Estimating Causal Effects Using Weighting-Based Estimators

Authored By *Yonghan Jung, Jin Tian, Elias Bareinboim*

Presented By *Alexander Shmakov*

COMPSCI 295, Winter 2020

Jung, Y., Tian, J., & Bareinboim, E. (2020). *Estimating Causal Effects Using Weighting-Based Estimators*. Proceedings of the AAAI Conference on Artificial Intelligence, 34(06), 10186-10193. https://doi.org/10.1609/aaai.v34i06.6579

Introduction Back-Door Criterion

We have learned extensively about the usefulness and power of the back-door criterion for **identifying** causal effects and back-door resolution for **estimating** a query.

Definition 3.3.1 (The Backdoor Criterion) *Given an ordered pair of variables (X, Y) in a directed acyclic graph G, a set of variables Z satisfies the* backdoor criterion *relative to (X, Y) if no node in Z is a descendant of X, and Z blocks every path between X and Y that contains an arrow into X.*

If a set of variables Z satisfies the backdoor criterion for X and Y, then the causal effect of X on Y is given by the formula

$$P(Y = y|do(X = x)) = \sum_z P(Y = y|X = x, Z = z)P(Z = z)$$

Introduction Back-Door Criterion

Some simpler notation, otherwise, we will get very confused later.

For any nodes x, y and set of nodes Z belonging to an SCM G satisfying the back-door criterion.
$P_x(y)$ is the intervention distribution, intervening on x.
$E_{P_x(y)}[Y]$ is the average value of the query from our SCM.

$$P_x(y) = \sum_z P(y|x,z)P(z)$$

$$\mathbb{E}_{P_x(y)}[Y] = \sum_z \mathbb{E}[Y|x,z]P(z)$$

We mainly focus on the expectation formula in this paper, because it must be used to compute useful values such as expected causal effect
Very pretty, very easy to understand, and revolutionized casual inference.

However, do you see some very restricting limitations for this equation?

Introduction Estimation Tractability

$$\mathbb{E}_{P_x(y)}[Y] = \sum_z \mathbb{E}[Y|x, z]P(z)$$

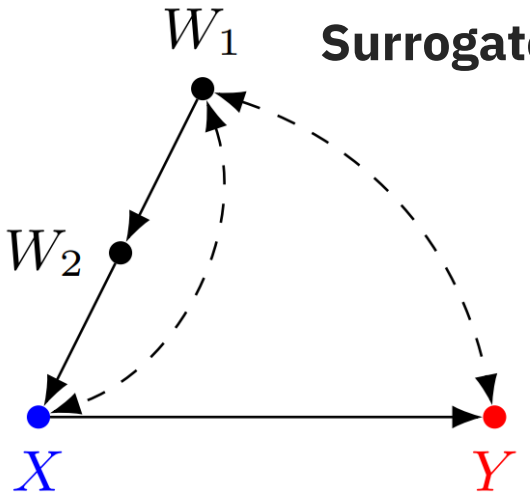
Computing this value requires us to sum (or integrate!) over all values of all covariates $\mathbf{z} \in \boldsymbol{Z}$.

Say that we have n covariates $Z = \{z_1, z_2, \ldots, z_n\}$ with each discrete covariate have k values, then the runtime for this sum is

$$\boldsymbol{O(k^n)}$$

This might be computable for 2 or 3 confounding variables,
but anything over 40 or 50 is impossible on any computer!

Introduction Back-Door Restriction



Surrogate Endpoints

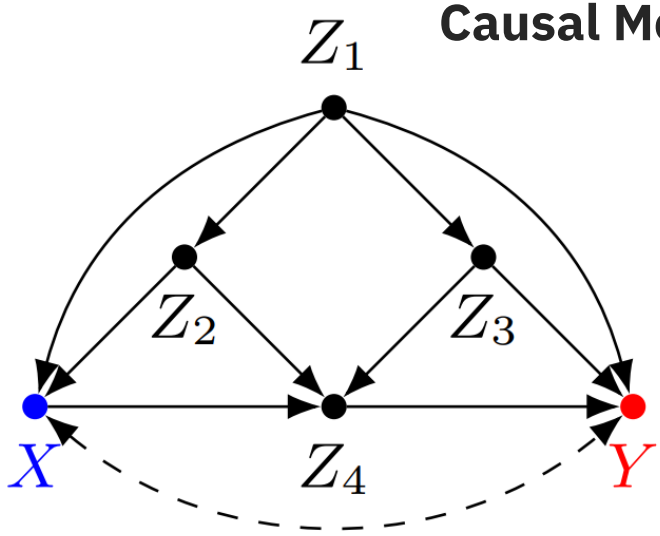
W_1: disease history

W_2: CD4 Cell Treatment

X: CD4 Cell Count

Y: HIV Progression

$P_x(y)$: $\left(\sum_{w_1} P\left(x, y|w_1, w_2\right) P\left(w_1\right)\right) / \left(\sum_{w_1} P\left(x|w_1, w_2\right) P\left(w_1\right)\right)$



