PML 15. Structural Causal Models & The do-Calculus

Probabilistic Machine Learning Reading Group

Fátima Sánchez-Cabo

June 10, 2026

Computational Systems Biomedicine Lab, CNIC

Contents

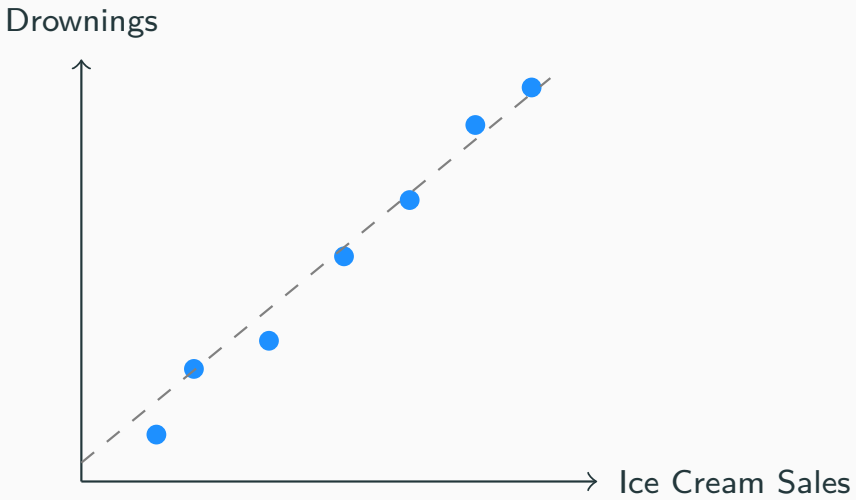
Contents

The Limits of Observational ML

- **Standard Machine Learning Goal:** Summarize relationships to predict on similar i.i.d. data using conditional probabilities: $P(Y|A = a)$.
- **Causal Inference Goal:** Predict behavior under external physical manipulation or policy intervention: $P(Y|do(A = a))$.
- **Fundamental Gap:** Multiple distinct physical mechanisms can generate the exact same joint observational distribution $P^{obs}(A, Y)$.
- **Epistemological Challenge:** Observational data alone cannot differentiate between causal structures like $A \rightarrow Y$, $Y \rightarrow A$, or $A \leftarrow U \rightarrow Y$.

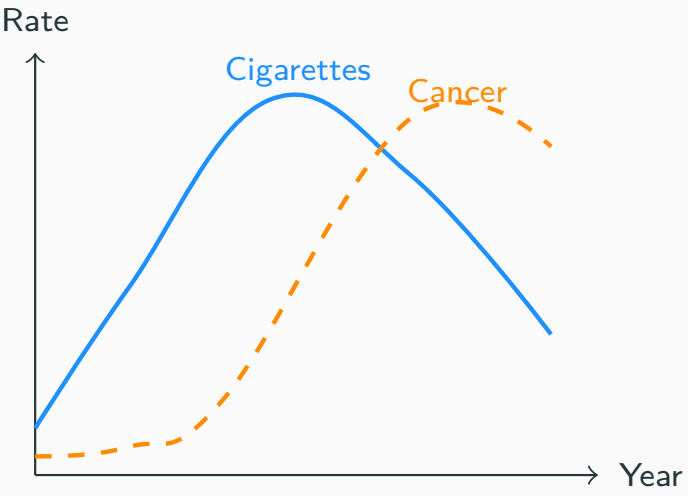
Correlation is Not Causation (I)

(a) Ice Cream & Drowning



Confounding by Season

(b) Smoking & Lung Cancer



Causal Lag (25 years)

Correlation is Not Causation (II)

(a) Drowning vs. Ice Cream

- Strong positive association observed across seasons.
- **Confounded by Temperature:** Warm weather increases both ice cream sales and swimming frequencies.
- Banning ice cream does *not* prevent drowning.

(b) Smoking vs. Lung Cancer

- Strong positive association with an approximate 25-year lag.
- Reflects a true underlying causal link.
- Intervening to ban or reduce smoking effectively reduces cancer mortality.

Remark: Predicting interventional outcomes requires moving beyond statistical association to an explicit structural hypothesis of the world.

High-Level Roadmap for Causal Estimation

1. **Causal Estimands**
   - Formally define the quantities of interest.
   - These summarize how the world changes under intervention (e.g., "Expected drownings if ice cream were banned").

