ASCI APR lecture *Causal Modelling*

Joris Mooij
j.m.mooij@uva.nl

Informatics Institute

University of Amsterdam

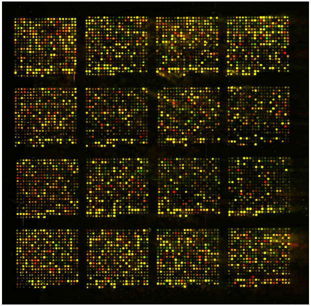
April 17th, 2015

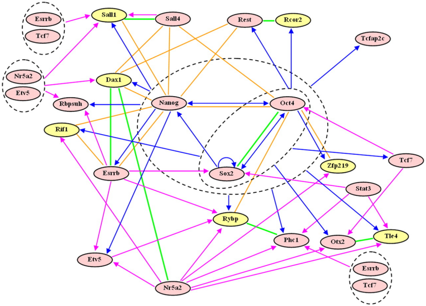
Outline

1. **Introduction**
2. Structural Causal Models
3. Causal Bayesian Networks
4. Back-door Criterion

Causality: ubiquitous in the sciences

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



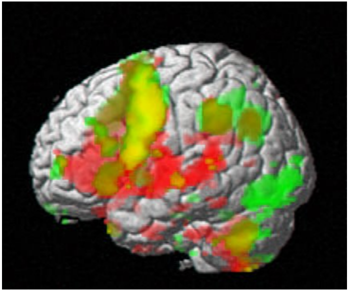


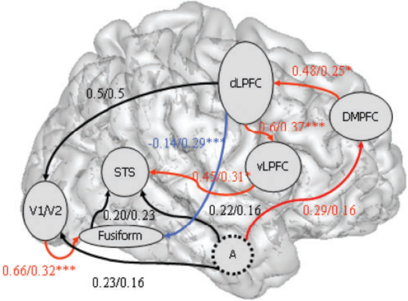
Causality: ubiquitous in the sciences

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

Causality: ubiquitous in the sciences

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





Causality: ubiquitous in the sciences

Economy:
Does austerity reduce national debt?

Causality: what is it?

Causality is central notion in:

- reasoning
- science
- policy decisions
- . . .

What is the "logic" of cause and effect?
(We don't learn this at school!)

Question: give a definition of cause and effect.

Hume on Causality

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

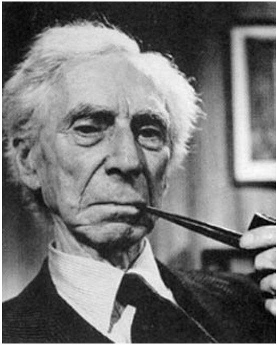
"Thus we remember to have seen that species of object we call *flame*, and to have felt that species of sensation we call *heat*. We likewise call to mind their constant conjunction in all past instances. Without any farther ceremony, we call the one *cause* and the other *effect*, and infer the existence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

Russell on Causality

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

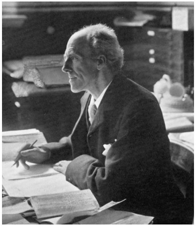


"All philosophers, of every school, imagine that causation is one of the fundamental axioms or postulates of science, yet, oddly enough, in advanced sciences such as gravitational astronomy, the word 'cause' never occurs. The law of causality, I believe, like much that passes muster among philosophers, 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*

Causality in Statistics

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

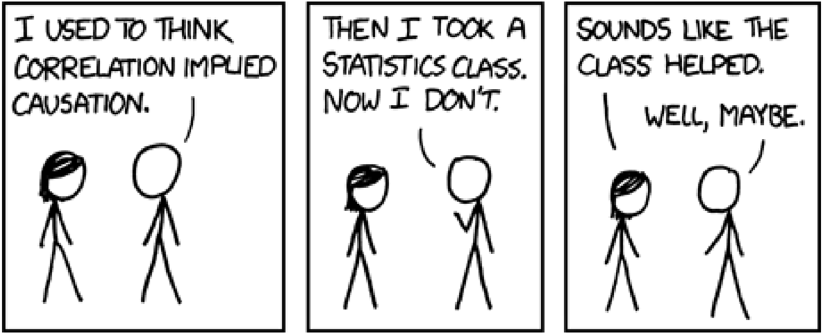


> “Beyond such discarded fundamentals as ‘matter’ and ‘force’ lies still another fetish amidst the inscrutable arcana of even modern science, namely, the category of cause and effect.”
>
> Karl Pearson, *The Grammar of Science*

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


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Randall Munroe, www.xkcd.org

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 data from a biobank (e.g., the EPD) 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:

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

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

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

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

Should we use this drug for treatment?

