Introduction to Causality

Riccardo Massidda

May 21, 2024

Department of Computer Science
University of Pisa

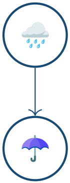
Probabilistic models, such as **Bayesian Networks**, enable the decomposition of joint probabilities

$$\mathbb{P}(🌧️, ☂️)$$

Probabilistic models, such as **Bayesian Networks**, enable the decomposition of joint probabilities

$$\mathbb{P}(🌧️, ☂️)$$

$$\mathbb{P}(☂️ \mid 🌧️) \cdot \mathbb{P}(🌧️)$$

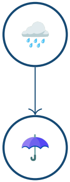


Probabilistic models, such as **Bayesian Networks**, enable the decomposition of joint probabilities

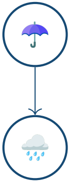
$$\mathbb{P}(🌧️, ☂️)$$

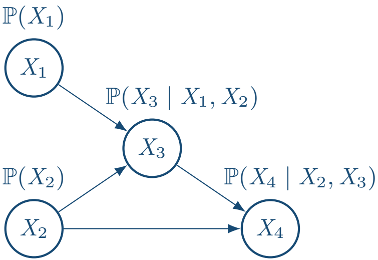
Causal ordering is **not** necessary.

$$\mathbb{P}(☂️ \mid 🌧️) \cdot \mathbb{P}(🌧️)$$



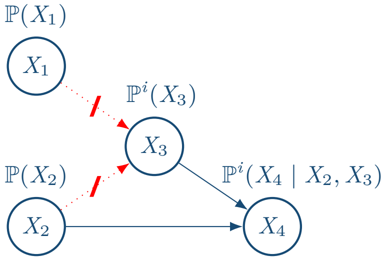
$$\mathbb{P}(🌧️ \mid ☂️) \cdot \mathbb{P}(☂️)$$





In a **Causal Bayesian Network**, edges represent **causal** relations.

Given causal ordering, we can represent external **interventions** on the model.

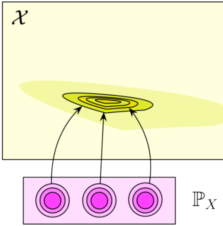


In a **Causal Bayesian Network**, edges represent **causal** relations.

Given causal ordering, we can represent external **interventions** on the model.

Independent Mechanism Principle (Peters et al. 2017)

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



(a) Probabilistic Model

(b) Causal Model

Figure 1: A probabilistic model represents a distribution $\mathbb{P}_{\boldsymbol{X}}$ on a set of random variables $\boldsymbol{X}$. For each intervention i, a causal model represents a distinct distribution $\mathbb{P}^i_{\boldsymbol{X}}$ on the same variables, where the observational distribution corresponds to the empty intervention. Illustration from Schölkopf et al. (2021).

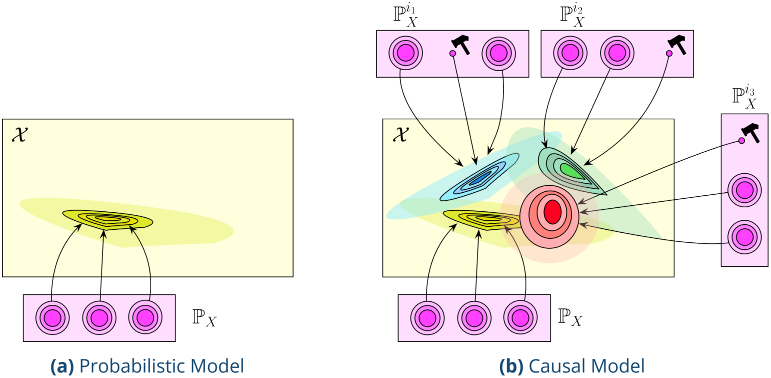


Figure 1: A probabilistic model represents a distribution $\mathbb{P}_{\boldsymbol{X}}$ on a set of random variables $\boldsymbol{X}$. For each intervention i, a causal model represents a distinct distribution $\mathbb{P}_{\boldsymbol{X}}^{i}$ on the same variables, where the observational distribution corresponds to the empty intervention. Illustration from Schölkopf et al. (2021).

