Causal Discovery

Alexis Bellot

Google DeepMind

Agenda

1. **Data Science**: Two paradigms
2. **Causal discovery**: What is the structure of the world and its relationship to data?
3. **Algorithms** for causal discovery
4. We have run a causal discovery algorithm, **now what?**

alexisbellot.github.io/Website/

Two Paradigms for Data Science

The **data-centric paradigm**:

> All wisdom comes from the sampling distribution of the data P. The challenge is to manipulate the distribution and ultimately fit the data in order to maximize success on the **training set**.

Two Paradigms for Data Science

The **data-centric paradigm**:

All wisdom comes from the sampling distribution of the data P. The challenge is to manipulate the distribution and ultimately fit the data in order to maximize success on the **training set**.

The **scientific paradigm**:

There is a world out there that we seek to model and understand. It is not about the data itself but about the **underlying mechanisms in the world**. What does the data tell me about the world out there?

Capabilities of Understanding

Capabilities of Understanding

1. Predict future events from present/past **observations**
2. Predict the consequences of hypothetical **actions**, such as treatment plans
3. Provide **explanations** (attribute reasons) for unanticipated events, why?
4. Design new informed experiments, seek new observations, **imagine** hypothetical scenarios

Typical questions

1. What **effect** can we expect from a given treatment given to patients with stage III cancer?
2. What fraction of health-care expenditure can be **attributed** to respiratory illnesses?
3. I have been suffering from obesity for two years, would my BMI be different **had I adhered** to a vegan diet?
4. Can hospital admission statistics prove systematic **discrimination** against a given minority group?

Typical questions

1. What **effect** can we expect from a given treatment in patients with stage III cancer?
2. What fraction of health-care expenditure can be **attributed** to respiratory illnesses?
3. I have been suffering from obesity for two years, would my BMI be different **had I** adhered to a vegan diet?
4. Can hospital admission statistics prove systematic **discrimination** against a given minority group?

$$Y = f(X), \qquad Y \leftarrow f(X)$$

The Origins of the Causal Revolution

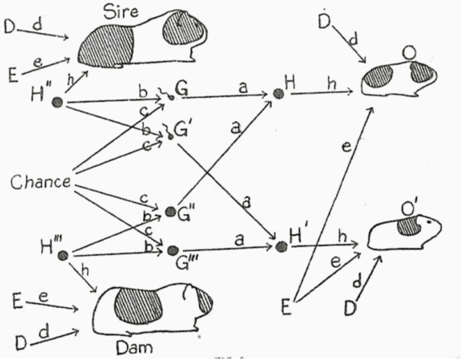
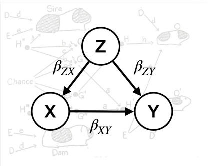


Figure: Path diagram showing the influence of heredity and environment on the inheritance of color in the guinea pig, reproduced from Wright (1920).

The Origins of the Causal Revolution



In a **linear Gaussian model**

$$Z \leftarrow U_Z, \quad X \leftarrow \beta_{ZX} Z + U_X, \quad Y \leftarrow \beta_{XY} X + \beta_{ZY} Z + U_Y, \quad U_Z, U_X, U_Y \sim \text{Gaussian}(0,1)$$

There is a **correspondence** between correlations in data P and path coefficients β

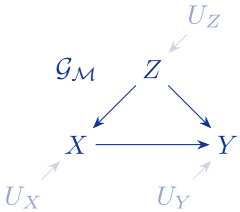
$$\mathbb{E}_P[ZX] = \mathbb{E}_P[Z \cdot (\beta_{ZX} Z + U_X)] = \beta_{ZX}$$
$$\mathbb{E}_P[ZY] = \beta_{XY}\beta_{ZX} + \beta_{ZY}$$
$$\mathbb{E}_P[XY] = \beta_{XY} + \beta_{ZX}\beta_{ZY}.$$

By solving this set of equations and inferring values for β, one begins to **understand** our system.

Path diagrams are the historical “parent” of **causal graphs**.

Think of causal graphs as **summaries** of the underlying model.

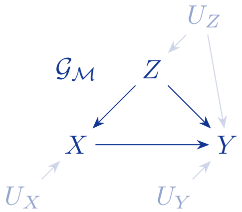
$$\mathcal{M} := \begin{cases} Z \leftarrow f_Z(U_Z) \\ X \leftarrow f_X(Z, U_X) \\ Y \leftarrow f_Y(X, Z, U_Y) \\ \\ P(U_Z, U_X, U_Y) \end{cases}$$



Path diagrams are the historical “parent” of **causal graphs**.

Causal graphs as **summaries** of the underlying model.

$$\mathcal{M} := \begin{cases} Z \leftarrow f_Z(U_Z) \\ X \leftarrow f_X(Z, U_X) \\ Y \leftarrow f_Y(X, Z, U_Y, U_Z) \\ \\ P(U_Z, U_X, U_Y) \end{cases}$$



Path diagrams are the historical "parent" of **causal graphs**.

Causal graphs as **summaries** of the underlying model.

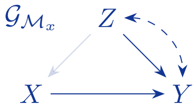
$$\mathcal{M} := \begin{cases} Z \leftarrow f_Z(U_Z) \\ X \leftarrow f_X(Z, U_X) \\ Y \leftarrow f_Y(X, Z, U_Y, U_Z) \\ \\ P(U_Z, U_X, U_Y) \end{cases}$$

$$\mathcal{G}_{\mathcal{M}} \quad \begin{tikzcd} & Z \arrow[dl] \arrow[dr] \arrow[dr, dashed, bend left, leftrightarrow] & \\ X \arrow[rr] & & Y \end{tikzcd}$$

Path diagrams are the historical “parent” of **causal graphs**.

Causal graphs as **summaries** of the underlying model.

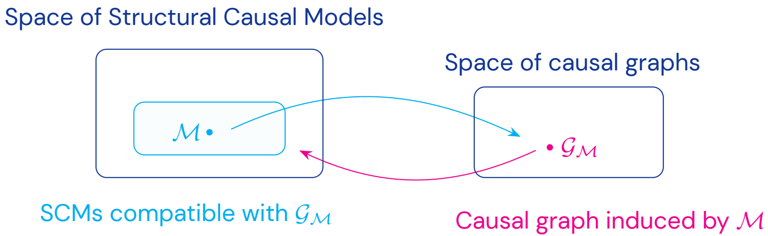
$$\mathcal{M}_x := \begin{cases} Z \leftarrow f_Z(U_Z) \\ X \leftarrow x \\ Y \leftarrow f_Y(X, Z, U_Y, U_Z) \\ \\ P(U_Z, U_X, U_Y) \end{cases}$$



$\mathbb{E}_P[Y \mid do(x)] = \mathbb{E}_{P_{\mathcal{M}_x}}[Y]$ stands for *the expectation of Y under a distribution for Y generated from $\mathcal{M}$ after fixing $X \leftarrow x$.*

Path diagrams are the historical “parent” of **causal graphs**.

Causal graphs as **summaries** of the underlying model.



