February 19,24 2025, Lecture Note: Causalality

Dakshitha Mandhalapu, Rohit Reddy Kesireddy

CS 516: Responsible Data Science and Algorithmic Fairness; Spring 2025
Abolfazl Asudeh; www.cs.uic.edu/∼ *asudeh/teaching/archive/cs516spring25/*

Introduction

This lecture centered on the fundamental principles of causality, emphasizing the distinction between observation and intervention. The discussion highlighted how passive observation reflects existing patterns and behaviors in the world, while interventions allow us to assess the impact of hypothetical actions. A key focus was on structural causal models, which provide a formal framework for understanding cause-and-effect relationships.

The analysis covered the construction of structural causal graphs (SCGs) and their role in distinguishing different types of causal structures, including chains, forks, and colliders. The discussion extended to confounding effects, backdoor paths, and the necessity of intervention techniques, such as the do-operator, to estimate causal effects accurately. Furthermore, the importance was emphasized on counterfactual reasoning in fairness and discrimination analysis, demonstrating its relevance in both theoretical and applied contexts.

Structural Causal Graphs

A **structural causal graph** (SCG) is a directed acyclic graph (DAG) used to represent cause-and-effect relationships between variables. Each node represents a variable, and directed edges represent causal influence.

Definition of a Structural Causal Model (SCM)

An SCM consists of:

- A set of **variables** $X_1, X_2, ..., X_d$
- **Structural assignments** $X_i := f_i(P_i, U_i)$ for each variable X_i
- **Exogenous noise variables** U_i, assumed to be independent
- A **causal graph**, which is a DAG where:
  - Parent nodes P_i are the direct causes of X_i
  - Directed edges indicate causal influence

Key Features of Structural Causal Graphs

- They specify **causal relationships**, not just correlations.
- They allow for **interventions** (via the do-operator).
- They help in **identifying confounders and mediators**.

Causal Graph Structures: Chains, Forks, and Colliders

(a) Chain (Mediator)

A **chain** is a sequence of variables where information flows from one variable to another through an intermediate node.

Example: $X \to Z \to Y$

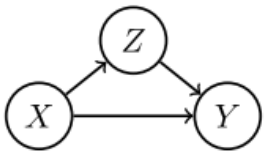

Figure 1: Example of a chain(mediator).

- Here, Z is a **mediator**, meaning it transmits the causal effect of X to Y.
- Mediators should not be controlled for when estimating the **total effect** of X on Y, as they are part of the causal mechanism.

Direct and Indirect Causal Effects in a Chain

- The **direct causal effect** of X on Y, controlling for Z, is given by:

$$\text{Cov}(X, Y|Z)$$

- The **indirect causal effect** of X on Y, mediated through Z, is given by:

$$\text{Cov}(X, Y) - \text{Cov}(X, Y|Z)$$

(b) Fork (Confounder)

A **fork** occurs when a variable Z is a **common cause** of two other variables X and Y.

Example: $X \leftarrow Z \to Y$

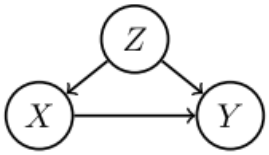

Figure 2: Example of a fork(confounder).

- Here, Z is a **confounder** because it creates a **spurious association** between X and Y.
- **Controlling for confounders** removes this bias and isolates the direct effect of X on Y.

Causal Effect in a Fork Structure

- If $\text{Cov}(X, Y) \neq 0$, it indicates a correlation between X and Y, which could be due to the influence of Z, rather than a direct causal relationship.
- This means that observing X provides information about Y, but it does not imply that X causes Y.
- **Conditioning on Z removes this correlation but does not imply a causal effect.**

Example: Ice Cream Sales and Electricity Bills

Consider an ice cream shop where sales X and electricity bills Y appear correlated, but both are influenced by the external factor of temperature Z:

- **Higher temperatures** lead to **more ice cream sales**.
- **Higher temperatures** also lead to **increased electricity usage** due to air conditioning.
- There is a correlation between ice cream sales and electricity bills, but neither causes the other. Instead, the temperature is the confounder creating this association.

(c) Collider (Selection Bias)

A **collider** occurs when two variables share a **common effect** Z, rather than a common cause.