Problems in formalizing causal reasoning: probabilities

Example (Simpson's paradox)

We collect data from a biobank (e.g., the EPD) 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:

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

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

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

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

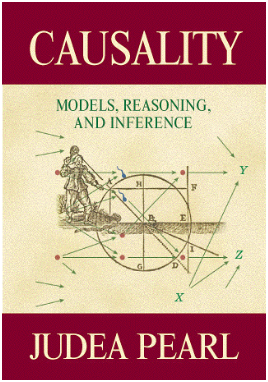
Should we use this drug for treatment?

Note

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

An important step forwards

Judea Pearl



CAUSALITY

MODELS, REASONING, AND INFERENCE

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 dealing with *observations*: conditioning.
- Pearl extends probability calculus by introducing a new operator for describing *interventions*, the **do-operator**.

Example (Do-operator)

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

Resolution:

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

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**:

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

2. **Action/observation exchange**:

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

3. **Ignoring actions**:

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

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

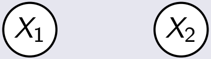
Outline

1. Introduction
2. **Structural Causal Models**
3. Causal Bayesian Networks
4. Back-door Criterion

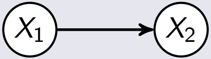
Causal graphs

- We can express the causal relationships between a set of variables $X_1, \ldots, X_N$ in terms of a directed graph, the causal graph.
- A directed edge $X_i \to X_j$ means that X_i is a direct cause of X_j (relative to $X_1, \ldots, X_N$), i.e., not mediated via other variables $X_1, \ldots, X_N$. X_i is called a parent of X_j, X_j is called a child of X_i.

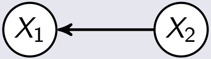
Example



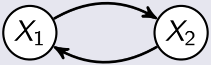
X_1 and X_2 are unrelated



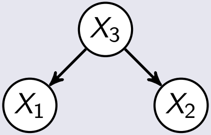
X_1 causes X_2



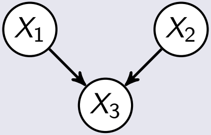
X_2 causes X_1



X_1 and X_2 cause each other



X_1 and X_2 have a common cause

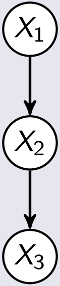


X_1 and X_2 have a common effect

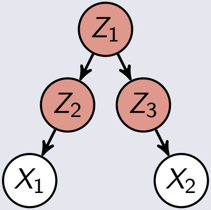
Causal graphs: ancestor relationships

- If $X_{i_1} \to X_{i_2} \to X_{i_3} \to \cdots \to X_{i_n}$ then we say that X_{i_1} is an ancestor of X_{i_n} and X_{i_n} is a descendant of X_{i_1}.
- If Z is an unobserved common ancestor of X_i and X_j then we call Z a confounder of X_i and X_j.

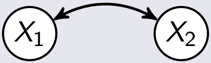
Example



X_1 is ancestor of X_3



Z_1 confounds X_1 and X_2



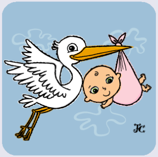
X_1 and X_2 are confounded

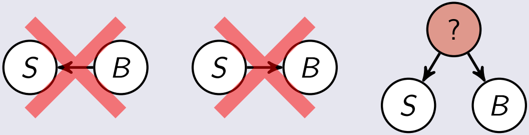
Confounders

- A correlation between X, Y may be explained by direct causal relation $X \to Y$ or $Y \to X$, or a confounder, or a combination of these.
- Another explanation of a correlation is selection bias.

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





Causal feedback

Definition: causal feedback

A SCM incorporates causal feedback if its graph contains a directed cycle

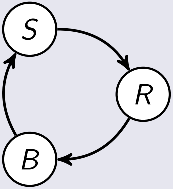
$$X_{i_0} \to X_{i_1} \to \cdots \to X_{i_n}, \qquad X_{i_0} = X_{i_n}$$

If it does not contain such a directed cycle, the model is called acyclic.

Example

In economy, causal feedback is often present:

R: risks taken by bank;
B: imminent bankruptcy;
S: saved by the government.



Structural Causal Models: Definition

Can be traced back to S. Wright's *path diagrams* (1921) and Structural Equation Models in the social sciences.

Structural Causal Models: Definition

Can be traced back to S. Wright's *path diagrams* (1921) and Structural Equation Models in the social sciences.

