Lecture 23: Causal Inference & Directed graphs

Sofia Olhede

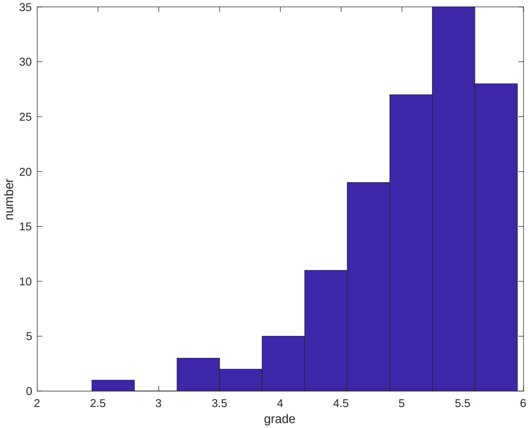
EPFL

December 8, 2020

1. Conditional independence

2. Fairness, Transparency and Accountability

Grades for midterm 2020



Distribution of midterm grades.

Conditional independence and graphs

- Theorem: if all events have positive probability then

$$X \perp\!\!\!\perp Y \mid Z \Rightarrow Y \perp\!\!\!\perp X \mid Z \tag{1}$$

$$X \perp\!\!\!\perp Y \mid Z \text{ and } U = h(X) \Rightarrow U \perp\!\!\!\perp Y \mid Z \tag{2}$$

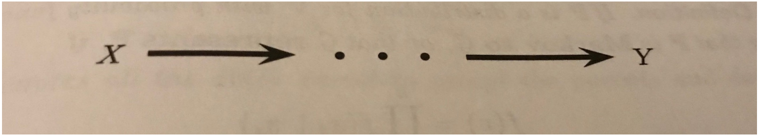
$$X \perp\!\!\!\perp Y \mid Z \text{ and } U = h(X) \Rightarrow X \perp\!\!\!\perp Y \mid (Z, U) \tag{3}$$

$$X \perp\!\!\!\perp Y \mid Z \text{ and } X \perp\!\!\!\perp W \mid (Y, Z) \Rightarrow X \perp\!\!\!\perp (W, Y) \mid Z \tag{4}$$

$$X \perp\!\!\!\perp Y \mid Z \text{ and } X \perp\!\!\!\perp Z \mid Y \Rightarrow X \perp\!\!\!\perp (Y, Z). \tag{5}$$

Conditional independence and graphs II

- A directed graph $\mathcal{G}$ consists of a set of vertices V with an edge set E of ordered vertices.
- In our discussion each vertex will represent a random variable.
- If $(X, Y) \in E$ then there is an arrow pointing from X to Y.
- If an arrow connects two variables X and Y (in either direction) then we say that X and Y are adjacent.
- If there is an arrow from X to Y then X is a parent of Y and Y is a child of X.
- The set of all parents of X is written as $\pi(X)$.
- A directed path between two variables is a set of arrows all pointing in the same direction linking one variable to the other as:



From 'All of Statistics': Figure 1.

Conditional independence and graphs III

- The sequence of adjacent variables that start from X and end with Y but are ignoring the direction of arrows is an undirected path.
- We say that X is an ancestor of Y if there is a directed path from X to Y. We also say that Y is a descendent of X.
- A configuration of the type:

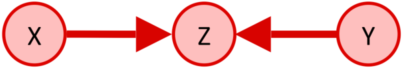


Figure 2. is a collider at the node Z. Other forms are not colliders.

- The collider property depends on what path one takes, so it is associated with a path.
- When variables that are pointed into the collider are not adjacent to each other then we say it is an unshielded collider.
- A directed path that starts end ends at the same node is a cycle. An acyclic graph has no cycles.

Directed Acyclic Graphs (DAGs)

- Let $\mathcal{G}$ be a DAG with vertices $V = (X_1, \ldots, X_k)$.
- Defn: If $\Pr\{\cdot\}$ is a distribution for vertex V with probability function $f(\cdot)$ then we say that $\Pr\{\cdot\}$ is <u>Markov to $\mathcal{G}$</u> if

$$f(v) = \prod_{i=1}^{k} f(x_i|\pi_i),$$

  where π_i are the parents of X_i. The set of distributions represented by $\mathcal{G}$ is denoted by $M(\mathcal{G})$.

