TECHNISCHE UNIVERSITÄT DARMSTADT

AIML Lab

Winter Semester 2025/26 Lecture

Causality for AI & ML

"Causal Perspectives"

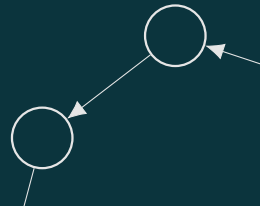
Prof. Dr. Kristian Kersting

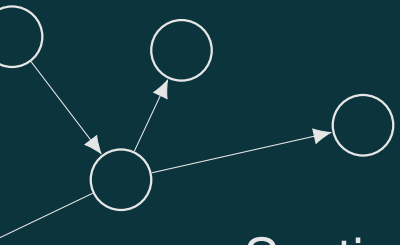
Moritz Willig — Today's speaker

Tim Woydt

Florian Busch

Matej Zečević

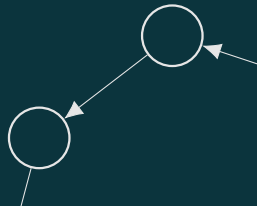




Section

1

Attribution



Fair Attribution

Shapley Values have their origin in Cooperative Game Theory.

$\rightarrow$ How to fairly distribute payouts among actors according to their contribution?

Shapley, Lloyd S. "A value for n-person games." (1953): 307-317.

Fair Attribution

Shapley Values have their origin in Cooperative Game Theory.

$\rightarrow$ How to fairly distribute payouts among actors according to their contribution?

- Consider the set of all actors N.
- Consider *coalitions* of actors, $S \subseteq N$.
- Consider the existence of a *value function*, $v : S \rightarrow \mathbb{R}$, that assigns an expected reward for any coalition $S \subseteq N$ (and particularly $v(\emptyset) = 0$)

Shapley, Lloyd S. "A value for n-person games." (1953): 307-317.

Fair Attribution

Shapley Values have their origin in Cooperative Game Theory.
$\rightarrow$ How to fairly distribute payouts among actors according to their contribution?

- Consider the set of all actors N.
- Consider *coalitions* of actors, $S \subseteq N$.
- Consider the existence of a *value function*, $v : S \rightarrow \mathbb{R}$, that assigns an expected reward for any coalition $S \subseteq N$ (and particularly $v(\emptyset) = 0$)

How to define a fair payout?
General considerations on an attribution function φ should involve:

- Gives high reward to actors that increase the reward more.
- Gives lower reward to actors that have less impact.
- Gives no reward to actors that never contribute to the reward.

Shapley, Lloyd S. "A value for n-person games." (1953): 307-317.

What Defines a Fair Attribution?

1. **Efficiency.** The value of the whole coalition is distributed to the individual actors:

$$\sum_{i \in N} \varphi_i(v) = v(N)$$

What Defines a Fair Attribution?

1. **Efficiency.** The value of the whole coalition is distributed to the individual actors:
$$\sum_{i \in N} \varphi_i(v) = v(N)$$

2. **Symmetry.** Actors i, j that contribute equally should receive equal attribution:
$$\forall S \in (N \setminus \{i, j\}).(v(S \cup \{i\}) = v(S \cup \{j\})) \Rightarrow (\varphi_i(v) = \varphi_j(v))$$

What Defines a Fair Attribution?

1. Efficiency. The value of the whole coalition is distributed to the individual actors:

$$\sum_{i \in N} \varphi_i(v) = v(N)$$

2. Symmetry. Actors i, j that contribute equally should receive equal attribution:

$$\forall S \in (N \setminus \{i, j\}).(v(S \cup \{i\}) = v(S \cup \{j\})) \Rightarrow (\varphi_i(v) = \varphi_j(v))$$

3. Additivity/Linearity.: When a game can be decomposed into two independent value functions v, w, attribution should behave linearly:

$$\varphi_i(v + w) = \varphi_i(v) + \varphi_i(w)$$

What Defines a Fair Attribution?

