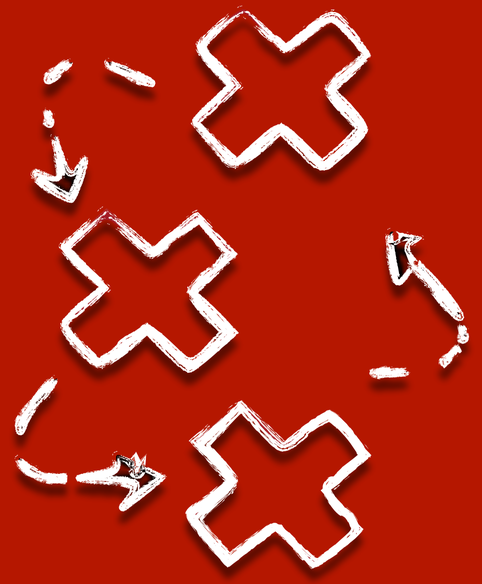


Causal Data Science

Lecture 8.2: Introduction to causal discovery

Lecturer: Sara Magliacane

UvA - Spring 2024

Overview on where we are in the course

1	Introduction
2	Probability recap
3	Causal graphs
4	Interventions
5	Covariate adjustment
6	Frontdoor criterion, Instrumental variables
7	Counterfactuals and potential outcomes
8	Estimating causal effects, Missing data
9	Constraint based structure learning
10	Score based structure learning
11	Advanced structure learning and transportability
12	Causality-inspired ML

Background on causal graphs

We know the causal graph, how do we estimate causal effects?

What happens if the graph is unknown?

Common assumptions

- If P is **Markov** and **faithful** to G, then for any disjoint $\mathbf{A}, \mathbf{B}, \mathbf{C} \subseteq \mathbf{V}$:

$$\mathbf{A} \perp_G \mathbf{B} \,|\, \mathbf{C} \iff X_{\mathbf{A}} \perp\!\!\!\perp_P X_{\mathbf{B}} \,|\, X_{\mathbf{C}}$$

Common assumptions

- If P is **Markov** and **faithful** to G, then for any disjoint $\mathbf{A}, \mathbf{B}, \mathbf{C} \subseteq \mathbf{V}$:

$$\mathbf{A} \perp_G \mathbf{B} \mid \mathbf{C} \iff X_{\mathbf{A}} \perp\!\!\!\perp_P X_{\mathbf{B}} \mid X_{\mathbf{C}}$$

- **Causal sufficiency** - no **latent** confounders (common causes), no selection bias

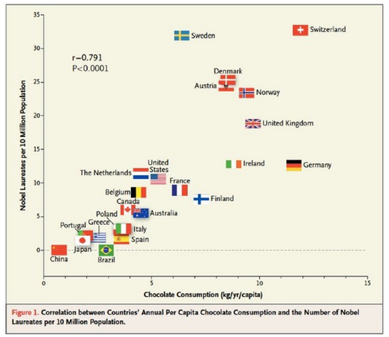
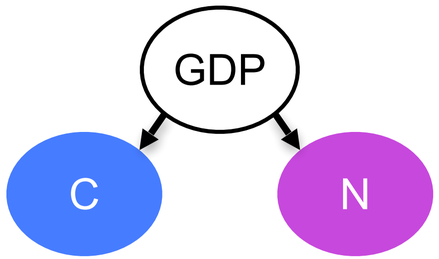


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.



Common assumptions

- If P is **Markov** and **faithful** to G, then for any disjoint $\mathbf{A}, \mathbf{B}, \mathbf{C} \subseteq \mathbf{V}$:

$$\mathbf{A} \perp_G \mathbf{B} \mid \mathbf{C} \iff X_{\mathbf{A}} \perp\!\!\!\perp_P X_{\mathbf{B}} \mid X_{\mathbf{C}}$$

- **Causal sufficiency** - no latent confounders (common causes), no selection bias
- **Acyclicity** - the underlying graph is acyclic
- Cycles + causal insufficiency: sigma separation, Joint Causal Inference

Recap: Factorizing joint distributions

- Given any **ordering** of the variables $(X_1, \ldots, X_p)$ we can write:

$$P(X_1, \ldots, X_p) = P(X_1)P(X_2 \mid X_1)\ldots P(X_p \mid X_1, \ldots, X_{p-1})$$

