Identification of Causal Structure with Latent Variables Based on Higher Order Cumulants

Wei Chen[1], Zhiyi Huang[1], Ruichu Cai[1,2*], Zhifeng Hao[1,3], Kun Zhang[4,5]
[1] School of Computer Science, Guangdong University of Technology, Guangzhou, China
[2] Peng Cheng Laboratory, Shenzhen, China
[3] College of Science, Shantou University, Shantou, China
[4] Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA, United States
[5] Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates
{chenweidelight, huangzhiyichn, cairuichu}@gmail.com, zfhao@gdut.edu.cn, kunz1@cmu.edu

Abstract

Causal discovery with latent variables is a crucial but challenging task. Despite the emergence of numerous methods aimed at addressing this challenge, they are not fully identified to the structure that two observed variables are influenced by one latent variable and there might be a directed edge in between. Interestingly, we notice that this structure can be identified through the utilization of higher-order cumulants. By leveraging the higher-order cumulants of non-Gaussian data, we provide an analytical solution for estimating the causal coefficients or their ratios. With the estimated (ratios of) causal coefficients, we propose a novel approach to identify the existence of a causal edge between two observed variables subject to latent variable influence. In case when such a causal edge exits, we introduce an asymmetry criterion to determine the causal direction. The experimental results demonstrate the effectiveness of our proposed method.

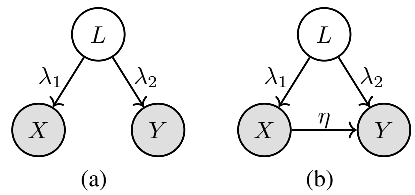


Figure 1: A causal structure with two observed variables X and Y affected by one latent variable L, where λ_1, λ_2 and η represent the causal strength between L and X, L and Y, X and Y, respectively.

Introduction

Inferring causal relationships between observed variables with latent variables is of significant importance and has been applied in many fields (Sachs et al. 2005; Wang and Drton 2020; Tramontano, Monod, and Drton 2022; Morioka and Hyvarinen 2023). The Latent Variables Linear Non-Gaussian Acyclic Model (LvLiNGAM) (Hoyer et al. 2008b; Entner and Hoyer 2011; Tashiro et al. 2014) is one of the most prominent approaches for this problem. Notably, LvLiNGAM can be transformed into a canonical model, wherein each latent variable is the cause of a minimum of two children and has no parents (Hoyer et al. 2008b). Therefore, the identification of the structure involving one latent variable and two observed variables (as shown in Figure 1) is one of the fundamental problems in causal discovery.

Various methods are proposed to identify the causal structure with latent variables. One of the typical methods is based on conditional independence tests, such as FCI (Spirtes, Meek, and Richardson 1995) and its variants (Colombo et al. 2012), but they can only identify up to Markov equivalent class. Under the measurement assumption, some approaches use rank constraints (Silva et al. 2006; Kummerfeld and Ramsey 2016; Huang et al. 2022), and Generalized Independence Noise (GIN) condition (Xie et al. 2020, 2022) to identify the relationships between latent variables and observed variables. By utilizing the non-Gaussianity, PraceLiNGAM (Tashiro et al. 2014), MLCLiNGAM (Chen et al. 2021) and RCD (Maeda and Shimizu 2020) can only identify some causal structure among observed variables that are not directly affected by the latent variables. However, the above approaches fail to identify the structures given in Figure 1. Although lvLiNGAM leverages the Overcomplete Independence Component Analysis technique (ICA) (Eriksson and Koivunen 2004; Lewicki and Sejnowski 2000), it is hard to obtain an optimal result, which would lead to wrong causal relations. In summary, how to fully identify the causal relationships between two observed variables influenced by one latent variable is still an open problem.

To overcome the aforementioned challenge, two problems should be considered: 1) How to detect whether there exists a causal edge between two observed variables? 2) Furthermore, when such an edge exists, how to determine its direction? Fortunately, we notice that certain types of non-Gaussianity can be measured by higher-order cumulants, which can also capture some data distribution information (Hyvärinen, Karhunen, and Oja 2001). Thus, utilizing higher-order cumulants can yield additional information to address the aforementioned issues.

To address the first problem, we find that some specific combinations of higher-order joint cumulants of two observed variables can be used to detect whether there exists

*Corresponding author.

a causal edge between them. In Figure 1 (a), the joint cumulants of X and Y can be approximately regarded as the multiplication of powers of coefficients and the cumulants of their shared components, which can be expressed as follows:

$$\mathcal{C}_{i,j}(X,Y)=\lambda_1^i\lambda_2^j\mathcal{C}_{i+j}(L), i,j>0, \tag{1}$$

where $\mathcal{C}_{i,j}(X,Y)$ represents cumulant $Cum(\underbrace{X,\ldots}_{i \text{ times}},\underbrace{Y,\ldots}_{j \text{ times}})$. This distinction arises from the fact that X and Y only share one latent component L in Figure 1(a), whereas they share two components L and the noise of X in Figure 1(b). Then, asymptotically, if and only if no causal directed edge exists between X and Y, the following constraint holds:

$$\mathcal{C}_{i+1,j+1}(X,Y)^2=\mathcal{C}_{i,j+2}(X,Y)\mathcal{C}_{i+2,j}(X,Y). \tag{2}$$

The next problem is how to determine the causal direction between two observed variables. Interestingly, considering data related to observed variables X and Y generated according to Figure 1(b), leveraging the third-order cumulants of X and Y, one have

$$\begin{aligned}&\mathcal{C}_3(X)=\lambda_1^3\mathcal{C}_3(L)+\mathcal{C}_3(E_X),\\&\mathcal{C}_3(Y)\neq(\lambda_1\eta+\lambda_2)^3\mathcal{C}_3(L)+\eta^3\mathcal{C}_3(E_X),\end{aligned} \tag{3}$$

where E_X and E_Y represent noise terms associated with X and Y respectively, and L denotes a latent variable. This discrepancy arises from the presence of an additional term E_Y in Y, which is distinct from X. Thus, we can develop an asymmetry criterion to determine the causal direction between two variables in the presence of latent variables.

