Causal graphical models

Causality

Christina Heinze-Deml

Spring 2019

Announcements

- Series 2 will be uploaded later today

- Next week:
  - In-class exercise from 11-12
  - Please setup Jupyter and R (see website for details)

Last week

- Internal vs. external validity
- Graph terminology
- Directed acyclic graph (DAG) models
- Markov properties
- d-separation

DAG models

- Let $G = (\boldsymbol{V}, E)$ be a DAG and p be the distribution of $X_{\boldsymbol{V}}$
- The pair (G, p) is a DAG model or a Bayesian network if

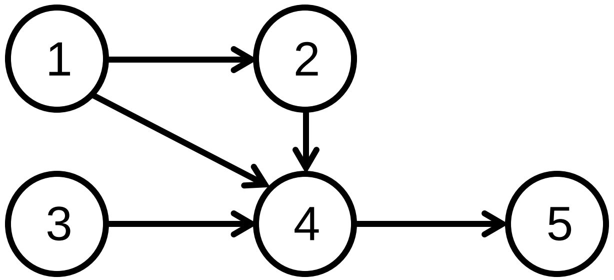
$$p(x_{\boldsymbol{V}}) = \prod_{i \in \boldsymbol{V}} p(x_i | x_{\mathrm{pa}(i)})$$

- If p factorizes according to G, d-separations in G imply conditional independencies in p

Graph terminology

- A non-endpoint node i is a collider on a path if the path contains $\rightarrow i \leftarrow$ (arrows collide at i).
- Otherwise, it is a non-collider on the path.

- Collider status is always relative to a path
  - 4 is a collider on the path (3,4,1)
  - 4 is a non-collider on the path (3,4,5)



d-separation

- A path between i to j is blocked by a set $\boldsymbol{S}$ (not containing i or j) if at least one of the following holds:
  - There is a non-collider on the path that is in $\boldsymbol{S}$; or
  - There is a collider on the path such that neither this collider nor any descendants are in $\boldsymbol{S}$.
- A path that is not blocked is active.

- If all paths between $i \in \boldsymbol{A}$ and $j \in \boldsymbol{B}$ are blocked by $\boldsymbol{S}$, then $\boldsymbol{A}$ and $\boldsymbol{B}$ are d-separated by $\boldsymbol{S}$. Otherwise they are d-connected given $\boldsymbol{S}$.

