Introduction to Causal Inference

Ruoxu Tan

Outline

1. Background and History
2. Neyman–Rubin Potential Outcomes
3. Pearl's Causal Graphs

Why Study Causality?

What caused an event? What would happen if we took a particular action?

- **Medicine.** Does smoking cause lung cancer? Does quitting lower the risk?
- **Economics.** Does one more year of schooling raise future earnings?
- **Education.** Does a new teaching method improve test scores?

Causal inference asks how changing one factor affects an outcome.

Outline

1. Background and History
   - Examples
   - History
2. Potential Outcomes
   - Potential Outcomes and Causal Effects
   - Randomized Experiments
   - Observational Studies
3. Causal Graphs
   - Causal Graph Basics
   - SCMs, Interventions, and Counterfactuals
   - Do-Calculus, Backdoor Criterion, and Ignorability
   - Causal Discovery

The Texas Sharpshooter Fallacy

A Texan fires a series of shots at a wall at random.

He **paints a target around the tightest cluster** and claims perfect aim.

The Texas Sharpshooter Fallacy

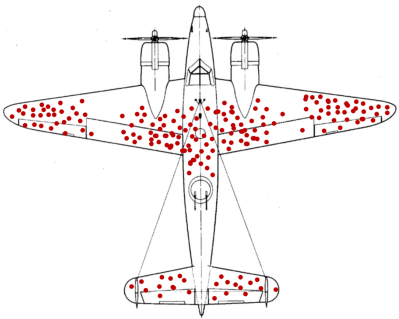
A Texan fires a series of shots at a wall at random.

He **paints a target around the tightest cluster** and claims perfect aim.

- He chooses where to put the target **after** seeing the bullet holes.
- Even purely random shots can form striking clusters.

Picking an explanation after seeing the data can mislead.

Survivorship Bias in Returning Aircraft

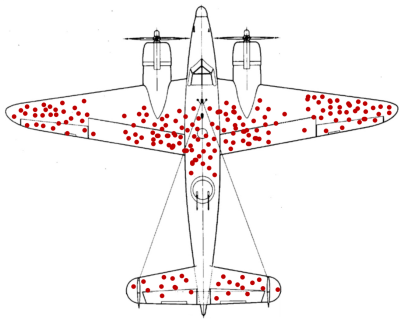


Bullet holes on returning aircraft (schematic)

Where should we add armor?

Where there are more bullet holes, or fewer?

Survivorship Bias in Returning Aircraft



Bullet holes on returning aircraft (schematic)

Where should we add armor?
Where there are more bullet holes, or fewer?

- We observe only the planes that **returned**.
- Planes hit in critical areas may never enter the sample.
- If areas are equally likely to be hit, **those with fewer holes** may need more armor.

The process that determines what we observe can bias the data.

Survivorship Bias in Everyday Life

- **Dropping out to start a business.** Gates and Zuckerberg made headlines. Failed ventures rarely do.
- **Fund rankings.** Excluding funds that have shut down can overstate past returns.
- **Longevity.** Some people who live to old age smoke or drink. Others with the same habits die younger.
- **Old buildings.** Buildings that still stand omit the many that collapsed or were demolished.

Ice Cream and Drowning

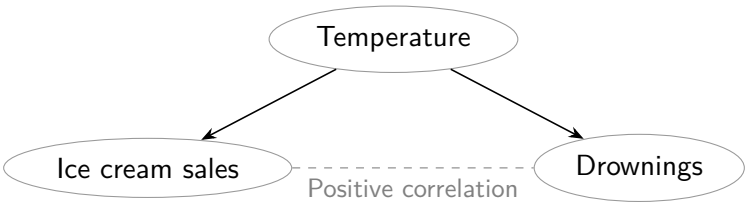
Ice cream sales and drownings both rise in summer.

Does **selling more ice cream cause more drownings**?

Ice Cream and Drowning

Ice cream sales and drownings both rise in summer.

Does **selling more ice cream cause more drownings**?

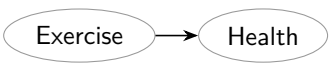


- Temperature may affect ice cream consumption and swimming. It is a **confounder**.
- **Correlation alone cannot establish causation.**

One Association, Several Causal Explanations

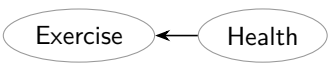
People who exercise more tend to be healthier. What could explain this?

Exercise affects health



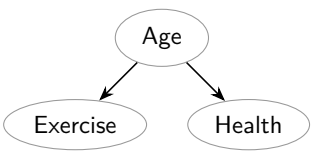
Exercise may improve health.

Health affects exercise



Healthy people may exercise more easily.

A common cause



Age may affect both.

A causal explanation specifies the directions of the arrows.

Which Kidney Stone Treatment Works Better?

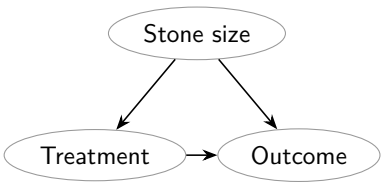
Compare open surgery (A) with a less invasive treatment (B).

The table shows successes out of patients treated and the success rate.

	A	B
Overall	273/350 (78.0%)	**289/350 (82.6%)**
Small stones	**81/87 (93.1%)**	234/270 (86.7%)
Large stones	**192/263 (73.0%)**	55/80 (68.8%)

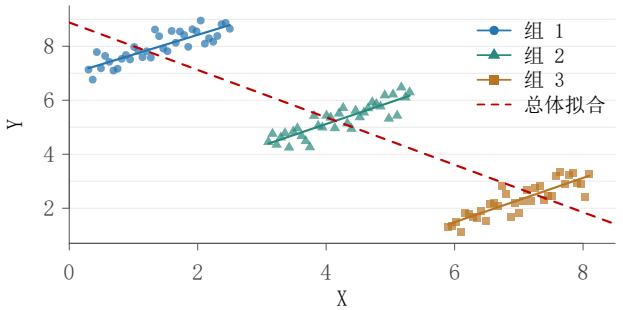
B succeeds more often overall, but A succeeds more often in each subgroup.

- Large stones: $263/350 = 75.1\%$ in A and $80/350 = 22.9\%$ in B.
- Larger stones are harder to treat. The groups have a **different case mix**.
- Overall rates average the subgroup rates with **different weights**.
- **Stone size** is a confounder.



One possible causal explanation

Simpson's Paradox



- Separate fits have positive slopes.
- The pooled fit has a negative slope.
- Pooling can reverse the trend in the same data.

Confounding or different weights can reverse the overall trend.

Lessons from the Examples

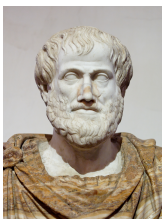
Causal inference is difficult for several reasons.

- The Texas sharpshooter and survivorship bias illustrate how **selective observation** can mislead.
- Ice cream and drowning, and Simpson's paradox, show how **confounding** can mislead.

Outline

1. Background and History
   - Examples
   - History
2. Potential Outcomes
   - Potential Outcomes and Causal Effects
   - Randomized Experiments
   - Observational Studies
3. Causal Graphs
   - Causal Graph Basics
   - SCMs, Interventions, and Counterfactuals
   - Do-Calculus, Backdoor Criterion, and Ignorability
   - Causal Discovery

Aristotle's Four Causes



Aristotle
384–322 BCE

Aristotle proposed **four kinds of causes**.

Each answers a different question about why a statue exists.

- **Material** What is it made of? Marble.
- **Formal** What is its form? The statue's shape and structure.
- **Efficient** What made it? The sculptor's work.
- **Final** What is its purpose? Commemoration, decoration, or another use.

Hume and Necessary Connection



David Hume
1711–1776

Hume asked what experience can tell us about causation and induction.

- A billiard ball starts moving after impact. We see this sequence repeat.
- Can we observe the **necessary connection** between cause and effect?
- Inferring that the future will resemble the past is the **problem of induction**.
- Observed regularity and causal necessity raise different questions.

Fisher and Randomized Experiments



Ronald Aylmer Fisher
1890–1962

Fisher developed **experimental design** and **analysis of variance** in agricultural trials.