Preliminary

Latent-Variable Linear Non-Gaussian Acyclic Model

In this paper, we consider the data over observed variables $\mathbf{V}$ that may be affected by the latent variables $\mathbf{L}$. Specifically, these data are generated from a linear causal model with non-Gaussian noises, which can be formalized as:

$$\mathbf{V}=\mathbf{B}\mathbf{V}+\mathbf{\Lambda}\mathbf{L}+\mathbf{E}, \tag{4}$$

where $\mathbf{B}$ is the causal strength matrix, $\mathbf{\Lambda}$ is the causal strength matrix from $\mathbf{L}$ to $\mathbf{V}$, $\mathbf{E}$ is a non-Gaussian noise term and each $E_i\in\mathbf{E}$ is independent of others. This model is also termed Latent-Variable Linear Non-Gaussian Acyclic Model (LvLiNGAM) (Hoyer et al. 2008b).

In the linear case, many researchers are likely to transform the above model into OICA model (Lewicki and Sejnowski 2000; Eriksson and Koivunen 2004), $\mathbf{V}=\mathbf{A}\mathbf{S}$, where $\mathbf{A}$ is represented as the mixing matrix and $\mathbf{S}$ denotes the latent independent components. For example, in Figure 1(b), according to the data generation process described by Eq. (4), X and Y can be formalized as:

$$\begin{aligned}&X=\lambda_1L+E_X=\alpha_1L+\beta_1E_X,\\&Y=\lambda_2L+\eta X+E_Y=\alpha_2L+\beta_2E_X+\gamma_2E_Y,\end{aligned} \tag{5}$$

where λ_1 and λ_2 are the causal strengths from L to X and Y, respectively; η represents the causal strength from X to Y. For clearly, $\alpha_1=\lambda_1$ and $\alpha_2=\eta\lambda_1+\lambda_2$ are the mixing coefficients of L on X and Y, respectively; β_1 and $\beta_2=\eta$ are the mixing coefficients of E_X on X and Y, respectively; and γ_2 is the mixing coefficients of E_Y on Y. E_X and E_Y are two independent noise terms.

In the perspective of the matrix, Eq. (5) can be transformed into the form of an OICA model over two observed variables X and Y as:

$$\begin{aligned}\underbrace{\begin{bmatrix}X\\Y\end{bmatrix}}_{\mathbf{V}}&=\underbrace{\begin{bmatrix}0&0\\\eta&0\end{bmatrix}}_{\mathbf{B}}\underbrace{\begin{bmatrix}X\\Y\end{bmatrix}}_{\mathbf{V}}+\underbrace{\begin{bmatrix}\lambda_1\\\lambda_2\end{bmatrix}}_{\mathbf{\Lambda}}\underbrace{[L]}_{\mathbf{L}}+\underbrace{\begin{bmatrix}E_X\\E_Y\end{bmatrix}}_{\mathbf{E}}\\&=\underbrace{\begin{bmatrix}\alpha_1&\beta_1&0\\\alpha_2&\beta_2&\gamma_2\end{bmatrix}}_{\mathbf{A}}\underbrace{\begin{bmatrix}L\\E_X\\E_Y\end{bmatrix}}_{\mathbf{S}}.\end{aligned} \tag{6}$$

Cumulants

The cumulants is a measure to capture the (joint) probability distribution information from data. The definition of cumulants of a random vector X is formalized as:

Definition 1 (**Cumulants** (Brillinger 2001))**.** Let $X=(X_1,X_2,\ldots,X_n)$ be a random vector of length n. The k-th order cumulant tensor of X is defined as a $n\times\cdots\times n$ (k times) table, $\mathcal{C}^{(k)}$, whose entry at position $(i_1,\cdots,i_k)$ is

$$\begin{aligned}&\mathcal{C}^{(k)}_{i_1,\cdots,i_k}=Cum(X_{i_1},\ldots,X_{i_k})\\&=\sum_{(D_1,\ldots,D_h)}(-1)^{h-1}(h-1)!\mathbb{E}\left[\prod_{j\in D_i}X_j\right]\cdots\mathbb{E}\left[\prod_{j\in D_h}X_j\right],\end{aligned} \tag{7}$$

where the sum is taken over all partitions $(D_1,\ldots,D_h)$ of the set $\{i_1,\ldots,i_k\}$.

For convenience, we use $\mathcal{C}_i(X)$ to denote $Cum(\underbrace{X,\ldots}_{i \text{ times}})$, and use $\mathcal{C}_{i,j}(X,Y)$ to denote $Cum(\underbrace{X,\ldots}_{i \text{ times}},\underbrace{Y,\ldots}_{j \text{ times}})$. For example, $\mathcal{C}_5(X)$ represents $Cum(X,X,X,X,X)$, $\mathcal{C}_{2,3}(X,Y)$ represents $Cum(X,X,Y,Y,Y)$.

Note that the first-order cumulant is the mean, and the second-order cumulant is the variance. If each variable X_i has zero mean, then the sum of the partitions with size 1 is 0 and can be omitted. For example, suppose X and Y are generated by Eq. (6), the 3rd and 5th order (joint) cumulants of X and Y are:

$$\begin{aligned}\mathcal{C}_3(X)&=\alpha_1^3\mathcal{C}_3(L)+\beta_1^3\mathcal{C}_3(E_X),\\\mathcal{C}_{2,1}(X,Y)&=\alpha_1^2\alpha_2\mathcal{C}_3(L)+\beta_1^2\beta_2\mathcal{C}_3(E_X),\\\mathcal{C}_{1,2}(X,Y)&=\alpha_1\alpha_2^2\mathcal{C}_3(L)+\beta_1\beta_2^2\mathcal{C}_3(E_X),\\\mathcal{C}_3(Y)&=\alpha_2^3\mathcal{C}_3(L)+\beta_2^3\mathcal{C}_3(E_X)+\gamma_2^3\mathcal{C}_3(E_Y),\\\mathcal{C}_{2,3}(X,Y)&=\alpha_1^2\alpha_2^3\mathcal{C}_3(L)+\beta_1^2\beta_2^3\mathcal{C}_3(E_X),\\\mathcal{C}_{3,2}(X,Y)&=\alpha_1^3\alpha_2^2\mathcal{C}_3(L)+\beta_1^3\beta_2^2\mathcal{C}_3(E_X).\end{aligned} \tag{8}$$

The equations above reveal that the higher-order cumulants imply the property of the generation process of the observed data (even when latent variables are present).

