Local Causal Discovery with Linear non-Gaussian Cyclic Models

Haoyue Dai[*1] **Ignavier Ng**[*1] **Yujia Zheng**[1] **Zhengqing Gao**[2] **Kun Zhang**[1,2]
[1]Carnegie Mellon University [2]Mohamed bin Zayed University of Artificial Intelligence

Abstract

Local causal discovery is of great practical significance, as there are often situations where the discovery of the global causal structure is unnecessary, and the interest lies solely on a single target variable. Most existing local methods utilize conditional independence relations, providing only a partially directed graph, and assume acyclicity for the ground-truth structure, even though real-world scenarios often involve cycles like feedback mechanisms. In this work, we present a general, unified local causal discovery method with linear non-Gaussian models, whether they are cyclic or acyclic. We extend the application of independent component analysis from the global context to independent subspace analysis, enabling the exact identification of the equivalent local directed structures and causal strengths from the Markov blanket of the target variable. We also propose an alternative regression-based method in the particular acyclic scenarios. Our identifiability results are empirically validated using both synthetic and real-world datasets.

1 INTRODUCTION

Causal discovery aims to identify causal relations among variables from data. In many real-world scenarios, it is not essential to determine the causal structure across all variables. Rather, the primary interest is often in unveiling the causes and effects related to specific target variables. Allocating resources to estimate a global structure for such narrowed objectives can be computationally excessive. This is exemplified in scRNA-seq data, where attempting global causal discovery to derive the gene regulatory network amongst approximately 20k genes is not only computationally expensive but also often redundant (Levine and Davidson, 2005). Local causal discovery, emphasizing the causal relations of a target variable and its neighbors, stands out as a more grounded and efficient approach. Additionally, techniques like divide-and-conquer and parallelization, when applied through local causal discovery, can often enhance the efficiency of identifying the global causal structure (Ma et al., 2023).

Building on this motivation, several studies have delved into the discovery of local structures within a select subset of variables (Margaritis and Thrun, 1999; Yin et al., 2008; Zhou et al., 2010; Niinimaki and Parviainen, 2012; Wang et al., 2014; Gao and Ji, 2015; Gao et al., 2017; Ling et al., 2020; Ng et al., 2021; Yu et al., 2021; Gupta et al., 2023). The distinction in this line of research lies in the estimation approaches used to estimate these local structures, such as parent-child sets. These approaches range from testing conditional independence relations to employing likelihood-based score functions. With appropriate tests or scores, they can offer nonparametric guarantees. Yet, without parametric assumptions, both independence tests and score functions cannot uniquely determine all directions, leading to some edges being undirected.

Moreover, most existing work in local causal discovery assume that there are no cycles in the ground-truth structure. This constrains its applicability given that cycles frequently appear in real-world contexts. These cycles can arise from various origins, including feedback mechanisms in biological systems (Benito et al., 2007), electrical engineering (Mason, 1953), or economic processes (Haavelmo, 1943). Such cyclic relationships can have profound implications, reshaping our understanding of the systems under consideration. Furthermore, in local context, one often cannot make the assumption of global acyclicity, since there is no way for the acyclicity beyond the considered subset of variables to be testable. While there has been a steady progress on causal discovery with cycles (Spirtes, 1995; Richardson, 1996; Lacerda et al., 2008; Hyttinen et al., 2012; Mooij and Heskes, 2013; Ghassami et al., 2020), none have offered methodologies with theoretical guarantees in the context of local search.

[*]Equal contribution. Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

Contributions. To our knowledge, this work is the first to tackle local causal discovery in cyclic models, crucial for gene regulatory networks with prevalent feedback loops and numerous genes. Moreover, we allow intersecting cycles, a known challenging case. By leveraging non-Gaussianity, our approach determines causal directions and strengths, standing in contrast to most previous local methods that only identify partially directed edges. Notably, this work offers a unified perspective on acyclic (Shimizu et al., 2006, 2011) and cyclic (Lacerda et al., 2008) cases within the local context. We establish identifiability guarantees for all proposed methods, and our theoretical results have been validated in both synthetic and real-world data.

2 Problem Setup

2.1 Notations, Definitions, and the Goal

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a directed graph with the vertex set $\mathcal{V} = [d] \coloneqq \{1, 2, \ldots, d\}$ and the edge set $\mathcal{E}$. Denote a directed edge from vertex i to vertex j as $i \rightarrow j$.

Random variables $\mathbf{X} = (X_i)_{i=1}^d$ are generated by a linear non-Gaussian (LiNG) structural equation model (SEM) (Lacerda et al., 2008) w.r.t. the graph $\mathcal{G}$, described in the matrix form as

$$\mathbf{X} = \mathbf{B}\mathbf{X} + \mathbf{E}, \tag{1}$$

where $\mathbf{E} = (E_i)_{i=1}^d$ are mutually independent non-Gaussian *exogenous noise* components, and $\mathbf{B}$ is the *adjacency matrix*, with the entry $\mathbf{B}_{j,i}$ representing the *direct causal effect* of X_i on X_j. $\mathbf{B}_{j,i} \neq 0$ if and only if $i \rightarrow j \in \mathcal{E}$. Solving for $\mathbf{X}$ in Equation (1) gives

$$\mathbf{X} = \mathbf{A}\mathbf{E}, \text{ with } \mathbf{A} \coloneqq (\mathbf{I} - \mathbf{B})^{-1}, \tag{2}$$

i.e., $\mathbf{X}$ can also be expressed directly as a linear combination of the noises, through the *mixing matrix* $\mathbf{A}$. Following (Lacerda et al., 2008), we allow cycles in $\mathcal{G}$ under some mild assumptions (see Section 3), and interpret $\mathbf{X}$ as the equilibrium of the dynamic system.

For a vertex $i \in \mathcal{V}$, denote its *Markov blanket* (MB) in $\mathcal{G}$ as $\mathrm{mb}_{\mathcal{G}}(i) \coloneqq \mathrm{pa}_{\mathcal{G}}(i) \cup \mathrm{ch}_{\mathcal{G}}(i) \cup \mathrm{sps}_{\mathcal{G}}(i)$, the union of its *parents* $\mathrm{pa}_{\mathcal{G}}(i) \coloneqq \{j \in \mathcal{V} : j \rightarrow i \in \mathcal{E}\}$, *children* $\mathrm{ch}_{\mathcal{G}}(i) \coloneqq \{j \in \mathcal{V} : i \rightarrow j \in \mathcal{E}\}$, and *spouses* $\mathrm{sps}_{\mathcal{G}}(i) \coloneqq \{j \in \mathcal{V} \backslash (\mathrm{pa}_{\mathcal{G}}(i) \cup \mathrm{ch}_{\mathcal{G}}(i)) : \mathrm{ch}_{\mathcal{G}}(i) \cap \mathrm{ch}_{\mathcal{G}}(j) \neq \emptyset\}$. Assuming faithfulness, $\mathrm{mb}_{\mathcal{G}}(i)$ corresponds to the minimal set of variables conditioned on which all other variables are independent of X_i. Consequently, it is an appropriate starting point for the local search on vertex i.

For a target vertex T, we provide a method in Appendix A to efficiently estimate $\mathrm{mb}_{\mathcal{G}}(T)$ from $\mathbf{X}$ even in the presence of cycles. Specifically, we generalize the method developed by Loh and Bühlmann (2014) in the acyclic case based on inverse covariance matrix of the distributions. Hence in the following main results (Sections 3 and 4), we assume the oracle $\mathrm{mb}_{\mathcal{G}}(T)$ is available, and focus on the problem of further discovering causal effects related to T from $T, \mathrm{mb}_{\mathcal{G}}(T)$, and their corresponding variables.

2.2 LiNG SEM and its Global Estimation

Our definition of a linear non-Gaussian (LiNG) cyclic model precisely follows (Lacerda et al., 2008). We allow cycles in $\mathcal{G}$, interpret $\mathbf{X}$ as the equilibrium of the dynamic system. We allow overlapped cycles, but only assume that there are no "self-loops", i.e., $\mathbf{B}$ has all zeros in the diagonal, because by trivial scaling, any equilibrium even with self-loops can be equivalently entailed by another LiNG model without self-loops, as long as the self-loop strengths $\mathbf{B}_{i,i} \neq 1$. Moreover, we assume no cycles with strength exactly 1, i.e., $\mathbf{B}$ has no eigenvalues of 1, rendering $\mathbf{I} - \mathbf{B}$ invertible. See Section 1.2 of (Lacerda et al., 2008) for details.

Recall that Equation (2), $\mathbf{X} = \mathbf{A}\mathbf{E}$, is in the exact form of independent component analysis (ICA) (Comon, 1994; Hyvärinen and Oja, 2000), where observed data $\mathbf{X}$ (signals) is an unknown linear invertible mixture of unknown non-Gaussian independent components $\mathbf{E}$ (blind sources). When all the variables in $\mathbf{X}$ are involved, namely, with *causal sufficiency*, ICA can estimate a *demixing matrix* $\mathbf{W}$ to separate $\mathbf{X}$ into independent components $\mathbf{W}\mathbf{X}$. It is shown that $\mathbf{W}$ identifies $\mathbf{A}^{-1} = \mathbf{I} - \mathbf{B}$ up to rows permutation and scaling. Interestingly, with the structural constraint of zero diagonals in $\mathbf{B}$ (i.e., diagonal ones in $\mathbf{A}^{-1}$), these indeterminacies can be further reduced. A row permutation is called *admissible* if it makes $\mathbf{W}$ have diagonal ones with corresponding scaling. When $\mathcal{G}$ is acyclic, ICA-LiNGAM (Shimizu et al., 2006) shows that the admissible permutation is unique, resulting in the exact identification of $\mathbf{B}$. This is because for acyclic $\mathcal{G}$, its $\mathbf{B}$ can be simultaneously row and column permuted to be strictly lower triangular.

Lacerda et al. (2008) generalizes ICA-LiNGAM to cyclic cases with almost a same algorithmic procedure: it begins with an ICA on $\mathbf{X}$ to obtain a demixing matrix $\mathbf{W}$, and then identifies the *admissible* row permutations. The key distinction is that, in the presence of cycles in $\mathcal{G}$, there can be multiple admissible permutations. Denote the set of adjacency matrices recovered by all admissible permutations as $\mathcal{B}$, i.e.,

Definition 1. For a LiNG model $\mathbf{X} = \mathbf{B}\mathbf{X} + \mathbf{E}$, denote

$$\mathcal{B} \coloneqq \{\mathbf{B}' : \mathbf{B}' = \mathbf{I} - \mathbf{P}^{\pi}\mathbf{D}\mathbf{A}^{-1}, \mathrm{diag}(\mathbf{B}') = \mathbf{0}\},$$

where $\mathbf{A}^{-1} = \mathbf{I} - \mathbf{B}$, $\mathbf{D}$ is an d-dim scaling matrix, and $\mathbf{P}^{\pi}$ is a permutation matrix (see Section 3 for details) with π enumerating permutations of $\mathcal{V} = [d]$.

Two different LiNG models from $\mathcal{B}$ entail a same equilibrium distribution and are termed *distributionally equivalent*, though they share different graph structures; see Figure 8 in Appendix B for an example. Note that with linearity and non-Gaussianity, no two different acyclic SEMs are distributionally equivalent, guaranteeing the unique identification of $\mathbf{B}$ in LiNGAM, but there are different cyclic SEMs that are distributionally equivalent, and thus the true $\mathbf{B}$ can be identified up to an equivalence class. Lacerda et al. (2008) shows that $\mathcal{B}$, defined above from all admissible permutations, characterizes exactly the LiNG equivalence class for $\mathbf{X}$.

3 LOCAL LING DISCOVERY

We develop a local causal discovery method based on independent subspace analysis (ISA), which enables the exact identification of the equivalent local directed structures and causal strengths from the MB of the target variable. We first explain how the commonly used ICA approach for discovering global causal structure (Shimizu et al., 2006) fails in local context. We then describe the key identifiability result of ISA that we exploit, and provide a specific characterization of the ISA solution. Finally, we describe our proposed Local ISA-LiNG method, which involves (1) performing ISA on the local variables, (2) finding *admissible* permutations on the ISA solutions, and (3) identifying local structures and coefficients from the permuted solutions. We prove that, interestingly, with only local variables, our proposed algorithm can identify exactly what can be identified globally with all variables.

3.1 Independent Subspace Analysis

Having introduced the cyclic LiNG and its ICA-based global estimation method, we now turn to our local case. When only a subset of variables (e.g., a target T and its $\mathrm{mb}_{\mathcal{G}}(T)$) is involved, the main challenge lies in causal insufficiency: with hidden confounders, ICA cannot demix mutually independent components.

Example 1. In Figure 1(i), consider a target $T = 4$ with $\mathrm{mb}_{\mathcal{G}}(T) = \{2, 3\}$. With a confounder X_1 outside of T's MB, ICA is not applicable on $\{X_2, X_3, X_4\}$, as these three signals mix four sources ($\{E_1, E_2, E_3, E_4\}$), and no invertible matrix $\mathbf{W} \in Gl(3)$ can separate out any three mutually independent components. △

Such an issue is typically pronounced in overcomplete ICA (OICA) (Hyvärinen and Oja, 2000), where the number of observed signals is less than the number of mixed sources. There are indeed work on LiNGAM with hidden confounders using OICA (Hoyer et al., 2008), but OICA is known to be both computationally and statistically ineffective. In this work, our methods do not involve OICA, away from the difficulties of trying to separate out that many mutually independent sources from only a few signals. Instead, we only seek the separation "as independent as possible", and show that it is informative enough. To achieve this, independence subspace analysis (ISA) (Hyvärinen and Hoyer, 2000; Theis, 2006) comes into play.

Definition 2. An m-dim random vector $\mathbf{Z}$ is called *irreducible* if it contains no lower-dim independent components, i.e., no invertible matrix $\mathbf{W} \in Gl(m)$ can decompose $\mathbf{WZ} = (\mathbf{Z}'_1, \mathbf{Z}'_2)$ into independent $\mathbf{Z}'_1 \perp\!\!\!\perp \mathbf{Z}'_2$.

Definition 3. For an m-dim random vector $\mathbf{Y}$, an invertible matrix $\mathbf{W}$ is called an *independent subspace analysis (ISA)* solution of $\mathbf{Y}$ if $\mathbf{WY} = (\mathbf{Z}_1^\intercal, \ldots, \mathbf{Z}_k^\intercal)^\intercal$ consists of mutually independent, irreducible random vectors $\mathbf{Z}_i$. The corresponding partition $\mathbf{\Gamma_W}$ of indices $[m]$ is called the *ISA partition* associated with $\mathbf{W}$.

Given a random vector $\mathbf{Y}$ with existing covariance and no Gaussian components, Theis (2006) shows that an ISA solution of $\mathbf{Y}$ exists and, similar to ICA, is unique except for general scaling and permutation.

Before stating the result of ISA, we first introduce some notations. For a permutation $\pi : \{1, \ldots, m\} \mapsto \{1, \ldots, m\}$ of m elements, let π_i be the i-th element in π, and $\pi[j]$ be the index of element j in π, i.e., $\pi_{\pi[j]} = j$. For an ordered subset $\mathbf{S} \subset [m]$, denote $\pi_{\mathbf{S}} \coloneqq (\pi_i : i \in \mathbf{S})$ and $\pi[\mathbf{S}] \coloneqq (\pi[j] : j \in \mathbf{S})$, where $(\cdot)$ means ordered sets. Define the $m \times m$ permutation matrix $\mathbf{P}^\pi$ by $\mathbf{P}^\pi_{i,j} = 1$ if $\pi_i = j$ and 0 otherwise. Given a partition $\mathbf{\Gamma}$ of $[m]$, an $m \times m$ block diagonal matrix $\mathbf{D}$ is said to be a *general scaling matrix* consistent with $\mathbf{\Gamma}$, if $\forall \mathbf{S} \in \mathbf{\Gamma}$, $\mathrm{rank}(\mathbf{D}_{\mathbf{S},\mathbf{S}}) = |\mathbf{S}|$, and $\mathbf{D}_{\mathbf{S},[m]\setminus\mathbf{S}} = \mathbf{0}$. Here the notation like $\mathbf{D}_{\mathbf{S}_1,\mathbf{S}_2}$ means the submatrix of $\mathbf{D}$ with rows and columns indexed by ordered sets $\mathbf{S}_1$ and $\mathbf{S}_2$ respectively. Subscripts on random vectors denotes indexing similarly, e.g., $\mathbf{X_S} = (X_i : i \in \mathbf{S})$. We have:

Theorem 1 (Indeterminacy of ISA; Theorem 1.8 of (Theis, 2006))**.** *Given an m-dim random vector $\mathbf{Y}$, if both $\mathbf{W}_1$ and $\mathbf{W}_2$ are ISA solutions of $\mathbf{Y}$ with partitions $\mathbf{\Gamma}_{\mathbf{W}_1}, \mathbf{\Gamma}_{\mathbf{W}_2}$, then there exists a permutation π of $[m]$ and a general scaling matrix $\mathbf{D}$ consistent with $\mathbf{\Gamma}_{\mathbf{W}_1}$ s.t. $\mathbf{W}_2 = \mathbf{P}^\pi\mathbf{D}\mathbf{W}_1$, and $\forall \mathbf{S}_1 \in \mathbf{\Gamma}_{\mathbf{W}_1}$, $\exists \mathbf{S}_2 \in \mathbf{\Gamma}_{\mathbf{W}_2}$, with $\mathbf{S}_2$ and $\pi[\mathbf{S}_1]$ having the same elements.*

ISA can be identified up to such indetermincaies, and can be estimated as efficiently as square ICA (Theis, 2006). ICA can then be viewed as a special case of ISA, where all subspaces are of one-dimension.