Causal Mediators

Z_1: Age

Z_2: Diet

Z_3: Smoking

Z_4: Blood Test Results

X: BMI

Y: Cancer Presence

$P_x(y)$: $\sum_{\mathbf{z}} P(z_4|x, \mathbf{z}^{(3)})P(\mathbf{z}^{(3)}) \sum_{x'} P(y|x', \mathbf{z})P(x'|\mathbf{z}^{(3)})$

Both graphs are identifiable from the do-calculus, but this calculation must be done manually.

However, they are not identifiable by any simple algorithm such as back-door or front-door!

Introduction Paper Claims

Existing methods are simple to compute from the graph and produce simple expressions

Back-Door $$\mathbb{E}_{P_x(y)}[Y] = \sum_z \mathbb{E}[Y|x,z]P(z)$$

Front-Door $$\mathbb{E}_{P_{\mathbf{x}}(\mathbf{y})}[Y] = \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{x}) \sum_{\mathbf{x}'} \mathbb{E}[Y|\mathbf{x}',\mathbf{z}]P(\mathbf{x}')$$

The goal of this paper is two-fold

1. Generalize the backdoor criterion as much as possible to apply to many other graphs.
2. Determine an efficient way to estimate the resulting expectation.

We will do this by introducing

1. The multi-outcome sequential back-door (*mSBD*) and surrogate criterions.
2. Weighting-based Empirical Estimators.

But first some groundwork on weighted distributions.

Weighted Distributions Basics

We will introduce an operation on distributions to modify them with arbitrary weights.

Suppose that $P(\mathbf{v})$ is a distribution over some space $\mathbf{V}$. Then we can define a new distribution over the same space modified by a **weight function**.

Definition 1 (Weighted distribution $P^{\mathcal{W}}(\mathbf{v})$)**.** Given a distribution $P(\mathbf{v})$ and a weight function $\mathcal{W}(\mathbf{v}) > 0$, a weighted distribution $P^{\mathcal{W}}(\mathbf{v})$ is given by

$$P^{\mathcal{W}}(\mathbf{v}) \equiv \frac{\mathcal{W}(\mathbf{v}) P(\mathbf{v})}{\sum_{\mathbf{v}'} \mathcal{W}(\mathbf{v}') P(\mathbf{v}')}. \tag{2}$$

Why is this useful?

By simply using some algebra, we can re-interpret the previous back-door expectation as a weighted modification of a regular conditional distribution. This is known as the inverse probability weighting.

Proposition 1. *Let* $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$. *If the causal effect* $P_{\mathbf{x}}(\mathbf{y})$ *is identifiable by the BD adjustment, then* $P_{\mathbf{x}}(\mathbf{y}) = P^{\mathcal{W}}(\mathbf{y}|\mathbf{x})$ *where* $\mathcal{W} = \frac{P(\mathbf{x})}{P(\mathbf{x}|\mathbf{z})}$, *and*

$$\mathbb{E}_{P_{\mathbf{x}}(\mathbf{y})}[\mathbf{Y}] = \mathbb{E}_{P^{\mathcal{W}}(\mathbf{y}|\mathbf{x})}[\mathbf{Y}|\mathbf{X} = \mathbf{x}]. \tag{3}$$

Weighted Distributions Proof of Back-Door IPW

A quick proof of this statement from the axioms of Bayesian Statistics (Pearl 2000, Ch. 3.6.1)

Back Door Definition

$$P(y|do(x)) = \sum_z P(Y = y|X = x, Z = z)P(Z = z)$$

Multiply Top and Bottom

$$P(y|do(x)) = \sum_z \frac{P(Y = y|X = x, Z = z)P(X = x|Z = z)P(Z = z)}{P(X = x|Z = z)}$$

Consolidate Numerator

$$P(y|do(x)) = \sum_z \frac{P(Y = y, X = x, Z = z)}{P(X = x|Z = z)}$$

Therefore, to move between $P(y|do(x))$ and $P(y|x) = P(y, x, z)/P(x)$, multiply by

$$\mathcal{W} = \frac{P(\mathbf{x})}{P(\mathbf{x}|\mathbf{z})}$$

Weighted Distributions Weighted Operators

We can even realize that since the expectation is what we really care about, we can define an object to track the weighted expectation. The authors use this to construct an *algebra* of **weighted operators** with **composition**.

