Comparative Analysis of Double Machine Learning Methods Across Causal Inference Frameworks

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.

Ideally, the project should provide actionable recommendations for selecting the most appropriate DML method based on the causal scenario, data generating process (linear vs. non-linear), noise levels, sample size, sparsity of causal covariates, number of causal and non-causal covariates. 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 main DML models used to evaluate are OLS, Random Forest, Neural Network, LASSO, and Elastic Net with correct and incorrect specifications. The causal scenario evaluated (which are based on different data generating processes) are Backdoor Path, Frontdoor Path, and Instrumental Variables. Covariates dimension, noise levels, and sample size are also varied to evaluate differences in performance for each combination of scenario, model, and specification.

The main metric evaluated is the capacity to recover the true Average Treatment Effect (ATE) in each scenario. In order to achieve this goal, we calculate bias, variance, mean squared error (MSE) and confidence interval coverage of the ATE estimates.

Below, we highlight some of the questions we are able to answer with the proposed simulations:

1. Which model performs better for different ratio of number of causal (d_c) and non-causal covariates (d_a)?
2. Which model performs better for high-dimensional data?
3. Which model performs better for each of the causal inference scenarios (Backdoor, Frontdoor, and IV)?
4. Which model performs better overall for the estimation of treatment, outcome, and instrument?
5. Which model has the best convergence rate?
6. Which model performs better for different noise levels (σ_z, σ_t, σ_y)?

7. Do bias, variance, MSE and confidence interval coverage metrics point to the same ranking of models?
8. Which model performs better for different levels of sparsity between covariates (level of correlation between covariates)?
9. Which model performs better when there is misspecification in the nuisance functions?
10. Which model performs better when there is unobserved confounders (U_i) not corrected by instrument (a type of misspecification)?

2 Double Machine Learning

Brief introduction of DML history

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, and $X_i \in \mathbb{R}^d$ is a vector of covariates.

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

$$\tau = \mathbb{E}[Y_i(1) - Y_i(0)], \tag{2.1}$$

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

Probably expand explanation on this section, specifically on the issues in achieving potential outcome estimates.

2.2 Identification via Conditional Expectations

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

$$\tau = \mathbb{E}[\mu_1(X_i) - \mu_0(X_i)], \tag{2.2}$$

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 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.

More explanation on what are nuisance functions.

2.3.1 Nuisance Parameter Estimation

We define the following nuisance functions:

$$m(X_i) = \mathbb{E}[Y_i|X_i], \tag{2.3}$$
$$g(X_i) = \mathbb{E}[T_i|X_i], \tag{2.4}$$
$$\pi(X_i) = \mathbb{P}(T_i = 1|X_i) \quad \text{(propensity score)} \tag{2.5}$$

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 (we make the regularity conditions explicity in Appendix 5.1). DML 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:

$$\psi(Y_i, T_i, X_i; \eta) = \left(\frac{T_i - g(X_i)}{\pi(X_i)(1 - \pi(X_i))} \right) (Y_i - m(X_i)) + (m_1(X_i) - m_0(X_i)) - \tau, \tag{2.6}$$

where $g(X_i) = \pi(X_i)$ in case of binary treatment and $m_t(X_i) = \mathbb{E}[Y_i|T_i = t, X_i]$ for $t \in \{0, 1\}$.

2.3.3 Estimation Procedure

The estimation of DML models contains several steps:

i. Estimate Nuisance Functions: Use one of the outlined ML models to obtain estimators $\hat{m}(X_i)$ and $\hat{g}(X_i)$ for the nuisance functions.

ii. Compute the Score Function: Evaluate the orthogonal score $\psi(Y_i, T_i, X_i; \hat{\eta})$ using the estimated nuisance parameters.

iii. Estimate τ: Solve the empirical moment condition which yields $\hat{\tau}$:

$$\frac{1}{n} \sum_{i=1}^{n} \psi(Y_i, T_i, X_i; \hat{\eta}) = 0. \tag{2.7}$$

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):

$$\begin{aligned}\psi(Y_i, T_i, X_i; \hat{\eta}^{(-k)}) = &\left(\frac{T_i - \hat{g}^{(-k)}(X_i)}{\hat{\pi}^{(-k)}(X_i)}\right)(Y_i - \hat{m}^{(-k)}(X_i)) \\ &+ \hat{m}_1^{(-k)}(X_i) - \hat{m}_0^{(-k)}(X_i) - \tau^{(-k)}.\end{aligned} \tag{2.8}$$

3. Aggregate: Combine the estimates from all folds:

$$\hat{\tau} = \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\hat{\tau}_i^{(-k)} = \frac{1}{K}\sum_{k=1}^{K}\hat{\tau}^{(-k)} \tag{2.9}$$

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:

$$\mathbb{P}[Y(t)] = \sum_{C} \mathbb{P}[Y \mid T, C]\mathbb{P}[C], \tag{3.1}$$

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 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.

