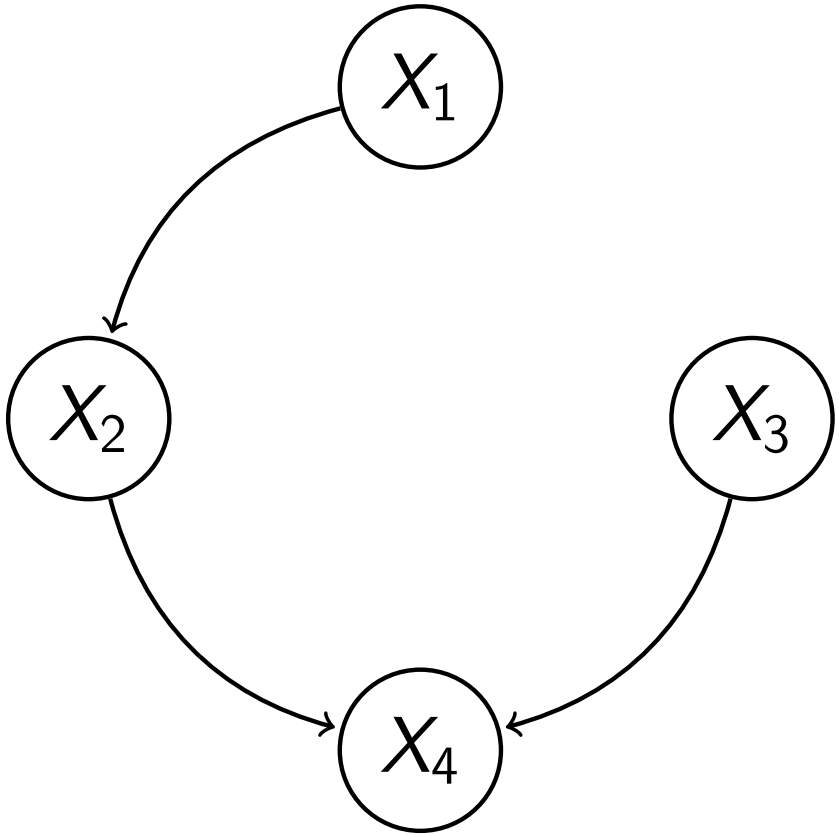
- $N_i \overset{iid}{\sim} \mathcal{N}$
- G has no cycles

- Acyclic
- Start with the source nodes and follow them to other nodes

SCM modeling interventions

SCM $(\boldsymbol{S}, \boldsymbol{P}^N)$. $P_{do(X_1:=0)} \neq P(.|X_1)$

$$X_1 := 0$$
$$X_2 := f_2(X_1, N_2)$$
$$X_3 := f_3(N_3)$$
$$X_4 := f_4(X_2, X_3, N_4)$$



- $N_i \overset{iid}{\sim} \mathcal{N}$
- G has no cycles

- Interventions induce a new distribution
- **Intervention is not the same as conditioning**

Prerequisites: Intuition

- Causal Model needs to entail a joint distribution on the random variables, which is possible when we do not have cycles in the relationship graph
- Acyclicity allows us to start sampling from the source nodes (nodes without any ancestors) and the work our way through the child nodes to draw samples
- Even when there are cycles in the graph of random variables, it is often possible to arrive at a joint distribution from which one could sample, if the system would arrive at a stable distribution after numerous iterations of sampling by arbitrarily starting at any node and working our way through it's descendants

Causal Language and Reasoning

Interventions: Intuition

- Interventions in an SCM are modifications of the original SCM
- Intervened random variables loose their incoming edges in the graph
- Interventions ($do(T := A)$) are different from conditioning ($(.|T = A)$ i.e.

$$P_{do(T:=A)}(.) \neq P(.|T = A)$$

Causal Language and Reasoning

Rules of Causal Calculus: Prerequisites

- G is a graph on the random variables
- X, Y, Z, W are sets of random variables
- New operator $do(.)$ for "doing", with the meaning of surgically setting values of variables
- $G_{\bar{X}}$ is the graph G with incoming edges to X removed
- $G_{\underline{X}}$ is the graph G with outgoing edges from X removed

Causal Language and Reasoning

Rules of Causal Calculus

- **Rule 1: Ignoring observations**

$$P(y|do\{x\}, z, w) = P(y|do\{x\}, w) \quad \textbf{if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\bar{X}}}$$

- **Rule 2: Action/Observation exchange**

$$P(y|do\{x\}, do\{z\}, w) = P(y|do\{x\}, z, w) \quad \textbf{if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\bar{X}\underline{Z}}}$$

