Journal of Machine Learning Research 21 (2020) 1-24 Submitted 4/19; Revised 11/19; Published 3/20

Learning Linear Non-Gaussian Causal Models in the Presence of Latent Variables

Saber Salehkaleybar SALEH@SHARIF.EDU
Department of Electrical Engineering
Sharif University of Technology, Tehran, Iran

AmirEmad Ghassami GHASSAM2@ILLINOIS.EDU
Department of ECE
University of Illinois at Urbana-Champaign
Urbana, IL 61801

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

Kun Zhang KUNZ1@CMU.EDU
Department of Philosophy
Carnegie Mellon University
Pittsburgh, PA 15213

Editor: Isabelle Guyon

Abstract

We consider the problem of learning causal models from observational data generated by linear non-Gaussian acyclic causal models with latent variables. Without considering the effect of latent variables, the inferred causal relationships among the observed variables are often wrong. Under faithfulness assumption, we propose a method to check whether there exists a causal path between any two observed variables. From this information, we can obtain the causal order among the observed variables. The next question is whether the causal effects can be uniquely identified as well. We show that causal effects among observed variables cannot be identified uniquely under mere assumptions of faithfulness and non-Gaussianity of exogenous noises. However, we are able to propose an efficient method that identifies the set of all possible causal effects that are compatible with the observational data. We present additional structural conditions on the causal graph under which causal effects among observed variables can be determined uniquely. Furthermore, we provide necessary and sufficient graphical conditions for unique identification of the number of variables in the system. Experiments on synthetic data and real-world data show the effectiveness of our proposed algorithm for learning causal models.

Keywords: Causal Discovery, Structural Equation Models, Non-Gaussianity, Latent Variables, Independent Component Analysis.

1. Introduction

One of the primary goals in empirical sciences is to discover causal relationships among a set of variables of interest in various natural and social phenomena. Such causal relationships can be recovered by conducting controlled experiments. However, performing controlled experiments is often expensive or even impossible due to technical or ethical reasons. Thus, it is vital to develop statistical methods for recovering causal relationships from non-experimental data.

Probabilistic graphical models are commonly used to represent causal relations. Alternatively, Structural Equation Models (SEM) which further specify mathematical equations among the variables can be used to represent probabilistic causal influences. Linear SEMs are a special class of SEMs where each variable is a linear combination of its direct causes and an exogenous noise. Since non-linear approaches often need large sample sizes to produce reliable results (see explanations in (Shimizu, 2014), Section 2.5) and the relationships between variables are approximately linear in many situations (after preprocessing, if needed), linear SEMs are widely used in practice, on the raw data or preprocessed data with proper nonlinear transformations. Under the causal sufficiency assumption, by utilizing conventional causal structure learning algorithms such as PC (Spirtes et al., 2000) and IC (Pearl, 2009), we can identify a class of models that are equivalent in the sense that they represent the same set of conditional independence assertions obtained from data. If we have background knowledge about the data-generating mechanism, we may further narrow down the possible models that are compatible with the observed data (Peters et al., 2016; Ghassami et al., 2018; Salehkaleybar et al., 2018; Zhang et al., 2017; Peters and Bühlmann, 2013; Zhang and Hyvärinen, 2009; Salehkaleybar et al., 2017; Hoyer et al., 2009; Janzing et al., 2012). For instance, Shimizu et al. (2006) proposed a linear non-Gaussian acyclic model (LiNGAM) discovery algorithm that can identify causal structure uniquely, thanks to the assumption of non-Gaussian distributions for the exogenous noises in the linear SEM model. However, LiNGAM algorithm and its regression-based variant (DirectLiNGAM) (Shimizu et al., 2011) rely on the causal sufficiency assumption, i.e., no unobserved common causes exist for any pair of variables that are under consideration in the model.

In the presence of latent variables, Hoyer et al. (2008) showed that linear SEM can be converted to a canonical form where each latent variable has at least two children and no parents. Such latent variables are commonly called "latent confounders". Furthermore, they proposed a solution which casts the problem of identifying causal effects among observed variables into an overcomplete Independent Component Analysis (ICA) problem (Hyvärinen et al., 2004) and returns multiple causal structures that are observationally equivalent. The time complexity of searching such structures can be as high as $\binom{p}{p_o}$ where p_o and p are the number of observed and total variables in the system, respectively. Entner and Hoyer (2010) proposed a method that identifies a partial causal structure among the observed variables by recovering all the unconfounded sets[1] and then learning the causal effects for each pair of variables in the set. However, their method may return an empty unconfounded set if latent confounders are the cause of most of observed variables in the system such as the simple example of Figure 1. Chen and Chan (2013) showed that a causal order and causal effects among observed variables can be identified if the latent confounders have Gaussian distribution and exogenous noises of observed variables are simultaneously super-Gaussian or sub-Gaussian. In (Tashiro et al., 2014), the ideas in DirectLiNGAM was extended to the case where latent confounders exist in the system. The proposed solution first tries to find a root variable (a variable with no parents). Then, the effect of such variable is removed by regressing it out. This procedure continues until any variable and its residual becomes dependent. Subsequently, a similar iterative procedure is used to find a sink variable and remove its effect from other variables. How-

1. A set of variables is called unconfounded if there is no variable outside the set which is confounder of some variables in the set. In Figure 1, variable V_3 is a confouder of variables V_1 and V_2 but it is not observable. Thus, the set of variables V_1 and V_2 is not unconfounded.

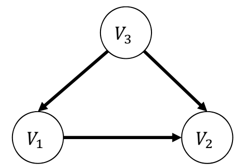

Figure 1: An example of causal graphs: V_1 and V_2 are observed variables while V_3 is latent.

ever, this solution may not recover causal order in some causal graphs such as the one in Figure 1.[2] Shimizu and Bollen (2014) proposed a Bayesian approach for estimating the causal direction between two observed variables when the sum of non-Gaussian independent latent confounders has a multivariate t-distribution. They compute log-marginal likelihoods to infer causal directions.

There exist work in the literature that tried to recover causal structure among observed variables in the presence of latent variables for the settings other than linear non-Gaussian model. In general cases, Spirtes et al. (2000) proposed Fast Causal Inference (FCI) algorithm that can identify some causal paths in the presence of latent variables by performing conditional independence test without assuming constraints on the causal mechanism (e.g., linearity). However, it cannot guarantee the existence of causal paths in some cases such as the one where a pair of observed variables has a direct causal influence from one to the other and there is also a confounder for them. Elidan and Friedman (2005) proposed a method to learn Bayesian networks with latent variables based on information bottleneck concept. In the proposed method, the structure of network is learnt for a given number of hidden variables by a scored based approach with a structural expectation maximization approach. In the literature of exploratory factor analysis, there exist work such as (Jennrich and Bentler, 2011), which proposed a bi-factor analysis for the case with at most two latent variables in the system. In the field of Markov random model, Chandrasekaran et al. (2010) considered Gaussian Markov random field model with latent variables and tried to identify conditional independences between observed variables given all variables in the system by considering a sparsity assumption on the conditional graphical model between the observed variables. Spirtes et al. (1995) utilized an extension of "Verma constraints" to learn causal strcutures in nested Markov models with latent variables. Kummerfeld and Ramsey (2016) proposed a method to learn causal structure by examining the rank of submatrices of correlation matrix for the specific class of measurement model where each observed variable has exactly one latent parent.

Rather surprisingly, although the causal structure is in general not fully identifiable in the presence of latent variables, we will show that the causal order among the observed variables is still identifiable under the faithfulness assumption. In order to obtain a causal order, we first check whether there exists a causal path between any two observed variables. Subsequently, from this information, we obtain a causal order among them. Having established a causal order, we aim to figure out whether the causal effects are uniquely identifiable from observational data. We show by an example that causal effects among observed variables is not uniquely identifiable even if the faithfulness assumption holds true and the exogenous noises are non-Gaussian. We propose a method to identify the set of all possible causal effects efficiently in time that are compatible with the observational data. Furthermore, we present some structural conditions on the causal graph un-

2. In Figure 1, the root variable (V_3) is latent and the regressor of sink variable V_2 and the residual are not independent without considering the latent variable V_3 in the set of regressors. Thus, no root or sink variable can be identified in the system.

der which causal effects among the observed variables can be identified uniquely. We also provide necessary and sufficient graphical conditions under which the number of latent variables is uniquely identifiable. One of the applications of determining the number of latent variables from the observational data is in psychometrics, where the analysis of testing data often requires to estimate how many latent variables, the items are measuring (Silva and Scheines, 2005; Kummerfeld and Ramsey, 2016).

The rest of this paper is organized as follows. In Section 2, we define the problem of identifying causal orders and causal effects in linear causal systems with latent variables. In Section 3, we propose our approach to learn the causal order among the observed variables and provide necessary and sufficient graphical conditions under which the number of latent variables is uniquely identifiable. In Section 4, we present a method to find the set of all possible causal effects which are consistent with the observational data and give conditions under which causal effects are uniquely identifiable. We conduct experiments to evaluate the performance of proposed solutions in Section 5 and conclude in Section 6.

2. Problem Definition

2.1. Notations

In a directed graph $G = (\mathcal{V}, E)$ with the vertex set $\mathcal{V} = \{V_1, \cdots, V_p\}$ and the edge set E, we denote a directed edge from V_i to V_j by (V_i, V_j). A directed path $P = (V_{i_0}, V_{i_1}, \cdots, V_{i_k})$ in G is a sequence of vertices of G where there is a directed edge from V_{i_j} to $V_{i_{j+1}}$ for any $0 \leq j \leq k-1$. We define the set of variables $\{V_{i_1}, \cdots, V_{i_{k-1}}\}$ as the intermediate variables on the path P. We say that a path is a latent path if all the intermediate variables on the path are latent. We use notation $V_i \rightsquigarrow V_j$ to show that there exists a directed path from V_i to V_j. If there is a directed path from V_i to V_j, V_i is ancestor of V_j and that V_j is a descendant of V_i. More formally, $anc(V_i) = \{V_j | V_j \rightsquigarrow V_i\}$ and $des(V_i) = \{V_j | V_i \rightsquigarrow V_j\}$. Each variable V_i is an ancestor and a descendant of itself.

