ASCI lecture *Causal Modelling*

Joris Mooij
j.m.mooij@uva.nl

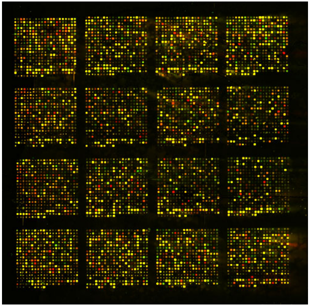
Informatics Institute

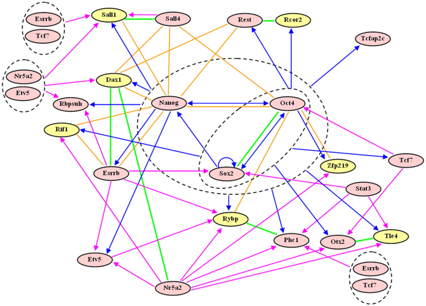
University of Amsterdam

April 14th, 2016

Outline

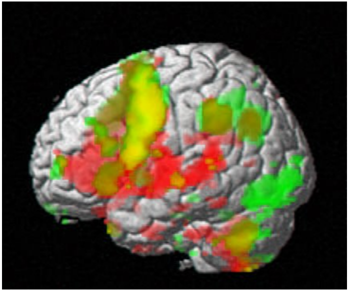
Genetics:
how to infer gene regulatory networks from micro-array data?

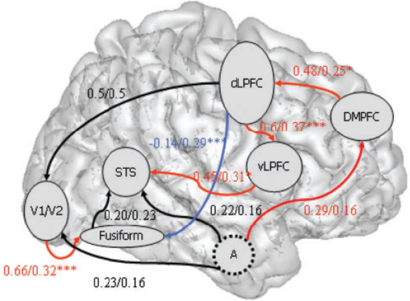




Social sciences:
does playing violent computer games cause aggressive behavior?

Neuroscience:
how to infer functional connectivity networks from fMRI data?





Economy:
Does austerity reduce national debt?

Politics:
Do extra bombings on IS targets reduce or increase the likelihood of terrorist attacks?

Causality: what is it?

Causality is central notion in science, decision-taking and daily life.

How to reason formally about cause and effect?
(We don't learn this at school, and only very rarely at university!)

Question: give a definition of cause and effect.

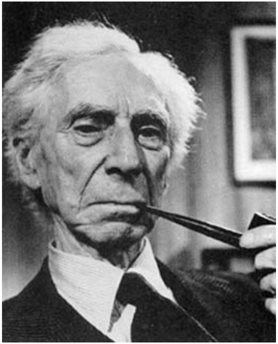
The subject of *causality* has a long history in philosophy. For example, this is what Hume had to say about it:

"Thus we remember to have seen that species of object we call *flame*, and to have felt that species of sensation we call *heat*. We likewise call to mind their constant conjunction in all past instances. Without any farther ceremony, we call the one *cause* and the other *effect*, and infer the existence of the one from that of the other."

David Hume, *Treatise of Human Nature*

But: does the rooster's crow really cause the sun to rise?

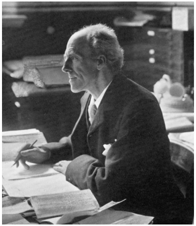
Some philosophers even proposed to abandon the concept of causality completely.



"All philosophers, of every school, imagine that causation is one of the fundamental axioms or postulates of science, yet, oddly enough, in advanced sciences such as gravitational astronomy, the word 'cause' never occurs. The law of causality, I believe, like much that passes muster among philosophers, is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm."

Bertrand Russell, *On The Notion Of Cause*

Karl Pearson (one of the founders of modern statistics, well-known from his work on the *correlation coefficient*) writes:



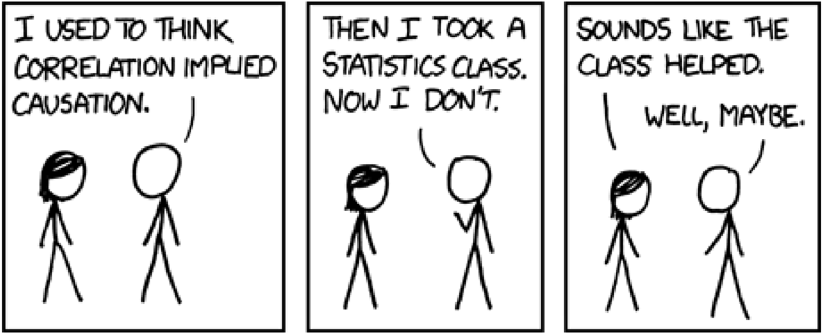
"Beyond such discarded fundamentals as 'matter' and 'force' lies still another fetish amidst the inscrutable arcana of even modern science, namely, the category of cause and effect."