- For example $P(X, Y, Z)$ can be equivalently factorized as:
  - $P(X, Y, Z) = P(X)P(Y \mid X)P(Z \mid X, Y)$
  - $P(X, Z, Y) = P(X)P(Z \mid X)P(Y \mid X, Z)$
  - $P(Z, Y, X) = P(Z)P(Y \mid Z)P(X \mid Y, Z)$ **...**

Recap: Exploiting conditional independences

- We can **simplify** the factorisation by using **conditional independences:**

$$X_i \perp\!\!\!\perp X_j | X_{\mathbf{Z}} \implies P(X_i | X_j, X_{\mathbf{Z}}) = P(X_i | X_{\mathbf{Z}})$$

- For example: $X \perp\!\!\!\perp Y | Z$:
  - $P(X, Y, Z) = P(X)P(Y|X)P(Z|X, Y)$
  - $P(X, Z, Y) = P(X)P(Z|X)P(Y|X, Z)$
  - $P(Z, Y, X) = P(Z)P(Y|Z)P(X|Y, Z)$ ...

Recap: Exploiting conditional independences

- We can **simplify** the factorisation by using **conditional independences:**

$$X_i \perp\!\!\!\perp X_j | X_{\mathbf{Z}} \implies P(X_i | X_j, X_{\mathbf{Z}}) = P(X_i | X_{\mathbf{Z}})$$

- For example: $X \perp\!\!\!\perp Y | Z$:
  - $P(X, Y, Z) = P(X)P(Y|X)P(Z|X, Y)$
  - $P(X, Z, Y) = P(X)P(Z|X)P(Y|X, Z) = P(X)P(Z|X)P(Y|Z)$
  - $P(Z, Y, X) = P(Z)P(Y|Z)P(X|Y, Z) = P(Z)P(Y|Z)P(X|Z)$

Recap: Bayesian networks

- We have a set of random variables $X_1, \ldots, X_p$ with joint $P(X_1, \ldots, X_p)$
- We have a DAG G, s.t. **each random variable X_i** is represented by **node i**

- We then say $P(X_1, \ldots, X_p)$ **factorizes over G** if

$$P(X_1, \ldots, X_p) = \prod_{i \in V} P(X_i \mid \mathbf{X}_{\mathrm{Pa}_G(i)})$$

- A **Bayesian network** (BN) is the tuple (G, P) s.t. P **factorizes over** G

Global Markov Property

- If (G, P) is a Bayesian network with a DAG $G = (\mathbf{V}, \mathbf{E})$, i.e. P **factorizes according to** G, then for any disjoint $\mathbf{A}, \mathbf{B}, \mathbf{C} \subseteq \mathbf{V}$:

$$\mathbf{A} \perp_G \mathbf{B} \,|\, \mathbf{C} \implies X_{\mathbf{A}} \perp\!\!\!\perp_P X_{\mathbf{B}} \,|\, X_{\mathbf{C}}$$

If P has a density (e.g. no deterministic relations) [Lauritzen 1996]

- **d-separations** that can be read purely from a graph imply **conditional independences** in the random variables and data generated by the graph

- In general some of the independences in the data are not represented in the graph, i.e. we have more edges that create “extra d-connections”

Terminology - I-maps

- (G, P) is a Bayesian network with a DAG $G = (\mathbf{V}, \mathbf{E})$. We say that G **is an I-map of P** (independence map), for any disjoint $\mathbf{A}, \mathbf{B}, \mathbf{C} \subseteq \mathbf{V}$:
  - $\mathbf{A} \perp_G \mathbf{B} \mid \mathbf{C} \implies X_{\mathbf{A}} \perp\!\!\!\perp_P X_{\mathbf{B}} \mid X_{\mathbf{C}}$ *(Global Markov Property)*

Terminology - I-maps

- (G, P) is a Bayesian network with a DAG $G = (\mathbf{V}, \mathbf{E})$. We say that G **is an I-map of P** (independence map), for any disjoint $\mathbf{A}, \mathbf{B}, \mathbf{C} \subseteq \mathbf{V}$:
  - $\mathbf{A} \perp_G \mathbf{B} \mid \mathbf{C} \implies X_{\mathbf{A}} \perp\!\!\!\perp_P X_{\mathbf{B}} \mid X_{\mathbf{C}}$ *(Global Markov Property)*

Many distributions factorize according to many graphs, e.g. any P factorizes according to all fully connected graphs

We want to narrow down the candidate graphs to a smaller set containing the true causal graph

What about sparsity/minimality?