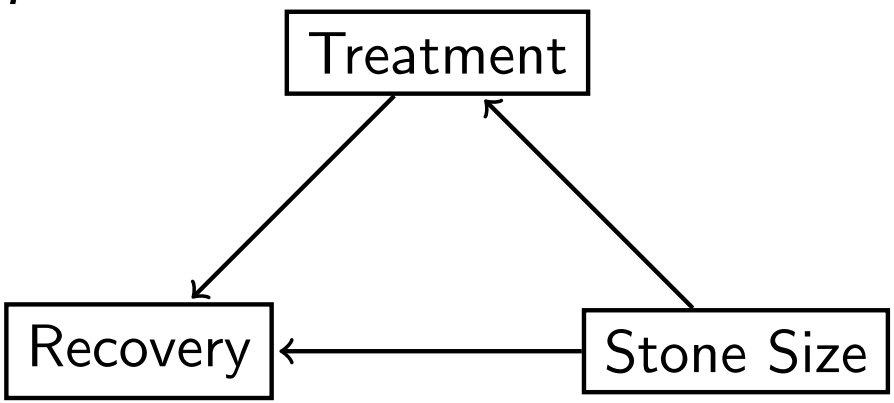
- **Rule 3: Ignoring actions**

$$P(y|do\{x\}, do\{z\}, w) = P(y|do\{x\}, w) \quad \textbf{if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\bar{X}\bar{Z}(\bar{W})}}$$

Example: Kidney Stones

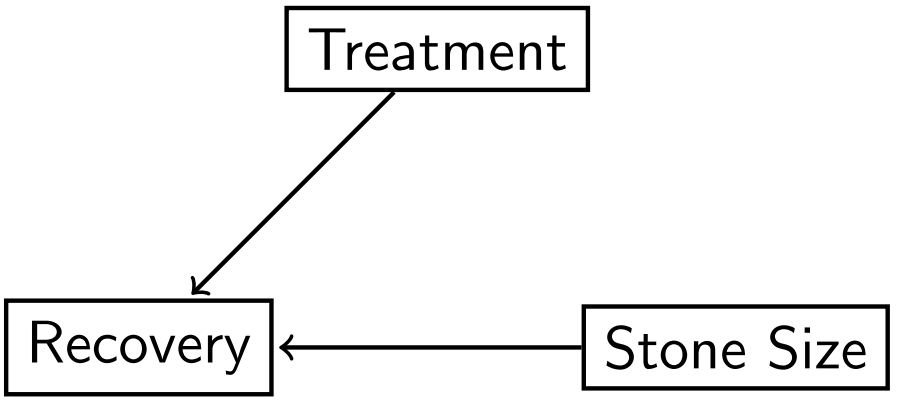
	Treament A	Treatment B
Small Stones $\frac{352}{700} = 0.51$	$\frac{81}{87} = 0.93$	$\frac{234}{270} = 0.87$
Large Stones $\frac{343}{700} = 0.49$	$\frac{192}{263} = 0.73$	$\frac{55}{80} = 0.69$
	$\frac{273}{350} = 0.78$	$\frac{289}{350} = 0.83$

P



$P(R, T, S) = P(S)P(T|S)P(R|T, S)$

$P_{do(T:=A)}$



Calculate $P_{do(T:=A)}(R = 1)$

Example: Kidney Stones

$$
\begin{aligned}
P_{do(T:=A)}(R=1) &= \sum_s P_{do(T:=A)}(R=1, S=s, T=A) \\
&= \sum_s P_{do(T:=A)}(R=1|S=s, T=A) P_{do(T:=A)}(S=s, T=A) \\
&= \sum_s P_{do(T:=A)}(R=1|S=s, T=A) P_{do(T:=A)}(S=s) \\
&= \sum_s P(R=1|S=s, T=A) P(S=s) \\
&= .51 \times 0.93 + 0.49 \times 0.73 \\
&= 0.832 \qquad \neq P(R=1|T=A) \\
&> 0.78 \approx P_{do(T:=B)}(R=1)
\end{aligned}
$$

Take home message from the example - 1

Remember, this is observational data as opposed to a randomized control trial (RCT) and it should be surprising to the reader that it is even possible to learn the relationship between the random variables (in this case, T (treatment), and R (recovery)) from simple observational data. In this case we have derived the probability of recovery, when treated with A, B.

