A Gentle Introduction to Casual Models

Fabio Massimo Zennaro
fabiomz@ifi.uio.no

University of Oslo

OsloMET
October 8th, 2019

1. Introduction

Foreword

- Study of *causes* as a scientific endeavour.
- Intuitive meaning of causality as *interventional causality* - true meaning of causality beyond the scope of this talk.

Outline

1. *Motivation*: why do we care about causality - an example.
2. *Statistics and Causality*: how the two fields relate.
3. *Models*: models to answer causal questions - BN, CBN and SCM.
4. *Problems*: what questions can we ask.

2. Motivation

A Motivating Example [6]

Assume we monitored the number of *ice-creams sold* (Ice) and the number of *thefts* (Thf) in our town:

Ice	Thf
195	39
137	27
14	6
61	14
130	27
137	29
...	...

What can we infer from this data?

A Motivating Example

Ice	Thf
195	39
137	27
14	6
...	...

✓ We can learn how the variables are *correlated*

$$Ice \uparrow, Thf \uparrow$$

✓ We can *predict* one variable from one another

$$Thf = f(Ice)$$

$$Ice = f(Thf)$$

A Motivating Example

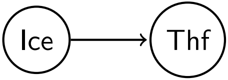
So, what if stop the sale of ice-creams?

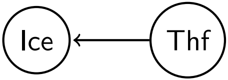
Ice	Thf
0	20
0	19
0	36
...	...

- × We fail to affect the number of thefts...
- × We miss information on the *directionality* of the relationship between the variables..

A Motivating Example

In order to *intervene* successfully, we need to know which variable affect which:

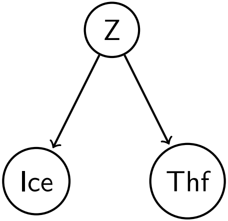




× Neither of the model seem to work...

A Motivating Example

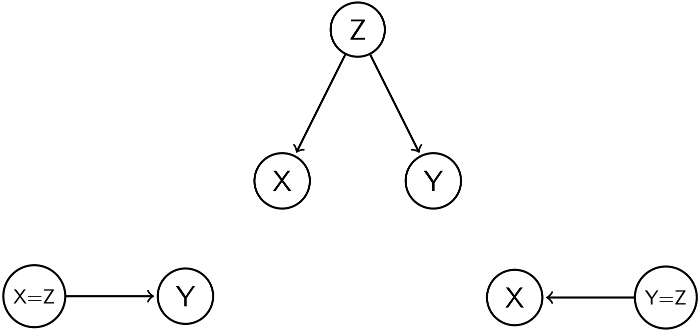
There likely is a *common cause* (Z) between the variables, such as the temperature



We have a **confounder** between *Ice* and *Thf*.

Aside: Reichenbach common cause principle [11]

Given two statistically dependent random variables X and Y there is a random variable Z that influences both.



3. Statistics and Causality

The Statistics-Causality Chasm [7, 2]

Statistics	**Causality**
Association	Cause
Correlation	Causation
Non-directionality	Directionality
Prediction	Action
Observation	Intervention

The Statistics-Causality Bridge [7]

How to bridge the gap between the domains is one of the main objective of the theory.

$$\begin{array}{c} \text{statistical} \\ \text{formalism} \end{array} \longrightarrow \begin{array}{c} \text{causality} \\ \text{formalism} \end{array}$$

$$\begin{array}{c} \text{observational} \\ \text{domain} \end{array} \longrightarrow \begin{array}{c} \text{interventional} \\ \text{domain} \end{array}$$

We need to rely on **assumptions**.

No causes in, no causes out.

Approaches to Causality [SEP][3][7][11]

From different fields: statistics, econometrics, epidemiology, social psychology, computer science.