3.2 One Specific ISA Characterization

Given a vertex subset $\mathbf{S} \subset \mathcal{V}$ and the corresponding variables $\mathbf{X_S}$, Section 3.1 shows that although ICA on

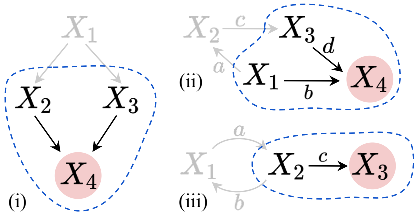


Figure 1: For Examples 1 to 3. On each $\mathcal{G}$, the target T is colored red, and its $\mathrm{mb}_{\mathcal{G}}(T)$ is circled by blue and colored dark. Same marks apply henceforth.

$\mathbf{X_S}$ may not be applicable, an ISA solution exists and is unique up to some indeterminacies. However, what is such an ISA solution? In the causally sufficient (i.e, ICA) case, a demixing matrix $\mathbf{A}^{-1} = \mathbf{I} - \mathbf{B}$ follows naturally from Equation (2), while this is less obvious in the ISA case. Below we give a specific characterization.

Theorem 2 (One characterization of ISA in LiNG model)**.** *Assume* $\mathbf{X}$ *follows a LiNG SEM* $\mathbf{X} = \mathbf{AE}$*. For any vertex subset* $\mathbf{S} \subset \mathcal{V}$*, the inverse of the principal submatrix of the mixing matrix* $\mathbf{A}$ *indexed by* $\mathbf{S}$*, denoted by* $\mathbf{A}^{-1}_{\mathbf{S},\mathbf{S}}$*, is an ISA solution of* $\mathbf{X_S}$*.*

Theorem 2 characterizes a specific ISA matrix $\mathbf{A}^{-1}_{\mathbf{S},\mathbf{S}}$ that separates $\mathbf{X_S}$ "as independent as possible", i.e., $\mathbf{A}^{-1}_{\mathbf{S},\mathbf{S}}\mathbf{X_S}$ produces irreducible independent subspaces. The proof is in Appendix D.1. However, before delving into further identification of $\mathbf{A}^{-1}_{\mathbf{S},\mathbf{S}}$, let us first closely examine and understand what it represents.

Recall that in ICA, the adjacency matrix $\mathbf{B}$ that represents the causal structure and strengths can be directly read off of the demixing characterization, $\mathbf{A}^{-1} = \mathbf{I} - \mathbf{B}$. However, in ISA, the local adjacencies may not be as so straightforward. $\mathbf{A}^{-1}_{\mathbf{S},\mathbf{S}}$ is the Schur complement of $[d]\backslash\mathbf{S}$ block in $\mathbf{I} - \mathbf{B}$. Typically, $\mathbf{A}^{-1}_{\mathbf{S},\mathbf{S}}$ does not equal $\mathbf{I} - \mathbf{B}_{\mathbf{S},\mathbf{S}}$, and $\mathbf{I} - \mathbf{B}_{\mathbf{S},\mathbf{S}}$ is not an ISA solution either:

Example 2. In Figure 1(ii), consider $\mathbf{S} = (1, 3, 4)$, i.e., a target $T = 4$ and its $\mathrm{mb}_{\mathcal{G}}(T) = \{1, 3\}$. The ISA characterization $\mathbf{A}^{-1}_{\mathbf{S},\mathbf{S}}$ separates $\mathbf{X_S}$ into three independent irreducible subspaces (1-dim components):

$$\mathbf{A}^{-1}_{\mathbf{S},\mathbf{S}}\mathbf{X_S} = \begin{bmatrix} 1 & 0 & 0 \\ -ac & 1 & 0 \\ -b & -d & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_3 \\ X_4 \end{bmatrix} = \begin{bmatrix} E_1 \\ cE_2 + E_3 \\ E_4 \end{bmatrix} \begin{matrix} \} \\ \} \\ \} \end{matrix},$$

but the local adjacencies $\mathbf{I} - \mathbf{B}_{\mathbf{S},\mathbf{S}} \neq \mathbf{A}^{-1}_{\mathbf{S},\mathbf{S}}$, and by

$$(\mathbf{I} - \mathbf{B}_{\mathbf{S},\mathbf{S}})\mathbf{X_S} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -b & -d & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_3 \\ X_4 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_3 \\ E_4 \end{bmatrix} \begin{matrix} \left.\vphantom{\begin{matrix} X_1 \\ X_3 \end{matrix}}\right\} \\ \} \end{matrix},$$

it is not an ISA solution, as it produces only two independent subspaces, of which the first one $(X_1^\intercal, X_3^\intercal)^\intercal$ is not irreducible with a decomposition $\begin{bmatrix} 1 & 0 \\ -ac & 1 \end{bmatrix}$. △

Write the matrix inverse in block form we will have:

$$\mathbf{I} - \mathbf{A}^{-1}_{\mathbf{S},\mathbf{S}} = \mathbf{B}_{\mathbf{S},\mathbf{S}} + \mathbf{B}_{\mathbf{S},\bar{\mathbf{S}}}(\mathbf{I} - \mathbf{B}_{\bar{\mathbf{S}},\bar{\mathbf{S}}})^{-1}\mathbf{B}_{\bar{\mathbf{S}},\mathbf{S}}, \tag{3}$$

where $\bar{\mathbf{S}} := \mathcal{V}\backslash\mathbf{S}$. By Equation (3), the (i, j)-th entry of $\mathbf{I} - \mathbf{A}^{-1}_{\mathbf{S},\mathbf{S}}$ corresponds not only to the direct causal effect from j to i, but also the total causal effect from j to i through all other variables outside of $\mathbf{S}$.

With this in mind, we note an issue on diagonals: while the global demixing matrix $\mathbf{A}^{-1} = \mathbf{I} - \mathbf{B}$ always has diagonal ones as we assume no self-loops, it may not be the case locally. Specifically, if $\mathcal{G}$ is acyclic, $\mathbf{A}^{-1}_{\mathbf{S},\mathbf{S}}$ still has diagonal ones, but this does not hold for cyclic $\mathcal{G}$:

Example 3. Consider the LiNG SEM in Figure 1(iii).

$$\mathbf{B} = \begin{bmatrix} 0 & b & 0 \\ a & 0 & 0 \\ 0 & c & 0 \end{bmatrix}; \quad \mathbf{A} = \frac{1}{1 - ab} \begin{bmatrix} 1 & b & 0 \\ a & 1 & 0 \\ ac & c & 1 - ab \end{bmatrix}.$$

Let $\mathbf{S} = (2, 3)$, i.e., $T = 3$ and its $\mathrm{mb}_{\mathcal{G}}(T) = \{2\}$,

$$\mathbf{A}^{-1}_{\mathbf{S},\mathbf{S}} = \begin{bmatrix} 1 - ab & 0 \\ -c & 1 \end{bmatrix},$$

where the diagonal entry on X_2 is not one. This is because X_2 is on a cycle outside of $\mathbf{S}$, which, from the local view of $\mathbf{S}$, is equivalent to a self-loop on X_2. The strength of this "self-loop" is thus unidentifiable. △

3.3 Local LiNG Identification from ISA

Having defined a specific ISA characterization $\mathbf{A}^{-1}_{\mathbf{S},\mathbf{S}}$, we are now left with the task to post-process any general ISA solution to this specific one (and its equivalence class, if any). Recall that in the global ICA case, the adjacency matrix equivalence class $\mathcal{B}$ directly stems from row permutations on any demixing matrix $\mathbf{W}$. However, with ISA, we face more complex cross-rows indeterminacies (Theorem 1). How to reduce them? Moreover, even if $\mathbf{A}^{-1}_{\mathbf{S},\mathbf{S}}$ is exactly recovered, a challenge lies still in translating it back into LiNG model parameters, as it may not directly represent adjacencies and may be unidentifiable due to external paths (Examples 2 and 3). Then, what is identifiable locally, and how? We address these questions below.

Consider a target vertex T and its $\mathrm{mb}_{\mathcal{G}}(T)$. Let $\mathbf{S}$ be $\{T\} \cup \mathrm{mb}_{\mathcal{G}}(T)$ with m elements. W.l.o.g., we rename vertices s.t. $\mathbf{S}$ reads 1 to m, i.e., $\mathbf{S} = [m] \subset [d] = \mathcal{V}$. Assume a LiNG SEM $\mathbf{X} = \mathbf{BX} + \mathbf{E} = \mathbf{AE}$. Perform ISA on $\mathbf{X_S}$, we obtain a solution $\mathbf{W}$ and the associated subspace partition $\mathbf{\Gamma_W}$. By Theorems 1 and 2, $\mathbf{W}$ can be row-permuted and subspace-scaled into $\mathbf{A}^{-1}_{\mathbf{S},\mathbf{S}}$.

3.3.1 Admissible Permutations

So the first step is to "re-permute" $\mathbf{W}$. Since columns of $\mathbf{W}$ correspond exactly to X_1 through X_m, rows

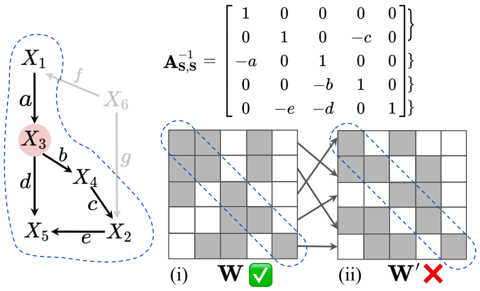

X_1
X_6
f
a
g
X_3
b
d
X_4
c
X_5
e
X_2
$\mathbf{A}_{\mathbf{S},\mathbf{S}}^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -c & 0 \\ -a & 0 & 1 & 0 & 0 \\ 0 & 0 & -b & 1 & 0 \\ 0 & -e & -d & 0 & 1 \end{bmatrix}$
(i) $\mathbf{W}$ ✅
(ii) $\mathbf{W}'$ ❌

Figure 2: For Example 4: row permutation on ISA matrices with nonzero diagonals can be entirely incorrect.

permutation of $\mathbf{W}$ can be seen as assigning names to each row, thereby forming their one-to-one correspondence to X_1 through X_m also. Intuitively, rows within a same multi-dim subspace always correspond to variables with common hidden confounders and are thus mutually unidentifiable. However, different subspaces collectively, especially singleton subspaces (components), should be re-identified to their correct locations. Nonetheless, we first note that the nonzero-diagonal permutation as in ICA, is incorrect here:

Example 4. Consider an acyclic $\mathcal{G}$ as in Figure 2. Let $\mathbf{S}$ be $(1, 2, 3, 4, 5)$, i.e., a target $T = 3$ and its $\mathrm{mb}_{\mathcal{G}}(T)$. The true but unknown $\mathbf{A}_{\mathbf{S},\mathbf{S}}^{-1}$ is provided for reference. We are only given an ISA output $\mathbf{W}$, as in (i), and its $\mathbf{\Gamma_W} = \{(1, 2), (3), (4), (5)\}$. Actually, $\mathbf{W}$ is just scaled from $\mathbf{A}_{\mathbf{S},\mathbf{S}}^{-1}$ without permutation, though this is unknown. Comparing $\mathbf{W}$ to $\mathbf{A}_{\mathbf{S},\mathbf{S}}^{-1}$, we notice that within the subspace of the 1st and 2nd rows, the nonzero entries are mixed by the general scaling. If we were to still follow the "admissible" criteria of nonzero diagonals as in ICA, we see that $\mathbf{W}$ is already satisfied (and is indeed correct). However, another permutation $\mathbf{W}'$ as in (ii), is also satisfied but is entirely incorrect (even on singletons' locations), leading to incorrect edges like 2 → 4, 3 → 1.

Why does incorrect permutation (ii) occur? Note that $\mathbf{A}_{\mathbf{S},\mathbf{S}}^{-1}$ possesses a unique row permutation (itself) with nonzero diagonals, so the blame falls on the scaling to 1st and 2nd rows with more nonzeros. Fortunately, these spurious nonzeros reveal themselves via rank deficiency. In (ii), even though nonzero diagonals exist, the diagonal block of the first subspace, $\mathbf{W}'_{(2,4),(2,4)}$, is proportional to $[1, \ -c]$ and has rank 1. Inspired by this, we can eliminate spurious nonzeros by forcing not the nonzero diagonal entries as in ICA, but the invertible diagonal blocks, formally described below. △

Definition 4. Given an ISA solution $\mathbf{W}$ and the associated partition $\mathbf{\Gamma_W}$, a permutation π is called *admissible*, if $\forall \mathbf{S}_i \in \mathbf{\Gamma_W}$, $\mathrm{rank}((\mathbf{P}^\pi \mathbf{W})_{\pi[\mathbf{S}_i],\pi[\mathbf{S}_i]}) = |\mathbf{S}_i|$.

Admissible permutations defined in Definition 4 are "sound and complete". See Appendix D.4 for formal definition and proof. Roughly speaking, all such admissible rows permutations correspond exactly to all rows permutations on $\mathbf{A}_{\mathbf{S},\mathbf{S}}^{-1}$ with nonzero diagonals (viewing each subspace collectively), which then correspond exactly to the LiNG equivalence class on $\mathbf{S}$.

3.3.2 Identifiable Local Causal Effects

Having admissible permutations, now we proceed to identify local causal structures and coefficients. As demonstrated in Examples 2 and 3, ISA matrices is not overall reliable. However, note that the misidentification of an X_i on both examples can be attributed to an incoming path (either in a cycle or not) outside of $\mathbf{S}$. This immediately sparks us that if all of i's parents are included in $\mathbf{S}$, such issues should not arise:

Lemma 1. *Given an ISA solution $\mathbf{W}$ and $\mathbf{\Gamma_W}$ on $\mathbf{X_S}$, $\forall i \in \mathbf{S}$, if $\mathrm{pa}_{\mathcal{G}}(i) \subset \mathbf{S}$, then its exogenous noise component E_i is separated out, i.e., $\exists j \in [m]$ s.t. $(j) \in \mathbf{\Gamma_W}$ and $(\mathbf{W}\mathbf{X_S})_j = cE_i$ with a scaling factor c. Moreover, the incoming causal strengths to X_i are identified up to c, i.e., the row vector $\mathbf{W}_j = c(\mathbf{I} - \mathbf{B}_{\mathbf{S},\mathbf{S}})_i$.*

Lemma 1 becomes especially helpful in our case: by definition of MB, for any of T and its children, all its parents are included in $\{T\} \cup \mathrm{mb}_{\mathcal{G}}(T)$, thus blocking all confounding paths, enabling recovery of exogenous noise, and moreover, the exact causal strengths. Once all edges into T and T's children are identified, we've attained the goal of local causal discovery, as these edges include all edges to and from T. As for other variables in MB, e.g., parents and spouses, they may be entangled within subspaces and remain unidentifiable, but this does not pose a concern anymore.

By Lemma 1, T and its children produce independent components (1-dim subspaces) by ISA. But conversely, an unconfounded parent or spouse can also produce a 1-dim subspace. Then, which of these components correspond exactly to our main focus, T and its children? The answer can be read off of T's column in $\mathbf{W}$:

Lemma 2. *Given an ISA solution $\mathbf{W}$ and $\mathbf{\Gamma_W}$ on $\mathbf{X_S}$ for $\mathbf{S} = \{T\} \cup \mathrm{mb}_{\mathcal{G}}(T)$. Denote by $\mathbf{C} \coloneqq \mathrm{supp}(\mathbf{W}_{:,T}) = (i \in [m] : \mathbf{W}_{i,T} \neq 0)$. Then $\forall i \in \mathbf{C}$, $\mathbf{W}_i$ must produce a single component, i.e., $(i) \in \mathbf{\Gamma_W}$. Moreover, $\{\pi[\mathbf{C}] : \pi$ admissible to $\mathbf{W}\} = \{\mathrm{supp}(\mathbf{B}'_{:,T}) : \mathbf{B}' \in \mathcal{B}\}$.*

In essence, Lemma 2 interprets the nonzero row indices on T's column vector in $\mathbf{W}$ as T and T's children. Note that there can be multiple directed graphs in the LiNG equivalence class, leaving different choices of $\mathbf{S}$ subsets as T's children. Any such choice can be interpreted by an admissible row permutation, and vice versa.

