TECHNISCHE UNIVERSITÄT DARMSTADT

AIML Lab

Winter Semester 2025/26 Lecture

Causality for AI & ML

"Structural Causal Models"

Prof. Dr. Kristian Kersting

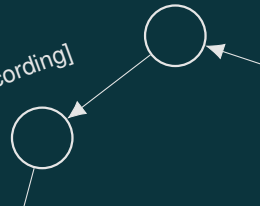
Moritz Willig — Today's speaker

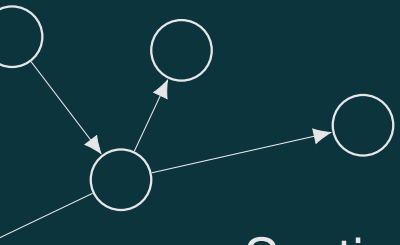
Tim Woydt

Florian Busch

Matej Zečević

[Rerecording]

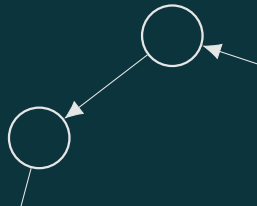




Section

0

Recap PGM



Distributions - Recap

Joint probability $P(X, Y)$

Conditional probability $P(X|Y) := \frac{P(X,Y)}{P(Y)}$ (for $P(Y) > 0$)

Chain Rule $P(X, Y) = P(X)P(Y|X) = P(Y)P(X|Y)$ ($n!$ possibilities)

Marginalization $P(X) = \sum_y P(X, Y = y)$

Bayes' rule $P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$

(Marginal) Independence $X \perp Y :\Leftrightarrow P(X|Y) = P(X) \Leftrightarrow P(X, Y) = P(X)P(Y)$

Conditional Independence $X \perp Y \mid Z :\Leftrightarrow P(X|Y, Z) = P(X|Z)$

Local Markov Assumption & I-maps - Recap

The DAG $\mathcal{G}$ represents a factorization of the joint probability distribution $P(\mathcal{V})$ under Local Markov Assumption, i.e.

$$X \perp Y \mid pa(X) \text{ for all } Y \notin de(X)$$

with $pa(X)$ denoting the set of parents and $de(X)$ the set of descendants of X with respect to the DAG $\mathcal{G}$.

$\mathcal{G}$ is called an independence map (I-map) of P if P factorizes over $\mathcal{G}$.

Up to its parameters, a BN can model any joint distribution that factorizes like

$$P(\mathcal{V}) = \prod_{X \in \mathcal{V}} P(X|pa(X))$$

d-separation - Recap

For a DAG $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ and $\mathbf{X}, \mathbf{Y}, W \subseteq V$ we say that $\mathbf{X}$ and $\mathbf{Y}$ are d-separated given W ($d_{\mathcal{G}}(\mathbf{X}, \mathbf{Y} | W) = 1$) if there is no active trail between any $X \in \mathbf{X}$ and any $Y \in \mathbf{Y}$ while observing W.

A trail is an undirected path in $\mathcal{G}$ that never visits a node twice and is called active while observing W if for each consecutive triplet $X - Z - Y$ one of the following holds: (We can think of active trails as information flow)

(a) $X \to Z \to Y$ (chain) and $Z \notin W$ (Y unobserved)

(b) $X \leftarrow Z \leftarrow Y$ (chain) and $Z \notin W$ (Y unobserved)

(c) $X \leftarrow Z \to Y$ (fork) and $Z \notin W$ (Y unobserved)

(d) $X \to Z \leftarrow Y$ (collider/v-structure) and $Z \in W$ or $de(Z) \cap W \neq \emptyset$ (Y or some descendent observed)

Probabilities & Bayesian Networks - Exercises

1. Show for independent events A, B, C that
   a) $P(A \vee B) = P(A) + (B) - P(A)P(B)$
   b) and similarly calculate $P(A \vee B \vee C)$.

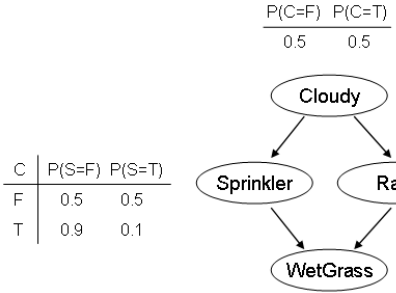
2. Show for any (set of) random variables Z that the following three conditions for $X \perp Y \mid Z$ are equivalent:
   i) $P(X, Y|Z) = P(X|Z)P(Y|Z)$
   ii) $P(X|Y, Z) = P(X|Z)$
   iii) $P(Y|X, Z) = P(Y|Z)$

3. Use the Local Marcov assumption to show that for the DAG $\langle\{V, W, X, Y, Z\}, \{(V, X), (W, X), (X, Y), (X, Z)\}\rangle$ the joint distribution factorizes as
$$P(V, W, X, Y, Z) = P(Z|X)P(Y|X)P(P|V, W)P(V)P(W).$$

Probabilities & Bayesian Networks - Exercises

4. Consider the water sprinkler Bayes net with binary nodes.

   a) Provide a minimal factorization for $P(C, S, R, W)$ w.r.t. the DAG.
   b) Compute $P(S = t | W = t)$.
   c) Compute $P(W = t)$.