Outline

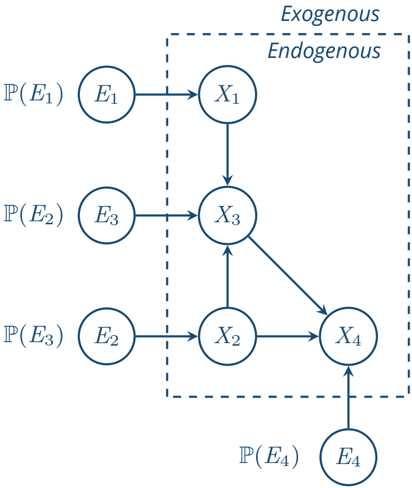
1. Structural Causal Models
2. Causal Reasoning
3. Causal Discovery
4. Causal Abstraction
5. Causal Representation Learning

Structural Causal Models

A Structural Causal Model

$$\mathcal{M} = (\boldsymbol{X}, \boldsymbol{E}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{E}}),$$

specifies the deterministic mechanisms $\boldsymbol{f}$ between a set of endogenous variables $\boldsymbol{X}$ and a set of exogenous variables $\boldsymbol{E}$ with distribution $\mathbb{P}_{\boldsymbol{E}}$.

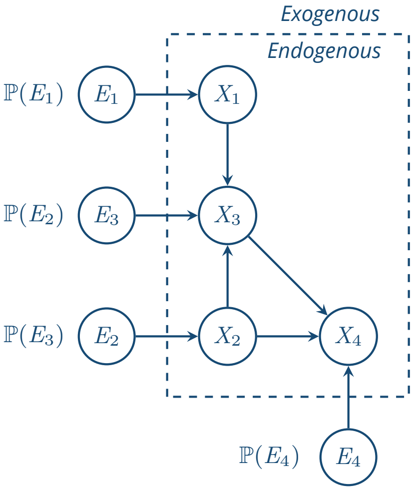


To each *endogenous* variable $X \in \boldsymbol{X}$, we assign an *exogenous* variable $E_X \in \boldsymbol{E}$.
The endogenous mechanism f_X of X is then defined as a function

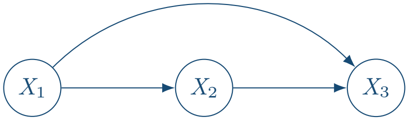
$$f_X : \mathcal{D}(\mathrm{Pa}(X) \cup E_X) \to \mathcal{D}(X).$$

Due to acyclicity, we can define the model reduction

$$\mathcal{M} : \mathcal{D}(\boldsymbol{E}) \to \mathcal{D}(\boldsymbol{X}).$$



Given the exogenous distribution $\mathbb{P}_{\boldsymbol{E}}$, the deterministic mechanisms $\boldsymbol{f}$ induce a distribution on the endogenous variables $\mathbb{P}_{\boldsymbol{X}}$.

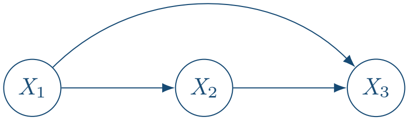


Structural Causal Model

$$
\begin{aligned}
X_1 &= E_1 \\
X_2 &= a \cdot X_1 + E_2 \\
X_3 &= b \cdot X_1 + c \cdot X_2 + E_3 \\
E_1, E_2, E_3 &\sim \mathcal{N}(0, \mathbf{I})
\end{aligned}
$$

Linear Gaussian SCMs

Given the exogenous distribution $\mathbb{P}_{\boldsymbol{E}}$, the deterministic mechanisms $\boldsymbol{f}$ induce a distribution on the endogenous variables $\mathbb{P}_{\boldsymbol{X}}$.



Structural Causal Model

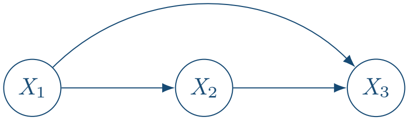
$$X_1 = E_1$$
$$X_2 = a \cdot X_1 + E_2$$
$$X_3 = b \cdot X_1 + c \cdot X_2 + E_3$$
$$E_1, E_2, E_3 \sim \mathcal{N}(0, \mathbf{I})$$

Causal Bayesian Network

$$X_1 \sim \mathcal{N}(0, 1)$$
$$X_2 \mid X_1 \sim \mathcal{N}(a \cdot X_1, 1)$$
$$X_3 \mid X_1, X_2 \sim \mathcal{N}(b \cdot X_1 + c \cdot X_2, 1)$$

Linear Gaussian SCMs

Given the exogenous distribution $\mathbb{P}_{\boldsymbol{E}}$, the deterministic mechanisms $\boldsymbol{f}$ induce a distribution on the endogenous variables $\mathbb{P}_{\boldsymbol{X}}$.