Terminology - I-maps

- (G, P) is a Bayesian network with a DAG $G = (\mathbf{V}, \mathbf{E})$. We say that G **is an I-map of P** (independence map), for any disjoint $\mathbf{A}, \mathbf{B}, \mathbf{C} \subseteq \mathbf{V}$:
  - $\mathbf{A} \perp_G \mathbf{B} \mid \mathbf{C} \implies X_{\mathbf{A}} \perp\!\!\!\perp_P X_{\mathbf{B}} \mid X_{\mathbf{C}}$ *(Global Markov Property)*
- **Causal minimality/Minimal I-map:** an I-map in which we cannot remove edges, otherwise it stops being an I-map

Terminology - I-maps

- (G, P) is a Bayesian network with a DAG $G = (\mathbf{V}, \mathbf{E})$. We say that G **is an I-map of P** (independence map), for any disjoint $\mathbf{A}, \mathbf{B}, \mathbf{C} \subseteq \mathbf{V}$:
  - $\mathbf{A} \perp_G \mathbf{B} \,|\, \mathbf{C} \implies X_{\mathbf{A}} \perp\!\!\!\perp_P X_{\mathbf{B}} \,|\, X_{\mathbf{C}}$ *(Global Markov Property)*
- **Causal minimality/Minimal I-map:** an I-map in which we cannot remove edges, otherwise it stops being an I-map
  - $G = (\mathbf{V}, \mathbf{E})$ is an I-map of P
  - $G' = (\mathbf{V}, \mathbf{E}')$ such that $\mathbf{E}' \subset \mathbf{E}$ is **not** an I-map of P

$$\exists \mathbf{A}, \mathbf{B}, \mathbf{C} \text{ s.t. } \mathbf{A} \perp_{G'} \mathbf{B} \,|\, \mathbf{C} \quad \text{and} \quad X_{\mathbf{A}} \not\perp\!\!\!\perp_P X_{\mathbf{B}} \,|\, X_{\mathbf{C}}$$

Constructing a Minimal I-Map

1. Choose any **ordering** of the variables $(X_1, \ldots, X_p)$
2. Write the factorisation via chain rule, **simplify** factorisation as much as possible with **conditional independences:**

$$X_i \perp\!\!\!\perp X_j \,|\, X_{\mathbf{Z}} \implies P(X_i \,|\, X_j, X_{\mathbf{Z}}) = P(X_i \,|\, X_{\mathbf{Z}})$$

Constructing a Minimal I-Map

1. Choose any **ordering** of the variables $(X_1, \ldots, X_p)$
2. Write the factorisation via chain rule, **simplify** factorisation as much as possible with **conditional independences:**

   $$X_i \perp\!\!\!\perp X_j \,|\, X_{\mathbf{Z}} \implies P(X_i \,|\, X_j, X_{\mathbf{Z}}) = P(X_i \,|\, X_{\mathbf{Z}})$$

3. **Create a BN** in which each X_i is caused by the variables in $P(X_i \,|\, X_{\mathbf{Pa(i)}})$

Constructing a Minimal I-Map

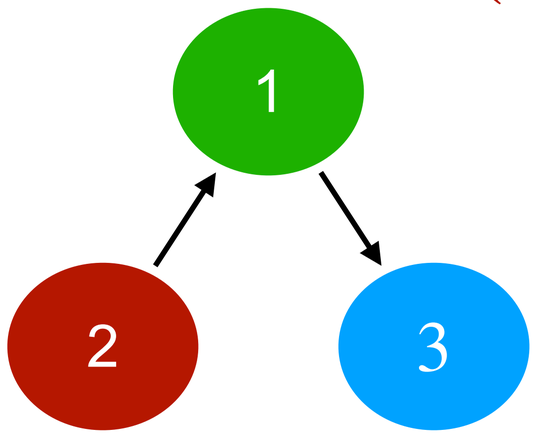
1. Choose any **ordering** of the variables $(X_1, \ldots, X_p)$
2. Write the factorisation via chain rule, **simplify** factorisation as much as possible with **conditional independences:**

   $$X_i \perp\!\!\!\perp X_j \,|\, X_{\mathbf{Z}} \implies P(X_i \,|\, X_j, X_{\mathbf{Z}}) = P(X_i \,|\, X_{\mathbf{Z}})$$

3. Create a BN in which each X_i is caused by the variables in $P(X_i \,|\, X_{\mathbf{Pa(i)}})$

If $X_3 \perp\!\!\!\perp X_2 \,|\, X_1$

$$P(X_2, X_1, X_3) = P(X_2)P(X_1 \,|\, X_2)P(X_3 \,|\, X_1, \cancel{X_2})$$



Constructing a Minimal I-Map (not unique!)