1. Efficiency. The value of the whole coalition is distributed to the individual actors:

$$\sum_{i \in N} \varphi_i(v) = v(N)$$

2. Symmetry. Actors i, j that contribute equally should receive equal attribution:

$$\forall S \in (N \setminus \{i, j\}).(v(S \cup \{i\}) = v(S \cup \{j\})) \Rightarrow (\varphi_i(v) = \varphi_j(v))$$

3. Additivity/Linearity.: When a game can be decomposed into two independent value functions v, w, attribution should behave linearly:

$$\varphi_i(v + w) = \varphi_i(v) + \varphi_i(w)$$

4. Dummy/Null Player. An actor i that does not contribute to any coalition gets zero attribution:

$$(\forall S \in N.v(S \cup \{i\}) = v(S)) \Rightarrow (\varphi_i(v) = 0)$$

Shapley Values

An actor joining a coalition improves reward by $\Delta v(S, i) = v(S \cup \{i\}) - v(S)$, assuming that $v(S \cup \{i\}) \geq 0$.

Shapley Values

An actor joining a coalition improves reward by $\Delta v(S, i) = v(S \cup \{i\}) - v(S)$, assuming that $v(S \cup \{i\}) \geq 0$.

Shapley Value for a player i in a coalition game (v, N):

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \Delta v(S, i) = \frac{1}{|N|} \sum_{S \subseteq N \setminus \{i\}} \binom{|N| - 1}{|S|}^{-1} \Delta v(S, i)$$

Shapley Values

An actor joining a coalition improves reward by $\Delta v(S, i) = v(S \cup \{i\}) - v(S)$, assuming that $v(S \cup \{i\}) \geq 0$.

Shapley Value for a player i in a coalition game (v, N):

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \Delta v(S, i) = \frac{1}{|N|} \sum_{S \subseteq N \setminus \{i\}} \binom{|N| - 1}{|S|}^{-1} \Delta v(S, i)$$

Interpretation

1. Consider actors joining a coalition S one-by-one.
2. Every player takes its respective reward $v(S \cup \{i\}) - v(S)$.
3. Shapley values compute the average contribution to every possible way (permutation) that the respective coalition could have been formed.

Shapley Values

An actor joining a coalition improves reward by $\Delta v(S, i) = v(S \cup \{i\}) - v(S)$, assuming that $v(S \cup \{i\}) \geq 0$.

Shapley Value for a player i in a coalition game (v, N):

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \Delta v(S, i) = \frac{1}{|N|} \sum_{S \subseteq N \setminus \{i\}} \binom{|N| - 1}{|S|}^{-1} \Delta v(S, i)$$

Interpretation

1. Consider actors joining a coalition S one-by-one.
2. Every player takes its respective reward $v(S \cup \{i\}) - v(S)$.
3. Shapley values compute the average contribution to every possible way (permutation) that the respective coalition could have been formed.

Shapley values are the only solution that fulfill all properties on efficiency, symmetry, additivity and null players.

Shapley, Lloyd S. "A value for n-person games." (1953): 307-317.

Extension to Continuous Variables

Shapley values can be extended to continuous variables.
Still defined over discrete players (indices $i \in \{1..d\}$),
with every player having an associated value $x_i \in \mathbb{R}$.

Extension to Continuous Variables

Shapley values can be extended to continuous variables.
Still defined over discrete players (indices $i \in \{1..d\}$),
with every player having an associated value $x_i \in \mathbb{R}$.

$\rightarrow$ The input domain of v becomes continuous, $v : \mathbb{R}^d \rightarrow \mathbb{R}$.

Extension to Continuous Variables

Shapley values can be extended to continuous variables.
Still defined over discrete players (indices $i \in \{1..d\}$),
with every player having an associated value $x_i \in \mathbb{R}$.

$\rightarrow$ The input domain of v becomes continuous, $v : \mathbb{R}^d \rightarrow \mathbb{R}$.

SHapley Additive exPlanations (SHAP): For some predictive function $f(\mathbf{x})$, attribute individual features $x_i \in \mathbf{x}$ towards the prediction:

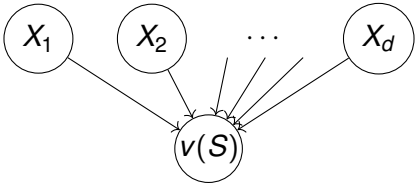
$$v(S) = \mathbb{E}[f(x) \mid x_S] = \int f(x_S, x_{\bar{S}}) p(x_{\bar{S}} | x_S) dx_{\bar{S}}$$

$\rightarrow$ Basically, marginalizing out all variables $\bar{S}$ not part of the coalition S.