2. **Identification**
   - Map causal estimands to quantities estimable from observational data.
   - Involves codifying causal knowledge (e.g., "Effect equals drownings after adjusting for month").

3. **Estimation**
   - Estimate the observable quantity using finite data samples.
   - Leverages non-parametric and ML predictive models to compute the identified quantity efficiently.

Lesson Focus

In this chapter, we'll mainly focus on the estimation of the causal effect of an intervention averaged over all members of a population, known as the **average treatment effect or ATE**

Contents

Contents

- The Epistemological Gap & SCMs
- Causal Formalism
  - Structural Causal Models
  - Causal DAGs Graph Surgery
  - Identification Ill-Posedness
  - Counterfactuals
- Randomized Control Trials
- Confounder Adjustment
  - Causal and Statistical Estimand, and Identification
  - ATE Estimation with Observed Confounders
  - Uncertainty Quantification & Matching
  - Practical Considerations and Procedures

Structural Causal Models: Factorization Limits

The Setup: Consider observing four variables: Smoking (A), Lung Cancer (Y), Health Consciousness (H), and Genetic Predisposition (G). Suppose we observe a dataset of these variables drawn independently and identically from a population,

$$(A_i, Y_i, H_i) \; P^{obs}$$

- A mathematically valid joint probability factorization can be written as:

$$P^{obs}(G, H, A, Y) = P^{obs}(G)P^{obs}(H|A)P^{obs}(Y|H, A)P^{obs}(G|A, H, Y$$

- The Issue: This ordering is completely agnostic to the underlying physical mechanisms. For instance, health consciousness (H) causally precedes and influences the decision to smoke (A).

Structural Causal Models: Formal Assignments

Formally, an SCM describes a mechanistic data-generating process via an ordered sequence of equations:

$$
\begin{aligned}
G &\leftarrow f_G(\xi_0) \\
H &\leftarrow f_H(\xi_1) \\
A &\leftarrow f_A(H, \xi_2) \\
Y &\leftarrow f_Y(G, H, A, \xi_3)
\end{aligned}
$$

Where:

- f_G, f_H, f_A, f_Y the (unknown) functions are fixed, describing invariant physical deterministic relationships in the real world.
- $\xi_0, \xi_1, \xi_2, \xi_3$ are unmeasured causes, modeled as independent random noise variables representing unique individual-level traits.

Structural Causal Models imply probabilistic models

But not the other way around. For example, our example SCM implies probabilistic model for the observed data based on the factorization:

$$P^{obs}(G, H, A, Y) = P^{obs}(G)P^{obs}(H)P^{obs}(A|H)P^{obs}(Y|H, A)$$

Thus we could sample from the SCM in the same way we would from a probabilistic model: draw a set of noise variables ξ and evaluate each assignment operation in the SCM in order.

Structural Causal Models and do-calculus

- Interventions are represented by replacing assignment statements. For example, if we were interested in the distribution of Y in the hypothetical scenario that smoking were eliminated, we could set the second line of the SCM to be $A \leftarrow 0$. We would denote this by

$$P(Y|do(A = 0), H)$$

.

- Because the f functions in the SCM are assumed to be invariant mechanistic relationships, the SCM encodes the assumption that this edited SCM generates data that we would see if we really applied this intervention in the world. Thus, the ordering of statements in an SCM imply substantive assumptions about how the world changes in response to interventions

Contents

Causal DAGs and Graph Surgery

- **Abstraction Level:** Causal DAGs isolate the structural connection framework from the explicit functional mechanics (f).
- **The Mechanism Invariance Assumption:** Intervening on a variable A eliminates its dependencies on its natural parents.
- **Graph Surgery (do-operator):** An intervention setting $A = a$ physically mutates the graph by completely removing all incoming parent edges into A.
- **DAGs relate to SCM** as Probabilistic Graphical Models (PGM) are related to probabilistic models
- **SCMs obey Markov property** If $X \leftarrow Y \rightarrow Z$ then $X \perp\!\!\!\perp Z|Y$
- **SCMs recall d-separation**
- **Y is a collider** if $X \rightarrow Y \leftarrow Z$ then $X \perp\!\!\!\perp Z|Y$

Origins of Statistical Dependencies

Without specifying a full SCM, a causal DAG abstracts how two variables (A and D) become statistically associated:

1. **Causal Paths (Directed Ancestry)**
   - There is a directed path from ancestor A to descendant D ($A \to \cdots \to D$).
   - Interventions on A actively propagate through the network to change D: $P(D \mid do(A = a)) \neq P(D \mid do(A = a'))$
2. **Non-Causal Paths (Common Confounding)**
   - A and D share a common cause C via a backdoor path ($A \leftarrow C \to D$).
   - Interventions on A do *not* alter D:

$$P(D \mid do(A = a)) = P(D \mid do(A = a')) \implies D \perp\!\!\!\perp do(A)$$

   - Yet, A carries information from C which impacts in D, hence: $P(D \mid A = a) \neq P(D \mid A = a')$

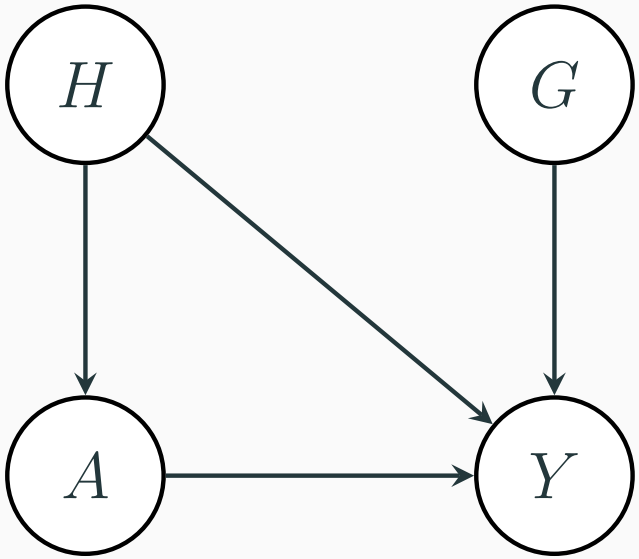
The Graph Surgery Principle

A causal DAG isolates structural dependencies without forcing the specification of full parametric functional forms of an SCM.

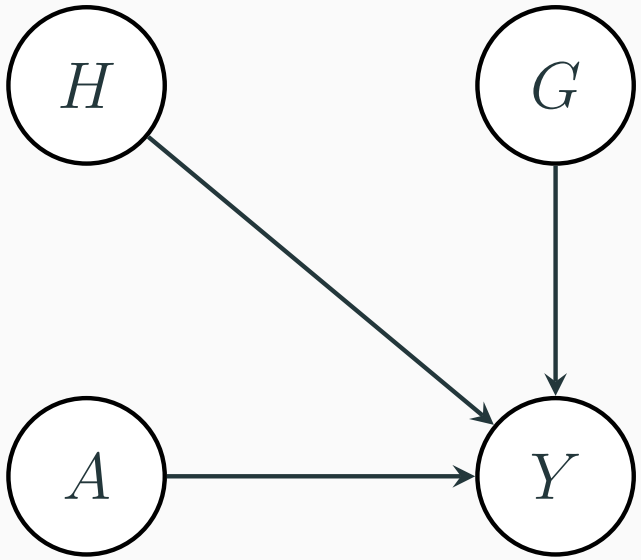
- **Mechanism Erasure via Intervention**
  - When an equation in an SCM is replaced by a fixed value assignment $(A \leftarrow a)$, treatment A is entirely stripped of its prior causal dependencies.
  - Consequently, after an intervention, none of the system's other background variables function as active causes of A.
- **Graphical Definition of Graph Surgery**
  - Setting an intervention $do(A = a)$ physically deletes all incoming parental edges to node A in the graph.
  - Downstream inference is then evaluated by conditioning on $A = a$ using the modified conditional independence structure of the post-surgery graph.

Visualizing Graph Surgery: G vs. $G_{\overline{A}}$

Original Causal Graph G



Interventional Graph $G_{\overline{A}}$



Pearl's do-Notation and ATE Semantics

- **Mathematical Definition of the Interventional Query**
  - $P(X \mid do(A = a))$ denotes the distribution of X evaluated under the mutilated post-surgery graph where all incoming edges to A are deleted.
  - The interventional expectation is defined as:

$$\mathbb{E}[X \mid do(A = a)] \triangleq \mathbb{E}_{P(X|do(A=a))}[X]$$

  - This formalizes causal goals like the Average Treatment Effect (ATE):

