# CAUSAL DISCOVERY FROM MULTIPLE DOMAINS UNDER BIJECTIVE FIXED-CAUSE FUNCTIONALS

**Kasra Jalaldoust** **Saber Salehkaleybar** **Negar Kiyavash**

**ABSTRACT**

We consider the problem of causal discovery in a multi-environment setting where the functional relations remain invariant among the environments while the distribution of exogenous noises may vary. We show that the correct causal relationships are identifiable under less restrictive assumptions on the structural causal model compared to previous work. In particular, we assume that for each given value of the cause variable, the functional relation between the exogenous noise and the effect is bijective and it is differentiable in both directions. This assumption generalizes a variety of models including additive noise, LiNGAM, and post-non-linear model. We present our identifiability result based on an equivalence relation between random variables in multi-environment setting, and propose a statistical method for causal discovery by leveraging an independence criterion from our identifiability result. Experiments on various synthetic and real-world datasets validate our theoretical findings.

## 1 Introduction

Recovering causal relations is central to many scientific fields. Causal relationships between random variables are commonly modeled as a directed acyclic graph (DAG), where there is a directed edge from variable $X$ to variable $Y$, if $X$ is a direct cause of $Y$. Performing controlled experiments can recover the causal relationships. However, it is not always possible to perform controlled experiments, for instance, due to technical or ethical issues. This limits us to observational data that is collected passively from the variables in the system. From the observational data, the true causal graph can be identified up to a Markov equivalence class, which is the class of DAGs representing the same set of conditional independence relations among the observed variables [1]. In order to identify the Markov equivalence class, one can use constraint-based algorithms such IC and IC* [1], PC, FCI [2] or score-based methods such as greedy equivalence search [3]. Moreover, within the framework of structural causal models (SCM), the true graph can be identified uniquely by considering some additional assumptions on the data generation mechanisms such as non-Gaussianity of noise [4], or non-linearity of functional relations [5]. These models mainly consider additive noise, and they basically exploit the independence relation between the residual and the cause.

In recent years, causal discovery using data collected from multiple environments has gained attention. This setting is commonly formulated as having joint i.i.d observations of variables in multiple datasets, such that the causal modules may change across datasets. [6] introduced the notion of "invariant prediction". In this model, it is assumed that for every environment $e \in \mathcal{E}$, the response variable $Y^e$ has a fixed set of predictors $X_{S^*}, S^* \subset \{1, 2, \dots, p\}$ as its "invariant" predictors, such that the residual distribution is invariant across the environments when $X_{S^*}$ is the set of explanatory variables. In particular, they assumed that causal mechanisms can be modeled as linear structural equation model (SEM) and showed that invariant predictors of a target variable are identifiable, as long as the distribution of exogenous noise corresponding to the target variable does not change across the environments. The proposed method has high computational complexity and the output may not contain all the parents of the target variable. Later, [7] extended this work to non-linear functional relations, still assuming additive noise. [8] used invariant causal prediction assumption to handle sequentially ordered form of data in which the environmental properties may change due to the index. In [9], Ghassami et al. considered invariance of functional relations between variables and their parents in the linear SEM while the distribution of exogenous noises may vary among the environments. They introduced a notion of completeness of causal inference for this setting and proposed a complete algorithm which is computationally intensive. They also presented a heuristic algorithm which improved the computational complexity. [10] (and previously, in [11]),

proposed a more generalized model for the multi-environment causal discovery problem. The changes in functional relations are modeled by adding deterministic functions from the environment index $C$ to the parameters of causal mechanisms. The variable $C$ is also modeled as a random variable of the system. It is assumed that the parameters of functional relation of variable $V_i$, denoted by $\theta_i(C)$, changes independently from the other variables' parameters. However, $\theta_i(C)$ and $\theta_j(C)$ for any $i \neq j$ are deterministic functions of the same random variable $C$ and it seems rare to be independent unless one of them is constant. In [12], Ghassami et al. proposed a method to identify the causal relation between two random variables in linear SEM where the causal coefficients and/or the distribution of exogenous noises may vary. They also extend this method for a network of variable in linear SEM.

**Contribution.** We present an identifiability result in the multi-environment setting where the functional relation between the exogenous noise and the effect variable is bijective for any fixed value of cause. In our model, we do not require exogenous noise to be additive. Therefore our method is capable of causal discovery even in cases with complex dependencies between the exogenous noise and the cause. These dependencies often exist in real-world applications and can cause the methods which are based on independence of the exogenous noise and the cause to fail. In particular, we introduce the notion of "similarity" between random variables in the multi-environment setting and present our identifiability result by investigating properties of "similar" random variables. Based on our identifiability result, we present a causal discovery method by exploiting an independence relation which enables statistical hypothesis testing for causal direction determination. Further, we extend our results to multivariate and mixed-type data in the multi-environment setting.

The structure of this paper is as follows: In Section 2, we present our identifiability result and propose a statistical hypothesis testing based on it. In Section 3, we extend our results to multivariate and mixed-type data settings. In Section 4, we describe our causal discovery algorithm which is based on statistical hypothesis testing. In Section 5, we provide experiment results and compare our method with previous work. Finally, we conclude the paper in Section 6.

## 2 Main Result

Consider a directed acyclic graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with vertex set $\mathcal{V}$ and edge set $\mathcal{E}$. We define a structural causal model with graph $\mathcal{G}$ where the set of vertices represents random variable $\mathcal{V} = \{X_1, \cdots, X_n\}$ and $(X_i, X_j) \in \mathcal{E}$ if $X_i$ is a direct cause of $X_j$. Each variable $X_i$ depends on its parent, $PA_i$, in graph $\mathcal{G}$ by equation $X_i = f_i(PA_i, E_i)$ where $E_i$ is the exogenous noise of variable $X_i$ and it is independent of other variables in the model. In the following, to get more insight about our method, we first present our identifiability result for the bivariate causal model and then extend it to the multi-variate case. In particular, we consider a structural causal model with the graph $X \to Y$ such that $Y = f(X, E)$ where $f : \mathcal{X} \times \mathcal{E} \to \mathcal{Y}$. The random variable $E$ is the exogenous noise associated with the effect $Y$, and it is independent from $X$. Furthermore, we do not consider any confounding variables in the model. The cause and the effect and the noise random variables are functions with the sample space $\Omega$ as their domain and take values in $\mathcal{X}$,$\mathcal{Y}$, and $\mathcal{E}$, respectively.

