# Detecting and Measuring Confounding Using Causal Mechanism Shifts

Abbavaram Gowtham Reddy
Vineeth N Balasubramanian

NeurIPS 2024

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

NEURAL INFORMATION PROCESSING SYSTEMS

# Preliminaries: Confounding Variables

## What is a confounding variable?

- Confounding variables induces spurious associations.

# Preliminaries: Confounding Variables

## What is a confounding variable?

- Confounding variables induces spurious associations.

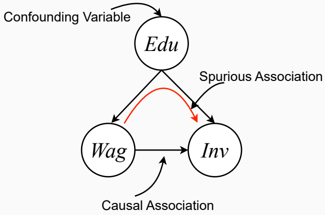



**Figure 1:** Edu: Education, Wag: Wages/Income, Inv: Investments.

# Preliminaries: Confounding Variables

## What is a confounding variable?

- Confounding variables induces spurious associations.

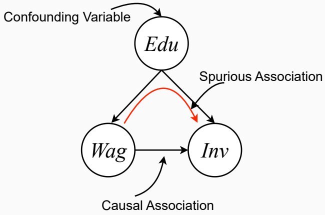



**Figure 1:** Edu: Education, Wag: Wages/Income, Inv: Investments.

## Why study confounding?

- Distinguish between causal and spurious associations.

# Preliminaries: Confounding Variables

## What is a confounding variable?

- Confounding variables induces spurious associations.

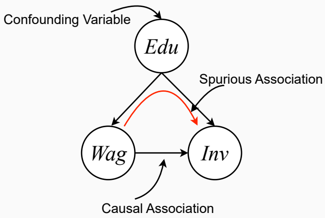



**Figure 1:** Edu: Education, Wag: Wages/Income, Inv: Investments.

## Why study confounding?

- Distinguish between causal and spurious associations.
- Estimate causal effects by adjusting confounding variables.

# Preliminaries: Causal Mechanisms

What are causal mechanisms and how do they shift?

- $\mathbb{P}(X|\mathbf{PA}_x)$ is the causal mechanism of $X$.

# Preliminaries: Causal Mechanisms

What are causal mechanisms and how do they shift?

- $\mathbb{P}(X|\mathbf{PA}_x)$ is the causal mechanism of $X$.
- $\mathbb{P}^c(X|\mathbf{PA}_x) \neq \mathbb{P}^{c'}(X|\mathbf{PA}_x) \implies$ mechanism change.
- $c, c'$ are known as context/environments/domains, etc.

# Preliminaries: Causal Mechanisms

What are causal mechanisms and how do they shift?

- $\mathbb{P}(X|\mathbf{PA}_x)$ is the causal mechanism of $X$.
- $\mathbb{P}^c(X|\mathbf{PA}_x) \neq \mathbb{P}^{c'}(X|\mathbf{PA}_x) \implies$ mechanism change.
- $c, c'$ are known as context/environments/domains, etc.
- Interventions create contexts.

# Preliminaries: Causal Mechanisms

What are causal mechanisms and how do they shift?

- $\mathbb{P}(X|\mathbf{PA}_x)$ is the causal mechanism of $X$.
- $\mathbb{P}^c(X|\mathbf{PA}_x) \neq \mathbb{P}^{c'}(X|\mathbf{PA}_x) \implies$ mechanism change.
- $c, c'$ are known as context/environments/domains, etc.
- Interventions create contexts.
- Interventions can be soft or hard.
- Hard intervention: $X$ is set to $x$.
- Soft intervention: $\mathbb{P}(X)$ is changed to $\tilde{\mathbb{P}}(X)$.

# Preliminaries: Causal Mechanisms

What are causal mechanisms and how do they shift?

- $\mathbb{P}(X|\mathbf{PA}_x)$ is the causal mechanism of $X$.
- $\mathbb{P}^c(X|\mathbf{PA}_x) \neq \mathbb{P}^{c'}(X|\mathbf{PA}_x) \implies$ mechanism change.
- $c, c'$ are known as context/environments/domains, etc.
- Interventions create contexts.
- Interventions can be soft or hard.
- Hard intervention: $X$ is set to $x$.
- Soft intervention: $\mathbb{P}(X)$ is changed to $\tilde{\mathbb{P}}(X)$.