We denote vectors and matrices by boldface letters. The vectors $\mathbf{A}_{i,:}$ and $\mathbf{A}_{:,i}$ represent i-th row and column of matrix $\mathbf{A}$, respectively. The (i,j) entry of matrix $\mathbf{A}$ is denoted by $[\mathbf{A}]_{i,j}$. For $n \times m$ matrix $\mathbf{A}$ and $n \times p$ matrix $\mathbf{B}$, the notation $[\mathbf{A}, \mathbf{B}]$ denotes the horizontal concatenation. For $n \times m$ matrix $\mathbf{A}$ and $p \times m$ matrix $\mathbf{B}$, the notation $[\mathbf{A}; \mathbf{B}]$ shows the vertical concatenation.

2.2. System Model

Consider a linear SEM among a set of variables $\mathcal{V} = \{V_1, \cdots, V_p\}$:

$$\mathbf{V} = \mathbf{A}\mathbf{V} + \mathbf{N}, \tag{1}$$

where the vectors $\mathbf{V}$ and $\mathbf{N}$ denote the random variables in $\mathcal{V}$ and their corresponding exogenous noises, respectively. The entry (i,j) of matrix $\mathbf{A}$ shows the strength of direct causal effect of variable V_j on variable V_i. We assume that the causal relations among random variables can be represented by a directed acyclic graph (DAG). Thus, the variables in $\mathcal{V}$ can be arranged in a causal order, such that no latter variable causes any earlier variable. We denote such a causal order on the variables by k in which $k(i), i \in \{1, \cdots, p\}$ shows the position of variable V_i in the causal order. $\mathbf{A}$ can be converted to a strictly lower triangular matrix by permuting its rows and columns simultaneously based on the causal order.

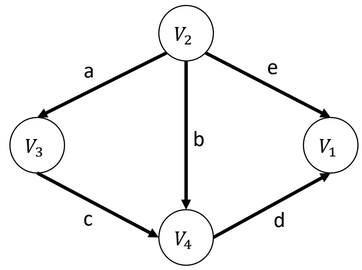

Figure 2: Causal graph of Example 1.

Example 1 *Consider the following linear SEM with four random variables* $\{V_1, \cdots, V_4\}$*:*

$$\begin{bmatrix} V_1 \\ V_2 \\ V_3 \\ V_4 \end{bmatrix} = \begin{bmatrix} 0 & e & 0 & d \\ 0 & 0 & 0 & 0 \\ 0 & a & 0 & 0 \\ 0 & b & c & 0 \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \\ V_3 \\ V_4 \end{bmatrix} + \begin{bmatrix} N_1 \\ N_2 \\ N_3 \\ N_4 \end{bmatrix},$$

where a, b, c, d *and* e *are some constants (see Figure 2). A causal order in this SEM model would be:* $k(1) = 4, k(2) = 1, k(3) = 2, k(4) = 3$. *Hence, matrix* $\mathbf{PAP}^T$ *is strictly lower triangular where* $\mathbf{P}$ *is a permutation matrix associated with* k *defined by the following non-zero entries:* $\{(k(i), i)|1 \leq i \leq 4\}$.

We split random variables in $\mathbf{V}$ into an observed vector $\mathbf{V_o} \in \mathbb{R}^{p_o}$ and a latent vector $\mathbf{V_l} \in \mathbb{R}^{p_l}$ where p_o and p_l are the number of observed and latent variables, respectively. Without loss of generality, we assume that first p_o entries of $\mathbf{V}$ are observable, i.e. $\mathbf{V_o} = [V_1, \cdots, V_{p_o}]^T$ and $\mathbf{V_l} = [V_{p_o+1}, \cdots, V_p]^T$. Therefore,

$$\begin{bmatrix} \mathbf{V_o} \\ \mathbf{V_l} \end{bmatrix} = \begin{bmatrix} \mathbf{A_{oo}} & \mathbf{A_{ol}} \\ \mathbf{A_{lo}} & \mathbf{A_{ll}} \end{bmatrix} \begin{bmatrix} \mathbf{V_o} \\ \mathbf{V_l} \end{bmatrix} + \begin{bmatrix} \mathbf{N_o} \\ \mathbf{N_l} \end{bmatrix}, \tag{2}$$

where $\mathbf{N_o}$ and $\mathbf{N_l}$ are the vectors of exogenous noises of $\mathbf{V_o}$ and $\mathbf{V_l}$, respectively. Furthermore, we have: $\mathbf{A} = [\mathbf{A_{oo}}, \mathbf{A_{ol}}; \mathbf{A_{lo}}, \mathbf{A_{ll}}]$.

The causal order among all variables k, induces a causal order k_o among the observed variables as follows: For any two observed variables V_i, V_j, $1 \leq i, j \leq p_o$, $k_o(i) < k_o(j)$ if $k(i) < k(j)$. Similarly, k induces a causal order among latent variables. We denote this causal order by k_l. It can be easily shown that $\mathbf{A_{oo}}$ and $\mathbf{A_{ll}}$ can be converted to strictly lower triangular matrices by permuting rows and columns simultaneously based on causal orders k_o and k_l, respectively.

Example 2 *In Example 1, suppose that only variables* V_1 *and* V_2 *are observable. Then, the causal order among observed variables would be:* $k_o(1) = 2$ *and* $k_o(2) = 1$. *Thus,* $\mathbf{PA_{oo}P}^T$ *is a strictly lower triangular matrix where* $\mathbf{P} = [0, 1; 1, 0]$. *For the latent variables,* $k_l(3) = 1$ *and* $k_l(4) = 2$.

In the remainder of this section, we briefly describe LiNGAM algorithm, which is capable of recovering the matrix $\mathbf{A}$ uniquely if all variables in the model are observable and exogenous noises are non-Gaussian (Shimizu et al., 2006). The vector $\mathbf{V}$ in Equation (1) can be written as a linear combination of exogenous noises as follows:

$$\mathbf{V} = \mathbf{B}\mathbf{N}, \tag{3}$$

where $\mathbf{B} = (\mathbf{I} - \mathbf{A})^{-1}$. The above equation fits into the standard linear Independent Component Analysis (ICA) framework, where independent non-Gaussian components are all variables in $\mathbf{N}$. By utilizing statistical techniques in ICA (Hyvärinen et al., 2004), matrix $\mathbf{B}$ can be identified up to scaling and permutations of its columns. More specifically, the independent components of ICA as well as the estimated $\mathbf{B}$ matrix are not uniquely determined because permuting and rescaling them does not change their mutual independence. So without knowledge of the ordering and scaling of the noise terms, the following general ICA model for $\mathbf{V}$ holds:

$$\mathbf{V} = \tilde{\mathbf{B}}\tilde{\mathbf{N}}, \tag{4}$$

where $\tilde{\mathbf{N}}$ contains independent components and these components (resp. the columns of $\tilde{\mathbf{B}}$) are a permuted and rescaled version of those in $\mathbf{N}$ (resp. the columns of $\mathbf{B}$). In what follows, we use $\mathbf{B}$ for matrix $\mathbf{B} = (\mathbf{I} - \mathbf{A})^{-1}$ while $\tilde{\mathbf{B}}$ is the mixing matrix for the ICA model, as given in (4). Hence $\tilde{\mathbf{B}}$ can be written as:

$$\tilde{\mathbf{B}} = \mathbf{B}\mathbf{P}\mathbf{\Lambda},$$

where $\mathbf{P}$ is a permutation matrix and $\mathbf{\Lambda}$ is a diagonal scaling matrix. Yet the corresponding causal model, represented by $\mathbf{A}$, can be uniquely identified because of its acyclicity constraint. In particular, the inverse of $\mathbf{B}$ can be converted uniquely to a lower triangular matrix having all-ones on its diagonal by some scaling and permutation of the rows.

3. Identifying Causal Orders among Observed Variables

Since the graph with adjacency matrix $\mathbf{A}$ is acyclic, there exists an integer d such that $\mathbf{A}^d = 0$. Thus, we can rewrite $\mathbf{B}$ in the following form:

$$\mathbf{B} = (\mathbf{I} - \mathbf{A})^{-1} = \sum_{k=0}^{d-1} \mathbf{A}^k. \tag{5}$$

It can be seen that there exists a causal path of length k from the exogenous noise of variable V_i to variable V_j if entry (j, i) of matrix $\mathbf{A}^k$ is nonzero. We define $[\mathbf{B}]_{j,i}$ as the total causal effect of variable V_i on variable V_j.

Assumption 1 *(Faithfulness assumption) The total causal effect from variable V_i to V_j is nonzero if there is a causal path from V_i to V_j. Thus, we have: $[\mathbf{B}]_{j,i} \neq 0$ if $V_i \rightsquigarrow V_j$.*

In the following lemma, we list two consequences of the faithfulness assumption that are immediate from the definition.

Lemma 1 *Under the faithfulness assumptions, for any two observed variables V_i and V_j, $1 \leq i, j \leq p_o$, the following holds:*
(i) Suppose that $V_i \rightsquigarrow V_j$. If $[\mathbf{B}]_{i,k} \neq 0$ for some $k \neq j$, then $[\mathbf{B}]_{j,k} \neq 0$.
(ii) If there is no causal path between V_i and V_j, then $[\mathbf{B}]_{i,j} = 0$ and $[\mathbf{B}]_{j,i} = 0$.

Based on Equation (2), we can write $\mathbf{V_o}$ in terms of $\mathbf{N_o}$ and $\mathbf{N_l}$ as follows

$$\mathbf{V_o} = (\mathbf{I} - \mathbf{D})^{-1}\mathbf{N_o} + (\mathbf{I} - \mathbf{D})^{-1}\mathbf{A_{ol}}(\mathbf{I} - \mathbf{A_{ll}})^{-1}\mathbf{N_l}, \tag{6}$$

where $\mathbf{D} = \mathbf{A_{oo}} + \mathbf{A_{ol}}(\mathbf{I} - \mathbf{A_{ll}})^{-1}\mathbf{A_{lo}}$. Let $\mathbf{B_o} := (\mathbf{I} - \mathbf{D})^{-1}$, $\mathbf{B_l} := (\mathbf{I} - \mathbf{D})^{-1}\mathbf{A_{ol}}(\mathbf{I} - \mathbf{A_{ll}})^{-1}$, and $\mathbf{N} := [\mathbf{N_o}; \mathbf{N_l}]$. Thus, $\mathbf{V_o} = \mathbf{B}'\mathbf{N}$ where $\mathbf{B}' := [\mathbf{B_o}, \mathbf{B_l}]$. This equation fits into a linear over-complete ICA where the exogenous noises are non-Gaussian and the number of observed variables is less than the number of variables in the system. The following proposition asserts when the columns of matrix $\mathbf{B}'$ are still identifiable up to permutations and scaling.