There are different models for multi-environment setting in the literature of causal inference. These models assume that either the functional relations or the distribution of exogenous noises in the system are invariant across different environments. Furthermore, as discussed in the related work, each approach comes with a set of additional assumptions about the structural causal model. To model multiple environment setting, we assume that there is a finite set of probability measure functions $\mathcal{M} = \{\mathbb{P}^1, \mathbb{P}^2, \ldots, \mathbb{P}^m\}$, where $\mathbb{P}^e$ represents the stochastic behavior of the system in the $e$-th environment. We consider a single $\sigma$-algebra, $\Sigma$, defined on $\Omega$ such that for each $\mathbb{P}^e \in \mathcal{M}$, the triple $(\Omega, \Sigma, \mathbb{P}^e)$ is a probability space. Random variables $X, Y, E$ can have different distribution functions under each probability measure, but the distribution of $Y$ depends on the distributions of $X$ and $E$ through the invariant functional relation $f$. For each random variable $V : \Omega \to \mathcal{V}$ and any point $v \in \mathcal{V}$, we denote the probability density under $\mathbb{P}^e \in \mathcal{M}$ as $p^e_V(v)$. Similarly, for a pair of random variables $V$ and $W$, we denote the conditional probability density $V$ given $W$ as $p^e_{V|W}(v|w)$.

For multivariable functions like $g : \mathcal{A} \times \mathcal{B} \to \mathcal{C}$, and the points $a \in \mathcal{A}, b \in \mathcal{B}$, we frequently write $g(., b)$ or $g(a, .)$ to denote $g_b : \mathcal{A} \to \mathcal{C}$ and $g_a : \mathcal{B} \to \mathcal{C}$, respectively which are defined as $g_b(a) = g_a(b) = g(a, b)$. We view conditional probability densities as multivariable functions.

### 2.1 Assumptions

In the remain of this section, we only consider the bivariate case with variables $X, Y$ and exogenous noise $E$ all of which could be real-valued scalars or vectors. In Section 3, we extend our results to multivariate case as well as discrete-valued noise terms.

We assume that the joint density function $p^e_{X,Y}$ exists under each $\mathbb{P}^e \in \mathcal{M}$. In addition, we assume the set of probability measures $\mathcal{M}$ are mutually absolutely continuous, i.e., each subset of the sample space either has zero probability under

Table 1: Noise Models

| **Model** | **Functional Form** | **Fixed-Cause Functional** $f(x,.)$ |
|---|---|---|
| Additive noise model [5] | $f(X,E) = g(X) + E$ | Affine: $g(x) + E$ |
| LiNGAM [4] | $f(X,E) = u.X + E$ | Affine: $u.x + E$ |
| Post-non-linear model [14] | $f(X,E) = h(g(X) + E)$ | monotone $\circ$ affine: $h(g(x) + E)$ |
| Heteroscedastic noise model | $f(X,E) = g(X) + h(X)E$ | Affine : $g(x) + h(x)E$ |

every $\mathbb{P}^e \in \mathcal{M}$, or it has a positive probability under every $\mathbb{P}^e \in \mathcal{M}$; This technical assumption simplifies the results and the notation. The function $f$ is assumed to be invariant across the environments. Consequently, the joint densities $\{p^e_{X,Y}\}_{e=1}^m$ have the same support.

The exogenous noises are assumed to be mutually independent, that is, $E \perp\!\!\!\perp X$ under each $\mathbb{P}^e \in \mathcal{M}$, or, $p^e_{E|X}(.|x) = p^e_E$ for all $e$. This assumption is implied by independent mechanisms principle, commonly applied in the literature [13]. Lastly, we assume causal sufficiency, i.e., no unobserved confounders exist.

**Definition 1.** A fixed-cause functional at $x \in \mathcal{X}$ is denoted by $f(x,.)$ which represents the functional relation between the exogenous noise $E$ and the effect variable $Y$, for a specific value of cause, $x$.

**Definition 2.** A bijective function $g : \mathcal{A} \to \mathcal{B}$ is a diffeomorphism if it is differentiable and also has a differentiable inverse.

**Assumption 1.** For every $x \in \mathcal{X}$, we assume that the fixed-cause functional $f(x,.)$ is a diffeomorphism.

This is not a restrictive assumption. Note that all the classical models in the Table 1 satisfy Assumption 1.

In practice, we do not have access to function $f$ and can not directly test whether Assumption 1 holds. Instead, we are given $n_e$ i.i.d. samples $\{(x^e_l, y^e_l)\}_{l=1}^{n_e}$ for each $\mathbb{P}^e \in \mathcal{M}$. We derive an independence relation which allows us to test for Assumption 1 given the samples.

### 2.2 Identifiability Result

**Definition 3.** Define $\tilde{Y}_x$ as a $\mathcal{Y}$-valued random variable with the same distribution as $Y$ conditioned on $X = x$, under each $\mathbb{P}^e \in \mathcal{M}$. More precisely, for each measurable subset $u \subset \mathcal{Y}$ and for each $1 \leq e \leq m$, we have

$$\mathbb{P}^e(\tilde{Y}_x \in u) = \mathbb{P}^e(Y \in u | X = x). \tag{1}$$

Since we assumed the joint probability density function $p_{X,Y}$ exists, this is equivalent to

$$p^e_{\tilde{Y}_x}(y) = p^e_{Y|X}(y|x), \tag{2}$$

for each $1 \leq e \leq m$ and each $y \in \mathcal{Y}$.

**Definition 4.** Consider two $\mathcal{V}$-valued random variables $V$ and $V'$. $V$ and $V'$ are "*identical*" (denoted by $V \overset{I}{=} V'$) if and only if for every $\mathbb{P}^e \in \mathcal{M}$ and for each $u \subset \mathcal{V}$,

$$\mathbb{P}^e(V \in u) = \mathbb{P}^e(V' \in u). \tag{3}$$

**Definition 5.** Let $A$ and $B$ be two $\mathcal{A}$-valued and $\mathcal{B}$-valued random variables, respectively. $A$ is "*similar*" to $B$ (denoted by $A \sim B$) if and only if there exists a diffeomorphism $g : \mathcal{A} \to \mathcal{B}$ such that $B \overset{I}{=} g(A)$.