Take home message from the example - 2

In a more general setting we would learn the causal effect - the change in expected value of the outcome variable to a change in the treatment variable. The limitations and bounds on the estimates of causal effect is a major part of causal inference research. The keen reader would have noticed that the above method was non-parametric - that is we did not learn the functional form of the relationship between treatment variable and outcome variable, only it's probability (by extension its expected value). If the input X and output Y are confounded by Z, then the expected value of outcome under treatment is:

$E_{do(X)}(Y) = \sum_Z Y * P(Y|Z,X) * P(Z)$

Take home message from the example - 3

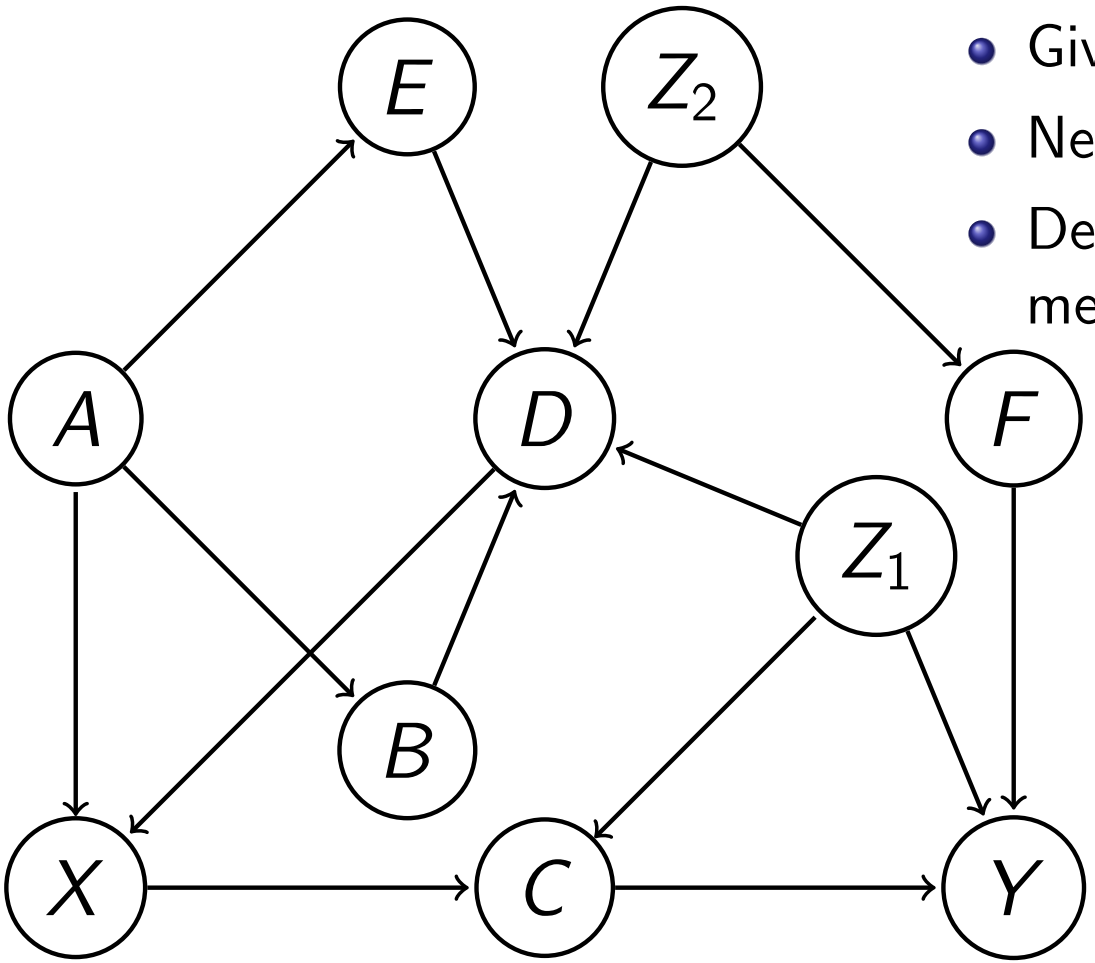
This example also demonstrates Simpson's paradox. Overall, it would appear that treatment A is a poorer choice, success rate of .78 vs .83. However, if we condition on the kidney stone size, we see that treatment A is more successful (.93 vs .87 for small stone, and .73 vs .69 for large stone). This paradox is known as Simpson's paradox, whose practical counterpart is the adjustment problem or covariate selection problem. This problem of identifying the variables to be measured (also called controlled) is now easier if we know their graphical relationship, and in fact can be performed mechanistically, as shown below.