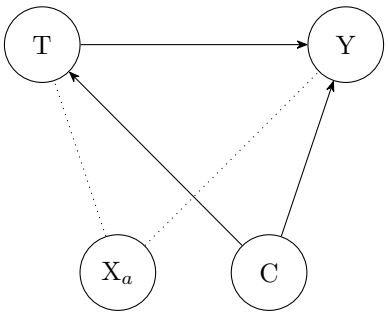

Figure 2: DAG of Backdoor Path

3.2.1 Backdoor Adjustment Scenario in DML

In Backdoor Adjustment with DML, the ideal $m(X_i) = \mathbb{E}[Y_i|X_i]$ and $g(X_i) = \mathbb{E}[T_i|X_i]$ from Equations (2.3) and (2.4) are:

$$m_{\text{BA}}(C_i) = \mathbb{E}[Y_i|C_i], \tag{3.2}$$
$$g_{\text{BA}}(C_i) = \mathbb{E}[T_i|C_i], \tag{3.3}$$

In our simulations we also test the habilit of the ML models to estimate ATE under the following scenarios. In these scenarios, the superscript w_i on the nuisance functions denotes the i-th scenario which involves misspecifications of the nuisance functions.

Inclusion of C_i and $X_{a,i}$ on every nuisance function

$$m_{\mathrm{BA}}^{\mathrm{w}_1}(C_i, X_{a,i}) = \mathbb{E}[Y_i|C_i, X_{a,i}], \tag{3.4}$$
$$g_{\mathrm{BA}}^{\mathrm{w}_1}(C_i, X_{a,i}) = \mathbb{E}[T_i|C_i, X_{a,i}], \tag{3.5}$$

Equations (3.4) and (3.5) represent a scenario in which one would not be aware that X_a does not cause T and Y.

Only partial inclusion of C_i and inclusion of $X_{a,i}$ on every nuisance function

$$m_{\mathrm{BA}}^{\mathrm{w}_2}(C_i^p, X_{a,i}) = \mathbb{E}[Y_i|C_i^p, X_{a,i}], \tag{3.6}$$
$$g_{\mathrm{BA}}^{\mathrm{w}_2}(C_i^p, X_{a,i}) = \mathbb{E}[T_i|C_i^p, X_{a,i}], \tag{3.7}$$

Equations (3.6) and (3.7) also represent a scenario in which one would not be aware that X_a does not cause T and Y and also does not include all causal cofounders. C_i^p is subset of C_i included, where $C_i \in \mathbb{R}^{d_c}$ and $C_i^p \in \mathbb{R}^{d_{cp}}$ and $d_c > d_{cp}$.

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:

$$\mathbb{P}[Y(t)] = \sum_M \mathbb{P}[M \mid T = t] \sum_{t'} \mathbb{P}[Y \mid M, T = t']\mathbb{P}[T = t']. \tag{3.8}$$

For instance, the average treatment effect in case $M, T \in \{0, 1\}$:

$$\tau = [\mathbb{P}(M = 1 \mid T = 1) - \mathbb{P}(M = 1 \mid T = 0)] \times [\mathbb{E}[Y \mid M = 1] - \mathbb{E}[Y \mid M = 0]] \tag{3.9}$$

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 4 illustrates the frontdoor adjustment involving the treatment T, mediator M, outcome Y, observed confounders C and features with no causal association X_a. C and X_a would not be present in a typical frontdoor adjustment DAG, but are included due to their relevance in the proposed simulations.

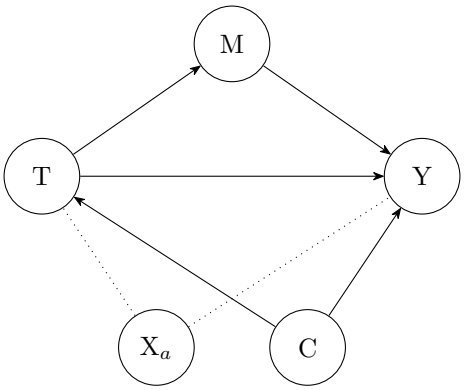

Figure 3: DAG of Frontdoor Path

3.3.1 Frontdoor Adjustment Scenario in DML

In our simulations, we use a binary mediator $M_i \in \{0, 1\}$, similar to the binary treatment T_i.