Randomization Assign treatments by chance to make comparisons fair.

Replication Repeat comparisons across plots to measure variability.

Blocking Group similar plots and randomize within blocks to improve precision.

Neyman and Potential Outcomes



Jerzy Neyman
1894–1981

Neyman used **potential outcomes** in his 1923 work on agricultural experiments.

- $Y_i(1)$ is the yield of plot i **with fertilizer**.
- $Y_i(0)$ is the yield of the same plot **without fertilizer**.

The **causal effect** of fertilizer on the plot is

$$\tau_i = Y_i(1) - Y_i(0),$$

the difference between two possible outcomes for the same plot.

Rubin and Observational Studies



Donald B. Rubin
1943–

In the 1970s, Rubin developed a potential outcomes framework for randomized and nonrandomized treatments.

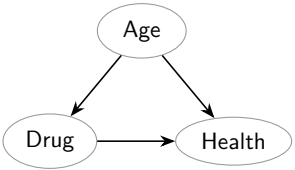
- Decisions about medical treatment or schooling are often nonrandom.
- Treated and control units may come from **different distributions**.
- When can we infer causal effects from observational data?

Pearl's Causal Graphs



Judea Pearl
1936–

- Represent causal mechanisms with graphs and structural equations.
- Use explicit assumptions to assess identification.
- Use do-calculus to reason about interventions.



The graph encodes assumptions about causal mechanisms.

From Philosophy to Statistical Frameworks

Period	Figure	Ideas and methods
Ancient Greece	Aristotle	Four kinds of causal explanation.
18th century	Hume	Regularity, necessary connection, and induction.
1920s–1930s	Fisher	Randomized experimental design and inference.
1923	Neyman	Potential outcomes for treatment effects in experiments.
1970s onward	Rubin	Potential outcomes in experiments and observational studies.
1990s onward	Pearl	Causal graphs, structural models, and do-calculus.

The Fundamental Problem of Causal Inference

Student	Received tutoring?	Score with tutoring $Y(1)$	Score without tutoring $Y(0)$
Li	Yes	85 (observed)	**?**
Wang	No	**?**	75 (observed)

- $85 - 75$ compares two students. A before-and-after comparison also changes time and experience.
- For each student, **only one potential outcome is observed**.

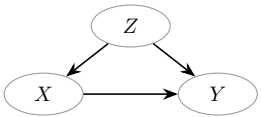
Potential Outcomes and Causal Graphs

Neyman–Rubin Potential Outcomes

$$Y(1), \quad Y(0), \qquad Y(1) - Y(0)$$

- Specify which potential outcomes to compare.
- Describe how treatment is assigned.
- Estimate effects under assumptions such as ignorability.

Pearl's Graphs and Structural Models



- Specify causal mechanisms and assumptions.
- Write interventions as $\mathrm{do}(Z = z)$.
- Ask when data can identify causal direction.

The two frameworks express the same core concepts in different ways.

Outline

1. Background and History
   - Examples
   - History
2. Potential Outcomes
   - Potential Outcomes and Causal Effects
   - Randomized Experiments
   - Observational Studies
3. Causal Graphs
   - Causal Graph Basics
   - SCMs, Interventions, and Counterfactuals
   - Do-Calculus, Backdoor Criterion, and Ignorability
   - Causal Discovery

Observation, Intervention, and Counterfactuals

Does exercise improve health?

Observation How does health differ between program participants and nonparticipants?

Compare the two groups as observed.

Intervention How would health change if the same population were assigned to exercise or to continue its usual activities?

Compare outcomes under two assignments.

Counterfactual Li joined the program. How healthy would he be without it?

Ask about the same person under a different treatment.

Research Question and Notation

How does a 12-week program of three 30-minute brisk walks per week affect community adults' health at week 12 compared with usual activity?

Symbol	Meaning
i	Unit (an adult in the study)
$Z \in \{0, 1\}$	Treatment (exercise program participation)
Y	Outcome (health score at week 12)
X	Pretreatment covariate (e.g., baseline health score)

Z is random, and $z \in \{0, 1\}$ is fixed.

Observing $Z = z$ and **setting Z to z** are different operations.

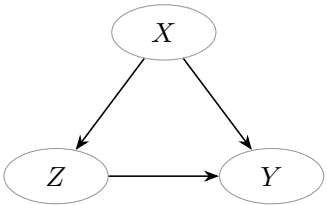
A Simple Structural Model

Suppose health scores are generated as follows.

$$
\begin{aligned}
X &\sim \text{Unif}(-10, 10), \\
Z &= \mathbf{1}\{X > 0\}, \\
Y &= 60 + 10Z + X.
\end{aligned}
$$

- X is baseline health relative to a reference score.
- People with better baseline health join the program.
- Baseline health and exercise determine later health.

The equations represent assumed causal mechanisms.



A common cause of treatment and outcome.

Intervention and Potential Outcomes

Set $Z := z$, keeping the distribution of X and the outcome mechanism fixed.
The **potential outcome** $Y(z)$ is the outcome under z.

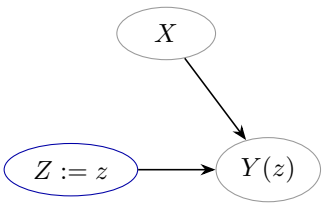
$$Y(z) = 60 + 10z + X, \qquad z \in \{0, 1\}.$$
$$Y(0) = 60 + X, \qquad Y(1) = 70 + X.$$

Intervention and Potential Outcomes

Set $Z := z$, keeping the distribution of X and the outcome mechanism fixed.
The **potential outcome** $Y(z)$ is the outcome under z.

$$Y(z) = 60 + 10z + X, \qquad z \in \{0, 1\}.$$
$$Y(0) = 60 + X, \qquad Y(1) = 70 + X.$$

- At fixed z, $Y(z)$ varies with X.
- Both potential outcomes share the **same X**.
- At $x = 4$, the scores are 64 and 74. With exercise, $Y = 74$.



The intervention sets Z.

Potential Outcomes and Conditional Distributions

In this model, $\mathbb{E}(X) = 0$ and

$$\mathbb{E}(X \mid Z = 1) = 5, \qquad \mathbb{E}(X \mid Z = 0) = -5.$$

Treatment	Intervention mean $\mathbb{E}\{Y(z)\}$	Observed group mean $\mathbb{E}(Y \mid Z = z)$
$z = 0$	60	55
$z = 1$	70	75
Difference	10	20

Potential Outcomes and Conditional Distributions

In this model, $\mathbb{E}(X) = 0$ and

$$\mathbb{E}(X \mid Z = 1) = 5, \qquad \mathbb{E}(X \mid Z = 0) = -5.$$

Treatment	Intervention mean $\mathbb{E}\{Y(z)\}$	Observed group mean $\mathbb{E}(Y \mid Z = z)$
$z = 0$	60	55
$z = 1$	70	75
Difference	10	20

- Intervention keeps the target population's distribution of X. Conditioning selects groups with different X distributions.
- **$Y(z)$ and $Y \mid Z = z$ have different distributions**. Why?

Stable Unit Treatment Value Assumption (SUTVA)

1. **No interference.** Unit i's outcome depends only on its own treatment. If $z_i = z_i'$, then $Y_i(\mathbf{z}) = Y_i(\mathbf{z}')$, where $\mathbf{z} = (z_1, \ldots, z_n)$ is the treatment vector.

2. **Consistency.** The observed outcome equals the potential outcome under the treatment received.

$$Z_i = z \quad \Longrightarrow \quad Y_i = Y_i(z).$$

Therefore the observed outcome is

$$\boldsymbol{Y_i = Y_i(Z_i) = Z_i Y_i(1) + (1 - Z_i) Y_i(0).}$$

- The intervention in $Y_i(z)$ must match the treatment received.
- This is a condition on outcomes. Estimator consistency concerns convergence as the sample size grows.

Potential Outcomes and the Fundamental Problem

Full table (illustrative)

Unit	$Y_i(0)$	$Y_i(1)$
Li	64	74
Wang	56	66
Zhang	67	77
Zhao	58	68

Observed data

Unit	Z_i	$Y_i(0)$	$Y_i(1)$
Li	1	?	74
Wang	0	56	?
Zhang	1	?	77
Zhao	0	58	?