- Denote d-separation by $\perp$

Example

$X \rightarrow Z \rightarrow Y$

eg fire → smoke → alarm

$X \perp Y \mid Z$

$X \leftarrow Z \leftarrow Y$

$X \leftarrow Z \rightarrow Y$

eg shoe size ← age of child → reading ability

$X \perp Y \mid Z$

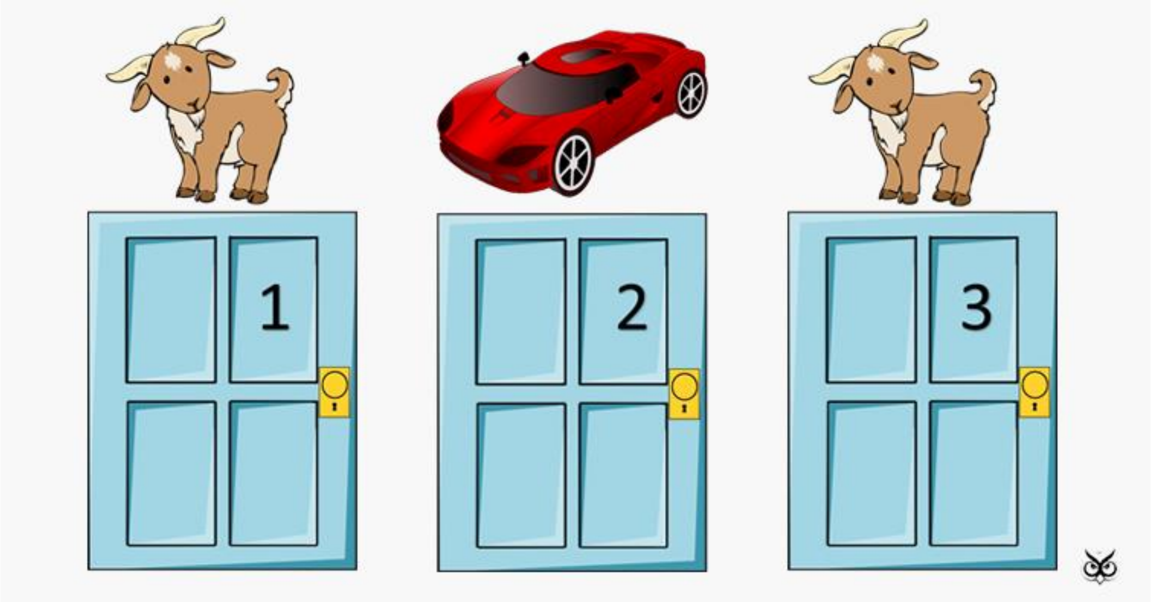
$X \rightarrow Z \leftarrow Y$

eg talent → celebrity ← beauty

$X \not\perp Y \mid Z$

Example: Monty Hall

- Quiz show hosted by Monty Hall
- Setting:
  - 3 closed doors with 1 car and 2 goats
  - You pick one door
  - Monty Hall opens one of the remaining doors with a goat
  - Then he asks you whether you want to switch
- How do you decide?



Picture from https://www.norwegiancreations.com/2018/10/bayes-rule-and-the-monty-hall-problem/

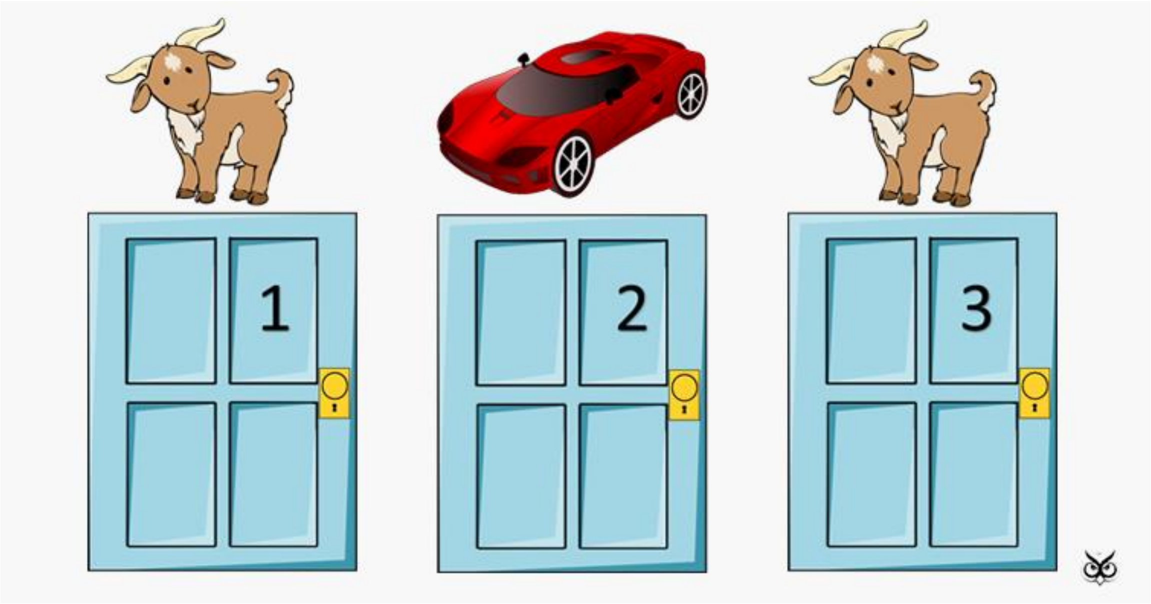
Example: Monty Hall

- Can show with Bayes rule:
  - If switch: success probability is 2/3
  - If stay: success probability is 1/3
- Instance of “selection bias”
  - “collider bias”, “Berkson’s paradox”