**Next:** Detecting and measuring confounding using mechanism shifts.

# Method: Detecting and Measuring Confounding

- Let $\mathbf{X}$ be a set of observed variables.

# Method: Detecting and Measuring Confounding

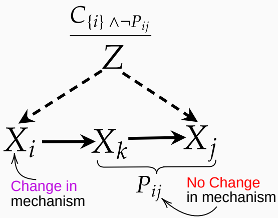
- Let $\mathbf{X}$ be a set of observed variables.
- $\mathbf{X}_S \subset \mathbf{X}$ be a set of variables indexed by $S$.

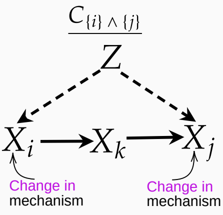
# Method: Detecting and Measuring Confounding

- Let $\mathbf{X}$ be a set of observed variables.
- $\mathbf{X}_S \subset \mathbf{X}$ be a set of variables indexed by $S$.
- $\mathbf{C}_{S \wedge \neg R}$ be the contexts with mechanism changes for $\mathbf{X}_S$ but not $\mathbf{X}_R$.

# Method: Detecting and Measuring Confounding

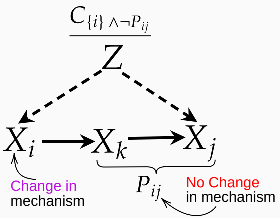
- Let $\mathbf{X}$ be a set of observed variables.
- $\mathbf{X}_S \subset \mathbf{X}$ be a set of variables indexed by $S$.
- $\mathbf{C}_{S \wedge \neg R}$ be the contexts with mechanism changes for $\mathbf{X}_S$ but not $\mathbf{X}_R$.
- Consider three sets of contexts: $\mathbf{C}_{\{i\} \wedge \neg P_{ij}}$ , $\mathbf{C}_{\{j\} \wedge \neg P_{ji}}$, $\mathbf{C}_{\{i\} \wedge \{j\}}$.



# Method: Detecting and Measuring Confounding

- Let $\mathbf{X}$ be a set of observed variables.
- $\mathbf{X}_S \subset \mathbf{X}$ be a set of variables indexed by $S$.
- $\mathbf{C}_{S \wedge \neg R}$ be the contexts with mechanism changes for $\mathbf{X}_S$ but not $\mathbf{X}_R$.
- Consider three sets of contexts: $\mathbf{C}_{\{i\} \wedge \neg P_{ij}}$ , $\mathbf{C}_{\{j\} \wedge \neg P_{ji}}$, $\mathbf{C}_{\{i\} \wedge \{j\}}$.




| Settings | Confounding Definition | Required Contexts | Intervention |
|---|---|---|---|
| 1 | Directed Information & Noncollapsibility | $\mathbf{C}_{\{i\} \wedge \neg P_{ij}}$ $\mathbf{C}_{\{j\} \wedge \neg P_{ji}}$ | Hard / Structural |
| 2 & 3 | Mutual Information | $\mathbf{C}_{\{i\} \wedge \{j\}}$ | Soft / Parametric |

# Setting 1: Detecting and Measuring Confounding

## Directed Information

$$I(X_i \to X_j) := D_{KL}(\mathbb{P}(X_i|X_j)||\mathbb{P}(X_i|do(X_j))|\mathbb{P}(X_j)) := \mathbb{E}_{\mathbb{P}(X_i,X_j)} \log \frac{\mathbb{P}(X_i|X_j)}{\mathbb{P}(X_i|do(X_j))}$$