In the Frontdoor Adjustment scenario with DML, we need to account for the mediator M_i when estimating the ATE. The identification of the causal effect involves modeling the relationships between T_i, M_i, and Y_i.

The ideal nuisance functions for DML in this scenario are:

$$m_{\text{FA}}(M_i, C_i) = \mathbb{E}[Y_i \mid M_i, C_i], \tag{3.10}$$
$$h_{\text{FA}}(T_i) = \mathbb{E}(M_i \mid T_i), \tag{3.11}$$
$$g_{\text{FA}}(C_i) = \mathbb{E}(T_i \mid C_i), \tag{3.12}$$

Here $h_{\text{FA}}(M_i, T_i, C_i)$ is the mediator model, representing the probability of the mediator given treatment.

To adapt the orthogonal score function in the presence of the mediator, we modify Equation (2.6) to incorporate the mediator's effect. The adapted orthogonal score function is:

$$\psi(Y_i, T_i, M_i, C_i; \eta) = \left(\frac{T_i - g_{\text{FA}}(C_i)}{g_{\text{FA}}(C_i)(1 - g_{\text{FA}}(C_i))} \right) (M_i - h_{\text{FA}}(T_i)) (Y_i - m_{\text{FA}}(M_i, C_i)) + \delta_M \delta_Y(C_i) - \tau \tag{3.13}$$

where:

$$\delta_M = h_{\text{FA}}(T_i = 1) - h_{\text{FA}}(T_i = 0), \tag{3.14}$$
$$\delta_Y(C_i) = m_{\text{FA}}(M_i = 1, C_i) - m_{\text{FA}}(M_i = 0, C_i), \tag{3.15}$$

and η represents the collection of nuisance functions.

In this score function:

- The first term adjusts for the treatment assignment, similar to the original score function, but now includes the mediator.
- The product $(M_i - h_{\text{FA}}(T_i))\,(Y_i - m_{\text{FA}}(M_i, C_i))$ captures the interaction between the mediator and the outcome.
- The term $\delta_M \delta_Y(C_i)$ represents the estimated causal effect based on the mediator and outcome models.

There are doubts about the orthogonal score function for the frontdoor adjustment.
We are not sure if the score function should be the one presented in Equation (3.13) or if it should be presented as the following:

$$m_{\text{FA}}(M_i, C_i) = \mathbb{E}[Y_i \mid M_i, C_i], \tag{3.16}$$
$$h_{\text{FA}}(T_i) = \mathbb{E}(M_i \mid T_i), \tag{3.17}$$
$$g_{\text{FA}}(C_i) = \mathbb{E}(T_i \mid C_i), \tag{3.18}$$

$$\psi(Y_i, T_i, M_i, C_i; \eta) = \left(\left(\frac{M_i - h_{\text{FA}}(T_i)}{h_{\text{FA}}(T_i)(1 - h_{\text{FA}}(T_i))} \right) (Y_i - m_{\text{FA}}(M_i, C_i)) \right) - \tau \tag{3.19}$$

$$\psi(Y_i, T_i, M_i, C_i; \eta) = \left(\left(\frac{M_i - h_{\text{FA}}(T_i)}{h_{\text{FA}}(T_i)(1 - h_{\text{FA}}(T_i))} \right) (Y_i - m_{\text{FA}}(M_i, C_i)) \right) \cdot (h_{\text{FA}}(T_i = 1) - h_{\text{FA}}(T_i = 0)) - \tau \tag{3.20}$$

Furthermore, in the frontdoor adjustment, I am not entirely sure if T should have a direct influence on Y. If T also has a direct influence on Y, the DAG would look like the following:

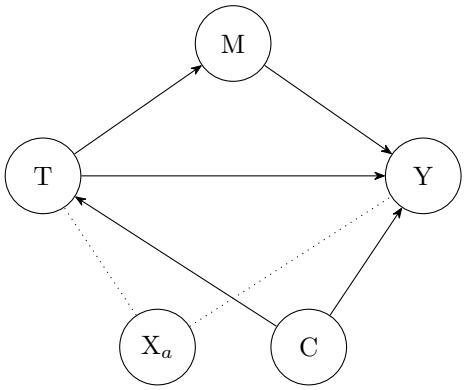

Figure 4: DAG of Frontdoor Path

In our simulations we also test the hability of the ML models to estimate ATE under the following scenario.