Intuition

The intuition of our method is based on the following observations. Take the causal graph between two observed variables X and Y, given in Figure 1(a) as an example. X and Y are affected by a latent variable, and there is no causal directed edge between them. The joint cumulant of X and Y can be approximately regarded as the multiplication of powers of coefficients and the cumulants of their shared components, which can be expressed as follows:

$$\begin{aligned} \mathcal{C}_{1,3}(X,Y) &= \lambda_1\lambda_2^3\mathcal{C}_4(L), \\ \mathcal{C}_{2,2}(X,Y) &= \lambda_1^2\lambda_2^2\mathcal{C}_4(L), \\ \mathcal{C}_{1,3}(X,Y) &= \lambda_1^3\lambda_2\mathcal{C}_4(L). \end{aligned} \tag{9}$$

Multiplying the above joint cumulants, one can obtain

$$\mathcal{C}_{2,2}(X,Y)^2 = \mathcal{C}_{1,3}(X,Y)\mathcal{C}_{3,1}(X,Y). \tag{10}$$

When there is a causal directed edge between two observed variables as given in Figure 1(b), the joint cumulants that contain two different terms about $\mathcal{C}_4(L)$ and $\mathcal{C}_4(E_X)$ are:

$$\begin{aligned} \mathcal{C}_{1,3}(X,Y) &= \lambda_1(\lambda_1\eta+\lambda_2)^3\mathcal{C}_4(L) + \eta^3\mathcal{C}_4(E_X), \\ \mathcal{C}_{2,2}(X,Y) &= \lambda_1^2(\lambda_1\eta+\lambda_2)^2\mathcal{C}_4(L) + \eta^2\mathcal{C}_4(E_X), \\ \mathcal{C}_{3,1}(X,Y) &= \lambda_1^3(\lambda_1\eta+\lambda_2)\mathcal{C}_4(L) + \eta^1\mathcal{C}_4(E_X). \end{aligned} \tag{11}$$

If we compute the constraint as Eq. (10), the term $\lambda_1^2(\lambda_1\eta + \lambda_2)^2\eta^2\mathcal{C}_4(Y)\mathcal{C}_4(E_X)$ appears on the left-hand side of Eq. (10), but it is absent from the right-hand side. This discrepancy leads to the violation of Eq. (10).

Furthermore, if we find the violation of the Eq. (10), that means there is a causal edge between two observed variables. The following question is how to determine its direction. Consider the causal graph given in Figure 1(b), where X is a parent of Y and both are affected by a latent confounder L. The data generation process of X and Y is described by Eq. (5). Note that in this case, X and Y share two latent independent components, namely L and E_X. Next, we will show that by utilizing fifth-order cumulants, the causal direction between two observed variables can be identified.

First, considering the covariance (i.e., the second-order cumulant), the second-order (joint) cumulants are:

$$\begin{aligned} Var(X) &= \alpha_1^2Var(L) + \beta_1^2Var(E_X), \\ Cov(X,Y) &= \alpha_1\alpha_2Var(L) + \beta_1\beta_2Var(E_X), \\ Var(Y) &= \alpha_2^2Var(L) + \beta_2^2Var(E_X) + \gamma_2^2Var(E_Y). \end{aligned} \tag{12}$$

We observed that the variance of the cause variable X only contains all the variance of shared components, e.g., $Var(L)$ and $Var(E_X)$, but not the effect variable Y. It shows that the parameter estimation of the shared components sheds light on the identification of causal direction. Although we can find that $Var(E_Y)$ is only in $Var(Y)$, we cannot uniquely determine the variance of L, E_X and E_Y, and estimate all the parameters by using only the three equations above, as the result in (Ledermann 1937; Bekker and ten Josephus Berge 1997).

Note that the covariance is the second-order cumulant, the idea that naturally arises is whether higher-order cumulants can be used to solve the problem. Thus, we consider the third order cumulant of X and Y:

$$\begin{aligned} \mathcal{C}_3(X) &= \alpha_1^3\mathcal{C}_3(L) + \beta_1^3\mathcal{C}_3(E_X), \\ \mathcal{C}_3(Y) &= \alpha_2^3\mathcal{C}_3(L) + \beta_2^3\mathcal{C}_3(E_X) + \gamma_2^3\mathcal{C}_3(E_Y). \end{aligned} \tag{13}$$

Suppose we can estimate the terms $\alpha_1^3\mathcal{C}_3(L)$, $\beta_1^3\mathcal{C}_3(E_X)$, $\alpha_2^3\mathcal{C}_3(L)$, and $\beta_2^3\mathcal{C}_3(E_X)$, then we have the asymmetries that

$$\begin{aligned} \mathcal{R}_{X\to Y} &= \mathcal{C}_3(X) - \alpha_1^3\mathcal{C}_3(L) - \beta_1^3\mathcal{C}_3(E_X) = 0, \\ \mathcal{R}_{Y\to X} &= \mathcal{C}_3(Y) - \alpha_2^3\mathcal{C}_3(L) - \beta_2^3\mathcal{C}_3(E_X) \neq 0. \end{aligned} \tag{14}$$

Interestingly, by using higher-order cumulants, the term $\alpha_1^3\mathcal{C}_3(L)$ can be obtained by the following equation through the estimated $\frac{\alpha_2}{\alpha_1}$ and $\frac{\beta_2}{\beta_1}$:

$$\alpha_1^3\mathcal{C}_3(L) = \frac{\mathcal{C}_{1,2}(X,Y) - \frac{\beta_2}{\beta_1}\times\mathcal{C}_{2,1}(X,Y)}{\frac{\alpha_2}{\alpha_1}\times(\frac{\alpha_2}{\alpha_1}-\frac{\beta_2}{\beta_1})}, \tag{15}$$

where $\frac{\alpha_2}{\alpha_1} \neq \frac{\beta_2}{\beta_1}$. The methods of estimating $\frac{\alpha_2}{\alpha_1}$ and $\frac{\beta_2}{\beta_1}$ will be introduced in the next section. Similarly, $\beta_1^3\mathcal{C}_3(E_X)$, $\alpha_2^3\mathcal{C}_3(L)$, and $\beta_2^3\mathcal{C}_3(E_X)$ can be obtained. Roughly speaking, we can determine the causal relationship only by utilizing the higher-order cumulants. Note that the insights into inferring causal relationships between two observed variables can be attained not only through third-order cumulants but also by harnessing cumulant orders higher than three.

