Introduction to Causal Inference & Causal Discovery

Overview

Causal inference:

- Causal graphical models
- Interventions (the "do" operator)
- Example: Study of Kidney Stone Treatments
- Backdoor criterion
- The ladder of causation
- Counterfactuals

Causal discovery:

- Markov equivalence
- Faithfulness
- Structure identifiability
- Constraint-based methods
- Score-based methods

Causal Inference

Causal graphical models (CGM)

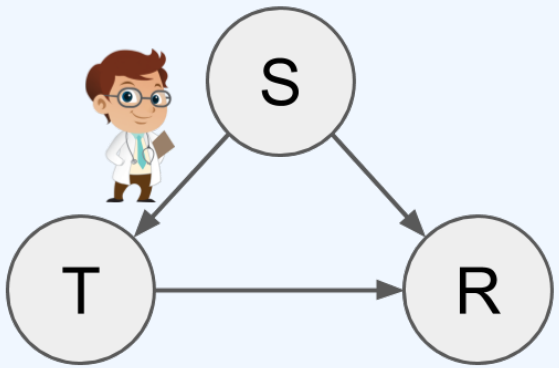
- A causal graphical model (CGM) is a pair $(p, \mathcal{G})$ s.t.
- $\mathcal{G}$ is a **directed acyclic graph** (DAG)
- $p \in \mathcal{L}(\mathcal{G})$, i.e. p factorizes according to $\mathcal{G}$.
- $\mathcal{G}$ describes **causal relationships** between variables, i.e., how the system reacts to **interventions**.

Causal graphical models (CGM)

- A causal graphical model (CGM) is a pair $(p, \mathcal{G})$ s.t.
- $\mathcal{G}$ is a **directed acyclic graph** (DAG)
- $p \in \mathcal{L}(\mathcal{G})$, i.e. p factorizes according to $\mathcal{G}$.
- $\mathcal{G}$ describes **causal relationships** between variables, i.e., how the system reacts to **interventions**.

Example: Kidney stone treatment

$T =$ Treatment $\in \{A, B\}$
$S =$ Stone size $\in \{\text{small}, \text{large}\}$
$R =$ Patient recovered $\in \{0, 1\}$



$$p(S, T, R) = p(S)p(T \mid S)p(R \mid S, T)$$

The "do" operator

Throughout, we will assume **perfect deterministic** interventions.

Definition (The “do” operator)

Given a causal graphical model $(p, \mathcal{G})$,

$$p(x \mid do(x'_k)) := \delta(x_k, x'_k) \prod_{i \neq k} p(x_i \mid x_{\pi_i^{\mathcal{G}}})$$

- Thus, $p(x \mid do(x'_k))$ is a "new" distribution over X_V.

The "do" operator

Throughout, we will assume **perfect deterministic** interventions.

Definition (The “do” operator)

Given a causal graphical model $(p, \mathcal{G})$,

$$p(x \mid do(x'_k)) := \delta(x_k, x'_k) \prod_{i \neq k} p(x_i \mid x_{\pi_i^{\mathcal{G}}})$$

- Thus, $p(x \mid do(x'_k))$ is a "new" distribution over X_V.
- Can compute marginals, e.g. $p(x_i | do(x'_k)) = \sum_{x_{V \setminus \{i\}}} p(x | do(x'_k))$

The "do" operator

Throughout, we will assume **perfect deterministic** interventions.

Definition (The “do” operator)

Given a causal graphical model $(p, \mathcal{G})$,

$$p(x \mid do(x'_k)) := \delta(x_k, x'_k) \prod_{i \neq k} p(x_i \mid x_{\pi_i^{\mathcal{G}}})$$

- Thus, $p(x \mid do(x'_k))$ is a "new" distribution over X_V.
- Can compute marginals, e.g. $p(x_i|do(x'_k)) = \sum_{x_{V \setminus \{i\}}} p(x|do(x'_k))$
- ... and conditionals, e.g. $p(x_i|x_j, do(x'_k)) = \frac{p(x_i, x_j|do(x'_k))}{p(x_j|do(x'_k))}$

The "do" operator

Throughout, we will assume **perfect deterministic** interventions.

Definition (The "do" operator)

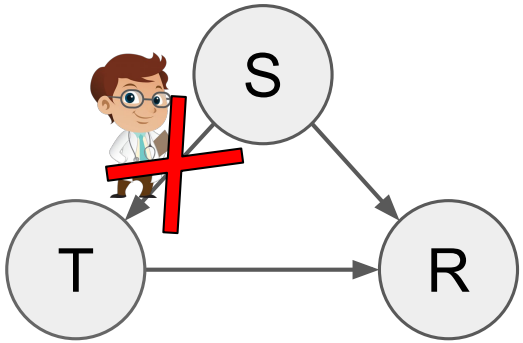
Given a causal graphical model $(p, \mathcal{G})$,

$$p(x \mid do(x'_k)) := \delta(x_k, x'_k) \prod_{i \neq k} p(x_i \mid x_{\pi_i^{\mathcal{G}}})$$

- Thus, $p(x \mid do(x'_k))$ is a "new" distribution over X_V.
- Can compute marginals, e.g. $p(x_i|do(x'_k)) = \sum_{x_{V\setminus\{i\}}} p(x|do(x'_k))$
- ... and conditionals, e.g. $p(x_i|x_j, do(x'_k)) = \frac{p(x_i,x_j|do(x'_k))}{p(x_j|do(x'_k))}$
- **Remark:** $p(x_{V\setminus\{k\}} \mid do(x_k)) = \prod_{i\neq k} p(x_i \mid x_{\pi_i^{\mathcal{G}}})$.

