Multi-Domain Causal Discovery in Bijective Causal Models

Kasra Jalaldoust KASRA@CS.COLUMBIA.EDU
Department of Computer Science, Columbia University

Saber Salehkaleybar S.SALEHKALEYBAR@LIACS.LEIDENUNIV.NL
Leiden Institute of Advanced Computer Science (LIACS), Leiden University

Negar Kiyavash NEGAR.KIYAVASH@EPFL.CH
College of Management of Technology, École Polytechnique Fédérale de Lausanne (EPFL)

Editors: Biwei Huang and Mathias Drton

Abstract

We consider the problem of causal discovery (a.k.a., causal structure learning) in a multi-domain setting. We assume that the causal functions are invariant across the domains, while the distribution of the exogenous noise may vary. Under causal sufficiency (i.e., no confounders exist), we show that the causal diagram can be discovered under less restrictive functional assumptions compared to previous work. What enables causal discovery in this setting is bijective generation mechanisms (BGM), which ensures that the functional relation between the exogenous noise E and the endogenous variable Y is bijective and differentiable in both directions at every level of the cause variable $X = x$. BGM generalizes a variety of models including additive noise model, LiNGAM, post-nonlinear model, and location-scale noise model. Further, we derive a statistical test to find the parents set of the target variable. 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 represented by a graph called causal diagram in which 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 due to technical or ethical considerations. This limits us to use the observational data that is collected passively from the variables in the system in order to recover the causal diagram. From the observational data, without parametric assumptions, the true causal diagram can be identified up to a Markov equivalence class (MEC), which is the class of graphs representing the same set of conditional independence relations among the observed variables (Pearl, 2009a). In order to identify the MEC, one can use constraint-based algorithms such IC and IC* Pearl (2009a), PC, FCI Spirtes et al. (2000) or score-based methods such as greedy equivalence search Chickering (2003). Moreover, within the framework of structural causal models (SCM), the true graph can be recovered uniquely by considering some additional assumptions on the data generation mechanisms such as non-Gaussianity of noise Shimizu et al. (2006), or non-linearity of functional relations Hoyer et al. (2009). The aforementioned results rely on Markovianity assumption, which states that there exist no exogenous variables that affect multiple endogenous variables, i.e., no confounders; throughout this paper we assume Markovianity.

In recent years, causal discovery using data collected from multiple domains has gained attention (Jaber et al., 2020; Perry et al., 2022; Huang et al., 2019a; Peters et al., 2016b; Mooij et al., 2020;

© 2025 K. Jalaldoust, S. Salehkaleybar & N. Kiyavash.

Yang et al., 2018; Squires et al., 2020). This setting is commonly formulated as having i.i.d samples of the variables across different domains, such that the causal modules may change across datasets, and these changes may or may not be known to the learner. As an example from applications, in (Ackerer et al., 2020) heteroskedasticity has been observed in food expenditure as a function of income for Belgian working class households. From the methods, Peters et al. (2016a) introduced the notion of "invariant prediction", where it is assumed that for every domain $e \in \mathcal{E}$, the response variable Y^i has a fixed set of predictors $X_{S^*}, S^* \subset \{1, 2, \ldots, p\}$ as its "invariant" predictors, such that the residual distribution is invariant across the domains when X_{S^*} is the set of explanatory variables. The task is then finding the set of covariates S^* using observations from multiple domains. Although, correctly recovering the set S^* is contingent on (1) having access to large enough set of domains and (2) invariance of the mechanism that determines Y based on X_{S^*} across all domains. The proposed method searches exhaustively over all possible subsets of the covariates, thus, has high computational and statistical complexity. The output is guaranteed to be contained by X_{S^*} with high probability, however, exact recovery requires further assumptions on diversity of the domains. Later, Heinze-Deml et al. (2017) extended this work to non-linear functional relations, yet still assuming additive noise. Pfister et al. (2019) used invariant causal prediction assumption to handle sequentially ordered data where heterogeneity stems from the varying index of the datapoints. Ghassami et al. (2017) considered invariance of functional relations between variables and their parents in the linear SCM while the distribution of exogenous noises may vary among the domains. The authors introduced a notion of completeness of causal discovery for this setting and proposed a complete algorithm which is computationally intensive. They also presented a heuristic algorithm that improved upon the complete baseline in computational complexity. Huang et al. (2019b) (and previously, Zhang et al. (2017)), proposed a more generalized model for the multi-domain causal discovery problem where the changes in causal mechanisms are modeled by adding deterministic functions from the domain index C to the parameters of causal mechanisms. The variable C itself is also modeled as a random variable in the proposed SCM. It is assumed that the parameters of the functional relation of variable V_i, denoted by $\theta_i(C)$, change independently from the other variables' parameters. However, this assumption may be violated as $\theta_i(C)$ and $\theta_j(C)$ are deterministic functions of the same random variable C, and they are unlikely to be independent unless one of them is constant. Ghassami et al. (2018) proposed a method to identify the causal relation between two random variables in linear SCM where the causal coefficients and/or the distribution of exogenous noises may vary. They also extend this method for a network of variables in linear SCM. The noise model we consider in this work was introduced by Nasr-Esfahany et al. (2023) as bijective generateion mechanisms (BGM) for the purpose of couterfactual identification, where the graph is assumed to be known; in our analysis, we study a complementary task that is the identification of the graph itself under BGM assumption. Immer et al. (2023) and Strobl and Lasko (2023) showed that location-scale noise model is identifiable, though their analysis is limited to bivariate case with both X, Y being real-valued; On the other hand, our noise model is less restrictive, since the location-scale is a special case in our model, although we require observations from more than one domain to make inferences. An extension of LiNGAM known as multi-group LiNGAM (MGL) is implemented for multi-domain setting. (Guo et al., 2023) introduce the "Causal de Finetti" framework, demonstrating how exchangeable data and the independent causal mechanism assumption can be leveraged to uniquely identify invariant causal structures using multi-domain data. (Reizinger et al., 2025) introduce an Identifiable Exchangeable Mechanisms framework that unifies causal discovery, independent component analysis, and causal representation learning by relaxing the conditions for identifying causal structures in exchangeable