Proposed Method

Parameters Estimation

To obtain $\mathcal{R}_{X\to Y}$ and $\mathcal{R}_{Y\to X}$, we need to first estimate the parameters on the ratio of mixing coefficients $\frac{\alpha_2}{\alpha_1}$ and $\frac{\beta_2}{\beta_1}$.

Given the causal structure in Figure 1(b), we can obtain the fifth-order (joint) cumulants of X and Y as:

$$\begin{aligned} \mathcal{C}_{4,1}(X,Y) &= \alpha_1^4\alpha_2\mathcal{C}_5(L) + \beta_1^4\beta_2\mathcal{C}_5(E_X), \\ \mathcal{C}_{3,2}(X,Y) &= \alpha_1^3\alpha_2^2\mathcal{C}_5(L) + \beta_1^3\beta_2^2\mathcal{C}_5(E_X), \\ \mathcal{C}_{2,3}(X,Y) &= \alpha_1^2\alpha_2^3\mathcal{C}_5(L) + \beta_1^2\beta_2^3\mathcal{C}_5(E_X), \\ \mathcal{C}_{1,4}(X,Y) &= \alpha_1\alpha_2^4\mathcal{C}_5(L) + \beta_1\beta_2^4\mathcal{C}_5(E_X). \end{aligned} \tag{16}$$

Note that in the above equations, α_1, α_2, β_1 and β_2 are not zero. Let $\theta_\alpha = \frac{\alpha_2}{\alpha_1}$ and $\theta_\beta = \frac{\beta_2}{\beta_1}$. Then, by combining the above equations, we can obtain θ_α and θ_β by an analytical solution of the following quadratic equation directly:

$$\begin{aligned} &\left(\mathcal{C}_{3,2}(X,Y)^2 - \mathcal{C}_{2,3}(X,Y)\mathcal{C}_{4,1}(X,Y)\right)\theta^2 \\ &+ \left(\mathcal{C}_{1,4}(X,Y)\mathcal{C}_{4,1}(X,Y) - \mathcal{C}_{2,3}(X,Y)\mathcal{C}_{3,2}(X,Y)\right)\theta \\ &+\mathcal{C}_{2,3}(X,Y)^2 - \mathcal{C}_{1,4}(X,Y)\mathcal{C}_{3,2}(X,Y) \\ &=0. \end{aligned} \tag{17}$$

The solution to the above equation is as follows:

$$a\theta^2 + b\theta + c = 0 \Rightarrow \theta^* = \frac{-b\pm\sqrt{b^2-4ac}}{2a}, \tag{18}$$

where $a = \mathcal{C}_{3,2}(X,Y)^2 - \mathcal{C}_{2,3}(X,Y)\mathcal{C}_{4,1}(X,Y) \neq 0$, $b = \mathcal{C}_{1,4}(X,Y)\mathcal{C}_{4,1}(X,Y) - \mathcal{C}_{2,3}(X,Y)\mathcal{C}_{3,2}(X,Y)$, and

$c = \mathcal{C}_{2,3}(X,Y)^2 - \mathcal{C}_{1,4}(X,Y)\mathcal{C}_{3,2}(X,Y)$. Practically, we cannot obtain the exact value of θ_α and θ_β, but they are one of the element in $\frac{-b+\sqrt{b^2-4ac}}{2a}$, $\frac{-b-\sqrt{b^2-4ac}}{2a}$ and $\theta_\alpha \neq \theta_\beta$. This result can still help us infer causal relationships.

The process of deriving a quadratic equation is provided in the appendix. Note that other cumulants higher than order 5 can also be used to estimate parameters. Thus, we can conclude the theorem of parameters estimation by using generalized order cumulants as follows:

Theorem 2. *Assume that two observed variables X and Y are generated by Eq. (4). Suppose there exists $i, j > 0$, such that $\mathcal{C}_{i+3,j}(X,Y)$, $\mathcal{C}_{i+2,j+1}(X,Y)$, $\mathcal{C}_{i+1,j+2}(X,Y)$ and $\mathcal{C}_{i,j+3}(X,Y)$ are not equal to zero. If X and Y are affected by two independent components with different ratios of mixing coefficients, then the ratio of mixing coefficients θ_1 and θ_2 of two independent components on X and Y can be estimated by the analytical solution of the following equation:*

$$\begin{aligned}&\left(\mathcal{C}_{i+2,j+1}(X,Y)^2 - \mathcal{C}_{i+1,j+2}(X,Y)\mathcal{C}_{i+3,j}(X,Y)\right)\theta^2\\ &+\mathcal{C}_{i,j+3}(X,Y)\mathcal{C}_{i+3,j}(X,Y)\theta\\ &-\mathcal{C}_{i+1,i+2}(X,Y)\mathcal{C}_{i+2,j+1}(X,Y)\theta\\ &+\mathcal{C}_{i+1,j+2}(X,Y)^2 - \mathcal{C}_{i,j+3}(X,Y)\mathcal{C}_{i+2,j+1}(X,Y)\\ &=0.\end{aligned} \tag{19}$$

Note that θ_1 and θ_2 can be obtained using cumulant orders higher than five. However, this study demonstrates the minimum required higher-order cumulant for parameter estimation, highlighting that at least fifth-order cumulants are necessary for accurate parameter estimation.

Because Eq. (19) is a quadratic equation, it would fail when $\mathcal{C}_{i+2,j+1}(X,Y)^2 - \mathcal{C}_{i+1,j+2}(X,Y)\mathcal{C}_{i+3,j}(X,Y) = 0$. Interestingly, under this condition, these two observed variables are affected by the same latent variable, and they are not the cause of each other. This condition is also a specific case in Eq. (2), as demonstrated in the following section. Besides, when there is no edge between X and Y influenced by one latent variable, we can estimate the mixing coefficients of the latent variable on X and Y, denoted as $\hat{\alpha}_1$ and $\hat{\alpha}_2$, by the method proposed in the work (Cai et al. 2023):

$$\begin{aligned}\hat{\alpha}_1 &= \sqrt{\frac{\mathcal{C}_{i+1,j}(X,Y)}{\mathcal{C}_{i,j+1}(X,Y)} \cdot \mathcal{C}_{1,1}(X,Y)},\\ \hat{\alpha}_2 &= \frac{\mathcal{C}_{1,1}(X,Y)}{\hat{\alpha}_1}.\end{aligned} \tag{20}$$

Identifiability

Based on the above analysis, we provide the identification results for the causal structure over two observed variables with one latent variable in this section.