Example: $X \rightarrow Z \leftarrow Y$

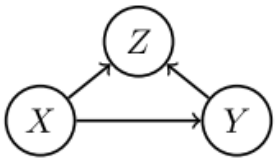

Figure 3: Example of a collider.

- **Conditioning on a collider** creates a **spurious association** between X and Y.
- This phenomenon is known as **Berkson's paradox** or **collider bias**.
- **Example:** If hospital admission Z is influenced by two diseases X and Y, observing admission can create a false correlation between X and Y.

Collider Bias and Measurement Bias

- **Collider bias** occurs when mistakenly conditioning on Z, leading to **spurious correlations** between X and Y.
- **Measurement bias** results from sample selection based on a collider, leading to **sampling bias** in the data.

Example 1: Berkson's Law (Survey Response Bias)

- Suppose a study measures the relationship between **interest in politics** X and **social engagement** Y.
- However, only individuals who **respond to a survey** Z are included in the dataset.
- If both **interest in politics** and **social engagement** influence whether someone responds, then conditioning on respondents introduces a **spurious correlation** between X and Y.
- This creates a **false association** that may not exist in the full population.

Example 2: GRE Scores, Performance, and Admissions

- Consider the relationship between **GRE scores** X and **graduate school performance** Y.
- Admissions decisions Z depend on both **GRE scores** and other factors like **undergraduate GPA**.
- If only **admitted students** are analyzed, then conditioning on Z creates a bias:
  - Among admitted students, those with lower GRE scores may have stronger GPAs, making GRE appear less predictive.
  - This leads to the **incorrect conclusion** that GRE scores do not correlate with performance, when in reality, the full population might show a strong relationship.

Interventions and Causal Effects

The Do-Operator

Interventions allow us to analyze causal effects by **manipulating variables** in the model to isolate causal influence from mere association.

- The **do-operator** $do(X := x)$ represents an intervention setting X to a specific value **independently** of its natural causes.
- This eliminates all **incoming edges** to X in the causal graph, allowing us to estimate causal effects without confounding.
- The new modified distribution after intervention is denoted as $P(Y|do(X = x))$, which is distinct from simple conditional probabilities.

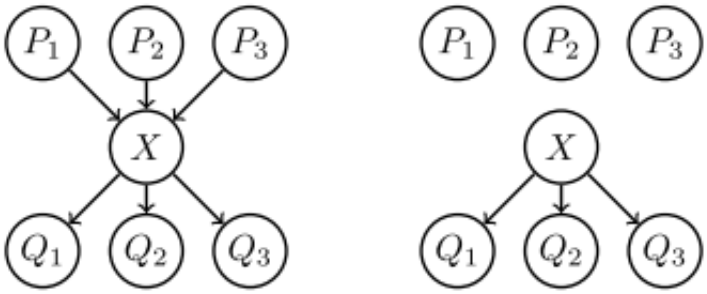

Figure 4: Graph before and after substitution.

Distinguishing Observations from Interventions

- **Observation (Conditional Probability):** $P(Y|X = x)$ tells us how Y varies given X, but does not imply causation due to potential confounding.
- **Intervention (Causal Effect):** $P(Y|do(X = x))$ describes what happens to Y when we actively change X, isolating true causal effects.

Formula for Causal Effect:

$$E[Y|do(X = 1)] - E[Y|do(X = 0)]$$

This difference represents the **average treatment effect (ATE)**, quantifying the expected change in Y due to an intervention on X.

Causal Effects and Confounding

Types of Causal Effects

- **Total Effect (TE):** Measures the overall influence of X on Y, accounting for both direct and indirect paths.
$$TE = E[Y|do(X = 1)] - E[Y|do(X = 0)]$$
- **Direct Effect (DE):** Measures the effect of X on Y while keeping mediators constant.
$$DE = E[Y|do(X = 1), Z = z] - E[Y|do(X = 0), Z = z]$$
- **Indirect Effect (IE):** Captures the impact of X on Y through a mediator Z.
$$IE = TE - DE$$

Backdoor Criterion and Adjustment Formula

Confounding occurs when a third variable, known as a confounder, influences both the treatment variable X and the outcome Y. This creates a spurious association, making it difficult to determine the true causal effect of X on Y. A backdoor path is a non causal path that connects X and Y via a common cause (confounder). These paths create bias in observational studies. To estimate the true causal effect of X on Y, we must block all backdoor paths.