Using different approaches: randomization, potential outcomes (Neyman, Rubin), structural causal models (Pearl, Halpern, Dawid), single-world intervention graphs (Richardson, Rubin), counterfactual logic (Brigg)

4. Causal Models

Pearl's Causality Ladder [8, 9, 13, 11]

3. **Counterfactuals**	What would have Y been, had X been x' when instead it was x? $p\left(Y_{do(X=x')} \middle\| Y=y, X=x\right)$ Structural causal models
2. **Causal Effects**	What is the effect of X on Y? $p\left(Y \middle\| do\left(X=x\right)\right)$ Causal Bayesian networks
1. **Associative Relationships**	How does Y relate to X? $p(Y\|X)$ Bayesian networks

Level 0: Directed Acyclic Graphs [1, 10]

3. **Counterfactuals**	What would have Y been, had X been x' when instead it was x? $p\left(Y_{do(X=x')} \| Y=y, X=x\right)$ Structural causal models
2. **Causal Effects**	What is the effect of X on Y? $p\left(Y \| do\left(X=x\right)\right)$ Causal Bayesian networks
1. **Associative Relationships**	How does Y relate to X? $p(Y \| X)$ Bayesian networks
0. **Mathematical Models**	Directed acyclic graphs

Directed Acyclic Graphs

A **directed acyclic graph** is a tuple:

$$\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$$

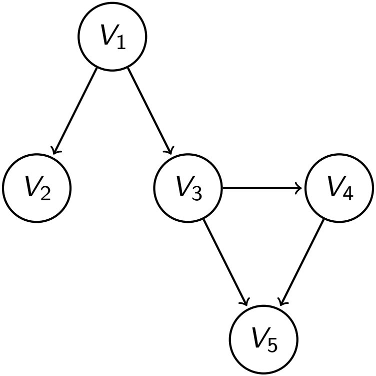
where:

- $\mathcal{V}$ is a set of *nodes* (*vertices*)
- $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is a set of *edges* (*arcs*)

such that:

- edges are *directed*;
- there are *no cycles*.

Directed Acyclic Graphs



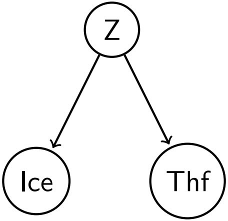
A DAG is a purely *mathematical structure*.

Assumptions of DAGs

- **Directionality**: edges have a direction.
- **Acyclicity**: no loops in the graph.

Example: DAG

$$\mathcal{V} = \{\mathrm{Ice}, \mathrm{Thf}, Z\}$$
$$\mathcal{E} = \{(Z, \mathrm{Ice}), (Z, \mathrm{Thf})\}$$



Level 1: Bayesian Networks [1, 12, 10, 11]

3. **Counterfactuals**	What would have Y been, had X been x' when instead it was x? $p\left(Y_{do(X=x')} \mid Y=y, X=x\right)$ Structural causal models
2. **Causal Effects**	What is the effect of X on Y? $p\left(Y \mid do\left(X=x\right)\right)$ Causal Bayesian networks
1. **Associative Relationships**	How does Y relate to X? $p(Y \mid X)$ Bayesian networks
0. **Mathematical Models**	Directed acyclic graphs

Bayesian Networks

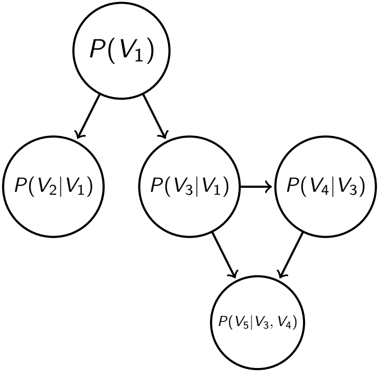
A **Bayesian network** (**belief network**) is a DAG endowed with a joint probability distribution $P(\mathbf{V})$ that respects the *Markov factorization property* wrt DAG:

$$P(V_1, V_2, ..., V_n) = \prod_{i=1}^{n} P(V_i | Pa(V_i)).$$

This is equivalent to:

- each variable is independent of its non-descendents given its parents;
- d-separation implies conditional independence.

Bayesian Networks



$$P(\mathbf{V}) = P(V_5|V_3, V_4) \cdot P(V_4|V_3) \cdot P(V_3|V_1) \cdot P(V_2|V_1) \cdot P(V_1)$$

A BN merges the *mathematical structure* of a DAG with *probability*.

d-Separation

Two nodes X and Y are *d-separated* by a set of nodes $\mathbf{Z}$ if:

1. all *chains* and *forks* between X and Y are blocked by the nodes in $\mathbf{Z}$;
2. no *collider* between X and Y is blocked by the nodes in $\mathbf{Z}$.

$$X \rightarrow Z_i \rightarrow Y \qquad\qquad X \leftarrow Z_j \rightarrow Y$$

$$X \rightarrow W \leftarrow Y$$

d-separation provides a graphical criterion to assess conditional independencies.