Causal Inference

Systematically deducing causal statements from assumptions and data.

(1) Query

e.g. $\mathbb{E}[Y \mid do(x)]$

(2) Data distribution

e.g. $P(x, y, z)$

(3) Assumption

$$\begin{tikzcd} X \arrow[r] \arrow[rr, bend left, dashed, leftrightarrow] & Z \arrow[r] & Y \end{tikzcd}$$

$\longrightarrow$

Causal inference engine

$\longrightarrow$

Uniquely computable from P?

If so, how?

e.g. $\sum_z P(z \mid x) \sum_{x'} P(y \mid x', z) P(x')$

What if you cannot confidently make assumptions?

Answer 1 – Give up … In general, you **need** some domain knowledge to answer questions that relate to "understanding" (Bareinboim et al., 2022, Pearl's Causal Hierarchy).

What if you cannot confidently make assumptions?

Answer 1 – Give up … In general, you **need** some domain knowledge to answer questions that relate to “understanding” (Bareinboim et al., 2022, Pearl’s Causal Hierarchy).

Answer 2 – Conduct **sensitivity analysis**. How much would different causal assumptions influence my conclusions?

What if you cannot confidently make assumptions?

Answer 1 – Give up … In general, you **need** some domain knowledge to answer questions that relate to “understanding” (Bareinboim et al., 2022, Pearl’s Causal Hierarchy).

Answer 2 – Conduct **sensitivity analysis**. How much would different causal assumptions influence my conclusions?

Answer 3 – **Learn from data** as much as as possible about the causal graph.

What if you cannot confidently make assumptions?

Answer 1 – Give up … In general, you need some domain knowledge to answer questions that relate to “understanding” (Bareinboim et al., 2022, Pearl’s Causal Hierarchy)

Answer 2 – Conduct sensitivity analysis. How much do different causal assumptions influence my conclusions.

Answer 3 – **Learn from data** as much as possible about the causal graph.

1. Understand the implications that causal graphs have on the data you observe.
2. Reverse engineer these implications to determine what set of graphs are plausible.

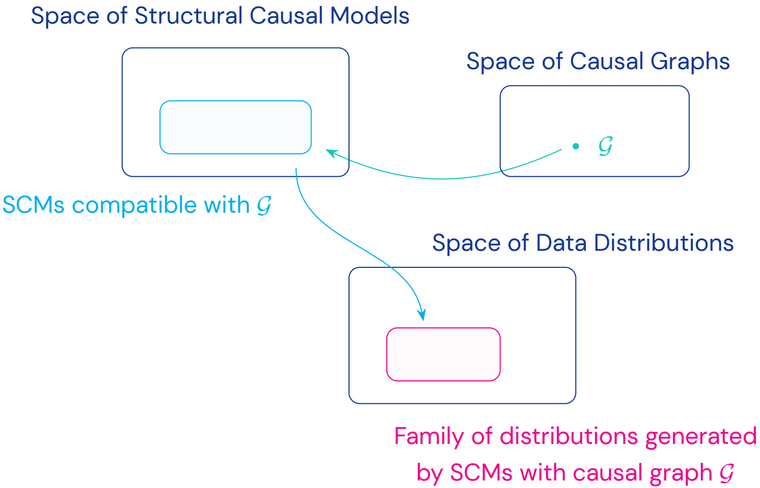
Important distinction to keep in mind

Causal inference involves **predicting the value of a causal effect** of interest, typically given a causal graph and data.

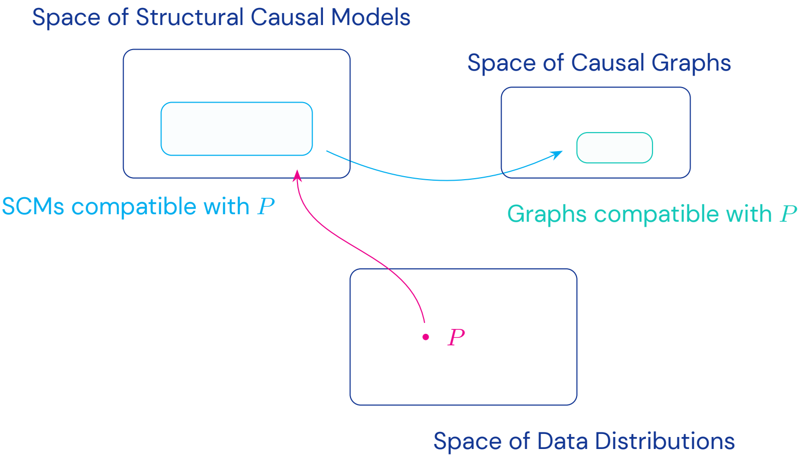
Causal discovery involves **learning the causal graph** from data.

How does data relate to causal models?

What does the causal graph tell us about data?



What does data tell us about the causal graph?



Fundamental Law of Conditional Independence

$$
\begin{tikzcd}
X \arrow[r] & R \arrow[r] \arrow[d] & S \arrow[r] & T \arrow[d] & U \arrow[l] & V \arrow[l] \arrow[r] \arrow[d] & Y \\
& W & & P & & Q &
\end{tikzcd}
$$

Causal graphs can be used to read off **conditional independencies** in the distribution of data P using the **d-separation criterion** (Pearl, 1988).

Fundamental Law of Conditional Independence

d-separation criterion. Given a causal graph $\mathcal{G}$,

$$(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_{\mathcal{G}} \Rightarrow (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_P.$$

Conditional independence is an equality relation between probabilities that can be verified with data.

$(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_P$ means $P(\mathbf{x} \mid \mathbf{z}, \mathbf{y}) = P(\mathbf{x} \mid \mathbf{z})$ for any $\mathbf{x}, \mathbf{z}, \mathbf{y}$.

d-separation in graphs

Rule 1. A and B are d–connected, if there is an **unblocked path** between them, that is a path that does not contain **colliders**. If no such path exists, we say that A and B are d–separated.

$$
\begin{tikzcd}
X \arrow[r] & R \arrow[r] \arrow[d] & S \arrow[r] & T \arrow[d] & U \arrow[l] & V \arrow[l] \arrow[r] \arrow[d] & Y \\
 & W & & P & & Q &
\end{tikzcd}
$$

X and T are not d–separated, denoted $(X \not\perp\!\!\!\perp T)_{\mathcal{G}}$.

d-separation in graphs

Rule 1. A and B are d–connected, if there is an **unblocked path** between them, that is a path that does not contain **colliders**. If no such path exists, we say that A and B are d–separated.

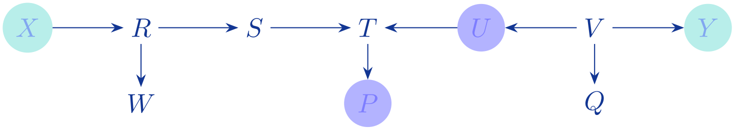
$$\begin{tikzcd}
X \arrow[r] & R \arrow[r] \arrow[d] & S \arrow[r] & T \arrow[d] & U \arrow[l] & V \arrow[l] \arrow[r] \arrow[d] & Y \\
& W & & P & & Q &
\end{tikzcd}$$

X and U are d–separated, denoted $(X \perp\!\!\!\perp U)_{\mathcal{G}}$.

d-separation in graphs

Rule 2. A and B are d–connected, **conditioned** on a set of nodes $\mathbf{Z}$, if there is a collider–free path between A and B that traverses no member of $\mathbf{Z}$.