Definition (Pearl, 2000)

A Structural Causal Model (SCM) is defined by:

1. N observed random variables $X_1, \dots, X_N$ and N latent random variables $E_1, \dots, E_N$

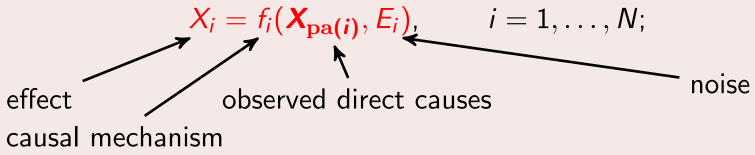
Structural Causal Models: Definition

Can be traced back to S. Wright's *path diagrams* (1921) and Structural Equation Models in the social sciences.

Definition (Pearl, 2000)

A Structural Causal Model (SCM) is defined by:

1. N observed random variables $X_1, \dots, X_N$ and N latent random variables $E_1, \dots, E_N$
2. N structural equations:



$$X_i = f_i(\boldsymbol{X}_{\mathbf{pa}(\boldsymbol{i})}, E_i), \qquad i = 1, \dots, N;$$

effect

causal mechanism

observed direct causes

noise

where the subsets $\mathrm{pa}(i) \subseteq \{1, \dots, N\}$ define the observed direct causes of X_i (the parents of X_i),

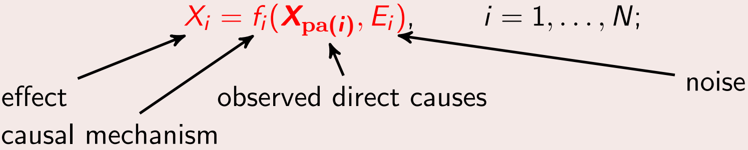
Structural Causal Models: Definition

Can be traced back to S. Wright's *path diagrams* (1921) and Structural Equation Models in the social sciences.

Definition (Pearl, 2000)

A Structural Causal Model (SCM) is defined by:

1. N observed random variables $X_1, \dots, X_N$ and N latent random variables $E_1, \dots, E_N$
2. N structural equations:



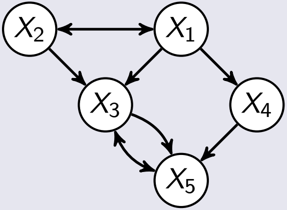
$$X_i = f_i(\boldsymbol{X}_{\mathbf{pa}(\boldsymbol{i})}, E_i), \qquad i = 1, \dots, N;$$

   where the subsets $\mathrm{pa}(i) \subseteq \{1, \dots, N\}$ define the observed direct causes of X_i (the parents of X_i),
3. a joint probability distribution $p(E_1, \dots, E_N)$ on latent variables.

Structural Causal Models: Example

Example

i	$\mathrm{pa}(i)$	$X_i = f_i(\boldsymbol{X}_{\mathrm{pa}(i)}, E_i)$
1	$\emptyset$	$X_1 = f_1(E_1)$
2	$\emptyset$	$X_2 = f_2(E_2)$
3	$\{1,2\}$	$X_3 = f_3(X_1, X_2, E_3)$
4	$\{1\}$	$X_4 = f_4(X_1, E_4)$
5	$\{3,4\}$	$X_5 = f_5(X_3, X_4, E_5)$



$$p(E_1, \dots, E_5) = p(E_1, E_2)p(E_3, E_5)p(E_4)$$

- Directed arrows (from X_j to X_i if $j \in \mathrm{pa}(i)$) correspond with functional dependences and are interpreted as direct causal relations.
- Bidirected arrows between noise variables indicate statistical dependences between noise variables.
- Usually, noise variables are not depicted explicitly.

Modeling interventions in a SCM

For a *causal* model, we need to specify how we model *interventions*.

Interventions in SCMs

An intervention $\mathrm{do}(\boldsymbol{X}_I = \xi_I)$ on a set of variables $\boldsymbol{X}_I$ with $I \subseteq \{1, \dots, N\}$, forcing them to attain the value ξ_I, changes the structural equations as follows:

Original SCM $\mathcal{M}$:

$$\begin{aligned} X_i &= f_i(\boldsymbol{X}_{\mathrm{pa}(i)}, E_i) & \forall i \in I \\ X_j &= f_j(\boldsymbol{X}_{\mathrm{pa}(j)}, E_j) & \forall j \notin I \end{aligned}$$

$$p(\boldsymbol{E}) = \dots$$

Intervened SCM $\mathcal{M}_{\xi_I}$:

$$\begin{aligned} X_i &= \xi_i & \forall i \in I \\ X_j &= f_j(\boldsymbol{X}_{\mathrm{pa}(j)}, E_j) & \forall j \notin I \end{aligned}$$

$$p(\boldsymbol{E}) = \dots$$

Modeling interventions in a SCM

For a *causal* model, we need to specify how we model *interventions*.

