Causality

Talk 2: Directed acyclic graph (DAG) models

Yao Zhang

Intelligent Information Processing Research Group,

Faculty of Electrical Engineering and Computer Science,

Ningbo University

Note: The following slides are primarily adapted from the course materials[1].

Nov 26, 2025

[1] C. Heinze-Deml. Causality. URL: https://stat.ethz.ch/lectures/ss21/causality.php.

ETH zürich

Directed acyclic graph (DAG) models

Causality
Christina Heinze-Deml
Spring 2021

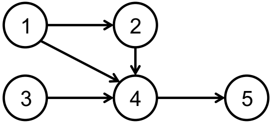
1

Today

- Graph terminology
- Directed acyclic graph (DAG) models
- Markov properties
- d-separation

Graph terminology

- A graph $G = (\boldsymbol{V}, \boldsymbol{E})$ consists of vertices (nodes) $\boldsymbol{V}$ and edges $\boldsymbol{E}$
- There is at most one edge between every ordered pair of vertices

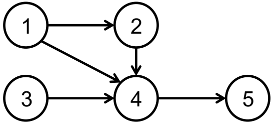


- Two vertices are adjacent if there is an edge between them
- If all edges are directed $(i \rightarrow j)$, the graph is called directed
- A path between i and j is a sequence of distinct vertices $(i, \dots, j)$ such that successive vertices are adjacent
- A directed path from i to j is a path between i and j where all edges are pointing towards j, i.e., $i \rightarrow \cdots \rightarrow j$

13

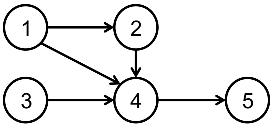
Graph terminology

- A cycle is a path $(i, j, \ldots, k)$ plus an edge between k and i
- A directed cycle is a directed path $(i, j, \ldots, k)$ from i to k, plus an edge $k \to i$
- A directed acyclic graph (DAG) is a directed graph without directed cycles



14

Graph terminology

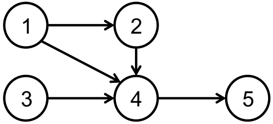


- If $i \to j$, then i is a parent of j, and j is a child of i
- If there is a directed path from i to j, then i is an ancestor of j and j is a descendant of i
- Each vertex is also an ancestor and descendant of itself
- The sets of parents, children, descendants and ancestors of i in G are denoted by $\mathrm{pa}(i, G)$, $\mathrm{ch}(i, G)$, $\mathrm{desc}(i, G)$, $\mathrm{an}(i, G)$
- We omit G if the graph is clear from the context

17

Graph terminology

- We write sets of vertices in bold face
- The previous definitions are applied disjunctively to sets
  - Example: $\mathrm{pa}(\boldsymbol{S}) = \bigcup_{k \in \boldsymbol{S}} \mathrm{pa}(k)$



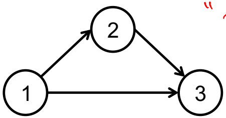
- The non-descendants of $\boldsymbol{S}$ are the complement of $\mathrm{desc}(\boldsymbol{S})$:

$$\mathrm{nondesc}(\boldsymbol{S}) := \boldsymbol{V} \setminus \mathrm{desc}(\boldsymbol{S})$$

18

Graph terminology

- We call G fully connected if all pairs of nodes are adjacent
- How many possibilities for a fully connected DAG?



"full DAG"

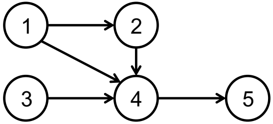
$1 \to 2 \to 3,\; 1 \to 3$ | $2 \to 1 \to 3,\; 2 \to 3$ | $3 \to 1 \to 2,\; 3 \to 2$

$1 \to 3 \to 2,\; 1 \to 2$ | $2 \to 3 \to 1,\; 2 \to 1$ | $3 \to 2 \to 1,\; 3 \to 1$

$p!$ possibilities ; $p = |V|$

DAGs and random variables

- Each vertex represents a random variable: vertex i represents random variable X_i
- If $\boldsymbol{A} \subseteq \boldsymbol{V}$, then $X_{\boldsymbol{A}} \coloneqq \{X_i : i \in \boldsymbol{A}\}$



- Edges denote relationships between pairs of variables (we will make this more precise)

26

Factorization of the joint density

- We can connect a distribution with density f to a DAG in the following way:
- We always have:

$$f(x_1, \dots, x_p) = f(\underline{x_1}) f(\underline{x_2 | x_1}) \dots f(x_p | x_1, \dots x_{p-1})$$ "chain rule"