Your door

Location of car

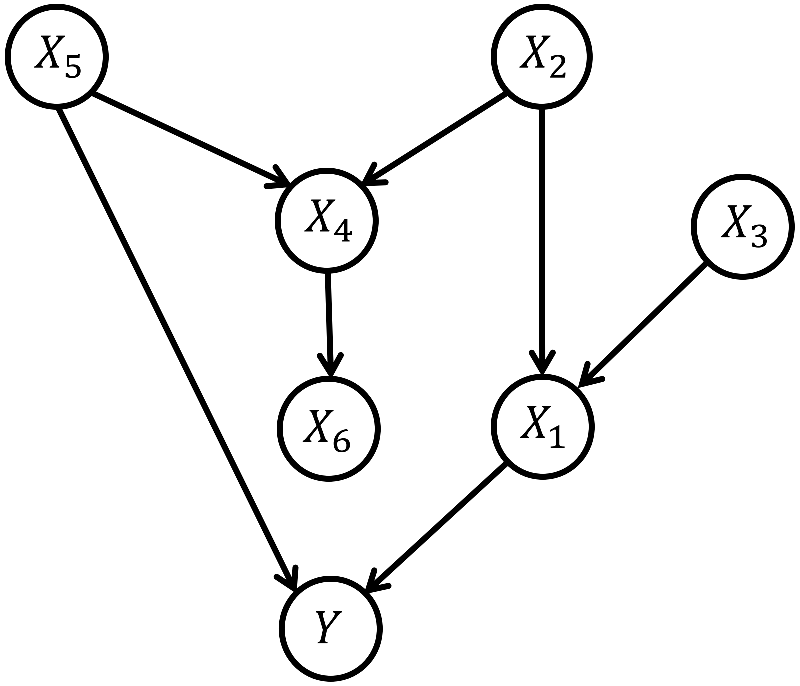
Door opened



Picture from https://www.norwegiancreations.com/2018/10/bayes-rule-and-the-monty-hall-problem/

Clicker question – d-separation

- Consider the following graph G. Assume $p(x_V)$ factorizes according to G.



$X_2 \perp\!\!\!\perp X_5$? Yes

$X_2 \perp\!\!\!\perp X_5 | X_6$? No

$X_2 \perp\!\!\!\perp X_5 | X_3$? Yes

$X_3 \perp\!\!\!\perp X_5 | X_2$? Yes

$X_2 \perp\!\!\!\perp X_3 | X_5$? Yes

$Y \perp\!\!\!\perp X_3$? No

$Y \perp\!\!\!\perp X_3 | X_1$? yes

DAG models

- A DAG model or Bayesian network is a combination (G, f), where G is a DAG and f is a distribution that factorizes according to G

- DAG models can be used for various purposes:
  - Estimating the joint density from low order conditional densities
  - Reading off conditional independencies from the DAG
  - Probabilistic reasoning (expert systems)
  - Causal inference

Probabilistic reasoning

- Conditional probabilities are rather counterintuitive for many people
- DAGs allow us to obtain conditional probabilities efficiently, using a “message passing” algorithm.
  - See R script “Graphical models”
  - We won’t discuss the details behind these algorithms

