ETH zürich

Towards structure learning

Causality

Christina Heinze-Deml

Spring 2019

Last week

- Counterfactuals
- Potential outcomes
- (Propensity score) matching

Today

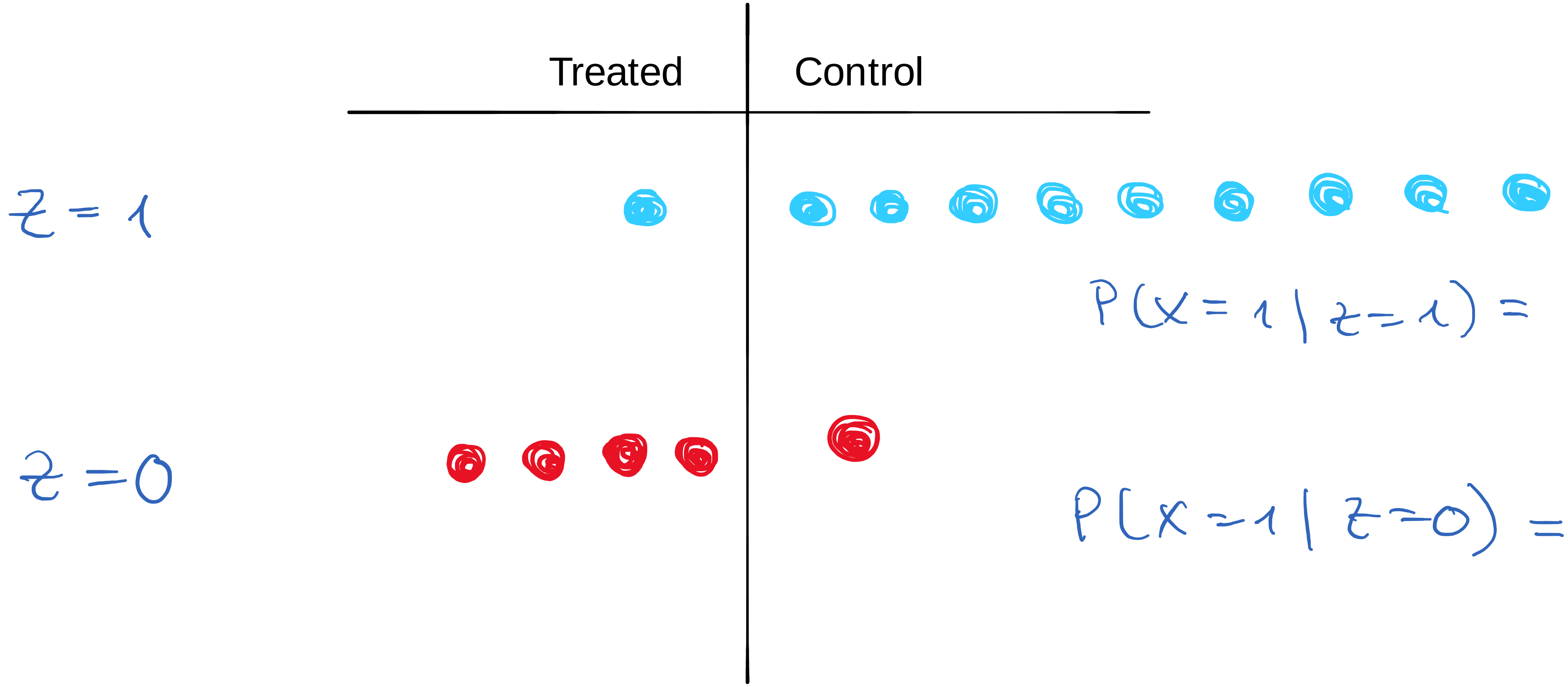
- Inverse probability weighting
- “Towards structure learning”

ETH zürich

Inverse probability weighting

- Example
  - Single binary confounder Z
  - Suppose propensity score $P(X = 1|Z = 1) = 0.1$
    - Among people with $Z = 1$, only 10% receive the treatment
  - Suppose propensity score $P(X = 1|Z = 0) = 0.8$
    - Among people with $Z = 0$, 80% receive the treatment

