Generalization of LiNGAM that allows confounding

Joe Suzuki and Tian-Le Yang

January 30, 2024

Abstract

LiNGAM determines the variable order from cause to effect using additive noise models, but it faces challenges with confounding. Previous methods maintained LiNGAM's fundamental structure while trying to identify and address variables affected by confounding. As a result, these methods required significant computational resources regardless of the presence of confounding, and they did not ensure the detection of all confounding types. In contrast, this paper enhances LiNGAM by introducing LiNGAM-MMI, a method that quantifies the magnitude of confounding using KL divergence and arranges the variables to minimize its impact. This method efficiently achieves a globally optimal variable order through the shortest path problem formulation. LiNGAM-MMI processes data as efficiently as traditional LiNGAM in scenarios without confounding while effectively addressing confounding situations. Our experimental results suggest that LiNGAM-MMI more accurately determines the correct variable order, both in the presence and absence of confounding.

1 Introduction

When multiple events occur, determining which is the cause and which is the effect is a problem we commonly encounter in daily life. This paper examines the challenge of finding the order from causes to effects when dealing with multiple random variables.

Similar to, yet distinct from, this type of causal inference is the problem of identifying the structure of a Bayesian Network (BN) [Pearl, 2009, Peter Spirtes, 1993, Bollen, 1989, Hyvärinen and Smith, 2013]. A BN represents the conditional independence between random variables using a Directed Acyclic Graph (DAG). In a BN, each vertex represents a variable, and directed edges are drawn from vertices corresponding to $X_1, \ldots, X_m$ to the vertex corresponding to X when the joint distribution is expressed as the product of conditional probabilities $P(X|X_1, \ldots, X_m)$. However, the structure of a BN can be represented differently, depending on how the joint distribution is factorized, and the direction of the arrows does not necessarily indicate the direction of causality. For instance, $P(X, Y) = P(X)P(Y|X) = P(X|Y)P(Y)$ can be written in such a way that it is unclear whether the correct representation is $X \to Y$ or $Y \to X$. We refer to different graph expression that shares the same distribution, such as $X \to Y$ and $Y \to X$ as *Markov equivalent* models.

Kano and Shimizu [2003] proposed a causal inference framework called the Additive Noise Model. Consider two variables X, Y, with means of zero that can be expressed as $Y = aX + e$ using some constant a and noise e. The Additive Noise Model posits that X and e are independent, with X being the cause and Y the effect. Some constants a' and noise e' may exist, such as $X = a'Y + e'$. However, both[1] $X \perp\!\!\!\perp e$ and $Y \perp\!\!\!\perp e'$ (i.e., the order is not identifiable) is equivalent to X and Y being Gaussian. For more details, refer to section 2.2 of this paper. Assuming a non-Gaussian distribution, a procedure was developed to determine which of X or Y is the cause and which is the effect. This approach is known as LiNGAM (Linear non-Gaussian Acyclic Model [Shimizu et al., 2006, 2011, Hyvärinen and Smith, 2013]).

However, the Additive Noise Model is considered to have a limited scope of application. It does not encompass cases where neither $X \perp\!\!\!\perp e$ nor $Y \perp\!\!\!\perp e'$ is true. This paper, along with existing studies in causal inference, refers to such cases as 'the presence of confounding.' As the number of variables p increases, the likelihood of encountering situations without confounding diminishes. Even when focusing specifically on LiNGAM, a significant amount of research in causal inference accommodates the presence of confounding. While existing studies will be discussed later, this paper offers a fundamentally different perspective.

Most of these studies (Entner and Hoyer [2010],Tashiro et al. [2014],Maeda and Shimizu [2020],Salehkaleybar et al. [2020],Wang and Drton [2023]) aim to identify variables affected by confounding, eliminate their influence, and then apply the traditional LiNGAM model, which does not account for confounding. As a result, they presuppose a specific type of confounding and require computation times that grow exponentially with the

[1] We write $X \perp\!\!\!\perp Y$ to denote that X and Y are independent

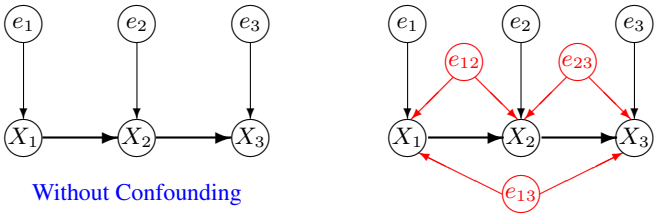


Figure 1: Confounders in red of the right figure affect more than one variable. The previous methods can deal with only non-consecutive variables such as X_1 and X_3.

number of variables p. However, there has been scant discussion regarding the challenges of causal inference in the presence of confounding, specifically its computational intensity, which often renders it impractical.

This paper explores an approach to finding the order of variables without identifying those affected by confounding, by extending LiNGAM. It specifically aims to quantify the magnitude of confounding among p variables and to determine the sequence of these variables that minimizes this confounding. For each of the $p!$ possible orders $X_1 \to \ldots \to X_p$, a corresponding series of noise terms $e_1 = X_1$, $e_i = X_i - \sum_{j=1}^{i-1} \beta_{i,j} X_j$ $(i = 2, \ldots, p)$ is identified, where $\beta_{i,j}$ are constants. Traditional LiNGAM presupposes that in one specific order of variables, $e_1, \ldots, e_p$ are independent, and this configuration represents the true model. This paper proposes to quantify the degree of confounding using the Kullback-Leibler (KL) divergence between the joint distribution of these noise terms $P(e_1, \ldots, e_p)$ and the product of their marginal probabilities $P(e_1) \cdots P(e_p)$. The assumption is that the order of variables $X_1 \to \ldots \to X_p$ that minimizes the KL divergence represents the true. This approach encompasses the existing Additive Noise Model and LiNGAM as particular instances.

Regarding computation time, finding the order that minimizes confounding is addressed through the shortest path problem, as the KL divergence representing confounding can be expressed as the sum of $p - 1$ mutual information quantities. Despite this, the worst-case computational complexity remains $O(p!)$. Nevertheless, when confounding is minimal, the computational demand decreases substantially, and in cases with no confounding, the procedure completes in a timeframe comparable to that of the standard LiNGAM.

When applying LiNGAM, the order is sequentially determined from the causal variables to the resultant ones. This method works well without confounding, but it is not optimal when noise is not independent, i.e., when confounding is present. For instance, consider the variables X, Y, Z; even if X is independent of the residuals Y_X, Z_X (the residuals of Y, Z after removing the influence of X), it is possible that both Y_X, Z_{XY} and Z_X, Y_{XZ} are far from independent, where Y_{XZ} and Z_{XY} are the residuals of Y and Z after removing the influence of X, Z, and X, Y, respectively. In such cases, a greedy search approach can fail. When conducting a causal search with a finite sample, it becomes impossible to reliably distinguish whether confounding is present compared to the true distribution. Therefore, issues related to greedy search can always potentially arise. However, a search based on the shortest path problem enables a more global approach, thereby avoiding these issues.

This method's versatility is further enhanced because it does not require prior knowledge of the confounding characteristics or assumptions. On the other hand, for example, the existing procedures preclude the case that confounders affect consecutive variables, which is fatal in real applications (Figure 1).

While existing methods demand exponential computation time regardless of the data, the approach proposed in this paper can be executed in a significantly shorter duration for most cases. Additionally, unlike current methods, it obviates the need for the BN structure learning procedure, simplifying implementation.

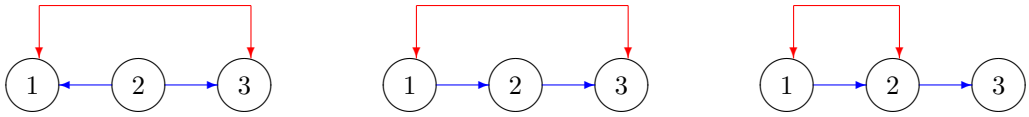
In this paper, we refer our proposed procedure to the *LiNGAM-MMI* because it minimizes the total sum of MI along the path. This paper provides the following contributions:

1. It quantifies confounding by using the KL divergence between the distributions of the noise and defines the true order as the one that minimizes it.

2. The optimal order is determined by formulating the problem as a shortest path problem, with the distance measured by the KL divergence. This global search approach requires relatively small computation, especially when the confounding is small.

3. Experiments show that the proposed procedure with the copula mutual information outperforms DirectLiNGAM [Shimizu et al., 2006] with the Hilbert-Schmidt Independence Criterion (HSIC) [Gretton et al., 2007] even in the absence of confounding due to the global nature of the search and becomes more pronounced when p (the number of variables) is large.

The organization of this paper is as follows: Section 2 provides the necessary background information for understanding the results and places them in the context of existing work. Section 3 presents the main results



(a) Ancestral (b) Non-Ancestral, Bow-free (c) Non-Bow-free

Figure 2: Ancestral and bow-free graphs. (a) The vertices 1 and 3 are siblings, but none of them are ancestors of the other. (b) The vertices 1 and 3 are siblings, but the former is an ancestor of the latter. (c) The vertices 1 and 2 consist of a bow.

of this study, with a particular focus on the underlying principles and methodologies. Section 4 illustrates these concepts with an example and experiments, assessing the effectiveness of LiNGAM-MMI. Section 5 concludes the paper with a summary of the findings and a discussion of potential directions for future research.

1.1 Related Work

In a Directed Acyclic Graph (DAG), connecting the directed edges $u \to v$ and $v \to w$ creates a directed path from u to w. Within this framework, u is considered an ancestor of w. This interpretation holds irrespective of the number of directed edges involved; the origin of a directed path is consistently seen as the ancestor of its endpoint. Moving on to mixed graphs, these incorporate edges with arrows in both directions, referred to as bidirectional edges such as $u \leftrightarrow v$, into a DAG. In such graphs, the simultaneous presence of directed edges and bidirectional edges is permissible (these are known as bows). A graph is defined as an ancestral graph if it does not include any pair of vertices that are simultaneously ancestors and siblings of each other (Figure 2). In existing research, ancestral graphs, or their variants such as bow-free graphs, often represent confounding between two variables by using bidirected edges, like $u \leftrightarrow v$, between vertices.