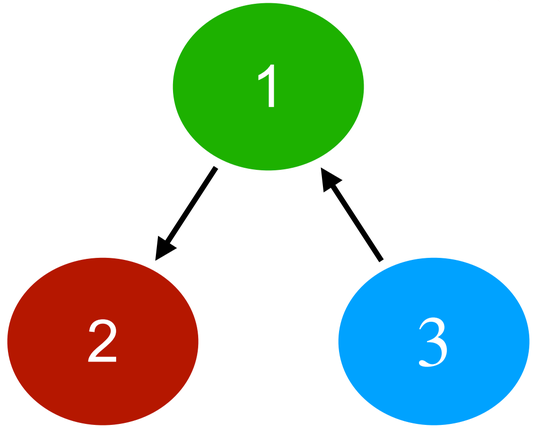
1. Choose any **ordering** of the variables $(X_1, \ldots, X_p)$
2. Write the factorisation via chain rule, **simplify** factorisation as much as possible with **conditional independences:**

   $$X_i \perp\!\!\!\perp X_j \,|\, X_{\mathbf{Z}} \implies P(X_i \,|\, X_j, X_{\mathbf{Z}}) = P(X_i \,|\, X_{\mathbf{Z}})$$

3. Create a BN in which each X_i is caused by the variables in $P(X_i \,|\, X_{\mathbf{Pa(i)}})$

If $X_3 \perp\!\!\!\perp X_2 \,|\, X_1$

$$P(X_3, X_1, X_2) = P(X_3)P(X_1 \,|\, X_3)P(X_2 \,|\, X_1, \cancel{X_3})$$



Minimal I-Map example

- Modified classic example: Discipline (D), Intelligence (I), SAT score (S), Grade (G)

$p(D, I, S, G)$

$D \perp\!\!\!\perp I$

$D \perp\!\!\!\perp G \quad D \perp\!\!\!\perp I \mid G \quad D \perp\!\!\!\perp G \mid I, S$

$G \perp\!\!\!\perp S \mid I \quad D \perp\!\!\!\perp G \mid I \quad S \perp\!\!\!\perp G \mid I, D$

Minimal I-Map example

- Modified classic example: Discipline (D), Intelligence (I), SAT score (S), Grade (G)

$p(D, I, S, G)$ $D \perp\!\!\!\perp I$ $p(D) \cdot p(I|D) \cdot p(S|I, D) \cdot p(G|S, I, D)$

$D \perp\!\!\!\perp G$ $D \perp\!\!\!\perp I|G$ $D \perp\!\!\!\perp G|I, S$

$G \perp\!\!\!\perp S|I$ $D \perp\!\!\!\perp G|I$ $S \perp\!\!\!\perp G|I, D$

Minimal I-Map example

- Modified classic example: Discipline (D), Intelligence (I), SAT score (S), Grade (G)

$p(D, I, S, G)$ $D \perp\!\!\!\perp I$ $p(D) \cdot p(I|D) \cdot p(S|I,D) \cdot p(G|S,I,D)$

$D \perp\!\!\!\perp G$ $D \perp\!\!\!\perp I|G$ $D \perp\!\!\!\perp G|I,S$

$G \perp\!\!\!\perp S|I$ $D \perp\!\!\!\perp G|I$ $S \perp\!\!\!\perp G|I,D$

$p(D) \cdot p(I) p(S|I,D) p(G|I)$

Minimal I-Map example

- Modified classic example: Discipline (D), Intelligence (I), SAT score (S), Grade (G)

$p(D, I, S, G)$

$D \perp\!\!\!\perp I$ $\quad p(D) \cdot p(I|D) \cdot p(S|I,D) \cdot p(G|S,I,D)$

$D \perp\!\!\!\perp G$ $\quad D \perp\!\!\!\perp I|G \quad D \perp\!\!\!\perp G|I,S$

$G \perp\!\!\!\perp S|I$ $\quad D \perp\!\!\!\perp G|I \quad S \perp\!\!\!\perp G|I,D$

$p(D) \cdot p(I) \, p(S|I,D) \, p(G|I)$

Minimal I-Map example

- Modified classic example: Discipline (D), Intelligence (I), SAT score (S), Grade (G)