Definition 2 (Weighing operator $\mathcal{B}$)**.** Given a weight function $\mathcal{W}(\mathbf{v}) > 0$, a function $h(\mathbf{Y})$, a set of variables $\mathbf{X} = \mathbf{x}$, the weighting operator $\mathcal{B}[h(\mathbf{Y}) \mid \mathbf{x}; \mathcal{W}]$ is defined by

$$\mathcal{B}[h(\mathbf{Y}) \mid \mathbf{x}; \mathcal{W}] \equiv \mathbb{E}_{P^{\mathcal{W}}(\mathbf{y}|\mathbf{x})}[h(\mathbf{Y})|\mathbf{x}] = \sum_{\mathbf{y}} h(\mathbf{y}) P^{\mathcal{W}}(\mathbf{y}|\mathbf{x}).$$

Definition 5 (Composition of weighting operators)**.** Given two weighting operators $\mathcal{B}_1(\mathbf{x}) \equiv \mathcal{B}[h_z(\mathbf{Z}) \mid \mathbf{x}; \mathcal{W}_1]$ and $\mathcal{B}_2(\mathbf{z}) \equiv \mathcal{B}[h_y(\mathbf{Y}) \mid \mathbf{z}; \mathcal{W}_2]$, the composition of $\mathcal{B}_1$ and $\mathcal{B}_2$ is defined by

$$(\mathcal{B}_1 \circ \mathcal{B}_2)(\mathbf{x}) \equiv \mathcal{B}[\mathcal{B}_2(\mathbf{z}) \mid \mathbf{x}; \mathcal{W}_1]. \tag{15}$$

We can again re-write our back-door expression as simply

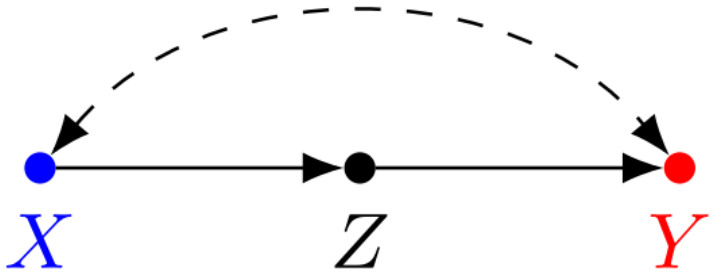
$$\mathbb{E}_{P_{\mathbf{x}}(\mathbf{y})}[\mathbf{Y}] = \mathcal{B}[\mathbf{Y} \mid \mathbf{x}; \mathcal{W}], \text{ where } \mathcal{W} = \frac{P(\mathbf{x})}{P(\mathbf{x}|\mathbf{z})}$$

Weighted Distributions Example

Immediately, we may use this machinery to re-interpret the front-door criterion as simply a combination of two backdoor criterions in two different directions!

The base front-door expression

$$\mathbb{E}_{P_{\mathbf{x}}(\mathbf{y})}[Y] = \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{x}) \sum_{\mathbf{x}'} \mathbb{E}[Y|\mathbf{x}', \mathbf{z}] P(\mathbf{x}')$$



Through some algebraic manipulation, re-order to view as two different backdoor distributions,

$$P_{\mathbf{x}}(\mathbf{y}) = \sum_{\mathbf{z}} \underbrace{P_{\mathbf{x}}(\mathbf{z})}_{\texttt{BD}=\emptyset} \underbrace{P_{\mathbf{z}}(\mathbf{y})}_{\texttt{BD}=\{\mathbf{X}\}},$$

$$\mathbb{E}_{P_{\mathbf{x}}(\mathbf{y})}[\mathbf{Y}] = \mathbb{E}_{P_{\mathbf{x}}(\mathbf{z})}\left[\mathbb{E}_{P_{\mathbf{z}}(\mathbf{y})}[\mathbf{Y}]\right]$$

To get that the front-door is simply a composition of two weighted operators!