Causal Language and Reasoning

Adjustment Set

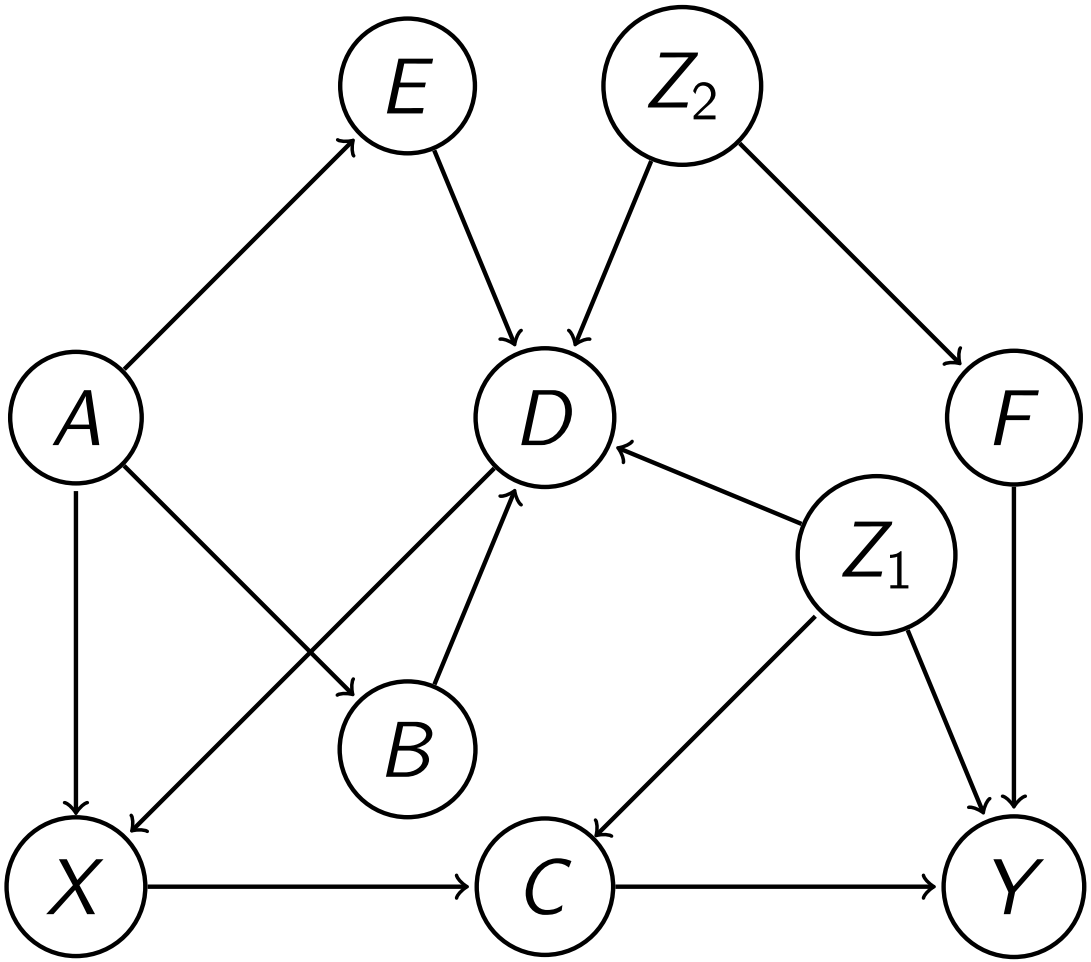
The natural next step is to explore the structure of graphs and the random variables that will need to be controlled for in order to allow computation of the causal effect of a random variable, X on the outcome of interest, Y. The set of random variables which when controlled for (measured), allow learning this relationship are called a valid adjustment set. Note that there could be many valid adjustment sets; it is not a unique set by any means. Even in a rather complicated graph, one could answer these questions mechanistically.