Lundberg, S.M. and Lee, S.I., 2017. A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

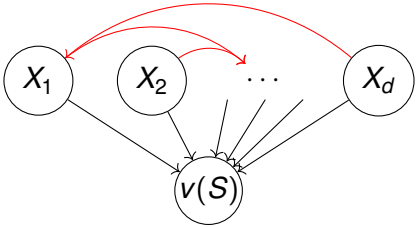
Considerations

Independence: Standard SHAP often assumes feature independence:



Considerations

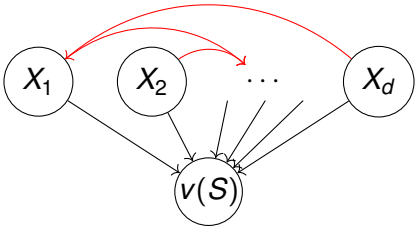
Independence: Standard SHAP often assumes feature independence:



What if individual factors cause each other: education $\rightarrow$ income $\rightarrow$ buying power.

Considerations

Independence: Standard SHAP often assumes feature independence:

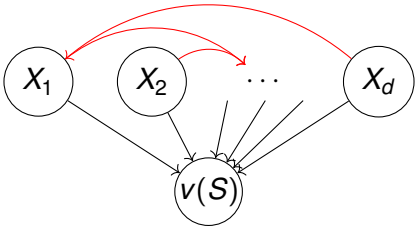


What if individual factors cause each other: education $\rightarrow$ income $\rightarrow$ buying power.

Spurious Explanations: SHAP might attribute all influence to ‘income’, but miss the stronger total causal effects of ‘education’.

Considerations

Independence: Standard SHAP often assumes feature independence:



What if individual factors cause each other: education $\rightarrow$ income $\rightarrow$ buying power.

Spurious Explanations: SHAP might attribute all influence to ‘income’, but miss the stronger total causal effects of ‘education’.

Off-Manifold Estimates: Sampling features independently can lead to “unrealistic” data points that the model never saw during training.

Interventional Expectations

Causal Shapley Values replace the conditional expectation with the *interventional expectation* using the *do*-operator.

→ "Explain *mechanisms* rather than just *data statistics*."

$$v_{causal}(S) = \mathbb{E}[f(x) \mid do(X_S = x_S)] = \int f(x_S, x_{\bar{S}}) \cdot \underbrace{p(x_{\bar{S}} \mid do(x_S))}_{\text{Interventional Density}} \, dx_{\bar{S}}$$

Distinction:

- **Conditional:** $P(y \mid x)$. Propagates information up and down the causal graph (allows for confounding and mediators).
- **Interventional:** $P(y \mid do(x))$. Cuts upstream links. Only propagates effects to descendants.

Heskes, T., Sijben, E., Bucur, I.G. and Claassen, T., 2020. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. Advances in neural informationprocessing systems, 33, pp.4778-4789.

Incorporating Structure

Some permutations of actors i, j, k joining a coalition $\{i, j, k\}$ might violate the underlying causal graph.

- If $x_i \to x_j \leftarrow x_k$, then j joining before i and k might not be allowed.

Incorporating Structure

Some permutations of actors i, j, k joining a coalition $\{i, j, k\}$ might violate the underlying causal graph.

- If $x_i \rightarrow x_j \leftarrow x_k$, then j joining before i and k might not be allowed.

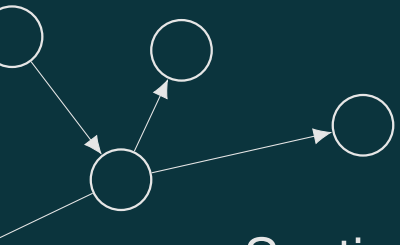
If the DAG is known we can use **Asymmetric Shapley Values**.

Instead of all permutations π being weighted equally $w(\pi) = 1/(N!)$, only consider permutations that **respect the causal ordering** of the DAG.

$\rightarrow$ Incompatible orderings receive zero weight.

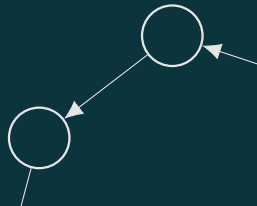
$$\leq_{\mathcal{G}} \subseteq \pi_S \Leftrightarrow w(\pi_S) \neq 0$$

Frye, C., Rowat, C. and Feige, I., 2020. Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. Advances in neural information processing systems, 33, pp.1229-1239.



Section 2

Potential Outcomes



Potential Outcomes

The **Rubin-Neyman Causal Model** $\rightarrow$ Potential Outcomes.

While Pearlian causality focuses on graphs, Rubin's framework focuses on **units**.

$\rightarrow$ Widely used in modern statistics and clinical trials.

