Causal Inference from Observational Data Based on Graphical Models

Guido Consonni

O'Bayes 2025, June 8, Athens

Outline

1. From Statistical to Causal Models
2. Causal Modeling
3. Causal Discovery
4. Causal Reasoning
5. Conclusions

Awards

Turing award – 2011

Judea Pearl

Awards

Turing award – 2011

Judea Pearl

"For fundamental contributions to artificial intelligence through the development of a calculus for probabilistic and causal reasoning"

Awards

Turing award – 2011

Judea Pearl

"For fundamental contributions to artificial intelligence through the development of a calculus for probabilistic and causal reasoning"

The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2021

(David Card)

Joshua D. Angrist and Guido W. Imbens

Awards

Turing award – 2011

Judea Pearl

"For fundamental contributions to artificial intelligence through the development of a calculus for probabilistic and causal reasoning"

The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2021

(David Card)

Joshua D. Angrist and Guido W. Imbens

"For their methodological contributions to the analysis of causal relationships"

Rousseeuw Prize for Statistics 2022

James Robins, Miguel Hernán, Thomas Richardson, Andrea Rotnitzky, Eric Tchetgen Tchetgen

"For their pioneering work on Causal Inference with applications in Medicine and Public Health"

Causality and AI/ML

"The kind of causal inference seen in natural human thought can be "algorithmitized" to help produce human-level machine intelligence"
Judea Pearl, 2019, *Communications of the ACM*

Causality and AI/ML

"The kind of causal inference seen in natural human thought can be "algorithmitized" to help produce human-level machine intelligence"
Judea Pearl, 2019, *Communications of the ACM*

"Some of the hard open problems of machine learning and AI are intrinsically related to causality, and progress may require advances in our understanding of how to model and infer causality from data"
Bernhard Schölkopf, 2022, *International Congress of Mathematicians*

Causality and AI/ML

"The kind of causal inference seen in natural human thought can be "algorithmitized" to help produce human-level machine intelligence"
Judea Pearl, 2019, *Communications of the ACM*

"Some of the hard open problems of machine learning and AI are intrinsically related to causality, and progress may require advances in our understanding of how to model and infer causality from data"
Bernhard Schölkopf, 2022, *International Congress of Mathematicians*

Some recent papers

Some recent papers

Preserving Causal Constraints in Counterfactual Explanations for Machine Learning Classifiers
NeurIPS, 2019

Interpretable ML; feasible counterfactuals

Some recent papers

Preserving Causal Constraints in Counterfactual Explanations for Machine Learning Classifiers
NeurIPS, 2019

Interpretable ML; feasible counterfactuals

Causal Reasoning and Large Language Models: Opening a New Frontier for Causality
Transactions on Machine Learning Research, 2024

"Behavorial" study of LLMs to benchmark their capability in generating causal arguments

Some recent papers

Preserving Causal Constraints in Counterfactual Explanations for Machine Learning Classifiers
NeurIPS, 2019

Interpretable ML; feasible counterfactuals

Causal Reasoning and Large Language Models: Opening a New Frontier for Causality
Transactions on Machine Learning Research, 2024

"Behavorial" study of LLMs to benchmark their capability in generating causal arguments

Improving the accuracy of medical diagnosis with causal machine learning
Nature communications, 2020

we reformulate diagnosis as a counterfactual inference task and derive counterfactual diagnostic algorithms.In medical diagnosis a doctor aims to explain a patient's symptoms by determining the diseases causing them, while existing diagnostic algorithms are purely associative

Robust Agents Learn Causal World Models

International Conference on Learning Representations, 2024

"Any agent capable of satisfying a regret bound for a large set of distributional shifts must have learned an approximate causal model of the data generating process"

Robust Agents Learn Causal World Models

International Conference on Learning Representations, 2024

"Any agent capable of satisfying a regret bound for a large set of distributional shifts must have learned an approximate causal model of the data generating process"

Explaining the Behavior of Black-Box Prediction Algorithms with Causal Learning

Transactions on Machine Learning Research, 2025

Causal approaches to post-hoc explainability for black-box prediction models(e.g. deep neural networks trained on image pixel data)

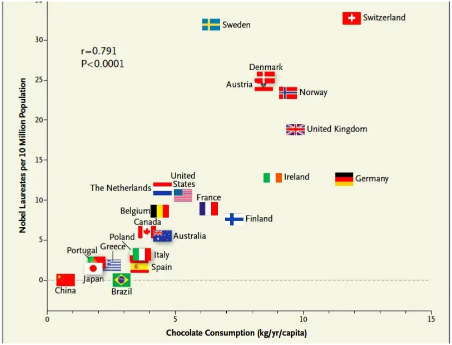
Table of Contents

1. From Statistical to Causal Models
2. Causal Modeling
3. Causal Discovery
4. Causal Reasoning
5. Conclusions

Correlation vs causation

Correlation does not imply causation

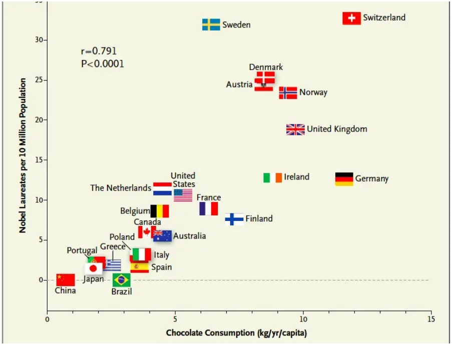
Chocolate and Nobel prize winners



Correlation vs causation

Correlation does not imply causation

Chocolate and Nobel prize winners

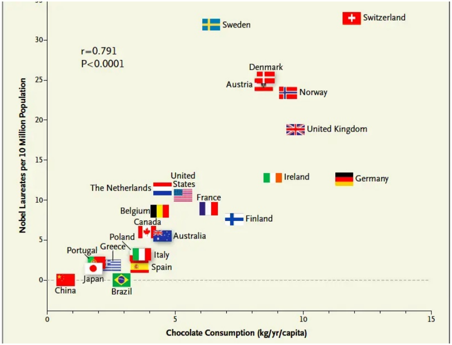


Understanding causation

- Manipulability
- Intervention

Correlation vs causation

Correlation does not imply causation
Chocolate and Nobel prize winners



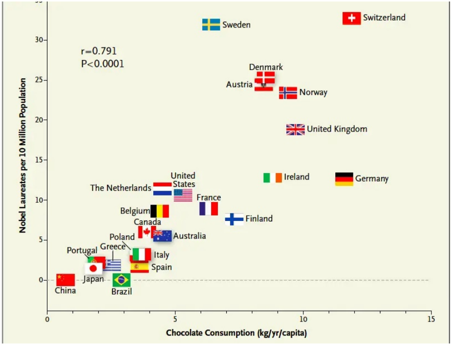
Understanding causation

- Manipulability
- Intervention

J. Woodward (2001). *Causation and manipulability*
J. Pearl (2009). *Causality: models, reasoning, and inference*. 2nd edn

Correlation vs causation

Correlation does not imply causation
Chocolate and Nobel prize winners



Understanding causation

- Manipulability
- Intervention

J. Woodward (2001). *Causation and manipulability*
J. Pearl (2009). *Causality: models, reasoning, and inference*. 2nd edn

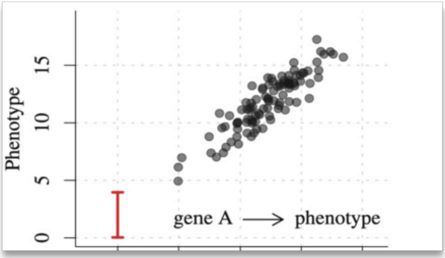
- Epidemiology
  J. M. Robins, M. A. Hernan, and B. Brumback (2000)
- Agriculture
  S. Wright (1921)
- Econometrics
  T. Haavelmo (1944); K. D. Hoover (2001)

Causal effect

Definition

A random variable X has a causal effect on a random variable Y if there exist $x \neq x'$ such that the distribution of Y after intervening on X and setting it to x differs from the distribution of Y after setting X to x'

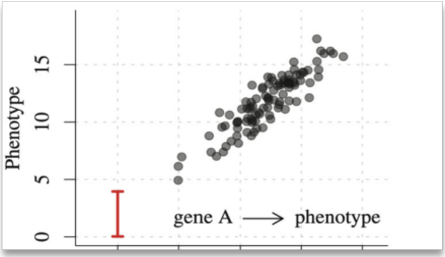
Causal effect

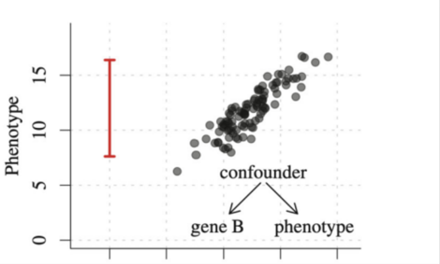


Definition

A random variable X has a causal effect on a random variable Y if there exist $x \neq x'$ such that the distribution of Y after intervening on X and setting it to x differs from the distribution of Y after setting X to x'

Causal effect





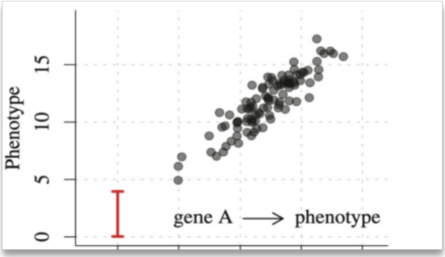
Definition

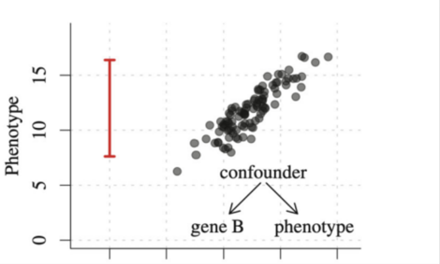
A random variable X has a causal effect on a random variable Y if there exist $x \neq x'$ such that the distribution of Y after intervening on X and setting it to x differs from the distribution of Y after setting X to x'

Causal effect

Definition

A random variable X has a causal effect on a random variable Y if there exist $x \neq x'$ such that the distribution of Y after intervening on X and setting it to x differs from the distribution of Y after setting X to x'



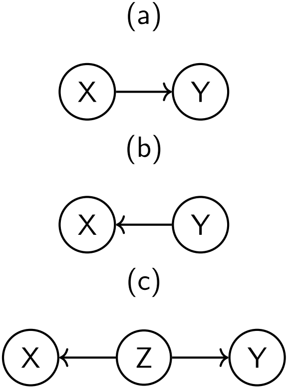


Gene A is correlated with the phenotype, and so is gene B
However only gene A ha a causal effect on the phenotype

Correlation and Causation: what's the connection?