The "do" operator

- Back to our example

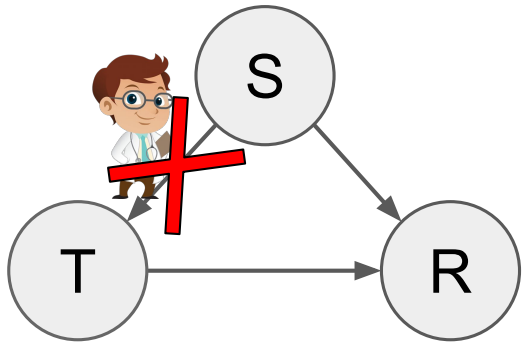


$$P(S,R \mid do(T)) = P(S)\underbrace{\cancel{P(T|S)}}P(R|S,T)$$

The decision of taking treatment T does not depend on S anymore

The "do" operator

- Back to our example



$$P(S,R \mid do(T)) = P(S)\underbrace{\cancel{P(T|S)}}\, P(R|S,T)$$

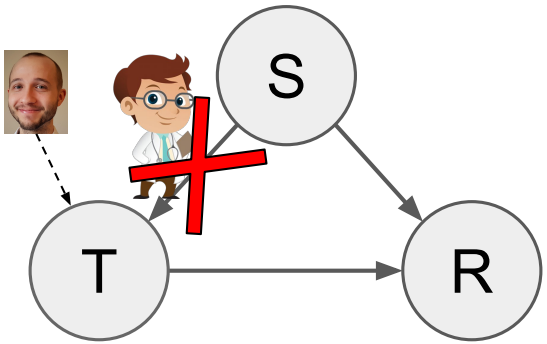
The decision of taking treatment T does not depend on S anymore

- Notice $p(\cdot \mid do(x'_k)) \in \mathcal{L}(\mathcal{G}')$, where $\mathcal{G}'$ is the **mutilated graph**, i.e.

$$\mathcal{G}' = (V, E') \quad E' = \{(i,j) \in E \mid j \neq k\}$$

The "do" operator

- Back to our example



$$P(S,R \mid do(T)) = P(S)\cancel{P(T|S)}P(R|S,T)$$

The decision of taking treatment T does not depend on S anymore

- Notice $p(\cdot \mid do(x'_k)) \in \mathcal{L}(\mathcal{G}')$, where $\mathcal{G}'$ is the **mutilated graph**, i.e.

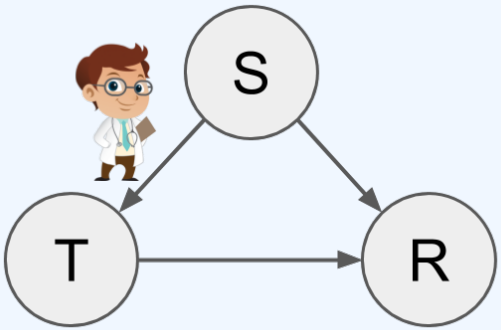
$$\mathcal{G}' = (V, E') \quad E' = \{(i,j) \in E \mid j \neq k\}$$

Different types of interventions

Intervening on the treatment T

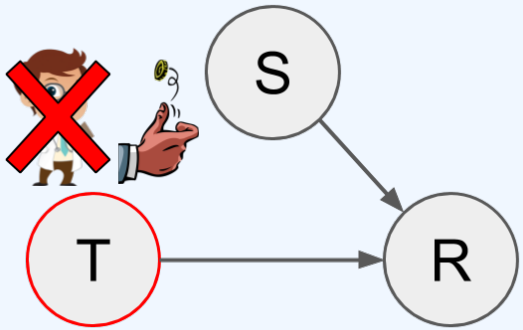
T = Treatment $\in \{A, B\}$
S = Stone size $\in \{\text{small}, \text{large}\}$
R = Patient recovered $\in \{0, 1\}$

Observations



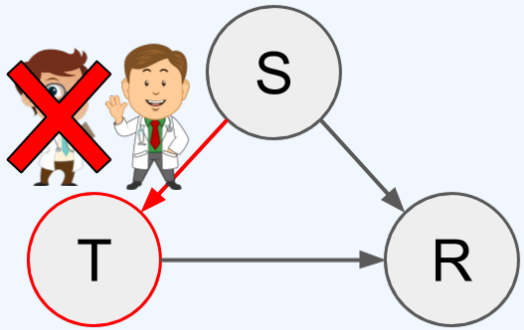
$p(S)p(T \mid S)p(R \mid S, T)$

Perfect intervention



$p(S)\tilde{p}(T)p(R \mid S, T)$

Imperfect intervention



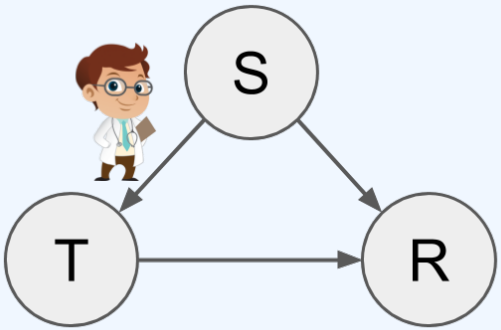
$p(S)\tilde{p}(T \mid S)p(R \mid S, T)$

Different types of interventions

Intervening on the treatment T

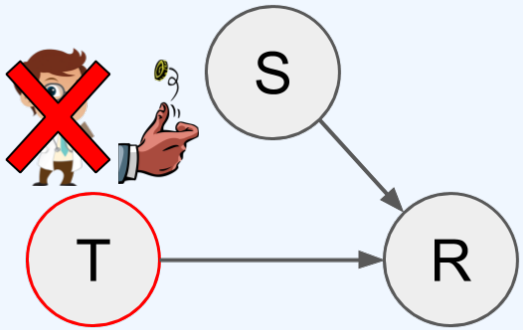
T = Treatment $\in \{A, B\}$
S = Stone size $\in \{\text{small}, \text{large}\}$
R = Patient recovered $\in \{0, 1\}$

Observations



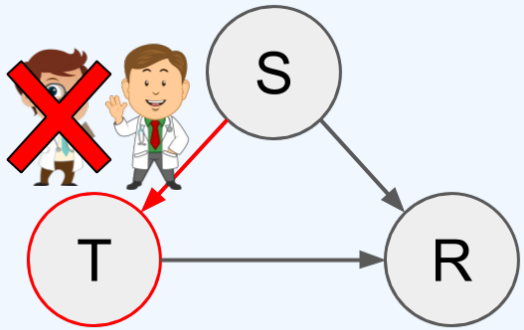
$$p(S)p(T \mid S)p(R \mid S, T)$$

Perfect intervention



$$p(S)\tilde{p}(T)p(R \mid S, T)$$

Imperfect intervention

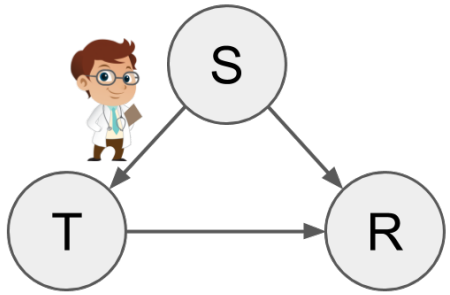


$$p(S)\tilde{p}(T \mid S)p(R \mid S, T)$$

Definition presented previously is a perfect intervention with $\tilde{p}(T) := \delta(T, T')$.
It is sometimes called a **perfect deterministic intervention**.

