Detecting and Measuring Confounding Using Causal Mechanism Shifts

Abbavaram Gowtham Reddy
Indian Institute of Technology Hyderabad
cs19resch11002@iith.ac.in

Vineeth N Balasubramanian
Indian Institute of Technology Hyderabad
vineethnb@iith.ac.in

Abstract

Detecting and measuring confounding effects from data is a key challenge in causal inference. Existing methods frequently assume causal sufficiency, disregarding the presence of unobserved confounding variables. Causal sufficiency is both unrealistic and empirically untestable. Additionally, existing methods make strong parametric assumptions about the underlying causal generative process to guarantee the identifiability of confounding variables. Relaxing the causal sufficiency and parametric assumptions and leveraging recent advancements in causal discovery and confounding analysis with non-i.i.d. data, we propose a comprehensive approach for detecting and measuring confounding. We consider various definitions of confounding and introduce tailored methodologies to achieve three objectives: (i) detecting and measuring confounding among a set of variables, (ii) separating observed and unobserved confounding effects, and (iii) understanding the relative strengths of confounding bias between different sets of variables. We present useful properties of a confounding measure and present measures that satisfy those properties. Our empirical results support the usefulness of the proposed measures.

1 Introduction

Understanding the underlying causal generative process of a set of variables is crucial in many scientific studies for applications in treatment and policy designs [44]. While randomized controlled trials (RCTs) and causal inference through active interventions are ideal choices for understanding the underlying causal model [19, 12, 13, 55], RCTs and/or active interventions are often impossible/infeasible, and some times unethical [50, 6]. Research efforts in causal inference hence rely on observational data to study causal relationships [44, 59, 65, 18, 41]. However, recovering the underlying causal model purely from observational data is challenging without further assumptions; this challenge is further exacerbated in the presence of unmeasured confounding variables.

A confounding variable is a variable that *causes* two other variables, resulting in a spurious association between those two variables. As exemplified with *Simpson's paradox* [58] and many other studies [20, 1, 31], the presence of confounding variables is an important quantitative explanation for why *correlation does not imply causation*. It is challenging to observe and measure all confounding variables in a scientific study [60, 44]. Identifying *latent* or *unobserved* confounding variables is even more challenging, and misinterpretation presents various challenges in downstream applications, such as discovering causal structures from observational data. Numerous methods operate under the assumption of *causal sufficiency* [45, 4, 60, 8, 51, 65], implying the non-existence of unobserved confounding variables. Causal sufficiency presupposes that all pertinent variables required for causal inference have been observed. However, this may not be a practical or testable assumption.

The study of confounding has various applications, chief among them being causal discovery - identifying the causal relationships among variables [38, 40, 63]. It is also useful for determining whether a set of observed confounding variables is sufficient to adjust for estimating causal effects [29], measuring the extent to which statistical correlation between variables can be attributed to confound-

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

ing [24, 25, 62], and verifying the comparability of treatment and control groups in non-randomized interventional studies [16].

A fundamental problem in causal inference tasks lies in detecting hidden confounding variables from observational data alone. However, this is non-trivial and poses various challenges. For example, a key issue is that given a marginal distribution over observed variables, there are infinitely many joint distributions corresponding to causal graphs involving unobserved variables [56]. To tackle such challenges, recent endeavors show that using data from different environments helps in improved causal discovery [40, 38, 33, 45, 23], detecting causal mechanism shifts [36], and detecting unobserved confounding [29, 38]. However, such recent efforts often subsume confounding detection under causal discovery, focusing primarily on identifying confounding factors while overlooking other useful information, such as the relative strength of confounding between variable sets and the distinction between observed and unobserved confounding within a variable set. We seek to address these gaps in this work.

We focus exclusively on the problem of studying confounding from multiple perspectives, including (i) detecting and measuring confounding among a set of variables, (ii) assessing the relative strengths of confounding among different sets of observed variables, and (iii) distinguishing between observed and unobserved confounding among a set of variables. *The primary focus of causal inference often lies in verifying the presence or absence of confounding rather than determining the exact value of the measured confounding. However, we leverage the measured confounding to assess the relative strengths of confounding between sets of variables. To achieve the above objectives, we utilize data from various contexts, where each context results from shifts in the causal mechanisms of a set of variables [38, 45].* This allows us to propose different measures of confounding based on the available context information. Our contributions can be summarized as follows.

- For various definitions of confounding, we propose corresponding measures of confounding and present useful properties of the proposed measures. To our knowledge, this is the first comprehensive study that examines various aspects of observed and unobserved confounding using data from multiple contexts without making parametric or causal sufficiency assumptions.
- We study pair-wise confounding, confounding among multiple variables, how to separate unobserved confounding from overall confounding, and present ways to assess relative confounding.
- We present an algorithm for detecting and measuring confounding using data from multiple contexts. Experimental results are performed to verify theoretical analysis.

2 Related Work

The study of confounding has typically been embedded as part of causal discovery algorithms in most existing work. Causal discovery methods can be categorized according to several criteria, including the type of data utilized (observational versus interventional/experimental), parametric versus non-parametric approaches, or whether they relax causal sufficiency assumptions [65, 59]. Considering our focus in this work on studying confounding comprehensively by going beyond observed confounding variables, we discuss literature that are directed towards methods that relax the causal sufficiency assumption and rely on experimental data.

Causal Discovery via Observational Data, Relaxing Causal Sufficiency: Constraint based causal discovery algorithms produce equivalence class of graphs that satisfy a set of conditional independence constraints [60, 11, 9, 42]. Other methods such as [2, 28, 27] reduce the problem complexity by assuming a parametric form of the underlying causal model (e.g., variables are jointly Gaussian in Chandrasekaran et al. [7]), thereby returning unique causal graphs. Nested Markov Models (NMMs) [56, 57, 49, 14] allow identifiability of causal models with latent factors by using (pairwise) Verma constraints. A recent approach using differentiable causal discovery [2] combines NMMs with the differentiable constraint [66] to discover a partially directed causal network and likely confounded nodes. Unlike these methods, our focus in this work is on detecting and measuring confounding under various settings, instead of recovering the entire causal graph or equivalence class.

Causal Discovery Using Data From Multiple Environments: Given access to a set of observed confounding variables, very recent work [29] presented testable conditional independence tests that are violated only when there is unobserved confounding. However, their analysis is focused towards the downstream causal effect estimation. We aim to provide a unified framework for studying and measuring confounding under different types of contextual information available.

Other methods [33, 23] learn an equivalence class of graphs when data from observational and interventional distribution are available. Confounding has also shown to be detected in linear models with non-Gaussian variables [20]. In linear models, a spectral analysis method was proposed in [25] to understand to what extent the statistical correlation between a set of variables on a target variable can be attributed to confounding. See Tab. 4 of [40] for an overview of causal discovery methods that use data from multiple environments or contexts. Under the specific assumptions of causal sufficiency and sparse mechanism shift, a method was proposed in [45] to reduce the size of a given Markov equivalence class using mechanism shift score. A differentiable causal discovery method was proposed in [4] to use interventional data to recover interventional Markov equivalence class. While these methods use data from different contexts, they assume the absence of unobserved confounding variables; we instead focus on capturing both observed and unobserved confounding.

Measuring and Interpreting Confounding: Earlier efforts in the field have studied different measures for observed confounding, each tailored to address specific challenges [44, 15, 35, 3, 39, 30, 43, 34]. Such measures have also been refined to address specific issues [24, 54]; for e.g., a method to correct the non-linearity effect present in confounding estimates via the exposure–outcome association with and without adjustment for confounding was proposed in [24]. In contrast, we measure the effects of both observed and *unobserved* confounding. Motivated from the *ignorability* property in potential outcomes framework [61, 26], the divergence between nominal and complete propensity density has been considered as an indicative of hidden confounding [26]. To the best of our knowledge, the efforts closest to ours are [38, 40], which study confounding using data from multiple contexts without the causal sufficiency assumption. However, they *do not measure confounding* and detect confounding only as a step to discover the causal graph. Ours is a more general framework for studying and measuring confounding from multiple perspectives.

In regression models, certain difference thresold between the coefficients of treatment variable before and after adjusting for the possible confounding is considered as the indication for the presence of confounding. This process of choosing a threshold is also called *change-in-estimate* criterion. Typical threshold used in literature is 10% [54, 32, 5].

3 Background and Problem Setup

Let $\mathbf{X}$ be a set of observed variables and $\mathbf{Z}$ be a set of unobserved or latent variables. The values of $\mathbf{X}, \mathbf{Z}$ can be real, discrete, or mixed. Let $\mathcal{G}$ be the underlying directed acyclic graph (DAG) among the variables $\mathbf{V} = \mathbf{X} \cup \mathbf{Z}$. Directed edges among the variables in $\mathbf{V}$ indicate direct causal influences. Assume that the set of unobserved variables $\mathbf{Z}$ are jointly independent and are exogenous to $\mathbf{X}$ (i.e., $Z_i \perp\!\!\!\perp Z_j$ and $X_k \not\to Z_j \;\; \forall i, j, k$). In this setting, any two nodes $X_i, X_j \in \mathbf{X}$ sharing a common parent $Z_k \in \mathbf{Z}$ are said to be confounded, and Z_k is said to be a confounding variable. For a node $X_i \in \mathbf{X}, \mathbf{PA}_i = \{X_j \in \mathbf{X} | X_j \to X_i\} \cup \{Z_j \in \mathbf{Z} | Z_j \to X_i\}$ denotes the set of parents of X_i.