Algorithm 1 Local ISA-LiNG

```
Input: A target $T \in \mathcal{V}$, its oracle MB $\mathrm{mb}_{\mathcal{G}}(T)$, and data $\mathbf{X}$. Assume w.l.o.g. $\mathbf{S} := \{T\} \cup \mathrm{mb}_{\mathcal{G}}(T) = [m]$
Output: A set of directed weighted edge sets
Initialize the output set $\mathcal{K} := \emptyset$;
Obtain an ISA solution $\mathbf{W}$ with $\mathbf{\Gamma_W}$ on $\mathbf{X_S}$;
Set $\mathbf{C} := \mathrm{supp}(\mathbf{W}_{:,T}) = (i \in [m] : \mathbf{W}_{i,T} \neq 0)$;
for any permutation $\pi$ admissible to $\mathbf{W}, \mathbf{\Gamma_W}$ do
    Initialize $\mathbf{K} := \emptyset$;
    Set $\mathbf{W}' := \mathbf{P}^{\pi}\mathbf{W}$;
    Set scaling matrix $\mathbf{D}$: $\forall \mathbf{S}_i \in \mathbf{\Gamma_W}$, $\mathbf{D}_{\pi[\mathbf{S}_i],\pi[\mathbf{S}_i]} := (\mathbf{W}'_{\pi[\mathbf{S}_i],\pi[\mathbf{S}_i]})^{-1}$ and $\mathbf{D}_{\pi[\mathbf{S}_i],[m]\setminus\pi[\mathbf{S}_i]} := \mathbf{0}$;
    Set $\mathbf{B}' := \mathbf{I} - \mathbf{D}\mathbf{W}'$;
    for $i \in \mathbf{C}$ do
        Assert $(i) \in \mathbf{\Gamma_W}$;
        Add to $\mathbf{K}$ a weighted edge denoted as $(j \to \pi[i], \mathbf{B}'_{\pi[i],j})$, for each $j \in [m]$ with $\mathbf{B}'_{\pi[i],j} \neq 0$;
    end for
    Set $\mathcal{K} := \mathcal{K} \cup \{\mathbf{K}\}$;
end for
Return $\mathcal{K}$;
```

3.3.3 The Local ISA-LiNG Algorithm

Finally, we have the local ISA-LiNG Algorithm 1. Below we give an illustrative example on how it works:

Example 5. Consider the example in Figure 3. There are two graphs in the global equivalence class $\mathcal{B}$, as shown in the upper row. Let $\mathbf{S}$ be $(1, 2, 3, 4, 5)$, i.e., a target $T = 3$ and its $\mathrm{mb}_{\mathcal{G}}(T)$. An ISA on $\mathbf{X_S}$ gives $\mathbf{W}$ with nonzero patterns as in the lower left matrix, where specifically, the striped entries are nonzero but rank deficient (see Definition 4). The 3rd (T-th) column has three nonzero entries, corresponding to T and its two children, which are yet unknown and can be different in different equivalent graphs. Two admissible rows permutations (the lower row) reveal their variable correspondences, with all edges into T and its children (dark edges in the graphs) recovered correctly for all equivalent graphs with different global directed cycles, while note these local edges themselves are acyclic. △

Theorem 3 (Correctness of local ISA-LiNG)**.** *For any $T \in \mathcal{V}$, let $\mathcal{K}$ be set of weighted edge sets returned by Algorithm 1 on T, $\mathrm{mb}_{\mathcal{G}}(T)$, and $\mathbf{X}$. We have:*

$$\mathcal{K} = \{\{(i \to j, \mathbf{B}'_{j,i}) : \forall j \in \{T\} \cup \mathrm{ch}_{\mathcal{G}'}(T), \forall i \in \mathrm{pa}_{\mathcal{G}'}(j)\} : \forall \mathbf{B}' \in \mathcal{B}, \text{ and the graph } \mathcal{G}' \text{ defined by } \mathbf{B}'\}.$$

The local ISA-LiNG algorithm (Algorithm 1) correctly identifies all the causal effects into the target T and all its children, for all LiNG models that equivalently entails the distribution of $\mathbf{X}$. That is, with only local variables, we identify exactly what can be identified globally. Note that this identification is unique (i.e., the returned $\mathcal{K}$ consists of a single item) if and only if none of T and $\mathrm{ch}_{\mathcal{G}}(T)$ is part of any cycles in $\mathcal{G}$ (including the case where $\mathcal{G}$ is acyclic).

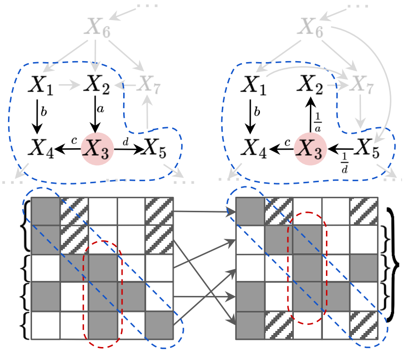


Figure 3: For Example 5, to illustrate Algorithm 1.

3.4 With the Notion of Stability

The $\mathcal{B}$ defined in Definition 1 characterizes the entire global LiNG equivalence class, yet not all models within it are "stable". In dynamical systems, stability refers to "the dissipation of the effects of one-time noise in models" (Lacerda et al., 2008). Applied to causal models, a model is "stable" when any infinitely long path (after traversing loops) result in zero causal effect. For example, in a simple cycle with two variables and two edges both carrying weights 2, the model is unstable, with the cycle product of $4 > 1$ leading to explosion. When both weights are 0.5, the model remains LiNG equivalent to the former one but achieves stability, with the cycle product of $0.25 < 1$. Formally, a global LiNG model is said to be *stable* when its adjacency matrix $\mathbf{B}$ is convergent, i.e., $\lim_{t\to\infty} \mathbf{B}^t = \mathbf{0}$. Note that here the entry $(\mathbf{B}^t)_{i,j}$ represents the summed causal effect from j to i along all paths of length t.

In practical global causal discovery scenarios, an often-made assumption is the stability of the underlying LiNG model, and people usually focus on identifying the stable LiNG model(s), instead of the entire equivalence class $\mathcal{B}$. This is straightforward in ICA-LiNG (Lacerda et al., 2008): as the entire $\mathcal{B}$ can be recovered first, we then only need to check the convergence of each item within $\mathcal{B}$. However, when we only have local variables, can we still recover the local part corresponding to the global stable model(s)?

The answer is affirmative but with constraints: our method can still correctly find stable solutions locally, as long as this local stable solution is identifiable. Denote the stable sub-equivalence class as $\mathcal{B}^* := \{\mathbf{B} \in \mathcal{B} : \lim_{t\to\infty} \mathbf{B}^t = \mathbf{0}\}$. When cycles in the ground-truth $\mathcal{G}$ are disjoint, there exists a unique global stable model, i.e., $|\mathcal{B}^*| = 1$. Let $\mathbf{B}^*$ be this unique stable adjacency matrix, and $\mathcal{G}^*$ be the corresponding graph. In this case, simply by adhering to an additional *local stability* condition, the stable solution can be identified locally:

Corollary 1 (Identifying stable solutions locally, with disjoint cycles)**.** *Suppose the cycles are disjoint in $\mathcal{G}$. Consider a modified version of Algorithm 1 in which the line "**if** $\mathbf{B}'$ is not convergent: **skip**" is added between lines 8 and 9. Then, this modified version of Algorithm 1 will yield a single local model, corresponding exactly to the unique global stable model. That is, the returned $\mathbb{K}$ consists of a single item $\mathbf{K}$, with*

$$\mathbf{K} = \{(i \to j, \mathbf{B}^*_{j,i}) : \forall j \in \{T\} \cup \mathrm{ch}_{\mathcal{G}^*}(T), \forall i \in \mathrm{pa}_{\mathcal{G}^*}(j)\}.$$

Corollary 1 is valid as here stability is determined sufficiently and necessarily by the cycle products, which is preserved locally. However, when some cycles in $\mathcal{G}$ intersect, the situation becomes more complex. Globally, there may be none or multiple global stable models in $\mathcal{B}^*$. Locally, while in this case, our Algorithm 1 can still identify local correspondences of all equivalent solutions (Theorem 3), the exact identification of the global stable solutions from local variables alone becomes inherently impossible. Intuitively, this is because that external cycles appear as self-loops on the local variables. More details are in Appendix D.

4 REGRESSION-BASED VARIANT

In Section 3 we propose a local ISA-based method suitable for both acyclic and cyclic graphs. In this section, we focus on the specific scenario where there are no cycles in $\mathcal{G}$, i.e., $\mathbf{X}$ follows a linear non-Gaussian *acyclic* model (LiNGAM (Shimizu et al., 2006)), and propose an alternative local regression-based method. The relationship between this section and Section 3 can be likened to that of Direct-LiNGAM (Shimizu et al., 2011) and ICA-LiNG (Lacerda et al., 2008) in the global context, with the former utilizing non-Gaussianity by ICA, and the latter by Darmois-Skitovitch theorem (Darmois, 1953; Skitovitch, 1953).

Acyclicity renders the existence of a *causal ordering*, i.e., vertices $\mathcal{V}$ can be ordered so that no later vertex has a direct edge onto any earlier variable. When all the variables in $\mathbf{X}$ are involved, namely, with *causal sufficiency*, Shimizu et al. (2011) gives the method named Direct-LiNGAM to uniquely identify the DAG $\mathcal{G}$ by estimating its causal ordering: Regress X_j on X_i, if the residual is statistically independent with the regressor X_i, then i is causally earlier than j. If such an independence holds for X_i on all its pairwise regressions with the remaining X_js, i must be a *root* vertex. Subroots are then recursively identified in a same way, forming the causal ordering.

However, when only a subset of variables (as of here, $\{T\} \cup \mathrm{mb}_{\mathcal{G}}(T)$) is involved, Direct-LiNGAM does not work, as causal sufficiency is violated, and there can be no independent residual due to hidden confounders, just like the absent independent components in ICA.

Example 6. In Figure 1(i), with a confounder X_1 outside of T's MB, neither regressing X_2 on X_3 nor the converse results in independent residuals, making it impossible to identify any "local root" in $\mathrm{mb}_{\mathcal{G}}(T)$. △

While identifying "local roots" is impossible, can we reverse our perspective from top-down to bottom-up and identify "local leaves" instead? Interestingly, the answer seems affirmative: In Example 6, regressing X_4 on $\{X_2, X_3\}$, the residual is exactly the exogenous noise E_4 and is independent to $\{X_2, X_3\}$. Formally, for any vertex subset $\mathbf{S} \subset \mathcal{V}$, we denote the corresponding random vector as $\mathbf{X_S} \coloneqq [X_i : i \in \mathbf{S}]^\intercal$. Perform ordinary least square error linear regression of a random variable X_i on a random vector $\mathbf{X_S}$, the asymptotic coefficients of fit is $\beta_{\mathbf{S}\to i} \coloneqq \mathrm{cov}(\mathbf{X_S}, \mathbf{X_S})^{-1} \mathrm{cov}(\mathbf{X_S}, X_i)$, where for $j \in \mathbf{S}$, $\beta^j_{\mathbf{S}\to i}$ is the coefficient on X_j. Denote the regression residual as $R_{\mathbf{S}\to i} \coloneqq X_i - \beta^\intercal_{\mathbf{S}\to i}\mathbf{X_S}$. Denote i's descendants in $\mathcal{G}$ as $\mathrm{des}_{\mathcal{G}}(i)$. We have:

Lemma 3. *For any $i \in \mathcal{V}, \mathbf{S} \subset \mathcal{V}\backslash\{i\}$, if $R_{\mathbf{S}\to i} \perp\!\!\!\perp \mathbf{X_S}$, i.e., independent residual, then $\mathbf{S} \cap \mathrm{des}_{\mathcal{G}}(i) = \emptyset$.*

Lemma 3 generalizes regressions in (Shimizu et al., 2011) from single variables to multi-dim vectors, but with a same idea: independent residuals imply causal ordering. While as in Example 6, independent residuals may be absent for "local roots" due to confounders (echoed as multi-dim subspaces in ISA), they must exist for "local leaves" (echoed as the 1-dim components in ISA). This is because, again, as in Lemma 1, that parents of T and its children are included in the MB, thus blocking all confounding paths, enabling recovery of exogenous noise and the exact causal strengths:

Lemma 4. *$\forall i, \mathbf{S}$ in $\mathcal{V}$, if $\mathrm{pa}_{\mathcal{G}}(i) \subset \mathbf{S} \subset \mathcal{V}\backslash \mathrm{des}_{\mathcal{G}}(i)$, then $\forall j \in \mathbf{S}$, $\beta^j_{\mathbf{S}\to i} = \mathbf{B}_{i,j}$, and $R_{\mathbf{S}\to i} = E_i$ (so $\perp\!\!\!\perp \mathbf{X_S}$).*

Lemma 4 holds generally for linear acyclic SEMs, echoing the local Markov property: given all its parents, a variable is independent of other non-descendants, enabling accurate estimation of direct effects to it. Lemmas 3 and 4 then readily leads to Algorithm 2.

The basic idea of Algorithm 2 is to recursively identify "local leaves", i.e., all T's children until T, via an "inverse causal ordering". Independent residuals must exist for these variables, as all their parents are included locally (lines 7-9). Lines 3-6 serve to avoid errors due to spouses, which could be hidden confounded:

Example 7. In Figure 4(i), $T = 1$, $\mathrm{mb}_{\mathcal{G}}(T) = \{2, 3, 4, 5\}$. After X_5 is first identified and removed, the remaining two "last leaves" X_3, X_4 are confounded by X_6 hidden outside of $\mathrm{mb}_{\mathcal{G}}(T)$, and thus neither can produce independent residual. The iteration cannot proceed, unless these two spouses are removed. △

Algorithm 2 Inverse Direct-LiNGAM

```
Input: A target vertex $T \in \mathcal{V}$, its oracle MB $\mathrm{mb}_{\mathcal{G}}(T)$, and their corresponding variables in $\mathbf{X}$
Output: A set of directed edges with weights
1: Initialize the remaining vertex set $\mathbf{U} := \{T\} \cup \mathrm{mb}_{\mathcal{G}}(T)$, and the output edge set $\mathbf{K} := \emptyset$;
2: while $\mathbf{U} \neq \emptyset$ do
3:     if $\exists k \in \mathbf{U} \backslash \{T\}$ s.t. $\beta^{T}_{\mathbf{U}\backslash\{k\}\to k} = 0$ then
4:         Set $\mathbf{U} := \mathbf{U}\backslash\{k\}$;
5:         continue to line 2;
6:     end if
7:     Assert $\exists j \in \mathbf{U}$ s.t. $R_{\mathbf{U}\backslash\{j\}\to j} \perp\!\!\!\perp \mathbf{X}_{\mathbf{U}\backslash\{j\}}$;
8:     Set $j$ as any one found in line 7;
9:     Add to $\mathbf{K}$ an edge $i \to j$ with weight $\beta^{i}_{\mathbf{U}\backslash\{j\}\to j}$ for each $i \in \mathbf{U}\backslash\{j\}$ with $\beta^{i}_{\mathbf{U}\backslash\{j\}\to j} \neq 0$;
10:    break if $j = T$; Otherwise set $\mathbf{U} := \mathbf{U}\backslash\{j\}$;
11: end while
12: Return $\mathbf{K}$;
```

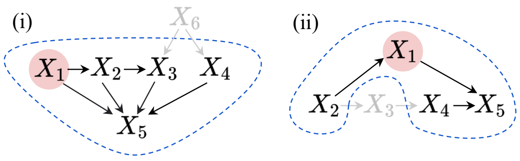


Figure 4: Examples to illustrate Algorithm 2.

Even without confounders, edges produced by independent residuals may still be incorrect due to spouses:

Example 8. In Figure 4(ii), $T = 1$, $\mathrm{mb}_{\mathcal{G}}(T) = \{2, 4, 5\}$. After X_5 is first identified and removed, though the "last leaf" X_4 produces independent residual regressing on X_1, X_2, due to hidden X_3, the coefficient on X_1 is nonzero, yielding an incorrect edge 1 → 4. Correction requires removing this spouse X_4. △

With spouses corrected, Algorithm 2 accurately estimate all edges into T and its children, including edges adjacent to T (the purpose of local search). Formally,

Theorem 4 (Correctness of Algorithm 2)**.** *For any $T \in \mathcal{V}$, let $\mathbf{K}$ be the weighted edge set returned by Algorithm 2 on T, $\mathrm{mb}_{\mathcal{G}}(T)$, and $\mathbf{X}$. We have:*

$$\mathbf{K} = \{(i \to j, \mathbf{B}_{j,i}) : \forall j \in \{T\} \cup \mathrm{ch}_{\mathcal{G}}(T), \forall i \in \mathrm{pa}_{\mathcal{G}}(j)\}.$$

Theorem 4 is similar to Theorem 3, except that the DAG can be uniquely identified. See Appendix D.8 for the proof, and Appendix C for also an alternative post-processing of ISA, with the same "ordering" idea here.

5 EXPERIMENTS

We assess the effectiveness of our method for cyclic and acyclic cases in Sections 5.1 and 5.2, respectively. We provide an analysis of how our method performs under different sample sizes in Section 5.3, and an experiment on real data in Section 5.4. The implementation details and running times are discussed in Appendices E and F.1, respectively.

5.1 Cyclic Case

We conduct experiments to illustrate the output of our method, by adopting the left cyclic graph in Figure 3 as ground truth. We simulate 2000 samples from the LiNG SEM in Equation (1), of which the nonzero weights of $\mathbf{B}$ are sampled uniformly from $[-0.9, -0.5] \cup [0.5, 0.9]$, and each exogenous noise E_i is sampled uniformly from $[-c_i, c_i]$ to the power of 5, with c_i sampled randomly from $[0.75, 1.25]$.