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

- **Conditioning vs. Intervening Matrix Factorization**
  - P(Y, H, G $\mid A = a) = P(Y \mid H, G, A = a)P(G)P(H \mid A = a)$
  - P(Y, H, G $\mid do(A = a)) = P(Y \mid H, G, A = a)P(G)P(H)$

Pearl's do-Notation and ATE Semantics

- **The Foundational Population Shift**
  The standard conditional distribution describes a population where health consciousness H has the distribution that we observe among individuals with smoking status A = a, while the interventional distribution described a population where health consciousness H follows the marginal distribution among all individuals
  - $P(H \mid A = \text{smoker})$ shifts mass toward lower values of health consciousness (H) based entirely on passive observation.
  - $P(H)$ preserves the total population marginal profile. The do-distribution thus tests a hypothesis of how smoking affects even the highly health-conscious subpopulation who choose not to smoke in the observed data.

Contents

The Core Challenge of Identification

The Ill-Posedness of Causal Inference

Infinitely many different underlying SCMs can generate identical joint observational data distributions P^{obs}. Data alone cannot distinguish between structures without external assumptions. We have to augment the data with knowledge.

- **Causal Estimand:** The target mathematical query containing unobserved intervention behaviors (e.g., $\mathbb{E}[Y|do(A=1)]$).
- **Statistical Estimand:** A rearranged formula written entirely in terms of observable conditional probabilities from the data.
- **Identification:** The strict process of proving that structural assumptions are sufficient to map a causal estimand uniquely to a statistical estimand.

Statistical Estimands Framework

- Under the specific structural parental properties of the SCM of our example, the Population Average Treatment Effect (ATE):

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

- Can be proven by construction to be uniquely identified as the following **Statistical Estimand** (τ^{ATE}):

$$\tau^{ATE} \triangleq \mathbb{E}\Big[\mathbb{E}[Y|H, A=1] - \mathbb{E}[Y|H, A=0]\Big]$$

- **Crucial Insight:** The right-hand side contains absolutely no do-operators and can be directly calculated from data distributions.

The Core Challenge of Identification

The Ill-Posedness of Causal Inference

Every causal identification strategy relies on assumptions that have some content that cannot be validated in the observed data. If the assumptions used to identify causal quantities could be validated, that would imply that the causal estimand was identifiable from the observed data alone.

Contents

Counterfactuals and the Causal Hierarchy

SCM let us formalize and study a hierarchy of different kinds of query about a system under consideration:

- **Observational queries:** Are smoking and lung cancer associated?
- **Interventional queries:** How much does smoking increase the probability of a cancer in a given population?
- **Conterfactual queries:** Would Alice have developed cancer if she hadn´t smoked?

Counterfactuals and the Causal Hierarchy

- **Interventional Queries:** Prospective, population-level inquiries regarding data behavior under active structural manipulation. Evaluated with marginal potential outcome tracking.
- **Counterfactual Queries:** Retrospective, individual-level queries modeling how a past event would have changed had an alternate path been chosen.
- **Potential Outcome Notation:** Individual i's response to treatment setting a keeping personal noise vectors $\xi_{3,i}$ entirely frozen. In our smoking example:

$$Y_i(a) \triangleq f_Y(G_i, H_i, a, \xi_{3,i})$$

Conterfactual vs interventional queries

- Just because a query can be written in terms of potential outcomes does not make it a conterfactual query.
- The key dividing line between interventional and conterfactual queries is whether the query requires knowing the joint distribution of potential outcomes within individuals, or whether marginal distributions of potential outcomes across individuals will suffice
- While the population-level $ATE = \mathbb{E}[Y_i(1) - \mathbb{E}[Y_i(0)]]$ requires only marginal information, true counterfactual definitions dictate knowing the unified joint behavior of multiple physical worlds.

Contents

Enforcing the No-Common-Causes Structure

- The simplest context for causal estimation occurs when there are absolutely no common causes influencing both treatment A and outcome Y. In the real world, systems are rarely unconfounded; however, we can design an experiment to physically enforce this no-common-causes condition.
- By assigning each participant to either the treatment or control group completely at random:
  - Treatment assignment does not depend on any baseline traits.
  - All incoming parental edges to A are eliminated by design.
- This structural isolation makes the interventional distribution equal to the raw observational conditional distribution:

$$P(Y \mid do(A = a)) = P(Y \mid A = a)$$

ATE Identification and the Sample Estimator

- Randomized Control Trials are primarily used to study the Population Average Treatment Effect (ATE):

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

- Under a valid randomization design, this unobserved causal parameter simplifies into an observational target parameter τ^{RCT}:

$$\tau^{RCT} = \mathbb{E}[Y \mid A = 1] - \mathbb{E}[Y \mid A = 0]$$

- **The Standard Sample Estimator:** Given finite data, the empirical difference in group means is defined as:

$$\hat{\tau}^{RCT} \triangleq \frac{1}{n_A} \sum_{i:A_i=1} Y_i - \frac{1}{n - n_A} \sum_{i:A_i=0} Y_i$$

where n_A represents the number of units who received treatment.