Adding bidirected edges allows every ancestral graph G to be transformed into a maximal ancestral graph (MAG) while preserving the conditional independence relations in G. However, such an MAG is not unique, and another Markov equivalent exists. They can be compactly represented by a partial ancestral graph (PAG) [Ali et al., 2009]. Spirtes et al. [1995] proposed the Fast Causal Inference algorithm (FCI) to estimate the PAG corresponding to the underlying causal graph. Zhang [2008] added additional orientation rules such that the output of FCI is complete. Colombo et al. [2012], Claassen et al. [2013] and Chen et al. [2021] followed in this direction. However, only an equivalence class of graphs a PAG represents can be discovered.

In contrast, Shimizu et al. [2006] show that when the true model is a recursive linear SEM with non-Gaussian errors, the exact graph - not just an equivalence class - can be identified from observational data using independent component analysis (ICA). Instead of ICA, the subsequent DirectLiNGAM [Shimizu et al., 2011] and Pairwise LiNGAM [Hyvärinen and Smith, 2013] methods use an iterative procedure to estimate a causal ordering.

Hoyer et al. [2008] consider the setting where a LiNGAM model generates the data, but some variables are unobserved. Using overcomplete ICA, they show that the canonical DAG can be identified when all parent-child pairs in the observed set are unconfounded. Shimizu and Bollen [2014] point out that "current versions of the overcomplete ICA algorithms are unreliable since they often suffer from local optima." To avoid using overcomplete ICA and improve practical performance, Entner and Hoyer [2010] and Tashiro et al. [2014] both propose procedures that test subsets of the observed variables and seek to identify as many pairwise ancestral relationships as possible. For more details, refer to section 2.5 of this paper.

Maeda and Shimizu [2020] proposes the Repetitive Causal Discovery (RCD) method for discovering mixed graphs. In contrast to the approach by Tashiro et al. [2014], RCD iteratively utilizes previously discovered structures to inform subsequent steps. Similar to Hoyer et al. [2008], Salehkaleybar et al. [2020] employ overcomplete ICA, which crucially requires all confounding to be linear. Wang and Drton [2023] introduce a method for identifying bow-free confounding, diverging from existing works that have assumed confounding to be ancestral. However, all these studies, including Wang and Drton [2023], deal only with non-consecutive confounding, as shown in Figure 1.

The top author of this paper has published the LiNGAM-MMI for the binary case [Suzuki and Inaoka, 2022] while this paper deals with the structure equation models [Kano and Shimizu, 2003].

2 Preliminaries

This section provides a background on covariance, independence, additive noise models, LiNGAM (Linear Gaussian Models), HSIC (Hilbert-Schmidt Information Criterion), and confounding factors.

2.1 Covariance and Independence

Let X, Y, e be zero mean random variables[2] related by

$$Y = aX + e \tag{1}$$

with $a \in \mathbb{R}$. We determine the constant a so that the covariance of X, e is zero:

$$\mathrm{Cov}[X, e] = \mathrm{Cov}[X, Y - aX] = 0 \,, \tag{2}$$

which means

$$a = \frac{\mathrm{Cov}[X, Y]}{V[X]} \,, \tag{3}$$

where $\mathrm{Cov}[\cdot, \cdot]$ and $V[\cdot]$ are the covariance and variance operations.

Let $N(\mu, \sigma^2)$ be the Gaussian distribution with mean μ and variance σ^2. We know that, in general, the converse of the implication

$$Z \perp\!\!\!\perp W \Longrightarrow \mathrm{Cov}[Z, W] = 0 \tag{4}$$

is not true. For example, suppose $Z \sim N(0, 1)$ and $W = Z^2$. They are not independent but

$$\mathrm{Cov}[Z, W] = \mathrm{Cov}[Z, Z^2] = E[(Z - 0)(Z^2 - 1)] = E[Z^3] - E[Z] = 0$$

because $E[Z] = E[Z^3] = 0$ and $E[Z^2] = 1$, where $E[\cdot]$ is the expectation operation. In this sense, (2) does not mean $X \perp\!\!\!\perp e$.

On the other hand, suppose $U, V \sim N(0, 1)$ with $U \perp\!\!\!\perp V$. Variables Z, W are said to be jointly Gaussian if there exists a matrix $A \in \mathbb{R}^{2\times 2}$ such that

$$\begin{bmatrix} Z \\ W \end{bmatrix} = A \begin{bmatrix} U \\ V \end{bmatrix} + \begin{bmatrix} E[Z] \\ E[W] \end{bmatrix} \,.$$

If Z, W are jointly Gaussian, then they are Gaussian. Moreover, if Z, W are jointly Gaussian, the converse ($\Longleftarrow$) of (4) holds. Just because Z, W are Gaussian does not imply the converse of (4). In fact, let $Z \in N(0, 1)$ and $R \in \{\pm 1\}$ equiprobable, and assume $Z \perp\!\!\!\perp R$. Then, Z and $W = ZR$ are not independent although $E[R] = E[Z] = E[ZR] = 0$ and

$$\mathrm{Cov}(Z, W) = E[(Z - 0)(ZR - 0)] = E[Z^2 R] = E[Z^2]E[R] = 0 \,.$$

In this paper, we refer to Z and W as 'Gaussian' only when they are jointly Gaussian, provided this does not lead to confusion. Thus, if X, e are Gaussian in (1), we have

$$\mathrm{Cov}[X, e] = 0 \Longleftrightarrow X \perp\!\!\!\perp e \,. \tag{5}$$

2.2 Additive Noise Model and LiNGAM

Kano and Shimizu [2003] considered a causal model in which cause X and effect Y are related by

$$\begin{cases} X = e_1 \\ Y = aX + e_2 \end{cases} \,, \tag{6}$$

where a is given by (3). In particular, they assumed that e_1, e_2 should be independent. Then, we write $X \to Y$. Similarly, we can construct the opposite model in which they are related by

$$\begin{cases} Y = e_1' \\ X = a'Y + e_2' \end{cases} \tag{7}$$

[2]Hereafter, we refer to the random variables simply as 'variables' when no confusion arises.

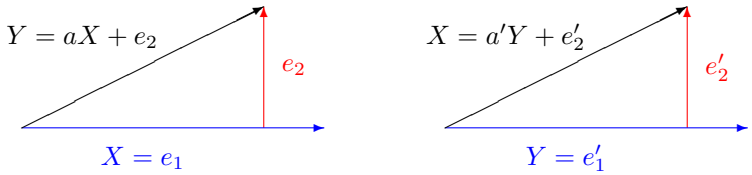


Figure 3: Additive Noise Model: Either $e_1 \perp\!\!\!\perp e_2$ ($X \to Y$) or $e'_1 \perp\!\!\!\perp e'_2$ ($Y \to X$) must hold, but not both.

with $e'_1 \perp\!\!\!\perp e'_2$, and write $Y \to X$. They assumed that exactly one of (6) with $e_1 \perp\!\!\!\perp e_2$ and (7) with $e'_1 \perp\!\!\!\perp e'_2$ happens, and identifies which of X, Y are the cause and effect, depending on which of (6) and (7) is correct. Note that a and

$$a' = \frac{\mathrm{Cov}[X, Y]}{V[Y]}$$

are determined so that

$$\begin{cases} \mathrm{Cov}[e_1, e_2] = 0 \\ \mathrm{Cov}[e'_1, e'_2] = 0 \end{cases} . \tag{8}$$

Suppose $a = 0$, which is equivalent to $a' = 0$. Then, X, Y are independent, and we cannot identify the order. Thus, we remove such a case from the beginning. We claim the order can be identified under $aa' \neq 0$ if and only if either of $e_1 e_2$ is not Gaussian.

In order to show the claim, suppose that e_1, e_2 are Gaussian. Then, from (6) and (7), we observe that X, Y, e'_1, e'_2 are Gaussian. Moreover, from (5), we have

$$\begin{cases} \mathrm{Cov}[e_1, e_2] = 0 \Longleftrightarrow e_1 \perp\!\!\!\perp e_2 \\ \mathrm{Cov}[e'_1, e'_2] = 0 \Longleftrightarrow e'_1 \perp\!\!\!\perp e'_2 \end{cases} . \tag{9}$$

From (8)(9), we conclude that both of $e_1 \perp\!\!\!\perp e_2$ and $e'_1 \perp\!\!\!\perp e'_2$ occur.

On the other hand, suppose that $e_1 \perp\!\!\!\perp e_2$ and $e'_1 \perp\!\!\!\perp e'_2$ occur simultaneously. Noting that (6) and (7) imply

$$\begin{cases} e'_1 = ae_1 + e_2 \\ e'_2 = (1 - aa')e_1 - a'e_2 \end{cases} , \tag{10}$$

we consider applying the following proposition.

Proposition 1 (Darmois [1953], Skitovitch [1953]) Let z, w, u, v be variables related by

$$\begin{cases} z = pu + qv \\ w = ru + sv \end{cases}$$

with $p, q, r, s \in \mathbb{R}$, and suppose $u \perp\!\!\!\perp v$ and $z \perp\!\!\!\perp w$. Then, if $pr \neq 0$, u is Gaussian, and if $qs \neq 0$, v is Gaussian.

Suppose $1 = aa' = \dfrac{\mathrm{Cov}[X, Y]^2}{V[X]V[Y]}$. Then, the relation between X, Y is deterministic, which is excluded in our discussion. Thus, under $a, a' \neq 0$, we have $a(1 - aa'), -a' \neq 0$ in (10). From Proposition 1, we conclude that $e_1.e_2$ are Gaussian. The claim can be summarized as follows:

Proposition 2 (Shimizu et al. [2006]) Under $a, a' \neq 0$, the order can be identified if and only if either of e_1, e_2 is non-Gaussian.

Shimizu et al. [2006] considered procedures (*LiNGAM*, linear non-Gaussian models) that identify the causal order based on Proposition 2.