Definition 2 *(Reducibility of a matrix) A matrix is reducible if two of its columns are linearly dependent.*

Proposition 3 *((Eriksson and Koivunen, 2004), Theorem 3) In the linear over-completer ICA problem, the columns of mixing matrix can be identified up to some scaling and permutation if it is not reducible.*

Lemma 4 *The columns of $\mathbf{B}'$ corresponding to any two observed variables are linearly independent.*

Proof Consider any two observed variables V_i and V_j. We know that $[\mathbf{B}']_{i,i}$ and $[\mathbf{B}']_{j,j}$ are non-zero. Furthermore, $\mathbf{B}'$ is a sub-matrix of $\mathbf{B}$. Hence, based on Lemma 1 (ii), if there is no causal path between V_i and V_j, we have: $[\mathbf{B}']_{i,j} = 0$ and $[\mathbf{B}']_{j,i} = 0$. Thus, $[\mathbf{B}']_{:,i}$ and $[\mathbf{B}']_{:,j}$ are not linearly dependent. Furthermore, if one of the variable is the ancestor of the another one, let say $V_i \in anc(V_j)$, according to Lemma 1 (i), $[\mathbf{B}']_{j,i} \neq 0$ while $[\mathbf{B}']_{i,j} = 0$. Thus, $[\mathbf{B}']_{:,i}$ and $[\mathbf{B}']_{:,j}$ are also not linearly dependent in this case and the proof is complete. ■

Although columns of $\mathbf{B}'$ corresponding to the observed variables are pairwise linearly independent, a column corresponding to a latent variable V_i might be linearly dependent on a column corresponding to an observed or latent variable V_j (see Example 3). In that case, we can remove the column $[\mathbf{B}']_{:,i}$ and N_i from matrix $\mathbf{B}'$ and vector $\mathbf{N}$, respectively and replace N_j by $N_j + \alpha N_i$ where α is a constant such that $[\mathbf{B}']_{:,i} = \alpha[\mathbf{B}']_{:,j}$. We can continue this process until all the remaining columns are pairwise linearly independent. Let $\mathbf{B}''$ and $\mathbf{N}''$ be the resulting mixing matrix and exogenous noise vector, respectively. According to Lemma 4, all the columns of $\mathbf{B}'$ corresponding to observed variables are in $\mathbf{B}''$. We utilize matrix $\mathbf{B}''$ to recover a causal order among the observed variables.

Since matrix $\mathbf{B}''$ is not reducible, its column can be identified up to some scaling and permutation according to Proposition 3. Let $\tilde{\mathbf{B}}''$ be the recovered matrix containing columns of $\mathbf{B}''$. Consider any two observed variables V_i and V_j, i.e., $1 \leq i, j \leq p_o$. We extract two rows of $\tilde{\mathbf{B}}''$ corresponding to variables V_i and V_j. Let n_{0*} be the number of columns in $[\tilde{\mathbf{B}}''_{i,:}; \tilde{\mathbf{B}}''_{j,:}]$ whose first entries are zero but second entries are nonzero. Similarly, let n_{*0} be the number of columns that their first entries are nonzero but their second entries are zero. The following lemma asserts that the existence of a causal path between V_i and V_j can be checked from n_{0*} and n_{*0} (or equivalently, $\tilde{\mathbf{B}}''$).

Lemma 5 *Under the faithfulness assumption, the existence of a causal path between any two observed variable can be inferred from matrix $\tilde{\mathbf{B}}''$.*

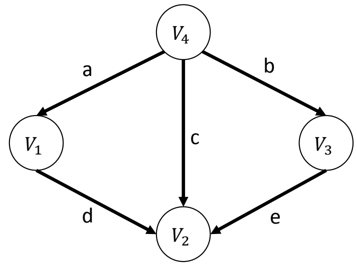

Figure 3: Causal graph of Example 3.

Proof. First, we show that if $V_i \rightsquigarrow V_j$, then $n_{0*} > 0$ and $n_{*0} = 0$. We know that matrix $[\tilde{\mathbf{B}}''_{i,:}; \tilde{\mathbf{B}}''_{j,:}]$ can be converted to $[\mathbf{B}''_{i,:}; \mathbf{B}''_{j,:}]$ by some permutation and scaling of its columns. Moreover, $\mathbf{B}''$ contains some of the columns of $\mathbf{B}'$ including all the columns corresponding to the observed variables. Thus, from Lemma 1, we know that if $[\mathbf{B}'']_{i,k} \neq 0$ for any $k \neq j$, then $[\mathbf{B}'']_{j,k} \neq 0$. Moreover, we have: $[\mathbf{B}'']_{j,j} \neq 0$ and $[\mathbf{B}'']_{i,j} = 0$. Hence, we can conclude that: $n_{0*} > 0$ and $n_{*0} = 0$.

If $n_{0*} > 0$ and $n_{*0} = 0$, then $V_i \rightsquigarrow V_j$. By contradiction, suppose that there is no causal path between V_i and V_i or $V_j \rightsquigarrow V_i$. The second case ($V_j \rightsquigarrow V_i$) does not happen due to what we just proved. Furthermore, from Lemma 1, we know that $[\mathbf{B}'']_{i,i} \neq 0$, $[\mathbf{B}'']_{i,j} = 0$. Therefore, $n_{*0} > 0$ which is in contradiction with our assumption. Hence, we can conclude that $n_{0*} > 0$ and $n_{*0} = 0$ if and only if $V_i \rightsquigarrow V_j$. ■

We can construct an auxiliary directed graph whose vertices are the observed variables and a directed edge exists from V_i to V_j if $V_i \rightsquigarrow V_j$ (which we can infer from n_{*0} and n_{0*}). Any causal order over the auxiliary graph is a correct causal order among the observed variables $\mathbf{V_o}$.

Example 3 *Consider the causal graph in Figure 3. Suppose that variables V_3 and V_4 are latent. $\mathbf{B}'$ would be:*

$$\begin{bmatrix} 1 & 0 & 0 & a \\ d & 1 & e & c + ad + be \end{bmatrix}.$$

We can remove the third column from $\mathbf{B}'$ and update the vector $\mathbf{N}$ to $[N_1; N_2 + eN_3; N_4]$. Thus, matrix $\mathbf{B}''$ is equal to:

$$\begin{bmatrix} 1 & 0 & a \\ d & 1 & c + ad + be \end{bmatrix},$$

which is not reducible. Without loss of generality, assume that the recovered matrix $\tilde{\mathbf{B}}''$ is equal to $\mathbf{B}''$. Therefore, $n_{0*} = 1$ and $n_{*0} = 0$. Hence, we can infer that there is a causal path from V_1 to V_2.

3.1. Recovering the Number of Variables in the System

According to Proposition 3, the number of variables in the system can be recovered if and only if matrix $\mathbf{B}'$ is not reducible. Furthermore, Equation (6) implies that matrix $\mathbf{B}'$ is not reducible if and only if the columns of matrix $[\mathbf{I}_{p_o \times p_o}, \mathbf{A_{ol}}(\mathbf{I} - \mathbf{A_{ll}})^{-1}]$ are not linearly independent. In the rest of this section, we will present equivalent necessary and sufficient graphical conditions under which

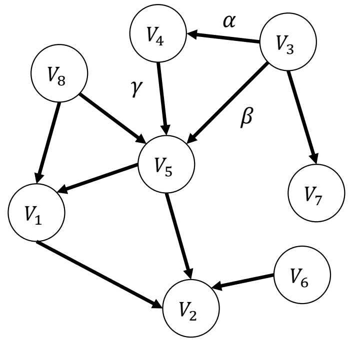

Figure 4: Causal graph of Example 5. V_1 and V_2 are the only observed variables.

the number of variables in the systems can be uniquely identified. But before that, we present a simple example where $[\mathbf{I}_{p_o \times p_o}, \mathbf{A_{ol}}(\mathbf{I} - \mathbf{A_{ll}})^{-1}]$ is reducible and give a graphical interpretation of it.

Example 4 *Consider a linear SEM with three variables V_1, V_2, and V_3 where $V_3 = N_3$, $V_1 = \alpha V_3 + N_1$, and $V_2 = \beta V_1 + N_2$. Thus, the corresponding causal graph would be: $V_3 \rightarrow V_1 \rightarrow V_2$. Suppose that V_3 is the only latent variable. Hence, $\mathbf{A_{ll}} = 0$, $\mathbf{A_{ol}} = [\alpha; 0]$, and $\mathbf{A_{ol}}(\mathbf{I} - \mathbf{A_{ll}})^{-1} = [\alpha; 0]$ which is linearly dependent on the first column of $\mathbf{I}$. In fact, latent variable V_3 can be absorbed in variable V_1 by changing the exogenous noise of V_1 from N_1 to $N_1 + \alpha N_3$. Thus, the number of variables in this model cannot be identified uniquely in this model.*

Definition 6 *(Absorbing) Variable V_i is said to be absorbed in variable V_j if the exogenous noise of V_i is set to zero $N_i \leftarrow 0$, and the exogenous noise of V_j is replaced by $N_j \leftarrow N_j + [\mathbf{B}]_{j,i} N_i$. We define absorbing a variable in $\emptyset$ by setting its exogenous noise to zero.*

Definition 7 *(Absorbablity) Let P'_{V_o} be the joint distribution of the observed variables after absorbing V_i in V_j. We say V_i is absorbable in V_j if $P'_{V_o} = P_{V_o}$.*

The following theorem characterizes the graphical conditions where a latent variable is absorbable. The proof of theorem is given in Appendix A.

Theorem 8
(a) A latent variable is absorbable in $\emptyset$ if and only if it has no observable descendant.
(b) A latent variable V_j is absorbable in variable V_i (observed or latent), if and only if all paths from V_j to its observable descendants go through V_i.