Gold Standard Status vs. Practical Limitations

- **The Gold Standard:** RCTs are uniquely powerful because the structural triangle layout shown on the right is explicitly broken by design.
- Core Limitations in Practice:
  - **Infeasibility:** Often restricted by high expenses, regulatory hurdles, or long timing delays.
  - **Selection Biases:** Trial cohorts often fail to mirror deployment populations (e.g., participants can skew younger or poorer).

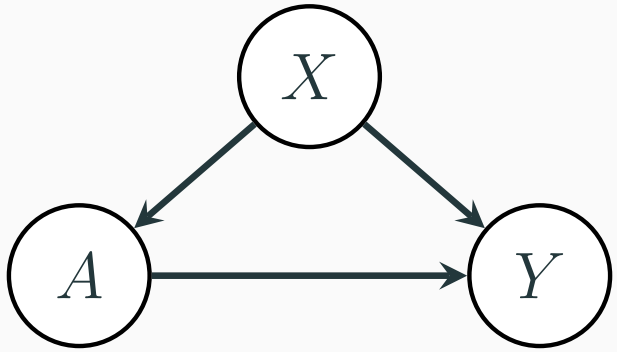
Contents

Contents

- The Epistemological Gap & SCMs
- Causal Formalism
  - Structural Causal Models
  - Causal DAGs Graph Surgery
  - Identification Ill-Posedness
  - Counterfactuals
- Randomized Control Trials
- Confounder Adjustment
  - Causal and Statistical Estimand, and Identification
  - ATE Estimation with Observed Confounders
  - Uncertainty Quantification & Matching
  - Practical Considerations and Procedures

The Triangle Confounding Framework

- **Observational Settings:** Unlike RCTs, observational studies suffer from **confounders**—common causes affecting both treatment A and outcome Y.



Confounder DAG

- **Causal Estimand:** We wish to learn the Average Treatment Effect (ATE), formalizing prospective policy queries:

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

- **Mechanism Invariance:** Because $do(A = a)$ only cuts the $X \to A$ edge, the structural conditional expectation mapping $A \to Y \leftarrow X$ remains completely stable.

Theorem 2: Backdoor Adjustment

Theorem (Adjustment with no unobserved confounders) *Suppose X contains all common causes of A and Y, and no variable in X is caused by A or Y. Assuming overlap ($0 < P(A = 1 \mid X = x) < 1$), the ATE is identified as:*

$$\tau = \mathbb{E}\Big[\mathbb{E}[Y \mid A = 1, X] - \mathbb{E}[Y \mid A = 0, X]\Big]$$

The Tower Property Proof:

$$\begin{aligned} ATE &= \mathbb{E}[Y \mid do(A = 1)] - \mathbb{E}[Y \mid do(A = 0)] \\ &= \mathbb{E}\Big[\mathbb{E}[Y \mid do(A = 1), X]\Big] - \mathbb{E}\Big[\mathbb{E}[Y \mid do(A = 0), X]\Big] \\ &= \mathbb{E}\Big[\mathbb{E}[Y \mid A = 1, X]\Big] - \mathbb{E}\Big[\mathbb{E}[Y \mid A = 0, X]\Big] \end{aligned}$$

The final step eliminates all do-operators because $P(Y \mid A, X)$ is identical in both the original and mutilated graphs.

