Causality
Backgrounds

Maciej Liśkiewicz

University of Lübeck

November, 2022

Causal Inference: Introduction

We will use material selected from different sources, including chapters of the following books:

- J. Pearl. Causality. Cambridge university press, 2009.
- J. Pearl, M. Glymour, and N.P. Jewell. Causal inference in statistics: A primer. John Wiley & Sons, 2016
- D. Koller and N. Friedman. Probabilistic graphical models: principles and techniques. MIT press, 2009.
- P. Spirtes, C.N. Glymour, R. Scheines, and D. Heckerman. Causation, prediction, and search. MIT press, 2000.

Additional source

- J. Peters, D. Janzing, and B. Schölkopf. Elements of causal inference: foundations and learning algorithms. MIT press, 2017.

Probabilities and Independencies

Basic axioms of probability calculus:

- $0 \leq P(A) \leq 1$
- $P(A \text{ or } B) = P(A) + P(B)$ if A and B are mutually exclusive
- $P(A) = P(A, B) + P(A, \neg B)$

More generally, if B_i, $i = 1, 2, \ldots, n$ is a set of exhaustive and mutually exclusive propositions (a partition) then we get the law of total probability:

$$P(A) = \sum_i P(A, B_i)$$

Probabilities and Independencies

- Conditional probabilities in terms of joint probability $P(A \mid B) = P(A, B)/P(B)$
- We say that A and B are independent if $P(A \mid B) = P(A)$
- Bayesian inference is based on the following inversion formula

$$P(H \mid e) = \frac{P(e \mid H) \cdot P(H)}{P(e)}$$

  for probability of hypothesis H upon obtaining evidence e

Probabilities and Independencies

- Conditional probabilities in terms of joint probability $P(A \mid B) = P(A, B)/P(B)$
- We say that A and B are independent if $P(A \mid B) = P(A)$
- Bayesian inference is based on the following inversion formula

$$P(H \mid e) = \frac{P(e \mid H) \cdot P(H)}{P(e)}$$

  for probability of hypothesis H upon obtaining evidence e
- If B_i, $i = 1, 2, \ldots, n$ is a partition then the probability of A can be computed as

$$P(A) = \sum_i P(A \mid B_i)P(B_i)$$

Probabilities and Independencies

- Conditional probabilities in terms of joint probability $P(A \mid B) = P(A, B)/P(B)$
- We say that A and B are independent if $P(A \mid B) = P(A)$
- Bayesian inference is based on the following inversion formula

$$P(H \mid e) = \frac{P(e \mid H) \cdot P(H)}{P(e)}$$

  for probability of hypothesis H upon obtaining evidence e
- If B_i, $i = 1, 2, \ldots, n$ is a partition then the probability of A can be computed as

$$P(A) = \sum_i P(A \mid B_i) P(B_i)$$

- For a set of n events, $E_1, E_2, \ldots, E_n$, the chain rule formula is stated as

$$P(E_1, E_2, \ldots, E_n) = P(E_n \mid E_{n-1}, \ldots, E_2, E_1) \ldots P(E_2 \mid E_1) P(E_1)$$

Probabilities and Independencies

Expectations

- For a random variable X and x from the domain of X we write $P(x)$ for $P(X = x)$
- The mean or expected value of X as

$$E(X) = \sum_{x} x \cdot P(x)$$

and the conditional version is defined as

$$E(X \mid y) = \sum_{x} x \cdot P(x \mid y)$$

Probabilities and Independencies

- The variance of X:

$$\sigma_X^2 = E((X - E(X))^2)$$

- The covariance of X and Y

$$\sigma_{XY} = E((X - E(X))(Y - E(Y)))$$

  the normalised version of which is called correlation coefficient

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

- The conditional variance, covariance, and correlation coefficient, given $Z = z$, are defined in a similar way; In particular: given $Z = z$

$$\rho_{XY|z} = \frac{\sigma_{XY|z}}{\sigma_{X|z} \sigma_{Y|z}}$$

Probabilities and Independencies

Expectations

Conditional Independence (CIs) and Graphoids

- Let $V = \{V_1, V_2, \ldots\}$ be a finite set of random variables.
- Let $P(\cdot)$ be a joint probability function over the variables in V, and let $X, Y, Z \subseteq V$.
- The sets X and Y are said to be conditionally independent given Z if

$$P(x \mid y, z) = P(x \mid z) \quad \text{if} \quad P(y, z) > 0$$

- This expresses the fact that learning Y does not provide additional information about X, once we know Z.