Inclusion of C_i and $X_{a,i}$ on every nuisance function

$$m^{\text{w1}}_{\text{FA}}(M_i, C_i, X_{a,i}) = \mathbb{E}[Y_i \mid M_i, C_i, X_{a,i}], \tag{3.21}$$
$$h^{\text{w1}}_{\text{FA}}(T_i, C_i, X_{a,i}) = \mathbb{E}(M_i \mid C_i, X_{a,i}), \tag{3.22}$$
$$g^{\text{w1}}_{\text{FA}}(C_i, X_{a,i}) = \mathbb{E}(T_i \mid C_i, X_{a,i}), \tag{3.23}$$

Equations (3.21), (3.22), and (3.23) represent a scenario where one is unaware that $X_{a,i}$ has no causal relation to T_i, M_i, or Y_i. Equation (3.22) represents scenarion where one is unaware that C has no causal relation to M_i.

Ignoring the mediator M_i

$$m_{\text{FA}}^{\text{w}_2}(C_i, X_{a,i}) = \mathbb{E}[Y_i \mid C_i, X_{a,i}], \tag{3.24}$$
$$g_{\text{FA}}^{\text{w}_2}(C_i, X_{a,i}) = \mathbb{E}[T_i \mid C_i, X_{a,i}], \tag{3.25}$$

Equations (3.24) and (3.25) represent a scenario where the mediator M_i is unavailable or ignored and one is unaware that $X_{a,i}$ has no causal relationship to T or Y. In this case, the orthogonal score function is the same as in Equation (2.6).

Considering the mediator M_i as a normal covariate

$$m_{\text{FA}}^{\text{w}_3}(C_i, M_i, X_{a,i}) = \mathbb{E}[Y_i \mid C_i, M_i, X_{a,i}], \tag{3.26}$$
$$g_{\text{FA}}^{\text{w}_3}(C_i, M_i, X_{a,i}) = \mathbb{E}[T_i \mid C_i, M_i, X_{a,i}], \tag{3.27}$$

Equations (3.26) to (3.27) also represent scenarios where one is unaware that $X_{a,i}$ has no causal relation to T_i or Y_i, and the scenario where M_i is available but not recognized as a mediator. In this case, the orthogonal score function is the same as in Equation (2.6).

3.4 Instrumental Variable

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

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 5 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. Again, the last two would not be present in a typical instrumental variable DAG, but are included due to their relevance in the proposed simulations.

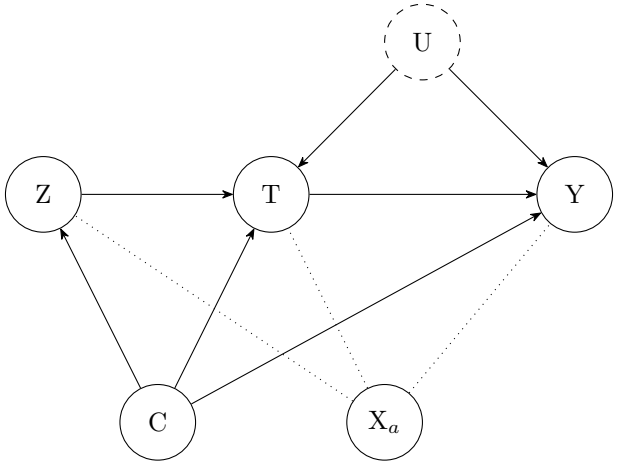

Figure 5: DAG of Instrumental Variable

3.4.1 Instrumental Variable Scenario in DML

To adapt the DML framework to the IV setting, we need to define appropriate nuisance functions and construct an orthogonal score function suitable for the IV context.

The ideal nuisance functions in this scenario are:

$$m_{\text{IV}}(C_i) = \mathbb{E}[Y_i \mid C_i], \tag{3.28}$$
$$q_{\text{IV}}(C_i) = \mathbb{E}[Z_i \mid C_i], \tag{3.29}$$
$$g_{\text{IV}}(C_i) = \mathbb{E}[T_i \mid C_i], \tag{3.30}$$

Where $m_{\text{IV}}(C_i)$ is the outcome model, capturing the expected outcome given covariates, $q_{\text{IV}}(C_i)$ is the instrument propensity score, representing the expected treatment given the instrument and covariates, and $g_{\text{IV}}(C_i)$ is the treatment model, representing the expected treatment given covariates.