Structural Causal Model

$$X_1 = 2 \cdot E_1$$
$$X_2 = a \cdot X_1 + 2 \cdot E_2$$
$$X_3 = b \cdot X_1 + c \cdot X_2 + 2 \cdot E_3$$
$$E_1, E_2, E_3 \sim \mathcal{N}(0, \mathbf{I} \cdot 1/2)$$

Causal Bayesian Network

$$X_1 \sim \mathcal{N}(0, 1)$$
$$X_2 \mid X_1 \sim \mathcal{N}(a \cdot X_1, 1)$$
$$X_3 \mid X_1, X_2 \sim \mathcal{N}(b \cdot X_1 + c \cdot X_2, 1)$$

Hard Intervention

Given an SCM

$$\mathcal{M} = (\boldsymbol{X}, \boldsymbol{E}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{E}}),$$

a subset of variables $\boldsymbol{V} \subset \boldsymbol{X}$ and a setting $\boldsymbol{v} \in \mathcal{D}(\boldsymbol{V})$, an hard intervention $i = (\boldsymbol{V} \leftarrow \boldsymbol{v})$ results in a SCM $\mathcal{M}^i = (\boldsymbol{X}, \boldsymbol{E}, \boldsymbol{f}^i, \mathbb{P}_{\boldsymbol{E}})$, where

$$f_X^i = \begin{cases} v_X & X \in \boldsymbol{V} \\ f_X & X \notin \boldsymbol{V}, \end{cases}$$

for each endogenous variable $X \in \boldsymbol{X}$.

Hard Intervention

Given an SCM

$$\mathcal{M} = (\boldsymbol{X}, \boldsymbol{E}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{E}}),$$

a subset of variables $\boldsymbol{V} \subset \boldsymbol{X}$ and a setting $\boldsymbol{v} \in \mathcal{D}(\boldsymbol{V})$, an hard intervention $i = (\boldsymbol{V} \leftarrow \boldsymbol{v})$ results in a SCM $\mathcal{M}^i = (\boldsymbol{X}, \boldsymbol{E}, \boldsymbol{f}^i, \mathbb{P}_{\boldsymbol{E}})$, where

$$f_X^i = \begin{cases} v_X & X \in \boldsymbol{V} \\ f_X & X \notin \boldsymbol{V}, \end{cases}$$

for each endogenous variable $X \in \boldsymbol{X}$.

Interventions can be: soft, stochastic, perfect, imperfect, ...

Causal Reasoning

Layer	Query	Model
Association	$P(Y \mid X = x)$	BN
Intervention	$P(Y \mid \mathrm{do}(X \leftarrow x))$	CBN
Counterfactual	$P(Y = y' \mid Y = y, X = x, \mathrm{do}(X \leftarrow x')), x \neq x'$	SCM

Layer	Query	Model
Association	$P(Y \mid X = x)$	BN
Intervention	$P(Y \mid \mathrm{do}(X \leftarrow x))$	CBN
Counterfactual	$P(Y = y' \mid Y = y, X = x, \mathrm{do}(X \leftarrow x')), x \neq x'$	SCM

Layer	Query	Model
Association	$P(Y \mid X = x)$	BN
Intervention	$P(Y \mid \mathrm{do}(X \leftarrow x))$	CBN
Counterfactual	$P(Y = y' \mid Y = y, X = x, \mathrm{do}(X \leftarrow x')), x \neq x'$	SCM

Average Treatment Effect

Suppose that we have a binary *treatment* variable X and an *outcome* variable Y. Then, we can compute the **Average Treatment Effect** as

$$\mathrm{ATE}(X, Y) = \mathbb{E}_{y \sim Y|\mathrm{do}(X \leftarrow 1)}\left[y\right] - \mathbb{E}_{y \sim Y|\mathrm{do}(X \leftarrow 0)}\left[y\right]$$

🔨

Suppose that we have a binary *treatment* variable X and an *outcome* variable Y. Then, we can compute the **Average Treatment Effect** as

$$\begin{aligned}\mathrm{ATE}(X, Y) &= \mathbb{E}_{y\sim Y|\mathrm{do}(X\leftarrow 1)}[y] - \mathbb{E}_{y\sim Y|\mathrm{do}(X\leftarrow 0)}[y] \\ &= \sum_{y\in\mathcal{D}(Y)} y \cdot p_Y^{\mathrm{do}(X\leftarrow 1)}(y) - \sum_{y\in\mathcal{D}(Y)} y \cdot p_Y^{\mathrm{do}(X\leftarrow 0)}(y).\end{aligned}$$

To compute $\mathrm{ATE}(X, Y)$ or any other causal estimate, we need to compute the interventional distribution $p_Y^{\mathrm{do}(X \leftarrow x)}$.