We first run the ICA-LiNG method by Lacerda et al. (2008) on all variables. To perform local causal discovery, we also run our Local ISA-LiNG method on target $T = 3$ and its MB $\{1, 2, 4, 5\}$. An example of the outputs by both methods, including the estimated edge weights, are provided in Figure 10 in Appendix F.2. One observes that our method correct identifies the edges according to Theorem 3, and that the estimated edge weights are close to the true ones.

With stability. To conduct a systematic validation, we restrict the cycles in the true graphs to be disjoint and the true B matrices to be stable using an accept-reject approach; that is, the spectral radius of B has to be strictly smaller than one. In this case, Corollary 1 indicates that the stable solution can be uniquely identified locally. We simulate 50-node directed cyclic graphs (DCGs) with maximum degree of 4, and 2000 samples from the LiNG SEM in Equation (1). We use the same setup described above for the edge weights and noise distributions. To perform local causal discovery, we randomly select a target T that is part of a cycle in the 50-node DCGs. Due to the lack of local causal discovery baselines that handle cyclic graphs, we compare our method with those for acyclic cases, including GSBN (Margaritis and Thrun, 1999), Local A* (Ng et al., 2021), CMB (Gao and Ji, 2015), and LDECC (Gupta et al., 2023). Note that GSBN and Local A* require information of two-step MBs (i.e., $\mathrm{mb}_{\mathcal{G}}(T)$ and MB of each variable in $\mathrm{mb}_{\mathcal{G}}(T)$), which are not directly comparable to our method that requires only $\mathrm{mb}_{\mathcal{G}}(T)$; thus, we consider modifications of these methods, described in Appendix E. We report the structural Hamming distance (SHD) of local DCG, which is explained in details in Appendix E.3.

We provide the results for the methods using estimated MB in Figure 5, and using oracle MB in Figure 11 in Appendix F.2. It is observed that our method achieves much lower SHD in both settings, thereby demonstrating its effectiveness for identifying the local structure.

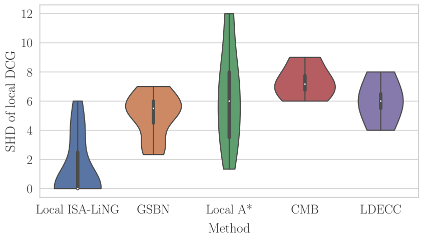


Figure 5: SHD of local DCG under estimated MB.

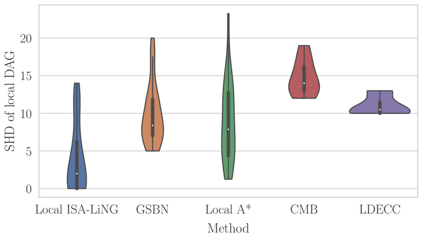


Figure 6: SHD of local DAG under estimated MB.

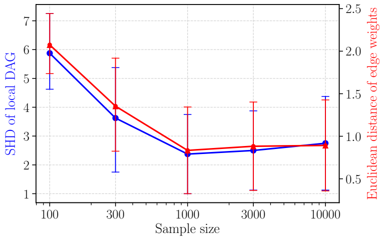


Figure 7: Local ISA-LiNG under oracle MB. X-axis is in log scale.

5.2 Acyclic Case

We consider the acyclic setting where the ground truths are DAGs. In the acyclic case , we use a more efficient post-processing procedure for demixing matrix $\mathbf{W}$, described in Appendix C. We simulate 50-node Erdös–Rényi (Erdös and Rényi, 1959) DAGs, and 2000 samples from the LiNG SEM in Equation (1) using the same setting (including edge weights and noise distributions) as that of Section 5.1. To perform local causal discovery, we consider target T from 50-node DAGs with expected degrees of 3 and 5, leading to roughly 14 and 20 variables in the MB $\mathrm{mb}_{\mathcal{G}}(T)$, respectively. We report the SHD of local DAG and partially DAG (PDAG), explained in Appendix E.3.

For degree of 3, the SHDs of local DAG for the methods using estimated MB are shown in Figure 12, while the complete results using estimated MB and oracle MB are given in Figures 12 and 13 in Appendix F.2, respectively, due to space limit. We provide the results for degree of 5 in Figure 14 in Appendix F.2. Similar to the cyclic case, our method achieves much lower SHD for both local DAG and PDAG as compared to the baselines. One also observes that GSBN and Local A* performs better than CMB and LDECC.

5.3 Analysis of Different Sample Sizes

We provide an analysis of the proposed method across sample sizes $n \in \{100, 300, 1000, 3000, 10000\}$, following the data generating procedure in Section 5.2. We report the SHD of local DAG and the Euclidean distance between the estimated edge weights and the true ones. The results using oracle MB is shown in Figure 7, while those using estimated MB are given in Figure 15 in Appendix F.2. As the sample size increases, both metrics decrease to small values close to zero, which help validate the asymptotic correctness of our method in terms of both structure and parameter estimation. This also demonstrates the possibility of reliable estimation even when the sample size is rather limited.

Moreover, we provide the scatter plots of the estimated and true edge weights in Figures 16 and 17 in Appendix F.2. For larger sample sizes, the data points are increasingly grouped onto the main diagonal, showing that the estimated weights become more accurate.

5.4 Real Data

We compare our method with GSBN and Local A* on a standard real-world dataset that collects continuous expression levels of proteins and phospholipids within human immunological cells (Sachs et al., 2005), characterized by 853 observational samples and a ground truth DAG with 11 variables and 17 edges. Here, we select PIP2, PIP3, and Akt as target variables, and compute the SHD of local DAG obtained by all three methods. As shown in the Table 1, our method achieves lower SHD in most cases. A detailed comparison of ground-truth and estimated local causal structures can be found in Figure 18 in Appendix F.2.

Table 1: SHD of different local causal discovery methods on real data by Sachs et al. (2005).

Target	Ours	GSBN	Local A*
PIP2	**1**	1.5	3.6
PIP3	**1**	**1**	4
Akt	**1**	1.3	1.3

6 CONCLUSION

We have expanded local causal discovery to include cyclic scenarios by generalizing the classic LiNGAM-based methods. Notably, while previous local search methods based on conditional independence tests or likelihood-based scores often fail to determine the direction of certain edges, our method leverages non-Gaussianity to enable more precise edge orientations. This leads to a more comprehensive representation of the causal graph, even in cyclic contexts. Additionally, we have established identifiability guarantees for all our proposed methods. These theoretical findings have been validated using various datasets in both synthetic and real-world settings. Future work includes characterizing the number of possible structures in the cyclic equivalence class estimated by our method.

Acknowledgements

The authors would like to thank reviewers for their helpful comments. This material is based upon work supported by the AI Research Institutes Program funded by the National Science Foundation under AI Institute for Societal Decision Making (AI-SDM), Award No. 2229881. The project is also partially supported by the National Institutes of Health (NIH) under Contract R01HL159805, and grants from Apple Inc., KDDI Research Inc., Quris AI, and Infinite Brain Technology.

References

J. Benito, H. Zheng, F. S. Ng, and P. E. Hardin. Transcriptional feedback loop regulation, function and ontogeny in Drosophila. In *Cold Spring Harbor symposia on quantitative biology*, volume 72, page 437. NIH Public Access, 2007.

P. Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

H. Dai, P. Spirtes, and K. Zhang. Independence testing-based approach to causal discovery under measurement error and linear Non-Gaussian models. *Advances in Neural Information Processing Systems*, 35:27524–27536, 2022.

G. Darmois. Analyse générale des liaisons stochastiques: etude particulière de l'analyse factorielle linéaire. *Revue de l'Institut international de statistique*, pages 2–8, 1953.

P. Erdös and A. Rényi. On random graphs I. *Publicationes Mathematicae*, 6:290–297, 1959.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics*, 9:432–41, 2008.

T. Gao and Q. Ji. Local causal discovery of direct causes and effects. *Advances in Neural Information Processing Systems*, 28, 2015.

T. Gao, K. Fadnis, and M. Campbell. Local-to-global Bayesian network structure learning. In *International Conference on Machine Learning*, pages 1193–1202. PMLR, 2017.

A. Ghassami, A. Yang, N. Kiyavash, and K. Zhang. Characterizing distribution equivalence and structure learning for cyclic and acyclic directed graphs. In *International Conference on Machine Learning*, pages 3494–3504. PMLR, 2020.

A. Gretton, K. Fukumizu, C. Teo, L. Song, B. Schölkopf, and A. Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems*, 2007.

S. Gupta, D. Childers, and Z. C. Lipton. Local causal discovery for estimating causal effects. In *Conference on Causal Learning and Reasoning*, pages 408–447. PMLR, 2023.

T. Haavelmo. The statistical implications of a system of simultaneous equations. *Econometrica, Journal of the Econometric Society*, pages 1–12, 1943.

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

A. Hyttinen, F. Eberhardt, and P. O. Hoyer. Learning linear cyclic causal models with latent variables. *The Journal of Machine Learning Research*, 13(1):3387–3439, 2012.

A. Hyvärinen and P. Hoyer. Emergence of phase-and shift-invariant features by decomposition of natural images into independent feature subspaces. *Neural computation*, 12(7):1705–1720, 2000.

A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

M. Janjic. A proof of generalized Laplace's expansion theorem. *Bull. Soc. Math. Banja Luka*, 15(2008): 5–7, 2008.

G. Lacerda, P. Spirtes, J. Ramsey, and P. O. Hoyer. Discovering cyclic causal models by independent components analysis. In *Conference on Uncertainty in Artificial Intelligence*, 2008.

M. Levine and E. H. Davidson. Gene regulatory networks for development. *Proceedings of the National Academy of Sciences*, 102(14):4936–4942, 2005.

Z. Ling, K. Yu, H. Wang, L. Li, and X. Wu. Using feature selection for local causal structure learning. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(4):530–540, 2020.

P.-L. Loh and P. Bühlmann. High-dimensional learning of linear causal networks via inverse covariance estimation. *Journal of Machine Learning Research*, 15(88):3065–3105, 2014.

S. Ma, J. Wang, C. Bieganek, R. Tourani, and C. Aliferis. Local causal pathway discovery for single-cell rna sequencing count data: a benchmark study. *Journal of Translational Genetics and Genomics*, 7 (1):50–65, 2023.

D. Margaritis and S. Thrun. Bayesian network induction via local neighborhoods. *Advances in neural information processing systems*, 12, 1999.

S. J. Mason. Feedback theory-some properties of signal flow graphs. *Proceedings of the IRE*, 41(9):1144–1156, 1953.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34:1436–1462, 2006.

J. Mooij and T. Heskes. Cyclic causal discovery from continuous equilibrium data. *arXiv preprint arXiv:1309.6849*, 2013.

I. Ng, Y. Zheng, J. Zhang, and K. Zhang. Reliable causal discovery with improved exact search and weaker assumptions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

T. Niinimaki and P. Parviainen. Local structure discovery in Bayesian networks. *arXiv preprint arXiv:1210.4888*, 2012.

P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5, 2011.

T. S. Richardson. *Discovering cyclic causal structure*. Carnegie Mellon [Department of Philosophy], 1996.

D. Rothenhäusler, C. Heinze, J. Peters, and N. Meinshausen. Backshift: Learning causal cyclic graphs from unknown shift interventions. *Advances in Neural Information Processing Systems*, 28, 2015.

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

R. Sanchez-Romero, J. D. Ramsey, K. Zhang, M. R. Glymour, B. Huang, and C. Glymour. Estimating feedforward and feedback effective connections from fMRI time series: Assessments of statistical methods. *Network Neuroscience*, 3(2):274–306, 2019.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

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

S. Shimizu, T. Inazumi, Y. Sogawa, A. Hyvärinen, Y. Kawahara, T. Washio, P. O. Hoyer, and K. Bollen. DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *Journal of Machine Learning Research*, 12 (Apr):1225–1248, 2011.

V. P. Skitovitch. On a property of the normal distribution. *DAN SSSR*, 89:217–219, 1953.

P. Spirtes. Conditional independence in directed cyclic graphical models representing feedback or mixtures. Technical report, Philosophy, Methodology and Logic Technical Report 59, CMU, 1994.

P. Spirtes. Directed cyclic graphical representations of feedback models. In *Conference on Uncertainty in Artificial Intelligence*, 1995.

F. Theis. Towards a general independent subspace analysis. In *Advances in Neural Information Processing Systems*, 2006.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

I. Tsamardinos, C. Aliferis, and A. Statnikov. Algorithms for large scale Markov blanket discovery. pages 376–381, 2003.

I. Tsamardinos, L. Brown, and C. Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65:31–78, 10 2006.

C. Wang, Y. Zhou, Q. Zhao, and Z. Geng. Discovering and orienting the edges connected to a target variable in a DAG via a sequential local learning approach. *Computational Statistics & Data Analysis*, 77:252–266, 2014.

J. Yin, Y. Zhou, C. Wang, P. He, C. Zheng, and Z. Geng. Partial orientation and local structural learning of causal networks for prediction. In *Causation and Prediction Challenge*, pages 93–105. PMLR, 2008.

K. Yu, Z. Ling, L. Liu, P. Li, H. Wang, and J. Li. Feature selection for efficient local-to-global Bayesian network structure learning. *ACM Transactions on Knowledge Discovery from Data*, 2021.

C. Yuan and B. Malone. Learning optimal Bayesian networks: A shortest path perspective. *Journal of Artificial Intelligence Research*, 48(1):23–65, 2013.

K. Zhang and L.-W. Chan. ICA with sparse connections. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 530–537. Springer, 2006.

Y. Zhou, C. Wang, J. Yin, and Z. Geng. Discover local causal network around a target to a given depth. In *Causality: Objectives and Assessment*, pages 191–202. PMLR, 2010.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] See Section 2 for mathematical setting, and Sections 3 and 4 for assumptions and algorithms.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes] See Section 5.3 and Appendix F.1 for the analysis.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes] The code is available on GitHub.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes] See Sections 3 and 4 for the assumptions.
   (b) Complete proofs of all theoretical results. [Yes] See Appendix D for the proofs.
   (c) Clear explanations of any assumptions. [Yes] See Section 2 and 3 for the explanations.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes] The code is available on GitHub.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes] See Appendix E.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes] See the figures of the experiments.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes] See Appendix E.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Local Causal Discovery with Linear non-Gaussian Cyclic Models: Supplementary Materials

A MARKOV BLANKET DISCOVERY FOR CYCLIC GRAPHS

The local causal discovery procedures presented in Sections 3 and 4 rely on knowledge about the MB of the target variable T i.e., its parents, children, and spouses. To the best of our knowledge, many existing MB estimation methods, such as those based on nonparametric conditional independence test, e.g., GSMB (Margaritis and Thrun, 1999), IAMB (Tsamardinos et al., 2003), and MMMB (Tsamardinos et al., 2006), focus on the Bayesian network (i.e., acyclic) setting. That is, it may not be immediately clear how their estimated MB relates to the true one $\mathrm{mb}_{\mathcal{G}}(T)$ in the presence of cycles, partly owing to the extra complications involved when handling cycles with conditional independence tests (Spirtes, 1994). In this section, we provide a method to estimate the MB of a variable from a linear cyclic SEM. Specifically, we build upon the method proposed by Loh and Bühlmann (2014) that, similar to methods based on conditional independence tests, makes the acyclicity assumption, and further generalize it to handle cycles.

We first define the moral graph of a directed cyclic graph the same way as that of a DAG. Specifically, the moral graph of directed graph $\mathcal{G}$ is an undirected graph that contains an edge between two nodes if (1) they are adjacent in $\mathcal{G}$, or (2) they share the same children. Clearly, the MB of a variable is simply its neighbors in the moral graph of $\mathcal{G}$. Here, we provide a method to estimate such moral graph, which informs us about $\mathrm{mb}_{\mathcal{G}}(T)$. Considering the linear SEM in Equation (1), the inverse covariance matrix of the distribution of variables $\mathbf{X}$ is given by $\mathbf{\Theta} = (\mathbf{I} - \mathbf{B})\mathbf{\Omega}^{-1}(\mathbf{I} - \mathbf{B})^{\intercal}$, where $\mathbf{\Omega} \coloneqq \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2) \coloneqq \mathrm{cov}(\mathbf{E})$ is the covariance matrix of exogenous noise components $\mathbf{E}$. Inspired by Loh and Bühlmann (2014, Assumption 1) in the acyclic case, we make the following assumption in the cyclic case.