The orthogonal score function for the IV scenario is different from the standard DML orthogonal score function. An appropriate orthogonal score function in the linear IV context is:

$$\psi_{\text{IV}}(W_i; \tau, \eta) = \left(Y_i - m_{\text{IV}}(C_i) - \tau[T_i - g_{\text{IV}}(C_i)]\right)\left(Z_i - q_{\text{IV}}(C_i)\right), \tag{3.31}$$

Is this orthogonal score function correct? I was not able to find another one that made more sense, but I would have thought that the variance of the propensity score should be dividing some of the parts of the orthogonal score function.

In our simulations we also test the habilit of the ML models to estimate ATE under the following scenarios. In these scenarios, the superscript w_i on the nuisance functions denotes the i-th scenario which involves misspecifications of the nuisance functions.

Inclusion of C_i and $X_{a,i}$ on every nuisance function

$$m_{\text{IV}}^{\text{w}_1}(C_i) = \mathbb{E}[Y_i \mid C_i, X_{a,i}], \tag{3.32}$$
$$q_{\text{IV}}^{\text{w}_1}(C_i) = \mathbb{E}[Z_i \mid C_i, X_{a,i}], \tag{3.33}$$
$$g_{\text{IV}}^{\text{w}_1}(C_i) = \mathbb{E}[T_i \mid C_i, X_{a,i}], \tag{3.34}$$

Equations (3.32), (3.33), and (3.34) represent a scenario in which one would not be aware that X_a does not cause T, Z and Y.

Only partial inclusion of C_i and inclusion of $X_{a,i}$ on every nuisance function

$$m_{\text{IV}}^{\text{w}_2}(C_i^p, X_{a,i}) = \mathbb{E}[Y_i \mid C_i, X_{a,i}], \tag{3.35}$$
$$q_{\text{IV}}^{\text{w}_2}(C_i^p, X_{a,i}) = \mathbb{E}[Z_i \mid, C_i, X_{a,i}], \tag{3.36}$$
$$g_{\text{IV}}^{\text{w}_2}(C_i^p, X_{a,i}) = \mathbb{E}[T_i \mid C_i, X_{a,i}], \tag{3.37}$$

Equations (3.35), (3.36), and (3.7) also represent a scenario in which one would not be aware that X_a does not cause T and Y and also does not include all causal cofounders. C_i^p is subset of C_i included, where $C_i \in \mathbb{R}^{d_c}$ and $C_i^p \in \mathbb{R}^{d_{cp}}$ and $d_c > d_{cp}$.

Z_i is treated as a normal cofounder

$$m_{\text{IV}}^{\text{w}_3}(C_i, Z_i) = \mathbb{E}[Y_i | C_i, Z_i], \tag{3.38}$$
$$g_{\text{IV}}^{\text{w}_3}(C_i, Z_i) = \mathbb{E}[T_i | C_i, Z_i], \tag{3.39}$$

Equations (3.40) and (3.41) represent a scenario in which one would treat Z as a normal covariate. In this case, the orthogonal score function would be the same as in the backdoor adjustment scenario of (2.6).

Z_i is treated as a normal cofounder and inclusion of $X_{a,i}$ on every nuisance function

$$m_{\text{IV}}^{\text{w}_4}(C_i, Z_i, X_{a,i}) = \mathbb{E}[Y_i | C_i, Z_i, X_{a,i}], \tag{3.40}$$
$$g_{\text{IV}}^{\text{w}_4}(C_i, Z_i, X_{a,i}) = \mathbb{E}[T_i | C_i, Z_i, X_{a,i}], \tag{3.41}$$

Equations (3.40) and (3.41) represent a scenario in which one would treat Z as a normal covariate and is not aware that X_a does not cause T, Z and Y. In this case, the orthogonal score function would also be the same as in the backdoor adjustment scenario of (2.6).

4 Simulation Methodology

4.1 Data Generating Process

We define the the backdoor path scenario as the typical scenario. The frontdoor path and instrumental variable scenarios are variations of the typical scenario.

4.1.1 Cofounders

The cofounders are generated with multivariate normal distribution using a sparse covariance matrix Σ_d and a vector of means $\boldsymbol{\mu}_d = \mathbf{0}$.