Valid Adjustment Set: An example



- Given: Causal Graph
- Needed: Effect of X on Y
- Decision: Random variables to measure

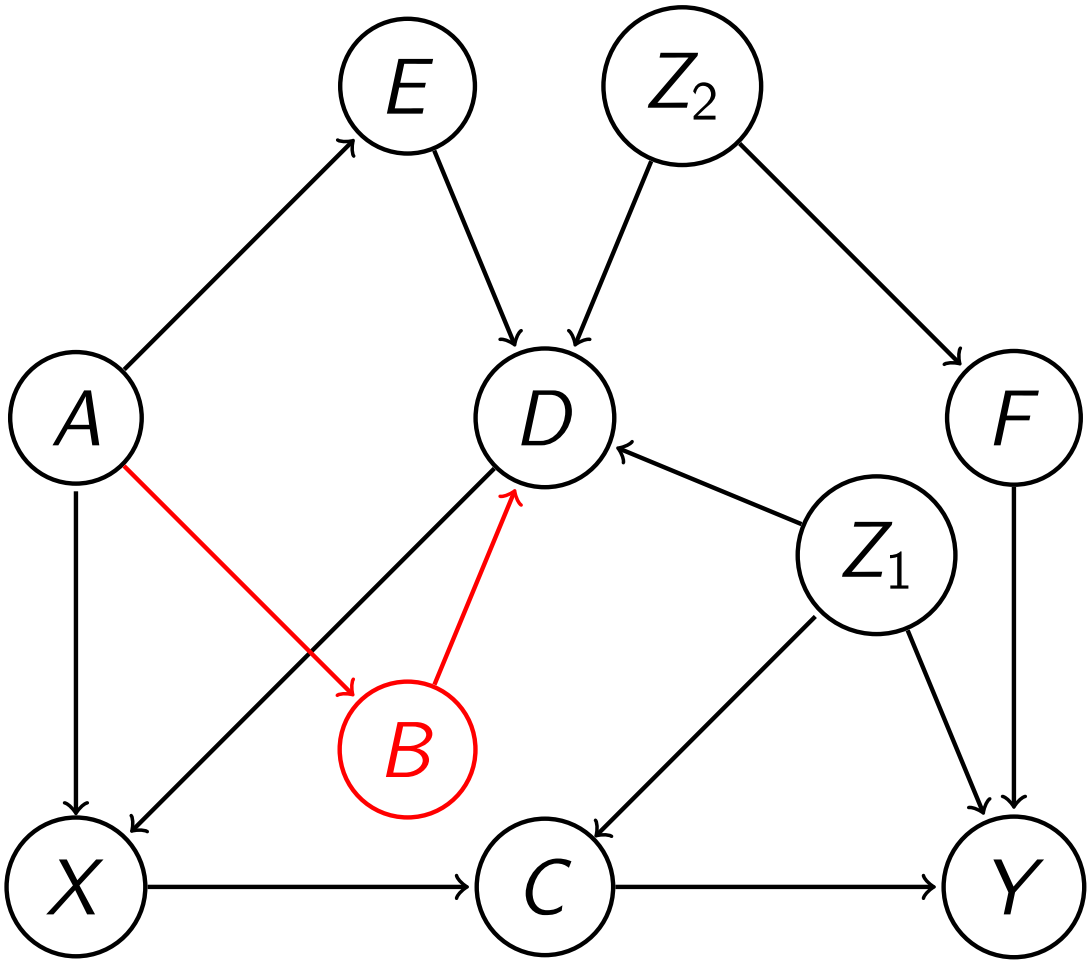
Valid Adjustment Set: Step 1



Test if Z_1 and Z_2 are a sufficient measurement set