Causal Identifiability

To compute $\mathrm{ATE}(X, Y)$ or any other causal estimate, we need to compute the interventional distribution $p_Y^{\mathrm{do}(X \leftarrow x)}$.

The distribution is **causally identifiable** if it can be computed from the observational distribution.

Causal Identifiability

To compute $\mathrm{ATE}(X, Y)$ or any other causal estimate, we need to compute the interventional distribution $p_Y^{\mathrm{do}(X \leftarrow x)}$.

The distribution is **causally identifiable** if it can be computed from the observational distribution.

In general, this is not possible, but we can use the **do-calculus** to identify the conditions under which it is possible.

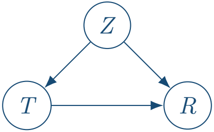
Causal Identifiability

To compute $\mathrm{ATE}(X, Y)$ or any other causal estimate, we need to compute the interventional distribution $p_Y^{\mathrm{do}(X \leftarrow x)}$.

The distribution is **causally identifiable** if it can be computed from the observational distribution.

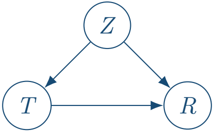
In general, this is not possible, but we can use the **do-calculus** to identify the conditions under which it is possible.

The do-calculus is a **complete** set of rules that can be easily applied to any causal graph.

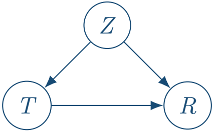


An hospital offers two distinct surgeries (T) for kidney stones which they assign depending on the whether the stones are large ($Z = 1$) or small ($Z = 0$). Depending on the size of the stones and the treatment, the success rate (R) varies.

Backdoor Criterion



	$Z=0$	$Z=1$
$T=a$	0.93 (81/87)	0.73 (192/263)
$T=b$	0.87 (234/270)	0.69 (55/80)



We can define the best treatment by computing the interventional success rate.

$$
\begin{aligned}
P(R=1 \mid \mathrm{do}(T \leftarrow a)) = \\
P(R=1 \mid T=a, Z=0)P(Z=0) \\
+P(R=1 \mid T=a, Z=1)P(Z=1) \\
= 0.93 \cdot 0.51 + 0.73 \cdot 0.49 = \mathbf{0.832},
\end{aligned}
$$

$$
\begin{aligned}
P(R=1 \mid \mathrm{do}(T \leftarrow b)) = \\
P(R=1 \mid T=b, Z=0)P(Z=0) \\
+P(R=1 \mid T=b, Z=1)P(Z=1) \\
= 0.87 \cdot 0.51 + 0.69 \cdot 0.49 = 0.782.
\end{aligned}
$$

By computing the conditional probabilities, we can easily see how conditioning differs from intervening.

$$P(R = 1 \mid T = a) = 0.780, \qquad P(R = 1 \mid \mathrm{do}(T \leftarrow a)) = \mathbf{0.832},$$
$$P(R = 1 \mid T = b) = \mathbf{0.830}. \qquad P(R = 1 \mid \mathrm{do}(T \leftarrow b)) = 0.782.$$

Given the probability of success when *observing* the treatments, we would have, arguably incorrectly, chosen treatment b.

Layer	Query	Model
Association	$P(Y \mid X = x)$	BN
Intervention	$P(Y \mid \mathrm{do}(X \leftarrow x))$	CBN
Counterfactual	$P(Y = y' \mid Y = y, X = x, \mathrm{do}(X \leftarrow x')), x \neq x'$	SCM

Example 3.4 (Eye disease) There exists a rather effective treatment for an eye disease. For 99% of all patients, the treatment works and the patient gets cured ($B = 0$); if untreated, these patients turn blind within a day ($B = 1$). For the remaining 1%, the treatment has the opposite effect and they turn blind ($B = 1$) within a day. If untreated, they regain normal vision ($B = 0$).

Which category a patient belongs to is controlled by a rare condition ($N_B = 1$) that is unknown to the doctor, whose decision whether to administer the treatment ($T = 1$) is thus independent of N_B. We write it as a noise variable N_T.

Assume the underlying SCM

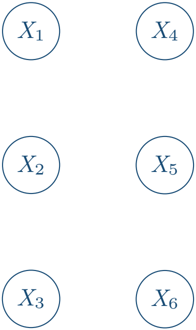
$$\mathfrak{C}: \begin{array}{rcl} T & := & N_T \\ B & := & T \cdot N_B + (1 - T) \cdot (1 - N_B) \end{array} \tag{3.5}$$

with Bernoulli distributed $N_B \sim \text{Ber}(0.01)$; note that the corresponding causal graph is $T \rightarrow B$.