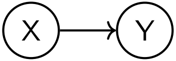
Principle

If two random variables X and Y are statistically dependent $X \not\perp\!\!\!\perp Y$ *then there exists a random variable* Z *which causally influences both of them and which explains all their dependence that is*
$X \perp\!\!\!\perp Y \mid Z$ *(c)*
As a special case, Z *may coincide with* X *or* Y *(a) or (b)*

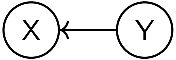


Chocolate consumption and # of Nobel laureates

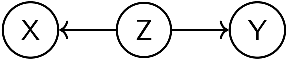
(a) ✗



(b) ✗



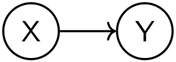
(c) ✓



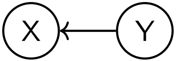
X: Chocolate consumption
Y: # Nobel laureates
Z: Economic factor

Chocolate consumption and # of Nobel laureates

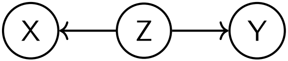
(a) ✗



(b) ✗

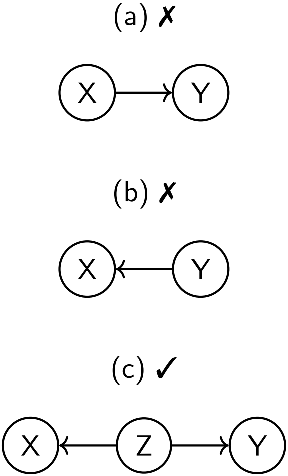


(c) ✓



X: Chocolate consumption
Y: # Nobel laureates
Z: Economic factor

Chocolate consumption and # of Nobel laureates

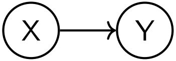


X: Chocolate consumption
Y: # Nobel laureates
Z: Economic factor

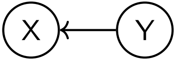
- The class of observational distributions over X and Y that can be realized by these models is the same in all three cases

Chocolate consumption and # of Nobel laureates

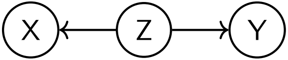
(a) ✗



(b) ✗



(c) ✓

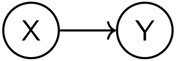


X: Chocolate consumption
Y: # Nobel laureates
Z: Economic factor

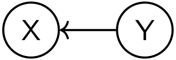
- The class of observational distributions over X and Y that can be realized by these models is the same in all three cases
- Cannot distinguish among a), b) and c) through passive observation i.e., in a purely data-driven way

Chocolate consumption and # of Nobel laureates

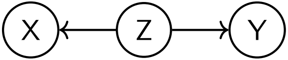
(a) ✗



(b) ✗



(c) ✓



X: Chocolate consumption
Y: # Nobel laureates
Z: Economic factor

- The class of observational distributions over X and Y that can be realized by these models is the same in all three cases
- Cannot distinguish among a), b) and c) through passive observation i.e., in a purely data-driven way
- Z latent confounder
  drives consumer spending and investment in education and research
  [from background knowledge]

Making Sense of Correlation

- Correlation is still useful

Making Sense of Correlation

- Correlation is still useful
- Causality is not always needed

Making Sense of Correlation

- Correlation is still useful
- Causality is not always needed
- Gene A and gene B remain useful features for making predictions

Making Sense of Correlation

- Correlation is still useful
- Causality is not always needed
- Gene A and gene B remain useful features for making predictions
- In a passive, or observational, setting

Making Sense of Correlation

- Correlation is still useful
- Causality is not always needed
- Gene A and gene B remain useful features for making predictions
- In a passive, or observational, setting
    - we measure the activities of certain genes
      and are asked to predict the phenotype

Making Sense of Correlation

- Correlation is still useful
- Causality is not always needed
- Gene A and gene B remain useful features for making predictions
- In a passive, or observational, setting
  - we measure the activities of certain genes
    and are asked to predict the phenotype

- However, if we want to answer interventional questions

Making Sense of Correlation

- Correlation is still useful
- Causality is not always needed
- Gene A and gene B remain useful features for making predictions
- In a passive, or observational, setting
  - we measure the activities of certain genes
    and are asked to predict the phenotype

- However, if we want to answer interventional questions
  - the outcome of a gene knockout experiment

Making Sense of Correlation

- Correlation is still useful
- Causality is not always needed
- Gene A and gene B remain useful features for making predictions
- In a passive, or observational, setting
  - we measure the activities of certain genes
    and are asked to predict the phenotype

- However, if we want to answer interventional questions
  - the outcome of a gene knockout experiment
  - the effect of a policy enforcing a job training program (or higher chocolate consumption)

Making Sense of Correlation

- Correlation is still useful
- Causality is not always needed
- Gene A and gene B remain useful features for making predictions
- In a passive, or observational, setting
  - we measure the activities of certain genes
    and are asked to predict the phenotype

- However, if we want to answer interventional questions
  - the outcome of a gene knockout experiment
  - the effect of a policy enforcing a job training program (or higher chocolate consumption)
- We need more than correlation

Making Sense of Correlation

- Correlation is still useful
- Causality is not always needed
- Gene A and gene B remain useful features for making predictions
- In a passive, or observational, setting
  - we measure the activities of certain genes
    and are asked to predict the phenotype

- However, if we want to answer interventional questions
  - the outcome of a gene knockout experiment
  - the effect of a policy enforcing a job training program (or higher chocolate consumption)
- We need more than correlation
- We need a causal model

Table of Contents

1. From Statistical to Causal Models
2. Causal Modeling
3. Causal Discovery
4. Causal Reasoning
5. Conclusions

Causal Graphical Model

Definition

A Causal Graphical Model (CGM) $\mathcal{M} = (G, p)$ over n random variables $X_1, \ldots, X_n$ consists of

- a directed acyclic graph (DAG) G in which directed edges $(X_j \to X_i)$ represent a direct causal effect of X_j on X_i;
- a joint distribution $p(X_1, \ldots, X_n)$ which is Markovian w.r.t. G

$$p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i \mid PA_i); \quad PA_i = \{X_j : (X_j \to X_i) \in G\}$$

Causal Graphical Model

Definition

A Causal Graphical Model (CGM) $\mathcal{M} = (G, p)$ over n random variables $X_1, \ldots, X_n$ consists of

- a directed acyclic graph (DAG) G in which directed edges $(X_j \to X_i)$ represent a direct causal effect of X_j on X_i;
- a joint distribution $p(X_1, \ldots, X_n)$ which is Markovian w.r.t. G

$$p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i \mid PA_i); \quad PA_i = \{X_j : (X_j \to X_i) \in G\}$$

PA_i is the set of parents, or direct causes, of X_i in G

Causal Graphical Model

Definition

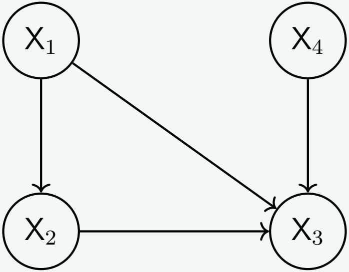
A Causal Graphical Model (CGM) $\mathcal{M} = (G, p)$ over n random variables $X_1, \ldots, X_n$ consists of

- a directed acyclic graph (DAG) G in which directed edges $(X_j \to X_i)$ represent a direct causal effect of X_j on X_i;
- a joint distribution $p(X_1, \ldots, X_n)$ which is Markovian w.r.t. G

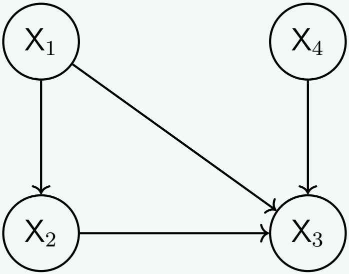
$$p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i \mid PA_i); \quad PA_i = \{X_j : (X_j \to X_i) \in G\}$$

PA_i is the set of parents, or direct causes, of X_i in G
Decomposition of the joint distribution into causal conditionals

Four variables



Four variables



$$P(X_1, X_2, X_3, X_4) =$$
$$P(X_1)\, P(X_4)\, P(X_2 \mid X_1)\, P(X_3 \mid X_1, X_2, X_4)$$

Causal Markov Condition

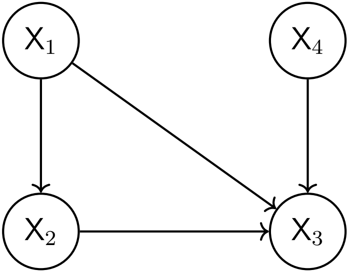
Definition

A joint distribution satisfies the causal Markov condition w.r.t. a DAG G if every variable is conditionally independent of its non-descendants in G given its parents in G

Causal Markov Condition

Definition

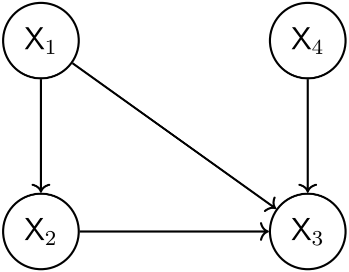
A joint distribution satisfies the causal Markov condition w.r.t. a DAG G if every variable is conditionally independent of its non-descendants in G given its parents in G