Example 5 *Consider a linear SEM with corresponding causal graph in Figure 4 where V_1 and V_2 are the only observed variables. V_7 satisfies condition (a) and its exogenous noise can be set to zero. Furthermore, V_3 and V_4 satisfy condition (b) with respect to V_5 and they can be absorbed in V_5 by setting the exogenous noise of V_5 to $N_5 + (\alpha\gamma + \beta)N_3 + \gamma N_4$. Finally, V_6 satisfies condition (b) and it can be absorbed in V_2. Note that V_8 and V_5 cannot be absorbed in V_1 or V_2.*

Definition 9 *We say a causal graph is minimal if none of its variables are absorbable.*

Based on above definition, a causal graph is minimal if none of the latent variables satisfy the conditions in Theorem 8. We borrowed the terminology of minimal causal graphs from Pearl (1988) for polytree causal structures. In (Pearl, 1988), a causal graph is called minimal if it has no redundant latent variables in the sense that the joint distribution without latent variables remains a connected tree. Later, Etesami et al. (2016) showed that in minimal latent directed information polytrees, each node has at least two children. The following lemma asserts that the same argument holds true for the non-absorbable latent variables in our setting. The proof of lemma is given in Appendix B.

Lemma 10 *A latent variable is non-absorbable if it has at least two non-absorbable children.*

The next theorem gives necessary and sufficient graphical conditions for non-reduciblity of matrix $\mathbf{B}'$. The proof of theorem is given in Appendix C.

Theorem 11 $\mathbf{B}'$ *is not reducible almost surely if and only if the corresponding causal graph G is minimal.*

Corollary 12 *Under faithfulness assumption and non-Gaussianity of exogenous noises, the number of variables in the system is identifiable almost surely if the corresponding graph is minimal.*

Proof. Based on Theorem 11, we know that matrix $\mathbf{B}'$ is not reducible almost surely if the corresponding causal graph G is minimal. Furthermore, according to Proposition 3, the number of variables in the systems is identifiable if matrix $\mathbf{B}'$ is not reducible. This completes the proof. ■

4. Identifying Total Causal Effects among Observed Variables

In this section, first, we will show by an example that total causal effects among observed variables cannot be identified uniquely under the faithfulness assumption and non-Gaussianity of exogenous noises.[3] However, we can obtain all the possible solutions. Furthermore, under some additional assumptions on linear SEM, we show that one can uniquely identify total causal effects among observed variables.

4.1. Example of non-Uniqueness of Total Causal Effects

Consider the causal graph in Figure 5 where V_i and V_j are observed variables and V_k is a latent variable. The direct causal effects from V_k to V_i, from V_k to V_j, and from V_i to V_j are α, γ, and β, respectively. We can write V_i and V_j based on the exogenous noises of their ancestors as follows:

$$\begin{aligned} V_i &= \alpha N_k + N_i, \\ V_j &= \beta N_i + (\alpha\beta + \gamma) N_k + N_j. \end{aligned} \tag{7}$$

Now, we construct a second causal graph depicted in Figure 5 where the exogenous noises of variables V_i and V_k are changed to αN_k and N_i, respectively. Furthermore, we set the direct causal effects from V_k to V_i, from V_k to V_j, and from V_i to V_j to 1, $-\gamma/\alpha$, and $\beta + (\gamma/\alpha)$, respectively. It

3. This example has also been studied in (Hoyer et al., 2008).

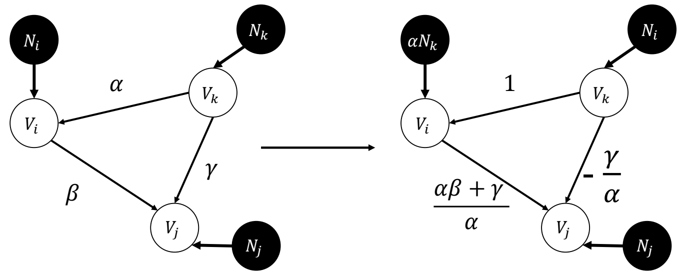

Figure 5: An example of non-identifiability of causal effects from observed variable V_i to observed variable V_j.

can be seen that equations in (7) do not change while the direct causal effect from V_i to V_j becomes $\beta + (\gamma/\alpha)$ in the second causal graph. Thus, we cannot identify causal effect from V_i to V_j merely by observational data from V_i and V_j. In Appendix D, we extend this example to the case where there might be multiple latent variables on the path from V_k to V_i and V_j, and from V_i to V_j.

The above example shows that causal effects may not be identified even by assuming non-Gaussianity of exogenous noises if we have some latent variables in the system. In the following, we first show that the set of all possible total causal effects can be identified. Afterwards, we will present a set of structural conditions under which we can uniquely identify total causal effects among observed variables.

4.2. Identifying the Set of All Possible Total Causal Effects

Since the subgraph corresponding to $\mathbf{A_{ll}}$ is a DAG, there exists an integer d_l such that $\mathbf{A_{ll}}^{d_l} = 0$. Hence, we can rewrite matrix $\mathbf{D}$ given in (6) as follows

$$\mathbf{D} = \mathbf{A_{oo}} + \sum_{k=0}^{d_l-1} \mathbf{A_{ol}}\mathbf{A_{ll}}^k\mathbf{A_{lo}}. \tag{8}$$

Lemma 13 *Matrix* $\mathbf{D}$ *in* (6) *can be converted to a strictly lower triangular matrix by permuting columns and rows simultaneously based on the causal order* k_o.

Proof. Let $\mathbf{P}$ be the permutation matrix corresponding to the causal order k_o. We want to show that $\mathbf{PDP}^T$ is strictly lower triangular. It suffices to prove $\mathbf{PA_{ol}A_{ll}}^k\mathbf{A_{lo}P}^T$ is strictly lower triangular for any $0 \leq k \leq d_l - 1$. Suppose that there exists a nonzero entry, (i,j), in $\mathbf{PA_{ol}A_{ll}}^k\mathbf{A_{lo}P}^T$ where $j \geq i$. Then, there should be a directed path from observed variable $V_{k_o^{-1}(j)}$ to $V_{k_o^{-1}(i)}$ of length $k+2$ through latent variables in the causal graph where $k_o^{-1}(i)$ is the index of an observed variable whose order is i in the causal order k_o. This means variable $V_{k_o^{-1}(j)}$ should come before variable $V_{k_o^{-1}(i)}$ in any causal order. But this violates the causal order k_o. ■

Previously, we showed that existence of a causal path between any two observed variables V_i and V_j can be determined by performing over-complete ICA. Let $des_o(V_i)$ be the set of all observed descendants of V_i, i.e., $des_o(V_i) = \{V_j | V_i \rightsquigarrow V_j, 1 \leq j \leq p_o\}$. We will utilize $des_o(V_i)$'s to enumerate all possible total causal effects among the observed variables.

Remark 14 *From Lemma 4, we have:* $des_o(V_i) \neq des_o(V_j)$ *for any* $1 \leq i, j \leq p_o$.

As we discussed in Section 3, under non-Gaussianity of exogenous noises, the columns of $\mathbf{B}''$ can be determined up to some scalings and permutations by solving an overcomplete ICA problem. Let p_r be the number of columns of $\mathbf{B}''$. Furthermore, without loss of generality, assume that variables $V_{p_o+1}, V_{p_o+2}, \cdots, V_{p_r}$ are the latent variables in the system whose corresponding columns remain in $\mathbf{B}''$.

Theorem 15 *Let* $r_i := |\{j : des_o(V_i) = des_o(V_j), 1 \leq j \leq p_r\}|$, *for any* $1 \leq i \leq p_o$. *Under the assumptions of faithfulness and non-Gaussianity of exogenous noises, the number of all possible* $\mathbf{D}$*'s that can generate the same distribution for* $\mathbf{V_o}$ *according to* (2)*, is equal to* $\Pi_{i=1}^{p_o} r_i$.

Proof. According to Proposition 3, under non-Gaussianity of exogenous noises, the columns of $\mathbf{B}''$ can be determined up to some scalings and permutations by solving an overcomplete ICA problem. Furthermore, for the column corresponding to the noise N_i, $1 \leq i \leq p_o$, we have r_i possible candidates with the same set of indicies of non-zero entries where all of them are pairwise linearly independent. Let $\mathbf{B'_o}$ be a $p_o \times p_o$ matrix by selecting one of the candidates for each column corresponding to noise N_i, $1 \leq i \leq p_o$. Thus, we have $\Pi_{i=1}^{p_o} r_i$ possible matrices.[4] Now, for each $\mathbf{B'_o}$, we just need to show that there exists an assignment for $\mathbf{A_{oo}}$, $\mathbf{A_{lo}}$, $\mathbf{A_{ol}}$, and $\mathbf{A_{ll}}$ such that they satisfy (6) and $\mathbf{A_{oo}}$ and $\mathbf{A_{ll}}$ can be converted to strictly lower triangular matrices with some simultaneous permutations of columns and rows.