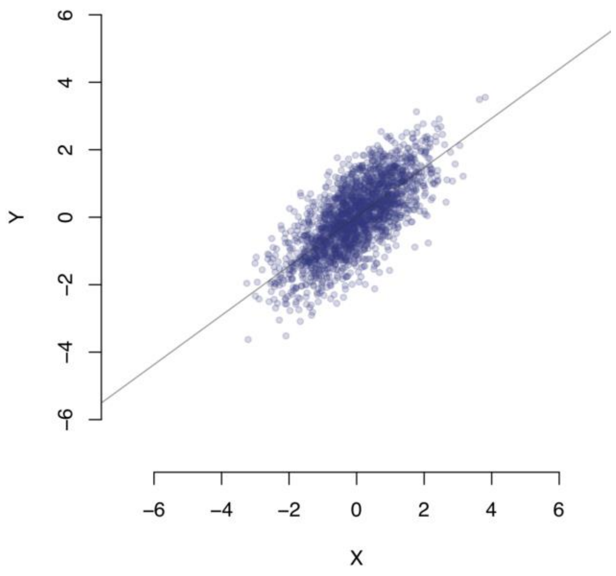
Today

- Selection bias
- Causal effects and do-operator
- Causal graphical models
- Structural equation models
- Path method

Classical regression models

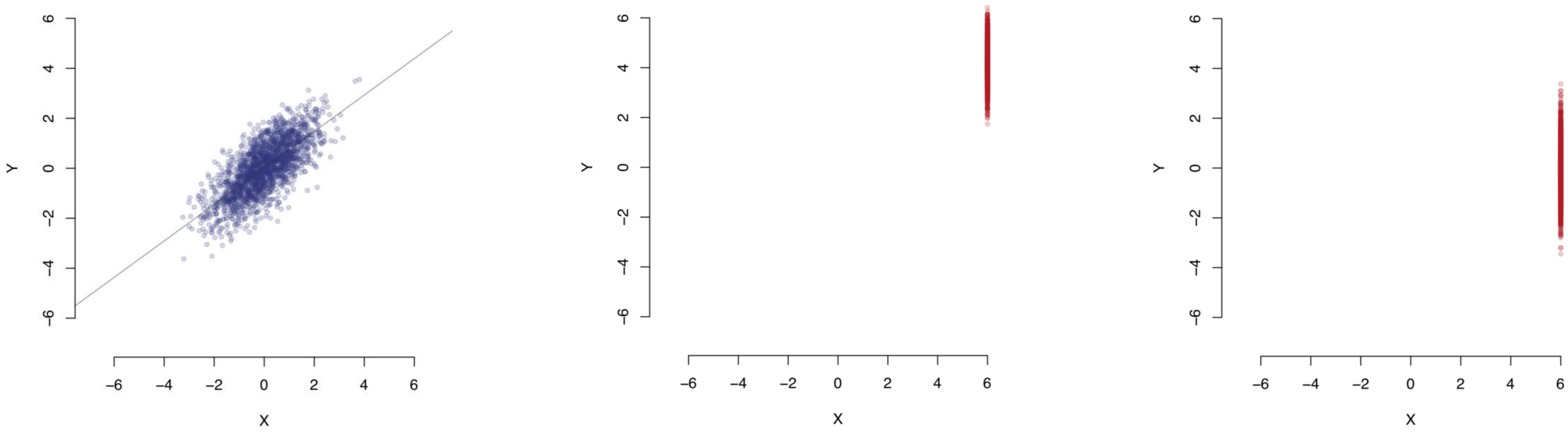
- We observe n i.i.d. observations of (X, Y) with distribution p
- Goal is to model certain aspects of $p(y|x)$, for example $E(Y|X = x)$
- Useful for prediction

↳ distribution of Y when we observe X=x



Classical regression models

- We observe n i.i.d. observations of (X, Y) with distribution p
- Goal is to model certain aspects of $p(y|x)$, for example $E(Y|X = x)$
- Useful for prediction – but what if we set X to e.g. 6?