- Each unit reveals only one potential outcome. The other remains unobserved.
- Time and experiences differ between before and after measurements, so these generally cannot fill the missing cell.

Individual and Average Effects

Individual effect Compare both outcomes for the same unit.

$$\tau_i = Y_i(1) - Y_i(0).$$

Finite-population effect Average over the specified n units.

$$\tau_n = \frac{1}{n}\sum_{i=1}^{n}\{Y_i(1) - Y_i(0)\}.$$

Population effect Average over the target population distribution.

$$\tau = \mathbb{E}\{Y(1) - Y(0)\}.$$

Average Treatment Effects

$$\text{ATE} \quad \tau = \mathbb{E}\{Y(1) - Y(0)\},$$
$$\text{ATT} \quad \tau_{\text{ATT}} = \mathbb{E}\{Y(1) - Y(0) \mid Z = 1\},$$
$$\text{CATE} \quad \tau(x) = \mathbb{E}\{Y(1) - Y(0) \mid X = x\}.$$

- **ATE.** What is the average effect in the target population?
- **ATT.** What is the average effect for actual participants?
- **CATE.** What is the average effect at a given baseline health score?

By iterated expectation, $\tau = \mathbb{E}\{\tau(X)\}$. The CATE averages over **people with the same** X.

Outline

1. Background and History
   - Examples
   - History
2. Potential Outcomes
   - Potential Outcomes and Causal Effects
   - Randomized Experiments
   - Observational Studies
3. Causal Graphs
   - Causal Graph Basics
   - SCMs, Interventions, and Counterfactuals
   - Do-Calculus, Backdoor Criterion, and Ignorability
   - Causal Discovery

Completely Randomized Experiment

Randomly assign n_1 of the n participants to exercise. The remaining $n_0 = n - n_1$ continue their usual activity, with $n_1, n_0 > 0$.

- All groups of size n_1 are equally likely, so each person has treatment probability n_1/n.

$$\mathbb{E}(Z_i) = P(Z_i = 1) = \frac{n_1}{n}, \qquad \mathbb{E}(1 - Z_i) = \frac{n_0}{n}.$$

- **Independence and positivity**

$$\{Y(0), Y(1)\} \perp Z, \qquad 0 < P(Z = 1) < 1.$$

Selecting two of four people gives six assignments, each with probability $1/6$.

(1)(1)(0)(0) (1)(0)(1)(0) (1)(0)(0)(1) (0)(1)(1)(0) (0)(1)(0)(1) (0)(0)(1)(1)

Estimating the ATE by a Difference in Means

The target is the average treatment effect (ATE) of the program on health scores.

$$\tau = \mathbb{E}\{Y(1) - Y(0)\}.$$

Estimating the ATE by a Difference in Means

The target is the average treatment effect (ATE) of the program on health scores.

$$\tau = \mathbb{E}\{Y(1) - Y(0)\}.$$

We observe one outcome Y_i per person and compute a sample mean for each group.

$$\overline{Y}_1 = \frac{1}{n_1}\sum_{i=1}^{n} Z_i Y_i \qquad \text{(Treated)},$$

$$\overline{Y}_0 = \frac{1}{n_0}\sum_{i=1}^{n} (1 - Z_i) Y_i \quad \text{(Non-treated)}.$$

The difference in sample means estimates the ATE.

$$\boldsymbol{\widehat{\tau} = \overline{Y}_1 - \overline{Y}_0.}$$

Why Is the Difference in Means Unbiased?

Since $\{\boldsymbol{Y(0), Y(1)}\} \perp \boldsymbol{Z}$, we have

$$\begin{aligned}\mathbb{E}(\overline{Y}_1) &= \frac{1}{n_1}\sum_{i=1}^{n}\mathbb{E}\{Y_i(1)\}\mathbb{E}(Z_i) \\ &= \frac{1}{n_1}\sum_{i=1}^{n}\mathbb{E}\{Y_i(1)\}\frac{n_1}{n} = \mathbb{E}\{Y(1)\}.\end{aligned}$$

Likewise, $\mathbb{E}(1 - Z_i) = n_0/n$, so

$$\mathbb{E}(\overline{Y}_0) = \frac{1}{n_0}\sum_{i=1}^{n}\mathbb{E}\{Y_i(0)\}\frac{n_0}{n} = \mathbb{E}\{Y(0)\}.$$

Subtracting gives

$$\boldsymbol{\mathbb{E}(\hat{\tau}) = \mathbb{E}(\overline{Y}_1) - \mathbb{E}(\overline{Y}_0) = \tau.}$$

Outline

1. Background and History
   - Examples
   - History
2. Potential Outcomes
   - Potential Outcomes and Causal Effects
   - Randomized Experiments
   - Observational Studies
3. Causal Graphs
   - Causal Graph Basics
   - SCMs, Interventions, and Counterfactuals
   - Do-Calculus, Backdoor Criterion, and Ignorability
   - Causal Discovery

Definition, Identification, and Estimation

Definition Which causal quantity do we want? For example, $\tau = \mathbb{E}\{Y(1) - Y(0)\}$.

Identification Under the assumptions, does the observed data distribution uniquely determine the quantity?

Estimation Compute an estimate $\widehat{\tau}$ from $(X_i, Z_i, Y_i)_{i=1}^n$.



Larger samples reduce sampling uncertainty. Identification still depends on assumptions that can be hard to verify.

Selection Bias in Observational Studies

In observational studies, $\{\boldsymbol{Y(0), Y(1)}\} \perp \boldsymbol{Z}$ often fails.

Let $\Delta = \mathbb{E}(Y \mid Z = 1) - \mathbb{E}(Y \mid Z = 0)$. By consistency,

$$\begin{aligned}\Delta &= \mathbb{E}\{Y(1) \mid Z = 1\} - \mathbb{E}\{Y(0) \mid Z = 0\} \\ &= \underbrace{\mathbb{E}\{Y(1) - Y(0) \mid Z = 1\}}_{\tau_{\mathrm{ATT}}} \\ &\quad + \underbrace{\mathbb{E}\{Y(0) \mid Z = 1\} - \mathbb{E}\{Y(0) \mid Z = 0\}}_{\text{selection bias}}.\end{aligned}$$

- The bias term compares mean health scores if **neither group joined the program**.
- People who join may already differ in health, diet, or health awareness.

Conditional Ignorability

Conditional ignorability assumes

$$\{Y(1), Y(0)\} \perp Z \mid X.$$

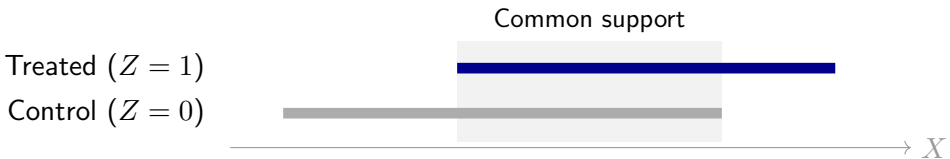
Given X, treatment assignment conveys no further information about potential outcomes.

- After adjusting for baseline health, diet may still affect both program participation and later health. This would violate ignorability.
- Ignorability allows an association between Y and Z given X. Covariate balance alone cannot establish independence of unobserved potential outcomes.
- Think of this as controlling for X.

Positivity

The **propensity score** is $e(x) = P(Z = 1 \mid X = x)$.

Positivity requires $0 < e(X) < 1$ almost surely in the target population.



- Even when positivity holds, propensity scores near 0 or 1 can make estimates unstable.

Deriving the Adjustment Formula

Under SUTVA, ignorability, and positivity, for $z = 0, 1$,

$$\begin{aligned}
\mathbb{E}\{Y(z)\} &= \mathbb{E}_X\left[\mathbb{E}\{Y(z) \mid X\}\right] && \text{(iterated expectation)} \\
&= \mathbb{E}_X\left[\mathbb{E}\{Y(z) \mid Z = z, X\}\right] && \text{(ignorability)} \\
&= \mathbb{E}_X\left[\mathbb{E}(Y \mid Z = z, X)\right] && \text{(consistency).}
\end{aligned}$$

Positivity ensures data support for both treatments across the target population's covariate values.