Directed Acyclic Graphs (DAGs) II

- For the DAG below $\Pr\{\cdot\} \in M(\mathcal{G})$ if and only if its pdf is of the form

$$f(x, y, z, w) = f(x)f(y)f(z|x, y)f(w|z).$$

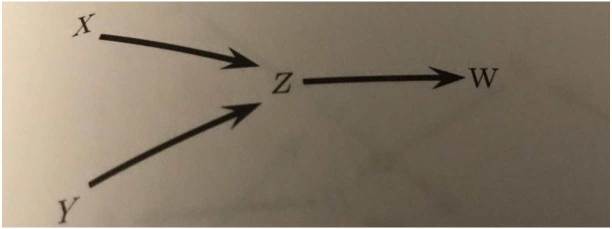


Figure 3.

Directed Acyclic Graphs (DAGs) III

- Theorem: A distribution $\Pr\{\cdot\} \in M(\mathcal{G})$ if and only if the following Markov condition holds: for every variable W

$$W \perp\!\!\!\perp \widetilde{W} \,|\, \pi_W, \tag{6}$$

  where $\widetilde{W}$ denotes all the other variables that are not parents or descendents of W.

- In Figure 3 the Markov condition implies that

$$X \perp\!\!\!\perp Y \quad \text{and} \quad W \perp\!\!\!\perp \{X, Y\} \,|\, Z.$$

- The Markov condition allows us to list some independence relations implies by a DAG. These relations might imply other independence relations.

Directed Acyclic Graphs (DAGs) IV

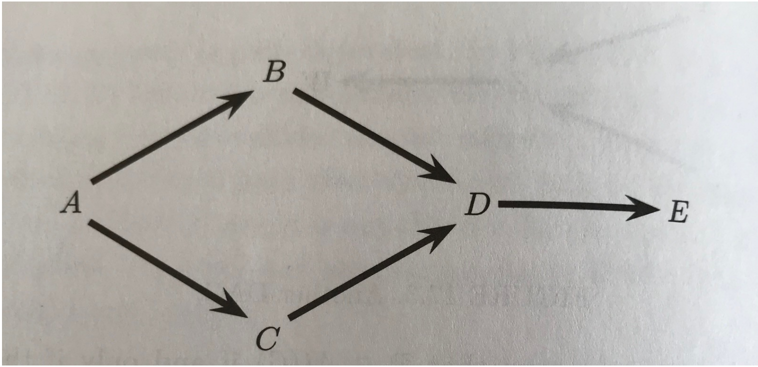


Figure 4.

- Figure 4 must have a pdf with the following structure:

$$f(a,b,c,d) = f(a)f(b|a)f(c|a)f(d|b,c)f(e|d)$$

$$D \perp\!\!\!\perp A \,|\, \{B, C\}, \quad E \perp\!\!\!\perp \{A, B, C\} \,|\, D, \quad B \perp\!\!\!\perp A \,|\, C.$$

Directed Acyclic Graphs (DAGs) V

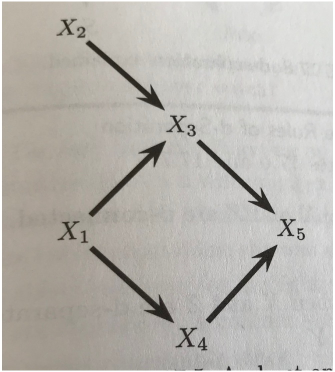


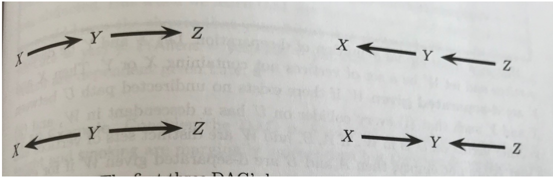
Figure 5.

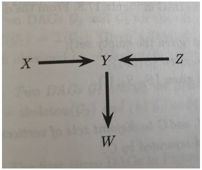
- Figure 5 implies:

$$X_1 \perp\!\!\!\perp X_2, \quad X_2 \perp\!\!\!\perp \{X_1, X_4\}, \quad X_3 \perp\!\!\!\perp X_4 \,|\, \{X_1, X_2\}$$
$$X_4 \perp\!\!\!\perp \{X_2, X_3\} \,|\, X_1, \quad X_5 \perp\!\!\!\perp \{X_1, X_2\} \,|\, \{X_3, X_4\}.$$

Directed Acyclic Graphs (DAGs) VI

- To go beyond the directly connected nodes we need the rules of *d*-Separation.
- Considering the DAGs in figs 6 & 7: when Y is not a collider X and Z are d–connected but they are *d*–separated given Y.
- If X and Z collide at Y then X and Z are *d*–separated but they are d–connected given Y.
- Conditioning on the descendant of a collider has the same effect as conditioning on the collider. Thus in Fig 7 X and Z are *d*–separated but they are *d*–connected given W.





Figures 6 & 7.