The Causal vs. Naive Integral Shift

Writing the statistical estimand explicitly highlights the structural difference from a naive association difference:

Causal Target Estimand (τ)

$$\tau = \int \mathbb{E}[Y \mid A = 1, X] P(X)\, dX - \int \mathbb{E}[Y \mid A = 0, X] P(X)\, dX$$

Naive Difference in Means (Association)

$$\mathbb{E}[Y \mid A = 1] - \mathbb{E}[Y \mid A = 0] =$$
$$\int \mathbb{E}[Y \mid A = 1, X] P(X \mid A = 1)\, dX - \int \mathbb{E}[Y \mid A = 0, X] P(X \mid A = 0)$$

Contents

Outcome Model Adjustment (g-Computation)

- **Definition:** Let the conditional expected outcome function be Q:

$$Q(a, x) \triangleq \mathbb{E}[Y \mid A = a, X = x]$$

- By substitution into Theorem 2, $\tau = \mathbb{E}[Q(1, X) - Q(0, X)]$. This yields the sample empirical average estimator:

$$\hat{\tau}^{Q} \triangleq \frac{1}{n} \sum_{i=1}^{n} \left(\hat{Q}(1, x_i) - \hat{Q}(0, x_i) \right)$$

- **Optimization:** To fit $\hat{Q}$, we minimize standard Mean Squared Error (MSE) loss using any predictive regressor.
- **Parametric Special Case:** If we assume a linear model $Q(A, X) = \beta_0 + \beta_A A + \beta_X X$, then $Q(1, X) - Q(0, X) = \beta_A$. The procedure collapses to reporting the regression coefficient β_A.

Propensity Score Adjustment (IPTW)

- **Definition:** Rather than modeling the outcome, we model the treatment selection mechanism via the propensity score $g(X)$:

$$g(x) \triangleq P(A = 1 \mid X = x)$$

- Using the identity $\mathbb{E}[YA \mid X] = \mathbb{E}[Y \mid A = 1, X]P(A = 1 \mid X)$, we can mathematically re-weight the statistical estimand as:

$$\tau = \mathbb{E}\left[\frac{YA}{g(X)} - \frac{Y(1-A)}{1-g(X)}\right]$$

- This implies the Inverse Probability of Treatment Weighted (**IPTW**) estimator:

$$\hat{\tau}^{IPTW} \triangleq \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i A_i}{\hat{g}(x_i)} - \frac{Y_i(1-A_i)}{1-\hat{g}(x_i)}\right)$$

Double Machine Learning

- **The Augmented IPTW (AIPTW) Estimator:**

$$\hat{\tau}^{AIPTW} \triangleq \frac{1}{n}\sum_{i=1}^{n}\left(\hat{Q}(1,X_i) - \hat{Q}(0,X_i) + \frac{A_i(Y_i - \hat{Q}(1,X_i))}{\hat{g}(X_i)} - \frac{(1 -$$

- **Double Robustness:** Consistency is fully preserved as long as *at least one* of the nuisance models ($\hat{Q}$ or $\hat{g}$) is correctly specified.
- **Product Error Rate:** The influence curve framing shows that the misestimation penalty is the product of nuisance model errors:

$$\mathbb{E}[\phi(\hat{Q},\hat{g}) - \phi(Q,g)] = \mathbb{E}\left[\frac{1}{g(X)}(\hat{g}-g)(\hat{Q}_1 - Q_1) + \frac{1}{1-g(X)}(\hat{g} - g\right.$$

- If this product is $o(\sqrt{n})$, we attain parametric $\sqrt{n}$ convergence rates even using flexible non-parametric ML algorithms.

Contents

Uncertainty Quantification

- From the asymptotic normality of the double ML estimator ($\sqrt{n}(\hat{\tau} - \tau) \xrightarrow{d} \mathcal{N}(0, \mathbb{E}[\phi^2])$), we can compute valid analytical variances directly:

$$\hat{\mathbb{V}}[\hat{\tau}] = \frac{1}{n} \sum_{i=1}^{n} \phi(X_i; \hat{Q}, \hat{g}, \hat{\tau})^2$$

- This yields traditional 95% Wald confidence intervals: $\hat{\tau} \pm 1.96\sqrt{\hat{\mathbb{V}}[\hat{\tau}]/n}$.