Assumption 1. *Let $\mathbf{B}$ and $\mathbf{\Omega}$ be the weighted adjacency matrix and noise covariance matrix, respectively, of the linear SEM in Equation (1). For every $j < i$, we have*

$$-\sigma_j^{-2}\mathbf{B}_{i,j} - \sigma_i^{-2}\mathbf{B}_{j,i} + \sum_{\ell \neq j,i} \sigma_\ell^{-2}\mathbf{B}_{j,\ell}\mathbf{B}_{i,\ell} = 0, \tag{4}$$

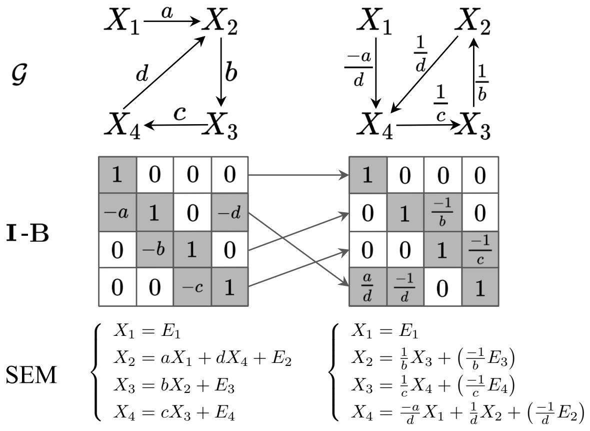
only if $\mathbf{B}_{i,j} = \mathbf{B}_{j,i} = 0$ and $\mathbf{B}_{j,\ell}\mathbf{B}_{i,\ell} = 0$ for all $\ell \neq j, i$.

As we will show in the proof, the LHS of Equation (4) is equal to $\mathbf{\Theta}_{j,i}$. It is worth noting that if the nonzero coefficients of $\mathbf{B}$ are randomly drawn from a distribution that is absolutely continuous with respect to Lebesgue measure, then the above assumption is only violated for a set of matrices $\mathbf{B}$ with zero Lebesgue measure. We then have the following proposition, with a proof given in Appendix D.9. Note that the proposition and its proof are built upon Loh and Bühlmann (2014, Theorem 2) in the acyclic case, which we generalize to the cyclic case.

Proposition 1. *Suppose $\mathbf{X}$ follows the linear SEM in Equation (1) with directed cyclic graph $\mathcal{G}$ and inverse covariance matrix $\mathbf{\Theta}$. Under Assumption 1, the structure defined by the support of $\mathbf{\Theta}$ is the same as the moral graph of $\mathcal{G}$.*

Asymptotically speaking, the true inverse covariance $\mathbf{\Theta}$ can be estimated by computing the inverse of empirical covariance matrix. For finite samples, Ravikumar et al. (2011) established high dimensional guarantee for estimating the support of $\mathbf{\Theta}$ using graphical Lasso (Friedman et al., 2008). An alternative approach (Meinshausen and Bühlmann, 2006) is to perform nodewise regression with Lasso (Tibshirani, 1996), which we adopt in this work. That is, we regress the target T on the other variables $[d] \setminus \{T\}$ with Lasso, from which the nonzero coefficients determine the MB of T.

B Illustrative Examples



$$\begin{array}{cccc} 1 & 0 & 0 & 0 \\ -a & 1 & 0 & -d \\ 0 & -b & 1 & 0 \\ 0 & 0 & -c & 1 \end{array} \qquad \begin{array}{cccc} 1 & 0 & 0 & 0 \\ 0 & 1 & \frac{-1}{b} & 0 \\ 0 & 0 & 1 & \frac{-1}{c} \\ \frac{a}{d} & \frac{-1}{d} & 0 & 1 \end{array}$$

$$\begin{cases} X_1 = E_1 \\ X_2 = aX_1 + dX_4 + E_2 \\ X_3 = bX_2 + E_3 \\ X_4 = cX_3 + E_4 \end{cases} \qquad \begin{cases} X_1 = E_1 \\ X_2 = \frac{1}{b}X_3 + (\frac{-1}{b}E_3) \\ X_3 = \frac{1}{c}X_4 + (\frac{-1}{c}E_4) \\ X_4 = \frac{-a}{d}X_1 + \frac{1}{d}X_2 + (\frac{-1}{d}E_2) \end{cases}$$

Figure 8: Example of two equivalent cyclic LiNG models.

This is an illustrative example of the global LiNG equivalence class $\mathcal{B}$ defined in Definition 1 of Section 2.2.

C POST-PROCESSING FOR LOCAL ISA-LING

In this section, we provide an alternative post-processing procedure to obtain the estimated structures and edge weights from the ISA solution $\mathbf{W}$, described in Algorithm 3. This procedure assumes that none of T and $\mathrm{ch}_{\mathcal{G}}(T)$ is part of any cycles in $\mathcal{G}$ (including the case where $\mathcal{G}$ is acyclic). The overall idea is similar to that of the regression-based approach described in Algorithm 2. That is, Algorithm 3 iteratively finds the "sink" node from the ISA solution that is not an ancestor of the other nodes in the remaining vertex set.

Algorithm 3 Alternative post-processing procedure of ISA solution

```
Input: A target vertex $T \in \mathcal{V}$, its oracle MB $\mathrm{mb}_{\mathcal{G}}(T)$, and ISA solution $\mathbf{W}$
Output: A set of directed edges with weights
1: Initialize the remaining vertex set $\mathbf{U}_1, \mathbf{U}_2 \coloneqq \{T\} \cup \mathrm{mb}_{\mathcal{G}}(T)$, and the output edge set $\mathbf{K} \coloneqq \emptyset$;
2: while $\mathbf{U}_1 \neq \emptyset$ do
3:     Assert $\exists j \in \mathbf{U}_1$ s.t. $\|\mathbf{W}_{\mathbf{U}_2,j}\|_0 = 1$;
4:     Set $j$ as any one found in line 7;
5:     Let $k \in \mathbf{U}_2$ be s.t. $\mathbf{W}_{k,j} = 1$;
6:     if $\mathbf{W}_{k,T} \neq 0$ then
7:         Add to $\mathbf{K}$ an edge $i \to j$ with weight $\mathbf{W}_{k,i}$ for each $i \in \mathbf{U}_1 \backslash \{j\}$ with $\mathbf{W}_{k,i} \neq 0$;
8:     end if
9:     break if $j = T$; Otherwise set $\mathbf{U}_1 \coloneqq \mathbf{U}_1 \backslash \{j\}$ and $\mathbf{U}_2 \coloneqq \mathbf{U}_2 \backslash \{k\}$;
10: end while
11: Return $\mathbf{K}$;
```

D PROOFS OF MAIN RESULTS

D.1 Proof of Theorem 2

Theorem 2 (One characterization of ISA in LiNG model)**.** *Assume* $\mathbf{X}$ *follows a LiNG SEM* $\mathbf{X} = \mathbf{A}\mathbf{E}$*. For any vertex subset* $\mathbf{S} \subset \mathcal{V}$*, the inverse of the principal submatrix of the mixing matrix* $\mathbf{A}$ *indexed by* $\mathbf{S}$*, denoted by* $\mathbf{A}^{-1}_{\mathbf{S},\mathbf{S}}$*, is an ISA solution of* $\mathbf{X}_{\mathbf{S}}$*.*

Proof. For convenience denote $\mathbf{W} := \mathbf{A}^{-1}_{\mathbf{S},\mathbf{S}}$. Write the ISA demixed subspaces as exogenous noise combinations:

$$\begin{aligned}
\mathbf{W}\mathbf{X}_{\mathbf{S}} &= \begin{bmatrix} \quad \mathbf{W} \quad \end{bmatrix} \cdot \mathbf{S}\left\{ \begin{bmatrix} \overbrace{\qquad\qquad \mathbf{A}_{\mathbf{S},:} \qquad\qquad}^{\mathcal{V} := [d]} \end{bmatrix} \right. \cdot \mathbf{E} \\
&= \begin{bmatrix} \mathbf{A}^{-1}_{\mathbf{S},\mathbf{S}} \end{bmatrix} \cdot \mathbf{S}\left\{ \left[\begin{array}{c|c} \overbrace{\mathbf{A}_{\mathbf{S},\mathbf{S}}}^{\mathbf{S}} & \overbrace{\qquad \mathbf{A}_{\mathbf{S},\bar{\mathbf{S}}} \qquad}^{\bar{\mathbf{S}} := \mathcal{V}\backslash\mathbf{S}} \end{array} \right] \right. \cdot \mathbf{E} \\
&= \mathbf{S}\left\{ \left[\begin{array}{c|c} \overbrace{\quad \mathbf{I} \quad}^{\mathbf{S}} & \overbrace{\qquad \cdots \qquad}^{\bar{\mathbf{S}} := \mathcal{V}\backslash\mathbf{S}} \end{array} \right] \right. \cdot \mathbf{E}
\end{aligned}$$

where $\mathbf{E} = (E_1^\intercal, \cdots, E_d^\intercal)^\intercal$ are the mutually independent exogenous non-Gaussian noise components.

To show that $\mathbf{W}$ is an ISA of $\mathbf{X}_{\mathbf{S}}$, we want to show that for any subspace $\mathbf{Z}_i \in (\mathbf{Z}_1^\intercal, \dots, \mathbf{Z}_k^\intercal)^\intercal = \mathbf{W}\mathbf{X}_{\mathbf{S}}$ with $m := |\mathbf{Z}_i| \geq 2$ (otherwise it's already a single component; $|\cdot|$ denotes dimension or cardinality), $\mathbf{Z}_i$ is irreducible (Definition 2), i.e., there exists no invertible matrix $\mathbf{H} \in Gl(m)$ s.t. $\mathbf{H}\mathbf{Z}_i$ produces two or more independent subspaces (random vectors). For convenience, we denote the row indices corresponding to the row-submatrix of $\mathbf{W}$ that produces the subspace $\mathbf{Z}_i$ as $\mathbf{M}$ ($\mathbf{M} \subset \mathbf{S}$), i.e., $\mathbf{Z}_i = \mathbf{W}_{\mathbf{M},:}\mathbf{X}_{\mathbf{S}}$. Similarly, rewrite it to noise combinations:

$$\begin{aligned}
\mathbf{W}\mathbf{X}_{\mathbf{S}} =\ & \mathbf{M}\left\{ \begin{bmatrix} \overbrace{\mathbf{W}_{\mathbf{M},:}}^{\mathbf{S}} \end{bmatrix} \right. \cdot \mathbf{S}\left\{ \begin{bmatrix} \overbrace{\qquad\qquad \mathbf{A}_{\mathbf{S},:} \qquad\qquad}^{\mathcal{V} := [d]} \end{bmatrix} \right. \cdot \mathbf{E} \\
=\ & \mathbf{M}\left\{ \left[\begin{array}{c|c} \overbrace{\quad \mathbf{I} \quad}^{\mathbf{M}} & \overbrace{\qquad \cdots \qquad}^{\bar{\mathbf{M}} := \mathcal{V}\backslash\mathbf{M}} \end{array} \right] \right. \cdot \mathbf{E} \qquad (5) \\
=:\ & \mathbf{C}_{\mathbf{M}} \cdot \mathbf{E},
\end{aligned}$$

where we denote the $m \times d$ rectangle mixing submatrix in Equation (5) as $\mathbf{C}_{\mathbf{M}}$. We do not use letter $\mathbf{A}$ for distinguishment, as it is multiplied by $\mathbf{W}$, and is different from submatrix from the original $\mathbf{A}$ mixing matrix.

Suppose for contradiction that $\mathbf{Z}_i$ is irreducible, i.e., there exists an invertible matrix $\mathbf{H} \in Gl(m)$ s.t. $\mathbf{H}\mathbf{Z}_i$ produces at least two independent subspaces, then, there must exist a partition $\mathbf{P}_1, \mathbf{P}_2$ of $[d]$ s.t.,

$$\begin{aligned}
\mathbf{H}\mathbf{Z}_i =\ & \mathbf{H} \cdot \mathbf{C}_{\mathbf{M}} \cdot \mathbf{E} \\
=\ & \left[\begin{array}{c} \mathbf{H}_1 \\ \hline \mathbf{H}_2 \end{array} \right] \cdot \mathbf{M}\left\{ \left[\begin{array}{c|c} \overbrace{\quad \mathbf{C}_{\mathbf{M},\mathbf{P}_1} \quad}^{\mathbf{P}_1} & \overbrace{\quad \mathbf{C}_{\mathbf{M},\mathbf{P}_2} \quad}^{\mathbf{P}_2 = \mathbf{M}\backslash\mathbf{P}_1} \end{array} \right] \right. \cdot \mathbf{E} \qquad (6) \\
=\ & \left[\begin{array}{c|c} \overbrace{\qquad \mathbf{0} \qquad}^{\mathbf{P}_1} & \overbrace{\qquad \cdots \qquad}^{\mathbf{P}_2 = \mathbf{M}\backslash\mathbf{P}_1} \\ \hline \cdots & \mathbf{0} \end{array} \right], \qquad (7)
\end{aligned}$$

i.e., $\mathbf{H}_1\mathbf{Z}_i$ and $\mathbf{H}_2\mathbf{Z}_i$ are linear combinations of disjoint sets of exogenous noise components, and thus by the Darmois-Skitovitch theorem (Darmois, 1953; Skitovitch, 1953), they are mutually independent.

By Equations (6) and (7) we have that row vectors of $\mathbf{H}_1$ lie in nullspace($\mathbf{C}^{\intercal}_{\mathbf{M},\mathbf{P}_1}$), and row vectors of $\mathbf{H}_2$ lie in nullspace($\mathbf{C}^{\intercal}_{\mathbf{M},\mathbf{P}_2}$). Also, since $\mathbf{C}_{\mathbf{M},\mathbf{M}} = \mathbf{I}$, rank($\mathbf{C}_{\mathbf{M}}$) $= m$ (i.e., full row rank), so,

$$\text{rank}(\mathbf{C}_{\mathbf{M},\mathbf{P}_1}) + \text{rank}(\mathbf{C}_{\mathbf{M},\mathbf{P}_2}) \geq \text{rank}(\mathbf{C}_{\mathbf{M},\mathbf{P}_1}|\mathbf{C}_{\mathbf{M},\mathbf{P}_2}) = m,$$

and thus

$$\begin{aligned} m - \text{nullity}(\mathbf{C}^{\intercal}_{\mathbf{M},\mathbf{P}_1}) + m - \text{nullity}(\mathbf{C}^{\intercal}_{\mathbf{M},\mathbf{P}_2}) &\geq= m, \\ \text{i.e., } \text{nullity}(\mathbf{C}^{\intercal}_{\mathbf{M},\mathbf{P}_1}) + \text{nullity}(\mathbf{C}^{\intercal}_{\mathbf{M},\mathbf{P}_2}) &\leq m \end{aligned}$$

Consider the following two cases:

1° When nullity($\mathbf{C}^{\intercal}_{\mathbf{M},\mathbf{P}_1}$) + nullity($\mathbf{C}^{\intercal}_{\mathbf{M},\mathbf{P}_2}$) $< m$, even when these two nullspaces are linearly independent, the number of their supports is less than m and there are not enough number of linearly independent row vectors to fill into $\mathbf{H}_1$ and $\mathbf{H}_2$ to form an invertible $\mathbf{H}$. Contradicted with our hypothesis.

2° When nullity($\mathbf{C}^{\intercal}_{\mathbf{M},\mathbf{P}_1}$) + nullity($\mathbf{C}^{\intercal}_{\mathbf{M},\mathbf{P}_2}$) $= m$, the above independence condition $\mathbf{H}_1\mathbf{Z}_i \perp\!\!\!\perp \mathbf{H}_2\mathbf{Z}_i$ is nontrivial (i.e., both are still random vectors with covariance, instead of a collapsing constant zero) only when:

$$\begin{cases} \text{nullity}(\mathbf{C}^{\intercal}_{\mathbf{M},\mathbf{P}_1}) > 0 \\ \text{nullity}(\mathbf{C}^{\intercal}_{\mathbf{M},\mathbf{P}_2}) > 0 \end{cases}$$

However, this is impossible:

Suppose for contradiction that $0 < \text{rank}(\mathbf{C}_{\mathbf{M},\mathbf{P}_1}), \text{rank}(\mathbf{C}_{\mathbf{M},\mathbf{P}_2}) < m$, then at least the $\mathbf{C}_{\mathbf{M},\mathbf{M}} = \mathbf{I}$ part must be separated, i.e., $\begin{cases} \mathbf{M} \not\subset \mathbf{P}_1 \\ \mathbf{M} \not\subset \mathbf{P}_2 \end{cases}$. Then, there must be a partition of $\mathbf{M}$ into into smaller respective subsets $(\mathbf{M}_u, \mathbf{M}_v)$ (we do not use $\mathbf{M}_1, \mathbf{M}_2$ to distinguish from the row indices for $\mathbf{H}_1, \mathbf{H}_2$) s.t. $\begin{cases} \mathbf{M}_u \subset \mathbf{P}_1 \\ \mathbf{M}_v \subset \mathbf{P}_2 \end{cases}$, then, since $\text{rank}(\mathbf{C}_{\mathbf{M},\mathbf{P}_1}) = |\mathbf{M}_u|$ and $\mathbf{C}_{\mathbf{M}_u,\mathbf{M}_u} = \mathbf{I}$, $\mathbf{C}_{\mathbf{M}_v,\mathbf{M}_u}$ must be all zeros. Further, since the $\mathbf{M}_v$ rows are linear combinations of the $\mathbf{M}_u$ rows, $\mathbf{C}_{\mathbf{M}_v,\mathbf{P}_1\backslash\mathbf{M}_u}$ must also be all zeros. Same applies to $\mathbf{C}_{\mathbf{M},\mathbf{P}_2}$. We have:

$$\mathbf{C}_{\mathbf{M},\mathbf{P}_1} = \begin{array}{c|c|c} & \overbrace{\mathbf{M}_u} & \overbrace{\mathbf{P}_1\backslash\mathbf{M}_u} \\ \mathbf{M}_u\{ & \mathbf{I} & \cdots \\ \hline \mathbf{M}_v\{ & \mathbf{0} & \mathbf{0} \end{array} \quad \text{and} \quad \mathbf{C}_{\mathbf{M},\mathbf{P}_2} = \begin{array}{c|c|c} & \overbrace{\mathbf{M}_v} & \overbrace{\mathbf{P}_2\backslash\mathbf{M}_v} \\ \mathbf{M}_u\{ & \mathbf{0} & \mathbf{0} \\ \hline \mathbf{M}_v\{ & \mathbf{I} & \cdots \end{array},$$

However, in this case, $\mathbf{C}_{\mathbf{M}_u,:}\mathbf{E} \perp\!\!\!\perp \mathbf{C}_{\mathbf{M}_v,:}\mathbf{E}$, as they share disjoint non-Gaussian $\mathbf{E}$ components. This contradicts with the initial hypothesis on a nontrivial subspace $\mathbf{Z}_i$, as $\mathbf{M}_u$ and $\mathbf{M}_v$ in $\mathbf{S}$ will not be mixed in $\mathbf{M}$, but rather produce two independence subspaces at the very beginning.