Karl Pearson, *The Grammar of Science*

Since then, many statisticians tried to avoid causal reasoning:

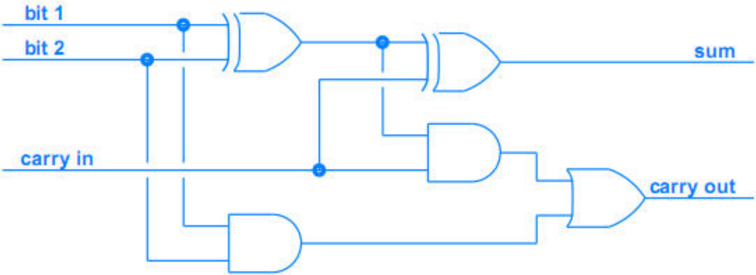
- "Considerations of causality should be treated as they have always been in statistics: preferably not at all." (Terry Speed, former president of the Biometric Society).
- "It would be very healthy if more researchers abandon thinking of and using terms such as cause and effect." (Prominent social scientist).

A modern philosopher on Causality



Randall Munroe, www.xkcd.org

Causality in engineering



Causality is a very useful concept in engineering.

Using causal reasoning, engineers can not only predict what happens when a system operators normally, but also when an external *intervention* changes part of the system.

Being able to predict what happens under interventions allows to exert *control*.

A formal theory of causality?

Question

Can we formalize causal reasoning?

Exercise 1

Please make Exercise 1...

Problems in formalizing causal reasoning: probabilities

Example (Simpson's paradox)

We collect electronic patient records to investigate the effectiveness of a new drug against a certain disease. It can happen that:

1. The probability of recovery is higher for patients that took the drug:

$$p(\text{recovery} \mid \text{drug}) > p(\text{recovery} \mid \text{no drug})$$

2. For **both** *male and female* patients, however, the relation is opposite:

$$p(\text{recovery} \mid \text{drug}, \text{male}) < p(\text{recovery} \mid \text{no drug}, \text{male})$$

$$p(\text{recovery} \mid \text{drug}, \text{female}) < p(\text{recovery} \mid \text{no drug}, \text{female})$$

Should we use this drug for treatment?

Problems in formalizing causal reasoning: probabilities

Example (Simpson's paradox)

We collect electronic patient records to investigate the effectiveness of a new drug against a certain disease. It can happen that:

1. The probability of recovery is higher for patients that took the drug:

$$p(\text{recovery} \mid \text{drug}) > p(\text{recovery} \mid \text{no drug})$$

2. For **both** *male and female* patients, however, the relation is opposite:

$$p(\text{recovery} \mid \text{drug}, \text{male}) < p(\text{recovery} \mid \text{no drug}, \text{male})$$

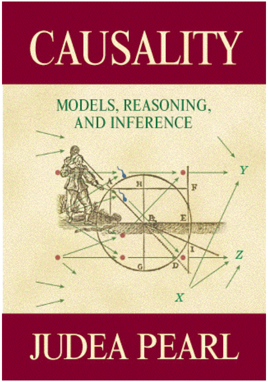
$$p(\text{recovery} \mid \text{drug}, \text{female}) < p(\text{recovery} \mid \text{no drug}, \text{female})$$

Should we use this drug for treatment?

Note

Fancy classifiers, deep learning and big data do not help us here!

Judea Pearl



ACM Turing Award 2011: "For fundamental contributions to artificial intelligence through the development of a calculus for probabilistic and causal reasoning."

Pearl's contribution: the do-operator

- Probability theory has a semantics for updating probabilities given *observations*: conditioning.
- Pearl extends probability calculus by introducing a new operator for describing *interventions*, the **do-operator**.

Example (Do-operator)

- $p(\text{lung cancer} \mid \text{smoke})$: the probability that somebody gets lung cancer, given (the observation) that the person smokes.
- $p(\text{lung cancer} \mid \text{do}(\text{smoke}))$: the probability that somebody gets lung cancer, if we *force* the person to smoke.

Resolution of Simpson's paradox:

- Simpson's paradox is only paradoxical if we misinterpret $p(\text{recovery} \mid \text{drug})$ as $p(\text{recovery} \mid \text{do}(\text{drug}))$.
- We should prescribe the drug if $p(\text{recovery} \mid \text{do}(\text{drug})) > p(\text{recovery} \mid \text{do}(\text{no drug}))$.

Do-calculus

Pearl recognized that the rules of probability theory do not suffice for causal reasoning. He formulated three additional rules (the "**do-calculus**"):

1. **Ignoring observations**:

$$p(\boldsymbol{y} \mid \mathrm{do}(\boldsymbol{x}), \boldsymbol{w}, \boldsymbol{z}) = p(\boldsymbol{y} \mid \mathrm{do}(\boldsymbol{x}), \boldsymbol{w}) \qquad \text{if } (\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{W})_{\mathcal{G}_{\overline{\boldsymbol{X}}}}$$

2. **Action/observation exchange**:

$$p(\boldsymbol{y} \mid \mathrm{do}(\boldsymbol{x}), \mathrm{do}(\boldsymbol{z}), \boldsymbol{w}) = p(\boldsymbol{y} \mid \mathrm{do}(\boldsymbol{x}), \boldsymbol{z}, \boldsymbol{w}) \qquad \text{if } (\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{W})_{\mathcal{G}_{\overline{\boldsymbol{X}}, \underline{\boldsymbol{Z}}}}$$