data. Another method for multi-domain causal discovery is ORION by (Mian et al., 2023), which uses minimum description length (MDL) principles to jointly discover causal networks and sources of heterogeneity across multiple environments that are modeled as hard and soft interventions.

Contribution

We present a causal discovery algorithm for the multi-domain setting under Markovianity assumption under Bijective generation mechanism (BGM) that generalizes some of the existing noise models including additive noise model (Hoyer et al., 2009), LiNGAM (Shimizu et al., 2006), post-nonlinear noise model (Zhang and Hyvarinen, 2012), and location-scale noise model (Immer et al., 2023; Strobl and Lasko, 2023). Since BGM does not require the exogenous noise to be additive, it captures cases with complex dependencies between the exogenous and endogenous variables, including the case of discrete observables. These dependencies often exist in real-world applications and can cause the methods which are based on independence of the regression residual and the cause to fail. We construct statistical tests to recover the true parent set of a target variable among a set of covariates and prove its soundness under BGM assumptions. Finally, we extend our results to multivariate and mixed-type data in the multi-domain setting.

The structure of this paper is as follows: In Section 3, we present our assumptions and the theoretical guarantees for discovery of the parent set, and propose a statistical testing scheme. We also present the extensions of our results to multivariate and mixed-type data setting. In Section 4, we describe our causal discovery algorithm and heuristics supported by our findings in Section 3. In Section 5 we provide experiment results and compare our method with previous work.

2. Preliminaries

Here we provide the background and notation necessary for the other sections.

Definition 1 *A Structural Causal Model is defined by the set of observable variables* $\mathbf{V} = \{V_1, V_2, \ldots, V_n\}$*, a set of mutually independent exogenous variables* $\mathbf{E} = \{E_1, E_2, \ldots, E_n\}$*, a set of functions (i.e., causal mechanisms)* $\mathcal{F} = \{f_1, f_2, \ldots, f_n\}$*, and a probability measure* $\mathbb{P}$*. In a Markovian SCM,* $V_i \leftarrow f_i(\mathbf{PA}_i, E_i)$*, where* $\mathbf{PA}_i \subseteq \mathbf{V} \setminus \{V_i\}$ *is called the parents of* V_i*. Moreover, the causal graph* G *is constructed over the vertex set* $\mathbf{V}$ *by adding directed edges from every* $V' \in \mathbf{PA}_i$ *to* V_i*. We assume that the SCM is non-recursive, i.e.,* $\mathcal{G}$ *is acyclic.*

In this paper, we first restrict ourselves to a bivariate case with the causal diagram $X \to Y$ which indicates that $Y \leftarrow f(X, E)$, and then we extend the results to the multivariate case. Due to Markovianity (a.k.a. causal sufficiency), we assume that X, E are independent random variables. The sets $\mathcal{X}$,$\mathcal{Y}$, and $\mathcal{E}$ denote the support of X, Y, E, respectively.

There exists a variety of models for multi-domain causal discovery in the literature (Jaber et al., 2020; Perry et al., 2022; Huang et al., 2019a; Peters et al., 2016b; Mooij et al., 2020; Yang et al., 2018; Squires et al., 2020). In order to identify the causal structure uniquely, most of these methods not only assume Markovianity, but also require that either the functional relations f_i or the distribution of exogenous variables $P(E)$ are invariant across the different domains. Furthermore, as discussed in the related work, each approach asserts additional assumptions about the noise models, some more restrictive than others.

Table 1: Noise Models

Model	**Functional Form**	**Fixed-Cause Functional** $f(x, \cdot)$
Additive noise model	$f(X, E) = g(X) + E$	Affine: $g(x) + E$
LiNGAM	$f(X, E) = \beta.X + E$	Affine: $\beta.x + E$
Post-nonlinear model	$f(X, E) = h(g(X) + E)$	monotone $\circ$ affine: $h(g(x) + E)$
location-scale noise model	$f(X, E) = g(X) + h(X)E$	Affine : $g(x) + h(x)E$

To model a multi-domain setting, we assume that the distribution of variables in each domain is governed by a probability measure; the set of these probability measures is denoted by $\mathcal{M} = \{\mathbb{P}^1, \mathbb{P}^2, \ldots, \mathbb{P}^m\}$, where $\mathbb{P}^i$ corresponds to the i-th domain. We assume that these probability measures are defined over the same σ-algebra Σ so that for each $\mathbb{P}^i \in \mathcal{M}$, the triple $(\Omega, \Sigma, \mathbb{P}^i)$ is a probability space. Random variables X, Y, E can have different distributions under each probability measure, but the distribution of Y depends on the distributions of X and E through the function $f : \mathcal{X} \times \mathcal{E} \to \mathcal{Y}$ that we assume to be invariant across the domains. For each random variable $V : \Omega \to \mathcal{V}$ (such as X and Y) and a value $v \in \mathcal{V}$ (such as $x \in \mathcal{X}$), we denote the distribution under $\mathbb{P}^i \in \mathcal{M}$ as $p^i_V(v)$. Similarly, for a pair of random variables $V, W \in \mathbf{V}$, we denote the distribution of V conditional on W as $p^i_{V|W}(v|w)$.