$\rightarrow$ 'What is the average treatment effect of some drug?'

Imbens, Guido W., and Donald B. Rubin. Causal inference in statistics, social, and biomedical sciences. Cambridge university press, 2015.

Potential Outcomes

The **Rubin-Neyman Causal Model** $\rightarrow$ Potential Outcomes.

While Pearlian causality focuses on graphs, Rubin's framework focuses on **units**.

$\rightarrow$ Widely used in modern statistics and clinical trials.

$\rightarrow$ 'What is the average treatment effect of some drug?'

Concerned with the 'fundamental problem of causal inference':

"We never observe both Y(0) and Y(1) for the same individual."

Imbens, Guido W., and Donald B. Rubin. Causal inference in statistics, social, and biomedical sciences. Cambridge university press, 2015.

The Fundamental Problem of Causal Inference

In the Potential Outcomes framework, every individual i has two 'potential' outcomes of their future:

- $Y_i(1)$: The outcome if individual i receives the treatment ($D = 1$).
- $Y_i(0)$: The outcome if individual i does *not* receive the treatment ($D = 0$).

$\rightarrow$ *"The drug was either administered to the patient ($D = 1$) or not ($D = 0$)."*

The Fundamental Problem of Causal Inference

In the Potential Outcomes framework, every individual i has two 'potential' outcomes of their future:

- $Y_i(1)$: The outcome if individual i receives the treatment ($D = 1$).
- $Y_i(0)$: The outcome if individual i does *not* receive the treatment ($D = 0$).

$\rightarrow$ *"The drug was either administered to the patient ($D = 1$) or not ($D = 0$)."*

We only ever observe either $Y_i(0)$ or $Y_i(1)$:

$$Y_i = \begin{cases} Y_i(0) & \textbf{if } D = 0 \\ Y_i(1) & \textbf{otherwise.} \end{cases}$$

The Fundamental Problem of Causal Inference

In the Potential Outcomes framework, every individual i has two 'potential' outcomes of their future:

- $Y_i(1)$: The outcome if individual i receives the treatment ($D = 1$).
- $Y_i(0)$: The outcome if individual i does *not* receive the treatment ($D = 0$).

$\rightarrow$ *"The drug was either administered to the patient ($D = 1$) or not ($D = 0$)."*

We only ever observe either $Y_i(0)$ or $Y_i(1)$:

$$Y_i = \begin{cases} Y_i(0) & \textbf{if } D = 0 \\ Y_i(1) & \textbf{otherwise.} \end{cases}$$

The unobserved outcome is the **counterfactual**.

Causal inference is essentially framed as a 'missing data' problem.

Estimating Causal Effects

Instead of the Pearlian $P(y|do(x))$, PO focuses on the difference between the two potential universes:

1. **Individual Causal Effect:** $\tau_i = Y_i(1) - Y_i(0)$
   Impossible to observe directly.
2. **Average Treatment Effect (ATE):**

$$ATE = \mathbb{E}[Y(1) - Y(0)]$$

3. **Conditional Average Treatment Effect (CATE):**

$$CATE(x) = \mathbb{E}[Y(1) - Y(0)|X = x]$$

The Fundamental Problem of Causal Inference

We are interested in the effectiveness of a treatment D.
Often quantified via the individual ATE $= \mathbb{E}[Y_i(1) - Y_i(0)]$

Problem: We never observe the other *potential outcome*.

i	D	Y	$Y(1)$	$Y(0)$	$Y(1) - Y(0)$
1	0	0	?	0	?
2	1	1	1	?	?
3	0	0	?	0	?
4	1	0	0	?	?
⋮	⋮	⋮	⋮	⋮	⋮

$\rightarrow$ We can never really compute $Yi(1) - Y_i(0)$.

The Fundamental Problem of Causal Inference

We are interested in the effectiveness of a treatment D.
Often quantified via the individual ATE $= \mathbb{E}[Y_i(1) - Y_i(0)]$

Problem: We never observe the other *potential outcome*.

i	D	Y	$Y(1)$	$Y(0)$	$Y(1) - Y(0)$
1	0	0	?	0	?
2	1	1	1	?	?
3	0	0	?	0	?
4	1	0	0	?	?
⋮	⋮	⋮	⋮	⋮	⋮

$\rightarrow$ We can never really compute $Yi(1) - Y_i(0)$.