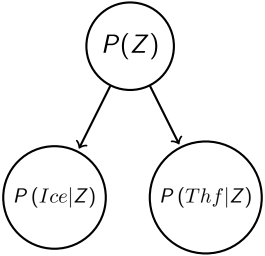
Assumptions of BNs

- **Markovianity**: d-separation implies independence.
- **Faithfulness**: independence implies d-separation.
- **Perfect map**: Markovianity and faithfulness.

Remarks

- BNs allow us to answer *associative relationships* questions (first step of the causal ladder).

Example: BN



$$\mathrm{Ice} \perp_D \mathrm{Thf}|Z$$

$$\mathrm{Ice} \perp \mathrm{Thf}|Z$$

$$P\left(\mathbf{V}\right) = P\left(\mathrm{Ice}, \mathrm{Thf}, Z\right) = P\left(\mathrm{Ice}|Z\right) P\left(\mathrm{Thf}|Z\right) P(Z)$$

Level 2: Causal Bayesian Networks [12, 10, 11]

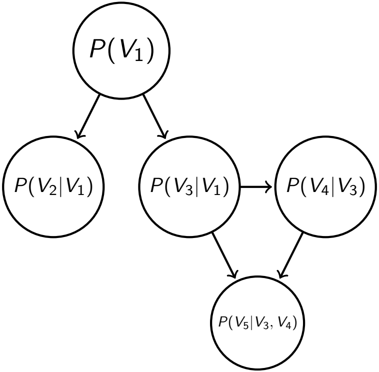
3. **Counterfactuals**	What would have Y been, had X been x' when instead it was x? $p\left(Y_{do(X=x')} \middle\| Y = y, X = x\right)$ Structural causal models
2. **Causal Effects**	What is the effect of X on Y? $p\left(Y \middle\| do\left(X = x\right)\right)$ Causal Bayesian networks
1. **Associative Relationships**	How does Y relate to X? $p(Y \| X)$ Bayesian networks
0. **Mathematical Models**	Directed acyclic graphs

Causal Bayesian Networks

A **causal Bayesian network** is a BN whose edges represent *causal relationsips*, such that:

- each variable is independent of its non-effects given its direct causes;
- each variable can be affected locally.

Causal Bayesian Networks



A CBN merges the *mathematical structure* of a DAG with *probability* and *causality*.

Assumptions of CBNs

- **Causal Markov assumption**: a node is independent of its non-effects given its direct causes.
- **Causal arrows**: arrows represent causal relationships.
- **Zero influence**: missing arrow means no causal relationship.
- **Common cause completeness**: all common causes are modeled.
- **Causal relationship completeness**: all causes among the variables in the model are present.
- **Autonomy**: external interventions act locally.

Remarks

- CBNs allow us to answer *causal effects* questions (second step of the causal ladder).
- We can formulate causal effect questions via *interventions*.

Interventions

An **intervention** is a new operation by which a variable is a set to a fixed value.

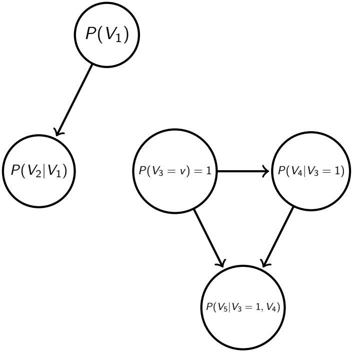
$$\mathrm{do}\,(X = x)$$

This reflect the *acting* of an observer on a system, with the assumption of **no side effects**.

Interventions

Graphically:

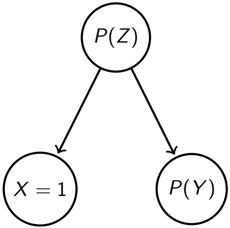
1. We fix the value of node on which we intervene;
2. We remove incoming arrows.



We obtained the new *intervened* (or *post-intervention*) model.

Interventions

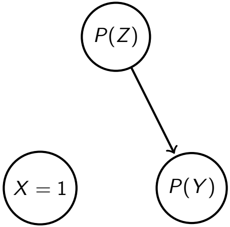
Conditioning $\neq$ Intervention



$$P(Y|X=1)$$

Distribution of Y when observing $X=1$.

Knowledge of $X=1$ allows inference on distribution of Z and then Y.