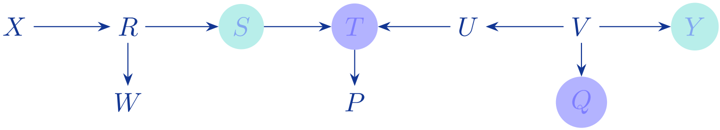
$(X \perp\!\!\!\perp Y \mid U, P)_{\mathcal{G}}$?



d-separation in graphs

Rule 3. If a **collider** is a member of the conditioning set $\mathbf{Z}$, or has a **descendant** in $\mathbf{Z}$, then it no longer blocks any path that traces this collider.

$(S \perp\!\!\!\perp Y \mid T, Q)_{\mathcal{G}}$?



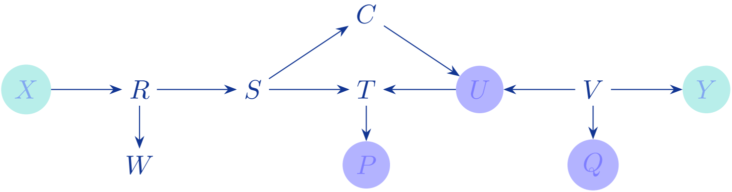
d-separation in graphs

Rule 3. If a **collider** is a member of the conditioning set $\mathbf{Z}$, or has a **descendant** in $\mathbf{Z}$, then it no longer blocks any path that traces this collider.

$(S \perp\!\!\!\perp Y \mid P, Q)_{\mathcal{G}}$?

$$
\begin{tikzcd}
X \arrow[r] & R \arrow[r] \arrow[d] & S \arrow[r] & T \arrow[d] & U \arrow[l] & V \arrow[l] \arrow[r] \arrow[d] & Y \\
 & W & & P & & Q &
\end{tikzcd}
$$

$(X \perp\!\!\!\perp Y \mid P, Q, U)_{\mathcal{G}}$?



Summary

1. Important property: causal graphs imply conditional independencies in data.

$$(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_{\mathcal{G}} \Rightarrow (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_P \text{ or equivalently } (\mathbf{X} \not\!\perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_P \Rightarrow (\mathbf{X} \not\!\perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_{\mathcal{G}}.$$

Statistical dependencies in data are **measurable traces** of the (unobserved) SCM.

2. This opens an avenue for **model testing**.

Model Testing Example: Smoking and lung cancer

Model Testing Example: Smoking and lung cancer

(In an alternative world) found gene (G) such that makes smoking (S) and cancer (C) independent.

That is, we collected some data $\{s^{(n)}, g^{(n)}, c^{(n)}\}_{n=1}^{N}$ and found empirically that $S \perp\!\!\!\perp C \mid G$, that is $P(S \mid C, G) = P(S \mid G)$.

Model Testing Example: Smoking and lung cancer

(In an alternative world) found gene (G) such that makes smoking (S) and cancer (C) independent.

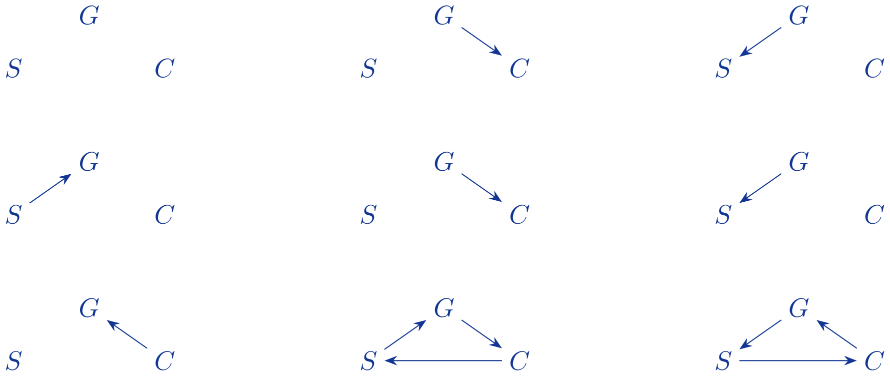
That is, we collected some data $\{s^{(n)}, g^{(n)}, c^{(n)}\}_{n=1}^{N}$ and found empirically that $S \perp\!\!\!\perp C \mid G$, that is $P(S \mid C, G) = P(S \mid G)$.

Causal discovery is the problem of looking for causal graphs $\mathcal{G}$ of three variables $\{S, C, G\}$ that could be reasonable candidates for this (in)dependence structure.

$(S \perp\!\!\!\perp C \mid G)_P, (S \not\perp\!\!\!\perp C)_P, (S \not\perp\!\!\!\perp G)_P, (G \not\perp\!\!\!\perp C)_P$

Many potential graphs can be **ruled out** as:

THM: $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_{\mathcal{G}} \Rightarrow (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_P$, or equivalently $(\mathbf{X} \not\perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_P \Rightarrow (\mathbf{X} \not\perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_{\mathcal{G}}$



$(S \perp\!\!\!\perp C \mid G)_P, (S \not\!\perp\!\!\!\perp C)_P, (S \not\!\perp\!\!\!\perp G)_P, (G \not\!\perp\!\!\!\perp C)_P$

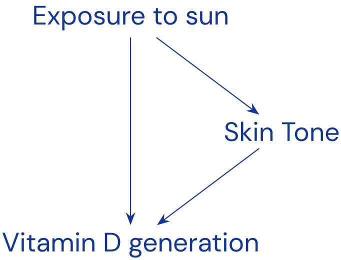
Others would be **weird / unexpected** causal explanations.

$$\begin{tikzcd} & G \arrow[dl] \arrow[dr] & \\ S \arrow[rr] & & C \end{tikzcd} \qquad \begin{tikzcd} & G \arrow[dl] \arrow[dr] & \\ S & & C \arrow[ll] \end{tikzcd} \qquad \begin{tikzcd} & G \arrow[dr] & \\ S \arrow[ur] \arrow[rr] & & C \end{tikzcd} \qquad \begin{tikzcd} & G \arrow[dl] & \\ S & & C \arrow[ul] \arrow[ll] \end{tikzcd}$$

Theoretically they are **not** excluded as:

$(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_{\mathcal{G}} \Rightarrow (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_P$ does not imply that $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_P \Rightarrow (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_{\mathcal{G}}$

Some natural systems likely display a statistical independence **without** an underlying structural separation.



However, *exposure to sun* has been observed to be **independent** of *vitamin D generation*.