Under which assumptions can we transform the equation to:

$$\mathbb{E}[Y|T = 1] - \mathbb{E}[Y|T = 0]$$

Assumptions of PO

To estimate ATE from observational data, the setting must satisfy:

1. **SUTVA:** "Stable Unit Treatment Value Assumption".
   "One units treatment does not affect another's outcome."
2. **Ignorability (Unconfoundedness):** $\{Y(0), Y(1)\} \perp D|X$.
   "Given covariates X, the treatment assignment is as good as random."
3. **Positivity (Overlap):** $0 < P(D = 1|X) < 1$.
   "Every person has a non-zero chance of being in either group."

1. SUTVA: ‘No-Interference’ Assumption

Stable Unit Treatment Value Assumption consists of two parts:

1. **No Interference:** The potential outcome of unit i is unaffected by the treatment assignment of unit j:
$$Y_i(d_1, d_2, \ldots, d_n) = Y_i(d_i)$$
2. **No Variations of Treatment:** There are no "hidden" versions of the treatment
$\rightarrow$ e.g., ‘Aspirin’ intake must be the same dose for everyone.

Counterexample (Spillover): In vaccine trials, if your neighbor is vaccinated ($D_j = 1$), your risk of infection $Y_i(0)$ decreases even if you are not.
$\rightarrow$ This violates SUTVA and biases the ATE.

2. Ignorability (Unconfoundedness)

also known as "Exchangeability".

"Given covariates X, the potential outcomes are independent of the actual treatment received":

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

$\rightarrow$ When conditioning on X, the treatment D is effectively randomized.

- **Connection to do-Calculus:** This is satisfied if X satisfies the *Backdoor Criterion* in a DAG.
- **Violation:** If a hidden confounder U affects both D and Y, and $U \notin X$, ignorability fails.

3. Positivity (Overlap)

"We cannot compare treated and control groups if certain types of people only ever appear in one group":

$$0 < P(D = 1 \mid X = x) < 1, \quad \forall x \text{ such that } P(X = x) > 0$$

3. Positivity (Overlap)

"We cannot compare treated and control groups if certain types of people only ever appear in one group":

$$0 < P(D = 1 \mid X = x) < 1, \quad \forall x \text{ such that } P(X = x) > 0$$

Example: In a study of "Smoking vs. Lung Cancer", if *every* person over 70 in your dataset is a smoker, you cannot calculate the effect of "not smoking" for that age group.

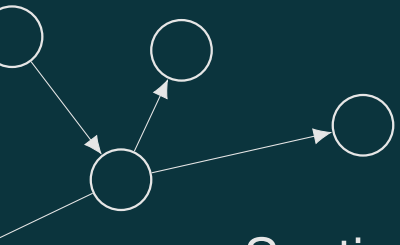
3. Positivity (Overlap)

"We cannot compare treated and control groups if certain types of people only ever appear in one group":

$$0 < P(D = 1 \mid X = x) < 1, \quad \forall x \text{ such that } P(X = x) > 0$$

Example: In a study of "Smoking vs. Lung Cancer", if *every* person over 70 in your dataset is a smoker, you cannot calculate the effect of "not smoking" for that age group.

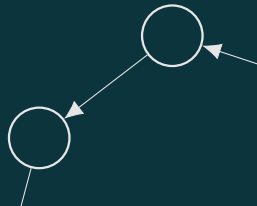
Consequence: If the propensity score $e(x)$ is 0 or 1, the denominator in weighting estimators becomes zero.



Section

3

Regression Discontinuity Design



Regression Discontinuity Design Intuition

Quasi-Experimental Methods interpret observational phenomena as interventions.

Grant Program Example: You oversee a grant program for students.

$\rightarrow$ Students need a score of 70 or above to enter the program.

Your organization wants to know whether the program actually helping students.

Hahn, J., Todd, P. and Van der Klaauw, W., 2001. Identification and estimation of treatment effects with a regression-discontinuity design. Econometrica, 69(1), pp.201-209.
Lee, D.S. and Lemieux, T., 2010. Regression discontinuity designs in economics. Journal of economic literature, 48(2), pp.281-355.

Regression Discontinuity Design Intuition

Quasi-Experimental Methods interpret observational phenomena as interventions.

Grant Program Example: You oversee a grant program for students.

→ Students need a score of 70 or above to enter the program.

Your organization wants to know whether the program actually helping students.

→ Comparing groups of students within and outside the program is biased.

→ Good students are expected to to perform above average anyway.

→ The rule for program admission are set. You can not intervene.