X → Y

Y → X

Classical regression models

- We observe n i.i.d. observations of (X, Y) with distribution p
- Goal is to model certain aspects of $p(y|x)$, for example $E(Y|X = x)$
- Useful for prediction

- Such analyses are generally not useful for policy or treatment decisions, since such decisions involve predictions in manipulated systems with post-intervention distributions different from p

Causal effect and do-operator

- Interventional definition of causal effect:
  X has a causal effect on Y if manipulating X changes the distribution of Y

- Mathematical notion of manipulation (see Pearl):
  - $do(X = x)$ (or shorthand $do(x)$) represents a hypothetical intervention where X is set to the value x, uniformly over the entire population
  - $p(y|do(X = x))$ is the distribution of Y after $do(X = x)$
  - $E(Y|do(X = x))$ is the expectation of Y after $do(X = x)$, etc

Conditioning on observing: $p(y \mid see(X=x)) = p(y \mid x)$ (ordinary conditioning)

vs.

Intervening: $p(y \mid do(X=x))$, also written as $p^{do(X=x)}(y)$

Causal effect and do-operator

- Mathematical definition of causal effect:
  X has a causal effect on Y if $p(y|do(X = x'))$ depends on x',
  i.e., if $\exists$ a and b so that $p(y|do(X = a)) \neq p(y|do(X = b))$

- Average causal effect:
  $\mathrm{ACE}(x, x') = \mathrm{E}(y|do(X = x)) - E(y|do(X = x'))$

eg X binary; $X=0$ (control), $X=1$ (treatment)

$ACE = E(y|do(x=1)) - E(y|do(x=0))$

Example

- Consider a rehabilitation program for prisoners. Participation in the program is voluntary.
  - $X = 1$ if prisoner participated in the program; $X = 0$ otherwise
  - $Y = 1$ if prisoner is rearrested within a year; $Y = 0$ otherwise

- $\mathrm{P}(Y = 1|X = 1)$: probability of re-arrest for prisoners who choose to participate
- $\mathrm{P}(Y = 1|do(X = 1))$: probability of re-arrest if program were compulsory for all prisoners
- Note that generally $\mathrm{P}(Y = 1|do(X = 1)) \neq \mathrm{P}(Y = 1|X = 1)$

Example

- Suppose $\mathrm{P}(Y = 1|X = 1) < \mathrm{P}(Y = 1|X = 0)$.
  - Re-arrest rate among prisoners who participated in the program is lower than among those who did not participate
  - Could be due to the program, due to the intrinsic motivation of the prisoners who chose to participate, due to a mixture of these two, or….

- Suppose $\mathrm{P}(Y = 1|do(X = 1)) < \mathrm{P}(Y = 1|do(X = 0))$.
  - Program lowers the re-arrest rate, i.e., program has a causal effect on the re-arrest rate
  - Manipulating X changes the distribution of Y
  - X is causal for Y

Frameworks

- Causal DAG models (Causal Bayesian networks)
- Structural equation models
- Potential outcomes

Causal Bayesian networks

- Let $G = (\boldsymbol{V}, E)$ be a DAG and p be the distribution of $X_{\boldsymbol{V}}$
- The pair (G, p) is a DAG model or a Bayesian network if

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

Causal Bayesian networks

- Let $G = (\boldsymbol{V}, E)$ be a DAG and p be the distribution of $X_{\boldsymbol{V}}$
- The pair (G, p) is a causal DAG model or a causal Bayesian network if for any $\boldsymbol{W} \subset \boldsymbol{V}$

$$p(x_V \mid do(x_W = x_W')) = \begin{cases} \prod_{i \in V \setminus W} p(x_i \mid x_{pa(i)}) & \text{if } x_W = x_W' \\ 0 & \text{otherwise} \end{cases}$$

$$= \prod_{i \in V \setminus W} p(x_i \mid x_{pa(i)}) \, \mathbb{1}\{x_W = x_W'\}$$

Causal Bayesian networks

- Let $G = (\boldsymbol{V}, E)$ be a DAG and p be the distribution of $X_{\boldsymbol{V}}$
- The pair (G, p) is a DAG model or a Bayesian network if