For a node X_i, $\mathbb{P}(X_i|\mathbf{PA}_i)$ is called the *causal mechanism* of X_i. The causal mechanism $\mathbb{P}(X_i|\mathbf{PA}_i)$ encodes how the variable X_i is influenced by its parents $\mathbf{PA}_i$. Following earlier work [22, 38, 45, 21, 46, 52], we make the following general assumption about the underlying causal mechanisms of data.

Assumption 3.1. ***(Independent Causal Mechanisms [44, 47])*** *A change in $\mathbb{P}(X_i|\mathbf{PA}_i)$ has no effect on and provides no information about $\mathbb{P}(X_j|\mathbf{PA}_j)$ $\forall j \neq i$.*

Identifying confounding from only observational data is challenging without further assumptions [28]. Hence, following earlier work [38, 29, 40], we assume that the data over the variables $\mathbf{X}$ is observed over multiple *contexts or environments*. While there are various ways of formulating/constructing contexts, in this paper, we assume that each context is created as a result of either *hard (a.k.a. structural)* interventions or *soft (a.k.a. parametric)* interventions on a subset $\mathbf{V}_S \subseteq \mathbf{V}$ of variables where S is a set of indices. Performing hard intervention on a variable V_i is the same as setting the value of V_i to a value v_i. Hard intervention on a variable V_i removes the influence of its parents $\mathbf{PA}_i$ on V_i. Performing soft intervention on a variable V_i is the same as changing the causal mechanism of V_i, $\mathbb{P}(V_i|\mathbf{PA}_i)$, with a new causal mechanism $\tilde{\mathbb{P}}(V_i|\mathbf{PA}_i)$. Soft intervention on a variable V_i does not remove the influence of its parents $\mathbf{PA}_i$ on V_i. The idea of explicitly considering context information and using different contexts as context variables to create extended causal graphs has been studied in the literature. Context variables are also called as *policy variables, decision variables, regime variables, domain variables, environment variables, etc.* [40, 45, 17, 22].

Let $\mathbf{C} = \{c_1, c_2, \ldots, c_n\}$ be the set of n contexts and let $\mathbb{P}^c(\mathbf{X}), c \in \mathbf{C}$, denotes the probability distribution of the observed variables $\mathbf{X}$ in the context c. Let $\mathbf{C}_{S \wedge R}$, where S, R are sets of indices, be

the set of contexts in which we observe mechanism changes for the set of variables $\mathbf{X}_{S\cup R}$. Similarly, let $\mathbf{C}_{S\wedge\neg R}$ be the set of contexts in which we observe mechanism changes for the set of variables $\mathbf{X}_S$ but not for the variables $\mathbf{X}_R$. We say that the causal mechanism of a variable X_i changes between two contexts c, c' if $\mathbb{P}^c(X_i|\mathbf{PA}_i) \neq \mathbb{P}^{c'}(X_i|\mathbf{PA}_i)$. Given the data over observed variables in each context, there exist methods for detecting mechanism shifts of each variable between the contexts [36, 38, 45, 37]. For example, the p-value$(\mathbb{P}^c(X_i|\mathbf{PA}_i^o) \neq \mathbb{P}^{c'}(X_i|\mathbf{PA}_i^o))$ where $\mathbf{PA}_i^o$ is the set of observed parents of X_i can be used to detect mechanism change for X_i between the contexts c, c' [38, 36]. Hence, we focus on detecting and measuring confounding among a set of variables, assuming that the causal mechanism shifts are observed among that set of variables.

Context information is not very useful if there is no restriction on how causal mechanisms are changed between the contexts [45, 38]. For example, the causal mechanisms of X_i and X_j both differing across all (or no) contexts would trivially satisfy Assumption 3.1, but reveal no information about the underlying causal mechanisms [10, 38]. Hence, following earlier work [45, 38, 17], we make the following assumptions.

Assumption 3.2. ***(Sparse Causal Mechanism Shift [53])*** *Causal mechanisms of variables change sparsely across contexts, i.e., if* $p := (\mathbb{P}^c(X_i|\mathbf{PA}_i) \neq \mathbb{P}^{c'}(X_i|\mathbf{PA}_i))$, *then* $0 < p < 0.5; \quad \forall c, c' \in \mathbf{C}$.

Assumption 3.2 implies that the causal mechanisms change infrequently across contexts. This assumption is more general because, in many scientific studies, for any given context, interventions typically affect only a few variables [53].

Assumption 3.3. ***(Markov Property under Mechanism Shifts [17])*** *The distribution* $\mathbb{P}(\mathbf{V})$ *is given by* $\mathbb{P}(\mathbf{V}) = \int \mathbb{P}^C(\mathbf{V})d\mathbb{P}(C) = \int \Pi_i \mathbb{P}^C(V_i|\mathbf{PA}_i)d\mathbb{P}(C)$. *In other words, variables* $\mathbf{V}$ *are assumed to be conditionally exchangeable, so that the same graph* $\mathcal{G}$ *applies in every context* $c \in \mathbf{C}$.

Assumption 3.4. ***(Causal Sufficiency Over*** $\mathbf{X} \cup \mathbf{Z}$***)*** *All common parents of any pair of observed nodes belong to the set* $\mathbf{X} \cup \mathbf{Z}$. *In other words, all relevant variables for detecting confounding and the unobserved confounding variables are already present in* $\mathbf{X} \cup \mathbf{Z}$.

Problem Statement: *Given data over the observed variables* $\mathbf{X}$ *in multiple contexts, each context resulting from a sparse causal mechanism shift of variables in* $\mathbf{V}$, *(i) can we identify which pairs or sets of variables are confounded and can we measure the confounding strength? (ii) can we isolate the confounding effects of observed and unobserved confounding variables? and (iii) can we study the relative strengths of confounding among different sets of variables?*

To address the above problem, in the next section, we consider various definitions of confounding and present appropriate confounding measures depending on the context information available.

4 Detecting and Measuring Confounding

In this section, we present methods for detecting and measuring confounding for various scenarios in which shifts in causal mechanisms are observed. Considering any three observed variables $X_i, X_j, X_o \in \mathbf{X}$ and an unobserved confounding variable $Z \in \mathbf{Z}$, we present measures of confounding depending on the information about mechanism shifts of X_i, X_j, X_o, Z. Each of the following subsections includes: (i) a definition of confounding, (ii) a corresponding definition of the confounding measure, (iii) a method for isolating the unobserved confounding measure from the overall confounding, (iv) an extension of the confounding measure to more than two variables, and (v) key properties of the proposed confounding measures. See Tab. 1 and Fig. 1 for an overview.

Settings	Confounding Definition Based On	Required Context Information	Type of Intervention
1	Directed Information [48] & Noncollapsibility [15, 43, 54]	$\mathbf{C}_{\{i\}\wedge\neg P_{ij}}$ $\mathbf{C}_{\{j\}\wedge\neg P_{ji}}$	Hard / Structural
2 & 3	Mutual Information	$\mathbf{C}_{\{i\}\wedge\{j\}}$	Soft / Parametric

Table 1: Summary of the various settings for detecting and measuring confounding between X_i, X_j. Here P_{ij} is the set of node indices that belong to a path from X_i to X_j including j.

4.1 Setting 1: Measuring Confounding Using Directed Information Between X_i, X_j.

In this setting, we use the fact that directed information does not vanish in the presence of a confounding variable [64, 48]. To this end, we leverage the interventional effects of X_i, X_j on each other to define a measure of confounding.

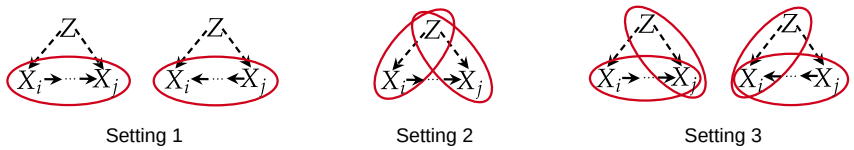

Figure 1: **Setting 1:** When contexts $\mathbf{C}_{\{i\}\wedge\neg P_{ij}}$ and $\mathbf{C}_{\{j\}\wedge\neg P_{ji}}$ are known where P_{ij} is the set of node indices that belong to a path from X_i to X_j including j, we leverage directed information from X_i to X_j and from X_j to X_i to define a measure of confounding (Defn. 4.4). **Setting 2:** Causal mechanism changes in Z introduces dependencies on the observed distributions of X_i, X_j. We leverage such dependencies to measure confounding when contexts $\mathbf{C}_{\{i\}\wedge\{j\}}$ are known (Defn. 4.6). **Setting 3:** If we know that there is a causal path from X_i to X_j, we leverage dependencies between the pairs (X_i, X_j) and (Z, X_j) to measure confounding. Similarly, if we know that there is a causal path from X_j to X_i, we leverage dependencies between the pairs (X_i, X_j) and (Z, X_i) to measure confounding (Defn. 4.7). Dashed arrows from Z indicate that Z is unobserved.