Hahn, J., Todd, P. and Van der Klaauw, W., 2001. Identification and estimation of treatment effects with a regression-discontinuity design. Econometrica, 69(1), pp.201-209.
Lee, D.S. and Lemieux, T., 2010. Regression discontinuity designs in economics. Journal of economic literature, 48(2), pp.281-355.

Regression Discontinuity Design Intuition

Quasi-Experimental Methods interpret observational phenomena as interventions.

Grant Program Example: You oversee a grant program for students.

→ Students need a score of 70 or above to enter the program.

Your organization wants to know whether the program actually helping students.

→ Comparing groups of students within and outside the program is biased.

→ Good students are expected to to perform above average anyway.

→ The rule for program admission are set. You can not intervene.

Considerations:

- There is nothing 'special' about the admission cut-off of 70.
- Continuity: Individuals near the cut-off have similar expected properties.

→ Differences in follow-up performance should only be due to the program.

Hahn, J., Todd, P. and Van der Klaauw, W., 2001. Identification and estimation of treatment effects with a regression-discontinuity design. Econometrica, 69(1), pp.201-209.
Lee, D.S. and Lemieux, T., 2010. Regression discontinuity designs in economics. Journal of economic literature, 48(2), pp.281-355.

Regression Discontinuity Design

Assumption: Units just above and just below the cutoff are effectively ‘randomized’.

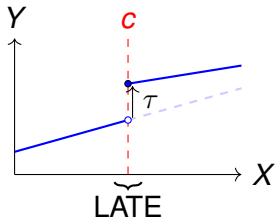
→ For units extremely close to the cutoff (e.g., 69.9 and 70.1 test score), their position is determined by idiosyncratic noise.

→ RDD estimates the *Local Average Treatment Effect* (**LATE**) near the cut-off.

Regression Discontinuity Design

Assumption: Units just above and just below the cutoff are effectively 'randomized'.
→ For units extremely close to the cutoff (e.g., 69.9 and 70.1 test score), their position is determined by idiosyncratic noise.
→ RDD estimates the *Local Average Treatment Effect* (**LATE**) near the cut-off.

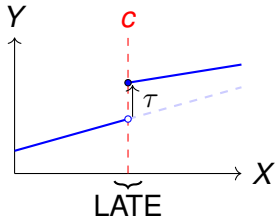
- **Running Variable (X):** The score determining assignment (e.g., Test Score).
- **Outcome (Y):** Long-term performance of the student.
- **Cutoff (c):** The threshold (e.g., Score ≥ 70).
- **Treatment (D):** Assigned if $X \geq c$.



Regression Discontinuity Design

Assumption: Units just above and just below the cutoff are effectively 'randomized'.
→ For units extremely close to the cutoff (e.g., 69.9 and 70.1 test score), their position is determined by idiosyncratic noise.
→ RDD estimates the *Local Average Treatment Effect* (**LATE**) near the cut-off.

- **Running Variable (X):** The score determining assignment (e.g., Test Score).
- **Outcome (Y):** Long-term performance of the student.
- **Cutoff (c):** The threshold (e.g., Score ≥ 70).
- **Treatment (D):** Assigned if $X \geq c$.



Local Average Treatment Effect (LATE) at the cutoff c:

$$\tau = \lim_{x \downarrow c} E[Y_i \mid X_i = x] - \lim_{x \uparrow c} E[Y_i \mid X_i = x]$$

Assumptions

"Near the cutoff, the only difference between treated and control is the treatment itself."

Assumptions

"Near the cutoff, the only difference between treated and control is the treatment itself."

Deterministic Rule: The cutoff c is usually a fixed policy
$\rightarrow$ e.g., a test score of 70, a poverty index, or a geographical border.

Assumptions

"Near the cutoff, the only difference between treated and control is the treatment itself."

Deterministic Rule: The cutoff c is usually a fixed policy
$\rightarrow$ e.g., a test score of 70, a poverty index, or a geographical border.

Continuity of Counterfactuals: All factors affecting Y must evolve smoothly across the threshold.
$\rightarrow$ Check that other variables (age, gender) do *not* 'jump' at c.

Assumptions

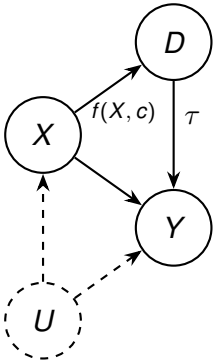
"Near the cutoff, the only difference between treated and control is the treatment itself."

