ECMA 31380 - Proposal Final Project

Fernando Rocha Urbano

Autumn 2024

1 Goal

The goal of this project is to evaluate the performance of different Double Machine Learning (DML) models in recovering the Average Treatment Effect (ATE) across various causal inference scenarios addressed in simulations.

The desired broader impact of the project is to provide actionable recommendations for selecting the most appropriate DML method based on the causal scenario, data structure, noise levels, and sample size. Additionally, we aim to create a benchmarking framework that practitioners can use to evaluate and compare causal inference estimates derived from the tested methods in different scenarios.

The DML methods compared in this study are:

- LASSO.
- Random Forest.
- Neural Networks.

The causal inference scenarios are:

- Backdoor adjustment.
- Frontdoor adjustment with and without access to mediators.
- Instrumental Variables with varying levels of instrument strength and correct vs. incorrect use of the IV.

We explore how each model implementation performs under varying conditions:

- Relationships as linear and non-linear between covariates and treatment and between covariates and target ($g(\cdot)$ and $m(\cdot)$).
- Noise levels (ε).
- Sample size (n).
- Size of high-dimensional impactful covariates (d_c).

- Size of high-dimensional irrelevant covariates (d_a).

For each of the scenario combinations, we show which of LASSO, Random Forest and Neural Networks provides more accurate ATE estimates based on the following metrics.

- Bias of $\hat{\text{ATE}}$.
- Variance of $\hat{\text{ATE}}$ across repeated simulations for a given scenario.
- Analytical Variance of the $\hat{\text{ATE}}$.
- MSE (Mean Squared Error).

2 Methodology

2.1 Problem Setup

Consider a random sample $(Y_i, T_i, X_i)_{i=1}^n$, where:

- $Y_i \in \mathbb{R}$ is the outcome variable.
- $T_i \in 0, 1$ is a binary treatment indicator.
- $X_i \in \mathbb{R}^p$ is a vector of covariates.

Our goal is to estimate the Average Treatment Effect (ATE), defined as:

$$\tau = \mathbb{E}[Y_i(1) - Y_i(0)], \tag{1}$$

where $Y_i(t)$ denotes the potential outcome for unit i under treatment $T_i = t$.

2.2 Identification via Conditional Expectations

Under the Conditional Independence Assumption (CIA) and overlap conditions, the ATE can be identified as:

$$\tau = \mathbb{E}[\mu_1(X_i) - \mu_0(X_i)], \tag{2}$$

where $\mu_t(X_i) = \mathbb{E}[Y_i|T_i = t, X_i]$ is the conditional expectation of the outcome given treatment and covariates.

2.3 Double Machine Learning (DML) Estimator

The DML framework aims to estimate τ while controlling for high-dimensional or complex relationships between Y_i, T_i, and X_i. The key idea is to use machine learning methods to estimate the nuisance parameters and then construct an estimator for τ that is robust to estimation errors in these nuisance parameters.

2.3.1 Nuisance Parameter Estimation

We define the following nuisance functions:

$$m(X_i) = \mathbb{E}[Y_i|X_i], \tag{3}$$
$$g(X_i) = \mathbb{E}[T_i|X_i], \tag{4}$$
$$\pi(X_i) = \mathbb{P}(T_i = 1|X_i) \quad \text{(propensity score)} \tag{5}$$

These functions can be estimated using flexible machine learning methods such as LASSO, Random Forests, or Neural Networks.

Under mild regularity conditions and appropriate rates of convergence for the nuisance estimators, the DML estimator is root-n consistent and asymptotically normal. It provides valid confidence intervals and hypothesis tests even when using complex machine learning methods for $\hat{m}(X)$ and $\hat{g}(X)$.

2.3.2 Orthogonal Score Function

To achieve robustness, we construct an orthogonal score function $\psi(Y_i, T_i, X_i; \eta)$, where $\eta = (m, g)$ represents the nuisance parameters. The orthogonal score satisfies the Neyman orthogonality condition, which ensures that small estimation errors in η have a negligible first-order impact on the estimation of τ.