$\Sigma_d \in \mathbb{R}^{d \times d}$, where $d := d_a + d_c$. d_a is the number of non-causal confounders previously represented as X_a and d_c is the number of causal confounders previously represented as C.

$$X \sim \mathcal{N}_d(\boldsymbol{\mu}_d, \Sigma_d) \tag{4.1}$$

The sparsity of the covariance matrix of the cofounders Σ_d is determined by α_d, which represents the probability that any covariance coefficient is zero. Larger α_d's result in sparser covariance matrices. To provide intuition on it, we report the average absolute correlation $|\rho_d|$ in the matrix.

After generating the data from the multivariate normal distribution defined as $X \in \mathbb{R}^{n \times d}$ we randomly separate the features of X in X_a and C. Following the logic specified above, $X_a \in \mathbb{R}^{n \times d_a}$ and $C \in \mathbb{R}^{n \times d_c}$.

In (3.6) and (3.7) "Only partial inclusion of C_i and inclusion of $X_{a,i}$ on every nuisance function", after generating cofounders, treatment, and target, we select a subset of the causal confounders C_i^p to be used in the nuisance functions. This would represent a scenario of unobserved confounders U_i. We define which causal confounders are not included in the nuisance function randomly, through the use of p_u, which represents the percentage of causal confounders not included in the nuisance functions. Meaning that $C_i^p \in \mathbb{R}^{n \times d_{cp}}$, where $d_{cp} = d_c - \lceil p_u \cdot d_c \rceil$.

4.1.2 Treatment

The probability of treatment is generated from a logic:

$$\mathbb{P}[T_i] = \frac{1}{1 + e^{(-f(C_i) + \varepsilon_{t,i})}} \tag{4.2}$$

where $f(C_i)$ is a function of the causal confounders C_i and $\varepsilon_{t,i}$ is a random error term generated from a normal distribution with mean 0 and variance σ_t.

$f(C_i)$ is a generic non-linear function explained in Section 4.1.5.

Treatment is later generated from a Bernoulli distribution with the probability of treatment from above in (4.2).

$$T_i \sim Bernoulli(\mathbb{P}[T_i]) \tag{4.3}$$

Should T_i follow a Bernoulli or just be a 1 or 0 based on the probability of the treatment? The second case would be something like

$$T_i = \begin{cases} 1 & \text{if } \mathbb{P}[Z_i] > 0.5 \\ 0 & \text{otherwise} \end{cases}$$

I believe Bernoulli is a better option.

4.1.3 Treatment Effect

The true individual treatment effect is generated from:

$$\tau_i(C_i) = f(C_i) \tag{4.4}$$

Again, $f(C_i)$ is a generic non-linear function explained in Section 4.1.5.

Meaning that there is heterogeneity in the treatment effect across the causal confounders C_i.

The true ATE (Average Treatment Effect) is the average of the treatment effect across the population:

$$\tau = \mathbb{E}[\tau(C_i)] = \frac{1}{n}\sum_{i=1}^{n}\tau_i(C_i) \tag{4.5}$$

Should it be $\mathbb{E}[\tau_i(C_i)]$ or $\mathbb{E}[\tau(C_i)]$? Not very crucial, but must be checked.

4.1.4 Outcome

The observed outcome for each individual is:

$$Y_i = Y_{i,0} + \tau_i(C_i)\cdot T_i + \varepsilon_{y,i}, \quad \varepsilon_{y,i} \sim \mathcal{N}(0, \sigma_y) \tag{4.6}$$

where $Y_{i,0}$ is the potential outcome if the treatment T_i was not applied, $\tau_i(C_i)$ is the treatment effect, and $\varepsilon_{y,i}$ is a random error term generated from a normal distribution with mean 0 and variance σ_y.

$$Y_{i,0} = f(C_i) \tag{4.7}$$

Once more, $f(C_i)$ is a generic non-linear function explained in Section 4.1.5.

The potential outcome for treatment and control are:

$$Y_i(1) = Y_{i,0} + \tau_i(C_i), \quad Y_i(0) = Y_{i,0}, \tag{4.8}$$

which can not be directly observed.

4.1.5 $f(C_i)$ Non-Linear Transformation

The function $f(C_i)$, used to calculate $Y_{i,0}$ and $\mathbb{P}[T_i]$ is a generic function composed non-linear relationships.