$$\mathbb{E}_{P_{\mathbf{x}}(\mathbf{y})}[\mathbf{Y}] = (\mathcal{B}_1 \circ \mathcal{B}_2)(\mathbf{x})$$

where $\mathcal{B}_1(\mathbf{x}) = \mathcal{B}[h(\mathbf{Z}) \mid \mathbf{x}; \mathcal{W}_1]$, $\mathcal{B}_2(\mathbf{z}) = \mathcal{B}[\mathbf{Y} \mid \mathbf{z}; \mathcal{W}_2]$,
$\mathcal{W}_1 = 1$, *and* $\mathcal{W}_2 = \frac{P(\mathbf{z})}{P(\mathbf{z}|\mathbf{x})}$.

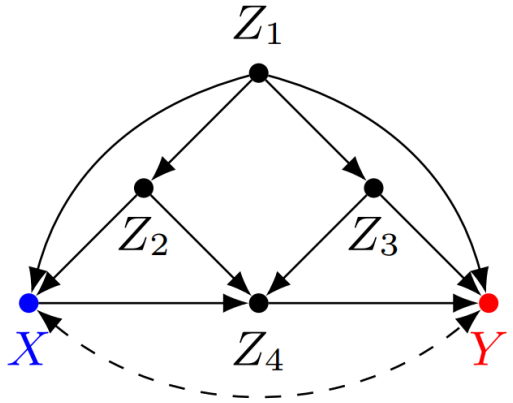
Weighted Distributions Algebra

Idea

Use this algebra of weighted distributions combined with composition to break down a causal graph into simpler units.

Homework Question! More info later.

In fact, this can be used to find a compact expectation for our causal mediator graph from before.



Problem

Back-door as a base still seems restrictive, what's the most general baseline criterion that we can solve explicitly?

Multi-outcome sequential back-door (mSBD)

mSBD Definition

Essentially a generalization of the back-door criterion to allow for a (topologically-ordered) **set of X nodes** and a **set of Y nodes.**

Definition 3 (Multi-outcome sequential back-door (mSBD) criterion)**.** Given the pair of sets $(\mathbf{X}, \mathbf{Y})$, let $\mathbf{X} = \{X_1, X_2, \cdots, X_n\}$ be topologically ordered as $X_1 < X_2 < \cdots < X_n$. Let $\mathbf{Y}_0 = \mathbf{Y} \setminus De(\mathbf{X})$ and $\mathbf{Y}_i = \mathbf{Y} \cap \left(De(X_i) \setminus De\left(\mathbf{X}^{\geq i+1}\right)\right)$ for $i = 1, 2, \cdots, n$. Let $ND\left(\mathbf{X}^{\geq i}\right)$ be the set of nondescendants of $\mathbf{X}^{\geq i}$. Then the sequence of variables $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_n)$ are said to be mSBD admissible relative to $(\mathbf{X}, \mathbf{Y})$ if it holds that $\mathbf{Z}_i \subseteq ND\left(\mathbf{X}^{\geq i}\right)$, and

$$\left(\mathbf{Y}^{\geq i} \perp\!\!\!\perp X_i | \mathbf{Y}^{(i-1)}, \mathbf{Z}^{(i)}, \mathbf{X}^{(i-1)}\right)_{G_{\underline{X_i}\overline{\mathbf{X}^{\geq i+1}}}}.$$

Essentially, we use the topological order on $\boldsymbol{X}$ to induce an ordering on $\boldsymbol{Y}$. Then, any current X_i must satisfy the back door criterion relative to every set of future $\boldsymbol{Y}^{\geq \boldsymbol{i}}$ relative to all previous observations and covariates $\left(X_{i-1}, \boldsymbol{Y}^{\boldsymbol{i-1}}, \boldsymbol{Z}^{\boldsymbol{i}}\right)$.

mSBD Adjustment

The definition gives use the identification rule.

Now we need to adjustment rule to get the resolution distributions. Luckily, from the previous explanation, it is just a larger back-door expression!