# Setting 1: Detecting and Measuring Confounding

## Directed Information

$$I(X_i \to X_j) := D_{KL}(\mathbb{P}(X_i|X_j)||\mathbb{P}(X_i|do(X_j))|\mathbb{P}(X_j)) := \mathbb{E}_{\mathbb{P}(X_i,X_j)} \log \frac{\mathbb{P}(X_i|X_j)}{\mathbb{P}(X_i|do(X_j))}$$

| | Graph | $I(X_i \to X_j)$ | $I(X_j \to X_i)$ |
|---|---|---|---|
| Uncnf. | $X_i \to X_j$ | $> 0$ | $= 0$ |
| Uncnf. | $X_j \to X_i$ | $= 0$ | $> 0$ |
| Confounded | $X_i \to X_j$<br>$Z \to X_i, Z \to X_j$ | $> 0$ | $> 0$ |
| Confounded | $X_j \to X_i$<br>$Z \to X_i, Z \to X_j$ | $> 0$ | $> 0$ |

**Table 1:** Directed information for various graphs.

# Setting 1: Detecting and Measuring Confounding

## Confounding Measure

Given the contexts $\mathbf{C}_{\{i\}\wedge\neg P_{ij}}$ and $\mathbf{C}_{\{j\}\wedge\neg P_{ji}}$, the measure of confounding $CNF(X_i, X_j)$ is defined as

$$CNF(X_i, X_j) := 1 - e^{-\min(I(X_i \to X_j), I(X_j \to X_i))}$$

# Setting 1: Detecting and Measuring Confounding

## Confounding Measure

Given the contexts $\mathbf{C}_{\{i\}\wedge\neg P_{ij}}$ and $\mathbf{C}_{\{j\}\wedge\neg P_{ji}}$, the measure of confounding $CNF(X_i, X_j)$ is defined as

$$CNF(X_i, X_j) := 1 - e^{-\min(I(X_i \to X_j), I(X_j \to X_i))}$$

- Why 'min'? Why exponential?

# Setting 1: Detecting and Measuring Confounding

**Confounding Measure**

Given the contexts $\mathbf{C}_{\{i\}\wedge\neg P_{ij}}$ and $\mathbf{C}_{\{j\}\wedge\neg P_{ji}}$, the measure of confounding $CNF(X_i, X_j)$ is defined as

$$CNF(X_i, X_j) := 1 - e^{-\min(I(X_i \to X_j), I(X_j \to X_i))}$$

- Why 'min'? Why exponential?
- Why directed information from both directions?

# Setting 1: Detecting and Measuring Confounding

**Confounding Measure**

Given the contexts $\mathbf{C}_{\{i\} \wedge \neg P_{ij}}$ and $\mathbf{C}_{\{j\} \wedge \neg P_{ji}}$, the measure of confounding $CNF(X_i, X_j)$ is defined as

$$CNF(X_i, X_j) := 1 - e^{-\min(I(X_i \rightarrow X_j), I(X_j \rightarrow X_i))}$$

- Why 'min'? Why exponential?
- Why directed information from both directions?
- Get $\mathbb{P}(X_j | X_i)$ using observational data.
- $\mathbb{P}(X_j | do(X_i)) = \mathbb{E}_{c \in \mathbf{C}_{\{i\} \wedge \neg P_{ij}}}[\mathbb{P}^c(X_j | X_i)]$

# Setting 1: Detecting and Measuring Confounding

## Conditional Directed Information

$$
\begin{aligned}
I(X_i \to X_j | X_o) &:= D_{KL}(\mathbb{P}(X_i|X_j, X_o) || \mathbb{P}(X_i|do(X_j), X_o) | \mathbb{P}(X_j, X_o)) \\
&:= \mathop{\mathbb{E}}_{\mathbb{P}(X_i, X_j, X_o)} \log \frac{\mathbb{P}(X_i|X_j, X_o)}{\mathbb{P}(X_i|do(X_j), X_o)}
\end{aligned}
$$

- Measure unobserved confounding by conditioning on observed confounding.
- Measure of conditional confounding can be calculated as

$$
CNF(X_i, X_j | X_o) := 1 - e^{-\min(I(X_i \to X_j | X_o), I(X_j \to X_i | X_o))}
$$

# Setting 1: Detecting and Measuring Confounding

- How to know whether a set $\mathbf{X}_S$ of variables share a common confounder?