For any random variable $V$ and diffeomorphism function $g$, density functions of $W = g(V)$ under $\mathbb{P}^e \in \mathcal{M}$ is $p^e_W(w) = \frac{p^e_V(g^{-1}(w))}{|det(J_g(g^{-1}(w)))|}$, where $J_g$ denotes the Jacobian matrix of the $g$ at each point in the support of $V$. As $g$ is a diffeomorphism, the matrix $J_g$ exists and is non-zero at every point of support of $\mathcal{V}$ [15]. Thus, based on Definitions 4 and 5, for any pair of similar random variable random variables $A \sim B$, we have

$$\frac{p^1_A(g^{-1}(b))}{p^1_B(b)} = \frac{p^2_A(g^{-1}(b))}{p^2_B(b)} = \cdots = \frac{p^m_A(g^{-1}(b))}{p^m_B(b)}. \tag{4}$$

It can be easily shown that any two continuous probability density functions can be transformed into each other by diffeomorphisms [16]. Therefore, similarity of two random variables is trivially true as long as $|\mathcal{M}| = 1$. However, it is a non-trivial relation when $|\mathcal{M}| > 1$, i.e., in multi-environment setting.

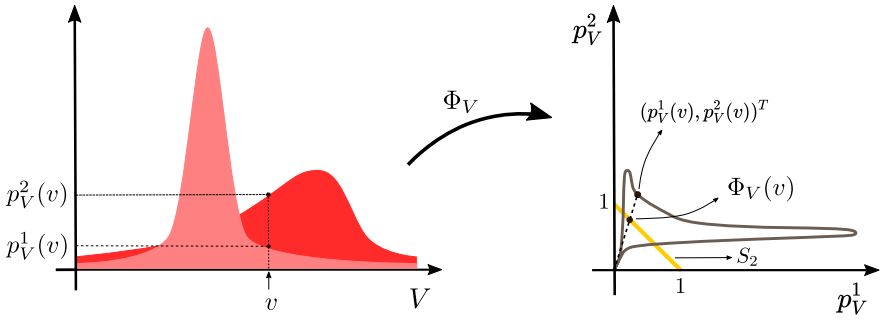


Figure 1: Left: Density functions $p_V^1, p_V^2$ for random variable $V$. Values of density functions at point $v$ are also shown in this plot. Right: Points on the gray curve are vectorization of values of density functions $p_V^1, p_V^2$ for each point $v$ in $\mathbb{R}$. Projecting the density vector $(p_V^1(v), p_V^2(v))^T$ on the simplex $S_2$ yields $\Phi_V(v)$ which is shown on the plot.

**Proposition 1.** Assumption 1 holds if and only if

$$\forall a, b \in \mathcal{X} : \tilde{Y}_a \sim \tilde{Y}_b. \tag{5}$$

We call this property "pairwise similarity of conditionals".

Assume that it is possible to check similarity of random variables. As Proposition 1 suggests, if we could find $a, b \in \mathcal{X}$ such that $\tilde{Y}_b \overset{I}{\neq} g(\tilde{Y}_a)$ for every diffeomorphism $g : \mathcal{Y} \to \mathcal{Y}$, we could reject $X \to Y$ under Assumption 1. Note that checking pairwise similarity of conditionals, even for a single pair $a, b \in \mathcal{X}$, seems to be impossible as there are countless candidates diffeomorphisms. In fact, Proposition 1 is just an identifiability result and it does not suggest how to check similarity of two random variables $\tilde{Y}_a$ and $\tilde{Y}_b$. In the next section, we introduce a practical method for statistically testing pairwise similarity of conditionals.

### 2.3 Statistical Hypothesis Testing

First, we introduce a theoretic indicator of similarity. Next, we propose an algorithm by which we can measure this indicator from the empirical data.

According to (4), for any two similar random variables $A$ and $B$ such that $B = g(A)$, the vectors $(p_A^1(g^{-1}(b)), \cdots, p_A^m(g^{-1}(b)))^T$ and $(p_B^1(b), \cdots, p_B^m(b))^T$ are aligned. Based on this observation, we define the following vector representation associated with random variables.

**Definition 6.** Consider $\mathcal{V}$-valued random variable $V$. The density-vectorization associated with $V$, $\Phi_V : \mathcal{V} \to S_m$, is defined as

$$\Phi_V(v) = \frac{h}{\|h\|_1}, \tag{6}$$

where $S_m$ is the simplex in $m$-dimensional space and

$$h = (p_V^1(v), p_V^2(v), \ldots, p_V^m(v))^T, \tag{7}$$

for every $v \in \mathcal{V}$.

**Definition 7.** Random variable $\Psi_V : \Omega \to S_m$ is called the "special random variable" associated with $V$ and it is defined as $\Psi_V := \Phi_V(V)$. Note that $\Phi_V$ is a deterministic function, but $\Psi_V$ is a random variable as it is a deterministic image of $V$.

**Proposition 2.** Consider $A$ and $B$, two $\mathcal{A}$-valued and $\mathcal{B}$-valued random variables, respectively. If $A \sim B$, then $\Psi_A \overset{I}{=} \Psi_B$.

The proposition suggests that if two random variables are similar, then the special random variables associated with them are identical. Note that this result does not introduce an equivalent condition for similarity of random variables. Instead, it can be used to reject the hypothesis of similarity whenever the latter condition about special random variables is violated[1].

[1]We conjecture that $\Psi_A \overset{I}{=} \Psi_B$ is a sufficient condition for $A$ and $B$ to be similar. We could not generate any counter-examples, but could not prove the statement either.

Proposition 1 implies that Assumption 1 is equivalent to pairwise similarity of conditionals. Moreover, according to Proposition 2, Assumption 1 implies that for every pair of values $a, b \in \mathcal{X}$, the special random variables associated with $\tilde{Y}_a$ and $\tilde{Y}_b$ are identical. Hence, if there exists $a, b \in \mathcal{X}$ for which $\Psi_{\tilde{Y}_a} \overset{I}{\neq} \Psi_{\tilde{Y}_b}$, then $X \not\to Y$ under our assumptions.

**Definition 8.** Let $\Gamma_{X\to Y}$ be an $S_m$-valued random variable obtained by taking a random sample $x \in \mathcal{X}$ according to the distribution of $X$ (in each environment), and then drawing a random sample from $\Psi_{\tilde{Y}_x}$. More formally,

$$\Gamma_{X\to Y} := \Psi_{\tilde{Y}_X} = \Phi_{\tilde{Y}_X}(\tilde{Y}_X). \tag{8}$$

**Proposition 3.** For every $a, b \in \mathcal{X}$, $\Psi_{\tilde{Y}_a} \overset{I}{=} \Psi_{\tilde{Y}_b}$ if and only if

$$\Gamma_{X\to Y} \perp\!\!\!\perp X, \text{ under each } \mathbb{P}^e \in \mathcal{M}. \tag{9}$$

**Corollary 1.** Combining this result with Propositions 1 and 2 implies that we can reject the causal direction $X \to Y$ if data does not statistically admit the above independence relation under some $\mathbb{P}^e \in \mathcal{M}$. Since if $\Gamma_{X\to Y} \not\perp\!\!\!\perp X$, under some $\mathbb{P}^e \in \mathcal{M}$, then there would be some $a, b \in \mathcal{X}$ such that $\Psi_{\tilde{Y}_a} \overset{I}{=} \Psi_{\tilde{Y}_b}$ does not hold. Thus, based on Proposition 2, $\tilde{Y}_a$ and $\tilde{Y}_b$ are not similar which violates Assumption 1 according to Proposition 1.