A common choice for the orthogonal score is:

$$\psi(Y_i, T_i, X_i; \eta) = \left(\frac{T_i - g(X_i)}{\pi(X_i)(1 - \pi(X_i))} \right) (Y_i - m(X_i)) + (m_1(X_i) - m_0(X_i)) - \tau, \tag{6}$$

where $g(X_i) = \pi(X_i)$ in case of binary treatment and:

$$m_t(X_i) = \mathbb{E}[Y_i|T_i = t, X_i] \quad \text{for } t \in \{0, 1\} \tag{7}$$

2.3.3 Estimation Procedure

The estimation proceeds in several steps:

1. Estimate Nuisance Functions: Use one of the three outline ML models to obtain estimators $\hat{m}(X_i)$ and $\hat{g}(X_i)$ for the nuisance functions.
2. Compute the Score Function: Evaluate the orthogonal score $\psi(Y_i, T_i, X_i; \hat{\eta})$ using the estimated nuisance parameters.
3. Estimate τ: Solve the empirical moment condition which yields $\hat{\tau}$:

$$\frac{1}{n} \sum_{i=1}^{n} \psi(Y_i, T_i, X_i; \hat{\eta}) = 0. \tag{8}$$

The orthogonal score function mitigates the impact of errors in $\hat{m}(X)$ and $\hat{g}(X)$ on the estimation of τ while also accomodating the use of modern ML techniques, allowing complex relationships between Y, X and T.

To prevent overfitting and ensure that the estimation error in the nuisance parameters does not bias the estimator of τ we employ cross-fitting.

2.3.4 Cross-Fitting

The steps of cross-fitting are:

1. Split the Sample: Divide the data into K folds $\{\mathcal{I}_k\}k = 1^K$.
2. For Each Fold:
   (a) Train Nuisance Estimators: Use data from all other folds:
   $$\mathcal{I}_{-k} = \bigcup j \neq k\mathcal{I}_j$$
   to estimate $\hat{m}^{(-k)}(X_i)$ and $\hat{g}^{(-k)}(X_i)$.
   (b) Compute Score Function: For observations in fold $\mathcal{I}_k$, compute $\psi(Y_i, T_i, X_i; \hat{\eta}^{(-k)})$ using the nuisance estimates from step (a):
   $$\begin{aligned}\psi(Y_i, T_i, X_i; \hat{\eta}^{(-k)}) = &\left(\frac{T_i - \hat{g}^{(-k)}(X_i)}{\hat{\pi}^{(-k)}(X_i)}\right)(Y_i - \hat{m}^{(-k)}(X_i)) \\ &+ \hat{m}_1^{(-k)}(X_i) - \hat{m}_0^{(-k)}(X_i) - \tau^{(-k)}.\end{aligned} \tag{9}$$
3. Aggregate: Combine the estimates from all folds:
   $$\hat{\tau} = \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\hat{\tau}_i^{(-k)} = \frac{1}{K}\sum_{k=1}^{K}\hat{\tau}^{(-k)} \tag{10}$$

3 Causal Inference Scenarios

The three most relevant scenarios of causal inference are the ones that require Instrumental Variables, Backdoor Adjustment, or Front Door Adjustment. A common way to represent the causal relation related to those is through the use of DAGs.

3.1 DAG (Directed Acyclic Graph)

Directed Acyclic Graph (DAG) serves as a representation of causal assumptions and a tool for deriving statistical properties of the variables involved.

It is composed of nodes (vertices) representing random variables or features and directed edges (arrows), indicating a direct influence or causal effect of T on Y.

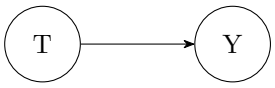

Figure 1: DAG Example

The acyclic characterist is due to absence of directed cycles; that is, there is no path where you can start at a node X and, by following directed edges, return to X.

3.2 Backdoor Adjustment

A backdoor path from treatment T to the outcome Y represents alternative routes through which association can flow from T to Y that are not due to the causal effect of T on Y. In the representation, the confounder C is a common cause of both T and Y, thus, a backdoor path.

In such case, the association between T and Y may be partially or entirely due to their mutual dependence on C rather than a direct causal effect, leading to biased causal estimates of the treatment if C is ignored.

The causal effect of T on Y can be expressed using the backdoor adjustment formula:

$$\mathbb{P}[Y(t)] = \sum_{C} \mathbb{P}[Y \mid T, C]\mathbb{P}[C], \tag{11}$$

Which serves a markov factorization, calculating with respect to the DAG structure.

For the backdoor adjustment to be valid, the following conditions must be satisfied:

1. No variable in the adjustment set is a descendant of the treatment T.
2. The adjustment set blocks all backdoor paths from T to Y (backdoor path is any path from T to Y that starts with an arrow into T).

DAG 3.2 illustrates the backdoor path involving the confounder C, treatment T, outcome Y, and features with non-causal association X_a which would not be present in a typical backdoor adjustment DAG, but play a relevant role in the simulations proposed.