3. **Ignoring actions**:

$$p(\boldsymbol{y} \mid \mathrm{do}(\boldsymbol{x}), \mathrm{do}(\boldsymbol{z}), \boldsymbol{w}) = p(\boldsymbol{y} \mid \mathrm{do}(\boldsymbol{x}), \boldsymbol{w}) \qquad \text{if } (\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{W})_{\mathcal{G}_{\overline{\boldsymbol{X}}, \overline{\boldsymbol{Z}(\boldsymbol{W})}}}$$

   where $\boldsymbol{Z}(\boldsymbol{W}) = \boldsymbol{Z} \setminus An_{\mathcal{G}_{\overline{\boldsymbol{X}}}}(\boldsymbol{W})$.

The do-calculus allows us to reason with (probabilistic) causal statements, given (partial) knowledge of the causal structure.

Outline

Causal relations

Definition

A causes B if changing A may lead to a change of B.

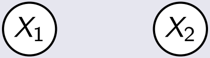
Causal relations

Definition

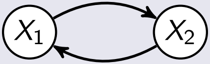
A causes *B* if changing *A* may lead to a change of *B*.

Causal graph represents the causal relationships between variables (nodes are variables, edges encode causal relations between variables).

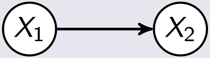
Example



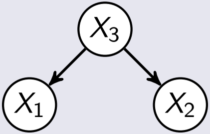
X_1 and X_2 are causally unrelated



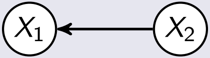
X_1 and X_2 cause each other



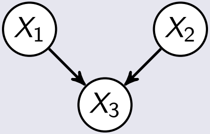
X_1 causes X_2



X_1 and X_2 have a common cause X_3



X_2 causes X_1



X_1 and X_2 have a common effect X_3

Direct causation

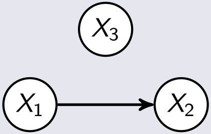
Let $\boldsymbol{V} = \{X_1, \dots, X_N\}$ be a set of variables.

Definition

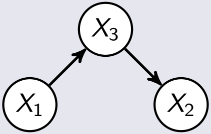
If X_i causes X_j even if all other variables $\boldsymbol{V} \setminus \{X_i, X_j\}$ are hold fixed at arbitrary values, then

- we say that X_i causes X_j directly with respect to $\boldsymbol{V}$
- we indicate this in the causal graph on $\boldsymbol{V}$ by a directed edge $X_i \to X_j$

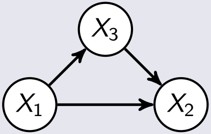
Example



X_1 causes X_2;
X_1 causes X_2 directly w.r.t. $\{X_1, X_2, X_3\}$



X_1 causes X_2;
X_1 does not cause X_2 directly w.r.t. $\{X_1, X_2, X_3\}$



X_1 causes X_2;
X_1 causes X_2 directly w.r.t. $\{X_1, X_2, X_3\}$

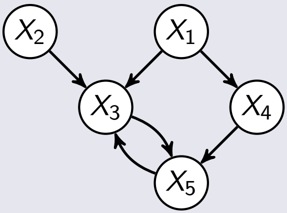
Terminology of directed graphs

Let $\mathcal{G}$ be a directed graph with nodes $\boldsymbol{V} = \{X_1, \dots, X_N\}$.

Definition

- If $X_i \to X_j$ we call X_i parent of X_j and X_j a child of X_i.
- If $X_i \to X_j$ or $X_j \to X_i$ then we call X_i and X_j adjacent.
- If $X_{i_1} \to X_{i_2} \to X_{i_3} \to \cdots \to X_{i_k}$ we say that there is a directed path from X_{i_1} to X_{i_k}.
- If there is a directed path from X_i to X_j (or if $X_i = X_j$), X_i is called a ancestor of X_j, and X_j is called a descendant of X_i.
- $\mathrm{An}_{\mathcal{G}}(\boldsymbol{X})$ denotes the set of all ancestors of nodes in subset $\boldsymbol{X} \subseteq \boldsymbol{V}$.

Example

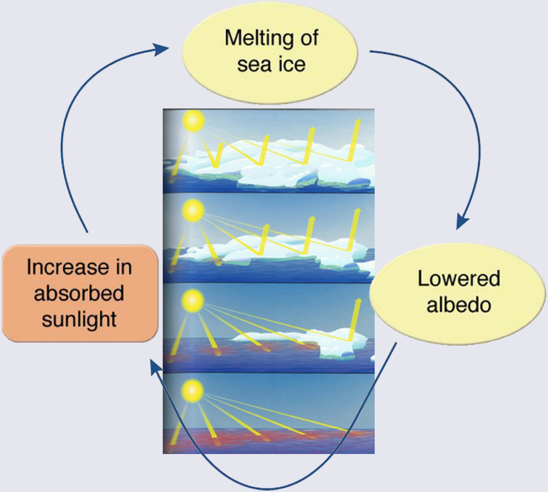


Causal interpretation

parent	= direct cause
child	= direct effect
ancestor	= cause
descendant	= effect

Feedback loops: Example

Example



Cycles, Feedback loops: Definitions

Let $\mathcal{G}$ be a directed graph with nodes $\boldsymbol{V} = \{X_1, \dots, X_N\}$.