Based on the above corollary, we can present our main result as follows:

**Theorem 1.** *If conditional density function of $Y$ given $X$ is continuous in all environments, then under Assumption 1, there is no causal direction from $X$ to $Y$ if $\Gamma_{X\to Y} \not\perp\!\!\!\perp X$, for some $\mathbb{P}^e \in \mathcal{M}$.*

In Section 4, we propose an algorithm for taking joint random samples from $X$ and $\Gamma_{X\to Y}$ based on observed data in order to perform statistical independence tests and infer causal relationships.

## 3 Extensions

In this section, we describe two extensions to our model. First, we discuss how our results extends to discrete case. Second, we discuss the extension to network of multiple variables. In these extensions, we modify both some of our assumptions in section 2.1 as well as our theoretical results to cover the extended settings[2].

### 3.1 Discrete Case

Assume that $E$, and consequently $Y$, are discrete random variables[3]. Continuity and differentiability are not defined for mappings between discrete sets. Instead, we work with bijective fixed-cause functionals, which yield the same results. Thus, we replace Assumption 1 with the following assumption:

**Assumption 2.** For every $x \in \mathcal{X}$, the fixed-cause functional $f(x, .)$ is bijective.

Like Definition 5, discrete variables $A, B$ are similar (denote it by $A \sim B$), if and only if there exists a bijection $g$ such that $B \overset{I}{=} g(A)$.

**Proposition 4.** For discrete-valued $E$, under Assumption 2, there is no causal direction from $X$ to $Y$ if $\Gamma_{X\to Y} \not\perp\!\!\!\perp X$, for some $\mathbb{P}^e \in \mathcal{M}$.

### 3.2 Multivariate Case

Consider a structural causal model with the set of observable variables $\mathbf{V} = \{V_i\}_{i=1}^N$ and the set of exogenous noises $\mathbf{E} = \{E_i\}_{i=1}^N$. Each variable $V_i \in \mathbf{V}$ is determined by an equation

$$V_i = f_i(\mathbf{PA}^i, E_i), \tag{10}$$

where $\mathbf{PA}^i \subseteq \mathbf{V} \setminus \{V_i\}$ is the parent set of the variable $V_i \in \mathbf{V}$. Similarly to the bivariate case, we assume that the set of exogenous noises $\mathbf{E}$ are mutually independent in every environment. We extend Assumption 1 to the multivariate case and assume the following for each variable $V_i \in \mathbf{V}$.

[2]Our finding could be expressed in measure theoretic language where the continuous and discrete cases (as well as the cases which are neither continuous nor discrete) can be unified. To make the presentation accessible to a larger audience, we decided to adopt the current presentation which avoids complex notation and technical discussions in measure theoretic formalism.

[3]Since we assume that $f(x, .)$ is a bijective function from $\mathcal{E}$ to $\mathcal{Y}$, $Y$ should take discrete values.

**Assumption 3.** For every value of $V_i$'s parents like $pa^i$, we assume that the fixed-cause functional $f_i(pa^i, .)$ is a diffeomorphism.

Note that Assumption 3 implies Assumption 1 for each variable and the set of its parents, i.e., the bivariate case where $X := \mathbf{PA}^i$ and $Y := V_i$.

**Proposition 5.** Under Assumption 3, if for each variable $V_i$, its conditional density function given any subset of variables $\mathbf{S} \subset \mathbf{V}$ is continuous in all environments, then $\mathbf{S} \neq \mathbf{PA}^i$ if

$$\Gamma_{\mathbf{PA}^i \to V_i} \not\perp\!\!\!\perp \mathbf{PA}^i, \text{ for some } \mathbb{P}^e \in \mathcal{M}. \tag{11}$$

We discussed the extensions to continuous multivariable case and discrete bivariate cases. Similarly, note that they can be extended to mixed-type (continuous or discrete) multivariable case.

## 4 Causal Discovery Algorithm

We assume that in each environment, there are $n_e$ i.i.d. samples drawn from $\mathbb{P}^e \in \mathcal{M}$. Denote these observations as $\{(x_l^e, y_l^e)\}_{l=1}^{n_e}$.[4]

### 4.1 Sampling From $\Gamma_{X \to Y}$

In this section, we describe our algorithm for the bivariate case. As discussed in Subsection 3.2, the algorithm can also be used for multivariate case.

If the conditional distribution functions $\{p^e_{Y|X}\}_{e=1}^m$ were available, we could obtain random samples from $\Gamma_{X \to Y}$ under each of the probability measure as followings. Note that $\tilde{Y}_X$ (Definition 8) and $Y$ are identically distributed. For a datapoint $(x, y)$ from say environment $e$, we would form the vector $w = (p^1_{Y|X}(y|x), p^2_{Y|X}(y|x), \ldots, p^e_{Y|X}(y|x))^T$ and project it on $S_m$ to obtain $\gamma := \frac{w}{\|w\|_1}$, which would be identically distributed as $\Gamma_{X \to Y}$ under $\mathbb{P}^e$. When the true conditional distribution functions are not available, assuming we have enough samples from the joint from joint observations of $X$ and $Y$, we estimate the conditional distribution function $p^e_{Y|X}$ in each of the environments: $\hat{p}^e_{Y|X}$.[5] Using these estimated conditionals, we can obtain joint samples of $(\Gamma_{X \to Y}, Y)$ as described above for each environment.

Clearly, this sampling routine yields "approximately accurately distributed" samples as $(\gamma, x)$. As the size of data increases (in all environments), we obtain arbitrarily accurately distributed samples of $(\Gamma_{X \to Y}, X)$ since we can estimate arbitrarily accurately estimated conditional distribution functions using sufficiently large samples in all environments.

### 4.2 Inference

As described in Subsection 4.1, we can obtain samples of $\Gamma_{X \to Y}$ in each environment. Using these samples, we perform an independence test to evaluate $\Gamma_{X \to Y} \perp\!\!\!\perp X$ under each $\mathbb{P}^e$.[6] According to Theorem 1, we shall reject $X \to Y$ if this independence relation is rejected under any probability measure $\mathbb{P}^e \in \mathcal{M}$. To aggregate the results of these $m$ independence tests (one in each of the $m$ environments), we consider the minimum $p$-values of the independence tests performed in each of the environments. This ensure that the output of the aggregation is small whenever the independence relation is rejected in at least one environment.