2.3 LiNGAM from data

Suppose that we observe i.i.d. (independent and identically distributed) data $x^n = (x_1, \ldots, x_n), y^n = (y_1, \ldots, y_n) \in \mathbb{R}^n$ for variables X, Y of sample size n. Then, We perform statistical independence testing to evaluate whether

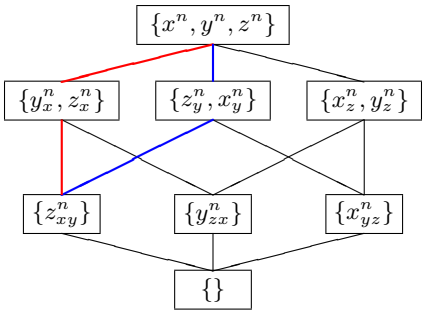


Figure 4: In order to obtain the residue $z_{xy}^n (= z_{yx}^n)$, compute $\{y_x^n, z_x^n\}$ first and then obtain $\{z_{xy}^n\}$ via (11), as shown in red, or compute $\{z_y^n, x_y^n\}$ first and then obtain $\{z_{yx}^n\}$ via (12), as shown in blue.

$x^n \perp\!\!\!\perp y_x^n$ or $y^n \perp\!\!\!\perp x_y^n$, examining which of $e_1 \perp\!\!\!\perp e_2$ and $e_1' \perp\!\!\!\perp e_2'$ is more likely, where

$$y_x^n := y^n - \frac{c(x^n, y^n)}{v(x^n)} x^n$$
$$x_y^n := x^n - \frac{c(x^n, y^n)}{v(y^n)} y^n$$

with $v(\cdot)$ and $c(\cdot, \cdot)$ the variance and covariance based on the samples x^n, y^n. For each of the two possibilities, we obtain the causal order as follows:

$$\begin{cases} x^n \perp\!\!\!\perp y_x^n & \Longrightarrow \quad X \to Y \\ y^n \perp\!\!\!\perp x_y^n & \Longrightarrow \quad Y \to X \end{cases}$$

Next, suppose that we observe i.i.d. data $z^n = (z_1, \ldots, z_n) \in \mathbb{R}^n$ for variable Z as well as the x^n, y^n. Then, we compare

$$x^n \perp\!\!\!\perp \{y_x^n, z_x^n\}\ ,\ y^n \perp\!\!\!\perp \{z_y^n, x_y^n\}\ ,\ z^n \perp\!\!\!\perp \{x_z^n, y_z^n\}\ ,$$

where the quantities z_x^n, z_y^n, x_z^n, y_z^n are defined similarly to x_y^n and y_x^n.

Then, suppose that $x^n \perp\!\!\!\perp \{y_x^n, z_x^n\}$ is the most likely among the three possibilities. We define the quantity

$$\begin{aligned} y_{xz}^n &:= y_x^n - \frac{c(y_x^n, z_x^n)}{v(z_x^n)} z_x^n \\ z_{xy}^n &:= z_x^n - \frac{c(y_x^n, z_x^n)}{v(y_x^n)} y_x^n \end{aligned} \tag{11}$$

to compare $y_x^n \perp\!\!\!\perp z_{xy}^n$ and $z_x^n \perp\!\!\!\perp y_{xz}^n$. If $y_x^n \perp\!\!\!\perp z_{xy}^n$ as well as $x^n \perp\!\!\!\perp \{y_x^n, z_x^n\}$ are the most likely, then we conclude that the order $X \to Y \to Z$ is the most likely among the six. Thus, we obtain the residue sequence (x^n, y_x^n, z_{xy}^n) by reducing the effects of the upper variables from the original samples x^n, y^n, z^n.

Note that we can show the equality $z_{xy}^n = z_{yx}^n$ for

$$z_{yx}^n := z_y^n - \frac{c(z_y^n, x_y^n)}{v(x_y^n)} x_y^n \tag{12}$$

(see Appendix A). We obtain the same value in two ways (Figure 4). Similarly, we have $x_{yz}^n = x_{zy}^n$ and $y_{zx}^n = y_{xz}$. In general, the order of suffices in the residues does not matter, although, in this paper, we do not provide its derivation.

For each of the six possibilities, we obtain the causal order as follows:

$$\begin{cases} x^n \perp\!\!\!\perp \{y_x^n, z_x^n\} & \begin{cases} y_x^n \perp\!\!\!\perp z_{xy}^n & \Longrightarrow \quad X \to Y \to Z \\ z_x^n \perp\!\!\!\perp y_{xz}^n & \Longrightarrow \quad X \to Z \to Y \end{cases} \\ y^n \perp\!\!\!\perp \{z_y^n, x_y^n\} & \begin{cases} z_y^n \perp\!\!\!\perp x_{yz}^n & \Longrightarrow \quad Y \to Z \to X \\ x_y^n \perp\!\!\!\perp z_{yx}^n & \Longrightarrow \quad X \to Z \to Y \end{cases} \\ z^n \perp\!\!\!\perp \{x_z^n, y_z^n\} & \begin{cases} x_z^n \perp\!\!\!\perp y_{zx}^n & \Longrightarrow \quad Z \to X \to Y \\ y_z^n \perp\!\!\!\perp x_{zy}^n & \Longrightarrow \quad Z \to Y \to X \end{cases} \end{cases}$$

The same procedure can be applied to any $p \geq 2$ variables rather than $p = 2, 3$.

2.4 Statistical Testing using HSIC

Let $\mathcal{X}$ be a set, and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive definite kernel. It is known [Suzuki, 2022] that there exists a unique reproducing kernel Hilbert space (RKHS) H, the closure $\overline{\{k(x,\cdot)\}_{x\in\mathcal{X}}}$ of the linear space generated by $\{k(x,\cdot)\}_{x\in\mathcal{X}}$. Let $(k_{\mathcal{X}}, H_{\mathcal{X}})$ and $(k_{\mathcal{Y}}, H_{\mathcal{Y}})$ be pairs of positive definite kernel and RKHS for sets $\mathcal{X}$ and $\mathcal{Y}$. We execute the test of independence between $k_{\mathcal{X}}(X,\cdot) \in H_{\mathcal{X}}$ and $k_{\mathcal{Y}}(Y,\cdot) \in H_{\mathcal{Y}}$ rather than between $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$, and define the quantity called Hilbert-Schmidt information criterion (HSIC) as follows:

$$\mathrm{HSIC}(X,Y) := \|E_{XY}[k_{\mathcal{X}}(X,\cdot)k_{\mathcal{Y}}(Y,\cdot)] - E_X[k_{\mathcal{X}}(X,\cdot)]E_Y[k_{\mathcal{Y}}(Y,\cdot)]\|_H^2 ,$$

where $E_X[\cdot]$, $E_Y[\cdot]$, and $E_{XY}[\cdot]$ denote the expectations with respect to X, Y, and (X,Y), respectively, and $\|\cdot\|_H$ is the norm defined in the tensored Hilbert space $H := H_{\mathcal{X}} \otimes H_{\mathcal{Y}}$. It is known that

$$HSIC(X,Y) = 0 \iff X \perp\!\!\!\perp Y \tag{13}$$

if the kernels are characteristic [Suzuki, 2022]. However, the $HSIC(X,Y)$ value is not known, and we infer whether $HSIC(X,Y) = 0$ or not by examining its estimate $HSIC_n(x^n, y^n)$ given realizations $\{(x_i, y_i)\}_{i=1}^n$ of (X,Y) with $x^n = (x_1, \ldots, x_n)$, $y^n = (y_1, \ldots, n_n)$.

In the test of independence, we see whether statistics $T := \mathrm{HSIC}_n(x^n, y^n)$ follows the distribution $f_{X\perp\!\!\!\perp Y}(t)$ under the null hypothesis $X \perp\!\!\!\perp Y$. Given a significance level $\alpha > 0$, if T is significantly large, i.e., $T > T_\alpha$, where

$$\alpha = \int_{T_\alpha}^{\infty} f_{X\perp\!\!\!\perp Y}(t)dt ,$$

we reject $X \perp\!\!\!\perp Y$. HSIC is well-known for its strong power in detecting non-independence and is often used in the LiNGAM procedure.

In the actual LiNGAM applications, to minimize computational costs, we compare $T = HSIC_n(x^n, y_x^n)$ and $T' = HSIC_n(y^n, x_y^n)$ rather than the p-values $\int_T^\infty f_{e_1\perp\!\!\!\perp e_2}(t)\,dt$ and $\int_{T'}^\infty f_{e_1'\perp\!\!\!\perp e_2'}(t)\,dt$ to choose between[3] $e_1 \perp\!\!\!\perp e_2$ and $e_1' \perp\!\!\!\perp e_2'$.

2.5 Confounding

We say that *confounding* exists for X, Y if neither $e_1 \perp\!\!\!\perp e_2$ nor $e_1' \perp\!\!\!\perp e_2'$ holds. Even if confounding exists, we may decide that $e_1 \perp\!\!\!\perp e_2$ is more likely than $e_1' \perp\!\!\!\perp e_2'$ if $HSIC_n(x^n, y_x^n)$ is smaller than $HSIC_n(y^n, x_y^n)$. However, this paper's main issue is identifying the order when more than two variables exist. For example, if the noises e_1, e_2, e_3 in

$$\begin{cases} X = e_1 \\ Y = aX + e_2 \\ Z = bX + cY + e_3 \end{cases}$$

are independent for some $a, b, c \in \mathbb{R}$, we identify the order as $X \to Y \to Z$. However, what if no independent noises exist for any order of X, Y, Z?

One might claim that the same strategy illustrated in Section 2.3 can be applied, such as comparing

$$HSIC_n(x^n, \{y_x^n, z_x^n\}), HSIC_n(y^n, \{z_y^n, x_y^n\}), HSIC_n(z^n, \{x_z^n, y_z^n\})$$

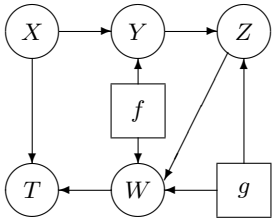
and, if the first value is the smallest, further compare $HSIC_n(y_x^n, z_{xy}^n)$ and $HSIC_n(z_x^n, y_{zx}^n)$, etc. But, in that case, what merits can be gained by following such a strategy? For example, what if both of $HSIC_n(y_x^n, z_{xy}^n)$ and $HSIC_n(z_x^n, y_{zx}^n)$ have large values, which suggests that both of $y_x^n \perp\!\!\!\perp z_{xy}^n$ and $z_x^n \perp\!\!\!\perp y_{zx}^n$ are unlikely? Then, we would admit that the decision either $X \to Y \to Z$ or $X \to Z \to Y$ would be wrong.