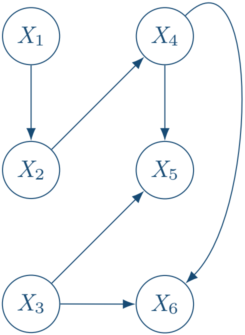
Now imagine a specific patient with poor eyesight comes to the hospital and goes blind ($B = 1$) after the doctor administers the treatment ($T = 1$). We can now ask the counterfactual question *"What would have happened had the doctor administered treatment $T = 0$?"*

Causal Discovery

Causal Discovery, or causal *learning*, consists of determining causal relations between variables $\boldsymbol{X}$ from their observational distribution $\mathbb{P}_{\boldsymbol{X}}$.



Causal Discovery, or causal *learning*, consists of determining causal relations between variables $\boldsymbol{X}$ from their observational distribution $\mathbb{P}_{\boldsymbol{X}}$.



Graph Identifiability

Given a set of assumptions $\boldsymbol{A}$, we say that the graph $\mathcal{G}$ of a causal model $\mathcal{M}$ is identifiable from the distribution $\mathbb{P}_{\boldsymbol{X}}$ whenever there does not exist another causal model $\mathcal{M}'$ satysfying $\boldsymbol{A}$ with a different graph $\mathcal{G}'$ but the same observational distribution $\mathbb{P}_{\boldsymbol{X}}$.

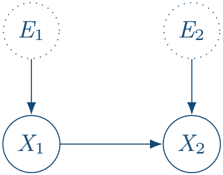
Lachapelle et al. 2019

Graph Identifiability

Given a set of assumptions $\boldsymbol{A}$, we say that the graph $\mathcal{G}$ of a causal model $\mathcal{M}$ is identifiable from the distribution $\mathbb{P}_{\boldsymbol{X}}$ whenever there does not exist another causal model $\mathcal{M}'$ satysfying $\boldsymbol{A}$ with a different graph $\mathcal{G}'$ but the same observational distribution $\mathbb{P}_{\boldsymbol{X}}$.

Lachapelle et al. 2019

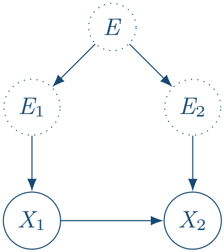
$$\boldsymbol{A} = \emptyset \implies \text{No identifiability.}$$



Causally Sufficient Model
$E_1 \perp\!\!\!\perp E_2$

A model is causally sufficient whenever there are no unobserved confounders. This equates to assuming that there is no selection bias and the exonous terms $\boldsymbol{E}$ are marginally independent, i.e.,

$$\forall i \neq j. \quad E_i \perp\!\!\!\perp E_j.$$

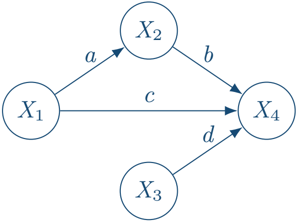


Causally Insufficient Model

$E_1 \not\perp\!\!\!\perp E_2$

A model is causally sufficient whenever there are no unobserved confounders. This equates to assuming that there is no selection bias and the exonous terms $\boldsymbol{E}$ are marginally independent, i.e.,

$$\forall i \neq j. \quad E_i \perp\!\!\!\perp E_j.$$

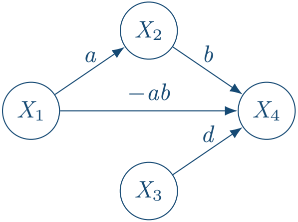


Faithful Model,

$$c \neq -ab$$

A model is causally faithfull whenever all conditional independences in the distribution $\mathbb{P}_{\boldsymbol{X}}$ imply d-separations in the graph $\mathcal{G}$, i.e.,

$$\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B} \mid \boldsymbol{C} \implies \boldsymbol{A} \perp\!\!\!\perp_{\mathcal{G}} \boldsymbol{B} \mid \boldsymbol{C}.$$

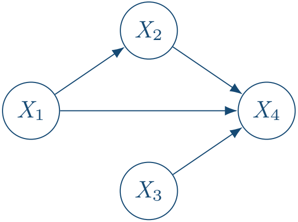


Faithfulness Violation,

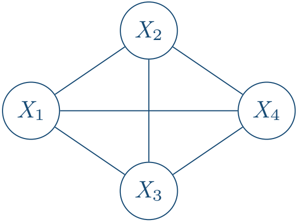
$$X_1 \perp\!\!\!\perp X_4 \quad \text{but} \quad X_1 \not\perp\!\!\!\perp_{\mathcal{G}} X_4.$$