P(C=F)	P(C=T)
0.5	0.5

C	P(S=F)	P(S=T)
F	0.5	0.5
T	0.9	0.1

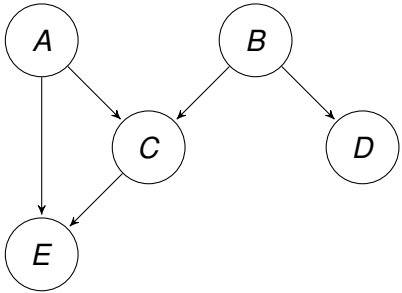
C	P(R=F)	P(R=T)
F	0.8	0.2
T	0.2	0.8

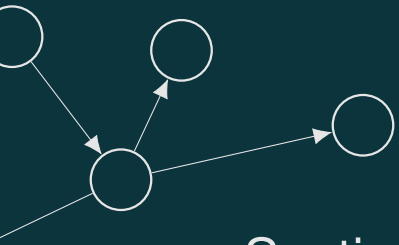
S R	P(W=F)	P(W=T)
F F	1.0	0.0
T F	0.1	0.9
F T	0.1	0.9
T T	0.01	0.99

Probabilities & Bayesian Networks - Exercises

5. Apply d-separation to determine which of the following conditional independencies hold for the DAG. For those that do not hold, name an active trail between the nodes.

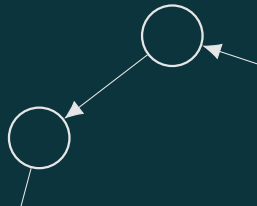
   a) $A \perp D \mid B, C$
   b) $A \perp B \mid E$
   c) $E \perp D \mid C$





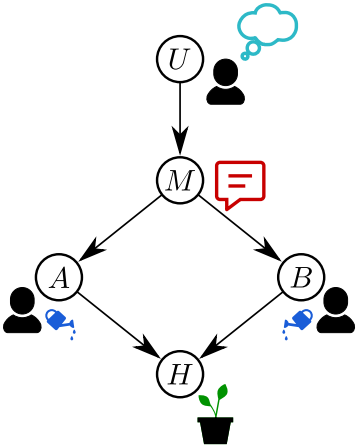
Section

1 Structural Causal Models



Causal Modeling

"Tom goes on vacation. Upon remembering that his plant needs to be taken care off while he is away, he sends a message his two friends. In case that the message is sent, both friends will take care of the flower."



Causal Modeling

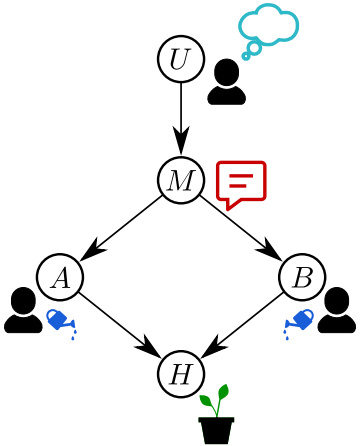
"Tom goes on vacation. Upon remembering that his plant needs to be taken care off while he is away, he sends a message his two friends. In case that the message is sent, both friends will take care of the flower."

$U \in \mathbb{B}$: Tom remembers to send a message.

$M \in \mathbb{B}$: Message is sent.

$A, B \in \mathbb{B}$: Friends read the message.

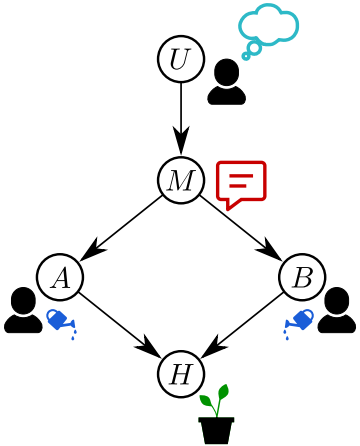
$H \in \mathbb{B}$: Plant is healthy.