From the above contradiction, every $\mathbf{Z}_i$ must be irreducible. So $\mathbf{A}^{-1}_{\mathbf{S},\mathbf{S}}$ is an ISA. □

Note that while we are not the first to use ISA in linear non-Gaussian models with latent variables, this work is, to the best of our knowledge, the first with an identifiability guarantee. As shown in Examples 1 to 4, the characterization and post-processing of ISA are highly nontrivial. Some prior works (Sanchez-Romero et al., 2019) simply treated ISA solutions like ICA solutions and applied the same post-processing, resulting in inaccuracies. Other works used ISA mainly for downstream steps e.g., OICA (Hoyer et al., 2008) or independence tests (Dai et al., 2022), but not directly for the LiNG model identification. We believe that the generalized characterization of ISA solutions provided in Theorem 2 can be helpful for future works on causal discovery with latent variables.

D.2 Proof of Lemma 1

Lemma 1. *Given an ISA solution* $\mathbf{W}$ *and* $\mathbf{\Gamma_W}$ *on* $\mathbf{X_S}$*,* $\forall i \in \mathbf{S}$*, if* $\text{pa}_{\mathcal{G}}(i) \subset \mathbf{S}$*, then its exogenous noise component* E_i *is separated out, i.e.,* $\exists j \in [m]$ *s.t.* $(j) \in \mathbf{\Gamma_W}$ *and* $(\mathbf{W}\mathbf{X_S})_j = cE_i$ *with a scaling factor* c*. Moreover, the incoming causal strengths to* X_i *are identified up to* c*, i.e., the row vector* $\mathbf{W}_j = c(\mathbf{I} - \mathbf{B}_{\mathbf{S},\mathbf{S}})_i$*.*

Proof. For any vertex i and vertex set $\mathbf{S}$ of $\mathcal{V}$ with $i \in \mathbf{S}$ and $\mathrm{pa}_{\mathcal{G}}(i) \subset \mathbf{S}$, we can write the variable X_i as

$$\begin{aligned} X_i &= \mathbf{A}_i \mathbf{E} && (8)\\ &= \mathbf{B}_{i,\mathrm{pa}_{\mathcal{G}}(i)} \mathbf{X}_{\mathrm{pa}_{\mathcal{G}}(i)} + E_i && (9)\\ &= \mathbf{B}_{i,\mathbf{S}} \mathbf{X}_{\mathbf{S}} + E_i && (10)\\ &= \mathbf{B}_{i,\mathbf{S}} \mathbf{A}_{\mathbf{S}} \mathbf{E} + E_i && (11)\end{aligned}$$

where subscripts of index of indices denote the corresponding row/column submatrices. Equation (9) to Equation (10) is trivial because i has no parents from outside of $\mathbf{S}$, i.e., $\mathbf{B}_{i,\mathbf{S}\setminus \mathrm{pa}_{\mathcal{G}}(i)} = \mathbf{0}$.

By Equation (8)=Equation (11), we have

$$\mathbf{A}_i = \mathbf{B}_{i,\mathbf{S}} \mathbf{A}_{\mathbf{S}} + \mathbb{1}_i^{|\mathbf{S}|}, \tag{12}$$

where $\mathbb{1}_i^{|\mathbf{S}|}$ denotes the row vector of dimension $|\mathbf{S}|$ with only the i-th indexed entry being one, and elsewhere zeros. Equation (12) tells that all noise components ("ancestors") coming into X_i, except for the E_i itself, must go through $\mathrm{pa}_G(i)$.

Keep only the columns of $\mathbf{S}$ on Equation (12), we have

$$\mathbf{A}_{i,\mathbf{S}} = \mathbf{B}_{i,\mathbf{S}} \mathbf{A}_{\mathbf{S},\mathbf{S}} + \mathbb{1}_i^{|\mathbf{S}|}, \tag{13}$$

By Equation (13) we have

$$(\mathbf{A}_{i,\mathbf{S}} - \mathbb{1}_i^{|\mathbf{S}|}) \mathbf{A}_{\mathbf{S},\mathbf{S}}^{-1} = \mathbf{B}_{i,\mathbf{S}}, \tag{14}$$

where note that $\mathbf{A}_{\mathbf{S},\mathbf{S}}^{-1}$ is exactly the ISA characterization (Section 3.2) for $\mathbf{X}_{\mathbf{S}}$. Expand Equation (14) we have

$$\begin{aligned} &\mathbf{A}_{i,\mathbf{S}} \mathbf{A}_{\mathbf{S},\mathbf{S}}^{-1} - \mathbb{1}_i^{|\mathbf{S}|} \mathbf{A}_{\mathbf{S},\mathbf{S}}^{-1} = \mathbf{B}_{i,\mathbf{S}}, \text{ i.e.,}\\ &\mathbb{1}_i^{|\mathbf{S}|} - (\mathbf{A}_{\mathbf{S},\mathbf{S}}^{-1})_i = \mathbf{B}_{i,\mathbf{S}}, && (15)\end{aligned}$$

Equation (15) tells that the i-th row of the ISA characterization $\mathbf{A}_{\mathbf{S},\mathbf{S}}^{-1}$ is exactly $\mathbb{1}_i^{|\mathbf{S}|} - \mathbf{B}_{i,\mathbf{S}}$, i.e., the i-th row of $\mathbf{I} - \mathbf{B}_{\mathbf{S},\mathbf{S}}$. In other words, $(\mathbf{A}_{\mathbf{S},\mathbf{S}}^{-1})_i \mathbf{A}_{\mathbf{S}} = \mathbb{1}_i^{|\mathbf{S}|}$. Then, substitute Equation (15) into the demixed subspaces,

$$(\mathbf{A}_{\mathbf{S},\mathbf{S}}^{-1})_i \mathbf{X}_{\mathbf{S}} = X_i - \mathbf{B}_{i,\mathbf{S}} \mathbf{X}_{\mathbf{S}} = E_i,$$

i.e., the indpendent component (1-dim subspace) of E_i is exactly recovered.

Finally, with the subspace-wise permutation and scaling indeterminacies of ISA (Theorem 1), there must be a row in any ISA solution $\mathbf{W}$ being proportional to $(\mathbf{I} - \mathbf{A}_{\mathbf{S},\mathbf{S}})_i$, and the decomposed component also. □

D.3 Proof of Lemma 2

Lemma 2. *Given an ISA solution* $\mathbf{W}$ *and* $\mathbf{\Gamma}_{\mathbf{W}}$ *on* $\mathbf{X}_{\mathbf{S}}$ *for* $\mathbf{S} = \{T\} \cup \mathrm{mb}_{\mathcal{G}}(T)$*. Denote by* $\mathbf{C} \coloneqq \mathrm{supp}(\mathbf{W}_{:,T}) = (i \in [m] : \mathbf{W}_{i,T} \neq 0)$*. Then* $\forall i \in \mathbf{C}$*,* $\mathbf{W}_i$ *must produce a single component, i.e.,* $(i) \in \mathbf{\Gamma}_{\mathbf{W}}$*. Moreover,* $\{\pi[\mathbf{C}] : \pi$ *admissible to* $\mathbf{W}\} = \{\mathrm{supp}(\mathbf{B}'_{:,T}) : \mathbf{B}' \in \mathcal{B}\}$*.*

The proof is apparent and is almost the same as the above for Lemma 1: since all of T's children is included (as the "all parents" in that of Lemma 1), all the weights outgoing from T can also be correctly estimated. This can also be seen from the expression of $(\mathbf{A}_{\mathbf{S},\mathbf{S}}^{-1})_{i,T}$ entries in Equation (3), that ISA has indeterminacies of subspace-wise permutations and scalings, and that rows permutations of $(\mathbf{A}_{\mathbf{S},\mathbf{S}}^{-1})_{i,T}$ with nonzero diagonal entries directly correspond to each of that on $\mathbf{A}_{\mathbf{S},\mathbf{S}}^{-1}$ (the equivalence class $\mathcal{B}$ in Definition 1).

D.4 Proof of Theorem 3

Theorem 3 (Correctness of local ISA-LiNG)**.** *For any $T \in \mathcal{V}$, let $\mathcal{K}$ be set of weighted edge sets returned by Algorithm 1 on T, $\mathrm{mb}_{\mathcal{G}}(T)$, and $\mathbf{X}$. We have:*

$$\mathcal{K} = \{\{(i \to j, \mathbf{B}'_{j,i}) : \forall j \in \{T\} \cup \mathrm{ch}_{\mathcal{G}'}(T), \forall i \in \mathrm{pa}_{\mathcal{G}'}(j)\} : \\ \forall \mathbf{B}' \in \mathcal{B}, \textit{ and the graph } \mathcal{G}' \textit{ defined by } \mathbf{B}'\}.$$

To show the correctness of Algorithm 1, we first show the correctness of the "admissible" block permutations defined in Definition 4. To put it formally, we have the following lemma:

Lemma 5. *Let $\mathbf{C}$ be an arbitrary $m \times m$ invertible matrix, $\mathbf{\Gamma}$ be an arbitrary partition of $[m]$.*

Denote by $\Pi_{\mathbf{C}}$ as the set of all the rows permutations of $\mathbf{C}$ that result in nonzero diagonal entries, i.e.,

$$\Pi_{\mathbf{C}} \coloneqq \{\pi : \mathbf{P}_\pi \mathbf{C} \textit{ has all the nonzero diagonal entries.}\},$$

Denote by $\Pi_{\mathbf{C};\mathbf{\Gamma}}$ as all the rows permutations that result in invertible diagonal blocks on general scaled $\mathbf{C}$, i.e.,

$$\Pi_{\mathbf{C};\mathbf{\Gamma}} \coloneqq \{\pi : \exists \mathbf{D}_{\mathbf{\Gamma}}, \forall \mathbf{S} \in \mathbf{\Gamma}, \mathrm{rank}((\mathbf{P}_\pi \mathbf{D}_{\mathbf{\Gamma}} \mathbf{C})_{\pi[\mathbf{S}],\pi[\mathbf{S}]}) = |\mathbf{S}|\},$$

where $\mathbf{D}_{\mathbf{\Gamma}}$ is any general scaling matrix (defined in Section 3.1) consistent with $\mathbf{\Gamma}$.

For any two permutations π and τ of $[m]$, we say they are groupwise equivalent regarding a partition $\mathbf{\Gamma}$ of $[m]$, denoted by $\pi \sim_{\mathbf{\Gamma}} \tau$, if and only if $\forall \mathbf{S} \in \mathbf{\Gamma}$, $\pi[\mathbf{S}]$ and $\tau[\mathbf{S}]$ have exactly the same elements. Then, $\Pi_{\mathbf{C}}$ and $\Pi_{\mathbf{C};\mathbf{\Gamma}}$ are equivalent up to groupwise permutations, i.e.,

1. *$\forall \tau \in \Pi_{\mathbf{C};\mathbf{\Gamma}},\ \exists \pi \in \Pi_{\mathbf{C}},\ s.t.\ \pi \sim_{\mathbf{\Gamma}} \tau$;*
2. *$\forall \pi \in \Pi_{\mathbf{C}}$, if $\forall \mathbf{S} \in \mathbf{\Gamma}$, $(\mathbf{P}_\pi \mathbf{C})_{\pi[\mathbf{S}],\pi[\mathbf{S}]}$ is invertible, then $\exists \tau \in \Pi_{\mathbf{C};\mathbf{\Gamma}},\ s.t.\ \pi \sim_{\mathbf{\Gamma}} \tau$.*

Lemma 5 tells that all permutations that can result in nonzero diagonal entries on an invertible matrix are groupwise equivalent to all permutations that in result in invertible diagonal blocks on the same matrix corresponding to a given partition of the row indices. Note that 1. is universally true, while 2. needs an additional mild assumption that the partition and nonzero-diagonal permutation itself result in invertible diagonal blocks. Consider a counterexample: $\mathbf{C} = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 1 & 3 \\ 1 & 0 & 1 \end{bmatrix}$ is invertible, and a partition of row indices $\mathbf{\Gamma} = \{(1,2),(3,)\}$. Clearly the identity π (i.e., $\mathbf{P}_\pi = \mathbf{I}$) is in $\Pi_{\mathbf{C}}$, with $\mathbf{C}$ already having nonzero diagonal entries. However, we cannot find any $\tau \in \Pi_{\mathbf{C};\mathbf{\Gamma}}$ with $\pi \sim_{\mathbf{\Gamma}} \tau$:

$$\mathbf{D}_{\mathbf{\Gamma}}\mathbf{C} = \begin{bmatrix} a & b & 0 \\ c & d & 0 \\ 0 & 0 & e \end{bmatrix} \begin{bmatrix} 1 & 1 & 2 \\ 1 & 1 & 3 \\ 1 & 0 & 1 \end{bmatrix} = \begin{bmatrix} a+b & a+b & 2a+3b \\ c+d & c+d & 2c+3d \\ e & 0 & e \end{bmatrix},$$

either $\tau = (1,2,3)$ or $(2,1,3)$ is not in $\Pi_{\mathbf{C};\mathbf{\Gamma}}$, because $\begin{bmatrix} a+b & a+b \\ c+d & c+d \end{bmatrix}$ is already not invertible itself. Therefore, we make an additional assumption for 2., which is, as we can see later, trivially satisfied for our choice of invertible $\mathbf{C}$.

Now we prove the correctness of Lemma 5:

Proof. First, $\Pi_{\mathbf{C}}$ is nonempty, because $\mathbf{C}$ is invertible, $\det(\mathbf{C}) = \sum_\pi \mathrm{sgn}(\pi) \Pi_{i=1}^m \mathbf{C}_{i,\pi_i} \neq 0$, then at least there is one π s.t. $\forall i = 1, \cdots, m$, $\mathbf{C}_{i,\pi_i} \neq 0$, and so the inverse of this π will do a rows permutation with nonzero diagonals. The nonemptiness of $\Pi_{\mathbf{C};\mathbf{\Gamma}}$ can be proved using a similar idea (i.e., $\forall \mathbf{\Gamma}, \forall \mathbf{C}, \exists \tau$ s.t. $\mathbf{P}_\tau \mathbf{D}_{\mathbf{\Gamma}} \mathbf{C}$ has invertible diagonal blocks), but using group decomposition of the determinant expression, called "Generalized Laplacian expansion" (Janjic, 2008).

To prove 1., for any $\tau \in \Pi_{\mathbf{C};\mathbf{\Gamma}}$, initialize a new empty π. For any group $\mathbf{S} \in \mathbf{\Gamma}$, by definition, the block $(\mathbf{P}_\tau \mathbf{D}_{\mathbf{\Gamma}} \mathbf{C})_{\tau[\mathbf{S}],\tau[\mathbf{S}]}$ is invertible. Note that $(\mathbf{P}_\tau \mathbf{D}_{\mathbf{\Gamma}} \mathbf{C})_{\tau[\mathbf{S}],\tau[\mathbf{S}]} = (\mathbf{P}_\tau)_{\tau[\mathbf{S}],\mathbf{S}} \cdot (\mathbf{D}_{\mathbf{\Gamma}})_{\mathbf{S},\mathbf{S}} \cdot \mathbf{C}_{\mathbf{S},\tau[\mathbf{S}]}$, so $\mathbf{C}_{\mathbf{S},\tau[\mathbf{S}]}$ must also be invertible. By above nonemptiness, we know that $\mathbf{C}_{\mathbf{S},\tau[\mathbf{S}]}$ can be row permutated to one with nonzero diagonal entries. Then, we set the corresponding indices as this permutation, i.e., set $(\mathbf{P}_\pi)_{\tau[\mathbf{S}],\mathbf{S}}$ as this permutation submatrix. Then $\mathbf{P}_\pi\mathbf{C}$ has nonzero diagonals, and for any $\mathbf{S}$, $\pi[\mathbf{S}]$ and $\tau[\mathbf{S}]$ have the same elements (row indices).