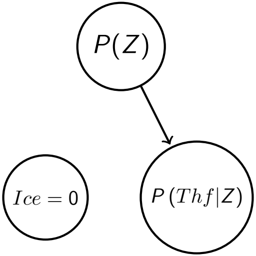


$$P(Y|\text{do}(X=1))$$

Distribution of Y when intervening to do $X=1$.

Knowledge of $\text{do}(X=1)$ does not affect the distribution of Z.

Example: CBN



$$P(\mathrm{Thf}, Z|\mathrm{do}(\mathrm{Ice} = 0)) = P(\mathrm{Thf}|Z)\, P(Z)$$

Level 3: Structural Causal Models

3. **Counterfactuals**	What would have Y been, had X been x' when instead it was x? $p\left(Y_{do(X=x')} \middle\| Y=y, X=x\right)$ Structural causal models
2. **Causal Effects**	What is the effect of X on Y? $p\left(Y \middle\| do\left(X=x\right)\right)$ Causal Bayesian networks
1. **Associative Relationships**	How does Y relate to X? $p(Y\|X)$ Bayesian networks
0. **Mathematical Models**	Directed acyclic graphs

Structural Causal Models [6, 10, 11]

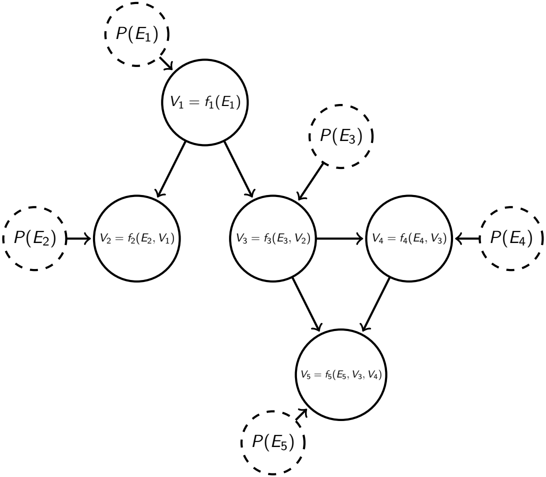
A **probabilistic structural causal model** is a CBN with *structural equations*.
It is defined as a tuple:

$$\mathcal{M} = \langle \mathcal{E}, \mathcal{V}, \mathcal{F}, \mathcal{P} \rangle$$

where:

- $\mathcal{E}$ is a set of *exogenous nodes* (*noise*);
- $\mathcal{V}$ is a set of *endogenous nodes* (*variables of interest*);
- $\mathcal{F}$ is a set of *structural functions*, one for each endogenous node;
- $\mathcal{P}$ is a set of *probability distributions*, one for each exogenous node.

Structural Causal Models



A SCM merges *algebraic formalism* and *graphical notation*.

Assumptions of SCMs

- **Autonomous functions**: each variable is governed by an autonomous function.
- **Independent noise**: each variable has a single independent source of noise (equivalent to *common cause completeness*).

Remarks

- SCMs allow us to answer *counterfactual* questions (third step of the causal ladder).
- SCMs are the model of choice for causal inference, and we will always refer to them from now on.

Counterfactuals

A **counterfactual** is an operation by which we compute a quantity of interest in an alternate world in which we perform an intervention.

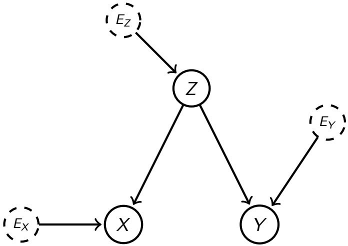
$$P\left(Y_{do(X=x')} | Y=y, X=x\right)$$

This reflect the *counterfactual question*: assuming we observed $Y=y$ and $X=x$, what would have Y been, had we acted on $do\left(X=x'\right)$?

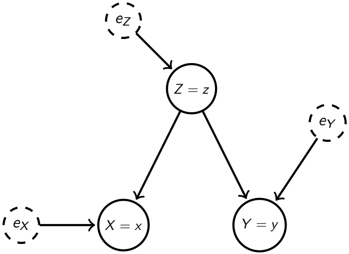
Counterfactuals

Evaluating a counterfactual $P\left(Y_{do(Z=z')} | Y=y, X=x, Z=z\right)$

1. *Abduction:* use observed variables to infer the value/distribution of exogenous variables.