Without further assumptions about the data generation mechanism, the true causal structure is identifiable up to Markov equivalence classes (skeleton[7] and v-structures of the causal graph)[13]. As we are mainly concerned with orienting the undirected edges of the skeleton graph, we seek to find the set of parents of the nodes in the graph. Let $V$ denote a variable in the SEM, and $\mathbf{A}$, the set of all the variables adjacent to $V$. Clearly, parents of $V$ denoted by $PA(V)$ (or for short $\mathbf{PA}$) is a subset of $\mathbf{A}$.

Let $L(\mathbf{S})$ be the minimum p-value of testing $\Gamma_{\mathbf{S} \to V} \perp\!\!\!\perp V$ over all of the environments. If $L(\mathbf{S})$ is lower than some threshold $c$, then there is enough evidence that in at least one environment, the independence relation $\Gamma_{\mathbf{S} \to V} \perp\!\!\!\perp V$ is violated. As a result of Theorem 1, $L(\mathbf{PA})$ should not be too small. Therefore, if $L(\mathbf{S}) < c$ we conclude that $\mathbf{S} \neq \mathbf{PA}$. In order to obtain a single subset of $\mathbf{A}$ as the inferred parent set, we propose the following heuristics:

[4] To cover the discrete extension from Subsection 3.1, we use the general term "distribution" to refer density functions in continuous settings and probability mass functions in discrete settings.

[5] We performed this estimation using NP package in R [17] in our implementation.

[6] In our implementation, we used d-variable HSIC test [18] provided in dHSIC package in R.

[7] The skeleton of a directed graph $G$ is an undirected graph which does not take the direction of edges of $G$ into account.

**H1**: Compute $L(\mathbf{S})$ for all $\mathbf{S} \subset \mathbf{A}$. To ensure that $\mathbf{PA}$ is contained in the output, return

$$\hat{\mathbf{PA}} := \bigcup_{\mathbf{S}\subset\mathbf{A}:L(\mathbf{S})>c} \mathbf{S}. \tag{12}$$

**H2**: If we know the size of the parent set apriori or have a bound on it, it suffices to explore subsets with that size or up to that bound, and return the subset with maximum value of $L$.

## 5 Experiments

In this section, we evaluate the performance of proposed method in synthetic and real data and compare it with related work.

### 5.1 Synthetic Data

We used a heteroscedastic noise model to generate our synthetic data in two environments. In this model, variable $Y$ is determined by the value of its parent $X$ and an exogenous noise $E$ by the following equation

$$Y = f(X) + g(X)E, \tag{13}$$

where we assumed $f(X) := \alpha^T X$, and $g(X)$ was randomly selected with equal probability from $\sqrt{|\beta^T X| + 1}$ or $\log(|\beta^T X| + 2)$, respectively. We focused on the setting discussed in Subsection 4.2, in which we observe variables $\mathbf{A} \cup \{V\}$, and the value of each variable is determined by the value of its parents according to (13). The coefficients $\alpha$ and $\beta$ for each structural equation were drawn randomly from $\mathcal{N}(0, 0.05)$ and *unif*$([-2, -1] \cup [1, 2])$, respectively. We considered two environments and in both environments, the exogenous noise was assumed normal. Let $\mu_e$ and $\sigma_e$ be the mean and the standard deviation of exogenous noise in environment $e \in \{1, 2\}$, respectively. We drew $\mu_1$ randomly from $\mathcal{N}(0, 1)$ and $\sigma_1$ was set to 2. We also set the parameters in the second environment to $\mu_2 = \mu_1 + $ *unif*$(\{-3, 3\})$ and $\sigma_2 = \sigma_1$*unif*$([\frac{2}{3}, \frac{3}{2}])$.

We compared the proposed methods with ICP [6], NLICP [7], MC [12], IB [12], LRE [9], and CD-NOD [10] which have been proposed previously for the multi-environment setting. ICP method assumes that the target variable is a linear function of a subset of predictors, where the coefficients may change across the environments while the distribution of additive exogenous noise is assumed to be invariant. Non-linear ICP allows the functional relation to be non-linear. LRE considers invariant linear functional relations while the distribution of additive exogenous noise might change among the environments. IB and MC extend LRE so that neither distribution of noise nor the linear functional relation is assumed to be invariant. CD-NOD recovers causal relations based on the assumption of independent changes of conditional probability $\mathbb{P}(\text{cause}|\text{effect})$. In our experiments, we also considered LiNGAM [4] which is a powerful causal discovery algorithm in single environment setting and assumes linear functional relations with additive non-Gaussian noise.

We evaluated the performance of our algorithms as well as previous work in two cases (each case comprised of 100 instances of synthetic dataset). We tuned the hyper-parameters of all algorithms (e.g. $c$ for our algorithm) with a training dataset before each experiment, and evaluated the performance on another testing dataset. In both experiments, H2 was initially fed with the size of true parent set.

**Bivariate case.** In this case, $|\mathbf{A}| = 1$. The adjacent variable was set as the parent of $V$ with probability $\frac{1}{2}$. We generated $1000$ samples in each of the environments. Accuracy of the algorithms are reported in Table 2. The proposed method (H1) indeed achieves the highest accuracy.

Table 2: Bivariate case: Accuracy of algorithms on detection of causal direction between two variables.

| **Alg.** | H1 | ICP 6 | NLICP 7 | LiNGAM 4 | MC 12 | IB 12 | LRE 9 | CD-NOD 10 |
|---|---|---|---|---|---|---|---|---|
| **Acc.** | 82% | 43% | 51% | 66% | 62% | 59% | 43% | 10% |

**Multivariate case.** In this case, $|\mathbf{A}|$ was selected uniformly at random from $\{2, 3, 4, 5\}$, and we picked $|\mathbf{PA}|$ uniformly from $\{1, 2, ..., |\mathbf{A}|\}$. We considered the fully connected skeleton among the variables $\mathbf{A} \cup \{V\}$ and the parents were determined according to a random topological order. We generated $10000$ samples in each of the environments and evaluated Precision, Recall, and F1-score for each of the algorithms. We repeated this procedure 100 times. Figure 2 depicts the performances of the various algorithms we compared. H1 outperformed all other algorithms with 20% margin of F1-score. H2 which has the advantage of prior knowledge about the size of parent set, has 12% improvement in precision over H1. (Please note that CD-NOD algorithm does not return any output in a reasonable time for the size of parents greater than one. Thus, its performance is reported only in the bivariate case.).