Causal Markov Condition

Definition

A joint distribution satisfies the causal Markov condition w.r.t. a DAG G if every variable is conditionally independent of its non-descendants in G given its parents in G

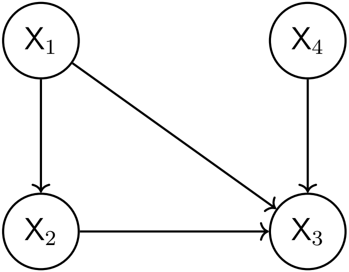


$$X_2 \perp\!\!\!\perp X_4 \mid X_1$$
$$X_4 \perp\!\!\!\perp \{X_1, X_2\}$$

Causal Markov Condition

Definition

A joint distribution satisfies the causal Markov condition w.r.t. a DAG G if every variable is conditionally independent of its non-descendants in G given its parents in G



$$X_2 \perp\!\!\!\perp X_4 \mid X_1$$
$$X_4 \perp\!\!\!\perp \{X_1, X_2\}$$

$p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i \mid PA_i)$ iff the Causal Markov Condition holds

Intervention on a causal DAG

Central idea

Intervening on a variable, by externally forcing it to take on a particular value,
renders it independent of its causes

Intervention on a causal DAG

Central idea

Intervening on a variable, by externally forcing it to take on a particular value,
renders it independent of its causes

and breaks their causal influence

Intervention on a causal DAG

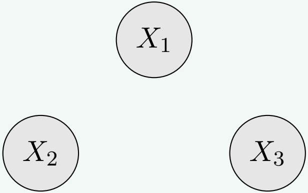
Central idea

Intervening on a variable, by externally forcing it to take on a particular value,
renders it independent of its causes

and breaks their causal influence

- do-operator
- graph-surgery

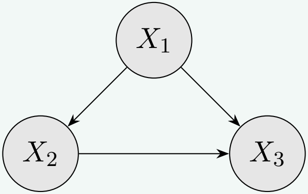
Three variables and a graph



Three variables and a graph

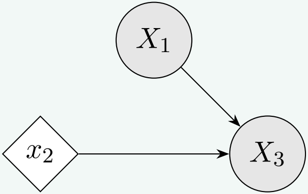
$$
\begin{tikzcd}
 & X_1 \arrow[dl] \arrow[dr] & \\
X_2 \arrow[rr] & & X_3
\end{tikzcd}
$$

From graph G to G'



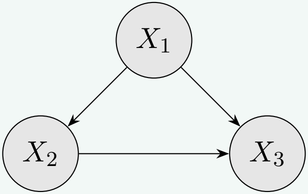
Starting graph G

From graph G to G'



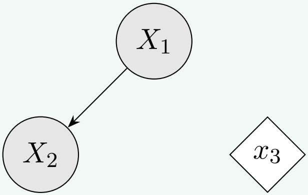
Post-intervention graph G' for $do(X_2 = x_2)$.

From graph G to G''



Starting graph G

From graph G to G''



Post-intervention graph G'' for $do(X_3 = x_3)$.

More on $do(X = x)$

- Intervention and Conditioning radically different

More on $do(X = x)$

- Intervention and Conditioning radically different
- Conditioning is passive

More on $do(X = x)$

- Intervention and Conditioning radically different
- Conditioning is passive
- Intervention is active

More on $do(X = x)$

- Intervention and Conditioning radically different
- Conditioning is passive
- Intervention is active
  - if a gene is knocked out, it is no longer influenced by other genes that were previously regulating it

More on $do(X = x)$

- Intervention and Conditioning radically different
- Conditioning is passive
- Intervention is active
  - if a gene is knocked out, it is no longer influenced by other genes that were previously regulating it
    instead, its activity is now solely determined by the intervention

More on $do(X = x)$

- Intervention and Conditioning radically different
- Conditioning is passive
- Intervention is active
  - if a gene is knocked out, it is no longer influenced by other genes that were previously regulating it
    instead, its activity is now solely determined by the intervention

Note

This is fundamentally different from conditioning, because passively observing the activity of a gene provides information about its driving factors (i.e., its direct causes)

More on $do(X = x)$

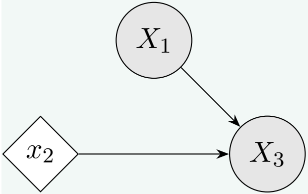
- Intervention and Conditioning radically different
- Conditioning is passive
- Intervention is active
  - if a gene is knocked out, it is no longer influenced by other genes that were previously regulating it
    instead, its activity is now solely determined by the intervention

Note

This is fundamentally different from conditioning, because passively observing the activity of a gene provides information about its driving factors (i.e., its direct causes)

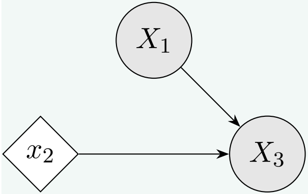
$$p(y \mid x) \neq p(y \mid do(X = x))$$

Example



DAG G' for $do(X_2 = x_2)$

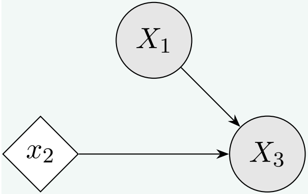
Example



DAG G' for $do(X_2 = x_2)$

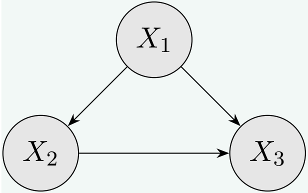
$$p(X_3|do(X_2 = x_2))$$
$$= \sum_{x_1} p(x_1)p(X_3|x_1, x_2)$$

Example



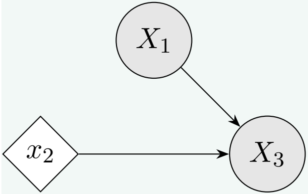
DAG G' for $do(X_2 = x_2)$

$$p(X_3 | do(X_2 = x_2)) = \sum_{x_1} p(x_1) p(X_3 | x_1, x_2)$$



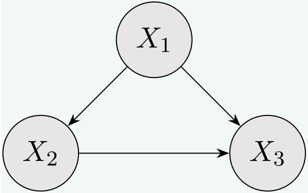
DAG G

Example



DAG G' for $do(X_2 = x_2)$

$$p(X_3|do(X_2 = x_2)) = \sum_{x_1} p(x_1)p(X_3|x_1, x_2)$$



DAG G

$$p(X_3 \mid x_2) = \sum_{x_1} p(x_1 \mid x_2)p(X_3 \mid x_1, x_2)$$

Structural Causal Model

Definition

An SCM $\mathcal{M} = (F, p_U)$ consists of

i) a set F of n assignments
(the structural equations)

$$F = \{X_i := f_i(PA_i, U_i),\, i = 1, \ldots, n\}$$

$PA_i \subseteq \{X_1, \ldots, X_n\} \setminus X_i$: causal parents
U_i's: noise variables

ii) a joint distribution $p_U(U_1, \ldots, U_n)$

Features of an SCM

Features of an SCM

- Each X_i is generated from other variables through a deterministic mechanism F

Features of an SCM

- Each X_i is generated from other variables through a deterministic mechanism F
- Randomness originates from U_i's
  - *stochasticity* of the process
  - *uncertainty* due to unmeasured parts

Features of an SCM

- Each X_i is generated from other variables through a deterministic mechanism F
- Randomness originates from U_i's
  - *stochasticity* of the process
  - *uncertainty* due to unmeasured parts
- $X_i := f_i(PA_i, U_i)$ asymmetry between LHS and RHS

Features of an SCM

- Each X_i is generated from other variables through a deterministic mechanism F
- Randomness originates from U_i's
  stochasticity of the process
  uncertainty due to unmeasured parts
- $X_i := f_i(PA_i, U_i)$ asymmetry between LHS and RHS
- In parametric linear form (linear f_i)
  SCMs are also known as structural equation models
  (path analysis)

Linking SCM's and CGM's

Definition

The causal graph G *induced* by an SCM is the directed graph with vertex set $\{X_1, \dots, X_n\}$ and a directed edge from each vertex in PA_i to X_i for all i.

Linking SCM's and CGM's

Definition

The causal graph G *induced* by an SCM is the directed graph with vertex set $\{X_1, \ldots, X_n\}$ and a directed edge from each vertex in PA_i to X_i for all i.

Example

SCM over $\{X_1, X_2, X_3\}$ with some $p_U(U_1, U_2, U_3)$

$$X_1 := f_1(U_1),\ X_2 := f_2(X_1, U_2),\ X_3 := f_3(X_1, X_2, U_3)$$

Linking SCM's and CGM's

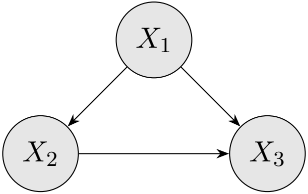
Definition

The causal graph G *induced* by an SCM is the directed graph with vertex set $\{X_1, \dots, X_n\}$ and a directed edge from each vertex in PA_i to X_i for all i.