When a confounder exists, estimating the variable order becomes more challenging. Most of the previous works [Entner, 2013, Tashiro et al., 2014] obtain the order for each maximal subset of variables that are not affected by any confounder and estimate the order among the whole variables by combining the orders among variables that are not affected by any confounder.

For the details of LvLiNGAM and ParceLiNGAM, see Appendix B.

These methods require us either to know a priori that the same type of confounding always exists or to spend exponential time with p (the number of variables) *even when no confounder exists*.

[3] Since $f_{e_1\perp\!\!\!\perp e_2}$ and $f_{e_1'\perp\!\!\!\perp e_2'}$ are different, it is not appropriate to compare $\widehat{HSIC}(e_1, e_2)$ and $\widehat{HSIC}(e_1', e_2')$ for determining which pair is more likely to be independent.



LvLiNGAM

	X	Y	Z	W
Y	◯			
Z	◯	◯		
W	◯	×	×	
T	◯	×	×	◯

ParceLiNGAM

	X	Y	Z	W
Y	◯			
Z	◯	◯		
W	◯	×	×	
T	◯	◯	◯	◯

Figure 5: Suppose we have five variables X, Y, Z, W, T are related by $Y = aX + f$, $Z = bY + g$, $W = cZ + f + g$, and $T = cX + dW$ for some $a, b, c, d \in \mathbb{R}$, and that the confounder f and g affect Y, W and Z, W, respectively. The table on the right shows what pair of variables the order can be recovered for LvLiNGAM and ParceLiNGAM. For example, in LvLiNGAM, the order between Y, Z can be recovered (denoted as "◯") even if Y is affected by f.

3 Extended LiNGAM

In this section, we propose to quantify the amount of confounding in terms of the noises $e_1, \ldots, e_p$, and extend the LiNGAM in the framework.

3.1 Quantification of Confounding

We define the amount of confounding by

$$K(e_1, \ldots, e_p) := E[\log \frac{P(e_1, \ldots, e_p)}{P(e_1) \cdots P(e_p)}] \tag{14}$$

when the variables $X_1, \ldots, X_p$ are related by

$$\begin{cases} X_1 &= e_1 \\ X_2 &= b_{2,1} X_1 + e_2 \\ \vdots & \vdots \qquad\qquad \vdots \\ X_p &= b_{p,1} X_1 + \cdots + b_{p,p-1} X_{p-1} + e_p \end{cases}$$

for some $b_{i,j} \in \mathbb{R}$, $i = 2, \ldots, p-1$, $j = 1, \ldots, i-1$. We interpret confounding as the divergence from the independence among the noises $e_1, \ldots . e_p$,

If there is no confunding, we have $P(e_1, \ldots, e_p) = P(e_1) \cdots P(e_p)$ and $K(e_1, \ldots, e_p) = 0$, which the original LiNGAM assumes for some order among $X_1, \ldots, X_p$.

We define the mutual information (MI) between X, Y

$$I(X, Y) := E[\log \frac{P(X, Y)}{P(X)P(Y)}] \,,$$

which takes non-negative values and satisfies

$$I(X, Y) = 0 \Longleftrightarrow X \perp\!\!\!\perp Y \,, \tag{15}$$

which is similar to (13). Then, the confounding can be expressed by the sum of the mutual information values as below:

$$\begin{aligned} & K(e_1, \ldots, e_p) \\ = \; & E[\log \frac{P(e_1, \ldots, e_p)}{P(e_1)P(e_2, \ldots, e_p)}] + E[\log \frac{P(e_2, \ldots, e_p)}{P(e_2) \cdots P(e_3, \ldots, e_p)}] + \cdots + E[\log \frac{P(e_{p-1}, e_p)}{P(e_{p-1})P(e_p)}] \\ = \; & I(e_1, \{e_2, \ldots, e_p\}) + I(e_2, \{e_3, \ldots, e_p\}) + \cdots + I(e_{p-1}, e_p) \end{aligned} \tag{16}$$

Moreover, since mutual information takes non-negative values, we have the equivalences

$$\begin{aligned} \text{no confounding} \quad & \Longleftrightarrow \quad K(e_1, \ldots, e_p) = 0 \\ & \Longleftrightarrow \quad I(e_1, \{e_2, \ldots, e_p\}) \cdots = I(e_{p-1}, e_p) = 0 \\ & \Longleftrightarrow \quad e_1, \ldots, e_p \text{ are independent} \end{aligned}$$

$$
\begin{aligned}
K_n(x^n, y^n_x, z^n_{xy}) &= I_n(x^n, \{y^n_x, z^n_x\}) + I_n(y^n_x, z^n_{xy}) \\
K_n(x^n, z^n_x, y^n_{xz}) &= I_n(x^n, \{y^n_x, z^n_x\}) + I_n(z^n_x, y^n_{xz}) \\
K_n(y^n, z^n_y, x^n_{yz}) &= I_n(y^n, \{z^n_y, x^n_y\}) + I_n(z^n_y, x^n_{yz}) \\
K_n(y^n, z^n_x, y^n_{xz}) &= I_n(y^n, \{z^n_y, x^n_y\}) + I_n(x^n_y, z^n_{xy}) \\
K_n(z^n, x^n_z, y^n_{zx}) &= I_n(z^n, \{x^n_z, y^n_z\}) + I_n(x^n_z, y^n_{zx}) \\
K_n(z^n, z^n_x, y^n_{xz}) &= I_n(z^n, \{x^n_z, y^n_z\}) + I_n(y^n_z, x^n_{yz})
\end{aligned}
$$

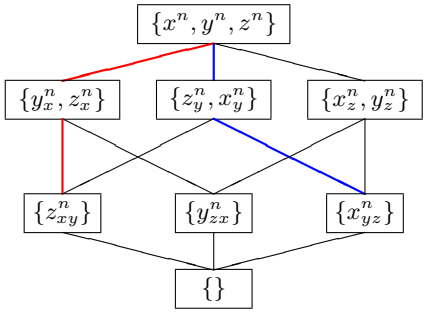


Figure 6: There are six paths corresponding to the six orders. We compare the sum of the distances from the top $\{x^n, y^n, z^n\}$ to the bottom $\{\}$ according to each path (order).

One might wonder why we do not minimize

$$HSIC(e_1, \{e_2, \ldots, e_p\}) + HSIC(e_2, \{e_3, \ldots, e_p\}) + \cdots + HSIC(e_{p-1}, e_p)$$

instead of minimizing (16). However, each term $HSIC(e_i, \{e_{i+1}, \ldots, e_p\})$ has a different deviation and the sum over $i = 1, \ldots, p-1$ does not provide any information about confounding if $p \geq 3$.

The additive noise model assumes that exactly one of $K(e_1, \ldots, e_p)$ is zero among the $p!$ orders. The extended framework assumes that exactly one order of p variables should minimize $K(e_1, \ldots, e_p)$ rather than diminish it.

We compare the identifiability conditions of the original and extended LiNGAM. The original LiNGAM requires that $e_1 \perp\!\!\!\perp e_2$ and $e'_1 \perp\!\!\!\perp e'_2$ should not happen at the same time. More precisely, it assumes the restrictive condition

$$I(e_1, e_2) = 0\,,\ I(e'_1, e'_2) > 0 \quad \text{or} \quad I(e_1, e_2) > 0\,,\ I(e'_1, e'_2) = 0\,.$$

Note that $I(e_1, e_2) = I(e'_1, e'_2) = 0$ is equivalent to that both of e_1, e_2 are Gaussian (Proposition 2), In this case, order identification is impossible for the original and extended LiNGAM. On the other hand, the extended LiNGAM only requires

$$I(e_1, e_2) \neq I(e'_1, e'_2)\,.$$

For the p variable case, the original LiNGAM requires $K(e_1, \ldots, e_p) = 0$ for exactly one order while the extended requires no tie-breaking occurs for the minimization of $K(e_1, \ldots, e_p)$, which does not seeme to be any constraint.

3.2 Compution of the Order that minimizes the confounding

This section considers minimizing MI values' sum (16). To this end, we prepare an ordered graph to obtain the shortest path. Given x^n, y^n, z^n, we can compare the six path as in Figure 6 for variables X, Y, Z ($p = 3$). For example, if we compare orders $X \to Y \to Z$ and $Y \to Z \to X$, we compute the sums over the paths in red and blue. Because the residues such as x^n, y^n_z are data rather than variables, we denote the estimated mutual information $I(\cdot, \cdot)$ and Kullback-Leibler (KL) divergence $K(\cdots)$ as $I_n(\cdot, \cdot)$ and $K_n(\cdots)$, respectively. Suppose we have DATA=$\{x^n, y^n, z^n\}$ as input. Then, we can compute the residues as in Figure 6 and an MI estimate value for each of the twelve edges (we assume $I_n(\{x^n_{yz}\}, \{\}) = I_n(\{y^n_{zx}\}, \{\}) = I_n(\{z^n_{xy}\}, \{\}) = 0$).

We regard the MI estimates as distances. Then, for each node v, we can compute the length $d(v)$ of the path from the top $\{X, Y, Z\}$ to v and the sum of the distances of the edges along the path. If multiple paths exist to a node, we choose the shortest path and store it in the node. For example, for the path $\{x^n, y^n, z^n\} \to \{y^n_x, z^n_x\} \to \{z^n_{xy}\} \to \{\}$, the sum of the distances is

$$I_n(x^n, \{y^n_x, z^n_x\}) + I_n(y^n_x, z^n_{xy}) + 0 = I_n(x^n, \{y^n_x, z^n_{xy}\}) + I_n(y^n_x, z^n_{xy}) + 0 = K_n(x^n, y^n_x, z^n_{xy})\,,$$

which is the estimated KL divergence of e_1, e_2, e_3 such that $X = e_1, Y = aX + e_2, Z = bX + cY + e_3$ for some constants a, b, c. We aim to find the shortest path from the top $\{x^n, y^n, z^n\}$ to the bottom $\{\}$.