$P(D, I, S, G)$

$D \perp\!\!\!\perp I$

$D \perp\!\!\!\perp G \quad D \perp\!\!\!\perp I \mid G \quad D \perp\!\!\!\perp G \mid I, S$

$G \perp\!\!\!\perp S \mid I \quad D \perp\!\!\!\perp G \mid I \quad S \perp\!\!\!\perp G \mid I, D$

$P(S, G, D, I)$

$P(S)\ P(G \mid S)\ P(D \mid S, G)\ P(I \mid D, S, G)$

Minimal I-Map example

- Modified classic example: Discipline (D), Intelligence (I), SAT score (S), Grade (G)

$P(D, I, S, G)$

$D \perp\!\!\!\perp I$

$D \perp\!\!\!\perp G$

$G \perp\!\!\!\perp S \mid I$

$D \perp\!\!\!\perp I \mid G$ $D \perp\!\!\!\perp G \mid I, S$

$D \perp\!\!\!\perp G \mid I$ $S \perp\!\!\!\perp G \mid I, D$

$P(S, G, D, I)$

$P(S)\ P(G \mid S)\ P(D \mid S, G)\ P(I \mid D, S, G)$

What happens if I remove S-> D, is it still an I-Map?

Now we have an extra d-separation $S \perp'_G D \mid G$ not matching the conditional independences

Minimal I-Map example

- Modified classic example: Discipline (D), Intelligence (I), SAT score (S), Grade (G)

$P(D, I, S, G)$

$D \perp\!\!\!\perp I$

$D \perp\!\!\!\perp G$

$G \perp\!\!\!\perp S \mid I$

$D \perp\!\!\!\perp I \mid G$ $D \perp\!\!\!\perp G \mid I, S$

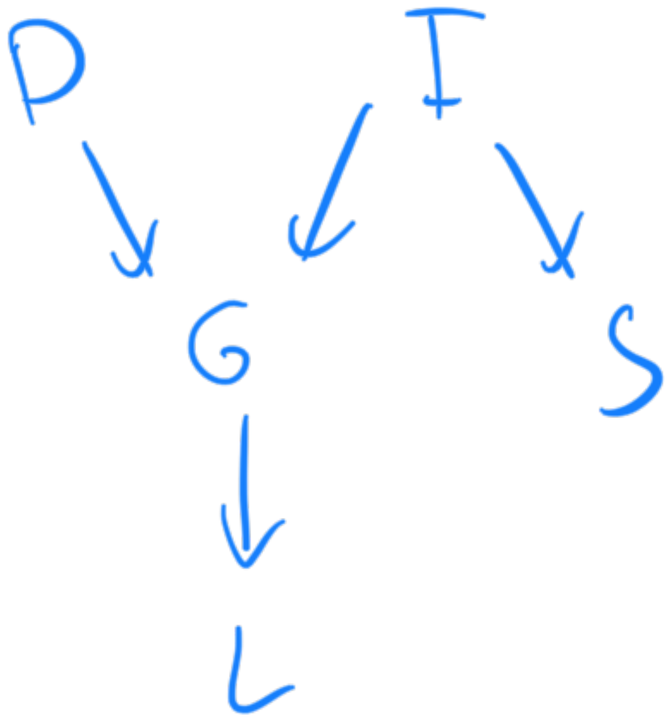
$D \perp\!\!\!\perp G \mid I$ $S \perp\!\!\!\perp G \mid I, D$

$P(S, G, D, I)$

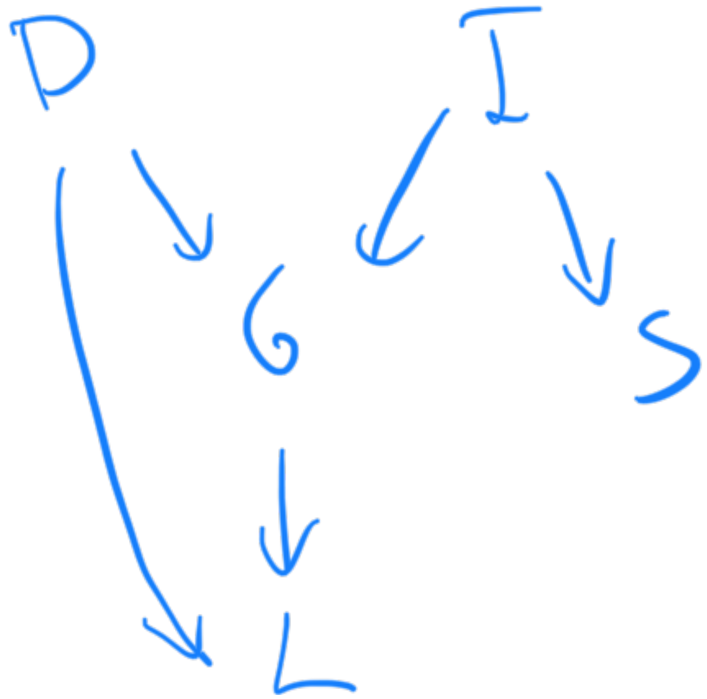
$P(S)\ P(G \mid S)\ P(D \mid S, G)\ P(I \mid D, S, G)$