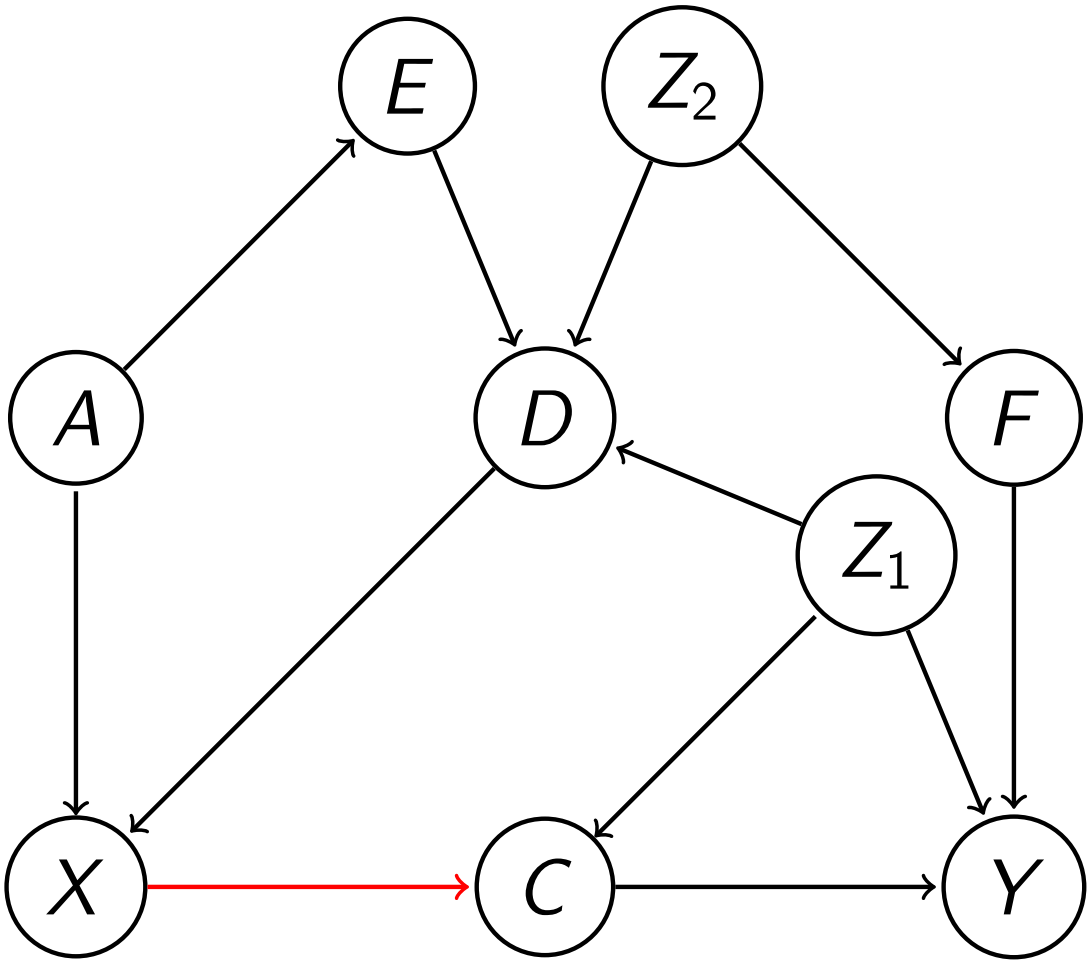
Step 1: Z_1 and Z_2 should not be descendants of X

Valid Adjustment Set: Step 2



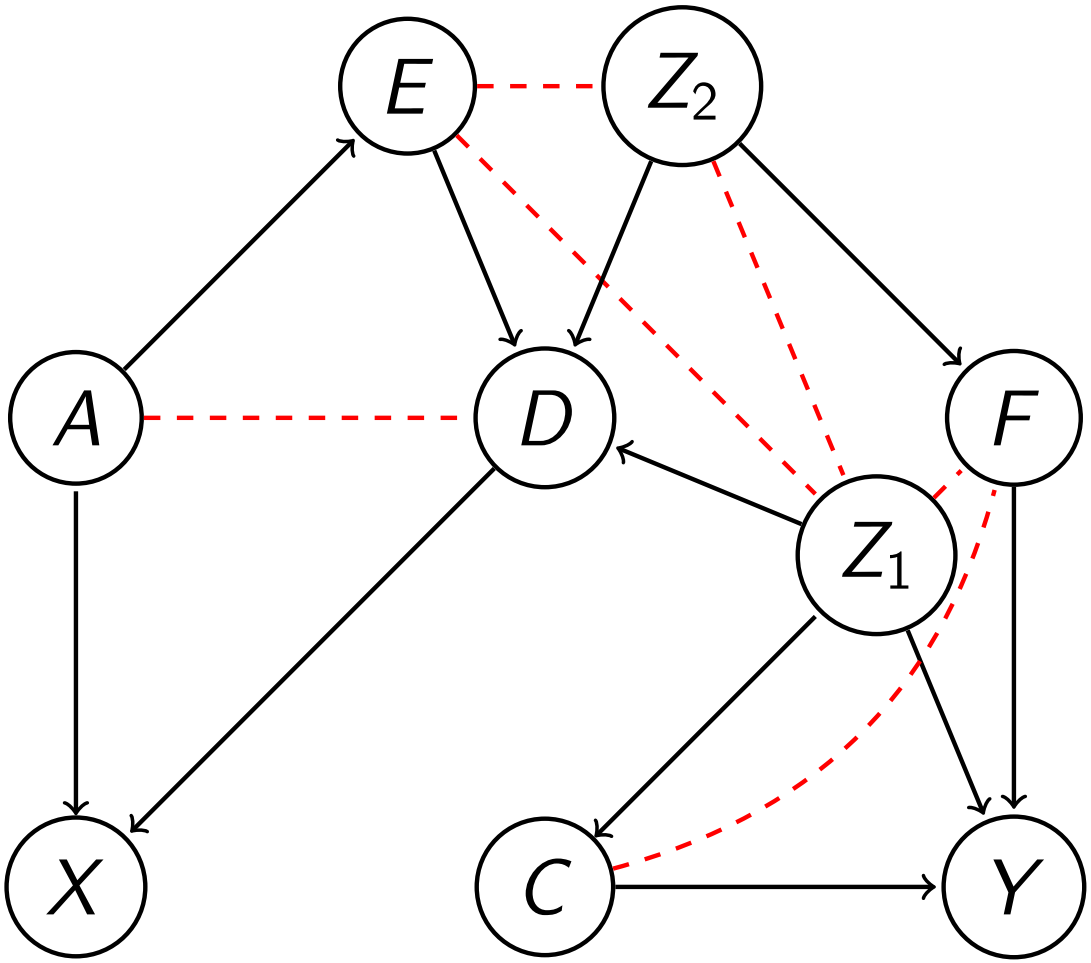
Step 2: Delete all non-ancestors of X, Y, Z_1, Z_2