Faithfulness

A distribution P is said to be **faithful** to $\mathcal{G}$ if

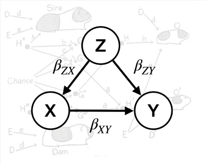
$$(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_P \Rightarrow (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_{\mathcal{G}}$$

Back to **Smoking example**

$$\begin{tikzcd} & G \arrow[dl] \arrow[dr] & \\ S \arrow[rr] & & C \end{tikzcd} \qquad \begin{tikzcd} & G \arrow[dl] \arrow[dr] & \\ S & & C \arrow[ll] \end{tikzcd} \qquad \begin{tikzcd} & G \arrow[dr] & \\ S \arrow[ur] \arrow[rr] & & C \end{tikzcd} \qquad \begin{tikzcd} & G \arrow[dl] & \\ S & & C \arrow[ul] \arrow[ll] \end{tikzcd}$$

are violations of faithfulness as $(S \perp\!\!\!\perp C \mid G)_P \not\Rightarrow (S \perp\!\!\!\perp C \mid G)_{\mathcal{G}}$

Why do we think Faithfulness is reasonable?



True underlying systems is a linear Gaussian model of the form,

$$
\begin{aligned}
Z &\leftarrow U_Z, \\
X &\leftarrow \beta_{ZX} Z + U_X, \\
Y &\leftarrow \beta_{XY} X + \beta_{ZY} Z + U_Y.
\end{aligned}
$$

Imagine we observe the independence $(X \perp\!\!\!\perp Y)_P$, that is $\mathbb{E}_P[XY] = 0$.

Violation of faithfulness would mean $(X \perp\!\!\!\perp Y)_P \not\Rightarrow (X \perp\!\!\!\perp Y)_{\mathcal{G}}$ which requires

$$\mathbb{E}_P[XY] = \beta_{XY} + \beta_{ZX}\beta_{ZY} = 0.$$

Under faithfulness,

$$(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_{\mathcal{G}} \iff (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_P$$

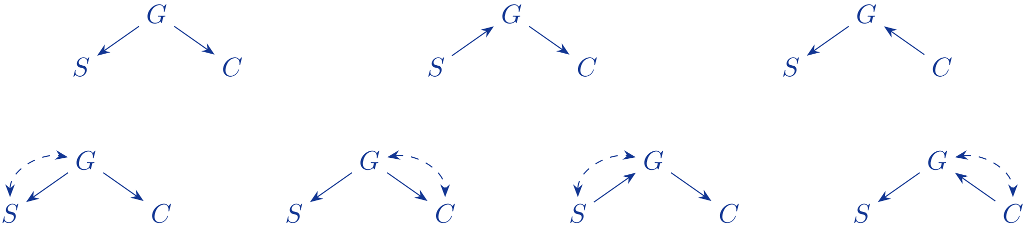
Back to **Smoking example**: $(S \perp\!\!\!\perp C \mid G)_P, (S \not\perp\!\!\!\perp C)_P, (S \not\perp\!\!\!\perp G)_P, (G \not\perp\!\!\!\perp C)_P$

$$\begin{tikzcd} & G \arrow[dl] \arrow[dr] & \\ S & & C \end{tikzcd} \qquad \begin{tikzcd} & G \arrow[dr] & \\ S \arrow[ur] & & C \end{tikzcd} \qquad \begin{tikzcd} & G \arrow[dl] & \\ S & & C \arrow[ul] \end{tikzcd}$$

Under faithfulness,

$$(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_{\mathcal{G}} \iff (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_P$$

Back to **Smoking example**: $(S \perp\!\!\!\perp C \mid G)_P, (S \not\perp\!\!\!\perp C)_P, (S \not\perp\!\!\!\perp G)_P, (G \not\perp\!\!\!\perp C)_P$



Colliders

Imagine we record a fourth variable: **the price of cigarettes P**.

In the data, we find that $(P \perp\!\!\!\perp G, C)_P$ and $(P \not\perp\!\!\!\perp G, C \mid S)_P$.

Under faithfulness, P must be d–separated from G and C.

$$\begin{tikzcd} P \arrow[dr] & & G \arrow[dl] \arrow[dr] & \\ & S & & C \end{tikzcd} \quad \begin{tikzcd} P \arrow[dr] & & G \arrow[dr] & \\ & S \arrow[ur] & & C \end{tikzcd} \quad \begin{tikzcd} P \arrow[dr] & & G \arrow[dl] & \\ & S & & C \arrow[ul] \end{tikzcd}$$

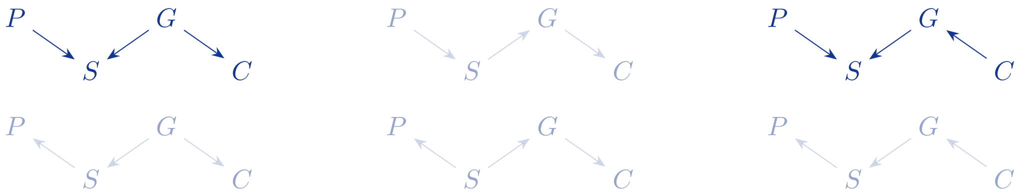
$$\begin{tikzcd} P & & G \arrow[dl] \arrow[dr] & \\ & S \arrow[ul] & & C \end{tikzcd} \quad \begin{tikzcd} P & & G \arrow[dr] & \\ & S \arrow[ul] \arrow[ur] & & C \end{tikzcd} \quad \begin{tikzcd} P & & G \arrow[dl] & \\ & S \arrow[ul] & & C \arrow[ul] \end{tikzcd}$$

Colliders

Imagine we record a fourth variable: **the price of cigarettes P**.

In the data, we find that $(P \perp\!\!\!\perp G, C)_P$ and $(P \not\perp\!\!\!\perp G, C \mid S)_P$.

Under faithfulness, P must be d–separated from G and C.



Colliders

Imagine we record a fourth variable: **the price of cigarettes P**.

In the data, we find that $(P \perp\!\!\!\perp G, C)_P$ and $(P \not\!\perp\!\!\!\perp G, C \mid S)_P$.

Under faithfulness, P must be d-separated from G and C.

$$\begin{tikzcd}
P \arrow[dr] & & G \arrow[dl] \arrow[dr] & \\
& S & & C
\end{tikzcd}
\qquad
\begin{tikzcd}
P \arrow[dr] & & G \arrow[dl] & \\
& S & & C \arrow[ul]
\end{tikzcd}
\qquad
\begin{tikzcd}
P \arrow[dr] & & G \arrow[dl] \arrow[dr, dash] & \\
& S & & C
\end{tikzcd}$$

Representation of equivalence class

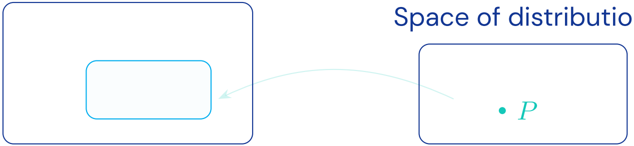
Algorithms

Most causal discovery algorithms are designed to **exploit faithfulness**

$$(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_P \iff (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_{\mathcal{G}}$$

Space of causal graphs

Space of distributions



Graphs consistent with an assumed **faithful** distribution P

Distribution with a set of observed independencies

Constraint-based Causal Discovery

Causal discovery based on independence testing

Constrained-based causal discovery algorithms explicitly **test for conditional independencies** to determine what edges we can rule out in the underlying graph.