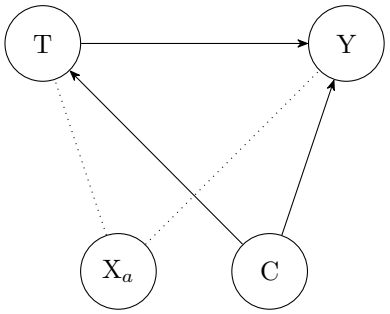

Figure 2: DAG of Backdoor Path

3.3 Frontdoor Adjustment

Frontdoor adjustment is a method used to estimate the causal effect of treatment T on outcome Y when there is unmeasured confounding that cannot be addressed using backdoor adjustment. It leverages a mediator M that lies on the causal path from T to Y.

The causal effect of T on Y can be expressed using the frontdoor adjustment formula:

$$\mathbb{P}[Y(t)] = \sum_{M} \mathbb{P}[M \mid T = t] \sum_{t'} \mathbb{P}[Y \mid M, T = t'] \mathbb{P}[T = t']. \tag{12}$$

For the frontdoor adjustment to be valid, the following conditions must be satisfied:

1. All causal paths from T to Y pass through M (i.e., there is no direct effect of T on Y bypassing M).
2. There are no unmeasured confounders between T and M.
3. All backdoor paths from M to Y are blocked by T (i.e., there are no unmeasured confounders between M and Y that are not affected by T).

DAG 3.3 illustrates the frontdoor adjustment involving the confounder C, treatment T, mediator M, outcome Y, and features with non-causal association X_a, which are included due to their relevance in the proposed simulations.

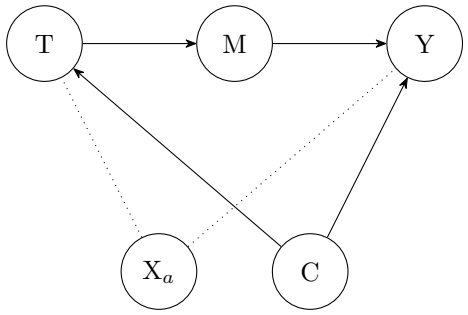

Figure 3: DAG of Frontdoor Path

3.4 Instrumental Variable

Instrumental variable (IV) estimation is a method used to estimate the causal effect of a treatment T on an outcome Y when there is unmeasured confounding that cannot be addressed using backdoor or frontdoor adjustments. This method leverages an instrument Z , which influences the treatment T but has no direct effect on the outcome Y except through T , and is independent of any unmeasured confounders U affecting both T and Y .

The causal effect of T on Y can be estimated using the instrumental variable formula:

$$\text{Causal Effect} = \frac{\text{Cov}(Z, Y)}{\text{Cov}(Z, T)}. \tag{13}$$

Alternatively, in terms of expectations for a binary instrument Z :

$$\text{Causal Effect} = \frac{\mathbb{E}[Y \mid Z = 1] - \mathbb{E}[Y \mid Z = 0]}{\mathbb{E}[T \mid Z = 1] - \mathbb{E}[T \mid Z = 0]}. \tag{14}$$

For the instrumental variable method to be valid, the following conditions must be satisfied:

1. Relevance: The instrument Z is associated with the treatment T (i.e., $\text{Cov}(Z, T) \neq 0$ or $Z \not\perp\!\!\!\perp T$).
2. Exclusion Restriction: The instrument Z affects the outcome Y only through its effect on the treatment T (i.e., there is no direct effect of Z on Y and no other pathways from Z to Y except through T).
3. Independence (Ignorability): The instrument Z is independent of any unmeasured confounders U that affect both T and Y (*i.e.*, $Z \perp\!\!\!\perp U$).

DAG 3.4 illustrates the instrumental variable setup involving the unobserved confounder U, instrument Z, treatment T, outcome Y, observed confounder C, and features with non-causal associations X_a, which are included due to their relevance in the proposed simulations.

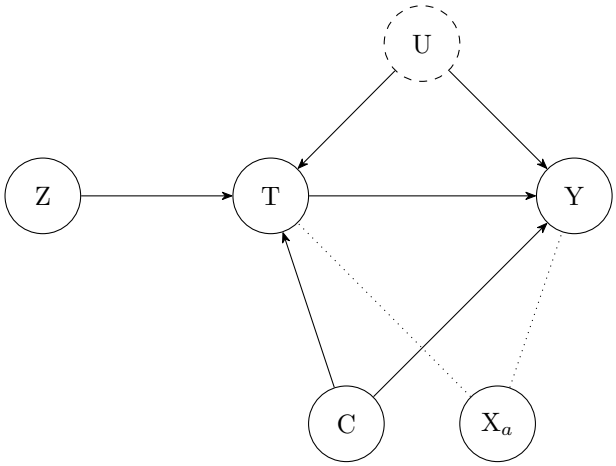

Figure 4: DAG of Instrumental Variable