Let $\mathbf{A_{lo}} = \mathbf{0}_{p_l \times p_o}$ and $\mathbf{A_{ll}} = \mathbf{0}_{p_l \times p_l}$. Assume that $\mathbf{B'_l}$ consists of the remaining columns which are not in $\mathbf{B'_o}$. We also add columns corresponding to latent absorbed variables to $\mathbf{B'_l}$. Now, we set $\mathbf{A_{oo}}$ and $\mathbf{A_{ol}}$ to $\mathbf{I} - \mathbf{B'_o}^{-1}$ and $\mathbf{B'_o}^{-1}\mathbf{B'_l}$, respectively. By these assignments, the proposed matrix $\mathbf{A} = [\mathbf{A_{oo}}, \mathbf{A_{ol}}; \mathbf{A_{lo}}, \mathbf{A_{ll}}]$ satisfies in (6). Thus, we just need to show that $\mathbf{I} - \mathbf{B'_o}^{-1}$ can be converted to a strictly lower triangular matrix by some permutations. To do so, first note that from Lemma 13, we know that matrix $\mathbf{D}$ can be converted to a strictly lower triangular matrix by a permutation matrix $\mathbf{P}$. Furthermore, based on this property of matrix $\mathbf{D}$, we have: $\mathbf{D}^{p_o} = \mathbf{0}$. Thus, we can write:

$$\mathbf{P}(\mathbf{I} - \mathbf{D})^{-1}\mathbf{P}^T = \sum_{k=0}^{p_o-1} \mathbf{P}\mathbf{D}^k\mathbf{P}^T = \sum_{k=0}^{p_o-1} (\mathbf{P}\mathbf{D}\mathbf{P}^T)^k.$$

Since matrix $(\mathbf{P}\mathbf{D}\mathbf{P}^T)^k$ is a lower triangular matrix for any $k \geq 0$, $(\mathbf{I} - \mathbf{D})^{-1}$ can be converted to a lower triangular matrix by permutation matrix $\mathbf{P}$. Furthermore, the set of nonzero entries of $\mathbf{B'_o}$ is the same as the one of $(\mathbf{I} - \mathbf{D})^{-1}$. Thus, $\mathbf{P}\mathbf{B'_o}\mathbf{P}^T$ is also a lower triangular matrix where all diagonal elements of it are equal to one. Hence, we can write $\mathbf{B'_o}$ in the form of $\mathbf{B'_o} = \mathbf{I} + \mathbf{B''_o}$ where $\mathbf{P}\mathbf{B''_o}\mathbf{P}^T$ is a strictly lower triangular matrix. Therefore, we have:

$$\mathbf{P}(\mathbf{I} - \mathbf{B'_o}^{-1})\mathbf{P}^T = \mathbf{P}(\mathbf{I} - \sum_{k=0}^{p_o-1} (-1)^k \mathbf{B''_o}^k)\mathbf{P}^T = \mathbf{P}(\sum_{k=1}^{p_o-1} (-1)^{k+1} \mathbf{B''_o}^k)\mathbf{P}^T, \quad (9)$$

4. Please note that diagonal entries of $\mathbf{B'_o}$ should be equal to one. Otherwise we can normalize each column to its on-diagonal entry.

Algorithm 1

Input: Collection of the sets $des_o(V_i), 1 \leq i \leq p_o$.
Run an over-complete ICA algorithm over observed variables $\mathbf{V_o}$ and obtain matrix $\tilde{\mathbf{B}}''$.
for $i = 1 : p_r$ **do**
  $I_i = \{k | [\tilde{\mathbf{B}}''_{:,i}]_k \neq 0\}$
  for $j = 1 : p_o$ **do**
    if $I_i = des_o(V_j)$ **then**
      $[\hat{\mathbf{B}}_\mathbf{o}]_{:,j} = \tilde{\mathbf{B}}''_{:,i} / [\tilde{\mathbf{B}}''_{:,i}]_j$
    end if
  end for
end for
Output: $\hat{\mathbf{B}}_\mathbf{o}$

where the last term shows that $\mathbf{I} - \mathbf{B_o'}^{-1}$ can be converted to a strictly lower triangular matrix and the proof is complete. ■

Comparing our results with (Hoyer et al., 2008), we can obtain all sets $des_o(V_i)$'s and determine which columns can be selected as corresponding columns of observed variables in $O(p_o^2 p_r)$ and then enumerate all the possible total causal effects while the proposed algorithm in (Hoyer et al., 2008) requires to search a space of $\binom{p_r}{p_o}$ different possible choices. Moreover, we can identify a causal order uniquely with the same time complexity by utilizing the method proposed in Section 3.

4.3. Unique Identification of Causal Effects under Structural Conditions

Based on Theorem 15, in this part, we propose a method to identify total causal effects uniquely under some structural conditions.

Assumption 2 *Assume that for any observed variables V_i and any latent variable V_k, we have: $des_o(V_k) \neq des_o(V_i)$.*

Assumption 2 is a very natural condition that one expects to hold for unique identifiability of causal effects. This is because if Assumption 2 fails, then based on Theorem 15, there are multiple sets of total causal effects that are compatible with the observed data.

Theorem 16 *Under Assumptions 1-2, and non-Gaussianity of exogenous noises, the total causal effect between any two observed variables can be identified uniquely.*

Proof. Let matrix $[\tilde{\mathbf{B}}'']_{p_o \times p_r}$ be the output of over-complete ICA problem whose columns are the columns in matrix $\mathbf{B}''$. We define I_i as the the set of indices of nonzero entries of column $\tilde{\mathbf{B}}''_{:,i}$, i.e. $I_i = \{k | [\tilde{\mathbf{B}}''_{:,i}]_k \neq 0\}$. We know that $I_i = des_o(V_j)$ if $\tilde{\mathbf{B}}''_{:,i}$ corresponds to the observed variable V_j. Moreover, under Assumption 2, any observed variable V_i and any variable V_j (observed or latent) have different sets $des_o(V_i)$ and $des_o(V_j)$. Thus, each set I_i is just equal to one of $des_o(V_i)$'s, let say $des_o(V_j)$. The column $\tilde{\mathbf{B}}''_{:,i}$ normalized by $[\tilde{\mathbf{B}}''_{:,i}]_j$ shows the total causal effects from variable j to other observed variables. ■

The description of the proposed solution in Theorem 16 is given in Algorithm 1. It is noteworthy that the example in Section 4.1 (given in Figure 5) violates the conditions in Theorem 16 since $des_o(V_k) = des_o(V_i)$. We have shown for this example that the causal effect from V_i to V_j cannot be identified uniquely.

5. Experiments

In this section, we first evaluate the performance of the proposed method in recovering causal orders from synthetic data, generated according to the causal graph in Figure 1. Our experiments show that the proposed method returns a correct causal order while, as we mentioned in Introduction section, previous methods proposed for linear non-Gaussian SEM with latent variables, might require additional assumptions in order to recover causal relations. More specifically, they do not have theoretical guarantee to recover the causal order or checking the existence of causal paths in our setting. Nevertheless, we evaluated the performances of lvLiNGAM (Hoyer et al., 2008), Pairwise lvLiNGAM (Entner and Hoyer, 2010), ParceLiNGAM (Tashiro et al., 2014), ICA-LiNGAM (Shimizu et al., 2006), Direct-LiNGAM (Shimizu et al., 2011) and FCI algorithm (Spirtes et al., 2000). We also consider another causal graph which satisfies Assumption 2 and demonstrate that the proposed method can return the correct causal effects. Next, we evaluate the performance of the proposed method for different number of variables in the system. Afterwards, for real data, we consider the daily closing prices of four world stock indicies and check the existence of causal paths between any two indicies. The results are compatible with common beliefs in economy.

5.1. Synthetic data

First, for the causal graph in Figure 1, we generated 1000 samples of observed variables V_1 and V_2 where nonzero entries of matrix $\mathbf{A}$ is equal to 0.9. We utilized the Reconstruction ICA (RICA) algorithm (Le et al., 2011) to solve the over-complete ICA problem as follows: Let $\mathbf{v_o}$ be a $p_o \times n$ matrix containing observational data where $[v_o]_{i,j}$ is j-th sample of variable V_i and n is the number of samples. First, the sample covariance matrix of $\mathbf{v_o}$ is eigen-decomposed, i.e., $1/(n-1)(\mathbf{v_o} - \bar{\mathbf{v}}_\mathbf{o})(\mathbf{v_o} - \bar{\mathbf{v}}_\mathbf{o})^T = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^T$ where $\mathbf{U}$ is the orthogonal matrix, $\mathbf{\Sigma}$ is a diagonal matrix, and $\bar{\mathbf{v}}_\mathbf{o}$ is the sample mean vector. Then, the observed data is pre-whitened as follows: $\mathbf{w} = \mathbf{\Sigma}^{-1/2}\mathbf{U}(\mathbf{v_o} - \bar{\mathbf{v}}_\mathbf{o})$. The RICA algorithm tries to find matrix $\mathbf{Z}$ that is the minimizer of the following objective function:

$$\underset{\mathbf{Z}}{\text{minimize}} \sum_{i=1}^{n}\sum_{j=1}^{p_r} g(\mathbf{Z}_{:,j}^T \mathbf{w}_{:,i}) + \frac{\lambda}{n}\sum_{i=1}^{n} \|\mathbf{Z}\mathbf{Z}^T\mathbf{w}_{:,i} - \mathbf{w}_{:,i}\|_2^2,$$

where parameter λ controls the cost of penalty term. We estimated matrix $\tilde{\mathbf{B}}''$ by $\mathbf{U}\mathbf{\Sigma}^{1/2}\mathbf{Z}^*$ where $\mathbf{Z}^*$ is the optimal solution of the above optimization problem.

In order to estimate the number of columns of $\tilde{\mathbf{B}}''$, we held out 250 of samples for model selection. More specifically, we solved the over-complete ICA problem for different number of columns, evaluated the fitness of each model by computing the objective function of RICA over the hold-out set, and selected the model with minimum cost. In order to check whether an entry is equal to zero, we used the bootstrapping method (Efron and Tibshirani, 1994), which generates 10 bootstrap samples by sampling with replacement from training data. For each bootstrap sample, we executed RICA algorithm to obtain an estimation of $\tilde{\mathbf{B}}''$. Since in each estimation, columns are in arbitrary permutation, we need to match similar columns in estimations of $\tilde{\mathbf{B}}''$. To do so, in each

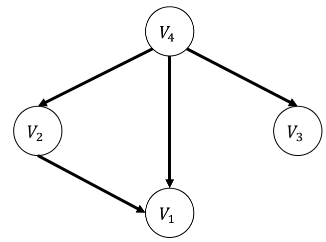

Figure 6: An example of causal graphs satisfying structural conditions.