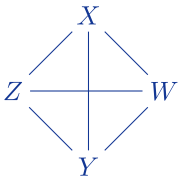
Two phases, starting from a fully connected (undirected) graph:

1. Remove edges: If two variables are conditionally independent remove edge (**skeleton**).
2. Orient edges.

Phase 1: Skeleton recovery

1. Recover **skeleton**: Start with complete graph, remove edge between any two nodes that can be made (conditionally) independent.

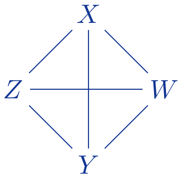
4 variables: X, Z, W, Y.



Phase 1: Skeleton recovery

1. Recover **skeleton**: Start with complete graph, remove edge between any two nodes that can be made (conditionally) independent.

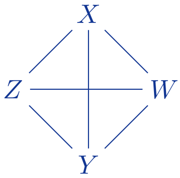
4 variables: X, Z, W, Y, and we find (empirically) that $(Z \perp\!\!\!\perp W \mid X)_P, (X \perp\!\!\!\perp Y \mid Z, W)_P$.

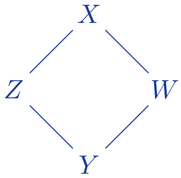


Phase 1: Skeleton recovery

1. Recover **skeleton**: Start with complete graph, remove edge between any two nodes that can be made (conditionally) independent.

4 variables: X, Z, W, Y, and we find (empirically) that $(Z \perp\!\!\!\perp W \mid X)_P, (X \perp\!\!\!\perp Y \mid Z, W)_P$.





Phase 2: Edge orientation

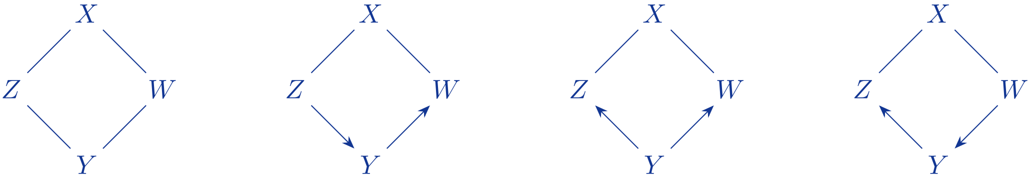
1. Recover **skeleton**: Start with complete graph, remove edge between any two nodes that can be made (conditionally) independent.
2. **Orient** edges as much as possible: look for v–structures.

4 variables: X, Z, W, Y, and we find (empirically) that $(Z \perp\!\!\!\perp W \mid X)_P, (X \perp\!\!\!\perp Y \mid Z, W)_P$.

Phase 2: Edge orientation

1. Recover **skeleton**: Start with complete graph, remove edge between any two nodes that can be made (conditionally) independent.
2. **Orient** edges as much as possible: look for v–structures

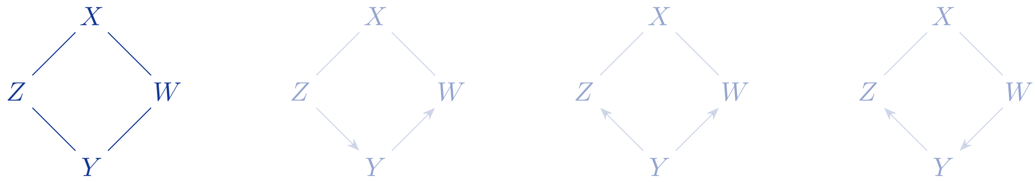
4 variables: X, Z, W, Y, and we find (empirically) that $(Z \perp\!\!\!\perp W \mid X)_P, (X \perp\!\!\!\perp Y \mid Z, W)_P$.



Phase 2: Edge orientation

1. Recover **skeleton**: Start with complete graph, remove edge between any two nodes that can be made (conditionally) independent.
2. **Orient** edges as much as possible: look for v–structures

4 variables: X, Z, W, Y, and we find (empirically) that $(Z \perp\!\!\!\perp W \mid X)_P, (X \perp\!\!\!\perp Y \mid Z, W)_P$.



All can be ruled out because $(Z \not\perp\!\!\!\perp W \mid Y, X)_P$.

Phase 2: Edge orientation

1. Recover **skeleton**: Start with complete graph, remove edge between any two nodes that can be made (conditionally) independent.
2. **Orient** edges as much as possible: look for v–structures

4 variables: X, Z, W, Y, and we find (empirically) that $(Z \perp\!\!\!\perp W \mid X)_P, (X \perp\!\!\!\perp Y \mid Z, W)_P$.

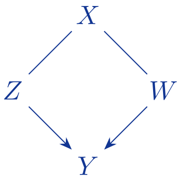
$$
\begin{tikzcd}
 & X \arrow[dl, dash] \arrow[dr, dash] & \\
Z \arrow[dr] & & W \arrow[dl] \\
 & Y &
\end{tikzcd}
$$

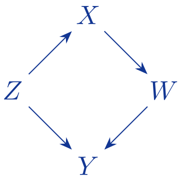
Is there anything else that can be established?

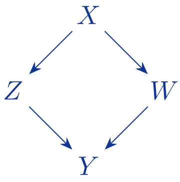
Phase 2: Edge orientation

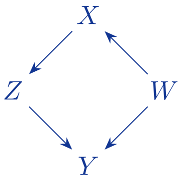
1. Recover **skeleton**: Start with complete graph, remove edge between any two nodes that can be made (conditionally) independent.
2. **Orient** edges as much as possible: look for v–structures

4 variables: X, Z, W, Y, and we find (empirically) that $(Z \perp\!\!\!\perp W \mid X)_P, (X \perp\!\!\!\perp Y \mid Z, W)_P$.









With some additional rules to orient edges, this algorithm is called **IC / PC algorithm** (Spirtes et al., 2000; Verma and Pearl, 1990).

Theorem. *Under an assumption of faithfulness, with an oracle for conditional independence, the IC/PC algorithm is guaranteed to recover the Markov equivalence class of the true graph.*

Good software packages for constraint-based causal discovery: `causal-learn` (Zheng et al., 2023) in python, `pcalg` (Kalisch et al., 2012) in R.

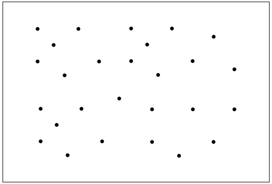
IC* / FCI algorithm in the presence of **unobserved confounding**.