Considering two observed variables that are affected by one latent confounder, we can detect whether there is a directed edge between them by the following theorem.

Theorem 3. *Assume that two observed variables X and Y are generated by Eq. (4), and X and Y are affected by the same latent variable L. Suppose there exists $i, j > 0$, such that $\mathcal{C}_{i,j+2}(X,Y)$, $\mathcal{C}_{i+1,j+1}(X,Y)$ and $\mathcal{C}_{i+2,j}(X,Y)$ are not equal to zero. Then there is no directed edge between X and Y, if and only if $\mathcal{C}_{i+1,j+1}(X,Y)^2 - \mathcal{C}_{i,j+2}(X,Y)\mathcal{C}_{i+2,j}(X,Y) = 0$.*

Theorem 3 provides a method for identifying whether there exists a directed edge between two observed variables in the presence of latent variables. If such a directed edge between two variables exists, a further question is how to identify the direction of the edge.

Fortunately, from the intuition provided in the previous section, we can solve this problem. Because we have no idea of the causal direction, we introduce S_1 and S_2 to be two shared independent components of X and Y. Let $\mathcal{R}_{X\to Y} = \mathcal{C}_3(X) - \alpha_1^3\mathcal{C}_3(S_1) - \beta_1^3\mathcal{C}_3(S_2)$ (and $\mathcal{R}_{Y\to X} = \mathcal{C}_3(Y) - \alpha_2^3\mathcal{C}_3(S_1) - \beta_2^3\mathcal{C}_3(S_2)$) be the causal direction criteria. Then we can obtain the following rule: if and only if $\mathcal{R}_{X\to Y} = 0$, then X is a parent of Y and both of them are affected by latent confounder. This rule can be summarized as:

Theorem 4. *Assume that two observed variables X and Y are generated by Eq. (4), and X and Y are affected by the same latent variable L. Then X is a cause of Y if and only if $\mathcal{R}_{X\to Y} = 0$.*

Theorem 4 provides a method for causal direction identification, achieved by the higher-order cumulants. Combining Theorem 3 and Theorem 4, we can identify the causal relationship between two observed variables that are affected by a latent confounder, which is guaranteed by the following theorem.

Theorem 5. *Assume that two observed variables X and Y are generated by Eq. (4), and X and Y are affected by the same latent variable L. The causal relationship between two observed variables can be identified by using the higher-order cumulants.*

The principles outlined in Theorem 5 provide a foundation for identifying causal structures within a two-variable context. It can be extended to scenarios involving more than two variables by transforming the causal structure into a canonical model, where each latent variable is the parent of two observed variables. Consequently, the criterion for identifying causal structures encompassing multiple variables stipulates that any pair of observed variables can be influenced by a maximum of one latent variable.

Learning Algorithm

Based on the identifiability results, we now consider a practical method for inferring causal structure from observed data with latent confounders.

We begin by considering the case where two observed variables X and Y with one latent confounder. Denote the statistic $s = \mathcal{C}_{3,2}(X,Y)^2 - \mathcal{C}_{2,3}(X,Y)\mathcal{C}_{4,1}(X,Y)$ as Theorem 3. It is straightforward that, first, test whether s is equal to zero to identify the presence of a directed edge between them. If such an edge exists, we further determine its direction by using Theorem 4.

In practice, the statistic wouldn't equal zero exactly, since the sample size is finite. To test whether the statistic s is equal to zero, we provide a procedure as following steps:

1) Sample data of size m from the original dataset $(\mathbf{X}, \mathbf{Y})$ with replacement.

2) Compute the statistic s.

3) Repeat steps 1 and 2 for B times to obtian $\{s^*_{m,1}, \ldots, s^*_{m,B}\}$.

4) Perform one sample sign test (Dixon and Mood 1946) on $\{s^*_{m,1}, \ldots, s^*_{m,B}\}$.

Similarly, we can use the above procedure to test whether $R_{X\rightarrow Y}$ is equal to zero. The reason we do this is that Non-Gaussian distribution contains a wide class of distribution, we can not find suitable statistics to capture the information of distribution without any prior information of distribution. Further, we conjecture the median can be asymptotically approximated to the mean, so we use the sign test to test whether the median of $\{s^*_{m,1}, \ldots, s^*_{m,B}\}$ is equal to zero.

The above procedure will result in one of several possible scenarios. First, if X and Y are only affected by a latent confounder, then we can infer that there is no directed edge between the two, and no further analysis is performed. Second, if there exists a directed edge between them, but both directional edges are accepted. We conclude that either model may be correct but we cannot infer it from the data. The positive result is when we are able to reject one of the directions and accept the other.

Simulations

Experiment Setups

To evaluate the performance of our proposed method, we conducted experiments on synthetic data, considering the following two cases:

[Case 1]: A causal graph over two observed variables that are affected by a latent variable, and there is no causal directed edge between observed variables, i.e., Figure 1(a).

[Case 2]: A causal graph over two observed variables that are affected by a latent variable, and there is a causal directed edge between observed variables, i.e., Figure 1(b).

The purposes of these experiments are: (1) Figure 1(a) is used to evaluate the performance on inferring the existence of the causal directed edge of different methods; (2) Figure 1(b) is used to evaluate the performance on inferring the direction of the causal directed edge of different methods.

Given the causal graphs, we randomly generated data according to Eq. (4), where the causal coefficient is sampled from a uniform distribution between $[0.8, 1.2]$. The noise terms are generated from three distinct distributions: Exponential Distribution, Gamma Distribution with shape parameter $k = 3$ and Gumbel distribution. For each model, the sample size N is varied among $\{5000, 10000, 50000, 100000\}$. For each setting, we generate 100 datasets.

Baseline Methods and Evaluation Metrics

In these experiments, we use Linear Non-Gaussian Acyclic Model (LiNGAM) (Shimizu et al. 2006), Additive Noise Model (ANM) (Hoyer et al. 2008a), LvLiNGAM (Hoyer et al. 2008b) and pairwise LvLiNGAM (Entner and Hoyer 2011) as the baseline methods. Among these methods, LiNGAM and ANM are two typical methods under the causal sufficiency assumption. LiNGAM leverages the ICA technique to estimate the mixing matrix and transformed it into the causal strength matrix, while ANM leverages the independence between the regression residual and the assumed cause to determine the causal direction. LvLiNGAM and Pairwise LvLiNGAM are two methods against the latent variables. Pairwise LvLiNGAM aims to capture partial information regarding causal relationships between variable pairs within data generated by LvLiNGAM. However, pairwise LvLiNGAM does not yield informative results when applied to synthetic data. Consequently, we only provide the results of LiNGAM, ANM, LvLiNGAM, and our proposed methods. For the test of our approach, we set $m = 0.8n$ and $B = 30$, where n is the sample size of the initial dataset.