# Setting 1: Detecting and Measuring Confounding

- How to know whether a set $\mathbf{X}_S$ of variables share a common confounder?

**Theorem**

*A set of observed variables $\mathbf{X}_S$ are jointly unconfounded if and only if there exists three variables $X_i, X_j, X_k \in \mathbf{X}_S$ such that $I(X_i \rightarrow X_j | X_k) = I(\{X_i X_k\} \rightarrow X_j)$.*

# Setting 1: Detecting and Measuring Confounding

- How to know whether a set $\mathbf{X}_S$ of variables share a common confounder?

**Theorem**

*A set of observed variables $\mathbf{X}_S$ are jointly unconfounded if and only if there exists three variables $X_i, X_j, X_k \in \mathbf{X}_S$ such that $I(X_i \to X_j | X_k) = I(\{X_i X_k\} \to X_j)$.*

- Joint confounding effect among a set $\mathbf{X}_S$ of variables is defined as

$$CNF(\mathbf{X}_S) = \sum_{i \in S} CNF(\mathbf{X}_{S \setminus \{i\}}, X_i)$$

# Settings 2 and 3: Detecting and Measuring Confounding

- **Setting 2:** Given $\mathbf{C}_{\{i\}\wedge\{j\}}$, use mutual information between $\mathbb{E}(X_i), \mathbb{E}(X_j)$.
- **Setting 3:** Given $\mathbf{C}_{\{i\}\cup\{j\}}$, use mutual information among $\mathbb{E}(X_i), \mathbb{E}(X_j)$,$\mathbb{E}(X_i|X_j), \mathbb{E}(X_j|X_i)$.

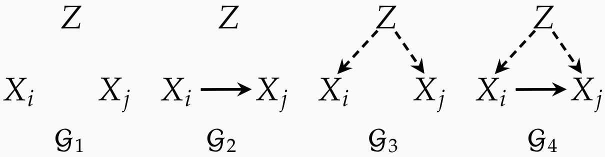
# Settings 2 and 3: Detecting and Measuring Confounding

- **Setting 2:** Given $\mathbf{C}_{\{i\}\wedge\{j\}}$, use mutual information between $\mathbb{E}(X_i), \mathbb{E}(X_j)$.
- **Setting 3:** Given $\mathbf{C}_{\{i\}\cup\{j\}}$, use mutual information among $\mathbb{E}(X_i), \mathbb{E}(X_j)$,$\mathbb{E}(X_i|X_j), \mathbb{E}(X_j|X_i)$.
- We propose pairwise, joint, and conditional confounding.

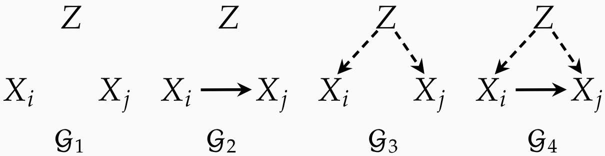
# Settings 2 and 3: Detecting and Measuring Confounding

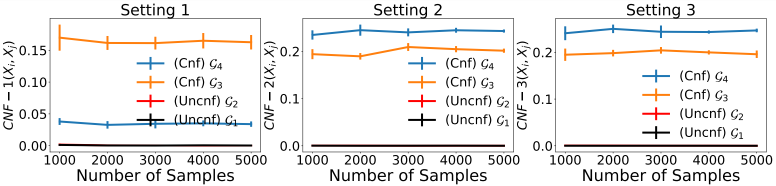
- **Setting 2:** Given $\mathbf{C}_{\{i\}\wedge\{j\}}$, use mutual information between $\mathbb{E}(X_i), \mathbb{E}(X_j)$.
- **Setting 3:** Given $\mathbf{C}_{\{i\}\cup\{j\}}$, use mutual information among $\mathbb{E}(X_i), \mathbb{E}(X_j)$,$\mathbb{E}(X_i|X_j), \mathbb{E}(X_j|X_i)$.
- We propose pairwise, joint, and conditional confounding.
- Symmetry: $CNF(X_i, X_j|X_o) = CNF(X_j, X_i|X_o)$.
- Positivity: $CNF(X_i, X_j|X_o) > 0$ if and only if there exists an unobserved confounding variable $Z$ between $X_i, X_j$.
- Monotonicity: $CNF(X_i, X_j) > CNF(X_k, X_l) \implies X_i, X_j$ are strongly confounded than $X_k, X_l$.