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

- The pair (G, p) is a causal DAG model or a causal Bayesian network if for any $\boldsymbol{W} \subset \boldsymbol{V}$

$$p(x_{\boldsymbol{V}} \,|do(x_{\boldsymbol{W}} = x'_{\boldsymbol{W}})) = \prod_{i \in \boldsymbol{V} \setminus \boldsymbol{W}} p\left(x_i \middle| x_{\mathrm{pa}(i)}\right) 1\{x_{\boldsymbol{W}} = x'_{\boldsymbol{W}}\}$$

Causal Bayesian networks

- Let $G = (\boldsymbol{V}, E)$ be a DAG and p be the distribution of $X_{\boldsymbol{V}}$
- The pair (G, p) is a DAG model or a Bayesian network if

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

- The pair (G, p) is a causal DAG model or a causal Bayesian network if for any $\boldsymbol{W} \subset \boldsymbol{V}$

$$p\left(x_{\boldsymbol{V} \setminus \boldsymbol{W}} \,\middle|\, do(x_{\boldsymbol{W}} = x'_{\boldsymbol{W}})\right) = \prod_{i \in \boldsymbol{V} \setminus \boldsymbol{W}} p(x_i | x_{\mathrm{pa}(i)}) \Big|_{x_{\boldsymbol{W}} = x'_{\boldsymbol{W}}}$$

Causal Bayesian networks

- The pair (G, p) is a causal DAG model or a causal Bayesian network if for any $\boldsymbol{W} \subset \boldsymbol{V}$

$$p(x_{\boldsymbol{V}} \mid do(x_{\boldsymbol{W}} = x'_{\boldsymbol{W}})) = \prod_{i \in \boldsymbol{V} \setminus \boldsymbol{W}} p\left(x_i \middle| x_{\mathrm{pa}(i)}\right) 1\{x_{\boldsymbol{W}} = x'_{\boldsymbol{W}}\}$$

- Modified factorization known as
  - “g-formula” (Robins)
  - “manipulation formula” (Spirtes, Glymour, Scheines)
  - “truncated factorization formula” (Pearl)

Causal Bayesian networks

- The truncated factorization formula implies that an intervention on X_j only changes $p(x_j|x_{\mathrm{pa}(j)})$; the other conditional distributions remain unchanged. This is also known as invariance.

Compare:

$$p(x_V) = \Big\{ \prod_{i \in V \setminus \{j\}} p(x_i \mid x_{pa(i)}) \Big\} \, p(x_j \mid x_{pa(j)})$$

$$p(x_V \mid do(X_j = x_j')) = \Big\{ \prod_{i \in V \setminus \{j\}} p(x_i \mid x_{pa(i)}) \Big\} \cdot \mathbb{1}\{x_j = x_j'\}$$

Causal Bayesian networks

- The truncated factorization formula implies that an intervention on X_j only changes $p(x_j|x_{\mathrm{pa}(j)})$; the other conditional distributions remain unchanged. This is also known as invariance.

eg: A: altitude
T: temperature

$p(a,t) = p(t|a)\, p(a)$

→ if we change A, then we assume that the physical mechanism $p(t|a)$ responsible for producing an average temperature is still in place (invariance)

→ holds independent of $p(a)$

Causal Bayesian networks

- The pair (G, p) is a causal DAG model or a causal Bayesian network if for any $\boldsymbol{W} \subset \boldsymbol{V}$

$$p(x_{\boldsymbol{V}} \,|do(x_{\boldsymbol{W}} = x'_{\boldsymbol{W}})) = \prod_{i \in \boldsymbol{V} \setminus \boldsymbol{W}} p\left(x_i \middle| x_{\mathrm{pa}(i)}\right) 1\{x_{\boldsymbol{W}} = x'_{\boldsymbol{W}}\}$$