Interventions in SCMs

An intervention $\mathrm{do}(\boldsymbol{X}_I = \boldsymbol{\xi}_I)$ on a set of variables $\boldsymbol{X}_I$ with $I \subseteq \{1, \dots, N\}$, forcing them to attain the value $\boldsymbol{\xi}_I$, changes the structural equations as follows:

Original SCM $\mathcal{M}$:

$$\begin{aligned} X_i &= f_i(\boldsymbol{X}_{\mathrm{pa}(i)}, E_i) & \forall i \in I \\ X_j &= f_j(\boldsymbol{X}_{\mathrm{pa}(j)}, E_j) & \forall j \notin I \\ p(\boldsymbol{E}) &= \dots \end{aligned}$$

Intervened SCM $\mathcal{M}_{\boldsymbol{\xi}_I}$:

$$\begin{aligned} X_i &= \xi_i & \forall i \in I \\ X_j &= f_j(\boldsymbol{X}_{\mathrm{pa}(j)}, E_j) & \forall j \notin I \\ p(\boldsymbol{E}) &= \dots \end{aligned}$$

- Interpretation: overriding default causal mechanisms that normally would determine the values of the intervened variables.

Modeling interventions in a SCM

For a *causal* model, we need to specify how we model *interventions*.

Interventions in SCMs

An intervention $\mathrm{do}(\boldsymbol{X}_I = \xi_I)$ on a set of variables $\boldsymbol{X}_I$ with $I \subseteq \{1, \dots, N\}$, forcing them to attain the value ξ_I, changes the structural equations as follows:

Original SCM $\mathcal{M}$:

$$\begin{aligned} X_i &= f_i(\boldsymbol{X}_{\mathrm{pa}(i)}, E_i) && \forall i \in I \\ X_j &= f_j(\boldsymbol{X}_{\mathrm{pa}(j)}, E_j) && \forall j \notin I \\ p(\boldsymbol{E}) &= \dots \end{aligned}$$

Intervened SCM $\mathcal{M}_{\xi_I}$:

$$\begin{aligned} X_i &= \xi_i && \forall i \in I \\ X_j &= f_j(\boldsymbol{X}_{\mathrm{pa}(j)}, E_j) && \forall j \notin I \\ p(\boldsymbol{E}) &= \dots \end{aligned}$$

- Interpretation: overriding default causal mechanisms that normally would determine the values of the intervened variables.
- In the graph of $\mathcal{M}$, the effect of the intervention is to remove all incoming arrows of intervened variables $\{X_i\}_{i \in I}$.

Modeling Interventions: Example

Example

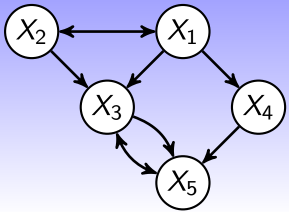
Observational (no intervention):

Structural causal model $\mathcal{M}$:

$$
\begin{aligned}
X_1 &= f_1(E_1)\\
X_2 &= f_2(E_2)\\
X_3 &= f_3(X_1, X_2, E_3)\\
X_4 &= f_4(X_1, E_4)\\
X_5 &= f_5(X_3, X_4, E_5)
\end{aligned}
$$

$$p(E_1, \dots, E_5) = p(E_1, E_2)p(E_3, E_5)p(E_4)$$

Causal graph $\mathcal{G}_{\mathcal{M}}$:



Modeling Interventions: Example

Example

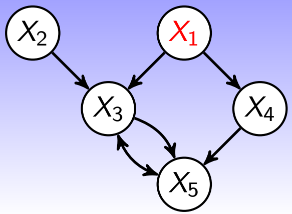
Intervention $\mathrm{do}(X_1 = \xi_1)$:

Structural causal model $\mathcal{M}_{\xi_1}$:

$$X_1 = \xi_1$$
$$X_2 = f_2(E_2)$$
$$X_3 = f_3(X_1, X_2, E_3)$$
$$X_4 = f_4(X_1, E_4)$$
$$X_5 = f_5(X_3, X_4, E_5)$$

$$p(E_1, \dots, E_5) = p(E_1, E_2)p(E_3, E_5)p(E_4)$$

Causal graph $\mathcal{G}_{\mathcal{M}_{\xi_1}}$:



Modeling Interventions: Example

Example

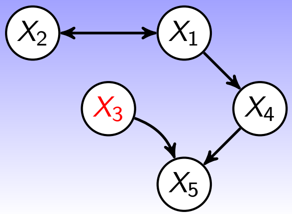
Intervention $\mathrm{do}(X_3 = \xi_3)$:

Structural causal model $\mathcal{M}_{\xi_3}$:

$$
\begin{aligned}
X_1 &= f_1(E_1) \\
X_2 &= f_2(E_2) \\
X_3 &= \xi_3 \\
X_4 &= f_4(X_1, E_4) \\
X_5 &= f_5(X_3, X_4, E_5)
\end{aligned}
$$

$$p(E_1, \dots, E_5) = p(E_1, E_2)p(E_3, E_5)p(E_4)$$

Causal graph $\mathcal{G}_{\mathcal{M}_{\xi_3}}$:



Exercise 2

Please make Exercise 2. . .

Outline

1. Introduction
2. Structural Causal Models
3. **Causal Bayesian Networks**
4. Back-door Criterion

Causal sufficiency

Definition: Confounder

A confounder is an unobserved variable that is an ancestor of at least two endogenous variables (a "hidden common cause").

Absence of confounders implies causal sufficiency.

Definition: Causal Sufficiency

If all noise variables in an SCM are jointly independent, i.e., if the joint probability distribution $p(\boldsymbol{E})$ of the noise variables factorizes:

$$p(\boldsymbol{E}) = \prod_{i=1}^{N} p(E_i)$$

then we say that the variables $\boldsymbol{X}$ are causally sufficient.

Markovian SCMs

Definition

An SCM $\mathcal{M}$ is called Markovian if

1. it is acyclic (“no causal feedback”);
2. it is causally sufficient (“no hidden common causes”).

Its causal graph is a Directed Acyclic Graph (DAG).

Markovian SCMs

Definition

An SCM $\mathcal{M}$ is called Markovian if

1. it is acyclic ("no causal feedback");
2. it is causally sufficient ("no hidden common causes").

Its causal graph is a Directed Acyclic Graph (DAG).

- Markovian SCMs are easier to handle than non-Markovian SCMs; this is why we will focus on these for the rest of this talk.
- Non-Markovian SCMs are an active research topic, and the theory for these cases is far from complete.
- Markovian SCMs are related to Causal Bayesian networks.

Bayesian Networks

Definition: Bayesian Network

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

- $\mathcal{G}$ is a Directed Acyclic Graph
- p is a probability distribution on the nodes $X_1, \dots, X_N$ of $\mathcal{G}$ such that

$$p(X_1, \dots, X_N) = \prod_{i=1}^{N} p(X_i \mid \boldsymbol{X}_{\mathrm{pa}(i)})$$

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

Causal Bayesian Networks

Definition: Causal Bayesian Network

A Bayesian Network is causal if:

- Arrows correspond with direct causal relations
- After an intervention $\mathrm{do}(\boldsymbol{X}_I = \boldsymbol{\xi}_I)$, the incoming arrows on $\boldsymbol{X}_I$ are removed and the probability distribution becomes:

$$p\big(X_1, \dots, X_N \,|\, \mathrm{do}(\boldsymbol{X}_I = \boldsymbol{\xi}_I)\big) = \prod_{\substack{i=1 \\ i \notin I}}^{N} p(X_i \,|\, \boldsymbol{X}_{\mathrm{pa}(i)}) \prod_{i \in I} \mathbf{1}_{[X_i = \xi_i]}$$

The Causal Markov Condition

Theorem: Causal Markov Condition

Any probability distribution induced by a Markovian SCM $\mathcal{M}$ can be factorized as:

$$p(X_1, \dots, X_N) = \prod_{i=1}^{N} p(X_i \mid \boldsymbol{X}_{\mathrm{pa}(i)})$$

The Causal Markov Condition

Theorem: Causal Markov Condition

Any probability distribution induced by a Markovian SCM $\mathcal{M}$ can be factorized as:

$$p(X_1, \dots, X_N) = \prod_{i=1}^{N} p(X_i \,|\, \boldsymbol{X}_{\mathrm{pa}(i)})$$

The proof proceeds by marginalization over the noise variables $\boldsymbol{E}$:

$$p(\boldsymbol{X}) = \int p(\boldsymbol{X}, \boldsymbol{E})\, d\boldsymbol{E} = \int \left(\prod_{i=1}^{N} \delta\big(X_i - f_i(\boldsymbol{X}_{\mathrm{pa}(i)}, E_i)\big)\right) \left(\prod_{i=1}^{N} p(E_i)\right) d\boldsymbol{E}$$

$$\overset{*}{=} \prod_{i=1}^{N} \int \delta\big(X_i - f_i(X_{\mathrm{pa}(i)}, E_i)\big)\, p(E_i)\, dE_i = \prod_{i=1}^{N} p(X_i \,|\, \boldsymbol{X}_{\mathrm{pa}(i)})$$

where we used the acyclicity in the step marked with a $*$.