Theorem 1 (mSBD adjustment). *If* $\mathbf{Z}$ *is mSBD admissible relative to* $(\mathbf{X}, \mathbf{Y})$, *then* $P_{\mathbf{x}}(\mathbf{y})$ *is identifiable and given by*[3]

$$P_{\mathbf{x}}(\mathbf{y}) = \sum_{\mathbf{z}} \prod_{k=0}^{n} P\left(\mathbf{y}_k | \mathbf{x}^{(k)}, \mathbf{z}^{(k)}, \mathbf{y}^{(k-1)}\right) \times \prod_{j=1}^{n} P\left(\mathbf{z}_j | \mathbf{x}^{(j-1)}, \mathbf{z}^{(j-1)}, \mathbf{y}^{(j-1)}\right). \quad (4)$$

Just a massive joint version of the back-door adjustment formula with an ordering on x and y!

mSBD Weighted Operators

Now express this massive expectation in terms of the weighted operator objects that we defined previously.

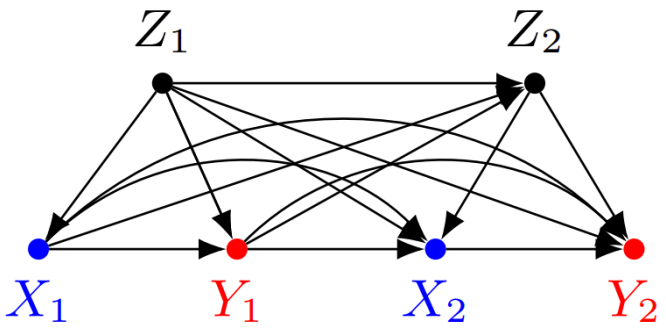
Theorem 2. *If* $\mathbf{Z}$ *is mSBD admissible relative to* $(\mathbf{X}, \mathbf{Y})$, *then*

$$\mathbb{E}_{P_{\mathbf{x}}(\mathbf{y})}\left[h\left(\mathbf{Y}\right)\right] = \mathcal{B}\left[h\left(\mathbf{Y}\right) \mid \mathbf{x}; \mathcal{W}\right], \text{ where} \tag{6}$$

$$\mathcal{W} = \mathcal{W}_{mSBD}(\mathbf{x}, \mathbf{y}, \mathbf{z}) \equiv \frac{P(\mathbf{x})}{\prod_{k=1}^{n} P\left(x_k | \mathbf{x}^{(k-1)}, \mathbf{y}^{(k-1)}, \mathbf{z}^{(k)}\right)}. \tag{7}$$

mSBD Example

Let's compute this expressions for a simple non-trivial graph.



$X = (X_1, X_2)$ satisfies the *mSBD* criterion to $Y = (Y_1, Y_2)$ with $Z = (Z_1, Z_2)$

The raw expectation

$$\mathbb{E}_{P_{x_1,x_2}(y_2)}[Y_2] = \sum_{z_1,z_2,y_1} \mathbb{E}[Y_2|x_1,x_2,z_1,z_2,y_1]\, P(y_1|x_1,z_1) \times P(z_1)P(z_2|x_1,z_1,y_1) \quad (5)$$

In terms of weighted operators

$$\mathbb{E}_{P_{x_1,x_2}(y_2)}[Y_2] = \mathcal{B}[Y_2 \mid \{x_1,x_2\}; \mathcal{W}], \quad (8)$$

where $\mathcal{W} = \dfrac{P(x_1,x_2)}{P(x_1|z_1)P(x_2|x_1,y_1,z_1,z_2)}$.

mSBD Surrogate Criterion

A small extension to mSBD to generalize it even more.

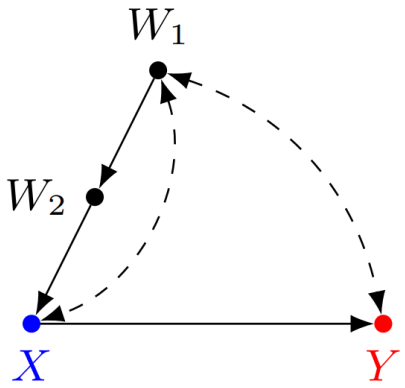
Allows us to find additional compatible graphs when the cause and effect are not mSBD but a surrogate variable r is.

Definition 4 (Surrogate criterion). $(\mathbf{R}, \mathbf{Z})$ is said to be surrogate admissible relative to $(\mathbf{X}, \mathbf{Y})$ if (1) $(\mathbf{Y} \perp\!\!\!\perp \mathbf{R}|\mathbf{X})_{G_{\overline{\mathbf{XR}}}}$; (2) $(\mathbf{Y} \perp\!\!\!\perp \mathbf{X}|\mathbf{R})_{G_{\underline{\mathbf{X}}\overline{\mathbf{R}}}}$; and (3) $\mathbf{Z}$ is mSBD admissible relative to $(\mathbf{R}, (\mathbf{X}, \mathbf{Y}))$.