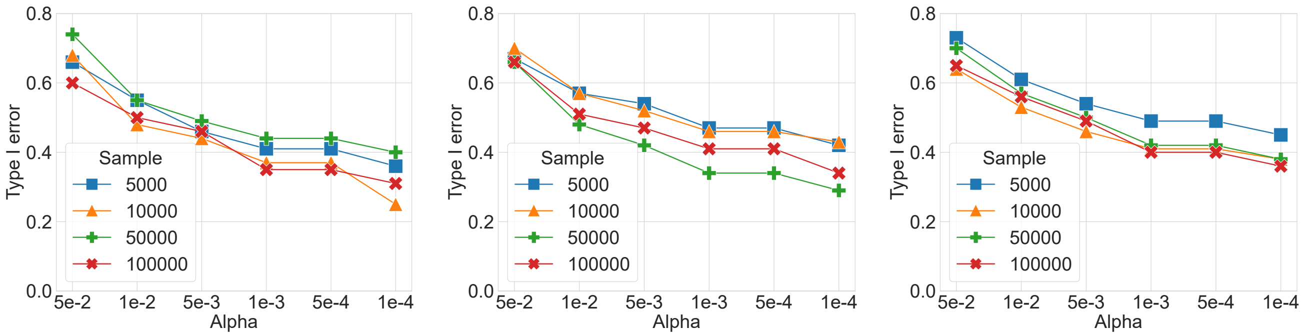
To evaluate the performance of different methods, we employ the accuracy score that the learned graph is the same as the true one as an evaluation metric. Besides, we also examine the probabilities of Type I error and Type II error of our hypothesis test procedure within specific simulation scenarios as sample size and significance level (α =0.05, 0.01, 0.005, 0.001, 0.0005 and 0.0001) change.

Experimental Results

Evaluation in Case 1. The experimental results of case 1 are given in Table 1. In this case, we just compare the result between LvLiNGAM and our methods due to ANM can not infer the existence of the causal directed edge. From the results, LvLiNGAM is unable to return the truth that there is no causal edge between observed variables. Because LvLiNGAM's performance depends on Overcomplete Independent Components Analysis, which usually gets stuck in local optima, it would infer redundant causal edges in practice. Our proposed method can determine that there is no edge between two observed variables in most cases, although the accuracy is not so high. These results rely on the sample size and the test method in practice, which are shown in the results on the Type I error and Type II errors.

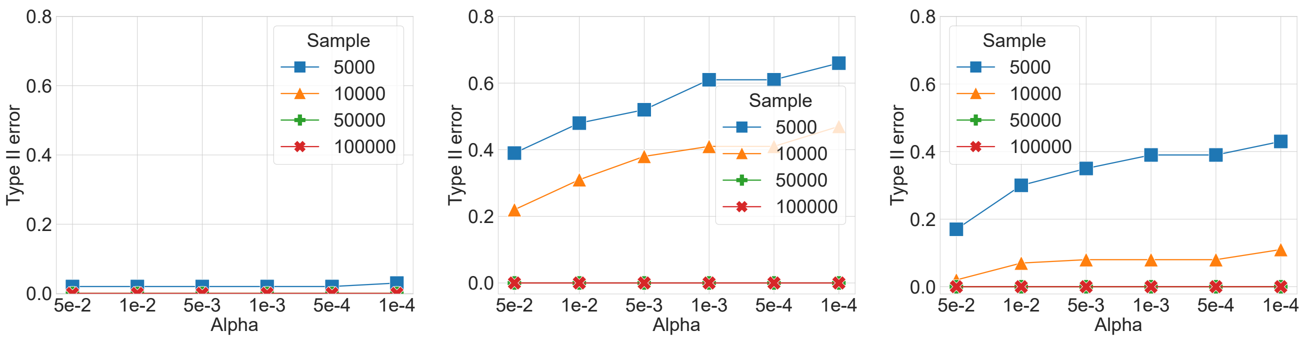
To examine the Type I error, we generated X and Y according to Figure 1(a) in which there exists no causal directed edge between X and Y. Figure 2 and Figure 3 show the resulting probability of Type I errors and that of Type II errors at different significance levels, respectively. From Figure 2, we find that the Type I error is around 0.3 under the significance level $\alpha = 0.0001$. The Type II error of our proposed method is sensitive to sample size. In detail, the Type II error is almost equal to zero even under the significance level $\alpha = 0.0001$ when the sample size is larger than 50000. So we use the results with a significance level of 0.0001 to compare baseline methods.

Evaluation in Case 2. The results of Case 2 are shown in Table 2. For large sample sizes, HSIC is not applicable because of its high time consumption and memory consumption, so ANM cannot return any result when the sample sizes are 50000 and 100000. For fairness, we compare the p-value of the hypothesis test in both directions for ANM and our method, which direction of p-value is lower is the correct one. From the results, we can see that ANM and LvLiNGAM obtain an accuracy score of around 0.5 in all



(a) Exponential Distribution (b) Gamma Distribution (c) Gumbel Distribution

Figure 2: The probability of Type I error varies across different distributions in the test for the presence of a causal edge.



(a) Exponential Distribution (b) Gamma Distribution (c) Gumbel Distribution

Figure 3: The probability of Type II error varies across different distributions in the test for the presence of a causal edge.