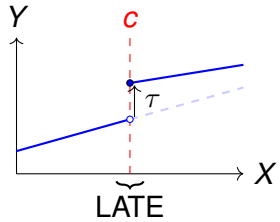
Deterministic Rule: The cutoff c is usually a fixed policy
$\rightarrow$ e.g., a test score of 70, a poverty index, or a geographical border.

Continuity of Counterfactuals: All factors affecting Y must evolve smoothly across the threshold.
$\rightarrow$ Check that other variables (age, gender) do *not* 'jump' at c.

No Manipulation (Non-Sorting): Units cannot precisely manipulate their running variable to 'choose' their side of the cutoff.
$\rightarrow$ *Test:* Check the density of X. If there is a 'clump' just above the cutoff, the design is broken.

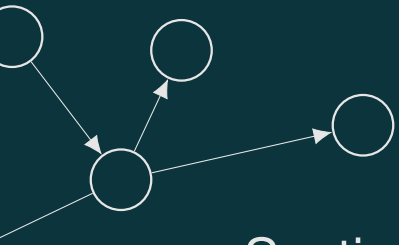
Graphical Perspective





Insight:

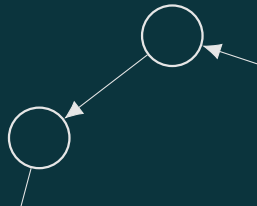
U does not point to D directly. By conditioning on X (at the limit $X \to c$), we block the back-door path through U.



Section

4

Granger Causality



Granger Causality

Strictly Temporal: Time-Series Predictive Causality.
$\rightarrow$ Leverage any predictive signal (whether causal or not).

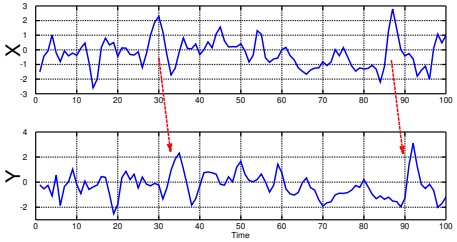
If some variable X 'Granger-Causes' Y, then past values of X contain information that helps predict Y beyond the past information contained in Y itself.

"Hearing the rooster crow helps us predict the sunrise."

Granger, C.W., 1969. Investigating causal relations by econometric models and cross-spectral methods. Econometrica: journal of the Econometric Society, pp.424-438.

Probabilistic Perspective

"Does X_i provide unique information about Y?"



Select feature X_i if $P(Y^{t+1} | \mathbf{X}^t) \neq P(Y^{t+1} | \mathbf{X}^t \setminus X_i^t)$

Figure: https://en.wikipedia.org/wiki/File:GrangerCausalityIllustration.svg, Creative Commons Attribution-Share Alike 3.0 Unported license.

Autoregression

"Does some X_i provide unique information about Y?"

Consider the time series X_i^t and Y^t.

Compare prediction error of two linear autoregressive models:

Baseline (Y onto itself):

$$Y^t = \sum_{i=1}^{k} \alpha_i Y^{t-i} + \epsilon_1^t$$

Extended Model (with additional X_i):

$$Y^t = \sum_{i=1}^{k} \alpha_i Y^{t-i} + \sum_{i=1}^{k} \beta_i X^{t-i} + \epsilon_2^t$$

Autoregression

"Does some X_i provide unique information about Y?"

Consider the time series X_i^t and Y^t.

Compare prediction error of two linear autoregressive models:

Baseline (Y onto itself):

$$Y^t = \sum_{i=1}^{k} \alpha_i Y^{t-i} + \epsilon_1^t$$

Extended Model (with additional X_i):

$$Y^t = \sum_{i=1}^{k} \alpha_i Y^{t-i} + \sum_{i=1}^{k} \beta_i X^{t-i} + \epsilon_2^t$$

X_i Granger-causes Y if the variance of the prediction error $\sigma^2(\epsilon_2)$ is *significantly smaller* than that of $\sigma^2(\epsilon_1)$.

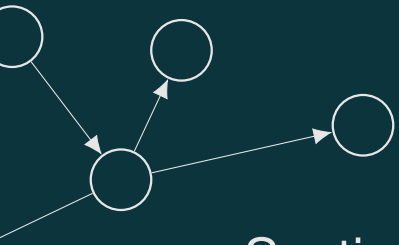
$\rightarrow$ Commonly determined via an F-Test.

Downsides of Granger Causality

Granger Causality is a prediction driven notion of causality.