Valid Adjustment Set: Step 3



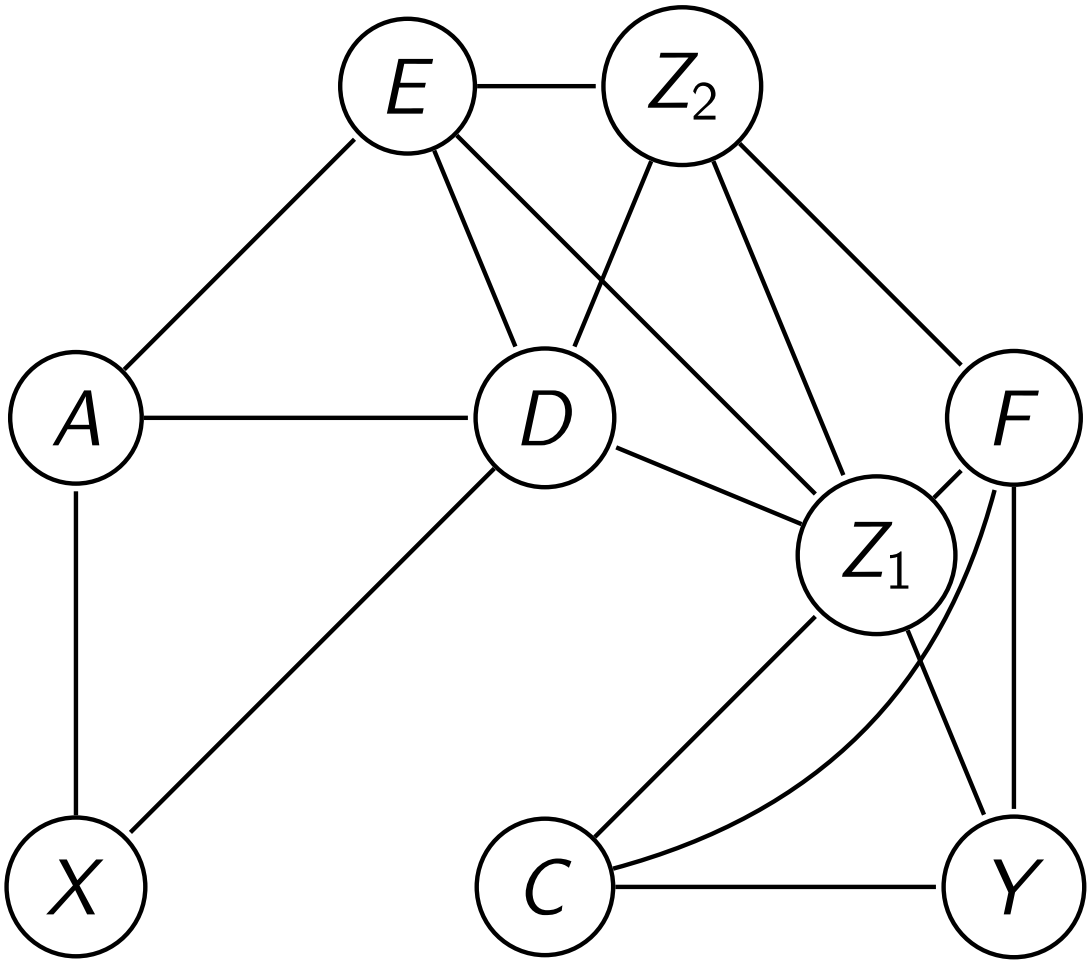
Step 3: Delete all arcs emanating from X