(Original model)

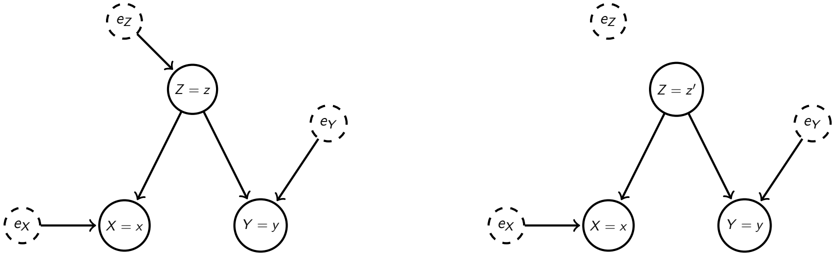


(Abduction)

Counterfactuals

Evaluating a counterfactual $P\left(Y_{do(Z=z')} | Y=y, X=x, Z=z\right)$

2. *Action:* intervene as requested in the counterfactual.

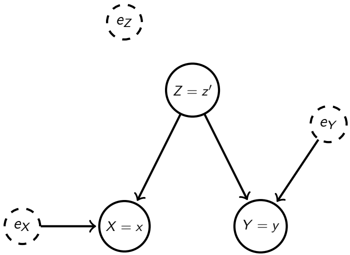


(Abducted model) (Action)

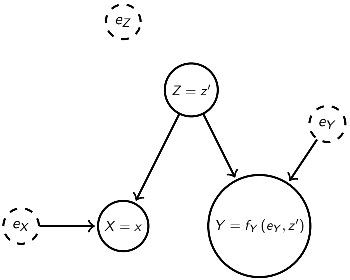
Counterfactuals

Evaluating a counterfactual $P\left(Y_{do(Z=z')} \middle| Y=y, X=x, Z=z\right)$

3. *Prediction:* compute the variable of interest in the counterfactual model.



(Acted model)



(Prediction)

Counterfactuals

Interventions $\neq$ Counterfactuals

$$P\left(Bet = Coin | \mathrm{do}\left(Bet = head\right)\right)$$

Probability of winning if we force the bet to head.

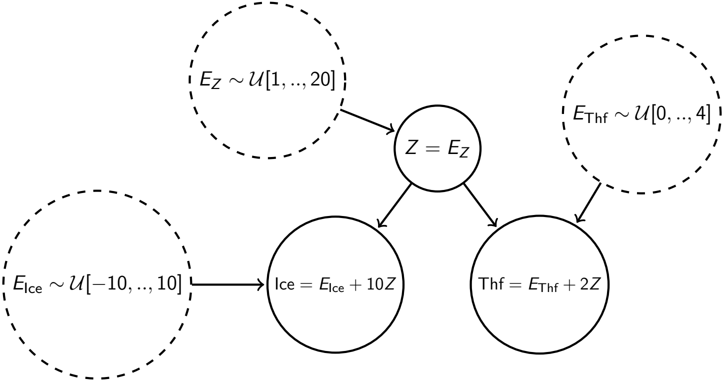
The outcome of the coin toss is still random, and the chance of winning half.

$$P\left(Bet = Coin_{do(Bet=head)} \middle| Coin = head, Bet = tail\right)$$

Probability of winning if we had forced the bet to head, having observed the outcome head and the bet tail.

We know with certainty the result of the bet.

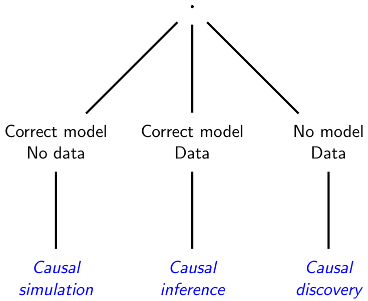
Example: SCM



$E_Z \sim \mathcal{U}[1,..,20]$
$Z = E_Z$
$E_{\text{Thf}} \sim \mathcal{U}[0,..,4]$
$E_{\text{Ice}} \sim \mathcal{U}[-10,..,10]$
$\text{Ice} = E_{\text{Ice}} + 10Z$
$\text{Thf} = E_{\text{Thf}} + 2Z$

5. Causal Problems

Causal Problems



Other challenging factors:

- *Model*: partially specified, hidden variables
- *Data*: observational/interventional, missing data

Casual Inference Approach [7]

1. **Define**: express the quantity of interest as a function of a generic model $Q(\mathcal{M})$;
2. **Assume**: define your specific model $\mathcal{M}*$;
3. **Identify**: evaluate if $Q(\mathcal{M})$ is identifiable in $\mathcal{M}*$;
4. **Estimate**: estimate, approximate or bound $Q(\mathcal{M}*)$.