Probabilities and Independencies

- We will use the notation

$$(X \perp\!\!\!\perp Y \mid Z)_P \quad \text{or simply} \quad (X \perp\!\!\!\perp Y \mid Z)$$

to denote the conditional independence of X and Y given Z; thus,

$$(X \perp\!\!\!\perp Y \mid Z)_P \quad \text{iff} \quad P(x \mid y, z) = P(x \mid z)$$

for all values x, y, z such that $P(y, z) > 0$

Probabilities and Independencies

- We will use the notation

$$(X \perp\!\!\!\perp Y \mid Z)_P \quad \text{or simply} \quad (X \perp\!\!\!\perp Y \mid Z)$$

  to denote the conditional independence of X and Y given Z; thus,

$$(X \perp\!\!\!\perp Y \mid Z)_P \quad \text{iff} \quad P(x \mid y, z) = P(x \mid z)$$

  for all values x, y, z such that $P(y, z) > 0$

- Recall

  $(X \perp\!\!\!\perp Y \mid Z)$ means: "in any state of knowledge Z X tells us nothing new about Y

Probabilities and Independencies

- We will use the notation

$$(X \perp\!\!\!\perp Y \mid Z)_P \quad \text{or simply} \quad (X \perp\!\!\!\perp Y \mid Z)$$

  to denote the conditional independence of X and Y given Z; thus,

$$(X \perp\!\!\!\perp Y \mid Z)_P \quad \text{iff} \quad P(x \mid y, z) = P(x \mid z)$$

  for all values x, y, z such that $P(y, z) > 0$
- Recall

  $(X \perp\!\!\!\perp Y \mid Z)$ means: "in any state of knowledge Z X tells us nothing new about Y
- Unconditional independence (also called marginal independence) will be denoted by

$$(X \perp\!\!\!\perp Y \mid \emptyset)_P \quad \text{or} \quad (X \perp\!\!\!\perp Y)_P$$

- Note that

$$(X \perp\!\!\!\perp Y \mid Z) \quad \Longrightarrow \quad \forall V_i \in X \ \forall V_j \in Y \ (V_i \perp\!\!\!\perp V_j \mid Z)$$

  but not necessarily the converse

Probabilities and Independencies

A list of properties satisfied by the CIs relation $(X \perp\!\!\!\perp Y \mid Z)$:

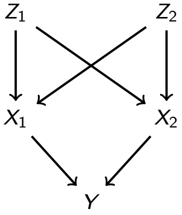
- **Symmetry:** $(X \perp\!\!\!\perp Y \mid Z) \implies (Y \perp\!\!\!\perp X \mid Z)$
- **Decomposition:** $(X \perp\!\!\!\perp YW \mid Z) \implies (X \perp\!\!\!\perp Y \mid Z)$
- **Weak union:** $(X \perp\!\!\!\perp YW \mid Z) \implies (X \perp\!\!\!\perp Y \mid ZW)$
- **Contraction:** $(X \perp\!\!\!\perp Y \mid Z) \& (X \perp\!\!\!\perp W \mid ZY) \implies (X \perp\!\!\!\perp YW \mid Z)$
- **Intersection:** $(X \perp\!\!\!\perp W \mid ZY) \& (X \perp\!\!\!\perp Y \mid ZW) \implies (X \perp\!\!\!\perp YW \mid Z)$

These properties were called *graphoid axioms* by Pearl and Paz (1987) and Geiger et al. (1990).