Inverse probability weighting

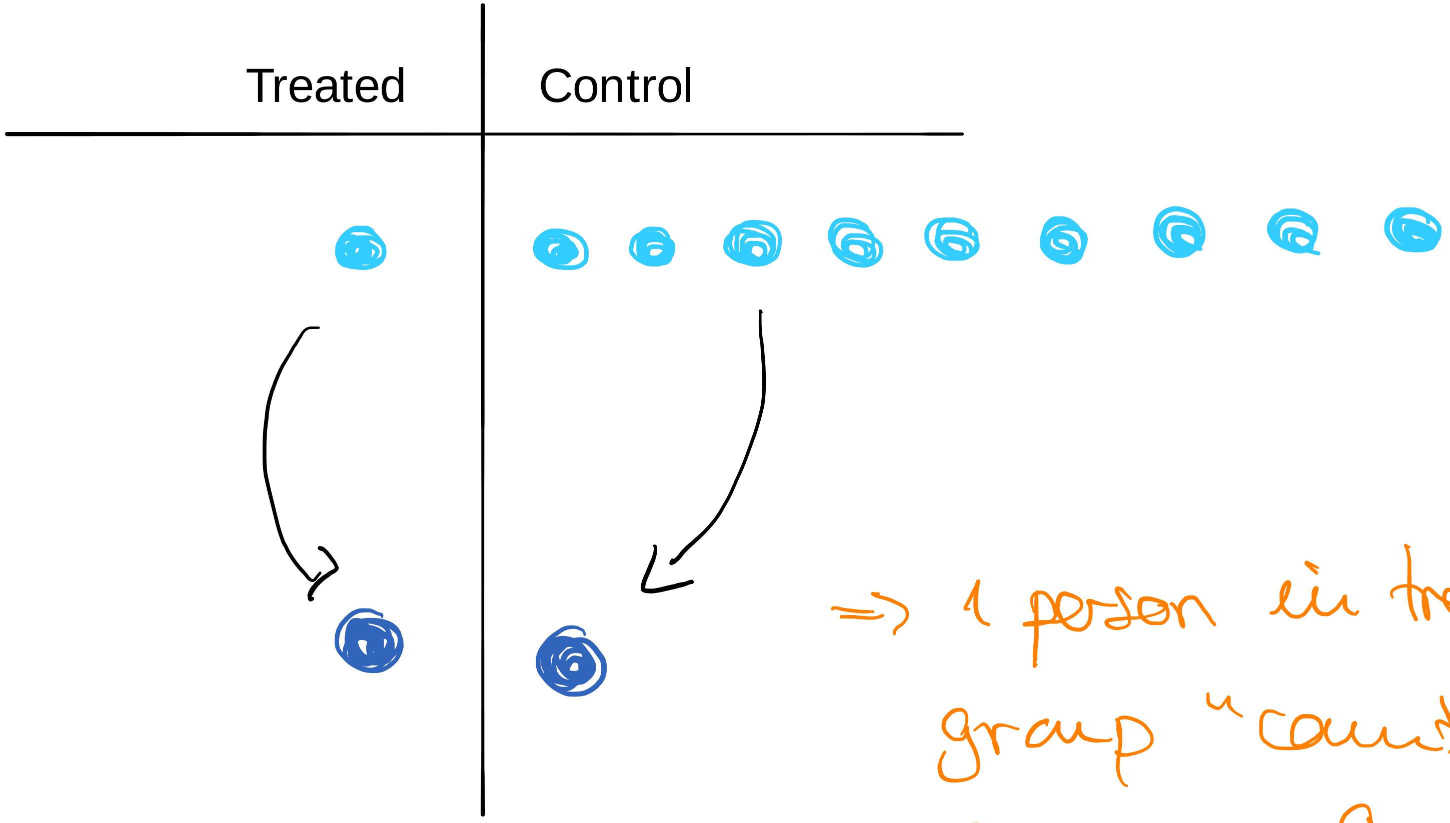


$P(x=1 \mid z=1) = 0.1$

$P(x=1 \mid z=0) = 0.8$

Inverse probability weighting

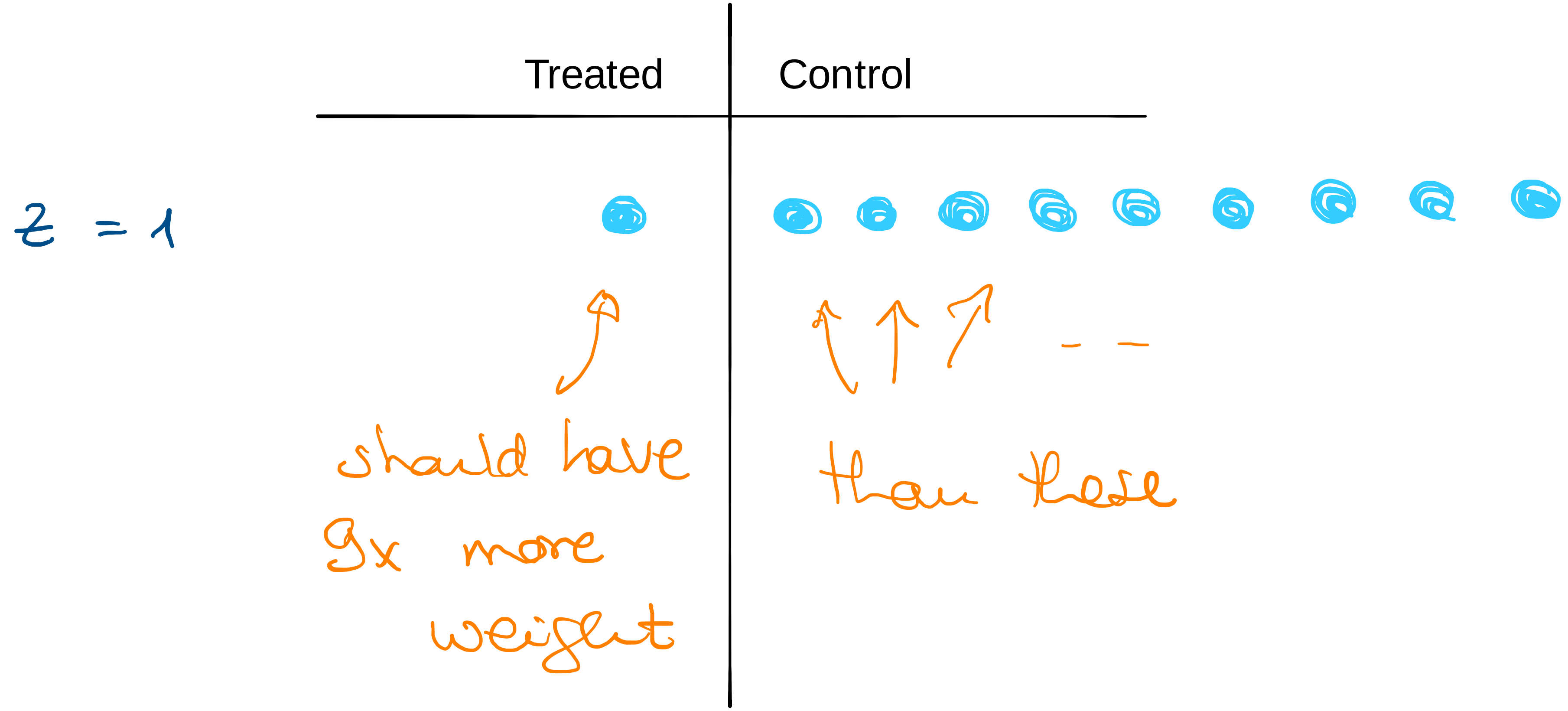
$z = 1$



Propensity score matching

$\Rightarrow$ 1 person in treatment group "counts" the same as 9 people from control

Inverse probability weighting



Inverse probability weighting

- Idea: rather than match, use all data but downweigh/upweigh observations
  - Weighting by the inverse of the probability of treatment received
  - For treated: weigh by the inverse of the propensity score $\pi = P(X = 1|Z)$
  - For control: weigh by the inverse of $1 - \pi = P(X = 0|Z)$