Causal Modeling

"Tom goes on vacation. Upon remembering that his plant needs to be taken care off while he is away, he sends a message his two friends. In case that the message is sent, both friends will take care of the flower."

$$
\begin{aligned}
U &:= \text{Bernoulli}(0.5) \\
M &:= U \\
A, B &:= M \\
H &:= A \vee B
\end{aligned}
$$



Structural Causal Model

A **Structural Causal Model** (SCM) is a tuple $\mathcal{M} = (\mathbf{V}, \mathbf{U}, \mathbf{F}, \mathcal{P}_{\mathbf{U}})$.

- $\mathbf{V}$ Set of Endogenous Variables.
- $\mathbf{U}$ Set of Exogenous Variables.
- $\mathbf{F}$ Structural Equations; $x_i := f_i(\text{pa}(x_i))$.
- $\mathcal{P}_{\mathbf{U}}$ Distribution of Exogenous Variables.

Structural Causal Model

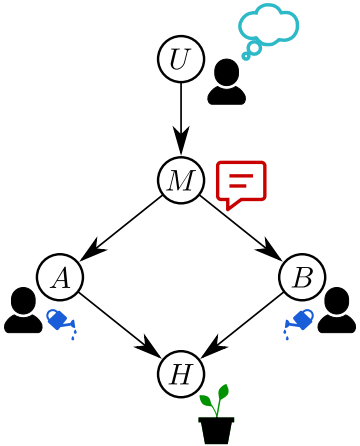
A **Structural Causal Model** (SCM) is a tuple $\mathcal{M} = (\mathbf{V}, \mathbf{U}, \mathbf{F}, \mathcal{P}_{\mathbf{U}})$.

- $\mathbf{V}$ Set of Endogenous Variables.
- $\mathbf{U}$ Set of Exogenous Variables.
- $\mathbf{F}$ Structural Equations; $x_i := f_i(\text{pa}(x_i))$.
- $\mathcal{P}_{\mathbf{U}}$ Distribution of Exogenous Variables.

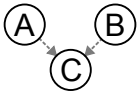
  - SCM induce a *directed acyclic graph* (DAG) $\mathcal{G}$ with vertices $\mathbf{X}$ and edges $\text{pa}(x_i) \to x_i$.
    - $\mathbf{X}$ is the set of all variables: $\mathbf{X} = \mathbf{V} \cup \mathbf{U}$
    - $\text{pa}(x_i)$ denotes the parents of x_i.

Flowering Plant SCM

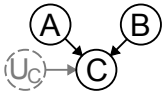
$$\mathcal{M} = \begin{cases} \mathbf{V} & = \{M, A, B, H \in \mathbb{B}\} \\ \mathbf{U} & = \{U \in \mathbb{B}\} \\ \mathbf{F} & = \begin{cases} f_M & := U \\ f_A & := M \\ f_B & := M \\ f_H & := A \vee B \end{cases} \\ \mathcal{P}_{\mathbf{U}} & = \{U = \text{Bernoulli}(0.5)\} \end{cases}$$



From Conditional Distributions to Equations



P(C=1)	B	A
0.9	0	0
0.8	0	1
0.4	1	0
1.0	1	1



C	U_C	B	A
1	<0.9	0	0
1	<0.8	0	1
1	<0.4	1	0
1	<1.0	1	1

Left is the conditional probability table $P(C|A,B)$ of the above graph *(with P(C=0) omitted)*. Notice the inherent uncertainty within the table.

We can make the noise explicit (right table) and write it as follows:

$$C := (\neg B \wedge \neg A \wedge U_C < 0.9) \vee (\neg B \wedge A \wedge U_C < 0.8) \vee \ldots$$

$$\text{where } U_C := \mathit{Uniform}[0,1)$$

The inverse way of structural equation to conditional distribution can be achieved via applying pushforward measure on the structural equations.

Factorization of the Joint Distribution

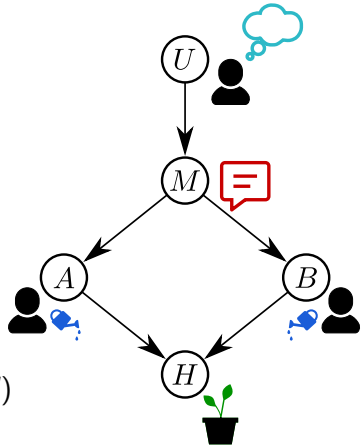
An SCM $\mathcal{M}$ entails a joint distribution $P_{\mathcal{M}}(X_1, \ldots, X_N)$ over all variables $X_1, \ldots, X_N$ that **factorizes according to the causal graph** $\mathcal{G}$:

$$P_{\mathcal{M}}(X_1, \ldots, X_N) = \prod_{i \in \{1..N\}} P(X_i | \operatorname{pa}(X_i))$$