Matching

- Models sample cell comparisons by mapping treated units to control units within specific covariate strata.
- For discrete domains, the estimator can be written as:

$$\hat{\tau}^{\text{matching}} = \sum_x \hat{P}(x) \left(\frac{1}{|\mathcal{A}_x|} \sum_{i \in \mathcal{A}_x} Y_i - \frac{1}{|\mathcal{C}_x|} \sum_{j \in \mathcal{C}_x} Y_j \right)$$

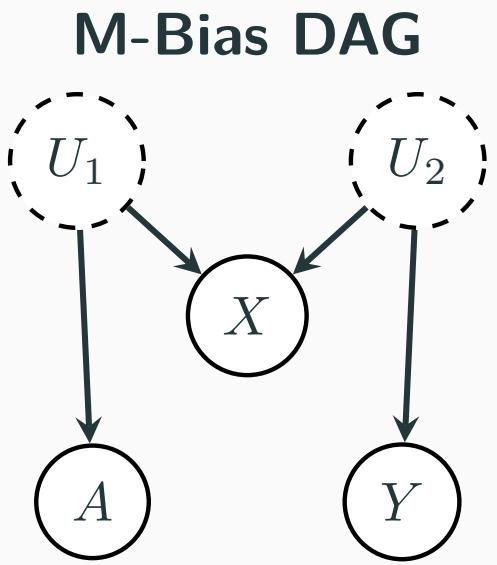
- Matching represents a non-parametric choice of outcome model adjustment where expectations are estimated via conditional cell samples.

Contents

- The Epistemological Gap & SCMs
- Causal Formalism
  - Structural Causal Models
  - Causal DAGs Graph Surgery
  - Identification Ill-Posedness
  - Counterfactuals
- Randomized Control Trials
- Confounder Adjustment
  - Causal and Statistical Estimand, and Identification
  - ATE Estimation with Observed Confounders
  - Uncertainty Quantification & Matching
  - Practical Considerations and Procedures

Confounder Selection Pitfalls

- **The Mediator Pitfall:** Conditioning on intermediate variables ($A \to M \to Y$) blocks legitimate causal pathways, wiping out portions of the total effect.
- **The M-Bias Collider Pitfall:** If an observed covariate X acts as a collider between two independent hidden factors ($U_1 \to X \leftarrow U_2$), conditioning on X opens a spurious backdoor path.



The VS Heuristic (2011): Explicitly condition on all pre-treatment variables that act as a cause of the treatment, the outcome, or both.

Overlap Truncation & Choice of Estimand

- **Overlap Failure Resolutions:** When regions contain extreme propensities ($\hat{g}(x) \approx 0$ or 1), we can discard data outside shared profiles. This restricts evaluation safely onto a specific, overlapping subpopulation.
- **Average Treatment Effect on the Treated (ATT):** Useful when treated units self-select from a small profile group, making treatment effects impossible to estimate for certain rare control profiles:

$$ATT \triangleq \mathbb{E}_{X|A=1}\Big[\mathbb{E}[Y \mid X, do(A=1)] - \mathbb{E}[Y \mid X, do(A=0)]\Big]$$

- Under Figure 36.3 and a weaker overlap restriction ($P(A=1 \mid X) < 1$), the ATT identifies as:

$$\tau^{ATT} = \mathbb{E}_{X|A=1}\Big[\mathbb{E}[Y \mid A=1, X] - \mathbb{E}[Y \mid A=0, X]\Big]$$

Overlap Truncation & Choice of Estimand

- **The Double ML Estimator for ATT:**

$$\hat{\tau}^{\text{ATT-AIPTW}} \triangleq \frac{1}{n}\sum_{i=1}^{n}\left(\frac{A_i(Y_i-\hat{Q}_0)}{P(A=1)} - \frac{(1-A_i)\hat{g}(X_i)(Y_i-\hat{Q}_0)}{P(A=1)(1-\hat{g}(X_i))}\right)$$

Contents

Summary and Practical Advice

- **No Empirical Validation Labels:** Unlike predictive ML where models are validated on out-of-sample data, causal estimands target fields we never explicitly observe simultaneously (the counterfactual arms).
- **Multi-Model Sensitivity Checks:** Run multiple distinct estimators (Outcome regression, IPTW, and Double ML) side-by-side using regularized learners.
  - If estimators agree qualitatively, it provides stability reassurance.
  - If they diverge significantly, check for underlying failures of overlap or misspecified nuisance distributions.
- **The Core Imperative:** Statistical machinery cannot bypass the need to justify the unobserved structural assumption that all confounders are fully recorded.

RL Reading Group

Reinforcement Learning Reading Group

Oct-Dec 2026

Next Reading Group

Join our next Reading Group on Reinforcement Learning!

Starting next autumn!

Connect & Subscribe

- **Website:**
  `https://rl-readinggroup.github.io/`
- **Mailing list:**
  `https://listas.csic.es/wws/subscribe/rl_reading_group?previous_action=edit_list_request`