More specifically $f(C_i)$ is the weighted sum of of different "q" transformations in the causal confounders C_i:

$$f(C) = \sum_{j=1}^{d_c}(\beta_j \cdot q_j(C_j)) \tag{4.9}$$

Where C_j is the the vector of the j-th causal confounder, β_j is a random coefficient drawn from a uniform distribution $U(-1, 1)$, and $q_j(C_j)$ is a random transformation of the j-th causal confounder. More specifically, the chosen transformation $q_j(C_j)$ is randomly chosen from the following set of possible transformations:

i. Linear: $q_j(C_{i,j}) = C_{i,j}$

ii. Quadratic: $q_j(C_{i,j}) = C_{i,j}^2$

iii. Cubic: $q_j(C_{i,j}) = C_{i,j}^3$

iv. Logarithmic: $q_j(C_{i,j}) = \log(|C_{i,j}| + 1)$

v. Exponential: $q_j(C_{i,j}) = \exp(\frac{1}{5}C_{i,j})$

vi. Sine: $q_j(C_{i,j}) = \sin(C_{i,j})$

vii. Cosine: $q_j(C_{i,j}) = \cos(C_{i,j})$

viii. Indicator: $q_j(C_{i,j}) = \mathbb{I}\{C_{i,j} > 0\} - \mathbb{I}\{C_{i,j} \leq 0\}$

ix. Piecewise: $q_j(C_{i,j}) = 2\mathbb{I}\{C_{i,j} < 0\} + \mathbb{I}\{0 \leq C_{i,j} < 1\} + \frac{1}{2}\mathbb{I}\{C_{i,j} > 1\}$

Should we keep the non-continuous transformations? I know that some of the theorems only apply for continuous functions.

For each simulation, $f(C)$ is calculated with some of the specified transformation sets and β's three times: one for the calculation of $Y_{i,0}$, one for the calculation of $\mathbb{P}[T_i]$, and one for the calculation of $\tau_i(C_i)$.

For some simulations, we only allow for linear transformations in $f(C)$, aiming to compare performance of the DML models in a simpler setting.

4.1.6 Parameter Tuning in the Data Generating Process

The data generating process has some parameters that can be tuned to generate different scenarios. We present the following variations:

i. $f(C_i)$ for Treatment Probability: linear or non-linear transformations (4.2).

ii. $f(C_i)$ for Treatment Effect: linear or non-linear transformations (4.4).

iii. $f(C_i)$ for Outcome: linear or non-linear transformations (4.7).

iv. d_c: number of causal confounders.

v. d_a: number of non-causal confounders.

vi. p_u: percentage of causal confounders not included in the nuisance functions in the backdoor adjustment scenario with wrong specification (3.6, 3.7).

vii. n: number of observations.

viii. α_d: sparsity of the covariance matrix of cofounders (X), which is a direct cause of average absolute correlation between cofounders $|\rho_d|$ (4.1).

ix. σ_t: variance of the error term ε_t in the treatment probability (4.2).

x. σ_y: variance of the error term ε_y in the outcome (4.6).

4.2 Differences in Data Generating Process for Instrumental Variable Scenario

The instrumental variable scenario provides some differences in the data generating process compared to the backdoor adjustment scenario, thus also allowing for more parameters variation.

4.2.1 Cofounders

The cofounders are again generated from a multivariate normal distribution using a sparse covariance matrix Σ_d and a vector of means $\boldsymbol{\mu}_d = \mathbf{0}$, as described in (4.1).

In this scenario, X is divided in $X_a \in \mathbb{R}^{n \times d_a}$, $C \in \mathbb{R}^{n \times d_c}$, and $U \in \mathbb{R}^{n \times d_u}$, where $d := d_a + d_c + d_u$.

U_i is used to generate $\mathbb{P}[T_i]$, and $\tau_i(C_i, U_i)$, and is not included in the nuisance functions.

4.2.2 Instrument

We define the instrument in both discrete and continuous forms. In both cases, the instrument is generated from C_i but not from U_i.

In the discrete case:

$$\mathbb{P}[Z_i] = \frac{1}{1 + e^{(-f(C_i) + \varepsilon_{z,i})}}, \quad \varepsilon_{z,i} \sim \mathcal{N}(0, \sigma_z) \tag{4.10}$$

$$Z_i \sim Bernoulli(\mathbb{P}[Z_i]) \tag{4.11}$$

Should Z_i follow a Bernoulli or just be a 1 or 0 based on the probability of the instrument? The second case would be something like

$$Z_i = \begin{cases} 1 & \text{if } \mathbb{P}[Z_i] > 0.5 \\ 0 & \text{otherwise} \end{cases}$$

I believe Bernoulli is a better option.