estimation, we divided all entries of a column by the entry with the maximum absolute value in that column. Then, we picked each column from the estimated mixing matrix, computed its l_2 distance from each column of another estimated mixing matrix, and matched to the one with a minimum distance. Afterwards, we used a t-test with confidence level of 95% to check whether an entry is equal to zero from the bootstrap samples. An estimation of $\tilde{\mathbf{B}}''$ from a bootstrap sample is given as follows:

$$\begin{bmatrix} -0.0272 & 0.5238 & 1 \\ 1 & 1 & 0.8579 \end{bmatrix}.$$

Moreover, experimental results showed the correct support of $\tilde{\mathbf{B}}''$, i.e., $[0,1,1;1,1,1]$ can be recovered with merely 10 bootstrap samples. Thus, there is a causal path from V_1 to V_2. Furthermore, for the causal graph $V_1 \leftarrow V_3 \rightarrow V_2$ in which V_3 is only the latent variable, we repeated the same procedure explained above. An estimation of $\tilde{\mathbf{B}}''$ from one of the bootstrap samples is given as follows:

$$\begin{bmatrix} 1 & -0.046 & 0.9838 \\ -0.031 & 1 & 1 \end{bmatrix}.$$

From experiments, the estimated support of $\tilde{\mathbf{B}}''$ from bootstrap samples was: $[0,1,1;1,0,1]$. Thus, we can conclude that there is no causal path between V_1 and V_2. Next, we considered the causal graph in Figure 6 where V_4 is the only latent variable. The direct causal effects of all directed edges are equal to 0.9. An estimation of $\tilde{\mathbf{B}}''$ from one of the bootstrap samples is given as follows:

$$\begin{bmatrix} -0.049 & 0.892 & 1 & 1 \\ -0.024 & 1 & 0.523 & -0.042 \\ 1 & -0.02 & 0.527 & -0.032 \end{bmatrix}.$$

Thus, we can deduce that there is only a causal path from V_2 to V_1. We can also estimate total causal effects between observed variables since this causal graph satisfies Assumption 2. The output of Algorithm 1 is:

$$\begin{bmatrix} 1 & 0.892 & -0.049 \\ -0.042 & 1 & -0.024 \\ -0.032 & -0.02 & 1 \end{bmatrix}.$$

which is close to the true causal effects. We evaluated previous methods for learning the causal graphs in Figure 1, Figure 6, and the causal graph $V_1 \leftarrow V_3 \rightarrow V_2$. Table 1 shows whether each of

	Figure 1	Figure 6	$V_1 \leftarrow V_3 \rightarrow V_2$
lvLiNGAM (Hoyer et al., 2008)	✓	×	✓
Pairwise lvLiNGAM (Entner and Hoyer, 2010)	×	×	✓
ParceLiNGAM (Tashiro et al., 2014)	×	×	×
ICA-LiNGAM (Shimizu et al., 2006)	✓	×	×
Direct-LiNGAM (Shimizu et al., 2011)	✓	×	×
FCI (Spirtes et al., 2000)	×	×	×
Proposed algorithm	✓	✓	✓

Table 1: Comparison of methods in recovering causal paths for the causal graphs in Figure 1, Figure 6, and the causal graph $V_1 \leftarrow V_3 \rightarrow V_2$.

p	10	15	20	25	30
$c = 2$	0.7	1.41	1.66	3.09	3.48
$c = 3$	0.76	1.48	1.75	3.33	3.84

Table 2: Running time (in seconds) of Algorithm 1 for differnet number of variables in the system and different graph densities $c = 2, 3$.

them can find all causal paths correctly. It can be seen that only the proposed algorithm is successful in recovering the causal paths in all considered causal graphs.

We generated 1000 DAGs of size p by first selecting a causal order among variables randomly and then connecting each pair of variables with probability $c/(p-1)$, where c is the average degree of each node. We generated data from a linear SEM where nonzero entries of matrix $\mathbf{A}$ were drawn uniformly from the range $[-0.9, -0.5] \cup [0.5, 0.9]$ and the exogenous noises followed a uniform distribution. In the remainder of this part, we assume that the number of latent variable is known. We first evaluated the running time of Algorithm 1 and compared it with the proposed algorithm in (Hoyer et al., 2008), which can provide all possible total causal effects. In the experiments, we selected $p_l = p/2$ variables randomly as latent variables. The running time of Algorithm 1 is given in Table 2 for $c = 2, 3$. In our experiments, the algorithm in (Hoyer et al., 2008) did not return any output in 10 minutes and it is only feasible on small graphs with fewer than six variables.

We evaluated the performance of the proposed algorithm and compared it with the previous ones, including Pairwise lvLiNGAM (Entner and Hoyer, 2010), ParceLiNGAM (Tashiro et al., 2014), LiNGAM (Shimizu et al., 2006), and Direct-LiNGAM (Shimizu et al., 2011), in the presence of latent variables. More specifically, we define precision of an algorithm as the fraction of correctly recovered causal paths among recovered causal paths between any two observed variables. We also define its recall as the fraction of recovered causal paths among actual causal paths between any two observed variables. Figure 7 shows presions and recalls of the mentioned algorithms for different number of variables $p = 10, 15, 20$, different number of observed variables, and different average degrees $c = 4, 7$. One can see that none of the algorithms has the best performance in all settings. However, the proposed algorithm and Pairwise lvLiNGAM (Entner and Hoyer, 2010) are the top two

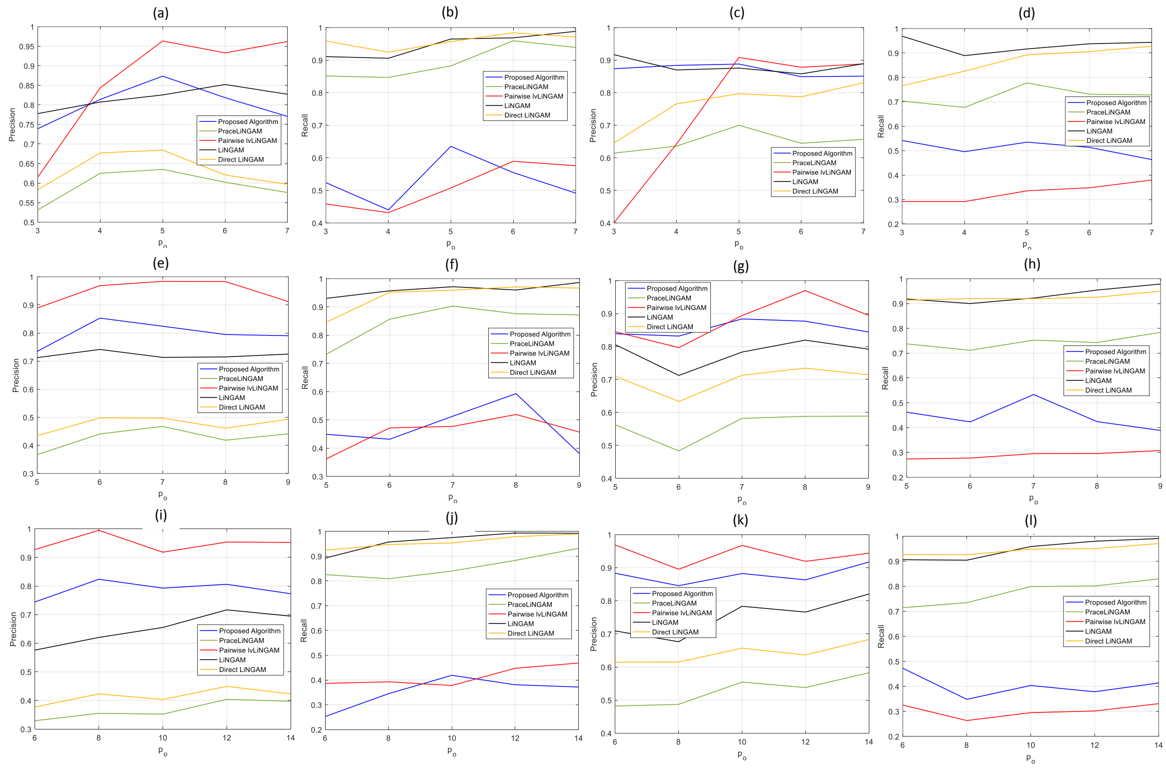

Figure 7: Precisions/Recalls of Pairwise lvLiNGAM (Entner and Hoyer, 2010), ParceLiNGAM (Tashiro et al., 2014), ICA-LiNGAM (Shimizu et al., 2006), Direct-LiNGAM (Shimizu et al., 2011) and the proposed algorithm in the presence of latent variables: (a) Precisions for $p = 10$, $c = 4$, (b) Recalls for $p = 10$, $c = 4$, (c) Precisions for $p = 10$, $c = 7$, (d) Recalls for $p = 10$, $c = 7$, (e) Precisions for $p = 15$, $c = 4$, (f) Recalls for $p = 15$, $c = 4$, (g) Precisions for $p = 15$, $c = 7$, (h) Recalls for $p = 15$, $c = 7$, (i) Precisions for $p = 20$, $c = 4$, (j) Recalls for $p = 20$, $c = 4$, (k) Precisions for $p = 20$, $c = 7$, (l) Recalls for $p = 20$, $c = 7$.

algorithms in terms of precision. Moreover, LiNGAM (Shimizu et al., 2006) and Direct-LiNGAM (Shimizu et al., 2011) have the best performance in terms of recall.

5.2. Real data

We considered the daily closing prices of the following world stock indicies from $10/12/2012$ to $10/12/2018$, obtained from Yahoo financial database: Dow Jones Industrial Average (DJI) in USA, Nikkei 225 (N225) in Japan, Euronext 100 (N100) in Europe, Hang Seng Index (HSI) in Hong Kong, and the Shanghai Stock Exchange Composite Index (SSEC) in China.

Let $c_i(t)$ be the closing price of i-th index on day t. We define the corresponding return by $R_i(t) := (c_i(t) - c_{i-1}(t))/c_{i-1}(t)$. We considered the returns of indicies as an observational data and applied the proposed method in Section 3 in order to check the existence of a causal path between any two indicies. Figure 8 depicts the causal relationships among the indicies. In this

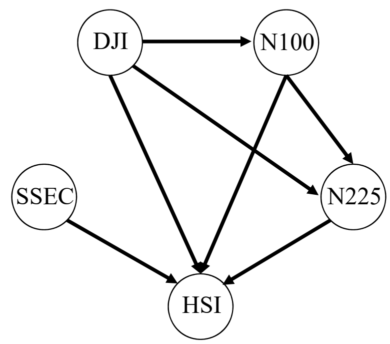

Figure 8: The causal relationships among five world stock indicies obtained from the proposed method in Section 3.

figure, there is a directed edge from index i to index j if we find a causal path from i to j. As can be seen, there are causal paths from DJI to HSI, N225, and N100 which is commonly known to be true in the stock market (Hyvärinen et al., 2010). Furthermore, HSI is influenced by all other indicies and SSEC only affects HSI which these findings are compatible with the previous results in (Hyvärinen et al., 2010).