Why should I care!?! (Kidney Stone Treatment)

T = Treatment $\in \{A, B\}$
S = Stone size $\in \{\text{small}, \text{large}\}$
R = Patient recovered $\in \{0, 1\}$



$$p(S)p(T \mid S)p(R \mid S, T)$$

	Overall	Patients with small stones	Patients with large stones
Treatment a: Open surgery	78% (273/350)	**93%** (81/87)	**73%** (192/263)
Treatment b: Percutaneous nephrolithotomy	**83%** (289/350)	87% (234/270)	69% (55/80)

(Example taken from *Element of Causal Inference* by Peters et al. p111)

Why should I care!?! (Kidney Stone Treatment)

Pay attention to these two questions...

Why should I care!?! (Kidney Stone Treatment)

Pay attention to these two questions...

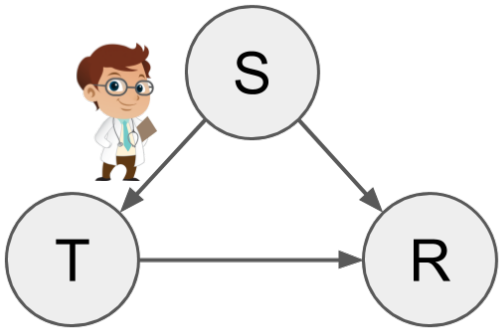
1- What is your chance of recovery knowing that the doctor gave you treatment A?

2- What is your chance of recovery if you decide to take treatment A?

(In both cases, assume you don't know the size of your stone)

Why should I care!?! (Kidney Stone Treatment)

T = Treatment $\in \{A, B\}$
Z = Stone size $\in \{\text{small}, \text{large}\}$
R = Patient recovered $\in \{0, 1\}$

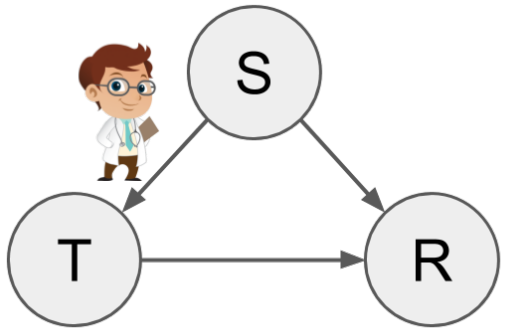


What is your chance of recovery knowing that the doctor gave you treatment A?

- Compute $P(R = 1 \mid T = A)$! (we know how to do that :D)
- Knowing that your doctor gave you treatment A tells you that you probably have a large kidney stone ... $P(S = \mathit{large} | T = A) = 0.75$
- ... which reduces your chance of recovery
  $P(R = 1 | T = A, S = \mathit{large}) = 0.73 < 0.93 = P(R = 1 | T = A, S = \mathit{small})$

Why should I care!?! (Kidney Stone Treatment)

T = Treatment $\in \{A, B\}$
Z = Stone size $\in \{\text{small}, \text{large}\}$
R = Patient recovered $\in \{0, 1\}$



What is your chance of recovery knowing that the doctor gave you treatment A?

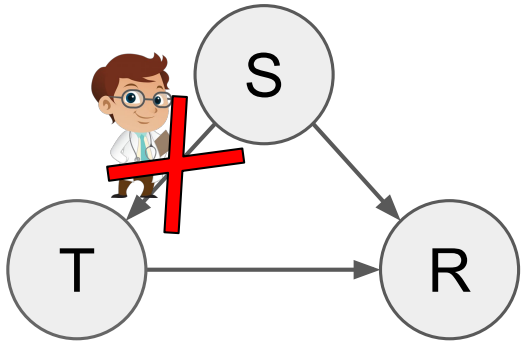
- Compute $P(R = 1 \mid T = A)$! (we know how to do that :D)
- Knowing that your doctor gave you treatment A tells you that you probably have a large kidney stone ... $P(S = large|T = A) = 0.75$
- ... which reduces your chance of recovery
  $P(R = 1|T = A, S = large) = 0.73 < 0.93 = P(R = 1|T = A, S = small)$

What is your chance of recovery if you decide to take treatment A?

- $P(R = 1 \mid do(T = A))$
- Your really don't know anything about your kidney stone

Why should I care!?! (Kidney Stone Treatment)

T = Treatment $\in \{A, B\}$
S = Stone size $\in \{\text{small}, \text{large}\}$
R = Patient recovered $\in \{0, 1\}$



$$P(S, R \mid do(T)) = P(S)\underbrace{\cancel{P(T|S)}}P(R|S, T)$$

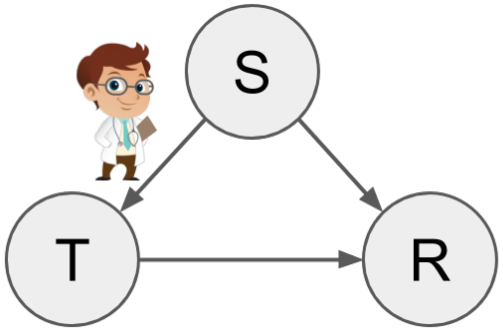
The decision of taking treatment T does not depend on S anymore

Then simply marginalize as usual:

$$P(R = 1|do(T = A)) = \sum_S P(R = 1, S|do(T = A))$$
$$= \sum_S P(R = 1|S, T = A)P(S) = 0,832$$

Why should I care!?! (Kidney Stone Treatment)

T = Treatment $\in \{A, B\}$
S = Stone size $\in \{\text{small}, \text{large}\}$
R = Patient recovered $\in \{0, 1\}$



What is your chance of recovery knowing that the doctor gave you treatment A?

$P(R = 1|T = A) = 0,78$ $P(R = 1|T = B) =$ **0,83**

What is your chance of recovery if you decide to take treatment A?