For multivariate functions like $g : \mathcal{A} \times \mathcal{B} \to \mathcal{C}$, and the points $a \in \mathcal{A}, b \in \mathcal{B}$, we write $g(\cdot, b)$ or $g(a, \cdot)$ to denote the functional $g_b : \mathcal{A} \to \mathcal{C}$ 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 multivariate functions defined over the support of the random variables.

3. Identification Result

We assume that the joint density function $p^i_{X,Y}$ exists under each $\mathbb{P}^i \in \mathcal{M}$. 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 every $\mathbb{P}^i \in \mathcal{M}$, or it has a positive probability under every $\mathbb{P}^i \in \mathcal{M}$; This assumption is known as positivity, it simplifies the results and the notation, and is standard in the literature (Rosenbaum and Rubin, 1983; Pearl, 2009b; Shpitser and Pearl, 2008; Peters et al., 2016b; Dawid, 2010; Imbens and Rubin, 2015). Thus, under positivity assumption, the densities $\{p^i_{X,Y}\}_{i=1}^m$ would all have the same support. We also assume that the function $f : \mathcal{X} \times \mathcal{E} \to \mathcal{Y}$ to be invariant across the domains.

We assume Markovianity, i.e., causal sufficiency in all domains, that is $E \perp\!\!\!\perp X$ in under every $\mathbb{P}^i \in \mathcal{M}$, or equivalently, $p^i_{E|X}(e|x) = p^i_E(e)$ for all values of the noise $E = e$. This assumption is also known as the principle of independent mechanisms is commonly imposed in the literature of multi-domain or single-domain causal discovery (Schölkopf et al., 2012; Peters et al., 2017; Lemeire and Janzing, 2013).

Definition 2 *A fixed-cause functional at $x \in \mathcal{X}$ is denoted by $f(x, \cdot)$ which represents the functional relation between the exogenous noise E and the effect variable Y, for a specific value of cause, x.*

Definition 3 *A bijective function $g : \mathcal{A} \to \mathcal{B}$ is a diffeomorphism if it is differentiable and also has a differentiable inverse.*

Below, we introduce the key assumption we use in this paper.

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

Notably BGM extends all classical noise models in the Table 1, since they all satisfy this Assumption 1. There are some recent works on learning causal directions by considering additional assumptions on the data generation mechansims. For example, in the CdF theorem (bivariate case) (Guo et al., 2023), it is assumed that there are two latent random variables corresponding to the causal mechanisms generating the cause and effect (given the cause). However, our approach does not take a Bayesian perspective and does not require such latent variables. Another line of research relies on nonlinear ICA, which often imposes additional assumptions on exogenous noise, such as the "variability" assumption (Hyvarinen et al., 2019). In contrast, our work does not require these assumptions. For instance, in (Reizinger et al., 2023), Assumption 2(v) states that the conditional distribution of certain pairs of exogenous noises must follow a von Mises-Fisher distribution.

We are given n_i i.i.d. samples $\{(x_l^i, y_l^i)\}_{l=1}^{n_i}$ for each $\mathbb{P}^i \in \mathcal{M}$, and our task is to decide whether $X \to Y$ is a valid causal diagram under BGM.

3.1. Necessary Conditions Implied by BGM

Definition 4 *Define $\tilde{Y}_x$ as a $\mathcal{Y}$-valued random variable with the same distribution as Y conditioned on $X = x$, under each $\mathbb{P}^i \in \mathcal{M}$. More precisely, for each measurable subset $u \subset \mathcal{Y}$ and for each $1 \leq e \leq m$, we have*

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

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

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

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

Definition 5 (Identical r.v.s) *Consider two $\mathcal{V}$-valued random variables V and V'. V and V' are "identical" (denoted by $V \overset{I}{=} V'$) if for every $\mathbb{P}^i \in \mathcal{M}$ and for each $u \subset \mathcal{V}$,*

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

Definition 6 (Similar r.v.s) *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 g, density function of $W = g(V)$ under $\mathbb{P}^i \in \mathcal{M}$ is $p^i_W(w) = \frac{p^i_V(g^{-1}(w))}{|det(J_g(g^{-1}(w)))|}$, 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 $\mathcal{V}$ (Stark and Woods, 1986). Thus, for any pair of similar random variable random variables $A \sim B$ (Definition 6), 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 in a single-domain setting, any two continuous probability density functions can be transformed into each other by diffeomorphisms Nelsen (2007). Therefore, similarity of two random variables is trivially true as long as $|\mathcal{M}| = 1$. However, in a multi-domain setting, i.e., $|\mathcal{M}| > 1$, similarity of random variables would be a non-trivial relation. Below, we prove a property that holds under BGM which would in turn enable causal discovery.

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".