Truncated factorization

Theorem: Truncated factorization

Any probability distribution induced by a Markovian SCM $\mathcal{M}$ can be factorized as:

$$p(X_1, \dots, X_N) = \prod_{i=1}^{N} p(X_i \mid \boldsymbol{X}_{\mathrm{pa}(i)})$$

After an intervention $\mathrm{do}(\boldsymbol{X}_I = \boldsymbol{\xi}_I)$, the probability distribution becomes:

$$p\big(X_1, \dots, X_N \mid \mathrm{do}(\boldsymbol{X}_I = \boldsymbol{\xi}_I)\big) = \prod_{\substack{i=1 \\ i \notin I}}^{N} p(X_i \mid \boldsymbol{X}_{\mathrm{pa}(i)}) \prod_{i \in I} \mathbf{1}_{[X_i = \xi_i]}$$

Truncated factorization

Theorem: Truncated factorization

Any probability distribution induced by a Markovian SCM $\mathcal{M}$ can be factorized as:

$$p(X_1, \dots, X_N) = \prod_{i=1}^{N} p(X_i \mid \boldsymbol{X}_{\mathrm{pa}(i)})$$

After an intervention $\mathrm{do}(\boldsymbol{X}_I = \boldsymbol{\xi}_I)$, the probability distribution becomes:

$$p\big(X_1, \dots, X_N \mid \mathrm{do}(\boldsymbol{X}_I = \boldsymbol{\xi}_I)\big) = \prod_{\substack{i=1 \\ i \notin I}}^{N} p(X_i \mid \boldsymbol{X}_{\mathrm{pa}(i)}) \prod_{i \in I} \mathbf{1}_{[X_i = \xi_i]}$$

- Each Markovian SCM induces a Causal Bayesian network.
- Conversely, for any given Causal Bayesian network, one can construct an equivalent Markovian SCM.
- SCMs are more general than Causal Bayesian Networks (can deal with confounders, feedback, allow us to define counterfactuals).

Outline

1. Introduction
2. Structural Causal Models
3. Causal Bayesian Networks
4. **Back-door Criterion**

Identifiability

- Suppose that we have i.i.d. data of the observational distribution $p(X, Y, \dots)$. From this, we can estimate $p(Y \mid X)$.

Identifiability

- Suppose that we have i.i.d. data of the observational distribution $p(X, Y, \dots)$. From this, we can estimate $p(Y \mid X)$.
- In general, however, $p(Y \mid \mathrm{do}(X)) \neq p(Y \mid X)$.

Identifiability

- Suppose that we have i.i.d. data of the observational distribution $p(X, Y, \dots)$. From this, we can estimate $p(Y \mid X)$.
- In general, however, $p(Y \mid \mathrm{do}(X)) \neq p(Y \mid X)$.
- How to estimate $p(Y \mid \mathrm{do}(X))$ from data?

Identifiability

- Suppose that we have i.i.d. data of the observational distribution $p(X, Y, \dots)$. From this, we can estimate $p(Y \mid X)$.
- In general, however, $p(Y \mid \mathrm{do}(X)) \neq p(Y \mid X)$.
- How to estimate $p(Y \mid \mathrm{do}(X))$ from data?
- Sometimes (given enough assumptions), $p(Y \mid \mathrm{do}(X))$ *can* be inferred from purely *observational* data $p(X, Y, \dots)$, without the need for actually performing the experiment $\mathrm{do}(X)$.

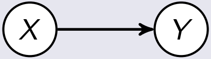
Identifiability

- Suppose that we have i.i.d. data of the observational distribution $p(X, Y, \dots)$. From this, we can estimate $p(Y \mid X)$.
- In general, however, $p(Y \mid \mathrm{do}(X)) \neq p(Y \mid X)$.
- How to estimate $p(Y \mid \mathrm{do}(X))$ from data?
- Sometimes (given enough assumptions), $p(Y \mid \mathrm{do}(X))$ *can* be inferred from purely *observational* data $p(X, Y, \dots)$, without the need for actually performing the experiment $\mathrm{do}(X)$.
- In that case, we say that $p(Y \mid \mathrm{do}(X))$ is identifiable.