Definition 4.1. ***(Directed Information [48]).*** *The directed information $I(X_i \to X_j)$ from $X_i \in \mathbf{X}$ to $X_j \in \mathbf{X}$ is defined as the conditional Kullback-Leibler divergence between the distributions $\mathbb{P}(X_i|X_j), \mathbb{P}(X_i|do(X_j))$. That is:*

$$I(X_i \to X_j) := D_{KL}(\mathbb{P}(X_i|X_j)||\mathbb{P}(X_i|do(X_j))) := \mathbb{E}_{\mathbb{P}(X_i,X_j)} \log \frac{\mathbb{P}(X_i|X_j)}{\mathbb{P}(X_i|do(X_j))} \tag{1}$$

Definition 4.2. ***(No Confounding [44])*** *When measuring the causal effect of a (treatment) variable X_i on a (target) variable X_j, the ordered pair (X_i, X_j) is unconfounded if and only if the directed information from X_j to X_i: $I(X_j \to X_i)$ is zero. Equivalently, $\mathbb{P}(X_j|X_i) = \mathbb{P}(X_j|do(X_i))$.*

A similar definition of confounding that relates the conditional distribution $\mathbb{P}(X_i|X_j)$ and interventional distribution $\mathbb{P}(X_i|do(X_j))$ is defined as follows.

Definition 4.3. ***(Noncollapsibility) [15, 43, 54]*** *The statistical association between two variables X_i and X_j is said to be noncollapsible if the association strength differs in each level/strata of other variable X_k. That is, if X_k is a confounding variable between X_i, X_j, we have $\mathbb{P}(X_j|X_i) \neq \mathbb{P}(X_j|do(X_i)) = \mathbb{E}_{X_k}(\mathbb{P}(X_j|X_i, X_k))$.*

From Defns. 4.1 and 4.2, for a pair of variables (X_i, X_j), observing $I(X_j \to X_i) > 0$ and $I(X_i \to X_j) > 0$ implies that $\mathbb{P}(X_j|do(X_i)) \neq \mathbb{P}(X_j|X_i)$ and hence the presence of confounding (see Tab. 2). Using the above properties of directed information, we measure *confounding* as follows.

Definition 4.4. ***(Confounding Measure 1)*** *When causal mechanism shifts of two variables $X_i, X_j \in \mathbf{X}$ are observed, resulting in different contexts, under the Assumptions 3.2-3.4, the measure of confounding $CNF\text{-}1(X_i, X_j)$ between X_i and X_j is defined as follows.*

$$CNF\text{-}1(X_i, X_j) := 1 - e^{-\min(I(X_i \to X_j), I(X_j \to X_i))} \tag{2}$$

For all the confounding measures, we use exponential transformation to limit the range of the measure between 0 and 1. Note that in a DAG, one of $I(X_i \to X_j), I(X_j \to X_i)$ is zero under no confounding (see Tab. 2 for a simple example with two and three node graphs). Hence $CNF\text{-}1(X_i, X_j)$ outputs zero when there is no confounding between X_i, X_j. Similarly $CNF\text{-}1(X_i, X_j)$ outputs positive real value in the range $(0, 1]$ when there is confounding. We leverage data from multiple contexts to evaluate $\mathbb{P}(X_i|X_j)$ and $\mathbb{P}(X_i|do(X_j))$ as follows. In this setting, we assume each context is generated as a result of *hard* interventions on a subset of variables. Let P_{ij} be the set of node indices that belong to a path from X_i to X_j including j, we use the contexts $\mathbf{C}_{\{i\}\wedge\neg P_{ij}}$ to evaluate $\mathbb{P}(X_j|do(X_i))$ as $\mathbb{P}(X_j|do(X_i)) = \mathbb{E}_{c\in\mathbf{C}_{\{i\}\wedge\neg P_{ij}}}[\mathbb{P}^c(X_j|X_i)]$. Intuitively, to compute the interventional effects of X_i on X_j, we need to observe mechanism changes only for X_i to account for the potential causal influence from X_i to X_j. In addition, none of the nodes in a causal path from X_i to X_j should be intervened. We use observational data to evaluate $\mathbb{P}(X_j|X_i)$.

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

Table 2: Directed information values in two and three node graphs. If X_i, X_j are confounded by Z, we observe positive directed information from both directions.

Proposition 4.1. ***(Identifiability of*** $\mathbb{P}(X_j|do(X_i))$***)*** *$\mathbb{P}(X_j|do(X_i))$ is identifiable from the set of contexts $\mathbf{C}_{\{i\}\wedge\neg P_{ij}}$. To detect and measure confounding between a pair of nodes X_i, X_j, it is enough to observe two sets of contexts $\mathbf{C}_{\{i\}\wedge\neg P_{ij}}$ and $\mathbf{C}_{\{j\}\wedge\neg P_{ji}}$. Thus, n sets of contexts are needed to detect and measure confounding between $\binom{n}{2}$ distinct pairs of nodes in a causal DAG with n nodes.*

When a confounding variable X_o between X_i, X_j is observed, and there may exist an unobserved confounding variable Z, it is crucial to detect and measure unobserved confounding effect [29]. We utilize conditional directed information to define the measure of unobserved confounding.

Definition 4.5. ***(Conditional Directed Information [48]).*** *The conditional directed information $I(X_i \to X_j|X_o)$ from X_i to X_j conditioned on X_o is defined as the conditional Kullback-Leibler divergence between the distributions $\mathbb{P}(X_i|X_j, X_o), \mathbb{P}(X_i|do(X_j), X_o)$ as follows.*

$$I(X_i \to X_j|X_o) \coloneqq D_{KL}(\mathbb{P}(X_i|X_j, X_o)||\mathbb{P}(X_i|do(X_j), X_o)) \coloneqq \mathop{\mathbb{E}}_{\mathbb{P}(X_i,X_j,X_o)} \log \frac{\mathbb{P}(X_i|X_j, X_o)}{\mathbb{P}(X_i|do(X_j), X_o)} \quad (3)$$

This measure can trivially be extended to the case where there exist multiple observed and unobserved confounding variables. The expression $\mathbb{P}(X_i|do(X_j), X_o)$ means conditioning on X_o in the interventional distribution $\mathbb{P}(X_i|do(X_j))$. Now, the conditional confounding can be measured as:

$$CNF\text{-}1(X_i, X_j|X_o) := 1 - e^{-\min(I(X_i \to X_j|X_o), I(X_j \to X_i|X_o))} \quad (4)$$

Intuitively, by conditioning on an observed confounding variable X_o, we control the association between X_i, X_j flowing via X_o and measure the influence via the unobserved confounding variables.

Beyond Pairwise Confounding: We now study when a set $\mathbf{X}_S$ of variables where $|\mathbf{X}_S| > 2$ are jointly confounded i.e., share a common confounding variable and how to measure the joint confounding among the variables $\mathbf{X}_S$.

Theorem 4.1. *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 $I(X_i \to X_j|X_k) = I(\{X_i, X_k\} \to X_j)$.*

We now define the measure of confounding among the variables in $\mathbf{X}_S$ as follows.

$$CNF\text{-}1(\mathbf{X}_S) = \sum_{i \in S} CNF\text{-}1(\mathbf{X}_{S\setminus\{i\}}, X_i) \quad (5)$$

Conditional confounding among a set of variables can be defined similarly to Eqn. 4. We now study some useful properties of the measure CNF-1.

Theorem 4.2. *For any three observed variables X_i, X_j, X_o and an unobserved confounding variable Z, the following statements are true for the measure CNF-1.*

1. ***(Reflexivity and Symmetry.)*** *$CNF\text{-}1(X_i, X_i|X_o) = 0$, $CNF\text{-}1(X_i, X_j|X_o) = CNF\text{-}1(X_j, X_i|X_o)$.*
2. ***(Positivity.)*** *$CNF\text{-}1(X_i, X_j) > 0$ if and only if X_i, X_j are confounded. Given an observed confounding variable X_o between X_i, X_j, $CNF\text{-}1(X_i, X_j|X_o) > 0$ if and only if there exists an unobserved confounding variable Z between X_i, X_j.*
3. ***(Monotonicity.)*** *$CNF\text{-}1(X_i, X_j) > CNF\text{-}1(X_k, X_l)$ implies that the pair of variables X_i, X_j are more strongly confounded than the pair of variables X_k, X_l in the sense of Defns. 4.2 and 4.3.*

4.2 Setting 2: Detecting and Measuring Confounding Using the Mechanism Shifts of Z.

The previous setting utilizes the interventional effects of $X_i(X_j)$ on $X_j(X_i)$ to define a measure of confounding between X_i, X_j. In this setting, we utilize the association between the observed marginal distributions of X_i, X_j under causal mechanism shifts of Z to measure confounding. To this end, similar to [38], we make the following assumption.