- Known as inverse probability of treatment weighting (IPTW)

Example

Inverse probability weighting

- Estimator

$$\hat{E}(Y|do(X=1)) = \frac{1}{n}\sum_i Y_i \, 1\{X_i = 1\} w_i$$

where $w_i = \frac{1}{\hat{\pi}_i} = \frac{1}{\hat{P}(X=1|Z_i)}$

- Equivalently for $\hat{E}(Y|do(X=0))$

- If propensity score $\pi_i = P(X=1|Z_i)$ is very small, weight will be very large
- Small estimation errors in $\hat{\pi}_i$ can lead to large estimation errors in $\hat{E}(Y|do(X=x))$

Inverse probability weighting

- More generally, consider
  - Observational distribution $p(x_{\boldsymbol{V}})$
  - Interventional distribution $p(x_{\boldsymbol{V}}|do(X_k \leftarrow \tilde{N}_k))$ with imperfect intervention
    - Shorthand: $p(x_{\boldsymbol{V}}|do(X_k \leftarrow \tilde{N}_k)) = \tilde{p}(x_{\boldsymbol{V}})$

- Factorizations agree except for the term of the intervened variable
  - Assume strictly positive densities

Inverse probability weighting

- Observational distribution $p(x_{\boldsymbol{V}})$
- Interventional distribution $p\left(x_{\boldsymbol{V}} \middle| do\left(X_k \leftarrow \widetilde{N}_k\right)\right) = \tilde{p}(x_{\boldsymbol{V}})$
- Interested in certain aspect $l(x_{\boldsymbol{V}})$, then:

$$E\left(l(x_{\boldsymbol{V}})|do\left(X_k \leftarrow \widetilde{N}_k\right)\right)$$

$$= \int l(x_{\boldsymbol{V}})\tilde{p}(x_{\boldsymbol{V}})dx_{\boldsymbol{V}}$$

$$= \int l(x_{\boldsymbol{V}})\frac{\tilde{p}(x_{\boldsymbol{V}})}{p(x_{\boldsymbol{V}})}p(x_{\boldsymbol{V}})dx_{\boldsymbol{V}}$$

$$= \int l(x_{\boldsymbol{V}})\frac{\tilde{p}\left(x_k|x_{\mathrm{pa}(k)}\right)}{p\left(x_k|x_{\mathrm{pa}(k)}\right)}p(x_{\boldsymbol{V}})dx_{\boldsymbol{V}}$$

Inverse probability weighting

- Given observations $x_V^1, \ldots, x_V^n$ drawn from observational distribution $p(x_V)$, can construct estimator for expectation under interventional distribution:

$$\hat{E}\left(l(x_V) | do\left(X_k \leftarrow \tilde{N}_k\right)\right) = \frac{1}{n} \sum_i l\left(x_V^i\right) w_i$$

where $w_i = \dfrac{\tilde{p}\left(x_k^i | x_{\mathrm{pa}(k)}^i\right)}{p\left(x_k^i | x_{\mathrm{pa}(k)}^i\right)}$

  - [See Series 4.]
  - Related to survey sampling, importance sampling, reinforcement learning
    - See Elements of Causal Inference, Chapter 8.2.

Towards structure learning

- Markov properties
- Causal minimality
- Faithfulness
- Markov equivalence

Markov properties

- Given a DAG $G = (\boldsymbol{V}, E)$, a distribution P with density p on $X_{\boldsymbol{V}}$ is said to satisfy:
  - The global Markov property wrt G if for all pairwise disjoint subsets $\boldsymbol{A}, \boldsymbol{B}$ and $\boldsymbol{S}$ of $\boldsymbol{V}$:

    $\boldsymbol{A}$ and $\boldsymbol{B}$ are d-separated by $\boldsymbol{S}$ in $G \Rightarrow X_{\boldsymbol{A}} \perp\!\!\!\perp X_{\boldsymbol{B}} | X_{\boldsymbol{S}}$ in P