Example

SCM over $\{X_1, X_2, X_3\}$ with some $p_U(U_1, U_2, U_3)$

$$X_1 := f_1(U_1),\; X_2 := f_2(X_1, U_2),\; X_3 := f_3(X_1, X_2, U_3)$$



induced DAG G

Difference between CGM and SCM

SCM allows for a rich class of causal models
including models with
cyclic causal relations
not obeying the causal Markov condition
(because of complex covariance structures between the noise terms)

Difference between CGM and SCM

SCM allows for a rich class of causal models
including models with
cyclic causal relations
not obeying the causal Markov condition
(because of complex covariance structures between the noise terms)

Further common assumptions

A1) Acyclicity: the induced graph G is a DAG

A2) Causal sufficiency/no hidden confounders: the U_i's are jointly independent, i.e.

$$p_U(U_1, \ldots, U_n) = p_{U_1}(U_1) \times \ldots p_{U_n}(U_n)$$

Difference between CGM and SCM

SCM allows for a rich class of causal models
including models with
cyclic causal relations
not obeying the causal Markov condition
(because of complex covariance structures between the noise terms)

Further common assumptions

A1) Acyclicity: the induced graph G is a DAG

A2) Causal sufficiency/no hidden confounders: the U_i's are jointly independent, i.e.

$$p_U(U_1, \ldots, U_n) = p_{U_1}(U_1) \times \ldots p_{U_n}(U_n)$$

Acyclicity and Causal sufficiency ensure that the distribution induced by an SCM *factorises* according to its induced causal graph G (and the causal Markov condition is satisfied w.r.t. G)

Interventions in SCM's

Definition

An intervention $do(X_i = x_i)$ in an SCM $\mathcal{M} = (F, p_U)$ is modeled by

- replacing the i-th structural equation in F by $X_i = x_i$
- remaining F_j's remain unchanged $(j \neq i)$

Result is the interventional SCM $\mathcal{M}^{do(X_i = x_i)} = (F', p_U)$.

Interventions in SCM's

Definition

An intervention $do(X_i = x_i)$ in an SCM $\mathcal{M} = (F, p_U)$ is modeled by

- replacing the i-th structural equation in F by $X_i = x_i$
- remaining F_j's remain unchanged $(j \neq i)$

Result is the interventional SCM $\mathcal{M}^{do(X_i = x_i)} = (F', p_U)$.

From $\mathcal{M}^{do(X_i = x_i)} = (F', p_U)$
deduce the interventional distribution $p(X_{-i} \,|\, do(X_i = x_i))$
and the intervention graph G'

Interventions in SCM

SCM $\mathcal{M} = (F, p_U)$

$$X_1 := f_1(U_1),\ X_2 := f_2(X_1, U_2),\ X_3 := f_3(X_1, X_2, U_3)$$

[1] This way of handling interventions coincides with that for CGMs

Interventions in SCM

SCM $\mathcal{M} = (F, p_U)$

$$X_1 := f_1(U_1),\ X_2 := f_2(X_1, U_2),\ X_3 := f_3(X_1, X_2, U_3)$$

SCM $\mathcal{M}^{do(X_2 = x_2)} = (F', p_U)$

$$X_1 := f_1(U_1),\ X_2 := x_2,\ X_3 := f_3(X_1, X_2, U_3)$$

[1] This way of handling interventions coincides with that for CGMs

Interventions in SCM

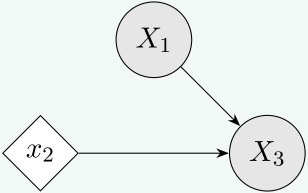
$SCM\ \mathcal{M} = (F, p_U)$

$$X_1 := f_1(U_1),\ X_2 := f_2(X_1, U_2),\ X_3 := f_3(X_1, X_2, U_3)$$

$SCM\ \mathcal{M}^{do(X_2=x_2)} = (F', p_U)$

$$X_1 := f_1(U_1),\ X_2 := x_2,\ X_3 := f_3(X_1, X_2, U_3)$$

Graph G' induced by $\mathcal{M}^{do(X_2=x_2)}$ [1]



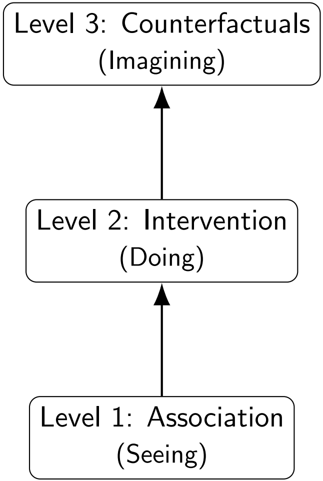
[1] This way of handling interventions coincides with that for CGMs

Seeing, Doing, Imagining

The ladder of causality

Seeing, Doing, Imagining

The ladder of causality



Seeing, Doing, Imagining

The ladder of causality

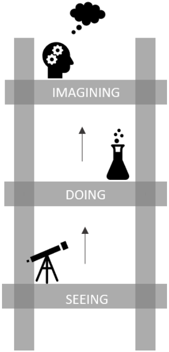
Level 3: Counterfactuals
(Imagining)

↑

Level 2: Intervention
(Doing)

↑

Level 1: Association
(Seeing)



Observations, interventions, counterfactuals

i) observation

Observations, interventions, counterfactuals

i) observation

passively seen or measured

Observations, interventions, counterfactuals

i) observation

   passively seen or measured

ii) intervention

Observations, interventions, counterfactuals

i) observation

passively seen or measured

ii) intervention

external manipulation or experimentation

Observations, interventions, counterfactuals

i) observation

passively seen or measured

ii) intervention

external manipulation or experimentation

iii) counterfactual

Observations, interventions, counterfactuals

i) observation
   passively seen or measured
ii) intervention
   external manipulation or experimentation
iii) counterfactual
   what *would* have been, given that something else was in fact observed

Observations, interventions, counterfactuals

i) observation

   passively seen or measured

ii) intervention

   external manipulation or experimentation

iii) counterfactual

   what *would* have been, given that something else was in fact observed

Issues with counterfactuals

Cannot be observed empirically

unfalsifiable

Observations, interventions, counterfactuals

i) observation
passively seen or measured

ii) intervention
external manipulation or experimentation

iii) counterfactual
what *would* have been, given that something else was in fact observed

Issues with counterfactuals

Cannot be observed empirically

unfalsifiable

unscientific (Popper, 1959)

problematic (Dawid, 2000)

Observations, interventions, counterfactuals

i) observation
   passively seen or measured
ii) intervention
   external manipulation or experimentation
iii) counterfactual
   what *would* have been, given that something else was in fact observed

Issues with counterfactuals

Cannot be observed empirically

unfalsifiable

unscientific (Popper, 1959)

problematic (Dawid, 2000)

Yet, humans seem to perform counterfactual reasoning in practice starting in early childhood (Buchsbaum et al., 2012)

Counterfactuals

"Given that patient X received treatment A and their health got worse, what would have happened if they had been given treatment B instead, *all else being equal*?"

Counterfactuals

"Given that patient X received treatment A and their health got worse, what would have happened if they had been given treatment B instead, *all else being equal*?"

- SCMs provide a suitable framework for counterfactual reasoning

Counterfactuals

"Given that patient X received treatment A and their health got worse, what would have happened if they had been given treatment B instead, *all else being equal*?"

- SCMs provide a suitable framework for counterfactual reasoning
- Observing what actually happened
  provides information about the *background state* of the system
  namely the noise terms $\{U_1, \ldots, U_n\}$ in an SCM

Counterfactuals

"Given that patient X received treatment A and their health got worse, what would have happened if they had been given treatment B instead, *all else being equal*?"

- SCMs provide a suitable framework for counterfactual reasoning
- Observing what actually happened provides information about the *background state* of the system namely the noise terms $\{U_1, \ldots, U_n\}$ in an SCM
- This differs from an intervention where such background information is not available

- Observing that treatment A did not work may tell us that the patient has a rare condition

- Observing that treatment A did not work may tell us that the patient has a rare condition
    - this provides information on their background state of health

- Observing that treatment A did not work may tell us that the patient has a rare condition
    - this provides information on their background state of health
- This, in turn, suggests that treatment B might have worked

- Observing that treatment A did not work may tell us that the patient has a rare condition
  - this provides information on their background state of health
- This, in turn, suggests that treatment B might have worked
- However, given that treatment A has been applied, patient's condition may have changed
  - so condition *"all else being equal"* fails

- Observing that treatment A did not work may tell us that the patient has a rare condition
    - this provides information on their background state of health
- This, in turn, suggests that treatment B might have worked
- However, given that treatment A has been applied, patient's condition may have changed
    - so condition *"all else being equal"* fails
- and B may no longer work in a future intervention *on this specific patient*

Definition (Counterfactuals in SCM's)

Given evidence $X = x$ observed from an SCM $\mathcal{M} = (F, p_U)$ the counterfactual SCM $\mathcal{M}^{X=x}$ is obtained by updating p_U to $p_{U \mid X=x}$

$$\mathcal{M}^{X=x} = (F, p_{U \mid X=x})$$

Definition (Counterfactuals in SCM's)

Given evidence $X = x$ observed from an SCM $\mathcal{M} = (F, p_U)$
the counterfactual SCM $\mathcal{M}^{X=x}$ is obtained by updating p_U to $p_{U \mid X=x}$

$$\mathcal{M}^{X=x} = (F, p_{U \mid X=x})$$

Counterfactuals are then computed by performing interventions in the counterfactual SCM $\mathcal{M}^{X=x}$

Computing counterfactuals with SCM: Example

$SCM\ \mathcal{M} = (F, p_U)$

Computing counterfactuals with SCM: Example

$SCM\ \mathcal{M} = (F, p_U)$

$X := U_X,\ Y := 3X + U_Y;\ U_X, U_Y \overset{iid}{\sim} N(0,1)$

Computing counterfactuals with SCM: Example

$SCM \ \mathcal{M} = (F, p_U)$

$X := U_X,\ Y := 3X + U_Y;\ U_X, U_Y \overset{iid}{\sim} N(0,1)$

We observe $X = 2$ and $Y = 6.5$
and want to answer the counterfactual question
"What would Y have been, had $X = 1$?"

Computing counterfactuals with SCM: Example

$SCM\ \mathcal{M} = (F, p_U)$

$X := U_X,\ Y := 3X + U_Y;\ U_X, U_Y \overset{iid}{\sim} N(0,1)$

We observe $X = 2$ and $Y = 6.5$
and want to answer the counterfactual question
"What would Y have been, had $X = 1$?"