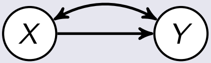
Identifiability

- Suppose that we have i.i.d. data of the observational distribution $p(X, Y, \dots)$. From this, we can estimate $p(Y \mid X)$.
- In general, however, $p(Y \mid \mathrm{do}(X)) \neq p(Y \mid X)$.
- How to estimate $p(Y \mid \mathrm{do}(X))$ from data?
- Sometimes (given enough assumptions), $p(Y \mid \mathrm{do}(X))$ *can* be inferred from purely *observational* data $p(X, Y, \dots)$, without the need for actually performing the experiment $\mathrm{do}(X)$.
- In that case, we say that $p(Y \mid \mathrm{do}(X))$ is identifiable.

Example



$p(Y|X) = p(Y|\mathrm{do}(X))$
identifiable



$p(Y|X) \neq p(Y|\mathrm{do}(X))$
not identifiable

Conditions for Identifiability

- Given enough modeling assumptions, the effects of interventions can sometimes be inferred from observational data alone!

Conditions for Identifiability

- Given enough modeling assumptions, the effects of interventions can sometimes be inferred from observational data alone!
- In many cases, the uncertainty about the model is too large (the set A of assumptions is too small) and experimentation becomes necessary.

Conditions for Identifiability

- Given enough modeling assumptions, the effects of interventions can sometimes be inferred from observational data alone!
- In many cases, the uncertainty about the model is too large (the set A of assumptions is too small) and experimentation becomes necessary.
- Can we find a condition which tells us when a causal effect $p(Y \mid \mathrm{do}(X))$ is identifiable?

Conditions for Identifiability

- Given enough modeling assumptions, the effects of interventions can sometimes be inferred from observational data alone!
- In many cases, the uncertainty about the model is too large (the set A of assumptions is too small) and experimentation becomes necessary.
- Can we find a condition which tells us when a causal effect $p(Y \mid \mathrm{do}(X))$ is identifiable?
- A sufficient condition is provided by Pearl's Back-door criterion. To state this, we first need some graph theoretical terminology.

Some graph-theoretical notions

Definition: path, directed path, ancestor and collider

Let $\mathcal{G}$ be a graph with directed ($\leftarrow$, $\rightarrow$) and bidirected ($\leftrightarrow$) edges.

- A path q is a sequence of consecutive edges (where the end node of each edge equals the start node of the next edge).

Some graph-theoretical notions

Definition: path, directed path, ancestor and collider

Let $\mathcal{G}$ be a graph with directed ($\leftarrow$, $\rightarrow$) and bidirected ($\leftrightarrow$) edges.

- A path q is a sequence of consecutive edges (where the end node of each edge equals the start node of the next edge).
- A path in which each edge is of the form $\dots \rightarrow \dots$ is called directed.

Some graph-theoretical notions

Definition: path, directed path, ancestor and collider

Let $\mathcal{G}$ be a graph with directed ($\leftarrow$, $\rightarrow$) and bidirected ($\leftrightarrow$) edges.

- A path q is a sequence of consecutive edges (where the end node of each edge equals the start node of the next edge).
- A path in which each edge is of the form $\dots \rightarrow \dots$ is called directed.
- If there is a directed path from X to Y, X is called a ancestor of Y.

Some graph-theoretical notions

Definition: path, directed path, ancestor and collider

Let $\mathcal{G}$ be a graph with directed ($\leftarrow$, $\rightarrow$) and bidirected ($\leftrightarrow$) edges.

- A path q is a sequence of consecutive edges (where the end node of each edge equals the start node of the next edge).
- A path in which each edge is of the form $\dots \rightarrow \dots$ is called directed.
- If there is a directed path from X to Y, X is called a ancestor of Y.
- A collider on a path q is a node X on q with precisely two "incoming" arrow heads: $\rightarrow X \leftarrow, \quad \rightarrow X \leftrightarrow, \quad \leftrightarrow X \leftarrow, \quad \leftrightarrow X \leftrightarrow$

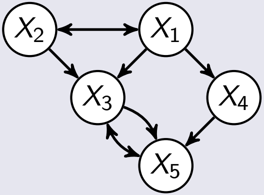
Some graph-theoretical notions

Definition: path, directed path, ancestor and collider

Let $\mathcal{G}$ be a graph with directed ($\leftarrow$, $\rightarrow$) and bidirected ($\leftrightarrow$) edges.