Factorization of the Joint Distribution

An SCM $\mathcal{M}$ entails a joint distribution $P_{\mathcal{M}}(X_1, \ldots, X_N)$ over all variables $X_1, \ldots, X_N$ that **factorizes according to the causal graph** $\mathcal{G}$:

$$P_{\mathcal{M}}(X_1, \ldots, X_N) = \prod_{i \in \{1..N\}} P(X_i | \text{pa}(X_i))$$



Example:

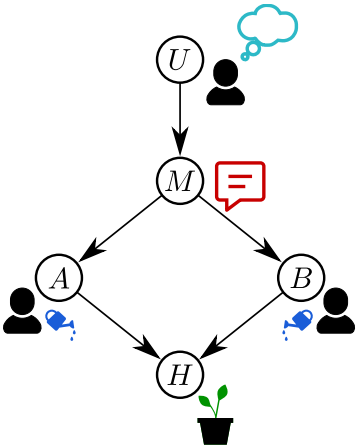
$$P(U, M, A, B, H) = P(H|A, B)P(A|M)P(B|M)P(M|U)P(U)$$

Independent Mechanisms Principle

Independent Mechanisms Principle

Changing one $p(X_i|\,\mathrm{pa}(i))$ does not influence any other $p(X_j|\,\mathrm{pa}(j))$ for $j \neq i$.

"Mechanisms do not inform each other"



Changing the way friend A receives the message has no influence on how friend B receives to the message.

Markovianity and Faithfulness

Markovianity: $P_{\mathcal{M}}$ is called Markovian to $\mathcal{G}_{\mathcal{M}}$ if all independencies implied by the graph also hold true in the distribution:

Markov Condition

$$(X \perp\!\!\!\perp Y|Z)_G \Rightarrow (X \perp\!\!\!\perp Y|Z)_P$$

Faithful: $\mathcal{G}_{\mathcal{M}}$ is called faithful to $P_{\mathcal{M}}$ if all independencies implied by the distribution also hold true in the graph:

Faithfulness

$$(X \perp\!\!\!\perp Y|Z)_G \Leftarrow (X \perp\!\!\!\perp Y|Z)_P$$

Independence Maps

I-Map: If a distribution is Faithful to a graph, the graph is a *I-map* (Independency map). It contains at least all the dependencies of the graph, *but might contain more*!

D-Map: If a distribution is Markovian to a graph, the graph is an *D-map* (Dependency map). It lists at least all the independencies of the distribution, *but might contain more*!

If Markovianity and Faithfulness are met it is called a *P-Map* (Perfect map):

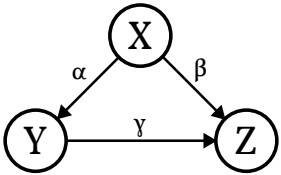
P-Map

$$(X \perp\!\!\!\perp Y|Z)_G \Leftrightarrow (X \perp\!\!\!\perp Y|Z)_P$$

The set of independencies in the graph is exactly identical to the set of independencies in the distribution.

Non-Faithful Graphs

Consider the following graph:



with structural equations:

$$X := U_X \qquad Y := \alpha X \qquad Z := \beta X + \gamma Y$$

Substituting equations: $Z := \beta X + \gamma Y = \beta X + \gamma\alpha X = (\beta + \gamma\alpha)X$

Whenever $\beta + \alpha\gamma = 0 \Rightarrow (X \perp\!\!\!\perp Z)_P$.
However, the graph contains the edge $X \to Z$ therefore $(X \not\perp\!\!\!\perp Z)_\mathcal{G}$ and $(X \perp\!\!\!\perp Z)_\mathcal{G} \neq (X \perp\!\!\!\perp Z)_P$!

Additive Noise Models (ANMs)

General causal models allow for arbitrary types of structural equations.
Additive noise models make particular assumptions on the type of equations:

$$f_i = \sum_{X_j \in \mathrm{pa}(x_i)} \alpha_{ij} x_j + \epsilon_i$$

with all $\epsilon_i \in \mathbf{U}$ and all ϵ_i mutually independent.

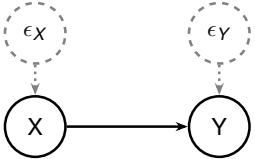
Additive Noise Models (ANMs)

General causal models allow for arbitrary types of structural equations.
Additive noise models make particular assumptions on the type of equations:

$$f_i = \sum_{X_j \in \mathrm{pa}(X_i)} \alpha_{ij} X_j + \epsilon_i$$

with all $\epsilon_i \in \mathbf{U}$ and all ϵ_i mutually independent.