Definition

$\mathcal{G}$ is cyclic if it contains a directed cycle

$$X_{i_1} \to X_{i_2} \to \cdots \to X_{i_k}, \qquad X_{i_1} = X_{i_k}$$

If it does not contain such a directed cycle, the graph is called acyclic. This is also known as a DAG (Directed Acyclic Graph).

Definition

If A causes B and B causes A, then we say that A and B are involved in a causal feedback loop.

Mutilated graphs

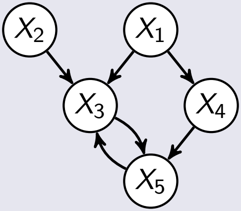
Definition

Given a directed graph $\mathcal{G} = (\boldsymbol{V}, \boldsymbol{E})$ and a subset $\boldsymbol{X} \subseteq \boldsymbol{V}$, we define

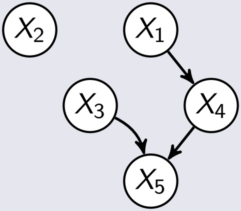
- $\mathcal{G}_{\overline{\boldsymbol{X}}}$ to be $\mathcal{G}$ without the incoming edges on nodes in $\boldsymbol{X}$;
- $\mathcal{G}_{\underline{\boldsymbol{X}}}$ to be $\mathcal{G}$ without the outgoing edges from nodes in $\boldsymbol{X}$.

Example

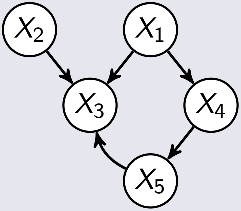
$\mathcal{G}$:



$\mathcal{G}_{\overline{X_3}}$:



$\mathcal{G}_{\underline{X_3}}$:



Perfect interventions

Definition

A perfect intervention $\mathrm{do}(\boldsymbol{X} = \boldsymbol{\xi})$ on a set of variables $\boldsymbol{X} \subseteq \boldsymbol{V}$ is an externally enforced change of the system that ensures that $\boldsymbol{X} = \boldsymbol{\xi}$ but leaves the rest of the system untouched.

The concept of perfect intervention assumes "modularity": the causal system can be divided into two parts, $\boldsymbol{X}$ and $\boldsymbol{V} \setminus \boldsymbol{X}$, and we can make changes to one part while keeping the other part intact.

Note

The causal graph $\mathcal{G}$ changes into $\mathcal{G}_{\overline{\boldsymbol{X}}}$ after a perfect intervention $\mathrm{do}(\boldsymbol{X} = \boldsymbol{\xi})$ (because none of the other variables can now cause $\boldsymbol{X}$).

Confounders: Example

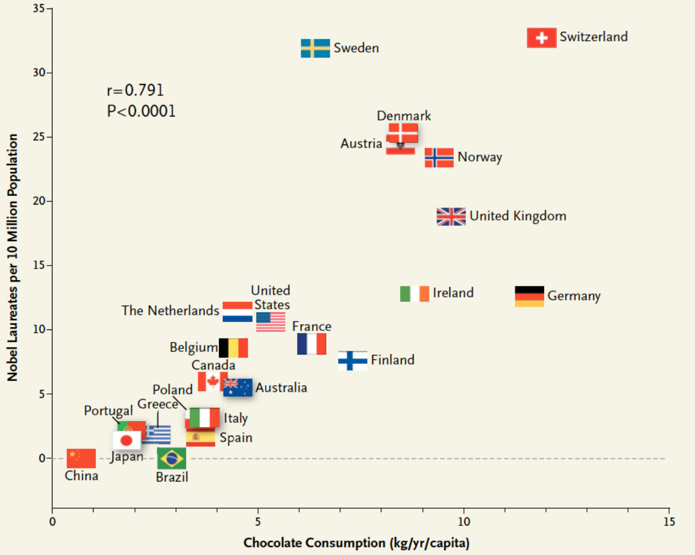


Figure 1. **Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**

Confounders: Definition

Definition

Let X, Y be observed variables and H an latent (unobserved) variable.
H confounds X and Y if:

1. there exists a directed path from H to X that does not contain Y
2. there exists a directed path from H to Y that does not contain X