Write $m_z(x) = \mathbb{E}(Y \mid Z = z, X = x)$. Then

$$\boldsymbol{\tau = \mathbb{E}_X\{m_1(X) - m_0(X)\}.}$$

This adjustment formula is the g-formula.

Common ATE Estimators

Under these assumptions, three common methods estimate $\tau = \mathbb{E}\{Y(1) - Y(0)\}$.

- Outcome regression via the g-formula
- Inverse probability weighting using the propensity score
- Doubly robust estimation, which combines the two

Outcome Regression

With $m_z(x) = \mathbb{E}(Y \mid Z = z, X = x)$, the ATE $\tau = \mathbb{E}\{m_1(X) - m_0(X)\}$ can be estimated from two outcome regressions.

1. Fit $\widehat{m}_1(x)$ in the treatment group and $\widehat{m}_0(x)$ in the control group.
2. Average the prediction differences over the target population.

$$\widehat{\tau}_{\text{reg}} = \frac{1}{n} \sum_{i=1}^{n} \{\widehat{m}_1(X_i) - \widehat{m}_0(X_i)\}.$$

- Misspecified models for the conditional means can bias the estimate.
- The underlying identity is $m_z(x) = \mathbb{E}_X[\mathbb{E}\{Y(z) \mid X\}]$.

Regression Coefficients and the ATE

Assume identification and a correctly specified conditional mean model. Consider a scalar X.

Additive linear model

$$m_z(x) = \alpha + \beta z + \gamma x \quad \Longrightarrow \quad \tau(x) = \beta, \quad \tau = \beta.$$

Model with a treatment interaction

$$m_z(x) = \alpha + \beta z + \gamma x + \delta z x \quad \Longrightarrow \quad \tau(x) = \beta + \delta x.$$

This gives

$$\tau = \beta + \delta \mathbb{E}(X), \qquad \widehat{\tau} = \widehat{\beta} + \widehat{\delta}\,\overline{X}.$$

Propensity Scores, Stratification, and Matching

With $e(x) = P(Z = 1 \mid X = x)$, ignorability implies

$$\{\boldsymbol{Y(1), Y(0)}\} \perp\!\!\!\perp \boldsymbol{Z} \mid \boldsymbol{e(X)}.$$

- **Stratification.** Group units by estimated propensity score, compare within strata, and pool the results.
- **Matching.** Pair treated and control units with similar scores.
- **Weighting.** Weight units by the inverse probability of the treatment they received.

Inverse Probability Weighting

Under SUTVA, ignorability, and positivity,

$$\mathbb{E}\left\{\frac{ZY}{e(X)}\right\} = \mathbb{E}\left[\frac{\mathbb{E}\{ZY(1) \mid X\}}{e(X)}\right] = \mathbb{E}\left[\frac{e(X)\mathbb{E}\{Y(1) \mid X\}}{e(X)}\right] = \mathbb{E}\{Y(1)\}.$$

The control mean follows analogously. Let $\widehat{e}_i = \widehat{e}(X_i)$. The IPW estimator is

$$\widehat{\tau}_{\text{IPW}} = \frac{1}{n}\sum_i \left\{\frac{Z_i Y_i}{\widehat{e}_i} - \frac{(1-Z_i)Y_i}{1-\widehat{e}_i}\right\}.$$

Let $w_{1i} = Z_i/\widehat{e}_i$ and $w_{0i} = (1-Z_i)/(1-\widehat{e}_i)$. Normalizing within groups gives

$$\widehat{\tau}_{\text{HIPW}} = \frac{\sum_i w_{1i}Y_i}{\sum_i w_{1i}} - \frac{\sum_i w_{0i}Y_i}{\sum_i w_{0i}}.$$

Doubly Robust Estimation

Let $\widehat{m}_{zi} = \widehat{m}_z(X_i)$ and $\widehat{e}_i = \widehat{e}(X_i)$.

$$\widehat{\tau}_{\text{AIPW}} = \frac{1}{n}\sum_{i=1}^{n}\Bigg[\underbrace{\widehat{m}_{1i} - \widehat{m}_{0i}}_{\text{outcome regression}} + \underbrace{\frac{Z_i(Y_i - \widehat{m}_{1i})}{\widehat{e}_i} - \frac{(1-Z_i)(Y_i - \widehat{m}_{0i})}{1-\widehat{e}_i}}_{\text{weighted residual correction}}\Bigg].$$

- Predict the conditional mean under each treatment.
- Use observed minus predicted outcomes to correct the regression estimate.
- This combination is called augmented inverse probability weighting (AIPW).

Double Robustness

For $\mu_1 = \mathbb{E}\{Y(1)\}$, let $\widetilde{m}_1, \widetilde{e}$ denote possibly misspecified parametric models.

$$\widetilde{\mu}_1 = \mathbb{E}\left[\widetilde{m}_1(X) + \frac{Z\{Y - \widetilde{m}_1(X)\}}{\widetilde{e}(X)}\right].$$

Under identification, $0 < \widetilde{e}(X) < 1$, and existence of the relevant expectations,

$$\widetilde{\mu}_1 - \mu_1 = \mathbb{E}\left[\underbrace{\frac{e(X) - \widetilde{e}(X)}{\widetilde{e}(X)}}_{A_1} \underbrace{\{m_1(X) - \widetilde{m}_1(X)\}}_{A_2}\right].$$

- Correct propensity scores give $A_1 = 0$. Correct outcome regression gives $A_2 = 0$. The control mean is analogous.
- ATE estimation is consistent if **the propensity score model or both outcome models are correct**.

Potential Outcomes Summary

Step	Key question
Definition	Which potential outcomes and causal effect are of interest?
Assumptions	Is treatment well defined? Is there interference? Is assignment ignorable?
Identification	How can we express the effect in terms of the observed distribution?
Estimation	How can we estimate it by stratification, regression, weighting, or a doubly robust method?

Further Topics

- Double/debiased machine learning extends bias correction to broader settings.
- Continuous or functional treatments
- Unmeasured confounding and violations of ignorability
- Other causal targets, such as quantiles and distribution functions

Outline

1. Background and History
   - Examples
   - History
2. Potential Outcomes
   - Potential Outcomes and Causal Effects
   - Randomized Experiments
   - Observational Studies
3. Causal Graphs
   - Causal Graph Basics
   - SCMs, Interventions, and Counterfactuals
   - Do-Calculus, Backdoor Criterion, and Ignorability
   - Causal Discovery

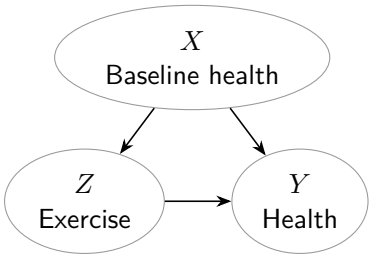
Why Use Causal Graphs?

We used ignorability to identify the average causal effect.

$$\{Y(0), Y(1)\} \perp Z \mid X.$$

- Which variables belong in X?
- How can we discover causal relationships?
- How can we represent unobserved common causes?

Causal graphs make **assumptions about causal structure** explicit.



Directed Acyclic Graphs

A **directed acyclic graph** (DAG) has directed edges and no directed cycles.

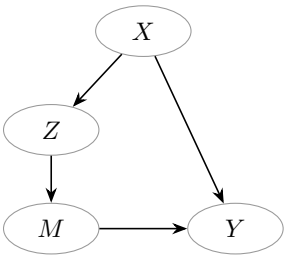
Nodes and edges Nodes represent variables. An edge $A \to B$ makes A a direct cause of B.

Parent and child In $A \to B$, A is the parent and B is the child.

Path A sequence of distinct nodes joined by edges, **regardless of edge direction**.

Directed path Every arrow points forward along the path.

Ancestor and descendant A directed path from A to B makes A an ancestor of B and B a descendant of A.



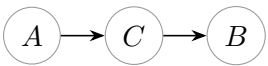
X, M are parents of Y.