6. Conclusions

We considered the problem of learning causal models from observational data in linear non-Gaussian acyclic models with latent variables. Under the faithfulness assumption, we proposed a method to check whether there exists a causal path between any two observed variables. Moreover, we gave necessary and sufficient graphical conditions to uniquely identify the number of variables in the system. From the information about the existence of a directed path, we could obtain a causal order among the observed variables. Additionally, we considered the problem of estimating total causal effects. We showed by an example that causal effects among observed variables cannot be identified uniquely even under the assumptions of faithfulness and non-Gaussianity of exogenous noises. However, we can identify all possible set of total causal effects that are compatible with the observational data efficiently in time. Furthermore, we presented structural conditions under which we can learn total causal effects among observed variables uniquely. Experiments on synthetic data and real-world data showed the effectiveness of our proposed algorithms on learning causal models. One of our future research directions is to extend the results to the case of cyclic linear SEMs. We believe that methods similar to the one proposed in this paper can recover some of the causal paths in the system. Another direction of future work entails developing causal structure learning algorithms for nonlinear SEM with latent variables by exploiting recent progress in non-linear ICA. In addition, it is desirable to develop a principled, efficient approach to selecting the optimal number of latent variables.

Acknowledgments

The authors thank the editor and reviewers for their supportive and insightful comments. Negar Kiyavash's work was in part supported by Navy grant N00014-19-1-2333. Kun Zhang would like to acknowledge the support by National Institutes of Health under Contract No. NIH-1R01EB022858-01, FAINR01EB022858, NIH-1R01LM012087, NIH-5U54HG008540-02, and FAIN- U54HG008540. The National Institutes of Health is not responsible for the views reported in this article.

Appendix A. Proof of Theorem 8

"if" part:
We say a directed path is latent if all the variables on the path except the endpoint are latent. The "if" parts of conditions in Theorem 8 can be rewritten as follows:
(a) Latent variable V_{p_o+j}, $1 \leq j \leq p_l$, is absorbable in $\emptyset$ if it has no observable descendant.
(b1) Latent variable V_{p_o+j}, $1 \leq j \leq p_l$, is absorbable in observed variable V_i, $1 \leq i \leq p_o$, if V_i is the only observed variable influenced by V_{p_o+j} through some latent paths.
(b2) Latent variable V_{p_o+j}, $1 \leq j \leq p_l$, is absorbable in latent variable V_{p_o+k}, $1 \leq k \leq p_l$, if all latent paths from V_{p_o+j} to observed variables go through V_{p_o+k}.
It is easy to show that conditions (b1) and (b2) are equivalent to "if" part of condition (b) in Theorem 8. From (6), we know that $\mathbf{V_o} = (\mathbf{I}-\mathbf{D})^{-1}[\mathbf{I}, \mathbf{A_{ol}}(\mathbf{I}-\mathbf{A_{ll}})^{-1}]\mathbf{N}$ where entry (i,j) of matrix $(\mathbf{I}-\mathbf{D})^{-1}\mathbf{A_{ol}}(\mathbf{I}-\mathbf{A_{ll}})^{-1}$ is the total causal effect of latent variable V_{p_o+j} to the observed variable V_i. This entry would be zero if no directed path exists from latent variable V_{p_o+j} to observed variable V_i. Now, we prove the correctness of above conditions:
(a) If a latent variable V_{p_o+j} has no observable descendant, then the j-th column of $\mathbf{A_{ol}}(\mathbf{I}-\mathbf{A_{ll}})^{-1}$ is all zeros. Hence, there would be no changes in $[\mathbf{I}, \mathbf{A_{ol}}(\mathbf{I}-\mathbf{A_{ll}})^{-1}]\mathbf{N}$ by setting N_{p_o+j} to zero. Therefore, there would be no change in P_{V_o}.
(b1) Since latent variable V_{p_o+j} only influences one observed variable through latent paths, $[\mathbf{A_{ol}}(\mathbf{I}-\mathbf{A_{ll}})^{-1}]_{:,j}$ has only one non-zero entry and therefore linearly dependent on one of columns of identity matrix, let say i-th column. Moreover, the total causal effect from V_{p_o+j} to V_i, i.e., $[\mathbf{B}]_{i,p_o+j}$ is equal to $[\mathbf{A_{ol}}(\mathbf{I}-\mathbf{A_{ll}})^{-1}]_{i,j}$ since there is no causal path from V_{p_o+j} to V_i that goes through an observed variable other than V_i. Thus, we replace N_i by $N_i + [\mathbf{A_{ol}}(\mathbf{I}-\mathbf{A_{ll}})^{-1}]_{i,j}N_{p_o+j}$ and set N_{p_o+j} to zero and there would be no change in $[\mathbf{I}, \mathbf{A_{ol}}(\mathbf{I}-\mathbf{A_{ll}})^{-1}]\mathbf{N}$.
(b2) Consider any observed variable V_i, $1 \leq i \leq p_o$. If all latent paths of V_{p_o+j} go though V_{p_o+k}, then $[\mathbf{A_{ol}}(\mathbf{I}-\mathbf{A_{ll}})^{-1}]_{i,j} = [\mathbf{A_{ol}}(\mathbf{I}-\mathbf{A_{ll}})^{-1}]_{i,k}[\mathbf{B}]_{p_o+k,p_o+j}$ since all the paths from V_{p_o+j} to V_{p_o+k} are latent. Thus, we can change N_{p_o+k} to $N_{p_o+k} + [\mathbf{B}]_{p_o+k,p_o+j}N_{p_o+j}$ and set N_{p_o+j} to zero and there would be no change in $[\mathbf{I}, \mathbf{A_{ol}}(\mathbf{I}-\mathbf{A_{ll}})^{-1}]\mathbf{N}$.
"only if" part:
Now, we prove that the conditions (a), (b1), and (b2) are the only absorbable case. It can be easily shown that an observed variable cannot be absorbed into any other observed or latent variables. Thus, it is just needed to consider the following cases:

- Absorbing a latent variable in an observed variable: Suppose that a latent variable V_j can be absorbed in an observed variable V_i. Furthermore, assume that V_j also influences other observed variable V_k through latent path(s). That is, there exist some paths that start from V_j and end in V_k without traversing, V_i. Let $\gamma \neq 0$ be the causal strength of such paths. Then,

$[\mathbf{B}]_{k,j} = [\mathbf{B}]_{k,i} \times [\mathbf{B}]_{i,j} + \gamma$. To absorb V_j in V_i, γ should be zero which would contradict the faithfulness assumption.

- Absorbing a latent variable in another latent variable: Suppose that a latent variable V_j can be absorbed in another latent variable V_i but for some observed variable V_k, all latent paths from V_j do not go through V_i. Let γ be the causal strength of such paths. Then, $[\mathbf{B}]_{k,j} = [\mathbf{B}]_{k,i} \times [\mathbf{B}]_{i,j} + \gamma$. To absorb V_j in V_i, γ should be zero which contradicts the faithfulness assumption.

Appendix B. Proof of Lemma 10

Suppose that a latent variable V_i has at least two non-absorbable children such as V_j and V_k. We need to consider three cases:

- If both of V_j and V_k are observed variables, then V_i is not absorbable according to Theorem 8.
- Suppose that V_j and V_k are latent variables. Each of them must reach at least two observed variables through latent paths (due to condition (b) in Theorem 8). Thus, V_i also reaches those observed variables through latent paths. Furthermore, all latent paths starting from V_i do not go through only one latent variable. Hence, none of the conditions in Theorem 8 are satisfied and V_i is not absorbable.
- One of V_j or V_k, let say variable V_j, is observed. V_k must reach an observed variable other than V_j through some latent paths. Otherwise, it is absorbable. Therefore, V_i is not absorbable since it does not satisfy any conditions in Theorem 8.

Appendix C. Proof of Theorem 11

If G is not minimal, then it can be easily seen that $\mathbf{B}'$ is also reducible. Now, suppose that G is minimal. We want to show that $\mathbf{B}'$ is also not reducible almost surely. By contradiction, suppose that $\mathbf{B}'$ is reducible. Then two columns of $[\mathbf{I}, \mathbf{A_{ol}}(\mathbf{I} - \mathbf{A_{ll}})^{-1}]$ must be linearly dependent. Now, two cases should be considered:

- One column of $\mathbf{A_{ol}}(\mathbf{I} - \mathbf{A_{ll}})^{-1}$, let say i-th column, and one column of $\mathbf{I}$ are linearly dependent. Hence, all the latent paths starting from latent variable V_{p_o+i} influences only one observed variable (Condition (b) in Theorem 8). Thus, G is not minimal which is a contradiction.
- Two columns of $\mathbf{A_{ol}}(\mathbf{I} - \mathbf{A_{ll}})^{-1}$, let say i, j are linearly dependent. If the corresponding columns have only one non-zero entry, then both of them can be absorbed in an observed variable (Condition (b) in Theorem 8). Thus, G is not minimal. Now, suppose that these columns have more than one nonzero entry each, let say entries k and l. Without loss of generality, suppose that V_{p_o+i} is the ancestor of V_{p_o+j} (the same argument still holds true if neither is an ancestor of the other). Let h_i be the maximum length of latent paths starting from latent variable V_{p_o+i}. By induction on h_i, we will show that i, j-th columns of $\mathbf{A_{ol}}(\mathbf{I} - \mathbf{A_{ll}})^{-1}$ are linearly dependent with measure zero. The case of $h_i = 1$ is trivial. Suppose that for

$h_i = r$, the statement holds true. We will prove it for $h_i = r + 1$. Let latent variable V_{p_o+u} be a child of V_{p_o+i} and assume some paths from V_{p_o+u} do not go through V_{p_o+j}. Let $[\mathbf{C}]_{i,j}$ be the total causal strength of only latent paths from V_j to V_i. We know that:

$$[\mathbf{C}]_{k,p_o+j}/[\mathbf{C}]_{l,p_o+j} = [\mathbf{C}]_{k,p_o+i}/[\mathbf{C}]_{l,p_o+i}. \tag{10}$$

Furthermore,

$$[\mathbf{C}]_{k,p_o+i} = [\mathbf{C}]_{k,p_o+u}[\mathbf{C}]_{p_o+u,p_o+i} + c', [\mathbf{C}]_{l,p_o+i} = [\mathbf{C}]_{l,p_o+u}[\mathbf{C}]_{p_o+u,p_o+i} + c'', \tag{11}$$

for some values c', c''. Moreover, $[\mathbf{C}]_{p_o+u,p_o+i} = [\mathbf{A}]_{p_o+u,p_o+i} + c'''$ for some c'''. Plugging (11) into (10), we have:

$$([\mathbf{C}]_{k,p_o+u}[\mathbf{C}]_{l,p_o+j} - [\mathbf{C}]_{k,p_o+j}[\mathbf{C}]_{l,p_o+u})[\mathbf{A}]_{p_o+u,p_o+i} = \\ [\mathbf{C}]_{l,p_o+j}c' - [\mathbf{C}]_{k,p_o+j}c'' - ([\mathbf{C}]_{k,p_o+u}[\mathbf{C}]_{l,p_o+j} - [\mathbf{C}]_{k,p_o+j}[\mathbf{C}]_{l,p_o+u})c'''.$$

The above equation holds with measure zero if $[\mathbf{C}]_{k,p_o+u}[\mathbf{C}]_{l,p_o+j} - [\mathbf{C}]_{k,p_o+j}[\mathbf{C}]_{l,p_o+u} \neq 0$ which is true with measure one from the induction hypothesis.