In Figure 7, we first compute the lengths of the edges from the top $\{X, Y, Z\}$ to $\{Y, Z\}, \{Z, X\}, \{X, Y\}$:

$$
\begin{aligned}
d(\{Y, Z\}) &:= I_n(x^n, \{y^n_x, z^n_x\})\,, \\
d(\{Z, X\}) &:= I_n(y^n, \{z^n_y, x^n_y\})\,, \text{ and} \\
d(\{X, Y\}) &:= I_n(z^n, \{x^n_z, y^n_z\})\,.
\end{aligned}
$$

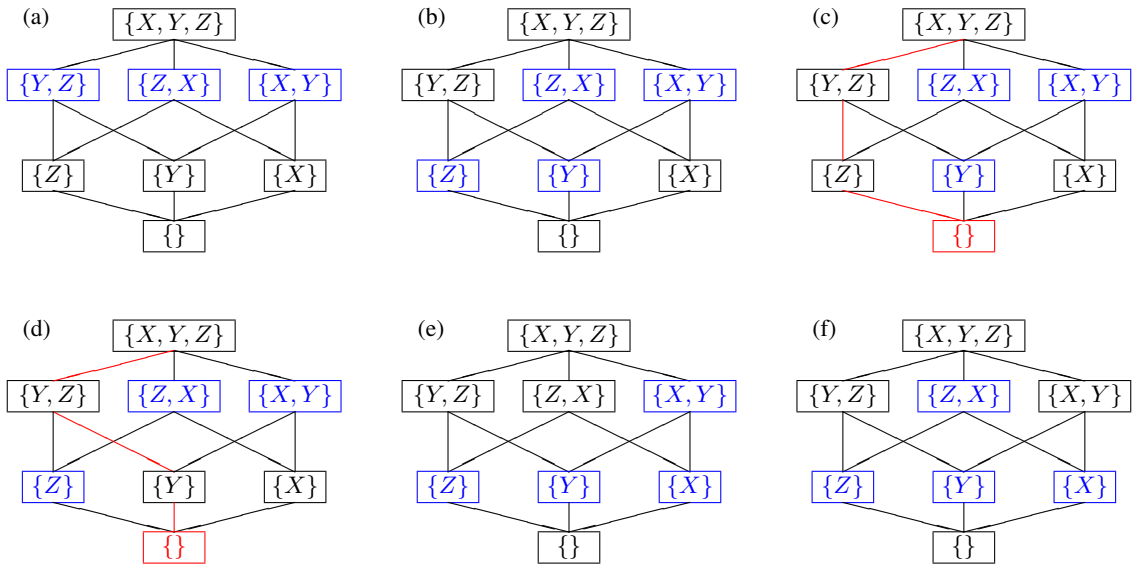


Figure 7: The ordered graph consists of the subsets of V, where the blue rectangles and red paths are the opened nodes and solutions, respectively.

We close the top node $\{X, Y, Z\}$ and open $\{Y, Z\}, \{Z, X\}, \{X, Y\}$ (Figure 7(a)). Suppose that $d(\{Y, Z\})$ is the smallest among the three nodes. Then, we compute $I_n(y_x^n, z_{xy}^n)$ and $I_n(z_x^n, y_{zx}^n)$ and obtain

$$d(\{Z\}) := d(\{Y, Z\}) + I_n(y_x^n, z_{xy}^n) \tag{17}$$

and $d(\{Y\}) := d(\{Y, Z\}) + I_n(z_x^n, y_{zx}^n)$, respectively. We close $\{Y, Z\}$ and open $\{Z\}$ and $\{Y\}$ (Figure 7 (b)).

If $d(\{Z\})$ is the smallest in Figure 7 (b), then $X \to Y \to Z$ is the shortest path (Figure 7 (c)); if $d(\{Y\})$ is the smallest in Figure 7 (b), then $X \to Z \to Y$ is the shortest path (Figure 7 (d)). On the other hand, if $d(\{Z, X\})$ is the smallest in Figure 7 (b), we compute $I_n(z_y^n, x_{yz}^n)$ and $I_n(x_y^n, z_{xy}^n)$, and we obtain $d(\{X\}) := d(\{Z, X\}) + I_n(z_y^n, x_{yz}^n)$ and

$$d(\{Z\}) := d(\{Z, X\}) + I_n(x_y^n, z_{xy}^n)\ . \tag{18}$$

We close $\{Z, X\}$ and open $\{X\}$ and $\{Z\}$. However, the values of (17) and (18) conflict; thus, we replace (17) by (18) if (18) is smaller (Figure 7 (e)). Finally, if $d(\{X, Y\})$ is the smallest in Figure 7 (b), we obtain the state as depicted in Figure 7 (f), in which the values of $d(\{Y\})$ conflict, and the shorter path is chosen from $\{X, Y, Z\}$ to $\{Y\}$.

We continue this procedure to obtain the distance $d(\{\})$ and the shortest path from the top $\{X, Y, Z\}$ to the bottom $\{\}$. The same procedure can be applied to any number p of variables, not just three.

The procedure (LiNGAM-MMI) is summarized below as Algorithm 1 below with input DATA and output SHORTEST_PATH. Let TOP and BOTTOM be the top and bottom nodes, and we define $append((u_1, \cdots, u_s), u_{s+1}) := (u_1, \cdots, u_s, u_{s+1})$ for the nodes $u_1, u_2, \cdots, u_{s+1}$.

Algorithm 1 Let $OPEN := \{TOP\}$, CLOSE:=$\{\}$, $\mathrm{path}(\mathrm{TOP}) := ()$, $r(\mathrm{TOP})$:=DATA, and repeat:

1. Suppose $d(v)$ is the smallest among $v \in$ OPEN and that $v_1, \ldots, v_m$ are connected to the v. Then, move $v \in$ OPEN to CLOSE;
2. For each $i = 1, \cdots, m$:
   (a) If $v_i \notin$ OPEN, compute the residue $r(v_i)$ of v_i from $r(v)$;
   (b) Compute the estimated MI mi via $r(v)$ and $r(v_i)$.
   (c) If either $v_i \notin$ OPEN or $\{v_i \in$ OPEN, and $d(v) + mi < d(v_i)\}$, then $d(v_i) = d(v) + mi$ and $\mathrm{path}(v_i) = append(\mathrm{path}(v), v_i)$
   (d) join v_i to OPEN if $v_i \notin$ OPEN for $j = 1, \ldots, m$.
3. If BOTTOM $\in$ OPEN, SHORTEST_PATH = $append(\mathrm{path}(v), \{\})$ and terminate.

Note that Algorithm 1 does not compute the residues and MI estimates at the beginning; instead, it calculates each step by step when necessary to reduce the computational complexity. In addition, the SHORTEST_PATH is expressed by a sequence of nodes such as $(\{X, Y, Z\}, \{Y, Z\}, \{Z\}, \{\})$ rather than variables separated by arrows, as in $X \to Y \to Z$.

Theorem 1 Algorithm 1 computes the causal order of random variables that minimizes the estimated KL divergence of the corresponding noise set.

Then, one might think that the proposed procedure takes an exponential time with the number p of variables. In fact, for the worst case, we cannot avoid such computation that LvLiNGAM [Entner, 2013] and ParceLiNGAM [Tashiro et al., 2014] require. However, we have a significant merit over them (Theorem 2 below).

Let $I_n(\cdot, \cdot)$ be an MI estimator such that $I_n(x^n, y^n) = 0 \iff X \perp\!\!\!\perp Y$ with probability one for the $\{(x_i, y_i)\}_{i=1}^n$, independent realizations of (X, Y) for $x^n = (x_1, \ldots, x_n)$ and $y^n = (y_1, \ldots, y_n)$. Such an estimator $I_n(\cdot, \cdot)$ is said to be consistent.

Theorem 2 Suppose we estimate the MI between variables in Algorithm 1 via a consistent $I_n(\cdot, \cdot)$, When no confounding exists, the original and proposed LiNGAM take the same number

$$\frac{p(p+1)}{2} - 1$$

of computing the MI estimates $I_n(\cdot, \cdot)$ with probability one.

Proof. Suppose that there exists an order among the p variables such that the total distance

$$I(e_1, \{e_2, \ldots, e_p\}) + I(e_2, \{e_3, \ldots, e_p\}) + \cdots + I(e_{p-1}, e_p) = 0$$

is zero, which is equivalent to

$$I(e_1, \{e_2, \ldots, e_p\}) = I(e_2, \{e_3, \ldots, e_p\}) = \cdots = I(e_{p-1}, e_p) = 0\,.$$

Then, we have the residues $e_1^n, \ldots, e_p^n \in \mathbb{R}^n$ of the associated order such that each of

$$I_n(e_1^n, \{e_2^n, \ldots, e_p^n\})\,,\ I(e_2^n, \{e_3^n, \ldots, e_p^n\})\,,\ \cdots\,,\ I(e_{p-1}^n, e_p^n)$$

almost surely converges to zero. Then, among the OPEN variables of Algorithm 1, the one with

$$\sum_{j=1}^{k} I_n(e_j^n, \{e_{j+1}^n, \ldots, e_p^n\}) = 0$$

is chosen for iteration $k = 1, \ldots, p$. ■

For example, suppose that in Figure 7, $X \to Y \to Z$ is the correct order and that no confounding exists. Then, we first compute and compare $d(\{Y, Z\})$, $d(\{Z, X\})$, and $d(\{X, Y\})$ to find that $d(\{Y, Z\}) = 0$ is the smallest. Then, we compute $d(\{Y\})$ and $d(\{Z\})$ and compare with $d(\{Z, X\})$ and $d(\{X, Y\})$ to find that $d(\{Z\}) = 0$ is the smallets. Thus, we computed 3+2=5 MI estimates for $p = 3$. On the other hand, the original LiNGAM procedure in Section 2.3 requires five estimates of the MI or HSIC estimates before obtaining the order.

The proposed procedure finds the best order among the $p!$ candidates. On the other hand, the original LiNGAM procedure in Section 2.3 searches the solution in a greedy (topdown) manner and successfully finds the correct order when no confounding exists. However, even when no confounding exists, the estimates of HSIC and MI show positive values if starting from data, which means that no one can see the border between whether confounding exists. Thus, we cannot assume that no confounding exists in general situations.