We are thus interested in

$$p^{\mathcal{M}^{X=2,Y=6.5;do(X=1)}}(Y)$$

Example ctd

Recall: $X := U_X$, $Y := 3X + U_Y$, $U_X, U_Y \overset{iid}{\sim} N(0,1)$

Example ctd

Recall: $X := U_X$, $Y := 3X + U_Y$, $U_X, U_Y \overset{iid}{\sim} N(0,1)$

- Update the noise distribution $p_U \to p_{U \mid X=2, Y=6.5}$

$$U_X \sim \delta(2),\; U_Y \sim \delta(0.5)$$

Example ctd

Recall: $X := U_X$, $Y := 3X + U_Y$, $U_X, U_Y \overset{iid}{\sim} N(0,1)$

- Update the noise distribution $p_U \to p_{U \mid X=2, Y=6.5}$

$$U_X \sim \delta(2),\, U_Y \sim \delta(0.5)$$

- Obtain the updated SCM $\mathcal{M}^{X=2, Y=6.5} = (F, p_{U \mid X=2, Y=6.5})$

Example ctd

Recall: $X := U_X$, $Y := 3X + U_Y$, $U_X, U_Y \overset{iid}{\sim} N(0,1)$

- Update the noise distribution $p_U \to p_{U \mid X=2, Y=6.5}$

$$U_X \sim \delta(2),\; U_Y \sim \delta(0.5)$$

- Obtain the updated SCM $\mathcal{M}^{X=2,Y=6.5} = (F, p_{U \mid X=2,Y=6.5})$
- Perform the intervention $do(X=1)$ on $\mathcal{M}^{X=2,Y=6.5}$

$$p^{\mathcal{M}^{X=2,Y=6.5;do(X=1)}}(Y) = \delta(3.5)$$

Example ctd

Recall: $X := U_X$, $Y := 3X + U_Y$, $U_X, U_Y \overset{iid}{\sim} N(0,1)$

- Update the noise distribution $p_U \to p_{U \mid X=2, Y=6.5}$

$$U_X \sim \delta(2),\, U_Y \sim \delta(0.5)$$

- Obtain the updated SCM $\mathcal{M}^{X=2,Y=6.5} = (F, p_{U \mid X=2,Y=6.5})$
- Perform the intervention $do(X=1)$ on $\mathcal{M}^{X=2,Y=6.5}$

$$p^{\mathcal{M}^{X=2,Y=6.5;do(X=1)}}(Y) = \delta(3.5)$$

- Above differs from the interventional distribution $Y \mid do(X=1) \sim N(3,1)$

Example ctd

2. the observed data $(X, Y) = (2, 6.5)$

1. the SCM $\mathcal{M}$ we start with

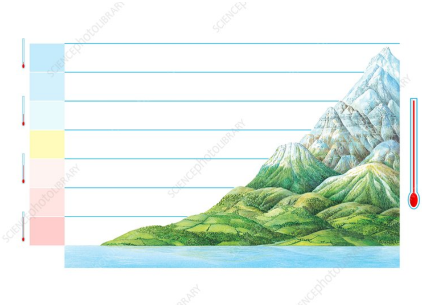
3. the intervention $do(X = 1)$

$$p^{\mathcal{M}^{(X,Y)=(2,6.5);do(X=1)}}(Y)$$

4. the distribut^n of the variable we are interested in

Factorizations

Altitude and Temperature



Factorizations

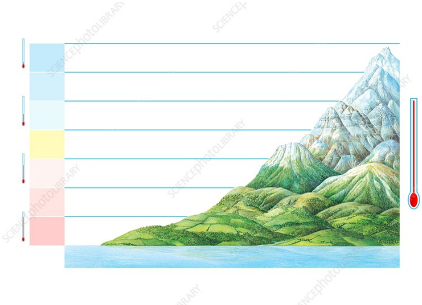
Altitude and Temperature

- Disentangled factorization

$$p(A, T) = p(A)p(T \mid A)$$

Factorizations

Altitude and Temperature



- Disentangled factorization

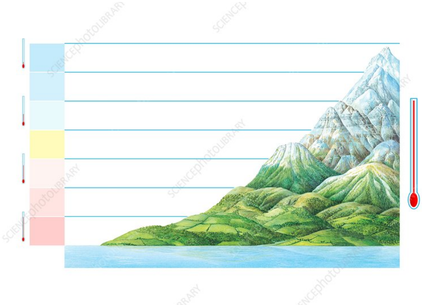
$$p(A, T) = p(A)p(T \mid A)$$

- Entangled factorization

$$p(A, T) = p(T)p(A \mid T)$$

Factorizations

Altitude and Temperature



Only in the disentangled factorization some components generalize across inteventions/domains

- Disentangled factorization

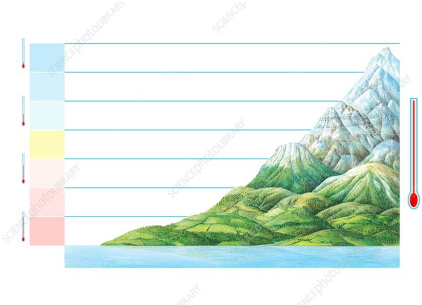
$$p(A, T) = p(A)p(T \mid A)$$

- Entangled factorization

$$p(A, T) = p(T)p(A \mid T)$$

Factorizations

Altitude and Temperature



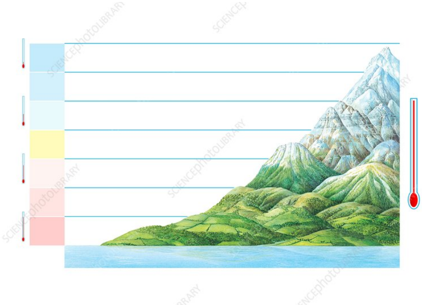
Only in the disentangled factorization some components generalize across inteventions/domains

- Austria and Switzerland (CH)

- Disentangled factorization
$$p(A,T) = p(A)p(T \mid A)$$
- Entangled factorization
$$p(A,T) = p(T)p(A \mid T)$$

Factorizations

Altitude and Temperature



Only in the disentangled factorization some components generalize across inteventions/domains

- Disentangled factorization

$$p(A,T) = p(A)p(T \mid A)$$

- Entangled factorization

$$p(A,T) = p(T)p(A \mid T)$$

- Austria and Switzerland (CH)

$$p_{Austria}(A,T) = p_{Austria}(A)p(T \mid A)$$
$$p_{CH}(A,T) = p_{CH}(A)p(T \mid A)$$

Factorizations

Altitude and Temperature

- Disentangled factorization

$$p(A,T) = p(A)p(T \mid A)$$

- Entangled factorization

$$p(A,T) = p(T)p(A \mid T)$$

Only in the disentangled factorization some components generalize across inteventions/domains

- Austria and Switzerland (CH)

$$p_{Austria}(A,T) = p_{Austria}(A)p(T \mid A)$$
$$p_{CH}(A,T) = p_{CH}(A)p(T \mid A)$$

$p(T \mid A)$ is likely to be the same across countries

Factorizations

Altitude and Temperature

- Disentangled factorization

$$p(A,T) = p(A)p(T \mid A)$$

- Entangled factorization

$$p(A,T) = p(T)p(A \mid T)$$

Only in the disentangled factorization some components generalize across inteventions/domains

- Austria and Switzerland (CH)

$$p_{Austria}(A,T) = p_{Austria}(A)p(T \mid A)$$
$$p_{CH}(A,T) = p_{CH}(A)p(T \mid A)$$

$p(T \mid A)$ is likely to be the same across countries
$p(A)$ is country-specific

Independent causal mechanisms

For a model to correctly predict the effect of interventions, it needs to have components that are robust when moving from an observational distribution to certain interventional distributions.

Independent causal mechanisms

For a model to correctly predict the effect of interventions, it needs to have components that are robust when moving from an observational distribution to certain interventional distributions.

Principle (Independent Causal Mechanisms (ICM))

The causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other.

Independent causal mechanisms

For a model to correctly predict the effect of interventions, it needs to have components that are robust when moving from an observational distribution to certain interventional distributions.

Principle (Independent Causal Mechanisms (ICM))

The causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other.

In the two-variable case, say (A, T), it reduces to independence between

- the cause distribution $p(A)$
- the mechanism producing the effect from the cause $p(T \mid A)$

Independent causal mechanisms

For a model to correctly predict the effect of interventions, it needs to have components that are robust when moving from an observational distribution to certain interventional distributions.

Principle (Independent Causal Mechanisms (ICM))

The causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other.

In the two-variable case, say (A, T), it reduces to independence between

- the cause distribution $p(A)$
- the mechanism producing the effect from the cause $p(T \mid A)$

Principle (Sparse Mechanism Shift)

Small distribution changes manifest in a sparse or local way in the causal/disentangled factorization; i.e., they should usually not affect all factors simultaneously.

Principle (Sparse Mechanism Shift)

Small distribution changes manifest in a sparse or local way in the causal/disentangled factorization; i.e., they should usually not affect all factors simultaneously.

In a non-causal/entangled factorization, many terms will be affected simultaneously if we change one of the physical mechanisms responsible for a system's statistical dependencies

Principle (Sparse Mechanism Shift)

Small distribution changes manifest in a sparse or local way in the causal/disentangled factorization; i.e., they should usually not affect all factors simultaneously.

In a non-causal/entangled factorization, many terms will be affected simultaneously if we change one of the physical mechanisms responsible for a system's statistical dependencies

In the Altitude-Temperature setting, if we change country

- we only need to change $p(A)$ if we use the causal factorization
- we need to change both $p(T)$ and $p(A \mid T)$ in the entangled factorization

Table of Contents

1. From Statistical to Causal Models
2. Causal Modeling
3. Causal Discovery
4. Causal Reasoning
5. Conclusions

Causal discovery

Causal discovery

- Domain knowledge often unavailable or incomplete

Causal discovery

- Domain knowledge often unavailable or incomplete
- Need to learn the causal DAG

Causal discovery

- Domain knowledge often unavailable or incomplete
- Need to learn the causal DAG
  Typically using observational (passive) data which are abundant

Causal discovery

- Domain knowledge often unavailable or incomplete
- Need to learn the causal DAG
  Typically using observational (passive) data which are abundant
- Hopeless?