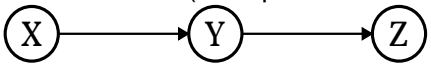
Implies separate independent noise terms per variable:



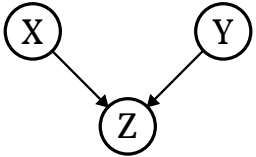
Peters, J., Mooij, J.M., Janzing, D. and Schölkopf, B., 2014. Causal discovery with continuous additive noise models. The Journal of Machine Learning Research, 15(1), pp.2009-2053.

Common Graph Substructures

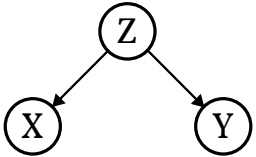
Causal Chain (a sequence of nodes connected in the same causal direction):



Collider/v-structure (a variable being affected by two or more variables):

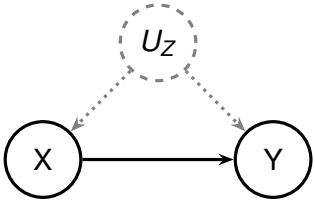


Fork/Confounder (a variable simultaneously affecting others):



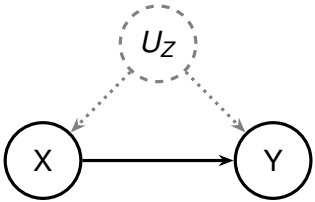
Hidden Confounding

Sometimes noise terms affect multiple variables:

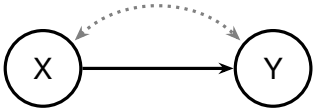


Hidden Confounding

Sometimes noise terms affect multiple variables:

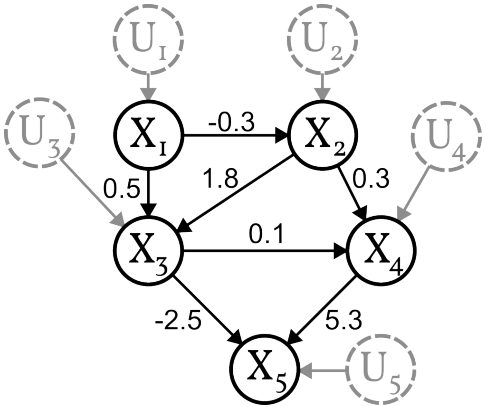


…occasionally drawn without the latent term:



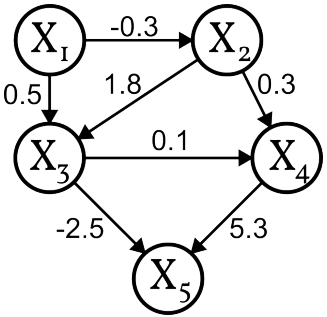
Additive Noise Model Example

ANM representation with weights at the edges:



Additive Noise Model Example

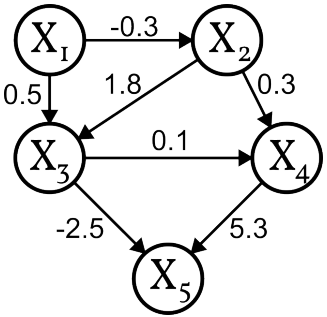
ANM representation with weights at the edges:



Independent noise variables are often omitted.

Additive Noise Model Example

ANM representation with weights at the edges:

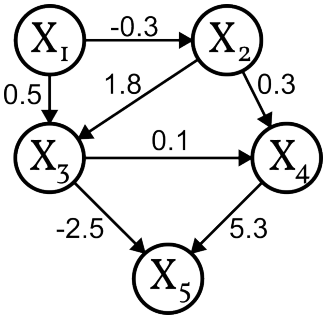


$$W = \begin{pmatrix} 0 & -0.3 & 0.5 & 0 & 0 \\ 0 & 0 & 1.8 & 0.3 & 0 \\ 0 & 0 & 0 & 0.1 & -2.5 \\ 0 & 0 & 0 & 0 & 5.3 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Easy to handle matrix representation.

Additive Noise Model Example

ANM representation with weights at the edges:



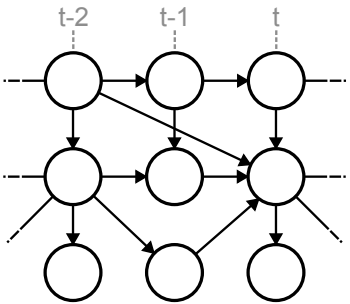
$$W = \begin{pmatrix} 0 & -0.3 & 0.5 & 0 & 0 \\ 0 & 0 & 1.8 & 0.3 & 0 \\ 0 & 0 & 0 & 0.1 & -2.5 \\ 0 & 0 & 0 & 0 & 5.3 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Easy to handle matrix representation.