$P(R = 1|do(T = A)) =$ **0,832** $P(R = 1|do(T = B)) = 0,782$

Why should I care!?! (Kidney Stone Treatment)

- What just happened? We showed

$$\underbrace{P(R=1|do(T=A))}_{\text{Never observed data from } p(T,S,R \mid do(T=A))} = \underbrace{\sum_S P(R=1|S,T=A)P(S)}_{\text{...Yet I can estimate the query, since there is no "do" here :D}}$$

Why should I care!?! (Kidney Stone Treatment)

- What just happened? We showed

$$\underbrace{P(R=1|do(T=A))}_{\text{Never observed data from } p(T,S,R \mid do(T=A))} = \underbrace{\sum_S P(R=1|S,T=A)P(S)}_{\text{...Yet I can estimate the query, since there is no "do" here :D}}$$

- Formally, this means $p(R=1 \mid do(T=A))$ is **identifiable from** $p(R,T,S)$ and $\mathcal{G}$ (our computations *critically* relied on the causal graph).

Why should I care!?! (Kidney Stone Treatment)

- What just happened? We showed

$$\underbrace{P(R=1|do(T=A))}_{\text{Never observed data from } p(T,S,R \mid do(T=A))} = \underbrace{\sum_S P(R=1|S,T=A)P(S)}_{\text{...Yet I can estimate the query, since there is no "do" here :D}}$$

- Formally, this means $p(R=1 \mid do(T=A))$ is **identifiable from** $p(R,T,S)$ and $\mathcal{G}$ (our computations *critically* relied on the causal graph).

- Turns out what we just did is an instance of the **backdoor criterion**...

Backdoor criterion

Theorem (Backdoor criterion)

$p(x_i \mid do(x_k)) = \sum_{x_S} p(x_i \mid x_k, x_S) p(x_S)$ *if*

1. S *contains no descendants of* x_k*, and*
2. S *blocks all paths from* x_i *to* x_k *entering* x_k *from "the backdoor", i.e. such that* $x_k \leftarrow ... \, x_i$

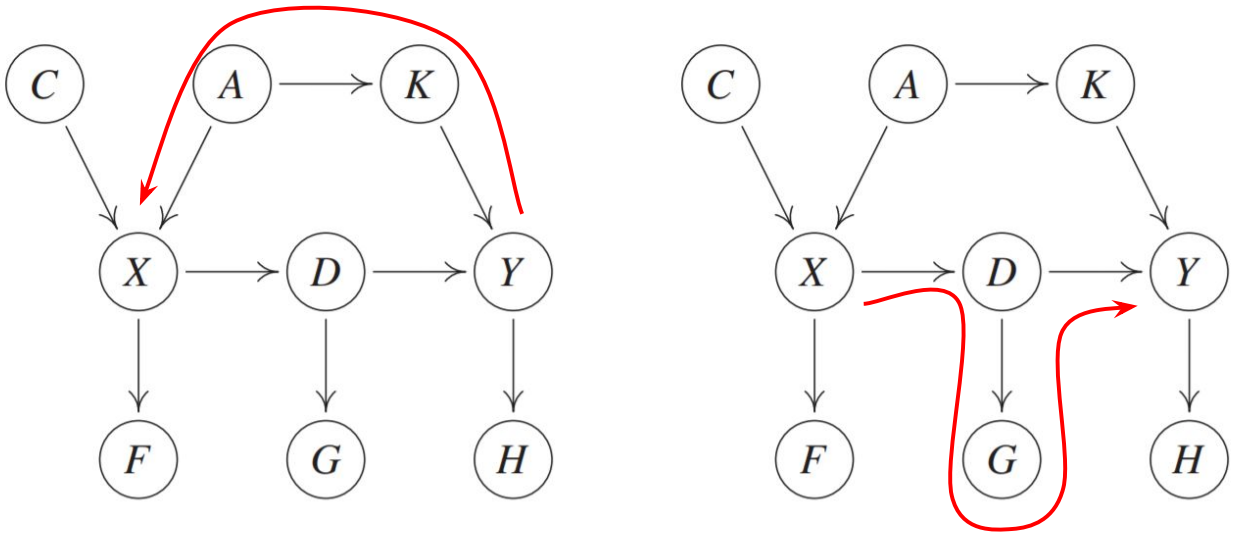
Backdoor criterion

Theorem (Backdoor criterion)

$p(x_i \mid do(x_k)) = \sum_{x_S} p(x_i \mid x_k, x_S) p(x_S)$ *if*

1. S *contains no descendants of* x_k*, and*
2. S *blocks all paths from* x_i *to* x_k *entering* x_k *from "the backdoor", i.e. such that* $x_k \leftarrow \ldots x_i$

Say we want to compute $p(y|do(x))$:



Left path: Only backdoor path. Blocked by $S = \{K\}$. **Right path:** Why we cannot include a descendant of X in S.

Backdoor criterion

Can all identifiable queries $p(x_i \mid do(x_k))$ be expressed with the backdoor criterion?

Backdoor criterion

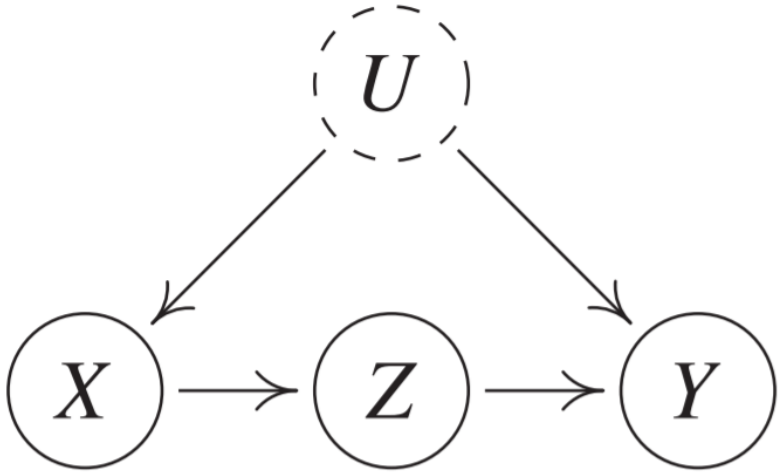
Can all identifiable queries $p(x_i \mid do(x_k))$ be expressed with the backdoor criterion?

Answer: No!

Backdoor criterion

Can all identifiable queries $p(x_i \mid do(x_k))$ be expressed with the backdoor criterion?

Answer: No!



- Since U is unobserved, we cannot apply the backdoor criterion...
- Turns out we can nevertheless identify $p(y|do(x))$ from $p(X, Z, Y)$ using the **front-door criterion**. Look it up!

Do-calculus

- Do-calculus is a set of **three rules** that can be applied to transform an interventional query (including a "do") into an observational expression (without any "do").
- Not enough time to present them...
- All identifiable queries can be found by a subsequent application of these rules, i.e. the rules are **complete**.