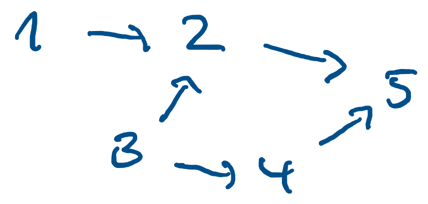
post-intervention distribution needed to define causal effects

conditional distributions that can be estimated from observational data

Causal Bayesian networks

- The modified factorizations represent factorizations wrt truncated graphs G_W, where all edges into $\boldsymbol{W}$ are removed

G

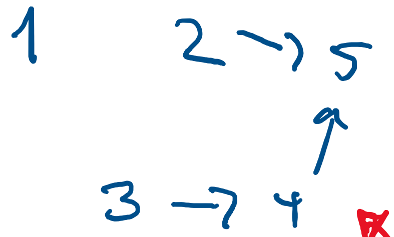


$$p(x_1, x_2, x_3, x_4, x_5) = p(x_1)\,p(x_3)\,p(x_2 \mid x_1, x_3)\,p(x_4 \mid x_3) \cdot p(x_5 \mid x_2, x_4)$$

Consider do-intervention on X_2

$$p(x_1, x_2, x_3, x_4, x_5 \mid do(X_2 = x_2')) = p(x_1)\,p(x_3)\,\mathbb{1}\{X_2 = x_2'\} \cdot p(x_4 \mid x_3)\,p(x_5 \mid x_2, x_4)$$

$G_{\{2\}}$

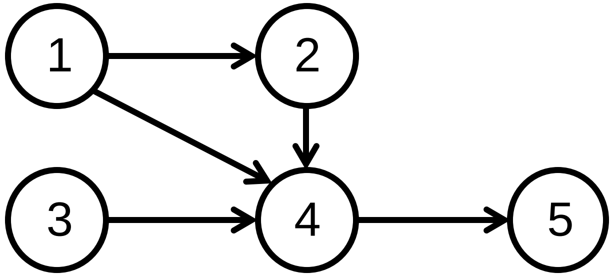


Truncated graph

$$p(x_1, x_3, x_4, x_5 \mid do(X_2 = x_2')) = p(x_1)\,p(x_3)\,p(x_4 \mid x_3) \cdot p(x_5 \mid x_2, x_4) \Big|_{x_2 = x_2'}$$

Causal Bayesian networks

- $X_{\mathrm{pa}(j)}$ can be interpreted as the direct causes of X_j
- Directed edges can be interpreted as direct causal effects



Example

- Consider the DAGs: $1 \to 2$ and $2 \to 1$
- Assume (X_1, X_2) are dependent
- Any distribution p of (X_1, X_2) factorizes wrt these two DAGs:
  $p(x_1, x_2) = p(x_1)p(x_2|x_1) = p(x_2)p(x_1|x_2)$.

- But the two DAGs are very different when interpreted causally:
  - The post-intervention distribution of X_1 is:
    - $p(x_1|do(x_2)) = p(x_1)$ for causal DAG $1 \to 2$
    - $p(x_1|do(x_2)) = p(x_1|x_2)$ for causal DAG $2 \to 1$
  - Thus, X_2 does not cause X_1 in the first DAG, but X_2 causes X_1 in the second graph

Examples

Causal DAGs

- Causal DAGs imply strong assumptions, allowing us to estimate post-intervention distributions from observational data

- How do we know the causal DAG?
  - Now: assume it is given, e.g. from background knowledge
  - Later: consider learning causal DAG (under some assumptions)
  - In any case, causal DAG provides clear framework to state causal assumptions for analysis
    - Allows for an honest debate about such assumptions
    - Can draw several possible causal DAGs, conduct the analysis for each of them and perform a sensitivity analysis

Frameworks

- Causal DAG models (Causal Bayesian networks)
- Structural equation models
- Potential outcomes