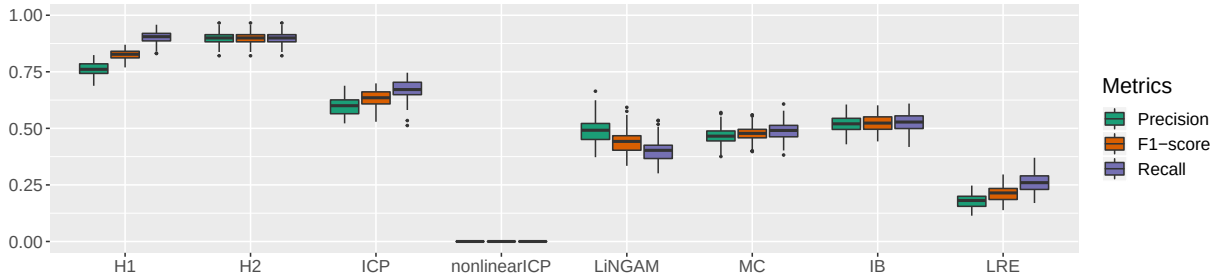


Figure 2: Multivariate case: Performance of our methods (H1 and H2) as well as previous work.

### 5.2 Real World Data

**CollegeDistance dataset.** We considered educational attainment data [19] which was collected from approximately 1100 high-school students. The data contains 13 features including gender, race, base year composite test score, family income, etc. As in [6] and [9], we split the observations into two groups (treated as two environments). This is done based on the distance to the closest 4-year college where the first group contains students within the 10 miles radius. The target variable is the number of years of education. We run our method for 100 trials and in about $80\%$ of the executions, the set with greatest value of $L(\mathbf{S})$ was $\mathbf{S} = \{\text{race}\}$. Thus, this variable was returned as the parent set of the target variable. This variable along with three other candidates was also selected in [9] as the parent set.

**Adult dataset.** This dataset is available at UCI Machine Learning Repository [20] and contains census information. We pre-processed the data by filtering or transforming some of the features before feeding them to our causal discovery algorithm. We chose sex as the environment variable, and working hour per week as the target variable. We considered the following features: age, race, marital status, level of education, level of income, work class, and country. Among these variables, the maximum value of $L(\mathbf{S})$ was achieved by $\mathbf{S} = \{\text{country}\}$. This is not surprising since working policy in each country has a direct impact on the working hour per week.

## 6 Conclusion

We studied causal discovery problem in multi-environment setting with invariant causal relations and varying noise distribution. We presented an identifiablity result assuming bijective fixed-cause functionals, which is less restrictive than previous approaches, and holds in several classical models. We defined "*similarity*" relation between random variables in settings with multiple probability measures and we introduced an independence relation which necessarily holds in case of causation. We compared the performance of our approach with previous work and in our experiments our method out performed them with a high margin. Future work includes extension of results to non-stationary time-series and applications in sequential decision-making.

## References


[1] Judea Pearl. *Causality*. Cambridge University Press, 2009.

[2] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

[3] David Maxwell Chickering. Optimal structure identification with greedy search. *J. Mach. Learn. Res.*, 3(null):507–554, March 2003.

[4] Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.*, 7:2003–2030, December 2006.

[5] Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pages 689–696, 2009.

[6] J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 78(5):947–1012, 2016. (with discussion).

[7] Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6, 06 2017.

[8] Niklas Pfister, Peter Bühlmann, and Jonas Peters. Invariant causal prediction for sequential data. *Journal of the American Statistical Association*, 114(527):1264–1276, 2019.

[9] AmirEmad Ghassami, Saber Salehkaleybar, Negar Kiyavash, and Kun Zhang. Learning causal structures using regression invariance. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3011–3021. Curran Associates, Inc., 2017.

[10] Biwei Huang, Kun Zhang, Jiji Zhang, Joseph Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. *arXiv preprint arXiv:1903.01672*, 2019.

[11] Kun Zhang, Biwei Huang, Jiji Zhang, Clark Glymour, and Bernhard Schölkopf. Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 1347–1353, 2017.

[12] AmirEmad Ghassami, Negar Kiyavash, Biwei Huang, and Kun Zhang. Multi-domain causal structure learning in linear systems. In *Advances in neural information processing systems*, pages 6266–6276, 2018.

[13] J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, USA, 2017.

[14] Kun Zhang and Aapo Hyvarinen. On the identifiability of the post-nonlinear causal model, 2012.

[15] Henry Stark and John W. Woods, editors. *Probability, Random Processes, and Estimation Theory for Engineers*. Prentice-Hall, Inc., USA, 1986.

[16] Roger B Nelsen. *An introduction to copulas*. Springer Science & Business Media, 2007.

[17] Tristen Hayfield and Jeffrey Racine. Nonparametric econometrics: The np package. *Journal of Statistical Software, Articles*, 27(5):1–32, 2008.

[18] Niklas Pfister, Peter Bühlmann, Bernhard Schölkopf, and Jonas Peters. Kernel-based tests for joint independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):5–31, 2018.

[19] Cecilia Elena Rouse. Democratization or diversion? the effect of community colleges on educational attainment. *Journal of Business & Economic Statistics*, 13(2):217–224, 1995.

[20] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

**Remark 1.** Suppose $g, f : \mathcal{A} \to \mathcal{B}$ are diffeomorphisms. Then $g^{-1}$ and $g \circ f$ are also diffeomorphisms.

**Lemma 1.** Similarity is an equivalence relation.

*Proof.* We should check the three properties of equivalence relations. Consider random variables $A$, $B$, and $C$ taking value in $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$, respectively. Assume that $A \sim B$ and $B \sim C$.

*Claim: $\sim$ is reflective.*

Define $g : \mathcal{A} \to \mathcal{A}$ such that $\forall x \in \mathcal{A} : g(x) = x$. It is bijective and continuously differentiable. Hence, it is a diffeomorphism. By Definition 5, $A \sim A$.