A model is causally faithfull whenever all conditional independences in the distribution $\mathbb{P}_{\boldsymbol{X}}$ imply d-separations in the graph $\mathcal{G}$, i.e.,

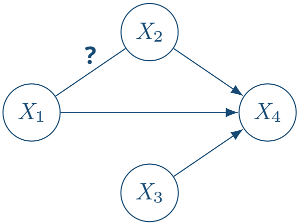
$$\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B} \mid \boldsymbol{C} \implies \boldsymbol{A} \perp\!\!\!\perp_{\mathcal{G}} \boldsymbol{B} \mid \boldsymbol{C}.$$



The PC algorithm is a constraint-based discovery method that iteratively removes and orients edges.



The PC algorithm is a constraint-based discovery method that iteratively removes and orients edges.



The PC algorithm is a constraint-based discovery method that iteratively removes and orients edges.

The algorithm returns the Markov equivalence class of the true causal graph, which can contain multiple graphs.

Model Class

Description	Mechanism	f	Id.
General SCM	$X := f_X(\mathrm{Pa}(X), E_X)$	—	✗

Description	Mechanism	f	Id.
General SCM	$X := f_X(\mathrm{Pa}(X), E_X)$	—	✗
ANM	$X := f_X(\mathrm{Pa}(X)) + E_X$	Nonlinear	✓
CAM	$X := \sum_{X' \in \mathrm{Pa}(X)} f(X') + E_X$	Nonlinear	✓
Gaussian ANM	$X := \langle \mathbf{w}, \mathrm{Pa}(X) \rangle + E_X$	Linear	✗
Non-Gaussian ANM	$X := f_X(\mathrm{Pa}(X)) + E_X$	Linear	✓
Gaussian Eq. Var	$X := \langle \mathbf{w}, \mathrm{Pa}(X) \rangle + E_X$	Linear	✓

Continuous Causal Discovery (CCD) approaches try to recast combinatorial discovery algorithms as optimization problems.

$$\min_{\mathcal{G}} \mathcal{S}(\mathcal{G}, \mathcal{D}_{\mathbf{X}})$$

s.t. $\mathcal{G}$ is acyclic.

There are two main open problems:

Continuous Causal Discovery (CCD) approaches try to recast combinatorial discovery algorithms as optimization problems.

$$\min_{\mathcal{G}} \mathcal{S}(\mathcal{G}, \mathcal{D}_{\mathbf{X}})$$

s.t. $\mathcal{G}$ is acyclic.

There are two main open problems:

- How to *efficiently* enforce acyclicity on the solution $\mathcal{G}$?

Continuous Causal Discovery (CCD) approaches try to recast combinatorial discovery algorithms as optimization problems.

$$\min_{\mathcal{G}} \mathcal{S}(\mathcal{G}, \mathcal{D}_{\mathbf{X}})$$

s.t. $\mathcal{G}$ is acyclic.

There are two main open problems:

- How to *efficiently* enforce acyclicity on the solution $\mathcal{G}$?
- How to encode assumptions in the *score* function $\mathcal{S}$?

Causal Abstraction

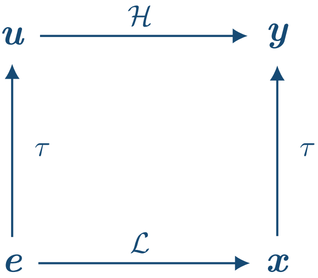
Traditional causal discovery algorithms require a large number of samples and are computationally expensive.

Traditional causal discovery algorithms require a large number of samples and are computationally expensive.

Can we represent a system with a *simpler* model at an *higher-level* of abstraction?

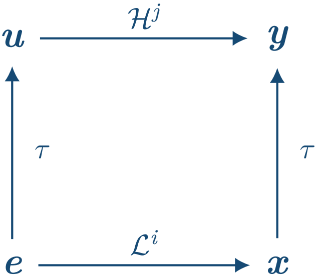
$$\boldsymbol{u} \xrightarrow{\mathcal{H}} \boldsymbol{y}$$

$$\boldsymbol{e} \xrightarrow{\mathcal{L}} \boldsymbol{x}$$



Given an abstraction function τ, an SCM $\mathcal{H}$ is an abstraction of $\mathcal{L}$ if the diagram commutes, i.e.,