To further elaborate on the above, we note that similarity of random variables requires existence of a diffeomorphism that transforms one to the other. In a multi-domain setting, not every two continuous random variables are similar, and therefore, we can test this property. According to Proposition 1, in a bijective generation mechanisms (i.e., Assumption 1), the variables that represent the conditional distributions of the effect variable Y are all similar.

Proposition 1 can be used as a basis to reject $X \to Y$ as follows: if we find $a, b \in \mathcal{X}$ such that $\tilde{Y}_a \not\sim \tilde{Y}_b$, that is $\tilde{Y}_b \overset{I}{\neq} g(\tilde{Y}_a)$ for any diffeomorphism $g : \mathcal{Y} \to \mathcal{Y}$, we can reject $X \to Y$. Note that verifying pairwise similarity of conditionals, even for a single pair $a, b \in \mathcal{X}$, is intractable in this way since there are countless candidates diffeomorphisms that one must consider. In fact, Proposition 1 is just an identifiability result and it does not provide a method on how to test for pairwise similarity property. Next, we introduce a practical method for statistically testing pairwise similarity of conditionals.

3.2. Statistical Tests for Similarity

First, we introduce a probabilistic indicator of similarity. Next, we propose an algorithm by which we can measure this indicator using finite samples.

According to Eq. (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 7 *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 $v \in \mathcal{V}$.

In words, the density-vectorization operator takes as input a point in the support of V, concatenates the density of V across the domains into a vector of m entries, and then normalizes it so that they sum up to one. The reason for using this operator is to summarize the relative density of V in each point of the support, since this quantity turns out to be an important invariant of this setting that we leverage for identification.

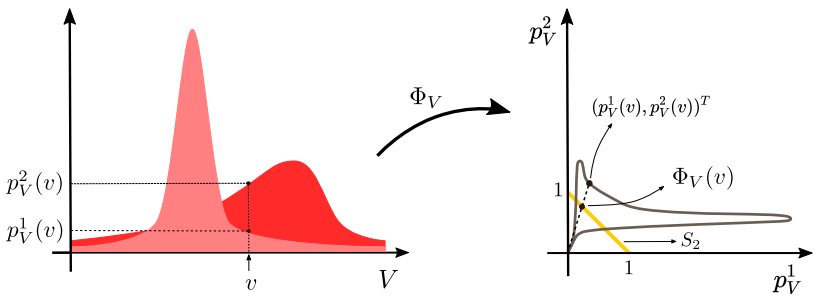

Figure 1: Left: Density functions p_V^1, p_V^2 for random variable V. Values of density functions at point v are also shown in this plot. Right: Points on the gray curve are vectorization of values of density functions p_V^1, p_V^2 for each point v in $\mathbb{R}$. Projecting the density vector $(p_V^1(v),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

Definition 8 *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 Φ_V is a deterministic function, but Ψ_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 $\Psi_A \overset{I}{=} \Psi_B$.*

In words, 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 is a necessary condition, and can be used to reject the hypothesis of similarity whenever the latter condition about special random variables is violated[1]. That being said, in case the distribution of variables does not change between the domains, then the density-vectorization operator collapses all points of the support to the point $\underbrace{\langle \frac{1}{m}, \frac{1}{m}, \ldots, \frac{1}{m} \rangle}_{m \text{ times}}$. This implies that Ψ_V is a constant, thus would be independent of all random variables. Therefore, if the distribution of variables in different domains are identical then the tests would be uninformative since no hypothesis may be rejected.

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. Motivated by this observation, we design a statistical test to reject pairwise similarity, thus, the causal diagram $X \to Y$ under BGM.

Definition 9 *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 domain), and then drawing a random sample from $\Psi_{\tilde{Y}_x}$. More formally,*

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

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

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

Corollary 1 *Combining the latter 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 $\mathbb{P}^i \in \mathcal{M}$. Since if $\Gamma_{X \to Y} \not\perp\!\!\!\perp X$, under some $\mathbb{P}^i \in \mathcal{M}$, 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 state our main result as follows:

Theorem 1 *Under Assumption 1 and for continuous random variables X and Y, the causal diagram $X \to Y$ can be rejected if $\Gamma_{X \to Y} \not\perp\!\!\!\perp X$, for some $\mathbb{P}^i \in \mathcal{M}$.*

In what follows, 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 3 as well as our theoretical results to cover the extended settings[2].

3.3. Discrete Case

Assume that E, and consequently Y, are discrete random variables[3]. Continuity and differentiability are not defined for mappings between discrete sets. Thus, we replace Assumption 1 with the following assumption:

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

Akin to Definition 6, 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 $\mathbb{P}^i \in \mathcal{M}$.*

In words, we can still test the causal diagram $X \to Y$ using the same exact method described for the continuous case.

3.4. Multivariate Case

In this subsection, we extend our results to a setting with more than two observed variables by extending the Assumption 1 to the multivariate case as follows.

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

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 $f(x, \cdot)$ is a bijective function from $\mathcal{E}$ to $\mathcal{Y}$, Y should take discrete values.

Notably Assumption 3 is an extension of Assumption 1 for each endogenous variable.

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 domains, then $\mathbf{S} \neq \mathbf{PA}_i$ if*

$$\Gamma_{\mathbf{PA}_i \to V_i} \not\perp\!\!\!\perp \mathbf{PA}_i, \text{ for some } \mathbb{P}_i \in \mathcal{M}. \tag{10}$$

In words, the conditional independence test in Eq. 10 can be used as a basis for rejecting the hypothesis of a certain set being the parent set of V_i. Similarly, note that this result can be extended to mixed-type (continuous or discrete) multivariable case. In what follows, we construct the discovery routine based on our findings.