Assumption 4.1. ***(Shift Faithfulness [38])*** *Let Z be a common parent for a set of variables $\boldsymbol{X}_S \subseteq \boldsymbol{X}$. Then each causal mechanism shift in Z between two contexts c, c' entails a causal mechanism change in each $X_i \in \boldsymbol{X}_S$ between the same contexts c, c'.*

One consequence of the Assumption 4.1 is that a change in the causal mechanism of Z induces correlations between the expectations of X_i, X_j in different contexts. To understand this, consider the following structural equations.

$$Z \sim \mathcal{N}(\mu(c), \sigma^2(c)) \qquad X_i \coloneqq \alpha Z + \epsilon_i \qquad X_j \coloneqq \beta X_i + \gamma Z + \epsilon_j \quad (6)$$

Where c denotes the context and ϵ_x and ϵ_y are noise variables with zero mean and have no additional restriction on the underlying probability distribution. The causal graph corresponding to this model has the nodes X_i, X_j, Z and edges: $Z \to X_i, Z \to X_j, X_i \to X_j$. It is easy to see that $\mathbb{E}(X_i) = \alpha\mu(c)$ and $\mathbb{E}(X_j) = (\alpha\beta + \gamma)\mu(c)$. Following Assumption 4.1, whenever there is a change in causal mechanism of Z (e.g., c changes to $\tilde{c}$ in Eqn. 6), there is a change in both $\mathbb{E}(X_i), \mathbb{E}(X_j)$. Additionally, since Z is a common cause of both X_i, X_j, there is a spurious association between $\mathbb{E}(X_i), \mathbb{E}(X_j)$. Subsequently, in the set of contexts $\mathbf{C}_{\{i\}\wedge\{j\}}$ the values $\mathbb{E}(X_i), \mathbb{E}(X_i)$ are spuriously associated. Under Assumptions 3.2 and 4.1, restricting our analysis to $\mathbf{C}_{\{i\}\wedge\{j\}}$ ensures that with high probability, the association between $\mathbb{E}(X_i)$, $\mathbb{E}(X_i)$ is due to the confounding variable Z. In this example, the association between $\mathbb{E}(X_i), \mathbb{E}(X_j)$ exists even if $\beta = 0$, i.e., $X_i \not\to X_j$. To define confounding measure, we create two random variables E_i^C, E_j^C which we define as $E_i^C = \mathbb{E}_{X_i \sim \mathbb{P}^c(X_i)}(X_i), E_j^C = \mathbb{E}_{X_j \sim \mathbb{P}^c(X_j)}(X_j)$ respectively where $c \in \mathbf{C}_{\{i\}\wedge\{j\}}$. Relying on the context information $\mathbf{C}_{\{i\}\wedge\{j\}}$ and utilizing the association between E_i^C and E_j^C, we define a confounding measure as follows.

Proposition 4.2. ***(Confounding Based on Mutual Information)*** *If two variables X_i, X_j are confounded by a variable Z, the induced random variables E_i^C, E_j^C as described above have non zero mutual information $I(E_i^C; E_j^C)$.*

Definition 4.6. ***(Confounding Measure 2)*** *When the causal mechanism shifts are observed for X_i, X_j in different contexts and the contexts $\mathbf{C}_{\{i\}\wedge\{j\}}$ are known, under the Assumptions 3.2-4.1, the measure of confounding $CNF\text{-}2(X_i, X_j)$ between X_i and X_j is defined as*

$$CNF\text{-}2(X_i, X_j) := 1 - e^{-I(E_i^C; E_j^C)} \tag{7}$$

To measure the unobserved confounding strength when we already observe a confounding variable X_o, we condition on the observed confounding variable X_o to define $CNF\text{-}2(X_i, X_j|X_o)$ as follows.

$$CNF\text{-}2(X_i, X_j|X_o) := 1 - e^{-I(E_i^C; E_j^C|X_o)} \tag{8}$$

Beyond Pairwise Confounding: Following earlier work [38], we utilize total correlation among triplets (E_i^C, E_j^C, E_k^C) of random variables in $\{E_i^C\}_{i\in S}$ to verify whether a set of variables $\mathbf{X}_S$ are jointly confounded. By Assumption 4.1, we know that the variables in $\mathbf{X}_S$ jointly confounded only if each pair $X_i, X_j;\ \ i, j \in S$ is pairwise confounded. If all three variables share the same latent confounding variable Z, then knowing about one of E_i^C, E_j^C, E_k^C explains away some of the association between the other two, so that we have $I(E_i^C, E_j^C|E_k^C) < I(E_i^C, E_j^C)$. However, for a triplet (X_i, X_j, X_k), it is possible that, rather than jointly confounded, there may be three disjoint confounding variables Z_{12}, Z_{13}, Z_{23} confounding each of the individual pairs: $(X_i, X_j), (X_j, X_k), (X_k, X_i)$. In general, for a set of variables of size s to permit such an equivalent explanation, we would need to have a total of $\binom{s}{2}$ confounding variables with $s(s-1)$ outgoing edges to obtain the same structure of pairwise confounding [38]. While this may plausibly occur for small sets of variables that appear to be pairwise correlated, we assume the true graph $\mathcal{G}$ to be causally minimal in the following sense.

Assumption 4.2. ***(Confounder Minimality [38])*** *For every subset $\mathbf{X}_S$ of at least $|S| \geq 4$ variables, there are at most $2|S|$ edges incoming into $\mathbf{X}_S$ from latent confounding variables with at least three children in $\mathbf{X}_S$.*

Assumption 4.2 ensures that variables that appear to be jointly confounded are indeed confounded. In other words, when a small number of latent variables suffice to explain the observed correlations, there should indeed exist only few confounding variables. With this assumption, we can guarantee that joint confounding can be identified from the total correlation.

Theorem 4.3. *Let $\mathbf{X}_S$ be a set of variables such that all $X_i, X_j \in \mathbf{X}_S$ are pairwise confounded. Then $\mathbf{X}_S$ is jointly confounded if and only if for each triple $X_i, X_j, X_k \in \mathbf{X}_S$ we have $I(E_i^C; E_j^C|E_k^C) < I(E_i^C; E_j^C)$.*

Now, the measure of joint confounding among a set of variables $\mathbf{X}_S$ can be defined using total correlation $T(E_i^C, \ldots, E_{|S|}^C)$ as follows. To evaluate the following expression, we need to use the contexts $\mathbf{C}_{\{1\}\cup\cdots\cup\{|S|\}}$ to ensure that with high probability, the association among the variables in $\mathbf{X}_S$ is due to the joint confounding variable Z.

$$CNF\text{-}2(\mathbf{X}_S) = 1 - e^{-T(E_i^C, \ldots, E_{|S|}^C)} \tag{9}$$

Theorem 4.4. *For any three observed variables X_i, X_j, X_o and an unobserved confounding variable Z, the following statements are true for the measure CNF-2.*

1. ***(Reflexivity and Symmetry.)*** *$CNF\text{-}2(X_i, X_i|X_o) = 1 - e^{-H(E_i^C|X_o)}$ $\forall i$ where $H(.|.)$ denotes conditional entropy and $CNF\text{-}2(X_i, X_j|X_o) = CNF\text{-}2(X_j, X_i|X_o)$.*
2. ***(Positivity.)*** *$CNF\text{-}2(X_i, X_j) > 0$ if and only if X_i, X_j are confounded. Given an observed confounding variable X_o between X_i, X_j, $CNF\text{-}2(X_i, X_j|X_o) > 0$ if and only if there exists an unobserved confounding variable Z between X_i, X_j.*
3. ***(Monotonicity.)*** *$CNF\text{-}2(X_i, X_j) > CNF\text{-}2(X_k, X_l)$ implies that the pair of variables X_i, X_j are more strongly confounded than the pair of variables X_k, X_l in the sense of Defn. 4.2.*

4.3 Setting 3: Observing the Causal Mechanism Shifts in Z and Known Causal Path Direction Between X_i and X_j

Similar to the previous settings, we utilize marginal and conditional distributions of X_i, X_j to define a measure of confounding. By prior knowledge, if we know the direction of causal path between X_i, X_j, we can utilize the causal direction to measure confounding as explained below. In addition to the notations E_i^C, E_j^C introduced in the previous setting, let us denote for each $c \in \mathbf{C}_{\{i\}\wedge\{j\}}, \mathbb{E}_{X_i\sim\mathbb{P}^c(X_i|X_j)}(X_i|X_j), \mathbb{E}_{X_j\sim\mathbb{P}^c(X_j|X_i)}(X_j|X_i)$ with E_{ij}^C, E_{ji}^C respectively. We now leverage dependency among these variables to define the measure of confounding. Intuitively, if $X_i \to X_j$ and if we observe a change in the causal mechanisms of both X_i, X_j due to the causal mechanism changes in Z, we also observe a change in the causal mechanism $\mathbb{P}(X_j|X_i)$.

Definition 4.7. ***(Confounding Measure 3)*** *When the causal mechanism shifts are observed for X_i, X_j and the causal direction between the nodes X_i, X_j is known, under the Assumptions 3.2-4.1, the measure of confounding $CNF\text{-}3(X_i, X_j)$ between $X_i \in \mathbf{X}$ and $X_j \in \mathbf{X}$ is defined as*

$$CNF\text{-}3(X_i, X_j) := \begin{cases} 1 - e^{-I(E_{ji}^C; E_j^C)} & if \quad X_i \to \cdots \to X_j \\ 1 - e^{-I(E_{ij}^C; E_i^C)} & if \quad X_j \to \cdots \to X_i \\ CNF\text{-}2(X_i, X_j) & Otherwise \end{cases} \tag{10}$$

To measure the unobserved confounding strength in the presence of an observed confounding variable X_o, similar to setting 2, we can modify Eqn. 10 to condition on the variable X_o.