M, Y are descendants of Z.

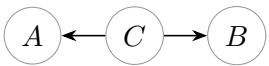
Three Basic Path Structures

On A–C–B, the effect of conditioning on C depends on the arrows.

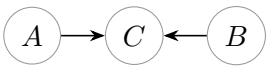
Chain



Common Cause



Collider



At a collider, both adjacent arrows on the path point into the node.

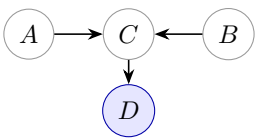
Path Blocking and d-Separation

Let S exclude the endpoints. A path is **blocked** by S if and only if

1. some **noncollider** on the path is in S, or
2. some **collider** on the path and all its descendants are outside S.

If S blocks **every path** between A and B, then S **d-separates** them.

$$A \perp_G B \mid S.$$

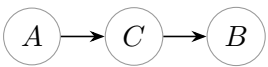


- $S = \varnothing$ blocks $A \to C \leftarrow B$.
- $S = \{C\}$ opens $A \to C \leftarrow B$.
- $S = \{D\}$ also opens the path by conditioning on a collider's descendant.

Three Basic Path Structures

On A–C–B, the effect of conditioning on C depends on the arrows.

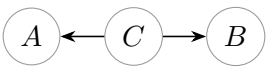
Chain



Open without C.

Blocked given C.

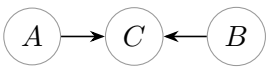
Common Cause



Open without C.

Blocked given C.

Collider



Blocked without C.

Open given C.

At a collider, both adjacent arrows point into the node. Open paths can carry association. Under the Markov property, blocking all paths implies conditional independence.

The Markov Property

The Markov property links graphical separation to conditional independence.

- The **global Markov property** holds when, for disjoint node sets A, B, S,

$$A \perp_G B \mid S \quad \Longrightarrow \quad A \perp B \mid S.$$

- For a joint density, the property is equivalent to the factorization

$$p(v_1, \ldots, v_d) = \prod_{j=1}^{d} p(v_j \mid \mathrm{pa}_j),$$

  where pa_j gives the values of the parents of V_j.

- For the chain $A \to C \to B$, the property gives

$$p(a, c, b) = p(a)p(c \mid a)p(b \mid c), \qquad A \perp B \mid C.$$

- **Faithfulness** gives the reverse implication. Under both assumptions, d-separation and conditional independence are equivalent.

Outline

1. Background and History
   - Examples
   - History
2. Potential Outcomes
   - Potential Outcomes and Causal Effects
   - Randomized Experiments
   - Observational Studies
3. Causal Graphs
   - Causal Graph Basics
   - SCMs, Interventions, and Counterfactuals
   - Do-Calculus, Backdoor Criterion, and Ignorability
   - Causal Discovery

Structural Causal Models

A structural causal model (SCM) specifies

1. **Structural equations** mapping parents and noise to each variable,

$$V_j := f_j(\mathrm{PA}_j, N_j), \qquad j = 1, \ldots, d.$$

2. A **joint noise distribution** in which $N_1, \ldots, N_d$ are mutually independent.

3. An **acyclic structure** so the equations determine the joint distribution in causal order.

Structural Causal Models

A structural causal model (SCM) specifies

1. **Structural equations** mapping parents and noise to each variable,

$$V_j := f_j(\mathrm{PA}_j, N_j), \qquad j = 1, \ldots, d.$$

2. A **joint noise distribution** in which $N_1, \ldots, N_d$ are mutually independent.
3. An **acyclic structure** so the equations determine the joint distribution in causal order.

- The resulting distribution is Markov with respect to the DAG.
- With only observed nodes, independent noise **rules out** unobserved common causes.

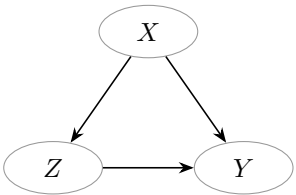
An SCM for Exercise and Health

The three noise variables are mutually independent.

$$N_X \sim \text{Unif}(-10, 10), \quad N_Z \sim \text{Unif}(0, 1), \quad N_Y \sim N(0, 1).$$

The structural equations are

$$\begin{aligned} X &:= N_X, \\ Z &:= \mathbf{1}\{N_Z < e(X)\}, \qquad e(x) = \frac{1}{2} + \frac{x}{40}. \\ Y &:= 60 + 10Z + X + N_Y, \end{aligned}$$



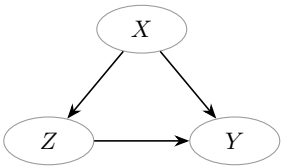
Causal graphs usually omit independent noise terms.

- Better baseline health makes participation in the exercise program more likely.
- $P(Z = 1 \mid X = x) = e(x) \in (1/4, 3/4)$.

Interventions and Incoming Edges

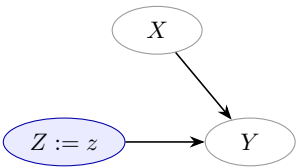
The intervention $\mathbf{do}(\boldsymbol{Z = z})$ replaces the equation for Z with $Z := z$.

Observed mechanism



$$Z := \mathbf{1}\{N_Z < e(X)\}.$$

After intervention



$$Y := 60 + 10z + X + N_Y.$$

- In the graph, remove all **incoming edges** of Z, keep its outgoing edges, and set $Z = z$.
- The **invariance assumption** preserves the other equations and noise distributions.
- Under the SCM, $\boldsymbol{Y(z) \Leftrightarrow Y \mid \mathbf{do}(Z = z)}$.

Observational and Interventional Distributions

- The SCM with independent noise gives formulas for both distributions.
- In the exercise example, $p(x, z, y) = p(x)p(z \mid x)p(y \mid z, x)$. This gives

$$p(y \mid \mathrm{do}(Z = z)) = \int p(y \mid z, x)\, p(x)\, \mathrm{d}x,$$

$$p(y \mid Z = z) = \int p(y \mid z, x)\, p(x \mid z)\, \mathrm{d}x.$$

- Intervention averages over the target population's X distribution. Conditioning on $Z = z$ uses the group's X distribution.

Counterfactual Health Without Exercise

The model is $X = N_X$, $Z = \mathbf{1}\{N_Z < e(X)\}$, $Y = 60 + 10Z + X + N_Y$.
Li had $X = 4$, $Z = 1$, and $Y = 77$. What if he had skipped the program?

Counterfactual Health Without Exercise

The model is $X = N_X$, $Z = \mathbf{1}\{N_Z < e(X)\}$, $Y = 60 + 10Z + X + N_Y$.
Li had $X = 4$, $Z = 1$, and $Y = 77$. What if he had skipped the program?

1. **Infer the background.** Use the observed outcome and its equation.

$$N_Y = 77 - 60 - 10 \times 1 - 4 = 3.$$

Counterfactual Health Without Exercise

The model is $X = N_X$, $Z = \mathbf{1}\{N_Z < e(X)\}$, $Y = 60 + 10Z + X + N_Y$.
Li had $X = 4$, $Z = 1$, and $Y = 77$. What if he had skipped the program?

1. **Infer the background.** Use the observed outcome and its equation.

$$N_Y = 77 - 60 - 10 \times 1 - 4 = 3.$$

2. **Intervene.** Set $Z := 0$ while keeping $X = 4$ and $N_Y = 3$.

3. **Predict the outcome.** Use the same outcome equation.

$$Y(0) = 60 + 10 \times 0 + 4 + 3 = 67.$$

Counterfactual Health Without Exercise

The model is $X = N_X$, $Z = \mathbf{1}\{N_Z < e(X)\}$, $Y = 60 + 10Z + X + N_Y$.
Li had $X = 4$, $Z = 1$, and $Y = 77$. What if he had skipped the program?

1. **Infer the background.** Use the observed outcome and its equation.

$$N_Y = 77 - 60 - 10 \times 1 - 4 = 3.$$

2. **Intervene.** Set $Z := 0$ while keeping $X = 4$ and $N_Y = 3$.