Causal discovery

- Domain knowledge often unavailable or incomplete
- Need to learn the causal DAG
  Typically using observational (passive) data which are abundant
- Hopeless?
- Surprisingly the problem becomes *easier* when the number of variables becomes *higher*

Causal discovery

- Domain knowledge often unavailable or incomplete
- Need to learn the causal DAG
  Typically using observational (passive) data which are abundant
- Hopeless?
- Surprisingly the problem becomes *easier* when the number of variables becomes *higher*
  because there are nontrivial *conditional independence* properties among the variables implied by a causal structure

Constraint-based methods

Basic idea

- Test which (conditional) independencies can be inferred from the data
- Try to find a graph which implies them

Constraint-based methods

Basic idea

- Test which (conditional) independencies can be inferred from the data
- Try to find a graph which implies them

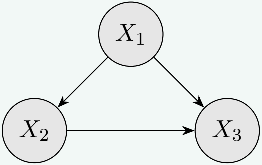
Assumption (Faithfulness)

The only (conditional) independencies satisfied by $p(\cdot)$ are those implied by the causal Markov condition

Example: faithfulness-SCM

$$X_1 := U_1$$
$$X_2 = \alpha X_1 + U_2$$
$$X_3 = \beta X_1 + \gamma X_2 + U_3$$

with $U_i \overset{iid}{\sim} \mathcal{N}(0, 1)$.



Causal DAG G

From the model:

$$X_3 = (\beta + \alpha\gamma)X_1 + \gamma U_2 + U_3$$

From the model:

$$X_3 = (\beta + \alpha\gamma)X_1 + \gamma U_2 + U_3$$

Thus, if $\beta + \alpha\gamma = 0$, then:

$$X_1 \perp\!\!\!\perp X_3$$

From the model:

$$X_3 = (\beta + \alpha\gamma)X_1 + \gamma U_2 + U_3$$

Thus, if $\beta + \alpha\gamma = 0$, then:

$$X_1 \perp\!\!\!\perp X_3$$

But this is **not** implied by the graph G.

From the model:

$$X_3 = (\beta + \alpha\gamma)X_1 + \gamma U_2 + U_3$$

Thus, if $\beta + \alpha\gamma = 0$, then:

$$X_1 \perp\!\!\!\perp X_3$$

But this is **not** implied by the graph G.

Faithfulness is violated

Markov equivalence

Definition (Markov equivalence)

Two DAGs are said to be Markov equivalent if they encode the same conditional independence (CI) statements.
The set of all DAGs encoding the same CI's is called a Markov equivalence class

Chains, forks and colliders

(a) Chains

$$X \to Y \to Z$$

$$X \leftarrow Y \leftarrow Z$$

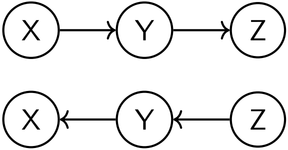
(b) Fork

$$\begin{tikzcd} & Y \arrow[dl] \arrow[dr] & \\ X & & Z \end{tikzcd}$$

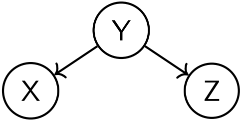
(a) and (b) imply $X \perp\!\!\!\perp Z \mid Y$
(and no others)

Chains, forks and colliders

(a) Chains

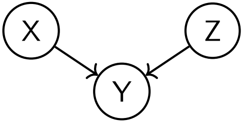


(b) Fork



(a) and (b) imply $X \perp\!\!\!\perp Z \mid Y$
(and no others)

(c) Collider
(v-structure)



(c) implies $X \perp\!\!\!\perp Z$
(but $X \not\!\perp\!\!\!\perp Z \mid Y$)

(a) and (b): same Markov equiv class
(c) singleton equivalence class

Markov equivalence: characterization

Result

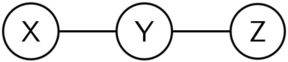
Two DAG's are Markov equivalent iff they have the same skeleton and the same v-structures

Markov equivalence: characterization

Result

Two DAG's are Markov equivalent iff they have the same skeleton and the same v-structures

Skeleton of Chains (a), Fork (b) and Collider (c)

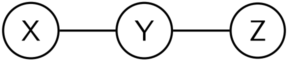


Markov equivalence: characterization

Result

Two DAG's are Markov equivalent iff they have the same skeleton and the same v-structures

Skeleton of Chains (a), Fork (b) and Collider (c)



v-structures

(a) and (b)	NO
(c)	YES

Constraint-based methods: computations

✓ **Skeleton estimation**

- Test $X \perp\!\!\!\perp Y \mid W$ for all $W \subseteq \mathbf{X} \setminus \{X, Y\}$.
- if no such W is found, connect X and Y
- Expensive

Constraint-based methods: computations

✓ **Skeleton estimation**

- Test $X \perp\!\!\!\perp Y \mid W$ for all $W \subseteq \mathbf{X} \setminus \{X, Y\}$.
- if no such W is found, connect X and Y
- Expensive

✓ **Edge orientation**

- Direct edges avoiding v-structures and cycles.

☞ **PC** (Spirtes et al. 2000) more efficient

☞ **FCI** (handles hidden confounders)

Constraint-based methods: computations

✓ **Skeleton estimation**

- Test $X \perp\!\!\!\perp Y \mid W$ for all $W \subseteq \mathbf{X} \setminus \{X, Y\}$.
- if no such W is found, connect X and Y
- Expensive

✓ **Edge orientation**

- Direct edges avoiding v-structures and cycles.

☞ **PC** (Spirtes et al. 2000) more efficient

☞ **FCI** (handles hidden confounders)

✗ **Limitations**

- Only a Markov equivalence class is recovered.
- CI testing is a hard problem

Constraint-based methods: computations

✓ **Skeleton estimation**

- Test $X \perp\!\!\!\perp Y \mid W$ for all $W \subseteq \mathbf{X} \setminus \{X, Y\}$.
- if no such W is found, connect X and Y
- Expensive

✓ **Edge orientation**

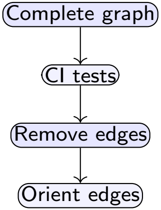
- Direct edges avoiding v-structures and cycles.

☞ **PC** (Spirtes et al. 2000) more efficient

☞ **FCI** (handles hidden confounders)

✗ **Limitations**

- Only a Markov equivalence class is recovered.
- CI testing is a hard problem



Score-based methods

- $\mathcal{G}$: set of DAG's over variables $\{X_1, \ldots, X_n\}$
- $D = \{\mathbf{x}_1 \ldots, \mathbf{x}_m\}$: observed data
- $S(G \mid D)$: score reflecting how well a G-graphical statistical model fits D

Score-based methods

- $\mathcal{G}$: set of DAG's over variables $\{X_1, \ldots, X_n\}$
- $D = \{\mathbf{x}_1 \ldots, \mathbf{x}_m\}$: observed data
- $S(G \mid D)$: score reflecting how well a G-graphical statistical model fits D
- Most methods assume a parametric model which factorises according to G

Score-based methods

- $\mathcal{G}$: set of DAG's over variables $\{X_1, \ldots, X_n\}$
- $D = \{\mathbf{x}_1 \ldots, \mathbf{x}_m\}$: observed data
- $S(G \mid D)$: score reflecting how well a G-graphical statistical model fits D
- Most methods assume a parametric model which factorises according to G
- $S_{BIC}(G \mid D) = \log p(D \mid G, \hat{\theta}^{MLE}) - \frac{k}{2} \log m$
  $k = \#$ of parameters

Score-based methods

- $\mathcal{G}$: set of DAG's over variables $\{X_1, \ldots, X_n\}$
- $D = \{\mathbf{x}_1 \ldots, \mathbf{x}_m\}$: observed data
- $S(G \mid D)$: score reflecting how well a G-graphical statistical model fits D
- Most methods assume a parametric model which factorises according to G
- $S_{BIC}(G \mid D) = \log p(D \mid G, \hat{\theta}^{MLE}) - \frac{k}{2} \log m$
  $k = \#$ of parameters
- $S_{BAYES}(G \mid D) = p(D \mid G) = \int_{\Theta} p(D \mid G, \theta) p(\theta \mid G) d\theta$

Score-based methods

- $\mathcal{G}$: set of DAG's over variables $\{X_1, \ldots, X_n\}$
- $D = \{\mathbf{x}_1 \ldots, \mathbf{x}_m\}$: observed data
- $S(G \mid D)$: score reflecting how well a G-graphical statistical model fits D
- Most methods assume a parametric model which factorises according to G
- $S_{BIC}(G \mid D) = \log p(D \mid G, \hat{\theta}^{MLE}) - \frac{k}{2} \log m$
  $k = \#$ of parameters
- $S_{BAYES}(G \mid D) = p(D \mid G) = \int_\Theta p(D \mid G, \theta) p(\theta \mid G) d\theta$

$$\hat{G} = \operatorname*{argmax}_{G \in \mathcal{G}} S(G \mid D)$$

Score-based methods

- $\mathcal{G}$: set of DAG's over variables $\{X_1, \ldots, X_n\}$
- $D = \{\mathbf{x}_1 \ldots, \mathbf{x}_m\}$: observed data
- $S(G \mid D)$: score reflecting how well a G-graphical statistical model fits D
- Most methods assume a parametric model which factorises according to G
- $S_{BIC}(G \mid D) = \log p(D \mid G, \hat{\theta}^{MLE}) - \frac{k}{2} \log m$
  $k = \#$ of parameters
- $S_{BAYES}(G \mid D) = p(D \mid G) = \int_\Theta p(D \mid G, \theta) p(\theta \mid G) d\theta$

$$\hat{G} = \underset{G \in \mathcal{G}}{\operatorname{argmax}}\, S(G \mid D)$$

With a prior $p(G)$ can also use the full posterior

$$p(G \mid D) \propto p(D \mid G) p(G)$$

Table of Contents

1 From Statistical to Causal Models

2 Causal Modeling

3 Causal Discovery

4 Causal Reasoning

5 Conclusions

Causal Reasoning

Causal reasoning starts from a known (or postulated) causal graph and answers causal queries of interest