The ladder of causation

You now know about the first two steps of Pearl's "ladder of causation".

Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y\|x)$	Seeing	What is? How would seeing X change my belief in Y?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y\|do(x), z)$	Doing Intervening	What if? What if I do X?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x\|x', y')$	Imagining, Retrospection	Why? Was it X that caused Y? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past 2 years?

Fig. 1. The Causal Hierarchy. Questions at level i can only be answered if information from level i or higher is available.

Taken from "The Seven Tools of Causal Inference with Reflections on Machine Learning" by Judea Pearl

Counterfactual

You need **structural causal models (SCM)**. Let $\mathcal{G}$ be a DAG:

$$X_1 := f_1(X_{\pi_1^{\mathcal{G}}}) + N_1 \tag{1}$$

$$X_2 := f_2(X_{\pi_2^{\mathcal{G}}}) + N_2 \tag{2}$$

$$\dots \tag{3}$$

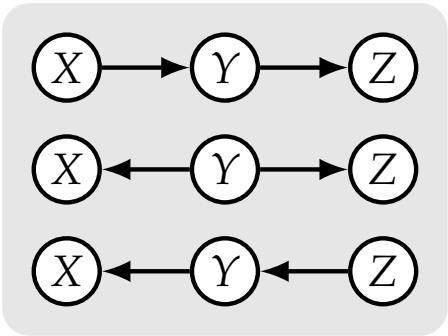
$$X_d := f_d(X_{\pi_d^{\mathcal{G}}}) + N_d \tag{4}$$

- This induces an **observational** distribution
- Can define **interventions** as well
- Can define **counterfactual** statements (not possible with a causal graphical model). See Section 6.4 in ECI.

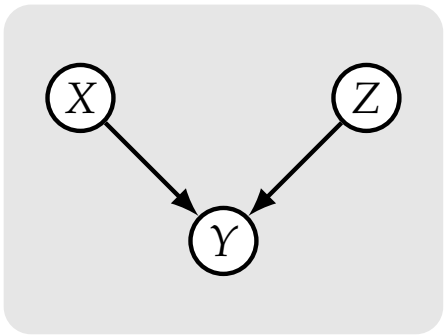
Causal Discovery

Markov Equivalence

- Recall: A Directed Graphical Model encodes the Conditional Independence of a distribution.
- Multiple DAGs may encode the same Conditional Independence statements.



$X \not\perp\!\!\!\perp Z$ and $X \perp\!\!\!\perp Z \mid Y$



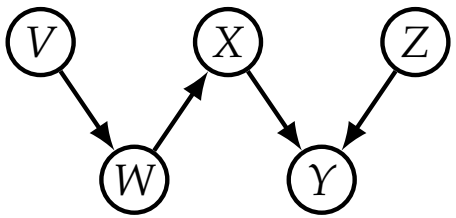
$X \perp\!\!\!\perp Z$ and $X \not\perp\!\!\!\perp Z \mid Y$

- Two DAGs encoding the same Conditional Independence statements are called **Markov Equivalent**.

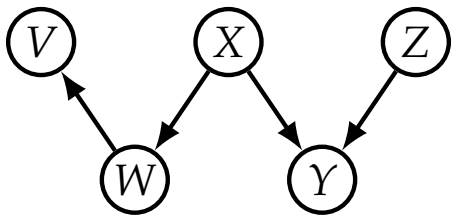
Markov Equivalence

Theorem (Verma & Pearl, 1991)

*Two DAGs G_1 and G_2 are **Markov Equivalent** if and only if they have the same skeleton and the same v-structures.*



G_1

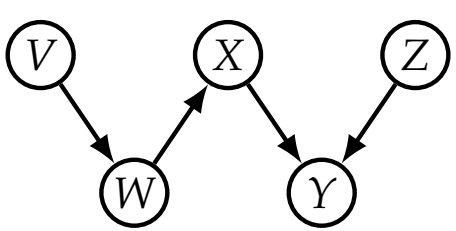


G_2

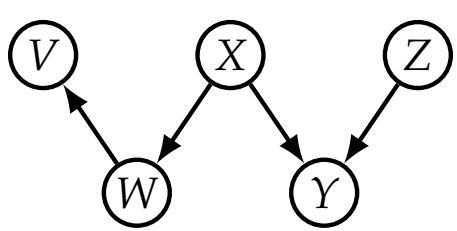
Markov Equivalence

Theorem (Verma & Pearl, 1991)

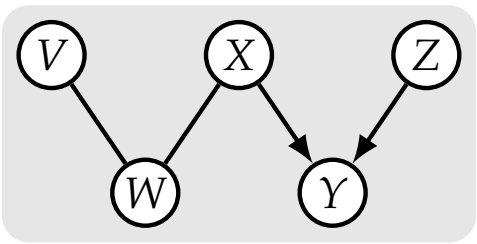
*Two DAGs G_1 and G_2 are **Markov Equivalent** if and only if they have the same skeleton and the same v-structures.*



G_1



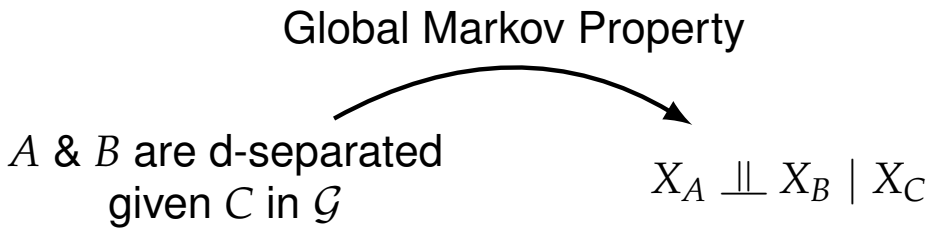
G_2



CPDAG

Markov Equivalence Classes can be represented as a **Completed Partially Directed Acyclic Graph** (CPDAG).

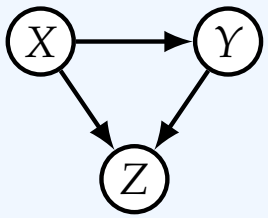
Faithfulness



Faithfulness

Faithfulness

Exercise: Violation of Faithfulness

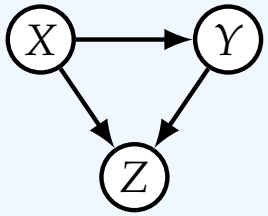


$$X := N_X$$
$$Y := X + N_Y$$
$$Z := X - Y + N_Z$$
with $N_X, N_Y, N_Z \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$

$p(X, Y, Z)$ is a Multivariate Normal distribution, where the only conditional independence statements are: $X \perp\!\!\!\perp Z$ and $X \not\perp\!\!\!\perp Z \mid Y$.