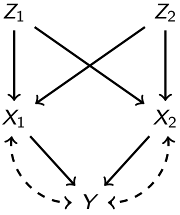
Graphs and Probabilities

Graphical Notation

- Let $G = (V, E)$ be a graph
- The vertices V will correspond to (random) variables
- The edges will denote a certain relationship between pairs of variables
- Each edge can be either
  - directed $V_i \rightarrow V_j$ or
  - undirected $V_i - V_j$
  - in some applications we will also use "bidirected" edges to denote the existence of unobserved common causes; For example:



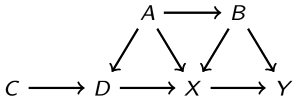
or



Graphs and Probabilities

Graphical Notation

- A path in a graph is a sequence of edges, e.g., $(C, D), (D, A), (A, B), (B, Y)$



- If every edge in a path is $A \to B$ we call it a directed path
- A graph that contains no directed cycles is called acyclic (DAG)
- We use the notions: parents, children, descendants, ancestors, spouses

Graphs and Probabilities

Graphical Notation

The role of graphs in probabilistic and statistical modeling

1. to provide convenient means of expressing substantive assumptions;
2. to facilitate economical representation of joint probability functions; and
3. to facilitate efficient inferences from observations

The second issue is nicely illustrated by the prominent model Bayesian Networks

Graphs and Probabilities

Bayesian Networks

- Task: to specify an arbitrary joint distribution, $P(x_1, \dots, x_n)$, for n variables
- The basic decomposition scheme offered by DAGs
- By the chain rule we get

$$P(x_1, \dots, x_n) = \prod_j P(x_j \mid x_1, \dots, x_{j-1})$$

- Suppose that the conditional probability of X_j is only sensitive to a small subset of the predecessors called

$$Pa_j$$

- Then $P(x_j \mid x_1, \dots, x_{j-1}) = P(x_j \mid pa_j)$

Graphs and Probabilities

Bayesian Networks

- Task: to specify an arbitrary joint distribution, $P(x_1, \ldots, x_n)$, for n variables
- The basic decomposition scheme offered by DAGs
- By the chain rule we get

$$P(x_1, \ldots, x_n) = \prod_j P(x_j \mid x_1, \ldots, x_{j-1})$$

- Suppose that the conditional probability of X_j is only sensitive to a small subset of the predecessors called

$$Pa_j$$

- Then $P(x_j \mid x_1, \ldots, x_{j-1}) = P(x_j \mid pa_j)$
- This allows to define a Bayesian network as an encoding of conditional independence relationships

$$P(x_1, \ldots, x_n) = \prod_j P(x_j \mid pa_j)$$

Graphs and Probabilities

Bayesian Networks

- For example, the DAG

$$
\begin{tikzcd}
X_1 \arrow[d] \arrow[drr] & & X_2 \arrow[d] \arrow[dll] \\
X_3 \arrow[dr] & & X_4 \arrow[dl] \\
& X_5 &
\end{tikzcd}
$$

- induces the following decomposition

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

Graphs and Probabilities

Bayesian Networks

- **Markov Compatibility** If a probability function P admits the factorization

$$P(x_1, \ldots, x_n) = \prod_j P(x_j \mid pa_j)$$

  relative to DAG G, we say that G represents P, that G and P are compatible, or that P is Markov relative to G.
- The core of the Bayesian network representation is a DAG G
- The second component is a set of “local” probability models that represent the dependence of each variable on its parents

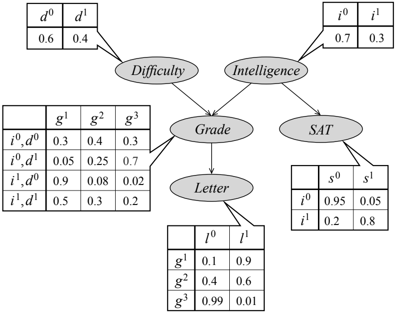
Graphs and Probabilities

Bayesian Network: Example 3.2.1 from Koller, Friedman (2009)

Consider the problem faced by a company trying to hire a recent college graduate:

- I – student's intelligence: *low, high*
- D – difficulty of the course: *easy, hard*
- G – student's grade in some course: $1, 2, 3$
- L – the quality of the recommendation letter : *strong, weak*
- S – the student's SAT score: *low, high*