Appendix D. An Example of non-Identifiability of Total Causal Effects

Let $P = (V_{i_0}, V_{i_1}, \cdots, V_{i_{r-1}}, V_{i_r})$ be a causal path of length r from variable V_{i_0} to variable V_{i_r}. We define the weight of path P, denoted by ω_p , as the product of direct causal strengths of edges on the path:

$$\omega_P = \prod_{s=0}^{r-1}[\mathbf{A}]_{i_{s+1},i_s}. \tag{12}$$

Suppose that Π_{V_i,V_j} be the set of all causal paths from variable V_i to variable V_j. It can be shown that the total causal effect from V_i to V_j can be computed by the following equation:

$$[\mathbf{B}]_{j,i} = \sum_{P\in\Pi_{V_i,V_j}} \omega_P. \tag{13}$$

Now, consider a causal graph in Figure 5 where V_i and V_j are observed variables and V_k is latent variable. There exist causal paths from V_k to V_i and V_j, and from V_i to V_j with the following properties:

- Let Π'_{V_k,V_j} be the causal paths from variable V_k to variable V_j where V_i is not on any of these paths. We assume that $\Pi'_{V_k,V_j} \neq \emptyset$.
- All intermediate variables in Π_{V_k,V_i}, Π'_{V_k,V_j} and Π_{V_i,V_j} are latent.

We can write V_i and V_j based on the exogenous noises of their ancestors as follows:

$$\begin{aligned} V_i &= \alpha N_k + \sum_{V_r\in anc(V_i)\backslash V_k} [\mathbf{B}]_{i,r}N_r, \\ V_j &= \beta N_i + \gamma N_k + \sum_{V_r\in anc(V_j)\backslash\{V_k,V_i\}} [\mathbf{B}]_{i,r}N_r, \end{aligned} \tag{14}$$

where $\alpha = \sum_{P \in \Pi_{V_k,V_i}} \omega_P$, $\beta = \sum_{P \in \Pi_{V_i,V_j}} \omega_P$, and $\gamma = \sum_{P \in \Pi'_{V_k,V_j}} \omega_P$.

Now, we construct a causal graph depicted in Figure 5 where the exogenous noises of variables V_i and V_k are changed to αN_k and N_i, respectively. Furthermore, we pick three paths $P_1 \in \Pi_{V_k,V_i}$, $P_2 \in \Pi'_{V_k,V_j}$, $P_3 \in \Pi_{V_i,V_j}$ where:

$$P_1 = (V_k, V_{u_1}, \cdots, V_i),$$
$$P_2 = (V_k, V_{u_2}, \cdots, V_j),$$
$$P_3 = (V_i, V_{u_3}, \cdots, V_j).$$

By our first property on the paths, we can find two paths P_1 and P_2 such that $V_{u_1} \neq V_{u_2}$. We also change matrix $\mathbf{A}$ to matrix $\mathbf{A}'$ where all the entries of $\mathbf{A}'$ are the same as $\mathbf{A}$ except three entries $[\mathbf{A}']_{u_1,k}$, $[\mathbf{A}']_{u_2,k}$, and $[\mathbf{A}']_{u_3,i}$. We will adjust these three entries such that the total causal effects from V_k to V_i, from V_k to V_j, and from V_i to V_j become 1, $-\gamma/\alpha$, and $\beta + \gamma/\alpha$, respectively. Moreover, these adjustments should not change the dependencies of observed variables V_i and V_j to the exogenous noises of their ancestors given in Equation (14). It can be shown that we can change the three mentioned causal effects to our desired values by the following adjustments:

$$[\mathbf{A}']_{u_1,k} = \frac{1 - \sum_{P \in \Pi_{V_k,V_i} \setminus \{P_1\}} \omega_P}{\omega_{P_1}/[\mathbf{A}]_{u_2,k}},$$
$$[\mathbf{A}']_{u_2,k} = \frac{-\gamma/\alpha - \sum_{P \in \Pi'_{V_k,V_j} \setminus \{P_2\}} \omega_P}{\omega_{P_2}/[\mathbf{A}]_{u_2,k}},$$
$$[\mathbf{A}']_{u_3,i} = \frac{\beta + \gamma/\alpha - \sum_{P \in \Pi_{V_i,V_j} \setminus \{P_3\}} \omega_P}{\omega_{P_3}/[\mathbf{A}]_{u_3,i}}.$$

Now, consider any latent variable V_u which is on one of the paths in Π_{V_k,V_i}, Π'_{V_k,V_j}, or Π_{V_i,V_j}. Changes in those mentioned three edges cannot affect the total causal effect from V_u to V_i or V_j since the edges (V_k, V_{u_1}), (V_k, V_{u_2}), and (V_i, V_{u_3}) are not a part of any paths from V_u to V_i or V_j. Thus, equations in (14) do not change while the total causal effect from V_i to V_j becomes $\beta + \gamma/\alpha$ in the second causal graph. It is noteworthy to mention that changes in the equations of latent variables are not important since we are not observing these variables.

References

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

Zhitang Chen and Laiwan Chan. Causality in linear nongaussian acyclic models in the presence of latent gaussian confounders. *Neural Computation*, 25(6):1605–1641, 2013.

Bradley Efron and Robert J Tibshirani. *An Introduction to the Bootstrap*. CRC press, 1994.

Gal Elidan and Nir Friedman. Learning hidden variable networks: The information bottleneck approach. *Journal of Machine Learning Research*, 6(Jan):81–127, 2005.

Doris Entner and Patrik O Hoyer. Discovering unconfounded causal relationships using linear non-gaussian models. In *JSAI International Symposium on Artificial Intelligence*, pages 181–195. Springer, 2010.

Jan Eriksson and Visa Koivunen. Identifiability, separability, and uniqueness of linear ica models. *IEEE Signal Processing Letters*, 11(7):601–604, 2004.

Jalal Etesami, Negar Kiyavash, and Todd Coleman. Learning minimal latent directed information polytrees. *Neural Computation*, 28(9):1723–1768, 2016.

AmirEmad Ghassami, Saber Salehkaleybar, Negar Kiyavash, and Elias Bareinboim. Budgeted experiment design for causal structure learning. In *International Conference on Machine Learning*, pages 1724–1733, 2018.

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.

Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*, pages 689–696, 2009.

Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*, volume 46. John Wiley & Sons, 2004.

Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11(May): 1709–1731, 2010.

Dominik Janzing, Joris Mooij, Kun Zhang, Jan Lemeire, Jakob Zscheischler, Povilas Daniušis, Bastian Steudel, and Bernhard Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182:1–31, 2012.

Robert I Jennrich and Peter M Bentler. Exploratory bi-factor analysis. *Psychometrika*, 76(4):537–549, 2011.

Erich Kummerfeld and Joseph Ramsey. Causal clustering for 1-factor measurement models. In *ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, pages 1655–1664. ACM, 2016.

Quoc V Le, Alexandre Karpenko, Jiquan Ngiam, and Andrew Y Ng. Ica with reconstruction cost for efficient overcomplete feature learning. In *Advances in Neural Information Processing Systems*, pages 1017–1025, 2011.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Jonas Peters and Peter Bühlmann. Identifiability of gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2013.

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

Saber Salehkaleybar, Jalal Etesami, and Negar Kiyavash. Identifying nonlinear 1-step causal influences in presence of latent variables. In *IEEE International Symposium on Information Theory*, pages 1341–1345. IEEE, 2017.

Saber Salehkaleybar, Jalal Etesami, Negar Kiyavash, and Kun Zhang. Learning vector autoregressive models with latent processes. In *International Conference on Machine Learning*, pages 4000–4007, 2018.

Shohei Shimizu. Lingam: Non-gaussian methods for estimating causal structures. *Behaviormetrika*, 41(1):65–98, 2014.

Shohei Shimizu and Kenneth 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*, 15(1):2629–2652, 2014.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(Oct):2003–2030, 2006.

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 (Apr):1225–1248, 2011.

Ricardo Silva and Richard Scheines. Generalized measurement models. Technical report, 2005.

Peter Spirtes, Christopher Meek, and Thomas Richardson. Causal inference in the presence of latent variables and selection bias. In *Uncertainty in Artificial Intelligence*, pages 499–506. Morgan Kaufmann Publishers Inc., 1995.

Peter Spirtes, Clark N Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, and Thomas Richardson. *Causation, Prediction, and Search*. MIT press, 2000.

Tatsuya Tashiro, Shohei Shimizu, Aapo Hyvärinen, and Takashi Washio. Parcelingam: a causal ordering method robust against latent confounders. *Neural Computation*, 26(1):57–83, 2014.

Kun Zhang and Aapo Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Uncertainty in Artificial Intelligence*, pages 647–655. AUAI Press, 2009.

Kun Zhang, Biwei Huang, Jiji Zhang, Clark Glymour, and Bernhard Schölkopf. Causal discovery in the presence of distribution shift: Skeleton estimation and orientation determination. In *International Joint Conference on Artificial Intelligence*, 2017.