Course Notes: A Crash Course on Causality – Week 2: Confounding and Directed Acyclic Graphs (DAGs)

Yingbo Li

07/09/2021

Table of Contents

Confounding

- Confounders: variables that affect both the treatment and the outcome
  - If we assign treatment based on a coin flip, since the coin flip doesn't affect the outcome, it's not a confounder
  - If older people are at higher risk of heart disease (the outcome) and are more likely to receive the treatment, then age is a confounder
- To control for confounders, we need to
  1. Identify a set of variables X that will make the ignorability assumption hold
     - Causal graphs will help answer this question
  2. Use statistical methods to control for these variables and estimate causal effects

Overview of graphical models

- Encode assumption about relationship among variables
  - Tells use which variables are independent, dependent, conditionally independent, etc

Terminology of graphs

- Directed graph: shows that A affects Y

$$A \longrightarrow Y$$

- Undirected graph: A and Y are associated with each other

$$A \longrightarrow\!\!\!\!\!- \; Y$$

- Nodes or vertices: A and Y
  - We can think of them as variables
- Edge: the link between A and Y
- Directed graph: all edges are directed
- Adjacent variables: if connected by an edge

Paths

$$\begin{tikzcd} W \arrow[r] & Z \arrow[r] \arrow[d] & B \\ & A \arrow[ur] & \end{tikzcd}$$

- A path is a way to get from one vertex to another, traveling along edges
- There are 2 paths from W to B:
  - $W \to Z \to B$
  - $W \to Z \to A \to B$

Directed Acyclic Graphs (DAGs)

- No undirected paths

$$\begin{tikzcd} Z \arrow[r] \arrow[d, dash] & B \\ A & \end{tikzcd}$$

- No cycles

$$\begin{tikzcd} Z \arrow[d] & B \arrow[l] \\ A \arrow[ur] & \end{tikzcd}$$

- This is a DAG

$$\begin{tikzcd} Z \arrow[r] \arrow[d] & B \\ A \arrow[ur] & \end{tikzcd}$$

More terminology

$$\begin{tikzcd}
A \arrow[r] & Z \arrow[r] \arrow[d] & D \\
 & B \arrow[ur] &
\end{tikzcd}$$

- A is Z's parent
- D has two parents, B and Z
- B is a child of Z
- D is a descendant of A
- Z is a ancestor of D

DAG example 1

$$
\begin{tikzcd}
A \arrow[r] & B & C \\
D \arrow[u] & &
\end{tikzcd}
$$

- C is independent of all variables

$$P(C \mid A, B, D) = P(C)$$

- B and C, D are independent, conditional on A

$$P(B \mid A, C, D) = P(B \mid A) \iff B \perp C, D \mid A$$

- B and D are marginally dependent

$$P(B \mid D) \neq P(B)$$

DAG example 2

$$
\begin{tikzcd}
A \arrow[rr, bend left] & B \arrow[r] & C \\
 & D \arrow[ul] \arrow[u] &
\end{tikzcd}
$$

- A and B are independent, conditional on C and D

$$P(A \mid B, C, D) = P(A \mid C, D) \iff A \perp B \mid C, D$$

- C and D are independent, conditional on A and B

$$P(D \mid A, B, C) = P(D \mid A, B) \iff D \perp C \mid A, B$$

Decomposition of joint distributions

1. Start with roots (nodes with no parents)
2. Proceed down the descendant line, always conditioning on parents

$$\begin{tikzcd}
A \arrow[r] & B & C \\
D \arrow[u] & &
\end{tikzcd}$$

- $P(A, B, C, D) = P(C)P(D)P(A \mid D)P(B \mid A)$

$$\begin{tikzcd}
A \arrow[rr, bend left] & B \arrow[r] & C \\
& D \arrow[ul] \arrow[u] &
\end{tikzcd}$$

- $P(A, B, C, D) = P(D)P(A \mid D)P(B \mid D)P(C \mid A, B)$

Compatibility between DAGs and distributions

- In the above examples, the DAGs admit the probability factorizations. Hence, the probability function and the DAG are compatible
- DAGs that are compatible with a particular probability function are not necessarily unique
  - Example1:

    $$A \longrightarrow B$$

  - Example 2:

    $$A \longleftarrow B$$

  - In both of the above examples, A and B are dependent, i.e., $P(A, B) \neq P(A)P(B)$

Types of paths

- Forks

$$D \longleftarrow E \longrightarrow F$$

- Chains

$$D \longrightarrow E \longrightarrow F$$

- Inverted forks

$$D \longrightarrow E \longleftarrow F$$

When do paths induce associations?