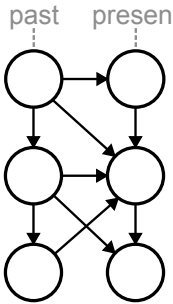
Remark: For every DAG there exists a permutation of the variable order $\pi(\mathbf{V})$ such that W becomes *strictly upper triangular*.

Causal Time Series Models

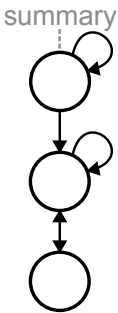
Full time causal graph



Extended Summary Causal Graph

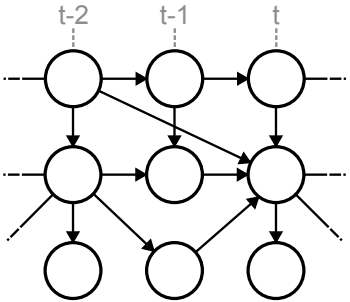


Summary Causal Graph

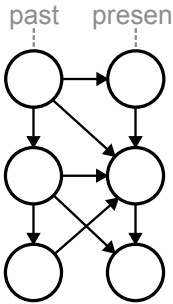


Causal Time Series Models

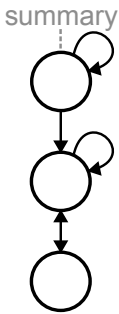
Full time causal graph



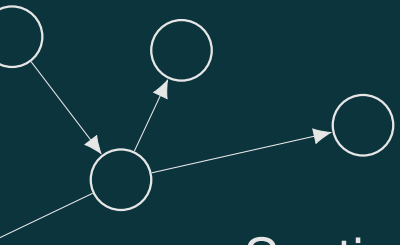
Extended Summary Causal Graph



Summary Causal Graph

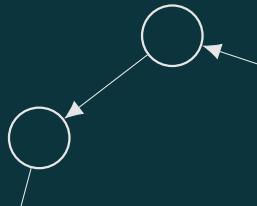


Can become cyclic!



Section 2

Identifiability of Causal Effects



Causal Inference

Task of Causal Inference: Can we answer a causal **query** $P(y|\text{do(x)})$, given the **causal graph** $\mathcal{G}$ and **observational data x**?

"What is the probability of the outcome $Y = y$ if I do set $X = x$?"

Causal Inference

Task of Causal Inference: Can we answer a causal **query** $P(y|\text{do(x)})$, given the **causal graph** $\mathcal{G}$ and **observational data x**?

"What is the probability of the outcome $Y = y$ if I do set $X = x$?"

Causal Inference

Task of Causal Inference: Can we answer a causal **query** $P(y|\text{do(x)})$, given the **causal graph** $\mathcal{G}$ and **observational data x**?

"What is the probability of the outcome $Y = y$ if I do set $X = x$?"

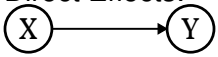
Average Treatment Effect (for binary outcome scenarios):

$$\text{ATE} = \mathbb{E}[P(y|do(X = 1)) - P(y|do(X = 0))]$$

The ATE is the *expected difference in outcome* that would result from, e.g., treating an individual compared to not treating them.

Inference on Simple Graphs I

Direct Effects:



Changing X does directly influence Y. No other factors are involved. While not having formally defined the do-operator, we assume:

$$P(y|do(x)) = P(y|x)$$

Inference on Simple Graphs II

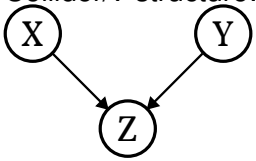
Causal Chain:

We can marginalize out Z to arrive at the previous scenario:

$$
\begin{aligned}
P(y|do(x)) &= \sum_z P(y, z|do(x)) \\
&= \sum_z P(y|z, do(x))P(z|do(x)) \\
&= \sum_z P(y|z, x)P(z|x) \\
&= \sum_z P(y, z|x) = P(y|x)
\end{aligned}
$$

Inference on Simple Graphs III

Collider/v-structure:



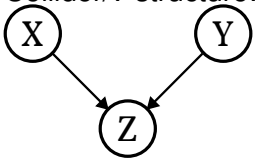
X is independent of Y:

$$P(y|do(x)) = P(y|x) = P(y)$$

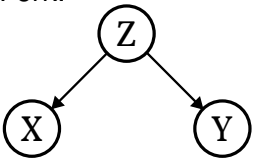
General Condition: Y must be a child (or generally an anchestor) of X.
If not fulfilled, there is no causal effect. The ATE is zero!