Structural equation models

- Let $X_{\boldsymbol{V}}$ be a collection of variables, and $G = (\boldsymbol{V}, E)$ be a DAG
- Each X_i is generated as a function of its graphical parents in G and noise ϵ_i:

$$X_i \leftarrow h_i\left(X_{\mathrm{pa}(i)}, \epsilon_i\right), \qquad i \in \boldsymbol{V}$$

  where $\epsilon_i, i \in \boldsymbol{V}$, are jointly independent

- The structural equations and the distribution of $\epsilon_{\boldsymbol{V}}$ yield a distribution p for $X_{\boldsymbol{V}}$

Structural equation models

- Then (G, p) is a causal Bayesian network if interventions are modelled as follows:
  - An intervention on X_j is modelled by replacing h_j. The generating mechanisms of the other variables (other structural equations) remain unchanged (invariance).
  - Thus, $do(X_j = x_j')$ is modelled by replacing

$$X_j \leftarrow h_j\left(X_{\mathrm{pa}(j)}, \epsilon_j\right) \qquad \text{by} \qquad X_j \leftarrow x_j'.$$

Intervention on X_2

$$\begin{cases} X_1 \leftarrow h_1(X_{pa(1)}, \varepsilon_1) \\ X_2 \leftarrow h_2(X_{pa(2)}, \varepsilon_2) \\ \vdots \\ X_p \leftarrow h_p(X_{pa(p)}, \varepsilon_p) \end{cases} \qquad \begin{cases} X_1 \leftarrow h_1(X_{pa(1)}, \varepsilon_1) \\ X_2 \leftarrow x_2' \\ \vdots \\ X_p \leftarrow h_p(X_{pa(p)}, \varepsilon_p) \end{cases}$$

Example

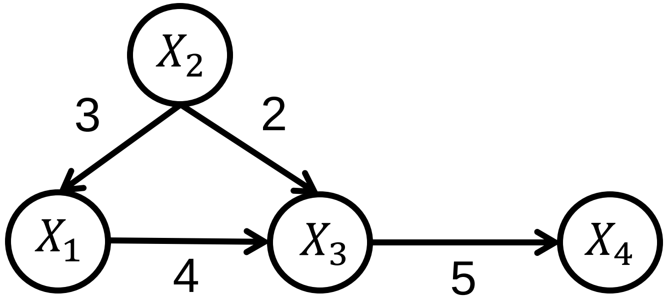
Clicker question – Observational and interventional distributions

Course outline

- Background and framework
- Using the known causal graph structure to identify causal effects
- Causal structure learning

Linear structural equation models

- Linear SEMs: all structural equations are linear and the noise is additive
- Example:
  - $X_1 \leftarrow 3X_2 + \varepsilon_1$
  - $X_2 \leftarrow \varepsilon_2$
  - $X_3 \leftarrow 4X_1 + 3X_2 + \varepsilon_3$
  - $X_4 \leftarrow 5X_3 + \varepsilon_4$



- What is the total causal effect of X_1 on X_4?
  - Path method:
    Increasing X_1 by 1 will on average increase X_3 by 4*1=4.
    Increasing X_3 by 4 will on average increase X_4 by 4*5=20.

Causal effects in linear SEMs via the path method

- Path method to compute the total causal effect of X_i on X_j in a linear SEM:
  - For each directed path from X_i to X_j, multiply the edge weights along the path
  - Sum up the results over all paths

- See R and note that it matches up with simulating from interventional distributions

Recap

- Concepts to know:
  - Selection bias
  - Causal effects and do-operator
  - Causal graphical models
  - Structural equation models
  - Path method

References and acknowledgments

- Slides adapted from M. Maathuis
- Some examples from
  - Script by J. Peters & N. Meinshausen (2018)
  - Peters, Janzing and Schölkopf (2017). Elements of Causal Inference.
  - Shalizi (2019). Advanced Data Analysis from an Elementary Point of View.