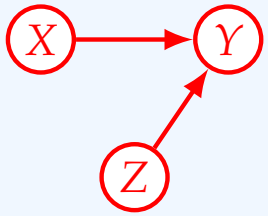
Faithfulness

Exercise: Violation of Faithfulness



$$X := N_X$$
$$Y := X + N_Y$$
$$Z := X - Y + N_Z$$
with $N_X, N_Y, N_Z \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$

Structure Learning $\Rightarrow$



$p(X, Y, Z)$ is a Multivariate Normal distribution, where the only conditional independence statements are: $X \perp\!\!\!\perp Z$ and $X \not\perp\!\!\!\perp Z \mid Y$.

Structure Identifiability

Theorem

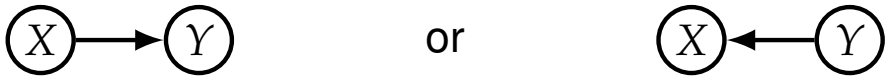
Assume that p is faithful wrt. $\mathcal{G}^0$. The Markov Equivalence class of $\mathcal{G}^0$, represented by its CPDAG, is identifiable from p.

Structure Identifiability

Theorem

Assume that p is faithful wrt. $\mathcal{G}^0$. The Markov Equivalence class of $\mathcal{G}^0$, represented by its CPDAG, is identifiable from p.

- Only the Markov Equivalence class is identifiable from observations, **not an individual graph**. Two Markov Equivalent graphs may lead to different causal conclusions!



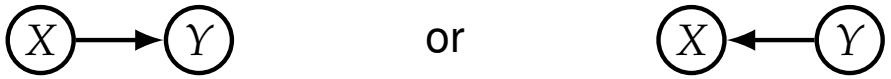
- Under different assumptions, an individual DAG may be identifiable

Structure Identifiability

Theorem

Assume that p is faithful wrt. $\mathcal{G}^0$. The Markov Equivalence class of $\mathcal{G}^0$, represented by its CPDAG, is identifiable from p.

- Only the Markov Equivalence class is identifiable from observations, **not an individual graph**. Two Markov Equivalent graphs may lead to different causal conclusions!



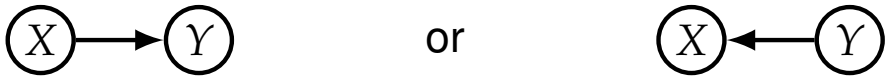
- Under different assumptions, an individual DAG may be identifiable
  - Additive Noise Model (ANM): $X_j := f_j(X_{\mathrm{Pa}_j}) + N_j$, $N_j \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, where f_j are nonlinear.

Structure Identifiability

Theorem

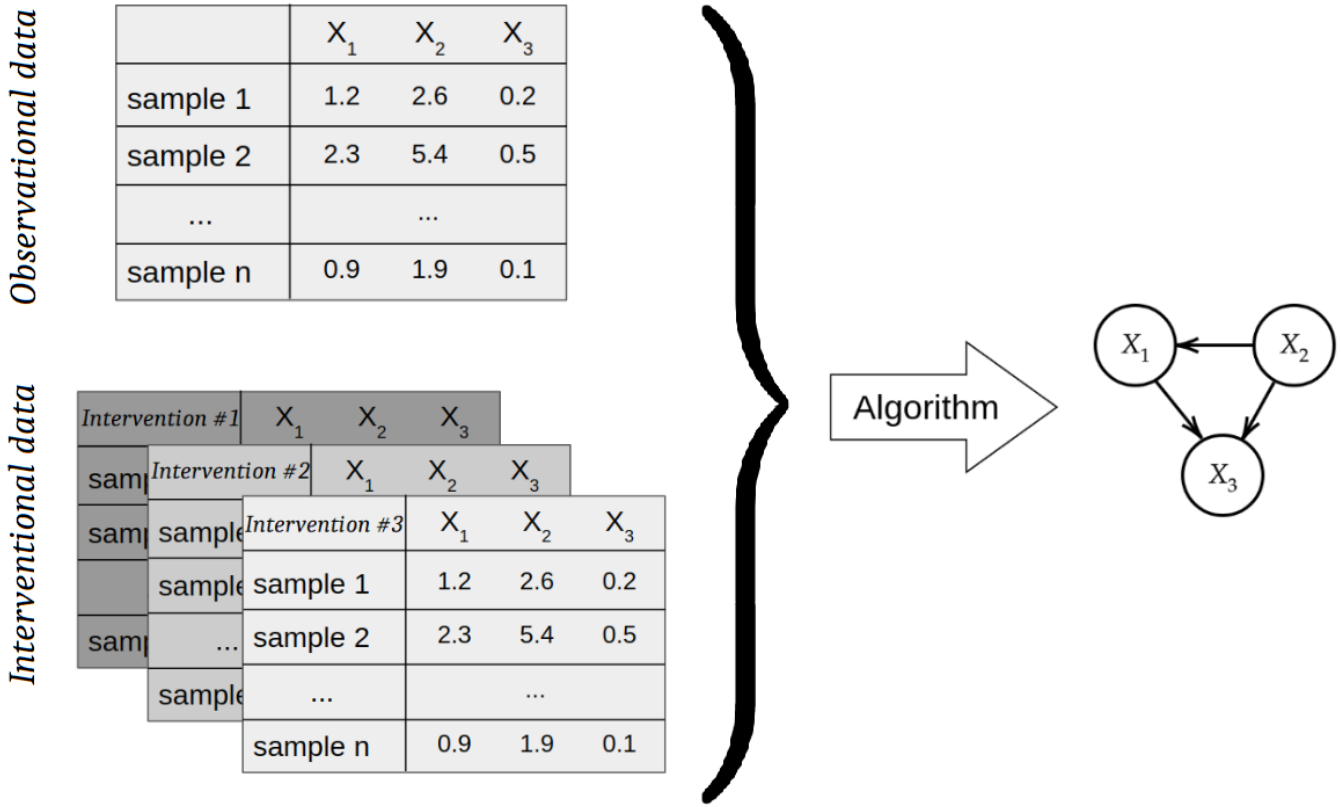
Assume that p is faithful wrt. $\mathcal{G}^0$. The Markov Equivalence class of $\mathcal{G}^0$, represented by its CPDAG, is identifiable from p.