Causal Reasoning

Causal reasoning starts from a known (or postulated) causal graph and answers causal queries of interest

Two steps

Causal reasoning starts from a known (or postulated) causal graph and answers causal queries of interest

Two steps

(i) *identify* the query, i.e., derive an estimand that only involves *observable* quantities

Causal Reasoning

Causal reasoning starts from a known (or postulated) causal graph and answers causal queries of interest

Two steps

(i) *identify* the query, i.e., derive an estimand that only involves *observable* quantities

(ii) *make inference* on the estimand using data

Definition (Treatment effect)

Outcome Y and binary treatment T

$$\tau := \mathbb{E}[Y \mid do(T = 1)] - \mathbb{E}[Y \mid do(T = 0)]$$

Definition (Treatment effect)

Outcome Y and binary treatment T

$$\tau := \mathbb{E}[Y \mid do(T = 1)] - \mathbb{E}[Y \mid do(T = 0)]$$

Outcome Y and continuous X

$$\tau(x') := \left[\frac{\partial}{\partial x}\mathbb{E}[Y \mid do(X = x)\right]_{x=x'}$$

Definition (Treatment effect)

Outcome Y and binary treatment T

$$\tau := \mathbb{E}[Y \mid do(T = 1)] - \mathbb{E}[Y \mid do(T = 0)]$$

Outcome Y and continuous X

$$\tau(x') := \left[\frac{\partial}{\partial x}\mathbb{E}[Y \mid do(X = x)]\right]_{x=x'}$$

Treatment effects involve interventional expressions

Definition (Treatment effect)

Outcome Y and binary treatment T

$$\tau := \mathbb{E}[Y \mid do(T = 1)] - \mathbb{E}[Y \mid do(T = 0)]$$

Outcome Y and continuous X

$$\tau(x') := \left[\frac{\partial}{\partial x} \mathbb{E}[Y \mid do(X = x)] \right]_{x=x'}$$

Treatment effects involve interventional expressions
Causal reasoning answers queries
using observational data together with a causal model

Identification

Given a causal graph and no hidden confounders

The causal effect can be identified through the interventional distribution

$$p(X_1, \ldots, X_n \mid do(X_i = x_i)) = \delta(x_i) \prod_{j \neq i} p(X_j \mid PA_j) \tag{g}$$

Identification

Given a causal graph and no hidden confounders

The causal effect can be identified through the interventional distribution

$$p(X_1, \ldots, X_n \mid do(X_i = x_i)) = \delta(x_i) \prod_{j \neq i} p(X_j \mid PA_j) \tag{g}$$

The interventional distribution of any X_h $(h \neq i)$ can be obtained by *marginalization*

Identification

Given a causal graph and no hidden confounders

The causal effect can be identified through the interventional distribution

$$p(X_1, \ldots, X_n \mid do(X_i = x_i)) = \delta(x_i) \prod_{j \neq i} p(X_j \mid PA_j) \tag{g}$$

The interventional distribution of any X_h $(h \neq i)$ can be obtained by *marginalization*

Remarks

Formula (g) has been named

- *g-formula*
  Robins (1986)
- *truncated factorization*
  Pearl (2000, 2009)

Identification

Given a causal graph and no hidden confounders

The causal effect can be identified through the interventional distribution

$$p(X_1, \ldots, X_n \mid do(X_i = x_i)) = \delta(x_i) \prod_{j \neq i} p(X_j \mid PA_j) \qquad \text{(g)}$$

The interventional distribution of any X_h $(h \neq i)$ can be obtained by *marginalization*

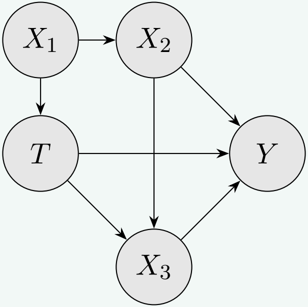
Remarks

Formula (g) has been named

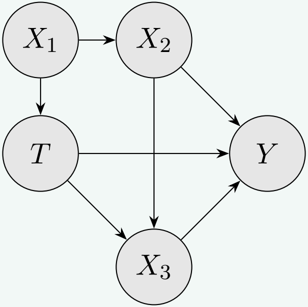
- *g-formula*
  Robins (1986)
- *truncated factorization*
  Pearl (2000, 2009)

It relies on the independence of causal mechanisms
i.e. intervening on a variable leaves the remaining causal conditionals unaffected

Evaluation of treatment effect with three covariates $\{X_1, X_2, X_3\}$



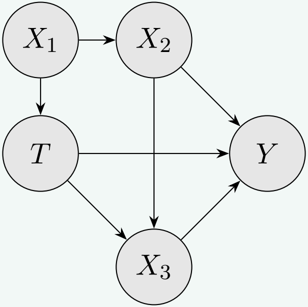
Evaluation of treatment effect with three covariates $\{X_1, X_2, X_3\}$



Factorization of interventional distribution

$$p(y, t, x_1, x_2, x_3 \mid do(T = t)) = \delta(t)p(x_1)p(x_2 \mid x_1)p(y \mid x_2, x_3, t)p(x_3 \mid x_2, t)$$

Evaluation of treatment effect with three covariates $\{X_1, X_2, X_3\}$



Factorization of interventional distribution

$$p(y, t, x_1, x_2, x_3 \mid do(T = t)) = \delta(t)p(x_1)p(x_2 \mid x_1)p(y \mid x_2, x_3, t)p(x_3 \mid x_2, t)$$

Target distribution $\quad p(y \mid do(T = t))$

Adjustment set

$$
\begin{aligned}
p(y \mid do(T=t)) &= \sum_{x_1,x_2,x_3} p(y,t,x_1,x_2,x_3 \mid do(T=t)) \\
&= \sum_{x_2}\sum_{x_1} p(x_2 \mid x_1)p(x_1) \sum_{x_3} p(y \mid x_2,x_3,t)p(x_3 \mid x_2,t) \\
&= \sum_{x_2} p(x_2)p(y \mid x_2,t)
\end{aligned}
$$

Adjustment set

$$
\begin{aligned}
p(y \mid do(T=t)) &= \sum_{x_1, x_2, x_3} p(y, t, x_1, x_2, x_3 \mid do(T=t)) \\
&= \sum_{x_2} \sum_{x_1} p(x_2 \mid x_1) p(x_1) \sum_{x_3} p(y \mid x_2, x_3, t) p(x_3 \mid x_2, t) \\
&= \sum_{x_2} p(x_2) p(y \mid x_2, t)
\end{aligned}
$$

x_2 is a valid adjustment set

More adjustment sets

It can be proved using graphical criteria or otherwise that

$$Y \perp\!\!\!\perp X_1 \mid \{T, X_2\} \tag{1.a}$$

$$X_2 \perp\!\!\!\perp T \mid X_1 \tag{1.b}$$

More adjustment sets

It can be proved using graphical criteria or otherwise that

$$Y \perp\!\!\!\perp X_1 \mid \{T, X_2\} \tag{1.a}$$

$$X_2 \perp\!\!\!\perp T \mid X_1 \tag{1.b}$$

$$\begin{aligned} p(y \mid do(T = t)) &= \sum_{x_1, x_2} p(x_1, x_2) p(y \mid x_1, x_2, t), \text{ using (1.a)} && \text{(2.a)} \\ &= \sum_{x_1} p(x_1) \sum_{x_2} p(x_2 \mid x_1, t) p(y \mid x_1, x_2, t), \text{ using (1.b)} \\ &= \sum_{x_1} p(x_1) p(y \mid x_1, t) && \text{(2.b)} \end{aligned}$$

More adjustment sets

It can be proved using graphical criteria or otherwise that

$$Y \perp\!\!\!\perp X_1 \mid \{T, X_2\} \tag{1.a}$$

$$X_2 \perp\!\!\!\perp T \mid X_1 \tag{1.b}$$

$$\begin{aligned}
p(y \mid do(T=t)) &= \sum_{x_1, x_2} p(x_1, x_2) p(y \mid x_1, x_2, t), \text{ using (1.a)} && (2.a) \\
&= \sum_{x_1} p(x_1) \sum_{x_2} p(x_2 \mid x_1, t) p(y \mid x_1, x_2, t), \text{ using (1.b)} \\
&= \sum_{x_1} p(x_1) p(y \mid x_1, t) && (2.b)
\end{aligned}$$

Both $\{x_1, x_2\}$ by (2.a) and $\{x_1\}$ by (2.b) are valid adjustment sets.
However $\{x_1, x_3\}$ is not.