# Results: Detecting and Measuring Confounding



# Results: Detecting and Measuring Confounding



# Results: Detecting and Measuring Confounding - Results

| | | Setting 1 | | | Setting 2 | | | Setting 3 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $N$, $\|C\|$ | **Sample Size** | **Precision** | **Recall** | **F1** | **Precision** | **Recall** | **F1** | **Precision** | **Recall** | **F1** |
| 10 | 100 | 0.64 | 0.97 | 0.77 | 0.67 | 0.83 | 0.74 | 0.64 | 0.72 | 0.68 |
| 10 | 200 | 0.64 | 1.0 | 0.78 | 0.67 | 0.83 | 0.74 | 0.70 | 0.79 | 0.74 |
| 10 | 300 | 0.64 | 1.0 | 0.78 | 0.67 | 0.83 | 0.74 | 0.65 | 0.76 | 0.70 |
| 10 | 400 | 0.64 | 1.0 | 0.78 | 0.67 | 0.83 | 0.74 | 0.67 | 0.83 | 0.74 |
| 10 | 500 | 0.64 | 1.0 | 0.78 | 0.67 | 0.83 | 0.74 | 0.67 | 0.83 | 0.74 |
| 15 | 100 | 0.81 | 0.95 | 0.88 | 0.80 | 0.85 | 0.82 | 0.80 | 0.79 | 0.80 |
| 15 | 200 | 0.82 | 1.0 | 0.90 | 0.80 | 0.85 | 0.82 | 0.80 | 0.85 | 0.82 |
| 15 | 300 | 0.82 | 1.0 | 0.90 | 0.80 | 0.85 | 0.82 | 0.80 | 0.85 | 0.82 |
| 15 | 400 | 0.82 | 1.0 | 0.90 | 0.80 | 0.85 | 0.82 | 0.80 | 0.85 | 0.82 |
| 15 | 500 | 0.82 | 1.0 | 0.90 | 0.80 | 0.85 | 0.82 | 0.80 | 0.84 | 0.82 |
| 20 | 100 | 0.68 | 0.95 | 0.80 | 0.68 | 0.88 | 0.77 | 0.69 | 0.84 | 0.76 |
| 20 | 200 | 0.69 | 1.0 | 0.82 | 0.68 | 0.88 | 0.77 | 0.68 | 0.87 | 0.76 |
| 20 | 300 | 0.69 | 1.0 | 0.82 | 0.68 | 0.88 | 0.77 | 0.67 | 0.86 | 0.75 |
| 20 | 400 | 0.69 | 1.0 | 0.82 | 0.68 | 0.88 | 0.77 | 0.68 | 0.87 | 0.76 |
| 20 | 500 | 0.69 | 1.0 | 0.82 | 0.68 | 0.88 | 0.77 | 0.68 | 0.87 | 0.76 |
| 25 | 100 | 0.83 | 0.96 | 0.89 | 0.83 | 0.91 | 0.87 | 0.83 | 0.89 | 0.86 |
| 25 | 200 | 0.83 | 1.0 | 0.91 | 0.83 | 0.91 | 0.87 | 0.82 | 0.90 | 0.86 |
| 25 | 300 | 0.83 | 1.0 | 0.91 | 0.83 | 0.91 | 0.87 | 0.83 | 0.91 | 0.87 |
| 25 | 400 | 0.83 | 1.0 | 0.91 | 0.83 | 0.92 | 0.87 | 0.83 | 0.91 | 0.87 |
| 25 | 500 | 0.83 | 1.0 | 0.91 | 0.83 | 0.91 | 0.87 | 0.83 | 0.91 | 0.87 |

**Table 2:** Results on synthetic datasets for settings 1,2,3.