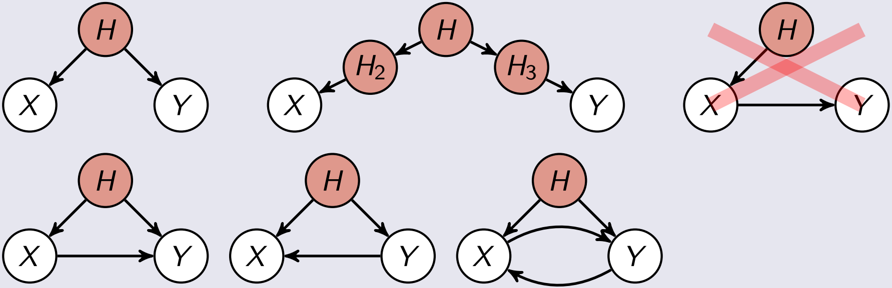
Confounders: Definition

Definition

Let X, Y be observed variables and H an latent (unobserved) variable.
H confounds X and Y if:

1. there exists a directed path from H to X that does not contain Y
2. there exists a directed path from H to Y that does not contain X

Example



(Conditional) independences

Definition: independence

Given two random variables X, Y, we write $X \perp\!\!\!\perp Y$ and say that X is independent of Y if

$$p(X, Y) = p(X)p(Y).$$

Intuitively, X is independent of Y if we do not learn anything about X when told the value of Y (or vice versa).

(Conditional) independences

Definition: independence

Given two random variables X, Y, we write $X \perp\!\!\!\perp Y$ and say that X is independent of Y if

$$p(X, Y) = p(X)p(Y).$$

Intuitively, X is independent of Y if we do not learn anything about X when told the value of Y (or vice versa).

Definition: conditional independence

Given a third random variable Z, we write $X \perp\!\!\!\perp Y \mid Z$ and say that X is (conditionally) independent from Y, given Z, if

$$p(X, Y \mid Z) = p(X \mid Z)p(Y \mid Z).$$

Intuitively, X is independent of Y if, given the value of Z, we do not learn anything new about X when told the value of Y.

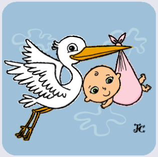
Reichenbach's Principle

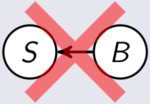
Reichenbach's Principle of Common Cause

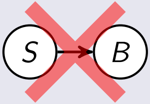
A dependence between X, Y implies that $X \to Y$, $Y \to X$, or there exists a confounder of X and Y (or any combination of these three).

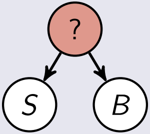
Example

- Significant correlation ($p = 0.008$) between human birth rate and number of stork populations in European countries [Matthews, 2000]
- Most people nowadays do not believe that storks deliver babies (nor that babies deliver storks)
- There must be some confounder explaining the correlation







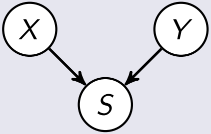


Selection Bias

Reichenbach's Principle may fail in case of selection bias.

If a data set is obtained by only including samples conditional on some event, *selection bias* may be introduced.

Example



X: the battery is empty
Y: the start engine is broken
S: the car does not start

- In general, X and Y are independent events: $X \perp\!\!\!\perp Y$.
- A car mechanic (who only observes cars for which $S = 1$) will observe a dependence between X and Y: $X \not\perp\!\!\!\perp Y \mid S$.
- When the car mechanic invokes Reichenbach's Principle without realizing that he is selecting on the value of S (maybe S is a latent variable), a wrong conclusion will be drawn.

Outline

Assumptions

For simplicity, in this lecture we restrict our attention to a subclass of causal models.

Causal Bayesian Networks: Assumptions

Causal Bayesian Networks are a class of causal models that incorporate the following assumptions:

1. No confounding
2. No feedback
3. No selection bias

Extensions of the theory that drop one or more of these assumptions exist (see e.g. the literature on Acyclic Directed Mixed Graphs, Semi-Markov Causal Models, Maximal Ancestral Graphs, Structural Equation Models, d-connection graphs). This is an active area of research.

Bayesian Networks

Definition

A Bayesian Network is a pair $(\mathcal{G}, p)$ where:

- $\mathcal{G}$ is a Directed Acyclic Graph
- p is a joint probability density on the nodes $X_1, \ldots, X_N$ of $\mathcal{G}$ s.t.

$$p(x_1, \ldots, x_N) = \prod_{i=1}^{N} p(x_i \mid \boldsymbol{x}_{\mathrm{pa}(i)})$$

where $\mathrm{pa}(i)$ are the parents of X_i in $\mathcal{G}$.

Definition

A Bayesian Network is causal if:

- Directed edges correspond with direct causal relations
- After a perfect intervention $\mathrm{do}(\boldsymbol{X}_I = \boldsymbol{x}_I)$, the incoming arrows on $\boldsymbol{X}_I$ are removed and the probability density becomes:

$$p\big(x_{\boldsymbol{V}\setminus\boldsymbol{I}} \mid \mathrm{do}(\boldsymbol{X}_I = \boldsymbol{x}_I)\big) = \prod_{i\in\boldsymbol{V}\setminus\boldsymbol{I}} p(x_i \mid \boldsymbol{x}_{\mathrm{pa}(i)})$$

In other words, a perfect intervention $\mathrm{do}(\boldsymbol{X}_I = \boldsymbol{x}_I)$ on a subset of variables $\boldsymbol{X}_I$ simply "divides out" the conditional densities $p(x_i \mid \boldsymbol{x}_{\mathrm{pa}(i)})$ from the joint density for all $i \in I$, and substitutes the variables $\boldsymbol{X}_I$ by their values $\boldsymbol{x}_I$.