	Sample	LiNGAM	ANM	LvLiNGAM	Ours
Exp.	5000	0.00	-	0.00	**0.64**
	10000	0.00	-	0.00	**0.75**
	50000	0.00	-	0.00	**0.60**
	100000	0.00	-	0.00	**0.69**
Gam.	5000	0.00	-	0.00	**0.58**
	10000	0.00	-	0.00	**0.57**
	50000	0.00	-	0.00	**0.71**
	100000	0.00	-	0.00	**0.66**
Gum.	5000	0.00	-	0.00	**0.55**
	10000	0.00	-	0.00	**0.62**
	50000	0.00	-	0.00	**0.62**
	100000	0.00	-	0.00	**0.64**

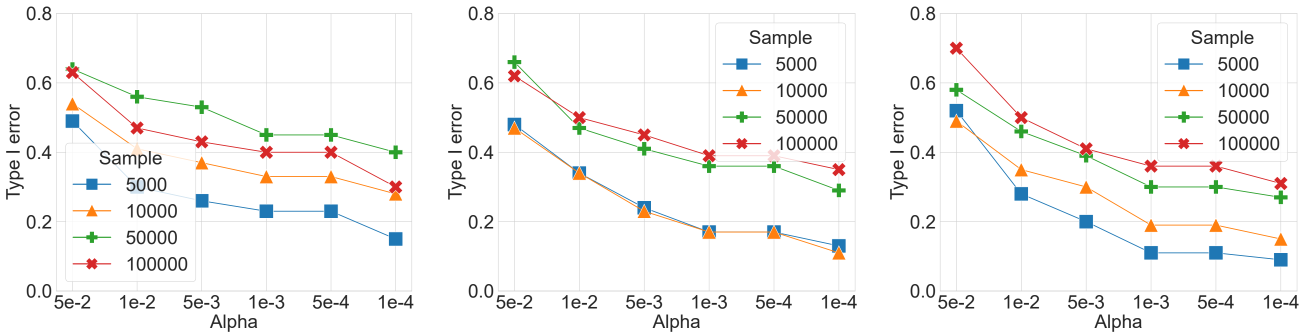
Table 1: Accuracy of the different methods varies across different distributions in Case 1. Here "-" denotes that the corresponding methods are not applicable.

	Sample	LiNGAM	ANM	LvLiNGAM	Ours
Exp.	5000	0.05	0.47	0.50	**0.79**
	10000	0.05	0.13	0.66	**0.88**
	50000	0.06	-	0.61	**0.91**
	100000	0.07	-	0.65	**0.92**
Gam.	5000	0.07	0.38	0.43	**0.55**
	10000	0.04	0.51	0.58	**0.71**
	50000	0.03	-	0.52	**0.85**
	100000	0.02	-	0.61	**0.89**
Gum.	5000	0.13	0.51	0.53	**0.70**
	10000	0.10	0.43	0.59	**0.74**
	50000	0.13	-	0.51	**0.90**
	100000	0.05	-	0.61	**0.92**

Table 2: Accuracy of the different methods varies across different distributions in Case 2. Here "-" denotes that the corresponding methods take too long to produce any results.

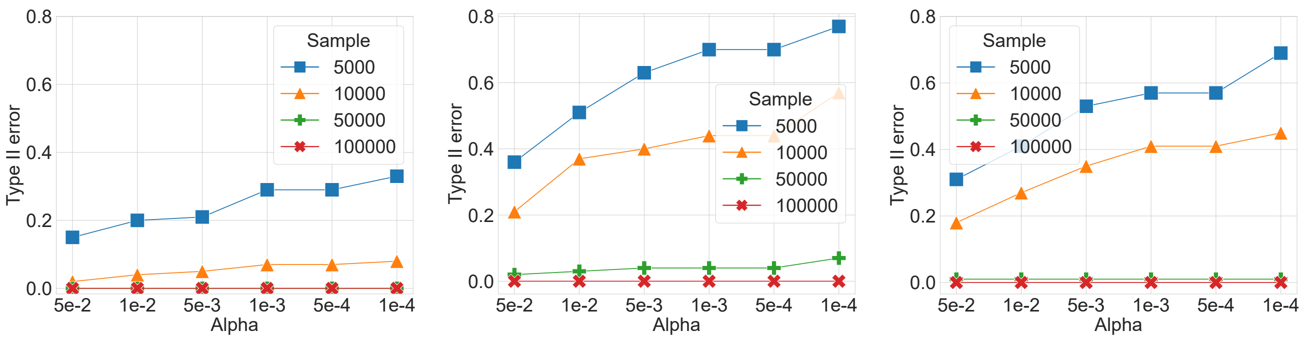
cases with varying sample sizes. As ANM does not take latent variables into account, it is unable to distinguish the causal direction. In the case of LvLiNGAM, its efficacy is intertwined with the Overcomplete ICA technique. However, this approach often encounters challenges by becoming trapped in local optima, leading to inaccurate results.

Besides, Figure 4 - 5 show the result of the probability of Type I errors and that of Type II errors at different significance levels, respectively. The results indicate that the Type I error increases when the sample size increases. The reason is that the hypothesis test tends not to reject the null hypothesis, which also leads to a high Type II error result.



(a) Exponential Distribution (b) Gamma Distribution (c) Gumbel Distribution

Figure 4: The probability of Type I error varies across different distributions in the test for the direction of the causal edge.



(a) Exponential Distribution (b) Gamma Distribution (c) Gumbel Distribution

Figure 5: The probability of Type II error varies across different distributions in the test of the direction of the causal edge.

Furthermore, we applied our proposed method to the datasets that are generated by nonlinear causal relationships, to relax the linear assumption. When considering an exponential noise distribution and a sample size of 100,000, our method achieved an accuracy of 0.84 in determining the causal direction between two observed variables. This demonstrates the robustness and potential applicability of our approach in scenarios beyond linear relationships.

Discussion and Conclusion

In this paper, we provide the identifiability theories for inferring causal relationships between two observed variables with latent variables, by utilizing higher-order cumulants. Based on these identifiability theories, we derive a method for causal discovery from observational data in the presence of latent variables. Specifically, it first detects whether there exists a causal edge between two observed variables. Then if such a causal edge exists, it determines the direction of the causal edge.

Compared with existing methods, the power of the identifiability results provided in this paper depends on the information from higher-order cumulants of non-Gaussian data. Interestingly, one can find the relations between our proposed criterion given in Eq. (2) is similar to tetrad constraint (Silva et al. 2006) under the non-Gaussian assumption. The tetrad constraint is satisfied among four pure observed variables that are affected by one latent variable. Note that in Eq. (2), when $i = 1$ and $j = 1$, one have $Cum(X, X, Y, Y)^2 = Cum(X, X, X, Y)Cum(X, Y, Y, Y)$. The combination of joint cumulants of X and Y can be used to eliminate the influence of the shared latent variable on them. To some extent, the satisfaction of this condition is facilitated by the non-Gaussian noise assumption and higher-order cumulants.