  - The local Markov property with respect to G if for all $j \in \boldsymbol{V}$:

    $$X_j \perp\!\!\!\perp X_{\mathrm{nondesc}(j) \setminus \mathrm{pa}(j)} \mid X_{\mathrm{pa}(j)}$$

  - The Markov factorization property with respect to G if

    $$p(x_{\boldsymbol{V}}) = \prod_{j \in \boldsymbol{V}} p(x_j | x_{\mathrm{pa}(j)})$$

Markov properties

- If P has a density (with respect to a product measure), then all three Markov properties are equivalent. We then simply say that:
  - P is Markov with respect to G
  - Equivalently: G is an independence map (I-map) of P

Lauritzen (1996), Theorem 3.27

Why is the Markov property important?

- The Markov property connects a distribution and a DAG
- In a Bayesian network (G, p), the Markov property holds by definition
- If $X_{\boldsymbol{V}}$ is generated from a SEM with DAG $G = (\boldsymbol{V}, \boldsymbol{E})$, then the distribution of $X_{\boldsymbol{V}}$ is Markov with respect to G
  - Regardless of the choice of the structural equations

Pearl (2009), Theorem 1.4.1

Minimal I-map

- Every distribution is Markov with respect to a full DAG
  - Equivalently: a full DAG is an I-map of any distribution
- We are not interested in such DAGs, but in DAGs that are in some sense sparse. This leads to the following definition:

- Definition: A DAG $G = (\boldsymbol{V}, \boldsymbol{E})$ is a minimal I-map of a distribution P if:
  - G is an I-map of P, and
  - $G' = (\boldsymbol{V}, \boldsymbol{E}')$ with $\boldsymbol{E}' \subset \boldsymbol{E}$ is not an I-map of P
  - P is then said to satisfy causal minimality with respect to G

Minimal I-map

- Minimal I-maps can be easily constructed (see week 2):
  - Take any ordering of the variables
  - Write out the corresponding full factorization
  - Simplify the terms as much as possible and draw the corresponding DAG

Example

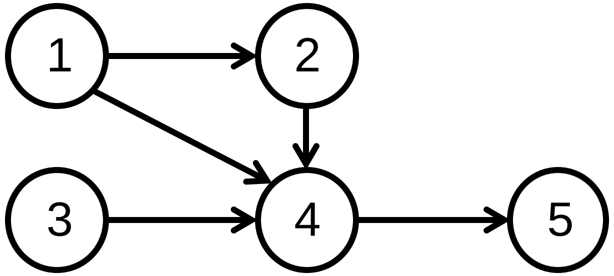
- Consider (X_1, X_2, X_3) and suppose that $X_1 \perp\!\!\!\perp X_3 | X_2$ is the only (conditional) independence:

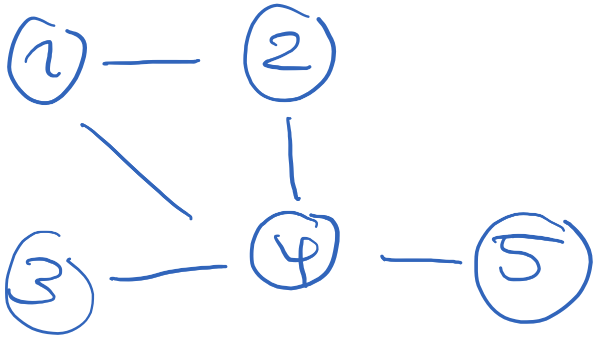
$$f(x_1|x_2, x_3) = f(x_1|x_2) \text{ and } f(x_3|x_1, x_2) = f(x_3|x_2)$$

  - Then $f(x_1, x_2, x_3) = f(x_1)f(x_2|x_1)f(x_3|x_1, x_2) = f(x_1)f(x_2|x_1)f(x_3|x_2)$
  - DAG: $1 \to 2 \to 3$

Graph terminology

- The skeleton of G does not take the directions of the edges into account: it is the graph $G = (\boldsymbol{V}, \tilde{\boldsymbol{E}})$ with $(i, j) \in \tilde{E}$ if $i \to j$ or $j \to i$ in G





Clicker question – I-maps and minimal I-maps

- Which of the following statements are correct?
  - A distribution does not have a unique I-map.
  - A distribution does not have a unique minimal I-map.
  - Different minimal I-maps of a distribution can have different skeletons.