3. **Predict the outcome.** Use the same outcome equation.

$$Y(0) = 60 + 10 \times 0 + 4 + 3 = 67.$$

If the data leave the noise uncertain, use its conditional distribution given the observations.
Individual counterfactuals require the structural equations in addition to the DAG.

Structural Models and Potential Outcomes

Setting $Z := z$ in the exercise model gives

$$Y(z) = f_Y(z, X, N_Y) = 60 + 10z + X + N_Y.$$

- For unit i, hold the background (x_i, n_{Yi}) fixed.

$$Y_i(z) = f_Y(z, x_i, n_{Yi}), \qquad z = 0, 1.$$

- In the population, the potential outcome follows the interventional distribution.

$$P\{Y(z) \in A\} = P\{Y \in A \mid \mathrm{do}(Z = z)\}.$$

- The two frameworks give the same average causal effect.

$$\mathbb{E}\{Y(1) - Y(0)\} = \mathbb{E}\{Y \mid \mathrm{do}(Z = 1)\} - \mathbb{E}\{Y \mid \mathrm{do}(Z = 0)\}.$$

- In a given SCM, both notations describe the same intervention. Individual counterfactuals also retain the unit's background.

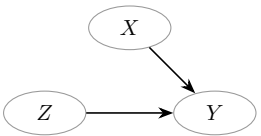
Outline

1. Background and History
   - Examples
   - History
2. Potential Outcomes
   - Potential Outcomes and Causal Effects
   - Randomized Experiments
   - Observational Studies
3. Causal Graphs
   - Causal Graph Basics
   - SCMs, Interventions, and Counterfactuals
   - Do-Calculus, Backdoor Criterion, and Ignorability
   - Causal Discovery

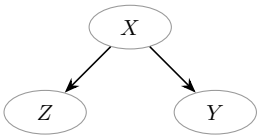
Identification with Do-Calculus

Identification question. Given a causal graph and assumptions, can we express $p(y \mid \mathrm{do}(Z = z))$ as a function of the observational distribution?

- Three rules transform probability expressions under d-separation conditions.
- Combine the rules with conditioning, sums, or integrals to eliminate do.



$G_{\bar{Z}}$ removes edges into Z.



$G_{\underline{Z}}$ removes edges out of Z.

Do-Calculus Rules 1 and 2

Rule 1. Delete an observation.

Delete all arrows **into** A. If A, W d-separate B and Y in this graph, drop the condition $B = b$.

$$p(y \mid \mathrm{do}(a), b, w) = p(y \mid \mathrm{do}(a), w).$$

Do-Calculus Rules 1 and 2

Rule 1. Delete an observation.

Delete all arrows **into** A. If A, W d-separate B and Y in this graph, drop the condition $B = b$.

$$p(y \mid \mathrm{do}(a), b, w) = p(y \mid \mathrm{do}(a), w).$$

Rule 2. Exchange an intervention for an observation.

Delete all arrows **into** A and **out of** B. If A, W d-separate B and Y in this graph, replace the intervention on B with the condition $B = b$.

$$p(y \mid \mathrm{do}(a), \mathrm{do}(b), w) = p(y \mid \mathrm{do}(a), b, w).$$

Do-Calculus Rule 3

Rule 3. Delete an intervention.

To check whether we can drop the intervention on B, modify the graph in two steps.

1. Delete all arrows **into** A.
2. In that graph, find the nodes in B with **no directed path to any node in** W. Delete all arrows into those nodes.

If A, W d-separate B and Y in the resulting graph, drop the intervention on B.

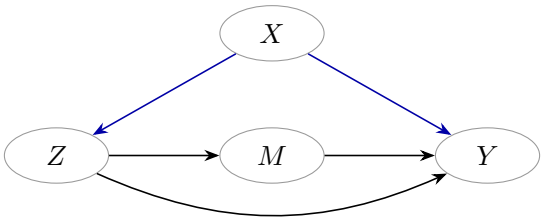
$$p(y \mid \mathrm{do}(a), \mathrm{do}(b), w) = p(y \mid \mathrm{do}(a), w).$$

If W is empty, step 2 removes all arrows into B.

In reverse, the rules insert an observation, replace an observation with an intervention, or insert an intervention.

Backdoor Paths

A **backdoor path** from Z to Y first goes **backward along an arrow** to a parent of Z.

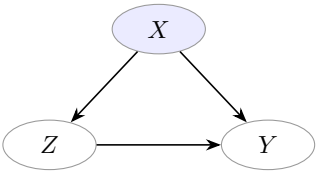


- $Z \to Y$ and $Z \to M \to Y$ are directed causal paths. Exercise can improve health directly or by improving sleep M.
- $Z \leftarrow X \to Y$ is a backdoor path. Baseline health affects exercise choices and later health.
- For the **total effect**, **block** backdoor paths and **keep** directed paths open.

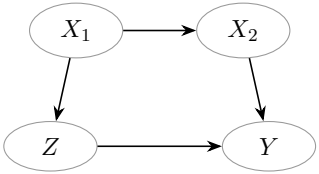
The Backdoor Criterion

A covariate set X that excludes Z and Y meets the **backdoor criterion** for (Z, Y) if and only if

1. X contains no descendants of Z, and
2. X blocks every backdoor path from Z to Y.



Adjusting for X blocks $Z \leftarrow X \rightarrow Y$.



Adjusting for either X_1 or X_2 blocks the backdoor path.

In the potential outcomes framework, X is the set of **confounding covariates**.

Deriving Backdoor Adjustment

Suppose X meets the backdoor criterion. Write $\mathrm{do}(z) = \mathrm{do}(Z = z)$ and apply the law of total probability.

$$p(y \mid \mathrm{do}(z)) = \int p(y \mid \mathrm{do}(z), x) p(x \mid \mathrm{do}(z)) \, \mathrm{d}x$$

$$\overset{(1)}{=} \int p(y \mid \mathrm{do}(z), x) p(x) \, \mathrm{d}x$$

$$\overset{(2)}{=} \int p(y \mid z, x) \, p(x) \, \mathrm{d}x.$$

1. Since X contains no descendants of Z, Rule 3 gives $p(x \mid \mathrm{do}(z)) = p(x)$.
2. Since X d-separates Z and Y in $G_{\underline{Z}}$, Rule 2 gives $p(y \mid \mathrm{do}(z), x) = p(y \mid z, x)$.

Backdoor Adjustment and Ignorability

Consider an SCM whose DAG represents the true causal structure, with no unobserved confounding.

$$X \text{ meets the backdoor criterion} \implies Y(z) \perp Z \mid X, \qquad z = 0, 1.$$

With SUTVA, the potential outcomes framework gives

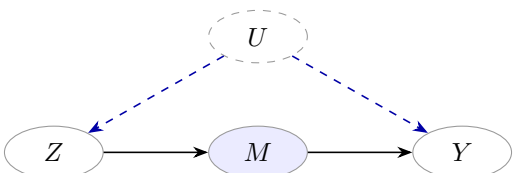
$$\begin{aligned}
\mathbb{E}\{Y(z)\} &= \mathbb{E}_X\big[\mathbb{E}\{Y(z) \mid X\}\big] \\
&= \mathbb{E}_X\big[\mathbb{E}\{Y(z) \mid Z = z, X\}\big] \\
&= \mathbb{E}_X\big[\mathbb{E}(Y \mid Z = z, X)\big].
\end{aligned}$$

- The backdoor criterion is **sufficient** for ignorability. The converse can fail.

Frontdoor Criterion

The **frontdoor criterion** for (Z, Y) requires an observed mediator M with three properties.

1. Every directed path from Z to Y passes through M.
2. There is no open backdoor path from Z to M.
3. Conditioning on Z blocks every backdoor path from M to Y.



Here M meets the criterion. With U unobserved, can't use the backdoor criterion.

Frontdoor Adjustment

Suppose M meets the frontdoor criterion and every pair of values of discrete Z and M has positive probability.

$$p(m \mid \mathrm{do}(z)) = p(m \mid z) \qquad \text{(2) No open backdoor path}$$

$$p(y \mid \mathrm{do}(m)) = \sum_{z'} p(y \mid m, z')\, p(z') \qquad \text{(3) Backdoor adjustment for } Z.$$

Since M intercepts every directed path from Z to Y, do-calculus combines the two distributions to give the **frontdoor adjustment formula**.