We can show that this is also a minimal I-Map (we cannot remove any edge without creating an extra d-separation not matching independences in the distribution)

D-separation reprise and minimal I-maps: Canvas quiz



Graph 1



Graph 2

Minimal I-Maps vs independences:

$p(D, I, S, G)$

$D \perp\!\!\!\perp I$

$D \perp\!\!\!\perp G$

$G \perp\!\!\!\perp S \mid I$

$D \perp\!\!\!\perp I \mid G$ $\quad D \perp\!\!\!\perp G \mid I, S$

$D \perp\!\!\!\perp G \mid I$ $\quad S \perp\!\!\!\perp G \mid I, D$

$D \perp_d I$ ✓

$D \perp_d G$ ✓

$G \perp_d S \mid I$ ✓

$D \not\perp_d I$ ×

$D \not\perp_d G$ ×

$G \not\perp_d S \mid I$ ×

Can we reduce the set of minimal I-maps even further, possibly making them fit "perfectly" the conditional independences?

Global Markov Property and Faithfulness

- If (G, P) is a Bayesian network with a DAG $G = (\mathbf{V}, \mathbf{E})$, i.e. **$P$ factorizes according to G**, then for any disjoint $\mathbf{A}, \mathbf{B}, \mathbf{C} \subseteq \mathbf{V}$:

$$\mathbf{A} \perp \mathbf{B} \,|\, \mathbf{C} \implies X_{\mathbf{A}} \perp\!\!\!\perp X_{\mathbf{B}} \,|\, X_{\mathbf{C}}$$

- The reverse implication is not true in general, but if it is **P is faithful to G**

$$X_{\mathbf{A}} \perp\!\!\!\perp X_{\mathbf{B}} \,|\, X_{\mathbf{C}} \implies \mathbf{A} \perp \mathbf{B} \,|\, \mathbf{C}$$

- If P is Markov and faithful to G, we say that **G is a perfect map of P**
  - In general minimal maps can also not be perfect maps

Perfect maps

- If P is Markov and faithful to G, we say that **G is a perfect map of P.**

  Then, for any disjoint $\mathbf{A}, \mathbf{B}, \mathbf{C} \subseteq \mathbf{V}$:

$$\mathbf{A} \perp_G \mathbf{B} \mid \mathbf{C} \iff X_{\mathbf{A}} \perp\!\!\!\perp_P X_{\mathbf{B}} \mid X_{\mathbf{C}}$$

Perfect maps - existence

- Not every distribution P has a perfect map

$$\begin{cases} X_1 = \epsilon_1 \\ X_2 = 3X_1 \\ X_3 = 4X_2 + \epsilon_3 \\ \epsilon_1, \epsilon_3 \sim N(0,1) \end{cases}$$

Perfect maps - existence

- Not every distribution P has a perfect map

$$\begin{cases} X_1 = \epsilon_1 \\ X_2 = 3X_1 \\ X_3 = 4X_2 + \epsilon_3 \\ \epsilon_1, \epsilon_3 \sim N(0,1) \end{cases}$$

$$\begin{aligned} X_2 \perp\!\!\!\perp X_3 \,|\, X_1 &\iff \forall x_1, x_2, x_3 : p(X_2 = x_2 \,|\, X_3 = x_3, X_1 = x_1) = p(X_2 = x_2 \,|\, X_1 = x_1) \\ &\iff p(X_2 = 3x_1 \,|\, X_1 = x_1, X_3 = x_3) = p(X_2 = 3x_1 \,|\, X_1 = x_1) \end{aligned}$$