*Claim: $\sim$ is symmetric.*

As $A \sim B$, from Definition 5, there exists a diffeomorphism $g$ such that $g(A) \overset{I}{=} B$. From Definition 4, for every $e$ and every $u \subset \mathcal{B}$

$$\mathbb{P}^e(g(A) \in u) = \mathbb{P}^e(B \in u). \tag{14}$$

Note that $g$ is a diffeomorphism, so $g^{-1}$ exists. Let $w$ be an arbitrary subset of $\mathcal{A}$. Let $u = g(w)$, hence $w = g^{-1}(u)$.

$$\begin{aligned}
\mathbb{P}^e(A \in w) &= \mathbb{P}^e(A \in g^{-1}(u)) && (15)\\
&= \mathbb{P}^e(g(A) \in u) && (16)\\
&= \mathbb{P}^e(B \in u) && (17)\\
&= \mathbb{P}^e(g^{-1}(B) \in g^{-1}(u)) && (18)\\
&= \mathbb{P}^e(g^{-1}(B) \in w). && (19)
\end{aligned}$$

Form Remark 1, we know that $g^{-1}$ is a diffeomorphims. As the choice of $w$ was arbitrary, from Definitions 4 and 5, we have

$$B \sim A. \tag{20}$$

*Claim: $\sim$ is transitive.*

Similarly, for every $e$ and every $u \subset \mathcal{B}$

$$\mathbb{P}^e(g(A) \in u) = \mathbb{P}^e(B \in u). \tag{21}$$

Also, for every $e$ and every $q \subset \mathcal{C}$,

$$\mathbb{P}^e(f(B) \in q) = \mathbb{P}^e(C \in q). \tag{22}$$

Note that both $g$ and $f$ are diffeomorphisms. Thus, they are invertible. Consider $q \subset \mathcal{C}$ arbitrarily. Let $u = f^{-1}(q)$ and $w = g^{-1}(u)$.

$$\begin{aligned}
\mathbb{P}^e((f \circ g)(A) \in q) &= \mathbb{P}^e(g(A) \in f^{-1}(q)) && (23)\\
&= \mathbb{P}^e(g(A) \in u) && (24)\\
&= \mathbb{P}^e(B \in u) && (25)\\
&= \mathbb{P}^e(f(B) \in f(u)) && (26)\\
&= \mathbb{P}^e(f(B) \in q) && (27)\\
&= \mathbb{P}^e(C \in q) && (28)
\end{aligned}$$

From Remark 1, we know that $f \circ g$ is a diffeomorphism. Therefore,

$$A \sim C. \tag{29}$$

□

**Proof of Proposition 1**

*Claim: Assumption 1 holds $\Rightarrow \forall a, b \in \mathcal{X} : \tilde{Y}_a \sim \tilde{Y}_b$.*

Fix the environment variable $e$ and consider $x \in \mathcal{X}$ arbitrarily. As a result of Assumption 1, $f(x, .)$ is invertible, and we have

$$p^e_{(Y|X)}(y|x) = p^e_{(E|X)}(f(x, .)^{-1}(y)|x) = p^e_E(f(x, .)^{-1}(y)), \tag{30}$$

where the first equation is because of the model $Y = f(X, E)$, and the second one is due to $X \perp\!\!\!\perp E$.

Let $g := f(x, .)$ which is a diffeomorphism, according to Assumption 1. As $Y = f(X, E)$, we have

$$\tilde{Y}_x \overset{I}{=} g(E). \tag{31}$$

Note that $\tilde{Y}_x$ is a random variable which has the same distribution as $Y$ given $X = x$. According to Definition 5,

$$\forall x \in \mathcal{X} : E \sim \tilde{Y}_x. \tag{32}$$

By Lemma 1, similarity is an equivalence relation, so we have

$$\forall a, b \in \mathcal{X} : \tilde{Y}_a \sim \tilde{Y}_b. \tag{33}$$

*Claim:* $\forall a, b \in \mathcal{X} : \tilde{Y}_a \sim \tilde{Y}_b \Rightarrow$ *Assumption 1 holds.*

Fix arbitrary $x_0 \in \mathcal{X}$. Define $\tilde{E}$ an independent random variable which is identically distributed as $\tilde{Y}_{x_0}$ ($E \overset{I}{=} \tilde{Y}_{x_0}$). Define $\tilde{f}$ so that $\tilde{f}(x, .) := g_x$ where $g_x$ is the diffeomorphism which forms the similarity between $\tilde{Y}_x$ and $\tilde{Y}_{x_0}$ (Definition 1). Assumption 1 holds for $\tilde{f}$.

**Proof of Proposition 2**

Fix the environment index $e$ and consider $s \subset S_m$ arbitrarily. Note that the events $\Psi_A \in s$ and $\Psi_B \in s$ should both be measurable under $\mathbb{P}^e$, as we intend to show

$$\mathbb{P}^e(\Psi_A \in s) = \mathbb{P}^e(\Psi_B \in s), \tag{34}$$

which results in $\Psi_A \overset{I}{=} \Psi_B$ (Definition 4).

Let $\alpha = \Phi_A^{-1}(s)$ and $\beta = \Phi_B^{-1}(s)$. Note that $\alpha \subset \mathcal{A}$ and $\beta \subset \mathcal{B}$. As $A \sim B$, according to Definition 5

$$\exists \text{ diffeomorphism } g : \mathcal{A} \to \mathcal{B} \text{ such that } g(A) \overset{I}{=} B \tag{35}$$

Consider $a \in \alpha$ arbitrarily. Let $b = g(a)$. We have

$$\begin{aligned}
\Phi_B(b) &= c.(p_B^1(b), p_B^2(b), \dots, p_B^m(b)) && (36)\\
&= c.\frac{1}{|det(J_g(a))|}.(p_A^1(a), p_A^2(a), \dots, p_A^m(a)) && (37)\\
&= c'.(p_A^1(a), p_A^2(a), \dots, p_A^m(a)), && (38)
\end{aligned}$$

where (36) holds according to Definition 6 and (37) holds due to (4).

As we defined $\alpha$, we have

$$\Phi_A(a) = c''.(p_A^1(a), p_A^2(a), \dots, p_A^m(a)) \in s. \tag{39}$$

Note that we also have

$$\Phi_B(b) = c'.(p_A^1(a), p_A^2(a), \dots, p_A^m(a)) \in s. \tag{40}$$

Thus, $\Phi_B(b)$ and $\Phi_A(a)$ are proportional with factor $\frac{c'}{c''}$. As they are both in $S_m$, this is possible only if $c'' = c'$ and $\Phi_B(b) = \Phi_A(a)$. Because the choice of $a$ was arbitrary, we have

$$g(\alpha) \subset \beta. \tag{41}$$

According to Lemma 1, $A \sim B \Rightarrow B \sim A$. Therefore, with the same arguments in the reverse direction, we get

$$g^{-1}(\beta) \subset \alpha, \tag{42}$$

which results in

$$g(\alpha) = \beta. \tag{43}$$

Finally,

$$\begin{aligned}
\mathbb{P}^e(\Psi_A \in s) &= \mathbb{P}^e(A \in \alpha) && (44)\\
&= \mathbb{P}^e(g(A) \in g(\alpha)) && (45)\\
&= \mathbb{P}^e(g(A) \in \beta) && (46)\\
&= \mathbb{P}^e(B \in \beta) && (47)\\
&= \mathbb{P}^e(\Psi_B \in s), && (48)
\end{aligned}$$

where (47) follows from $g(A) \overset{I}{=} B$ which is imposed by $A \sim B$.