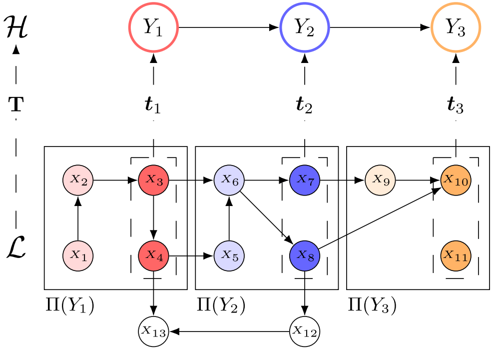
$$\tau \circ \mathcal{L} = \mathcal{H} \circ \tau.$$



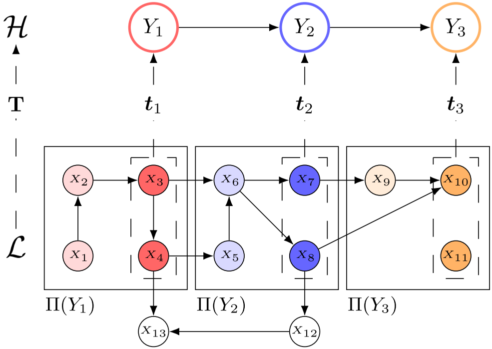
Given an abstraction function τ, an SCM $\mathcal{H}$ is an abstraction of $\mathcal{L}$ if the diagram commutes, i.e.,

$$\tau \circ \mathcal{L}^i = \mathcal{H}^j \circ \tau,$$

for any intervention i.

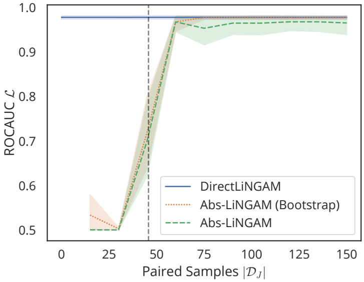


In the *linear* case, we can use abstraction to cluster larger causal graphs.

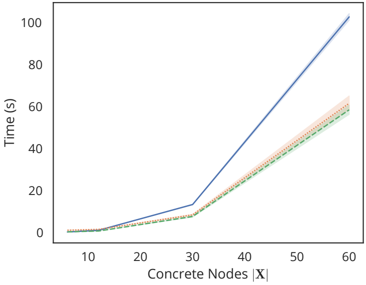


In the *linear* case, we can use abstraction to cluster larger causal graphs.

This improves the complexity of causal reasoning and allows for more interpretable models.



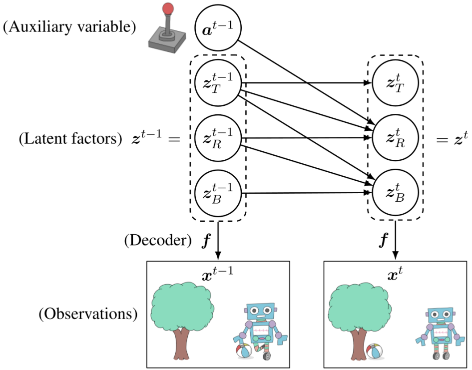
(a) Performance over Paired Samples $|\mathcal{D}_J|$



(b) Execution Time (s) over Graph Size $|\boldsymbol{X}|$

Introducing abstract information in the LiNGAM pipeline, we gain significant speedup (2x) in execution time (b, *right*) without performance loss (a, *left*).

Causal Representation Learning



In many contexts, we can assume that high-dimensional observations $\boldsymbol{x}$ are generated through a decoder function

$$f : \mathcal{D}(Z) \to \mathcal{D}(X)$$

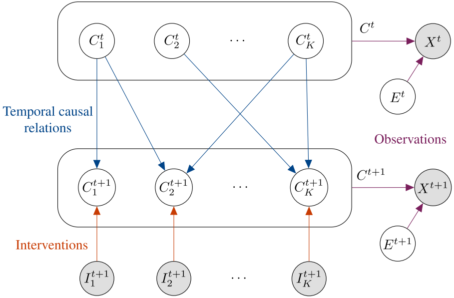
from a set of latent *causal* variables $\boldsymbol{Z}$.

Disentanglement

Factors are *statistically* independent. Altering a factor should only affect a single dimension of the data.

Causal Representation

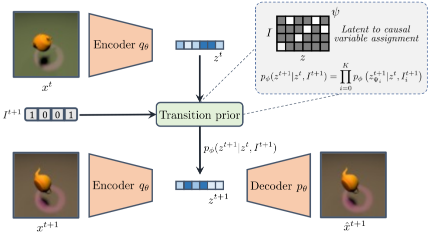
Factors are *causally* independent. Altering a factor might affect other factors, but we can independently manipulate them.