Score-based causal discovery

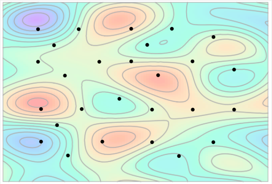
A **different** approach to causal discovery:

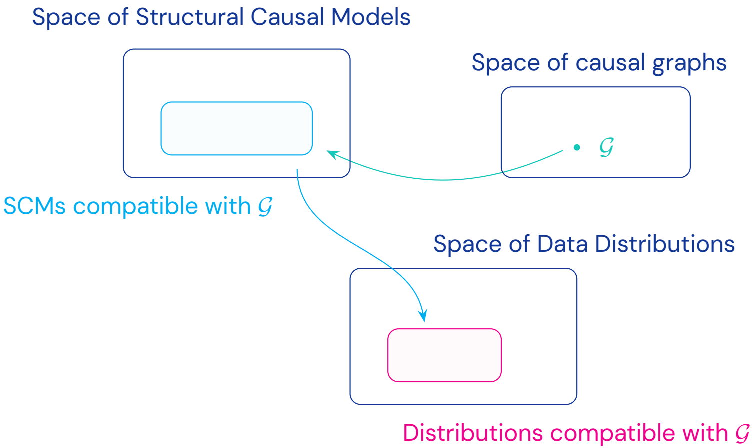
1. Define a criterion or **score** $\mathcal{S}$ to evaluate how well the causal graph fits the data.
2. **Search** over the space of causal graphs for a graph achieving the maximal score.

All possible causal graphs



Scores for all possible causal graphs

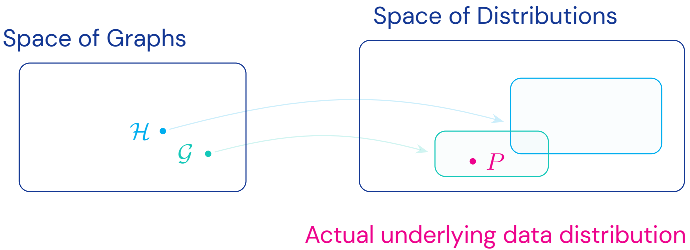




Each graph $\mathcal{G}$ is associated with a family of distributions $\{P_{\mathcal{M}}(\mathbf{v}) : \mathcal{M} \in \mathbb{M}(\mathcal{G})\}$.

What makes a good score?

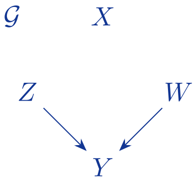
1. **Soundness**: Better score for valid causal explanation.

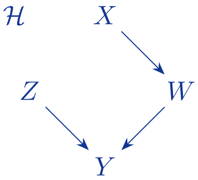


What makes a good score?

1. **Soundness**: Better score for valid causal explanation.

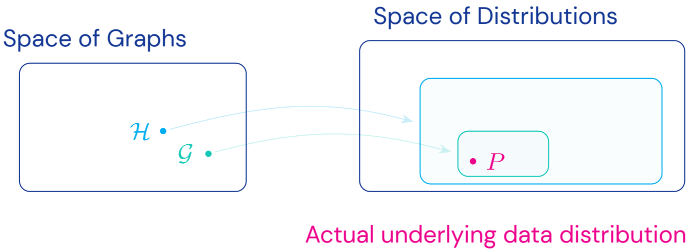
In the data, $X \not\perp\!\!\!\perp W$.





What makes a good score?

1. **Soundness**: Better score for valid causal explanation
2. **Parsimony**: Smaller models are preferred.



What makes a good score?

1. **Soundness**: Better score for valid causal explanation
2. **Parsimony**: Smaller models are preferred.

In the data, $Z \perp\!\!\!\perp X$.

$$
\begin{tikzcd}
\mathcal{G} & X \arrow[dl] \arrow[dr] & \\
Z \arrow[dr] & & W \arrow[dl] \\
& Y &
\end{tikzcd}
\qquad\qquad
\begin{tikzcd}
\mathcal{H} & X \arrow[dr] & \\
Z \arrow[dr] & & W \arrow[dl] \\
& Y &
\end{tikzcd}
$$

Score-based causal discovery is the product of a long legacy within the **Bayesian model selection** (Gelman et al., 1995) literature.

A score $\mathcal{S} : (\mathcal{G}, \mathbf{v}) \mapsto \mathbb{R}$.

The **marginal likelihood** as a score

$$P(\mathcal{G} \mid \mathbf{v}) \propto P(\mathcal{G}) \underbrace{P(\mathbf{v} \mid \mathcal{G})}_{\text{marginal likelihood}}$$

The marginal likelihood $P(\mathbf{v} \mid \mathcal{G})$ is difficult to compute.

Most methods attempt to **approximate** its value.

The **Bayesian information criterion** (BIC) for a candidate model $\mathcal{G}$ is an asymptotic approximation to the marginal likelihood.

It requires a parametric model for the distribution of variables $P(\mathbf{v} \mid \mathcal{G}, \boldsymbol{\theta})$.

The BIC is defined as

$$\mathcal{S}_{\mathsf{BIC}}(\mathbf{v}, \mathcal{G}) := -2 \underbrace{\log P(\mathbf{v} \mid \mathcal{G}, \hat{\boldsymbol{\theta}}_{\mathsf{MLE}})}_{\text{log-likelihood of the data}} + \underbrace{|\boldsymbol{\theta}| \log n}_{\text{Penalty for models with more parameters}}$$

The BIC is (asymptotically) **sound** and **parsimonious** for scoring causal graphs (without unobserved confounding) (Haughton, 1988).

BIC: Example

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

Consider scoring the causal graph $\mathcal{G}$,
assuming the underlying SCM is **linear** and **Gaussian**,

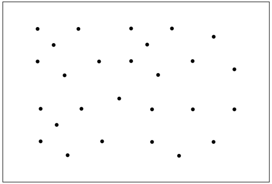
$$\begin{bmatrix} Z \\ W \\ Y \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \theta_{YZ} & \theta_{YW} & 0 \end{bmatrix} \begin{bmatrix} Z \\ W \\ Y \end{bmatrix} + \begin{bmatrix} U_Z \\ U_W \\ U_Y \end{bmatrix}, \qquad U_i \sim \mathcal{N}(0, \sigma_i^2), \quad i \in \{Z, W, Y\}$$

A total of 5 parameters: $(\theta_{YZ}, \theta_{YW}, \sigma_Z^2, \sigma_W^2, \sigma_Y^2)$.
Maximum likelihood estimates and log-likelihood can be computed in closed-form.