The experimental results reveal that our proposed method achieves favorable performance, particularly with larger sample sizes (around 100,000). This also reflects that the test method requires a large sample size to approximate the true joint cumulants of the variables. Notably, if we can find a distribution to approximate the combination of joint cumulants during the test process, it is plausible to design a more dependable test method that does not necessitate stringent sample size requirements. Moreover, one might consider the linear assumption to be overly restrictive and unsuitable for real-world scenarios. However, the experimental results obtained from data generated by additive nonlinear relationships demonstrate the potential applicability of our method even in nonlinear cases. If a more effective test can be devised, even with a limited sample size, it would facilitate the extension of our method to high-dimensional scenarios in practical applications. This will be the research direction of our next work.

Acknowledgments

We sincerely appreciate the insightful discussions from Shenlong Pan and anonymous reviewers, which greatly helped to improve the paper. This research was supported in part by National Key R&D Program of China (2021ZD0111501), Natural Science Foundation of China (62206064, 61876043, 61976052, 62206061), National Science Fund for Excellent Young Scholars (62122022), the major key project of PCL (PCL2021A12), Guangzhou Basic and Applied Basic Research Foundation (2023A04J1700). KZ was supported in part by the NSF-Convergence Accelerator Track-D award #2134901, by the National Institutes of Health (NIH) under Contract R01HL159805, by grants from Apple Inc., KDDI Research, Quris AI, and IBT, and by generous gifts from Amazon, Microsoft Research, and Salesforce.

References

Bekker, P. A.; and ten Josephus Berge. 1997. Generic global indentification in factor analysis. *Linear Algebra and its Applications*, 264: 255–263.

Brillinger, D. R. 2001. *Time series: data analysis and theory*. Society for Industrial and Applied Mathematics.

Cai, R.; Huang, Z.; Chen, W.; Hao, Z.; and Zhang, K. 2023. Causal Discovery with Latent Confounders Based on Higher-Order Cumulants. *International conference on machine learning*.

Chen, W.; Cai, R.; Zhang, K.; and Hao, Z. 2021. Causal Discovery in Linear Non-Gaussian Acyclic Model With Multiple Latent Confounders. *IEEE Transactions on Neural Networks and Learning Systems*.

Colombo, D.; Maathuis, M. H.; Kalisch, M.; and Richardson, T. S. 2012. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, 40(1): 294 – 321.

Dixon, W. J.; and Mood, A. M. 1946. The statistical sign test. *Journal of the American Statistical Association*, 41(236): 557–566.

Entner, D.; and Hoyer, P. O. 2011. Discovering unconfounded causal relationships using linear non-gaussian models. In *New Frontiers in Artificial Intelligence: JSAI-isAI 2010 Workshops, LENLS, JURISIN, AMBN, ISS, Tokyo, Japan, November 18-19, 2010, Revised Selected Papers 2*, 181–195. Springer.

Eriksson, J.; and Koivunen, V. 2004. Identifiability, separability, and uniqueness of linear ICA models. *IEEE Signal Processing Letters*, 11(7): 601–604.

Hoyer, P.; Janzing, D.; Mooij, J. M.; Peters, J.; and Schölkopf, B. 2008a. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21.

Hoyer, P. O.; Shimizu, S.; Kerminen, A. J.; and Palviainen, M. 2008b. Estimation of causal effects using linear non-Gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49(2): 362–378.

Huang, B.; Low, C. J. H.; Xie, F.; Glymour, C.; and Zhang, K. 2022. Latent hierarchical causal structure discovery with rank constraints. *Advances in Neural Information Processing Systems*, 35: 5549–5561.

Hyvärinen, A.; Karhunen, J.; and Oja, E. 2001. *Independent Component Analysis*. Wiley.

Kummerfeld, E.; and Ramsey, J. 2016. Causal clustering for 1-factor measurement models. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1655–1664.

Ledermann, W. 1937. On the rank of the reduced correlational matrix in multiple-factor analysis. *Psychometrika*, 2: 85–93.

Lewicki, M. S.; and Sejnowski, T. J. 2000. Learning overcomplete representations. *Neural Computation*, 12(2): 337–365.

Maeda, T. N.; and Shimizu, S. 2020. RCD: Repetitive causal discovery of linear non-Gaussian acyclic models with latent confounders. In *International Conference on Artificial Intelligence and Statistics*, 735–745. PMLR.

Morioka, H.; and Hyvarinen, A. 2023. Connectivity-contrastive learning: Combining causal discovery and representation learning for multimodal data. In *International Conference on Artificial Intelligence and Statistics*, 3399–3426. PMLR.

Sachs, K.; Perez, O.; Pe'er, D.; Lauffenburger, D. A.; and Nolan, G. P. 2005. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721): 523–529.

Shimizu, S.; Hoyer, P. O.; Hyvärinen, A.; Kerminen, A.; and Jordan, M. 2006. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10).

Silva, R.; Scheines, R.; Glymour, C.; Spirtes, P.; and Chickering, D. M. 2006. Learning the Structure of Linear Latent Variable Models. *Journal of Machine Learning Research*, 7(2).

Spirtes, P.; Meek, C.; and Richardson, T. 1995. Causal inference in the presence of latent variables and selection bias. In *Conference on Uncertainty in artificial intelligence*, 499–506.

Tashiro, T.; Shimizu, S.; Hyvärinen, A.; and Washio, T. 2014. ParceLiNGAM: A causal ordering method robust against latent confounders. *Neural computation*, 26(1): 57–83.

Tramontano, D.; Monod, A.; and Drton, M. 2022. Learning linear non-Gaussian polytree models. In *Uncertainty in Artificial Intelligence*, 1960–1969. PMLR.

Wang, Y. S.; and Drton, M. 2020. High-dimensional causal discovery under non-Gaussianity. *Biometrika*, 107(1): 41–59.

Xie, F.; Cai, R.; Huang, B.; Glymour, C.; Hao, Z.; and Zhang, K. 2020. Generalized independent noise condition for estimating latent variable causal graphs. *Advances in neural information processing systems*, 33: 14891–14902.

Xie, F.; Huang, B.; Chen, Z.; He, Y.; Geng, Z.; and Zhang, K. 2022. Identification of linear non-gaussian latent hierarchical structure. In *International Conference on Machine Learning*, 24370–24387. PMLR.