Beyond Pairwise Confounding: Using the Assumption 4.2, we have the following.

Theorem 4.5. *Let $\mathbf{X}_S$ be a set of variables such that all $X_i, X_j \in \mathbf{X}_S$ are pairwise confounded and the causal relationships among each pair X_i, X_j. Then $\mathbf{X}_S$ is jointly confounded if and only if for each triple $X_i, X_j, X_k \in \mathbf{X}_S$ we have $I(E_{ij}^C; E_{jk}^C|E_j^C) < I(E_{ij}^C; E_{jk}^C)$.*

Since we have access to random variables E_{ij}^C in addition to E_i^C, E_j^C, it is not straightforward to use all of them to measure joint confounding. To keep the measure simple, we let the measure of joint confounding among the variables $\mathbf{X}_S$ be the same as $CNF\text{-}2(\mathbf{X}_S)$. That is, $CNF\text{-}3(\mathbf{X}_S) = CNF\text{-}2(\mathbf{X}_S)$. Setting 3 is an alternative to Setting 2 when we know the direction of the causal path between X_i, X_j. Settings 2 and 3 act as complementary to each other in validating the correctness of our analysis.

Theorem 4.6. *For any three observed variables X_i, X_j, X_o and an unobserved confounding variable Z, the following statements are true for the measure CNF-3.*

1. ***(Reflexivity and Symmetry.)*** *$CNF\text{-}3(X_i, X_i|X_o) = 1 - e^{-H(E_i^C|X_o)}$ $\forall i$ where $H(.|.)$ denotes conditional entropy and $CNF\text{-}3(X_i, X_j|X_o) = CNF\text{-}3(X_j, X_i|X_o)$.*
2. ***(Positivity.)*** *$CNF\text{-}3(X_i, X_j) > 0$ if and only if X_i, X_j are confounded. Given an observed confounding variable X_o between X_i, X_j, $CNF\text{-}3(X_i, X_j|X_o) > 0$ if and only if there exists an unobserved confounding variable Z between X_i, X_j.*
3. ***(Monotonicity.)*** *$CNF\text{-}3(X_i, X_j) > CNF\text{-}3(X_k, X_l)$ implies that the pair of variables X_i, X_j are more strongly confounded than the pair of variables X_k, X_l in the sense of Defn. 4.2.*

5 Algorithm

Algorithm 1 outlines the procedures to measure confounding in all three settings and can be extended to the case where we evaluate conditional confounding and evaluating confounding among multiple variables. We present two real-world examples where our methods can be applied in Appendix § B.

Algorithm 1: Algorithm for evaluating pairwise CNF-1, CNF-2, CNF-3

Data: Context information $\mathbf{C}_{\{i\}\wedge\neg P_{ij}}, \mathbf{C}_{\{j\}\wedge\neg P_{ji}}, \mathbf{C}_{\{i\}\wedge\{j\}}$, Contextual Datasets $\{\mathcal{D}^c\}_{c\in\mathbf{C}}$.
Result: $CNF\text{-}1(X_i, X_j), CNF\text{-}2(X_i, X_j), CNF\text{-}3(X_i, X_j)$
Step 1: Evaluate $\mathbb{P}(X_i|X_j), \mathbb{P}(X_j|X_i)$ using observational data;
Step 2: Evaluate $\mathbb{P}(X_i|do(X_j))$ using $\{\mathcal{D}^c\}_{c\in\mathbf{C}_{\{j\}\wedge\neg P_{ji}}}$;
Step 3: Evaluate $\mathbb{P}(X_j|do(X_i))$ using $\{\mathcal{D}^c\}_{c\in\mathbf{C}_{\{i\}\wedge\neg P_{ij}}}$;
Step 4: Evaluate $I(X_i \to X_j), I(X_j \to X_i)$;
Step 5: $CNF\text{-}1(X_i, X_j) = 1 - e^{-\min(I(X_i\to X_j), I(X_j\to X_i))}$;
Step 6: Evaluate E_i^C, E_j^C using $\{\mathcal{D}^c\}_{c\in\mathbf{C}_{\{i\}\wedge\{j\}}}$;
Step 7: $CNF\text{-}2(X_i, X_j) = 1 - e^{-I(E_i^C; E_j^C)}$;
Step 8: Evaluate E_{ij}^C, E_{ji}^C using $\{\mathcal{D}^c\}_{c\in\mathbf{C}_{\{i\}\wedge\{j\}}}$;
Step 9: compute $CNF\text{-}3(X_i, X_j)$ according to Defn. 4.7;
return $CNF\text{-}1(X_i, X_j), CNF\text{-}2(X_i, X_j), CNF\text{-}3(X_i, X_j)$

6 Experiments and Results

We perform simulation studies to verify the correctness of the proposed measures. All the experiments are run on a CPU. We report the mean and standard deviation of results taken over five random seeds. Code to reproduce the results is presented in the supplementary material. Code is available at https://github.com/gautam0707/CD_CNF.

Measuring Confounding: In this set of experiments, we consider the following four causal structures made of three nodes X_i, X_j, Z: $\mathcal{G}_1$: Empty graph over Z, X_i, X_j i.e., nodes are isolated in the graph, $\mathcal{G}_2 : X_i \to X_j$, $\mathcal{G}_3 : Z \to X_i, Z \to X_j$, $\mathcal{G}_4 : Z \to X_i, Z \to X_j, X_i \to X_j$. In $\mathcal{G}_1, \mathcal{G}_2$, there is no confounding between X_i, X_j and in $\mathcal{G}_3, \mathcal{G}_4$ there is confounding effect of Z on X_i and X_j. Results in Fig. 2 show that our measures output zero when there is no confounding between X_i, X_j and output positive values when X_i, X_j are confounded by a confounding variable Z.

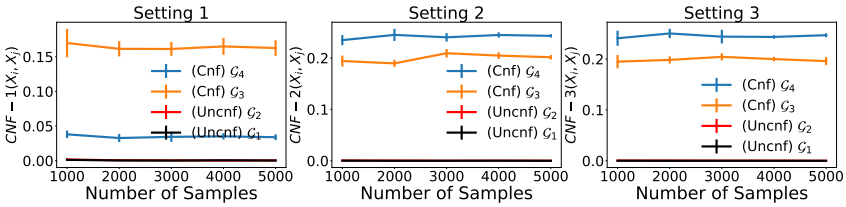

Figure 2: Measure of confounding between a pair of variables X_i, X_j. Our measures output zero when there is no confounding between X_i, X_j and output positive values when X_i, X_j are confounded.

Measuring Conditional Confounding: We consider the following two causal structures. $\mathcal{G}_5 : Z_1 \to X_i, Z_1 \to X_j, Z_2 \to X_i, Z_2 \to X_j, X_i \to X_j$. $\mathcal{G}_6 : Z \to X_i, Z \to X_j, X_i \to X_j$. In $\mathcal{G}_5$, X_i and X_j are confounded by two variables Z_1, Z_2. We measure conditional confounding between X_i, X_j conditioned on $\emptyset$, Z_1, and Z_2 respectively. Since confounding still exists in all of the above conditioning settings, CNF-2 correctly returns positive confounding value in all three cases (see Fig. 3 left). On the other hand, in $\mathcal{G}_6$, we measure conditional confounding

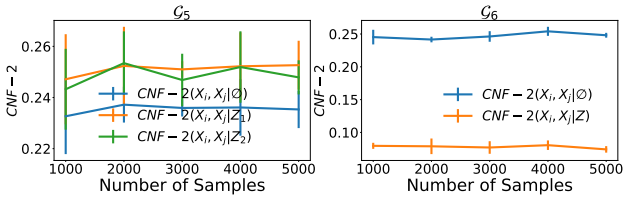

Figure 3: **Left:** Conditioning on one of $\emptyset, Z_1, Z_2$ will not remove confounding between X_i, X_j in $\mathcal{G}_5$. Hence CNF-2 returns positive values. **Right:** In $\mathcal{G}_6$, conditioning on $\emptyset$ does not remove the confounding effect of Z on X_i, X_j. Hence, we observe a positive value for $CNF\text{-}2(X_i, X_j|\emptyset)$. Conditioning on Z will block the confounding between X_i, X_j. Hence CNF-2 is closer to zero.

between X_i, X_j conditioning on empty set and Z. Since conditioning on Z will block the confounding association between X_i, X_j, CNF-2 returns confounding value closer to zero. However, the unconditioned confounding (conditioning on empty set) value is still large. These results empirically validate the correctness of the proposed measures.

Downstream Causal Effect Estimation: For the causal graphs $\mathcal{G}_3, \mathcal{G}_4$, we examine the impact of controlling for nodes identified using our method. We measure the causal effect of X_i on X_j with and without controlling for the detected confounding variable and report the absolute difference between the true and estimated causal effects in Tab. 3. The results show that controlling for the variables identified by our method reduces the bias in the estimated causal effects.