In reality, if p is large, assuming that no confounding exists among the p variables is hopeless.

Moreover, if we follow the greedy search in Section 2.3 while the procedure is efficient, we will face fatal situations. For example, suppose we wish to identify the order among X, Y, Z from x^n, y^n, z^n, and that $I_n(x^n, \{y_x^n, z_x^n\})$ is smaller than the other MI estimates. Then, we cannot decide that X is the top variable because both of $I_n(y_x^n, z_{xy}^n\})$ and $I_n(y_{xz}^n, z_x^n\})$ may be large. In general, $e_1 \perp\!\!\!\perp \{e_2, e_3\}$ does not mean $e_2 \perp\!\!\!\perp e_3$.

3.3 MI Estimation

Many MI estimation procedures exist, such as [Kraskov et al., 2004], [Belghazi et al., 2018]. The precision of the proposed procedure depends on the choice of the MI estimation method. Among them, we choose copula entropy [Ma and Sun, 2011], which shows the best performance in our preliminary experiments.

Copulas [Sklar, 1959] provide a framework to separate the dependence structure from the marginal distributions. Let

$$F_X(x) := \int_{-\infty}^{x} f_X(x)dx \ , \ F_Y(y) := \int_{-\infty}^{y} f_Y(y)dy \ , \text{ and } F_{XY}(x,y) := \int_{-\infty}^{x}\int_{-\infty}^{y} f_{XY}(x,y)dx$$

with $x, y \in \mathbb{R}$ be the distribution functions of variables X and (X, Y). Sklar [1959] proves that the existence of the function C

$$F_{XY}(x,y) = C(F_X(x), F_Y(y))$$

with $x, y \in \mathbb{R}$ for any variables X, Y. Then, for $u_X = F_X(x)$ and $u_Y = F_Y(y)$, we have

$$\frac{du_X}{dx} = f_X(x) \ , \ \frac{du_Y}{dy} = f_Y(y)$$

and

$$c(u_X, u_Y) := \frac{\partial^2 C(u_X.u_Y)}{\partial u_X \partial u_Y} = \frac{f_{XY}(x,y)}{f_X(x)f_Y(y)} \ .$$

Then, we have two expressions for the MI:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)\log\frac{f_{XY}(x,y)}{f_X(y)f_Y(y)}dxdy = \int_0^1\int_0^1 c(u_X,u_Y)\log c(u_X,u_Y)du_X du_Y \ . \tag{19}$$

We estimate the right-hand side (negated copula entropy) of (19) rather than the left, which we claim is much easier to compute. Even when X, Y are in high dimensions, the estimation of the right-hand side is always in the two-dimensional space. For the details, see Ma [2021]. For n data points, we take the following steps:

1. Computing

$$u_X^i = \frac{1}{n}\sum_{j=1}^{n} 1_{\{x^j \le x^i\}} \quad \text{and} \quad u_Y^i = \frac{1}{n}\sum_{j=1}^{n} 1_{\{y^j \le y^i\}} \ ,$$

   for $i = 1, \ldots, n$ from samples $\left\{x^i\right\}_{i=1}^n, \left\{y^i\right\}_{i=1}^n$

2. Using whatever entropy estimation method to estimate the entropy of samples $\left\{u^i\right\}_{i=1}^n$ where $u^i = \left\{(u_X^i, u_Y^i)\right\}$ (concatenation and yields variables in two dimensions).

In the second step, we use Kraskov et al. [2004] based on the KNN method.

One might think that an independence test based on the HSIC (Hilbert Schmidt Information Criterion [Gretton et al., 2007]) achieves better performance than the one based on MI. In the case $p = 2$, this may hold. However, the independence test requires us to execute a greedy search for $p \ge 3$, which may result in poor performance unless n is infinitely large and no confounder exists. In the next section, we can examine the performances of the original LiNGAM using the HSIC and the proposed LiNGAM.

4 Experiments

This section presents the experimental results.

We compare the proposed method to variants of the LiNGAM as specified in Table 1, focusing on causal order identification. For the implementation, we utilize Python packages: `copent` for copula entropy [Ma, 2021], and `lingam`, `gCastle` [Zhang et al., 2021]. When confounders are present, we also substitute RESIT with ParceLiNGAM [Tashiro et al., 2014].

For assessing causal order, we employ two criteria for performance comparison: Criterion A counts an error when the whole order of p variables is incorrect. Criterion B counts the pairwise errors and divides the count by $p(p-1)/2$. Criterion A evaluates the error rate more severely, particularly when p is large.

Table 1: The procedures used in the experiments and Figures 7-10

Pairwise	DirectLiNGAM using pairwise likelihood	Hyvärinen and Smith [2013]
Kernel	kernel-based DirectLiNGAM	Shimizu et al. [2011]
HSIC	DirectLiNGAM using HSIC	Hyvärinen and Smith [2013]
ICA	ICA-LiNGAM	Shimizu et al. [2006]
MMI	LiNGAM-MMI using copula entropy	Proposed

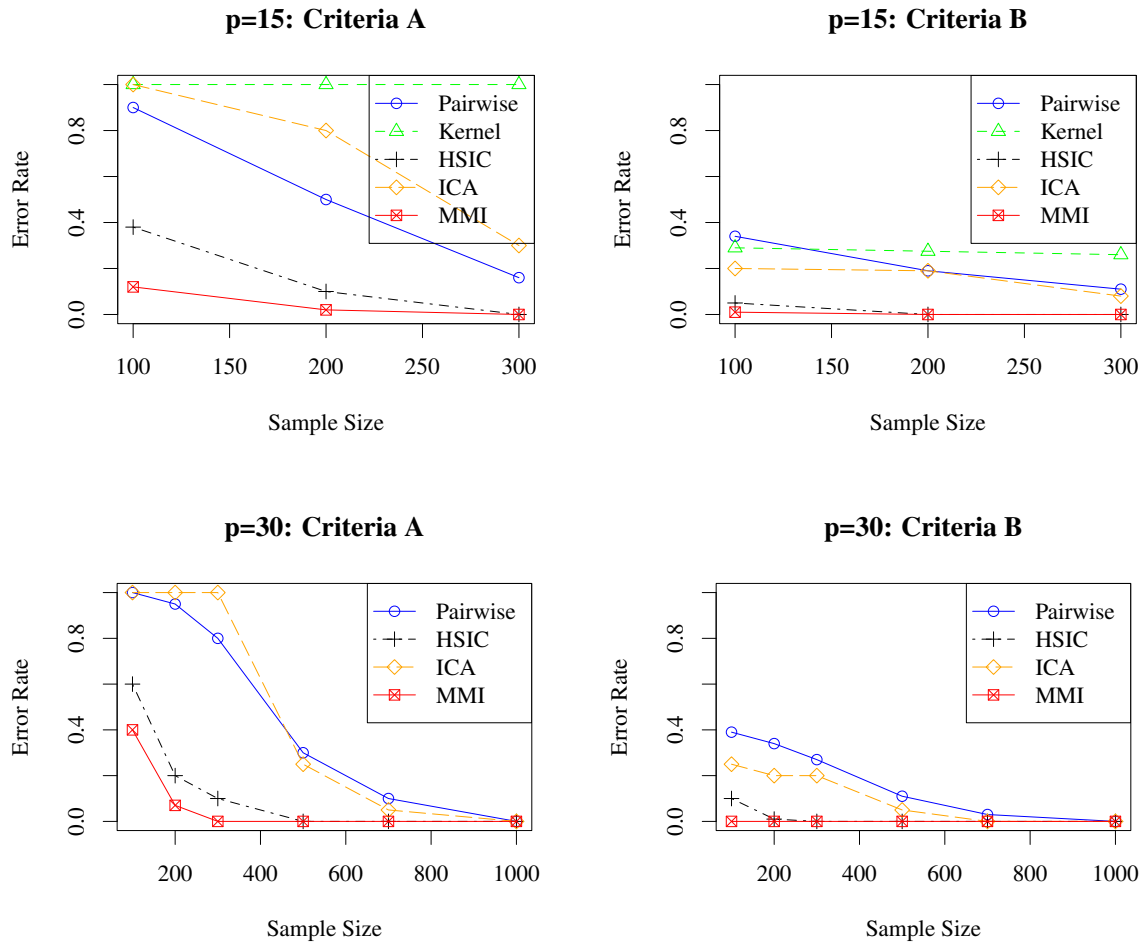


Figure 8: **No Confounder** Error ratio of Criteria A and B with sample size n (the lower, the better). We observe that the proposed LiNGAM-MMI performs better than the conventional methods.