4. Algorithm

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

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

In Subsection 3.4 we propose the algorithm in the general multivariate case. If the conditional distribution functions $\{p^i_{Y|X}\}_{i=1}^m$ 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 9) and Y are identically distributed, since the subscript X is randomized. For a datapoint (x, y) from say domain e, we form the vector $w = (p^1_{Y|X}(y|x), p^2_{Y|X}(y|x), \ldots, p^i_{Y|X}(y|x))^T$ and project it on S_m (the simplex in m dimensions) to obtain $\gamma := \frac{w}{\|w\|_1}$, which would be identically distributed as $\Gamma_{X \to Y}$ under $\mathbb{P}^i$. When the true conditional distribution functions are not available, assuming we have enough samples from the joint observations of X and Y, we estimate the conditional distribution function $p^i_{Y|X}$ in each of the domains: $\hat{p}^i_{Y|X}$.[5] Using these estimated conditionals, we can obtain joint samples of $(\Gamma_{X \to Y}, Y)$ as described above for each domain.

Upon consistent estimation of the conditional distributions, this sampling routine yields "approximately accurately distributed" samples as (γ, x). As the size of data increases (in all domains), we obtain more accurate samples of $(\Gamma_{X \to Y}, X)$, as we can estimate the conditional distributions more accurately across the domains.

4.2. Inference

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

4. To cover the discrete extension from Subsection 3.3, 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 Hayfield and Racine (2008) in our implementation, and the code is available in the supplementary material.
6. In our implementation, we used d-variable HSIC test Pfister et al. (2018) provided in dHSIC package in R.

domains. This ensure that the output of the aggregation is small whenever the independence relation is rejected in at least one domain.

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) (Peters et al., 2017). 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 SCM, 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 domains. If $L(\mathbf{S})$ is lower than some threshold c, then there is enough evidence that at least in one domain, the independence relation $\Gamma_{\mathbf{S}\to V} \perp\!\!\!\perp V$ is violated. As a result of Theorem 1, 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 methods:
H1: Compute $L(\mathbf{S})$ for all $\mathbf{S} \subset \mathbf{A}$. To ensure that $\mathbf{PA}$ is contained in the output, return

$$\hat{\mathbf{PA}} := \bigcup_{\mathbf{S}\subset\mathbf{A}:L(\mathbf{S})>c} \mathbf{S}. \tag{11}$$

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

Next, we evaluate the performance of proposed methods, and we explore use of our algorithms 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 domains. In this model, variable Y is determined by the value of its parent X and an exogenous noise E by the following equation

$$Y = f(X) + g(X)E, \tag{12}$$

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 Eq. (12). The coefficients α and β for each structural equation were drawn randomly from $\mathcal{N}(0, 0.05)$ and *unif*$([-2, -1] \cup [1, 2])$, respectively. We considered two domains and in both domains, the exogenous noise was assumed normal. Let μ_e and σ_e be the mean and the standard deviation of exogenous noise in domain $e \in \{1, 2\}$, respectively. We drew μ_1 randomly from $\mathcal{N}(0, 1)$ and σ_1 was set to 2. We also set the parameters in the second domain to $\mu_2 = \mu_1 + \textit{unif}(\{-3, 3\})$ and $\sigma_2 = \sigma_1 \textit{unif}([\frac{2}{3}, \frac{3}{2}])$. An instance of the data in bivariate case is shown in Figure 2.

We compared the proposed methods with ICP (Peters et al., 2016a), NLICP (Heinze-Deml et al., 2017), MC (Ghassami et al., 2018), IB (Ghassami et al., 2018), LRE (Ghassami et al., 2017), CD-NOD (Huang et al., 2019b), CdF (Guo et al., 2023), and multi-group LiNGAM (MGL) (Shimizu,

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

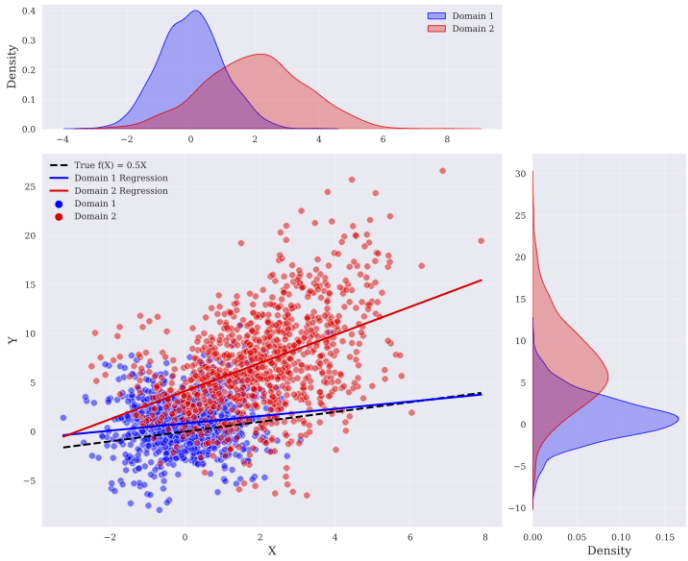

Figure 2: 1000 datapoints in two domains, the marginals of X and Y, the regression lines in each domain, and the true invariant component f.