Adjustment sets

Whenever

$$p(y \mid do(T = t)) = \sum_z p(z)p(y \mid z, t) \tag{3}$$

z is called a valid adjustment set

Adjustment sets

Whenever

$$p(y \mid do(T = t)) = \sum_z p(z)p(y \mid z, t) \tag{3}$$

z is called a valid adjustment set

Under causal sufficiency (no hidden variables) there exist graphical criteria to find valid adjustment sets

Adjustment sets

Whenever

$$p(y \mid do(T = t)) = \sum_z p(z)p(y \mid z, t) \qquad (3)$$

z is called a valid adjustment set

Under causal sufficiency (no hidden variables) there exist graphical criteria to find valid adjustment sets

To estimate the involved quantities in (3)
additional assumptions are required
in particular *overlap*:
for any t and feature values $\mathbf{x}$, $\mathbf{X}$, $0 < p(T = t \mid \mathbf{X} = \mathbf{x}) < 1$

Adjustment sets

Whenever

$$p(y \,|\, do(T = t)) = \sum_z p(z) p(y \,|\, z, t) \tag{3}$$

z is called a valid adjustment set

Under causal sufficiency (no hidden variables) there exist graphical criteria to find valid adjustment sets

To estimate the involved quantities in (3)
additional assumptions are required
in particular *overlap*:
for any t and feature values $\mathbf{x}$, $\mathbf{X}$, $0 < p(T = t \,|\, \mathbf{X} = \mathbf{x}) < 1$

Optimal adjustment sets

Henckel *et al* 2022

Estimation of treatment effect by regression adjustment

Z: adjustment set

Estimation of treatment effect by regression adjustment

Z: adjustment set
Expected value of the outcome Y following an intervention on T

$$
\begin{aligned}
\mathbb{E}[Y \mid do(T=t)] &= \sum_y y p(y \mid do(T=t)) \\
&= \sum_y y \sum_z p(z) p(y \mid z, t) \\
&= \sum_z p(z) \sum_y y p(y \mid z, t) = \sum_z p(z) \mathbb{E}[Y \mid z, t] \\
&= \sum_z p(z) f(z, t)
\end{aligned}
$$

Estimation of treatment effect by regression adjustment

Z: adjustment set
Expected value of the outcome Y following an intervention on T

$$
\begin{aligned}
\mathbb{E}[Y \mid do(T=t)] &= \sum_y y p(y \mid do(T=t)) \\
&= \sum_y y \sum_z p(z) p(y \mid z, t) \\
&= \sum_z p(z) \sum_y y p(y \mid z, t) = \sum_z p(z) \mathbb{E}[Y \mid z, t] \\
&= \sum_z p(z) f(z, t)
\end{aligned}
$$

(Average) Treatment Effect (ATE)

$$
\begin{aligned}
\tau &= \mathbb{E}[Y \mid do(T=1)] - \mathbb{E}[Y \mid do(T=0)] \\
&= \sum_z p(z) [f(z,1) - f(z,0)]
\end{aligned}
$$

$\{(y_i, t_i, z_i)\}_{i=1}^{m}$: observational sample

$\{(y_i, t_i, z_i)\}_{i=1}^m$: observational sample

$\hat{f}(z, t)$: estimator of $\mathbb{E}[Y \mid z, t]$ based on a regression model

$\{(y_i, t_i, z_i)\}_{i=1}^m$: observational sample

$\hat{f}(z, t)$: estimator of $\mathbb{E}[Y \mid z, t]$ based on a regression model

Regression adjusted plug-in estimator of ATE

$$\hat{\tau}_1 = \frac{1}{m} \sum_{i=1}^{m} (\hat{f}(z, 1) - \hat{f}(z, 0))$$

$\{(y_i, t_i, z_i)\}_{i=1}^m$: observational sample

$\hat{f}(z,t)$: estimator of $\mathbb{E}[Y \mid z,t]$ based on a regression model

Regression adjusted plug-in estimator of ATE

$$\hat{\tau}_1 = \frac{1}{m}\sum_{i=1}^{m}(\hat{f}(z,1) - \hat{f}(z,0))$$

An alternative robust estimator

$$\hat{\tau}_2 = \frac{1}{m_1}\sum_{i:t_i=1}(y_i - \hat{f}(z_i,0)) + \frac{1}{m_0}\sum_{i:t_i=0}(\hat{f}(z_i,1) - y_i)$$

m_1: # obs in the treatment group
m_0 # obs in the control group

Further methods to estimate ATE

- Matching and Weighting
- Propensity Score-Methods

Causal inference with unobserved confounders

In general this is not possible

Causal inference with unobserved confounders

In general this is not possible

In some particular situations ATE can still be estimated

Causal inference with unobserved confounders

In general this is not possible

In some particular situations ATE can still be estimated

- Front-Door Adjustment (Mediator)

Causal inference with unobserved confounders

In general this is not possible

In some particular situations ATE can still be estimated

- Front-Door Adjustment (Mediator)
- Instrumental Variables (IV)
  - *Mendelian randomization*: special case

Causal inference with unobserved confounders

In general this is not possible

In some particular situations ATE can still be estimated

- Front-Door Adjustment (Mediator)
- Instrumental Variables (IV)
  - *Mendelian randomization*: special case
- Regression Discontinuity Design

DAGs *versus* Potential Outcomes

Reference
Guido Imbens
Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics. *Journal of Economic Literature*, 2020

DAGs

- J Pearl
  - Precursor: S Wright (path analysis)
  - Computer Science, Statistics, Epidemiology, ML/AI
- Graph captures the way researchers think about causality
- Powerful way to illustrate assumptions
- Systematic way to answer causal queries (do-calculus)
- Useful in complex models (large number of variables)

Potential outcomes

- D Rubin
  - Precursors: R Fisher, J Neyman (RCT's)
  - Economics, Econometrics, Social sciences
- Critical assumptions (monotonocity, convexity) easier to explain and incorporate
- Connects well to economic theory
- Has established canonical identification strategies for problems with a small number of variables
- Deals nicely with heterogeneity, study designs, estimation

Imbens's Conclusions

Imbens's Conclusions

- The DAG approach fully deserves the attention of all researchers and users of causal inference.

Imbens's Conclusions

- The DAG approach fully deserves the attention of all researchers and users of causal inference.
- Two key questions:
  - ☞ Should it be the framework of choice for all causal questions, or at least in the social sciences?

Imbens's Conclusions

- The DAG approach fully deserves the attention of all researchers and users of causal inference.
- Two key questions:
  - ☞ Should it be the framework of choice for all causal questions, or at least in the social sciences?
  - ☞ Should it be the starting point for teaching about causality?

Imbens's Conclusions

- The DAG approach fully deserves the attention of all researchers and users of causal inference.
- Two key questions:
  - ☞ Should it be the framework of choice for all causal questions, or at least in the social sciences?
  - ☞ Should it be the starting point for teaching about causality?
- Imbens's answer to both questions is NO

Table of Contents

1. From Statistical to Causal Models
2. Causal Modeling
3. Causal Discovery
4. Causal Reasoning
5. Conclusions

- Causality is a key-topic
  A Gelman & A Vehtari, 2021

  What are the Most Important Statistical Ideas of the Past 50 Years?

- Causality is a key-topic
  A Gelman & A Vehtari, 2021

  What are the Most Important Statistical Ideas of the Past 50 Years?
- DAGs or Potential Outcomes
  Opportunity not a problem
  Each has its own merits

- Causality is a key-topic
  A Gelman & A Vehtari, 2021

  What are the Most Important Statistical Ideas of the Past 50 Years?
- DAGs or Potential Outcomes
  Opportunity not a problem
  Each has its own merits
- DAG approach resonates better within the Computer Science community
  ML/AI
  It features already in several research areas
  Promises to have a tremendous impact in the near future

- Causality is a key-topic
  A Gelman & A Vehtari, 2021

  What are the Most Important Statistical Ideas of the Past 50 Years?
- DAGs or Potential Outcomes
  Opportunity not a problem
  Each has its own merits
- DAG approach resonates better within the Computer Science community
  ML/AI
  It features already in several research areas
  Promises to have a tremendous impact in the near future
  It is a bridge between Statistics-Data Science-AI

A Few Topics

- Causal Representation Learning.

A Few Topics

- Causal Representation Learning.
  Learn variables from data

A Few Topics

- Causal Representation Learning.
  Learn variables from data
  Interventions, reasoning, planning
- Causal Auto encoder

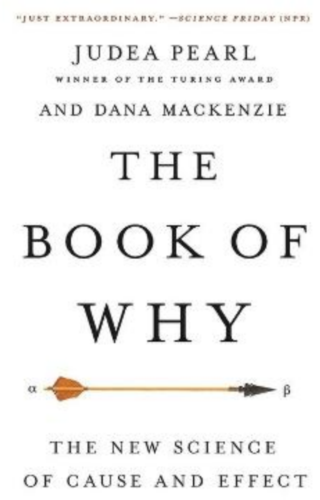
A Few Topics

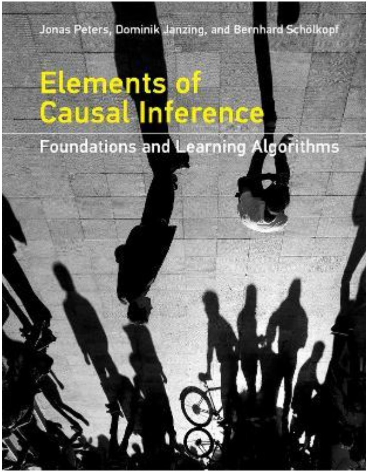
- Causal Representation Learning.
  Learn variables from data
  Interventions, reasoning, planning
- Causal Auto encoder
  Build a generative causal model
- Learning transferable mechanisms

A Few Topics

- Causal Representation Learning.
  Learn variables from data
  Interventions, reasoning, planning
- Causal Auto encoder
  Build a generative causal model
- Learning transferable mechanisms
  Solving multiple tasks
  in multiple environments

Selected References





From Statistical to Causal Learning

Bernhard Schölkopf
Max Planck Institute for Intelligent Systems, Tübingen, Germany
bs@tuebingen.mpg.de

Julius von Kügelgen
Max Planck Institute for Intelligent Systems, Tübingen, Germany
University of Cambridge, United Kingdom
jvk@tuebingen.mpg.de

April 4, 2022

Abstract

We describe basic ideas underlying research to build and understand artificially intelligent systems: from symbolic approaches via statistical learning to interventional models relying on concepts of causality. Some of the hard open problems of machine learning and AI are intrinsically related to causality, and progress may require advances in our understanding of how to model and infer causality from data.*

Mathematics Subject Classification 2020

Primary 68T05; Secondary 68Q32, 68T01, 68T10, 68T30, 68T37

Keywords

Causal inference, machine learning, causal representation learning