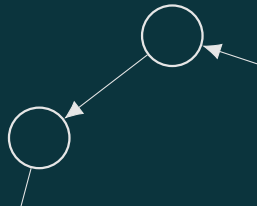
Easy to apply, but
does not handle confounders!

‘Granger causality does *not* imply that if X is intervened, Y will change.’



Section 5

Cyclic Causal Models



Why Cycles?

Most of this course assumed **Acyclic** Causal Models. But in nature we often encounter feedback cycles:

- **Biological Pathways:** Gene regulation networks where A inhibits B and B inhibits A.
- **Economic Equilibria:** Price affects Demand, which in turn affects Price.
- **Climate Systems:** Surface albedo and temperature create self-reinforcing cycles.

Problem: If $X \to Y$ and $Y \to X$ exist, the standard Markov Property and topological ordering of DAGs break down.

$\to$ We move from *recursive* models (step-by-step generation) to *non-recursive* models (simultaneous constraints).

Converging Systems

Where do cyclic SCMs actually come from?

$\rightarrow$ Cycles can be framed as projections of **Ordinary Differential Equations**.

$$\frac{d\mathbf{x}}{dt} = f(\mathbf{x}(t), \mathbf{u}, t)$$

Converging Systems

Where do cyclic SCMs actually come from?

→ Cycles can be framed as projections of **Ordinary Differential Equations**.

$$\frac{d\mathbf{x}}{dt} = f(\mathbf{x}(t), \mathbf{u}, t)$$

Equilibrium Points: If the system reaches a steady state ($\frac{d\mathbf{x}}{dt} = 0$), the differential equations collapse into Structural Equations.

$$0 = f(\mathbf{x}^*, \mathbf{u})$$

→ The fixed point $\mathbf{x}^*$ 'solves' the equations of $\mathcal{M}$.

Bongers, S. and Mooij, J.M., 2018. From random differential equations to structural causal models: The stochastic case. arXiv preprint arXiv:1803.08784, 3.
Iwasaki, Y. and Simon, H.A., 1994. Causality and model abstraction. Artificial intelligence, 67(1), pp.143-194.

Identification in Cycles

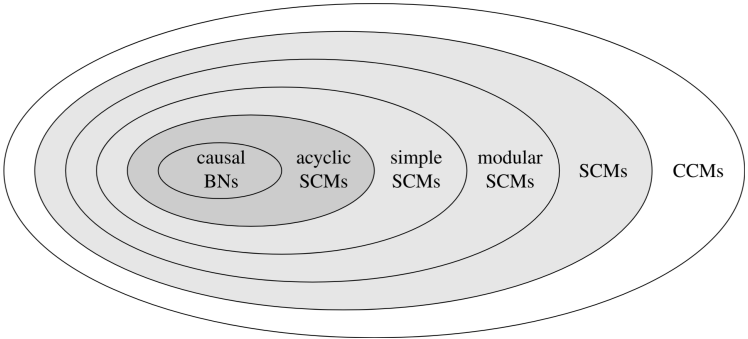
Discovery of cyclic causal models has been a long standing research topic.

In cyclic models, d-separation is often ‘too strong’.

$\rightarrow$ σ**-separation** [Forré and Mooji, 2017] is a criterion for cyclic faithfulness.

Spirtes, P., 1994. Conditional independence in directed cyclic graphical models for feedback. Carnegie Mellon [Department of Philosophy]. Forré, P. and Mooij, J.M., 2017. Markov properties for graphical models with cycles and latent variables. arXiv preprint arXiv:1710.08775.

Hierarchy of Structural Causal Models



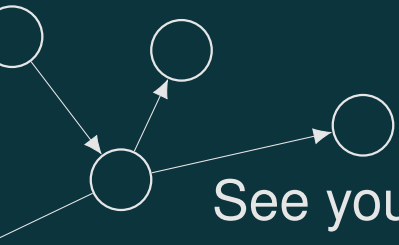
Simple SCM are uniquely solvable for (subsets of) the endogenous variables.

Modular SCM are closed under marginalization and intervention. Yield unique solutions under subsets of variables.

SCMs: Arbitrary structural equations between variables.

Constraint Causal Models: Contain 'instantaneous' equality constraints among variables.

Figure: Bongers, S., Forré, P., Peters, J. and Mooij, J.M., 2021. Foundations of structural causal models with cycles and latent variables. The Annals of Statistics, 49(5), pp.2885-2915.



See you next week!