Local Markov Condition

Theorem

For any (Causal) Bayesian Network with variables $\{X_1, \dots, X_N\}$*, the following "Local Markov Condition" holds:*

$$X_i \perp\!\!\!\perp X_{\mathrm{nd}(i)} \,|\, X_{\mathrm{pa}(i)}$$

for all $i = 1, \dots, N$*. Here,* $\mathrm{nd}(i)$ *are the non-descendants of* X_i*.*

Paths and colliders

Definition

Let $\mathcal{G}$ be a DAG with nodes $\boldsymbol{V} = \{X_1, \dots, X_N\}$.

- A path $X_{i_1} \dots X_{i_2} \dots X_{i_k}$ is a sequence of distinct nodes such that X_{i_j} and $X_{i_{j+1}}$ are adjacent (for $j = 1, \dots, k-1$).

Paths and colliders

Definition

Let $\mathcal{G}$ be a DAG with nodes $\boldsymbol{V} = \{X_1, \dots, X_N\}$.

- A path $X_{i_1} \dots X_{i_2} \dots X_{i_k}$ is a sequence of distinct nodes such that X_{i_j} and $X_{i_{j+1}}$ are adjacent (for $j = 1, \dots, k-1$).
- A collider on a path is a (non-endpoint) node X_{i_j} ($j = 2, \dots, k-1$) on the path with precisely two "incoming" arrow heads: $X_{i_{j-1}} \to X_{i_j} \leftarrow X_{i_{j+1}}$.

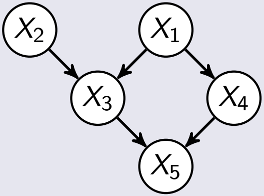
Paths and colliders

Definition

Let $\mathcal{G}$ be a DAG with nodes $\boldsymbol{V} = \{X_1, \dots, X_N\}$.

- A path $X_{i_1} \dots X_{i_2} \dots X_{i_k}$ is a sequence of distinct nodes such that X_{i_j} and $X_{i_{j+1}}$ are adjacent (for $j = 1, \dots, k-1$).
- A collider on a path is a (non-endpoint) node X_{i_j} ($j = 2, \dots, k-1$) on the path with precisely two "incoming" arrow heads: $X_{i_{j-1}} \to X_{i_j} \leftarrow X_{i_{j+1}}$.
- A non-collider on a path is any (non-endpoint) node X_{i_j} ($j = 2, \dots, k-1$) on the path which is not a collider.

Example



$X_1 \to X_3 \leftarrow X_1$ is not a path.
$X_2 \to X_3 \leftarrow X_1$ is a path.
$X_1 \to X_3 \to X_5 \leftarrow X_4 \leftarrow X_1$ is not a path.
The path $X_3 \to X_5 \leftarrow X_4$ contains a collider X_5.
The path $X_4 \leftarrow X_1 \to X_3$ contains no collider.

Blocked paths

Definition

Let $\mathcal{G}$ be a directed graph with nodes $\boldsymbol{V}$. Given a path p between nodes X and Y in $\boldsymbol{V}$, and a set of nodes $\boldsymbol{Z} \subseteq \boldsymbol{V} \setminus \{X, Y\}$, we say that $\boldsymbol{Z}$ blocks p if p contains

- a non-collider which is in $\boldsymbol{Z}$, or
- a collider which is *not* an ancestor of $\boldsymbol{Z}$.

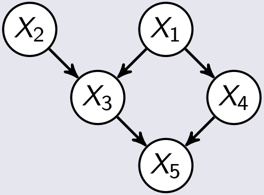
Blocked paths

Definition

Let $\mathcal{G}$ be a directed graph with nodes $\boldsymbol{V}$. Given a path p between nodes X and Y in $\boldsymbol{V}$, and a set of nodes $\boldsymbol{Z} \subseteq \boldsymbol{V} \setminus \{X, Y\}$, we say that $\boldsymbol{Z}$ blocks p if p contains

- a non-collider which is in $\boldsymbol{Z}$, or
- a collider which is *not* an ancestor of $\boldsymbol{Z}$.

Example



$X_3 \to X_5 \leftarrow X_4$ is blocked by $\emptyset$.
$X_3 \to X_5 \leftarrow X_4$ is blocked by $\{X_1\}$.
$X_3 \to X_5 \leftarrow X_4$ is not blocked by $\{X_5\}$.
$X_3 \leftarrow X_1 \to X_4$ is not blocked by $\emptyset$.
$X_2 \to X_3 \leftarrow X_1 \to X_4$ is blocked by $\{X_1\}$.
$X_2 \to X_3 \leftarrow X_1 \to X_4$ is not blocked by $\{X_5\}$.

d-separation

Let $\mathcal{G}$ be a directed graph with nodes $\boldsymbol{V}$.