Causal Graph	Not Controlling Confounding					Controlling Confounding				
	1000	2000	3000	4000	5000	1000	2000	3000	4000	5000
$\mathcal{G}_3$	0.55	0.57	0.55	0.52	0.52	0.06	0.02	0.007	0.03	0.009
$\mathcal{G}_4$	0.24	0.26	0.23	0.24	0.23	0.04	0.05	0.06	0.02	0.05

Table 3: Downstream application of causal effect estimation.

Binary Data - Erdös-Rényi Causal Graphs: To verify the performance of our method on a large scale, similar to [38], we generate causal graphs of various number nodes using Erdös-Rényi model. In these experiments, each context is a result of intervention on one node. This is the reason for having the same value for number of nodes N and number of contexts $|C|$. Sample size denotes the number of data points used in each context. We detect and measure whether each pair of nodes is confounded or not. We then calculate the *Precision, Recall*, and *F1* scores. Our confounding measures obtain good results across all settings.

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

7 Conclusions, Limitations, and Future Work

In this paper, based on the known causal mechanism shifts of observed variables, we propose three measures of confounding along with their conditional and multivariate variants. We also study key properties of these measures. Our measures complement each other depending on the available context information. We propose algorithms to compute the proposed measures and empirically verify their correctness. However, for the same confounded pair of variables, our metrics may yield different results depending on the chosen measure. As discussed in the introduction, the measures are intended to assess the relative strengths of confounding rather than for point-to-point comparison. The number of contexts required to evaluate the measure can be large because many contexts without changes in particular mechanisms are discarded. Identifying appropriate real-world datasets and applying the proposed measures to those datasets is an interesting area for future work, as is developing measures that efficiently use context information. Additionally, devising new definitions for confounding and proposing corresponding confounding measures is also an interesting future direction. We aim to pursue these ideas.

Acknowledgments

This work was partly supported by the Prime Minister's Research Fellowship (PMRF) program and a Google Research Scholar Award. We are grateful to the anonymous reviewers for their valuable feedback, which improved the presentation of the paper.

References

[1] John Aldrich. Correlations genuine and spurious in pearson and yule. *Statistical science*, pages 364–376, 1995.

[2] Rohit Bhattacharya, Tushar Nagarajan, Daniel Malinsky, and Ilya Shpitser. Differentiable causal discovery under unmeasured confounding. In *International Conference on Artificial Intelligence and Statistics*, pages 2314–2322, 2021.

[3] Norman E Breslow, Nicholas E Day, and Elisabeth Heseltine. Statistical methods in cancer research. 1980.

[4] Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. In *Advances in Neural Information Processing Systems*, volume 33, pages 21865–21877, 2020.

[5] Esben Budtz–Jørgensen, Niels Keiding, Philippe Grandjean, and Pal Weihe. Confounder selection in environmental epidemiology: Assessment of health effects of prenatal mercury exposure. *Annals of Epidemiology*, 17(1):27–35, 2007.

[6] Timothy A Carey and William B Stiles. Some problems with randomized controlled trials and some viable alternatives. *Clinical Psychology & Psychotherapy*, 23(1):87–95, 2016.

[7] Venkat Chandrasekaran, Pablo A Parrilo, and Alan S Willsky. Latent variable graphical model selection via convex optimization. In *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1610–1613. IEEE, 2010.

[8] David Maxwell Chickering. Learning equivalence classes of bayesian-network structures. *The Journal of Machine Learning Research*, 2:445–498, 2002.

[9] Diego Colombo, Marloes H Maathuis, Markus Kalisch, and Thomas S Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, pages 294–321, 2012.

[10] Shai Ben David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility theorems for domain adaptation. In *AISTATS*, pages 129–136, 2010.

[11] Mirthe Maria Van Diepen, Ioan Gabriel Bucur, Tom Heskes, and Tom Claassen. Beyond the markov equivalence class: Extending causal discovery under latent confounding. In *2nd Conference on Causal Learning and Reasoning*, 2023.

[12] Frederick Eberhardt and Richard Scheines. Interventions and causal inference. *Philosophy of science*, 74(5):981–995, 2007.

[13] Frederick Eberhardt, Clark Glymour, and Richard Scheines. On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables. *arXiv preprint arXiv:1207.1389*, 2012.

[14] Robin J Evans and Thomas S Richardson. Smooth, identifiable supermodels of discrete dag models with latent variables. *Bernoulli*, 25:848–876, 2019.

[15] Sander Greenland and Hal Morgenstern. Confounding in health research. *Annual review of public health*, 22(1):189–212, 2001.

[16] R.H.H. Groenwold, E. Hak, and A.W. Hoes. Quantitative assessment of unobserved confounding is mandatory in nonrandomized intervention studies. *Journal of Clinical Epidemiology*, 62(1): 22–28, 2009.

[17] Siyuan Guo, Viktor Tóth, Bernhard Schölkopf, and Ferenc Huszár. Causal de finetti: On the identification of invariant causal structure in exchangeable data. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[18] Gemma Hammerton and Marcus R Munafò. Causal inference with observational data: the need for triangulation of evidence. *Psychological medicine*, 51(4):563–578, 2021.

[19] Alain Hauser and Peter Bühlmann. Two optimal strategies for active learning of causal models from interventional data. *International Journal of Approximate Reasoning*, 55(4):926–939, 2014.

[20] Patrik O Hoyer, Shohei Shimizu, Antti J Kerminen, and Markus Palviainen. Estimation of causal effects using linear non-gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49(2):362–378, 2008.

[21] Biwei Huang, Kun Zhang, Jiji Zhang, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Behind distribution shift: Mining driving forces of changes and causal arrows. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 913–918. IEEE, 2017.

[22] Biwei Huang, Kun Zhang, Jiji Zhang, Joseph Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, 21(89):1–53, 2020.

[23] Amin Jaber, Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. Causal discovery from soft interventions with unknown targets: Characterization and learning. *Advances in neural information processing systems*, 33:9551–9561, 2020.

[24] Holly Janes, Francesca Dominici, and Scott Zeger. On quantifying the magnitude of confounding. *Biostatistics*, 11(3):572–582, 2010.

[25] Dominik Janzing and Bernhard Schölkopf. Detecting confounding in multivariate linear models via spectral analysis. *Journal of Causal Inference*, 6(1):20170013, 2018.

[26] Andrew Jesson, Alyson Rose Douglas, Peter Manshausen, Maëlys Solal, Nicolai Meinshausen, Philip Stier, Yarin Gal, and Uri Shalit. Scalable sensitivity and uncertainty analyses for causal-effect estimates of continuous-valued interventions. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[27] David Kaltenpoth and Jilles Vreeken. Nonlinear causal discovery with latent confounders. In *International Conference on Machine Learning*, pages 15639–15654, 2023.

[28] David Kaltenpoth and Jilles Vreeken. Causal discovery with hidden confounders using the algorithmic Markov condition. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, pages 1016–1026, 2023.

[29] Rickard Karlsson and Jesse Krijthe. Detecting hidden confounding in observational data using multiple environments. *Advances in Neural Information Processing Systems*, 36, 2023.

[30] David G Kleinbaum, Kevin M Sullivan, and Nancy D Barker. *A pocket guide to epidemiology*. Springer, 2007.

[31] Charles Ksir and Carl L Hart. Correlation still does not imply causation. *The Lancet Psychiatry*, 3(5):401, 2016.

[32] Paul H Lee. Is a cutoff of 10% appropriate for the change-in-estimate criterion of confounder identification? *Journal of epidemiology*, 24(2):161–167, 2014.

[33] Adam Li, Amin Jaber, and Elias Bareinboim. Causal discovery from observational and interventional data across multiple environments. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[34] George Maldonado and Sander Greenland. Simulation study of confounder-selection strategies. *American journal of epidemiology*, 138(11):923–936, 1993.

[35] George Maldonado and Sander Greenland. Estimating causal effects. *International journal of epidemiology*, 31(2):422–429, 2002.

[36] Sarah Mameche, David Kaltenpoth, and Jilles Vreeken. Discovering invariant and changing mechanisms from data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1242–1252, 2022.

[37] Sarah Mameche, David Kaltenpoth, and Jilles Vreeken. Learning causal models under independent changes. *Advances in Neural Information Processing Systems*, 36, 2023.

[38] Sarah Mameche, Jilles Vreeken, and David Kaltenpoth. Identifying confounding from causal mechanism shifts. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS)*. PMLR, 2024.

[39] Olli S Miettinen and E Francis Cook. Confounding: essence and detection. *American journal of epidemiology*, 114(4):593–603, 1981.

[40] Joris M Mooij, Sara Magliacane, and Tom Claassen. Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, 21:1–108, 2020.

[41] Austin Nichols. Causal inference with observational data. *The Stata Journal*, 7(4):507–541, 2007.

[42] Juan Miguel Ogarrio, Peter Spirtes, and Joe Ramsey. A hybrid causal search algorithm for latent variable models. In *Proceedings of the Eighth International Conference on Probabilistic Graphical Models*, pages 368–379, 2016.

[43] Menglan Pang, Jay S Kaufman, and Robert W Platt. Studying noncollapsibility of the odds ratio with marginal structural and logistic regression models. *Statistical methods in medical research*, 25(5):1925–1937, 2016.