Theorem 3. *If* $(\mathbf{R}, \mathbf{Z})$ *is surrogate admissible relative to* $(\mathbf{X}, \mathbf{Y})$, *then*[4]

$$\mathbb{E}_{P_{\mathbf{x}}(\mathbf{y})}\left[h\left(\mathbf{Y}\right)\right] = \mathcal{B}\left[h\left(\mathbf{Y}\right) \mid \mathbf{x} \cup \mathbf{r}; \mathcal{W}_{mSBD}(\mathbf{r}, \mathbf{x} \cup \mathbf{y}, \mathbf{z})\right].$$

Example Surrogate endpoint graph from before.



$$\boldsymbol{R} = \{\boldsymbol{W_2}\}$$
$$\boldsymbol{Z} = \{\boldsymbol{W_1}\}$$

$$\mathbb{E}_{P_x(y)}\left[Y\right] = \mathcal{B}\left[Y \;\middle|\; \{w_2, x\}; \mathcal{W} = \frac{P\left(w_2\right)}{P\left(w_2|w_1\right)}\right]$$

Estimation Empirical Weighted Operators

Now we wish to estimate the expectation given a Dataset

Supposed we have gathered N samples in the form of $D_{obs} = \{V_i\}$, a dataset drawn from your chosen casual model.

Definition 8 (Empirical weighting operator $\widehat{\mathcal{B}}$)**.** Given $D_{obs} = \{\mathbf{V}_{(i)}\}_{i=1}^{N} \sim P(\mathbf{v})$, the empirical weighting operator $\widehat{\mathcal{B}}[h(\mathbf{Y}) \mid \mathbf{x}; \mathcal{W}](\mathbf{x}) \equiv g^*(\mathbf{x})$ is estimated by the weighted regression as follows:

$$g^* = \arg\min_{\widehat{g} \in \mathcal{R}} \sum_{i=1}^{N} \widehat{\mathcal{W}}(\mathbf{V}_{(i)}) \left(h(\mathbf{Y}_{(i)}) - \widehat{g}(\mathbf{X}_{(i)})\right)^2, \quad (22)$$

where $\widehat{\mathcal{W}}(\mathbf{v})$ is the empirically estimated $\mathcal{W}(\mathbf{v})$, and $\mathcal{R}$ is a class of regression functions (e.g., linear regressions).

$\widehat{W}$ will typically be the weights derived from our *mSBD* explicit baselines.

R is much more free-form but may typically be linear regression for continuous Y or logistic regression for a discrete Y. However, it may be any optimizable class of functions in general (neural networks, etc.)

Estimation Empirical Weighted Operators

Definition 8 (Empirical weighting operator $\widehat{\mathcal{B}}$)**.** Given $D_{obs} = \{\mathbf{V}_{(i)}\}_{i=1}^{N} \sim P(\mathbf{v})$, the empirical weighting operator $\widehat{\mathcal{B}}[h(\mathbf{Y}) \mid \mathbf{x}; \mathcal{W}](\mathbf{x}) \equiv g^*(\mathbf{x})$ is estimated by the weighted regression as follows:

$$g^* = \arg\min_{\widehat{g} \in \mathcal{R}} \sum_{i=1}^{N} \widehat{\mathcal{W}}(\mathbf{V}_{(i)}) \left(h(\mathbf{Y}_{(i)}) - \widehat{g}(\mathbf{X}_{(i)})\right)^2, \quad (22)$$

where $\widehat{\mathcal{W}}(\mathbf{v})$ is the empirically estimated $\mathcal{W}(\mathbf{v})$, and $\mathcal{R}$ is a class of regression functions (e.g., linear regressions).

Presenter Interpretation™

Replace the true functions in our SCM with our own custom deterministic functions (essentially converting it to a linear SCM or similar model) and optimize this model with weighted mean squared error.

What does this empirical operator give us?

This allows us to get rid of the expensive sum over the mediator variables and replace it with a parameterized model. This model only must estimate a representation of the expectation, not the entire distribution $P_x(y)$ so it should be relatively easy to train.