- If nodes A and B are on the ends of a path, then they are associated (via this path), if
  - Some information flows to both of them (aka Fork), or
  - Information from one makes it to the other (aka Chain)
- Example: information flows from E to A and B

$$\begin{tikzcd} & E \arrow[dl] \arrow[dr] & \\ D \arrow[d] & & G \arrow[d] \\ A & & B \end{tikzcd}$$

- Example: information from A makes it to B

$$A \longrightarrow G \longrightarrow D \longrightarrow F \longrightarrow B$$

Paths that do not induce association

- Information from A and B collide at G

$$A \longrightarrow G \longleftarrow B$$

- G is a collider
- A and B both affect G:
  - Information does not flow from G to either A or B
  - **So A and B are independent** (if this is the only path between them)
- If there is a collider anywhere on the path from A to B, then no association between A and B comes from this path

$$A \longrightarrow G \longleftarrow D \longleftarrow B$$

Blocking on a chain

- Paths can be blocked by conditioning on nodes in the path
- In the graph below, G is a node in the middle of a chain. If we condition on G, then we block the path from A to B

$$A \longrightarrow G \longrightarrow B$$

- For example, A is the temperature, G is whether sidewalks are icy, and B is whether someone falls
  - A and B are associated marginally
  - But if we conditional on the sidewalk condition G, then A and B are independent

Blocking on a fork

- Associations on a fork can also be blocked
- In the following fork, if we condition on G, then the path from A to B is block

$$A \leftarrow G \rightarrow B$$

No need to to block a collider

- The opposite situation occurs if a conllider is blocked

$$A \longrightarrow G \longleftarrow B$$

- In the following inverted fork
  - Originally A and B are not associated, since information collides at G
  - But if we condition on G, then A and B become associated
- Example: A and B are the states of two on/off switches, and G is whether the lightbulb is lit up.
  - The two switches A and B are determined by two independent coin flips
  - G is lit up only if both A and B are in the on state
  - Conditional on G, the two switches are not independent: if G is off, then A must be off if B is on

d-separation

- A path is d-separated by a set of nodes C if
  - It contains a chain ($D \to E \to F$) and the middle part is in C, or
  - It contains a fork ($D \leftarrow E \to F$) and the middle part is in C, or
  - It contains an inverted fork ($D \to E \leftarrow F$), and the middle part is not in C, nor are any descendants of it
- Two nodes, A and B, are d-separated by a set of nodes C if it **blocks every path** from A to B. Thus

$$A \perp B \mid C$$

- Recall the ignorability assumption

$$Y^0, Y^1 \perp A \mid X$$

Confounders on paths

- A simple DAG: X is a confounder between the relationship between treatment A and outcome Y

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

- A slightly more complicated graph
  - V affects A directly
  - V affects Y indirectly, through W
  - Thus, V is a confounder

$$\begin{tikzcd}
& V \arrow[r] \arrow[dl] & W \arrow[dr] & \\
A \arrow[rrr] & & & Y
\end{tikzcd}$$

Frontdoor paths

- A frontdoor path from A to Y is one that **begins with an arrow emanating out of A**
- We do not worry about frontdoor paths, because they capture effects of treatment
- Example: $A \rightarrow Y$ is a frontdoor path from A to Y

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

- Example: $A \rightarrow Z \rightarrow Y$ is a frontdoor path from A to Y

$$
\begin{tikzcd}
& X \arrow[dl] \arrow[dr] & \\
A \arrow[r] & Z \arrow[r] & Y
\end{tikzcd}
$$

Do not block nodes on the frontdoor path

- If we are interested in the causal effect of A on Y, we should not control for (aka block) Z
  - This is because controlling for Z would be controlling for an affect of treatment

$$
\begin{tikzcd}
 & X \arrow[dl] \arrow[dr] & \\
A \arrow[r] & Z \arrow[r] & Y
\end{tikzcd}
$$

- Causal mediation analysis involves understanding frontdoor paths from A and Y

Backdoor paths

- Backdoor paths from treatment A to outcome Y are paths from A to Y that **travels through arrows going into** A
- Here, $A \leftarrow X \rightarrow Y$ is a backdoor path from A to Y

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

- Backdoor paths confound the relationship between A and Y, so they need to be blocked!
- To sufficiently control for confounding, we must identify a set of variables that block all backdoor paths from treatment to outcome
  - Recall the ignorability: if X is this set of variables, then $Y^0, Y^1 \perp A \mid X$

Criteria

- Next we will discuss two criteria to identify sets of variables that are sufficient to control for confounding
  - Backdoor path criterion: if the graph is known
  - Disjunctive cause criterion: if the graph is not known

Backdoor path criterion

- Backdoor path criterion: a set of variables X is sufficient to control for confounding if
  - It blocks all backdoor paths from treatment to the outcome, and
  - It does not include any descendants of treatment
- Note: the solution X is not necessarily unique

Backdoor path criterion: a simple example

$$\begin{tikzcd} & V \arrow[r] \arrow[dl] & W \arrow[dr] & \\ A \arrow[rrr] & & & Y \end{tikzcd}$$

- There is one backdoor path from A to Y
  - It is not blocked by a collider
- Sets of variables that are sufficient to control for confounding:
  - $\{V\}$, or
  - $\{W\}$, or
  - $\{V, W\}$