Alg.	H1	ICP	NLICP	LiNGAM	MC	IB	LRE	CD-NOD	CdF	MGL
Acc.	82%	43%	51%	66%	62%	59%	43%	10%	49%	63%

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

2012) which have been proposed previously for the multi-domain setting. ICP method assumes that the target variable is a linear function of a subset of predictors, where the coefficients may change across the domains 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 domains. 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 (Shimizu et al., 2006) which is a causal discovery algorithm in single domain 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. The code is made available here: https://github.com/jalaldoust/MD-BGM.

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 domains. Accuracy of the algorithms are

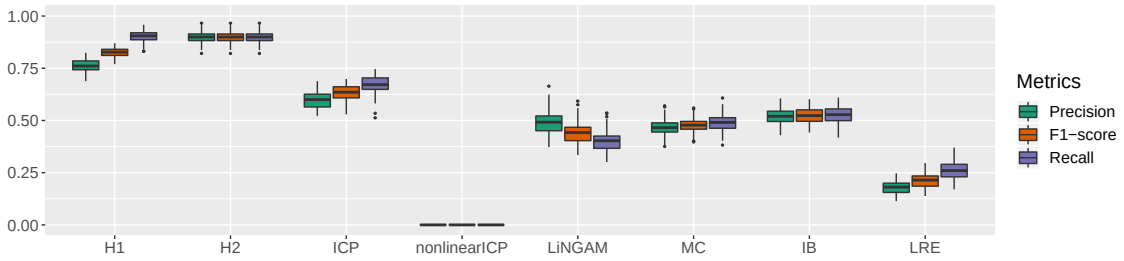

Figure 3: Multivariate case: Performance of our methods (H1 and H2) as well as previous work.

reported in Table 2. The proposed method (H1) indeed achieves the highest accuracy. It is worth to note that since we used only 100 instances of the problem, the standard deviation of the accuracies is no more than 5%, which still deems our algorithm as the most accurate with high confidence.

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 domains and evaluated Precision, Recall, and F1-score for each of the algorithms. We repeated this procedure 100 times. Figure 3 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).

5.2. Real World Data

CollegeDistance dataset. We considered educational attainment data (Rouse, 1995) 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 Peters et al. (2016a) and Ghassami et al. (2017), we split the observations into two groups (treated as two domains). 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 Ghassami et al. (2017) as the parent set.

Adult dataset. This dataset is available at UCI Machine Learning Repository Dua and Graff (2017) 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 domain 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.

5.3. Limitations

Our algorithm has few practical limitations. First, accurately estimating the density in each environment may require large amounts of data, and the minimum required data grows exponentially with the dimensionality of the variables unless additional assumptions are imposed. Second, our framework requires domain indices to be integers, and the causal discovery task depends on having sufficient data in all domains. Consequently, our method is most effective when the number of domains is small relative to the dataset size. Third, the computational cost is significant due to the combinatorial search over parent sets, computation of test statistics, and execution of independence tests. A theoretical limitation of our approach is that we do not claim completeness of the discovery algorithm—while we guarantee soundness (i.e., once a non-parent is identified, it is truly a non-parent), we do not ensure that all non-parents are rejected. Since completeness is not the focus of this work, the discussion of necessary faithfulness assumptions remains an open question.

6. Conclusion

We studied causal discovery problem in multi-domain setting with invariant causal relations and varying noise distribution. We presented an identifiablity result assuming bijective generation mechanisms, 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

Damien Ackerer, Natasa Tagasovska, and Thibault Vatter. Deep smoothing of the implied volatility surface. *Advances in Neural Information Processing Systems*, 33:11552–11563, 2020.

David Maxwell Chickering. Optimal structure identification with greedy search. *J. Mach. Learn. Res.*, 3(null):507–554, March 2003. ISSN 1532-4435. doi: 10.1162/153244303321897717. URL https://doi.org/10.1162/153244303321897717.

A. Philip Dawid. Beware of the dag! In *Proceedings of the Nineteenth Annual Conference on Uncertainty in Artificial Intelligence*, 2010.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

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. URL http://papers.nips.cc/paper/6894-learning-causal-structures-using-regression-invariance.pdf.

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.

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. URL https://openreview.net/forum?id=o4RtDFMSNL.

Tristen Hayfield and Jeffrey Racine. Nonparametric econometrics: The np package. *Journal of Statistical Software, Articles*, 27(5):1–32, 2008. ISSN 1548-7660. doi: 10.18637/jss.v027.i05. URL https://www.jstatsoft.org/v027/i05.

Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6, 06 2017. doi: 10.1515/jci-2017-0016.

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.

Biwei Huang, Kun Zhang, Mingming Gong, and Clark Glymour. Causal discovery and forecasting in nonstationary environments with state-space models. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 2901–2910. PMLR, 2019a.

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*, 2019b.

Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd international conference on artificial intelligence and statistics*, pages 859–868. PMLR, 2019.

Guido W Imbens and Donald B Rubin. *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge University Press, 2015.

Alexander Immer, Christoph Schultheiss, Julia E Vogt, Bernhard Schölkopf, Peter Bühlmann, and Alexander Marx. On the identifiability and estimation of causal location-scale noise models. In *International Conference on Machine Learning*, pages 14316–14332. PMLR, 2023.