- A set of variables $X_{\mathrm{pa}(j)}$ is said to be Markovian parents of X_j if it is a minimal subset of $\{X_1, \dots X_{j-1}\}$ such that $f(x_j | x_1, \dots, x_{j-1}) = f(x_j | x_{\mathrm{pa}(j)})$
  - Note: Markovian parents depend on the chosen ordering of the variables

27

Factorization of the joint density

- We always have:

$$f(x_1, \dots, x_p) = f(x_1) f(x_2|x_1) \dots f(x_p|x_1, \dots x_{p-1})$$

- A set of variables $X_{\mathrm{pa}(j)}$ is said to be Markovian parents of X_j if it is a minimal subset of $\{X_1, \dots X_{j-1}\}$ such that $f(x_j | x_1, \dots, x_{j-1}) = f(x_j | x_{\mathrm{pa}(j)})$
- Then

$$f(x_1, \dots, x_p) = \prod_{j=1}^{p} f(x_j | \underline{\underline{x_{\mathrm{pa}(j)}}})$$

"factorization property"

- We can draw a DAG accordingly
- The distribution is said to factorize according to this DAG

Notes week 2 - I

Consider (X_1, X_2, X_3) and suppose that $X_1 \perp\!\!\!\perp X_3 \mid X_2$

is the only (conditional) independence:

$$f(x_1 \mid x_2, x_3) = f(x_1 \mid x_2)$$

$$f(x_3 \mid x_1, x_2) = f(x_3 \mid x_2)$$

Then
$$\begin{aligned} f(x_1, x_2, x_3) &= f(x_1) f(x_2 \mid x_1) f(x_3 \mid x_1, x_2) \\ &= f(x_1) f(x_2 \mid x_1) f(x_3 \mid x_2) \end{aligned}$$
simplify

DAG: $1 \to 2 \to 3$

Or
$$\begin{aligned} f(x_3, x_2, x_1) &= f(x_3) f(x_2 \mid x_3) f(x_1 \mid x_2, x_3) \\ &= f(x_3) f(x_2 \mid x_3) f(x_1 \mid x_2) \end{aligned}$$
simplify

DAG: $3 \to 2 \to 1$

Or
$$f(x_1, x_3, x_2) = f(x_1) f(x_3 \mid x_1) f(x_2 \mid x_1, x_3)$$
cannot simplify further

DAG:
$$\begin{tikzcd} 1 \arrow[r] \arrow[rr, bend right] & 3 \arrow[r] & 2 \end{tikzcd}$$

Note: Markovian parents depend on the chosen ordering of the variables

Factorization of the joint density

- A distribution can factorize according to several DAGs
- Every distribution factorizes according to a full DAG
  - Note: there are *p!* possibilities

- Sometimes a distribution factorizes according to a sparse DAG
  - I.e., a DAG with few edges
  - E.g. first-order Markov chain:
    - $f(x_1, \dots, x_p) = f(x_1) f(x_2|x_1) \dots f(x_p|x_1, \dots, x_{p-1}) = f(x_1) f(x_2|x_1) \dots f(x_p|x_{p-1})$
    - DAG: $1 \to 2 \to \cdots \to p$

DAG models

- A DAG model or Bayesian network is a combination (G, P), where G is a DAG and P 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

Estimating the joint density

- Estimating the joint density of many variables is generally difficult
  - Example: The joint distribution of p binary variables requires $2^p - 1$ parameters

- But if you know that the distribution factorizes according to a DAG, then you only need to estimate $f(x_i | x_{\mathrm{pa}(i)})$ for $i = 1, \ldots, p$

- If the parent sets are small, this means we only need to estimate low order conditional densities

DAG models

- A DAG model or Bayesian network is a combination (G, P), where G is a DAG and P 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

Reading off conditional independencies: Markov property

- First-order Markov models: the future is independent of the past given the present

$$1 \to 2 \to \cdots \to (t-1) \to t \to (t+1)$$

$$X_{t+1} \perp\!\!\!\perp \{X_{t-1}, X_{t-2}, \dots, X_1\} \mid X_t$$

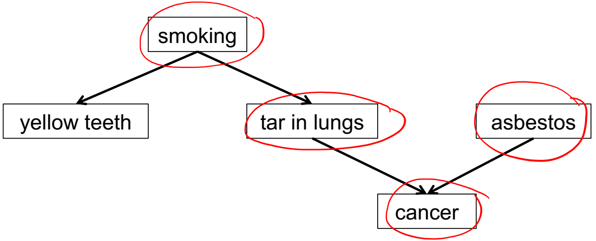
- In DAG models, we have a similar (local) Markov property
- Let $\boldsymbol{S}$ be any collection of nodes. Then:

$$X_{\boldsymbol{S}} \perp\!\!\!\perp X_{\mathrm{nondesc}(\boldsymbol{S}) \setminus \mathrm{pa}(\boldsymbol{S})} \mid X_{\mathrm{pa}(\boldsymbol{S})}$$