To prove 2., it sufficies to show that for any $\pi \in \Pi_\mathbf{C}$, there is also $\pi \in \Pi_{\mathbf{C};\mathbf{\Gamma}}$. For any $\mathbf{S} \in \mathbf{\Gamma}$, consider the principal submatrix $(\mathbf{P}_\pi\mathbf{C})_{\pi[\mathbf{S}],\pi[\mathbf{S}]}$, which is assumed to be invertible. Note that $(\mathbf{P}_\pi\mathbf{C})_{\pi[\mathbf{S}],\pi[\mathbf{S}]} = (\mathbf{P}_\pi)_{\pi[\mathbf{S}],\mathbf{S}} \cdot \mathbf{C}_{\mathbf{S},\pi[\mathbf{S}]}$, and so $\mathbf{C}_{\mathbf{S},\pi[\mathbf{S}]}$ is invertible. For any $\mathbf{D}_\mathbf{\Gamma}$, we have $(\mathbf{P}_\pi\mathbf{D}_\mathbf{\Gamma}\mathbf{C})_{\pi[\mathbf{S}],\pi[\mathbf{S}]} = (\mathbf{P}_\pi)_{\pi[\mathbf{S}],\mathbf{S}} \cdot (\mathbf{D}_\mathbf{\Gamma})_{\mathbf{S},\mathbf{S}} \cdot \mathbf{C}_{\mathbf{S},\pi[\mathbf{S}]}$, where each factor is invertible, so $(\mathbf{P}_\pi\mathbf{D}_\mathbf{\Gamma}\mathbf{C})_{\pi[\mathbf{S}],\pi[\mathbf{S}]}$ is also invertible. Then, $\pi \sim_\mathbf{\Gamma} \pi$ trivially. □

By setting the invertible matrix $\mathbf{C}$ as the ISA characterization $\mathbf{A}^{-1}_{\mathbf{S},\mathbf{S}}$ (Section 3.2), we know that the admissible post-processing of rows permutation on any general ISA solution matrices $\mathbf{W}$ will make all subspaces at the correct location. The additional assumption for 2. that diagonal blocks of $\mathbf{C}$ are invertible echoes the 'weak stability' assumption mentioned in (Hyttinen et al., 2012). For correctness, specifically, the 1-dim subspaces (independent components), including the T and T's children that we are interested in, can be identified at the correct location (for each LiNG model in the equivalence class $\mathcal{B}$). The last step left is to associate the $\mathbf{A}^{-1}_{\mathbf{S},\mathbf{S}}$ to the local adjacencies $\mathbf{B}_{\mathbf{S},\mathbf{S}}$. By Equation (3), though in general $\mathbf{A}^{-1}_{\mathbf{S},\mathbf{S}} \neq \mathbf{B}_{\mathbf{S},\mathbf{S}}$, they must be equal on the rows of T and T's children, as their parents are all included in $\mathbf{S}$. Finally, the correctness of Algorithm 1 is proved.

D.5 Proof of Corollary 1

Corollary 1 (Identifying stable solutions locally, with disjoint cycles)**.** *Suppose the cycles are disjoint in $\mathcal{G}$. Consider a modified version of Algorithm 1 in which the line "**if** $\mathbf{B}'$ is not convergent: **skip**" is added between lines 8 and 9. Then, this modified version of Algorithm 1 will yield a single local model, corresponding exactly to the unique global stable model. That is, the returned $\mathcal{K}$ consists of a single item $\mathbf{K}$, with*

$$\mathbf{K} = \{(i \to j, \mathbf{B}^*_{j,i}) : \forall j \in \{T\} \cup \mathrm{ch}_{\mathcal{G}^*}(T), \forall i \in \mathrm{pa}_{\mathcal{G}^*}(j)\}.$$

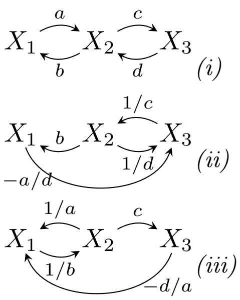
Proof. By the indeterminacy of ISA (Theorem 1), the LiNG's ISA characterization (Theorem 2), and the characterization of the LiNG global equivalence class (Definition 1), we know that for each admissible $\mathbf{B}'$ at the line 8 of Algorithm 1, there exists a ground-truth model $\mathbf{B} \in \mathcal{B}$ with the corresponding permutation, s.t. $\mathbf{B}' = \mathbf{I} - \mathbf{D}(((\mathbf{I}-\mathbf{B})^{-1})_{\mathbf{S},\mathbf{S}})^{-1}$, where $\mathbf{S} = \{T\} \cup \mathrm{mb}_\mathcal{G}(T)$, and $\mathbf{D}$ is the general scaling matrix for diagonal ones.

When the cycles in $\mathcal{G}$ are disjoint, all graphs in the LiNG equivalence class have disjoint cycles, and the unique global stable model $\mathbf{B}^*$ is the one where all cycles' products have absolute values less than one. As here stability is determined merely by the cycle products, to prove Corollary 1, we only need to show the following statement:

Consider a LiNG equivalence class $\mathcal{B}$ where cycles are disjoint. For any adjacency matrix $\mathbf{B} \in \mathcal{B}$ and its corresponding graph $\mathcal{G}$ and mixing matrix $\mathbf{A} \coloneqq (\mathbf{I}-\mathbf{B})^{-1}$, for any $\mathbf{S} \subset \mathcal{V}$ (not necessarily a vertex and its Markov blanket), we initialize a local graph termed $\mathcal{G}^{(\mathbf{S})}$ over $\mathbf{S}$ from the local adjacency matrix termed $\mathbf{B}^{(\mathbf{S})} \coloneqq \mathbf{I} - \mathbf{A}^{-1}_{\mathbf{S},\mathbf{S}}$. From Example 3 we know that $\mathbf{A}^{-1}_{\mathbf{S},\mathbf{S}}$ does not necessarily have all diagonals as ones, so we further remove all self-loops on $\mathcal{G}^{(\mathbf{S})}$. Then, all cycles in this local $\mathcal{G}^{(\mathbf{S})}$, if any, must be disjoint. Further, for each local cycle in $\mathcal{G}^{(\mathbf{S})}$ consisting of vertices $\mathbf{C}^{(\mathbf{S})} \subset \mathbf{S}$, we have the followings:

1. All vertices on the local cycle need no row scalings, i.e., $\forall i \in \mathbf{C}^{(\mathbf{S})}$, $(\mathbf{A}^{-1}_{\mathbf{S},\mathbf{S}})_{i,j} = 1$, and
2. There exists a global cycle in $\mathcal{G}$ consisting of vertices $\mathbf{C}$ with $\mathbf{C}^{(\mathbf{S})} \subset \mathbf{C}$ (in a consistent ordering), and,
3. The cycle product of this local cycle $\mathbf{C}^{(\mathbf{S})}$ on $\mathcal{G}^{(\mathbf{S})}$ equals the cycle product of the global cycle $\mathbf{C}$ on $\mathcal{G}$.

The above statement follows from Equation (3): the (i,j)-th entry of $\mathbf{I} - \mathbf{A}^{-1}_{\mathbf{S},\mathbf{S}}$ corresponds not only to the direct causal effect from j to i, but also the total causal effect from j to i through all other variables outside of $\mathbf{S}$. For the above point 1., a vertex i has non-unit diagonal $\mathbf{A}^{-1}_{\mathbf{S},\mathbf{S}} \neq 1$ only when i is involved in a cycle where all the remaining vertices in this cycle is not in $\mathbf{S}$ (see Example 3), so that this cycle appear as a self-loop on i relative to $\mathbf{S}$. As cycles are disjoint, i cannot belong to any cycle in $\mathcal{G}^{(\mathbf{S})}$. Point 1 shows that whenever a cycle can appear locally, the edge weights on this cycle must follow exactly from $\mathbf{A}^{-1}_{\mathbf{S},\mathbf{S}}$, without any scalings. Then, using the characterization in Equation (3), points 2 and 3 show that the cycle products can be preserved locally. A local stable model (with disjoint cycles and abs(cycle products)$<$ 1) must correspond to a global model with those stable cycles, which, in Algorithm 1's case, implies the correct local stable model among $\{T\} \cup \mathrm{ch}_{\mathcal{G}^*}(T)$. □



Remark. However, note that the above proof relies on the assumption that the cycles in $\mathcal{G}$ are disjoint. However, when some cycles in $\mathcal{G}$ intersect, things become more complex. The figure to the right shows an example of three cyclic graphs in a LiNG equivalence class with intersected cycles. Globally, there may be none (e.g., when $a = b = c = d = 0.8$) or multiple (e.g., both *(i)* and *(ii)* when $a = b = c = 0.8$, $d = -2$) global stable models. In this case, while our method can still identify local correspondings of all equivalent models (Theorem 3), and some unstable solutions may be partially eliminated (using the local stability constraint), we show that the exact identification of the global stable solutions from local variables alone becomes inherently impossible, because intuitively, external cycles appear as self-loops on the local variables. Note that when cycles intersect, the simple cycles' products cannot be related to the stability directly anymore: a LiNG model can be stable with some cycle products larger than one, and a LiNG model with all abs(cycle products) less than one can also be unstable. Even if we force to use cycle products to define stability (as some papers (Rothenhäusler et al., 2015) do), the abovementioned unidentifiability issue of global stable solutions from local variables still remains.

D.6 Proof of Lemma 3

Lemma 3. *For any $i \in \mathcal{V}, \mathbf{S} \subset \mathcal{V}\backslash\{i\}$, if $R_{\mathbf{S}\to i} \perp\!\!\!\perp \mathbf{X_S}$, i.e., independent residual, then $\mathbf{S} \cap \operatorname{des}_{\mathcal{G}}(i) = \emptyset$.*

The proof can be referred to (Shimizu et al., 2011), by using Darmois-Skitovitch theorem (Darmois, 1953; Skitovitch, 1953).

D.7 Proof of Lemma 4

Lemma 4. *$\forall i, \mathbf{S}$ in $\mathcal{V}$, if $\operatorname{pa}_{\mathcal{G}}(i) \subset \mathbf{S} \subset \mathcal{V}\backslash \operatorname{des}_{\mathcal{G}}(i)$, then $\forall j \in \mathbf{S}$, $\beta^j_{\mathbf{S}\to i} = \mathbf{B}_{i,j}$, and $R_{\mathbf{S}\to i} = E_i$ (so $\perp\!\!\!\perp \mathbf{X_S}$).*

The proof follows naturally from Lemma 1 (in the acyclic graph case).

D.8 Proof of Theorem 4

Theorem 4 (Correctness of Algorithm 2)**.** *For any $T \in \mathcal{V}$, let $\mathbf{K}$ be the weighted edge set returned by Algorithm 2 on T, $\operatorname{mb}_{\mathcal{G}}(T)$, and $\mathbf{X}$. We have:*

$$\mathbf{K} = \{(i \to j, \mathbf{B}_{j,i}) : \forall j \in \{T\} \cup \operatorname{ch}_{\mathcal{G}}(T), \forall i \in \operatorname{pa}_{\mathcal{G}}(j)\}.$$

Proof. At every iteration of Algorithm 2, consider a 'last' remaining vertex $j \in \mathbf{U}$ s.t. there exists no other $j' \in \mathbf{U}$ as j's descendant on $\mathcal{G}$. 1° if j is T or T's child, since $\operatorname{pa}_{\mathcal{G}}(j) \subset \operatorname{mb}_{\mathcal{G}}(T)$, and since none of $\operatorname{pa}_{\mathcal{G}}(j)$ can be removed earlier, we have $\operatorname{pa}_{\mathcal{G}}(j) \subset \mathbf{U}$. With all the parents in $\mathbf{U}$ and no descendants in $\mathbf{U}$, regress j on $\mathbf{U}\backslash\{j\}$ will produce independent residual, and the nonzero coefficients correspond to the true direct parents with true weights, i.e., $\beta^i_{\mathbf{U}\backslash\{j\}\to j} = \mathbf{B}_{j,i}$; 2° if j is T's spouse, since j is 'last', none of j and T's common children and descendants are in $\mathbf{U}$. Also since $\operatorname{pa}_{\mathcal{G}}(T) \subset \mathbf{U}$, every confounding path between j and T is blocked, and thus $\beta^T_{\mathbf{U}\backslash\{j\}\to j} = 0$; 3° it is impossible for j to be T's parents, since when T pops out, the program breaks. □

D.9 Proof of Proposition 1

We first state the following lemmas from Ng et al. (2021, Lemmas 1 & 2) (which were based on Loh and Bühlmann (2014)). The original lemmas in Ng et al. (2021) focus on linear acyclic SEM, but their proofs do not make use of the acyclicity constraint. Therefore, we restate the lemmas here for cyclic graphs, whose proofs are similar to the acyclic case and omitted.

Lemma 6. *Suppose $\mathbf{X}$ follows the linear SEM in Equation (1) with directed cyclic graph $\mathcal{G}$ and inverse covariance* *matrix* $\mathbf{\Theta}$*. The entries of* Θ *are given by*

$$\begin{aligned}\mathbf{\Theta}_{j,i} &= -\sigma_j^{-2}\mathbf{B}_{i,j} - \sigma_i^{-2}\mathbf{B}_{j,i} + \sum_{\ell\neq j,i}\sigma_\ell^{-2}\mathbf{B}_{j,\ell}\mathbf{B}_{i,\ell}, \qquad &\forall j \neq k,\\ \mathbf{\Theta}_{j,j} &= \sigma_j^{-2} + \sum_{\ell\neq j}\sigma_\ell^{-2}\mathbf{B}_{j,\ell}^2, &\forall j.\end{aligned}$$

Lemma 7. *Suppose* $\mathbf{X}$ *follows the linear SEM in Equation* (1) *with directed cyclic graph* $\mathcal{G}$ *and inverse covariance matrix* $\mathbf{\Theta}$*. Then, the structure defined by the support of* $\mathbf{\Theta}$ *is a subgraph of the moral graph of* $\mathcal{G}$*.*

We then provide the proof for the following proposition.

Proposition 1. *Suppose* $\mathbf{X}$ *follows the linear SEM in Equation* (1) *with directed cyclic graph* $\mathcal{G}$ *and inverse covariance matrix* $\mathbf{\Theta}$*. Under Assumption 1, the structure defined by the support of* $\mathbf{\Theta}$ *is the same as the moral graph of* $\mathcal{G}$*.*

Proof. By Lemma 7, the structure defined by the support of $\mathbf{\Theta}$ is a subgraph of the moral graph of $\mathcal{G}$. By Assumption 1 and Lemma 6, if $\mathbf{\Theta}_{j,i} = 0$, then we have $\mathbf{B}_{i,j} = \mathbf{B}_{j,i} = 0$ and $\mathbf{B}_{j,\ell}\mathbf{B}_{i,\ell} = 0$ for all $\ell \neq j, i$, which, by definition, indicates that i and j are not adjacent in the moral graph of $\mathcal{G}$. This indicates that the moral graph of $\mathcal{G}$ is a subgraph of the structure defined by the support of $\mathbf{\Theta}$. □

E SUPPLEMENTARY EXPERIMENTS DETAILS

We provide additional details for the experiments conducted in Section 5. Specifically, we provide the implementation details of our method and the baselines in Appendices E.1 and E.2, respectively. We then describe in Appendix E.3 the performance metrics used in our experiments.

E.1 Implementation Details of Local ISA-LiNG

The proposed Local ISA-LiNG method involves estimating the demixing matrix with ISA (see Algorithm 1). One could use the ISA procedure developed by Theis (2006). In this work, we use ICA to first estimate the components, and then use independence test, i.e. the Hilbert-Schmidt independence criterion (HSIC) (Gretton et al., 2007), to identify the subspaces from the components estimated by ICA. Such a procedure is found to work well in practice. For the HSIC test in our experiments, we use a significance level of 0.05. Furthermore, as suggested by Lacerda et al. (2008), we adopt ICA with sparse connection (Zhang and Chan, 2006) for the specific ICA procedure. In the acyclic case , we use a more efficient post-processing procedure to identify the edges from the demixing matrix $\mathbf{W}$, described in Appendix C.

We perform nodewise regression with Lasso (Meinshausen and Bühlmann, 2006) to estimate the MB of each target variable (see Appendix A for more details). Moreover, when applying Algorithms 1 and 3 to identify the edges from the estimated demixing matrix, we use a threshold of 0.05 to set the entries with small absolute values to zero. All experiments are conducted on 2 CPUs with 4GB of memory. The code is available at `https://github.com/MarkDana/local-ling-discovery`.

E.2 Implementation Details of Existing Methods

We provide the implementation details of several existing methods considered in our experiments. All experiments are conducted on 2 CPUs with 4GB of memory. For Local A* and GSBN, we perform nodewise regression with Lasso (Meinshausen and Bühlmann, 2006) to estimate the MB of each target variable, similar to our Local ISA-LiNG method. For CMB and LDECC, we use their default method for MB estimation. In the following, we describe the implementation details of each method.

ICA-LiNG. As suggested by Lacerda et al. (2008), we use ICA with sparse connection (Zhang and Chan, 2006) for the specific ICA procedure, and a threshold of 0.05 for the demixing matrix, similar to our method.