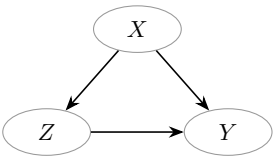
$$\begin{aligned} p(y \mid \mathrm{do}(z)) &= \sum_{m} p(m \mid \mathrm{do}(z))\, p(y \mid \mathrm{do}(m)) \\ &= \sum_{m} p(m \mid z) \sum_{z'} p(y \mid m, z')\, p(z'). \end{aligned}$$

The observed distribution of (Z, M, Y) identifies the **total causal effect** even with unobserved confounding between Z and Y.

Confounders, Mediators, and Colliders

The target is the **total causal effect** of Z on Y.

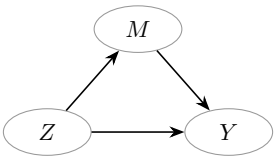
Confounder



Adjusting for X blocks the backdoor path.

Baseline health affects exercise choices and later health.

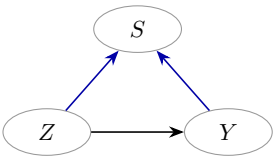
Mediator



Adjusting for M blocks part of the effect.

Better sleep mediates the effect of exercise on health.

Collider



$Z \to S \leftarrow Y$ is blocked. Conditioning on S opens it.

$Z \to Y$ stays open.

Exercise and health affect follow-up status S.

Outline

1. Background and History
   - Examples
   - History
2. Potential Outcomes
   - Potential Outcomes and Causal Effects
   - Randomized Experiments
   - Observational Studies
3. Causal Graphs
   - Causal Graph Basics
   - SCMs, Interventions, and Counterfactuals
   - Do-Calculus, Backdoor Criterion, and Ignorability
   - Causal Discovery

Estimating Effects and Learning Graphs

We learn directed acyclic graphs from observational data, assuming no unobserved common causes.

- **Causal effect estimation** uses a given graph and identification assumptions to estimate an intervention effect.
- **Causal discovery** learns causal structure from data under additional assumptions.

The Markov property and faithfulness link d-separation to conditional independence.

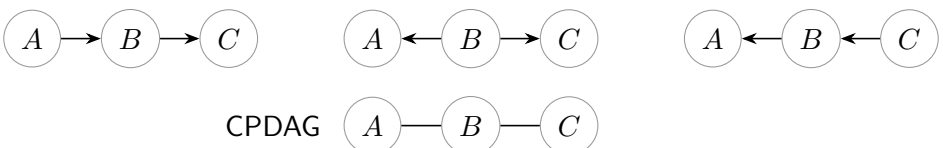
$$A \perp_G B \mid S \quad \underset{\text{Faithfulness}}{\overset{\text{Markov property}}{\rightleftarrows}} \quad A \perp B \mid S.$$

Which edges and directions can we recover from conditional independence alone?

Markov Equivalence and CPDAGs

Two DAGs are **Markov equivalent** exactly when they have the same skeleton and the same unshielded colliders. In an unshielded collider, the two parents are nonadjacent.

These three DAGs share a skeleton, contain no colliders, and imply $A \perp C \mid B$.



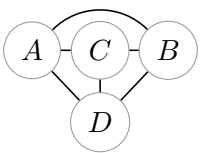
A **CPDAG** (completed partially directed acyclic graph) directs edges whose orientation is fixed across the class and leaves reversible edges undirected.

The collider $A \to B \leftarrow C$ implies $A \perp C$ and belongs to a different class.

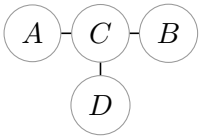
PC: Learning the Skeleton

PC tests conditional independence in rounds of increasing conditioning-set size. For each edge, choose the set from the current neighbors of either endpoint, excluding the other endpoint. When removing an edge, record the **separating set** Sep.

Complete graph



Learned skeleton

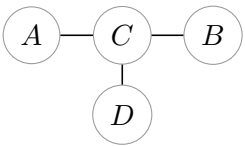


1. **Marginal tests,** $|S| = 0$
   Since $A \perp B$, remove $A - B$ and set $\mathrm{Sep}(A, B) = \varnothing$.
2. **Conditional tests,** $|S| = 1$
   Since $A \perp D \mid C$ and $B \perp D \mid C$, remove $A - D$ and $B - D$. Both separating sets are $\{C\}$.

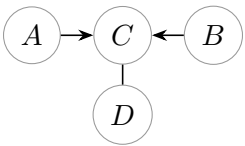
PC: Orienting Edges

For a triple $A - C - B$ in the skeleton with nonadjacent endpoints, orient C as a collider if $C \notin \mathrm{Sep}(A, B)$.

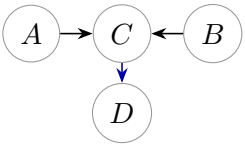
Undirected skeleton



Orient the collider



Apply orientation rules



Here $\mathrm{Sep}(A, B) = \varnothing$. Choosing $D \to C$ would create an extra unshielded collider $A \to C \leftarrow D$. Meek (UAI, 1995) gives the rule that forces $C \to D$. All three edges of the resulting CPDAG are directed.

PC: Testing Conditional Independence

For jointly Gaussian data, conditional independence is equivalent to zero partial correlation. See Kalisch and Bühlmann (JMLR, 2007).

$$U \perp V \mid S \quad \Longleftrightarrow \quad \rho_{UV \cdot S} = 0.$$

Here is how to test $A \perp D \mid C$.

PC: Testing Conditional Independence

For jointly Gaussian data, conditional independence is equivalent to zero partial correlation. See Kalisch and Bühlmann (JMLR, 2007).

$$U \perp V \mid S \quad \Longleftrightarrow \quad \rho_{UV \cdot S} = 0.$$

Here is how to test $A \perp D \mid C$.

1. Regress A and D separately on C and save the residuals.
2. Let $\widehat{\rho}_{AD \cdot C}$ be the sample correlation of the residuals. Under the null, the Fisher statistic

$$Z = \sqrt{n-4}\, \operatorname{arctanh}(\widehat{\rho}_{AD \cdot C})$$

   is approximately standard normal.
3. If $p > \alpha$, fail to reject independence. PC removes $A - D$ and records $\{C\}$ as the separating set.

Mistakenly retaining $A - B$ creates a triangle and prevents PC from using $A - C - B$ to orient the collider.

GES: Scoring Candidate Graphs

Greedy Equivalence Search (GES) moves between Markov equivalence classes, following Chickering (JMLR, 2002). For linear Gaussian models, one score to maximize is

$$S(G) = -\frac{1}{2}\,\mathrm{BIC}(G) = \ell(\widehat{\theta}_G) - \frac{d_G}{2}\log n,$$

where d_G is the number of free parameters. DAGs in the same class have the same score.

GES: Scoring Candidate Graphs

Greedy Equivalence Search (GES) moves between Markov equivalence classes, following Chickering (JMLR, 2002). For linear Gaussian models, one score to maximize is

$$S(G) = -\frac{1}{2}\,\mathrm{BIC}(G) = \ell(\widehat{\theta}_G) - \frac{d_G}{2}\log n,$$

where d_G is the number of free parameters. DAGs in the same class have the same score.

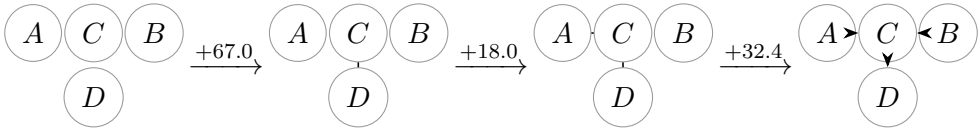
Four-node example Let $C = A + B + N_C$ and $D = C + N_D$, with A, B, N_C, N_D independent standard normal variables. For illustration, use $n = 100$ and set the sample covariance equal to the model covariance.