Perfect maps - existence

- Not every distribution P has a perfect map

$$\begin{cases} X_1 = \epsilon_1 \\ X_2 = 3X_1 \\ X_3 = 4X_2 + \epsilon_3 \\ \epsilon_1, \epsilon_3 \sim N(0,1) \end{cases}$$

$$X_2 \perp\!\!\!\perp X_3 \,|\, X_1 \iff \forall x_1, x_2, x_3 : p(X_2 = x_2 \,|\, X_3 = x_3, X_1 = x_1) = p(X_2 = x_2 \,|\, X_1 = x_1)$$
$$\iff p(X_2 = 3x_1 \,|\, X_1 = x_1, X_3 = x_3) = p(X_2 = 3x_1 \,|\, X_1 = x_1)$$

$$X_2 \not\perp_G X_3 \,|\, X_1$$

Perfect maps - existence

- Not every distribution P has a perfect map

$$\begin{cases} X_1 = \epsilon_1 \\ X_2 = 3X_1 \\ X_3 = 4X_2 + \epsilon_3 \\ \epsilon_1, \epsilon_3 \sim N(0,1) \end{cases}$$

$$X_2 \perp\!\!\!\perp X_3 \,|\, X_1 \iff \forall x_1, x_2, x_3 : p(X_2 = x_2 \,|\, X_3 = x_3, X_1 = x_1) = p(X_2 = x_2 \,|\, X_1 = x_1)$$
$$\iff p(X_2 = 3x_1 \,|\, X_1 = x_1, X_3 = x_3) = p(X_2 = 3x_1 \,|\, X_1 = x_1)$$

$$X_2 \not\perp_G X_3 \,|\, X_1$$

$$X_2 \not\perp\!\!\!\perp X_3 \quad X_1 \not\perp\!\!\!\perp X_3 \quad X_1 \perp\!\!\!\perp X_3 \,|\, X_2$$

Perfect maps - existence

- Not every distribution P has a perfect map

There doesn't exist a DAG that is a perfect map for P

$$\begin{cases} X_1 = \epsilon_1 \\ X_2 = 3X_1 \\ X_3 = 4X_2 + \epsilon_3 \\ \epsilon_1, \epsilon_3 \sim N(0,1) \end{cases}$$

$$X_2 \perp\!\!\!\perp X_3 \,|\, X_1 \iff \forall x_1, x_2, x_3 : p(X_2 = x_2 \,|\, X_3 = x_3, X_1 = x_1) = p(X_2 = x_2 \,|\, X_1 = x_1)$$
$$\iff p(X_2 = 3x_1 \,|\, X_1 = x_1, X_3 = x_3) = p(X_2 = 3x_1 \,|\, X_1 = x_1)$$

$$X_2 \not\perp_G X_3 \,|\, X_1$$

$$X_2 \not\perp\!\!\!\perp X_3 \quad X_1 \not\perp\!\!\!\perp X_3 \quad X_1 \perp\!\!\!\perp X_3 \,|\, X_2$$

Perfect maps - Markov equivalence

- If P is Markov and faithful to G, we say that **G is a perfect map of P.**

  Then, for any disjoint $\mathbf{A}, \mathbf{B}, \mathbf{C} \subseteq \mathbf{V}$:

$$\mathbf{A} \perp \mathbf{B} \,|\, \mathbf{C} \iff X_{\mathbf{A}} \perp\!\!\!\perp X_{\mathbf{B}} \,|\, X_{\mathbf{C}}$$

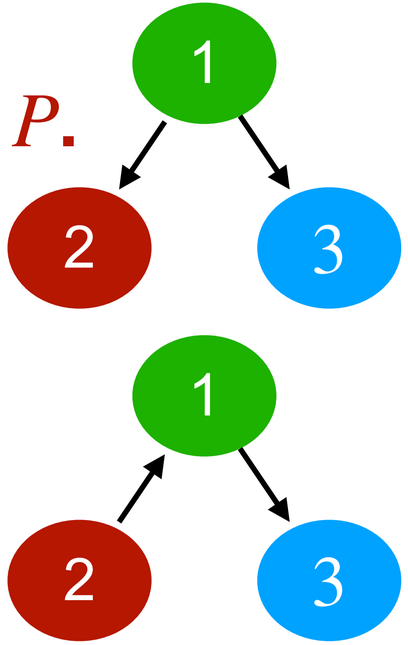
- In general there are multiple DAGs that can describe the same d-separations (and independences), i.e. there are multiple perfect maps