**Proof of Proposition 3**

*Claim: If for every $a, b \in \mathcal{X}$, $\Psi_{\tilde{Y}_a} \overset{I}{=} \Psi_{\tilde{Y}_b}$, then $\Gamma_{X\to Y} \perp\!\!\!\perp X$, under each $\mathbb{P}^e \in \mathcal{M}$.*

Fix the environment index $e$. Consider any $s \subset S_m$ arbitrarily. Consider $a, b \in \mathcal{X}$ arbitrarily. Then,

$$\mathbb{P}^e(\Gamma_{X\to Y} \in s | X = a) = \mathbb{P}^e(\Psi_{\tilde{Y}_X} \in s | X = a) \tag{49}$$
$$= \mathbb{P}^e(\Psi_{\tilde{Y}_a} \in s) \tag{50}$$
$$= \mathbb{P}^e(\Psi_{\tilde{Y}_b} \in s) \tag{51}$$
$$= \mathbb{P}^e(\Psi_{\tilde{Y}_X} \in s | X = b) \tag{52}$$
$$= \mathbb{P}^e(\Gamma_{X\to Y} \in s | X = b), \tag{53}$$

where (49) is from Definition 8 and (50) is from imposing the condition $X = a$. Moreover, (51) is from the assumption $\Psi_{\tilde{Y}_a} \overset{I}{=} \Psi_{\tilde{Y}_b}$ and Definition 4. As the distribution of $\Gamma_{X\to Y}$ is invariant for every condition on $X$, we conclude $\Gamma_{X\to Y} \perp\!\!\!\perp X$ under $\mathbb{P}^e$ and as the choice of $e$ was arbitrary, this holds for every $e$.

*Claim: If $\Gamma_{X\to Y} \perp\!\!\!\perp X$, under each $\mathbb{P}^e \in \mathcal{M}$, then for every $a, b \in \mathcal{X}$, $\Psi_{\tilde{Y}_a} \overset{I}{=} \Psi_{\tilde{Y}_b}$.*

Fix the environment index $e$. Consider any $s \subset S_m$ arbitrarily. Consider $a, b \in \mathcal{X}$ arbitrarily. Then,

$$\mathbb{P}^e(\Psi_{\tilde{Y}_a} \in s) = \mathbb{P}^e(\Psi_{\tilde{Y}_X} \in s | X = a) \tag{54}$$
$$= \mathbb{P}^e(\Gamma_{X\to Y} \in s | X = a) \tag{55}$$
$$= \mathbb{P}^e(\Gamma_{X\to Y} \in s | X = b) \tag{56}$$
$$= \mathbb{P}^e(\Psi_{\tilde{Y}_b} \in s). \tag{57}$$

Due to arbitrary choice of $s$, according to Definition 4, we have

$$\Psi_{\tilde{Y}_a} \overset{I}{=} \Psi_{\tilde{Y}_b} \tag{58}$$

**Proof of Theorem 1**

In continuous case, we have Assumption 1, along with other technical assumptions mentioned in the Section 2.1. Proposition 1 states that under this set of assumptions, if $X$ causes $Y$ in our model, we have

$$\forall a, b \in \mathcal{X} : \tilde{Y}_a \sim \tilde{Y}_b. \tag{59}$$

Proposition 2 states that for every two similar random variables $A \sim B$, the corresponding special random variables are identical, i.e.,

$$\Psi_A \overset{I}{=} \Psi_B. \tag{60}$$

Being identical means that they have the same distribution function under every measure $\mathbb{P}^e \in \mathcal{M}$. From these two results, we can imply the following condition from our set of assumptions, in case of $X$ causing $Y$ in our model.

$$\forall a, b \in \mathcal{X} : \Psi_{\tilde{Y}_a} \overset{I}{=} \Psi_{\tilde{Y}_b}. \tag{61}$$

Proposition 3 states that from the condition (61), we can imply the following independence criterion in all environments (i.e., for every $\mathbb{P}^e \in \mathcal{M}$):

$$\Gamma_{X\to Y} \perp\!\!\!\perp X. \tag{62}$$

Therefore, our set of assumptions implies that if the causal direction is from $X$ to $Y$, then the above independence should hold in all environments. As a result, violation of this independence criterion in at least one of the environments, rejects the hypothesis that $X$ causes $Y$.

**Proof of Proposition 4**

We should prove equivalent results from Propositions 1, 2, and 3 for the discrete case. We define a discrete notion of similarity.

$$A \overset{d}{\sim} B \iff \exists \text{ bijective } g \text{ such that } g(A) \overset{I}{=} B. \tag{63}$$

Note that Definition 4 should not be changed as it work in both discrete and continuous cases. We state our identifiability result in discrete case as follows:

*Assumption 2 holds if and only if* $\forall a, b \in \mathcal{X} : \tilde{Y}_a \overset{d}{\sim} \tilde{Y}_b$.

The same reasoning in the proof of Proposition 1 yields this result.

The result equivalent to Proposition 2 is as follows:

*For any two discrete-valued random variables* $A$ *and* $B$*, if* $A \overset{d}{\sim} B$*, then* $\Psi_A \overset{I}{=} \Psi_B$.

To prove this result, it suffices to change the proof of Proposition 2 by replacing $|det(J_g(a))|$ with 1 (in Equation (37)), and use $p^e_V$ to denote the probability mass function (instead of probability density function).

Finally, we state the result equivalent to Proposition 3, for discrete case:

*In the discrete case, for every* $a, b \in \mathcal{X}$*, we have* $\Psi_{\tilde{Y}_a} \overset{I}{=} \Psi_{\tilde{Y}_b}$ *if and only if* $\Gamma_{X \to Y} \perp\!\!\!\perp X$.

The exact same reasoning in Proposition 3 works for proving this result. Finally, Proposition 4 is proved in the same way as Theorem 1.

**Proof of Proposition 5**

Proposition 5 is a restatement of Theorem 1 for multivariate case. By setting $X := \mathbf{PA}^i$ and $Y := V_i$, the result is immediately implied from Theorem 1.