Backdoor path criterion: a collider example

$$\begin{tikzcd}
V \arrow[dd] \arrow[dr] & & W \arrow[dl] \arrow[dd] \\
& M & \\
A \arrow[rr] & & Y
\end{tikzcd}$$

- There is one backdoor path from A to Y
  - It is blocked by a collider M, so there is no confounding
- If we condition on M, then it open a path between V and W

$$\begin{tikzcd}
V \arrow[dd] \arrow[dr] \arrow[rr, dash, dashed] & & W \arrow[dl] \arrow[dd] \\
& \boxed{M} & \\
A \arrow[rr] & & Y
\end{tikzcd}$$

- Sets of variables that are sufficient to control for confounding:
  - $\{\}, \{V\}, \{W\}, \{M, V\}, \{M, W\}, \{M, V, W\}$
  - But not $\{M\}$

Backdoor path criterion: a multi backdoor paths example

$$
\begin{tikzcd}
 & V \arrow[dl] \arrow[dr] & \\
Z \arrow[r] & A \arrow[r] & Y \\
 & W \arrow[ul] \arrow[u] &
\end{tikzcd}
$$

- First path: $A \leftarrow Z \leftarrow V \rightarrow Y$
  - No collider on this path
  - So controlling for either Z, V, or both is sufficient
- Second path: $A \leftarrow W \rightarrow Z \leftarrow V \rightarrow Y$
  - Z is a collider
  - So controlling Z opens a path between W and V
  - We can block $\{V\}$, $\{W\}$, $\{Z, V\}$, $\{Z, W\}$, or $\{Z, V, W\}$
- To block both paths, it's sufficient to control for
  - $\{V\}$, $\{Z, V\}$, $\{Z, W\}$, or $\{Z, V, W\}$
  - But not $\{Z\}$ or $\{W\}$

Disjunctive cause criterion

- For many problems, it is difficult to write down accurate DAGs
- In this case, we can use the disjunctive cause criterion: control for all observed causes of the treatment, the outcome, or both
- If there exists a set of observed variables that satisfy the backdoor path criterion, then the variables selected based on the disjunctive cause criterion are sufficient to control for confounding
- Disjunctive cause criterion does not always select the smallest set of variable to control for, but it is conceptually simple

Example

- Observed pre-treatment variables: $\{M, W, V\}$
- Unobserved pre-treatment variables: $\{U_1, U_2\}$
- Suppose we know: W, V are causes of A, Y or both
- Suppose M is not a cause of either A or Y
- Comparing two methods for selecting variables
  1. Use all pre-treatment covariates: $\{M, W, V\}$
  2. Use variables based on disjunctive cause criterion: $\{W, V\}$

Example continued: hypothetical DAG 1

$$\begin{tikzcd} M & V \arrow[l] \arrow[r] \arrow[dl] & W \arrow[dr] & \\ A \arrow[rrr] & & & Y \end{tikzcd}$$

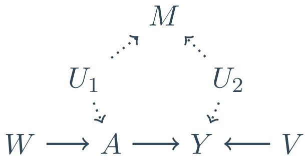
1. Use all pre-treatment covariates: $\{M, W, V\}$
   - Satisfy backdoor path criterion? Yes
2. Use variables based on disjunctive cause criterion: $\{W, V\}$
   - Satisfy backdoor path criterion? Yes

Example continued: hypothetical DAG 2

$$\begin{tikzcd}
V \arrow[dd] \arrow[dr] & & W \arrow[dd] \arrow[dl] \\
& M & \\
A \arrow[rr] & & Y
\end{tikzcd}$$

1. Use all pre-treatment covariates: $\{M, W, V\}$
   - Satisfy backdoor path criterion? Yes
2. Use variables based on disjunctive cause criterion: $\{W, V\}$
   - Satisfy backdoor path criterion? Yes

Example continued: hypothetical DAG 3



1. Use all pre-treatment covariates: $\{M, W, V\}$
   - Satisfy backdoor path criterion? No
2. Use variables based on disjunctive cause criterion: $\{W, V\}$
   - Satisfy backdoor path criterion? Yes

Example continued: hypothetical DAG 4

$$
\begin{tikzcd}
M & & \\
U_1 \arrow[r, dotted] \arrow[d, dotted] & W \arrow[d] & U_2 \arrow[l, dotted] \arrow[dl, dotted] \\
A \arrow[r] & Y & V \arrow[l]
\end{tikzcd}
$$

1. Use all pre-treatment covariates: $\{M, W, V\}$
   - Satisfy backdoor path criterion? No
2. Use variables based on disjunctive cause criterion: $\{W, V\}$
   - Satisfy backdoor path criterion? No

References

- Coursera class: “A Crash Course on Causality: Inferring Causal Effects from Observational Data”, by Jason A. Roy (University of Pennsylvania)
  - https://www.coursera.org/learn/crash-course-in-causality