39

$X_S \perp\!\!\!\perp X_{nondesc(S)\setminus pa(S)} \mid X_{pa(S)}$

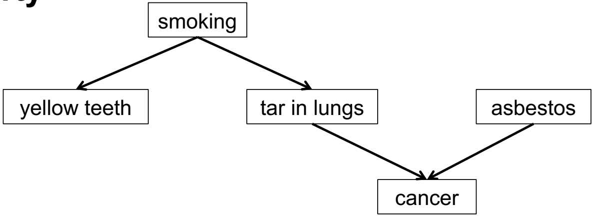
Example



- Take $\boldsymbol{S} = \{\text{yellow teeth}\}$ and apply the local Markov property
- Then:
  - $\mathrm{pa}(\text{yellow teeth}) = \{\text{smoking}\}$
  - $\mathrm{nondesc}(\text{yellow teeth}) = \{\text{smoking}, \text{tar}, \text{cancer}, \text{asbestos}\}$
- Hence, $\text{yellow teeth} \perp\!\!\!\perp \{\text{tar}, \text{cancer}, \text{asbestos}\} \mid \text{smoking}$ in any distribution that factorizes according to this DAG

40

Markov property



- Is $\text{tar} \perp\!\!\!\perp \text{asbestos} \mid \text{cancer}$?
- The local Markov property cannot be used to read off arbitrary conditional (in)dependencies
  - For this we have d-separation

41

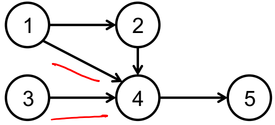
Graph terminology

- Need new 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

- Is 4 a collider in the given graph? ⇐ bad question

- 4 is a collider on the path (3, 4, 1)
- 4 is a non-collider (3, 4, 5)

⇒ collider status is always relative to a path!



42

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 $\perp$

Global Markov property

- Definition:
  A distribution P with density p satisfies the global Markov property with respect to a DAG G if:
  $$\boldsymbol{A} \text{ and } \boldsymbol{B} \text{ are d-separated by } \boldsymbol{S} \text{ in } G \Rightarrow X_{\boldsymbol{A}} \perp\!\!\!\perp X_{\boldsymbol{B}} \mid X_{\boldsymbol{S}} \text{ in } P$$

- Theorem (Pearl, 1988):
  A distribution P with density p satisfies the global Markov property with respect to G if and only if p factorizes according to G.

Notes week 2 - II

G :

$$\begin{tikzcd}
& 1 \arrow[dl] & \\
2 \arrow[dr] & & 3 \arrow[ul] \arrow[dl] \\
& 4 &
\end{tikzcd}$$

A distribution P is said to satisfy

* the global Markov property wrt G if

$$X_2 \perp\!\!\!\perp X_3 \mid X_1$$

and

$$X_1 \perp\!\!\!\perp X_4 \mid X_2, X_3$$

* the local Markov property wrt G if

$$X_2 \perp\!\!\!\perp X_3 \mid X_1$$

$$X_4 \perp\!\!\!\perp X_1 \mid X_2, X_3$$

* the Markov factorization property wrt G if

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

Faithfulness

- 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 $\boldsymbol{V}$:

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

Example

$X \rightarrow Z \rightarrow Y$	$X \perp Y \mid Z$	e.g. fire → smoke → alarm
$X \leftarrow Z \leftarrow Y$		
$X \leftarrow Z \rightarrow Y$	$X \perp Y \mid Z$	e.g. shoe size ← age of child → reading skills
$X \rightarrow Z \leftarrow Y$	$X \not\perp Y \mid Z$	e.g. talent → celebrity ← beauty

48

DAG models

- A DAG model or Bayesian network is a combination (G, P), where G is a DAG and P 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 `02_graphical_models.R`
  - We won’t discuss the details behind these algorithms

Discussion

Any comments or questions?

We may not always find an answer, and since we're not very familiar with causality, we will need to dedicate more time to this topic.