Adding $A \to C$ to the empty graph reduces the residual variance of C from 3 to 2 and adds one coefficient.

$$\Delta S = \frac{100}{2}\log\frac{3}{2} - \frac{1}{2}\log 100 \;\approx\; 18.0.$$

GES: Forward and Backward Search

Forward phase Start from the empty graph. Take the valid insertion with the largest positive score gain and update the CPDAG.



- In step 2, $A - C$ and $B - C$ tie. The example takes $A - C$.
- In step 3, adding $B - C$ gains 32.4 with the collider and 18.0 for the same skeleton with no collider.

Backward phase Once no insertion improves the score, consider valid deletions. The three possible deletions here change the score by about -32.4, -32.4, and -67.0. Keep the current CPDAG.

Additive Noise Models

An additive noise model (ANM) writes each variable as a function of its parents plus noise.

$$V_j = f_j(\mathrm{PA}_j) + N_j, \qquad N_1, \ldots, N_d \text{ mutually independent.}$$

For two variables, compare the models in the two directions.

Forward: $X \to Y$

$$Y = f(X) + N_Y, \qquad N_Y \perp X.$$

Reverse: $Y \to X$

$$X = g(Y) + N_X, \qquad N_X \perp Y.$$

- Fit f and g by regression. Test whether each residual is independent of its predictor.
- The direction is **identifiable** when the joint distribution admits an ANM with independent noise in exactly one direction.

Additive Noise Models: Identifiability

Suppose the forward model is $Y = f(X) + N$, with $N \perp X$.

1. **Nonlinear function, Gaussian input and noise.** Both X and N are nondegenerate Gaussian, and f is nonlinear and three times continuously differentiable.

2. **Linear function, non-Gaussian component.** Restrict both directions to linear models. If $f(x) = ax$ with $a \neq 0$ and at least one of the nondegenerate variables X, N is non-Gaussian, the direction is identifiable.

3. **Changing shape of the reverse conditional distribution.** Under a reverse ANM, $X \mid Y = y$ varies only by a shift. Changes in its variance (when defined) or support width rule out the reverse direction.

In the linear Gaussian case, both directions admit an ANM and the direction is unidentified.

Additive Noise Models: Two Examples

Quadratic with Gaussian noise

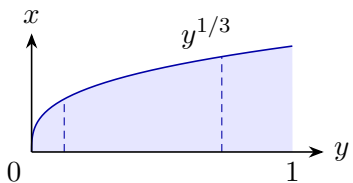
$Y = X^2 + N$, with $X, N \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$.

- **Forward.** $Y - X^2 = N \perp X$.
- **Reverse.** Symmetry gives $\mathbb{E}[X \mid Y] = 0$, so the residual is X. It depends on Y because $\mathbb{E}[Y \mid X] = X^2$ varies with X.

Cubic with uniform noise

$Y = X^3 + N$, with $X, N \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(0,1)$.

- **Forward.** $Y - X^3 = N \perp X$.
- **Reverse.** For $0 < y < 1$, the support of $X \mid Y = y$ is $[0, y^{1/3}]$. Its width varies with y.



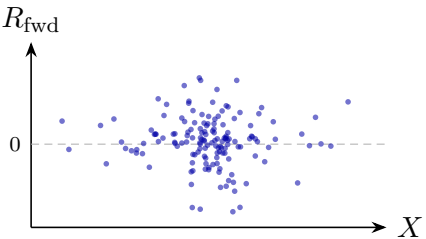
In both examples, only $X \to Y$ admits independent additive noise.

LiNGAM: Direction from Non-Gaussian Noise

Shimizu et al. (JMLR, 2006) use non-Gaussian noise to orient a linear ANM. Let $X = N_X$ and $Y = X + N_Y$, with independent, centered Laplace noise of variance 1.

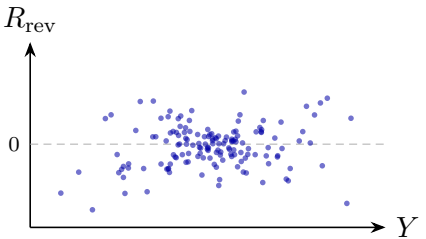
Forward regression of Y on X

$$R_{\text{fwd}} = Y - X = N_Y, \qquad R_{\text{fwd}} \perp X$$



Reverse regression of X on Y

$$R_{\text{rev}} = X - Y/2 = (N_X - N_Y)/2$$

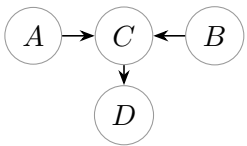


Y and R_{rev} are uncorrelated. Since $\text{Cov}(Y^2, R_{\text{rev}}^2) = 1.5$, they are **dependent**.

DirectLiNGAM: Finding Root Nodes

Shimizu et al. (JMLR, 2011) find roots by regression. Let A, B, N_C, N_D be independent and non-Gaussian.

$$C = A + B + N_C, \qquad D = C + N_D.$$



1. **Find a root.** Regress the other variables on a candidate. Choose one independent of each residual.
2. **Replace with residuals.** Choosing A first gives

$$B' = B, \quad C' = B + N_C, \quad D' = B + N_C + N_D.$$

3. **Fit the edges.** Repeat to get A, B, C, D. Regress each variable on all its predecessors and keep the edges with nonzero coefficients.

RESIT: Finding Sinks and Pruning Parent Edges

Peters et al. (JMLR, 2014) first find sinks, then prune parent edges. Let $M = X^2 + U$ and $Y = M + V$. The variables X, U, V are independent, with X standard normal and U, V identically distributed centered Laplace noises.



Find sinks

- Regress each remaining variable on the others. Choose the variable whose residual depends least on its predictors jointly.
- Remove Y using $V \perp (X, M)$, then M using $U \perp X$. Reverse the order to obtain X, M, Y.

Prune parent edges

- Treat earlier variables as candidate parents. Omit a parent and refit. Delete its edge if the new residual is jointly independent of all earlier variables.
- Regressing Y on M alone still gives $V \perp (X, M)$, so delete $X \to Y$.

What Do the Methods Identify?

Method	Evidence	Key conditions	Output
PC	Conditional independence	Markov property, faithfulness	Equivalence class
GES	Fit and complexity	Faithfulness, consistent score	Equivalence class
LiNGAM	Non-Gaussian noise	Linear, independent non-Gaussian noise	DAG
ANM / RESIT	Residual independence	Identifiable additive noise model	DAG

Question. If the CPDAG is $A - B - C$, can we identify $p(c \mid do(a))$?

What Do the Methods Identify?

Method	Evidence	Key conditions	Output
PC	Conditional independence	Markov property, faithfulness	Equivalence class
GES	Fit and complexity	Faithfulness, consistent score	Equivalence class
LiNGAM	Non-Gaussian noise	Linear, independent non-Gaussian noise	DAG
ANM / RESIT	Residual independence	Identifiable additive noise model	DAG

Question. If the CPDAG is $A - B - C$, can we identify $p(c \mid do(a))$?

- In $A \to B \to C$, B is a mediator and $p(c \mid do(a)) = p(c \mid a)$.
- In $A \leftarrow B \to C$, B is a common cause and $p(c \mid do(a)) = p(c)$.

Two DAGs in the same equivalence class can imply different interventional distributions.

References

Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press.

Ding, P. (2024). *A First Course in Causal Inference*. Chapman & Hall/CRC.

Kalisch, M. and Bühlmann, P. (2007). Estimating High-Dimensional Directed Acyclic Graphs with the PC-Algorithm. *JMLR*, 8, 613–636.

Meek, C. (1995). Causal Inference and Causal Explanation with Background Knowledge. *UAI*, 403–410.

Chickering, D. M. (2002). Optimal Structure Identification with Greedy Search. *JMLR*, 3, 507–554.

Shimizu, S. et al. (2006). A Linear Non-Gaussian Acyclic Model for Causal Discovery. *JMLR*, 7, 2003–2030.

Shimizu, S. et al. (2011). DirectLiNGAM: A Direct Method for Learning a Linear Non-Gaussian Structural Equation Model. *JMLR*, 12, 1225–1248.

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. (2014). Causal Discovery with Continuous Additive Noise Models. *JMLR*, 15, 2009–2053.