The joint distribution has 48 entries. The corresponding example Bayesian network:



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Reasoning Pattern in BNs: $P(H = h \mid E = e)$

Graphs and Probabilities

d-Separation

***d*-Separation** A path π in a DAG G is said to be d-separated (or blocked) by a set of nodes Z if and only if

1. π contains a chain $i \rightarrow m \rightarrow j$ or a fork $i \leftarrow m \rightarrow j$ such that the middle node m is in Z, or
2. π contains an inverted fork (or collider) $i \rightarrow m \leftarrow j$ such that the middle node m is not in Z and such that no descendant of m is in Z.

Graphs and Probabilities

d-Separation

***d*-Separation** A path π in a DAG G is said to be d-separated (or blocked) by a set of nodes Z if and only if

1. π contains a chain $i \rightarrow m \rightarrow j$ or a fork $i \leftarrow m \rightarrow j$ such that the middle node m is in Z, or
2. π contains an inverted fork (or collider) $i \rightarrow m \leftarrow j$ such that the middle node m is not in Z and such that no descendant of m is in Z.

A set Z is said to d-separate X from Y in G, denoted as

$$(X \perp\!\!\!\perp Y \mid Z)_G$$

if and only if Z blocks every path from a node in X to a node in Y.

Graphs and Probabilities

d-Separation

- DAG $G = (V, E)$

$$\begin{tikzcd}
 & & & A \arrow[rr] \arrow[dl] \arrow[dr] & & B \arrow[dl] \arrow[dr] & \\
C \arrow[rr] & & D \arrow[rr] & & X \arrow[rr] & & Y
\end{tikzcd}$$

Graphs and Probabilities

d-Separation

- DAG $G = (V, E)$

$$
\begin{tikzcd}
 & & & A \arrow[rr] \arrow[dl] \arrow[dr] & & B \arrow[dl] \arrow[dr] & \\
C \arrow[rr] & & D \arrow[rr] & & X \arrow[rr] & & Y
\end{tikzcd}
$$

- A path: $V_1, \ldots, V_k$ s.t. $V_i \to V_{i+1}$ or $V_i \leftarrow V_{i+1}$ is in E for all $1 \leq i < k$.

Graphs and Probabilities

d-Separation

- DAG $G = (V, E)$

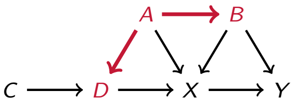
$$
\begin{tikzcd}
 & & A \arrow[rr] \arrow[dl] \arrow[dr] & & B \arrow[dl] \arrow[dr] & \\
C \arrow[r] & D \arrow[rr] & & X \arrow[rr] & & Y
\end{tikzcd}
$$

- A path: $V_1, \ldots, V_k$ s.t. $V_i \rightarrow V_{i+1}$ or $V_i \leftarrow V_{i+1}$ is in E for all $1 \leq i < k$.
- X on π is called a collider if π contains $\rightarrow X \leftarrow$.

Graphs and Probabilities

d-Separation

- DAG $G = (V, E)$

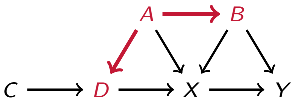


- A path: $V_1, \ldots, V_k$ s.t. $V_i \rightarrow V_{i+1}$ or $V_i \leftarrow V_{i+1}$ is in E for all $1 \leq i < k$.
- X on π is called a collider if π contains $\rightarrow X \leftarrow$.
- D and B are *d*-connected if there is a path π betwen D and B which does not contain a collider.

Graphs and Probabilities

d-Separation

- DAG $G = (V, E)$

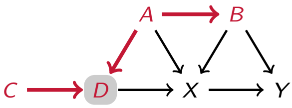


- A path: $V_1, \ldots, V_k$ s.t. $V_i \to V_{i+1}$ or $V_i \leftarrow V_{i+1}$ is in E for all $1 \leq i < k$.
- X on π is called a collider if π contains $\to X \leftarrow$.
- D and B are d-connected if there is a path π betwen D and B which does not contain a collider.
- Note: C and B are not d-connected.