Casual Inference of Causal Effects from Observational Data [10, 11]

Can we identify causal effect of X_1 on Y given observational data from the model? *Confounders* make evaluation non-trivial.

1. **Define**: $P(Y|\text{do}(X_1 = x))$;
2. **Assume**: we defined our SCM of interest;
3. **Identify**: use *do-calculus* / *ID algorithm*;
4. **Estimate**:
   - *Truncated factorization*
   - *Adjustment formula*
   - *Inverse probability weighting*
   - *Propensity score*

Truncated factorization

The **truncated factorization** (or **g-formula**) simply computes the *joint distribution* in the interventional model.

By Markovianity, in the original model:

$$P(\mathbf{X}) = \prod_i P(X_i | pa(X_i))$$

Markovianity holds in the intervened model $\mathcal{M}$ where $\mathrm{do}(X_1 = x)$ too:

$$\begin{aligned} P_{\mathcal{M}}(\mathbf{X}) &= \prod_i P_{\mathcal{M}}(X_i | pa(X_i)) \\ &= P_{\mathcal{M}}(X_1) \prod_{i \neq 1} P_{\mathcal{M}}(X_i | pa(X_i)) \end{aligned}$$

Truncated factorization

$$P_{\mathcal{M}}(X_1) \prod_{i \neq 1} P_{\mathcal{M}}(X_i | pa(X_i)) = \begin{cases} \prod_{i \neq 1} P_{\mathcal{M}}(X_i | pa(X_i)) & \text{if } X_1 = x \\ 0 & \text{otherwise} \end{cases}$$

$$\underbrace{\begin{cases} \prod_{i \neq 1} P_{\mathcal{M}}(X_i | pa(X_i)) & \text{if } X_1 = x \\ 0 & \text{otherwise} \end{cases}}_{\text{interventional}} = \underbrace{\begin{cases} \prod_{i \neq 1} P(X_i | pa(X_i)) & \text{if } X_1 = x \\ 0 & \text{otherwise} \end{cases}}_{\text{observational}}$$

- If X_1 has no parents then intervention $=$ conditioning!
- Truncated factorization is a very general formula (it can deal with multiple interventions)

Adjustment formula

An **adjustment formula** computes the causal effect without evaluating the whole joint interventional distribution, but considering only the *confounders* of the quantities of interest.

We want a set of nodes $\mathbf{Z}$ such that:

$$\underbrace{P\left(Y|\mathrm{do}\left(X_1 = x\right)\right)}_{\text{interventional}} = \underbrace{\sum_{\mathbf{Z}} P\left(Y|X_1, \mathbf{Z}\right) P\left(\mathbf{Z}\right)}_{\text{observational}}$$

The set $\mathbf{Z}$ has to be chosen properly.

Adjustment formula

If we observe all the parents of X_1, we can use **parent adjustment**:

$$\mathbf{Z} = pa\left(X_i\right)$$

If we do not observe all the parents of X_1, we can use the **backdoor criterion** and find $\mathbf{Z}$ such that:

1. no node in $\mathbf{Z}$ is a descendant of X_1;
2. $\mathbf{Z}$ blocks every path between Y and X_1 containing an arrow into X_1.

Inverse Probability Weighting

If we can easily estimate conditional probabilities we can use **inverse probability weighting**.

By adjustment and Bayes:

$$P(Y|\mathrm{do}(X_1 = x)) = \sum_{\mathbf{Z}} P(Y|X_1, \mathbf{Z}) P(\mathbf{Z}) = \sum_{\mathbf{Z}} \frac{P(Y, X_1, \mathbf{Z})}{P(X_1|\mathbf{Z})}$$

Propensity score

If estimating for $\mathbf{Z}$ proves challenging (high-dimensionality), we can rely on a **propensity score**.