Amin Jaber, Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. Causal discovery from soft interventions with unknown targets: Characterization and learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 9551–9561. Curran Associates, Inc., 2020.

Jan Lemeire and Dominik Janzing. Revisiting causal discovery by assuming the independence of cause and mechanism. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2013.

Osman Mian, Michael Kamp, and Jilles Vreeken. Information-theoretic causal discovery and intervention detection over multiple environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 9171–9179, 2023.

Joris M Mooij, Sara Magliacane, and Tom Claassen. Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, 21(99):1–108, 2020.

Arash Nasr-Esfahany, Mohammad Alizadeh, and Devavrat Shah. Counterfactual identifiability of bijective causal models. In *International Conference on Machine Learning*, pages 25733–25754. PMLR, 2023.

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

Judea Pearl. *Causality*. Cambridge University Press, 2009a. doi: 10.1017/CBO9780511803161.

Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009b.

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*, volume 35, pages 10904–10917. Curran Associates, Inc., 2022.

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, 2016a. URL http://arxiv.org/abs/1501.01332. (with discussion).

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

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016b.

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.

Niklas Pfister, Peter Bühlmann, and Jonas Peters. Invariant causal prediction for sequential data. *Journal of the American Statistical Association*, 114(527):1264–1276, 2019. doi: 10.1080/01621459.2018.1491403. URL https://doi.org/10.1080/01621459.2018.1491403.

Patrik Reizinger, Yash Sharma, Matthias Bethge, Bernhard Schölkopf, Ferenc Huszár, and Wieland Brendel. Jacobian-based causal discovery with nonlinear ica. *Transactions on Machine Learning Research*, 2023.

Patrik Reizinger, Siyuan Guo, Ferenc Huszár, Bernhard Schölkopf, and Wieland Brendel. Identifiable exchangeable mechanisms for causal structure and representation learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=k03mB41vyM.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

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

Bernhard Schölkopf, Dominik Janzing, and Jonas Peters. Semi-supervised learning, causality, and the conditional independence assumption. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012.

Shohei Shimizu. Joint estimation of linear non-gaussian acyclic models. *Neurocomputing*, 81: 104–107, 2012.

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. ISSN 1532-4435.

Ilya Shpitser and Judea Pearl. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9:1941–1979, 2008.

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

Chandler Squires, Yuhao Wang, and Caroline Uhler. Permutation-based causal structure learning with unknown intervention targets. *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, 124:909–918, 2020.

Henry Stark and John W. Woods, editors. *Probability, Random Processes, and Estimation Theory for Engineers*. Prentice-Hall, Inc., USA, 1986. ISBN 013711706X.

Eric V Strobl and Thomas A Lasko. Identifying patient-specific root causes with the heteroscedastic noise model. *Journal of Computational Science*, 72:102099, 2023.

Karren D Yang, Ari Katcoff, and Caroline Uhler. Characterizing and learning equivalence classes of causal dags under interventions. *Proceedings of the 35th International Conference on Machine Learning*, 80:5541–5550, 2018.

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

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. doi: 10.24963/ijcai.2017/187. URL https://doi.org/10.24963/ijcai.2017/187.

Appendix

Remark 1

Suppose $g, f : \mathcal{A} \to \mathcal{B}$ are diffeomorphisms. Then g^{-1} and gof 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 6, $A \sim A$.

Claim: $\sim$ is symmetric.

As $A \sim B$, from Definition 6, there exists a diffeomorphism g such that $g(A) \overset{I}{=} B$. From Definition 5, for every e and every $u \subset \mathcal{B}$

$$\mathbb{P}^i(g(A) \in u) = \mathbb{P}^i(B \in u). \tag{13}$$

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

$$\begin{aligned}
\mathbb{P}^i(A \in w) &= \mathbb{P}^i(A \in g^{-1}(u)) && (14)\\
&= \mathbb{P}^i(g(A) \in u) && (15)\\
&= \mathbb{P}^i(B \in u) && (16)\\
&= \mathbb{P}^i(g^{-1}(B) \in g^{-1}(u)) && (17)\\
&= \mathbb{P}^i(g^{-1}(B) \in w). && (18)
\end{aligned}$$

Form Remark 1, we know that g^{-1} is a diffeomorphims. As the choice of w was arbitrary, from Definitions 5 and 6, we have

$$B \sim A. \tag{19}$$

Claim: $\sim$ is transitive.

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

$$\mathbb{P}^i(g(A) \in u) = \mathbb{P}^i(B \in u). \tag{20}$$

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

$$\mathbb{P}^i(f(B) \in q) = \mathbb{P}^i(C \in q). \tag{21}$$

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

$$\begin{aligned}
\mathbb{P}^i((f \circ g)(A) \in q) &= \mathbb{P}^i(g(A) \in f^{-1}(q)) && (22)\\
&= \mathbb{P}^i(g(A) \in u) && (23)\\
&= \mathbb{P}^i(B \in u) && (24)\\
&= \mathbb{P}^i(f(B) \in f(u)) && (25)\\
&= \mathbb{P}^i(f(B) \in q) && (26)\\
&= \mathbb{P}^i(C \in q) && (27)
\end{aligned}$$

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

$$A \sim C. \tag{28}$$

■

Proof of Proposition 1

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

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

$$p^i_{(Y|X)}(y|x) = p^i_{(E|X)}(f(x,\cdot)^{-1}(y)|x) = p^i_E(f(x,\cdot)^{-1}(y)), \tag{29}$$