Local A*. The original Local A* method (Ng et al., 2021) applies exact search strategy like A* (Yuan and Malone, 2013) on the target T, its MB $\mathrm{mb}_{\mathcal{G}}(T)$, and MB of each variable in $\mathrm{mb}_{\mathcal{G}}(T)$. In the estimated structure, the method then identifies all (1) undirected edges involving T and (2) v-structures involving T to be the final local structure around T. In our modification, we apply A* on only T and its MB $\mathrm{mb}_{\mathcal{G}}(T)$. In this case, the modified method identifies all (1) undirected edges involving T and (2) v-structures of which T is not the collider to be the final estimated local structure around T. The resulting algorithm performs local discovery on only T and its MB $\mathrm{mb}_{\mathcal{G}}(T)$. Here, we use BIC score (Schwarz, 1978) for the A* search.

GSBN. The original GSBN method (Margaritis and Thrun, 1999) applies certain rules based on conditional independence tests to identify all (1) undirected edges involving T and (2) v-structures of which T is the collider. The latter requires information of two-step MBs, i.e., $\mathrm{mb}_{\mathcal{G}}(T)$ and MB of each variable in $\mathrm{mb}_{\mathcal{G}}(T)$. In our modification, we adopt different but similar rules to identify (1) undirected edges involving T and (2) v-structures of which T is not the collider:

1. For each $X \in \mathrm{mb}_{\mathcal{G}}(T)$, determine X to be a (direct) neighbor of T if $T \not\perp\!\!\!\perp X | S$ for all $S \subseteq \mathrm{mb}_{\mathcal{G}}(T) \setminus \{Y\}$.

2. Given the neighbors of target T identified in the first step, we use the following rule to identify the v-structures: for each $Z \in \mathrm{mb}_{\mathcal{G}}(T)$ and Y being a neighbor of T, determine $T \rightarrow Y \leftarrow Z$ to be a v-structure if $T \not\perp\!\!\!\perp Z | S \cup \{Y\}$ for all $S \subseteq \mathrm{mb}_{\mathcal{G}}(T) \setminus \{Y, Z\}$.

The resulting algorithm performs local discovery on only T and its MB $\mathrm{mb}_{\mathcal{G}}(T)$. Here, we use Fisher Z test with significance level of 0.05 for identifying conditional independence relations.

CMB. For the CMB method (Gao and Ji, 2015), we use an implementation through the pyCausalFS package.

LDECC. We use the default implementation provided by Gupta et al. (2023).

E.3 Performance Metrics

In the acyclic case, we report two performance matrics, namely SHD of local DAG and PDAG. In the cyclic case, we report the SHD of local DCG. These metrics are explained in details below. For each setting, the metrics are calculated over 8 random simulations.

SHD of local DAG. For this metric, the ground-truth and estimated structures contain all incoming directed edges of T and its children. We then compute the SHD between the ground-truth and estimated local structures. Such a metric is used to validate our method (see Theorem 3) that can estimate more edges and directions than the baselines. Note that the estimated output by Local A* and GSBN may contain undirected edges; therefore, we enumerate all possible combinations of directed edges from these undirected ones, and compute the final averaged SHD for these two methods.

SHD of local PDAG. Since Local A* and GSBN return PDAG around the target T, we design this metric specifically for these baselines. In particular, the ground-truth and estimated local structures contain (1) undirected edges involving T and (2) v-structures of which T is not the collider. We then compute the SHD between the ground-truth and estimated local structures. Note that since our method returns DAG around the target T that contains additional edges, we convert it into the same format of PDAG as well.

SHD of local DCG. This metric is similar to the SHD of local DAG explained above, except that the ground truth may contain cycles.

F SUPPLEMENTARY EXPERIMENTS RESULTS

F.1 Running Time

We report the running times for different sizes of MBs. Specifically, we follow the data generating procedure in Section 5.2. We generate random DAGs with expected degrees of 3, 5, and 7, and select target with a relatively large MB. The running times of different methods, including those with and without oracle MBs, are shown in Figure 9. Note that, for Local A*, instances with MBs exceeding 19 variables are omitted due to the long running time. It is observed that our method has a longer running time than that of LDECC and CMB. Furthermore, when size of MB increases, the running time of our method is shorter and increases much slowly compared to GSBN and Local A*.

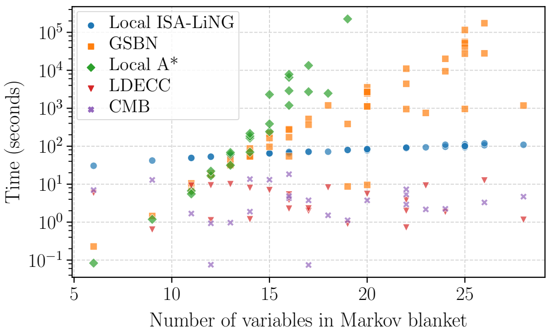
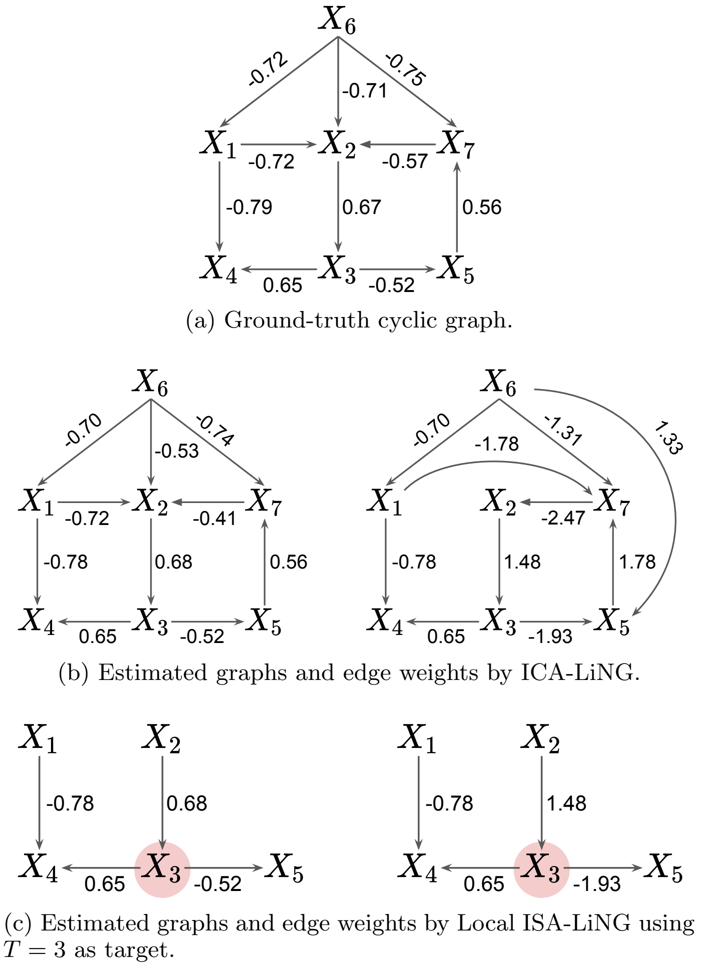


Figure 9: Running time of different methods. Y-axis is in log scale.

F.2 Additional Figures

This section provides additional figures for Section 5, namely Figures 10, 11, 12, 13, 14, 15, 16, 17, and 18.



(a) Ground-truth cyclic graph.

(b) Estimated graphs and edge weights by ICA-LiNG.

(c) Estimated graphs and edge weights by Local ISA-LiNG using $T = 3$ as target.

Figure 10: Ground truth and estimated cyclic graphs and edge weights with 2000 samples.

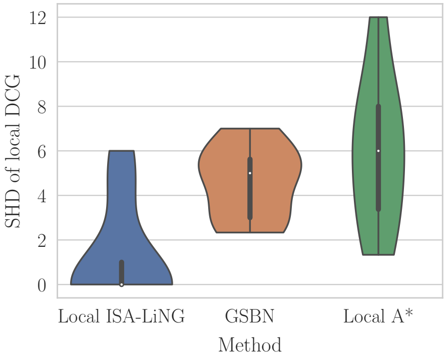
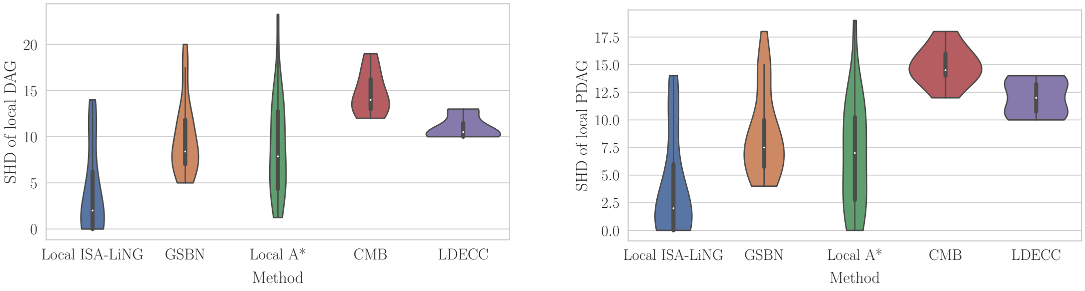


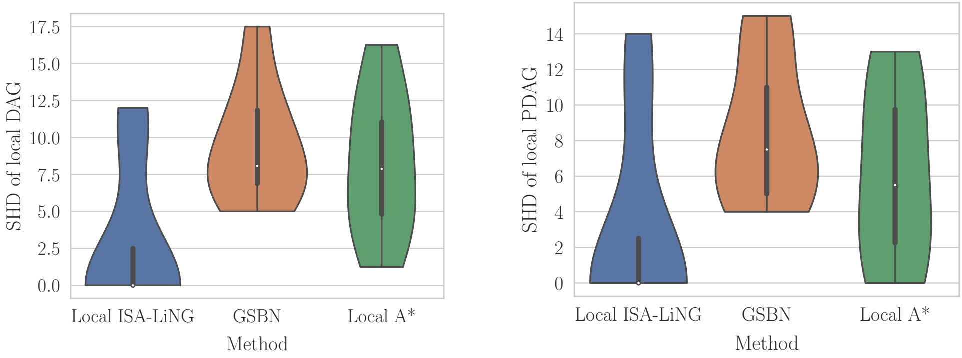
Figure 11: SHD of local DCG under oracle MB.



(a) SHD of local DAG.

(b) SHD of local PDAG.

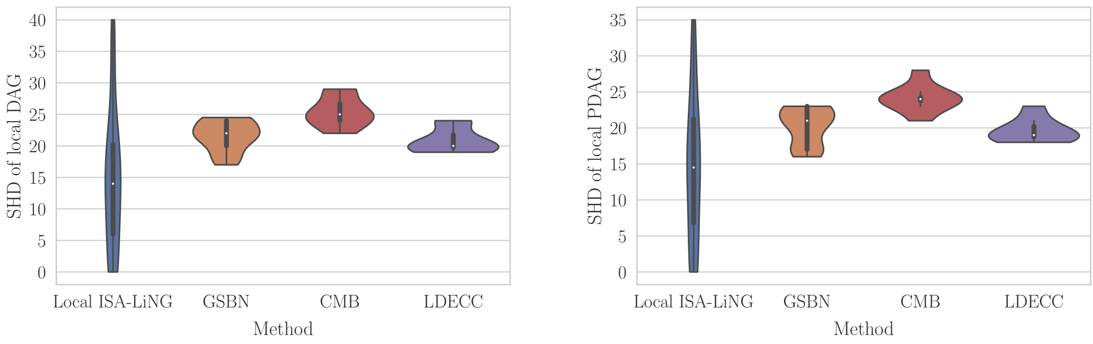
Figure 12: Results of local causal discovery with 2000 samples and degree of 3 under estimated MB.



(a) SHD of local DAG.

(b) SHD of local PDAG.

Figure 13: Results of local causal discovery with 2000 samples and degree of 3 under oracle MB.



(a) SHD of local DAG.

(b) SHD of local PDAG.

Figure 14: Results of local causal discovery with 2000 samples and degree of 5 under estimated MB.

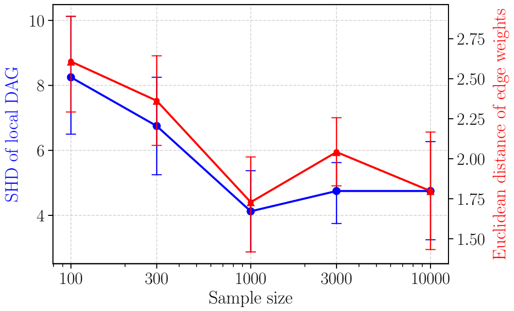
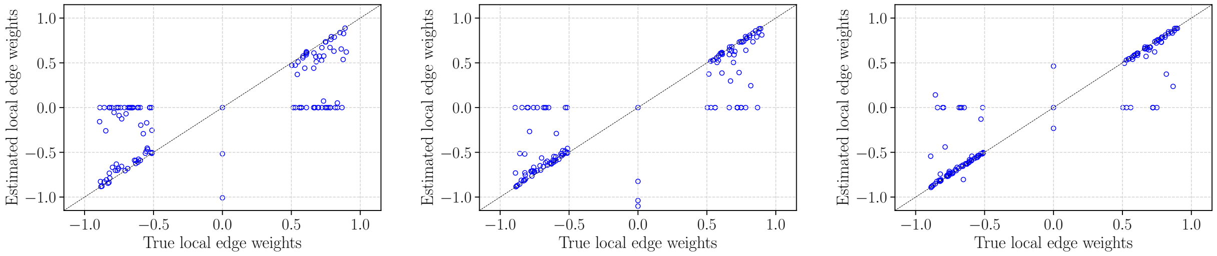


Figure 15: Results of Local ISA-LiNG with MB estimated by Lasso. X-axis is visualized in log scale.

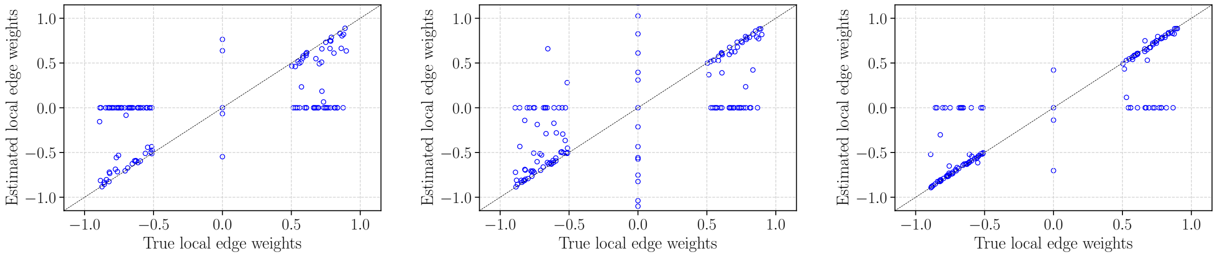


(a) 100 samples.

(b) 300 samples.

(c) 1000 samples.

Figure 16: Edge weights estimated by Local ISA-LiNG under oracle MB.



(a) 100 samples.

(b) 300 samples.

(c) 1000 samples.

Figure 17: Edge weights estimated by Local ISA-LiNG under estimated MB.

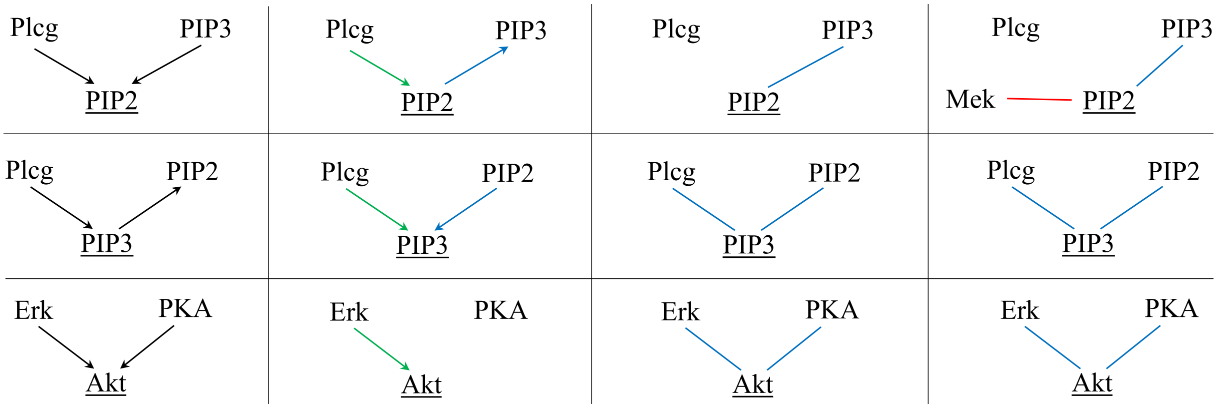


Figure 18: Result of local causal discovery on real-world dataset by Sachs et al. (2005). The first column showcases the ground-truth local causal graphs. From the second to the last column, each column corresponds to the local causal graphs recovered by (1) Local ISA-LiNG, (2) Local A*, and (3) GSBN, respectively. We use underlined vertices to denote target variables. For the second column, green and blue arrows denote correct directed edges and wrong directed edges discovered by our method, respectively. For the third and fourth columns, blue lines denote correct undirected edges discovered by the baselines.