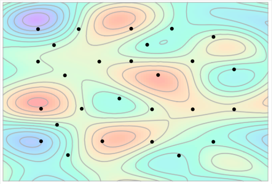
$$\mathcal{S}_{\text{BIC}} = -2\,\mathbf{log}\,P(z, w, y \mid \hat{\theta}_{YZ}, \hat{\theta}_{YW}, \hat{\sigma}_Z^2, \hat{\sigma}_W^2, \hat{\sigma}_Y^2) + 5\,\mathbf{log}\,n$$

Searching in the space of graphs

All possible causal graphs



Scores for all possible causal graphs



Number of DAGs with 2 variables: **3**

Number of DAGs with 3 variables: **25**

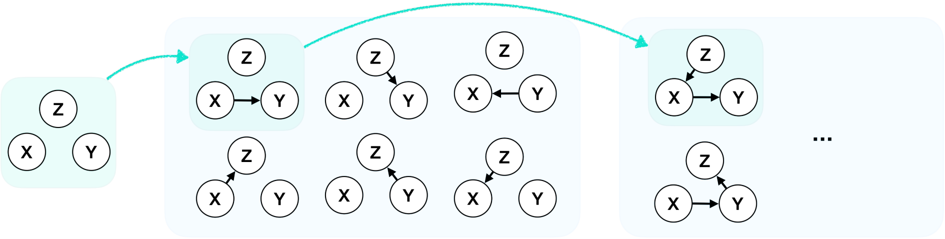
Number of DAGs with 4 variables: **543**

Number of DAGs with 5 variables: **29281**

Greedy search

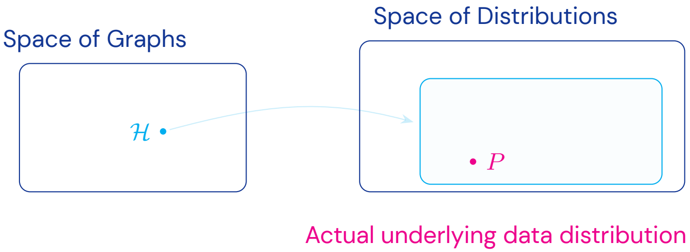
Progressively explore the space of DAGs by making **local** moves (Meek, 1997).

1. Evaluate / score **neighbouring** graphs
2. Move to highest scoring candidate graph



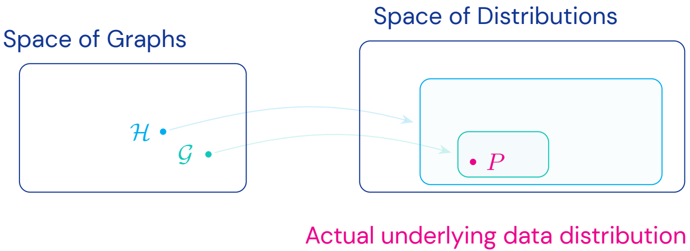
2 Phases in Greedy search algorithm

First, **add** edges until score cannot be improved.



2 Phases in Greedy search algorithm

Second, **remove** edges until score cannot be improved.



Greedy Equivalence Search

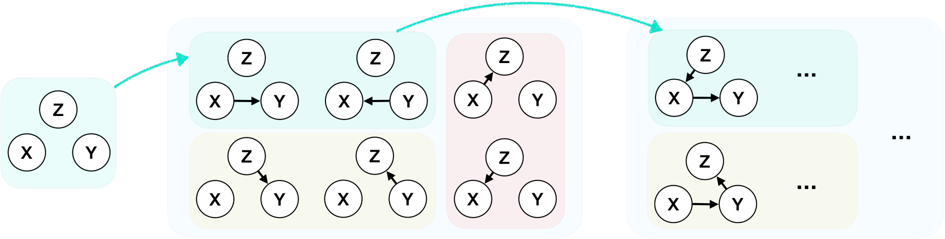
Progressively explore the space of **equivalence classes** (Meek, 1997).

1. Evaluate / score neighbouring equivalence classes
2. Move to highest scoring candidate equivalence class

Greedy Equivalence Search

Progressively explore the space of equivalence classes (Meek, 1997).

1. Evaluate / score neighbouring **equivalence classes**
2. Move to highest scoring candidate **equivalence class**



Greedy Equivalence Search

Theorem (Chickering, 2002). *Under an assumption of faithfulness, the equivalence class returned by Greedy Equivalence Search (GES) coincides with the equivalence class of the true causal graph asymptotically.*

Search with gradient-based optimization

Search with gradient-based optimization

A Directed Acyclic Graph (DAG) can be modelled by an **adjacency matrix**.

$$\begin{bmatrix} Z \\ W \\ Y \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \theta_{YZ} & \theta_{YW} & 0 \end{bmatrix} \begin{bmatrix} Z \\ W \\ Y \end{bmatrix} + \begin{bmatrix} U_Z \\ U_W \\ U_Y \end{bmatrix}$$

Is equivalent to saying

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

\+ Linear functional relations

Search with gradient-based optimization

A Directed Acyclic Graph (DAG) can be modelled by an **adjacency matrix**.

$$\begin{bmatrix} X_1 \\ \vdots \\ X_k \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & \dots \\ \vdots & \ddots & \\ w_{k1} & & w_{kk} \end{bmatrix} \begin{bmatrix} X_1 \\ \vdots \\ X_k \end{bmatrix} + \begin{bmatrix} U_1 \\ \vdots \\ U_k \end{bmatrix}$$

Presumably we could recover a good estimate of W by running linear regressions, and interpret non-zero entries as the presence of an edge.

W to be a valid DAG must be **acyclic**.

Search with gradient-based optimization

Learning causal graphs can be thought of as parameter **optimization** under **constraints**.

$$\max_{W \in \mathbb{R}^{k \times k}} \text{Score}(W, \mathbf{X}), \quad \text{subject to } W \text{ being a DAG.}$$

Acyclicity

What does an acyclic W look like?

$$W = \begin{bmatrix} 0 & w_{12} & w_{13} \\ w_{21} & 0 & w_{23} \\ w_{31} & w_{32} & 0 \end{bmatrix}$$

$w_{ij} = 0$ if and only if $X_j \to X_i$ not in $\mathcal{G}_W$.

One useful note: W encodes the **paths of length 1** in $\mathcal{G}_W$, i.e. $w_{11} = 0$ means that there is no path of length 1 that starts and ends at X_1.

Acyclicity

What does an acyclic W look like?

$$W^2 = \begin{bmatrix} 0 & w_{12} & w_{13} \\ w_{21} & 0 & w_{23} \\ w_{31} & w_{32} & 0 \end{bmatrix} \begin{bmatrix} 0 & w_{12} & w_{13} \\ w_{21} & 0 & w_{23} \\ w_{31} & w_{32} & 0 \end{bmatrix} = \begin{bmatrix} w_{12}w_{21} + w_{13}w_{31} & \cdots \\ \vdots & \ddots \\ & \end{bmatrix}$$

What does it mean for the first diagonal entry to be zero?

Diagonal entries of W^2 give **paths of length 2** starting and ending at the same node.

Acyclicity

A square matrix W that does **not have cycles of any length** satisfies the following equality (Zheng et al., 2018),

$$\text{Trace}(W + W^2 + W^3 + \dots) = 0$$

Equivalent to,

$$\text{Trace}\left(I + W + \frac{1}{2!}W^2 + \frac{1}{3!}W^3 + \dots\right) = \text{Trace}(I)$$

Equivalent to,

$$\text{Trace}\left(\mathbf{exp}\, W\right) - d = 0$$

Search with gradient-based optimization

Learning causal graphs can be thought of as parameter **optimization** under **constraints**.

$$\max_{W \in \mathbb{R}^{k \times k}} \text{Score}(W, \mathbf{X}), \quad \text{subject to } W \text{ being a DAG.}$$

written,

$$\max_{W \in \mathbb{R}^{k \times k}} \text{Score}(W, \mathbf{X}) + \lambda \cdot (\text{Trace}(\mathbf{exp}\, W) - d)$$

Causal discovery with two variables only

Typically, if $X \not\perp\!\!\!\perp Y$ in a system of two variables we cannot establish anything about their causal structure, that is,

$$X \text{ ———— } Y$$

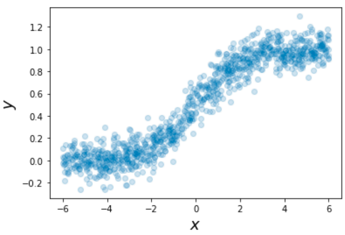
Under some conditions, we can find, however, an **asymmetry** in data generated by a model $X \to Y$ or by a model $Y \to X$.