- Only the Markov Equivalence class is identifiable from observations, **not an individual graph**. Two Markov Equivalent graphs may lead to different causal conclusions!



- Under different assumptions, an individual DAG may be identifiable
  - Additive Noise Model (ANM): $X_j := f_j(X_{\mathrm{Pa}_j}) + N_j$, $N_j \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, where f_j are nonlinear.
  - Using **interventional data** (i.e. data resulting from controlled experiments).

Causal Structure Learning (Causal Discovery)

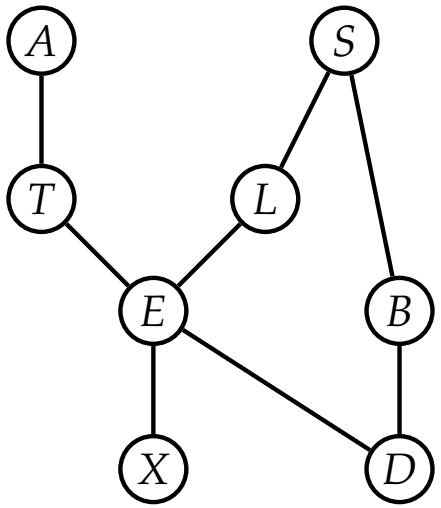


How to recover the (CP)DAG using a dataset $\mathcal{D}$?

Constraint-based methods

Step 1: Identify the skeleton

For each pair of nodes X & Y, and $\mathbf{A} \subseteq \mathbf{V} \backslash \{X, Y\}$, test if $X \perp\!\!\!\perp_{\mathcal{D}} Y \mid \mathbf{A}$.
If there is no set $\mathbf{A}$ s.t. $X \perp\!\!\!\perp_{\mathcal{D}} Y \mid \mathbf{A}$, then add an edge X — Y.



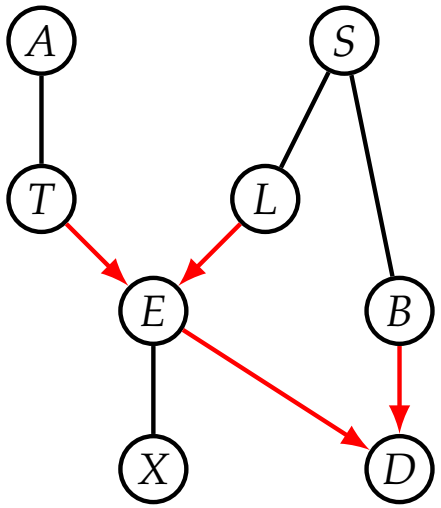
Constraint-based methods

Step 1: Identify the skeleton

For each pair of nodes X & Y, and $\mathbf{A} \subseteq \mathbf{V} \backslash \{X, Y\}$, test if $X \perp\!\!\!\perp_{\mathcal{D}} Y \mid \mathbf{A}$. If there is no set $\mathbf{A}$ s.t. $X \perp\!\!\!\perp_{\mathcal{D}} Y \mid \mathbf{A}$, then add an edge $X - Y$.

Step 2: Identify the v-structures

For each structure $X - Z - Y$ with no edge between X & Y, orient $X \rightarrow Z \leftarrow Y$ iff $Z \notin \mathbf{A}$, where $\mathbf{A}$ is such that $X \perp\!\!\!\perp_{\mathcal{D}} Y \mid \mathbf{A}$.



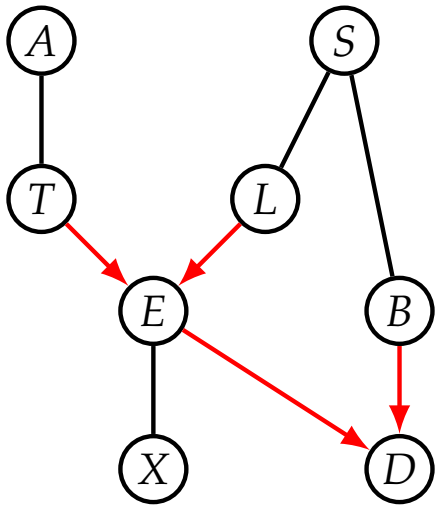
Constraint-based methods

Step 1: Identify the skeleton

For each pair of nodes X & Y, and $\mathbf{A} \subseteq \mathbf{V} \backslash \{X, Y\}$, test if $X \perp\!\!\!\perp_{\mathcal{D}} Y \mid \mathbf{A}$. If there is no set $\mathbf{A}$ s.t. $X \perp\!\!\!\perp_{\mathcal{D}} Y \mid \mathbf{A}$, then add an edge $X - Y$.

Step 2: Identify the v-structures

For each structure $X - Z - Y$ with no edge between X & Y, orient $X \rightarrow Z \leftarrow Y$ iff $Z \notin \mathbf{A}$, where $\mathbf{A}$ is such that $X \perp\!\!\!\perp_{\mathcal{D}} Y \mid \mathbf{A}$.