Directed Acyclic Graphs (DAGs) VII

- Defn: X and Y are <u>d–separated given W</u> if there exists no undirected path U between X and Y such that (i) every collider on U has a descendent in W, and (ii) no other vertex on U is in W. If A, B and W are distinct sets of vertices and A and B are non-empty, then A and B are d–separated given W if for every $X \in A$ and $Y \in B$, X and Y are d–separated given W. Sets of vertices that are not d–separated are d–connected.
- Theorem: Let A, B and C be disjoint sets of vertices. Then $A \perp\!\!\!\perp B \mid C$ if and only if A and B are d–separated by C.
- Graphs may appear to look different but in fact imply the same independence relations.
- For $\mathcal{G}$ a DAG let $\mathcal{I}(\mathcal{G})$ be all the independence statements implied by the graph.
- For $\mathcal{G}_1$ and $\mathcal{G}_2$ both DAGs on the same vertices, then they are <u>Markov equivalent</u> if $\mathcal{I}(\mathcal{G}_1) = \mathcal{I}(\mathcal{G}_2)$.

Directed Acyclic Graphs (DAGs) VIII

- Given DAG $\mathcal{G}$ we write as skeleton($\mathcal{G}$) the undirected graph obtained by replacing the arrows in the graph by undirected edges.
- Theorem: Two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent if and only if i) skeleton($\mathcal{G}_1$)=skeleton($\mathcal{G}_2$) and ii) $\mathcal{G}_1$ and $\mathcal{G}_2$ have the same unshielded colliders.
- Estimation for DAGs. What does the DAG framework bring?
- Assume we have a parametric model $f(x \mid \pi_x, \theta_x)$ The likelihood can then be written as

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f(V_i \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{j=1}^{m} f(X_{ij} \mid \pi_j, \, \theta_j).$$

here X_{ij} is the value of X_j for the ith data point.

Fairness and Algorithms

- Why is there a concern in using automated decision–making in society?
  Dignity;
  Fairness;
  Accountability.
- This discussion kicked off with the study of a Recidivism prediction instrument (considered for use/used in pre-trial decision-making, parole decisions, and in some US states even sentencing).
- A US company Northpointe Inc. developed a RPI called COMPAS. This was studied by a US civil liberties group ProPublica, and was accused of being racist.
- What does 'fair' mean anyway? ? Most of the following is from Chouldechova (2017).

Fairness and Algorithms II

- Let $S(x)$ denote a risk score based on some covariates $\mathrm{X} = \mathrm{x} \in \mathbb{R}^p$.
- Assume that the population can be split into two groups with each person getting a label R in $\{g, b\}$.
- Each assessed person is given an indicator Y which tells you if the person would re-offend or not.
- $S(x)$ is converted to Y by using a threshold s_{HR}.
- How can we tell if $S(x)$ is any good?
- Defn 1: A score $S(x)$ is well–calibrated if it reflects the same likelihood of recidivism irrespective of the persons' group membership. That is, if for all values s we have

$$\Pr\{Y = 1 \,|\, S = s, R = g\} = \Pr\{Y = 1 \,|\, S = s, R = b\}.$$

Fairness and Algorithms III

- Defn 2: (Predictive parity): A score $S(x) = s$ satisfies predictive parity at a threshold s_{HR} if the likelihood of recidivism among high-risk offenders is the same irrespective of group membership. That is:

$$\Pr\{Y = 1 \mid S > s_{HR}, R = g\} = \Pr\{Y = 1 \mid S > s_{HR}, R = b\}.$$

- Predictive parity at a given threshold s_{HR} is not equivalent to well-calibration.
- Defn 3: (Error rate balance). A score $S(x) = s$ satisfies error rate balance at a threshold s_{HR} if the false positive and false negative error rates are equal across groups. That is, if,

$$\Pr\{S > s_{HR} \mid Y = 0, R = g\} = \Pr\{S > s_{HR} \mid Y = 0, R = b\} \quad (7)$$
$$\Pr\{S \leq s_{HR} \mid Y = 1, R = g\} = \Pr\{S \leq s_{HR} \mid Y = 1, R = b\}. \quad (8)$$

The first line are the group-specific false positive rates, and the second line are the group-specific false negative rates.

Fairness and Algorithms IV

- Defn 4: (Statistical parity). A score $S(x) = s$ satisfies statistical parity at a threshold s_{HR} if the proportion of of individuals classified as high-risk is the same for each group. This mathematically is

$$\Pr\{S > s_{HR} \mid R = g\} = \Pr\{S > s_{HR} \mid R = b\}.$$

This is considered as the "group fairness" condition.