Backdoor Criterion

The backdoor criterion provides a method to identify confounding variables that must be controlled. A set of variables Z satisfies the backdoor criterion if:

- No node in Z is a descendant of X. We should not control for variables that are affected by X
- Z blocks all backdoor paths between X and Y. Conditioning on Z prevents non causal influences from affecting our estimate of P(Y | do(X))

Adjustment Formula

If a set of variables Z satisfies the backdoor criterion, we can estimate the causal effect using the adjustment formula:

$$P(Y|do(X)) = \sum_z P(Y|X, Z = z)P(Z = z)$$

This allows us to express the interventional probablity P(Y | do(X)) using observational data.

Example: Smoking and Lung Cancer

Consider the relationship between smoking (X) and lung cancer (Y). Suppose a genetic factor (Z) influences both smoking behavior and susceptibility to lung cancer.

- If we calculate P(Y | X), we might overestimate the effect due to genetic predisposition.
- By controlling for Z we remove the bias introduced by confounding.

Counterfactuals and Potential Outcomes

Counterfactual Thinking

Counterfactuals ask What would have happened if X had been different?Counterfactual analysis is crucial in fairness and discrimination studies.

Example:

- A company rejects a job applicant.
- Counterfactual question: Would the applicant have been accepted if they belonged to a different demographic group?

Potential Outcomes Framework

The Potential Outcomes Model defines counterfactuals formally. For a binary treatment $X \in \{0, 1\}$, each unit i has two potential outcomes:

- $Y_i(1)$: The outcome if $X = 1$ (e.g., admitted to a program).
- $Y_i(0)$: The outcome if $X = 0$ (e.g., rejected from a program).

However, we only observe one of these outcomes for each individual. The causal effect is:

$$\text{Average Treatment Effect (ATE)} = E[Y(1)] - E[Y(0)]$$

Key Assumptions

- Stable Unit Treatment Value Assumption : The treatment of one unit does not affect the outcome of another.
- Ignorability: The treatment assignment is independent of potential outcomes, given some covariates.

Counterfactual Fairness

A decision rule Y satisfies Counterfactual Fairness if:

$$P(Y_A|X, A = a) = P(Y_A|X, A = a')$$

This means that changing a protected attribute (e.g., race, gender) should not alter the prediction.

Example:

- A university's admission decision should remain the same regardless of whether an applicant is Black or White, male or female, disabled or not given the same qualifications.

Structural Discrimination and Policy Implications

What is Structural Discrimination?

Structural discrimination refers to systemic biases embedded in policies, institutions, and societal norms. Unlike individual bias, which results from explicit discrimination, structural discrimination operates indirectly through existing social and economic structures.

Example:

- A university systematically underfunds departments with more female students.
- Women applying to graduate school are overrepresented in less-funded departments, leading to lower acceptance rates.

Even if no explicit discrimination occurs during admissions, the structural disadvantage leads to unequal outcomes.

Examples of Structural Discrimination

Case 1: Griggs v. Duke Power Co. (1971)

- A company required a high school diploma for employment.
- The rule disproportionately excluded Black applicants.
- The Supreme Court ruled that "neutral policies" can be discriminatory if they have disparate impact.

Case 2: Harvard Admissions Case

- Plaintiffs argued that Asian-American applicants were systematically rated lower on subjective personality traits.
- Even though no explicit racial quota existed, structural biases in evaluation criteria led to racial disparities.

Policy Implications

To mitigate structural bias, policies must be causally fair:

1. Identifying structural barriers through causal analysis.
2. Redesigning selection mechanisms to reduce systemic disadvantage.
3. Ensuring equal access to opportunities (e.g., improving STEM representation among women).

Conclusion

Causal fairness requires moving beyond correlation to understand the true impact of decisions. Key takeaways:

- Backdoor Criterion helps identify confounders that must be controlled.
- Counterfactual Fairness ensures decisions are independent of protected attributes.
- Structural Discrimination highlights the importance of system-wide interventions to promote fairness.

Understanding causality allows for better decision-making in AI systems, hiring practices, college admissions, and public policy. By integrating causal reasoning into fairness evaluations, we can move toward more equitable and transparent decision-making frameworks.