In the continuous case:

$$Z_i = f(C_i) + \varepsilon_{z,i}, \quad \varepsilon_{z,i} \sim \mathcal{N}(0, \sigma_z) \tag{4.12}$$

4.2.3 Treatment

The $\mathbb{P}[T_i]$ previously addressed in the backdoor adjustment scenario with (4.2) becomes:

$$\mathbb{P}[T_i] = \frac{1}{1 + e^{(-f(C_i, U_i, Z_i) + \varepsilon_{t,i})}}, \quad \varepsilon_{t,i} \sim \mathcal{N}(0, \sigma_t) \tag{4.13}$$

A similar procedure is done for the individual treatment effect from (4.4):

$$\tau_i(C_i, U_i) = f(C_i, U_i) \tag{4.14}$$

Should the τ_i be a function of Z_i as well? I am almost sure that no. I give an explanation below:

In a standard instrumental variables (IV) framework, the key assumption is that the instrument Z affects the outcome Y only through its influence on the treatment T, thus influencing probability of treatment, but having no effect on the potential outcomes.

4.2.4 Outcome

Differently from the treatment, the data generating process of the outcome is not a function of the instrument Z_i. Nonetheless, it is influenced by the unobserved confounders U_i, thus still having a data generating process different from the backdoor adjustment scenario (4.7).

$$Y_{i,0} = f(C_i, U_i) \tag{4.15}$$
$$Y_i = Y_{i,0} + \tau_i(C_i, U_i) \cdot T_i + \varepsilon_{y,i}, \quad \varepsilon_{y,i} \sim \mathcal{N}(0, \sigma_y) \tag{4.16}$$

4.2.5 Parameter Tuning in the Data Generating Process for Instrumental Variable Scenario

In the instrumental variable scenario, we have the possibility to tune all the previously mentioned parameters in (4.2) as well as:

i. $f(C_i)$ for Instrument: linear or non-linear transformations (4.10 and 4.12).

ii. σ_z: variance of the error term ε_z in the instrument generation (4.10 and 4.12).

iii. d_u: number of unobserved confounders (which substitutes the percentage of causal confounders p_u not included in the nuisance functions in the backdoor adjustment scenario with wrong specification).

5 Appendix

5.1 Mild Regularity Conditions for Double Machine Learning

i. Smoothness of the Target Function: The functions $m(X)$ and $g(X)$, representing the outcome and treatment models, should belong to a sufficiently smooth function class (e.g., Hölder class or Sobolev space). This ensures that they can be approximated well by machine learning methods.

ii. Boundedness and Regularity: The outcome Y, treatment T, and covariates X should have bounded support or satisfy moment conditions, such as:

$$\mathbb{E}[|Y|^2] < \infty, \quad \mathbb{E}[|T|^2] < \infty.$$

Additionally, $m(X)$ and $g(X)$ should have bounded derivatives in some cases.

iii. Sparsity or Complexity Control: The nuisance estimators $\hat{m}(X)$ and $\hat{g}(X)$ should satisfy complexity restrictions, such as sparsity in high-dimensional settings or controlled VC dimensions, to ensure valid estimation and inference.

iv. Consistency and Rates of Convergence: The nuisance estimators $\hat{m}(X)$ and $\hat{g}(X)$ must converge to the true $m(X)$ and $g(X)$ at sufficiently fast rates (typically faster than $n^{-1/4}$ in terms of mean squared error).

v. Independence or Weak Dependence: Observations should be independent and identically distributed (i.i.d.) or satisfy weak dependence conditions, such as mixing properties for time-series data.

vi. Overlap Condition (Positivity): The propensity score $\pi(X) = \mathbb{P}(T = 1|X)$ must be bounded away from 0 and 1:

$$0 < c \leq \pi(X) \leq 1 - c \quad \text{for some } c > 0.$$

vii. Orthogonality of the Moment Function: The estimating equation or moment function used in DML should satisfy a doubly-robust property, meaning that small estimation errors in $\hat{m}(X)$ and $\hat{g}(X)$ do not affect the asymptotic distribution of the estimator.

viii. Cross-Fitting: Proper cross-fitting should be employed to ensure the orthogonality of the estimating equations and to avoid overfitting. This mitigates potential overfitting by using disjoint data splits for estimating nuisance parameters and evaluating the final moment function.