Definition

Given two distinct nodes $X, Y \in \boldsymbol{V}$ and a set of nodes $\boldsymbol{Z} \subseteq \boldsymbol{V} \setminus \{X, Y\}$, we say that X and Y are *d*-separated by $\boldsymbol{Z}$ iff all paths between X and Y are blocked by $\boldsymbol{Z}$.
For three disjoint subsets $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z} \subseteq \boldsymbol{V}$ of nodes, we say that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are *d*-separated by $\boldsymbol{Z}$ iff all paths between any node in $\boldsymbol{X}$ and any node in $\boldsymbol{Y}$ are blocked by $\boldsymbol{Z}$.

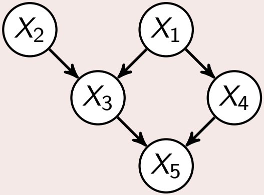
d-separation

Let $\mathcal{G}$ be a directed graph with nodes $\boldsymbol{V}$.

Definition

Given two distinct nodes $X, Y \in \boldsymbol{V}$ and a set of nodes $\boldsymbol{Z} \subseteq \boldsymbol{V} \setminus \{X, Y\}$, we say that X **and** Y **are** ***d*-separated by** $\boldsymbol{Z}$ iff all paths between X and Y are blocked by $\boldsymbol{Z}$.
For three disjoint subsets $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z} \subseteq \boldsymbol{V}$ of nodes, we say that $\boldsymbol{X}$ **and** $\boldsymbol{Y}$ **are** ***d*-separated by** $\boldsymbol{Z}$ iff all paths between any node in $\boldsymbol{X}$ and any node in $\boldsymbol{Y}$ are blocked by $\boldsymbol{Z}$.

Example



X_2 and X_1 are *d*-separated by $\emptyset$.
X_2 and X_1 are *d*-separated by X_4.
X_2 and X_1 are not *d*-separated by X_5.
X_3 and X_4 are not *d*-separated by $\emptyset$.
X_3 and X_4 are *d*-separated by X_1.
X_3 and X_4 are not *d*-separated by $\{X_1, X_5\}$.

Global Markov Condition

Theorem

In any (Causal) Bayesian Network, the following "Global Markov Condition" holds:

$$\boldsymbol{X}, \boldsymbol{Y} \text{ d-separated by } \boldsymbol{Z} \qquad \Longrightarrow \qquad \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$$

for all disjoint subsets $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$ *of nodes.*

In other words, we can read of conditional independences from the graph of a Bayesian Network by using the Global Markov Condition.

Outline

Identifiability

Given i.i.d. data of the observational distribution $p(x, y, \dots)$.
From this we can estimate $p(y \mid X = x)$.

Question

Can we also estimate $p(y \mid \text{do}(X = x))$ from the observational data?

Given enough assumptions, the answer is yes. In that case, we do not have to actually perform the intervention experiment!

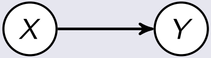
Definition

If a quantity like $p(y \mid \text{do}(X = x))$ can be expressed in terms of the observational distribution $p(x, y, \dots)$, we say that it is identifiable (from the observational distribution).

Example

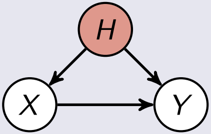
Is $p(y \mid \mathrm{do}(X = x))$ identifiable?

identifiable:



$p(y \mid \mathrm{do}(X = x)) = p(y \mid X = x)$

not identifiable:



$p(y \mid \mathrm{do}(X = x)) \neq p(y \mid X = x)$

Indeed, for the graph with the latent variable H:

$$p(y \mid \mathrm{do}(X = x)) = \int p(h) p(y \mid x, h) \, dh$$

which is generally different from

$$p(y \mid X = x) = \frac{\int p(h) p(x \mid h) p(y \mid x, h) \, dh}{\int p(h) p(x \mid h) p(y \mid x, h) \, dh \, dx}.$$

Adjustment for covariates

- For a Causal Bayesian Network in which all variables are observed:

$$p(y \mid \mathrm{do}(X = x), \boldsymbol{x}_{\mathrm{pa}(X)}) = p(y \mid x, \boldsymbol{x}_{\mathrm{pa}(X)})$$

  and therefore:

$$p(y \mid \mathrm{do}(X = x)) = \int p(y \mid x, \boldsymbol{x}_{\mathrm{pa}(X)}) p(\boldsymbol{x}_{\mathrm{pa}(X)}) \, d\boldsymbol{x}_{\mathrm{pa}(X)}$$

Adjustment for covariates

- For a Causal Bayesian Network in which all variables are observed:

$$p(y \mid \mathrm{do}(X = x), \boldsymbol{x}_{\mathrm{pa}(X)}) = p(y \mid x, \boldsymbol{x}_{\mathrm{pa}(X)})$$

  and therefore:

$$p(y \mid \mathrm{do}(X = x)) = \int p(y \mid x, \boldsymbol{x}_{\mathrm{pa}(X)}) p(\boldsymbol{x}_{\mathrm{pa}(X)}) \, d\boldsymbol{x}_{\mathrm{pa}(X)}$$

- So $p(y \mid \mathrm{do}(X = x))$ is identifiable in Causal Bayesian Networks without latent variables.