We need interventional samples to learn causal representations such as TempoRal Intervened Sequences (TRIS),

$$\mathcal{D} = \{(\mathbf{x}_t, i, \mathbf{x}_{t+1})\},$$

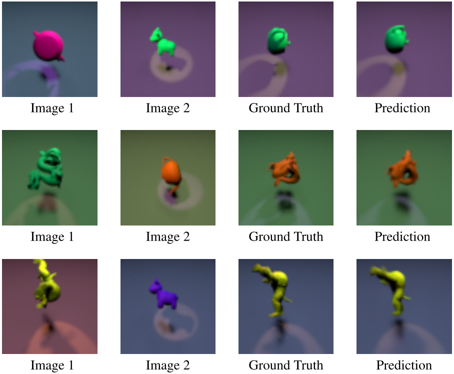
where we can observe the state of the model before *and* after an intervention i.



We need interventional samples to learn causal representations such as TempoRal Intervened Sequences (TRIS),

$$\mathcal{D} = \{(\mathbf{x}_t, i, \mathbf{x}_{t+1})\},$$

where we can observe the state of the model before *and* after an intervention i.

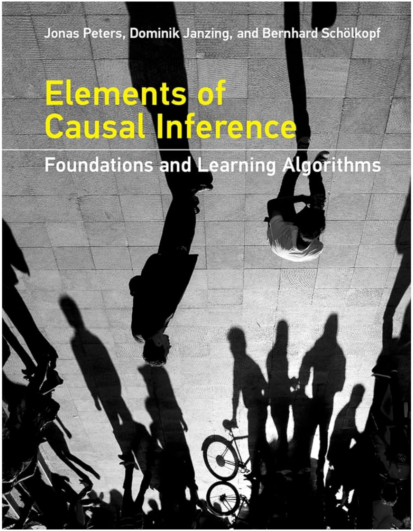


We need interventional samples to learn causal representations such as TempoRal Intervened Sequences (TRIS),

$$\mathcal{D} = \{(\mathbf{x}_t, i, \mathbf{x}_{t+1})\},$$

where we can observe the state of the model before *and* after an intervention i.

Conclusion

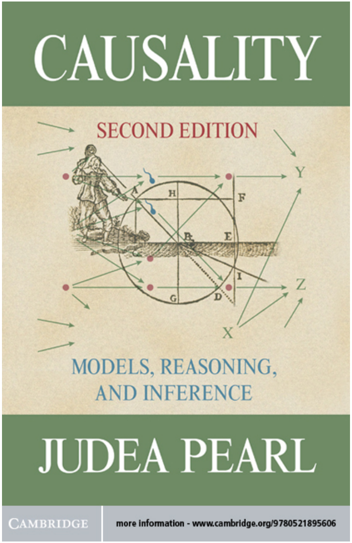


Elements of Causal Inference: Foundations and Learning Algorithms

Jonas **Peters**, Dominik Janzing, Bernhard Schölkopf

The MIT Press, 2017

Main References



Causality:
Models, Reasoning, and Inference

Judea **Pearl**

Cambridge University Press, 2009

Conclusion 👋

Conclusion 👋

- Causal *Graphical Models* are a powerful tool to represent causal relationships between variables.

Conclusion

👋

- Causal *Graphical Models* are a powerful tool to represent causal relationships between variables.
- Efficient causal *discovery* from observational data is a challenging problem.

Conclusion

- Causal *Graphical Models* are a powerful tool to represent causal relationships between variables.
- Efficient causal *discovery* from observational data is a challenging problem.
- Causal *Abstraction* and Causal *Representation Learning* are promising research directions for future applications.

Conclusion 👋

- Causal *Graphical Models* are a powerful tool to represent causal relationships between variables.
- Efficient causal *discovery* from observational data is a challenging problem.
- Causal *Abstraction* and Causal *Representation Learning* are promising research directions for future applications.

Riccardo Massidda
Università di Pisa
Dept. of Computer Science
Office 298B, 2nd Floor

pages.di.unipi.it/massidda/
riccardo.massidda@phd.unipi.it

References

Lachapelle, Sébastien et al. (2019). **"Gradient-based neural dag learning".** In: *arXiv preprint arXiv:1906.02226*.

Lippe, Phillip et al. (2022). **"CITRIS: Causal identifiability from temporal intervened sequences".** In: *International Conference on Machine Learning*. PMLR, pp. 13557–13603.

Peters, Jonas, Dominik Janzing, and Bernhard Schölkopf (2017). ***Elements of causal inference: foundations and learning algorithms*.** The MIT Press.

Schölkopf, Bernhard et al. (2021). **"Towards causal representation learning".** In: *arXiv preprint arXiv:2102.11107*.