Inference on Simple Graphs III

Collider/v-structure:



Fork:



X is independent of Y:

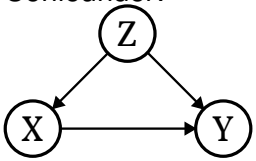
$$P(y|do(x)) = P(y|x) = P(y)$$

General Condition: Y must be a child (or generally an anchestor) of X.
If not fulfilled, there is no causal effect. The ATE is zero!

Also resolves the Fork structure.

Inference on Simple Graphs IV

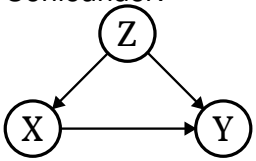
Confounder:



The query $P(y|do(x))$ considers a scenario were we set X freely and without external influence. However, in out observed data X is influenced by Z!

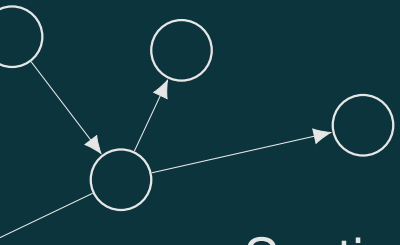
Inference on Simple Graphs IV

Confounder:



The query $P(y|do(x))$ considers a scenario were we set X freely and without external influence. However, in out observed data X is influenced by Z!

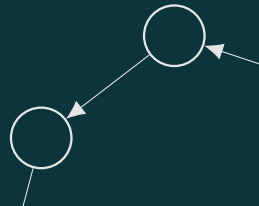
We need to adjust for the effects of Z!



Section

3

Interventions



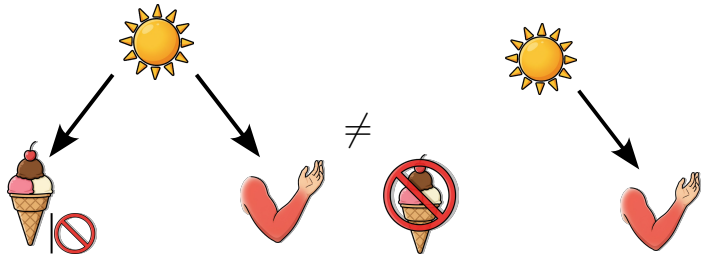
Interventions

Sometimes observation is not enough.

Sometimes agent actions alter the causal graph.

Reminder: Conditioning $\neq$ Intervening

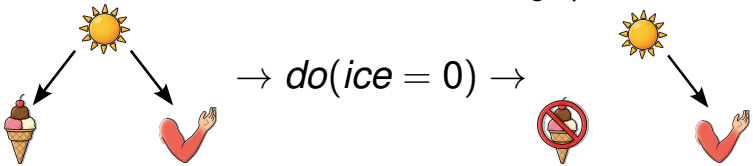
Conditioning and intervening are not the same!



Conditioning: Consider only those days where no ice-cream was sold:
$p(\text{sun-burns}|\text{ice-cream} = 0)$.
Intervening: Forbid all ice-cream sales:
$p(\text{sun-burns}|do(\text{ice-cream} = 0))$.

Interventions and do-Operator

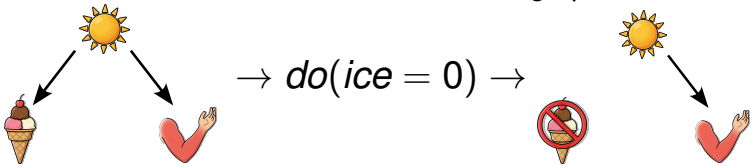
From the unintervened to the intervened graph:



The do-operator forcibly sets *ice* to 0, therefore *cutting all influence to the parents* of the intervened node.

Interventions and do-Operator

From the unintervened to the intervened graph:



The do-operator forcibly sets *ice* to 0, therefore *cutting all influence to the parents* of the intervened node.

do-Operator

For an SCM $\mathcal{M} = (\mathbf{V}, \mathbf{U}, \mathbf{F}, \mathcal{P}_{\mathbf{U}})$ the do-operator $do(X_i = c)$ replaces the structural equation $f_i \in \mathbf{F}$ with the assignment $f_i := c$.