- We call these DAGs **Markov equivalent** and we **cannot distinguish them from observational data alone** (or without further assumptions)

Perfect maps - Markov equivalence

- If P is Markov and faithful to G, we say that **G is a perfect map of P.**
  Then, for any disjoint $\mathbf{A}, \mathbf{B}, \mathbf{C} \subseteq \mathbf{V}$:

$$\mathbf{A} \perp \mathbf{B} \,|\, \mathbf{C} \iff X_{\mathbf{A}} \perp\!\!\!\perp X_{\mathbf{B}} \,|\, X_{\mathbf{C}}$$



- In general there are multiple DAGs that can describe the same d-separations (and independences), i.e. there are multiple perfect maps

- We call these DAGs **Markov equivalent** and we **cannot distinguish them from observational data alone** (or without further assumptions)

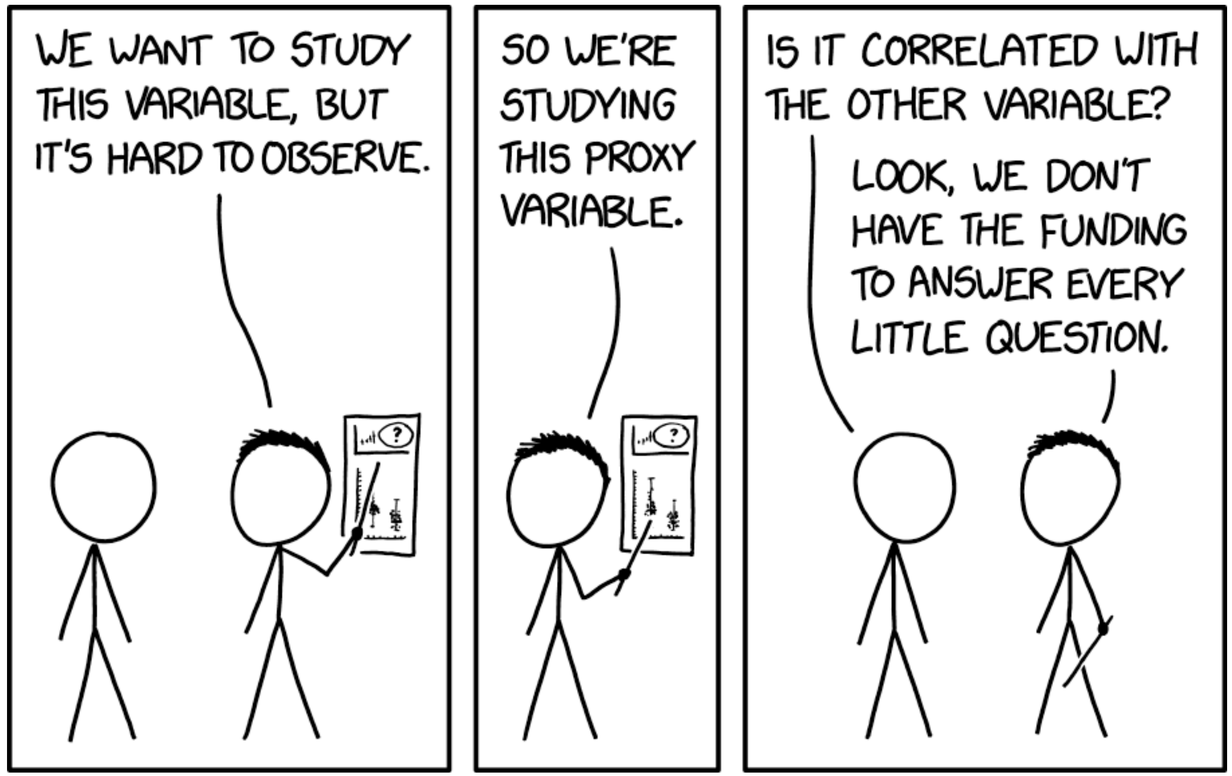
Next class - constraint-based causal discovery

- **In a nutshell:** we perform a set of conditional independence tests on the data and use them to constrain the possible graphs using d-separation

- In general, we can narrow down the possible graphs only up to their **Markov equivalence class (MEC)**

- The output of the algorithms we will see (e.g. SGS, PC) is a **CPDAG**, a mixed graph in which directed edges represent causal relations on which all DAGs in the MEC agree - these relations are **identifiable**

Questions?



https://xkcd.com/2652/