[44] Judea Pearl. *Causality*. Cambridge university press, 2009.

[45] Ronan Perry, Julius Von Kügelgen, and Bernhard Schölkopf. Causal discovery in heterogeneous environments under the sparse mechanism shift hypothesis. In *Advances in Neural Information Processing Systems*, pages 10904–10917, 2022.

[46] Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 15(1): 2009–2053, 2014.

[47] Jonas Peters, Dominik Janzing, and Bernhard Schlkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017.

[48] Maxim Raginsky. Directed information and pearl's causal calculus. In *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 958–965, 2011.

[49] Thomas S Richardson, Robin J Evans, James M Robins, and Ilya Shpitser. Nested markov properties for acyclic directed mixed graphs. *The Annals of Statistics*, 51(1):334–361, 2023.

[50] Robert William Sanson-Fisher, Billie Bonevski, Lawrence W. Green, and Cate D'Este. Limitations of the randomized controlled trial in evaluating population-based health interventions. *American Journal of Preventive Medicine*, 33(2):155–161, 2007.

[51] Mauro Scanagatta, Cassio P de Campos, Giorgio Corani, and Marco Zaffalon. Learning bayesian networks with thousands of variables. *Advances in neural information processing systems*, 28, 2015.

[52] B Schölkopf, D Janzing, J Peters, E Sgouritsa, K Zhang, and J Mooij. On causal and anticausal learning. In *29th International Conference on Machine Learning (ICML 2012)*, pages 1255–1262. International Machine Learning Society, 2012.

[53] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[54] Noah A Schuster, Jos WR Twisk, Gerben Ter Riet, Martijn W Heymans, and Judith JM Rijnhart. Noncollapsibility and its role in quantifying confounding bias in logistic regression. *BMC medical research methodology*, 21:1–9, 2021.

[55] Karthikeyan Shanmugam, Murat Kocaoglu, Alexandros G Dimakis, and Sriram Vishwanath. Learning causal graphs with small interventions. *Advances in Neural Information Processing Systems*, 28, 2015.

[56] Ilya Shpitser, Robin J Evans, Thomas S Richardson, and James M Robins. Introduction to nested markov models. *Behaviormetrika*, 41:3–39, 2014.

[57] Ilya Shpitser, Robin J Evans, and Thomas S Richardson. Acyclic linear sems obey the nested markov property. In *Uncertainty in artificial intelligence: proceedings of the... conference. Conference on Uncertainty in Artificial Intelligence*, volume 2018. NIH Public Access, 2018.

[58] Edward H Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241, 1951.

[59] Peter Spirtes and Kun Zhang. Causal discovery and inference: concepts and recent methodological advances. In *Applied informatics*, volume 3, pages 1–28. Springer, 2016.

[60] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. 2000.

[61] Zhiqiang Tan. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637, 2006.

[62] Tyler J VanderWeele and Ilya Shpitser. On the definition of a confounder. *Annals of statistics*, 41(1):196, 2013.

[63] Y. Samuel Wang and Mathias Drton. Causal discovery with unobserved confounding and non-gaussian data. *Journal of Machine Learning Research*, 24(271):1–61, 2023.

[64] Aleksander Wieczorek and Volker Roth. Information theoretic causal effect quantification. *Entropy*, 21(10), 2019.

[65] Alessio Zanga, Elif Ozkirimli, and Fabio Stella. A survey on causal discovery: Theory and practice. *International Journal of Approximate Reasoning*, 151:101–129, 2022.

[66] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

Appendix

A Proofs

Proposition 4.1. ***(Identifiability of*** $\mathbb{P}(X_j|do(X_i))$***)*** *$\mathbb{P}(X_j|do(X_i))$ is identifiable from the set of contexts $\mathbf{C}_{\{i\}\wedge\neg P_{ij}}$. To detect and measure confounding between a pair of nodes X_i, X_j, it is enough to observe two sets of contexts $\mathbf{C}_{\{i\}\wedge\neg P_{ij}}$ and $\mathbf{C}_{\{j\}\wedge\neg P_{ji}}$. Thus, n sets of contexts are needed to detect and measure confounding between $\binom{n}{2}$ distinct pairs of nodes in a causal DAG with n nodes.*

Proof. Since the set of contexts $\mathbf{C}_{\{i\}\wedge P_{ij}}$ consist of data with all possible interventions on X_i, if a context c is generated by performing intervention on X_i with the value x_i, the expression $\mathbb{P}(X_j|do(X_i = x_i))$ is equal to the expression $\mathbb{P}(X_j|X_i = x_i)$ in that context c.

From Defn. 4.4, to detect and measure confounding between the pair of variables X_i, X_j, we need to evaluate $\mathbb{P}(X_j|do(X_i))$ and $\mathbb{P}(X_i|do(X_j))$. To this end, from the previous paragraph, we need two sets of contexts $\mathbf{C}_{\{i\}\wedge P_{ij}}$ and $\mathbf{C}_{\{j\}\wedge P_{ji}}$. Following these observations, it is enough to have n sets of contexts to detect and measure confounding between $\binom{n}{2}$ distinct pairs of nodes. □

Theorem 4.1. *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 $I(X_i \to X_j|X_k) = I(\{X_i, X_k\} \to X_j)$.*

Proof. Consider three variables X_i, X_j, X_k in the underlying causal graph. Consider the conditional directed information between X_i, X_j given X_k and the subsequent manipulations as follows.

$$
\begin{aligned}
I(X_i \to X_j|X_k) &:= \mathbb{E}_{\mathbb{P}(X_i,X_j,X_k)} \log \frac{\mathbb{P}(X_i|X_j, X_k)}{\mathbb{P}(X_i|do(X_j), X_k)} \\
&= \mathbb{E}_{\mathbb{P}(X_i,X_j,X_k)} \log \left(\frac{\mathbb{P}(X_i, X_k|X_j)}{\mathbb{P}(X_i, X_k|do(X_j))} \times \frac{\mathbb{P}(X_k|do(X_j))}{\mathbb{P}(X_k|X_j)} \right) \\
&= \mathbb{E}_{\mathbb{P}(X_i,X_j,X_k)} \log \frac{\mathbb{P}(X_i, X_k|X_j)}{\mathbb{P}(X_i, X_k|do(X_j))} - \mathbb{E}_{\mathbb{P}(X_j,X_k)} \log \frac{\mathbb{P}(X_k|X_j)}{\mathbb{P}(X_k|do(X_j))} \\
&= I(\{X_iX_k\} \to X_j) - I(X_k \to X_j)
\end{aligned}
$$

Since $I(X_k \to X_j) \geq 0$, we have $I(X_i \to X_j|X_k) \leq I(\{X_iX_k\} \to X_j)$. Equality holds only when X_k, X_j are unconfounded. □

Theorem 4.2. *For any three observed variables X_i, X_j, X_o and an unobserved confounding variable Z, the following statements are true for the measure CNF-1.*

1. ***(Reflexivity and Symmetry.)*** *CNF-1$(X_i, X_i|X_o) = 0$, CNF-1$(X_i, X_j|X_o) = CNF$-1$(X_j, X_i|X_o)$.*
2. ***(Positivity.)*** *CNF-1$(X_i, X_j) > 0$ if and only if X_i, X_j are confounded. Given an observed confounding variable X_o between X_i, X_j, CNF-1$(X_i, X_j|X_o) > 0$ if and only if there exists an unobserved confounding variable Z between X_i, X_j.*
3. ***(Monotonicity.)*** *CNF-1$(X_i, X_j) > CNF$-1(X_k, X_l) implies that the pair of variables X_i, X_j are more strongly confounded than the pair of variables X_k, X_l in the sense of Defns. 4.2 and 4.3.*

Proof. **Reflexivity:** From the definition of directed information, $I(X_i \to X_i|X_o) = \mathbb{E}_{\mathbb{P}(X_i,X_j,X_o)} log \frac{\mathbb{P}(X_i|X_o)}{\mathbb{P}(X_i|X_o)} = 0$ and hence CNF-1$(X_i, X_j|X_o) = 1 - e^0 = 0$.

Symmetry: Even if $I(X_i \to X_j|X_o)$ is not symmetric, the expression '$\min(I(X_i \to X_j|X_o), I(X_j \to X_i|X_o))$' is symmetric and hence CNF-1$(X_i, X_j|X_o)$ is symmetric.

Positivity: If X_i, X_j are confounded, irrespective of the direction of the causal path between X_i and X_j, we have $\mathbb{P}(X_i|X_j) \neq \mathbb{P}(X_i|do(X_j))$ and $\mathbb{P}(X_j|X_i) \neq \mathbb{P}(X_j|do(X_i))$. Hence $I(X_i \to X_j) > 0$ and $I(X_j \to X_i) > 0$. We now have CNF-1$(X_i, X_j) > 0$. The above statement is true even if there is no causal path between the nodes X_i, X_j. The above statements are valid even after conditioning on an observed confounding variable X_o if there is an unobserved confounding between X_i, X_j.