Estimation Empirical Operator Algebra

Theorem 6 (Consistency of the composition). *Let* $\widehat{\mathcal{B}}_1(\mathbf{x})$ *and* $\widehat{\mathcal{B}}_2(\mathbf{z})$ *be consistent estimators of* $\mathcal{B}_1(\mathbf{x})$ *and* $\mathcal{B}_2(\mathbf{z})$. *Let the function class* $\mathcal{R}_1$ *of* $\widehat{\mathcal{B}}_1$ *be a compact space. Then,* $\left(\widehat{\mathcal{B}}_1 \circ \widehat{\mathcal{B}}_2\right)(\mathbf{x})$ *is a consistent estimator of* $(\mathcal{B}_1 \circ \mathcal{B}_2)(\mathbf{x})$.

This approximate operator is compatible with the true operator algebra!

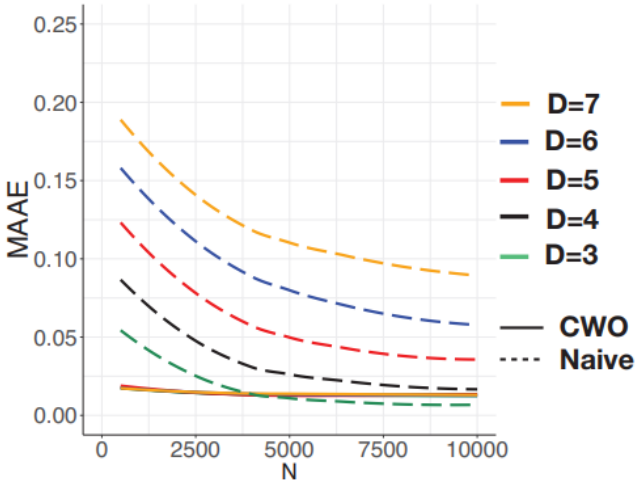
One of the major results from this paper is the proof for this theorem, which allows us to combine the decomposition techniques we discovered before with these cheap empirical expectation estimates!

This allows us to find a decomposition of the true SCM and then estimate each component independently.

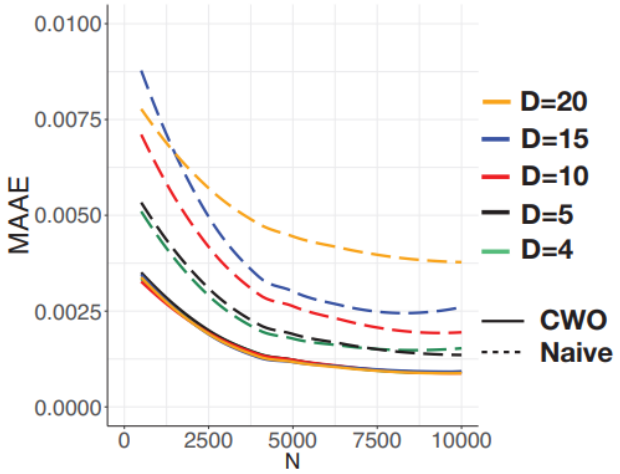
Results Error Curves For Example Graphs

Evaluate Composition of Weighting Operators (CWO)

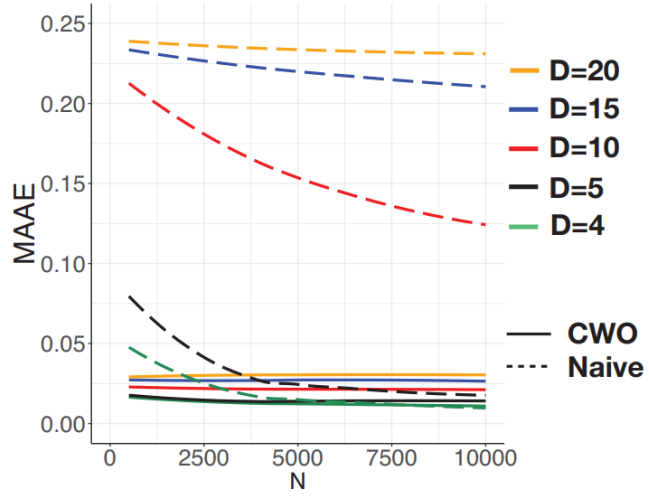
- Compare to the manually derived and sampled formula (Naïve).
- Evaluate using Median Average Absolute Error – the median of AAE across 100 random graphs.
- Vary the dimensionality of the covariate variables (Z and W).