Example: Asymmetry in bi-variate associations

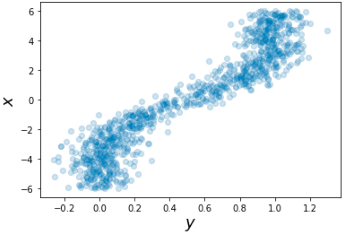
$$\begin{tikzcd} U_X \arrow[d] & U_Y \arrow[d] \\ X \arrow[r] & Y \end{tikzcd}$$

SCM for X, Y is

$$Y \leftarrow \text{logistic}(X) + U_Y, \quad X \leftarrow U_X, \quad U_X \sim \mathcal{U}(-6, 6), \quad U_Y \sim \mathcal{N}(0, 0.01)$$



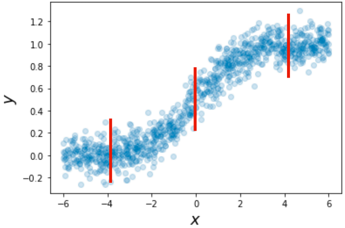
(a) Y as a function of X.



(b) X as a function of Y.

Example: Asymmetry in bi-variate associations

$$\begin{tikzcd} U_X \arrow[d] & U_Y \arrow[d] \\ X \arrow[r] & Y \end{tikzcd}$$



(a) Y as a function of X.

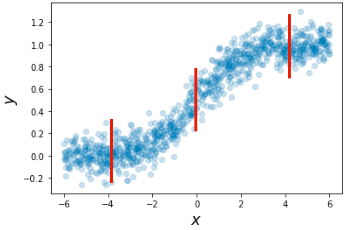
Fit $Y \approx f(X)$.

Look at the **residuals** $\hat{U}_Y := Y - f(X)$.

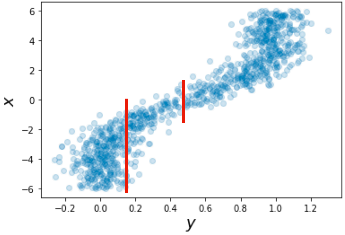
You find that approximately $\hat{U}_Y \perp\!\!\!\perp X$.

Example: Asymmetry in bi-variate associations

$$\begin{tikzcd} U_X \arrow[d] & U_Y \arrow[d] \\ X \arrow[r] & Y \end{tikzcd}$$



(a) Y as a function of X.



(b) X as a function of Y.

Fit $X \approx f(Y)$.

Look at the **residuals** $\hat{U}_X := X - f(Y)$.

You find that approximately $\hat{U}_X \not\perp\!\!\!\perp Y$.

We expect regression in one direction to give independent residuals but not in the other! (Shimizu et al., 2006)

Criterion for **inferring causal direction**:

If residuals are independent of regression covariate, then correct causal direction.

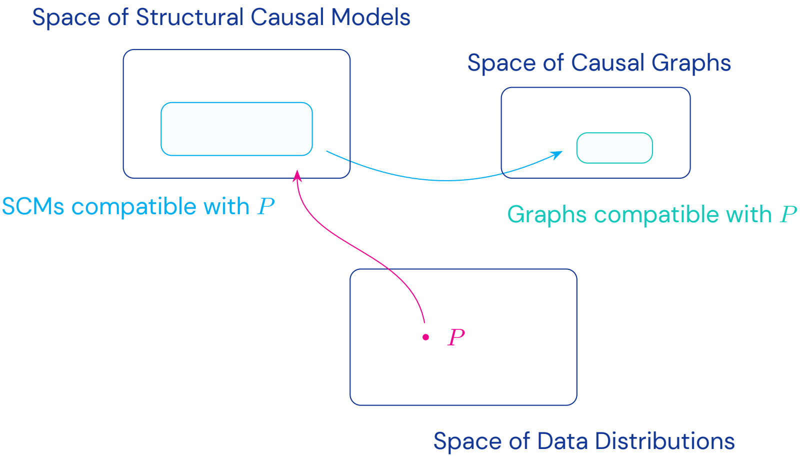
Works for (unconfounded) **additive noise models** of the form,

$$Y \leftarrow f(X) + U, \qquad X \perp\!\!\!\perp U$$

- f is non-linear or,
- U is non-Gaussian

Summary and Aspirations

Causal Discovery



End-to-end Causal Inference

Systematically deducing causal statements from an equivalence class and data.

(1) Query

e.g. $\mathbb{E}[Y \mid do(x)]$

(2) Data distribution → **Causal inference engine** → **Uniquely computable from P?**

If so, how?

e.g. $P(x, y, z)$

(3) Assumption

e.g. Equivalence class,
output of causal discovery algorithm

Bibliography I

E. Bareinboim, J. D. Correa, D. Ibeling, and T. Icard. On pearl's hierarchy and the foundations of causal inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, page 507–556. Association for Computing Machinery, NY, USA, 1st edition, 2022.

D. M. Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

D. M. Haughton. On the choice of a model to fit data from an exponential family. *The annals of statistics*, pages 342–355, 1988.

Bibliography II

M. Kalisch, M. Mächler, D. Colombo, M. H. Maathuis, and P. Bühlmann. Causal inference using graphical models with the r package pcalg. *Journal of statistical software*, 47: 1–26, 2012.

C. Meek. *Graphical Models: Selecting causal and statistical models*. PhD thesis, Carnegie Mellon University, 1997.

J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan kaufmann, 1988.

S. Shimizu, P. O. Hoyer, A. Hyvärinen, A. Kerminen, and M. Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

Bibliography III

P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman. *Causation, prediction, and search*. MIT press, 2000.

T. Verma and J. Pearl. Causal networks: Semantics and expressiveness. In *Machine intelligence and pattern recognition*, volume 9, pages 69–76. Elsevier, 1990.

S. Wright. The relative importance of heredity and environment in determining the piebald pattern of guinea-pigs. *Proceedings of the National Academy of Sciences*, 6(6):320–332, 1920.

X. Zheng, B. Aragam, P. K. Ravikumar, and E. P. Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

Bibliography IV

Y. Zheng, B. Huang, W. Chen, J. Ramsey, M. Gong, R. Cai, S. Shimizu, P. Spirtes, and K. Zhang. Causal-learn: Causal discovery in python. *arXiv preprint arXiv:2307.16405*, 2023.