Valid Adjustment Set: Step 4



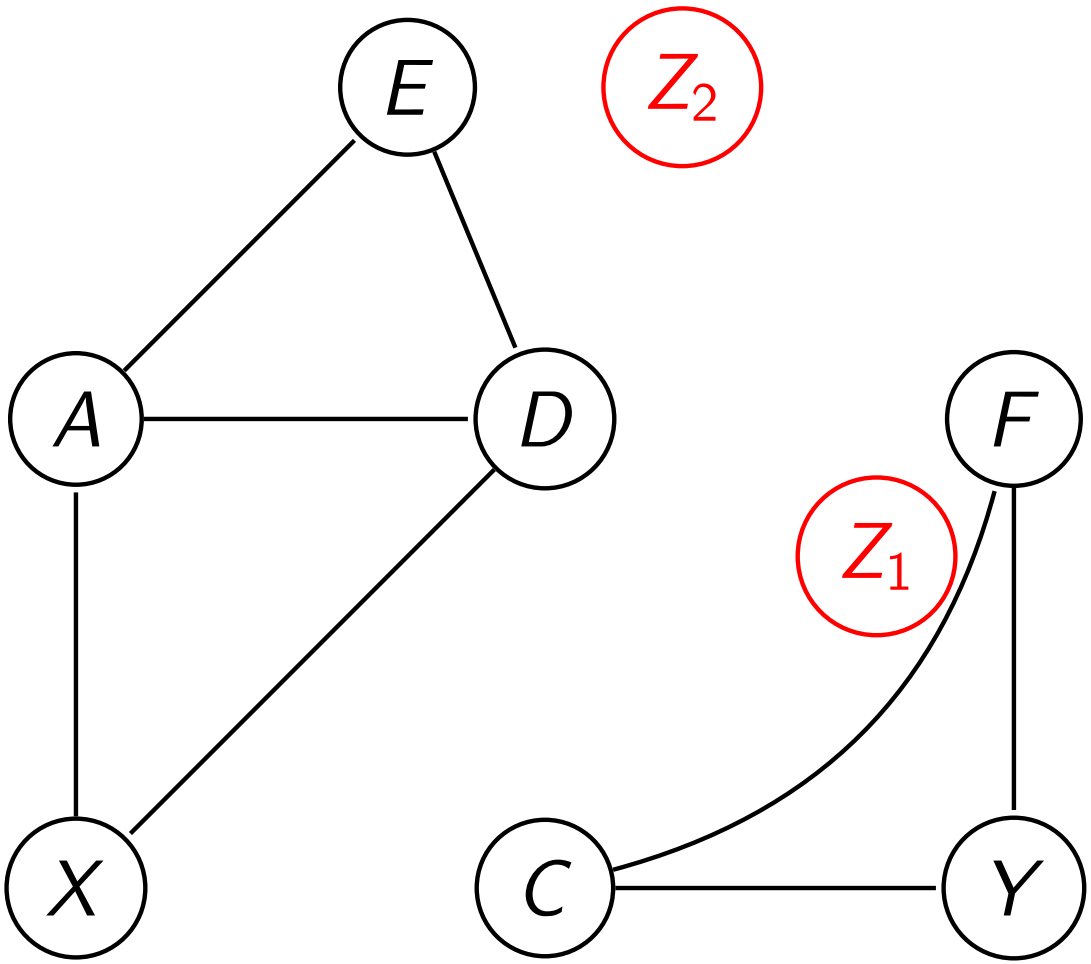
Step 4: Connect any two parents sharing a common child

Valid Adjustment Set: Step 5



Step 5: Strip arrow heads from all edges

Valid Adjustment Set: Step 6



Step 6: Delete Z_1 and Z_2

Z_1 and Z_2 are sufficient measurements if X and Y are disconnected

Causal Language and Reasoning

Subtleties

- Causal Markov Condition (CMC): Conditional independences of the Probability distribution can be represented as d-separatedness of the Graphical model
- Faithfulness: vice-versa, that is the d-separation in the Graph translates to conditional independences in the Probability distribution

Causal Language and Reasoning

Neyman-Rubin Potential Outcome framework

- Uses counter factuals to estimate Average Treament Effect (ATE)
- Assumptions involve: Consistency, Ignorability, Positivity
- Has been shown to have direct relation to SCM

Questions

Thank you