Monotonicity: Without loss of generality, assume that the inequality $CNF\text{-}1(X_i, X_j) > CNF\text{-}1(X_k, X_l)$ is a result of $I(X_i \to X_j) > I(X_k \to X_l)$. That is, the KL divergence between $\mathbb{P}(X_i|X_j)$ and $\mathbb{P}(X_i|do(X_j))$ is greater than the kl divergence between $\mathbb{P}(X_k|X_l)$ and $\mathbb{P}(X_k|do(X_l))$. That is, the pair of distributions $\mathbb{P}(X_k|X_l)$ and $\mathbb{P}(X_k|do(X_l))$ are closer to each other compared to the pair $\mathbb{P}(X_i|X_j)$ and $\mathbb{P}(X_i|do(X_j))$. As a result, X_k, X_l are closer to being *not confounded* in the sense of Defns. 4.2 and 4.3. □

Proposition 4.2. ***(Confounding Based on Mutual Information)*** *If two variables X_i, X_j are confounded by a variable Z, the induced random variables E_i^C, E_j^C as described above have non zero mutual information $I(E_i^C; E_j^C)$.*

Proof. There are two sources of dependency between E_i^C, E_j^C. If X_i, X_j are causally related in the underlying causal model generating the data, there will be a dependency between E_i^C, E_j^C in the context $C_{\{i\}\wedge\{j\}}$ as the interventions are soft. On the other hand, as per the Assumption 4.1, any shift in the causal mechanism of Z leads to a change in both the mechanisms of X_i, X_j leading to a dependency. Hence the random variables E_i^C, E_j^C have non-zero mutual information. □

Theorem 4.3. *Let $\mathbf{X}_S$ be a set of variables such that all $X_i, X_j \in \mathbf{X}_S$ are pairwise confounded. Then $\mathbf{X}_S$ is jointly confounded if and only if for each triple $X_i, X_j, X_k \in \mathbf{X}_S$ we have $I(E_i^C; E_j^C|E_k^C) < I(E_i^C; E_j^C)$.*

Proof. Following the Assumption 4.2, when three variables X_i, X_j, X_k are confounded by as single confounding variable Z, conditioning on one of E_i^C, E_j^C, E_k^C explains away some of the dependency between other two. Hence we have $I(E_i^C; E_j^C|E_k^C) < I(E_i^C; E_j^C)$ for all triples i, j, k. □

Theorem 4.4. *For any three observed variables X_i, X_j, X_o and an unobserved confounding variable Z, the following statements are true for the measure CNF-2.*

1. ***(Reflexivity and Symmetry.)*** *$CNF\text{-}2(X_i, X_i|X_o) = 1 - e^{-H(E_i^C|X_o)}$ $\forall i$ where $H(.|.)$ denotes conditional entropy and $CNF\text{-}2(X_i, X_j|X_o) = CNF\text{-}2(X_j, X_i|X_o)$.*
2. ***(Positivity.)*** *$CNF\text{-}2(X_i, X_j) > 0$ if and only if X_i, X_j are confounded. Given an observed confounding variable X_o between X_i, X_j, $CNF\text{-}2(X_i, X_j|X_o) > 0$ if and only if there exists an unobserved confounding variable Z between X_i, X_j.*
3. ***(Monotonicity.)*** *$CNF\text{-}2(X_i, X_j) > CNF\text{-}2(X_k, X_l)$ implies that the pair of variables X_i, X_j are more strongly confounded than the pair of variables X_k, X_l in the sense of Defn. 4.2.*

Proof. **Reflexivity:** from the definition of mutual information, $I(E_i^C; E_i^C|X_o) = H(E_i^C|X_o) - H(E_i^C|E_i^C, X_o) = H(E_i^C|X_o)$. Substituting in the definition of $CNF\text{-}2(X_i, X_j)$, result follows.

Symmetry: The result follows from the 'symmetry' property of mutual information.

Positivity: If X_i, X_j are confounded, from the Assumption 4.1, E_i^C, E_j^C are dependent random variables. Hence the mutual information is positive. The result follows after substituting some positive value for $I(E_i^C; E_j^C)$ in the definition of $CNF\text{-}2(X_i, X_j)$. The same argument goes for conditional confounding.

Monotonicity: from the definition of $CNF\text{-}2(X_i, X_j)$, $CNF\text{-}2(X_i, X_j) > CNF\text{-}2(X_k, X_l)$ implies $I(E_i^C; E_j^C) > I(E_k^C; E_l^C)$. From the Defn. 4.2, X_i, X_j have higher mutual information than the pair X_k, X_l and hence X_i, X_j are more strongly confounded than X_k, X_l. □

Theorem 4.5. *Let $\mathbf{X}_S$ be a set of variables such that all $X_i, X_j \in \mathbf{X}_S$ are pairwise confounded and the causal relationships among each pair X_i, X_j. Then $\mathbf{X}_S$ is jointly confounded if and only if for each triple $X_i, X_j, X_k \in \mathbf{X}_S$ we have $I(E_{ij}^C; E_{jk}^C|E_j^C) < I(E_{ij}^C; E_{jk}^C)$.*

Proof. Following the Assumption 4.2, when three variables X_i, X_j, X_k are confounded by as single confounding variable Z, conditioning on E_k^C explains away some of the dependency between E_{ij}^C, E_{jk}^C. Hence we have $I(E_{ij}^C; E_{jk}^C|E_j^C) < I(E_{ij}^C; E_{jk}^C)$ for all triples i, j, k. □

Theorem 4.6. *For any three observed variables X_i, X_j, X_o and an unobserved confounding variable Z, the following statements are true for the measure CNF-3.*

1. ***(Reflexivity and Symmetry.)*** *$CNF\text{-}3(X_i, X_i|X_o) = 1 - e^{-H(E_i^C|X_o)}$ $\forall i$ where $H(.|.)$ denotes conditional entropy and $CNF\text{-}3(X_i, X_j|X_o) = CNF\text{-}3(X_j, X_i|X_o)$.*
2. ***(Positivity.)*** *$CNF\text{-}3(X_i, X_j) > 0$ if and only if X_i, X_j are confounded. Given an observed confounding variable X_o between X_i, X_j, $CNF\text{-}3(X_i, X_j|X_o) > 0$ if and only if there exists an unobserved confounding variable Z between X_i, X_j.*
3. ***(Monotonicity.)*** *$CNF\text{-}3(X_i, X_j) > CNF\text{-}3(X_k, X_l)$ implies that the pair of variables X_i, X_j are more strongly confounded than the pair of variables X_k, X_l in the sense of Defn. 4.2.*

Proof. **Reflexivity:** from the definition of mutual information, $I(E_{ii}^C; E_i^C|X_o) = I(E_i^C; E_i^C|X_o) = H(E_i^C|X_o) - H(E_i^C|E_i^C, X_o) = H(E_i^C|X_o)$. Substituting in the definition of $CNF\text{-}3(X_i, X_j)$, result follows.

Symmetry: Since we rely on the direction of the causal path between X_i, X_j, for a given pair of nodes X_i, X_j, we have $CNF\text{-}3(X_i, X_j) = CNF\text{-}3(X_j, X_i)$ from Defn. 4.7.

Positivity: If X_i, X_j are confounded and $X_i \to X_j$, from the Assumption 4.1, E_{ji}^C, E_j^C are dependent random variables. Hence the mutual information E_{ji}^C, E_j^C is positive. The result follows after substituting positive value for $I(E_{ji}^C; E_j^C)$ in the definition of $CNF\text{-}3(X_i, X_j)$. The same argument goes for conditional confounding.

Monotonicity: from the definition of $CNF\text{-}3(X_i, X_j)$, without loss of generality, $CNF\text{-}3(X_i, X_j) > CNF\text{-}3(X_k, X_l)$ implies $I(E_{ji}^C; E_j^C) > I(E_{lk}^C; E_l^C)$. From the Defn. 4.2, X_i, X_j have higher mutual information and hence are more strongly confounded than X_k, X_l. □

B Real-world Examples

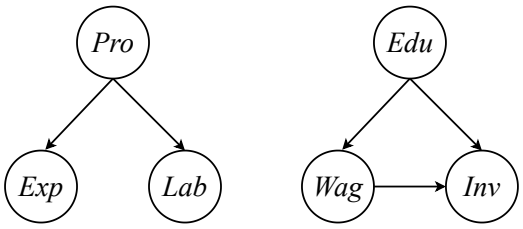

Figure 4: Two real-world examples where our method can be applied. Here *Pro*: Production Volume, *Exp*: Exports, *Lab*: Total Labor Required, *Edu*: Education, *Wag*: Wages, *Inv*: Investments. We can perform interventions on the above variables and any combination thereof to obtain context-specific data. We can use such data to identify and measure confounding by applying our methods.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All the claims made in the abstract and introduction are supported with theory and experiments in the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See § 7 for the discussion on the limitations of this work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: For each of the theoretical results, proofs and the assumptions used are presented in the Appendix § A.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Experimental setup is presented in main paper and code to reproduce the results is made public. See § 6.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code to reproduce the results is made public. See § 6.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Experimental setup is presented in main paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Each experiment is repeated for several random seeds.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Experimental setup is presented in main paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Our research confirms with NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: To the best of our knowledge, there are no detrimental impacts of our work.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our paper poses no risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All papers, data, and code used are cited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Code is documented and provided in supplementary material. Code will be made publicly available.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.