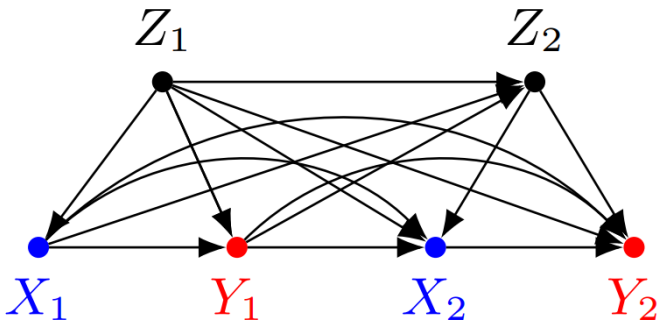
(a) mSBD (Fig. 2a)

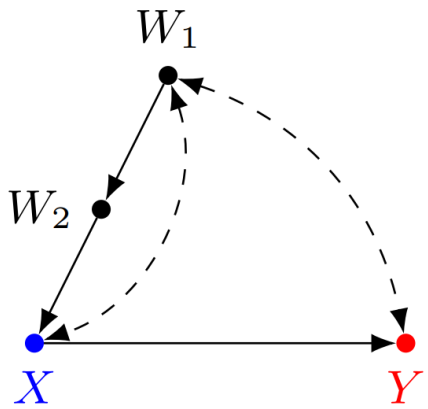


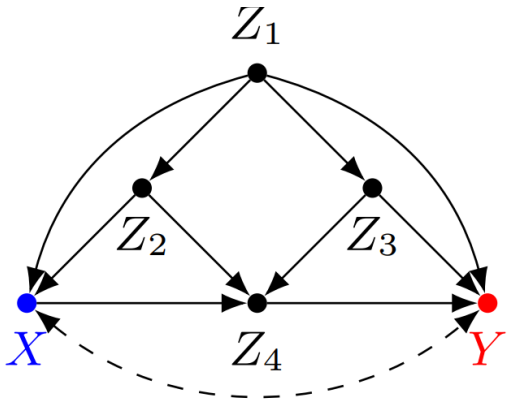
(b) Surrogate endpoints



(c) Causal mediators (Fig. 1b)

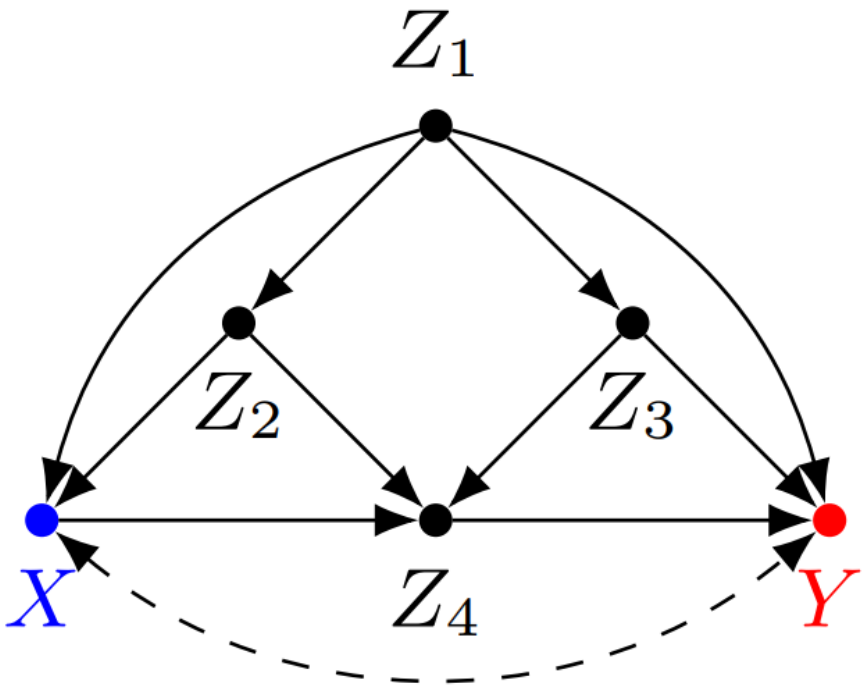






Homework

Construct a composition of weighted operators model for the following graph using back-door weighted operators as the components.



Z_1: Age

Z_2: Diet

Z_3: Smoking

Z_4: Blood Test Results

X: BMI

Y: Cancer Presence

Thank You

Jung, Y., Tian, J., & Bareinboim, E. (2020). *Estimating Causal Effects Using Weighting-Based Estimators*. Proceedings of the AAAI Conference on Artificial Intelligence, 34(06), 10186-10193. https://doi.org/10.1609/aaai.v34i06.6579