A propensity score is a function $\ell = g(\mathbf{Z})$ such that $P(\mathbf{Z}|\ell) = P(\mathbf{Z}|X,\ell)$. Then:

$$\begin{aligned}
P(Y|\mathrm{do}(X_1 = x)) &= \sum_{\mathbf{Z}} P(Y|X_1,\mathbf{Z})\, P(\mathbf{Z}) \\
&= \sum_{\mathbf{Z}} \sum_{\ell} P(Y|X_1,\mathbf{Z})\, P(\mathbf{Z}|\ell)\, P(\ell) \\
&= \sum_{\mathbf{Z}} \sum_{\ell} P(Y|\ell, X_1,\mathbf{Z})\, P(\mathbf{Z}|X,\ell)\, P(\ell) \\
&= \sum_{\ell} P(Y|\ell, X_1)\, P(\ell)
\end{aligned}$$

Do-calculus

Formally, any identifiable intervention may be computed via **do-calculus**.

Complete set of rules to manipulate interventional quantities:

1. *Insertion/deletion of observations*;
2. *Action/observation exchange*;
3. *Insertion/deletion of actions*.

Observation on causal inference of causal effects

- Several other techniques exist for estimating causal effects (*randomized experiments*, *matching*, *natural experiments*, *instrumental variables*...) [Athey]
- Different techniques may vary on their *feasibility* and their *statistical properties* (bias, variance).
- In general, identifiability of an interventional query is solved via *ID algorithm* [14, 13].

Casual Inference of Counterfactual Effects [10, 6]

Can we identify what Y would have been, if X had been x' instead of x while keeping everything elso constant?

1. **Define**: $P\left(Y_{\mathrm{do}(X=x')} | Y = y, X = x\right)$;
2. **Assume**: we defined our SCM of interest;
3. **Identify**: use *IDC algorithm*;
4. **Estimate**:
   - *Twin networks*
   - *Mediation formula*

Observation on causal inference of counterfactual effects

- Several quantities may be of interest (*effect of the treatment on the treated*, *direct effects*, *probability of necessity*, *probability of sufficiency*) [10].
- Important role in fairness [4].

Graph discovery from data [11, 5]

Given observational data, can we identify the graphical causal model $\mathcal{M}$ that generated the data?

- For each probabilistic SCM there is a *single* pdf underlying it.
- For each pdf there is a set of SCMs encoding it (**Markov equivalence class**)

Approaches to graph discovery

- **Independence-based** or **constraint-based**: exploit independences in the data to find a Markov equivalence class (*PC, SGS, PC-stable*);
- **Score-based** or **fitness-based**: use a loss function to greedily rank models (*GES*);
- **Assumption-based**: use prior knowledge to restrict the space of models (*ANMs, LiNGAMs*).

6. Conclusions

Conclusions

- We can formally express causal statements.
- There are different formalisms to do it. SCMs is one of them:
  - It is general.
  - It helps making assumptions explicit.
  - It eases reasoning via graphs.
- Causality will likely have an important role in learning.

Conclusions

"More has been learned about causal inference in the last few decades than the sum total of everything that had been learned about it in all prior recorded history"

(Gary King, Harvard, 2014)

Thanks!

Thank you for listening!

References I

[1] David Barber. *Bayesian reasoning and machine learning*. Cambridge University Press, 2012.

[2] A Philip Dawid. Statistical causality from a decision-theoretic perspective. *Annual Review of Statistics and Its Application*, 2:273–303, 2015.

[3] Andrew Gelman. Causality and statistical learning. *American Journal of Sociology*, 117(3):955–966, 2011.

[4] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4069–4079, 2017.

[5] Marloes H Maathuis, Preetam Nandy, and P Btihlmann. A review of some recent advances in causal inference., 2016.

[6] Judea Pearl. *Causality*. Cambridge university press, 2009.

References II

[7] Judea Pearl. An introduction to causal inference. *The international journal of biostatistics*, 2010.

[8] Judea Pearl. Theoretical impediments to machine learning with seven sparks from the causal revolution. *arXiv preprint arXiv:1801.04016*, 2018.

[9] Judea Pearl. Sufficient causes: Revisiting oxygen, matches, and fires. *Journal of Causal Inference*, 2019.

[10] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: a primer*. John Wiley & Sons, 2016.

[11] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

[12] Richard Scheines. An introduction to causal inference. 1997.

References III

[13] Ilya Shpitser and Judea Pearl. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9(Sep):1941–1979, 2008.

[14] Jin Tian and Judea Pearl. A general identification condition for causal effects. In *Aaai/iaai*, pages 567–573, 2002.