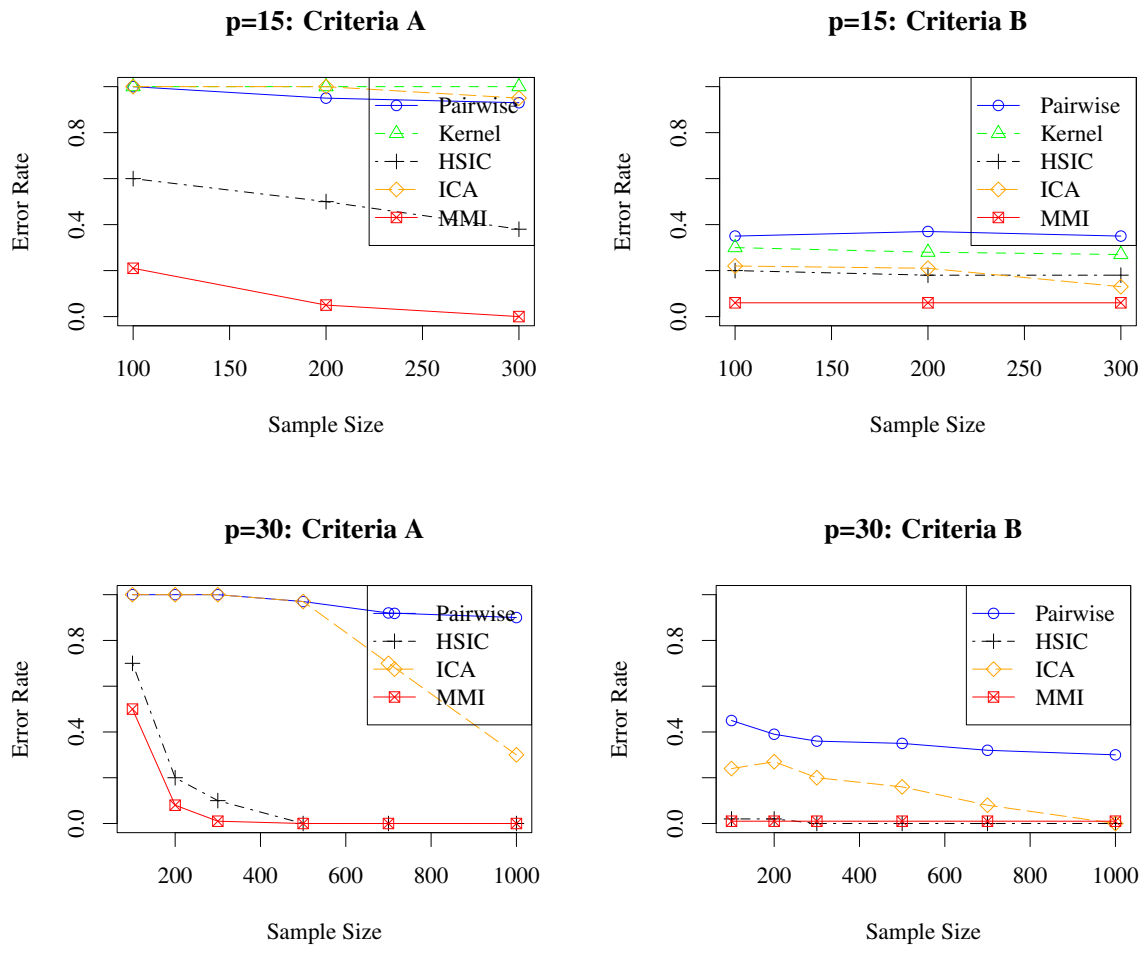


Figure 9: **With Confounder** Error ratio of Criteria A and B with sample size n (the lower, the better). We observe that the proposed LiNGAM-MMI performs better than the conventional methods.

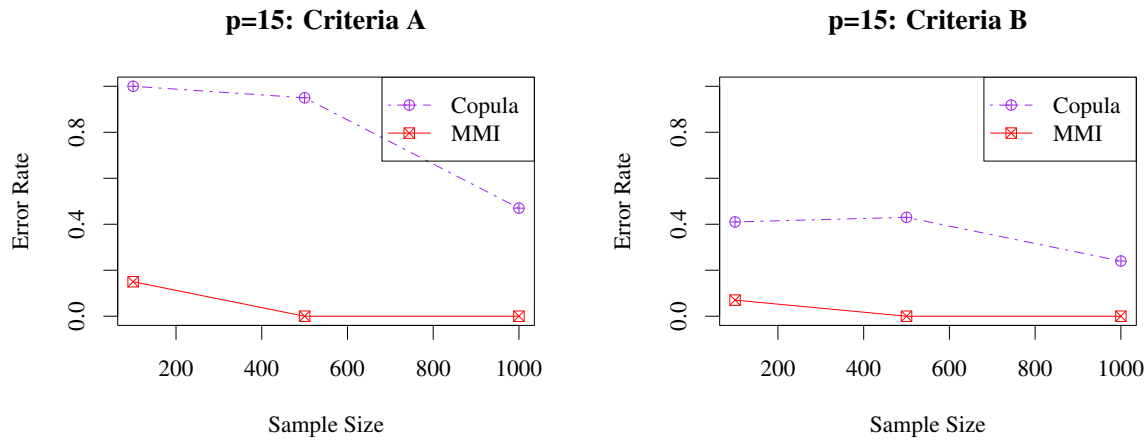


Figure 10: **No Confounder** Error ratio of Criterion A and B with sample size n (the lower, the better) when copula entropy was applied to the proposed and existing methods.

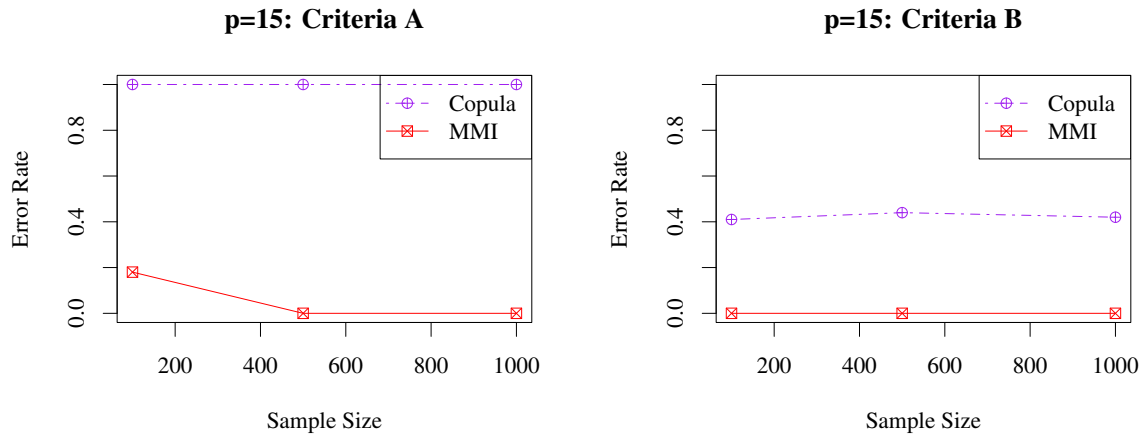


Figure 11: **With Confounder** Error ratio of Criterion A and B with sample size n (the lower, the better) when copula entropy was applied to the proposed and existing methods.

4.1 Simulations

We estimate the causal order for $p = 15, 30$. The true causal order is $x_0 \to x_1, \ldots, \to x_{p-1} \to x_p$ with $x_i = x_{i-1} + e_i, i = 1, \ldots, p$. When considering confounders, we assume they are $f_k \sim 2U(0,1) - 1, k = 1, \ldots, 8$ at different positions. For $p = 30$, we set the confounder with f_1 on $\{x_1, x_2\}$, f_2 on $\{x_5, x_6\}$, f_3 on $\{x_8, x_9\}$, f_4 on $\{x_{10}, x_{11}\}$, f_5 on $\{x_{13}, x_{14}\}$, f_6 on $\{x_{15}, x_{16}\}$, f_7 on $\{x_{20}, x_{21}\}$, f_8 on $\{x_{25}, x_{26}\}$, respectively. For $p = 15$, we set the confounder as f_1 on $\{x_2, x_3\}$, f_2 on $\{x_5, x_6\}$, f_3 on $\{x_8, x_9\}$, f_4 on $\{x_{12}, x_{13}\}$, respectively.

We execute the existing and proposed procedures for $n = 100, 200, 300, 500, 700, 1000$, $p = 30$ and $n = 100, 200, 300$, $p = 15$, we compare the performances of Criteria A and B. For the whole procedure, we take the arithmetic average over 50 trials. The results in Figure 8 and Figure 9 show that our proposed method performs best whether the sample size n is small or large. We observe that the DirectLiNGAM with HSIC shows competitive performances with the LiNGAM-MMI only when confounding is absent and the sample size is large. In particular, even when no confounding exists, the DirectLiNGAM with HSIC often failed for $p = 30$ and a small sample size. The reason seems to be that when the sample size is small, the situation is far from confounding-free.

In addition, we compare our LiNGAM-MMI with the DirectLiNGAM with copula entropy (the same MI measure as our method). We show the results in Figure 10 and Figure 11, which suggests that our method still performs better even though the DirectLiNGAM uses copula entropy, which suggests that the greedy search fails to give correct orders often.

4.2 Real data

We examined the order of the variables in the General Social Survey data set [4], taken from a sociological data repository, used in Shimizu et al. [2011], Maeda and Shimizu [2020]. See the actual causal structure in Figure 12. In general, we do not know the true order among variables. However, in this example, from the meanings of the six variables, we may assume that

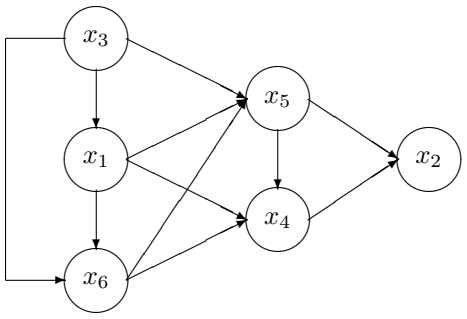
$$X_3 \to X_1 \to X_6 \to X_5 \to X_4 \to X_2 \,.$$

We observed that the LiNGAM-MMI obtains the same order. We cannot claim that the LiNGAM-MMI always shows better performances only from this example but rather illustrate how to apply it to real data. For seeking the causal structure besides the order, we need to obtain the parent sets of each variable using structure learning such as the PC algorithm [Kalisch and Bühlman, 2007].

5 Concluding Remarks

This paper extends LiNGAM itself without explicitly searching for variables influenced by confounding. Specifically, it proposes a method (LiNGAM-MMI) that quantifies the magnitude of confounding using KL divergence

[4] http://www.norc.org/GSS+Website/



x_1: Father's Occupation
x_2: Son's Income
x_3: Father's Education
x_4: Son's Occupation
x_5: Son's Education
x_6: Number of Siblings

Figure 12: Causal relations by LiNGAM-MMI in GSS dataset. Red arrows: correct; dashed ones: wrong; other: appropriate.

and determines the order of variables to minimize it. This approach formulates the problem using the shortest path problem and successfully finds globally optimal solutions efficiently. The proposed LiNGAM-MMI completes processing in the same amount of time as the existing LiNGAM when there is no confounding, as it does not address confounding. The experiments support the merits of the LiNGAM-MMI.

Although this paper has extended the estimation of Mutual Information (MI) using copulas, applying estimators with higher accuracy is expected to improve performance further.