Graphs and Probabilities

d-Separation

- DAG $G = (V, E)$



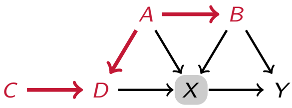
$W = \{D\}$

- A path: $V_1, \ldots, V_k$ s.t. $V_i \rightarrow V_{i+1}$ or $V_i \leftarrow V_{i+1}$ is in E for all $1 \leq i < k$.
- X on π is called a collider if π contains $\rightarrow X \leftarrow$.
- D and B are *d*-connected if there is a path π betwen D and B which does not contain a collider.
- Note: C and B are not *d*-connected.
- C and B are *d*-connected by a set W if there is π between them on which
  - every non-collider is not in W and
  - every collider is an ancestor of W.

Graphs and Probabilities

d-Separation

- DAG $G = (V, E)$



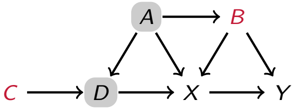
$W = \{X\}$

- A path: $V_1, \ldots, V_k$ s.t. $V_i \rightarrow V_{i+1}$ or $V_i \leftarrow V_{i+1}$ is in E for all $1 \leq i < k$.
- X on π is called a collider if π contains $\rightarrow X \leftarrow$.
- D and B are *d*-connected if there is a path π betwen D and B which does not contain a collider.
- Note: C and B are not *d*-connected.
- C and B are *d*-connected by a set W if there is π between them on which
  - every non-collider is not in W and
  - every collider is an ancestor of W.

Graphs and Probabilities

d-Separation

- DAG $G = (V, E)$



$W = \{A, D\}$

- A path: $V_1, \ldots, V_k$ s.t. $V_i \to V_{i+1}$ or $V_i \leftarrow V_{i+1}$ is in E for all $1 \le i < k$.
- X on π is called a collider if π contains $\to X \leftarrow$.
- D and B are d-connected if there is a path π betwen D and B which does not contain a collider.
- Note: C and B are not d-connected.
- C and B are d-connected by a set W if there is π between them on which
  - every non-collider is not in W and
  - every collider is an ancestor of W.
- W d-separates C and B if they are not d-connected by W.

Graphs and Probabilities

d-Separation

Theorem (*d*-separation vs. conditional independence (Verma, Pearl))

For any three disjoint subsets of nodes X, Y, Z in a DAG G and for all probability functions P, we have:

(1) $(X \perp\!\!\!\perp Y \mid Z)_G \quad \Longrightarrow \quad (X \perp\!\!\!\perp Y \mid Z)_P$ *whenever G and P are compatible; and*

(2) *if $(X \perp\!\!\!\perp Y \mid Z)_P$ holds in all distributions compatible with G, it follows that $(X \perp\!\!\!\perp Y \mid Z)_G$*

Graphs and Probabilities

d-Separation

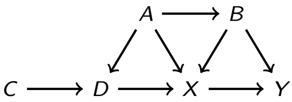
Theorem (d-separation vs. conditional independence (Verma, Pearl))

For any three disjoint subsets of nodes X, Y, Z in a DAG G and for all probability functions P, we have:

(1) $(X \perp\!\!\!\perp Y \mid Z)_G \quad \Longrightarrow \quad (X \perp\!\!\!\perp Y \mid Z)_P$ *whenever G and P are compatible; and*

(2) *if $(X \perp\!\!\!\perp Y \mid Z)_P$ holds in all distributions compatible with G, it follows that $(X \perp\!\!\!\perp Y \mid Z)_G$*

For example, for any P over $V = \{A, B, C, D, X, Y\}$ compatible with:



we have, e.g.: $(C \perp\!\!\!\perp B)_P$ and $(C \perp\!\!\!\perp X \mid \{D, A\})_P$

Literatur

- J. Pearl (2009), Ch.1
- J. Pearl, M. Glymour, and N.P. Jewell (2016), Ch. 1,2
- D. Koller and N. Friedman (2009), Ch.3