Intervened SCM

Unintervened SCM:

$$\mathcal{M} = \begin{cases} \mathbf{V} & = \{Sun, Ice, SunBurn\} \\ \mathbf{U} & = \{U_{Sun} \in [0..1]\} \\ \mathbf{F} & = \begin{cases} f_{Sun} & := U_{Sun} \\ f_{Ice} & := Sun^2 \\ f_{SunBurns} & := 0.5\,Sun^2 \end{cases} \\ \mathcal{P}_{\mathbf{U}} & = \{U_{Sun} = \text{Uniform}(0, 1)\} \end{cases}$$

Intervened SCM

Intervened SCM $do(ice = 0)$:

$$\mathcal{M} = \begin{cases} \mathbf{V} & = \{Sun, Ice, SunBurn\} \\ \mathbf{U} & = \{U_{Sun} \in [0..1]\} \\ \mathbf{F} & = \begin{cases} f_{Sun} & := U_{Sun} \\ f_{Ice} & := 0 \\ f_{SunBurns} & := 0.5 Sun^2 \end{cases} \\ \mathcal{P}_{\mathbf{U}} & = \{U_{Sun} = \text{Uniform}(0, 1)\} \end{cases}$$

Consequences of Interventions

Interventions alter the structural equation, and therefore the conditional distribution, of the *intervened* variable.

Remember the independent mechanisms principle!
The mechanisms, and therefore the *conditional* distributions, of the *descendants* of the intervened variable stay the same. Only their *marginal* distributions change due to the altered input distribution coming from their parents!

Hard-Interventions

Previous interventions were of the form $do(X_i = c)$ where c is a constant. Such interventions are called **hard-interventions**.

- Fix the value of the target variable.
- Thereby, cutting all parent edges $\text{pa}(X_i) \to X_i$.

Truncated Factorization

For arbitrary interventions $\forall X_i \in \mathbf{V}.\forall c \in \text{dom}(X_i).do(X_i = c)$, SCM define a family of joint distributions.

Every hard intervention $do(X_j = c)$ entails a particular **Truncated Factorization**:

$$P_{\mathcal{M}}(X_1, \ldots, X_N | do(X_j = c)) = \left(\prod_{i \in \{1..N\} \setminus \{j\}} P(X_i | \text{pa}(X_i)) \right) \cdot P(X_j | do(X_j = c))$$

if $X_i = c$ and 0 otherwise.

Truncated Factorization

For arbitrary interventions $\forall X_i \in \mathbf{V}.\forall c \in \text{dom}(X_i).do(X_i = c)$, SCM define a family of joint distributions.

Every hard intervention $do(X_j = c)$ entails a particular **Truncated Factorization**:

$$P_{\mathcal{M}}(X_1, \ldots, X_N | do(X_j = c)) = \left(\prod_{i \in \{1..N\} \setminus \{j\}} P(X_i | \text{pa}(X_i))\right) \cdot P(X_j | do(X_j = c))$$

if $X_i = c$ and 0 otherwise.

Compare to the unintervened factorization of slide 11:

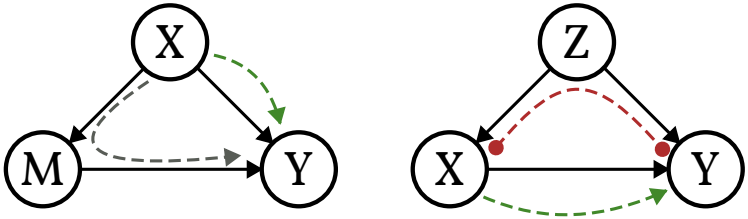
$$P_{\mathcal{M}}(x_1, \ldots, x_N) = \prod_{i \in \{1..N\}} P(x_i | \text{pa}(X_i))$$

Soft-Interventions

Other types of interventions are possible.

1. Stochastic Interventions - replace by a distribution instead of a fixed constant. E.g., $f_i := \mathcal{N}(0, 1)$.
2. Mechanism Change - arbitrary replacement of equations.
3. Shift-Intervention - offset the original value by a fixed constant: $f_i := f'_i + c$.
4. Conditional Interventions - dynamic treatment regimes.

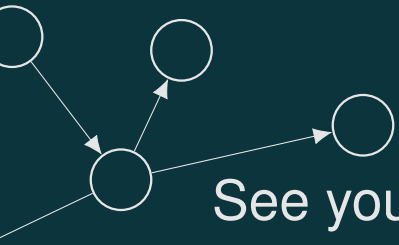
Direct-, Indirect, Total Effects



Causal Effects can be decomposed into direct-, indirect and total effects.

(Natural) Direct Effects (NDE) [green] travel along the direct edge $X \to Y$.
(Natural) Indirect Effects (NIE) [gray] travel along any directed path $X \to M_i \to \ldots M_k \to Y$ via intermediate *mediators*.
Non-Causal Associations [red]. Open path when non-conditioning on a confounder.

The **total effect** (TE) is the sum of direct and indirect effects: $TE = NDE + NIE$.



See you next week!