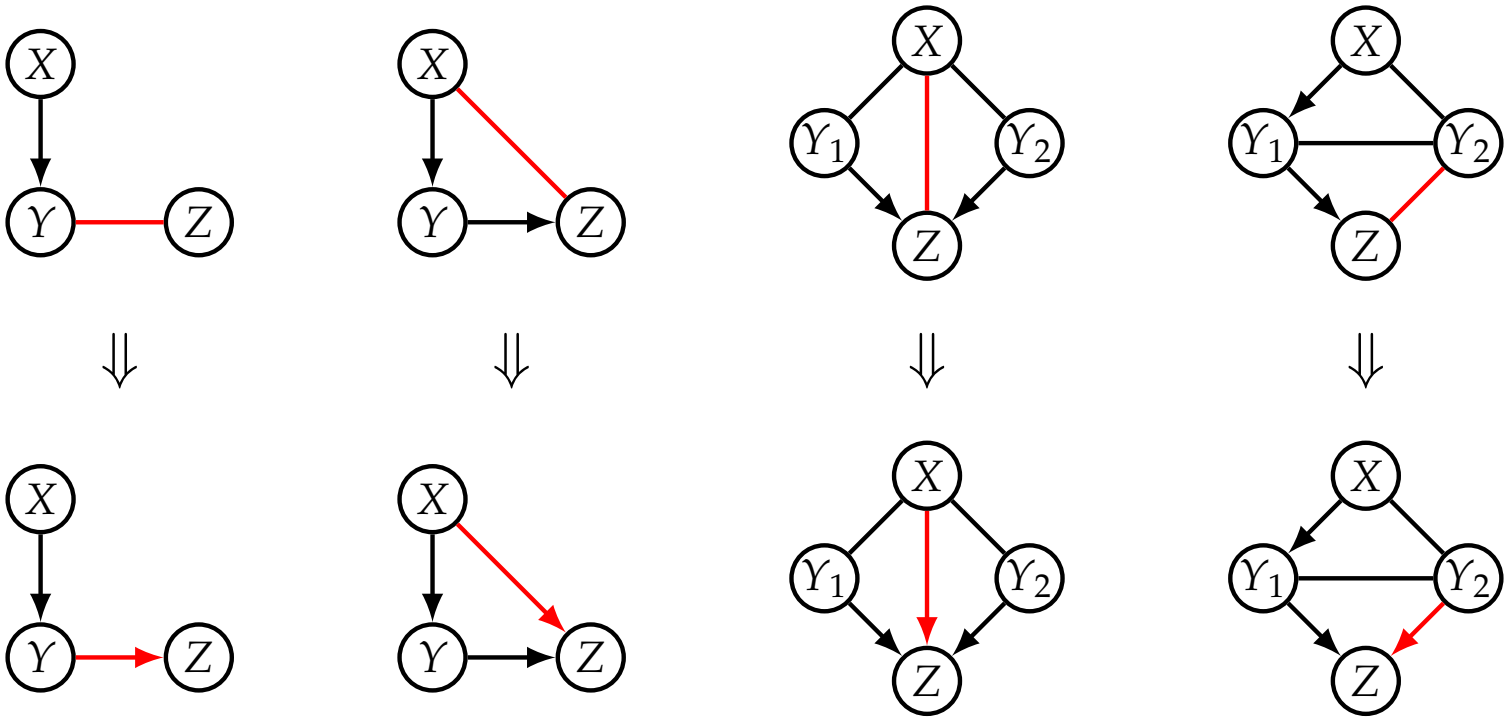


IC Algorithm

Constraint-based methods

Step 2': Additional orientations

Use **Meek's orientation rules** to orient some of the remaining edges.



Rule 1 Rule 2 Rule 3 Rule 4

Score-based methods

- Idea: treat the problem of learning the structure of the DAG as a **model selection problem**

$$\max_{\mathcal{G} \in \text{DAG}} \text{score}(\mathcal{G} \mid \mathcal{D})$$

- Recall: choices of scores

Score-based methods

- Idea: treat the problem of learning the structure of the DAG as a **model selection problem**

$$\max_{\mathcal{G} \in \text{DAG}} \text{score}(\mathcal{G} \mid \mathcal{D})$$

- Recall: choices of scores
  - **Likelihood score**:

$$\text{score}_L(\mathcal{G} \mid \mathcal{D}) = \log p(\mathcal{D} \mid \hat{\theta}_{\mathcal{G}}^{\text{MLE}}, \mathcal{G})$$

Score-based methods

- Idea: treat the problem of learning the structure of the DAG as a **model selection problem**

$$\max_{\mathcal{G} \in \text{DAG}} \text{score}(\mathcal{G} \mid \mathcal{D})$$

- Recall: choices of scores
  - **Likelihood score**:

$$\text{score}_L(\mathcal{G} \mid \mathcal{D}) = \log p(\mathcal{D} \mid \hat{\theta}_{\mathcal{G}}^{\text{MLE}}, \mathcal{G})$$

  - **Bayesian score**:

$$\text{score}_B(\mathcal{G} \mid \mathcal{D}) = \log p(\mathcal{D} \mid \mathcal{G}) + \log p(\mathcal{G})$$

Score-based methods

- Idea: treat the problem of learning the structure of the DAG as a **model selection problem**

$$\max_{\mathcal{G} \in \text{DAG}} \text{score}(\mathcal{G} \mid \mathcal{D})$$

- Recall: choices of scores
  - **Likelihood score**:

$$\text{score}_L(\mathcal{G} \mid \mathcal{D}) = \log p(\mathcal{D} \mid \hat{\theta}_{\mathcal{G}}^{\text{MLE}}, \mathcal{G})$$

  - **Bayesian score**:

$$\text{score}_B(\mathcal{G} \mid \mathcal{D}) = \log p(\mathcal{D} \mid \mathcal{G}) + \log p(\mathcal{G})$$

  - **Bayesian Information Criterion (BIC)**:

$$\text{score}_{BIC}(\mathcal{G} \mid \mathcal{D}) = \log p(\mathcal{D} \mid \hat{\theta}_{\mathcal{G}}^{\text{MLE}}, \mathcal{G}) - \frac{\log N}{2} \text{Dim}[\mathcal{G}]$$

Score-based methods

$$\max_{\mathcal{G} \in \text{DAG}} \text{score}(\mathcal{G} \mid \mathcal{D})$$

- How to search over the space of DAGs?
- The number of DAGs over n nodes is **super-exponential** in n: $2^{\Theta(n^2)}$.

Score-based methods

$$\max_{\mathcal{G} \in \text{DAG}} \text{score}(\mathcal{G} \mid \mathcal{D})$$

- How to search over the space of DAGs?
- The number of DAGs over n nodes is **super-exponential** in n: $2^{\Theta(n^2)}$.

Theorem

*Let $G_{\leq d} = \{\mathcal{G}\ a\ DAG \mid every\ node\ has\ at\ most\ d\ parents\}$. Finding a DAG in $G_{\leq d}$ that maximizes a score is **NP-hard** for $d \geq 2$.*

Score-based methods

$$\max_{\mathcal{G} \in \text{DAG}} \text{score}(\mathcal{G} \mid \mathcal{D})$$

- How to search over the space of DAGs?
- The number of DAGs over n nodes is **super-exponential** in n: $2^{\Theta(n^2)}$.

Theorem

*Let $G_{\leq d} = \{\mathcal{G}\ a\ DAG \mid every\ node\ has\ at\ most\ d\ parents\}$. Finding a DAG in $G_{\leq d}$ that maximizes a score is **NP-hard** for $d \geq 2$.*

- Heuristic solutions:
  - **Greedy algorithms**: Hill climbing, GES
  - **Genetic algorithms**
  - **Constrained continuous optimization**: NOTEARS, Gran-DAG, DCDI, etc...