- We will discuss the answers at hand of the following example:
  Consider (X_1, X_2, X_3) and suppose that $X_1 \perp\!\!\!\perp X_2 | X_3$ is the only (conditional) independence.

Faithfulness and perfect maps

- Given a DAG $G = (\boldsymbol{V}, \boldsymbol{E})$, a distribution P on $X_{\boldsymbol{V}}$ is said to be faithful with respect to G if for all pairwise disjoint subsets $\boldsymbol{A}, \boldsymbol{B}$ and $\boldsymbol{S}$ of $\boldsymbol{V}$:

$$X_{\boldsymbol{A}} \perp\!\!\!\perp X_{\boldsymbol{B}} | X_{\boldsymbol{S}} \text{ in } P \Rightarrow \boldsymbol{A} \text{ and } \boldsymbol{B} \text{ are d-separated by } \boldsymbol{S} \text{ in } G$$

- If a distribution P is Markov and faithful with respect to a DAG G, then G is said to be a perfect map of P. In this case, we have that for all pairwise disjoint subsets $\boldsymbol{A}, \boldsymbol{B}$ and $\boldsymbol{S}$ of $\boldsymbol{V}$:

$$X_{\boldsymbol{A}} \perp\!\!\!\perp X_{\boldsymbol{B}} | X_{\boldsymbol{S}} \text{ in } P \Leftrightarrow \boldsymbol{A} \text{ and } \boldsymbol{B} \text{ are d-separated by } \boldsymbol{S} \text{ in } G$$

Faithfulness and perfect maps

- If a distribution P is Markov and faithful with respect to a DAG G, then G is said to be a perfect map of P. In this case, we have that for all pairwise disjoint subsets $\boldsymbol{A}, \boldsymbol{B}$ and $\boldsymbol{S}$ of $\boldsymbol{V}$:

$$X_{\boldsymbol{A}} \perp\!\!\!\perp X_{\boldsymbol{B}} | X_{\boldsymbol{S}} \text{ in } P \Leftrightarrow \boldsymbol{A} \text{ and } \boldsymbol{B} \text{ are d-separated by } \boldsymbol{S} \text{ in } G$$

- Combination of the Markov and faithfulness assumptions creates one-to-one link between d-separation in the DAG and conditional independence in P
- This will turn out to be very convenient for structure learning
- Not every distribution has a perfect map

Examples

Perfect maps and Markov equivalence

- Definition: Two DAGs G_1 and G_2 are Markov equivalent if they describe the same set of d-separation relationships, i.e., for all pairwise disjoint subsets $\boldsymbol{A}, \boldsymbol{B}$ and $\boldsymbol{S}$ of $\boldsymbol{V}$, we have:

  $\boldsymbol{A}$ and $\boldsymbol{B}$ are d-separated by $\boldsymbol{S}$ in G_1 $\Leftrightarrow$ $\boldsymbol{A}$ and $\boldsymbol{B}$ are d-separated by $\boldsymbol{S}$ in G_2

- A perfect map (if it exists) is unique up to Markov equivalence

Recap

- Concepts to know:
  - Inverse probability weighting
  - Markov properties
  - Causal minimality
  - Faithfulness
  - Markov equivalence

ETH zürich

References and acknowledgments

- Slides adapted from M. Maathuis
- Inverse probability weighting
  - Peters, Janzing and Schölkopf (2017). Elements of Causal Inference. Chapter 8.2.1
- Markov properties, faithfulness and causal minimality
  - Peters, Janzing and Schölkopf (2017). Elements of Causal Inference. Chapter 6.5

- Optional reading:
  - Pearl (2009). Causality: Models, Reasoning and Inference. Chapter 1.
  - Lauritzen (1996). Graphical Models. Chapter 3.