- A path q is a sequence of consecutive edges (where the end node of each edge equals the start node of the next edge).
- A path in which each edge is of the form $\dots \rightarrow \dots$ is called directed.
- If there is a directed path from X to Y, X is called a ancestor of Y.
- A collider on a path q is a node X on q with precisely two "incoming" arrow heads: $\rightarrow X \leftarrow$, $\rightarrow X \leftrightarrow$, $\leftrightarrow X \leftarrow$, $\leftrightarrow X \leftrightarrow$

Example



The sequence $X_1 \rightarrow X_3 \leftarrow X_1$ is not a path.
The sequence $X_1 \leftrightarrow X_2 \rightarrow X_3$ is a path.
X_1, X_2, X_3 and X_4 are ancestors of X_5.
The path $X_3 \rightarrow X_5 \leftarrow X_4$ contains a collider X_5.
The path $X_1 \leftrightarrow X_2 \rightarrow X_3$ contains no collider.

More graph theory: blocking paths

Definition: blocking paths

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

- a non-collider which is in S, or
- a collider which is *not* an ancestor of S.

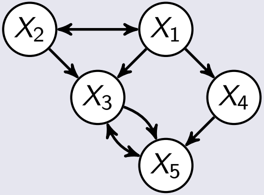
More graph theory: blocking paths

Definition: blocking paths

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

- a non-collider which is in S, or
- a collider which is *not* an ancestor of S.

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

Adjustment for covariates

- In the Markovian case, by using truncated factorization, we can show:

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

and therefore:

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

Adjustment for covariates

- In the Markovian case, by using truncated factorization, we can show:

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

  and therefore:

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

- So $p(Y \mid \mathrm{do}(X))$ is identifiable (in the Markovian case).

Adjustment for covariates

- In the Markovian case, by using truncated factorization, we can show:

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

and therefore:

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

- So $p(Y \mid \mathrm{do}(X))$ is identifiable (in the Markovian case).
- 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 $p(\boldsymbol{X})$?

Adjustment for covariates

- In the Markovian case, by using truncated factorization, we can show:

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

  and therefore:

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

- So $p(Y \mid \mathrm{do}(X))$ is identifiable (in the Markovian case).
- 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 $p(\boldsymbol{X})$?
- A sufficient condition is given by Pearl's *Back-door criterion*.

The Back-door Criterion

Theorem: Back-door criterion

A set S of nodes is "admissible" or "sufficient" for adjustment if

1. no element of S is a descendant of X
2. the elements of S block all back-door paths $X \leftarrow \dots Y$ and $X \leftrightarrow \dots Y$ (paths between X and Y with an arrow pointing to X).

In that case,

$$p(Y \mid \mathrm{do}(X)) = \int p(Y \mid X, \boldsymbol{X}_S) p(\boldsymbol{X}_S) \, d\boldsymbol{X}_S$$

The Back-door Criterion

Theorem: Back-door criterion

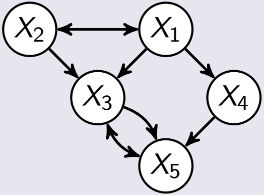
A set S of nodes is “admissible” or “sufficient” for adjustment if

1. no element of S is a descendant of X
2. the elements of S block all back-door paths $X \leftarrow \dots Y$ and $X \leftrightarrow \dots Y$ (paths between X and Y with an arrow pointing to X).

In that case,

$$p(Y \mid \mathrm{do}(X)) = \int p(Y \mid X, \boldsymbol{X}_S) p(\boldsymbol{X}_S) \, d\boldsymbol{X}_S$$

Example



$\{X_1\}$ is sufficient for adjustment to find the causal effect of X_4 on X_5.
$\{X_1\}$ is sufficient for adjustment to find the causal effect of X_2 on X_5.
No set is sufficient for adjustment to find the causal effect of X_3 on X_5.

Exercise 3

Please make Exercise 3. . .

Causal reasoning vs. probabilistic reasoning

Statistics, (most of) Machine Learning

- About associations (correlation between smoking and lung cancer)
- Models the distribution of the data
- Predicting by conditioning (if we *know that somebody smokes*, what is the probability that he/she will get lung cancer?)

Causality

- About causation (smoking causes lung cancer)
- Models the mechanism that generates the data
- Predicting results of interventions (if we *force somebody to smoke*, what is the probability that he/she will get lung cancer?)

Observing $\neq$ intervening: $\quad p(Y \mid X) \neq p(Y \mid \mathrm{do}(X))$

Further reading

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.