Appendix A: Proof of $z^n_{xy} = z^n_{yx}$

From definitions of $v(\cdot)$ and $c(\cdot,\cdot)$, we have

$$\begin{aligned}
c(y^n_x, z^n_x) &= c(y^n - \frac{c(x^n,y^n)}{v(x^n)}x^n, z^n - \frac{c(x^n,z^n)}{v(x^n)}x^n) = c(y^n,z^n) - \frac{c(x^n,y^n)c(x^n,z^n)}{v(x^n)} \\
v(y^n_x) &= v(y^n - \frac{c(x^n,y^n)}{v(x^n)}x^n) = v(y^n) + \frac{c(x^n,y^n)^2}{v(x^n)} - 2\frac{c(x^n,y^n)^2}{v(x^n)} = v(y^n) - \frac{c(x^n,y^n)^2}{v(x^n)} \\
\frac{c(y^n_x, z^n_x)}{v(y^n_x)} &= \frac{v(x^n)c(y^n,z^n) - c(x^n,y^n)c(x^n,z^n)}{v(x^n)v(y^n) - c(x^n,y^n)^2}\,.
\end{aligned}$$

Thus, we can express (11) as

$$\begin{aligned}
z^n_{xy} &= z^n - \frac{c(x^n,z^n)}{v(x^n)}x^n - \frac{v(x^n)c(y^n,z^n) - c(x^n,y^n)c(x^n,z^n)}{v(x^n)v(y^n) - c(x^n,y^n)^2}\{y^n - \frac{c(x^n,y^n)}{v(x^n)}x^n\} \\
&= z^n - \frac{c(x^n,z^n)\{v(x^n)v(y^n) - c(x^n,y^n)^2\} - c(x^n,y^n)\{v(x^n)c(y^n,z^n) - c(x^n,y^n)c(x^n,z^n)\}}{v(x^n)\{v(x^n)v(y^n) - c(x^n,y^n)^2\}}x^n \\
&\quad - \frac{v(x^n)c(y^n,z^n) - c(x^n,y^n)c(x^n,z^n)}{v(x^n)v(y^n) - c(x^n,y^n)^2}y^n \\
&= z^n - \frac{v(y^n)c(x^n,z^n) - c(x^n,y^n)c(y^n,z^n)}{v(x^n)v(y^n) - c(x^n,y^n)^2}x^n - \frac{v(x^n)c(y^n,z^n) - c(x^n,y^n)c(x^n,z^n)}{v(x^n)v(y^n) - c(x^n,y^n)^2}y^n
\end{aligned}$$

which coincides with (12) since they are symmetric between x^n and y^n.

Appendix B: LvLiNGAM and ParceLiNGAM

LvLiNGAM [Entner, 2013] considers e_{12}, e_{13}, etc., as well as e_1, e_2, e_3 as noise variables, and obtains the orders of variable pairs that are not affected by any confounder such as e_{12}, e_{13} (Figure 1), and estimates the order of the whole variables by combining those pairwise orders.

ParceLiNGAM [Tashiro et al., 2014] divides the variable set V into upper, middle, and lower variable sets (we denote them as U, M, L, respectively, such that $V = U \cup M \cup L$) by top-down and bottom-up causal searches, where the upper and lower variable sets are the maximal subsets that contain no confounder but the top and bottom variables, respectively. Let $F_U(u)$ and $F_L(l)$ be the Fisher estimations between $u \in U$ and its lower variables and

between $l \in L$ and its upper variables, respectively. Then, the following quantity can be evaluated for V:

$$F := \sum_{u \in U} F_U(u) + \sum_{l \in L} F_L(l) \ .$$

For each subset S of V, ParceLiNGAM estimates U, M, L and computes F for S rather than V. Then, the order between variables $v, v' \in V$ can be determined based on S that determines the order between v and v' and maximizes F.

However, ParceLiNGAM has several drawbacks. First of all, it is possible that the order cannot be determined based on any $S \subseteq V$ for some $v, v' \in V$. For example, if V consists of X, Y and a confounder exist for them, ParceLiNGAM does not work at all (Figure 5). Secondly, the evaluation is for each pair of variables rather than for each path from the top to the bottom, so we are not sure that the estimation is correct. Moreover, Fisher estimation between variable r and variable set R evaluates the sum of pairwise independences between r and each element in R, and using it is not appropriate unless the variables in R are mutually independent. Finally, the computation is huge: 2^p subsets should be considered for p variables, so that ParceLiNGAM can be used only when p is small.

References

R. Ayesha Ali, Thomas S. Richardson, and Peter Spirtes. Markov equivalence for ancestral graphs. *The Annals of Statistics*, 37(5B):2808 – 2837, 2009. doi: 10.1214/08-AOS626. URL https://doi.org/10.1214/08-AOS626.

MI Belghazi, A Baratin, S Rajeshwar, S Ozair, and Y Bengio. Mutual information neural estimation. In *International conference on machine learning*, 2018.

K. A. Bollen. Structural equations with latent variables. John Wiley & Sons, 1989.

Wenyu Chen, Mathias Drton, and Ali Shojaie. Causal structural learning via local graphs, 2021.

Tom Claassen, Joris M. Mooij, and Tom Heskes. Learning sparse causal models is not np-hard. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, UAI'13, page 172–181, Arlington, Virginia, USA, 2013. AUAI Press.

Diego Colombo, Marloes H. Maathuis, Markus Kalisch, and Thomas S. Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, 40(1):294 – 321, 2012. doi: 10.1214/11-AOS940. URL https://doi.org/10.1214/11-AOS940.

G. Darmois. Analyse générale des liaisons stochastiques: etude particulière de l'analyse factorielle linéaire. *Revue de l'Institut International de Statistique / Review of the International Statistical Institute*, 21(1/2):2–8, 1953. ISSN 03731138. URL http://www.jstor.org/stable/1401511.

Doris Entner. *Causal Structure Learning and Effect Identifcation in Linear Non-Gaussian Models and Beyond*. PhD thesis, University of Helsinki, 01 2013.

Doris Entner and Patrik O Hoyer. On causal discovery from time series data using fci. *Probabilistic graphical models*, pages 121–128, 2010.

Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola. A kernel statistical test of independence. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper/2007/file/d5cfead94f5350c12c322b5b664544c1-Paper.

Patrik O. Hoyer, Shohei Shimizu, Antti J. Kerminen, and Markus Palviainen. Estimation of causal effects using linear non-gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49(2):362–378, 2008. ISSN 0888-613X. doi: https://doi.org/10.1016/j.ijar.2008.02.006. URL https://www.sciencedirect.com/science/article/pii/S0888613X08000212. Special Section on Probabilistic Rough Sets and Special Section on PGM'06.

Aapo Hyvärinen and Stephen M. Smith. Pairwise likelihood ratios for estimation of non-gaussian structural equation models. *Journal of Machine Learning Research*, 14(4):111–152, 2013. URL http://jmlr.org/papers/v14/hyvarinen13a.html.

Markus Kalisch and Peter Bühlman. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(3), 2007.

Y. Kano and S. Shimizu. "Causal inference using non-normality". In *The International Symposium on Science of Modeling: The 30th Anniversary of the Information Criterion*, pages 261–270, Washington DC, 12 2003.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Phys. Rev. E*, 69:066138, Jun 2004. doi: 10.1103/PhysRevE.69.066138. URL https://link.aps.org/doi/10.1103/PhysRevE.69.066138.

Jian Ma. copent: Estimating copula entropy and transfer entropy in r. *arXiv:2005.14025*, 2021.

Jian Ma and Zengqi Sun. Mutual information is copula entropy. *Tsinghua Science & Technology*, 16(1):51–54, 2011. ISSN 1007-0214. doi: https://doi.org/10.1016/S1007-0214(11)70008-6. URL https://www.sciencedirect.com/science/article/pii/S1007021411700086.

Takashi Nicholas Maeda and Shohei Shimizu. Rcd: Repetitive causal discovery of linear non-gaussian acyclic models with latent confounders. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 735–745. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/maeda20a.html.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009. ISBN 052189560X.

Richard Scheines Peter Spirtes, Clark Glymour. *Causation, Prediction, and Search*. Springer New York, NY, 1993.

Saber Salehkaleybar, AmirEmad Ghassami, Negar Kiyavash, and Kun Zhang. Learning linear non-gaussian causal models in the presence of latent variables. *Journal of Machine Learning Research*, 21(39):1–24, 2020. URL http://jmlr.org/papers/v21/19-260.html.

Shohei Shimizu and K. Bollen. Bayesian estimation of causal direction in acyclic structural equation models with individual-specific confounder variables and non-gaussian distributions. *Journal of machine learning research : JMLR*, 15:2629–2652, 2014.

Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen;rinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(72):2003–2030, 2006. URL http://jmlr.org/papers/v7/shimizu06a.html.

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O. Hoyer, and Kenneth Bollen. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research*, 12(33):1225–1248, 2011. URL http://jmlr.org/papers/v12/shimizu11a.html.

W. P. Skitovitch. On a property of the normal distribution. *Doklady Akademii Nauk SSSR*, 89:217–219, 1953.

Martin J. Sklar. Fonctions de repartition a n dimensions et leurs marges. In *Publications de l'Institut de Statistiquede l'Université de Paris*, 1959.

Peter Spirtes, Christopher Meek, and Thomas Richardson. Causal inference in the presence of latent variables and selection bias. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, page 499–506, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. ISBN 1558603859.

J Suzuki. *Kernel Methods for Machine Learning with Math and R: 100 Exercises for Building Logic*. Springer, 2022.

Joe Suzuki and Yusuke Inaoka. Causal order identification to address confounding: binary variables. *Behaviormetrika*, 2022.

Tatsuya Tashiro, Shohei Shimizu, Aapo Hyvärinen, and Takashi Washio. ParceLiNGAM: A Causal Ordering Method Robust Against Latent Confounders. *Neural Computation*, 26(1):57–83, 01 2014. ISSN 0899-7667.

Y. Samuel Wang and Mathias Drton. Causal discovery with unobserved confounding and non-gaussian data. *Journal of Machine Learning Research*, 24(271):1–61, 2023. URL http://jmlr.org/papers/v24/21-1329.html.

Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896, 2008. doi: 10.1016/j.artint.2008.08.001.

Keli Zhang, Shengyu Zhu, Marcus Kalander, Ignavier Ng, Junjian Ye, Zhitang Chen, and Lujia Pan. gcastle: A python toolbox for causal discovery. *ArXiv*, abs/2111.15155, 2021.