Adjustment for covariates

- For a Causal Bayesian Network in which all variables are observed:

$$p(y \mid \mathrm{do}(X = x), \boldsymbol{x}_{\mathrm{pa}(X)}) = p(y \mid x, \boldsymbol{x}_{\mathrm{pa}(X)})$$

  and therefore:

$$p(y \mid \mathrm{do}(X = x)) = \int p(y \mid x, \boldsymbol{x}_{\mathrm{pa}(X)}) p(\boldsymbol{x}_{\mathrm{pa}(X)}) \, d\boldsymbol{x}_{\mathrm{pa}(X)}$$

- So $p(y \mid \mathrm{do}(X = x))$ is identifiable in Causal Bayesian Networks without latent variables.
- Which other sets (instead of the parents of X) could we use to express the causal effect on Y of intervening on X in terms of the observed distribution?

Adjustment for covariates

- For a Causal Bayesian Network in which all variables are observed:

$$p(y \mid \mathrm{do}(X = x), \boldsymbol{x}_{\mathrm{pa}(X)}) = p(y \mid x, \boldsymbol{x}_{\mathrm{pa}(X)})$$

  and therefore:

$$p(y \mid \mathrm{do}(X = x)) = \int p(y \mid x, \boldsymbol{x}_{\mathrm{pa}(X)}) p(\boldsymbol{x}_{\mathrm{pa}(X)}) \, d\boldsymbol{x}_{\mathrm{pa}(X)}$$

- So $p(y \mid \mathrm{do}(X = x))$ is identifiable in Causal Bayesian Networks without latent variables.
- Which other sets (instead of the parents of X) could we use to express the causal effect on Y of intervening on X in terms of the observed distribution?
- A sufficient condition is given by Pearl's Back-door criterion.

The Back-door Criterion

The following result is known as the “Back-door Criterion”:

Theorem

A set $\boldsymbol{S}$ of nodes is “admissible” for adjustment to find the causal effect of X on Y, if :

1. $X, Y \notin \boldsymbol{S}$;
2. *no element of $\boldsymbol{S}$ is a descendant of X;*
3. *$\boldsymbol{S}$ blocks all back-door paths $X \leftarrow \dots Y$.*

In that case,

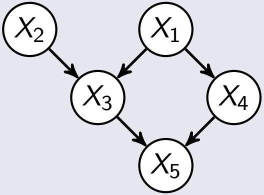
$$p(y \mid \mathrm{do}(X = x)) = \int p(y \mid x, \boldsymbol{s}) p(\boldsymbol{s}) \, d\boldsymbol{s}.$$

For the special case $\boldsymbol{S} = \emptyset$, this simply should be read as:

$$p(y \mid \mathrm{do}(X = x)) = p(y \mid x).$$

The Back-door Criterion: Example

Example



- $\{X_1\}$ is admissible for adjustment to find the causal effect of X_4 on X_5.
- $\emptyset$ is admissible for adjustment to find the causal effect of X_2 on X_5.
- $\{X_1\}$ is admissible for adjustment to find the causal effect of X_2 on X_5.
- $\{X_1, X_4\}$ is admissible for adjustment to find the causal effect of X_2 on X_5.
- $\{X_3\}$ is not admissible for adjustment to find the causal effect of X_2 on X_5.
- $\{X_1, X_3\}$ is admissible for adjustment to find the causal effect of X_5 on X_2.

Exercise 2

Please make Exercise 2...

Causal vs. probabilistic reasoning

Traditional statistics, machine learning

- About **associations** (stork population and human birth rate are correlated)

Causality

- About **causation** (storks do not causally affect human birth rate)

Traditional statistics, machine learning

- About **associations** (stork population and human birth rate are correlated)
- Model the **distribution** of the data

Causality

- About **causation** (storks do not causally affect human birth rate)
- Model the **mechanism** that generates the data

Causal vs. probabilistic reasoning

Traditional statistics, machine learning

- About **associations** (stork population and human birth rate are correlated)
- Model the **distribution** of the data
- Predict given **observations** (if we **observe** a certain number of storks, what is our best estimate of human birth rate?)

Causality

- About **causation** (storks do not causally affect human birth rate)
- Model the **mechanism** that generates the data
- Predict results of **interventions** (if we **change** the number of storks, what will happen with the human birth rate?)

Thank you for your attention!

- Pearl, J. (1999).
  Simpson's paradox: An anatomy.
  Technical Report R-264, UCLA Cognitive Systems Laboratory.
- Pearl, J. (2000).
  Causality: Models, Reasoning, and Inference.
  Cambridge University Press.
- Pearl, J. (2009).
  Causal inference in statistics: An overview.
  Statistics Surveys, 3:96–146.
- Spirtes, P., Glymour, C., and Scheines, R. (2000).
  Causation, Prediction, and Search.
  The MIT Press.