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

$$\tilde{Y}_x \overset{I}{=} g(E). \tag{30}$$

Note that $\tilde{Y}_x$ is a random variable which has the same distribution as Y given $X = x$. According to Definition 6,

$$\forall x \in \mathcal{X} : E \sim \tilde{Y}_x. \tag{31}$$

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

$$\forall a, b \in \mathcal{X} : \tilde{Y}_a \sim \tilde{Y}_b. \tag{32}$$

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 $\tilde{f}(x, \cdot) := g_x$ 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 domain 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 $\mathbb{P}^i$, as we intend to show

$$\mathbb{P}^i(\Psi_A \in s) = \mathbb{P}^i(\Psi_B \in s), \tag{33}$$

which results in $\Psi_A \overset{I}{=} \Psi_B$ (Definition 5).

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 6

$$\exists \text{ diffeomorphism } g : \mathcal{A} \to \mathcal{B} \text{ such that } g(A) \overset{I}{=} B \tag{34}$$

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

$$\begin{aligned}
\Phi_B(b) &= c.(p_B^1(b), p_B^2(b), \ldots, p_B^m(b)) && (35)\\
&= c.\frac{1}{|det(J_g(a))|}.(p_A^1(a), p_A^2(a), \ldots, p_A^m(a)) && (36)\\
&= c'.(p_A^1(a), p_A^2(a), \ldots, p_A^m(a)), && (37)
\end{aligned}$$

where (35) holds according to Definition 7 and (36) holds due to (4).

As we defined α, we have

$$\Phi_A(a) = c''.(p_A^1(a), p_A^2(a), \ldots, p_A^m(a)) \in s. \tag{38}$$

Note that we also have

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

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

$$g(\alpha) \subset \beta. \tag{40}$$

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

$$g^{-1}(\beta) \subset \alpha, \tag{41}$$

which results in

$$g(\alpha) = \beta. \tag{42}$$

Finally,

$$\begin{aligned}
\mathbb{P}^i(\Psi_A \in s) &= \mathbb{P}^i(A \in \alpha) && (43)\\
&= \mathbb{P}^i(g(A) \in g(\alpha)) && (44)\\
&= \mathbb{P}^i(g(A) \in \beta) && (45)\\
&= \mathbb{P}^i(B \in \beta) && (46)\\
&= \mathbb{P}^i(\Psi_B \in s), && (47)
\end{aligned}$$

where (46) 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 $\mathbb{P}^i \in \mathcal{M}$.

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

$$\begin{aligned}
\mathbb{P}^i(\Gamma_{X\to Y} \in s | X = a) &= \mathbb{P}^i(\Psi_{\tilde{Y}_X} \in s | X = a) && (48)\\
&= \mathbb{P}^i(\Psi_{\tilde{Y}_a} \in s) && (49)\\
&= \mathbb{P}^i(\Psi_{\tilde{Y}_b} \in s) && (50)\\
&= \mathbb{P}^i(\Psi_{\tilde{Y}_X} \in s | X = b) && (51)\\
&= \mathbb{P}^i(\Gamma_{X\to Y} \in s | X = b), && (52)
\end{aligned}$$

where (48) is from Definition 9 and (49) is from imposing the condition $X = a$. Moreover, (50) is from the assumption $\Psi_{\tilde{Y}_a} \overset{I}{=} \Psi_{\tilde{Y}_b}$ and Definition 5. 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 $\mathbb{P}^i$ and as the choice of e was arbitrary, this holds for every e.

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

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

$$\begin{aligned}
\mathbb{P}^i(\Psi_{\tilde{Y}_a} \in s) &= \mathbb{P}^i(\Psi_{\tilde{Y}_X} \in s | X = a) && (53)\\
&= \mathbb{P}^i(\Gamma_{X\to Y} \in s | X = a) && (54)\\
&= \mathbb{P}^i(\Gamma_{X\to Y} \in s | X = b) && (55)\\
&= \mathbb{P}^i(\Psi_{\tilde{Y}_b} \in s). && (56)
\end{aligned}$$

Due to arbitrary choice of s, according to Definition 5, we have

$$\Psi_{\tilde{Y}_a} \overset{I}{=} \Psi_{\tilde{Y}_b} \tag{57}$$

Proof of Theorem 1

In continuous case, we have Assumption 1, along with other technical assumptions mentioned in the Section 3. Proposition 1 states that under this set of assumptions, if X causes Y in our model, we have

$$\forall a, b \in \mathcal{X} : \tilde{Y}_a \sim \tilde{Y}_b. \tag{58}$$

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

$$\Psi_A \overset{I}{=} \Psi_B. \tag{59}$$

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

$$\forall a, b \in \mathcal{X} : \Psi_{\tilde{Y}_a} \overset{I}{=} \Psi_{\tilde{Y}_b}. \tag{60}$$

Proposition 3 states that from the condition (60), we can imply the following independence criterion in all domains (i.e., for every $\mathbb{P}^i \in \mathcal{M}$):

$$\Gamma_{X\to Y} \perp\!\!\!\perp X. \tag{61}$$

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

$$A \overset{d}{\sim} B \iff \exists \text{ bijective } g \text{ such that } g(A) \overset{I}{=} B. \tag{62}$$

Note that Definition 5 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 (36)), and use p_V^i 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.